IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 24, 2020, accepted March 13, 2020, date of publication March 17, 2020, date of current version March 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981403

Energy Efficient 3-D UAV Control for Persistent Communication Service and Fairness: A Deep Reinforcement Learning Approach

HANG QI[1], ZHIQUN HU[2], HAO HUANG[1], XIANGMING WEN[1], AND ZHAOMING LU[1]

[1]Beijing Key Laboratory of Network System Architecture and Convergence, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

[2]School of Computing and Information Engineering, Hubei University, Wuhan 430062, China

Corresponding author: Zhiqun Hu (zhiqunhu520@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61901163, in part by the Beijing Natural Science Foundation under Grant L192034, and in part by the Beijing Laboratory of Advanced Information Network.

ABSTRACT Recently, unmanned aerial vehicles (UAVs) as flying wireless communication platform have attracted much attention. Benefiting from the mobility, UAV aerial base stations can be deployed quickly and flexibly, and can effectively establish Line-of-Sight communication links. However, there are many challenges in UAV communication system. The first challenge is energy constraint, where the UAV battery lifetime is in the order of fraction of an hour. The second challenge is that the coverage area of UAV aerial base station is limited and the commercial UAV is usually expensive. Thus, covering a large target region all the time with sufficient UAVs is quite challenging. To solve above challenges, in this paper, we propose energy efficient and fair 3-D UAV scheduling with energy replenishment, where UAVs move around to serve users and recharge timely to replenish energy. Inspired by the success of deep reinforcement learning, we propose a UAV Control policy based on Deep Deterministic Policy Gradient (UC-DDPG) to address the combination problem of 3-D mobility of multiple UAVs and energy replenishment scheduling, which ensures energy efficient and fair coverage of each user in a large region and maintains the persistent service. Simulation results reveal that UC-DDPG shows a good convergence and outperforms other scheduling algorithms in terms of data volume, energy efficiency and fairness.

INDEX TERMS UAV communication, energy efficiency, fairness, energy replenishment, deep reinforcement learning, DDPG.

I. INTRODUCTION

Unmanned aerial vehicle (UAV) as flying wireless communication platform is a promising technology to enhance the wireless network with its inherent attributes such as mobility, flexibility and adaptive altitude [1]. For example, UAVs can act as mobile aeriel base stations (BSs) to provide on-the-fly communication, which can significantly improve the coverage of ground wireless devices and boost the capacity of wireless networks. Compared to the terrestrial BS, the advantage of using UAV-based aerial BS is that it can fast deploy communication infrastructure to provide cost-effective connectivity when communication networks are disrupted by a natural disaster [2] or areas are poorly covered by terrestrial networks [3]. In addition, UAV-based aerial BSs can effectively establish Line-of-Sight (LoS) communication links by adjusting their location and are likely to have better communication channels than terrestrial networks.

Although UAV-based aerial BS has huge advantages, UAV communication system still faces many challenges. The first challenge is that battery operated UAVs [5] usually have limited on-board energy due to the aircraft's size and weight constraint. The energy constraint has a large impact on the endurance and performance of UAV systems, where the battery lifetime is in the order of fraction of an hour (typically 15-30 minutes) for most commercial consumer-grade UAVs [4], [5]. Thus, energy efficient based communication strategies have been studied to prolong the UAV network service lifetime under a broad range of aspects [6]–[15].

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenhui Yuan.

The authors in [6]–[8] optimized the UAV placement and trajectory of single UAV to maximize the coverage and throughput by using the minimum transmission power. The works in [9]–[12] further studied the optimal 3-D locations of multiple UAVs to maximize the downlink coverage performance with minimum transmission power. In [13], the authors determined the optimal 3-D mobile trajectory of multiple UAVs with a minimum energy consumption to adapt to the time-varying nature of user density. Energy-aware control protocols [14] minimize the unnecessary maneuvers of mobile devices and energy-aware network layer protocols are extensively surveyed in [5] to reduce battery consumption or conserve power. However, above works do not consider the energy replenishment of UAVs, which is inevitable in practice. Up to now, the works on energy replenishment through scheduling of UAVs are quite limited. In [15], the authors presented deployment strategies for multiple UAVs to maximize the stationary coverage of a target region to guarantee the continuity of the service by energy replenishment operations at ground charging stations. However, this work just addressed stationary coverage where the UAVs are at fixed positions.

As we know, UAVs can work as BSs to provide wireless communication for ground users by carrying current wireless technologies, such as LTE or Wi-Fi. On the other hand, the cost of each commercial UAV is usually several thousand dollars. Thus, covering a large target region all the time with sufficient UAVs is quite challenging due to the limited communication range and relatively high costs, which is the second challenge. Therefore, UAVs moving around to ensure each user to be covered is of paramount importance. In [16], the authors proposed a deep reinforcement learning (DRL) framework to control the mobile trajectory of a group of UAVs to achieve fairness coverage while maintaining their connectivity with minimum energy consumption. However, this work assumes all UAVs have the same altitude. Besides, it does not consider on-board circuit power, communication power, 3-D deployment of aerial BSs and energy replenishment.

Different from the aforementioned existing works under the assumption of either 2-D or stationary UAV coverage, inspired by the success of DRL, we propose a UAV Control policy based on Deep Deterministic Policy Gradient (DDPG) algorithm [17] (UC-DDPG) to address the combination problem of 3-D mobility of multiple UAVs and energy replenishment scheduling, which ensures energy efficient and fair coverage of each ground user in a large target region, while maintaining the persistent service.

The main contributions of this paper are listed as follows:

- Detailed UAVs communication system models are built, including channel model, data rate model and energy model.
- In contrast with other papers, this paper takes UAV battery lifetime and energy replenishment into account, so as to maintain persistent service.
- In order to improve energy efficiency and guarantee service fairness, we develop a 3-D UAV deployment scheduling algorithm based on DDPG algorithm, which takes the residual energy of UAV, circuit power, communication power, mobility power and hover power into account.

The rest of paper is organized as follows. Section II presents the related works. Section III presents system models and problem definition. The preliminaries of DDPG is introduced in Section IV. The proposed UC-DDPG is detailedly introduced in Section V. In Section VI, the convergence and performance of the proposed algorithm are verified by numerical results. Finally, Section VII concludes the paper.

II. RELATED WORK

A. UAV 3-D DEPLOYMENT AND COVERAGE

Recently, most studies have been carried out to optimize the deployment of UAV BSs, aiming at coverage range, number of active UAVs, transmit power. In [6], the authors optimized the UAV position to satisfy the rate requirement of users in the entire high-rise building with minimum total transmit power. In [18] and [19], the hovering altitude of the UAV can be determined to maximize the radio coverage on the ground. An optimum placement of multiple UAVs is further investigated in [20] to maximize the number of covered users in the target region. Similarly, the studies in [9] and [21] investigate the optimal 3-D placement of UAVs to maximize the coverage while minimizing the transmitting power of the UAVs. In [22] and [23], the authors minimize the number of UAVs that must be deployed for covering all the ground terminals in the target region. The works in [10] proposed a framework to achieve energy-efficient uplink data collection from ground IoT devices by jointly optimizing the 3-D placement, device-UAV association and uplink power control in single time slot. Then, the works optimize the UAV's mobile trajectory by allowing UAVs to dynamically update their locations depending on the time-varying device's activation process, where the total energy consumption of the UAVs while updating their location is minimized. In [13], the optimal placement of multiple UAVs, such as altitude and coverage radius, is derived in a single time slot when the transmit power of UAV equals to their on-board circuit power. Then, the optimal placement updating problem in multiple time slot duration is also addressed to achieve near minimal energy consumption in polynomial time.

B. UAV ENERGY MANAGEMENT

The problem of prolonging the UAV working time has been extensively studied in the existing literature, such as energy-based protocol [7], [8], [14], [24], [25] to reduce the UAV's energy consumption, and replenishment strategies [15], [26]–[29] by leveraging the presence of charging infrastructures on the ground. The authors in [7] studied the optimization of the throughput of a relay-based UAV system by jointly controlling the UAVs trajectory as well as the source/relay transmit power. Later, the authors extend the

work in [7] to optimize the energy efficiency of the relay-based UAV system by optimizing the UAV's trajectory [8]. In [14], the authors illustrate that energy efficiency based protocol minimizes the unnecessary maneuvers, which can be implemented via carefully controlling the movement of UAVs and optimizing the communication strategies with the minimum energy expenditure. In [24], the authors propose the use of passive scanning for the mobiles and periodic beaconing for UAVs as access points, where a cooperative game theory is used to provide effective coverage for mobile users. The authors in [25] design an energy efficient traveling path algorithm considering the peculiar feature of UAVs, such as the available energy, weight, maximum speed, *etc*.

