Deep-Reinforcement-Learning-Based Optimal Transmission Policies for Opportunistic UAV-Aided Wireless Sensor Network

Yitong Liu, *Student Member, IEEE*, Junjie Yan, *Student Member, IEEE*, and Xiaohui Zhao

***Abstract*—When there are unmanned aerial vehicles (UAVs) performing their specifically assigned tasks in the air, some of them still have available resources to access different ground communication networks to improve their communication performance, especially for the wireless sensor network. Technically, when they execute their own given missions with predetermined trajectories, they can also provide opportunistic assistance for terrestrial networks at the same time. In this article, we solve an opportunistic UAV-assisted data transmission problem in a wireless sensor network from a novel perspective. In consideration of UAVs dynamic behaviors, varying transmission tasks, and real-time matching between UAVs and sensor clusters, we propose to jointly optimize UAV scheduling and power control aiming to obtain optimal policies to maximize the network data transmission in a long run under the opportunistic access mode. We reformulate this optimization problem as a Markov decision process (MDP) and take deep reinforcement learning (DRL) as our tool to obtain solutions. We develop a DQN-based and a deep deterministic policy gradient (DDPG)-based optimization approaches to adjust the power allocation of cluster heads, and the scheduling and bandwidth allocation of UAVs during their missions over the covered area to improve the whole network data transmission performance. Simulation results demonstrate the validity and superiority of our proposed approaches compared with other benchmark policies in different perspectives.**

***Index Terms*—Deep reinforcement learning (DRL), opportunistic unmanned aerial vehicle (UAV) transmission, resource allocation, scheduling, wireless sensor network.**

I. Introduction

THE INCREASING development of next-generation wireless networks sets an unparalleled criterion for high-quality services. With the advantages of maneuverability and high probability of Line-of-Sight (LoS) transmission, unmanned aerial vehicles (UAVs) are envisioned to play a key role in future wireless communication systems. As a useful extension, UAVs-assisted communication improves the performance of terrestrial networks and replenishes conventional infrastructures with more flexible connectivity. The practical applications, such as aerial base station [1], real-time transmission relay [2], flying edge computing cloudlets [3], etc., are all important and remarkable technologies for developing the potential ability of UAVs. As one of the promising technologies, it is foreseeable that UAVs-assisted communication will contribute a lot to our society constantly.

Due to the wide range of applications, relevant research on UAV-enabled networks have been conducted quite thoroughly. There are two main research directions for air-ground integrated wireless networks with UAVs, i.e., the UAV-assisted terrestrial communications and the cellular-connected UAV communications [4]. On the one hand, many studies assume that the considered networks exclusively employ some dedicated UAVs to assist terrestrial users. In this case, UAVs actually serve as aerial infrastructures to provide necessary services or assistance by consuming part of their own propulsion energy. To better meet the demands, we usually design optimal flying trajectories for UAVs. On the other hand, UAVs are regarded as aerial users, like the terrestrial ones in the air–ground-integrated networks. Since UAVs also have their own missions to accomplish, the competitions about time and spectrum access may be bound to burst among users both in aerial and ground. In this case, the aforementioned data transmission advantages of UAVs seem to have an adversary impact on the communication of terrestrial networks. Thus, the UAVs serving as aerial communication platforms or aerial mission realizers are the two completely independent application scenarios in the existing works. In fact, if we efficiently manage the utilization of UAVs in the sense of opportunistic access for data transmission, the UAVs with their missions can also help a lot for terrestrial users. The combination of the advantages of both research directions is necessary and meaningful to be studied, which is an independent problem compared with the aforementioned ones.

In this article, we aim to propose a UAVs-aided wireless sensor network from a novel perspective through an opportunistic access of UAVs under the consideration of efficient use of UAVs in certain communication scenarios. In fact, various types of UAV missions, for instance, goods delivering, area patrolling, video shooting, or target detection, exist simultaneously. We notice there is only a small proportion of UAVs providing specialized communication services and most of UAVs hardly need to communicate with others or they only require little feedback information. Their mobility is according to a preset trajectory or by trajectory scheduling with a small amount of control messages. In other words, the precious communication resources and abilities of these UAVs are almost under developed and not well used, which is an

Manuscript received 18 September 2021; revised 1 December 2021; accepted 7 January 2022. Date of publication 12 January 2022; date of current version 25 July 2022. This work was supported by the National Natural Science Foundation of China under Grant 61571209. *(Corresponding author: Xiaohui Zhao.)*

The authors are with the College of Communication Engineering, Jilin University, Changchun 130012, Jilin, China (e-mail: yitong19@mails.jlu.edu.cn; yanjj19@mails.jlu.edu.cn; xhzhao@jlu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2022.3142269

interesting research problem for us. To overcome the above drawbacks, when executing the specific missions, these UAVs with available communication resources can provide assistance to enhance the terrestrial network communication performance opportunistically. Therefore, the analysis on this mode of UAV utilization is of great importance.

Different from the application of UAVs as aerial base stations, the opportunistic UAVs are generally dynamic and we cannot directly interfere their original trajectories for our communication demand. When these UAVs execute their original tasks with their designed flight paths across air, they are widely deployed in various scenarios and easy to access everywhere. They can contribute their most idle communication resources to assist terrestrial wireless communication networks without increasing extra flight energy consumption while keeping normal execution of their specific missions. Through this application mode, these UAVs can take advantages of aerial positions and LoS channel conditions to assist terrestrial data transmission. Therefore, the utilization of the opportunistic UAVs will convert the fixed heavy communication pressure into opportunities for air–ground collaborative communication optimization [4], which alleviates resource waste.

Recently, there have been more extensive researches to address the data transmission challenges in UAV-assisted communication networks. In the existing studies, how to design an optimal trajectory for better data transmission is one of the most significant research interests. The most common optimization goal is to obtain a larger amount of data under some constraints. In [5], subject to the practical mobility and the information causality constraints at the relay, a throughput maximization problem of a mobile relaying system is solved by optimizing the source and relay transmit power along with the relay trajectory. In [6], a UAV collects data from multiple sensor nodes in a UAV-enabled wireless sensor network. The UAV communication scheduling and 3-D trajectory are jointly optimized to maximize the minimum average data collection rate. To analyze the delay-constrained communication scenarios, Wu and Zhang [7] jointly optimized the UAV trajectory and OFDMA resource allocation to maximize the minimum average throughput under the constraints on the minimum instantaneous rate required for each user. Considering the feasible deployment of UAVs in a space–air–ground three-tier heterogeneous network for supporting IoT applications, Wang *et al.* [8] formulated a two-stage joint hovering altitude and power control for UAV networks. In [9], UAV communication and NOMA are combined for constructing high capacity IoT uplink transmission systems. To maximize the system capacity, the subchannel assignment, the IoT nodes uplink transmit power, and the UAV flying heights are jointly optimized.

Moreover, many studies have been devoted to the efficient use of UAVs in data transmission tasks. To analyze a practical UAV-aided data collection scenario with energy harvesting, Fu *et al.* [10] considered the UAV trajectory, hovering height, and the wireless power charging to improve the energy efficiency. Nazib and Moh [11] proposed a special scheme working in mountainous areas to optimize the trajectory of the data collection route, which consumes less energy compared with the conventional approaches. In addition, minimizing transmission time is also an interesting research problem to be discussed. Wang *et al.* [12] solved the problem by optimizing the UAV-sensor association mechanism and adjusting data collection of UAVs to shorten the task execution time. Wang *et al.* [13] minimized the task required execution time via jointly optimizing the trajectory of UAV and the transmission scheduling for all ground terminals. Considering more influences, Li *et al.* [14] jointly optimized the trajectory, altitude, velocity of UAV, and the links with ground users to minimize the total mission time. Furthermore, to keep data freshness is a critical requirement for the data collection performance. Samir *et al.* [15] jointly optimized the UAV trajectories and scheduling policies to maintain a minimum Age of Information (AoI) under the minimum throughput constraints. Similarly, the trajectory planning is performed in [16] to find the routes to improve the freshness of information collected from all sensor nodes. Through the path planning, Pan *et al.* [17] solved the energy wasting problem and reduced the required number of UAVs. The work in [18] presents an essential tradeoff between the aerial cost containing the propulsion energy and the operation consumption of all UAVs, and the ground cost from the energy consumption of all sensor nodes. It aims to optimize the UAV trajectory jointly with wake-up time allocation and the transmit power of all sensor nodes to minimize the weighted sum of these two costs. Some key technologies of the multi-UAVs network are discussed in detail in [19], such as the stable flying ad-hoc network structure formulation, UAV network protocol architectures, distributed gateway-selection algorithms, and cloud-based stability-control mechanisms. In summary, there have been plenty of significant works on the deployment of multi-UAVs, since the utilization of multiple UAVs can provide better performance for the ground wireless communication networks. These aforementioned works can contribute instructive references for our follow-up study. From these studies, we can conclude that UAVs as opportunistic relays can also provide more communication service for terrestrial networks to realize better performance.

