drones

MDPI

Article

Joint Resource Allocation and Drones Relay Selection for Large-Scale D2D Communication Underlaying Hybrid VLC/RF IoT Systems

Xuewen Liu [1,†], Shuman Huang [2,†], Kaisa Zhang [3,*], Saidiwaerdi Maimaiti [4,*], Gang Chuai [2], Weidong Gao [2], Xiangyu Chen [2], Yijian Hou [2] and Peiliang Zuo [1]

[1] Department of Electronics and Communication Engineering, Beijing Electronics Science and Technology Institute, Beijing 100070, China; xuewen_liu1990@bupt.edu.cn (X.L.); zplzpl88@bupt.cn (P.Z.)
[2] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China; huangsm@bupt.edu.cn (S.H.); chuai@bupt.edu.cn (G.C.); gaoweidong@bupt.edu.cn (W.G.); xychen324@bupt.edu.cn (X.C.); houyijian@bupt.edu.cn (Y.H.)
[3] School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[4] School of Physics and Electrical Engineering, Kashi University, Kashi 844008, China
* Correspondence: kaisa@bupt.edu.cn (K.Z.); saidi713@163.com (S.M.)
† These authors are co-first authors of this article.

Abstract: Relay-aided Device-to-Device (D2D) communication combining visible light communication (VLC) with radio frequency (RF) is a promising paradigm in the internet of things (IoT). Static relay limits the flexibility and maintaining connectivity of relays in Hybrid VLC/RF IoT systems. By using a drone as a relay station, it is possible to avoid obstacles such as buildings and to communicate in a line-of-sight (LoS) environment, which naturally aligns with the requirement of VLC Systems. To further support the application of VLC in the IoT, subject to the challenges imposed by the constrained coverage, the lack of flexibility, poor reliability, and connectivity, drone relay-aided D2D communication appears on the horizon and can be cost-effectively deployed for the large-scale IoT. This paper proposes a joint resource allocation and drones relay selection scheme, aiming to maximize the D2D system sum rate while ensuring the quality of service (QoS) requirements for cellular users (CUs) and D2D users (DUs). First, we construct a two-phase coalitional game to tackle the resource allocation problem, which exploits the combination of VLC and RF, as well as incorporates a greedy strategy. After that, a distributed cooperative multi-agent reinforcement learning (MARL) algorithm, called WoLF policy hill-climbing (WoLF-PHC), is proposed to address the drones relay selection problem. Moreover, to further reduce the computational complexity, we propose a lightweight neighbor–agent-based WoLF-PHC algorithm, which only utilizes historical information of neighboring DUs. Finally, we provide an in-depth theoretical analysis of the proposed schemes in terms of complexity and signaling overhead. Simulation results illustrate that the proposed schemes can effectively improve the system performance in terms of the sum rate and outage probability with respect to other outstanding algorithms.

Keywords: drone relay-aided D2D communication; resource allocation; drones relay selection; visible light communication (VLC); WoLF policy hill-climbing (WoLF-PHC)

Citation: Liu, X.; Huang, S.; Zhang, K.; Maimaiti, S.; Chuai, G.; Gao, W.; Chen, X.; Hou, Y.; Zuo, P. Joint Resource Allocation and Drones Relay Selection for Large-Scale D2D Communication Underlaying Hybrid VLC/RF IoT Systems. *Drones* **2023**, *7*, 589. https://doi.org/10.3390/drones7090589

Academic Editor: Emmanouel T. Michailidis

Received: 1 August 2023
Revised: 14 September 2023
Accepted: 15 September 2023
Published: 19 September 2023

1. Introduction

With the wide application of the Internet of Things (IoT) in various fields such as city, industry, and transportation, a constant emergence of IoT devices (IoDs) are connected via the Internet to exchange information about themselves and their surroundings. It is expected that the number of IoDs will increase to 75.4 billion by 2025, more than 9-fold the number in 2017 [1]. The proliferation of IoDs puts higher demands on the spectrum, data rate, and latency for IoT communications. In response, device to device (D2D) communication, where two nearby devices can exchange information directly, has been widely

employed in IoT networks to improve spectrum efficiency and data rates, along with reducing transmission delays [2–4]. Depending on whether radio frequency (RF) resources are shared between D2D users (DUs) and traditional cellular users (CUs), D2D communication can be classified into two categories: underlay and overlay communication. In particular, underlay D2D communication has been proven to provide a higher spectrum efficiency and match spectrum sharing nature in IoT networks [5]. However, it will inevitably lead to mutual interference between DUs and CUs. In addition to the absence of electro-magnetic interference with existing RF systems, the emerging visible light communication (VLC) offers many advantages, such as a broad spectrum, innate security, and license-free deployment [6]. Yet, VLC is also susceptible to blockage and has severe path attenuation. Therefore, combining RF and VLC bands for D2D communication has been regarded as an enticing solution to mitigate the interference and overcrowding of the RF spectrum, thus boosting the system capacity [7–9].

However, the envisioned benefits of D2D communication may be limited by long distances, obstacles, and inferior channel conditions, especially for VLC-D2D communication. As a result, D2D communication is not well-suited for IoT applications that require wide coverage and high reliability [10]. A promising response to this dilemma is to implement relay-aided D2D communication, which is able to extend the communication range as well as improve both reliability and flexibility [11]. That is, D2D communication can be extended to a relay-aided manner when IoDs that need to communicate are far away from each other or are blocked by obstacles. Such relay-aided systems are feasible for the large-scale IoT without extra construction costs, like massive machine-type communication, the cognitive IoT, and wireless sensing, as there are a large number of available IoDs (e.g., sensors, actuators, drones) that can act as relays [12–14]. Unmanned aerial vehicles (UAVs) have been widely used in both military and civilian applications [15,16]. Renhai Feng [17] considers unmanned aerial vehicles (UAVs) to relay the maintenance data by visible light communication (VLC) under the requirements of ultra-reliability and low-latency. Zhiyu Zhu [18] and Yining Wang [19] enable UAVs to determine their deployment and user association to minimize the total transmit power with VLC. In [20], the authors optimized the UAV-assisted VLC framework that aims at minimizing the required number of UAVs first and minimizing the total transmitted power second. In [21], the authors consider UAVs equipped with a VLC access point and the coordinated multipoint (CoMP) capability to maximize the total data rate and minimize the total communication power consumption simultaneously. In [22], the authors describe a UAV-assisted outdoor VLC system to provide high-speed and high-capacity communication for some users who are blocked by natural disasters or mountains, in where the UAV is set as a communication relay. However, to my knowledge, there is no research on using drones as relays for joint resource allocation and relay selection in a hybrid VLC/RF IoT system for D2D communication.

Accordingly, we concentrate on a large-scale drone relay-aided D2D communication underlaying hybrid VLC/RF IoT system, where multiple CU, DU, and drone relays coexist. Different from existing research, in this paper, we innovatively introduced drones as relay stations to address the challenges posed by the constrained coverage, poor reliability, and the lake of flexibility. More concretely, each DU corresponds to a pair of IoDs with inferior channel condition, so a drone relay is needed to aid the communication. And this relay can be selected from other idle IoDs. Besides, each DU and its relay are allowed to utilize either VLC resource or orthogonal RF resource of a certain CU. Obviously, there are two main variables that determine the system performance. One is the resource allocation for each DU, which also actually decides the resource used by its corresponding relay. For large-scale IoT, the number of DUs is typically higher than that of CUs. This means that although the VLC resource is included, some DUs still share the same resource, causing mutual interference among the DUs, the corresponding relays, and perhaps a certain CU. Hence, how to fully leverage the combination of VLC and RF to alleviate the mutual interference is a crucial issue. Another variable is the drone relay selection for each DU. For large-scale IoT, a DU has multiple available relays. However, different relays bring not

only different communication gains to the DU, but also different interferences to the users sharing the same resource. Thus, how to select the relays to improve the overall system performance is another important issue.

So far, there have been some works on resource allocation [23–27], relay selection [28–30], and their joint optimization [31–37] for relay-aided D2D communication. However, there is no research on using drones as relays for joint resource allocation and relay selection for large-scale D2D communication in hybrid VLC/RF IoT system. Static relay limits the flexibility and maintaining connectivity of relays in Hybrid VLC/RF IoT Systems. By using a drone as a relay station, it is possible to avoid obstacles such as buildings and to communicate in a line-of-sight (LoS) environment, which naturally aligns with the requirement of VLC Systems. In addition, most works mainly consider small-scale scenarios in which sharing resources among DUs is not required. Most of the optimization methods proposed in these works are not suitable for large-scale scenarios, where resource allocation and relay selection become more difficult for the following reasons.

1. Large-scale relay-aided D2D communication causes resource shortages, leaving each resource shared by multiple DUs. The arising complex interference relationships make the resource allocation for one DU also impact the performance of other DUs who share the same resource. This motivates us to view the resource allocation process for each DU as finding the optimal set of DUs for each resource, in which the mutual interference is minimal.
2. Similarly, the interference relationships make the relay selection for all DUs within the same set co-dependent. This prompts us to further consider allowing these DUs to cooperate with each other for a higher collective gain.
3. Large-scale IoDs deployment inherently exacerbates the time complexity and signaling overhead required for the optimization process, especially when it comes to relay selection. The optimization schemes are desired to have low complexity due to practical applications.

Against this background, we present a joint optimization of resource allocation and drones relay selection for large-scale relay-aided D2D underlaying hybrid VLC/RF IoT system, aiming to maximize the D2D system sum rate while ensuring the quality of service (QoS) requirements for CUs and DUs. First, inspired by the aforementioned perspective of finding the optimal DU set for each resource, the resource allocation problem can be modeled as a coalitional game. In particular, we construct a two-phase coalitional game that allows each DU to explore and finally join a coalition while guaranteeing QoS. The different coalitions that eventually form are exactly the optimal sets of DUs for different resources. Afterwards, with a large number of DUs and available relays, we regard each DU as an agent that can autonomously select a proper relay through learning. In this way, the relay selection problem is modeled as a multi-agent problem and thus can be solved in a distributed manner. Furthermore, given the aforementioned co-dependency in relay selection among the DUs in the same coalition, we propose two cooperative relay selection schemes based on multi-agent reinforcement learning (MARL) with low complexity, termed WoLF policy hill-climbing (WoLF-PHC). These two proposed schemes can not only overcome the inherent non-stationary of the multi-agent environment, but also encourage the DUs to cooperate for a higher system sum rate. The main contributions of this paper are summarized as follows:

1. The model of the drone relay-aided D2D communication underlaying hybrid VLC/RF system for the large-scale IoT is given. Aiming to maximize the sum rate of the D2D system while ensuring QoS, the joint optimization problem of resource allocation and drones relay selection is formulated. The problem has a nonconvex and combinatorial structure that makes it difficult to be solved in a straightforward way. Thus, we divide it into two subproblems and solve them sequentially.
2. From the perspective of finding the optimal DU set for each resource, we construct a two-phase coalitional game to tackle the resource allocation problem. Specifically, we leverage the combination of VLC and RF to ensure QoS in the coalition initialization

phase. We also incorporate a greedy strategy into the coalition formation phase to obtain the global optimal sets of DUs.

3. In order to eliminate co-dependency, we first propose a cooperative WoLF-PHC-based relay selection scheme, where the agents in the same coalition share a common reward. Meanwhile, in any coalition, each agent's policy can use the historical action information of other agents to overcome the non-stationary of the environment. Interestingly, combining the results of the resource allocation, we find that only the historical information of neighboring agents is sufficient to alleviate the instability. Hence, a lightweight neighbor–agent-based WoLF-PHC algorithm with curtailed complexity is further proposed.
4. We provide a theoretical analysis of the proposed schemes in terms of complexity and signaling overhead. Also, we provide numerical results to indicate that the proposed schemes outperform the considered benchmarks in terms of the sum rate and outage probability. Moreover, we investigate the trade-off between the sum rate performance and computational complexity.

The rest of this paper is organized as follows. Section 2 is the related works. In Section 3, the system model is given and the problem is formulated. In Sections 4 and 5, we present the proposed resource allocation and relay selection schemes, respectively. The complexity and signaling overhead, and the simulation results are shown and analyzed in Sections 6 and 7, respectively. Finally, Section 8 concludes the paper.

