drones

MDPI

Article

Three-Dimensional Trajectory and Resource Allocation Optimization in Multi-Unmanned Aerial Vehicle Multicast System: A Multi-Agent Reinforcement Learning Method

Dongyu Wang [1,*], Yue Liu [1], Hongda Yu [1] and Yanzhao Hou [2]

[1] The Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China; leomoon@bupt.edu.cn (Y.L.); yhd0905@bupt.edu.cn (H.Y.)
[2] Shenzhen Institute, Beijing University of Posts and Telecommunications, Shenzhen 518055, China; houyanzhao@bupt.edu.cn
* Correspondence: dy_wang@bupt.edu.cn

Abstract: Unmanned aerial vehicles (UAVs) are able to act as movable aerial base stations to enhance wireless coverage for edge users with poor ground communication quality. However, in urban environments, the link between UAVs and ground users can be blocked by obstacles, especially when complicated terrestrial infrastructures increase the probability of non-line-of-sight (NLoS) links. In this paper, in order to improve the average throughput, we propose a multi-UAV multicast system, where a multi-agent reinforcement learning method is utilized to help UAVs determine the optimal altitude and trajectory. Intelligent reflective surfaces (IRSs) are also employed to reflect signals to solve the blocking problem. Furthermore, since the UAV's onboard power is limited, this paper aims to minimize the UAVs' energy consumption and maximize the transmission rate for edge users by jointly optimizing the UAVs' 3D trajectory and transmit power. Firstly, we deduce the channel capacity of ground users in different multicast groups. Subsequently, the K-medoids algorithm is utilized for the multicast grouping problem of edge users based on transmission rate requirements. Then, we employ the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm to learn an optimal solution and eliminate the non-stationarity of multi-agent training. Finally, the simulation results show that the proposed system can increase the average throughput by 14% approximately compared to the non-grouping system, and the MADDPG algorithm can achieve a 20% improvement in reducing the energy consumption of UAVs compared to traditional deep reinforcement learning (DRL) methods.

Keywords: unmanned aerial vehicles (UAVs); trajectory optimization; power allocation; multicast; intelligent reflecting surface (IRS); multi-agent deep deterministic policy gradient (MADDPG)

Citation: Wang, D.; Liu, Y.; Yu, H.; Hou, Y. Three-Dimensional Trajectory and Resource Allocation Optimization in Multi-Unmanned Aerial Vehicle Multicast System: A Multi-Agent Reinforcement Learning Method. *Drones* **2023**, *7*, 641. https://doi.org/10.3390/drones7100641

Academic Editors: Diego González-Aguilera and Shiva Raj Pokhrel

Received: 30 August 2023
Revised: 28 September 2023
Accepted: 18 October 2023
Published: 19 October 2023

1. Introduction

Recently, unmanned aerial vehicles (UAVs) have attracted much attention due to their flexible position adjustment capability, which enhances the probability of line-of-sight (LoS) communication links to ground users (GUs). In wireless communication systems, UAVs can provide storage and computational resources to alleviate the communication pressure of the whole network. UAVs can be used as relays between GUs and ground base stations (GBSs) in a store-carry-forward manner. They can also serve as aerial base stations to improve the coverage of GBSs. Additionally, overloaded GBSs can offload the traffic to UAVs. Therefore, UAVs are widely employed in scenarios such as disaster management, traffic monitoring, and emergency rescue, assuming the roles of data coverage, data collection, and information transmission.

In this paper, UAVs are considered to act as aerial base stations. When utilized as an aerial mobile base station to serve GUs, UAVs possess the following advantages [1]:

1. On-demand deployment: While conventional terrestrial base stations are fixed and immovable, UAVs are able to be deployed more flexibly and on-demand in accordance with GUs' locations.
2. Better communication quality: Instead of more obstacle blocking between GBSs and GUs, the air-to-ground link is dominated more by the LoS link.
3. Mobility over time: UAVs have the capacity to move over time, adjusting their positions to satisfy the demands of GUs and enhance communication performance.

Meanwhile, the research about using UAVs as aerial base stations has encountered challenges, as follows:

1. Trajectory design complexity: Since UAVs can move in multiple dimensions and need to be positioned to meet the needs of GUs; optimal deployment and trajectory design strategies need to be addressed.
2. Energy limitation: How to optimize the performance with the restricted energy needs to be taken into account because UAVs consume energy both during their flight and communication.
3. Signal blocking: In practical applications, the air-to-ground link is more likely to be blocked by territorial obstacles when UAVs fly at low altitudes in complicated environments.

Motivated by the above advantages and challenges, this paper considers a multi-cell cellular network in which UAVs act as aerial base stations to expand coverage and improve communication quality for edge GUs at a long distance. Furthermore, IRSs are deployed on the surface of ground buildings to reflect the UAVs' signals to GUs.

For GUs, various users have different transmission rate requirements. Some GUs request high-latency-sensitivity services such as virtual reality and Internet of Vehicles (IoV), while others only request fundamental data computing services. In order to improve service efficiency, multicast channels with fixed transmitters have been extensively studied in wireless communications [2]. The capacity of a multicast channel is determined by the data rate of the GU with the worst channel quality in order for all users to successfully decode the common information from the transmitter. On one hand, the transmission rate is constrained if all the edge GUs are considered as one multicast group, which leads to more energy consumed by UAVs for hovering and transmission. On the other hand, when the UAV communicates with each GU in unicast mode, a higher transmission rate can be guaranteed. However, the UAV will consume more energy to travel longer distances. In order to trade off the two schemes, the GUs are divided into multiple multicast groups based on the transmission rate requirement in this paper. As a result, the average throughput of the entire system is improved since the transmission rate of each multicast group is determined by the GU with the worst channel condition within the group. UAVs can act as mobile transmitters [2] and leverage Orthogonal Frequency Division Multiple Access (OFDMA) technology to achieve multicast communication, thereby reducing energy consumption while overcoming user bottlenecks.

In contrast to conventional fixed GBSs, UAVs can adjust their positions horizontally and vertically over time, thus improving the wireless channel quality of users with different locations and transmission rate requirements. Therefore, optimizing the deployment location and trajectory of UAVs has become an important topic. The majority of current research focuses on optimizing UAV flight trajectory with a fixed altitude, ignoring the impact of NLoS links when UAVs fly at a low altitude. In terms of vertical altitude, for GUs with high transmission rate requirements, UAVs can appropriately shorten the communication link distance, thus increasing link capacity or saving transmit power. However, as shown in Figure 1, the lower the UAV's altitude is, the more dominant the NLoS link will be [3]. Signals emitted by the aerial base station and received by the UAV propagate through the LoS link until reaching the urban environment, where additional loss is incurred due to shadowing and scattering caused by obstacles, such as buildings. Therefore, the 3D trajectory optimization problem needs to be tackled to design the optimal altitude over time.

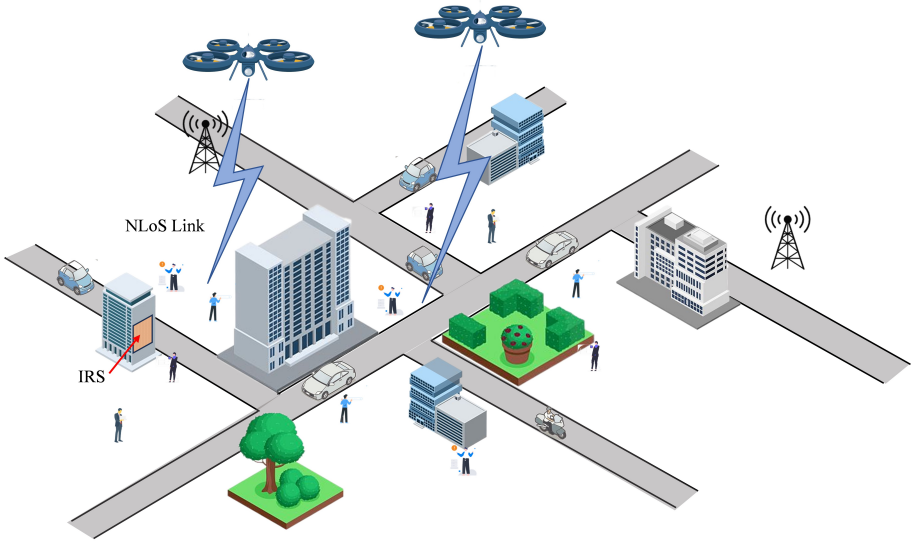

Figure 1. Signal propagation of UAV in urban environment.

In summary, this paper investigates the joint optimization problem of UAV 3D trajectory and power allocation in edge regions in a multi-UAV multicast network. GUs are categorized into multiple multicast groups, and each UAV serves one group at a time slot. The objective of this paper is to minimize the energy consumption of UAVs. To achieve this goal, GUs are first divided into different multicast groups depending on the relative distances and transmission rate requirements. This grouping issue is resolved by the K-medoids algorithm [4,5]. Next, it is important to find the 3D trajectory and power allocation that minimize UAVs' energy consumption. This is known as an NP-hard problem because it involves coupled optimization variables, such as the UAVs' association with the multicast group, the transmission power, the UAV data rate, and the UAV trajectory. The problem also consists of several non-convex constraints. Since multiple UAVs are considered in the system and each UAV needs to learn a policy, we utilize multi-agent deep reinforcement learning (MADRL) to solve the problem. MADRL can address the problem that the environment becomes non-stationary from the perspective of any individual agent with each agent's policy updated [6]. For each agent, the Deep Deterministic Policy Gradient (DDPG) algorithm is applied which can learn policies in high-dimensional and continuous action spaces [7]. The main contributions in this paper are summarized as follows:

1. We propose a multi-UAV multicast system assisted by IRSs in which we formulate the multicast grouping problem and an optimization problem aiming to minimize the UAVs' energy consumption.
2. We utilize the K-medoids algorithm [4,5] to solve the multicast grouping problem, and the MADRL framework is employed to efficiently obtain the optimal 3D trajectory and power allocation scheme and eliminate the non-stationarity of multi-agent training.
3. The performance can be evaluated by the provided simulation results. We verify that the proposed system can improve the average throughput, and the MADDPG algorithm can reduce the energy consumption of UAVs effectively compared to traditional DRL methods.