The authors in [26] and [27] study the continuous coverage problem for mobile targets, where [26] properly allocates the charging slots for replenishing energy while [27] replaces the UAV that runs out of energy by a new one during the coverage process. Shakhatreh *et al.* [28] improve the model in [26] to the scenario with multiple UAVs. Considering the on-board circuit power and mobility power, the UAV control for scheduling fly or recharge has been investigated to guarantee persistent coverage of a target area by exploiting characteristics of fixed terrestrial charging infrastructures on the ground [15], [29]. The works in [30], [31] describe the design of reliable charging station for UAV. In [30], the ground charging station is designed to achieve the reliable recharging process, while a guidance system enabling the UAV to land on a charging station is described in [31].

The study for extracting energy from the environmental forces has also been applied in UAVs system. In [32], [33] and [34], the authors plan a path for UAV to extend the flight duration by exploiting the wind energy, while [34] considers the uncertainty of the wind field and the variation with respect to time. The authors in [35], [36] study the wireless power transfer techniques enabled by radio frequency signals to charge the UAVs. Similarly, the laser energy harvesting system is expected to efficiently prolong the UAV's flight duration, where a laser transmitter sends laser beams to charge a fixed-wing UAV in flight [37]–[39]. In [40], the authors present a rotational energy harvester using a brushless Direct-Current (DC) generator to harvest ambient energy from the propellers of the UAVs in order to prolong the UAV's flight duration.

C. DEEP REINFORCEMENT LEARNING IN WIRELESS NETWORK

DRL has recently attracted much attention from wireless communication field and is used to solve various problems [49]. In [50], an artificial intelligence framework (AIF) for smart wireless network management was proposed. DCRQN [51] which is a novel Wi-Fi handoff management scheme based on Deep Q-Network (DQN) [52] effectively improves the data rate during the handoff process. In [53], the authors presented DeepNap, which uses a DQN to learn effective BS sleeping policies and reduces the energy consumption of Wi-Fi networks. In [54], a novel channel allocation algorithm based on DRL was presented, which improves spectrum efficiency and decreases the co-channel interference for multi-beam satellite systems. The authors in [55] proposed to use DRL to obtain the optimal interference alignment (IA) user selection policy in the cache-enabled opportunistic IA networks. In [56], a DRL approach was proposed to maximize the channel utility for multi-user wireless networks with less computation and limited observations. In UAV communication networks, a DRL framework for multi-user access control is proposed in [57], which effectively improves system throughput. Recently, a Q-learning [59] based framework [58] is proposed for quality of experience (QoE) driven deployment and movement of UAV-BSs, and shows good performance and low complexity.

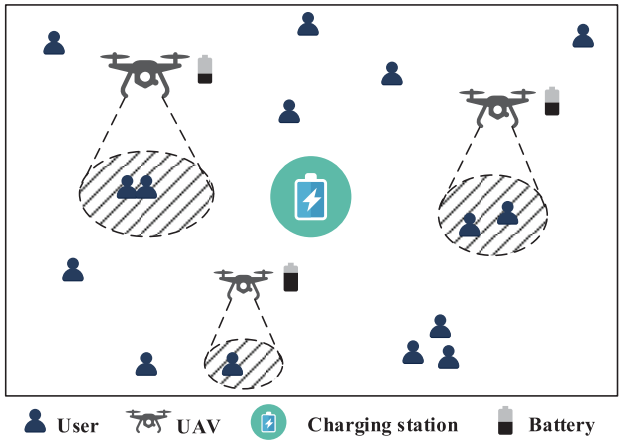


FIGURE 1. The scenario of UAV communication system.

III. SYSTEM MODELS AND PROBLEM DEFINITION

In this section, we introduce the system models of the paper in the first place, including the scenario, channel model, data rate model as well as energy model. Then, the problem definition of energy efficient 3-D UAV control for persistent communication service and fairness is presented.

A. SCENARIO

We consider a rectangular geographical area of size $a \times b\, m^2$, as shown in Figure 1, within which a set $\mathcal{K} = \{1, 2, \ldots, K\}$ of K ground users are distributed. In this system, a set $\mathcal{A} = \{1, 2, \ldots, N\}$ of N rotary wing UAVs are deployed to provide communication coverage to the ground users in the target area. Because the number of UAVs is limited, the users cannot be completely covered by hovering UAVs. As a result, the UAVs should move around to provide service for all users. The total service time is T. The locations of user $k \in \mathcal{K}$ and UAV $i \in \mathcal{A}$ are, respectively, given by (x_k, y_k) and (x_i, y_i, h_i), where $x_i, x_k \in [0, a]$ and $y_i, y_k \in [0, b]$. The height of the UAV, h_i, belongs to $[h_{min}, h_{max}]$, where h_{min} and h_{max} are the minimum and maximum allowed height of UAV, respectively. A charging station S_E locates at the center of the plane where the altitude is h_{min}, and it can be used to recharge the UAV's

battery at a speed of C_{S_E} Watt. We assume all the UAVs start with the fully charged batteries, and the battery capacity is E_{max}. Assume ϖ is the angle of the sensing cone. With h_i, the radius of the cover area can be given by [15], [18],

$$R(h_i) = h_i \cdot \tan(\frac{\varpi}{2}). \tag{1}$$

For simplicity, T is divided into consecutive time slots $\{t_0, t_1, \ldots, t_{end}\}$ of length equal to t_{slot}, and there is a control center which can collect information from UAVs and command UAVs. In t_j, UAV i can fly, hover, serve users and replenish energy, which is determined by the commands of control center. The users which are in the coverage of UAV i can be served by UAV i simultaneously. The users are assigned to different channel where the channel bandwidth is B, and there is no interference between them. If a user is covered by multiple UAVs, it will connect to the first UAV which provides communication service. The residual energy of UAV i at the beginning of time slot t_j is denoted by E_{i,t_j}. If a UAV replenishes energy in charging station S_E, it will not serve users.

B. CHANNEL MODEL

1) AIR-TO-GROUND PATH LOSS MODEL

According to [18], the air-to-ground (A2G) path loss model can be characterized into LoS links and non Line-of-Sight (NLoS) links, which can be given respectively by

$$\begin{cases} L_{LoS}^{ik} = 20\log(\frac{4\pi f_c d_{ik}}{c}) + \eta_{LoS}, \\ L_{NLoS}^{ik} = 20\log(\frac{4\pi f_c d_{ik}}{c}) + \eta_{NLoS} \end{cases} \tag{2}$$

where f_c is the carrier frequency; d_{ik} is the distance between the UAV i and the user k, given by $d_{ik} = \sqrt{(x_i - x_k)^2 + (y_i - y_k)^2 + {h_i}^2}$; c is the speed of light; η_{LoS} and η_{NLoS} are the mean value of the excessive path loss on the top of the free space for LoS and NLoS links, determined by environment (suburban, urban, dense urban, highrise urban or others).

For A2G communications, each transmitter-receiver pair will typically have a LoS link with a given probability, which depends on the environment, location of the users and the UAV as well as the elevation angle [18]. Therefore, we have the LoS probability [18], [20]

$$P_{LoS}^{ik}(\vartheta_{ik}) = \frac{1}{1 + \psi \exp(-\zeta(\vartheta_{ik} - \psi))} \tag{3}$$

where ψ and ζ are constant values which depend on the environment and $\vartheta_{ik} = \frac{180}{\pi}\arcsin(\frac{h_i}{d_{ik}})$ is the elevation angle. Note that, the NLoS probability is $P_{NLoS}^{ik}(\vartheta_{ik}) = 1 - P_{LoS}^{ik}(\vartheta_{ik})$.