2. Related Works

With the potential to substantially increase system capacity, the novel D2D concept combining VLC and RF communication was first proposed in [7]. In [38], a survey on D2D Communication for 5 GB/6G Networks about concept, applications, challenges, and future directions have been discussed. In [39], the authors provide a V2I and V2V collaboration framework to support emergency communications in the ABS-aided internet of vehicles. Up to now, several works have been proposed to study the resource allocation for D2D communication in hybrid VLC/RF systems. In [8], an iterative two-stage resource allocation algorithm was proposed based on the analysis of the interference generated by D2D transmitters and those received by D2D receivers. With only limited channel state information (CSI), the authors in [9] attempted to implement a quick band selection between VLC and RF using deep neural networks. On this basis, refs. [25,26] included a millimeter wave into the hybrid VLC/RF bands and formulated the multi-band selection problem as a multi-armed bandit problem. However, the above works only considered the overlay mode instead of the multiple DUs coexisting in the underlay mode, which is an essential use-case in future networks. Only our previous work [40] considered this use-case for D2D underlay communication and solved the resource allocation problem using the coalitional game. The main difference between our work and previous work is that D2D communication is extended to a relay-aided manner, which gives rise to new problems.

The relay-aided D2D communication appeared due to the demand to extend the communication range as well as enhance both reliability and flexibility. As a matter of fact, jointly optimizing resource allocation and relay selection for relay-aided D2D communication in traditional RF systems has been widely studied. In [31], the joint optimization problem of mode selection, power control, channel allocation, and relay selection was decomposed into four subproblems and solved individually, aiming to maximize the total throughput. However, the authors in [32] first addressed the power control problem separately, and then solved the remaining joint problem using an improved greedy algorithm. Similarly, ref. [33] addressed the power control problem first so that the remaining joint problem could be converted into the tractable integer–linear programming problem. In [34], taking into account both willingness and social attributes, a social-aware relay selection algorithm was proposed, and then a greedy-based resource allocation scheme was presented. Furthermore, in order to motivate users acting as relays, ref. [35] assumed that the relays involved in assisting D2D communication could harvest energy

from RF signals and formulated the optimization problem as a three-dimensional resource–power–relay problem. The authors in [36] focused on an energy efficiency optimization problem of relay-aided D2D communication under simultaneous wireless information and the power transfer paradigm. Besides, ref. [37] derived an energy efficient oriented joint relay selection and resource allocation solution for mobile edge computing systems by using convex optimization techniques. Despite all this, these research works took into consideration neither large-scale nor hybrid VLC/RF scenarios. Moreover, most of the above works on relay selection adopted either the brute-force algorithm based on designated regions or the distance-based algorithm, which have high computational complexities and are not suitable for large-scale applications.

Given the dynamics of practical networks, reinforcement learning (RL) techniques have been introduced to provide a solution to the relay selection problem. Ref. [41] developed a centralized hierarchical deep RL-based relay selection algorithm to minimize the total transmission delay in mmWave vehicular networks. Ref. [42] presented a multi-featured actor-critic RL algorithm to maximize the data delivery ratio in energy-harvesting wireless sensor networks. Also, ref. [43] incorporated the prioritized experience replay into a deep deterministic policy algorithm and minimized outage probability without any prior knowledge of CSI. The above works modeled the policy search process as a Markov decision process, which is true if different agents update their policies independently at different times. Nevertheless, if two or more agents update their policies at the same time, a non-stationary multi-agent environment may occur [44]. How to reduce the action space and computational complexity of multi-agent systems to improve the training speed while ensuring a stationary multi-agent environment is a key issue.

In summary, there are four drawbacks in the above studies:

(1) The above works only considered the overlay mode instead of the multiple DUs coexisting in the underlay mode, which is an essential use-case in future networks.
(2) Although some works focus on jointly optimizing resource allocation and relay selection for relay-aided D2D communication, these works did not take the large-scale IoT or hybrid VLC/RF scenarios into consideration.
(3) Static relay is adopted in existing research, which limits the flexibility and maintaining connectivity of relays in Hybrid VLC/RF IoT Systems. The dynamic relay-assisted D2D communication system with wide coverage, high flexibility, good reliability, and strong connectivity needs to be constructed.
(4) Most of the joint optimization methods proposed in these works are not suitable for large-scale scenarios, and new methods with low complexities and signaling overhead are forced to be developed.

3. System Model and Problem Formulation

3.1. System Description

We consider a drone relay-aided D2D communication underlaying hybrid VLC/RF system for the large-scale IoT, as shown in Figure 1, which consists of M CU, N DU, and R drone relays uniformly distributed in a square room. Note that a DU represents a D2D pair, consisting of a transmitter (DU-TX) and a receiver (DU-RX). Let $\mathcal{N} = \{1, \ldots, n, \ldots, N\}$, $\mathcal{S} = \{1, \ldots, s, \ldots, N\}$, and $\mathcal{D} = \{1, \ldots, d, \ldots, N\}$ denote the set of DUs, DU-TXs, and DU-RXs, respectively. Similarly, $\mathcal{M} = \{1, \ldots, m, \ldots, M\}$ and $\mathcal{R} = \{1, \ldots, r, \ldots, R\}$ represent the set of CU and drone relays, respectively. In this paper, we assume that each CU has been pre-allocated an orthogonal uplink RF resource, i.e., CU m has occupied the RF resource c_m. Combined with the VLC resource c_{M+1}, there are $M+1$ available resources and their set is denoted as $\mathcal{C} = \{c_1, \ldots, c_m, \ldots, c_M, c_{M+1}\}$. Meanwhile, each DU is allowed to reuse a resource from the set $\mathcal{C}$. To describe whether DU $n \in \mathcal{N}$ reuses resource $c_m \in \mathcal{C}$, we introduce a decision matrix for resource allocation:

$$\boldsymbol{\beta} \in (\beta_n^{c_m})_{N*(M+1)} \tag{1}$$

where $\beta_n^{c_m}$ is a binary variable. Specifically, $\beta_n^{c_m} = 1$ denotes that DU n reuses resource c_m; otherwise, $\beta_n^{c_m} = 0$.

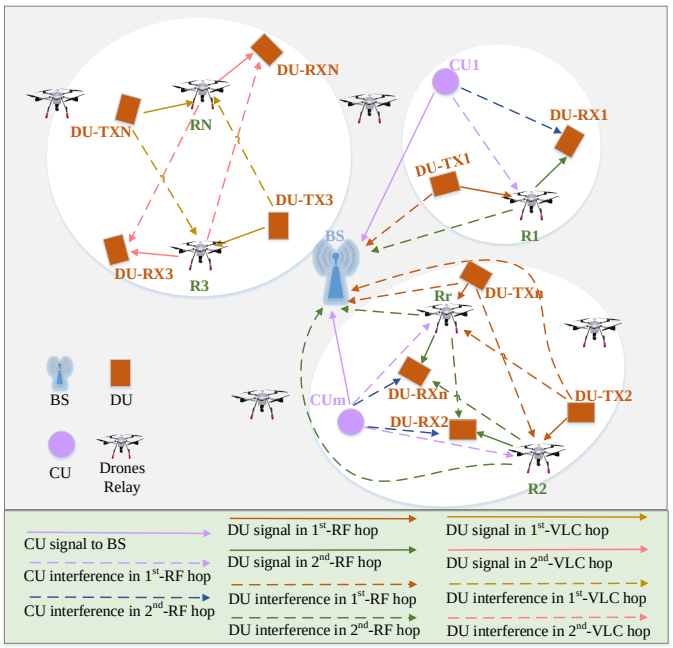

Figure 1. Relay-aided D2D communication underlaying hybrid VLC/RF system model.

For the sake of practicality, it is supposed that each DU can only select one relay for assistance and each relay is allowed to be attached to, at most, one DU at a time. To describe whether DU $n \in \mathcal{N}$ transmits data with the help of relay $r \in \mathcal{R}$, we introduce another decision matrix for relay selection:

$$\boldsymbol{\alpha} \in (\boldsymbol{\alpha_{n,r}})_{N*R} \tag{2}$$

where $\alpha_{n,r}$ is a binary variable. Specifically, $\alpha_{n,r} = 1$ denotes that relay r aids DU n; otherwise, $\alpha_{n,r} = 0$. Furthermore, the drone relays involved in the aid utilize the mixed VLC/RF decode-and-forward protocol with a half duplex mode to transfer data, thus dividing the data transmission into two-hops: (1) DU-TX s transfers data to the corresponding relay r by reusing resource $c_m \in \mathcal{C}$; (2) the drone relay r forwards the data to the corresponding DU-RX d by reusing c_m.

In such a system, we focus on the analysis of the interference caused by resource sharing. On the one hand, the DUs using the VLC resource are exposed to the interference from the other DUs and their corresponding relays operating in VLC. On the other hand, users who share the same RF resource, including one CU, several DUs, and their corresponding relays, interfere with each other. Note that users exploiting different resources are not mutually interfered. All types of interference are sketched in Figure 1. A detailed analysis of the interference in two typical communication modes: VLC-D2D and RF-D2D, will be presented in the following.

3.2. VLC-D2D Communication Mode

In the VLC-D2D communication mode, it is supposed that DU $n = (s, d), n \in \mathcal{N}, s \in \mathcal{S}$, $d \in \mathcal{D}$ utilizing resource c_{M+1} communicates in VLC with the assistance of relay $r \in \mathcal{R}$. According to [45], the VLC channel gain is given as:

$$G_{i,j} = \frac{(k+1)A}{2\pi d_{i,j}^2} cos^k(\phi) g_f g_c(\psi) cos(\psi) \tag{3}$$

where A is the detector area; $d_{i,j}$ denotes the distance between the source i and destination j; ϕ and ψ represent the angle of irradiance and incidence, respectively; $k = \frac{-1}{log_2(cos(\Phi_{1/2}))}$ is the Lambertian order and $\Phi_{1/2}$ is the half-intensity radiation angle; g_f is the gain of the optical filter; and $g_c(\cdot)$ is the optical concentrator gain, which is a function of ψ and is denoted as:

$$g_c(\psi) = \begin{cases} \frac{l^2}{sin^2(\Psi)} & 0 \leq \psi \leq \Psi \\ 0 & \psi > \Psi \end{cases} \tag{4}$$

where l is the refractive index and Ψ is the semi-angle of the field-of-view of the photodiode.

In the 1st-hop VLC-D2D link, the signal to the interference plus noise ratio (SINR) of relay r from DU-TX s is expressed as:

$$\gamma_{s,r}^{c_{M+1}} = \frac{\kappa^2 P_s^{\mathrm{V}} G_{s,r}}{B_{\mathrm{V}} N_{\mathrm{V}} + I_{s,r}^{c_{M+1}}} \tag{5}$$

where P_s^{V} is the transmitted optical power of s; κ is the efficiency of converting the optical signal to an electrical signal; N_{V} is the noise power spectral density in VLC and B_{V} is the bandwidth of VLC; and $I_{s,r}^{c_{M+1}}$ is the interference at r when receiving the signal from s, which comes from other DU-TXs sharing VLC resource c_{M+1}. However, $I_{s,r}^{c_{M+1}}$ is difficult to be expressed exactly because the set of DUs sharing the same resource may be different at different periods. Inspired by [46], we replace the exact form $\gamma_{s,r}^{c_{M+1}}$ with the expected form $\overline{\gamma}_{s,r}^{c_{M+1}}$, which can be approximately shown as:

$$\overline{\gamma}_{s,r}^{c_{M+1}} = \frac{\kappa^2 P_s^{\mathrm{V}} G_{s,r}}{B_{\mathrm{V}} N_{\mathrm{V}} + \mathbf{E}[I_{s,r}^{c_{M+1}}]} \tag{6}$$

where the symbol $\mathbf{E}[\cdot]$ indicates the expectation of $[\cdot]$, which can be formulated as:

$$\mathbf{E}[I_{s,r}^{c_{M+1}}] = \sum_{s' \in A_{M+1} \backslash s} \frac{\kappa^2 P_{s'}^{\mathrm{V}} G_{s',r}}{|A_{M+1}| - 1} \tag{7}$$

where A_{M+1} represents the set of DUs utilizing c_{M+1}, and the operator $|A|$ denotes the cardinality of set A. Since DU n corresponds its DU-TX s one by one, $n \in A$ and $s \in A$ are regarded as the equivalent hereinafter.

