drones

MDPI

Article

An Energy-Effective and QoS-Guaranteed Transmission Scheme in UAV-Assisted Heterogeneous Network

Jinxi Zhang [1], Weidong Gao [2], Gang Chuai [2,*] and Zhixiong Zhou [3,*]

[1] Beijing Kupei Sports Culture Corporation Limited, Beijing 100091, China
[2] Department of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[3] Institute for Sport Performance and Health Promotion, Capital University of Physical Education and Sports, Beijing 100088, China
* Correspondence: chuai@bupt.edu.cn (G.C.); zhouzhixiong@cupes.edu.cn (Z.Z.)

Abstract: In this article, we consider a single unmanned aerial vehicle (UAV)-assisted heterogeneous network in a disaster area, which includes a UAV, ground cellular users, and ground sensor users. The cellular data and sensing data are transmitted to UAVs by cellular users and sensor users, due to the outage of the ground wireless network caused by the disaster. In this scenario, we aim to minimize the energy consumption of all the users, to extend their communication time and facilitate rescue. At the same time, cellular users and sensor users have different rate requirements, hence the quality of service (QoS) of the users should be guaranteed. To solve these challenges, we propose an energy-effective relay selection and resource-allocation algorithm. First, to solve the problem of insufficient coverage of the single UAV network, we propose to perform multi-hop transmission for the users outside the UAV's coverage by selecting suitable relays in an energy-effective manner. Second, for the cellular users and sensor users inside the coverage of the UAV but with different QoS requirements, we design a non-orthogonal multiple access (NOMA)-based transmission scheme to improve spectrum efficiency. Deep reinforcement learning is exploited to dynamically adjust the power level and allocated sub-bands for inside users to reduce energy consumption and improve QoS satisfaction. The simulation results show that the proposed NOMA transmission scheme can achieve 9–17% and 15–32% performance gain on the reduction of transmit power and the improvement of QoS satisfaction, respectively, compared with state-of-the-art NOMA transmission schemes and orthogonal multiple access scheme.

Keywords: UAV communication; Internet of Things; relay selection; resource allocation; deep reinforcement learning

Citation: Zhang, J.; Gao, W.; Chuai, G.; Zhou, Z. An Energy-Effective and QoS-Guaranteed Transmission Scheme in UAV-Assisted Heterogeneous Network. *Drones* **2023**, *7*, 141. https://doi.org/10.3390/drones7020141

Academic Editor: Carlos Tavares Calafate

Received: 27 December 2022
Revised: 17 January 2023
Accepted: 16 February 2023
Published: 17 February 2023

1. Introduction

One of the important application scenarios of the fifth-generation mobile communication network is massive machine-type communication [1]. Worldwide, massive sensor devices are widely deployed to perform the tasks of environment monitoring, surveillance, and data collection to help build intelligent homes and intelligent cities, which greatly facilitate human life and improve manufacturing efficiency [2,3]. The IoT and traditional cellular networks form a terrestrial heterogeneous network, which can provide users with stable communication services.

UAV-assisted air–ground heterogeneous networks will be an important network architecture and deployment scenario in 5G and beyond [4,5]. When natural disasters occur and ground networks are destroyed due to the damage of communication infrastructure, UAV communication provides a promising solution to restore the communication of ground users by quickly setting up an air base station (BS) and establishing UAV-assisted air-ground heterogeneous networks [6]. Compared with ground BS, UAV communication has

the advantages of flexible deployment, high mobility, and strong line-of-sight (LoS) paths, and has been widely used in military, public and civilian fields [7].

However, there are still many challenges in the air–ground heterogeneous network. On the one hand, although the deployment of the UAV can restore the communication of the target area, there are still areas that are not illuminated.The ground cellular users (CUs) and IoT users (or sensor users (SUs)) located in the coverage hole cannot build direct communication with the UAV due to poor channel conditions. In this case, device-to-device (D2D) technology [8] can be employed to establish multi-hop communication for users outside the coverage area, then the data is transmitted through relays in a decode and forward (DAF) manner. To save energy, relay selection should be performed in an energy-effective way. In other words, the users outside the coverage area should comprehensively consider their own energy cost and rate requirements, and select the relay that can provide the required rate and consume as little energy as possible.

On the other hand, NOMA transmission provides an effective solution to satisfy the transmission requirement for different users using limited system bandwidth. Compared with orthogonal multiple access (OMA), NOMA can increase the spectrum efficiency by allowing multiple users sharing the same channel resource. Reducing energy consumption and guaranteeing the QoS of users are the key problems in NOMA-enabled networks. Since ground CUs and IoT devices are energy-constrained, power control is a crucial issue to be handled. In addition, to facilitate rescue, the rate requirement of users should be guaranteed. Therefore, the trade-off between reducing energy consumption and ensuring the QoS of users in the air–ground heterogeneous network should be achieved, which pose a huge challenge to the traditional orthogonal resource allocation scheme. However, in the literature, there is little research focusing on reducing energy consumption while guaranteeing the different QoS requirements of users in NOMA-enabled heterogeneous networks. Hence, the optimization of relay selection and resource allocation in NOMA transmission is needed.

Motivated by the aforementioned analysis, we propose an energy-effective relay selection and NOMA-based resource allocation for a UAV-assisted heterogeneous network. First, to tackle the problem of limited coverage of a single UAV network, we propose to associate inside users and outside users by formulating a many-to-one matching game [9,10], which considers the power consumption and the QoS requirements of outside users. Hence, the serving area of the UAV can be effectively extended and the relay selection is performed at the minimum energy cost. To relay the data for outside users, different QoS requirements are needed for the relays, hence the QoS requirements for users are diversified.

Second, to serve multiple users using limited resources, we adopt NOMA transmission scheme to increase spectrum efficiency. The interference in the network is incurred due to the non-orthogonal resource allocation in NOMA transmission. Thus, the sub-band allocation scheme should be carefully designed and uplink transmit power should be fine-tuned to alleviate the interference and guarantee the QoS of users. To this end, we propose a deep reinforcement learning (DRL)-based resource allocation algorithm to select appropriate power level and sub-band for ground users, which can achieve the trade-off between reducing energy consumption and guaranteeing user's QoS.

The main contributions of this paper can be summarized as follows: (1) we model the problem of relay selection and resource allocation for NOMA transmission in UAV-assisted heterogeneous networks and propose an energy-effective relay selection and NOMA-based resource allocation algorithm. (2) We perform the relay selection for outside users and inside users by designing a low-complexity many-to-one matching algorithm, which can save energy and guarantee the QoS of the users. (3) To satisfy the diverse QoS requirements of inside users, we propose an DRL-based power and sub-band allocation algorithm, which can achieve the balance of saving energy and guaranteeing QoS. (4) The performance of the proposed algorithm is validated under different network parameters. Simulation results demonstrate that the proposed algorithm can achieve better performance than

state-of-the-art NOMA schemes and OMA schemes in terms of energy consumption and QoS satisfaction.

The rest of this paper is organized as follows: the related works are concluded in Section 2. In Section 3, single UAV-assisted heterogeneous network models and problem formulation are presented in detail. In Section 4, we illustrate our proposed energy-effective relay selection and DRL-based NOMA transmission scheme. The simulation results are presented and discussed in Section 5. Finally, we draw the conclusions in Section 6.

2. Related Works

In the literature, UAV-assisted air–ground heterogeneous networks have been extensively studied. To further increase the spectrum efficiency, the combination of NOMA and UAV communication has been studied. In [11], the authors studied the NOMA transmission model in UAV-assisted IoT systems. First, the authors used matching game to optimize the resource block (RB) allocation in the system, and then successive convex approximation was used to optimize the transmission power and UAV's height. In [12], the authors considered a terrestrial heterogeneous IoT where multiple ground SUs and CUs coexist, and proposed an effective successive interference cancellation (SIC)-free NOMA transmission scheme to optimize RB allocation and power allocation. In [13], the authors proposed a game theory-based NOMA scheme to maximize energy efficiency (EE) in NOMA-based fog UAV wireless networks. In [14], a channel gain-based NOMA scheme was proposed and network EE was optimized using alternating optimization. In [15], the authors proposed jointly optimizing UAV trajectory planning and sub-slot allocation to maximize the sum-rate of IoT devices in UAV-assisted IoT. The authors of [16] studied the multi-NOMA-UAV assisted IoT system to increase the number of served IoT nodes and improve the system EE. In [17], the power allocation was optimized in NOMA clustering-enabled UAV-IoT. In [18], the joint optimization of UAV deployment and power allocation was proposed to maximize the sum-rate of ground users.

Due to the scarcity of spectrum resources and energy constraints, resource allocation and power optimization have also become the focus of research [19–24]. The authors in [19–23] studied the ground data collection assisted by UAV and minimized the energy consumption of IoT devices in an UAV-assisted ground IoT by optimizing the UAV trajectories [19,20] and 3D deployment [23]. The authors in [21] also considered the UAV-assisted IoT, aiming to minimize the total time of data collection for the UAV to save energy. By applying alternation optimization and successive convex approximation, the UAV's trajectory and transmit power have been optimized, thereby saving energy while ensuring the users' QoS. The authors in [22] optimized the UAV's trajectory in terms of flight speed and acceleration, and obtained the optimal solution of UAV trajectory and uplink transmission power under two modeling problems of maximizing the minimum average rate and maximizing EE. The authors in [24] comprehensively considered the power optimization, RB allocation and UAV location optimization under the SAG-IoT architecture to achieve the optimization of energy efficiency.