Therefore, the average path loss of A2G channel can be expressed as

$$L^{ik}(h_i, \vartheta_{ik}) = 20\log(\frac{4\pi f_c d_{ik}}{c}) + P_{LoS}^{ik}(\vartheta_{ik})\eta_{LoS} + P_{NLoS}^{ik}(\vartheta_{ik})\eta_{NLoS}. \tag{4}$$

2) SMALL-SCALE FADING CHANNEL MODEL

We consider the small-scale channel fading following Rician distribution [42]. A Rician distribution is an adequate choice due to the possible combination of LoS and multiple scatters that can be experienced at the receiver. The complex channel gain between the pair of the UAV and its user is denoted by g_{ik}. Then the probability distribution function (PDF) of g_{ik} can be expressed as [41]

$$f_{g_{ik}}(x) = \frac{2(1 + K(\vartheta_{ik}))x}{\Omega} e^{(-K(\vartheta_{ik}) - \frac{(1+K(\vartheta_{ik})x^2)}{\Omega})} \times I_0(2x\sqrt{\frac{K(\vartheta_{ik})(1 + K(\vartheta_{ik}))}{\Omega}}), \quad x \geqslant 0 \tag{5}$$

where g_{ik} is the complex channel gain between the pair of the UAV and its user; $I_0(\cdot)$ is the zero-order modified Bessel function of the first kind, which can be defined by $I_0(x) = \sum_{n=0}^{\infty} \frac{(x/2)^{2n}}{n!\Gamma(n+1)}$; $K(\vartheta_{ik})$ is the Rician factor defined as the ratio of the power in the LoS component to the power in the NLoS multiple scatters. Ω is the average fading power where $\Omega = 1$.

Based on [43], the Rician factor can be modeled as a non-increasing function of ϑ_{ik}, which can be expressed as

$$K(\vartheta_{ik}) = \kappa_0 \cdot \exp[\frac{2}{\pi}\ln(\frac{\kappa_{\frac{\pi}{2}}}{\kappa_0})\vartheta_{ik}] \tag{6}$$

where $\kappa_0 = K(0)$ and $\kappa_{\frac{\pi}{2}} = K(\frac{\pi}{2})$.

Then, the expectation of K can be estimated according to [44],

$$\begin{aligned} \overline{K} &= E[K(\vartheta)] \approx \int_0^{\frac{\pi}{2}} \vartheta \cdot \kappa_0 \cdot \exp[\frac{2}{\pi}\ln(\frac{\kappa_{\frac{\pi}{2}}}{\kappa_0})\vartheta]d\vartheta \\ &= \frac{\pi^2 \kappa_{\frac{\pi}{2}}}{4\ln(\frac{\kappa_{\frac{\pi}{2}}}{\kappa_0})}[1 - \frac{1}{\ln(\frac{\kappa_{\frac{\pi}{2}}}{\kappa_0})} + \frac{\kappa_0}{\kappa_{\frac{\pi}{2}}\ln(\frac{\kappa_{\frac{\pi}{2}}}{\kappa_0})}]. \end{aligned} \tag{7}$$

C. DATE RATE MODEL

Assume the allocated transmit power to the interested user k is P_t. The instantaneous signal-to-noise ratio (SNR) between the UAV i and the user k can be modeled as

$$\gamma_{ik}^{SNR} = \frac{10^{\log_{10} P_t - L^{ik}(h_i, \vartheta_{ik})/10}}{\sigma^2} g_{ik} \tag{8}$$

where σ^2 is the noise power.

Then, the PDF of γ_{ik}^{SNR} is obtained by introducing a change of variables in the expression for the PDF $f_{g_{ik}}(x)$, yielding

$$f_{\gamma_{ik}^{SNR}}(\gamma) = \frac{f_{g_{ik}}(\sqrt{\Omega\gamma/\bar{\gamma}_{ik}})}{2\sqrt{\gamma\bar{\gamma}_{ik}/\Omega}}$$

where $\overline{\gamma}_{ik} = \frac{10^{\log_{10} P_t - L^{ik}(h_i, \vartheta_{ik})/10}}{\sigma^2}$ is the average SNR. Therefore, the PDF of γ_{ik}^{SNR} can be given by [41]

$$f_{\gamma_{ik}^{SNR}}(\gamma) = \frac{(1 + K(\vartheta_{ik}))}{\overline{\gamma}_{ik}} e^{(-K(\vartheta_{ik}) - \frac{(1+K(\vartheta_{ik}))\gamma}{\overline{\gamma}_{ik}})} \times I_0(2\sqrt{\frac{K(\vartheta_{ik})(1 + K(\vartheta_{ik}))\gamma}{\overline{\gamma}_{ik}}}), \quad \gamma \geqslant 0. \tag{9}$$

Based on the Shannon-Hartley theorem, the average data rate between the UAV-BS i and the user k can be given by

$$R_{ik} = \int_0^\infty B\log_2(1+\gamma) f_{\gamma_{ik}^{SNR}}(\gamma) d\gamma$$

which can be approximated to [45]

$$R_{ik} = (B\log_2 e)(\ln(1+\overline{\gamma}_{ik}) - \frac{E[\gamma^2] - \overline{\gamma}_{ik}^2}{2(1+\overline{\gamma}_{ik})^2}). \quad (10)$$

To evaluate (10), the second moment of γ is required, which can be formulated as [45]

$$E[\gamma^2] = \int_0^\infty \gamma^2 f_{\gamma_{ik}^{SNR}}(\gamma) d\gamma. \quad (11)$$

Then, combining (9) and (10), a second-order approximation for R_{ik} can be attained as

$$R_{ik} \approx \frac{B}{\ln 2}[\ln(1+\overline{\gamma}_{ik}) - \frac{\overline{\gamma}_{ik}^2 \begin{pmatrix} 2e^{-K(\vartheta_{ik})}(3K(\vartheta_{ik})^2 + 3K(\vartheta_{ik}) + 1) \\ -(1+K(\vartheta_{ik}))^2 \end{pmatrix}}{2(1+\overline{\gamma}_{ik})^2(1+K(\vartheta_{ik}))^2}]. \quad (12)$$

D. ENERGY MODEL

The total energy consumption of the UAV network includes communication energy and propulsion energy. The communication energy is needed due to the radiation, signal processing and other circuitry while the propulsion energy is required to ensure that the UAV remains aloft as well as for supporting its mobility. However, the propulsion energy is different according to the UAV's flying state. In this subsection, the energy models including communication energy, hover energy as well as mobility energy are illustrated.

1) COMMUNICATION ENERGY

Assume the on-board circuit power is set to be P_{cu}. Since the UAVs fly on the target areas to serve users, the corresponding communication time for the UAV i to the users is depended on the control policy. Let t_{com} denote the duration that UAV i communicates with the users and n_{i,t_j} denote the number of the served users by UAV i in t_j. Then, at t_j, the communication energy of UAV i, E_i^C, can be given by

$$E_i^C(t_j) = (n_{i,t_j}P_t + P_{cu})t_{com}. \quad (13)$$

2) HOVER ENERGY

According to the [47], [48], the hover energy consumption of UAV can be derived using power consumption of a multirotor helicopter, which is approximately linearly proportional to the weight of its battery and payload. Then, the hover power in Watt by the UAV can be given by [48]

$$P_{hover} = \frac{MG^{\frac{3}{2}}}{\sqrt{2\rho\pi\beta^2}} \quad (14)$$

where M is the number of rotors of the helicopter; $G = (W+m)g$ is the thrust in Newton, given the frame weight W in kg, the battery and payload weight m in kg, and gravity g in N/kg. ρ is the fluid density of the air in kg/m^3, and β is the rotor disk radius in m.

Therefore, the energy consumed of UAV i in hover at t_j can be computed as

$$E_i^H(t_j) = P_{hover}t_{hover} \quad (15)$$

where t_{hover} denotes the duration that UAV i hovers in t_j.

3) MOBILITY ENERGY

Let P_h, P_a and P_d denote the mobility power in the horizontal direction, ascending power and descending power, respectively. Similarly, v_h, v_a and v_d represent the velocity in the horizontal direction, ascending velocity and descending velocity, respectively. Assume the UAV i updates its location in the considered area at t_j. Then, the mobility energy of UAV i at t_j can be given by [13]

$$E_i^M(t_j) = P_h\frac{d(i,t_j)}{v_h} + I(\Delta h(i,t_j))P_a\frac{\Delta h(i,t_j)}{v_a} - (1-I(\Delta h(i,t_j)))P_d\frac{\Delta h(i,t_j)}{v_d} \quad (16)$$

where $d(i,t_j)$ and $\Delta h(i,t_j)$ are the horizontal moving distance and the variation of the height of the UAV i at t_j, respectively. Then, the effective horizontal and vertical (ascending or descending) velocities will be $v_h = \upsilon\sin\varphi$ and $v_a = v_d = \upsilon\cos\varphi$ with $\varphi = \arctan(\frac{d(i,t_j)}{\Delta h(i,t_j)})$, where υ denotes the velocity of the UAV. $I(\Delta h(i,t_j))$ is the indicator function, which can be expressed as

$$I(\Delta h(i,t_j)) = \begin{cases} 1 & \Delta h(i,t_j) \geqslant 0, \\ 0 & \Delta h(i,t_j) < 0. \end{cases} \quad (17)$$

The power consumption of the horizontal direction can be given by [8], [10]

$$P_h = P_p + P_I \quad (18)$$

where P_p is the parasitic power for overcoming the parasitic drag due to the aircraft's skin friction, form drag, *etc*, which can be given by [10], [46]

$$P_p = \frac{1}{2}\rho C_{D_0} S {v_h}^3 + \frac{\pi}{4}M\rho c_b C_{D_0} w^3\beta^4(1+3(\frac{v_h}{w\beta})^2) \quad (19)$$

where C_{D_0} is the drag coefficient, c_b is the rotor chord, S is the reference area (frontal area of the UAV), w is the angular velocity.

And P_I is the induced power for overcoming the lift-induced drag due to the wings redirecting air to generate the lift for compensating the aircraft's weight. According to [8], [10], P_I can be given by

$$P_I = G\sqrt{\frac{\lambda - {v_h}^2}{2}} \quad (20)$$

where $\lambda = \sqrt{{v_h}^4 + (\frac{G}{\pi\rho\beta^2})^2}$.

Similarly, P_a and P_d can be given by [46]

$$P_a = \frac{G}{2}v_a + \frac{G}{2}\sqrt{{v_a}^2 + \frac{2G}{\pi\rho\beta^2}}$$
$$\text{and} P_d = \frac{G}{2}v_d - \frac{G}{2}\sqrt{{v_d}^2 - \frac{2G}{\pi\rho\beta^2}} \quad (21)$$

respectively.

E. PROBLEM DEFINITION

We purpose to design a control algorithm which commands how each UAV acts in each time slot. The targets of the control algorithm include: 1) providing persistent service in T; 2) maximizing communication data volume; 3) minimizing the UAVs energy consumption; 4) guaranteeing the fairness of the users. For the first target, because the battery lifetime of UAVs is much less than T, a replenishment policy should be designed to guarantee persistent service. In order to achieve the second objective, intuitively, appropriate communication locations where there is a good channel environment and the users are covered as many as possible should be found. For the third target, the flight path of UAVs should be designed carefully to reduce needless energy consumption, *e.g.*, the UAV moves to some places without any user. Lastly, for the sake of guaranteeing the fairness of the users, the UAVs should serve all users as evenly as possible, rather than only serve part of users. In summary, it is a quite sophisticated task and traditional optimization algorithms are unsuitable. Recently, DRL has received extensive attention in the field of wireless communication [49]. DRL can learn the best policy by real-time interacting with the environment and only very minimal prior knowledge is needed, which applies to designing the UAV control algorithm. DDPG, which is the state-of-the-art DRL algorithm, shows good performance in solving complex tasks [17]. In the following sections, we will detailedly present the control algorithm based on DDPG.

IV. PRELIMINARIES ON DDPG

This section gives a brief description of reinforcement learning (RL) and DDPG. For a comprehensive presentation, please refer to [59] and [17].

Figure 2 shows the basic form of RL. RL is learning how to map state to action, so as to maximize a numerical reward. The learner, *i.e.* the agent, is not told which actions to take, but instead must discover which actions yield the most reward by trying them. The agent observes the state s of the environment, and executes an action a according to the policy. Then, the agent receives a reward r and observes a new state s'. The above process is repeated until the end of the agent-environment interaction, and a complete interaction process is referred to as an episode. This information, (s, a, r, s'), is used to improve the agent's policy, and the episode will be repeated until the policy converges to the optimal policy. Q-Learning and SARSA [59] are the most common algorithms in RL.

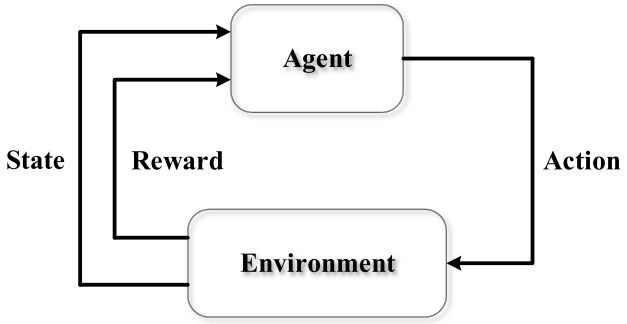


FIGURE 2. The basic component and form of reinforcement learning.

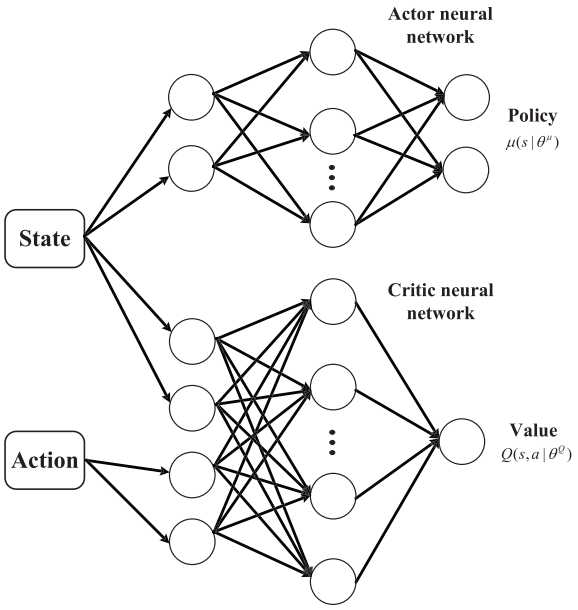


FIGURE 3. The basic structure of DDPG.

However, RL is unsuitable and inapplicable to the complex tasks which have continuous and high dimensional state spaces or action spaces. DRL embraces the advantage of deep neural network (DNN) to train learning process, thereby improving the learning speed and the performance of RL algorithms. DDPG [17], which is a model-free off-policy actor-critic DRL algorithm, can learn policies in continuous, high dimensional state spaces and action spaces. As shown in Figure 3, the DDPG algorithm maintains a parameterized actor neural network $\mu(s \mid \theta^\mu)$ which specifies the current policy by deterministically mapping states to a specific action. The parameterized critic neural network $Q(s, a \mid \theta^Q)$ is learned using the Bellman equation as in Q-learning. The actor is updated by applying the chain rule to the expected return from the start distribution J with respect to the actor parameters as follows:

$$\begin{aligned}\nabla_{\theta^\mu} J &\approx \mathbb{E}[\nabla_{\theta^\mu} Q(s, \mu(s \mid \theta^\mu) \mid \theta^Q)] \\ &= \mathbb{E}[\nabla_a Q(s, \mu(s) \mid \theta^Q) \nabla_{\theta^\mu} \mu(s \mid \theta^\mu)]. \end{aligned} \quad (22)$$

Specially, experience replay and target network are introduced in DDPG to guarantee the convergence, which is

inspired by DQN [52]. In experience replay, a replay buffer with a finite size is used to store the sample (s, a, r, s'). When the replay buffer was full, the oldest samples were discarded. The actor and critic are updated by sampling a minibatch randomly from the replay buffer. Experience replay breaks the correlations between samples and therefore reduces the variance of learning. In target network, a copy of the actor and critic networks, $Q'(s, a \mid \theta^{Q'})$ and $\mu'(s \mid \theta^{\mu'})$, is created. $Q'(s, a \mid \theta^{Q'})$ and $\mu'(s \mid \theta^{\mu'})$ are used to calculate the target values, and their weights are then updated by having them slowly track the original networks: $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$ with $\tau \ll 1$, which greatly improving the stability of learning.

V. UAV CONTROL BASED ON DDPG

In the section, we design a UAV control policy based on DDPG. In this problem, the agent is the control center, and UC-DDPG is implemented in the control center. The basic three elements (state, action and reward) of RL are designed as follow.

A. STATE

In time slot t_j, state s_j is defined as

$$\begin{aligned} s_j = \{&x_1, y_1, h_1, x_2, y_2, h_2, \ldots, x_N, y_N, h_N, \\ &E_{1,t_j}, E_{2,t_j}, \ldots, E_{N,t_j}, \\ &data_{1,t_j}, data_{2,t_j}, \ldots, data_{K,t_j}\}, \end{aligned} \tag{23}$$

where (x_i, y_i, h_i) denotes the location of UAV i at the beginning of the time slot t_j and E_{i,t_j} denotes the residual energy of UAV i at the beginning of t_j. $data_{k,t_j}$ denotes accumulative received data volume of user k before t_j.

As shown above, s_j is a vector with the size of $4N + K$ and consisted of three parts. The first part is the locations of all UAVs and the second part is the residual energy of each UAV. The last part is the accumulative received data volume of each user. Specially, all the elements in state s_j are normalized to accelerate the process of learning. In detail, x_i, y_i, h_i and E_{i,t_j} are divided by their corresponding maximum, *i.e.* a, b, h_{max} and E_{max}. $data_{k,t_j}$ is divided by $\sum_k data_{k,t_j}$ which is the total received data volume by all users.

B. ACTION

Obviously, action a_j is

$$\begin{aligned} a_j = \{&\varphi_1, \phi_1, d_1, \varphi_2, \phi_2, d_2, \ldots, \varphi_N, \phi_N, d_N\}, \\ &\varphi_i \in [0, \pi], \\ &\phi_i \in [0, 2\pi), \\ &d_i \in [0, d_{max}]. \end{aligned} \tag{24}$$

In the start of t_j, the UAV flies according to action a_j with a fixed velocity v. φ_i and ϕ_i are the polar angle and the azimuthal angle of the UAV i flight direction, respectively. d_i is the flight distance and d_{max} is the largest allowed flight distance. If a UAV flies off the border, it will stay at the border. After the flight, in the remaining time of t_j, if the UAV is not in charging station, it will hover and provide communication service for covered users. Otherwise, the UAV will charge until the end of t_j.

C. REWARD

The reward r_j is calculated at the end of the time slot t_j and is designed as

$$r_j = \begin{cases} 7^{JFI} \cdot \dfrac{data_{t_j}}{E_{t_j}} + \sum_i I_i & \\ \cdot 10(-\dfrac{5}{6}\bar{E}_{i,t_j}^2 - \dfrac{7}{6}\bar{E}_{i,t_j} + 1) & \forall E_{i,t_{j+1}} > 0, \\ -20 & \exists E_{i,t_{j+1}} = 0. \end{cases} \tag{25}$$

If the residual energy of any UAV at the end of t_j is larger than 0, the first line formula will be used to calculate the reward. Thereinto, E_{t_j} denotes the energy consumed by all UAVs in t_j, where $E_{t_j} = \sum_{i=1}^{N}[E_i^C(t_j) + E_i^H(t_j) + E_i^M(t_j)]$. $data_{t_j}$ represents the data volume received by all users in t_j. I_i is the indicator function, which can be expressed as

$$I_i = \begin{cases} 1 & \text{UAV } i \text{ replenished energy in } t_j, \\ 0 & \text{else.} \end{cases} \tag{26}$$

$\bar{E}_{i,t_j}$ denotes the normalized residual energy of UAV i at the beginning of t_j, and is equal to $\frac{E_{i,t_j}}{E_{max}}$. *JFI* is Jain's Fairness Index, which is used to estimate the fairness. *JFI* is defined by

$$JFI = \frac{(\sum_{k=1}^{K} data_{k,t_{j+1}})^2}{K\sum_{k=1}^{K} data_{k,t_{j+1}}^2}, \tag{27}$$

and $JFI \in [\frac{1}{K}, 1]$. The fairer the service is, the larger *JFI* is. The first part of the first line formula, *i.e.* $7^{JFI} \cdot \frac{data_{t_j}}{E_{t_j}}$, can be interpreted as fairness times energy efficiency. The larger the energy efficiency and the fairness are, the larger the first part is. The second part of the first line formula, *i.e.* $\sum_i I_i \cdot 10(-\frac{5}{6}\bar{E}_{i,t_j}^2 - \frac{7}{6}\bar{E}_{i,t_j} + 1)$, is used to stimulate the agent to learn replenishment policy. It is expected that the UAV replenishes energy when it has low energy rather than high energy. As a result, this part is developed. If $\bar{E}_{i,t_j}$ is less than 60%, the value will be positive. If not, the value will be negative.

Otherwise, the second line formula is used, where there is a UAV without energy in t_j. The agent will receive the "penalization" which is -20 in our implementation.

D. UC-DDPG

The pseudo-code of UC-DDPG is presented in Algorithm 1. In the first place, we randomly initialize actor neural network μ and critic neural network Q, and build target network Q' and μ'. A replay buffer *RB* with a fixed size is also built.

Algorithm 1 UC-DDPG

Initialization:

Randomly initialize actor neural network $\mu(s \mid \theta^{\mu})$ and critic neural network $Q(s, a \mid \theta^{Q})$.

Initialize target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^{Q}$, $\theta^{\mu'} \leftarrow \theta^{\mu}$.

Initialize replay buffer *RB*.

Algorithm:

1: **for** episode in $\{1, 2, 3, \ldots\}$ **do**

2: Initialize a Gaussian noise $\mathcal{N}$ with mean 0 and variance $var \leftarrow 5$.

3: Randomly initialize the positions of all UAVs.

4: Initialize the UAVs with fully charged battery.

5: Receive initial observation state s_0.

6: **for** t_j in $\{t_0, t_1, \ldots t_{end}\}$ **do**

7: Select action $a_j = \mu(s_j \mid \theta^{\mu}) + \mathcal{N}$.

8: $var \leftarrow var \times 0.995$

9: All UAVs execute action according to a_j.

10: UAVs serve users or replenish energy.

11: Calculate reward r_j according to equation (25).

12: Get new state s_{j+1}.

13: Store (s_j, a_j, r_j, s_{j+1}) in *RB*.

14: Sample a random minibatch of L samples (s_l, a_l, r_l, s_{l+1}) from *RB*.

15: Set $y_l = r_l + \gamma Q'(s_{l+1}, \mu'(s_{l+1} \mid \theta^{\mu'}) \mid \theta^{Q'})$.

16: Update critic by minimizing the loss:

$$loss = \frac{1}{L} \sum_{l} (y_l - Q(s_l, a_l \mid \theta^{Q}))^2.$$

17: Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^{\mu}} J \approx \frac{1}{L} \sum_{l} \nabla_a Q(s_l, \mu(s_l) \mid \theta^{Q}) \nabla_{\theta^{\mu}} \mu(s_l \mid \theta^{\mu}).$$

18: Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1 - \tau)\theta^{Q'},$$
$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1 - \tau)\theta^{\mu'}.$$

19: **if** there is a UAV without power **then**

20: break

21: **end if**

22: **end for**

23: **end for**

At the start of each episode, the positions of all UAVs are randomly initialized and all UAVs have fully charged batteries. A random noise $\mathcal{N}$ which is used to balance exploration and exploitation is initialized. In the early stage, the policy is far from optimal, and various actions need to be explored. As the algorithm is iterated, the policy gradually converges. Therefore, it is needed to decrease exploration and increase exploitation. In our implementation, we use Gaussian noise with mean 0 and variance *var*. The value of *var* is 5 in the beginning and times 0.995 after each time slot. The episode terminates when there is a UAV without energy or the length of this episode is longer than T.

At the beginning of the time slot t_j, the agent determines actions a_j according to the actor network μ, current state s_j and the Gaussian noise $\mathcal{N}$, where $a_j = \mu(s_j) + \mathcal{N}$. Then, the actions are distributed to each UAV and the UAVs execute received actions. Corresponding flight energy consumption is calculated according to equation (16). During the flight, the UAV does not serve users. After the flight, if the UAV is not in charging station, the UAV will hover and provide communication service for users in the remaining time of t_j. Otherwise, the UAV will charge until the end of t_j. At the end of t_j, the communication energy consumption, hover energy consumption, charging energy and the data volume are calculated according to the system model in Section III. Then, the reward r_j is calculated according to equation (25) and a new state s_{j+1} is observed. The tuple, (s_j, a_j, r_j, s_{j+1}), is stored in the replay buffer *RB*. Next, a minibatch with a size of L is randomly sampled from the *RB*. *loss* is calculated according to the target network μ', Q', the critic network Q and the samples in the minibatch, as shown in lines 15-16. The critic network parameters are updated by minimizing *loss* (line 16) and the actor network parameters are updated by policy gradient (line 17). Finally, the target networks μ', Q' are updated by slowly tracking the original networks, as shown in line 18.

VI. SIMULATION AND PERFORMANCE EVALUATION

In this section, we present simulation to evaluate the performance of UC-DDPG. The simulation runs are performed with TensorFlow 1.12 [60]. We consider the users are randomly located within a square area with a size of $400m \times 400m$. The minimum and maximum allowed height of UAV are $20m$ and $50m$ respectively. The number of UAVs is 2 and we test the performance of UC-DDPG under three different user numbers (10, 15 and 20). The charging station is a cuboid with the size of $20m \times 20m \times 15m$. The UAVs communicate in an urban environment with $\psi = 12.08, \zeta = 0.11, \eta_{LoS} = 1.6\,dB$ and $\eta_{NLoS} = 23\ dB$ at 2GHz carrier frequency [20]. The size of replay buffer *RB* is 30000 and the size of minibatch L is 64. The total service time T is 2 hours and the time slot t_{slot} is 60 seconds. Therefore, there are 120 time slots in T. The main simulation parameters are listed in Table 1. Both actor network and critic network are feed-forward fully connected neural network and their parameters are listed in Table 2 and Table 3.

In particular, Random flight and Hilbert curve flight are used as benchmark.

Random flight. Random flight is very straightforward. In time slot t_j, for each UAV i, the flight direction polar angle φ_i, the azimuthal angle ϕ_i and flight distance d_i are randomly selected in $[0, \pi]$, $[0, 2\pi)$ and $[0, d_{max}]$ respectively. Similarly, if a UAV flies off the border, it will stay at the border.

Hilbert curve flight. Hilbert curve flight is a traversal algorithm. The altitude of all UAVs is same and fixed, which is $35m$ in our simulation. UAVs fly along the 3rd-order Hilbert

TABLE 1. Key parameters of simulation.

Notation	Definition	Vaule
T	Total service time	2 hours
ϖ	The angle of the sensing cone	90°
ψ	The parameter of A2G path loss model	12.08
ζ	The parameter of A2G path loss model	0.11
M	Number of the rotors	4
ρ	A fluid density of the air	1.2 kg/m^3
β	A rotor disk radius	0.25 m
W	Weight of the frame	1.5 kg
m	Weight of the battery and payload	2 kg
g	Gravity	9.8 N/kg
B	Channel bandwidth	10 KHz
f_c	Carrier frequency	2 GHz
η_{LoS}	Additional path loss to free space for LoS	1.6 dB
η_{NLoS}	Additional path loss to free space for NLoS	23 dB
σ^2	Noise Power	-100 dBm
P_t	Transmit power	5 W
P_{cu}	On-board circuit power	0.01 W
v	The velocity of the UAV	10 m/s
E_{max}	Battery capacity	250 kJ
C_{SE}	Charging power	400 W
t_{slot}	The length of the time slot	60 s
C_{D_0}	The drag coefficient	0.025
c_b	The rotor chord	0.022 m
S	The reference area	0.192 m^2
d_{max}	The largest allowed flight distance	100 m
τ	The parameter used to update Q' and μ'	0.001

TABLE 2. Parameters of actor neural network.

Name	Number	Size	Activation Function
Input Layer	1	$4N + K$	NA
Hidden Layer	2	300, 300	ReLU, ReLU6
Output Layer	1	$3N$	Tanh

TABLE 3. Parameters of critic neural network.

Name	Number	Size	Activation Function
Input Layer	1	$7N + K$	NA
Hidden Layer	2	300, 300	ReLU, ReLU6
Output Layer	1	1	NA

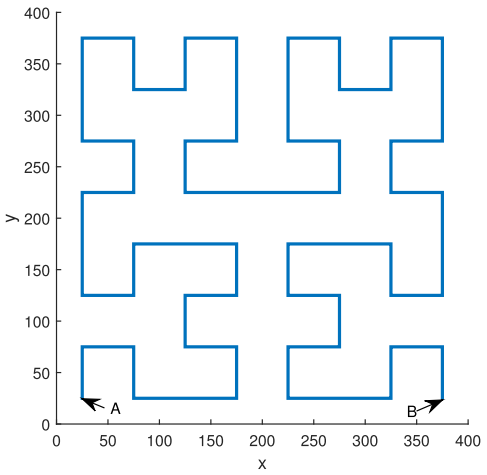


FIGURE 4. 3rd-order Hilbert curve with the unit length of 50*m*.

curve, as shown in Figure 4. The two UAVs start from points A and B respectively. In each time slot, the UAVs fly a unit distance which is 50m in our implementation. If the UAV reaches the endpoint, it will double back. Specially, if the residual energy of the UAV is less than 20%E_{max} at the start of time slot, the UAV will fly to charging station to replenish energy until the battery is full. Then the UAV returns to the original position and serves users sequentially.

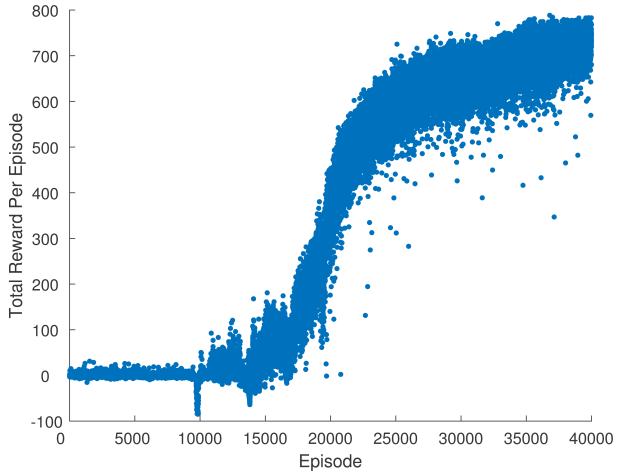


FIGURE 5. The total reward of each episode during training.

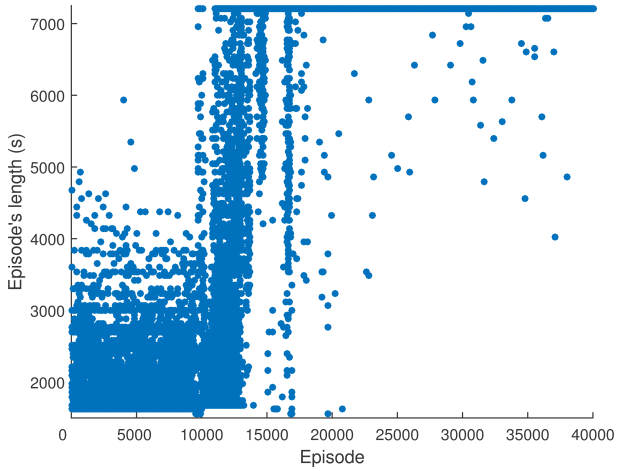
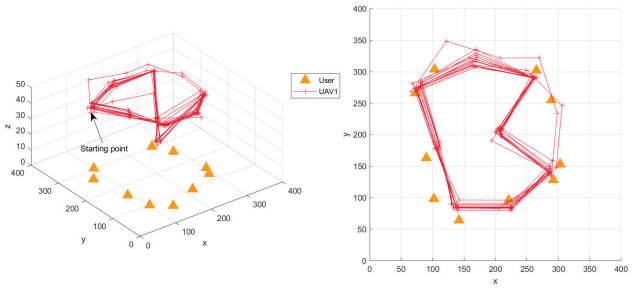


FIGURE 6. The length of each episode during training.

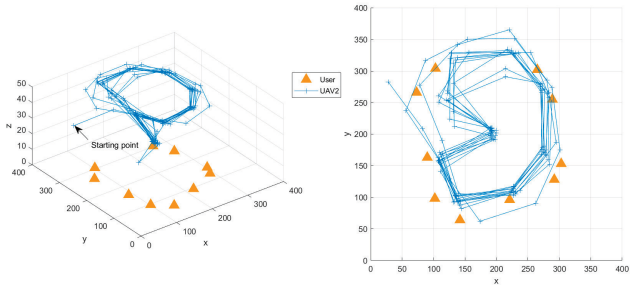
A. TRAINING

For each scenario with different numbers of users, we execute training with 40000 episodes. We show the training process by the example of 10 users, and the training processes in all scenarios are similar. The total reward of each episode and the length of each episode are shown in Figure 5 and Figure 6.

It can be observed that the Total Reward Per Episode (TRPE) is small and almost unchanged in the first 10000 episodes. Then, the TRPE shows fluctuation between 10000th episode and 17000th episode. For the rest of episodes, it gradually increases and stabilizes. Figure 5 and Figure 6 correspond to each other. In the first 10000 episodes, the service time of each episode is short, which results in the small TRPE. After that, we can see that the agent learns how to charge between 10000th episode and



(a) The flight trajectory of UAV1. The right is the top view.



(b) The flight trajectory of UAV2. The right is the top view.

FIGURE 7. **The flight trajectories of all UAVs.**

17000th episode. Finally, the agent knows how to charge and can provide persistent service in 7200 seconds, which is the length of T. It is interesting that the agent increases the TRPE by learning the way of charging firstly. As shown in Figure 6, the agent has learned to charge in 17000th episode, but the TRPE of 17000th episode is not high. Afterwards, the agent increases the TRPE by learning how to serve user fairly and energy efficiently.

After training, the flight trajectories of all UAVs which are generated by the actor network μ are shown in Figure 7. It seems that the agent learned a sort of cyclic trajectory. All UAVs circle and repeat the pattern of flying, serving or charging. We can also see that the UAVs do not fly to the places without users to avoid wasting energy. It is a pity that the agent did not learn to let two UAVs collaborate effectively, such as serving different users separately to further reduce flight energy consumption. The research on effective cooperation will be put into our future work.

B. ENERGY EFFICIENCY

Data volume and energy efficiency of different scenarios are depicted in Figure 8 and Figure 9, where energy efficiency equals to data volume divided by energy consumption. We use the actor network which trained 40000 episodes to determine actions. We can see that UC-DDPG has the best performance. Compared with Hilbert curve flight, UC-DDPG gets about twice the amount of data and the energy efficiency. As expected, Random flight has the worst performance. In Random flight, the UAV flies randomly and does not know replenishing energy, therefore it runs out of power soon. As a result, Random flight has the minimal data

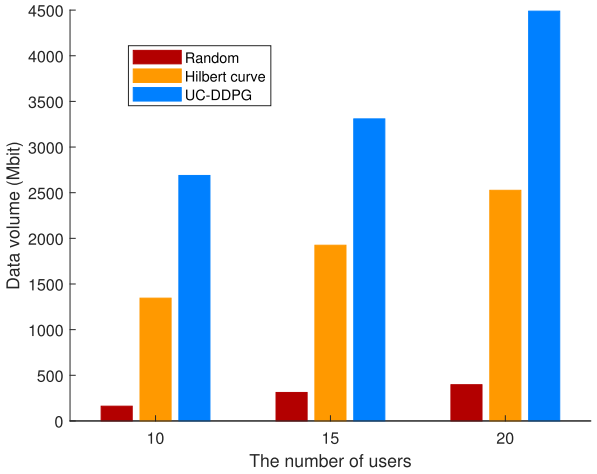


FIGURE 8. **Comparisons of data volume over different algorithms and user number.**

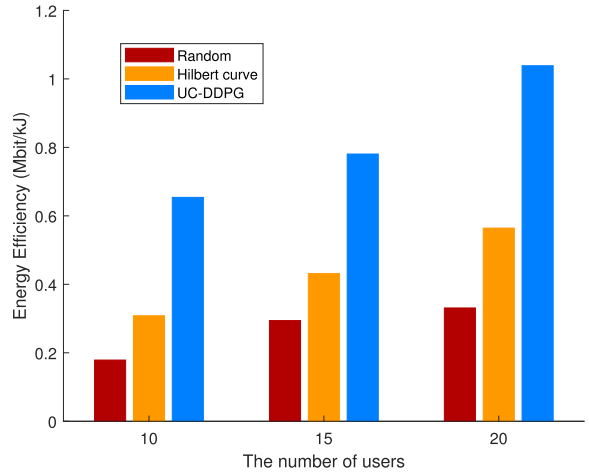


FIGURE 9. **Comparisons of energy efficiency over different algorithms and user number.**

volume and the lowest energy efficiency. It can be observed that the data volume and the energy efficiency of Random flight increase with the growth of number of users. It is because that UAVs have higher probability of covering users when the number of users increases. Hilbert curve flight has better performance than Random flight, but is not as good as UC-DDPG. In Hilbert curve flight, the UAVs traverse all places. However, there may be some places without any user. Consequently, the UAVs waste energy in vain. On the other hand, the positions used to serve users in Hilbert curve flight may be energy inefficient due to the poor channel environment, which leads to further deterioration of energy efficiency. In contrast, as shown in Figure 7, UC-DDPG avoids places where there are no users by training, and learns the better service positions, which increases the data volume and improves the energy efficiency.

C. FAIRNESS

Figure 10 shows the fairness in different scenarios. We can see that both Hilbert curve flight and UC-DDPG have pretty

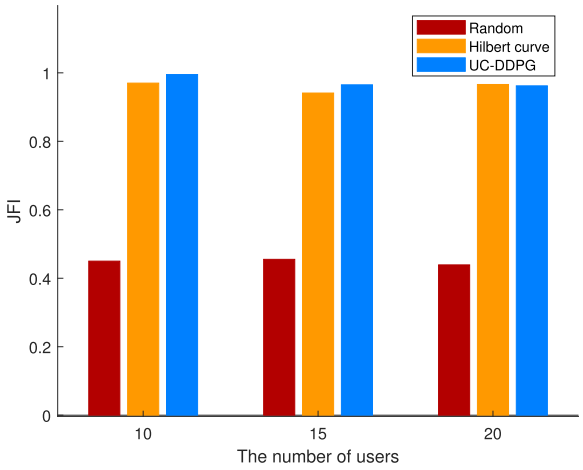


FIGURE 10. Comparisons of fairness over different algorithms and user number.

good fairness and Random flight has inferior fairness. It is not strange that the fairness has no relation with the number of users. Because in Random flight, the UAVs randomly fly, it usually has very low fairness. Hilbert curve flight traverses all places, thus it always produces good fairness. In UC-DDPG, JFI in reward r_j and normalized data volume in state s_j guide the agent to learn how to serve users fairly. After training, the agent learned fair service policy.

VII. CONCLUSION

The energy efficient, fair 3-D deployment and energy replenishment policy of multiple UAVs are jointly studied in this paper. Firstly, we build detailed channel model, data rate model and energy model. Then, inspired by the success of DRL, we propose a UAV control policy based on DDPG which is a deep actor-critic algorithm. The state, action and reward of RL are carefully designed under the consideration of energy efficiency, fairness and persistence. A lot of training ensures the performance of UC-DDPG. Simulation results show that UC-DDPG has good convergence and outperforms other scheduling algorithms (Random flight and Hilbert curve flight) in terms of data volume, energy efficiency and fairness. In future work, we plan to use multi-agent DRL to improve the cooperation between UAVs.

REFERENCES

[1] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8052–8066, Dec. 2017.

[2] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervas. Comput.*, vol. 16, no. 1, pp. 24–32, Jan. 2017.

[3] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[4] M. Asadpour, B. Van den Bergh, D. Giustiniano, K. Hummel, S. Pollin, and B. Plattner, "Micro aerial vehicle networks: An experimental analysis of challenges and opportunities," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 141–149, Jul. 2014.

[5] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[6] H. Shakhatreh, A. Khreishah, and B. Ji, "Providing wireless coverage to high-rise buildings using UAVs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[7] 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.

[8] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[9] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1647–1650, Aug. 2016.

[10] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.

[11] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

[12] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Optimal transport theory for power-efficient deployment of unmanned aerial vehicles," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[13] J. Lu, S. Wan, X. Chen, Z. Chen, P. Fan, and K. B. Letaief, "Beyond empirical models: Pattern formation driven placement of UAV base stations," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3641–3655, Jun. 2018.

[14] 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.

[15] A. Trotta, M. D. Felice, F. Montori, K. R. Chowdhury, and L. Bononi, "Joint coverage, connectivity, and charging strategies for distributed UAV networks," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 883–900, Aug. 2018.

[16] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. ICLR*, 2016, pp. 1–14.

[18] 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.

[19] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Joint sum-rate and power gain analysis of an aerial base station," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[20] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–5.

[21] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 434–437, Aug. 2017.

[22] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[23] E. Kalantari, H. Yanikomeroglu, and A. Yongacoglu, "On the number and 3D placement of drone base stations in wireless cellular networks," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–6.

[24] S. Koulali, E. Sabir, T. Taleb, and M. Azizi, "A green strategic activity scheduling for UAV networks: A sub-modular game perspective," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 58–64, May 2016.

[25] C. D. Franco and G. Buttazzo, "Energy-aware coverage path planning of UAVs," in *Proc. IEEE Int. Conf. Auto. Robot Syst. Competitions*, Apr. 2015, pp. 111–117.

[26] B. D. Song, J. Kim, J. Kim, H. Park, J. R. Morrison, and D. H. Shim, "Persistent UAV service: An improved scheduling formulation and prototypes of system components," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, May 2013, pp. 915–925.

[27] L. Di Puglia Pugliese, F. Guerriero, D. Zorbas, and T. Razafindralambo, "Modelling the mobile target covering problem using flying drones," *Optim. Lett.*, vol. 10, no. 5, pp. 1021–1052, Jun. 2016.

[28] H. Shakhatreh, A. Khreishah, J. Chakareski, H. B. Salameh, and I. Khalil, "On the continuous coverage problem for a swarm of UAVs," in *Proc. IEEE 37th Sarnoff Symp.*, Sep. 2016, pp. 130–135.

[29] A. Trotta, M. Di Felice, K. R. Chowdhury, and L. Bononi, "Fly and recharge: Achieving persistent coverage using small unmanned aerial vehicles (SUAVs)," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[30] J. Kim, B. D. Song, and J. R. Morrison, "On the scheduling of systems of UAVs and fuel service stations for long-term mission fulfillment," *J. Intell. Robotic Syst.*, vol. 70, nos. 1–4, pp. 347–359, Apr. 2013.

[31] J. Leonard, A. Savvaris, and A. Tsourdos, "Energy management in swarm of unmanned aerial vehicles," in *Proc. Int. Conf. Unmanned Aircr. Syst.*, Jul. 2013, pp. 124–133.

[32] A. Chakrabarty and J. Langelaan, "Energy maps for long-range path planning for small- and micro- UAVs," in *Proc. AIAA Guid., Navigat., Control Conf.*, Aug. 2009, pp. 1–13.

[33] J. W. Langelaan, "Tree-based trajectory planning to exploit atmospheric energy," in *Proc. Amer. Control Conf.*, Jun. 2008, pp. 2328–2333.

[34] W. H. Al-Sabban, L. F. Gonzalez, and R. N. Smith, "Wind-energy based path planning for unmanned aerial vehicles using Markov decision processes," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 784–789.

[35] N. Zlatanov, D. W. Kwan Ng, and R. Schober, "Capacity of the two-hop relay channel with wireless power transfer from relay to source and processing cost," *IEEE Trans. Wireless Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[36] J. Xu and R. Zhang, "Energy beamforming with one-bit feedback," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5370–5381, Oct. 2014.

[37] T. J. Nugent and J. T. Kare, "Laser power for UAVs," LaserMotive, Kent, WA, USA, White Paper, 2010.

[38] M. C. Achtelik, J. Stumpf, D. Gurdan, and K.-M. Doth, "Design of a flexible high performance quadcopter platform breaking the MAV endurance record with laser power beaming," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 5166–5172.

[39] J. Ouyang, Y. Che, J. Xu, and K. Wu, "Throughput maximization for laser-powered UAV wireless communication systems," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[40] R. Sowah, M. A. Acquah, A. R. Ofoli, G. A. Mills, and K. M. Koumadi, "Rotational energy harvesting to prolong flight duration of quadcopters," in *Proc. 2015 IEEE Ind. Appl. Soc. Annu. Meeting*, Dec. 2015, pp. 1–7.

[41] M. K. Simon and M. Alouini, "A unified approach to the performance analysis of digital communication over generalized fading channels," *Proc. IEEE*, vol. 86, no. 9, pp. 1860–1877, Sep. 1998.

[42] S. O. Rice, "Statistical properties of a sine wave plus random noise," *Bell Syst. Tech. J.*, vol. 27, no. 1, pp. 109–157, Jan. 1948.

[43] M. M. Azari, F. Rosas, K.-C. Chen, and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 330–344, Jan. 2018.

[44] X. Yuan, Z. Feng, W. Xu, W. Ni, J. A. Zhang, Z. Wei, and R. P. Liu, "Capacity analysis of UAV communications: Cases of random trajectories," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7564–7576, Aug. 2018.

[45] D. da Costa and S. Aissa, "Capacity analysis of cooperative systems with relay selection in Nakagami-*m* fading," *IEEE Commun. Lett.*, vol. 13, no. 9, pp. 637–639, Sep. 2009.

[46] J. G. Leishman, *Principles of Helicopter Aerodynamics (Cambridge Aerospace Series)*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[47] A. Thibbotuwawa, Z. A. Banaszak, G. Bocewicz, and P. Nielsen, "Energy consumption in unmanned aerial vehicles: A review of energy consumption models and their relation to the UAV Routing," in *Proc. 38th Inf. Syst. Archit. Technol. Conf.*, Jan. 2018, pp. 173–184.

[48] K. Dorling, J. Heinrichs, G. G. Messier, and S. Magierowski, "Vehicle routing problems for drone delivery," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 1, pp. 70–85, Jan. 2017.

[49] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 2019.

[50] G. Cao, Z. Lu, X. Wen, T. Lei, and Z. Hu, "AIF: An artificial intelligence framework for smart wireless network management," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 400–403, Feb. 2018.

[51] Z. Han, T. Lei, Z. Lu, X. Wen, W. Zheng, and L. Guo, "Artificial intelligence-based handoff management for dense WLANs: A deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 31688–31701, 2019.

[52] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[53] J. Liu, B. Krishnamachari, S. Zhou, and Z. Niu, "DeepNap: Data-driven base station sleeping operations through deep reinforcement learning," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4273–4282, Dec. 2018.

[54] S. Liu, X. Hu, and W. Wang, "Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems," *IEEE Access*, vol. 6, pp. 15733–15742, 2018.

[55] Y. He, C. Liang, F. R. Yu, N. Zhao, and H. Yin, "Optimization of cache-enabled opportunistic interference alignment wireless networks: A big data deep reinforcement learning approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[56] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[57] Y. Cao, L. Zhang, and Y.-C. Liang, "Deep reinforcement learning for multi-user access control in UAV networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

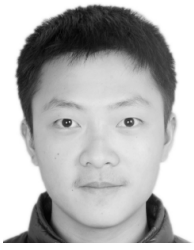
[58] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.

[59] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[60] *Tensorflow.Org*. Accessed: Jan. 2020. [Online]. Available: https://www.tensorflow.org/

HANG QI received the B.Sc. degree in communication engineering from Shandong University, Jinan, China. He is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include machine learning, design of MAC protocol for wireless networks, and optimization in wireless networks.

ZHIQUN HU received the B.E. degree in communication engineering from the Hubei University of Technology, Wuhan, China, in 2012, and the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2018. Since 2018, she has been with Hubei University, where she is currently an Assistant Professor of computer science and information engineering. Her main research interests include wireless communications, deep reinforcement learning, performance analysis, V2X, and UAV communication.



HAO HUANG received the B.Sc. degree in communication engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China. He is currently pursuing the Ph.D. degree in communication engineering with the Beijing University of Posts and Telecommunications, China. His current research interests include reinforcement learning and intelligent transportation.

XIANGMING WEN received the B.E., M.S., and Ph.D. degrees in electrical engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China. He is currently the Vice President of BUPT, where he is also a Professor with the Communication Network Center and the Director of the Beijing Key Laboratory of Network System Architecture and Convergence. He is also the Principle Investigator of more than 18 projects, including the National Key Project of Hi-Tech Research and Development Program of China (863 Program) and the National Natural Science Foundation of China. He is also the Vice Director of the Organization Committee of the China Telecommunication Association. In the last five years, he has authored more than 100 published articles. His current research include broadband mobile communication theory, multimedia communications, and information processing

ZHAOMING LU received the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2012. He is currently an Associate Professor with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His research interests include open wireless networks, QoE management in wireless networks, software defined wireless networks, cross-layer design for mobile video applications, and network-assisted autonomous driving.