drones

MDPI

Article

Energy-Efficient Device-to-Device Communications for Green Internet of Things Using Unmanned Aerial Vehicle-Mounted Intelligent Reflecting Surface

Fangqing Tan [1], Shuo Pang [1], Yashuai Cao [2,*], Hongbin Chen [1] and Tiejun Lv [3]

[1] Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing, Guilin University of Electronic Technology, Guilin 541004, China; tfqing@guet.edu.cn (F.T.); chbscut@guet.edu.cn (H.C.)
[2] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[3] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
* Correspondence: caoys@tsinghua.edu.cn

Abstract: The Internet of Things (IoT) serves as a crucial element in interconnecting diverse devices within the realm of smart technology. However, the energy consumption of IoT technology has become a notable challenge and an area of interest for researchers. With the aim of achieving an IoT with low power consumption, green IoT has been introduced. The use of unmanned aerial vehicles (UAVs) represents a highly innovative approach for creating a sustainable green IoT network. UAVs offer advantages in terms of flexibility, mobility, and cost. Moreover, device-to-device (D2D) communication is essential in emergency communications, due to its ability to support direct communication between devices. The intelligent reflecting surface (IRS) is also a hopeful technology which reconstructs the radio propagation environment and provides a possible solution to reduce co-channel interference resulting from spectrum sharing for D2D communications. The investigation in this paper hence focuses on energy-efficient UAV-IRS-assisted D2D communications for green IoT. In particular, a problem of optimization aimed at maximizing the system's average energy efficiency (EE) is formulated, firstly, by simultaneously optimizing the power coefficients of all D2D transmitters, the UAV's trajectory, and the base station (BS)'s active beamforming, along with the IRS's phase shifts. Second, to address the problem, we develop a multi-agent twin delayed deep deterministic policy gradient (MATD3)-based scheme to find a near-optimal solution, where D2D transmitters, the BS, and the UAV cooperatively learn to improve EE and suppress the interference. To conclude, numerical simulations are conducted to assess the availability of the proposed scheme, and the simulation results demonstrate that the proposed scheme surpasses the baseline approaches in both convergence speed and EE performance.

Keywords: intelligent reflecting surface; device-to-device communication; unmanned aerial vehicle; multi-agent deep reinforcement learning

Citation: Tan, F.; Pang, S.; Cao, Y.; Chen, H.; Lv, T. Energy-Efficient Device-to-Device Communications for Green Internet of Things Using Unmanned Aerial Vehicle-Mounted Intelligent Reflecting Surface. *Drones* **2024**, *8*, 122. https://doi.org/10.3390/drones8040122

Academic Editors: Bo Rong and Michel Kadoch

Received: 15 February 2024
Revised: 14 March 2024
Accepted: 20 March 2024
Published: 26 March 2024

1. Introduction

The Internet of Things (IoT) as an important information network has been widely used in many areas of everyday life such as smart home, smart transportation, smart city, etc. [1,2]. Nevertheless, there are a vast quantity of IoT devices worldwide contributing to a significant increase in energy usage, placing additional strain on the global electric grid and exacerbating environmental changes. Efforts have been made to extensively explore technical solutions that promote efficient consumption in order to achieve the objective of green communications and enhance the utilization efficiency of transmission energy [3]. In recent years, with the increase in IoT devices, device-to-device (D2D) technology has been developed as an emerging technology to enhance spectrum efficiency (SE) and address the growing lack of spectrum resource scarcity [4]. The D2D technology can be widely

used in various scenarios, including IoT [5], emergency communications [6], and social networks [7]. Recent research has shown that D2D technology has remarkable advantages in improving SE and enhancing the user experience [8].

Green IoT, a novel paradigm emphasizing energy efficiency (EE) to address the environmental impact of IoT technologies for achieving sustainable and environmentally friendly development, is regarded as a prospective research avenue in the IoT domain [9]. However, as a crucial technology in IoT, the constrained onboard battery capacity hinders the realization of the full potential of D2D communication. Enhancing the system EE of D2D communication through effective resource management remains an ongoing research focus. Furthermore, in some scenarios with long distances or numerous obstacles, the communication between D2D devices can be severely affected. For instance, the authors of [10] focused on the unmanned aerial vehicle (UAV) location and resource allocation in UAV-assisted D2D communication for industrial IoT, and confirmed that the proposed decomposition-based algorithm can improve the system EE when compared with other benchmarks. The use of intelligent reflecting surfaces (IRSs) offers a promising option to address this issue, which can eliminate interference between D2D pairs and thus enhance EE by reconstructing the wireless communication environment [11–13]. Specifically, the IRS consists of multiple electromagnetic reflecting elements that are passive in nature. By altering the amplitude and phase of the incoming signal, it achieves a beamforming gain through reflection [14]. The deployment of IRS on urban building surfaces presents a significant improvement opportunity for guaranteeing quality of service in base station (BS) coverage blind spots. This improvement is achieved by establishing reflective line-of-sight (LoS) links. Furthermore, the IRS-assisted system offers several advantages over traditional relay systems, including low cost and low power consumption [15]. Moreover, D2D communications in cellular networks is subject to mutual interference from cellular users and other D2D devices. Using IRS to effectively regulate each reflection element's amplitude or phase shift coefficient, and resource allocation of the D2D communication network can effectively mitigate the interference [16]. Unlike traditional approaches like alternating optimization, deep reinforcement learning (DRL) does not require prior knowledge and exhibits lower complexity and better performance. Recently, some research has already applied it to IRS-assisted wireless communication systems. For example, in [17], the authors introduce the single-agent deep deterministic policy gradient (DDPG) to optimize the passive beamforming of IRS in MISO systems and the authors of [18] extend this method to IRS-assisted MIMO systems. In [19], a multi-agent reinforcement learning (MARL)-based scheme was proposed for the joint design of passive beamforming and power control. Furthermore, to address the optimization problem of the mixed action space in IRS-assisted D2D communication networks, the authors of [20] proposed a novel multi-agent multi-pass deep Q-networks algorithm using centralized training and a decentralized execution (CTDE) framework.

Additionally, considering the UAV's high mobility, it can act as a relay and be deployed at high altitudes to provide LoS links between ground terminals. The advantages of both IRS and UAV in improving communications and networks have been demonstrated [21]; many works have started to investigate the potentials of IRS mounted on a UAV, termed aerial UAV-IRS [22–26]. Because of the high mobility of the UAV, UAV-assisted communication systems often exhibited time-varying and dynamic characteristics, thereby makes the trajectory optimization and resource allocation for such systems intricate. The rapid advancement in machine learning has brought reinforcement learning to the forefront as a promising solution for tackling these challenges. Consequently, several studies have been dedicated to exploring DRL in UAV-IRS-assisted communication systems. The authors of [23] employed DDPG and double deep Q-learning network (DDQN) algorithms to tackle the challenge of trajectories and phase shift optimization in an IRS-assisted UAV network, and the numerical results demonstrated that DDPG-based continuous control achieves a better performance. The researchers in [24] introduced an innovative SWIPT method that involved optimizing resource allocation and a twin delayed DDPG (TD3)-based algorithm

was used to obtain the solution to the problem. The authors of [25] considered the outdated channel state information (CSI) and developed a novel twin-agent DDPG algorithm to optimize radio resource allocation. However, the schemes based on single-agent DRL left out of consideration the interaction between multiple communicating nodes, which leads to poor performance in cooperative tasks such as D2D communication and vehicle communication. In addition, single-agent DRL requires a centralized data center to collect the status information of all agents and carries out centralized training, which results in a significant communication overhead. In contrast, a localized observation-based MARL algorithm has been proposed for communication systems [26,27]. It was demonstrated in [28] that the MARL algorithm can achieve better performance and robustness in UAV-IRS-assisted systems compared to single-agent reinforcement learning.