The research on relay-based transmission has also been carried out. The network performance exploiting UAVs [25–27] and mobile devices as relays [28–30] are studied respectively. The authors in [28] exploited ground devices to dispatch the files transmitted by UAV by establishing ground D2D links and adopted graph theory to optimize the resource allocation in NOMA-enabled transmission. The authors in [29,30] designed a multi-hop communication algorithm for the ground IoT, and obtained the outage probability for relay links.

In addition, the security problem in UAV-assisted network was widely studied [31,32]. The control schemes based on blockchain and artificial intelligence were proposed to secure drone networking [33,34]. In [35], an enhanced authentication protocol for drone communications, and the authors proved that their algorithm could better fight drone capture attacks. In [36], a lightweight mutual authentication scheme based on physical unclonable functions for UAV-ground BS authentication was proposed. In [37], the authors studied the physical layer security

issue in UAV-assisted cognitive relay system, and proposed alternate optimization-based algorithm to maximize average worst-case secrecy rate.

The research comparison of energy saving for users and coverage expand for UAVs in the existing literature has been listed in Table 1. It can be seen that these studies either did not reduce the energy consumption of ground users while guaranteeing the user QoS, or did not adopt the NOMA transmission scheme to increase spectrum efficiency. Therefore, the research in the existing literature cannot guarantee the users' QoS and save energy under insufficient coverage and limited resources of UAV. In addition, energy saving in relay selection is not considered enough, and few studies focus on user QoS and energy consumption in NOMA-enabled air-ground networks. The existing research is performed on the premise that the system bandwidth is sufficient and the users' QoS can be guaranteed. However, for the scenario with limited system bandwidth and massive users, it is difficult to ensure the QoS requirements of all the users. Motivated by this, we propose our energy-effective relay selection and resource allocation algorithm in NOMA UAV-assisted heterogeneous networks.

Table 1. Comparison of related works in the existing literature.

Article	Method	Advantage	Limitation
[11]	Matching game and alternative optimization	Maximized uplink capacity	Not energy saving and user's QoS is not guaranteed
[12]	Message-passing algorithm	Successive interference cancellation-free	Relay selection is not optimized
[13]	Matching game	Maximized EE	User's QoS is not guaranteed
[14]	Alternating optimization	Maximized EE	The coverage extension of UAV is not considered
[15]	Alternating optimization	Maximized EE	Not energy saving
[16]	Alternating optimization	Maximized EE	The number of served users is limited
[17]	Deep reinforcement learning	Maximized sum-rate	Not energy saving
[18]	Particle swarm optimization and dynamic power allocation	Maximized sum-rate	Not energy saving
[19]	Optimal transport theory	Optimal trajectories and minimized energy consumption	NOMA and multi-hop transmission is not considered
[20–22]	Alternating optimization	Optimal trajectories and minimized energy consumption	NOMA and multi-hop transmission is not considered
[23,24]	Matching theory and alternating optimization	Optimal 3D deployment and transmit power of UAVs	NOMA and multi-hop transmission is not considered
[25–27]	Joint transmit power and trajectory optimization	Optimized trajectory for UAVs	The number of served users is limited
[28]	Graph theory	Minimized flying time and maximized system capacity	Not energy saving
[29,30]	Shortest-path routing algorithm	Shortest path for multi-hop D2D links	NOMA transmission is not considered

3. System Model

In this paper, we consider a single UAV-assisted heterogeneous network, as shown in Figure 1. After the disaster, the ground BS in the disastrous area is out of service. To deliver the emergency calls and messages of ground CUs and the monitoring data of IoT users (sensor users), a rotary-wing UAV is deployed to hover above the target area and acts as aerial BS to serve ground users. In this paper, we focus on the scenario where the serving range of the UAV can not cover all the users in the target area. Therefore, we ignore the impact of surrounding UAVs and consider the single UAV network. Without loss of generality, we consider that there are some users that locate outside the coverage area and cannot directly connect to the UAV. Therefore, the D2D-enabled multi-hop transmission is

employed. For simplicity, we consider the number of hops is two, that is, a user outside the UAV coverage transmits its data to the relay who is in the coverage area of the UAV, then the data will be delivered to the UAV from the relay.In addition, we assume that the channel resource in the network is limited. In this case, OMA transmission fails to provide service to a large number of ground users, and we adopt NOMA scheme to transmit data for inside users. The reason is that NOMA allows multiple users to transmit on the same channel resource, which greatly improves the spectrum efficiency. This enables UAV to serve a large number of users with limited resources.

In Figure 1, the coverage of UAV is partitioned into multiple rings. Clustering is performed by selecting the users from different rings, e.g., inside CU #1, inside CU #2 and inside SU #1 form a NOMA cluster. The users in the same NOMA cluster occupy a unique part of the system bandwidth, and there is no interference among the users in different clusters.

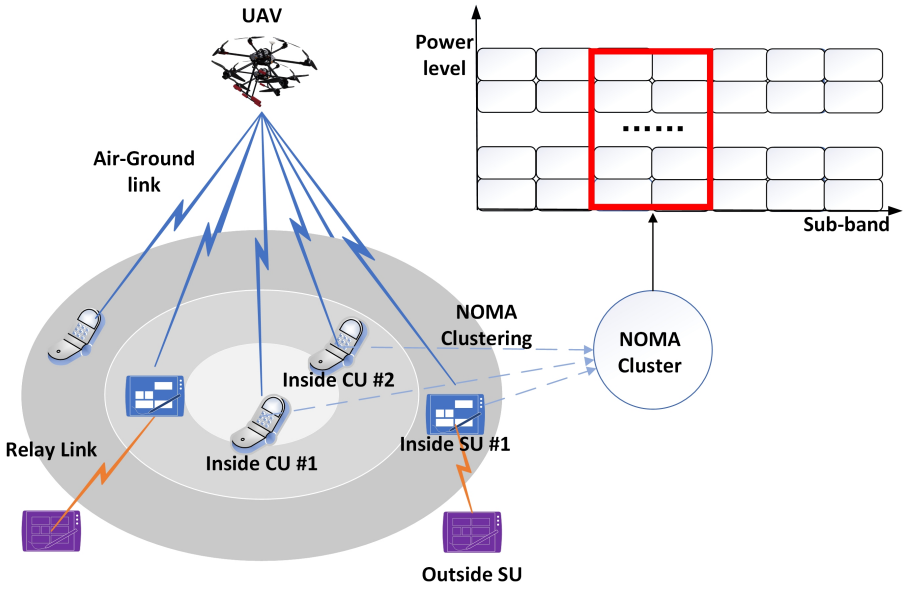

Figure 1. Network architecture.

We denote the set of ground users as $\mathcal{U}$, which is comprised of CUs and SUs denoted as $\mathcal{U}_C$ and $\mathcal{U}_I$, i.e., $\mathcal{U} = \mathcal{U}_C \cup \mathcal{U}_I$. The set of SUs are divided into two parts, $\mathcal{U}_I^{in}$ and $\mathcal{U}_I^{out}$, which corresponds to the SUs inside and outside the UAV coverage. We consider that the outside users only include SUs and each outside SU can only select an inside SU as its relay. However, the proposed model and algorithm can be extended to other complicated scenarios.We delineate the UAV coverage based on path loss. Specifically, when the path-loss of a local user to the UAV is less than the predefined threshold [38], the user is regarded as an inside user and can be served by the UAV. Otherwise, it is regarded as an outside user which can only communicate with UAVs through two-hop relay links. For simplicity, we assume the CUs and SUs have different QoS requirements, which are denoted as R_{min}^C and R_{min}^I. In addition, we denote the positions of the UAV and user $k \in \mathcal{U}$ as $[x_u, y_u, h]$ and $[x_k, y_k]$, where $[x_u, y_u]$ is the horizontal location of the UAV and h is the height of the UAV, which is fixed after the deployment.

Hence, the distance between the UAV and user k is calculated as:

$$d_{k,u} = \sqrt{(x_u - x_k)^2 + (y_u - y_k)^2 + h^2} \tag{1}$$

Then, the probability of the line-of-sight link between the UAV and user k is calculated as:

$$P_{LoS}^{k,u} = \frac{1}{1 + aexp(-b[\epsilon_{k,u} - a])} \tag{2}$$

where a and b are the constants determined by the environment. $\epsilon_{k,u}$ is the elevation angle between the UAV and user k, which is calculated as: $\epsilon_{k,u} = arctan(h/\sqrt{(x_u - x_k)^2 + (y_u - y_k)^2})$.

Hence, the path-loss between the UAV and user k is calculated as:

$$h_{k,u} = 20\log_2(\frac{4\pi f_c d_{k,u}}{c}) + P_{LoS}^{k,u}\zeta_{LoS} + (1 - P_{LoS}^{k,u})\zeta_{NLoS} \tag{3}$$

where f_c denotes the carrier frequency, c is the speed of light, ζ_{LoS} and ζ_{NLoS} denote the attenuation loss in LoS and NLoS links.

For the transmission of the air–ground uplink, we adopt NOMA to increase the spectrum efficiency. In a NOMA-based system, a user will suffer the potential interference from the users that occupy the same spectrum resource. In our proposed scenario, it is assumed that the UAV is equipped with a SIC receiver that can demodulate the target signal from the superimposed signal. At the same time, it is assumed that the demodulation sequence is from the users with the highest received power to the users with the lowest received power. For user k, the interference comes from the users whose received power is lower than user k. Therefore, the uplink signal-to-interference-plus-noise ratio (SINR) of user k at sub-band n is calculated as:

$$\gamma_{k,u,n} = \frac{p_k h_{k,u} g_{k,u,n}}{\sum_{k' \in \mathcal{U}_{n,k}} p_{k'} h_{k',u} g_{k',u,n} + \sigma^2} \tag{4}$$

where p_k is the uplink transmit power of user k, $\mathcal{U}_{n,k}$ is the set of users who reuse sub-band n and whose received power level at the UAV is lower than that of user k, $g_{k,u,n}$ is the small-scale fading between the UAV and user k at sub-band n, which follows the exponential distribution with unit mean. σ^2 is the noise power.