Interestingly and importantly, the utilization of opportunistic UAV is an independent problem for researchers to solve. From our point of view, we can exploit the potential of UAVs, which are already assigned to specific tasks to help the data transmission of the ground network. Based on their predetermined trajectories, we will discuss the opportunistic access between UAVs and ground users rather than flight route planning, which is efficient and practical in certain communication applications. However, the related study is quite few and requires deep consideration. The survey in [4] systematically explores the opportunistic assistance of UAVs for ground networks from a new perspective, including the opportunistic data dissemination, collection, caching, computing, and forwarding. Different promising research directions and related optimization frameworks are enumerated in detail for future study. Liu *et al.* [20] investigated the problem of opportunistic UAV transmission in a D2D communication network where UAVs assist the transmission of D2D users while performing flying missions with the given trajectories. In this model, the

authors develop a dynamic relay mode and a ferry transmission mode for different service demands. By optimizing mode choice, time allocation, and channel competition, the results show that proper schemes for opportunistic UAV transmission can improve global transmission performance significantly. In [21], a real-time relay assignment and a channel allocation are optimized to maximize the long-term average total transmission rate in an opportunistic UAV-assisted dynamic network. Liu *et al.* [22] focused on the problems in dynamic flying ad-hoc networks. According to the topology change, the appropriate idle UAVs and transmission modes are selected to increase opportunistic transmission data. Liu *et al.* [23] developed an air–ground collaborative online planning method to improve the opportunistic data collection performance.

Inspired by the above research contributions, we are interested in the opportunistic UAVs-assisted data transmission maximization in a wireless sensor network. We specially work on the sensor network divided into several clusters based on the location distribution. Considering their needs in data transmission, the clustering can reduce the interference among sensors and the energy consumption of the whole network. The data transmission tasks aggregate at the cluster head first and the aggregated data are sent to the opportunistic UAVs later on. Afterward, the corresponding UAVs forward these data to a data center. Due to the embedded battery in the sensors with energy harvesting devices, the terrestrial wireless sensor network can work independently. The main goal of our optimization problem is to maximize the network data transmission by jointly optimizing UAV scheduling, bandwidth allocation, as well as transmit power. There are several challenges in our problem. For example, it is a mixed-integer nonlinear programming (MINLP) problem. Since the arrival of the sensing data, the energy harvesting of sensor nodes, and the positions of the available UAVs are all dynamic or stochastic, it is difficult or even impossible to obtain complete knowledge about this dynamic network for finding optimal solutions by using convex optimization methods. Thus, we adopt the deep reinforcement learning (DRL) approach to cope with our optimization problem. The contributions of this article are summarized as follows.

1) We propose a novel model for the opportunistic UAVs-aided wireless sensor network consisting of sensor clusters with energy harvesting. Based on our model, we improve the data transmission performance of the terrestrial network without deploying additional communication infrastructure or extra UAVs, which efficiently exploits the idle resource of hovering UAVs with their LoS communication channels. This proposed communication model reduces the transmission interference and the energy consumption. We also propose a K-means-based algorithm for the clustering of randomly distributed wireless sensors. All sensor nodes in the covered area form several transmission clusters with heads for the communication among nodes and UAVs for better transmission performance.
2) Our optimization problem is a MINLP, we cannot solve it directly. We propose to reformulate it as a discrete-time Markov decision process (MDP) based on the real-time information about the amount of transmission task, the harvested energy, the battery power, and the channel condition. Under this MDP, we take the DRL method to deal with the reformulated problem. The designed agent in our DRL approach selects an optimized scheme involving UAV scheduling, bandwidth allocation, and cluster head transmit power. The flexible scheduling and proper power control can effectively enhance the performance of the network. Note that we focus on the UAV opportunistic access rather than flight route planning, which is an independent problem to solve.
3) By using DRL, we develop a DQN-based and a deep deterministic policy gradient (DDPG)-based algorithms to maximize the data transmission and compare them with four baseline policies to demonstrate the superiority of our proposed approaches from different perspectives. The DQN-based algorithm is typical in DRL and well used for the optimization problem with discrete space. Differently, the DDPG-based algorithm can take actions from continuous space to significantly improve the system performance and appear easier to implement in practical environment. Simulation results prove that both proposed DQN and DDPG-based algorithms are valid and effective for our proposed network.

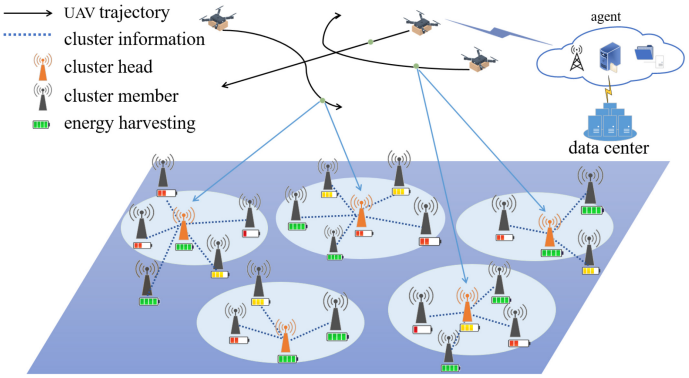


Fig. 1. System architecture.

The remainder of this article is organized as follows. Section II introduces our communication scenario, the selected air-to-ground channel model, and our optimization problem. Section III briefly introduces our clustering algorithm and reviews the DRL background. Then, we propose our reformulated MDP, DRL-based algorithms, and other baseline approaches in detail. The simulation results and computation complexity of our DDPG-based algorithm are illustrated and analyzed in Section IV. Finally, we conclude our article in Section V.

II. System Model

Fig. 1 illustrates an opportunistic UAVs-assisted wireless sensor network architecture consisting of many sensors randomly distributed on the ground while UAVs are flying above. We assume that this communication scenario may happen in a farmland, battlefield, disaster zone, or environmental

protection monitoring where necessary communication infrastructures are hard to build or maintain, and the communication links are prone to be fragile or even unavailable sometimes. Since this sensor network continuously generates data, which ought to be processed and analyzed at the data center, we need to establish fundamental communication links for data transmission. While executing specific missions, some of the aforementioned UAVs still have excess communication resource and ability for working as relays to forward data opportunistically. By this means, the transmitted messages in the area could be perceived and the data transmission of the area could be restored or maintained. Moreover, in a general case, though the communication links can keep available all the time, the utilization of opportunistic data forwarding can still improve the network performance to some extent and make sense.

We suppose that there are certain sensors equipped with energy harvesting devices in the target area, and they are able to harvest energy from renewable energy sources, such as RF [24] or solar. The energy will be stored in their batteries. Although energy harvesting is available, the performance of a sensor node is also constrained by its battery capacity. Therefore, it is unwise for all ground sensors sending data independently within a limited time. Due to the battery restriction, without considering the transmission management among nodes and the communication framework, the data transmission task will suffer severe interference and performance degradation, so that a decline in transmission performance is inevitable in the system. For this reason, clusters of data uploading can be formed among vicinal sensors. In our system, the sensors can be grouped into clusters according to their geographical distribution and the cluster members will aggregate data to the cluster head periodically. Because of the randomness of sensing data, the data transmission tasks gathered at the cluster head vary with time. When UAVs are available, the transmitted data will be uploaded by the cluster heads while the other cluster members remain silent. This communication pattern will reduce transmission interference and save energy, which can improve the communication performance of the system.

A. Communication Scenario Description

We consider a wireless network consisting of many sensor nodes with energy harvesting devices. For easier management and better transmission, we organize the network into clusters and we will introduce this clustering process in the next section. The number of clusters is denoted as $\mathcal{N} = \{1, \ldots, n, \ldots, N\}$. For an arbitrary cluster head n, its coordinate is denoted as $(x_n; y_n)$. In each cluster, there exists only one cluster head. Due to the sensor nodes equipped with the energy harvesting devices, the cluster head is able to keep alive within a complete period. The cluster head has necessary knowledge of its members. The number of nodes in cluster n is denoted as $\mathcal{R} = \{1, \ldots, r, \ldots, R\}$. We select an arbitrary sensor node in the cluster n as $S_{r,n}$. We model the channel gain between the nodes and the cluster head as $h^t_{s_{r,n}} = U^t d^{-\beta}_{r,n}$, where U^t is randomly generated according to the Rayleigh distribution, $d_{r,n}$ denotes the distance between the cluster member and the cluster head, and β is the path-loss exponent. The nodes in the network have an active (generating data packet) probability $H_{r,n}$ and the silent probability is $1 - H_{r,n}$. A binary variable $\varphi^t_{r,n}$ indicates the node is active when its value is equal to 1 or 0 otherwise. During a time slot, each sensor node only sends no more than one data packet to its corresponding cluster head by TDMA. The data transmission time slot is divided equally to all the active nodes in a cluster. Then, the cluster head aggregates all the data packets received into a packet for later transmission. The data collected in cluster n during time slot t is denoted as DG^t_n, which is calculated by

$$R^t_{S_{r,n}} = \begin{cases} 0, & \text{if } \varphi^t_{r,n} = 0 \\ W_0 \log\left(1 + \frac{P^t_{S_{r,n}} h^t_{S_{r,n}}}{\sigma_0^2}\right), & \text{else} \end{cases} \tag{1}$$

$$DG^t_n = \frac{\tau}{\sum_{r=1}^{R} \varphi^t_{r,n}} \sum_{r=1}^{R} R^t_{S_{r,n}} \tag{2}$$

where ${\sigma_0}^2$ denotes the power spectral density of the additive white Gaussian noise (AWGN) at the receivers. W_0 is the bandwidth of the cluster head n. $P^t_{S_{r,n}}$ is the transmission power of node, which is determined by the available battery energy at this time. The battery reserves part of its energy for keeping the corresponding node alive and the rest is used for data transmission within the cluster.

We assume that each cluster head broadcasts its current state periodically over a control channel (CCH) to provide real-time information of the ground network, precisely the concerned service request, the location and energy messages, the channel information, and the amount of current transmission task. Besides, the UAVs periodically aggregate the information of the ground network as well as their own mobility features to coordinate with the control unit where an artificial intelligence (AI) agent resides [15]. The agent monitors the CCH of the network in a long term and makes decisions based on the obtained information. In addition, the UAVs and the cluster heads will consult the agent for the power allocation and scheduling decisions in their following traveling. Since there are no control message exchanges between the UAVs and the cluster heads, their communication links are established on the service channels (SCHs). Because the CCH and SCHs use different frequency bands, the power control and data transmission can be conducted simultaneously.

Moreover, the set of available opportunistic UAVs is denoted as $\mathcal{M} = \{1, \ldots, m, \ldots, M\}$. The position of any UAV m can be indicated as $(x_m; y_m; z_m)$. Time is discretized into slots as $\{t_1, \ldots, t_k, \ldots, t_N\}$. We suppose the length of a time slot denoted as τ and it is sufficiently short. Thus, the position of UAV can be considered approximately unchanged within a single time slot. The time allocation in a period is illustrated in Fig. 2. At the beginning of the period, the clustering on the sensor nodes is made and the cluster heads are determined. Then, the data transmission starts. After a complete period, this process repeats to realize whole time data transmission. Since the proposed network works in the same way in each period, we only concentrate on the data transmission

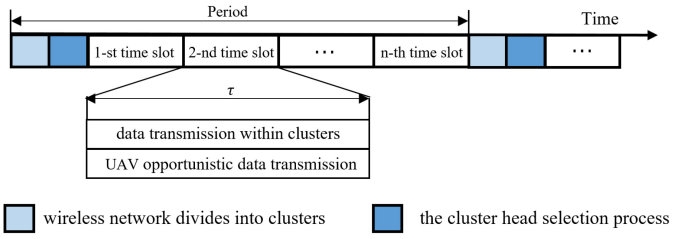


Fig. 2. Time slot segmentation for clustering and data transmission.

performance within a complete period. In each time slot, the data transmission within clusters and the UAV opportunistic transmission are at the same time. The transmission tasks aggregated at a cluster head will be transmitted in the next time slot. For the UAVs, they will match with the corresponding clusters according to the given schemes and each UAV can provide service for a certain number of clusters synchronously by OFDMA. We denote the service threshold of UAV as S_{th}, and let $\alpha^t_{m,n}$ be a binary variable to indicate cluster n served by UAV m in time slot t if $\alpha^t_{m,n} = 1$, or 0 otherwise. The bandwidth of UAVs divided for multiusers and the real-time uploading power of the cluster heads are both dynamic. A designed power control scheme is required for efficient transmission. They will adjust the transmit power and bandwidth allocation according to an optimized scheme. After UAVs receive data, they will forward the messages to the data center later to restore the broken communication links and enhance the capacity of the data transmission in this area.

In this article, we aim to obtain proper transmission policies by jointly optimizing the UAV scheduling, the bandwidth allocation, and the transmit power of the cluster heads. Since the UAV positions, the real-time transmission tasks, and the situation of these clusters are dynamic in different time slot, a flexible scheduling strategy can lead to better transmission performance and the proper power control will raise the energy efficiency.

B. A2G Channel Model

Air-to-ground (A2G) communication has gained much attention in both industry and academia recently for opening a promising way toward 5G applications. We apply one of the most widely investigated A2G channel model, which considers the probability of both LoS and NLoS links in this work. According to [25], the path loss between cluster head n and UAV m at time slot t is denoted as

$$L^t_{m,n} = \frac{\eta_{\text{LoS}} - \eta_{\text{NLoS}}}{1 + a\exp\left[-b(\theta^t_{m,n} - a)\right]} + 20\log\left(\frac{4\pi f_c d^t_{m,n}}{c}\right) + \eta_{\text{NLoS}} \tag{3}$$

where η_{LoS}, η_{NLoS}, a, and b are the constants determined by communication environment, such as rural, suburban, urban, or high-rise urban areas. Their values vary in different environments. Particularly, η_{LoS} and η_{NLoS} denote the mean additional loss for LoS and NLoS links, respectively. f_c denotes the center frequency of the carrier and c is the velocity of light. $\theta^t_{m,n}$ is the elevation angle from cluster head n to UAV m and $d^t_{m,n}$ is the distance between them. More specifically, $d^t_{m,n}$ and $\theta^t_{m,n}$ are

$$d^t_{m,n} = \sqrt{(x_n - x_m)^2 + (y_n - y_m)^2 + z_m^2} \tag{4}$$

$$\theta^t_{m,n} = \sin^{-1}\left(\frac{z_m}{d^t_{m,n}}\right) \tag{5}$$

where z_m is the flight altitude of UAV. The channel gain in time slot t can be calculated by

$$h^t_{m,n} = 10^{-L^t_{m,n}/10}. \tag{6}$$

Thus, the instantaneous data transmission rate expression from cluster head n to UAV m at time slot t is given by

$$R^t_{m,n} = \alpha^t_{m,n} W^t_{m,n} \log\left(1+\gamma^t_{m,n}\right) \tag{7}$$

$$\gamma^t_{m,n} = \frac{P^t_{m,n} h^t_{m,n}}{\sigma^2} \tag{8}$$

where $\alpha^t_{m,n}$ indicates whether the cluster is served. $W^t_{m,n}$ is the bandwidth allocation of UAV m to cluster n in the current time slot. $\gamma^t_{m,n}$ is the signal-to-noise ratio (SNR) between cluster head n and UAV m. $\sigma^2 = W^t_{m,n} N_0$ where N_0 denotes the power spectral density of the AWGN at the receivers. Because of the mobility of UAV, the real-time path loss and transmission rate are varying, which has impacts on the successful data transmission. In this model, during the flight of UAV, we design a proper matching scheme and a power allocation algorithm among UAVs and clusters to enhance the transmission performance of our system. For clear presentation, the notations applied in this problem are mainly summarized in Table I.

TABLE I
TABLE OF NOTATIONS

Parameters	Description
N	Number of clusters
M	Number of opportunistic UAVs
$(x_n; y_n)$	Position of cluster head
$(x_m; y_m; z_m)$	Position of opportunistic UAVs
τ	Length of a time slot
S_{th}	Service threshold of UAVs
$B_{\max}$	Battery capacity
$P_{\max}$	Maximum transmit power
$\alpha^t_{m,n}$	indicates whether cluster n is served by UAV m
$d^t_{m,n}$	Distance between node n to UAV m
$\theta^t_{m,n}$	Elevation angle from node n to UAV m
$h^t_{m,n}$	Pathloss between node n to UAV m
N_0	Noise power spectral density
$\gamma^t_{\text{m,n}}$	SNR between node n and UAV m
$R^t_{m,n}$	Transmission rate between node n and UAV m
$W^t_{m,n}$	Bandwidth allocation of UAV m to node n
DG^t_n	the transmission task aggregated at cluster n

C. Optimization Problem Formulation

This section presents the problem formulation. With the assistance of the opportunistic UAV, the amount of total transmitted data of the network in each time slot can be given by

$$D(t) = \sum_{n=1}^{N} \min\left(\sum_{m=1}^{M} R^t_{m,n}\tau, DG^t_n\right) \tag{9}$$

where M and N denote the numbers of UAVs and clusters, respectively. DG_n^t is the randomly arrived transmission task at each cluster head as an upper limit for current data transmission. The objective of this article is to optimize the transmission scheduling and the power control of the network, including the bandwidth allocation of UAV and the transmit power of cluster heads of sensor nodes. We cope with this problem under the mobility of UAVs and necessary constraints. Our optimization problem can be formulated as

$$\mathcal{OP}: \max_{\alpha_{m,n}^t, W_{m,n}^t, P_{m,n}^t} \lim_{T\to+\infty} \frac{1}{T}\sum_{t=0}^{T-1} D(t) \tag{10}$$

$$\text{s.t. } C1: \sum_{n=1}^{N} \alpha_{m,n}^t \le S_{th} \quad \forall m \quad \forall t \tag{10a}$$

$$C2: \sum_{m=1}^{M} \alpha_{m,n}^t \le 1 \quad \forall n \quad \forall t \tag{10b}$$

$$C3: \sum_{n=1}^{N} W_{m,n}^t \le W_m^t \ell \quad \forall n \quad \forall t \tag{10c}$$

$$C4: 0 \le P_{m,n}^t \le \min\left\{ B_{m,n}^t + E_{m,n}^t, B_{\max} \right\} \tag{10d}$$

$$C5: 0 \le B_{m,n}^t \le B_{\max} \tag{10e}$$

$$C6: 0 \le E_{m,n}^t \le E_{\max}. \tag{10f}$$

Our optimization objective (10) represents the long-term average total data transmission of this proposed network, where $B_{\max}$ and $E_{\max}$ are the upper limits of battery capacity and the harvested energy of the cluster heads. S_{th} is the service threshold of UAVs. $P_{m,n}^t$ and $W_{m,n}^t$ indicate the current power allocation for a cluster head and UAV bandwidth, respectively. Constraints $C1$ and $C2$ ensure that each UAV only schedules at most S_{th} clusters and each cluster is only scheduled by one UAV at most in a time slot. Constraint $C3$ insures that the sum of the allocated bandwidth cannot exceed the available bandwidth of the UAV. W_m^t is the total bandwidth of UAV m, and ℓ indicates the ratio of the remaining bandwidth resource, which is available for the opportunistic assistance. Constraints $C4$ and $C5$ guarantee that the uploading power is limited by the battery capacity of cluster head. Constraint $C6$ indicates that the harvested energy of sensor nodes cannot be infinite.

This optimization problem is challenging to solve, since it is a MINLP. Besides, because of the mobility of the UAVs and randomness of the arrival tasks, it is not practical or even impossible to obtain complete knowledge about the network. In order to find solutions for this problem, we take DRL to interact with the changing and complicated environment and make proper decisions due to its ability for tackling this issue. In the next section, we will introduce our proposed sensor clustering algorithm and DRL-based algorithms for the formulated optimization.

III. Solutions

A. Sensor Node Clustering

For easier management and better data transmission, we organize the wireless sensors of this network into clusters. Since the key of our research focuses on the optimization problem (10) rather than wireless network protocol, we briefly introduce our clustering process. There are two important tasks that should be finished in the clustering: 1) the cluster formation and 2) the cluster head selection. We apply a K-means-based algorithm, one of the most classical unsupervised learning algorithms to form the clusters among vicinal sensor nodes, which is shown in Table II. This algorithm partitions the whole network into the given number of clusters based on the geographical distribution of sensors with an iterative way where each sensor node only belongs to only one cluster. After the clustering is completed, the clustering message is sent to all the nodes, and then the nodes in a cluster will run for the cluster head. In a cluster, there exists only one cluster head. Considering the long-term existence of the sensor network, there may be some inevitable accidents, such as certain out of function sensor nodes or network topology variation because of changing environment. We suggest to redetermine the cluster formation and cluster head to adapt to these uncertainties by the K-means-based algorithm after a complete period. Simply, we choose the cluster head according to their positions because of our mentioned possible applications. There is an advantage of this clustering principle in the data transmission for a cluster head who is the closest to all the other nodes.

B. Deep Reinforcement Learning

When we use DRL, we adopt an AI agent located at the data center that interacts with the dynamic environment in a sequence of actions, observations, and rewards. This AI agent is capable of observing the network transmission and learning policies in scheduling and power control among the deployed UAVs and nodes. Meanwhile, the time-varying messages about the wireless sensor network can also be observed by the UAVs, since they periodically aggregate the ground network information and their own mobility features, and coordinate with the agent. With the obtained information and the constructed neural networks, by taking DRL, the AI agent can finally select optimal schemes for this UAV-aided network. Precisely, at each time slot t, the AI agent decides an action for each opportunistic UAV and the cluster head. The deployed UAVs will connect the corresponding cluster heads and then conduct their bandwidth allocation according to the given action. Note that the speed and trajectory of the UAVs cannot be influenced by the decided action, because our system model adopts the opportunistic UAV utilization whose mobility only obeys their previously made mission arrangement. On the other hand, the terrestrial sensor nodes use the corresponding transmit power to achieve optimal power control. After performing each selected action, the system receives a step reward to indicate how much this selected action contributes to the objective and then it will traverse to the next state. Since the attained reward depends on the entire previous sequence of actions and states from the environment, the influence of current action can only be found after many time slots. During the above process, the agent continually observes the environment changes and modifies the system state representation. In order to maximize the network data transmission by the opportunistic UAVs, the agent regulates the transmission policy to

make better decisions, which leads to steady communication performance improvement in the long run.

We propose our DRL-based algorithms to obtain effective scheduling and power control strategies for the performance improvement in the whole network, since DRL is one of the most popular machine learning algorithms, which fully combines the decision-making capabilities of reinforcement learning (RL) and the representation advantages of deep learning (DL). RL with multiple hidden layers and an appropriate input–output pattern representation provides a possible vision for solving complex optimization problems. RL is an intelligent learning paradigm based on the MDP. In an MDP, an agent interacts with corresponding environment and continuously evaluates action values to find a best policy in an action set $\pi^* = \{a_1, \ldots, a_t, \ldots, a_N\}$, which maps a state set $S = \{s_1, \ldots, s_t, \ldots, s_N\}$ to a legitimate action and allows the agent to act optimally in the environment. In this way, it is able to maximize all discounted cumulative future rewards.

Importantly, the DRL approach improves the performance of RL in many respects. The classic DQN algorithm achieves better performance compared with separate RL and DL by combining a Q-learning algorithm with a deep neural network (DNN). The action-value function $Q(.)$ in Q-learning updates according to

$$Q^{\pi}(s_t, a_t) \leftarrow Q^{\pi}(s_t, a_t) + \eta\left\{r_t + \lambda \max_{a+1} Q^{\pi}(s_{t+1}, a_{t+1}) - Q^{\pi}(s_t, a_t)\right\} \tag{11}$$

where η is the learning rate and λ is the discount factor in the algorithm.

A DQN algorithm not only maintains the advantages of decision making in the Q-learning approach but also utilizes DNNs to approximate the action-value function $Q(.)$ to evaluate a reward r_t obtained by applying a certain policy π, i.e.,

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left\{\sum_{t=1}^{N} r_t | s_t, a_t, \pi_t\right\} \tag{12}$$

$$Q^{\pi}(s_t, a_t) = \mathbb{E}\left\{r_t + \lambda Q^{\pi}(s_{t+1}, a_{t+1})\right\} \tag{13}$$

where π_t is the policy selected by the agent, consisting of sequential decisions about the decided action at each time slot, and $\mathbb{E}(.)$ represents the expectation calculation. Thus, (12) denotes the expected value of the cumulative reward obtained by executing an action from the given policy under the current state. Moreover, the function $Q(.)$ defined by (13) is the Bellman equation, which presents a recursive relationship to obtain an optimal policy.

There are two sets of identical DNN in a DQN algorithm to train in a stable way. A estimated network Q constantly changes its parameters to minimize the loss function value while the target network Q' is usually fixed and only updates the parameters at certain intervals. The loss function of a DQN algorithm is expressed as

$$\text{Loss}(w) = [r_t + \lambda \max_{a_{t+1}} Q'(s_{t+1}, a_{t+1}|w') - Q(s_t, a_t|w)]^2 \tag{14}$$

where w denotes the estimated network parameter and w' stands for the target network parameter.

For the stable performance of an algorithm, an experience replay is applied in a DQN. We know when the training data in a DNN are correlated, it probably leads to convergence difficulty of a model and the loss values appear continuous fluctuation. To solve these problems, a random minibatch of transition from experience replay buffer is sampled for each training process. It significantly mitigates the correlations among consecutive transitions and increases their independency. Importantly, the DQN algorithm lays a solid foundation for the follow-up DRL development.

Even though the DQN algorithm is simple and effective, and it has achieved high performance in DRL for some simple applications, it still has some limitations. Its model is prone to overestimation and it is hard to handle continuous-valued control tasks, which limits its applications to certain extent. Although by discretizing continuous action space, a DQN algorithm can complete continuous control tasks approximately, and its huge dimensions may cause some problems. Additionally, the Degree of Freedom (DoF) can expand action space size exponentially and make the neural network hard to train. In our model, if we discretize the bandwidth and transmit power allocation into 10 DoF, respectively, the action space size is about one million, which is difficult for model training. Thus, for these reasons, we will propose a DQN-based algorithm and an actor–critic-based DDPG algorithm to demonstrate their effectiveness for our problem in consideration of their real application and comparison.

A DDPG algorithm consists of two neural networks: 1) an actor network and 2) a critic network. In this algorithm, a DNN turns as an actor to choose an action and a deep Q-network mentioned above behaves as a critic network, which interacts with the actor and tells whether the action is appropriate. Repeating this process, the actor gradually finds out how to choose proper action in each state while the critic constantly iterates to improve the state-action values. These two neural networks, respectively, output specific action values and Q values of the current state–action and update the corresponding neural network weights by a stochastic gradient descent (SGD) method. Moreover, both the critic network and the actor network are composed of two subnets: 1) an online network and 2) a target network, under the same architecture illustrated in Fig. 3. These four neural networks contain many layers with corresponding parameters. Here, the parameters of the actor and the critic in the online networks are denoted as θ^{μ} and θ^{Q}, respectively, while $\theta^{\mu'}$ and $\theta^{Q'}$ for the target network, respectively. Back to our case, the UAV scheduling and power allocation in the network are determined by this given structure. The objective of the actor neural network is to continuously gain knowledge about how to map the state of the environment to the best action, which can maximize the cumulative reward sum. Instead of using a loss function, the actor network utilizes the gradient $J(\theta^{\mu}) = \mathbb{E}\{R_t|s_t, a_t\}$. The policy of the actor in online network is updated by

$$\nabla_{\theta^{\mu}} J = \mathbb{E}_{s_t \sim \rho^{\beta}}\left\{\nabla_a Q\Big(s, a|\theta^{Q}\Big)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^{\mu}} \mu(s|\theta^{\mu})|_{s=s_t}\right\} \tag{15}$$

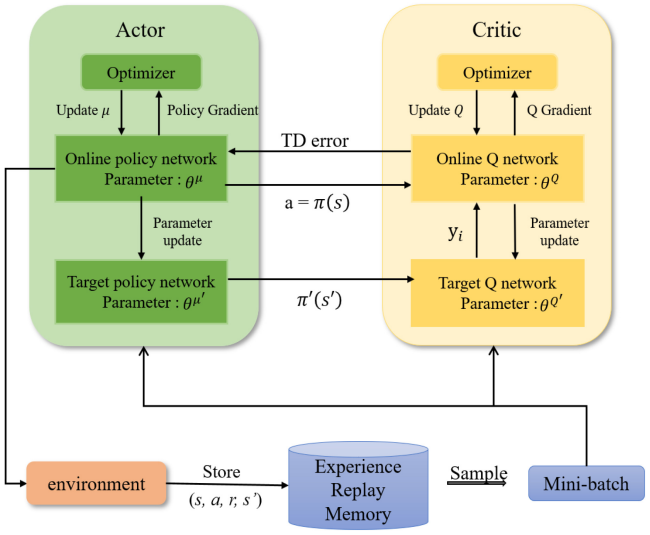


Fig. 3. Block diagram of DDPG.

where $Q(s, a|\theta^Q)$ is the optimal expected long-term return locally and immediately available for each state–action pair. ρ^β denotes the state distribution of s_t. The objective function is denoted as

$$J(\theta^\mu) = \mathbb{E}_{\theta^\mu}\left\{r_1 + \lambda r_2 + \cdots + \lambda^{t-1} r_t\right\}, \lambda \in [0, 1]. \quad (16)$$

The critic network is a deep Q-network trained to approximate a real Q-table using neural networks. Moreover, just like the DQN algorithm, it also takes an experience replay and target network to assist its convergence. The loss of the critic network is equal to the difference between two sides of the Bellman equation, so that the critic can be updated through minimizing the temporal difference error

$$\text{Loss}(\theta^Q) = \left[r_t + \lambda Q'\left(s_{t+1}, a_{t+1}|\theta^{Q'}\right) - Q\left(s_t, a_t|\theta^Q\right)\right]^2. \quad (17)$$

In the following, we introduce the procedures of an entire training in our DDPG algorithm. First, the actor decides an action $a_t = \mu(s_t|\theta^\mu) + N_t$ based on the current online network $\mu(s_t|\theta^\mu)$ and a random noise N_t. After executing a_t in the facing environment, it will go to the next state s_{t+1} and obtain a reward r_t. The tuples (s_t, a_t, r_t, s_{t+1}) generated by state transitions are stored in the experience replay buffer and work as the training set for the online network. A minibatch consisting of N samples from the buffer can be randomly selected to update both the actor network and the critic network. Using the available training data, we calculate the loss according to (17). We also take the SGD approach to update the corresponding neural network parameters. The online neutral networks can be updated by (15) and (17). Finally, the target network is updated with a small constant ξ in our DDPG algorithm, i.e.,

$$\theta^{Q'} \leftarrow \xi\theta^Q + (1 - \xi)\theta^{Q'} \quad (18a)$$

$$\theta^{\mu'} \leftarrow \xi\theta^\mu + (1 - \xi)\theta^{\mu'} \quad (18b)$$

where $\xi \in (0, 1)$ represents the rate at which the network is updated.

In the policy gradient (PG), it chooses a random action from a determined distribution. In contrast, our DDPG approach directly generates an optimal action a_t with given state s_t by a deterministic strategy $a_t = \mu(s_t|\theta^\mu)$, and an actor outputs a specific value rather than an action probability in DDPG, which leads to superiority in learning high-dimensional actions with fast convergence.

C. Markov Decision Process Model

To solve an optimization problem by the DRL method whose output only depends on past and current input, we need to know a series of causal information. Since our problem is a nonconvex mixed-integer optimization, we must reformulate it as a discrete-time MDP.

At the beginning of UAV-aided data transmission in our system, the agent observes the real-time sensor network environment and collects all the parameters associated with the state. In this model, the movement of UAV, the wireless fading channels, and the battery remaining energy all have the Markov properties. Therefore, the maximization of the data transmission of the system can be described as a discrete-time MDP. Mathematically, a complete MDP can be represented by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. Considering that the state transition probability $\mathcal{P}$ is unknown in our problem, we redefine our MDP as a triple $(\mathcal{S}, \mathcal{A}, \mathcal{R})$. A DRL-based framework is proposed to formulate our optimization problem into an MDP by defining states, actions, and rewards in this tuple within time slot t as follows.

1) State Space:

$$\mathcal{S}^t = \left[B_1^t, \ldots, B_n^t, EH_1^t, \ldots, EH_n^t, DG_1^t, \ldots, DG_n^t \right.$$
$$\left. C_{1,1}^t, \ldots, C_{m,n}^t \mid t = t_1, \ldots, t_N\right].$$

In this space, m and n denote a specific UAV and a cluster head, respectively. The agent obtains the battery power B_n^t, the energy harvesting message EH_n^t, and the transmission task amount DG_n^t of the cluster heads calculated by (2), as well as a series of path-loss information $C_{m,n}^t$ among nodes and UAVs at the real-time position at the current time slot t. According to these information, the agent can choose an action for the whole network.

2) Action Space:

$$\mathcal{A}^t = \left[\alpha_{1,1}^t, \ldots, \alpha_{m,n}^t, P_{1,1}^t, \ldots, P_{m,n}^t \right.$$
$$\left. W_{1,1}^t, \ldots, W_{m,n}^t \mid t = t_1, \ldots, t_N\right].$$

The action decided by the agent consists of the scheduling vector $\alpha_{m,n}^t$ and the power control $P_{m,n}^t$, and the bandwidth allocation $W_{m,n}^t$ in each time slot. $\alpha_{m,n}^t$ indicates the current matching choices about which clusters are served by the chosen UAVs. The transmit power of the cluster head and the UAV bandwidth allocation is regulated according to $P_{m,n}^t$ and $W_{m,n}^t$, respectively. Therefore, the action space determines the instant reward.

3) Immediate Return:

$$\mathcal{R}^t = [D(t) \mid t = t_1, \ldots, t_N].$$

In this article, our target is to maximize the data transmission amount of the network under certain constraints. Thus, we take the cumulative transmitted data $D(t)$ in time slot t as our immediate reward r_t calculated by (9) in the tuple.

After executing an action, the agent will obtain an immediate feedback as a standard for the following decision making. We hope to maximize the cumulative discounted reward in the long term, so that our proposed transmission strategies can be gradually evaluated and improved by continuous network training.

D. Proposed Algorithms

We expect to obtain a transmission strategy of scheduling and power control of the whole network to maximize network data transmission. Practically, the power control can be considered as a continuous task. The framework of DRL is very appropriate to capture infinite characteristics of action space. It is worth mentioning that our optimization is a problem in discrete-continuous hybrid action space where the matching process among the UAVs and cluster heads is a discrete action, but the power allocation of network is a continuous one. In order to cope with this problem and reduce the calculation complexity of the strategy, we approximate the hybrid space through discretization using the DQN algorithm by which the transmit power and bandwidth allocation are discretized into a series of fixed numbers to be selected. Then, we propose a DQN-based algorithm to solve problem (10) given in Table III where the size of the experience pool Z is X'. Moreover, it is also possible to relax the action space to a continuous set by using the DDPG algorithm to weaken the limitations of the DQN algorithm. We find a DDPG-based solution and take our proposed DQN-based approach as the parallel comparison algorithm. We summarize our DDPG-based UAV scheduling and power control algorithm in Table IV where the replay memory pool is Z with its size X and we take E episodes in the training.

To further investigate the effectiveness of our proposed algorithms and the impact of optimized factors on the optimization problem (10), we propose other four approaches for the performance comparison.

1) *Optimized Power Control With Greedy Scheduling (OPFS):* This approach is simplified from our proposed DDPG-based algorithm with only partial optimization. In this approach, at each time slot, we only optimize the power control of UAVs and cluster heads under the required constraints by the DDPG-based algorithm and we always select a fixed set of clusters for data transmission.
2) *Greedy Power Control With Optimized Scheduling (GPOS):* This approach is also simplified from our proposed DDPG-based algorithm with only partial optimization. In this approach, at each time slot, we only optimize the UAV scheduling under the required constraints by the DDPG-based algorithm. We adopt a greedy plan for power control in which the bandwidth of UAVs is equally allocated for their users and the cluster head consumes all the available energy remained in the battery regardless of future use.
3) *Greedy Power Control With Fixed Scheduling (GPFS):* In this approach, at each time slot, we select the fixed clusters with the greedy power control in which the bandwidth of UAVs is equally allocated for their users and the cluster head consumes all the available energy remained in the battery regardless of future use.
4) *Random Power Control With Random Scheduling (RPRS):* In this approach, at each time slot, we randomly select all the actions for the system.

All the simulation results are illustrated and analyzed in the next section.

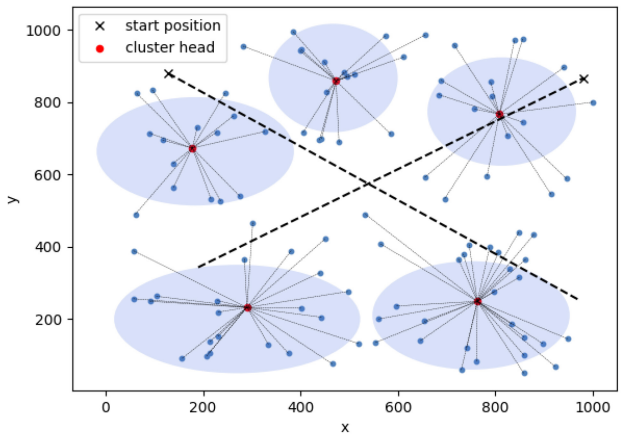


Fig. 4. Clustering result and flight directions of two UAVs.

IV. Simulation Results

In this section, we evaluate the performance of our proposed DQN and DDPG-based algorithms and compare them with four other approaches to show the superiority of our proposed algorithms and the impact of our optimized factors in this optimization problem. The numerical results demonstrate that our proposed algorithms are more effective to enhance the system transmission performance in the long term.

A. Simulation Setup

Our simulation runs under the environment of software and hardware configuration of i5 CPU, 8-GB RAM, Win10 operating system, Python 3.6 version, and TensorFlow 1.14.0. The training adopts a set of actor–critic neural networks with two hidden layers where the number of neurons in each layer is 100 and 400, respectively. The critic neural network also has two hidden layers where the number of neurons in each layer is 900 and 100, respectively. All the simulation parameters are given in Table V.

In our simulations, in consideration of the problem scale, we take 100 sensor nodes distributed randomly in a 1 km $\times$ 1 km area and use the proposed clustering algorithm in Table II to divide them into five clusters shown in Fig. 4 for the following simulations. Over the network, the opportunistic UAVs fly through the target area. We might as well assume there are two UAVs performing goods delivering tasks from one location to another following fixed trajectories restricted by warehouse layout factors. They move in a velocity of 10 m/s and the given directions from the initialized positions at the time. When reaching the edge of the target area, two UAVs are ready to serve the covered area. When they leave this area, the

TABLE II
K-MEANS-BASED CLUSTERING ALGORITHM

Algorithm 1

Input: Dataset $D = \{x_1, x_2, ..., x_m\}$, clustering number k.
Output: Clusters $C = \{C_1, C_2, ..., C_k\}$
1: Initialize mean vector $\{z_1, z_2, ..., z_m\}$ by randomly selecting k samples
2: **repeat**
3: let $C_i = \emptyset (1 \le i \le k)$
4: **for** $j = 1, 2, ..., m$ **do**
5: Calculate distance between sample x_j and each mean vector $z_i (1 \le i \le k)$ with $d_{ji} = \|x_j - z_i\|_2$
6: Determine clustering mark based on the distance $\lambda_j = \arg\min_{i \in \{1,2,...k\}} d_{ji}$
7: Divide sample into a cluster $C_{\lambda_j} = C_{\lambda_j} \cup \{x_j\}$
8: **end for**
9: **for** $i = 1, 2, ..., k$ **do**
10: Calculate new mean vector $z_i' = \frac{1}{|C_i|} \sum_{x \in C_i} x$
11: **if** $z_i \neq z_i'$ **then**
12: Update z_i with z_i'.
13: **end if**
14: **end for**
15: **until** none of z_i' is updated

TABLE III
DQN-BASED UAV SCHEDULING AND POWER CONTROL ALGORITHM

Algorithm 2

Input: parameters X', λ, η.
Output: w and optimal action a_t^* of time slot t.
1: Initialize replay memory pool Z with its size X'.
2: Initialize estimated network parameter w with random weight and target network with $w' \leftarrow w$.
3: **for** $episode = 1, 2, ..., E$ **do**
4: Initialize our scenario and observe environment state s_1 .
5: **for** $t = 1, 2, ..., N$ **do**
6: Take action $a_t = \arg\max_a Q(s_t, a, w)$ or choose an action with a certain probability.
7: Obtain reward r_t and observe next state s_{t+1}.
8: Store data (s_t, a_t, r_t, s_{t+1}) in Z.
9: **if** Z is full, **do**
10: Sample sets of data (s_t, a_t, r_t, s_{t+1}) from Z randomly.
11: Update estimated network by (14).
12: Update target network with $w' \leftarrow w$.
13: **end for**
14: **end for**
15: Return w.
16: Choose optimal action $a_t^* = \arg\max_a Q(s_t, a, w)$ at time slot t.

service finishes. We do not pay more attention to this process since it not difficult to get the necessary information about the edges of the area.

B. Results and Analysis

There are certain factors that may affect the performance of the UAVs-assisted sensor network. In this section, we are going to testify the superiority of our proposed algorithms and investigate how these factors affect the transmission performance.

We first analyze the cumulative reward and the convergence performance of our proposed two algorithms. To demonstrate the superiority of our proposed algorithms, the DQN-based

TABLE IV
DDPG-BASED UAV SCHEDULING AND POWER CONTROL ALGORITHM

Algorithm 3

Input: parameters X, ξ, λ, η.
Output: θ^μ and optimal action a_t^* of time slot t.
1: Initialize replay memory pool Z with its size X.
2: Initialize online network parameters θ^μ, θ^Q with random weights and target network with $\theta^{\mu'} \leftarrow \theta^\mu$ and $\theta^{Q'} \leftarrow \theta^Q$.
3: **for** $episode = 1, 2, ..., E$ **do**
4: Initialize our scenario and observe environment state s_1.
5: **for** $t = 1, 2, ..., N$ **do**
6: Take action $a_t = \mu(s_t|\theta^\mu) + N_t$, where N_t is a random noise for action exploration.
7: Obtain reward r_t and observe next state s_{t+1}.
8: Store data (s_t, a_t, r_t, s_{t+1}) in Z.
9: **if** Z is full, **do**
10: Sample sets of data (s_t, a_t, r_t, s_{t+1}) from Z randomly.
11: Update critic of the online network by (17).
12: Update actor of the online network by (15).
13: Update target network by (18).
14: **end for**
15: **end for**
16: Return θ^μ.
17: Choose optimal action $a_t^* = \mu(s_t|\theta^\mu)$ at time slot t.

TABLE V
SIMULATION PARAMETERS

Parameter	Value
Number of sensor nodes	100
UAV altitude	100 m
Available bandwidth for each UAV	5 MHz
Carrier frequency f_c	2 GHz
Noise power spectrum density N_0	-174 dBm/Hz
Battery capacity $B_{\max}$	1 J
Maximum energy arrival $E_{\max}$	0.8 J
Transmit power $P_{\max}$	$[0, 1]$ W
Length of time slot τ	0.5 s
UAV velocity v	10 m/s
Size of replay memory Z	6000
Batch size	32
(a, b, η_{LoS}, η_{NLoS})	(9.61, 0.16, 1, 20)

algorithm divides the continuous action space into a discrete action space and the power allocation is limited by finite discrete quantities. As shown in Fig. 5, the cumulative reward of the DDPG-based solution is higher than that of the DQN-based algorithm. For the proposed DDPG-based algorithm, the reward increases quickly at the first 300 episodes and becomes almost constant. The reasons are that the network lacks of information about highly rewarding experience at the beginning of the training, and the action decisions are mostly made randomly. However, based on the continuous training process, the DRL agent gradually learns how to adopt proper actions to maintain higher rewards; however, the rewards cannot keep increasing because of the constraints. We can also observe that the reward of the DQN-based approach increases from about 400–520 Mb occasionally, but roughly fluctuates around 500 Mb, which is behind that of the DDPG-based algorithm. This is because that its bandwidth allocation and uploading power control of the cluster head are restricted by the finite discrete quantities in each time slot while the DDPG-based

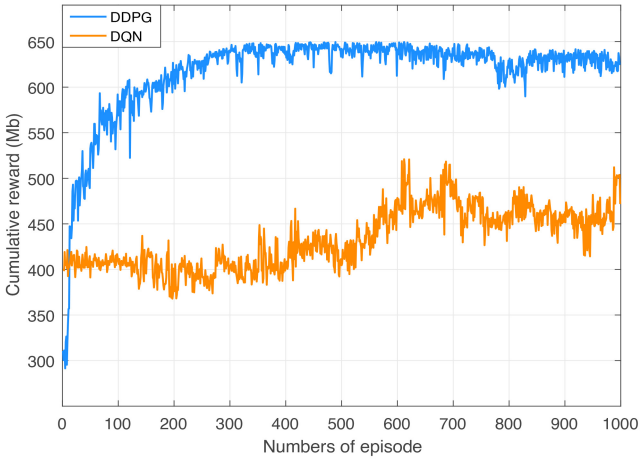


Fig. 5. Reward of DDPG and DQN-based algorithms.

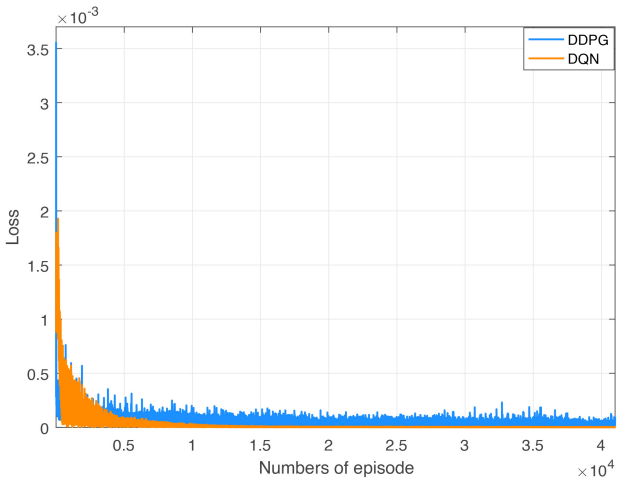


Fig. 6. Loss of DDPG and DQN-based algorithms.

algorithm selects better actions without quantization errors. Therefore, the DQN-based approach can only achieve proper performance under these constraints and it is less better than the DDPG-based algorithm designed based on the continuous action space.

The loss of the DDPG and DQN-based algorithms is illustrated in Fig. 6. When the loss value appears quite small after some periods of training episodes, the current neural network parameters and performance tend to be stable. The DQN-based approach needs more steps than the DDPG-based algorithm to reach a stable status, because its high DoF increases the action space size exponentially, and the obtained proper schemes require to try a large number of discrete actions. We find that DDPG-based algorithm converges after about 14 000 training steps while the DQN-based algorithm converges after 25 000 steps, which shows that the DDPG-based algorithm converges faster than the DQN-based algorithm. However, the loss of the DDPG-based algorithm is obviously larger than that of the DQN-based algorithm. This is due to the fact that the DQN-based algorithm only needs to update one state–action neural network; however, there are two neural networks to be updated in the DDPG-based algorithm with inevitable errors. This tells us that if the problem scale is smaller, or with less

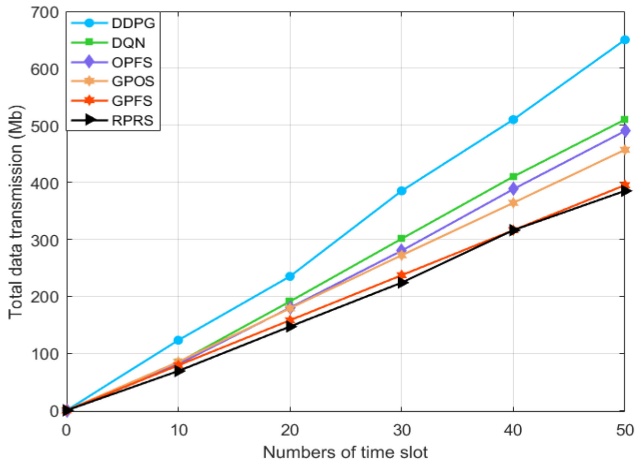


Fig. 7. Performance comparison of different approaches.

sensors, the DQN-based algorithm can also provide satisfied solutions.

In Fig. 7, the transmission performance of six approaches is illustrated within 50 time slots for an episode. We can see that our proposed DDPG-based algorithm demonstrates the best performance, followed by the DQN-based algorithm, OPFS algorithm, GPOS algorithm, and the rest two traditional algorithms, namely, the GPFS and the RPRS approaches. We know that the OPFS and GPOS algorithms are simplified from the original DDPG approach but only conduct partial optimization. Comparing the performance of the DDPG-based algorithm with those of the OPFS and GPOS algorithms, respectively, we can see that our power control optimization and UAV scheduling are both valid and effective. Since the OPFS algorithm shows a better performance comparing with that of the GPOS algorithm, it concludes that a proper power control may be more effective and influential than UAV scheduling in our system. Due to the DDPG-based algorithm, which does not divide an action space into a discrete one, it usually performs better than the DQN-based algorithm. Considering the two partial optimized approaches, it is reasonable to see that they perform worse than the DQN-based algorithm. Furthermore, when we compare the GPFS and the RPRS algorithms with other DRL-based algorithms, we find that in some cases, the RPRS algorithm may perform better than the GPFS one because of its randomness. However, these two algorithms are always inferior to the DRL-based algorithms, which can learn from the environment to adjust the actions intelligently. Similarly, we can find out the effectiveness of the power control optimization by observing the performance gap between the GPFS and OPFS algorithms, and realize the superiority of the scheduling optimization by comparing the performance between the GPOS and GPFS algorithms. Therefore, these numerical results testify that our proposed two approaches can enhance the network transmission performance, especially the DDPG-based one.

The transmission performance with different battery capacities $B_{\max}$ is shown in Fig. 8 with the quantization step 0.25 J. From Fig. 8, we can see that the transmission performance of all the algorithms improves as $B_{\max}$ increases, which is consistent with actual situation. Particularly, the performance

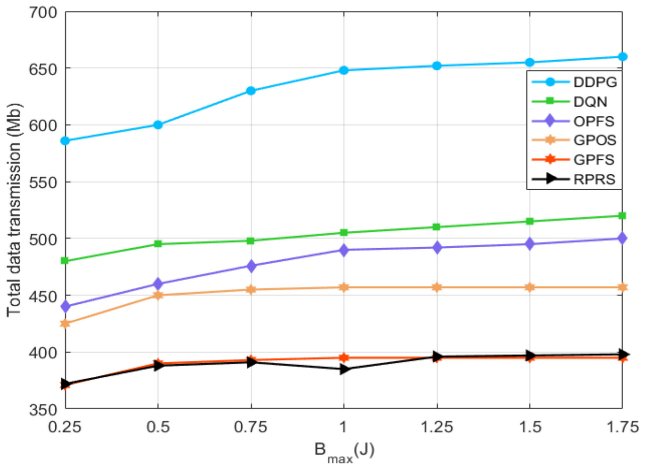


Fig. 8. Performance comparison of different approaches with battery capacity.

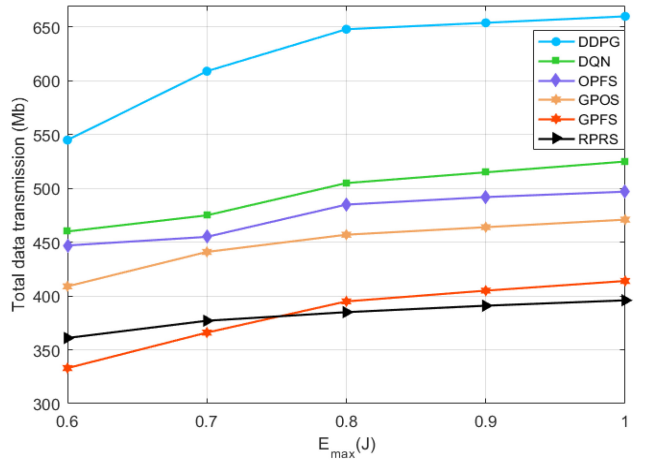


Fig. 9. Performance comparison of different approaches with maximum harvested energy.

of the DDPG-based algorithm is always obviously leading in all the compared algorithms. As the battery capacity increases from 0.25 to 1 J, the increase of total data transmission is clear, but this increase slows down and seems constant later. This is because the performance of the transmission capacity is limited by the harvested energy in the cluster heads whose maximum value is 0.8 J in each time slot. When the battery capacity increases from 1.25 to 1.75 J, which is much larger than the harvested energy limit, the changes hardly make sense. Since the GPOS, GPFS, and RPRS algorithms adopt the traditional schemes rather than DRL-based algorithms on power control, their transmission performance almost do not change after the battery capacity reaches a certain threshold.

Similarly, the transmission performance with the varied maximum harvested energy $E_{\max}$ is shown in Fig. 9. With 0.1 J increase in the energy harvesting process successively, we can observe that the performance behavior is almost the same as that in Fig. 8. Under different maximum harvested energy $E_{\max}$, the total data transmission of our proposed DDPG-based algorithm is always the highest. Furthermore, we find that when the maximum harvested energy $E_{\max}$ is lower than 0.75 J, and the performance of the GPFS algorithm is inferior to that of

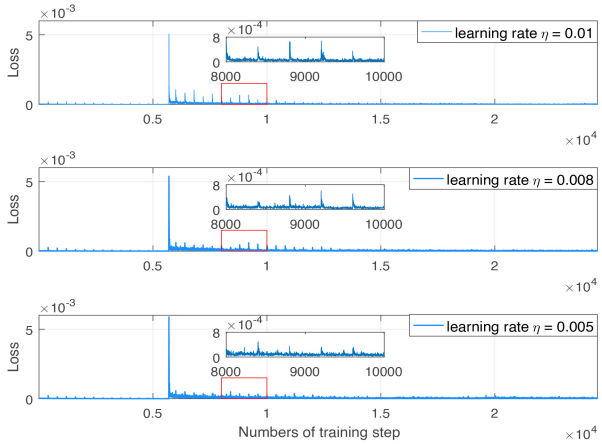


Fig. 10. Loss comparison with different learning rate of DDPG algorithm.

the RPRS one with a certain probability. When the set $E_{\max}$ for the clusters becomes higher, the GPFS approach will behave better comparing with the RPRS algorithm. Moreover, from Figs. 8 and 9, we also know that $B_{\max}$ and $E_{\max}$ have certain ceiling values that their increase cannot obviously make the total data transmission improve even higher. This implies that we do not require to set much higher $B_{\max}$ or $E_{\max}$ for the performance improvement of the system, which is very suitable and practical to real communication applications.

Finally, we discuss the impacts made by two adjusted hyperparameters in the DDPG-based algorithm, namely, the learning rate η and the experience replay buffer capacity X since we consider more about this algorithm. Fig. 10 illustrates the loss of the DDPG-based algorithm with different η. Since η represents the ability of the impact on the parameters of a neural network based on the consideration of output errors or loss, an optimal η varies with the changing environment and requirement. Its value normally decays with training progress. During a training process, the learning rate affects a neural network convergence speed. In this figure, for our problem, we can see that when the selected η varies from 0.01 and 0.008 to 0.005, the algorithm converges at around 12 000 steps, 14 000 steps, and 15 000 steps, respectively. This means a higher η can lead to faster convergence; however, its loss fluctuation is quite clear.

The cumulative reward influenced by the experience replay buffer capacity X of the DDPG-based algorithm is presented in Fig. 11. During the network training, if X reaches its limited value, the oldest memory will be deleted. When X is set to a very high value, the excessive memory will slow down the training process and may lead to unsatisfactory results. As illustrated in this figure, when we adjust X from 4000 to 6000 and 10 000, we can see when X is 6000, the cumulative reward achieves the best performance relatively. This tells us that varying X is not linear with the increase of the cumulative reward. We should carefully choose this parameter for the efficiency of the algorithm.

C. Complexity Analysis

In this section, we only make the complexity analysis concerning floating point operations per second (FLOPS) of

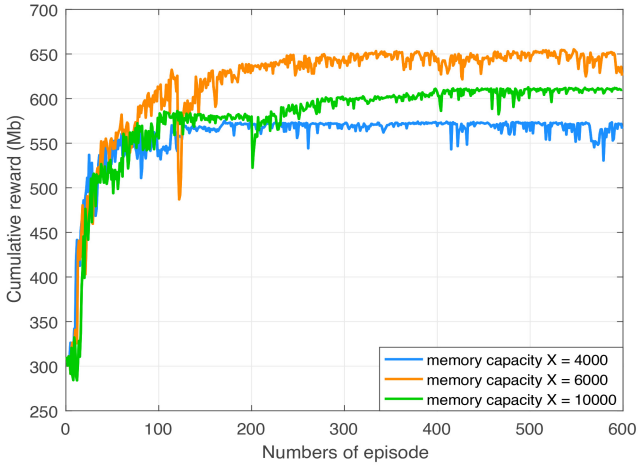


Fig. 11. Reward comparison with different memory capacity of DDPG algorithm.

the proposed DDPG-based algorithm. In this algorithm, the training network consists of two sets of actor networks and two sets of critic networks, which are fully connected. Our input states are the information for wireless communications rather than images or videos. Therefore, the neural network contains no convolution layer. For a dot product of a P vector and an $P \times Q$ matrix, its number of FLOPS computation is $(2P-1)Q$. Furthermore, we also need to consider the computations of activation layers. When calculating the FLOPS, we simply count some certain operations as a single FLOP, for example, an addition, a subtraction, a multiplication, a division, a square root, an exponentiation, and more. We derive the computations of activation layers and find out that the computation is Q with Q inputs for Relu layers, $6 \times Q$ for tanh layers, and $4 \times Q$ for the sigmoid layer [26]. According to the aforementioned information, we can draw a conclusion that the computation complexity of J layers fully connected actor net and K layers fully connected critic net can be calculated by

$$\begin{aligned} & v_{\text{activation}} u_i + 2 \times \sum_{j=0}^{J-1} u_{\text{actor},j} u_{\text{actor},j+1} \\ & +2 \times \sum_{k=0}^{K-1} u_{\text{critic},k} u_{\text{critic},k+1} \\ & = O\left(\sum_{j=0}^{J-1} u_{\text{actor},j} u_{\text{actor},j+1} + \sum_{k=0}^{K-1} u_{\text{critic},k} u_{\text{critic},k+1} \right) \end{aligned}$$

where u_i denotes the unit number in the ith layer, u_0 denotes the input size, and $v_{\text{activation}}$ denotes the corresponding parameters referring to the type of the activation layer. For our problem, $J = 2$ and $K = 2$. Moreover, the DQN algorithm is almost with the same order of computation complexity, since the principle calculations are in the used neural networks.

V. Conclusion

In this article, we designed a novel opportunistic UAVs-aided wireless sensor network to maximize its data transmission performance without the increase of extra flights specially for communication. In order to cope with this optimization problem under certain constraints, we proposed a DQN-based approach and a DDPG-based approach to realize the maximization of long-term data transmission by real-time UAV scheduling, bandwidth allocation, and power control schemes. After the comparison with other four benchmark approaches, we found that our proposed approaches perform better in data transmission performance with the consideration of the convergence speed, loss, harvested energy, and battery capacity. When we discussed two hyperparameters η and X for learning rate and memory capacity, respectively, in the DDPG-based approach, we also obtained some useful and practical conclusions for the use of our proposed algorithm. In addition, we found that the DQN-based algorithm is much more suitable to the relevantly small scale optimization problems or the problems defined on discrete action space while the DDPG-based counterpart is more resilient to these problems.

References

[1] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[2] P. Zhan, K. Yu, and A. L. Swindlehurst, "Wireless relay communications with unmanned aerial vehicles: Performance and optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 3, pp. 2068–2085, Jul. 2011.

[3] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2049–2063, Mar. 2018.

[4] D. Liu *et al.*, "Opportunistic UAV utilization in wireless networks: Motivations, applications, and challenges," *IEEE Commun. Mag.*, vol. 58, no. 5, pp. 62–68, May 2020.

[5] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.

[6] C. You and R. Zhang, "3D trajectory optimization in Rician fading for UAV-enabled data harvesting," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3192–3207, Jun. 2019.

[7] Q. Wu and R. Zhang, "Common throughput maximization in UAV-enabled OFDMA systems with delay consideration,"*IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6614–6627, Dec. 2018.

[8] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, "Joint UAV hovering altitude and power control for space-air-ground IoT networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1741–1753, Apr. 2019.

[9] R. Duan, J. Wang, C. Jiang, H. Yao, Y. Ren, and Y. Qian, "Resource allocation for multi-UAV aided IoT NOMA uplink transmission systems,"*IEEE Internet Things J.*, vol. 6, no. 4, pp. 7025–7037, Aug. 2019.

[10] S. Fu *et al.*, "Energy-efficient UAV enabled data collection via wireless charging: A reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 10209–10219, Jun. 2021.

[11] R. A. Nazib and S. Moh, "Energy-efficient and fast data collection in UAV-aided wireless sensor networks for hilly terrains," *IEEE Access*, vol. 9, pp. 23168–23190, 2021.

[12] Y. Wang, Z. Hu, X. Wen, Z. Lu, and J. Miao, "Minimizing data collection time with collaborative UAVs in wireless sensor networks," *IEEE Access*, vol. 8, pp. 98659–98669, 2020.

[13] Z. Wang, G. Zhang, Q. Wang, K. Wang, and K. Yang, "Completion time minimization in wireless-powered UAV-assisted data collection system," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 1954–1958, Jun. 2021.

[14] J. Li *et al.*, "Joint optimization on trajectory, altitude, velocity, and link scheduling for minimum mission time in UAV-aided data collection," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1464–1475, Feb. 2020.

[15] M. Samir, C. Assi, S. Sharafeddine, D. Ebrahimi, and A. Ghrayeb, "Age of information aware trajectory planning of UAVs in intelligent transportation systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12382–12395, Nov. 2020.

[16] J. Liu, P. Tong, X. Wang, B. Bai, and H. Dai, "UAV-aided data collection for information freshness in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2368–2382, Apr. 2021.

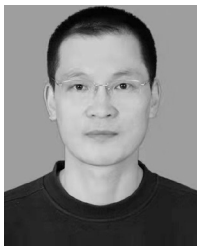
[17] Y. Pan, Y. Yang, and W. Li, "A deep learning trained by genetic algorithm to improve the efficiency of path planning for data collection with multi-UAV," *IEEE Access*, vol. 9, pp. 7994–8005, 2021.
[18] C. Zhan and Y. Zeng, "Aerial-ground cost tradeoff for multi-UAV-enabled data collection in wireless sensor networks,"*IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1937–1950, Mar. 2020.
[19] J. Wang, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones,"*IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, Sep. 2017.
[20] D. Liu *et al.*, "Opportunistic utilization of dynamic multi-UAV in device-to-device communication networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 1069–1083, Sep. 2020.
[21] X. Zhong, Y. Guo, N. Li, and S. Li, "Joint relay assignment and channel allocation for opportunistic UAVs-aided dynamic networks: A mood-driven approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15019–15034, Dec. 2020.
[22] D. Liu, J. Wang, Y. Xu, Y. Xu, Y. Yang, and Q. Wu, "Opportunistic mobility utilization in flying ad-hoc networks: A dynamic matching approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 728–731, Apr. 2019.
[23] D. Liu *et al.*, "Opportunistic data collection in cognitive wireless sensor networks: Air-ground collaborative online planning," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8837–8851, Sep. 2020.
[24] J. Ren, J. Hu, D. Zhang, H. Guo, Y. Zhang, and X. Shen, "RF energy harvesting and transfer in cognitive radio sensor networks: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 104–110, Jan. 2018.
[25] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
[26] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.

Yitong Liu (Student Member, IEEE) received the bachelor's degree in electronic information science and technology from Shandong University, Weihai, China, in 2019. She is currently pursuing the master's degree in information and communication engineering with Jilin University, Changchun, China.

Her research interests include UAV communication, dynamic resource allocation, and wireless communication.

Junjie Yan (Student Member, IEEE) received the bachelor's degree in communication engineering from the College of Communication Engineering, Jilin University, Changchun, China, in 2017, where she is currently pursuing the Ph.D. degree.

Her research interests mainly include intelligent edge computing, UAV communications, and artificial intelligence.



Xiaohui Zhao received the Ph.D. degree in applied mathematics and control theory from the Université de Technologie de Compiègne, Compiègne, France, in 1993.

He has been a Senior Visiting Scholar for half a year to the Laboratoire d'Informatique, Université de Pierre et Marie Curie, Paris, France, in 2006. He is currently a Professor of communication engineering with Jilin University, Changchun, China. His research interests include wireless communication, cognitive radio, and adaptive signal processing.