Furthermore, there are few works that focus on the system EE maximization for UAV-IRS-assisted D2D networks in the downlink scenario. For example, Ref. [29] considered a multiple UAV-IRS-enhanced anti-jamming D2D communication network and maximized the achievable sum rate by optimizing the IRS mode selection and phase shift, where each UAV is equipped with an IRS as an aerial relay to forward signals from multiple D2D users. In [30], the authors considered a scenario in which a UAV was equipped with an active IRS-assisted terahertz band D2D communication; the maximum system sum rate was achieved through reasonable power control and beamforming. In [31], the author proposed a distributed deep learning algorithm for realizing power control and channel selection. However, the previous studies did not take into account the impact caused by the movement of the UAV. Additionally, in [32], the researchers investigated a D2D communication system in the uplink scenario that utilized UAV-IRS assistance; they further employed a DDQN-based algorithm that optimized both the UAV's flight trajectory and the IRS's phase shift. But, in practical applications, D2D communication is usually considered as a distributed cooperative scenario. Using traditional centralized algorithms requires real-time access to global information, inevitably resulting in a significant amount of communication overhead [33]. In general, there has not been sufficient investigation into the integration of UAV-IRSs into D2D communications for the downlink scenario.

Motivated by the potential advantages and features of UAV-IRS, this paper explores energy-efficient UAV-IRS-assisted D2D systems, where the IRS is mounted on a rotary-wing UAV to serve as an aerial relay. In this paper, we aim to maximize the system's average EE by jointly optimizing the considered system's resource management along with the movement of the UAV-IRS. In order to obtain a near-optimal solution, we come up with a multi-agent DRL (MADRL)-based approach. The primary contributions of this paper are the following:

- We investigate the downlink of D2D communications assisted by UAV-IRS, in which the BS, the UAV-IRS, and all D2D pairs collaborate to achieve improved EE performance. Specifically, the UAV carries the IRS to establish LoS links between various communication nodes. To maximizing the system EE over time in a changing environment, we formulate an optimization problem that involves optimizing the UAV's trajectory, the BS's active beamforming, the IRS's passive beamforming, and the D2D transmitters' power control.
- To solve the proposed EE maximization problem, we use a Markov Game to model the cooperative task considering each D2D pair, BS, and UAV-IRS. Consequently, the resource allocation and trajectory optimization problem is addressed using multi-agent twin delayed DDPG (MATD3) [34]. To enhance its learning efficiency, prioritized experience replay (PER) is employed. In addition, the algorithm's complexity is thoroughly examined.
- The availability of the proposed algorithm is validated through simulations, with numerical results demonstrating that the proposed scheme outperforms benchmark schemes in terms of convergence speed and EE performance.

The remaining part of this paper is organized as follows. The system model is presented in Section 2. The EE maximization optimization problem is formulated in Section 3.

In addition, Section 4 introduces the MATD3-based algorithm. Section 5 provides the numerical results of the simulation, while Section 6 concludes the paper.

2. System Model

Figure 1 depicts our considered UAV-IRS-assisted D2D communication underlying cellular network in the downlink scenario, including an UAV-IRS, a BS, L cellular users (CUs), and K D2D pairs. Specifically, a rotary-wing UAV carries a passive IRS as an aerial relay to forward communication between D2D devices. In addition, we denote the sets of all CUs and D2D pairs as $\mathcal{L} = \{1, \ldots, l, \ldots, L\}$ and $\mathcal{K} = \{1, \ldots, k, \ldots, K\}$, respectively. Each D2D communication pair includes a transmitter with only one antenna and a receiver with only one antenna, and it is assumed that all D2D communication pairs and CUs share the same spectrum resources. Let D_k^t and D_k^r represent the k-th D2D transmitter and the k-th receiver, respectively. The UAV-IRS has a uniform rectangular array (URA) comprising $M = M_y \times M_z$ reflection elements, while the BS is equipped with a uniform linear array (ULA) consisting of N_t antennas.

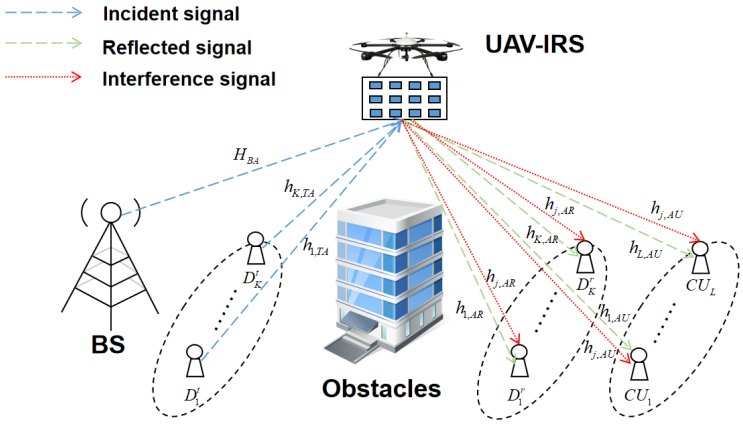


Figure 1. UAV-IRS-assisted D2D communication network.

Let $\boldsymbol{u}_{k,t} = [x_{k,t}, y_{k,t}, 0]^T$ and $\boldsymbol{u}_{k,r} = [x_{k,r}, y_{k,r}, 0]^T$ represent the coordinates of D_k^t and D_k^r, respectively. The BS is located at the coordinates $\boldsymbol{u}_B = [x_B, y_B, 0]^T$. Furthermore, we make the assumption that the rotary-wing UAV maintains a constant altitude characterized by H_u. The total time period T is partitioned into N equal time slots, represented by δ_n for each slot. Therefore, let $\boldsymbol{q}[n] = [x_u[n], y_u[n], H_u]^T$ represent the UAV location information at the n-th time slot. The movement of the UAV must comply with the following restrictions:

$$\|\boldsymbol{q}[n+1] - \boldsymbol{q}[n]\| \leq v_{max}\delta_n, n = 0, \cdots, N, \tag{1a}$$

$$x_{min} \leq x_u[n] \leq x_{max}, y_{min} \leq y_u[n] \leq y_{max}, n = 0, \cdots, N, \tag{1b}$$

$$\boldsymbol{q}[0] = [0, 0, H_u]^T, \tag{1c}$$

where (1a) stands for the mobility constraints of the UAV including start point and end point, v_{max} represents the drone's maximum flying velocity, (1b) sets the constraint for the flying range of the UAV, and (1c) specifies the initial location.

In this paper, it is assumed that the presence of obstacles results in the absence of direct links between any two nodes. The channel coefficients from transmitter D_k^t to the UAV-IRS, from the UAV-IRS to receiver D_k^r, from the BS to the UAV-IRS, and from the UAV-IRS to the CU l are denoted by $\boldsymbol{h}_{k,TA}[n] \in \mathbb{C}^{M\times 1}$, $\boldsymbol{h}_{k,AR}[n] \in \mathbb{C}^{M\times 1}$, $\boldsymbol{H}_{BA}[n] \in \mathbb{C}^{M\times N_t}$, and $\boldsymbol{h}_{l,AU}[n] \in \mathbb{C}^{M\times 1}$, respectively. It is presumed that the channels between any ground devices and the UAV-IRS are regarded as LoS links.

Considering that the UAV-IRS-D2D link and the UAV-IRS-CU link may be blocked by obstacles, the path loss between the UAV-IRS and other communication nodes can be modeled as a probabilistic LoS path model [35]. It can be described as

$$P_{LoS}(\theta) = \frac{1}{1 + a_1 \exp(-b_1[\theta - a_1])}, \tag{2}$$

where a_1 and b_1 are fixed values that vary based on the specific conditions or circumstances, and $\theta = \arctan(\frac{h}{d})$ is the elevation angle, where h and d are the altitude intercept and the projector range between the UAV-IRS and the BS/users. Then, the path loss can be described as

$$PL = (\eta_{LoS} P_{LoS}(\theta) + P_{NLoS}(\theta)\eta_{NLoS}) \times \beta_0 d^{-\alpha_0}, \tag{3}$$

where $\beta_0 = \left(\frac{4\pi f}{c}\right)^{-2}$ is the constant coefficient related to the antenna gain and frequency.