Further, the uplink transmission rate of inside user k (CU or SU) is calculated as:

$$R_{k,u} = B_a \sum_{n \in \mathcal{N}} a_{n,k} \log_2(1 + \gamma_{k,u,n}) \tag{5}$$

where $a_{n,k} \in 0, 1$ is the binary indicator to show that whether user k transmit on sub-band n. B_a is the bandwidth of a sub-band of the uplink air-ground channel.

For the relay link, the transmission rate is calculated as:

$$R_{k,M_k} = B_r \log_2(1 + \frac{p_k h_{k,M_k}}{\sigma^2}) \tag{6}$$

where M_k is the associated relay for outside user k, $h_{k,M_k} = g_{k,M_k} d_{k,M_k}^{-\alpha}$ is the channel gain from outside user k to relay M_k, g_{k,M_k} is the small-scale fading between outside user k and relay M_k, which follows the exponential distribution with unit mean. d_{k,M_k} is the distance between user k and relay M_k, B_r is the bandwidth of a resource block in the relay link.

It is worthwhile to note that the interference among the relay links is not considered as the orthogonal frequency is employed for different links. In addition, we assume the relay links and the air-ground links reuse the same frequency band. However, the interference from the relay links to the UAV can be neglected due to low transmission power and long distance.

In this paper, we characterize the QoS requirement of CUs and SUs as the minimum transmission rate. For CU-k and SU-k, the QoS requirement is represented as the minimum transmission rate $R_{min}^{C,k}$ and $R_{min}^{I,k}$, respectively. For each relay node k, its total QoS requirement is constituted by the transmission rate requirements of its own air-ground link and all the outside users connecting to k.

$$R_{min}^{I,k}(relay) = R_{min}^{I,k} + \sum_{m \in \mathcal{M}_k} R_{min}^{I,m} \tag{7}$$

where $\mathcal{M}_k$ is the set of outside sensor users choosing k as the relay node.

In this paper, we aim to reduce the energy consumption while guaranteeing the QoS of the users. For the users in $\mathcal{U}_I^{in}$, the transmit power and resource allocation should be optimized to alleviate the intra-cell interference and increase user rate. For the users in $\mathcal{U}_I^{out}$, the relay selection should also be performed in an energy-effective manner. The transmit power of all users is used to characterize the power consumption. In addition, the QoS satisfaction is characterized using the indicator function to denote whether the QoS requirement of a user is satisfied. We design the target problem as follows:

$$\begin{aligned} \underset{p_k, a_{n,k}}{maximize}\ & \omega \sum_{k\in\mathcal{U}} \eta / p_k + (1-\omega)\Big(\sum_{k\in\mathcal{U}_C} \mathbb{I}(R_{k,u} \geq R_{min}^{C,k}) + \sum_{k\in\mathcal{U}_I^{in}} \mathbb{I}(R_{k,u} \geq R_{min}^{I,k})\Big) \\ s.t.\ & C1: p_{min}^C \leq p_k \leq p_{max}^C, \forall k \in \mathcal{U}_C \\ & C2: p_{min}^I \leq p_k \leq p_{max}^I, \forall k \in \mathcal{U}_I \\ & C3: \sum_{n\in\mathcal{N}} a_{n,k} = 1, \forall k \in \mathcal{U}_I^{in} \end{aligned} \tag{8}$$

where p_k is the transmit power of user k, and $R_{min}^{C,k}$ and $R_{min}^{I,k}$ are the QoS requirements for CU-k and SU-k, respectively. $\mathbb{I}$ is the QoS indicator function. $\omega \in (0,1)$ is the weighting factor to characterize the importance of power consumption and users' QoS satisfaction. η is the tuning coefficient to adjust the transmit power and QoS indicator to the same order of magnitude. C1 and C2 denote the power constraint for CUs and SUs, p_{min}^C and p_{min}^I are the minimum transmit power for CU and SU, p_{max}^C and p_{max}^I are the maximum transmit power for CU and SU. C3 indicates that each inside user can only occupy one sub-band. The symbols used in the paper are described in Table 2.

Table 2. Symbol table.

Symbol	Description	Symbol	Description
$\mathcal{U}$	The set of ground users	ζ_{LoS}	The attenuation loss in LoS links
$\mathcal{U}_C$	The set of cellular users	ζ_{NLoS}	The attenuation loss in NLoS links
$\mathcal{U}_I$	The set of IoT users	$\gamma_{k,u,n}$	The uplink signal-to-interference-plus-noise ratio (SINR) of user k at sub-band n
$\mathcal{U}_I^{in}$	The set of inside IoT users	p_k	The uplink transmit power of user k
$\mathcal{U}_I^{out}$	The set of outside IoT users	$\mathcal{U}_{n,k}$	The set of interfering users of user k in sub-band n
$R_{min}^{C,k}$	The QoS requirement of cellular user k	$g_{k,u,n}$	The small-scale fading between the UAV and user k at sub-band n
$R_{min}^{I,k}$	The QoS requirement of IoT user k	$R_{k,u}$	The uplink transmission rate of inside user k
$d_{k,u}$	The distance between the UAV and user k	M_k	The associated relay for outside user k
$P_{LoS}^{k,u}$	The probability of LoS link between the UAV and user k	R_{k,M_k}	The transmission rate of outside user k when connected to relay M_k
$\epsilon_{k,u}$	The elevation angle between the UAV and user k	$a_{n,k}$	The binary indicator to show that whether user k transmit on sub-band n
$h_{k,u}$	The path-loss between the UAV and user k	B_a	The bandwidth of a sub-band in the uplink air-ground channel

4. Proposed Algorithm

As can be seen, problem (8) is computationally hard. Let us assume that the uplink transmission power of inside users can be discretized into N_p levels and the total number

of sub-bands in the network is N_{sub}. For inside users, the worst case for the number of combination of power level and sub-band is $(\mathcal{U}_C + \mathcal{U}_I^{in})^{N_p * N_{sub}}$. For outside users, the worst case for the number of combination of power level is $(\mathcal{U}_I^{out})^{N_p}$. Hence, the total number of combination of power level and sub-band in the network is $(\mathcal{U}_I^{out})^{N_p} + (\mathcal{U}_C + \mathcal{U}_I^{in})^{N_p * N_{sub}}$, which makes solving problem (8) time consuming.

Due to the huge number of combinations of power and sub-bands, we decompose the target problem and optimize the target function for inside users and outside users, respectively. The flowchart of our proposed algorithm is shown in Figure 2. First, the relay selection is performed to select relay for each outside SU and determine the association between outside SUs and inside SUs. Next, based on the association result, DRL-based power and resource selection algorithm is performed for inside users. Finally, the optimal power and resource selection strategy for inside users is obtained.

In this section, we first propose an energy-efficient relay selection scheme for outside users exploiting the many-to-one matching game, and the energy consumption and QoS requirements in the relay selection process are considered. After determining the relay for each outside user, we perform the joint power and sub-band selection algorithm for the inside users. The deep reinforcement learning is adopted to adjust the power and sub-band selection dynamically in different environment.

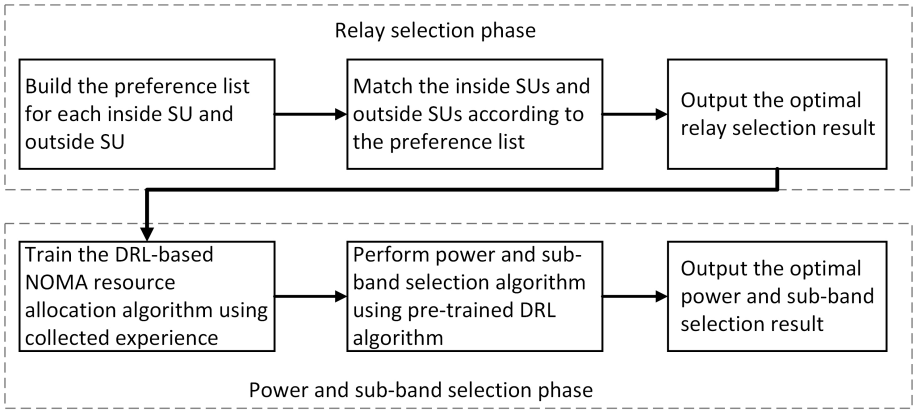

Figure 2. The flowchart of proposed algorithm.

4.1. Energy-Efficient Relay Selection Algorithm

To restore the connection for outside users, we first associate the outside SUs and relays (i.e., inside SUs) in an energy-effective manner. The essence of our scheme is to determine the association scheme according to the consumed energy of outside users. We derive the preference list for the relays and outside SUs, respectively, by calculating the utility function related to the consumed energy and sorting them in a descending order. After each outside user proposes their preferred relays, the relays accept the application according to the pre-built preference list and the maximum number of accepted users for each relay. The outside users are more inclined to select the relay that consume as little energy as possible.