In the 2nd-hop VLC-D2D link, the expected SINR of the corresponding DU-RX d from relay r is expressed as:

$$\overline{\gamma}_{r,d}^{c_{M+1}} = \frac{\kappa^2 P_r^{\mathrm{V}} G_{r,d}}{B_{\mathrm{V}} N_{\mathrm{V}} + \mathbf{E}[I_{r,d}^{c_{M+1}}]} \tag{8}$$

where P_r^{V} is the transmitted optical power of r; $\mathbf{E}[I_{r,d}^{c_{M+1}}]$ is the expected interference generated by the relays assisting other DUs who reuse c_{M+1} and is measured as:

$$\mathbf{E}[I_{r,d}^{c_{M+1}}] = \sum_{n' \in A_{M+1} \backslash n} \sum_{r' \in \mathcal{R} \backslash r} \alpha_{n',r'} \frac{\kappa^2 P_{r'}^{\mathrm{V}} G_{r',d}}{|A_{M+1}| - 1}. \tag{9}$$

Due to the average, peak, and non-negative constraints on the modulated optical signals, the classical Shannon equation can not be applied to the VLC. Although the exact capacity of the VLC channel remains unknown, the dual-hop achievable data rate of DU n can be approximated as [47]:

$$R_n^{c_{M+1}}(\alpha) = \frac{1}{2} B_{\mathrm{V}} \log_2(1 + \frac{e}{2\pi} min(\overline{\gamma}_{s,r}^{c_{M+1}}, \overline{\gamma}_{r,d}^{c_{M+1}})). \tag{10}$$

3.3. RF-D2D Communication Mode

In the RF-D2D communication mode, we assume that DU $n = (s, d), n \in \mathcal{N}, s \in \mathcal{S}$, $d \in \mathcal{D}$ is assisted by $r \in \mathcal{R}$ with reusing the RF resource $c_m, m \in \mathcal{M}$. Moreover, we follow the 3GPP recommendation for indoor D2D communication as defined in [48], i.e., the D2D indoor path loss model is formulated as:

$$PL = 16\log_{10}(d_{i,j}) + 89.5. \tag{11}$$

In the 1st-hop RF-D2D link, the SINR of relay r from DU-TX s is shown as:

$$\gamma_{s,r}^{c_m} = \frac{P_s^{\mathrm{R}} H_{s,r}}{B_{\mathrm{R}} N_{\mathrm{R}} + I_{s,r}^{c_m}} \tag{12}$$

where P_s^{R} is the transmitted power of s; N_{R} is the noise power spectral density in RF and B_{R} is the bandwidth of RF; H is the RF channel gain; and $I_{s,r}^{c_m}$ is the sum interference received by r, which contains two parts. The first part is the interference from other DU-TXs sharing c_m, which is denoted as $I_{s,r}^{c_m}(D)$. Similar to the interference $I_{s,r}^{c_{M+1}}$, the interference in this part cannot be accurately described due to the dynamics of resource allocation. The second part is the interference from CU m, denoted as $I_{s,r}^{c_m}(C)$, which is also difficult to calculate exactly. This is because CUs do not always transmit data to the base station (BS) but probabilistically. Consequently, we use the expected form $\overline{\gamma}_{s,r}^{c_m}$ instead of $\gamma_{s,r}^{c_m}$, which is given by:

$$\overline{\gamma}_{s,r}^{c_m} = \frac{P_s^{\mathrm{R}} H_{s,r}}{B_{\mathrm{R}} N_{\mathrm{R}} + \mathbf{E}[I_{s,r}^{c_m}]} \tag{13}$$

where $\mathbf{E}[I_{s,r}^{c_m}]$ denotes the expected sum interference and can be written as:

$$\begin{aligned} \mathbf{E}[I_{s,r}^{c_m}] &= \mathbf{E}[I_{s,r}^{c_m}(D)] + \mathbf{E}[I_{s,r}^{c_m}(C)] \\ &= \sum_{s' \in A_m \setminus s} \frac{P_{s'}^{\mathrm{R}} H_{s',r}}{|A_m| - 1} + \mu_m P_m^{\mathrm{R}} H_{m,r} \end{aligned} \tag{14}$$

where P_m^{R} is the transmitted power of CU m; μ_m represents the communication activity probability of m; and A_m is the set of DUs exploiting c_m.

In the 2nd-hop RF-D2D link, the expected SINR of the corresponding DU-RX d from relay r is indicated as:

$$\overline{\gamma}_{r,d}^{c_m} = \frac{P_r^{\mathrm{R}} H_{r,d}}{B_{\mathrm{R}} N_{\mathrm{R}} + \mathbf{E}[I_{r,d}^{c_m}]} \tag{15}$$

where P_r^{R} is the transmitted power of relay r; $\mathbf{E}[I_{r,d}^{c_m}]$ is the expected sum interference and is calculated as:

$$\mathbf{E}[I_{r,d}^{c_m}] = \sum_{n' \in A_m \setminus n} \sum_{r' \in \mathcal{R} \setminus r} \alpha_{n',r'} \frac{P_{r'}^{\mathrm{R}} H_{r',d}}{|A_m| - 1} + \mu_m P_m^{\mathrm{R}} H_{m,d}. \tag{16}$$

Here, the data rate of DU n can be measured by Shannon's capacity formula:

$$R_n^{c_m}(\alpha) = \frac{1}{2} B_{\mathrm{R}} \log_2(1 + min(\overline{\gamma}_{s,r}^{c_m}, \overline{\gamma}_{r,d}^{c_m})). \tag{17}$$

Similarly, the expected SINR at BS b from CU m in the 1st-hop is shown as:

$$\overline{\gamma}_m^{(1)} = \frac{P_m^{\mathrm{R}} H_{m,b}}{B_{\mathrm{R}} N_R + \sum_{s \in A_m} \frac{P_s^{\mathrm{R}} H_{s,b}}{|A_m|}}. \tag{18}$$

And the expected SINR at BS b from CU m in the 2nd-hop can be given by:

$$\overline{\gamma}_m^{(2)} = \frac{P_m^{\mathrm{R}} H_{m,b}}{B_{\mathrm{R}} N_{\mathrm{R}} + \sum_{n \in A_m} \sum_{r \in \mathcal{R}} \alpha_{n,r} \frac{P_r^{\mathrm{R}} H_{r,b}}{|A_m|}}. \tag{19}$$

Therefore, the data rate of CU m can be calculated as:

$$R_m(\alpha) = \frac{1}{2} B_{\mathrm{R}} \left[\log_2(1 + \overline{\gamma}_{m,b}^{(1)}) + \log_2(1 + \overline{\gamma}_{m,b}^{(2)}) \right]. \tag{20}$$

3.4. Problem Formulation

Our goal is to maximize the D2D system sum rate while ensuring the QoS requirements for the CUs and DUs. Thus, the joint optimization problem of resource allocation and relay selection is formulated as:

$$\max_{\alpha,\beta} \sum_{n=1}^{N} \left[(1 - \beta_n^{c_{M+1}}) \sum_{m=1}^{M} \beta_n^{c_m} R_n^{c_m}(\alpha) + \beta_n^{c_{M+1}} R_n^{c_{M+1}}(\alpha) \right] \tag{21a}$$

$$\text{s.t. } R_m(\alpha) \geq R_{th}^{C}, \forall m \in \mathcal{M} \tag{21b}$$

$$R_n^{c_m}(\alpha) \geq R_{th}^{D}, \forall n \in \mathcal{N}, c_m \in \mathcal{C} \tag{21c}$$

$$\beta_n^{c_m} \in \{0, 1\}, \quad \forall n \in \mathcal{N}, c_m \in \mathcal{C} \tag{21d}$$

$$\sum_{c_m \in \mathcal{C}} \beta_n^{c_m} = 1, \ \forall n \in \mathcal{N} \tag{21e}$$

$$\alpha_{n,r} \in \{0, 1\}, \quad \forall n \in \mathcal{N}, r \in \mathcal{R} \tag{21f}$$

$$\sum_{r \in \mathcal{R}} \alpha_{n,r} = 1, \quad \forall n \in \mathcal{N}, \sum_{n \in \mathcal{N}} \alpha_{n,r} \leq 1, \ \forall r \in \mathcal{R}. \tag{21g}$$

Constraint (21b) guarantees the QoS of the CUs while R_{th}^{C} denotes the rate threshold of the CU link. By analogy, constraint (21c) guarantees the QoS of the DUs while R_{th}^{D} denotes the rate threshold of the D2D link. Constraint (21d) shows that the resource allocation decision $\beta_n^{c_m}$ is a binary variable. Constraint (21e) ensures that each DU only reuses one resource. Constraint (21f) indicates that the relay selection decision $\alpha_{n,r}$ is a binary variable. Constraint (21g) further ensures that each DU only employs one relay and each relay aids at most one DU.

It is clear that the formulated problem possesses a non-convex and combinational structure that makes it intractable to solve in polynomial time. Since both $\alpha_{n,r}$ and $\beta_n^{c_m}$ are 0-1 integer variables, an intuitive method is to enumerate all possible combination policies and find out the optimal resource allocation and relay selection strategy. Nevertheless, the time complexity of the exhaustive method is $\mathcal{O}\left(A_R^N (M+1)^N\right)$, which cannot work out for large-scale scenarios. To address the problem efficiently, we decompose the optimization problem into two subproblems, i.e., resource allocation and relay selection, and tackle them sequentially.

4. Coalitional Game Based Resource Allocation

Since an appropriate resource allocation solution has a large positive impact on the system throughput improvement, we first address the resource allocation problem under random relay selection to approach the maximum throughput quickly. It is worth noting that the relays are randomly selected from the candidate relays, which are described in Section 5.1. In this section, with random relays, a two-phase coalitional game is introduced to solve the resource allocation problem.

4.1. Coalitional Game Formulation

We model the resource allocation problem as a coalitional game. Specifically, each CU forms a coalition representing one RF resource, and an empty coalition is used to represent

the VLC resource. Next, each DU independently and randomly chooses to join a coalition, which means that the DU shares the same resource with other users in the chosen coalition.

In the game, $G = (\mathcal{I}, \mathcal{V}, \mathcal{F})$ with a non-transferable utility is defined. The set of players $\mathcal{I} = \mathcal{M} \cup \mathcal{N}$ consists of both the CUs and DUs. The coalition structure is denoted by the set $\mathcal{F} = \{F_1, F_2, \ldots, F_m, \ldots, F_{M+1}\}$, where F_m is the m-th coalition, and all coalitions are disjoint. That is, we have $F_i \cap F_j = \varnothing$ for any $i \neq j$, and $\cup_{m=1}^{M+1} F_m = \mathcal{I}$. The characteristic function $\mathcal{V}$ denotes the coalition utility, which is expressed as:

$$\mathcal{V}(F_m) = \begin{cases} 0, \ R_m < R_{th}^C, \exists m \in F_m \ or \ R_n^{c_m} < R_{th}^D, \exists n \in F_m \\ \sum_{n \in F_m} R_n^{c_m}, \ otherwise \end{cases}. \tag{22}$$

Given the two coalitions F_i and F_j, if the switch operation of DU n can increase the total throughput of the system, DU n will leave its current coalition F_i and participate in the new coalition F_j. We say that DU n prefers being a member of F_j to F_i, which is denoted by $F_j \rhd F_i$. Thereby, the transfer rule is as:

$$\begin{aligned} F_j \rhd F_i &\iff \mathcal{V}(F_i \setminus \{n\}) + \mathcal{V}(F_j \cup \{n\}) > \mathcal{V}(F_i) + \mathcal{V}(F_j) \\ &\iff \mathcal{CS}(\mathcal{F}') > \mathcal{CS}(\mathcal{F}) \end{aligned} \tag{23}$$

where $\mathcal{CS}(\mathcal{F}) = \sum_{m=1}^{M+1} \mathcal{V}(F_m)$ denotes the sum rate of the current coalition structure $\mathcal{F} = \{F_1, \ldots, F_i, \ldots, F_j, \ldots, F_{M+1}\}$, and $\mathcal{F}' = \{F_1, \ldots, F_i \setminus \{n\}, \ldots, F_j \cup \{n\}, \ldots, F_{M+1}\}$ is the new coalition structure.

According to Equation (23), the D2D system reaches the maximum throughput when all DUs no longer perform switch operations. At this time, the final evolutionary coalition structure $\mathcal{F}_{fin}$ is the solution of the resource allocation problem. More concretely, the different coalitions in $\mathcal{F}_{fin}$ are exactly the optimal sets of DUs for different resources.

4.2. Coalition Formation Algorithm

Based on the coalition structure and transfer rule described above, we need to try our best to satisfy the QoS requirements for all players so that more switch operations can be performed to search for the global optimal solution. Therefore, we construct a two-phase coalitional game composed of the coalition initialization phase and coalition formation phase.

To ensure the QoS of the CUs and DUs, we design the following process for the coalition initialization phase by leveraging the combination of VLC and RF.

(1) Initialize coalitions. In the relay-aided D2D network, the advantage of using the VLC band is more prominent, as it can provide a high data rate. To be specific, VLC signals are strongly attenuated with distance, so the interferences from other DU-TXs and relays operating in the VLC are naturally suppressed. Moreover, VLC signals are closely related to the D2D peer's orientation in terms of irradiance and incidence angles, thus reducing the amount of interferences received. Accordingly, all DUs choose to be members of the coalition F_{M+1}.

(2) Environment sensing. It is obvious that the DUs with a long distance or misaligned orientation are not good candidates for utilizing the VLC resource. In addition, the DUs in close proximity are also unsuitable for reusing the VLC resource due to the heavy interference generated. In this regard, we can filter these DUs that require reallocated resources by observing the data rate received per DU, which intuitively reflects the above environmental factors.