$\Theta = \text{diag}\left[\beta_1 e^{j\theta_1} \cdots \beta_M e^{j\theta_M}\right]^H$ represents the passive beamforming of the IRS, in which $\beta_m, \theta_m, \forall m \in \{1, 2, \ldots M\}$ are the amplitude and phase shift coefficients of the m-th reflection element, respectively. In this paper, the primary focus of the optimization adjustment lies in the phase shift. Therefore, the amplitude coefficient is fixed to a value of one, i.e., $\beta_m = 1$. The reflected interference channel from other transmitters can be represented as $h_{j,k,TA} \in \mathbb{C}^{M \times 1}, j \neq k$. Subsequently, the received signal at receiver D_k^r is

$$\begin{aligned} y_{D,k}[n] = & \underbrace{\boldsymbol{h}_{k,AR}^H[n]\Theta[n]\boldsymbol{h}_{k,TA}[n]\sqrt{p_k[n]}s_k}_{\text{The desired signal}} + n_d \\ & + \underbrace{\sum_{j \neq k}^{K} \boldsymbol{h}_{k,AR}^H[n]\Theta[n]\boldsymbol{h}_{j,TA}[n]\sqrt{p_j[n]}s_j}_{\text{Interference from other D2D transmitters}} \\ & + \underbrace{\sum_{l=1}^{L} (\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n])\boldsymbol{\omega}_l[n]s_l}_{\text{Interference from BS}}, \end{aligned} \tag{4}$$

where $\boldsymbol{\omega}_l[n] \in \mathbb{C}^{N_t \times 1}, \forall l \in \mathcal{L}$, and p_k, s_k represent the active beamforming vector at the BS, the power coefficient of the transmitter D_k^t, and the transmit data from the D_k^t to D_k^r, respectively. $n_d \sim \mathcal{N}(0, \sigma_d^2)$ denotes the AWGN noise. The corresponding SINR at D_k^r is

$$\gamma_{D,k}[n] = \frac{|(\boldsymbol{h}_{k,AR}^H[n]\Theta[n]\boldsymbol{h}_{k,TA}[n])|^2 p_k[n]}{\sum_{j \neq k}^{K} |(\boldsymbol{h}_{k,AR}^H[n]\Theta[n]h_{j,TA}[n])|^2 p_j[n] + I_{\mathrm{C}}}, \tag{5}$$

where $I_{\mathrm{C}} = \sum_{l=1}^{L} |\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n]\boldsymbol{\omega}_l[n]|^2 + \sigma_d^2$.

In addition, the signal received at CU l can be denoted as

$$\begin{aligned} y_{C,l}[n] = & \underbrace{\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n]\boldsymbol{\omega}_l[n]s_l}_{\text{The desired signal}} + n_0 \\ & + \underbrace{\sum_{k=1}^{K} \boldsymbol{h}_{l,AR}^H[n]\Theta[n]\boldsymbol{h}_{k,TA}[n]\sqrt{p_k[n]}s_k}_{\text{Intereference from D2D transmitters}} \\ & + \underbrace{\sum_{j \neq l}^{L} (\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n])\boldsymbol{\omega}_j[n]s_j}_{\text{Intereference from other CUs}}, \end{aligned} \tag{6}$$

where $n_0 \sim \mathcal{N}(0, \sigma_0^2)$ denotes the AWGN noise at CU l. Accordingly, the corresponding SINR at CU l is

$$\gamma_{C,l}[n] = \frac{|(\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n])\boldsymbol{\omega}_l[n]|^2}{\sum_{k=1}^{K}|(\boldsymbol{h}_{l,AR}^H[n]\Theta[n]\boldsymbol{h}_{k,TA}[n])|^2 p_k[n] + I_{Cl}}, \tag{7}$$

where $I_{Cl} = \sum_{j\neq l}^{L}|(\boldsymbol{h}_{l,AU}^H[n]\Theta[n]\boldsymbol{H}_{BA}[n])\boldsymbol{\omega}_j[n]|^2 + \sigma_0^2$. Thus, the achievable rate of receiver D_k^r and CU l in the n-th time slot can be described as

$$R_{D,k}[n] = \log_2(1 + \gamma_{D,k}[n]), \tag{8a}$$

$$R_{C,l}[n] = \log_2(1 + \gamma_{C,l}[n]), \tag{8b}$$

respectively.

The propulsion power consumption of a rotor-craft UAV is modeled as [36]

$$\begin{aligned} E_{\mathrm{UAV}}[n] = \delta_n \Big[& P_0\left(1 + \frac{3(V[n])^2}{U_{\mathrm{tip}}^2}\right) + \frac{1}{2}d_0\rho s A (V[n])^3 \\ & + P_i\left(\sqrt{1 + \frac{(V[n])^4}{4v_0^4}} - \frac{(V[n])^2}{2v_0^2}\right)^{\frac{1}{2}}\Big], \end{aligned} \tag{9}$$

where $V[n] = \frac{\sqrt{\|\boldsymbol{q}[n]-\boldsymbol{q}[n-1]\|^2}}{\delta_n}$ stands for the UAV's instantaneous speed in the n-th time slot; P_0 and P_i are the constant blade profile power and induced power, respectively; U_{tip} and v_0 are the constant blade tip speed and mean rotor-induced velocity during hover, respectively; ρ, s, and A are the air density, rotor solidity, and rotor disc area, respectively; and d_0 stands for the fuselage drag ratio.

Considering the energy expenditure of the system circuit and IRS elements, the system EE is expressed as

$$EE[n] = \frac{\sum_{k=1}^{K} R_{D,k}[n] + \sum_{l=1}^{L} R_{C,l}[n]}{\sum_{l=1}^{L}\|\boldsymbol{\omega}_l[n]\|^2 + \sum_{k=1}^{K} p_k[n] + E_{\mathrm{CR}}[n]}, \tag{10}$$

where $E_{\mathrm{CR}}[n] = E_{\mathrm{CIR}} + E_{\mathrm{IRS}} + E_{\mathrm{UAV}}[n]$, with E_{CIR} and E_{IRS} being the power consumption of the circuits and the IRS, respectively.

3. Problem Formulation

The aim of this paper is to enhance the average system EE by simultaneously optimizing the flight path of the UAV, $Q = \{\boldsymbol{q}[n], \forall n\}$, the passive beamforming matrix at the IRS, $\Theta = \{\Theta[n], \forall n\}$, the power allocation coefficients for D2D transmitters, $P = \{p_k[n], \forall k, n\}$, and the BS's active beamforming matrix, $\Omega = \{\boldsymbol{\omega}_l[n], \forall l, n\}$. The optimization problem can be represented as

$$P1: \max_{Q,P,\Omega} \quad \frac{1}{N}\sum_{n=1}^{N} EE[n], \tag{11a}$$

$$s.t. (1a), (1b), (1c), \tag{11b}$$

$$\sum_{l=1}^{L}\|\boldsymbol{\omega}_l[n]\|^2 \leq P_{\mathrm{BS}}, \tag{11c}$$

$$p_k[n] \leq P_{\mathrm{D2D}}, \forall k, \tag{11d}$$

$$\theta_m \in (0, 2\pi], \beta_m = 1, \forall m \tag{11e}$$

$$R_{D,k}[n] \geq R_{th,k}, \forall k, \tag{11f}$$

$$R_{C,l}[n] \geq R_{th,l}, \forall l, \tag{11g}$$

where (11b) is the UAV's movement constraint; (11c) and (11d) denote the power constraints at D2D transmitters and the BS, with P_{D2D} and P_{BS} being the maximal transmit power of each transmitter and the BS, respectively; constraints (11f) and (11g) are the QoS constraints of each D2D pair and CU, respectively. The non-convex constraints (11b) and (11e) make the problem intractable to solve. To tackle this challenge, the following section will utilize a MADRL algorithm.

4. The Proposed Solution

In this section, we begin by employing the Markov Game framework to model the optimization problem $P1$. Then, we will introduce the various elements of the multi-agent environment. Since a fully decentralized MARL-based algorithm faces the problem of a non-stationary environment and is difficult to converge, a MATD3 approach based on the CTDE framework is adopted.