The utility for user $i \in \mathcal{U}_I^{out}$ to connect to a candidate relay $j \in \mathcal{U}_I^{in}$ and the utility for relay j to accept applicant i are calculated as:

$$U_i(j) = U_j(i) = (p_{i,j}^{min} + p^{cir})^{-1} \tag{9}$$

where p^{cir} is the static circuit energy consumption at SU, $p_{i,j}^{min}$ is the minimum transmit power for user i to reach the QoS requirement when connected to relay j, which is calculated as:

$$p_{i,j}^{min} = 2^{(R_{min}^I / B_r - 1)} \sigma^2 / h_{i,j} \tag{10}$$

In our paper, the association between the outside users and inside users is formulated as a many-to-one matching game. In the proposed matching game, the players are modeled

as outside users and inside users. Each player has a preference list and the final association result is obtained based on the preference order:

Definition 1 (preference order). *For each outside user (or relay), the preference order $\triangleright_i$ is defined as a complete, reflexive, and transitive binary relation over the entire set of relays (or outside users).*

For each outside user $i \in \mathcal{U}_I^{out}$, the preference relation over all the possible relays is defined as follows:

$$j \triangleright_i j' \Leftrightarrow U_i(j) > U_i(j') \tag{11}$$

which means that user i is more likely to select j rather than j' as its relay.

For each relay $j \in \mathcal{U}_I^{in}$, the preference relation over all the outside users is defined as follows:

$$i \triangleright_j i' \Leftrightarrow U_j(i) > U_j(i') \tag{12}$$

which means that relay j is preferable for acting as the relay of user i.

The detailed process of our proposed energy-effective relay-selection algorithm is summarized in Algorithm 1. At the beginning, the outside users and relays exchange their channel state information and other related information. Each outside user and relay builds the preference list according to the utility, calculated as (9). Then, each outside user makes the connection request to its preferred relay. After receiving the proposal, each relay ranks the applicants according to the preference list. The maximum number of outside users that can be accepted by the relay is set to N_{max}. Each relay accepts its preferred applicant and the remaining applicants will be rejected. Then, each rejected applicant proposes the next preferable relay. The association process terminates when all the outside users are associated with a relay.

Algorithm 1: Energy-efficient relay selection for outside users.

Initialize the preference list for each outside user and relay according to (9), and the set of outside users accepted by relay as: $\mathcal{U}_I^{out,rejected} = \mathcal{U}_I^{out}$.
while $\mathcal{U}_I^{out,rejected} \neq \emptyset$ **do**
  foreach *outside user* $i \in \mathcal{U}_I^{out}$ **do**
    Request to connect to its preferred relay as indicated by (11).
  end
  foreach *relay* $j \in \mathcal{U}_I^{in}$ **do**
    Sorts the applicants in a descending order according to the preference list as indicated by (12);
    while *the number of accepted users at relay j not exceed* N_{max} **do**
      accept the preferred applicant, add it into $\mathcal{U}_I^{out,accepted}$ and delete it from the preference list of j;
    end
  end
  Each rejected user updates the preference list by deleting the preferred relay.
end

4.2. Deep Reinforcement Learning-Based Power and Sub-Band Selection for NOMA Transmission

In this subsection, we perform power and sub-band selection for inside users in a single UAV-assisted heterogeneous network. After executing the relay selection algorithm for outside users, the D2D links between the outside users and the inside relays are established. Therefore, the inside users have different QoS requirements in terms of transmission rate. In this case, the allocation of limited resource among inside users has a significant impact on system performance.

Therefore, we adopt NOMA scheme for the transmission of inside users, which can increase spectrum efficiency by allowing multiple users share the same channel resource.

To reduce the complexity of SIC, NOMA clustering is performed by dividing the inside users into different clusters according to their geometry locations [39]. The users in the same cluster perform the joint power and sub-band selection to reduce the power consumption while ensuring the QoS requirements of the users can be reached. In the NOMA clustering, each cluster occupies an orthogonal frequency band (i.e., sub-band) and there is no interference among the users from different clusters.

However, the resource allocation and interference cancellation inside the NOMA cluster is a crucial issue to be handled. When an external environment (i.e., channel state information, transmission power, or resource occupancy status) changes, traditional NOMA schemes need iterative calculation or to obtain the power and channel selection scheme. In addition, the traditional NOMA schemes fail to guarantee the QoS of users with limited resources.

To solve this problem, we resort to deep reinforcement learning (DRL) for dynamic and adaptive resource allocation to achieve the goal of reducing energy consumption while guaranteeing the users' QoS. Deep reinforcement learning has the general intelligence to solve complex problems, and can automatically obtain the optimal resource allocation strategy from the changing external environment. DRL is built on the learning and prediction function of reinforcement learning, and uses deep neural networks to provide an autonomous decision-making mechanism for learning agents by forming a powerful approximation function. Compared with traditional NOMA schemes, the proposed NOMA scheme can adaptively obtain the optimal power and channel selection strategy for the inside users using a pre-trained strategy.

Therefore, we first give a brief introduction to DRL, then we illustrate the framework and implementation of our proposed DRL-based power and sub-band selection.

4.2.1. Basis of Deep Reinforcement Learning

The target problem (8) involves competitive relation between users, which is not convex and cannot be solved by traditional optimization technique. In this case, reinforcement learning (RL) can effectively solve the problem through interacting with the unknown environment and improve the decision-making to maximize the target value.

RL is a method or framework for learning, prediction, and decision-making, which has the natural advantage of automatically obtaining the optimal strategy through the interaction between agents and the environment. The basic elements of reinforcement learning can be represented by a five-tuple $(\mathcal{S}, \mathcal{A}, \pi, p, \mathcal{R})$. The agents learn and make decisions by sensing the state of the environment $\mathcal{S}$. In the $t-th$ time slot, the agent executes an action following the strategy π to select a sub-band and transmission power level according to current state $s_t \in \mathcal{S}$. The transition of different states and the obtained rewards are both stochastic, which can be modeled as a Markov decision process (MDP). When the agent takes action $a_t \in \mathcal{A}$, the state transitions from s_t to s_{t+1} and the acquisition of $r_t \in \mathcal{R}$ can be characterized by the conditional transition probability $p(s_{t+1}, r_t | s_t, a_t)$.

The main goal of reinforcement learning is to find the strategy that maximizes the cumulative discounted reward, which not only considers immediate rewards, but also takes into consideration the discounted future rewards:

$$r_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k} \tag{13}$$

where R_t is the instantaneous reward received by the agent at time slot t. γ is the discount factor, when γ is close to 0, the agent is more concerned about short-term rewards; when γ is close to 1, long-term rewards is considered more important.

In this paper, we adopt the most commonly-used reinforcement learning method, Q-learning, to solve the MDP problem. In Q-learning, a Q-table is constructed to reflect the strategy, which stores the Q-value of different state-action pairs. According to the Bellman

equation and the temporal difference learning method [40], after the agent observes the state s_t at time t and executes an action a_t, the corresponding Q-value is updated as:

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta(r_{t+1} + \gamma max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (14)$$

where η is the learning rate. With the state transitions in the learning process, the Q-table gradually stabilizes, and the optimal strategy function for each agent can be obtained.

However, Q-learning (QL) can only effectively solve the problems whose states and actions are discrete and limited due to that the Q-tables can only record the Q-value of a limited number of state–action pairs. However, in many practical problems or tasks, the number of states and actions is large, which makes QL inefficient in solving the problem. In this case, deep Q-learning (DQL) can be used to learn the strategy from the state transitions in high-dimensional and continuous state space successfully using deep learning (DL), which has been widely studied and applied in UAV networks [41,42].

The core idea of DQL is approximating a complex nonlinear Q-function $Q(\mathcal{S}, \mathcal{A})$ using deep Q-network (DQN):

$$Q(s_t, a_t; \theta) \approx Q^*(s_t, a_t) \quad (15)$$

where $Q(s_t, a_t; \theta)$ is the Q-value function approximated by DQN with parameter θ, $Q^*(s_t, a_t)$ is the Q-value function with the best future reward.

In DQN, the input is the state vector, and the output is the value function vector containing the Q-value of each action under the state. At time t, the agent observes the state s_t, chooses the action a_t following the strategy, and receives the reward r_t. Then, the optimal DQN parameter can be calculated by minimizing the following loss function:

$$\mathcal{L}(\theta) = ([r_t + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}|\theta)] - Q(s_t, a_t|\theta)) \quad (16)$$

In addition, to overcome the problems of non-convergence and instability, experience replay and target network [43] are introduced into DQN to stabilize the learning process. To reduce the gap between the target Q-value obtained through experience replay and the Q-value calculated by the primary Q-network, the parameter θ of the neural network is updated through the back propagation using the gradient-descent method [40]. Hence, the loss function to be minimized is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \in \mathcal{B}}([r_t + \gamma max_{a_{t+1}} Q(s_{t+1}, a_{t+1}|\theta')] - Q(s_t, a_t|\theta)) \quad (17)$$

where $\mathcal{B}$ is the mini-batch set selected from replay buffer. The detailed structure of DQN is shown in Figure 3.

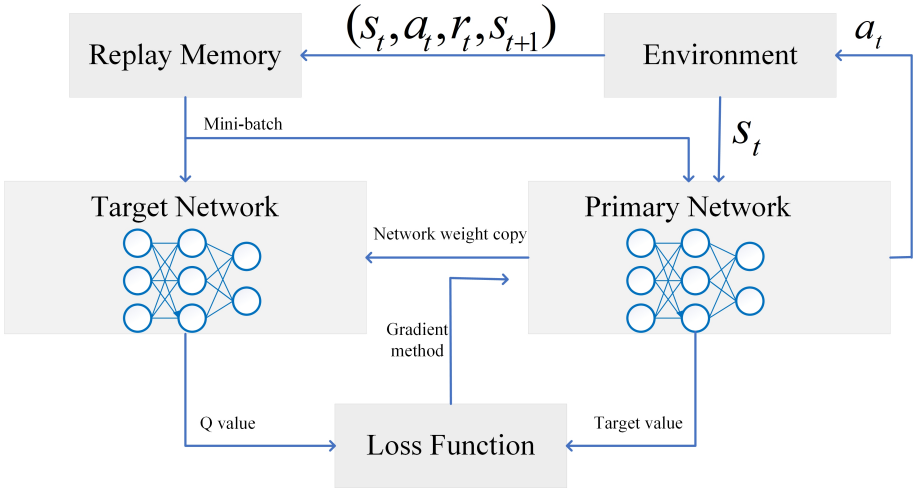

Figure 3. Structure of DQN.