(3) Guarantee the QoS. More concretely, we sort the data rate achieved by each DU in descending order and filter out those with data rates below the threshold R_{th}^D. In other words, these DUs are more appropriate to exploit the RF resources. To this end, a priority sequence $S_n = (n_1, n_2, \ldots, n_k, \ldots, n_M), n_k \in \mathcal{M}$ is designed to guide the switch operation of DU n, where n_k with the smaller subscript k indicates that DU n suffers less interference from CU n_k. For simplicity, the priority order is determined by the distance d_{nm} between the n-th DU and m-th CU. The farther the d_{nm} is, the higher

priority of DU n sharing the resource of CU m is. Note that if CU n_k no longer meets the threshold R_{th}^{C} due to DU n joining the coalition F_{n_k}, then DU n should switch to the next coalition $F_{n_{k+1}}$.

In the traditional coalition formation algorithm [49], a randomly selected DU n performs switch operations in a random order based on a randomly initialized coalition partition. According to the transfer rule in Equation (23), DU n leaves the current coalition F_i and joins the new coalition F_j only when the system profit increases. However, it only refers to the local information and may deviate from the global optimal solution. Furthermore, the existence of users who do not satisfy the QoS demands compromises the coalition utility, which may adversely affect the decision to switch operations. To this end, by allowing DUs to carry out some exploratory operations with a chance probability, we introduce a greedy strategy to search for the global maximum throughput of the D2D system in our coalition formation phase. Considering the convergence rate of the algorithm, the chance probability should decrease gradually with the increase of the number of switch operations. Moreover, it should also depend on the system profit generated by the switch operation. More concretely, it is recommended to reduce the probability of performing the exploratory operation when the system penalty is high, i.e., the system profit is highly negative. In this regard, the chance probability P_c is designed as [50]:

$$P_c(L_t) = \exp\Big(\frac{\mathcal{CS}(\mathcal{F}') - \mathcal{CS}(\mathcal{F})}{L_t}\Big) \tag{24}$$

where $\mathcal{CS}(\mathcal{F}') - \mathcal{CS}(\mathcal{F})$ denotes the system profit, $L_t = \frac{L_0}{log_2(t+1)}$ is the function of the current number of switch operations t, and L_0 is the constant value.

The detailed process of the coalition formation algorithm for resource allocation is shown in Algorithm 1.

4.3. Theoretical Analysis

In this subsection, we provide the theoretical analysis in terms of convergence, stability, and optimability.

Convergence: Starting from any initial coalition structure $\mathcal{F}_{in}$, the proposed coalition formation algorithm will always converge to a final coalition structure $\mathcal{F}_{fin}$.

Proof: For a given number of the CUs and DUs, the total number of the possible coalition structure is finite. As stated before, to improve the D2D system sum rate, each switch operation is allowed to sacrifice the immediate profit with a chance probability. Nevertheless, the probability will approach zero as the number of switch operations increases, denoted by $\lim_{t \to +\infty} P_c(L_t) = 0$, if $\mathcal{CS}(\mathcal{F}') < \mathcal{CS}(\mathcal{F})$. That is, every switch operation will eventually contribute to a higher profit, thus ensuring the convergence to a final coalition structure.

Stability: The final coalition structure of our algorithm $\mathcal{F}_{fin} = \{F_1, F_2 \ldots, F_m, \ldots, F_{M+1}\}$ is Nash-stable, which means that for any $n \in \mathcal{N}, n \in F_m \subset \mathcal{F}_{fin}$, the condition $F_m \rhd F_{m'}$, $\forall F_{m'} \subset \mathcal{F}_{fin}, F_{m'} \neq F_m$ is always satisfied.

Proof: Supposing that $\mathcal{F}_{fin}$ is not Nash-stable, there is at least a $n \in \mathcal{N}, n \in F_m$ and a new coalition $F_{m'}, F_{m'} \subset \mathcal{F}_{fin}, F_{m'} \neq F_m$ that conform to the transfer rule $F_{m'} \rhd F_m$, and then a new coalition structure $\mathcal{F}_{fin'}, \mathcal{F}_{fin'} \neq \mathcal{F}_{fin}$ is formed. This is in contradiction with the premise that $\mathcal{F}_{fin}$ is the final coalition structure. Therefore, the final coalition structure $\mathcal{F}_{fin}$ is Nash-stable.

Optimality: The Nash-stable coalition structure obtained by this algorithm corresponds to the optimal system solution.

Proof: Regarding the renewal of the coalition structure as the evolution of the Markov chain, we can prove that the Markov chain will enter a stationary state with the increase of the number of iterations, so as to guarantee the optimability. The detailed proof can be referred to [50].

Algorithm 1: The Coalition Formation Algorithm for Resource Allocation

Initialization phase:
Initialize coalition structure $\mathcal{F}_{in}$: $F_m = \{m\}, m \in \mathcal{M}, F_{M+1} = \mathcal{N}$;
Collect data rate $R_n^{c_{M+1}}$ of DU n in coalition F_{M+1};
Sort DUs in descending order of $R_n^{c_{M+1}}$, and get the set of DUs with data rate below R_{th}^D, denoted as $\mathcal{N}_{th}$;
for DU $n \in \mathcal{N}_{th}$ **do**
  According to priority sequence S_n, DU n joins coalition $F_{n_k}, k \leftarrow 1$;
  while $R_{n_k} < R_{th}^C$ **do**
    if $k < M$ **then**
      DU n joins coalition $F_{n_k}, k \leftarrow k + 1$;
    else
      DU n joins back coalition F_{M+1};
      break
    end if
  end while
end for
Get the initial coalition structure $\mathcal{F}_{in}$;
Formation phase:
Set the current structure to $\mathcal{F}_{cur} \leftarrow \mathcal{F}_{in}$, $t \leftarrow 0$;
repeat
  Uniformly randomly choose DU $n \in \mathcal{N}$ and denote its coalition as F_i;
  Uniformly randomly choose another coalition $F_j \subset \mathcal{F}_{cur}, F_j \neq F_i$;
  if The switch operation from F_i to F_j satisfying $F_j \rhd F_i$ **then**
    The chosen DU leaves coalition F_i, and joins coalition F_j;
    Update $t \leftarrow t + 1$ and current structure as follows:
    $\mathcal{F}_{cur} \leftarrow \mathcal{F}_{cur} \setminus (F_i \cup F_j) \bigcup (F_i \setminus \{n\}) \bigcup (F_j \cup \{n\})$;
  else
    Draw a random number P uniformly distributed in $(0, 1]$, and calculate the chance probability P_c;
    if $P < P_c$ **then**
      Allow D_n to join F_j, update $t \leftarrow t + 1$ and current structure as:
      $\mathcal{F}_{cur} \leftarrow \mathcal{F}_{cur} \setminus (F_i \cup F_j) \bigcup (F_i \setminus \{n\}) \bigcup (F_j \cup \{n\})$;
    end if
  end if
until The coalition structure converges to the final Nash-stable $\mathcal{F}_{fin}$.

5. MARL-Based Relay Selection

After obtaining the resource allocation solution, we discuss how to select the optimal relay for each DU to further improve the D2D system sum rate. Considering the large number of DUs and relays, it may not be practical to accomplish relay selection with a centralized method due to its high signaling overhead. In this regard, each DU can be considered as an agent and independently selects a relay for assistance, which constitutes a multi-agent system. However, the interferences among some DUs for the large-scale IoT make the relay selection for these DUs co-dependent. That is, a DU needs to consider the relay selection behaviors of other DUs within the same coalition when selecting a relay. In this section, in order to eliminate the co-dependency, we introduce a distributed cooperative MARL-based algorithm, named WoLF-PHC, which encourages the DUs to cooperate for a higher system sum rate.

5.1. Modeling of Multi-Agent Environments

By solving the resource allocation problem in Section 4, N DUs are grouped into $M+1$ coalitions. Note that DUs in different coalitions are not mutually interfered, which implies that the DUs in coalition $F_m \subset \mathcal{F}_{fin}$ do not need to consider the relay selection behaviors of the DUs in other coalitions $F_{m'}, \forall F_{m'} \subset \mathcal{F}_{fin}, F_{m'} \neq F_m$. Hence, without a loss of generality, we focus on the relay selection problem for the DUs in coalition F_m and conduct the modeling analysis hereafter.

In the formulation of the MARL problem, all DUs as agents are independently refining their relay selection policies according to their own observations of the global environment state. Nevertheless, if two or more agents update their policies at the same time, the multi-agent environment appears non-stationary, which violates the Markov hypothesis required for the convergence of RL [51]. Here, we consider modeling the problem as a partially observable Markov game. Formally, the multi-agent Markov game in F_m is formalized by the 5-tuple $< \mathcal{N}^m, \mathcal{S}^m, \mathcal{Z}^{n_m}, \mathcal{A}^{n_m}, R^m >$, where $\mathcal{N}^m \subset \mathcal{N}$ is the set of DUs in F_m, and $|\mathcal{N}^m|$ is the total number of DUs in F_m; $\mathcal{S}^m$ is the global environment state space; $\mathcal{Z}^{n_m}$ is the local observation space for DU $n_m \in \mathcal{N}^m$, determined by the observation function O as $\mathcal{Z}^{n_m} = O(\mathcal{S}^m, n_m)$; $\mathcal{A}^{n_m}$ is the action space for DU n_m; R^m is the immediate reward that is shared by all DUs in F_m to promote cooperative behavior among them. As depicted in Figure 2, at each step t, given the current environment state $\mathcal{S}_t^m$, each DU n_m takes an action $a_t^{n_m}$ from its action space $\mathcal{A}_t^{n_m}$ according to the observation $\mathcal{Z}_t^{n_m}$ and its current policy $\pi(a_t^{n_m}|\mathcal{Z}_t^{n_m})$, forming a joint action $\mathbf{a}_t^m$. Thereafter, the environment generates an immediate reward R_{t+1}^m and evolves to the next state $\mathcal{S}_{t+1}^m$. Then, each DU receives a new observation $\mathcal{Z}_{t+1}^{n_m}$. To be specific, at each step t, for DU n_m, the observation space $\mathcal{Z}_t^{n_m}$, action space $\mathcal{A}^{n_m}$, and reward R_{t+1}^m are defined as follows:

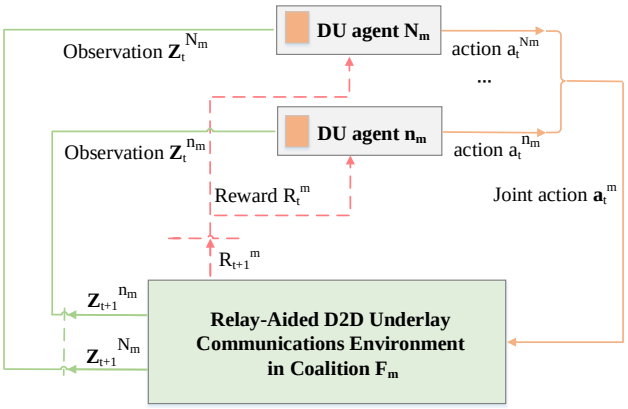

Figure 2. The agent–environment interaction in the MARL formulation of the relay selection in relay-aided networks.

Observation space: The state space observed by n_m can be described as $\mathcal{Z}_t^{n_m} = \mathbf{a}_{t-1}^m$, which includes the historical actions of all DUs in F_m at the previous step. One of the motivations behind this is that if we know the actions taken by all agents, the multi-agent environment becomes stationary [52]. Furthermore, each DU can fully learn to cooperate with other DUs to achieve the global optimal reward in this way.

Action space: The action space of n_m can be described as $\mathcal{A}^{n_m} = [r : \forall r \in \mathcal{R}]$, which represents that the DU can select a relay (The terms *select a relay* and *take an action* will be used interchangeably throughout the paper.) from the set of relays $\mathcal{R}$ for assistance. Accordingly, the dimension of the action space is the total number of relays R. In order to reduce computational complexity, we limit the number of available relays by delineating the area. For DU n_m, let the distance between DU-TX s_m and DU-RX d_m be D_{sd}^m. As shown in Figure 3, we create two circles of radius D_{sd}^m and place s_m and d_m at the center of each circle, thus forming an overlapping area. The relays that are located inside the overlapping area are considered as the candidate relays. Subsequently, $\mathcal{A}^{n_m}$ can be reduced to:

$$\mathcal{A}^{n_m} = [r : D_{sr}^m \leq D_{sd}^m, D_{rd}^m \leq D_{sd}^m, \forall r \in \mathcal{R}] \tag{25}$$

where D_{sr}^m denotes the distance between s_m and r; D_{rd}^m is the distance between r and d_m. Besides, we assume that the candidate relays for each DU do not overlap.