4.1. Markov Game Formulation

Since the transmitter in each D2D communication pair cannot directly communicate with other transmitters, the formulated problem $P1$ can be regarded as a Markov Game. In this setting, each communicating node serves as an agent and aims to optimize the long-term cumulative reward by utilizing observations and selecting actions based on its individual policy. Given the non-stationary nature of the environment, it is necessary for all agents to work cooperatively in order to maximize the shared reward. To ensure each communication node works cooperatively, the UAV-IRS, the BS, and each transmitter are considered as agents. Hence, there are $K+2$ agents in the system. Let I_k, I_b, and I_u represent the agents of each transmitter, the BS, and the UAV-IRS, respectively. Thus, the set of all agents can be defined as $I = \{I_1, \ldots, I_K, I_b, I_u\}$. The Markov Game for the considered UAV-IRS-assisted D2D communication scenario can be viewed as a tuple $(\{o_i\}_{i\in I}, \{a_i\}_{i\in I}, \mathcal{P}, r, \gamma)$, where the set of the observation space and action space of $K+2$ agents are denoted as $O = \{o_1, \ldots, o_K, o_b, o_u\}$ and $A = \{a_1, \ldots, a_K, a_b, a_u\}$, respectively. $\mathcal{P}$ refers to the probability of all agents performing actions by exploiting the current state and transitioning to the subsequent state, r is the reward function, and γ is the reward discount factor. In a Markov Game, each agent aim to maximize its own total expected reward $R_i = \sum_{n=0}^{n=N} \gamma r_i, 0 < \gamma \leq 1$. In order to solve the problem of non-stationarity in a multi-agent environment, it is assumed that the policies of all other agents are known. The specific design of observations, actions, and rewards are as follows.

4.1.1. Observation

The observations of I_k, I_b, and I_u are denoted as o_k, o_b, and o_u, respectively. Since each D2D transmitter only knows its local observations and partial interference information, to simplify the analysis process, it is assume that the CSI can be obtained by adopting the channel estimation method that is used in [37,38]. The observation o_k contains the CSI between D_k^t and D_k^r, and the interference information from other D2D transmitters and the BS at the $(n-1)$-th time slot, which can be represented as

$$\begin{aligned} o_k[n] = \{&\boldsymbol{h}_{k,AR}^H[n-1]\Theta[n-1]\boldsymbol{h}_{k,TA}[n-1], \\ &\boldsymbol{h}_{k,AR}^H[n-1]\Theta[n-1]\boldsymbol{h}_{j,TA}[n-1], \\ &\boldsymbol{h}_{l,AU}^H[n-1]\Theta[n-1]\boldsymbol{H}_{BA}[n-1]\}, \forall k \in \mathcal{K}. \end{aligned} \tag{12}$$

Similarly, the observation of the BS contains the CSI between the BS and CUs, and the interference channel information from D2D transmitters and the BS at the $(n-1)$-th time slot, and can be expressed as

$$\begin{aligned} o_b[n] = \{&\boldsymbol{h}_{l,AU}^H[n-1]\Theta[n-1]\boldsymbol{H}_{BA}[n-1], \\ &\boldsymbol{h}_{l,AR}^H[n-1]\Theta[n-1]\boldsymbol{h}_{k,TA}[n-1]\}, \forall l \in \mathcal{L}. \end{aligned} \tag{13}$$

Additionally, the observation o_u contains the UAV trajectory and the CSI from the UAV-IRS to other devices at the $n-1$th time slot, and is given as

$$\begin{aligned} o_u[n] =& \{q[n-1], \boldsymbol{h}_{k,AR}^H[n-1]\Theta[n-1]\boldsymbol{h}_{k,TA}[n-1], \\ & \boldsymbol{h}_{l,AU}^H[n-1]\Theta[n-1]\boldsymbol{H}_{BA}[n-1], \\ & \boldsymbol{h}_{l,AR}^H[n-1]\Theta[n-1]\boldsymbol{h}_{k,TA}[n-1]\}. \end{aligned} \tag{14}$$

4.1.2. Action Space

For agent I_k, the action $a_k[n] = \{p_k[n]\}, \forall k \in \mathcal{K}$ includes the power allocation coefficient of the k-th transmitter. For agent I_b, the action $a_b[n] = \{\omega_1[n], \ldots, \omega_l[n], \ldots, \omega_L[n]\}$ includes the active beamforming vector for all CUs. For agent I_u, the action $a_u[n] = \{\Theta[n], q[n]\}$ contains the passive phase shift matrix and the UAV's trajectory.

4.1.3. Reward Function

Considering our objective of improving the average EE, the reward function can be formulated as follows:

$$\begin{aligned} r(a[n], s[n]) = EE[n] + \eta_{r,k} \sum_{k=1}^{K} p_{r,k}[n] + \eta_{r,l} \sum_{l=1}^{L} p_{r,l}[n] \\ + \eta_b p_b[n] + \eta_g \sum_{k=1}^{K} p_{g,k}[n] + \eta_u p_u[n], \end{aligned} \tag{15}$$

where p_u, $p_{g,k}$, p_b, $p_{r,k}$, $p_{r,l}$ are the penalties when the constraints are not satisfied. Let $p_{r,k} = [R_{th,k} - R_{D,k}[n]]^+$ and $p_{r,l} = [R_{th,l} - R_{D,l}[n]]^+$ denote the QoS constraint penalty; Let $p_{g,k} = [P_{BS} - \sum_{l=1}^{L} \|\omega_k[n]\|^2]^+$ and $p_b = [P_{D2D} - p_k[n]]^+$ denote the maximum transmit power constraint penalty, respectively, in which $[x]^+ = \max\{0, x\}$; the UAV's trajectory constraint penalty is defined as p_u. The non-negative constants $\eta_{r,k}$, $\eta_{r,l}$, η_b, η_g, η_u represent the weight coefficients used to balance the different penalty functions.

4.2. MATD3 Approach

MATD3 is a multi-agent extended version of the single-agent TD3, which adopts the mode of the CTDE framework in the training process. As demonstrated in Figure 2, each agent first obtains the local observation $o_i[n]$ and feeds it into the actor network. Then, each agent obtains the action $a_i[n]$ to execute according to its own policy $\pi_i[n]$ at the time slot n, and interacts with the environment to obtain a new observation $o_i[n+1]$ and store $(o_i[n], a_i[n], r[n], o_i[n+1])$ in the experience buffer pool D. Subsequently, the critic network of each agent incorporates the global state, which encompasses the observations and actions of all other agents.

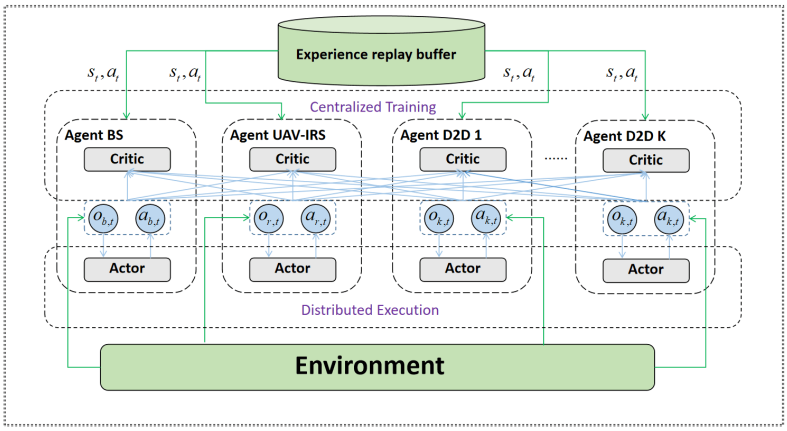


Figure 2. The structure of the MATD3-based algorithm.