4.2.2. Customized DQN Design

In the proposed reinforcement learning framework, the agents are the inside users, and the actions are all the combinations of power level and sub-band that can be selected by each agent. The uplink transmission power of inside users is discretized into N_p levels, and each cluster is assigned N_s to sub-bands, hence the dimension of the action space of each agent is $N_p N_s$. The observable environment state of each agent k at time slot t consists of the following parts:

- The instantaneous channel state information of all the sub-bands at time-slot $t-1$: $\boldsymbol{g_{t-1}} = (g_{k,u,1}^{t-1}, ..., g_{k,u,N_s}^{t-1})$, where N_s is the number of sub-bands of each cluster, and $g_{k,u,n}^{t-1}$ is the small-scale channel gain of sub-band n between the UAV and user k at time-slot $t-1$.
- The interference power level that user k receives in each sub-band at time-slot $t-1$: $\boldsymbol{I_{t-1}} = (I_{t-1}^1, ..., I_{t-1}^{N_s})$, where $I_{t-1}^n = \sum_{k' \in \mathcal{U}_{n,k}} p_{k'} h_{k',u}$ denotes the potential interference at sub-band n that agent k receives at time-slot $t-1$.
- The ACK indicator vector of users in the same cluster with the agent: $\boldsymbol{ACK_{t-1}} = (ACK_{t-1}^1, ..., ACK_{t-1}^{K_c})$, where $ACK_{t-1}^k = 1$ means that the QoS requirement of user k is reached and K_c is the total number of users in the same cluster.

Therefore, the observed state at time t for each agent can be expressed as $s_t = \{\boldsymbol{g_{t-1}}, \boldsymbol{I_{t-1}}, \boldsymbol{ACK_{t-1}}\}$.According to the agent's observed state s_t and the selected action a_t at time-slot t, the environment will transit to a new state s_{t+1}, and the agent will receive an immediate reward R_t. In this paper, we aim to reduce the energy consumption of inside users while ensuring that the QoS of outside users is not degraded. In the proposed reinforcement learning framework, the reward function is consistent with the objective function (8), which consists of two parts, namely the power consumption and the QoS indication summation of all the users in the same cluster. Specifically, the instantaneous reward function for agent k (inside CU or SU) at time-slot t is designed as:

$$R_t = \omega\Big(\eta / p_k + \sum_{m \in \mathcal{U}^k} \eta / p_m\Big) + (1-\omega)\Big(\sum_{m \in \mathcal{U}_C^k} \mathbb{I}(R_{u,m} \geq R_{min}^C) + \sum_{m \in \mathcal{U}_I^k} \mathbb{I}(R_{u,m} \geq R_{min}^I)\Big) \tag{18}$$

where $\mathcal{U}_C^k$ and $\mathcal{U}_I^k$ denote the CUs and SUs in the same cluster with user k, $\mathcal{U}^k = \mathcal{U}_C^k \cup \mathcal{U}_I^k$.

4.2.3. Implementation of Deep Q-Learning

In deep Q-learning, there are two stages: offline training and testing. Unlike DL, there is no concept of a training data set or test data set in DQN, and each agent learns the optimal strategy by collecting experience from the transition of states, and then the performance of learned policy is tested in realistic environment. In the offline training stage, the agent traverses the states as much as possible by interacting with the environment, continuously learns and improves its action selection strategy, and finally obtains a stable Q-function approximation. When the offline training process is completed, the agent already has the knowledge of which action to perform under different environment to get the best cumulative discounted reward. In the testing phase, the agent exploits the learned policy (approximation function) to guide the action selection in different environment state.

Due to the independence of the agents' execution of actions, when the actions are performed synchronously, the agent has no knowledge of the actions selected by other agents. In this case, the value of $\boldsymbol{I_{t-1}}$ in the input state vector is not latest updated, and the state observed by each agent cannot accurately represent the environment in real time. To this end, the agents are designed to update their actions asynchronously: in each time slot, only one agent performs the action selection. Under the asynchronous strategy, each agent can observe the environmental change caused by the behavior of other agents through recording $\boldsymbol{I_{t-1}}$ and $\boldsymbol{ACK_{t-1}}$ in the input state vector. In this way, the wrong action selection caused by inaccurate observation of the environment state can be mitigated.

The detailed process of implementing DQL is presented in Algorithm 2. At the beginning of training, the structure of two deep neural networks and other parameters are initialized. Then, each agent interacts with the environment in an alternative manner. The agent chooses an action according to the ϵ-greedy strategy, which is a commonly-used random strategy in deep reinforcement learning. The agent randomly selects an action with probability ϵ, and executes the strategy of softmax action selection with probability $1-\epsilon$ [40]. Under softmax policy, the possibility of the agent choosing action a is calculated as:

$$P(a_t = a) = \frac{e^{\delta Q(s_t,a)}}{\sum_{a' \in \mathcal{A}} e^{\delta Q(s_t,a')}} \tag{19}$$

where δ is the environment factor, $Q(s_t, a)$ is the Q-value of action a under state s_t.

Algorithm 2: Deep Q-learning for power and sub-band selection.

Initialize network structure θ and θ' with random weight and initialize the parameters: γ, ϵ, δ, M, M_B;
each agent randomly selects an action and observes the initial state s_0;
for $t = 1, 2, ...$ **do**
  for *each agent* **do**
    1) Choose action a_t by ϵ-greedy policy;
    2) Execute the action and receives reward r_t and the environment transits to the next state s_{t+1};
    3) Store (s_t, a_t, r_t, s_{t+1}) into the replay buffer;
    4) Randomly select the mini-batch set from the replay buffer and minimize the loss function (17) using gradient-descent method and update the primary network with θ;
  end
  Every T time slots, copy the weight of primary network to target network for each agent;
end

5. Numerical Results

In this section, we analyze the convergence and effectiveness of the proposed algorithm. We assume that the target area is a circle with radius of 500 m, where a UAV with fixed height is deployed as UAV-BS. Fixing the height and the coverage radius of the UAV, 128 users are randomly distributed in the considered area, and 64 users are located in the coverage of UAV served by air-ground links, while 64 users are located outside the coverage, which can only be served by relay links. The system bandwidth is 10 MHz, which can be shared by inside users and divided into multiple sub-bands. We set the number of available sub-bands for each cluster to four, which means each sub-band occupies a quarter of the bandwidth allocated to each cluster. Hence, the bandwidth of each sub-band is related to the number of clusters. For example, the bandwidth of each sub-band under four clusters is twice of that under eight clusters.

To characterize the performance gain of proposed NOMA, we compared the performance of the proposed algorithm with the following algorithms:

- Game theory-based NOMA (GTB-NOMA) [13]: the channel allocation scheme in NOMA transmission is based on game theory;
- Channel gain-based NOMA (CGB-NOMA) [14]: the channel allocation scheme in NOMA transmission is based on the channel gain difference between users;
- Time division-based NOMA (TD-NOMA) [17]: the devices are served by UAV in a time-division manner using deep reinforcement learning;
- Dynamic power allocation-based NOMA (DPA-NOMA) [18]: the user clustering in NOMA is optimized and a dynamic power allocation is proposed for NOMA users;

- OMA: the bandwidth is allocated to users equally and each inside user occupies a narrow orthogonal frequency band, which means there is no interference among users.

The simulation is conducted by using MATLAB/Simulink; environment modeling in MATLAB and simulink training with a deep reinforcement learning algorithm is adopted. In the proposed DQN, the neural network structure is constituted by an input layer, two fully connected layers and an output layer, and each fully connected layer has 50 neurons. The size of replay buffer and mini-batch are set to 30,000 and 32, respectively. At the beginning of the training stage, ϵ is initialized to 0.9 for extensive exploration of all the possible actions under different states, and gradually decreases to 0.1 as the training progresses to speed up the convergence. The discount factor γ is set to 0.9, and the target network is updated by copying the weight from the primary network every 200 time slots. The execution of DQN is performed in epochs, where each epoch consists of 100 time slots and the final state of current epoch is delivered into the next epoch as the input state. The tuning factor η is set as follows: in the objective function (8), when the user's transmission power is set to 25 dBm and the user's QoS is satisfied, the power term of the inside users is equal to the QoS satisfaction term of the users. When $\omega = 0.5$, $\eta = 50$, and η is dynamically adjusted according to different value of ω. The detailed system parameters are shown in Table 3.

We first verify the convergence of the proposed NOMA algorithm by showing the change of the average real-time reward of clusters during the DQN training process. The number of users is fixed at 32. It can be seen from Figure 4 that the average reward under proposed NOMA outperforms OMA under different numbers of clusters, P. At the beginning of training, the agents tend to select random actions to traverse the environment and evaluate the Q-value of the possible actions. With the continuous increase of experience samples and the improvement of DQN's approximation of the Q-function, the real-time reward increases rapidly. As the learning progresses, the strategy gradually stabilizes, and the real-time reward converges. After several epochs, the DQN curve tends to converge, which means that the optimal strategy has been learned. Another observation is that in the case of $P = 4$, it takes a longer time (around 135 epochs) to converge than the case of $P = 8$. When $P = 4$, the users in the same cluster is increased to 16, the interaction process between each agent and the environment is more complicated due to the increased number of agents, and therefore the speed of learning the optimal strategy slows down.