The rest of the paper is organized as follows. The literature review is presented in Section 2. Section 3 introduces the system model. Section 4 states the problem description and formulates the optimization objective. Section 5 illustrates the solution, including the grouping algorithm and the optimization algorithm. The simulation results are discussed in Section 6. Finally, Section 7 concludes this paper. The methodology is shown in the following Figure 2.

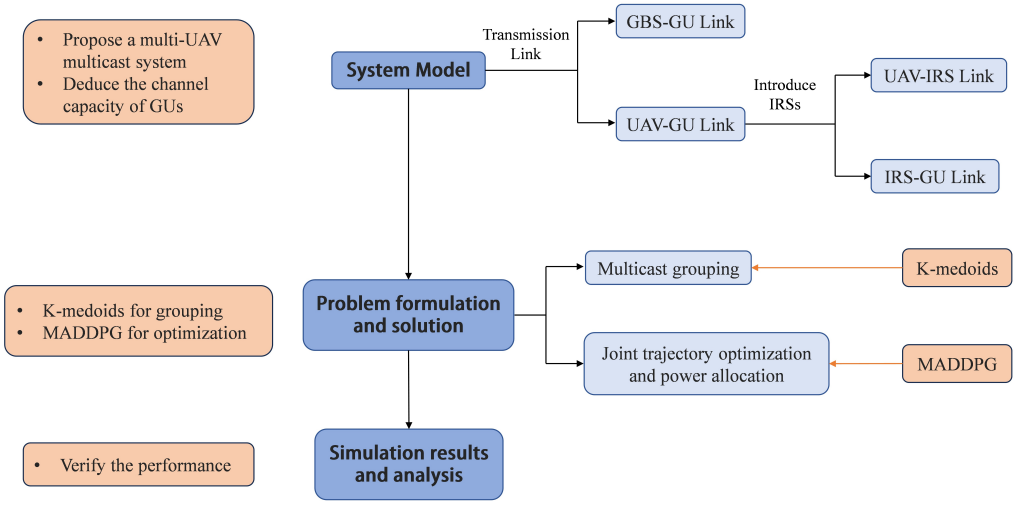

Figure 2. Flowchart of the methodology in this paper.

2. Related Works

In light of the challenges mentioned in Section 1, we investigated some of the literature on UAVs serving as aerial base stations in wireless communication systems. The literature can be reviewed from two perspectives: the problem statement and the optimization methods.

As for the problem statement, Deng et al. [8] investigated UAVs as aerial base stations to multicast public information to GUs and proposed a machine learning method to jointly optimize the multicast grouping and trajectory planning scheme. However, they solely considered a single-UAV network. Chen et al. [9] studied multiple UAVs as aerial base stations to enhance the coverage of cellular networks and proposed a decentralized joint trajectory and power control (DTPC) algorithm to minimize the UAVs' overall energy consumption. Heuristic algorithms were proposed to formulate the trajectory optimization problem to minimize the total travel time of the UAVs in the multi-cell network in [10,11], where UAVs are employed as flying base stations to serve GUs in collaboration with GBSs. However, none of the above papers considered the multicast grouping problem based on user features or a situation in which the UAV-GU communication link may be blocked in practical low-altitude environments. In [12–17], intelligent reflection surfaces (IRSs) were introduced into UAV systems to assist the transmission between UAVs and GUs. Nguyen et al. [18] deployed IRSs to combat air-to-ground (A2G) blockage events and derived closed-form expressions for Signal-to-Interference-Plus-Noise-Ratio (SINR) distributions. The literature shows that IRSs can reduce blockage effectively in the UAV-GU communication link.

As for optimization methods, heuristic algorithms are typically used to solve the trajectory optimization problem. In [10,11], a set of local search heuristic algorithms was proposed considering the curse of dimensionality problem. Xue et al. [19] proposed a heuristic algorithm based on alternating descent and successive convex approximation (SCA) to solve a joint 3D location and transmit power optimization problem. However, typical heuristic algorithms are more suitable for static optimization problems without taking historical data into consideration. In [20], the authors formulated the problem as Budgeted Multi-Armed Bandits (BMABs) to optimize the UAV trajectory and minimize battery consumption and used two Upper Confidence Bound (UCB) BMAB schemes to tackle the issue. Nowadays, with the development of artificial intelligence technology, more and more researchers use machine learning techniques to solve optimization issues. The interaction between the agent and the environment in reinforcement learning (RL) is more similar to the mechanism by which UAVs make decisions based on the observation. The Double Deep Q-Network (DDQN) and DDPG algorithms were utilized in [12] to solve the UAV trajectory optimization problem in the IRS-assisted UAV system. Fan et al. [21]

proposed a novel multi-agent DRL method with global–local rewards for UAVs' dynamic trajectory planning and data offloading decisions. In the multi-agent case, the intricate process of interactions between agents makes the environment constantly and dynamically changing. The non-stationarity will reduce the stability of the algorithm.

In summary, very little of the literature focuses on both the multicast grouping issue based on users' characteristics and the joint optimization problem. Since machine learning techniques have become the new research trend, MADRL can be utilized to handle the issues in traditional DRL methods.

3. System Model

3.1. System Model

We consider a multi-cell cellular network, as shown in Figure 3, where multiple UAVs serve as aerial base stations to provide access to edge GUs that cannot be effectively served by GBSs. The set of edge GUs is represented by $\mathcal{N} = \{n|n = 1, 2, \ldots, N\}$. The GUs are divided into K multicast groups, with different transmission rate requirements for each group. The set of multicast groups is denoted as $\mathcal{K} = \{k|k = 1, 2, \ldots, K\}$. N_k represents the number of GUs within the kth multicast group, satisfying $N = \sum_{k \in \mathcal{K}}^{N_k}, N_i \bigcap N_j = \varnothing, \forall i \neq j \in \mathcal{N}$, which means there's no overlap between groups. The location of the ith GU is denoted as $\mathbf{L}_i^{GU} = (x_i, y_i), i \in \mathcal{N}$, and the transmission rate is denoted as γ_i. The two characteristics are used as metrics to classify multicast groups. When grouping has been completed, in order to satisfy the transmission rate requirement of all users in each group, the transmission rate requirement of the kth group is denoted as $\gamma_k = \max_{i \in N_k} \gamma_i$.

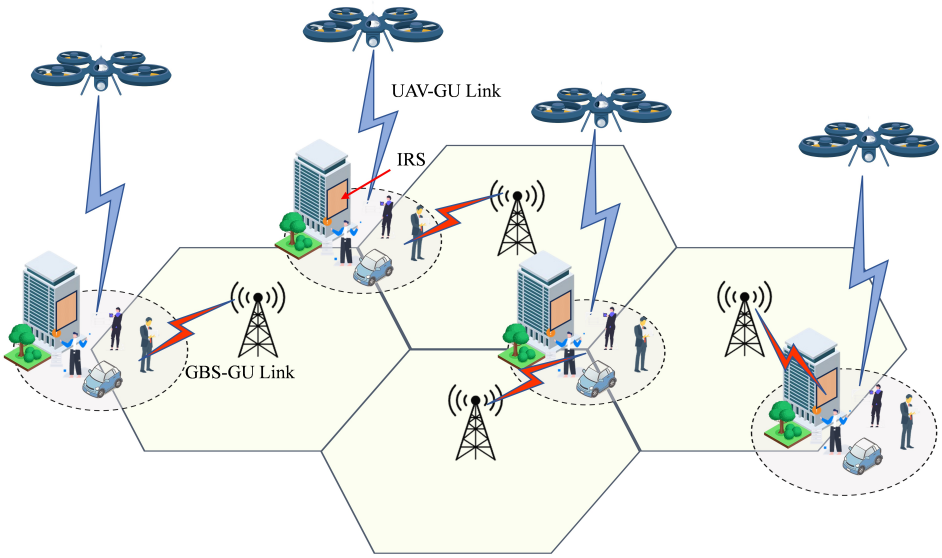

Figure 3. Multi-UAV multicast system model.

UAVs fly within the multi-cell edge region and serve terrestrial multicast groups. Let $\mathcal{T} = \{t|t = 1, 2, \ldots, T\}$ denote the set of service time slots, which is also termed as an episode. The length of each time slot is defined as Δt. Assume that there are U UAVs, and the set of UAVs is represented by $\mathcal{U} = \{u|u = 1, 2, \ldots, U\}$. $\mathbf{L}_{u,t}^{UAV} = (x_{u,t}, y_{u,t}, h_{u,t})$ denotes the trajectory of the uth UAV at time slot t. Each multicast group is served by the closest UAV at each time slot. If the closest UAV is already occupied by its closest group, the second closest UAV will work. We assume that there are sufficient UAVs to ensure that each multicast group is served by one UAV at each time slot.