Unlike multi-agent DDPG (MADDPG), MATD3 integrates two techniques: clipped double Q-learning (CDQ) and target policy smoothing. In the MATD3 algorithm, the training of the critic network is performed centrally, and input to the critic network includes both the observed states and the actions taken by other agents. The centralized training process assumes training taking place in the UAV-IRS, during which all communication nodes upload channel state information while forwarding data through UAV-IRS. The CDQ-learning technique is utilized to mitigate the issue of Q-value overestimation. Specifically, each agent comprises an evaluated actor network, two evaluated critic networks, a target actor network, and two target critic networks. Each evaluated actor network outputs action $a_i[n] = \pi_i(o_i[n]|\theta_i^{\mu}) + n_i[n]$ with the local observation $o_i[n]$ and the network parameter θ_i^{μ}. The evaluated critic networks output the Q values $Q_{i,1}(o_i[n], a_i[n]|\theta_{i,1}^{q})$ and $Q_{i,2}(o_i[n], a_i[n]|\theta_{i,1}^{q})$ to evaluate the action of the actor network with the local observation $o_i[n]$ and the action $a_i[n]$. Specifically, the centralized Q value is the minimum value between $Q_{i,1}$ and $Q_{i,2}$. During the training phase, a mini-batch of data is utilized to update the critic network parameter θ_i^{q} by minimizing the temporal difference (TD) error. Differently from the single-agent TD3, the critic input of each agent on MATD3 has additional information, such as the actions and observations of other agents, in addition to its own state–action information. TD3 adopts the target network approach to fix the Q-network in the TD target.

The target value y can be computed through

$$\begin{array}{c} y = r_i + \gamma \min\{Q_{i,1}(o', a_1', \ldots, a_I'|\theta_{i,1}^{q\prime}), \\ Q_{i,2}(o', a_1', \ldots, a_I'|\theta_{i,2}^{q\prime})\} \mid_{\boldsymbol{a'}=\{\pi_i(o_i'|\theta_i^{\mu\prime})+\epsilon, i\in\mathcal{I}\}} \end{array}, \tag{16}$$

where $\epsilon = \mathrm{clip}(\mathcal{N}(0, \sigma), -c, c)$ is the clipped Gaussian noise and c is a variable parameter.

The evaluated critic network is updated through minimizing the loss function:

$$\min L(\theta_{i,j}^{q}) = \mathbb{E}_{\boldsymbol{o},a,r,\boldsymbol{o'}}[(Q_i(\boldsymbol{o}, a_1, \ldots, a_I|\theta_{i,j}^{q}) - y)^2], i \in \mathcal{I}, j \in \{1, 2\}. \tag{17}$$

The evaluated actor network of each agent is updated through gradient descent:

$$\begin{array}{c} \nabla_{\theta_i^{\mu}} J_i(\pi_i) = \mathbb{E}_{o,a\sim D}\Big[\nabla\theta_i^{\mu}\pi_i\Big(o_i|\theta_i^{\mu}\Big)\times \\ \nabla_{a_i} Q_i\Big(\boldsymbol{o}, \boldsymbol{a}|\theta_{i,1}^{q}\Big)\Big|_{\boldsymbol{a}=\{\pi_i(o_i|\theta_i^{\mu}), i\in\mathcal{I}\}}\Big]. \end{array} \tag{18}$$

The target network parameters $\theta_{i,1}^{q\prime}$, $\theta_{i,2}^{q\prime}$ and $\theta_i^{\mu\prime}$ are updated by

$$\begin{array}{l} \theta_{i,1}^{q'} \leftarrow \theta_{i,1}^{q} + (1-\tau)\theta_{i,1}^{q'}, \\ \theta_{i,2}^{q'} \leftarrow \theta_{i,2}^{q} + (1-\tau)\theta_{i,2}^{q'}, \\ \theta_i^{\mu'} \leftarrow \theta_i^{\boldsymbol{\mu}} + (1-\tau)\theta_i^{\mu'}. \end{array} \tag{19}$$

4.3. Prioritized Experience Replay

The technique of classical experience replay involves randomly selecting samples from the experience replay buffer in a uniform manner. This is carried out to reduce the correlation between the samples. However, the importance of experience is ignored. In the case of sparse rewards, agents receive a reward only after executing multiple correct actions, resulting in limited transitions to encourage proper learning. Therefore, using random sampling of experiences in this scenario can result in reduced learning efficiency. Prioritized experience replay is adopted to solve such problems [39], which is an improved method of the experience replay buffer. It determines the order in which experience samples are extracted by introducing a priority and replay probability. Based on priority, high-priority experience samples are more likely to be extracted and used for training the agent.

In this paper, proportional prioritization is adopted to restore the priority of a transition, which can be expressed as follows:

$$p_i^t = |\delta_i^t| + \epsilon, \tag{20}$$

where δ_i is the TD error and ϵ is a small fixed value to avoid the probability of zero. The TD error represents the difference between the current Q value and the Q value that should be pursued in the next step, and higher TD errors will be assigned a higher priority. Thus, the sampling probability can be defined as follows:

$$P_i^t = \frac{(p_i^t)^\alpha}{\sum_k (p_i^k)^\alpha}, \tag{21}$$

where α is the hyperparameter that regulates the degree of priority.

Due to the introduction of bias through PER, which alters the data distribution, it becomes necessary to employ importance sampling to mitigate the impact. The weights for importance sampling can be defined as follows:

$$w_i^t = \left(\frac{1}{N_0} \cdot \frac{1}{P_i^t}\right)^\beta, \tag{22}$$

where β is the hyperparameter that regulates the degree of bias introduced by PER and N_0 corresponds to the number of existing transitions in the experience replay pool. Consequently, Equation (16) can be reconstructed as the loss function:

$$L(\theta_i^q) = \mathbb{E}_{\boldsymbol{o},a,r,\boldsymbol{o}'}[\omega_i(Q_i(\boldsymbol{o}, a_1, \ldots, a_I|\theta^q) - y)^2]. \tag{23}$$

Algorithm 1 outlines the training process of the optimization for resource allocation and trajectory using MATD3.

Algorithm 1: Joint resource allocation and trajectory using MATD3

Initialize the parameters of the evaluated actor network $\theta_i^{\boldsymbol{\mu}}$, and the evaluated critic networks $\theta_{i,1}^{\boldsymbol{q}}$ and $\theta_{i,2}^{\boldsymbol{q}}$, respectively;
Initialize target actor and critic network parameters $\theta_i^{\boldsymbol{\mu}'}$, $\theta_{i,1}^{\boldsymbol{q}'}$, and $\theta_{i,2}^{\boldsymbol{q}'}$, respectively;
Initialize prioritized experience reply buffer $\mathcal{D}$;
Initialize observation of each agent i;
for *i=0 to maximum training episode* $\mathcal{E}$ **do**
    Each agent i get local observation o_i;
    for *j=0 to maximum training step* $\mathcal{M}$ **do**
        Each agent i gets local observation o_i;
        Execute actions $A = \{a_1, \ldots, a_K, a_b, a_u\}$;
        Get new observation o_i' and reward r_i;
        Store transition (o_i, a_i, r_i, o_i') into $\mathcal{D}$;
        Sample a mini-batch of transitions (o_i, a_i, r_i, o_i') from $\mathcal{D}$ (21);
        Get the target value y (17);
        Update the parameters of the critic network by minimizing the loss function described in Equation (23);
        Update the parameters of the actor network using the policy gradient method outlined in Equation (18);
        Update target network parameters (19).
    end
end

4.4. Complexity Analysis

The time computation complexity depends on the network operations between two layers. Since our proposed algorithm includes two actor networks and four critic networks, its complexity is given by $O(2 \times \sum_{l=1}^{L_a} n_l^a \cdot n_{l+1}^a + 4 \times \sum_{u=1}^{L_c} n_u^c \cdot n_{u+1}^c)$, in which the n_l-th and the n_u-th are the number of operations of the l-th actor and the u-th critic network layers, respectively; L_a represents the number of layers in the actor network, while L_c represents the number of layers in the critic network. For the online execution phase, the time complexity of each actor network is $O(2 \times \sum_{l=1}^{L_a} n_l^a \cdot n_{l+1}^a)$.