Next, we explore the impact of several key factors on system performance, including the number of power levels and clusters, the number of users in the system, and the QoS requirement for users.

In Figures 5 and 6, we explore the power consumption and QoS satisfaction changes with the number of clusters, P, and the number of power levels, Q. In Figure 5, we plot the average transmit power for our proposed NOMA under different P and Q. When the number of clusters P is fixed, the power consumption decreases as the number of power levels Q increases. The reason for this is the agent is more declined to select a lower power level under the motivation to increase the reward. Under this circumstance, more power levels enable the agent to have more actions to select in the learning process, and the agent tends to reduce its own power consumption while ensuring QoS to increase the reward. When Q is fixed, the power consumption of $P = 4$ clusters is greater than that of $P = 8$. As the number of users in the cluster is reduced, the number of users in each sub-band is reduced, and the interference level among users sharing the same sub-band is reduced. In this case, complex SIC is unnecessary. Therefore, compared with $P = 8$, an upper transmission power is needed to meet the QoS requirements of users, thus increasing the energy consumption.

Table 3. Simulation Parameters.

Parameter	Value
Area radius	500 m
System bandwidth	10 MHz
Memory size, M	30,000
Mini-batch size, M_B	32
Discount rate, γ	0.9
Network update frequency, T	100
Pathloss exponent, α	2
Probability in ϵ-greedy, ϵ	$0.9 \rightarrow 0.1$
Attenuation factor for NLOS links, β	0.01
Shadow fading variance, σ_{SF}^2	6 dB
Environment constant, a	11.95
Environment constant, b	0.136
Temperature in softmax, δ	1
Tuning factor, η	50
Transmit power for CU, (p_{max}^C, p_{min}^C)	(27,20) dBm
Transmit power for SU, (p_{max}^I, p_{min}^I)	(23,18) dBm
Maximum number of accepted users, N_{max}	4
Pathloss threshold, γ	97.31
Minimum rate for CU, R_{min}^C	200 bit/s
Minimum rate for SU, R_{min}^I	80 bit/s

Figure 6 shows how the QoS satisfaction varies with different P and Q. QoS satisfaction is defined as the ratio of the number of users whose QoS requirement is satisfied to the total number of users. In the case of $P = 4$, when Q increases from 7 to 14, the agent has more feasible actions to select, thereby increasing the probability of increasing the transmission rate and reward. More power levels are beneficial to SIC, thereby increasing the QoS satisfaction. However, when the number of clusters is increased to eight, there is only a slight difference in the QoS satisfaction under two different power levels. This is because there are only four users in each cluster, sharing two sub-bands. Therefore, the interference between users is trivial and the interfering signals can be readily differentiated by SIC, which guarantees that most users can reach their QoS requirements.

In Figures 7 and 8, we evaluate the impact of the number of users on system performance. The average transit power per user and average QoS satisfaction per user under two algorithms are presented, respectively. It can be seen from the figures that as the number of users increases from 16 to 64, the average transmit power per user and QoS satisfaction per user both decrease. Compared with state-of-art NOMA schemes and OMA scheme, the proposed NOMA can achieve better performance in terms of transmit power and QoS satisfaction, which further validates the effectiveness of the proposed algorithm. For the proposed NOMA, as users increase, the number of users in each cluster increase. In this case, the interference within the cluster increases, resulting in the decrease in QoS satisfaction of the network. However, NOMA can provide a larger sub-band bandwidth than OMA, and reduce intra-cluster interference using SIC. Therefore, better performance can be achieved under NOMA compared with OMA. In addition, the proposed NOMA also outperforms state-of-art NOMA schemes thanks to the improved power and sub-band allocation strategy obtained by DRL. We also explore the effect of ω on system performance. As can be seen, when ω is increased from 0.3 to 0.7, the average transmit power decreases and the QoS satisfaction decreases. The reason is that the increased ω makes the reward function put more emphasis on saving energy, hence the transmit power decreases. On the other hand, the emphasis on QoS satisfaction is lowered, which results in degraded QoS satisfaction.

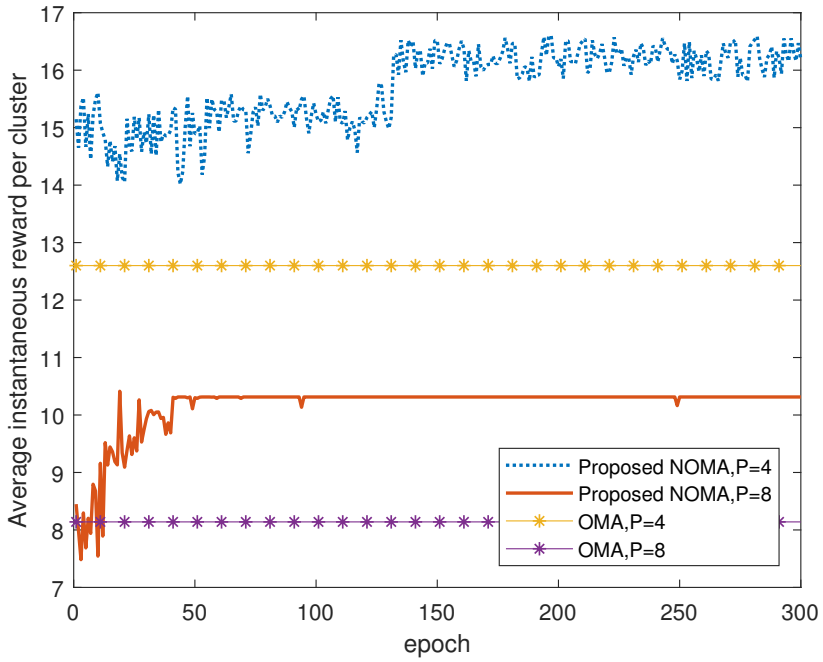

Figure 4. Convergence of the proposed algorithm.

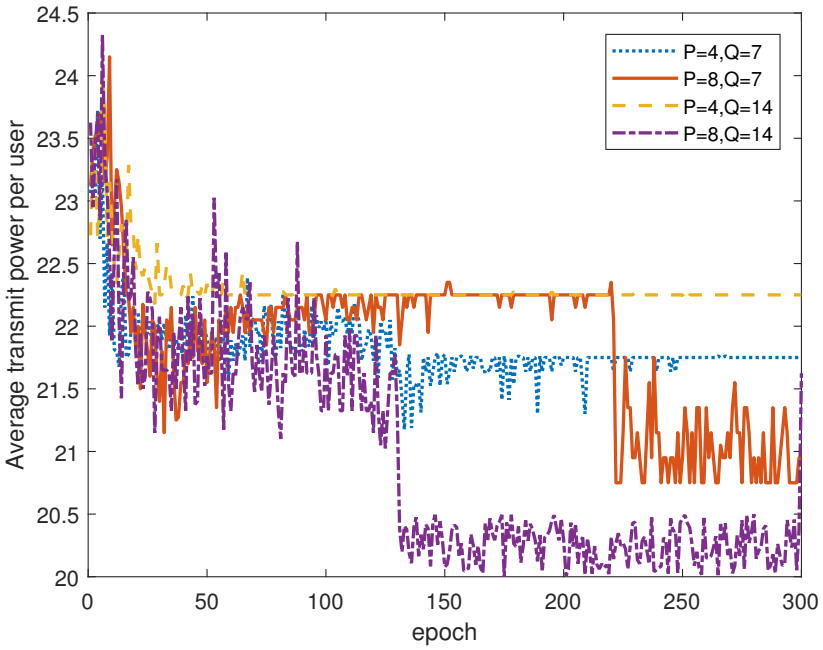

Figure 5. Impact of different number of clusters and power levels on average transmit power.

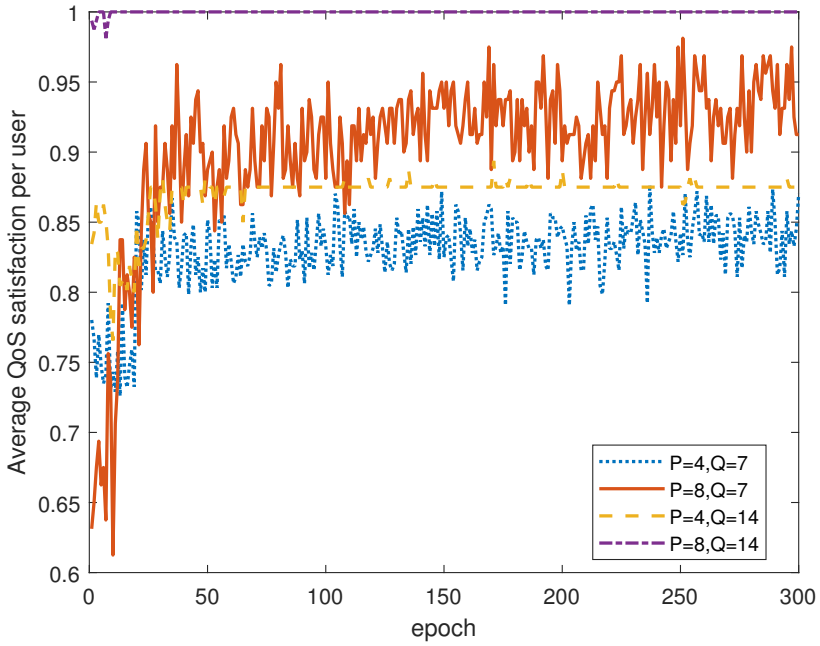

Figure 6. Impact of different number of clusters and power levels on QoS satisfaction.