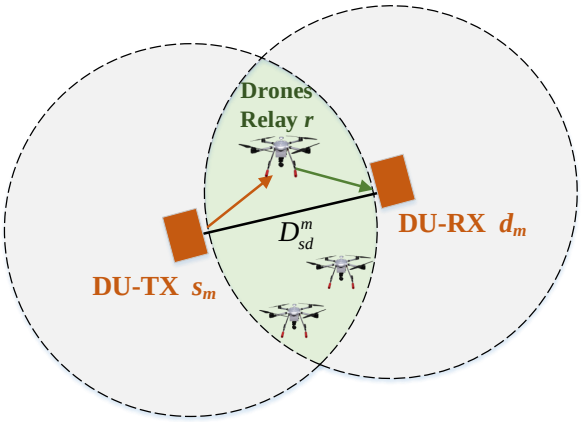

Figure 3. The delineated area of candidate relays.

Reward: To encourage each DU to learn to collaborate with other DUs and thus maximize the D2D system throughput, the DUs in the coalition F_m share a common reward R_{t+1}^m, which is defined as:

$$R_{t+1}^m = \sum_{n_m \in \mathcal{N}^m} R_{n_m}^{c_m}(\alpha_t^m) \tag{26}$$

where $\alpha_t^m \in (\alpha_{n_m,r})_{N_m * R}$ is the decision matrix for relay selection at the current step t in coalition F_m, which depends on the joint action $\mathbf{a}_t^m$. That is, if $a_t^{n_m} = r^*$, then we have $\alpha_{n_m,r^*} = 1$ and $\alpha_{n_m,r} = 0, \forall r \in \mathcal{R}, r \neq r^*$.

5.2. WoLF-PHC

In a multi-agent environment, each agent is part of the other agent's environment, leading to a non-stationary environment. Directly applying a classical single-agent RL (e.g., Q-learning and policy gradient) in the multi-agent case may cause severe oscillations and eventually make the results hard to converge [53]. In contrast, WoLF-PHC, as an extension of Q-learning, adopts the principle of fast learning when losing and slow learning when winning, which allows agents to learn moving targets with both guaranteed rationality and convergence [54]. Hence, we apply the WoLF-PHC to enable the DUs to learn their own relay selection decisions in a multi-agent system.

In the WoLF-PHC, each DU continuously interacts with the environment and other DUs in the same coalition to update the Q-value. To simplify the representation, for DU $n_m \in F_m$, the local observation $\mathcal{Z}_t^{n_m}$, action $a_t^{n_m}$, and action space $\mathcal{A}_t^{n_m}$ at the current step t are simply denoted as $\mathcal{Z}$, a and $\mathcal{A}$, respectively; the reward received R_{t+1}^m, new observation $\mathcal{Z}_{t+1}^{n_m}$, and action $a_{t+1}^{n_m}$ at the next step are denoted as R', $\mathcal{Z}'$ and a', respectively. Let $Q(\mathcal{Z}, a)$ be the estimated Q-value with action a in state $\mathcal{Z}$ during the learning process. As with the Q-learning algorithm, the update rule of the Q-value can be expressed as:

$$Q(\mathcal{Z}, a) \leftarrow Q(\mathcal{Z}, a) + \delta[R' + \beta \max_{a' \in \mathcal{A}} Q(\mathcal{Z}', a') - Q(\mathcal{Z}, a)] \tag{27}$$

where $\delta \in (0, 1]$ represents the learning rate, and $\beta \in (0, 1]$ is the discount factor.

To learn the optimal Q-value, the DU updates its own relay selection policy $\pi(\mathcal{Z}, a)$ that describes the probability of taking action a in state $\mathcal{Z}$. As a generalization of the widely used greedy algorithm, the policy hill-climbing (PHC) algorithm increases the probability of taking the highest valued action while it decreases the probability of other actions according to the learning parameter θ [55]. Moreover, the policy should be restricted to a legal probability distribution. Thus, the updated rule of policy $\pi(\mathcal{Z}, a)$ can be calculated as:

$$\pi(\mathcal{Z}, a) \leftarrow \pi(\mathcal{Z}, a) + \Delta_{\mathcal{Z},a} \tag{28}$$

where

$$\Delta_{\mathcal{Z},a} = \begin{cases} -\min(\pi(\mathcal{Z},a), \frac{\theta}{M-1}),\ a \neq \arg\max\limits_{\acute{a}\in\mathcal{A}} Q(\mathcal{Z},\acute{a}) \\ \sum_{\acute{a}\neq a} \min(\pi(\mathcal{Z},\acute{a}), \frac{\theta}{M-1}),\ otherwise \end{cases} \tag{29}$$

where M is a constant coefficient.

In essence, the key contribution of the WoLF-PHC is the variable learning parameter θ consisting of two parameters: θ^w and θ^l, with $\theta^w < \theta^l$. They are employed to update the policy, which depends upon whether the current policy $\pi(\mathcal{Z},a)$ is winning or losing. To this end, the average policy denoted as $\bar{\pi}(\mathcal{Z},a)$ is introduced to judge the win–lose of the current policy and can be formulated as:

$$\bar{\pi}(\mathcal{Z},a) \leftarrow \bar{\pi}(\mathcal{Z},a) + \frac{\pi(\mathcal{Z},a) - \bar{\pi}(\mathcal{Z},a)}{C(\mathcal{Z})} \tag{30}$$

where $C(\mathcal{Z})$ represents the number of occurrences of the state $\mathcal{Z}$ observed by the DU, which is updated by:

$$C(\mathcal{Z}) \leftarrow C(\mathcal{Z}) + 1. \tag{31}$$

By comparing the expected payoff of the current policy with that of the average policy over time, the DU can choose its appropriate learning parameter θ from θ^w and θ^l. If the expected value of the current policy is larger, θ^w is applied to update the policy cautiously; otherwise, θ^l is utilized to learn quickly. Accordingly, the learning parameter θ can be described as:

$$\theta = \begin{cases} \theta^w,\ \sum_{a\in\mathcal{A}} \pi(\mathcal{Z},a)Q(\mathcal{Z},a) > \sum_{a\in\mathcal{A}} \bar{\pi}(\mathcal{Z},a)Q(\mathcal{Z},a) \\ \theta^l,\ otherwise \end{cases} \tag{32}$$

The detailed process of the WoLF-PHC algorithm for relay selection is given in Algorithm 2.

Algorithm 2: The WoLF-PHC Algorithm for Relay Selection

Set δ, β, θ^w, θ^l;
for each coalition $F_m, m \in \mathcal{M} \cup \{M+1\}$ **do**
  Initialize for each DU $n_m \in F_m$:
  $Q(\mathcal{Z},a) \leftarrow 0$, $\pi(\mathcal{Z},a) \leftarrow \frac{1}{|A|}$, $C(\mathcal{Z}) \leftarrow 0$;
  for each step t **do**
    for each DU n_m **do**
      Receive current local observation $\mathcal{Z}$ and update $C(\mathcal{Z})$ by using (31);
      Select relay a at random with probability policy $\pi(\mathcal{Z},a)$;
    end for
    All DUs take actions and receive immediate reward R';
    for each DU n_m **do**
      Receive next observation $\mathcal{Z}'$;
      Update Q-value $Q(\mathcal{Z},a)$ as well as Q-table by using (27);
      Update average policy $\bar{\pi}(\mathcal{Z},a)$ by using (30);
      Update relay selection policy $\pi(\mathcal{Z},a)$ by using (28), (29) and (32);
      Update observation $\mathcal{Z} \leftarrow \mathcal{Z}'$;
    end for
  end for
  for each DU n_m **do**
    Find the optimal relay $r^* = \arg\max Q$, and set $\alpha_{n_m,r^*} = 1$;
  end for
end for
Output the optimal decision matrix: $\alpha^* \in (\alpha_{n,r^*}), \forall n \in \mathcal{N}$.

5.3. Neighbor–Agent-Based WoLF-PHC

In the WoLF-PHC, we define the observation space of agent n_m as the past joint action of all agents within coalition F_m, so as to guarantee the stability of the multi-agent environment. Before reselecting relays, when the number of the DUs and resources are 10 and four, we visualize the geographic location of all the DUs and the result of the resource allocation, as shown in Figure 4, where different colors represent different resources. It can be seen that the closer DUs use different resources, while the more distant DUs share the same resource. In other words, the DUs in a coalition are far apart from each other. In the case of limited range D2D communication, the interference between any candidate relay of DU n_m and a remote DU n_m^r can be considered the same and negligible. Similarly, the interference between any candidate relay of n_m^r and n_m can be considered the same. Thus, the relay selection decisions of n_m and n_m^r are independent of each other. That is, it is not necessary to have all agents' historical actions to ensure stability; only the actions of neighboring agents is enough. Accordingly, we propose a lightweight algorithm that allows the target agent to observe the actions of a fixed number of neighboring agents, named neighbor–agent-based WoLF-PHC.

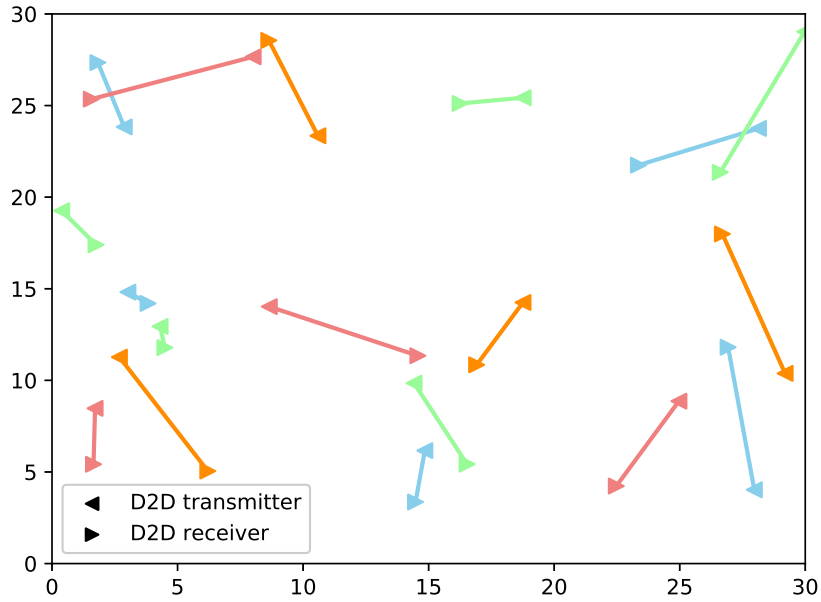

Figure 4. DUs geographic location and resource allocation results visualization.

In the neighbor–agent-based WoLF-PHC, for the target agent n_m, we define the nearest λ agents to target n_m within a coalition as the neighboring agents of n_m. Moreover, the observation space $\mathcal{Z}^{n_m}$ is changed from the joint action $\mathbf{a}_{t-1}^{m}$ to the joint action of neighbors $\mathbf{a}_{t-1}^{m,nb}$, where $\mathbf{a}_{t-1}^{m,nb} = \{a_{t-1}^{i}, i \in \mathcal{N}_{n_m}^{nb}\}$ comprises the actions of $\lambda + 1$ agents in the neighboring set $\mathcal{N}_{n_m}^{nb} \subset \mathcal{N}^m$, which incorporates n_m itself and its neighboring agents. Note that if $|\mathcal{N}_{n_m}^{nb}| = |\mathcal{N}^m|$, the neighbor–agent-based WoLF-PHC is the same as the WoLF-PHC.

6. Complexity and Signaling Overhead Analysis

6.1. Complexity Analysis

The complexity of the proposed joint resource allocation and relay selection algorithm can be analyzed from the following two parts.

One part of the complexity comes from the resource allocation scheme based on the coalitional game. The computational complexity of the resource allocation scheme is $\mathcal{O}(I_{in})$, where I_{in} is the number of inner iterations required to converge to the final coalition structure.

Another part of the complexity arises from the relay selection scheme based on the WoLF-PHC or neighbor–agent-based WoLF-PHC. For each agent $n \in \mathcal{N}$, the computational complexity is calculated as $\mathcal{O}(|\mathcal{Z}^n|^2|\mathcal{A}^n|)$, where $|Z^n| < N$ is the observation space size of n, and $|\mathcal{A}^n| < R$ is the action space size of n. As for the WoLF-PHC, the overall complexity is, at most, $\mathcal{O}(N(Z^*)^2\mathcal{A}^*)$, where $Z^* = \max_{n\in\mathcal{N}} |Z^n|$ denotes the largest size of the observation space, and $A^* = \max_{n\in\mathcal{N}} |A^n|$ is the largest size of the action space. As for the neighbor–agent-based WoLF-PHC, the overall complexity is at most $\mathcal{O}(N(\lambda+1)^2\mathcal{A}^*)$, where the setting parameter λ is much less than Z^* in general.

Therefore, the total complexity of our proposed algorithms is $\mathcal{O}(I_{in}N(\mathcal{Z}^*)^2\mathcal{A}^*)$ or $\mathcal{O}(I_{in}N(\lambda+1)^2\mathcal{A}^*)$. To obtain the global optimal solution, apart from solving subproblems sequentially, an ideal algorithm usually requires multiple outer iterations until the D2D sum rate no longer rises. As a result, the complexity of the ideally proposed algorithms (IPA) is $\mathcal{O}([I_{in}N(\mathcal{Z}^*)^2\mathcal{A}^*]^{I_{ou}})$ or $\mathcal{O}([I_{in}N(\lambda+1)^2\mathcal{A}^*]^{I_{ou}})$, where I_{ou} is the number of outer iterations.

However, the relays reselected by any agent come from its corresponding delineated area, i.e., the candidate relays are close to each other, which leads to less impact of reselecting relays on the resource allocation solution. In this way, the performance of our proposed algorithms with lower complexity is considered to be approximate that of IPA. Hence, it is more suitable to apply our proposed algorithms rather than IPA to large-scale scenarios.

6.2. Signaling Overhead Analysis

The signaling overhead of our proposed algorithm should also be analyzed in two parts.

On the one hand, since the resource allocation mechanism is implemented in a centralized manner, the signaling overhead mainly comes from the process of acquiring CSI, which can be classified into transmission and interference CSI. Concretely, in the relay-aided D2D network, the transmission CSI includes the links from CUs to the BS, from DU-TXs to the corresponding relays, and from these relays to DU-RXs; the interference CSI includes the links from CUs to the relays and DU-RXs, and from DU-TXs to the BS. When the number of CUs, DUs and relays are M, N and R, respectively, we can conclude that the signaling overhead for the CSI measurement in a centralized manner is $\mathcal{O}(2NR + MR + 2N + 2M)$ by using the evaluation method in [56]. In contrast, the signaling overhead for CSI measurements can be reduced to $\mathcal{O}(2NR)$ in a distributed manner, which usually comes at the expense of the global system performance. Note that the number of R is generally assumed to be larger than that of N and M, so as to ensure the reliability of relay-aided D2D communication. Thus, the difference in signaling overhead between these two manners is not significant.

On the other hand, the distributed relay selection mechanism is performed independently in each coalition without exchanging information among coalitions, which greatly reduces the signaling overhead. However, for the DUs within any coalition, in order to encourage the DUs to achieve the global optimal reward in a collaborative way, each DU needs to upload its own historical information to the BS, including the actions taken and data rate obtained. Then, the BS broadcasts the actions of all DUs within a coalition along with a common reward. All the above information exchanged between the DUs and BS are numerical data with a size of only a few kilobytes, which leads to a small signaling overhead. Consequently, this part of the overhead is negligible compared to that incurred by the former centralized resource allocation mechanism.

In summary, the signaling overhead of our proposed algorithm approximates $\mathcal{O}(2NR + MR + 2N + 2M)$.

7. Numerical Results

In this section, we present numerical results to evaluate our proposed algorithm. In our simulation, we consider a 30 m $\times$ 30 m room in which CUs utilize RF resources for uplink communication, and relay-assisted DUs want to implement the applications that require high rate communication; the DUs can choose either the VLC-D2D or RF-D2D

communication mode. Furthermore, the distance between the transmitter and receiver of each DU is uniformly distributed and the upper bound is 10 m, which makes cooperation gain obtained by the combination of VLC and RF the most notable [7]. Moreover, the idle relays available are evenly distributed and the number is fixed at 50 [57]. To model the realistic VLC-D2D communication channel, we assume that the irradiance and incidence angle follow a Gaussian distribution with a mean value of 0° and a standard deviation of 30° [7]. We repeat the simulations 200 times independently and average the results, thus mitigating the randomness of the above parameters. Considering the QoS requirements, the minimum rate thresholds of the CUs and DUs are set to 10 Mbps and 20 Mbps, respectively. Additional detailed simulation parameters can be seen in Table 1.

Table 1. Simulation Parameters.

General Parameters	
Communication activity probability of CUs, μ_m	0.7
RF Parameters	
System bandwidth in RF, B_{R}	20 MHz
Noise Power Spectral Density in RF, N_{R}	-174 dBm/Hz
Transmitted power of CUs/DUs/relays, $P_m^{\mathrm{R}}/P_s^{\mathrm{R}}/P_r^{\mathrm{R}}$	400/200/200 mW
VLC Parameters	
Physical area of photodiode, A	10^{-4} m^2
The gain of the optical filter, g_f	1
Refractive index, l	1.5
Half-intensity radiation angle, $\Phi_{1/2}$	60°
Field-of-view of photodiode, Ψ	60°
O/E conversion efficiency, κ	0.53 A/W
System bandwidth in VLC, B_{V}	20 MHz
Noise Power Spectral Density in VLC, N_{V}	10^{-21}A^2/Hz
Transmitted optical power of DUs/relays, $P_s^{\mathrm{V}}/P_r^{\mathrm{V}}$	200/200 mW
WoLF-PHC Parameters	
Learning rate, δ	0.2
Discount factor, β	0.8
Learning parameter (win), θ^w	0.4
Learning parameter (lose), θ^l	0.8

7.1. Performance Analysis of PCG-Based Resource Allocation

At first, by comparing with the exhaustive algorithm (EA), we further demonstrate the optimability of the proposed coalitional game (PCG)-based resource allocation in practice. Meanwhile, we give the performance comparison between the proposed joint resource allocation and relay selection algorithm, namely PCG-WP, and the corresponding IPA. In this case, we present the D2D system sum rate comparison under the above algorithms by varying the number of CUs and DUs. Given the high complexity of EA and IPA, we fix the number of DUs at eight in Figure 5, and fix the number of CUs at two in Figure 6. From these two figures, on the one hand, we can observe that the sum rate of the D2D system achieved by PCG is almost close to that implemented by EA, which demonstrate that our proposed PCG can achieve a sum rate close to EA, but with a lower complexity. On the other hand, the sum rate gap between PCG-WP and IPA is insignificant. Concretely, the sum rate of IPA is at most 10% larger than that of PCG-WP.

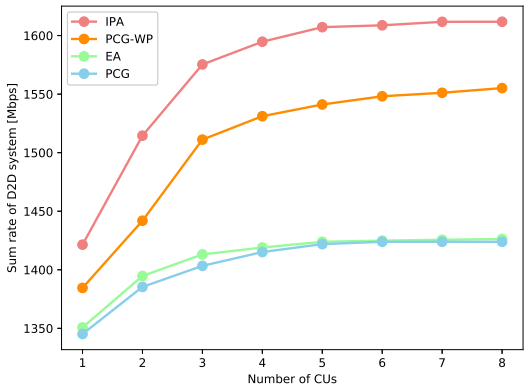

Figure 5. Sum rate under EA/PCG and IPA/PCG-WP vs. number of CUs.

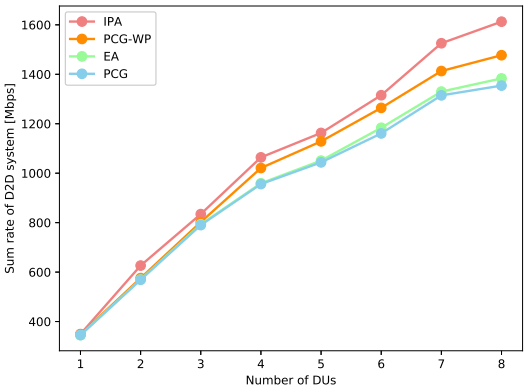

Figure 6. Sum rate under EA/PCG and IPA/PCG-WP vs. number of DUs.

7.2. Performance Analysis of WoLF-PHC for Relays Selection

Then, based on the final coalition structure obtained by PCG, we employ RL algorithms to reselect relays for the DUs in each coalition. In this simulation, we use Q-learning (QL) as a comparative algorithm to evaluate the convergence performance of our proposed WoLF-PHC (WP). In addition, we also show the convergence performance of the algorithms only exploiting local information, including the neighbor–agent-based WoLF-PHC (NWP) and the neighbor–agent-based Q-learning (NQL). For the sake of simplicity, we define the NWP working with λ neighboring users as NλWP, and the same goes for NλQL.

Figure 7 compares the convergence of the above four approaches in terms of the total reward performance when the number of DUs is 10 and the number of CUs is one. The total reward is the sum of the rewards received by all coalitions. From Figure 7, the proposed WP converges to the maximum total reward of about 1150 at nearly 2700 steps, while the N3WP converges to the close-to-maximum reward of about 1070 at a faster convergence rate of around 1500 steps. Therefore, the use of N3WP increases the convergence speed by approximately 44.4% in the case of a total reward loss of 6.9%. By contrast, both the QL and N3QL fail to converge and exhibit poor performance, despite the N3QL seeming to be more stable (less fluctuations) than the QL. On the one hand, capitalizing on the "wining or learning fast" mechanism, the WP-based approaches present a much better convergence performance than the QL-based approaches. On the other hand, the approaches that utilize local information (N3WP and N3QL) can greatly reduce the state space, thereby accelerating the convergence speed but sacrificing the tolerable performance, while the complexity of IPA grows exponentially. This result further confirms the feasibility of replacing IPA with PCG-WP.

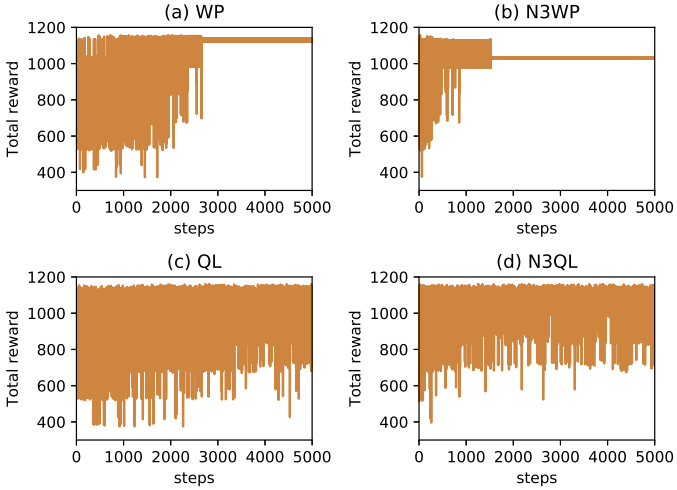

Figure 7. Convergence performance of QL and WP based algorithms.

7.3. Performance Comparison

Next, we compare the two proposed schemes, PCG-WP and PCG-N3WP, with the following five baseline schemes:

(1) Random algorithm (RA). In this scheme, each D2D pair assisted by a randomly selected relay can randomly use either the VLC resource or RF resource of any CU.
(2) PCG-based resource allocation and random relay selection (PCG-RD). For investigating the potential gain of the joint optimizing of resource allocation and relay selection, the PCG-RD that optimizes only the resource allocation is considered as a comparative algorithm.
(3) Random resource allocation and WP-based relay selection (RD-WP). Similar to the PCG-RD above, the RD-WP that optimizes only the relay selection is regarded as a comparative algorithm to analyze the joint optimization gain.
(4) Traditional coalitional game [49] and WP-based relay selection (TCG-WP). In this scheme, the resource allocation problem is addressed by the traditional coalitional game with random initialization and formation, and the WP method is used for relay selection.
(5) Best response dynamics (BRD) in [58]. Compared with our proposed cooperative scheme, each DU in this scheme is selfish and aims at maximizing its own throughput performance. In both the resource allocation and relay selection stage, every DU simultaneously optimizes its actions with respect to the action profile, which is composed of the actions played by the other DUs in the same coalition at the previous time.

In Figure 8, we evaluate the impact of the number of CUs on the D2D system sum rate under different schemes. Here, the number of DUs is enlarged to 14 and the number of CUs varies from one to eight. As the number of CUs increases, the performance of both the RA and RD-WP declines slightly and then levels off, although that of the RA exhibits slight fluctuations on the curve due to randomness. The performance degradation is due to the short distance (up to 10 m) between the transceivers of each D2D pair, which makes the VLC superior to the RF. When the number of CUs equals one, the probability of randomly selecting VLC resources for every DU is up to 50%, so the sum rate of both RA and RD-WP reaches the maximum. However, the performance of the remaining schemes improves with the increase in the number of CUs, thanks to the rational resource allocation. Among them, the BRD with the selfish nature exhibits the worst performance, while the cooperative PCG-WP obtains the best one. This is because each DU in BRD optimizes its own profit, regardless of the interference introduced to other DUs. When three CUs are involved, the sum rate of PCG-WP is larger than that of PCG-N3WP, TCG-WP, and BRD of about 5.2%, 13.3%, and 27.2%, respectively. As the number of CUs increases further, which implies that the number of DUs within a coalition decreases, PCG-N3WP becomes

enough to characterize global information and thus achieve almost the same throughput as PCG-WP. Meanwhile, the sum rate gap between PCG-WP and TCG-WP is gradually narrowing. This is due to the fact that the switch operations in TCG-WP are no longer limited by QoS constraints in the case of adequate resources. In addition, when the number of CUs is five, PCG-WP outperforms PCG-RD and RD-WP by about 19.0% and 44.5%, which highlights the gain of joint optimization.

Moreover, in Figure 9, we focus on the system performance in terms of the outage probability, which is calculated as the ratio of users who do not meet the QoS demands to the total system users. As can be seen, the outage probability declines sharply as the number of CUs increases. The underlying reason is that more CUs will naturally contribute to a lower interference. When the number of CUs equals one, BRD shows the worst-case due to the ping-pong effect between the VLC resource and RF resource of the CU. As the number of CUs grows, however, its performance surpasses that of RD because the probability of the ping-pong effect decreases. In combination with Figure 8, it can be noticed that BRD outperforms RD-WP in terms of the sum rate performance, while its outage performance is slightly worse than that of RD-WP. This is due to the selfish nature of BRD, i.e., improving the rate of some DUs at the expense of others. More importantly, PCG-N3WP achieves almost the same and lowest outage probability as PCG-WP. Note that TCG-WP initially exhibits a poor performance, and its performance exceeds that of our proposed PCG-WP and PCG-N3WP when the number of CUs is larger than seven. It makes sense that when resources are sufficient, an affordable individual DU performance can be sacrificed for the sake of the overall system performance in our schemes.

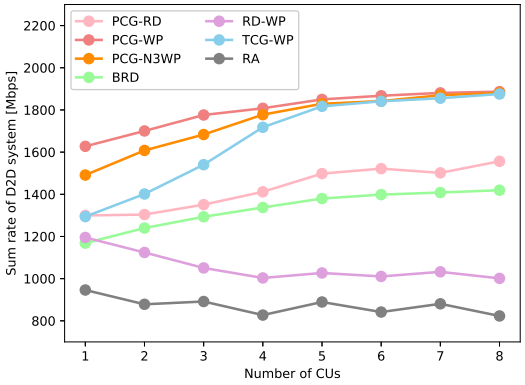

Figure 8. Sum rate of different methods vs. number of CUs.

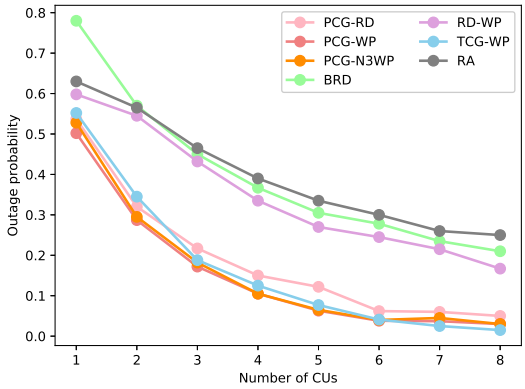

Figure 9. Outage probability of different methods vs. number of CUs.

Figure 10 depicts the comparison of the D2D system sum rate for different mechanisms in the resource-lacking system by fixing the number of CUs at two and varying the number of DUs from four to 18. It is shown that the increase in the number of DUs can boost the total throughput, and PCG-WP always achieves the highest total throughput. Moreover,

with the aggravation of traffic congestion, the gap between PCG-WP and other competitive schemes is growing. When 18 DUs are involved, PCG-WP results in a 129.2% higher total throughput than the baseline RA. Another observation is that when the number of DUs is larger than 12, the performance of all schemes except the proposed PCG-WP and PCG-N3WP shows little improvement. This can be inferred that without the effective joint gain of the resource allocation and relay selection, the gain from increasing the number of DUs alone no longer compensates for the loss from the resulting severe interference. Concretely, in the context of insufficient resources, PCG has more prominent advantages over TCG in finding the optimal solutions. The reason is that the QoS requirements of users restrict TCG to perform switch operations, which leads to deviation from the optimal solution. While PCG satisfies the QoS demands as much as possible in the initialization stage, the greedy policy further allows the system to explore more operations in the formation stage, so as to bring the sum rate performance enhancement. The last observation is that when the number of DU increases, the advantage of exploiting global information for relay selection becomes obvious.

In Figure 11, we can observe that the outage probability goes up as the number of DUs increases because of the fierce competition for resources and relays. In contrast to Figure 9, the performance of BRD is slightly better than that of RD-WP, which suggests that an efficient resource allocation scheme may be more important than an appropriate relay selection scheme in the resource-scarce environment, and vice versa. In addition, increasing the number of DUs makes the gap between PCG-WP and other algorithms become notable.

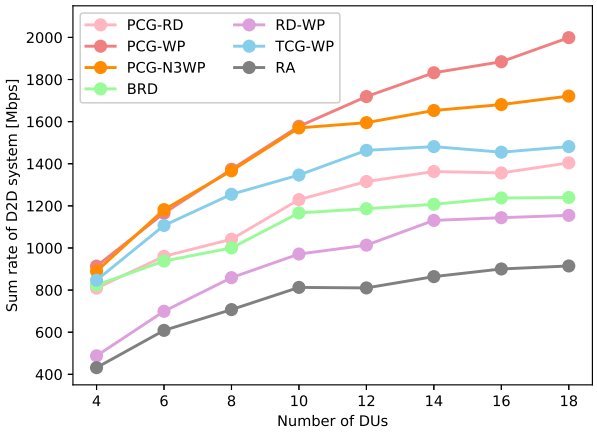

Figure 10. Sum rate of different methods vs. number of DUs.

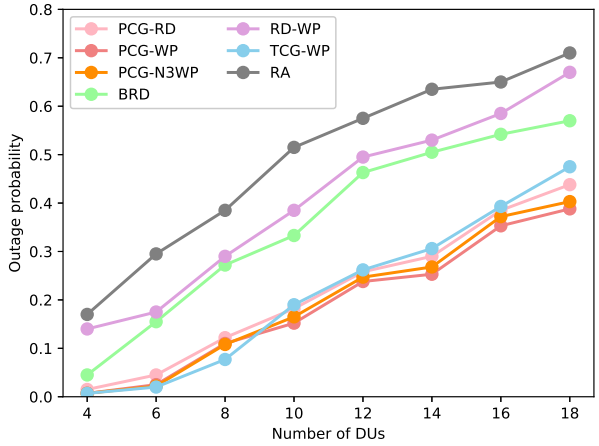

Figure 11. Outage probability of different methods vs. number of DUs.

Moreover, we study the impact of the number of neighboring users λ on the system performance in terms of the sum rate of the D2D system and convergence rate. The

convergence rate is indicated by the reciprocal of the number of iterations to converge. In Figure 12, the number of CUs remains as two and the number of DUs equals 18. As expected, if λ decreases, the sum rate decreases as well, while the convergence rate increases greatly. More specifically, the decrease in λ from seven to three decreases the sum rate by 10.3%, and also decreases the number of iterations to converge by 82.9%. Obviously, PCG-NWP trades a smaller throughput loss for a significantly faster convergence rate. In this regard, users can make a trade-off between throughput and convergence performance according to preferences and practical constraints.

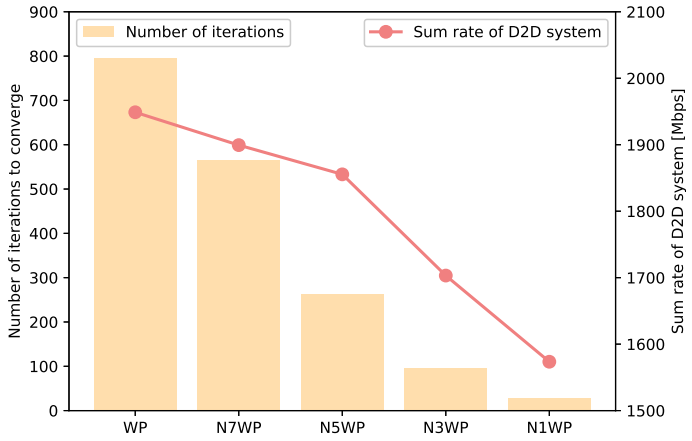

Figure 12. System performance for different number of neighboring users λ.

7.4. Summary of Main Results

In order to present the results of this article more clearly, this section summarizes the main conclusions as follows:

(1) In the stage of Resource Allocation, our proposed PCG can achieve a sum rate close to EA, but with a lower complexity.
(2) Compared with WoLF-PHC (WP), the neighbor–agent-based WoLF-PHC (N3WP) increases the convergence speed by approximately 44.4% in the case of a total reward loss of 6.9%.
(3) Our proposed WP presents a much better convergence performance than the QL-based approaches.
(4) The approaches that utilize local information (N3WP and N3QL) can greatly reduce the state space, thereby accelerating the convergence speed.
(5) Just randomly optimizing the Resource Allocation or Relays Selection policy cannot make the overall performance maximization. Appropriate methods applied to joint optimization are indispensable.
(6) In the resource-lacking system, our proposed WP or NWP shows greater advantages.

8. Conclusions

In this paper, we proposed an efficient joint resource allocation and drone relay selection algorithm with a low complexity and signaling overhead for large-scale IoT. With randomly selected relays from a delineated area, the two-phase coalitional game-based algorithm was proposed to solve the resource allocation problem. Then, the WoLF-PHC based algorithm was proposed to solve the relay selection problem. Meanwhile, the lightweight neighbor–agent-based WoLF-PHC was introduced to further reduce the complexity. Simulation results demonstrated that our algorithms outperformed the considered benchmarks, especially in traffic congestion scenarios. Moreover, the appropriate number of neighboring users can be chosen based on preferences and practical constraints when applying our relay selection algorithm.

Author Contributions: Conceptualization, data curation, investigation, methodology, resources, and software, X.L. and S.H.; formal analysis and supervision, K.Z., S.M. and G.C.; visualization and writing—original draft, W.G. and X.C.; validation and writing—review and editing, G.C., X.L., P.Z. and Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by the National Key Research and Development Project of China (Grant No. 2020YFB1806703), by the Fundamental Research Funds for the Central Universities (Grant No. 3282023010), and by the Key Projects of Kashgar University (Grant No. GCC2023ZK-004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

D2D	Device-to-Device
VLC	Visible Light Communication
IoT	Internet of Things
QoS	Quality of Service
WoLF-PHC	WoLF policy hill-climbing
MARL	Multi-agent Reinforcement Learning
UAVs	Unmanned Aerial Vehicles
VLC/RF	Visible Light Communication/Radio Frequency
RL	Reinforcement Learning
EA	Exhaustive Algorithm
PCG	Proposed Coalitional Game
RA	Random Algorithm
PCG-RD	PCG based Resource Allocation and Random Relay Selection
RD-WP	Random Resource Allocation and WP based Relay Selection
TCG-WP	Traditional Coalitional Game and WP based Relay Selection
BRD	Best Response Dynamics

References

1. Chen, X.; Ng, D.W.K.; Yu, W.; Larsson, E.G.; Al-Dhahir, N.; Schober, R. Massive Access for 5G and Beyond. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 615–637. [CrossRef]
2. Tehrani, M.N.; Uysal, M.; Yanikomeroglu, H. Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions. *IEEE Commun. Mag.* **2014**, *52*, 86–92. [CrossRef]
3. Bello, O.; Zeadally, S. Intelligent Device-to-Device Communication in the Internet of Things. *IEEE Syst. J.* **2016**, *10*, 1172–1182. [CrossRef]
4. Jameel, F.; Hamid, Z.; Jabeen, F.; Zeadally, S.; Javed, M.A. A Survey of Device-to-Device Communications: Research Issues and Challenges. *IEEE Commun. Surv. Tutor.*,**2019**, *20*, 2133–2168. [CrossRef]
5. Chen, G.; Tang, J.; Coon, J.P. Optimal Routing for Multihop Social-Based D2D Communications in the Internet of Things. *IEEE Internet Things J.* **2018**, *5*, 1880–1889. [CrossRef]
6. Demirkol, I.; Camps-Mur, D.; Paradells, J.; Combalia, M.; Popoola, W.; Haas, H. Powering the Internet of Things through Light Communication. *IEEE Commun. Mag.* **2019**, *57*, 107–113. [CrossRef]
7. Mach, P.; Becvar, Z.; Najla, M.; Zvanovec, S. Combination of visible light and radio frequency bands for device-to-device communication. In Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio (PIMRC), Montreal, QC, Canada, 8–13 October 2017; pp. 1–7.
8. Becvar, Z.; Najla, M.; Mach, P. Selection between Radio Frequency and Visible Light Communication Bands for D2D. In Proceedings of the 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal, 3–6 June 2018; pp. 1–7.
9. Najla, M.; Mach, P.; Becvar, Z. Deep Learning for Selection Between RF and VLC Bands in Device-to-Device Communication. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 1763–1767. [CrossRef]
10. Salim, M.M.; Wang, D.; Elsayed, H.A.E.A.; Liu, Y.; Elaziz, M.A. Joint Optimization of Energy-Harvesting-Powered Two-Way Relaying D2D Communication for IoT: A Rate–Energy Efficiency Tradeoff. *IEEE Internet Things J.* **2020**, *7*, 11735–11752. [CrossRef]
11. Zhang, C.; Ge, J.; Gong, F.; Jia, F.; Guo, N. Security–Reliability Tradeoff for Untrusted and Selfish Relay-Assisted D2D Communications in Heterogeneous Cellular Networks for IoT. *IEEE Syst. J.* **2020**, *14*, 2192–2201. [CrossRef]

12. Chai, J.; Feng, L.; Zhou, F.; Zhao, P.; Yu, P.; Li, W. Energy-Efficient Resource Allocation Based on Hypergraph 3D Matching for D2D-Assisted mMTC Networks. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–7.
13. Huang, H.; Hu, S.; Yang, T.; Yuan, C. Full-Duplex Nonorthogonal Multiple Access With Layers-Based Optimized Mobile Relays Subsets Algorithm in B5G/6G Ubiquitous Networks. *IEEE Internet Things J.* **2021**, *8*, 15081–15095. [CrossRef]
14. Ademaj, F.; Rzymowski, M.; Bernhard, H.-P.; Nyka, K.; Kulas, L. Relay-Aided Wireless Sensor Network Discovery Algorithm for Dense Industrial IoT Utilizing ESPAR Antennas. *IEEE Internet Things J.* **2021**, *8*, 16653–16665. [CrossRef]
15. Hou, X.; Wang, J.; Jiang, C.; Zhang, X.; Ren, Y.; Debbah, M. UAV-Enabled Covert Federated Learning. *IEEE Trans. Wirel. Commun.* **2023**. [CrossRef]
16. Wang, J.; Jiang, C.; Wei, Z.; Pan, C.; Zhang, H.; Ren, Y. Joint UAV Hovering Altitude and Power Control for Space-Air-Ground IoT Networks. *IEEE Internet Things J.* **2018**, *6*, 1741–1753. [CrossRef]
17. Feng, R.; Li, Z.; Wang, Q.; Huang, J. An ADMM-Based Optimization Method for URLLC-Enabled UAV Relay System. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 1123–1127. [CrossRef]
18. Zhu, Z.; Yang, Y.; Guo, C.; Chen, M.; Cui, S.; Poor, H.V. Power Efficient Deployment of VLC-enabled UAVs. In Proceedings of the 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, UK, 31 August–3 September 2020; pp. 1–6. [CrossRef]
19. Wang, Y.; Chen, M.; Yang, Z.; Luo, T.; Saad, W. Deep Learning for Optimal Deployment of UAVs With Visible Light Communications. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 7049–7063. [CrossRef]
20. Abdel-Rahman, M.J.; AlWaqfi, A.M.; Atoum, J.K.; Yaseen, M.A.; MacKenzie, A.B. A Novel Multi-Objective Sequential Resource Allocation Optimization for UAV-Assisted VLC. *IEEE Trans. Veh. Technol.* **2023**, *72*, 6896–6901. [CrossRef]
21. Maleki, M.R.; Mili, M.R.; Javan, M.R.; Mokari, N.; Jorswieck, E.A. Multi-Agent Reinforcement Learning Trajectory Design and Two-Stage Resource Management in CoMP UAV VLC Networks. *IEEE Trans. Commun.* **2022**, *70*, 7464–7476. [CrossRef]
22. Yu, L.; Liu, C.; Qian, J.; Wang, Y.; Wang, Z. Performance Analysis of Unmanned Aerial Vehicle Assisted Fiber-based Visible Light Communication System. *J. Phys. Conf. Ser.* **2022**, *2264*, 012009. [CrossRef]
23. Zhang, R.; Li, Y.; Wang, C.-X.; Ruan, Y.; Zhang, H. Energy efficiency of relay aided D2D communications underlaying cellular networks. In Proceedings of the 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 8–13 October 2017; pp. 1–5.
24. Dang, S.; Chen, G.; Coon, J.P. Outage Performance Analysis of Full-Duplex Relay-Assisted Device-to-Device Systems in Uplink Cellular Networks. *IEEE Trans. Veh. Technol.* **2017**, *66*, 4506–4510. [CrossRef]
25. Fouda, M.M.; Hashima, S.; Sakib, S.; Fadlullah, Z.M.; Hatano, K.; Shen, X. Optimal Channel Selection in Hybrid RF/VLC Networks: A Multi-Armed Bandit Approach. *IEEE Trans. Veh. Technol.* **2022**, *71*, 6853–6858. [CrossRef]
26. Hashima, S.; Fouda, M.M.; Sakib, S.; Fadlullah, Z.M.; Hatano, K.; Mohamed, E.M.; Shen, X. Energy-Aware Hybrid RF-VLC Multi-Band Selection in D2D Communication: A Stochastic Multi-Armed Bandit Approach. *IEEE Internet Things J.* **2022**, *9*, 18002–18014. . [CrossRef]
27. Zhang, Y.; Wang, J.; Zhang, L.; Zhang, Y.; Li, Q.; Chen, K. Reliable Transmission for NOMA Systems with Randomly Deployed Receivers. *IEEE Trans. Commun.* **2023**, *71*, 1179–1192. [CrossRef]
28. Zhang, C.; Ge, J.; Pan, M.; Gong, F.; Men, J. One Stone Two Birds: A Joint Thing and Relay Selection for Diverse IoT Networks. *IEEE Trans. Veh. Technol.* **2018**, *67*, 5424–5434. [CrossRef]
29. Dang, S.; Chen, G.; Coon, J.P. Multicarrier Relay Selection for Full-Duplex Relay-Assisted OFDM D2D Systems. *IEEE Trans. Veh. Technol.* **2018**, *67*, 7204–7218. [CrossRef]
30. Khuntia, P.; Hazra, R.; Goswami, P. A Bidirectional Relay-Assisted Underlay Device-to-Device Communication in Cellular Networks: An IoT Application for FinTech. *IEEE Internet Things J.* **2021** . [CrossRef]
31. Sun, J.; Zhang, Z.; Xing, C.; Xiao, H. Uplink Resource Allocation for Relay-Aided Device-to-Device Communication. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 3883–3892. [CrossRef]
32. Gao, P.; Yang, Z.; Pei, L.; Du, J.; Chen, M. Energy-Efficient Mode Selection and Resource Allocation for Relay-Assisted D2D Communications. In Proceedings of the 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
33. Liu, M.; Zhang, L.; Gautam, P.R. Joint Relay Selection and Resource Allocation for Relay-Assisted D2D Underlay Communications. In Proceedings of the 2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC), Lisbon, Portugal, 24–27 November 2019; pp. 1–6.
34. Tian, C.; Qian, Z.; Wang, X.; Hu, L. Analysis of Joint Relay Selection and Resource Allocation Scheme for Relay-Aided D2D Communication Networks. *IEEE Access* **2019**, *7*, 142715–142725. [CrossRef]
35. Salim, M.M.; Wang, D.; Liu, Y.; Elsayed, H.A.E.A.; Elaziz, M.A. Optimal Resource and Power Allocation With Relay Selection for RF/RE Energy Harvesting Relay-Aided D2D Communication. *IEEE Access* **2019**, *7*, 89670–89686. [CrossRef]
36. Gupta, A.; Singh, K.; Sellathurai, M. Time-Switching EH-Based Joint Relay Selection and Resource Allocation Algorithms for Multi-User Multi-Carrier AF Relay Networks. *IEEE Trans. Green Commun. Netw.* **2019**, *3*, 505–522. [CrossRef]
37. Li, Y.; Xu, G.; Yang, K.; Ge, J.; Liu, P.; Jin, Z. Energy Efficient Relay Selection and Resource Allocation in D2D-Enabled Mobile Edge Computing. *IEEE Trans. Veh. Technol.* **2020**, *69*, 15800–15814. [CrossRef]

38. Gismalla, M.S.M.; Azmi, A.I.; Salim, M.R.B.; Abdullah, M.F.L.; Iqbal, F.; Mabrouk, W.A.; Othman, M.B.; Ashyap, A.Y.I.; Supa'at, A.S.M. Survey on Device to Device (D2D) Communication for 5 GB/6G Networks: Concept, Applications, Challenges, and Future Directions. *IEEE Access* **2022**, *10*, 30792–30821. [CrossRef]
39. He, Y.; Wang, D.; Huang, F.; Zhang, R.; Gu, X.; Pan, J. A V2I and V2V Collaboration Framework to Support Emergency Communications in ABS-Aided Internet of Vehicles. *IEEE Trans. Green Commun. Netw.* **2023** . [CrossRef]
40. Huang, S.; Chuai, G.; Gao, W. Coalitional Games Based Resource Allocation for D2D Uplink Underlaying Hybrid VLC-RF Networks. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; pp. 2316–2321.
41. Zhang, H.; Chong, S.; Zhang, X.; Lin, N. A Deep Reinforcement Learning Based D2D Relay Selection and Power Level Allocation in mmWave Vehicular Networks. *IEEE Wirel. Commun. Lett.* **2020**, *9*, 416–419. [CrossRef]
42. Wang, T.; Wu, S.; Wang, Z.; Jiang, Y.; Ma, T.; Yang, Z. A Multi-Featured Actor-Critic Relay Selection Scheme for Large-Scale Energy Harvesting WSNs. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 180–184. [CrossRef]
43. Huang, C.; Chen, G.; Gong, Y.; Wen, M.; Chambers, J.A. Deep Reinforcement Learning-Based Relay Selection in Intelligent Reflecting Surface Assisted Cooperative Networks. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 1036–1040. [CrossRef]
44. Liu, X.; Chuai, G.; Wang, X.; Xu, Z.; Gao, W.; Zhang, K.; Zuo, P. QoE-driven Antenna Tuning in Cellular Networks With Cooperative Multi-agent Reinforcement Learning. *IEEE Trans. Mob. Comput.* **2023** . [CrossRef]
45. Petkovic, M.I.; Cvetkovic, A.M.; Narandzic, M.; Chatzidiamantis, N.D.; Vukobratovic, D.; Karagiannidis, G.K. Mixed RF-VLC Relaying Systems for Interference-Sensitive Mobile Applications. *IEEE Trans. Veh. Technol.* **2020**, *69*, 11099–1111. [CrossRef]
46. Zhong, X.; Guo, Y.; Li, N.; Chen, Y. Joint Optimization of Relay Deployment, Channel Allocation, and Relay Assignment for UAVs-Aided D2D Networks. *IEEE/ACM Trans. Netw.* **2020**, *28*, 804–817. [CrossRef]
47. Demir, M.S.; Uysal, M. A Cross-Layer Design for Dynamic Resource Management of VLC Networks. *IEEE Trans. Commun.* **2021**, *69*, 1858–1867. [CrossRef]
48. 3GPP TR 36.843. Study on LTE Device to Device Proximity Services; Radio Aspects. v12.0.1, Release 12. 2014. Available online: https://www.360docs.net/doc/0f7312484.html (accessed on 20 July 2023).
49. Chen, Y.; Ai, B.; Niu, Y.; Guan, K.; Han, Z. Resource Allocation for Device-to-Device Communications Underlaying Heterogeneous Cellular Networks Using Coalitional Games. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 4163–4176. [CrossRef]
50. 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]
51. 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]
52. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O.P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Proc. Adv. Neural Inf. Process. Syst.* **2017**, 6379–6390.
53. Hernandezleal, P.; Kartal, B.; Taylor, M.E. A survey and critique of multiagent deep reinforcement learning. *Auton. Agents Multi-Agent Syst.* **2019**, *33*, 750–797. [CrossRef]
54. Xiao, L.; Li, Y.; Liu, J.; Zhao, Y. Power control with reinforcement learning in cooperative cognitive radio networks against jamming. *J. Supercomput.* **2015**, *71*, 3237–3257. [CrossRef]
55. Bowling, M.; Veloso, M. Rational and convergent learning in stochastic games. In Proceedings of the 17th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021; Volume 2, pp. 1021–1026.
56. Kai, Y.; Wang, J.; Zhu, H.; Wang, J. Resource Allocation and Performance Analysis of Cellular-Assisted OFDMA Device-to-Device Communications. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 416–431. [CrossRef]
57. Wang, X.; Jin, T.; Hu, L.; Qian, Z. Energy-Efficient Power Allocation and Q-Learning-Based Relay Selection for Relay-Aided D2D Communication. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6452–6462. [CrossRef]
58. Asheralieva, A.; Miyanaga, Y. An Autonomous Learning-Based Algorithm for Joint Channel and Power Level Selection by D2D Pairs in Heterogeneous Cellular Networks. *IEEE Trans. Commun.* **2016**, *64*, 3996–4012. [CrossRef]