5. Simulation Results

In this section, we validate the efficacy of the proposed algorithm in optimizing resource allocation and designing the UAV trajectory. It is assumed that the UAV maintains a constant altitude of $H_u = 25$ m throughout the flight. For comparison, we compare the system EE performance under fixed position and random trajectory. The UAV-IRS's initial position is configured as (25 m, 25 m, 25 m), while the location of the BS is fixed at (0 m, 30 m, 5 m). Moreover, the position of the UAV in the fixed position scheme is consistent with the initial position of the proposed scheme. All CUs and D2D pairs are located around the initial UAV position. The simulation platform is based on AMD Ryzen 7735H, NVIDIA GeForce RTX4060, python3.7.4 and Torch-1.12.1. Each agent's actor and critic networks are built with three fully connected layers, comprising 512, 256, and 128 neural units, respectively. The energy consumption model parameters of the rotary-wing UAV are set based on [36]. Formal verification is crucial in IoT systems to ensure safety, security, and reliability by detecting errors, verifying complex interactions, and enhancing trust in the system's performance, scalability, and compliance with specifications [40,41]. In order to reduce the complexity of the experiment and considering that we currently only have a single drone, we have not yet addressed the issues of safe drone operation and carried out formal verification. For the relevant parameters of the probabilistic path loss model, we set $a = 9.61, n = 0.16$ [35]. According to [42], the total energy consumption of the system's circuit power and the IRS is set as $E_{\mathrm{IRS}} + E_{\mathrm{CIR}} = 4w$. For the hidden layer, the Relu activation function is applied and, for the output layer of actor networks, the Tanh activation function is employed. The number of training episodes is set as 5000 and each episode has 200 time steps. In this paper, each time step is treated as a time slot. Additional simulation parameters can be found in Table 1 [36,43].

Figure 3 illustrates the trajectory of the UAV after optimization. Obviously, the UAV will fly to a fixed area and fluctuate in a small range after the optimization. When it flies to the center of the map, the UAV hovers at the center of all users to enhance the average system EE while meeting the QoS constraints of all users.

Figure 4a shows the episode return variation during training with different learning rate settings. Observing the learning process under different learning rate settings, it becomes apparent that the MATD3-based algorithm gradually converges. Obviously, the learning curve tends to converge after 1000 episodes. Specifically, the proposed algorithm at the learning rate of 0.0001 converges faster and obtains a better performance than the proposed algorithm at the learning rate of 0.001 and 0.00001. The reason behind this is that the excessively large learning rate may result in excessively large weight updates, thereby causing the loss to become too small and missing the optimal solution. Furthermore, a too small learning rate also leads to poor performance, since it results in tiny updates to the parameter, making it difficult for the algorithm to effectively learn the characteristics of the environment and the reward function. Figure 4b illustrates the average EE versus episodes with the same parameter setting. This proves that the reward setting is consistent with our optimization goal. Finally, the proposed algorithm is superior when examining the converged average EE with the learning rate being set to 0.0001.

Table 1. Simulation parameters.

Parameters	Description	Value
σ_0, σ_d	Noise power	−169 dBm
f	Carrier frequency	2.8 GHz
P_{BS}	The BS's maximum transmit power	30 dBm
P_{D2D}	Transmitter's maximum transmit power	23 dBm
α_0	Path loss exponent	2.2
K	Number of D2D pairs	2
L	Number of CUs	2
U_{tip}	Blade tip speed	200 m/s
d_0	Drag coefficient of the UAV fuselage	0.3
ρ	The air density	1.225 kg/m^3
s	The solidity of the UAV's rotor	0.05
A	Rotor disc area of the UAV	0.79
P_0	The blade profile power of the UAV	580.65
P_i	The induced power of the UAV	790.6715
$R_{th,l}$	Minimum rate requirements of CU l	0.25 bps/Hz
$R_{th,k}$	Minimum rate requirements of D2D k	0.25 bps/Hz
v_{max}	UAV's max velocity	25 m/s
δ_n	Time slot length	0.1 s
η_{NLoS}	Additional attenuation coefficient of LoS	−20 dB
η_{LoS}	Additional attenuation coefficient of LoS	−1 dB
γ	Discount factor for rewards	0.95
τ	Soft update rate	0.01
$\mathcal{B}$	Batch size	256
$\mathcal{N}$	Experience buffer size	50,000
$\mathcal{M}$	Maximum training steps per episode	200
$\mathcal{E}$	Maximum training episodes	5000
α	The hyperparameter of PER	0.6
β	The hyperparameter of PER	0.4

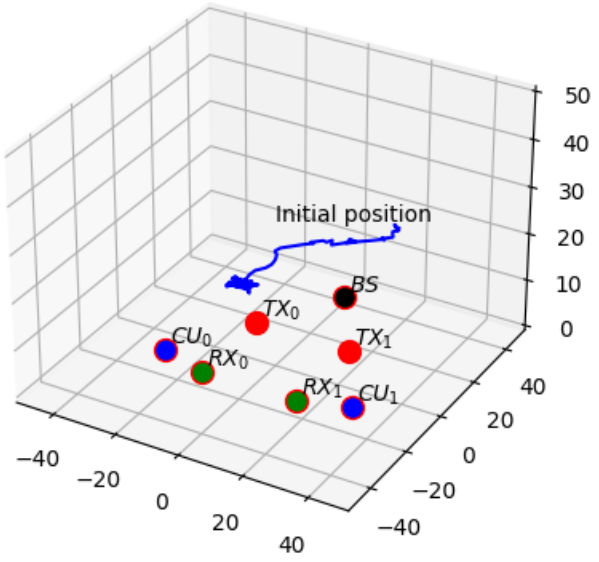
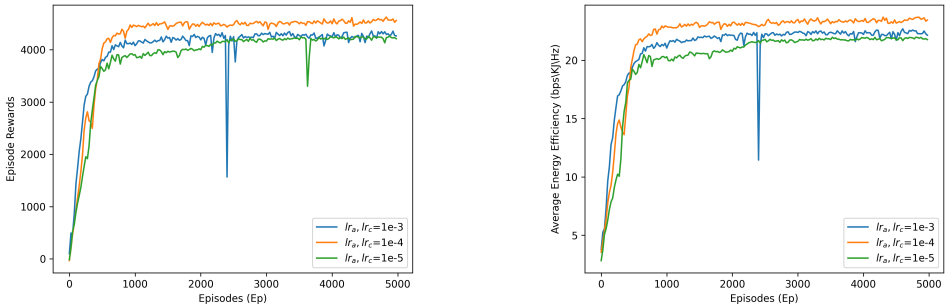


Figure 3. The UAV-IRS trajectory.

Figure 5 demonstrates the average EE versus the number of episodes of different schemes with $M = 25$. The benchmark schemes include DDPG, TD3, SD3 [44], MADDPG, MATD3 with fixed UAV position, and MATD3 with fixed phase shifts. The state input of all single-agent algorithms (DDPG, TD3, SD3) is set to global state information. The EE improvement achieved by simultaneously optimizing the resource allocation and the UAV trajectory is clearly superior to that of the fixed position and the fixed phase shift schemes. The main reason is that the reflective surface in the fixed position is difficult to meet the

rate constraint of all users, and another reason is that the UAV-IRS requires greater energy consumption to sustain a hover state and maintain a fixed position, compared to maintaining a flight state. It can also be observed that the MATD3-based scheme surpasses the MADDPG-based scheme and all single-agent-based schemes; the DDPG-based scheme performs the worst due to the presence of high input dimensions. Furthermore, since MATD3 introduces dual Q-value networks, it increases the complexity of the training process and then results in a slower convergence speed compared to the MADDPG-based scheme.



(a) (b)

Figure 4. The training process of MATD3 algorithm under different learning hyperparameters. (**a**) The episode return versus episodes. (**b**) The average EE versus episodes.

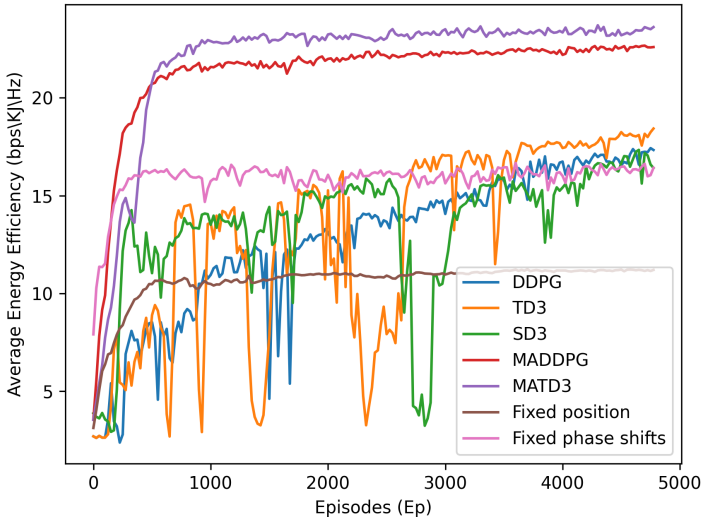


Figure 5. The average EE versus episodes.