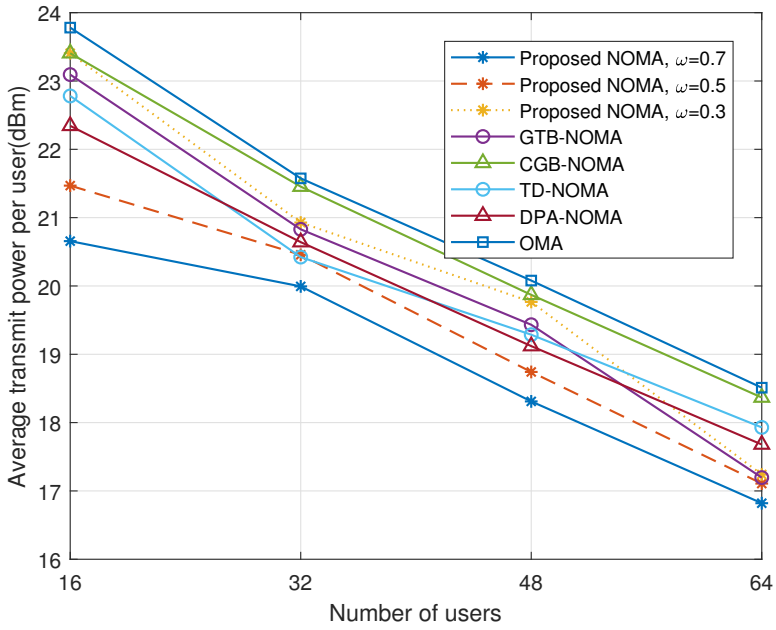

Figure 7. Impact of different number of users on average transmit power.

In Figure 9, we fix $R^C_{min} = 180$ bit/s and explore the impact of different R^I_{min} on system performance. We fix the number of clusters $P = 8$ and the number of power levels $Q = 7$, and investigate the power consumption of proposed NOMA. It can be seen from the figure that as the user's QoS requirement increases, the users' power consumption gradually increases. When R^I_{min} is increased from 80 bit/s to 200 bit/s, the power consumption increases. The reason for this is that as the required rate increases, it becomes more difficult to reach R^I_{min}, especially for inside users acting as relays whose rate requirements become even larger. Hence, the users can only increase the transmit power to increase the transmission rate, which leads to the growth of power consumption. However, as R^I_{min} increases from 200 bit/s to 240 bit/s, energy consumption decreases. The reason for this is that when R^I_{min} is too large, the QoS requirements of a large number of users cannot

be satisfied. From the perspective of increasing instantaneous reward, the agents tend to lower their transmit power, which leads to the reduced energy consumption.

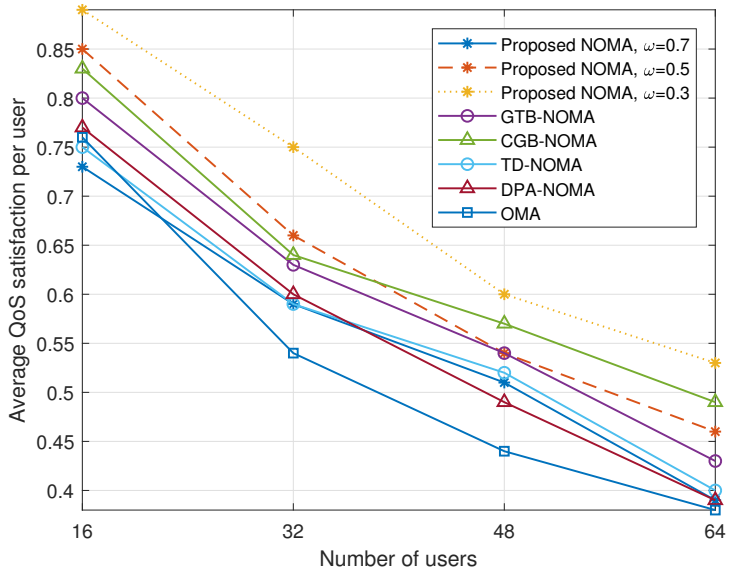

Figure 8. Impact of different number of users on QoS satisfaction.

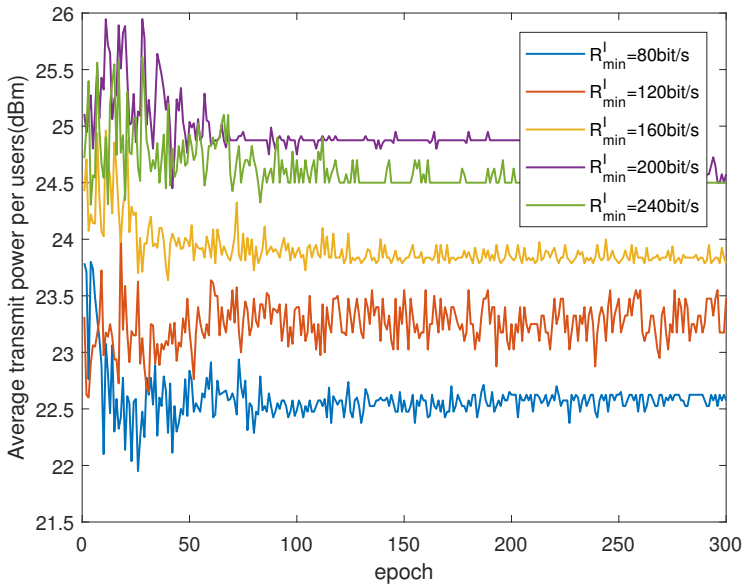

Figure 9. Impact of different QoS requirements on average transmit power.

6. Conclusions

In this paper, we consider a single-UAV heterogeneous network where sensor users and cellular users coexist and share the same frequency band. To enlarge the coverage area of the UAV, we propose to build relays for remote users that cannot be covered by the UAV-BS. In such a scenario, we optimize the network performance from the perspective of saving energy while guaranteeing the QoS. First, we propose an energy-effective relay-selection algorithm for outside users based on matching theory. Next, we propose a NOMA-based transmission scheme for inside users. Then, we formulate the problem of joint power and sub-band selection for inside users and use DQN to solve the problem. The simulation results show that our proposed NOMA transmission scheme can effectively decrease the energy consumption and improve the QoS satisfaction compared to benchmark schemes.

However, in this paper, we mainly focus on the scenario where relay transmission is required due to insufficient coverage in the single UAV-assisted network. For simplicity, we ignore the complex interference problem under multi-UAV networks. At the same time, we assume that the UAV is stationary rather than mobile. In the future, we plan to perform in-depth research of the following aspects: (1) the network scenario can be extended to a multi-UAV network, and the power and resource-allocation scheme under multiple UAV networks can be designed based on deep reinforcement learning. (2) The optimization of the UAV's trajectory can be designed, and the impact of UAV's movement on the user's energy consumption and QoS should be explored. (3) To further extend the coverage of UAV, we will investigate the case that the multi-hop (the number of hop > 2) transmission in relay transmission can be considered, and the transmission delay, user energy consumption, and QoS should be simultaneously considered.

Author Contributions: Conceptualization, J.Z. and G.C.; methodology, J.Z. and Z.Z.; software, J.Z.; validation, W.G., G.C. and Z.Z.; formal analysis, Z.Z.; investigation, J.Z.; resources, W.G.; data curation, W.G.; writing—original draft preparation, J.Z.; writing—review and editing, G.C.; visualization, W.G.; supervision, Z.Z.; project administration, Z.Z.; funding acquisition, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by grants from the National Key Research and Development Program of China (Grant No. 2020YFC2006200).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the editors and the reviewers for their constructive comments.

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

Abbreviations

The following abbreviations are used in this manuscript:

BS	base station
CU	cellular user
D2D	device-to-device
DL	deep learning
DQL	deep Q-learning
DQN	deep Q-network
DRL	deep reinforcement learning
DAF	decode and forward
EE	energy efficiency
IoT	Internet of Things
IoE	Internet of Everything
LoS	line-of-sight
MDP	markov decision process
NOMA	non-orthogonal multiple access
OMA	orthogonal multiple access
QoS	quality of service
QL	Q-learning
RB	resource block
RL	reinforcement learning
SE	spectrum efficiency
SU	sensor users
SINR	Signal-to-Interference-plus-Noise Ratio
UAV	unmanned aerial vehicle