Meanwhile, in order to solve the signal-blocking problem when UAVs fly at low altitudes, an IRS is deployed on the surface of the building near each multicast group to avoid the NLoS link between UAVs and GUs. Denote $\mathcal{R} = \{r | r = r_1, r_2, \ldots, r_K\}$ as the set of IRSs. r_k is the IRS corresponding to the kth multicast group, and its location is defined as $\mathbf{L}_{r_k}^{IRS} = (x_{r_k}, y_{r_k}, h_{r_k})$. For the IRS, we assume that a uniform planer array (UPA) consists of $M_c \times M_r$ passive reflection units (PRU). Each PRU can passively change its phase-shift with an independent reflection coefficient: $r_{r_k,m_c,m_r} = ae^{j\theta_{r_k,m_r,m_c}}, \forall k \in K$; $\forall m_r \in 1, 2, \ldots, M_r; \forall m_c \in 1, 2, \ldots, M_c$ where $a \in [0, 1]$ is the fixed reflection loss of the IRS and $\theta_{r_k,m_r,m_c} \in [-\pi, \pi)$ is the phase shift inserted at PRU (m_r, m_c) [12].

3.2. Transmission Model

The transmission model consists of two parts: the GBS-GU link and the UAV-GU link. UAVs provide supplemental services when the GBS is unable to meet the needs of the edge GUs. The service procedure of the uth UAV for the multicast group is shown in Figure 4. The uth UAV serves the kth multicast group at time slot t, during which other UAVs serve the other multicast groups. Over time, the UAV continues to fly to serve the next multicast group at time slot $t + 1$.

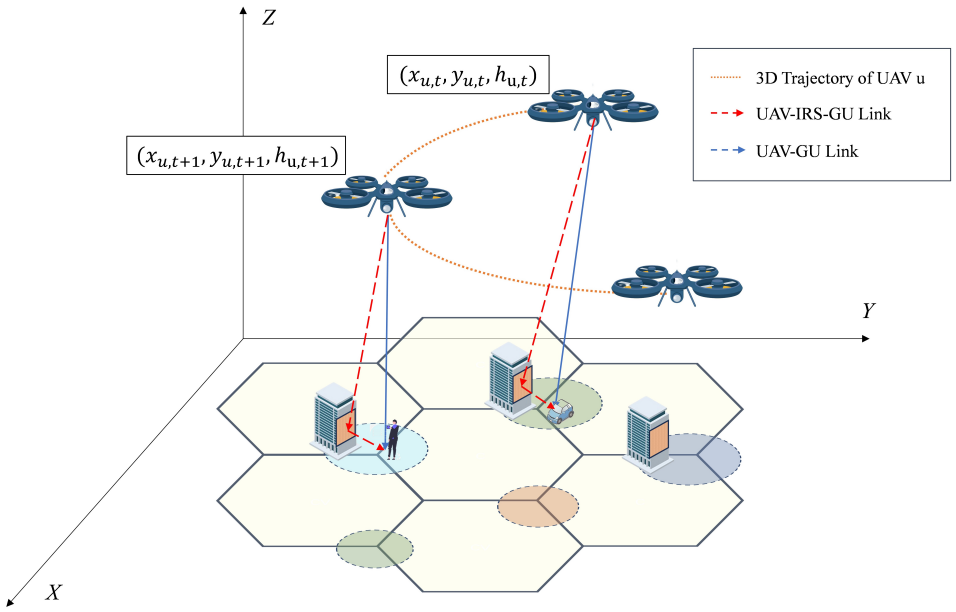

Figure 4. The service process of the uth UAV.

3.2.1. GBS-GU Link

In the GBS-GU link, the ith GU is served by the closest GBS. The GBS-GU link channel is assumed according to [22], where it can be assumed to be a fading channel with a distance-dependent path loss with the exponent $\delta \geqslant 2$ and an additional random term $\zeta_{g,i} \sim Exp(1)$ accounting for small-scale fading. As a result, the received signal-to-noise-ratio (SNR) at the ith GU from the GBS can be expressed as

$$SNR_i^G = \gamma_i^G = \frac{G_G P_G g_0}{\sigma^2 (r_{g,i}^2 + H_G^2)^{\delta/2}} \tag{1}$$

where $i \in \mathcal{N}$; G_G, H_G, r_i, and P_G denote a fixed antenna gain, the height of the GBS, the distance between the ith GU and the GBS, and the transmit power of the GBS, respectively; $g_0 = (\frac{c}{4\pi f_c})^2$ denotes the average channel power gain at a reference distance of $d_0 = 1m$;

and σ^2 denotes the noise power. According to Shannon's theorem, the transmission rate from the GBS to the ith GU is calculated as

$$R_{g,i} = B\log_2(1+\gamma_i^G\zeta_i) \tag{2}$$

where B is the total transmission bandwidth.

3.2.2. GBS-GU Link Outage Probability

An outage occurs when the GBS is unable to meet the ith GU's transmission rate requirement γ_i because of the small-scale fading between the GBS and GUs. The outage probability is expressed as

$$\begin{aligned} P_{g,i} &= Pr\{R_{g,i} < \gamma_i\} \\ &= Pr\{B\log_2(1+\gamma_i^G\zeta_{g,i}) < \gamma_i\} \\ &= Pr\{\zeta_{g,i} < \frac{2^{\gamma_i/B}-1}{\gamma_i^G}\} \\ &= 1 - exp(-\frac{2^{\gamma_i/B}-1}{\gamma_i^G}) \end{aligned} \tag{3}$$

3.2.3. UAV-GU Link

When GBSs are unable to provide reliable communication to the ith GU, the service is provided by UAVs in the multicast mode. In order to avoid NLoS links induced by shadowing and scattering caused by building complexes in urban environments, IRSs are introduced to reflect UAV-GU signals. As shown in Figure 5, the UAV-GU link can be replaced by the two UAV-IRS and IRS-GU links, both of which are LoS connections.

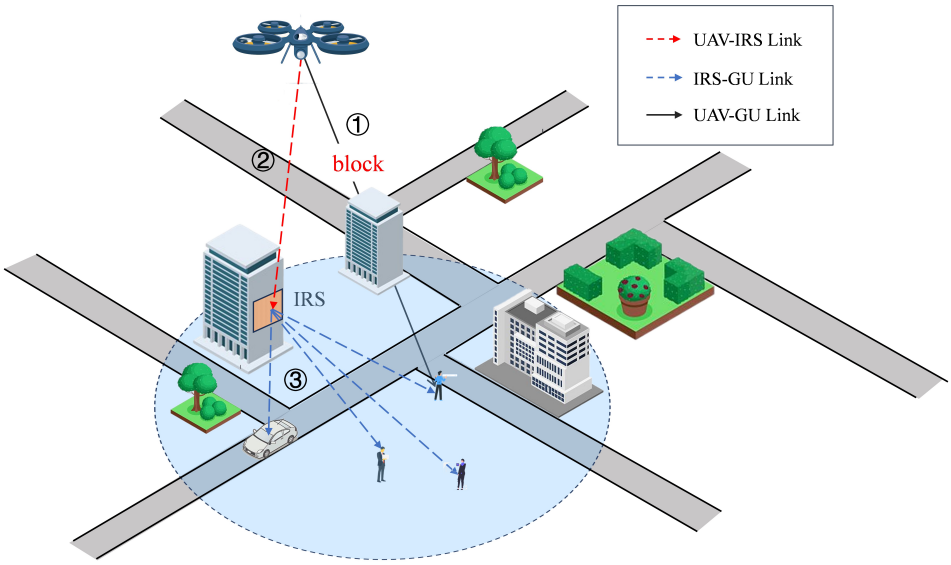

Figure 5. IRS reflects UAV-GU signals in one multicast group.

Similar to [22], we assume the corresponding antenna gain in direction (α, β) as

$$G_U(\alpha,\beta) = \begin{cases} G_0/\Theta_U^2, & -\Theta_U \leqslant \alpha \leqslant \Theta_U, -\Theta_U \leqslant \beta \leqslant \Theta_U \\ g_0 \approx 0, & otherwise \end{cases} \tag{4}$$

where $G_0 = \frac{30000}{2^2} * (\frac{\pi}{180})^2 \approx 2.2846$ and $\Theta_U \in (0, \frac{\pi}{2})$. Thus, the ground coverage region of the UAV's antenna main lobe corresponds to the disk region with a radius $r_{u,t} = h_{u,t}\tan\Theta_U$ that is centered on the projection of the UAV on the ground. Determining the beamwidth

Θ_U, the coverage radius can be adjusted by changing the UAV's altitude $h_{u,t}$ so that the GUs are located within the coverage area of the UAV.

When the signal reaches the urban environment, obstacles, such as building complexes, cause additional losses in the UAV-GU link. Define the outage probability at time slot t of the uth UAV with the ith GU in the kth multicast group as:

$$p_{u,i,t} = P_{NLoS} = 1 - \frac{1}{1 + Cexp\Big(- D\big(\arctan(\frac{h_{u,t}}{d_{u,i,t}}) - C\big)\Big)} \tag{5}$$

where $i \in N_k, k \in \mathcal{K}$; $d_{u,i,t} = \sqrt{(h_{u,t})^2 + (x_i - x_{u,t})^2 + (y_i - y_{u,t})^2}$ is the distance between the ith GU and the uth UAV's at time slot t; $\arctan(\frac{h_{u,t}}{d_{u,i,t}})$ denotes the elevation angle of the UAV to the ith GU; and D and C are constant values depending on the environment.

On one hand, when the UAV-GU link is unblocked, and assuming free space fading channel gain, the channel gain between the uth UAV and the ith GU at time slot t can be expressed as

$$g_{u,i,t}^{LoS} = g_0 d_{u,i,t}^{-\delta} \tag{6}$$

where $\delta = 2$ and $g_0 = (\frac{c}{4\pi f_c})^2$. On the other hand, when the UAV-GU link is blocked, IRS is applied, as shown in processes ② and ③ of Figure 5.

(1) UAV-IRS link

At time slot t, let $\omega_{1,u,r_k,t} = \frac{|x_{r_k} - x_{u,t}|}{d_{u,r_k,t}}$, $\omega_{2,u,r_k,t} = \frac{|y_{r_k} - y_{u,t}|}{d_{u,r_k,t}}$, and $\omega_{3,u,r_k,t} = \frac{|h_{r_k} - h_{u,t}|}{d_{u,r_k,t}}$ denote the cosine and sine of the horizontal angle of arrival (AoA) of the signal at the IRS r_k from the uth UAV and the sine of the vertical AoA of the signal at the IRS r_k, respectively [12]. $d_{u,r_k,t} = \sqrt{(h_{u,t} - h_{r_k})^2 + (x_{u,t} - x_{r_k})^2 + (y_{u,t} - y_{r_k})^2}$ denotes the Euclidean distance between the uth UAV and IRS r_k near the kth multicast group at time slot t. The channel gain between the uth UAV and IRS r_k at time slot t can be expressed as

$$g_{u,r_k,t} = \sqrt{g_0 d_{u,r_k,t}^{-\delta}} \cdot \Omega_{u,r_k,t} \tag{7}$$

where d_r and d_c denote the length and the width of each UPA, respectively; $\Omega_{u,r_k,t} = [1, e^{-j\frac{2\pi}{\lambda} d_r \omega_{1,u,r_k,t} \omega_{3,u,r_k,t}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M_r-1) d_r \omega_{1,u,r_k,t} \omega_{3,u,r_k,t}}]^T \otimes [1, e^{-j\frac{2\pi}{\lambda} d_c \omega_{2,u,r_k,t} \omega_{3,u,r_k,t}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M_c-1) d_c \omega_{2,u,r_k,t} \omega_{3,u,r_k,t}}]^T$ represents the reflecting array response vector of the IRS [23].

(2) IRS-GU link

Similar to the UAV-IRS link, we define the cosine and sine of the horizontal angle of arrival (AoA) of the signal at the ith GU from the IRS r_k as $\omega_{1,r_k,i} = \frac{|x_{r_k} - x_i|}{d_{r_k,i}}$ and $\omega_{2,r_k,i} = \frac{|y_{r_k} - y_i|}{d_{r_k,i}}$. $\omega_{3,r_k,i} = \frac{h_{r_k}}{d_{r_k,i}}$ is the sine of the vertical AoA. $d_{r_k,i} = \sqrt{(h_{r_k})^2 + (x_i - x_{r_k})^2 + (y_i - y_{r_k})^2}$ denotes the Euclidean distance between the IRS r_k and the ith GU in the k multicast group. The channel gain from the IRS r_k multicasting to the ith GU is

$$g_{r_k,i} = \sqrt{g_0 d_{r_k,i}^{-\delta}} \cdot \Omega_{r_k,i} \tag{8}$$

where $\Omega_{r_k,i} = [1, e^{-j\frac{2\pi}{\lambda} d_r \omega_{1,r_k,i} \omega_{3,r_k,i}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M_r-1) d_r \omega_{1,r_k,i} \omega_{3,r_k,i}}]^T \otimes [1, e^{-j\frac{2\pi}{\lambda} d_c \omega_{2,r_k,i} \omega_{3,r_k,i}}, \ldots, e^{-j\frac{2\pi}{\lambda}(M_c-1) d_c \omega_{2,r_k,i} \omega_{3,r_k,i}}]^T$.

(3) IRS-Assisted UAV-GU link

The channel gain of the UAV-GU link assisted by the IRS r_k is given by

$$g_{u,i,t}^{IRS} = a(g_{r_k,i})^T \cdot M_{r_k,t} \cdot g_{u,r_k,t} \tag{9}$$

where $M_{r_k,t} = diag(e^{j\theta^t_{r_k,1,1}}, \ldots, e^{j\theta^t_{r_k,m_r,m_c}}, \ldots, e^{j\theta^t_{r_k,M_r,M_c}})$ denotes the IRS reflection phase coefficient matrix.

Combining (5), (6) and (9), the achievable average channel gain and SNR at the ith GU are expressed as

$$g_{u,i,t} = (1 - p_{u,i,t})g^{LoS}_{u,i,t} + p_{u,i,t}g^{IRS}_{u,i,t} \tag{10}$$

$$SNR^U_{u,i,t} = \gamma^U_{u,i,t} = \frac{G_U P_{u,t} g_{u,i,t}}{\sigma^2} \tag{11}$$

where $P_{u,t}$ is the uth UAV's transmit power at time slot t and σ^2 denotes the thermal noise power, which is linearly proportional to the allocated bandwidth [24].

We utilize OFDMA technology for multicasting. The transmission rate for the kth multicast group at time slot t is decided by the GU with the worst channel quality within the group:

$$R_{u,k,t} = \min_{i \in N_k} \frac{B}{|N_k|} log_2(1 + \gamma^U_{u,i,t}) \tag{12}$$

where $|N_k|$ denotes the number of GUs in the kth multicast group, and B is the total transmission bandwidth. According to (2), (3) and (12), the channel capacity of the ith GU in the kth multicast group can be calculated by

$$R_{i,t} = \underbrace{(1 - P_{g,i})R_{g,i}}_{\text{GBS-GU}} + \underbrace{P_{g,i}R_{u,k,t}}_{\text{UAV-IRS-GU}}, i \in N_k \tag{13}$$

Therefore, the average throughput of the system is

$$throughput = \sum_{i \in N_k, k \in \mathcal{K}} R_{i,t} \tag{14}$$

4. Problem Formulation

4.1. Multicast Grouping

Our goal is to optimize the 3D trajectory of the UAVs based on the transmission rate requirements of different GUs while minimizing energy consumption. First, we divide the N GUs into K multicast groups based on the characteristics of the GUs. We assume that the GUs remain static during the grouping procedure. The characteristics of the ith GU is defined by $\boldsymbol{\phi}_i = \{\mathbf{L}^{GU}_i, \gamma_i\}, i \in N$, which denotes the location and transmission rate requirement, respectively. Let $x_{i,k} \in \{0, 1\}$ indicate the correspondence between the GUs and multicast groups. $x_{i,k} = 1$ indicates that the ith GU belongs to the kth multicast group. The multicast grouping problem can be formulated as

$$\mathbf{P1}: \min_{\mathbf{X}, \boldsymbol{\psi}} \sum_{k=1}^{K} \sum_{i=1}^{N} x_{i,k} \|\boldsymbol{\phi}_i - \boldsymbol{\psi}_k\|^2 \tag{15}$$

$$s.t. \ x_{i,k} \in \{0, 1\}, \forall i \in N \tag{16}$$

$$\textstyle\sum_{k=1}^{K} x_{i,k} = 1, \forall i \in N \tag{17}$$

$$\textstyle\sum_{i=1}^{N} x_{i,k} \geqslant S, \forall k \tag{18}$$

where $\boldsymbol{\psi}_k$ is the characteristics of the GU selected as the kth multicast group's center. The constraints in (17) guarantee that a GU can only be in one multicast group. The constraints in (18) ensure that the number of GU within a group cannot exceed S.

4.2. Trajectory Optimization and Resource Allocation

4.2.1. UAV Energy Consumption

UAVs are mostly battery-powered with limited energy storage, so we aim to minimize the energy consumption of UAVs. At time slot t, the energy consumption of a UAV consists

of the energy used to transmit signals to GUs and the energy for UAVs flying in the air. The hovering energy consumption can be neglected compared to both of them [9].

First, the delay of transmission from the uth UAV to the kth multicast group at time slot t can be denoted as $T_{u,k,t} = \frac{S_{u,k,t}}{R_{u,k,t}}$, $T_{u,k,t} \leqslant \Delta t$, where $S_{u,k,t}$ is the size of the transmission data; the transmission should be finished within the time slot. Therefore, the transmission energy consumption at time slot t can be calculated as

$$E_{u,m,t} = P_{u,t} T_{u,k,t} = P_{u,t} \frac{S_{u,k,t}}{R_{u,k,t}} \tag{19}$$

Then, according to [12], the flying propulsion consumption at time slot t is calculated as

$$E_{u,f,t} = \left(P_0 (1 + \frac{3\|\mathbf{v}_{u,t}^2\| V_{max}^2}{U_{tip}^2}) + P_1 v_{h,u,t} \right) \Delta t \tag{20}$$

where P_0 is the blade power; U_{tip} is the tip speed of the rotor; P_1 is the descending/ascending power; V_{max} is the achievable maximal speed of UAVs; $\mathbf{v}_{u,t} = (v_{x,u,t}, v_{y,u,t}, v_{h,u,t})$ denotes the normalized speed of the uth UAV at time slot t, where $\|\mathbf{v}_{u,t}\| \leqslant 1$ and $-1 \leqslant v_{x,u,t}, v_{y,u,t}, v_{h,u,t} \leqslant 1$. Therefore, the location of the uth UAV at time slot $t+1$ is

$$\mathbf{L}_{u,t+1}^{UAV} = \mathbf{L}_{u,t}^{UAV} + \mathbf{v}_{u,t} \Delta t \tag{21}$$

Furthermore, as for obstacle avoidance, we consider the collision between the UAVs. UAVs have to keep a minimum distance from each other at any time. In other words,

$$\|\mathbf{L}_{u,t}^{UAV} - \mathbf{L}_{u',t}^{UAV}\|^2 \geqslant d_{min}^2; \forall u \neq u' \in U, t \tag{22}$$

where d_{min} is the minimum distance UAVs should keep between each other.

In order to minimize the energy consumption in (19), the transmission rate $R_{u,k,t}$ needs to be maximized. The channel gain g_u, i, t also needs to be maximized according to (11) and (12). The probability of the NLoS link P_{NLoS} in (5) is a monotonically decreasing function as the UAV's altitude increases. When the altitude $h_{u,t}$ is elevated, its elevation angle to the GU increases, and the probability that the LoS link is dominated increases. However, as the UAV is elevated and the distance between the UAV and the GU increases, the channel gain $g_{u,i,t}^{LoS}$ decreases. Therefore, in order to achieve the goal of minimizing energy consumption, altitude optimization needs to be tackled.

4.2.2. Joint Trajectory Optimization and Power Allocation Problem

Based on the above analysis, to minimize the total energy consumption of multiple UAVs, the joint optimization problem about the 3D trajectory and power allocation can be formulated as

$$\mathbf{P2}: \min_{V,P} \sum_{t=1}^{T} \sum_{u=1}^{U} E_{u,m,t} + E_{u,f,t} \tag{23}$$

$$s.t \quad \|\mathbf{v_{u,t}}\| \leqslant \mathbf{1}; \tag{24}$$

$$P_{u,t} \leqslant P_{max}; \tag{25}$$

$$T_{u,k,t} \leqslant \Delta t; \tag{26}$$

$$\|\mathbf{L}_{u,t}^{UAV} - \mathbf{L}_{u',t}^{UAV}\|^2 \geqslant d_{min}^2; \forall u \neq u' \in U, t; \tag{27}$$

where $V = \{\mathbf{v}_{u,t} | u \in U\}$ indicates the speed of the UAVs and $P = \{P_{u,t} | u \in U\}$ indicates the transmit power of the UAVs. The constraints in (24) ensure the normalized speed, and (25) guarantees that the UAV's transmit power cannot exceed the maximal power. The constraints in (26) ensure that the data transmission at time slot t must be completed within it. The constraints in (27) guarantee that individual UAVs do not collide during their respective flights.

5. Proposed Solution

5.1. K-Medoids for Multicast Grouping

We utilize K-medoids, an improved algorithm of the K-means clustering algorithm, to solve the multicast grouing problem **P1**. In the K-medoids algorithm, when given the location of the cell edge GUs and the number of multicast groups K, we can find the corresponding multicast group centers and edge GUs contained in each multicast group [4,5].

First, the K medoids are randomly initialized, and the iteration is set as $t = 0$. Subsequently, the medoids are continuously updated in iterations. When the medoids and their characteristics $\boldsymbol{\psi}_k$ are given, **P1** can be simplified as

$$\min_{\mathbf{X}} \sum_{k=1}^{K} \sum_{i=1}^{N} x_{i,k} \|\boldsymbol{\phi}_i - \boldsymbol{\psi}_k\|^2, s.t.(16)(17)(18) \tag{28}$$

Problem (28) can be solved by the branch and bound method [5]. In iteration $t+1$, the medoid of each multicast group is updated to be the member point with the smallest criterion function, which is the distance of the user characteristics between one member point and the other member points. The medoid in the kth multicast group is updated as

$$\boldsymbol{\psi}_k^{t+1} = \arg\min_{\boldsymbol{\phi}_i^t} \|\boldsymbol{\phi}_i^t - \boldsymbol{\phi}_j^t\|^2, \forall i, j \in m_k; i \neq j \tag{29}$$

where m_k denotes the set of members in the kth group. Repeat the above process until all the medoids do not change. The algorithm outputs the optimal medoids' location $\mathbf{L}_k^{MG} = (x_k, y_k), \forall k \in \mathcal{K}$ and the GUs' location in the kth multicast group $\mathbf{L}_{m_k}^{GU} = (x_i, y_i)$, $i \in N_k, k \in K$.

The specific algorithm is shown in Algorithm 1.

Algorithm 1 Constrained K-medoids

Input: GUs' location $\mathbf{L^{GU}}$, number of groups K
Output: optimal grouping strategy $\mathbf{X}^*$
1: **Initialization:** set iteration t = 0 and randomly initialize $\boldsymbol{\psi}_k^0$.
2: **for** the iteration $t = 1, 2, 3, \ldots$
3: Given $\boldsymbol{\psi}_k^t$, solve (28) to find current optimal grouping strategy $\mathbf{X^t}$.
4: Update multicast group medoids $\boldsymbol{\psi}_k^{t+1}$ using (29).
5: **until** medoids no longer change.

5.2. MADDPG for Optimization Problem

After solving the grouping problem, the joint trajectory optimization and power allocation problem is solved based on the best grouping strategy. In this section, we propose solving **P2** with the MADDPG algorithm.

Each UAV acts as an agent that interacts with the environment over T time slots. At time slot t, an action a_t is generated based on the state s_t of the environment around the agent, and a reward r_t is obtained for judging the action a_t generated in the current state. The goal of each agent is to train a policy π that can generate the action a_t that makes the reward r_t the highest based on the current state. Subsequently, at time slot $t+1$, the state s_t transits into a new state s_{t+1} due to the action a_t, and the process is repeated till the end of an episode.

5.2.1. State, Action, and Reward

In order to solve **P2** with MADDPG, we define the state, action, and reward of each UAV, which acts as an agent at time slot t.

(1) State

For each agent, the state includes information perceived from the environment based on which the agent determines its action and evaluates long-term rewards. Since the transmission rate is related to the distance between the UAV and GU, each agent's perceived information includes the observed multicast groups and the locations of its member users along with the UAVs' locations. Then, the state of the uth UAV can be formulated as

$$\mathbf{s}_{u,t} = [\mathbf{L}_{1,t}^{UAV}, \mathbf{L}_{2,t}^{UAV}, \ldots, \mathbf{L}_{M_u,t}^{UAV}, \mathbf{L}_{m_1,t}^{GU}, \mathbf{L}_{m_2,t}^{GU}, \ldots, \mathbf{L}_{m_{M_k},t}^{GU}] \tag{30}$$

which includes the locations of the M_u closest neighbor UAVs that can be observed at the current time slot and the locations of GUs in the M_k closest multicast group.

(2) Action

In **P2**, we need to determine the speed and the power allocation of the UAV at time slot t. The action of the uth UAV can be formulated as

$$\mathbf{a}_{u,t} = [\mathbf{v}_{u,t}, P_{u,t}] \tag{31}$$

where (24) and (25) are satisfied.

(3) Reward

The reward function determines the objective of the RL problem. The objective function of **P2** is to minimize the energy consumption of the UAV, while the objective of RL is to maximize the reward. The reward of each agent can be set as negative energy consumption:

$$r_{u,t} = -(E_{u,m,t} + E_{u,f,t}) \tag{32}$$

5.2.2. MADDPG Algorithm

The 3D trajectory and power allocation to be optimized in Problem **P2** are both continuous variables. DDPG, based on an actor–critic architecture, is able to learn a deterministic policy for continuous actions, which can directly output the optimal action. However, in the multi-agent case, if each agent only considers its own observations and actions to learn its own policy using DDPG, the environment will become non-stationary from the perspective of any individual agent. This is due to the fact that the policies of other agents are also constantly updated and changing, and the sampled data of the agent does not follow a consistent probability distribution.

Therefore, Ref. [6] proposed a framework of centralized training and decentralized execution, which allows the critic to obtain the policy information of other agents during the training process. Only local information is needed when applying the actor to make decisions. The framework is shown in Figure 6. On one hand, the centralized critics Q_u that MADDPG trains for each agent use the observed states and actions of all agents as input. As a result, critics are able to capture changes in all agents' policies, thus eliminating non-stationarity [9]. The Q-value computed by the critic Q_u is used to update the corresponding actor's policy network π_u. On the other hand, when each agent is sufficiently trained, the actor can compute the policy independently based on the state, without the feedback of the critic and the state or action information of other agents. Hence, MADDPG facilitates fully decentralized execution. In general, MADDPG can be regarded as a centralized RL technique when training.

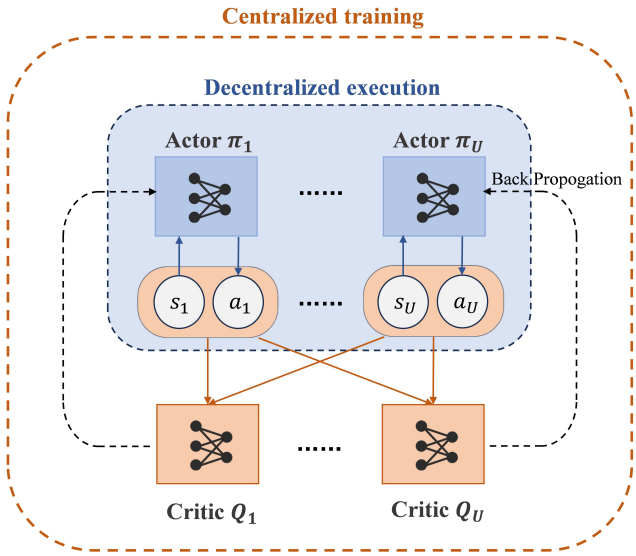

Figure 6. The centralized training and decentralized execution framework of MADRL.

As shown in Figure 7, the training process of MADDPG includes data sampling, model training, and parameter updating. We define the state set, the action set, and the reward set from all of the agents as $\mathbf{s}_t = \{\mathbf{s}_{1,t}, \mathbf{s}_{2,t}, \ldots, \mathbf{s}_{U,t}\}$, $\mathbf{a}_t = \{\mathbf{a}_{1,t}, \mathbf{a}_{2,t}, \ldots, \mathbf{a}_{U,t}\}$, and $\mathbf{r}_t = \{r_{1,t}, r_{2,t}, \ldots, r_{U,t}\}$, respectively. According to [9], setting a specific reward in (32) for each agent will lead to increasing training complexity, so we set the reward of each agent to be the cumulative reward of all agents, $\sum_{u=1}^{U} r_{u,t}$. The reward in (32) is changed to be

$$r_{u,t} = -\sum_{u=1}^{U} (E_{u,m,t} + E_{u,f,t}) \tag{33}$$

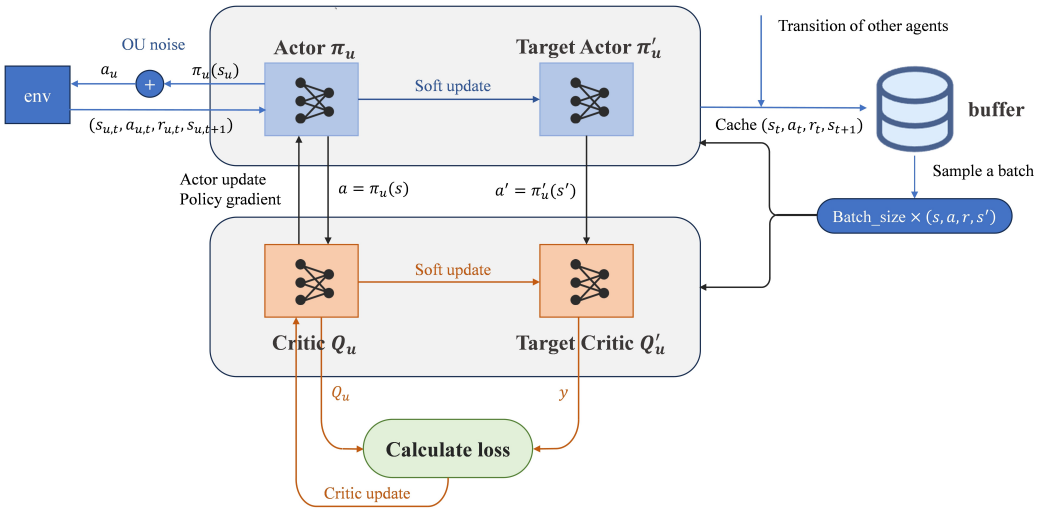

Figure 7. The training procedure of each agent (agent u) in MADDPG.

Each agent contains four networks during training. The actor network and target actor network share the same network architecture. The input is the current state $\mathbf{s_{u,t}}$ of agent u, and the output is the action $\mathbf{a}_{u,t}$. The critic network has the same network architecture

as the target critic network. The inputs are the states and actions of all the agents, and the output is the Q-value, which is defined as

$$Q_u^{\pi}(\mathbf{s}, \mathbf{a}) = \mathbb{E} \sum_{t=0}^{\infty} \gamma^t r_{u,t} \tag{34}$$

where γ denotes the discount factor.

Let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_U\}$ denote the policies of U agents, which are fitted, respectively, by U actor networks parameterized by $\boldsymbol{\theta}^{\pi} = \{\theta_1^{\pi}, \theta_2^{\pi}, \ldots, \theta_U^{\pi}\}$. As shown in Figure 7, the parameters of the actor are updated by policy gradient, which is calculated as

$$\nabla_{\theta_u} J(\theta_u) = \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\mathcal{D}}[\nabla_{\mathbf{a}_u} Q_u^{\pi}(\mathbf{s}, \mathbf{a})|_{\mathbf{a}_u=\boldsymbol{\pi}_u(\mathbf{s}_u)} \nabla_{\theta_u} \boldsymbol{\pi}_u(\mathbf{s}_u)] \tag{35}$$

where $\mathcal{D}$ represents the experience replay buffer. Each state transition $(\mathbf{s}_{u,t}, \mathbf{a}_{u,t}, r_{u,t}, \mathbf{s}_{u,t+1})$ of agent u with other agents is stored in the buffer.

U critic networks are parameterized by $\boldsymbol{\theta}^{Q} = \{\theta_1^{Q}, \theta_2^{Q}, \ldots, \theta_U^{Q}\}$. The updating process of the critic is shown in Figure 7. The gradient descent method is used to minimize the following loss function

$$\mathcal{L}(\theta_u^{Q}) = \mathbb{E}_{(\mathbf{s},\mathbf{a},\mathbf{r},\mathbf{s}')\sim\mathcal{D}}[Q_u^{\pi}(\mathbf{s}, \mathbf{a}) - y]^2 \tag{36}$$

where $y = r_u + \gamma Q_u'^{\pi'}(\mathbf{s}', \mathbf{a}')|_{\mathbf{a}_u'=\boldsymbol{\pi}_u'(\mathbf{s}_u)}$, $Q_u'^{\pi'}$ and $\boldsymbol{\pi}_u'$ denote the target critic and the target actor, respectively. After the actor and critic are updated, we use a soft update to update the two target networks as

$$\begin{aligned} \boldsymbol{\theta}^{\pi'} &\leftarrow \tau\boldsymbol{\theta}^{\pi} + (1-\tau)\boldsymbol{\theta}^{\pi'} \\ \boldsymbol{\theta}^{Q'} &\leftarrow \tau\boldsymbol{\theta}^{Q} + (1-\tau)\boldsymbol{\theta}^{Q'} \end{aligned} \tag{37}$$

where τ is typically set as $\tau = 0.001$.

The specific steps are shown in Algorithm 2.

Algorithm 2 MADDPG for Optimization Problem

Initialization: Randomly set $\boldsymbol{\theta}^{\pi}$, $\boldsymbol{\theta}^{\pi'}$, $\boldsymbol{\theta}^{Q}$, $\boldsymbol{\theta}^{Q'}$; $\mathcal{D} = \varnothing$.
for episode = 1 to max-episode-num **do**
  Reset the environment, and receive initial state $\boldsymbol{s} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_U\}$.
  for t = 1 to max-episode-length T **do**
    for each agent u, select action $\boldsymbol{a}_u = \boldsymbol{\pi}_u(\boldsymbol{s}_u) +$ noise.
    Execute actions $\boldsymbol{a} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots \boldsymbol{a}_U\}$ and observe reward $\boldsymbol{r} = \{r_1, r_2, \ldots r_U\}$ and next state $\boldsymbol{s'}$.
    Push $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s'})$ into replay buffer $\mathcal{D}$.
    $\boldsymbol{s} \leftarrow \boldsymbol{s'}$.
    if length of $\mathcal{D}$ larger than given length **then**
      for UAV agent u=1 to U **do**
        Sample a random batch of S samples $\{(\boldsymbol{s}^j, \boldsymbol{a}^j, \boldsymbol{r}^j, \boldsymbol{s'}^j)\}_{j=1,\ldots,S}$ from $\mathcal{D}$.
        Set $y^j = r_u^j + \gamma Q'^{\pi'}_u(\boldsymbol{s'}^j, \boldsymbol{a'}^j)|_{\boldsymbol{a}_u'^j=\boldsymbol{\pi}_u'(\boldsymbol{s}_u^j)}$.
        Update critic θ_u^{Q} by minimizing the loss $\mathcal{L}(\theta_u^{Q}) = \frac{1}{S}\sum_j (Q_u^{\pi}(\boldsymbol{s}^j, \boldsymbol{a}^j) - y^j)^2$.
        Update actor θ_u^{π} using the sampled policy gradient $\nabla_{\theta_u^{\pi}} J(\theta_u^{\pi}) \approx \frac{1}{S}\sum_j [\nabla_{\boldsymbol{a}_u^j} Q^{\pi}(\boldsymbol{s}^j, \boldsymbol{a}^j)|_{\boldsymbol{a}_u^j=\boldsymbol{\pi}_u(\boldsymbol{s}_u^j)} \nabla_{\theta_u^j} \boldsymbol{\pi}_u(\boldsymbol{s}_u^j)]$.
      end for
      Update target network parameters by $\boldsymbol{\theta}^{\pi'} \leftarrow \tau\boldsymbol{\theta}^{\pi} + (1-\tau)\boldsymbol{\theta}^{\pi'}$ and $\boldsymbol{\theta}^{Q'} \leftarrow \tau\boldsymbol{\theta}^{Q} + (1-\tau)\boldsymbol{\theta}^{Q'}$.
    end if
  end for
end for

6. Simulation Results

In this section, we simulate the performance of MADDPG in the multi-UAV multicast system, comparing it with DDPG, Deep Q-Network (DQN), DDQN, and a classical RL algorithm Upper Confidence Bound (UCB) designed for Budgeted Multi-Armed Bandits (BMABs) problems. This paper considers a three-dimensional space with four cellular cells, each equipped with a GBS at the center. A total of 100 GUs are randomly distributed in the targeted area. The GUs are determined whether or not they are served by a UAV acting as an aerial base station according to (3). The GUs are divided into five multicast groups according to their locations and transmission rate requirements. There are five UAVs to serve them, and the initial positions of the UAVs are set as $(400, 400, 200)$, $(400, -400, 200)$, $(-400, 400, 200)$, $(-400, -400, 200)$, and $(0, 0, 0)$, respectively. For the channel model, we set $C = 5$ and $D = 0.35$ in (5) [25]. The total transmission bandwidth and noise power are $B = 2$ MHz and $\sigma^2 = -100$ dBm, respectively. The maximal power, the UAV's maximal speed, and the rotor tip speed are set as $P_{max} = 500$ mW, $V_{max} = 30$ m/s, and $U_{tip} = 200$ m/s, respectively. We utilize PyTorch and Gym to model the environment and simulate the algorithms. In MADDPG, the actor and critic networks are both set as fully connected neural networks with [128, 64] neurons where the activation function of the two hidden layers is $ReLU$. We utilize $tanh$ as the activation function of the actor network's output layer, which can restrict the output action within $(-1, 1)$. For the optimizer, the Adam optimizer is applied to train the DNNs in the policy network and Q-Network. During the training, the length of an episode is set as $T = 25$. The capacity of the buffer and the batch size are 10^5 and 1024, respectively. The specific simulation settings are shown in Table 1.

Table 1. Parameters settings on simulation.

Parameter	Value
Number of ground users N	100
Number of UAVs K	5
Number of multicast groups k	5
Blocking parameters C, D	5, 0.35
Bandwidth B	2 MHz
UAV maximal power P_{max}	500 mW
Noise power σ^2	-100 dBm
Blade power P_0	$\frac{12\times30^3\times0.4^3}{8}\rho s G$
Descending/Ascending power $P1$	11.46
V_{max}, U_{tip}	30, 200
s, ρ, G	0.05, 1.225, 0.503 [26]
S, d_{min}	30, 20 m
Number of episodes	15,000
Length of an episode T	25
Batch size, Learning rate	1024, 0.001

In Figure 8, we plot the performance of Algorithm 1 to solve the multicast grouping problem. There are 160 edge GUs that cannot obtain reliable communication from the GBS in the figure and are classified into K = 8 multicast groups. As shown in Figure 8, there are seven cellular cells in the two-dimensional space of $(-500, 500) \times (-500, 500)$, and the red triangle in the center of each cell denotes the GBS. The scatters denote randomly distributed GUs in the space of $(-350, 350) \times (-350, 350)$, and scatters with the same color indicate that the users are classified into the same multicast group.

In Figure 9, we depict the average reward of different algorithms during 15,000 episodes. The average reward in each episode is set as the negative value of the average energy consumed per UAV within a time slot. We compare the training procedure of MADDPG with that of three DRL methods. For the discrete action space, we discretized the continuous action space, where the transmit power set is $P_{u,t} = P_{max} \cdot \{0.1, 0.2, \ldots, 1\}$ and the UAV's speed set is $v_{u,t} = \{[-1, -1, -1], [-1, -1, 1], [-1, 1, -1], [-1, 1, 1], [1, -1, -1], [1, -1, 1], [1, 1, -1], [1, 1, 1]\}$. As shown in Figure 9, the average reward of all algorithms

increases with the number of episodes. The discrete action space of DQN is only a subset of that of DDPG. As a result, DQN converges more quickly compared to DDPG, but the final performance of DDPG is slightly higher than that of DQN. DDQN performs similarly to DQN. As for MADDPG, since the framework of centralized training and distributed execution can solve the issue of non-stationarity, MADDPG can achieve a higher average reward and save the UAVs' energy more effectively.

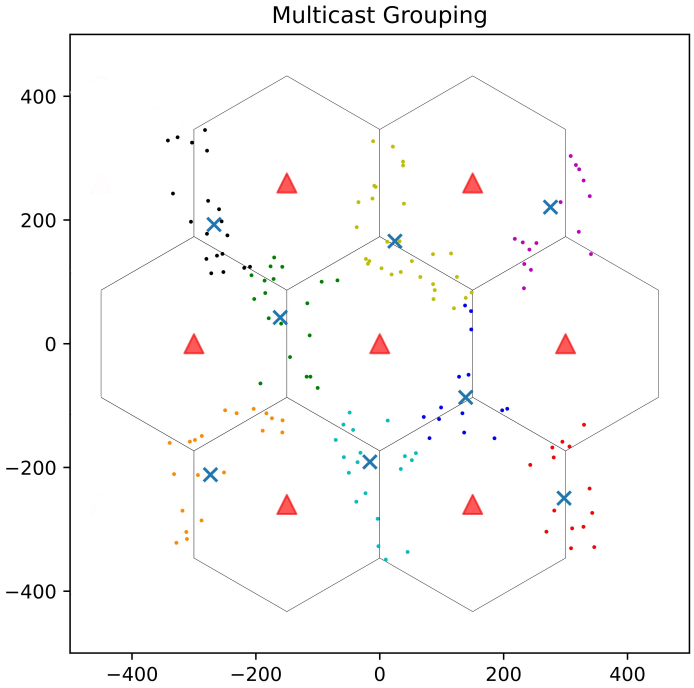

Figure 8. Multicast grouping results for $N = 160$ and $K = 8$. The triangles represent the ground base stations. The dots denote ground edge users. The crosses denote different multicast group centers.

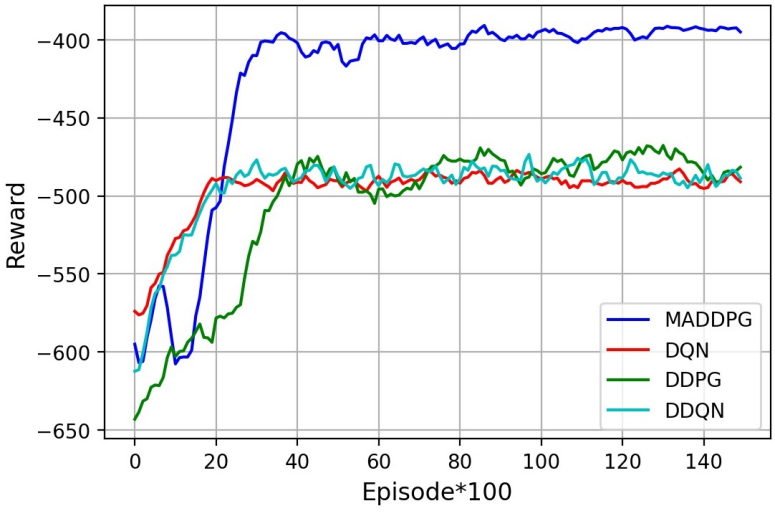

Figure 9. Training procedure of different algorithms for minimizing UAV energy consumption.

The optimal 3D trajectory of the UAV obtained by solving **P2** with MADDPG is plotted in Figure 10. The order in which each agent serves the multicast groups is basically the same, and different UAVs serve different multicast groups in the same time slot. For the sake of observation, only the trajectory of one UAV is plotted, and it can be seen that the UAVs decide the altitude that minimizes the energy consumption based on the served multicast group's transmission rate requirement.

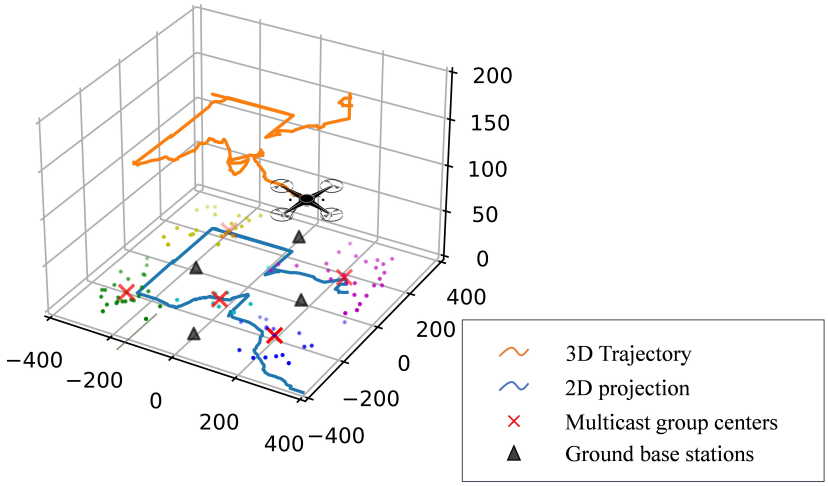

Figure 10. The 3D flying trajectory of UAV using MADDPG.

In Figure 11, we compare the average throughput of different algorithms when grouping GUs and not grouping. In contrast to communicating with all users by broadcasting, the average throughput is improved by multicast grouping considering the characteristics of the GUs' respective transmission rate requirements. In that case, the average throughput is not restricted to the GUs with the worst channel but only to those within the multicast group. Meanwhile, MADDPG also performs well in improving the GUs' throughput compared to DQN and DDPG.

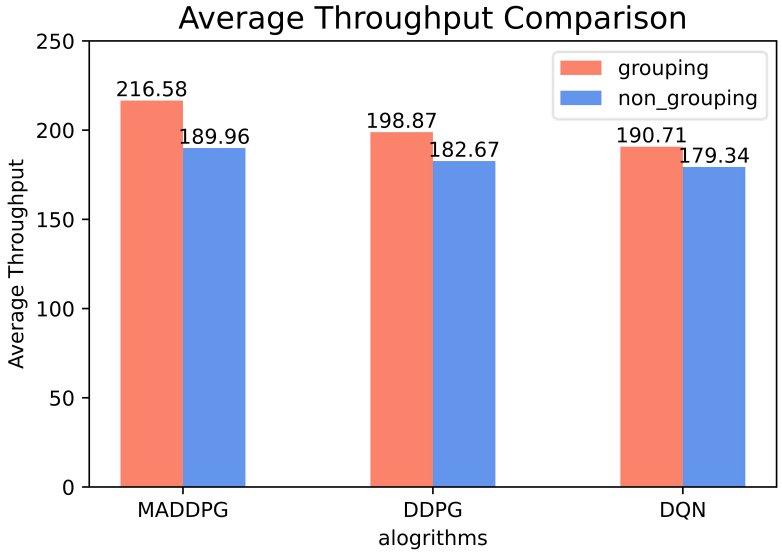

Figure 11. Comparison of average throughput per user of grouping and non-grouping.

In Figure 12, we plot the energy consumption as the number of multicast groups K varies under different algorithms. As K increases, UAVs need to travel longer distances to serve more multicast groups and consume more energy for transmission. However, if a smaller K is chosen, although the energy consumption decreases, the average throughput will also reduce. As a result, we chose an intermediate value $K = 5$ to make a trade-off. As for different algorithms, MADDPG performs better than DQN, DDQN, DDPG, and the classical RL method UCB in reducing the energy consumption of UAVs. The results are consistent with those in Figure 9.

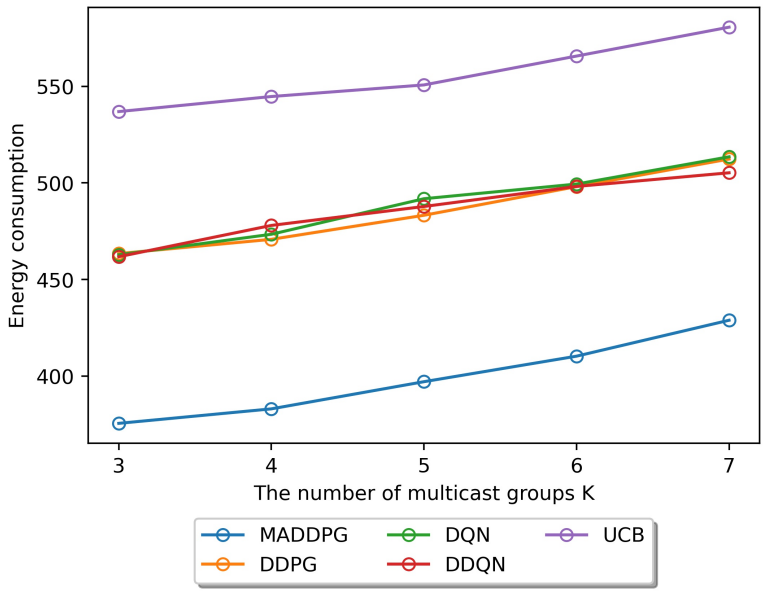

Figure 12. Energy consumption with varying number of multicast groups K and different algorithms.

7. Conclusions

This paper investigates the optimization of grouping, joint 3D trajectories, and power allocation in multi-UAV multicast systems, aiming to minimize the energy consumption of UAVs. To achieve this goal, this paper first solves the multicast grouping problem for users with different transmission rate requirements using the constrained K-medoids algorithm. Then, due to the problem of non-stationarity caused by multiple agents training in the traditional DRL algorithm, this paper adopts MADDPG to eliminate the non-stationarity and set the negative value of UAVs' energy consumption as the reward. Simulation results show that MADRL can effectively reduce the energy consumption of UAVs, and, at the same time, the combination of a multicast communication approach and using UAVs as aerial base stations can effectively improve the average throughput. Based on this work, we will investigate the performance of the proposed multi-UAV multicast system in other scenarios, such as vehicular environments.

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

Funding: The work is supported by the National Key R&D Program of China under Grant 2019YFE0114000; the Shenzhen Science and Technology Innovation Commission Free Exploring Basic Research Project Grant 2021Szvup012.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to thank the Editor-in-Chief, Editor, and anonymous reviewers for their valuable reviews.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of this study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

DDQN	Double Deep Q-Network
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Network
MADDPG	Multi-Agent Deep Deterministic Policy Gradient

References

1. Abubakar, A.I.; Ahmad, I.; Omeke, K.G.; Ozturk, M.; Ozturk, C.; Abdel-Salam, A.M.; Mollel, M.S.; Abbasi, Q.H.; Hussain, S.; Imran, M.A. A Survey on Energy Optimization Techniques in UAV-Based Cellular Networks: From Conventional to Machine Learning Approaches. *Drones* **2023**, *7*, 214. [CrossRef]
2. Wu, Y.; Xu, J.; Qiu, L.; Zhang, R. Capacity of UAV-Enabled Multicast Channel: Joint Trajectory Design and Power Allocation. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–7. [CrossRef]
3. Al-Hourani, A.; Kandeepan, S.; Lardner, S. Optimal LAP Altitude for Maximum Coverage. *IEEE Wirel. Commun. Lett.* **2014**, *3*, 569–572. [CrossRef]
4. Bradley, P.S.; Bennett, K.P.; Demiriz, A. Constrained k-means clustering. *Microsoft Res.* **2000**, *20*, 8. Available online: https://www.microsoft.com/en-us/research/publication/constrained-k-means-clustering/ (accessed on 1 August 2023).
5. Khamidehi, B.; Sousa, E.S. Trajectory Design for the Aerial Base Stations to Improve Cellular Network Performance. *IEEE Trans. Veh. Technol.* **2021**, *70*, 945–956. [CrossRef]
6. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In Proceedings of the 2017 Advances in Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 6379–6390. [CrossRef]
7. Hu, J.; Zhang, H.; Song, L.; Schober, R.; Poor, H.V. Cooperative internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning. *IEEE Trans. Commun.* **2020**, *68*, 6807–6821. [CrossRef]
8. Deng, C.; Xu, W.; Lee, C.-H.; Gao, H.; Xu, W.; Feng, Z. Energy Efficient UAV-Enabled Multicast Systems: Joint Grouping and Trajectory Optimization. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–7. [CrossRef]
9. Chen, B.; Liu, D.; Hanzo, L. Decentralized Trajectory and Power Control Based on Multi-Agent Deep Reinforcement Learning in UAV Networks. In Proceedings of the 2022 IEEE International Conference on Communications (ICC), Seoul, Republic of Korea, 16–20 May 2022; pp. 3983–3988. [CrossRef]
10. Lee, J.; Friderikos, V. Multiple UAVs Trajectory Optimization in Multicell Networks with Adjustable Overlapping Coverage. *IEEE Internet Things J.* **2023**, *10*, 9122–9135. [CrossRef]
11. Lee, J.; Friderikos, V. Path optimization for Flying Base Stations in Multi-Cell Networks. In Proceedings of the 2020 IEEE Wireless Communications and Networking Conference (WCNC), Seoul, Republic of Korea, 25–28 May 2020; pp. 1–6. [CrossRef]
12. Mei, H.; Yang, K.; Liu, Q.; Wang, K. 3D-Trajectory and Phase-Shift Design for RIS-Assisted UAV Systems Using Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2022**, *71*, 3020–3029. [CrossRef]
13. Wei, Z.; Cai, Y.; Sun, Z.; Ng, D.W.K.; Yuan, J.; Zhou, M.; Sun, L. Sum-Rate Maximization for IRS-Assisted UAV OFDMA Communication Systems. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 2530–2550. [CrossRef]
14. Ji, Z.; Yang, W.; Guan, X.; Zhao, X.; Li, G.; Wu, Q. Trajectory and Transmit Power Optimization for IRS-Assisted UAV Communication Under Malicious Jamming. *IEEE Trans. Veh. Technol.* **2022**, *71*, 11262–11266. [CrossRef]
15. Ge, Y.; Fan, J.; Zhang, J. Active Reconfigurable Intelligent Surface Enhanced Secure and Energy-Efficient Communication of Jittering UAV. *IEEE Internet Things J.* **2023**, *20*, 4962–4975 . [CrossRef]
16. Son, H.; Jung, M. Phase Shift Design for RIS-Assisted Satellite-Aerial-Terrestrial Integrated Network. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, 1–9. [CrossRef]
17. Feng, W.; Tang, J.; Wu, Q.; Fu, Y.; Zhang, X.; So, D.K.C.; Wong, K.K. Resource Allocation for Power Minimization in RIS-assisted Multi-UAV Networks with NOMA. *IEEE Trans. Commun.* **2023**. [CrossRef]
18. Nguyen, T.L.; Kaddoum, G.; Do, T.N.; Haas, Z.J. Channel Characterization of UAV-RIS-Aided Systems with Adaptive Phase-Shift Configuration. *IEEE Wirel. Commun. Lett.* **2023**. [CrossRef]
19. Xue, Z.; Wang, J.; Ding, G.; Wu, Q. Joint 3D Location and Power Optimization for UAV-Enabled Relaying Systems. *IEEE Access* **2018**, *6*, 43113–43124. [CrossRef]
20. Hosny, R.; Hashima, S.; Mohamed, E.M.; Zaki, R.M.; ElHalawany, B.M. Budgeted Bandits for Power Allocation and Trajectory Planning in UAV-NOMA Aided Networks. *Drones* **2023**, *7*, 518. [CrossRef]
21. Fan, W.; Luo, K.; Yu, S.; Zhou, Z.; Chen, X. AoI-driven Fresh Situation Awareness by UAV Swarm: Collaborative DRL-based Energy-Efficient Trajectory Control and Data Processing. In Proceedings of the 2020 IEEE/CIC International Conference on Communications in China (ICCC), Chongqing, China, 9–11 August 2020; pp. 841–846. [CrossRef]
22. Lyu, J.; Zeng, Y.; Zhang, Y. UAV-Aided Offloading for Cellular Hotspot. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 3988–4001. [CrossRef]

23. Wang, F.; Zhang, X. Active-IRS-Enabled Energy-Efficiency Optimizations for UAV-Based 6G Mobile Wireless Networks. In Proceedings of the 2023 57th Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, USA, 22–24 March 2023; pp. 1–6. [CrossRef]
24. Samir, M.; Chraiti, M.; Assi, C.; Ghrayeb, A. Joint Optimization of UAV Trajectory and Radio Resource Allocation for Drive-Thru Vehicular Networks. In Proceedings of the 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 15–18 April 2019; pp. 1–6. [CrossRef]
25. Wang, Y.; Hu, Z.; Wen, X.; Lu, Z.; Miao, J.; Qi, H. Three-Dimensional Aerial Cell Partitioning Based on Optimal Transport Theory. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020; pp. 1–6. [CrossRef]
26. Mei, H.; Wang, K.; Zhou, D.; Yang, K. Joint trajectory-task-cache optimization in UAV-enabled mobile edge networks for cyber-physical system. *IEEE Access* **2019**, *7*, 156476–156488. [CrossRef]