Furthermore, Figure 6 illustrates the correlation between the number of elements (M) in the IRS and the average EE. It can be observed that increasing M can enhance the system EE when it is small; however, the EE performance gain becomes smaller and smaller when M is very large. This is due to the increased dimensionality of the action space, which requires more time for the algorithm to achieve convergence. In the comparison, it is evident that the proposed MATD3-based approach obtains a higher EE than the other two benchmarks in the case of more IRS elements.

Figure 7 illustrates the relationship between the system's average EE and the maximum transmit power constraint , P_{D2D}, under various schemes. Specifically, in the fixed position method, without optimizing the movement, the UAV is fixed at a fixed location and during the entire time period. In the random trajectory method, without optimizing the trajectory, the UAV moves randomly. The average EE of the proposed scheme gradually improves with an increase in the maximum transmit power budget, P_{D2D}, as illustrated in Figure 7. This improvement can be attributed to the increase in the system sum rate achieved by allocating more transmit power. It can also observed that the random trajectory scheme

and fixed position scheme cannot gain improvement by increasing the transmit power. Furthermore, the figure demonstrates that, by deploying the IRS in a flexible manner, the performance of the system EE can be greatly enhanced.

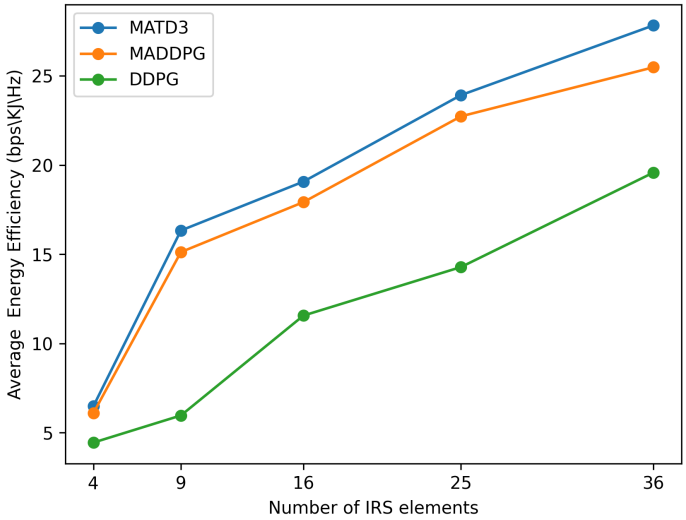


Figure 6. The average EE versus M.

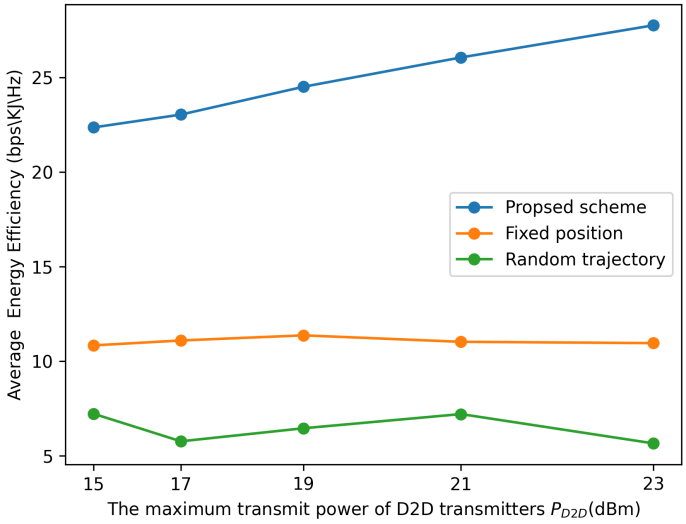


Figure 7. The average EE versus P_{D2D}.

6. Conclusions

This study investigates the application of UAV-IRS-assisted D2D communications to enhance system EE by jointly optimizing the UAV trajectory, D2D transmitter power coefficients, BS's beamforming vector, and IRS passive beamforming matrix, while the rate requirements of all cellular users and D2D users are simultaneously satisfied. Considering the distributed nature of D2D communication, we use MADRL to train each communication node and propose a training framework based on the CTDE framework. To address the formulated problem, we initially formulate it as a Markov Game and subsequently introduce a solution approach utilizing MATD3. The numerical results indicate that our proposed scheme greatly improves the EE. However, the scalability of MATD3 in real-world scenarios is a potential challenge. As the number of agents increases, the complexity of the interaction among them grows exponentially. This can lead to issues such as communication overhead, increased computational resources, and difficulties in maintaining a stable training process. Another point to note is that this work does not consider the onboard

energy constraints of the UAV and the channel allocation for D2D communications, which will be addressed in future works. In the current experiments, we have overlooked the complexity of the real environment. In future work, we will consider further designing path planning for security issues and conducting formal verification using real-time data.

Author Contributions: Conceptualization, F.T. and S.P.; methodology, S.P. and H.C.; software, S.P.; validation, F.T., and H.C.; formal analysis, F.T. and T.L.; investigation, Y.C.; resources, F.T.; data curation, H.C.; writing—original draft preparation, S.P.; writing—review and editing, F.T.; visualization, Y.C.; supervision, H.C.; project administration, Y.C. and T.L.; funding acquisition, F.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China under Grants 62261013 and 62061009, in part by the Guangxi Natural Science Foundation under Grants 2023JJA170030, and in part by Director Foundation of Guangxi Key Laboratory of Wireless Wideband Communication and Signal Processing under Grants GXKL06220104.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

UAV-IRS	Unmanned aerial vehicle-mounted intelligent reflecting surface
IRS	Intelligent reflecting surface
UAV	Unmanned aerial vehicle
MISO	Multiple input and single output
MIMO	Multiple input and multiple output
BS	Base station
CU	Cellular user
SINR	Signal-to-interference-plus-noise ratio
MDP	Markov decision process
LoS	Line-of-sight
NLoS	Non-line-of-sight
MADRL	Multi-agent deep reinforcement learning
DDPG	Deep deterministic policy gradient
TD3	Twin delayed deep deterministic policy gradient
MADDPG	Multi-agent deep deterministic gradient
MATD3	Multi-agent twin delayed deep deterministic policy gradient
EE	Energy efficiency
CTDE	Centralized training and decentralized execution
PER	Prioritized experience replay
QoS	Quality of service
CSI	Channel state information

References

1. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tut.* **2015**, *17*, 2347–2376. [CrossRef]
2. Lin, J.; Yu, W.; Zhang, N.; Yang, X.; Zhang, H.; Zhao, W. A survey on internet of things: Architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J.* **2017**, *4*, 1125–1142. [CrossRef]
3. Verma, S.; Kaur, S.; Khan, M.A.; Sehdev, P.S. Toward green communication in 6G-enabled massive internet of things. *IEEE Internet Things J.* **2021**, *8*, 5408–5415. [CrossRef]
4. Li, Z.; Guo, C. Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications. *IEEE Trans. Veh. Technol.* **2020**, *69*, 1828–1840. [CrossRef]
5. Bello, O.; Zeadally, S. Intelligent device-to-device communication in the internet of things. *IEEE Syst. J.* **2016**, *10*, 1172–1182. [CrossRef]
6. Li, Y.; Jin, D.; Yuan, J.; Han, Z. Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 3965–3977. [CrossRef]
7. Chen, X.; Proulx, B.; Gong, X.; Zhang, J. Exploiting social ties for cooperative D2D communications: A mobile social networking case. *IEEE/ACM Trans. Netw.* **2015**, *23*, 1471–1484. [CrossRef]

8. Deng, L.; Wu, G.; Fu, J.; Zhang, Y.; Yang, Y. Joint resource allocation and trajectory control for UAV-enabled vehicular communications. *IEEE Access* **2019**, *7*, 132806–132815. [CrossRef]
9. Zhu, C.; Leung, V.C.M.; Shu, L.; Ngai, E.C.-H. Green internet of things for smart world. *IEEE Access* **2015**, *3*, 2151–2162. [CrossRef]
10. Su, Z.; Feng, W.; Tang, J.; Chen, Z.; Fu, Y.; Zhao, N.; Wong, K.K. Energy-efficiency optimization for D2D communications underlaying UAV-assisted industrial IoT networks with SWIPT. *IEEE Internet Things J.* **2023**, *10*, 1990–2002. [CrossRef]
11. Jia, S.; Yuan, X.; Liang, Y.-C. Reconfigurable intelligent surfaces for energy efficiency in D2D communication network. *IEEE Wireless Commun. Lett.* **2021**, *10*, 683–687. [CrossRef]
12. Chen, Y.; Ai, B.; Zhang, H.; Niu, Y.; Song, L.; Han, Z.; Poor, H.V. Reconfigurable intelligent surface assisted device-to-device communications. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 2792–2804. [CrossRef]
13. Ji, Z.; Qin, Z.; Parini, C.G. Reconfigurable intelligent surface aided cellular networks with device-to-device users. *IEEE Trans. Commun.* **2022**, *70*, 1808–1819. [CrossRef]
14. Gong, S.; Lu, X.; Hoang, D.T.; Niyato, D.; Shu, L.; Kim, D.I.; Liang, Y.-C. Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey. *IEEE Commun. Surv. Tut.* **2020**, *22*, 2283–2314. [CrossRef]
15. Wu, Q.; Zhang, S.; Zheng, B.; You, C.; Zhang, R. Intelligent reflecting surface-aided wireless communications: A tutorial. *IEEE Trans. Commun.* **2021**, *65*, 3313–3351. [CrossRef]
16. Sultana A, Fernando X. Intelligent reflecting surface-aided device-to-device communication: A deep reinforcement learning approach. *Future Internet.* **2022**, *14*, 1–18. [CrossRef]
17. Huang, C.; Mo, R.; Yuen, C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J. Sel. Area Comm.* **2020**, *38*, 1839–1850. [CrossRef]
18. Feng, K.; Wang, Q.; Li, X.; Wen, C.-K. Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems. *IEEE Wireless Commun. Lett.* **2020**, *9*, 745–749. [CrossRef]
19. Zhang, J.; Li, J.; Zhang, Y.; Wu, Q.; Wu, X.; Shu, F.; Jin, S.; Chen, W. Collaborative intelligent reflecting surface networks with multi-agent reinforcement learning. *IEEE Commun. Surv. Tutor.* **2022**, *16*, 532–545. [CrossRef]
20. Guo, L.; Jia, J.; Chen, J.; Du, A.; Wang, X. Deep reinforcement learning empowered joint mode selection and resource allocation for RIS-aided D2D communications. *Neural Comput. Applic.* **2023**, *35*, 18231–18249. [CrossRef]
21. Jin, N.; Liao, Y.; Yang, G.; Liang, Y.-C.; Chen, X. Energy-efficient symbiotic cellular-UAV communication via aerial RIS: Joint trajectory design and resource optimization. In Proceedings of the 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall), London, UK, 26–29 September 2022; pp. 1–6.
22. Abdalla, A.S.; Marojevic, V. DDPG learning for aerial RIS-assisted MU-MISO communications. In Proceedings of the 2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Kyoto, Japan, 12–15 September 2022; pp. 701–706.
23. Mei, H.; Yang, K.; Liu, Q.; Wang, K. 3D-trajectory and phase-shift design for RIS-assisted UAV systems using deep reinforcement learning. *IEEE Trans. Veh. Technol.* **2022**, *71*, 3020–3029. [CrossRef]
24. Peng, H.; Wang, L.-C. Energy harvesting reconfigurable intelligent surface for UAV based on robust deep reinforcement learning. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 6826–6838. [CrossRef]
25. Guo, X.; Chen, Y.; Wang, Y. Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave UAV communications. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 1795–1799. [CrossRef]
26. Wang, D.; Liu, Y.; Yu, H.; Hou, Y. Three-dimensional trajectory and resource allocation optimization in multi-unmanned aerial vehicle multicast system: A multi-agent reinforcement learning method. *Drones* **2023**, *7*, 641. [CrossRef]
27. Budhiraja, I.; Kumar, N.; Tyagi, S. Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication. *IEEE Internet Things J.* **2021**, *8*, 3143–3156. [CrossRef]
28. Xu, J.; Kang, X.; Zhang, R.; Liang, Y.-C.; Sun, S. Optimization for master-UAV-powered auxiliary-aerial-IRS-assisted IoT networks: An option-based multi-agent hierarchical deep reinforcement learning approach. *IEEE Internet Things J.* **2022**, *9*, 22887–22902. [CrossRef]
29. Hou, Z.; Huang, Y.; Chen, J.; Li, G.; Guan, X.; Xu, Y.; Chen, R.; Xu, Y. Joint IRS selection and passive beamforming in multiple IRS-UAV enhanced anti-jamming D2D communication networks. *IEEE Internet Things J.* **2023**, *10*, 19558–19569. [CrossRef]
30. Farrag, S.; Maher, E.A.; El-Mahdy, A.; Dressler, F. Sum rate maximization of uplink active RIS and UAV-assisted THz mobile communications. In Proceedings of the 2023 19th International Conference on the Design of Reliable Communication Networks (DRCN), Vilanovai la Geltru, Spain, 17–20 April 2023; pp. 1–7.
31. You, Q.; Xu, Q.; Yang, X.; Sun, W.-B.; Wang, L. Distributed deep learning for RIS aided UAV-D2D communications in space-air-ground networks. In Proceedings of the 2023 IEEE/CIC International Conference on Communications in China (ICCC), Dalian, China, 10–12 August 2023; pp. 1–6.
32. Vishnoi, V.; Consul, P.; Budhiraja, I.; Gupta, S.; Kumar, N. Deep reinforcement learning based energy consumption minimization for intelligent reflecting surfaces assisted D2D users underlaying UAV network. In Proceedings of the IEEE INFOCOM 2023—IEEE Conference on Computer and Communications Workshops (INFOCOM WKSHPS), Hoboken, NJ, USA, 17–20 May 2023; pp. 1–6.
33. Li, T.; Zhu, K.; Luong, N.C.; Niyato, D.; Wu, Q.; Zhang, Y.; Chen, B. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2022**, *24*, 1240–1279. [CrossRef]

34. Ackermann, J.; Gabler, V.; Osa, T.; Sugiyama, M. Reducing overestimation bias in multi-agent domains using double centralized critics. *arXiv* **2019**, arXiv:1910.01465.
35. Al-Hourani, A.; Kandeepan, S.; Lardner, S. Optimal LAP altitude for maximum coverage. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 569–572. [CrossRef]
36. Zhan, C.; Lai, H. Energy minimization in internet-of-things system based on rotary-wing UAV. *IEEE Wireless Commun. Lett.* **2019**, *8*, 1341–1344. [CrossRef]
37. Wang, Z.; Liu, L.; Cui, S. Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6607–6620. [CrossRef]
38. Wu, Q.; Zhang, R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network. *IEEE Commun. Mag.* **2020**, *58*, 106–112. [CrossRef]
39. Schaul, T.; Quan, J.; Antonoglou, I.; Silver, D. Prioritized experience replay. *arXiv* **2015**, arXiv:1511.05952.
40. Krichen, M. A Survey on formal verification and validation techniques for internet of things. *Appl. Sci.* **2023**, *13*, 8122. [CrossRef]
41. Hofer-Schmitz, K.; Stojanović, B. Towards formal verification of IoT protocols: A Review. *Comput. Netw.* **2020**, *174*, 107233. [CrossRef]
42. Cui, J.; Liu, Y.; Nallanathan, A. Multi-agent reinforcement learning-based resource allocation for UAV networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 729–743. [CrossRef]
43. Nguyen, K.K.; Khosravirad, S.R.; da Costa, D.B.; Nguyen, L.D.; Duong, T.Q. Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning. *IEEE J. Sel. Topics Signal Process* **2022**, *16*, 358–368. [CrossRef]
44. Pan, L.; Cai, Q.; Huang, L. Softmax deep double deterministic policy gradients. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11767–11777.