References

1. Cisco. *Cisco Annual Internet Report (2018–2023)*; Cisco: San Francisco, CA, USA, 2020; pp. 1–41.
2. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2347–2376. [CrossRef]
3. Xu, L.D.; He, W.; Li, S. Internet of Things in Industries: A Survey. *IEEE Trans. Ind. Inform.* **2014**, *10*, 2233–2243. [CrossRef]
4. Samir, M.; Sharafeddine, S.; Assi, C.M.; Nguyen, T.M.; Ghrayeb, A. UAV Trajectory Planning for Data Collection from Time-Constrained IoT Devices. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 34–46. [CrossRef]
5. Pasandideh, F.; da Costa, J.P.J.; Kunst, R.; Islam, N.; Hardjawana, W.; Pignaton de Freitas, E. A Review of Flying Ad Hoc Networks: Key Characteristics, Applications, and Wireless Technologies. *Remote Sens.* **2022**, *14*, 4459. [CrossRef]
6. Zhang, J.; Chuai, G.; Gao, W. Energy-Efficient Optimization for Energy-Harvesting-Enabled mmWave-UAV Heterogeneous Networks. *Entropy* **2022**, *24*, 300. [CrossRef]
7. Mase, K.; Okada, H. Message communication system using unmanned aerial vehicles under large-scale disaster environments. In Proceedings of the 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong, China, 30 August–2 September 2015; pp. 2171–2176. [CrossRef]
8. Zhang, J.; Chuai, G.; Gao, W.; Saidi, M.; Si, Z. Coalition Game-Based Beamwidth Selection for D2D Users Underlying Ultra Dense mmWave Networks. In Proceedings of the 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Seoul, Republic of Korea, 6–9 April 2020; pp. 1–6. [CrossRef]
9. Liu, D.; Xu, Y.; Wang, J.; Xu, Y.; Anpalagan, A.; Wu, Q.; Wang, H.; Shen, L. Self-Organizing Relay Selection in UAV Communication Networks: A Matching Game Perspective. *IEEE Wirel. Commun.* **2019**, *26*, 102–110. [CrossRef]
10. Lhazmir, S.; Oualhaj, O.A.; Kobbane, A.; Ben-Othman, J. UAV for Energy-Efficient IoT Communications: Matching Game Approach. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6. [CrossRef]
11. Duan, R.; Wang, J.; Jiang, C.; Yao, H.; Ren, Y.; Qian, Y. Resource Allocation for Multi-UAV Aided IoT NOMA Uplink Transmission Systems. *IEEE Internet Things J.* **2019**, *6*, 7025–7037. [CrossRef]
12. Liu, M.; Yang, J.; Gui, G. DSF-NOMA: UAV-Assisted Emergency Communication Technology in a Heterogeneous Internet of Things. *IEEE Internet Things J.* **2019**, *6*, 5508–5519. [CrossRef]
13. Li, Y.; Zhang, H.; Long, K.; Choi, S.; Nallanathan, A. Resource Allocation for Optimizing Energy Efficiency in NOMA-based Fog UAV Wireless Networks. *IEEE Netw.* **2020**, *34*, 158–163. [CrossRef]
14. Sohail, M.F.; Leow, C.Y.; Won, S. Energy-Efficient Non-Orthogonal Multiple Access for UAV Communication System. *IEEE Trans. Veh. Technol.* **2019**, *68*, 10834–10845. [CrossRef]
15. Na, Z.; Liu, Y.; Shi, J.; Liu, C.; Gao, Z. UAV-Supported Clustered NOMA for 6G-Enabled Internet of Things: Trajectory Planning and Resource Allocation. *IEEE Internet Things J.* **2021**, *8*, 15041–15048. [CrossRef]
16. Liu, X.; Liu, Z.; Zhou, M. Fair Energy-Efficient Resource Optimization for Green Multi-NOMA-UAV assisted Internet of Things. *IEEE Trans. Green Commun. Netw.* **2021**. [CrossRef]
17. Mrad, A.; Al-Hilo, A.; Sharafeddine, S.; Assi, C. NOMA-Aided UAV Data Collection from Time-Constrained IoT Devices. In Proceedings of the ICC 2022—IEEE International Conference on Communications, Seoul, Republic of Korea, 4 April 2022; pp. 1–6. [CrossRef]
18. Abdel-Razeq, S.; Shakhatreh, H.; Alenezi, A.; Sawalmeh, A.; Anan, M.; Almutiry, M. PSO-based UAV deployment and dynamic power allocation for UAV-enabled uplink NOMA network. *Wirel. Commun. Mob. Comput.* **2021**, *5*, 1–17. [CrossRef]
19. Mozaffari, M.; Saad, W.; Bennis, M.; Debbah, M. Mobile Internet of Things: Can UAVs Provide an Energy-Efficient Mobile Architecture? In Proceedings of the 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, USA, 4–8 December 2016; pp. 1–6. [CrossRef]
20. Zhan, C.; Lai, H. Energy Minimization in Internet-of-Things System Based on Rotary-Wing UAV. *IEEE Wirel. Commun. Lett.* **2019**, *8*, 1341–1344. [CrossRef]
21. Gu, J.; Wang, H.; Ding, G.; Xu, Y.; Xue, Z.; Zhou, H. Energy-Constrained Completion Time Minimization in UAV-Enabled Internet of Things. *IEEE Internet Things J.* **2020**, *7*, 5491–5503. [CrossRef]
22. Eom, S.; Lee, H.; Park, J.; Lee, I. UAV-Aided Wireless Communication Designs with Propulsion Energy Limitations. *IEEE Trans. Veh. Technol.* **2020**, *69*, 651–662. [CrossRef]
23. Liu, Y.; Liu, K.; Han, J.; Zhu, L.; Xiao, Z.; Xia, X.G. Resource Allocation and 3-D Placement for UAV-Enabled Energy-Efficient IoT Communications. *IEEE Internet Things J.* **2021**, *8*, 1322–1333. [CrossRef]
24. Li, Z.; Wang, Y.; Liu, M.; Sun, R.; Chen, Y.; Yuan, J.; Li, J. Energy Efficient Resource Allocation for UAV-Assisted Space-Air-Ground Internet of Remote Things Networks. *IEEE Access* **2019**, *7*, 145348–145362. [CrossRef]
25. Lee, J.; Friderikos, V. Trajectory Planning for Multiple UAVs in UAV-aided Wireless Relay Network. In Proceedings of the ICC 2022—IEEE International Conference on Communications, Seoul, Republic of Korea, 4 April 2022; pp. 1–6. [CrossRef]
26. Li, B.; Zhao, S.; Zhang, R.; Yang, L. Joint Transmit Power and Trajectory Optimization for Two-Way Multihop UAV Relaying Networks. *IEEE Internet Things J.* **2022**, *9*, 24417–24428. [CrossRef]
27. Zhang, G.; Ou, X.; Cui, M.; Wu, Q.; Ma, S.; Chen, W. Cooperative UAV Enabled Relaying Systems: Joint Trajectory and Transmit Power Optimization. *IEEE Trans. Green Commun. Netw.* **2022**, *6*, 543–557. [CrossRef]

28. Wang, B.; Zhang, R.; Chen, C.; Cheng, X.; Yang, L.; Li, H.; Jin, Y. Graph-Based File Dispatching Protocol with D2D-Enhanced UAV-NOMA Communications in Large-Scale Networks. *IEEE Internet Things J.* **2020**, *7*, 8615–8630. [CrossRef]
29. Liu, X.; Gui, G.; Zhao, N.; Meng, W.; Li, Z.; Chen, Y.; Adachi, F. UAV Coverage for Downlink in Disasters: Precoding and Multi-hop D2D. In Proceedings of the 2018 IEEE/CIC International Conference on Communications in China (ICCC), Beijing, China, 16–18 August 2018; pp. 341–346. [CrossRef]
30. Liu, X.; Li, Z.; Zhao, N.; Meng, W.; Gui, G.; Chen, Y.; Adachi, F. Transceiver Design and Multihop D2D for UAV IoT Coverage in Disasters. *IEEE Internet Things J.* **2019**, *6*, 1803–1815. [CrossRef]
31. Krichen, M.; Adoni, W.Y.H.; Mihoub, A.; Alzahrani, M.Y.; Nahhal, T. Security Challenges for Drone Communications: Possible Threats, Attacks and Countermeasures. In Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH), Riyadh, Saudi Arabia, 22–24 May 2022; pp. 184–189. [CrossRef]
32. Pandey, G.K.; Gurjar, D.S.; Nguyen, H.H.; Yadav, S. Security Threats and Mitigation Techniques in UAV Communications: A Comprehensive Survey. *IEEE Access* **2022**, *10*, 112858–112897. [CrossRef]
33. Bera, B.; Chattaraj, D.; Das, A.K. Designing secure blockchain based access control scheme in iot-enabled internet of drones deployment. *Comput. Commun.* **2020**, *153*, 229–249. [CrossRef]
34. Gupta, R.; Kumari, A.; Tanwar, S. Fusion of blockchain and artificial intelligence for secure drone networking underlying 5g communications. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4176. [CrossRef]
35. Wu, T.; Guo, X.; Chen, Y.; Kumari, S.; Chen, C. Amassing the Security: An Enhanced Authentication Protocol for Drone Communications over 5G Networks. *Drones* **2022**, *6*, 10. [CrossRef]
36. Alladi, T.; Naren; Bansal, G.; Chamola, V.; Guizani, M. SecAuthUAV: A Novel Authentication Scheme for UAV-Ground Station and UAV-UAV Communication. *IEEE Trans. Veh. Technol.* **2020**, *69*, 15068–15077. [CrossRef]
37. Wang, Z.; Guo, J.; Chen, Z.; Yu, L.; Wang, Y.; Rao, H. Robust secure UAV relay-assisted cognitive communications with resource allocation and cooperative jamming. *J. Commun. Netw.* **2022**, *24*, 139–153. [CrossRef]
38. Lin, X.; Su, G.; Chen, B.; Wang, H.; Dai, M. Striking a Balance Between System Throughput and Energy Efficiency for UAV-IoT Systems. *IEEE Internet Things J.* **2019**, *6*, 10519–10533. [CrossRef]
39. Elbayoumi, M.; Kamel, M.; Hamouda, W.; Youssef, A. NOMA-Assisted Machine-Type Communications in UDN: State-of-the-Art and Challenges. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1276–1304. [CrossRef]
40. Sutton, R.; Barto, R. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, UK, 2018.
41. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174. [CrossRef]
42. Lin, Y.; Wang, M.; Zhou, X.; Ding, G.; Mao, S. Dynamic Spectrum Interaction of UAV Flight Formation Communication with Priority: A Deep Reinforcement Learning Approach. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 892–903. [CrossRef]
43. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.; Veness, J.; Bellemare, M.; Graves, A.; Riedmiller, M.; Fidjeland, A.; Ostrovski, G.; et al. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef]