Learning-based resilience guarantee for multi-UAV collaborative QoS management

Bai, Chengchao; Yan, Peng; Yu, Xiaoqiang; Guo, Jifeng

DOI
10.1016/j.patcog.2021.108166

Publication date
2022

Document Version
Final published version

Published in
Pattern Recognition

Pattern Recognition 122 (2022) 108166

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Learning-based resilience guarantee for multi-UAV collaborative QoS management

Chengchao Bai[a,1,*], Peng Yan[b,1], Xiaoqiang Yu[b], Jifeng Guo[b]

[a] Delft University of Technology, Stevinweg 1, Delft 2627 CN, the Netherlands
[b] School of Astronautics, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 8 December 2020
Revised 17 May 2021
Accepted 4 July 2021
Available online 1 September 2021

Keywords:
Unmanned business
Communication service
Multi-UAV
Deep reinforcement learning
QoS-aware
System resilience

ABSTRACT

Unmanned and intelligent technologies are the future development trend in the business field. It is of great significance for the connotation analysis and application characterization of massive interactive data. Particularly, during major epidemics or disasters, how to provide business services safely and securely is crucial. Specifically, providing users with resilient and guaranteed communication services is a challenging business task when the communication facilities are damaged. Unmanned aerial vehicles (UAVs), with flexible deployment and high maneuverability, can be used to serve as aerial base stations (BSs) to establish emergency networks. However, it is challenging to control multiple UAVs to provide efficient and fair communication quality of service (QoS) to users due to their limited communication service capabilities. In this paper, we propose a learning-based resilience guarantee framework for multi-UAV collaborative QoS management. We formulate this problem as a partial observable Markov decision process and solve it with proximal policy optimization (PPO), which is a policy-based deep reinforcement learning method. A centralized training and decentralized execution paradigm is used, where the experience collected by all UAVs is used to train the shared control policy. Each UAV takes actions based on the partial environment information it observes. In addition, the design of the reward function considers the average and variance of the communication QoS of all users. Extensive simulations are conducted for performance evaluation. The simulation results indicate that (1) the trained policies can adapt to different scenarios and provide resilient and guaranteed communication QoS to users, (2) increasing the number of UAVs can compensate for the lack of service capabilities of UAVs, (3) when UAVs have local communication service capabilities, the policies trained with PPO have better performance compared with the policies trained with other algorithms.

1. Introduction

New technological developments and application requirements have prompted continuous innovation in business models. Such as mobile payment, paperless transactions, and intelligent business services deeply integrated with unmanned robotics. Especially in today's rapid implementation of artificial intelligence methods represented by machine learning and deep learning, how to efficiently guarantee the acquisition of high-value information from massive commercial data and ensure the accuracy and effectiveness of services is critical [1,2]. At present, many scholars have conducted in-depth research on high-dimensional data analysis [3–6], risk assessment [7], and framework design [8–10]. However, with the widespread promotion of intelligent unmanned business models, the use and study of interactive data generated during the service process (such as the interaction data between unmanned business machine platforms and the environment) are still relatively small. Besides, these data are related to the quality and completion of business services. Therefore, it is precious to carry out data-based intelligent unmanned commercial application research.

Unmanned and intelligent technologies will be the inevitable elements in the business field in the future. The emergence of unmanned delivery, unmanned takeout, unmanned supermarkets, unmanned rentals, etc., gives good examples. Particularly, in view of the epidemic of COVID-19, providing safe, efficient, and customized business services has become the new focal point of development as well as a challenge. Specifically, communication service quality is a crucial problem. Therefore, using data analysis to realize the communication service guarantee of multiple intelligent business

* Corresponding author.
E-mail addresses: C.Bai@tudelft.nl (C. Bai), yanpeng@hit.edu.cn (P. Yan), 6111820504@hit.edu.cn (X. Yu), guojifeng@hit.edu.cn (J. Guo).
[1] Chengchao Bai and Peng Yan contribute equally to this work.

https://doi.org/10.1016/j.patcog.2021.108166
0031-3203/

nodes is the key to the promotion of commercialization. The application of unmanned technology is a promising approach to guarantee service optimality. Unmanned aerial vehicles (UAVs) greatly support communication services under resource constraints owing to their advantages of wide airspace and flexibility in networking.

In recent years, UAVs have witnessed a wide range of applications in many scenarios [11,12]. The flexible deployment, high maneuverability, and more importantly, significantly lower production costs of UAVs contribute to the increasingly important roles they play in various applications [13–15]. In particular, UAVs can serve as aerial base stations (BSs) to provide services to ground users. During catastrophic natural disasters, for example, when ground BSs are damaged, UAVs can be quickly deployed to establish emergency networks, providing wireless communication service to ground devices with surviving BSs [16].

UAVs serving as aerial BSs for ground user communication have the following advantages. First, the high flying height of UAVs allows easy establishment of line-of-sight (LoS) connections with ground users, improving the quality of communication [17]. Second, easy and fast deployment of UAVs for emergency communication services minimizes losses and effectively meets user requirements. Third, various types of UAVs such as rotary-wing and fixed-wing UAVs can be arranged to cater to the requirements of different users.

However, various challenges are encountered in enabling UAVs to provide efficient and fair communication quality of service (QoS) to users in a target region. First, when providing communication services to users over a large area, the UAVs usually have local communication service capabilities due to energy consumption and economic cost. Besides, because of limited resources, a sufficient number of UAVs cannot be deployed to serve the target region. Hence, UAVs are required to have high mobility to maintain the quality of communication services in the area. Second, real-time connection of UAVs is required because a UAV network usually has an extremely limited number of gateways, and a UAV that is outside the communication range of other UAVs will not be able to obtain state information from other UAVs, resulting in the loss of connection of the ground users associated with it with the external network. Third, UAVs must consider the impact of possible obstacles such as tall buildings on the quality of communication services. In addition, we must consider the environment in which there is rapid movement of and dynamic changes in the number of users. More importantly, fairness is critical for communication service because every user has an equal right to obtain high quality communication services. Considering the above factors, in order to provide high-quality communication services for ground users, UAVs must be able to achieve autonomous coordination as a team, where each UAV serves as a communication node.

To address the above problems, we formulate the control of each UAV as a partial observable Markov decision process (POMDP) and solve it using a deep reinforcement learning (DRL) method. The DRL method has performance comparable or superior to humans on a range of tasks [18–21]. It learns the optimal policy parameterized by a deep neural network through interaction with the environment, which can enable the reinforcement learning (RL) agent to adapt to changes in the environment. Thus, we use a state-of-the-art policy-based DRL method, proximal policy optimization (PPO) [22], to enable the UAVs to provide resilient and guaranteed communication services to users, that is, efficient and fair communication service to users and adapting to the changes in the environment. Our objective is to maximize the communication QoS for the user with the worst QoS. To achieve this objective, in the design of the reward function, we consider the average and variance of the communication QoS of all users. In addition, to improve the training speed and the robustness of the UAV control policy, a centralized training and decentralized execution paradigm is used in the training process, where the experience collected by all UAVs is used to train the shared control policy. In the UAV control process, the trained policy is copied to each UAV, which takes action according to the partial environment information it observes.

The main contributions of this work are as follows: (i) A DRL-based multi-UAV control policy is proposed, which enables UAVs to provide efficient and fair communication services to users and adapt to the changes in the environment. (ii) The simulation results show that increasing the number of UAVs can compensate for the lack of service capabilities of UAVs. Thus, there are two ways to increase the communication QoS: increasing the number of UAVs and expanding the communication service area of UAVs. (iii) Comparison with other selected algorithms indicates that when the UAVs have local communication service capabilities, the policy trained with PPO has the best performance on this problem.

The remainder of this paper is organized as follows. Section 2 reviews the related work, and Section 3 describes the system model. Section 4 introduces the details about the design of the DRL method. Simulation results and corresponding discussions are presented in Section 5. Finally, conclusions are presented in Section 6.

2. Related work

In this section, we provide an overview of the related work.

2.1. Multi-UAV collaborative QoS management

Application of multiple UAVs to collaboratively provide communication services to users has been extensively studied [23]. In [24], an online UAV scheduling scheme was proposed, which can schedule and manage UAVs online to guarantee reliable links with internet of vehicles. The optimal UAV trajectory can be designed to provide users with high-quality communication services. However, the UAV scheduling scheme is on the base station and its utility is constrainted by the communication range between the UAV and the base station. In [25], the analysis and representation of motion trajectories in a highly informative way is studied. In [26], the UAV trajectory is optimized to maximize the sum rate of the edge users served by the UAV. The optimal UAV trajectory is obtained by solve a mixed-integer nonconvex problem with an iterative algorithm. However, only one single UAV is considered, which has limited communication service capability. In [27], UAVs were used to serve vehicles on a highway, considering the situations when the vehicles move between two road-side units and the communication infrastructures are partially or totally damaged. The QoS of each vehicle was guaranteed by optimizing the UAV trajectory with a successive convex approximation method and optimizing the radio resource allocation with a Linear Programming method. In [28], the fair communication service between UAVs and users was investigated by maximizing the communication QoS of the user with the worst QoS. The objective was implemented by the joint optimization of the UAV trajectories, the transmission power, and the user scheduling using an effective iterative algorithm based on successive convex approximation and block coordinate descent techniques. In [29], multiple rotary-wing UAVs were used as aerial BSs to provide communication services to ground users. This problem was formulated as a multi-objective optimization problem and solved with particle swarm optimization-based techniques. However, in Samir et al. [27], Bejaoui et al. [28], Perabathini et al. [29], the UAVs have global perception capabilities and can know the locations of all users. The limited perception capabilities of the UAVs are not considered. In [30], the energy-efficient cooperative control policy of rechargeable multi-UAVs was considered for providing seamless coverage and long-term information services

for the nodes of Internet of Things with limited cruising duration. The problem was formulated as a mixed-integer nonconvex problem and solved by exploiting sequential convex optimization techniques. In [31], UAV trajectories were designed such that the fairness rate is maximized and their effect on users' spectral efficiency was studied. The UAVs were directed by a data-rate gradient calculated using the free-space path-loss channel model. In addition, the proposed method can adapt to changes in the user locations online and capture the interaction between multiple UAV trajectories using a central processing unit. However, both [30,31] do not consider the limitation of communication distance between UAVs and the obstruction of user communication services by obstacles in the environment.

2.2. UAV control for communication service using reinforcement learning

The use of a reinforcement learning method to enable UAVs to provide reliable and flexible communication services to users has been studied [32]. In [33], the UAV placement problem in a manned-and-unmanned (MUM) is considered. The positions of relay UAVs are guided to support the broken wireless links according to traffic QoS requirements and the link conditions. The deep Q-learning model is used to determine the optimal link between two UAV nodes. However, the UAV link selection is computed with a centralized manner and the performance is limited by the number of UAVs. In [34], the trajectory of an UAV is optimized to maximize the sum rate of the transmission during flying time. A Q-learning method is used to train the movement decision policy for the UAV. In [35], the optimal positioning of the aerial BS was considered by taking into account the user mobility and guaranteeing the minimum QoS. A Q-learning-based method was proposed to guide the UAV to the optimal position based on past experiences. However, in both [34,35], only a single UAV is used and multi-UAV is not considered. In [36], the trajectory design problem for UAVs in a UAV-to-Device communication scenario was studied, and a Q-learning-based multi-agent DRL method was proposed to solve this problem. In [37], the UAV trajectory design problem was studied to perform different real-time sensing tasks and an enhanced multi-UAV Q-learning algorithm was proposed to solve this problem. In [36,37], however, each UAV carries out tasks independently, without considering multi-UAV coordination. In [38], the problem of resource allocation of multiple UAVs for providing communication services to ground users was investigated. The uncertainty of environments was considered by formulating the resource allocation problem as a stochastic game. A multi-agent reinforcement learning framework based on Q-learning was proposed to determine the optimal control policy of each UAV according to its local observations. Compared with our work, the fairness of communication services for ground users is not considered. In [39], the trajectory design problem for UAVs is studied where a cellular Internet of UAVs are used to guarantee the QoS of the transmission of the sensory data. The problem is considered as a Markov decision problem (MDP) and solved with a multi-agent deep reinforcement learning method. In [40], the Q-learning method was used to solve the problem of UAV trajectory optimization under the constraint of QoS requirements in an energy-efficient manner. However, in both [39,40], the impact of obstacles in the environment on communication services is not considered. The work [17] is similar to ours. In [17], a DRL-based method, DRL-EC3 (DRL-based energy-efficient control for coverage and connectivity), was proposed for UAV control. This method maximizes an energy efficiency function and considers the communication coverage, fairness, energy consumption, and the connectivity of UAVs. However, the UAVs are assumed to have global perception capabilities and the limited perception capabilities of the UAVs is not considered.

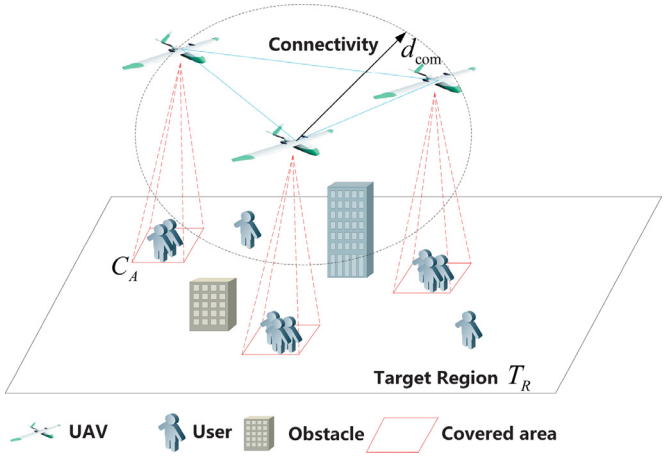


Fig. 1. UAV network providing communication services for ground users in a target region.

2.3. Comparison of our work with the related studies

Our work differs from the aforementioned related studies in the following aspects:

- We consider the resilient changes in the environment, including the number of users, the speed of users' movement, the number of UAVs, and the number of obstacles in the environment.
- We compare the effects of UAVs' communication service capabilities on the communication QoS of the users.
- Our approach is compared with two traditional task planning algorithms and a value-based reinforcement learning method to demonstrate its effectiveness.

3. System model

In this section, we describe the system model and formulate the QoS management problem as an optimization problem.

3.1. Problem description

We consider a team of N UAVs flying at different altitudes to maintain communication services for M users on the ground target region T_R. The UAVs can avoid collisions by flying at different altitudes. The communication range of the UAV is denoted as d_{com}. All UAVs can obtain the states of other UAVs through the communication network. The communication service area of UAVs is limited, which is denoted as C_A. All UAVs know the distribution of users and the quality of user services within their service area. Our goal is to determine an optimal control policy for the UAV team by optimizing the trajectories of the UAVs so that they can provide better communication QoS for users in the target area. The problem scenario is shown in Fig. 1.

3.2. UAV model

In this paper, we use a simplified fixed-wing UAV model as defined in Qiu and Duan [41]. UAV i is assumed to fly at a fixed altitude H_i and speed V_i. The kinematics of UAV i can be formulated as

$$\begin{cases} \dot{x}_i = V_i \cos\psi_i \\ \dot{y}_i = V_i \sin\psi_i \\ \dot{\psi}_i = (\psi_{ic} - \psi_i)/\tau_\psi \end{cases} \tag{1}$$

where $\mathbf{p}_i = (x_i, y_i)$ is the position of UAV i in the two-dimensional Cartesian coordinate system, ψ_i is the velocity heading angle, ψ_{ic} is

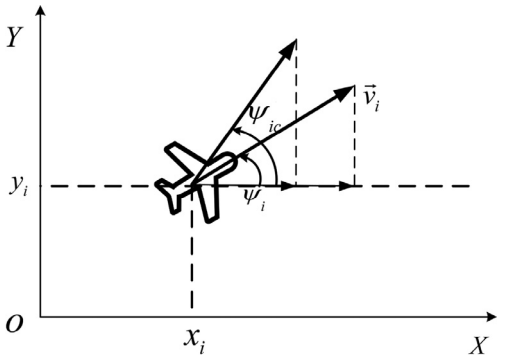


Fig. 2. UAV kinematics model.

the velocity heading angle command, and τ_ψ is the time constant related to the dynamics of the UAV. The kinematics model of UAV is shown as Fig. 2.

Considering the motion constraints of the UAV, the velocity heading angular rate is limited to

$$\left|\dot{\psi}_i\right| \le n_{\max} g / V_i \tag{2}$$

where $n_{\max}$ is the maximum lateral overload limit of the UAV and g is the acceleration of gravity.

3.3. User model

In this paper, we consider the user model as a random walk model in the target area. Each user k moves at a fixed speed V_k^{user} and in a random direction $\psi_k (\psi_k \in [0, 2\pi])$. The positions of all users are restricted to the target area, users cannot enter the obstacle area, and the positions of two users cannot overlap.

3.4. Communication model

The communication model between UAVs and users mainly refers to Bejaoui et al. [28]. We assume that the position of user k on the ground is $\mathbf{w}_k = (x_k, y_k)$, and the projection position of UAV i on the horizontal plane is $\mathbf{p}_i = (x_i, y_i)$, flying at a fixed altitude H_i. The distance between UAV i and user k is expressed by

$$d_{k,i} = \sqrt{{H_i}^2 + \|\mathbf{p}_i - \mathbf{w}_k\|^2} \tag{3}$$

For simplicity, we assume that the UAVs have LoS links to ground users if the LoS is not obstructed by an obstacle. In addition, the Doppler effect caused by relative motion is ignored. Thus, the free-space path loss model can be used to describe the communication channel between UAV i and user k, which can be expressed as

$$h_{k,i} = \frac{\beta_0}{{d_{k,i}}^2} \tag{4}$$

where β_0 represents the channel gain at a reference distance ($d = 1$ m).

In this study, all the UAVs are assumed to work in the same frequency band. The corresponding received signal to interference plus noise ratio (SINR) at user k from UAV i is computed as

$$\gamma_{k,i} = \frac{p_i \cdot h_{k,i}}{\sum_{j=1, j\neq i}^{N} p_j h_{k,j} + \sigma_0^2} \tag{5}$$

where p_i denotes the down-link transmit power of UAV i, σ_0^2 represents the variance of additive white Gaussian noise at the receiver and the term $\sum_{j=1, j\neq i}^{N} p_j h_{k,j}$ stands for the co-channel interference received by user k during its communication with UAV i. The down-link transmit power of UAV i is a constant with the value of the maximum allowable transmit power $P_{\max}$. At the same time, to consider the effect of obstacles in the target area, when the LoS between UAV i and user k is obstructed by an obstacle, user k cannot receive the communication service of UAV i, i.e., $\gamma_{k,i} = 0$.

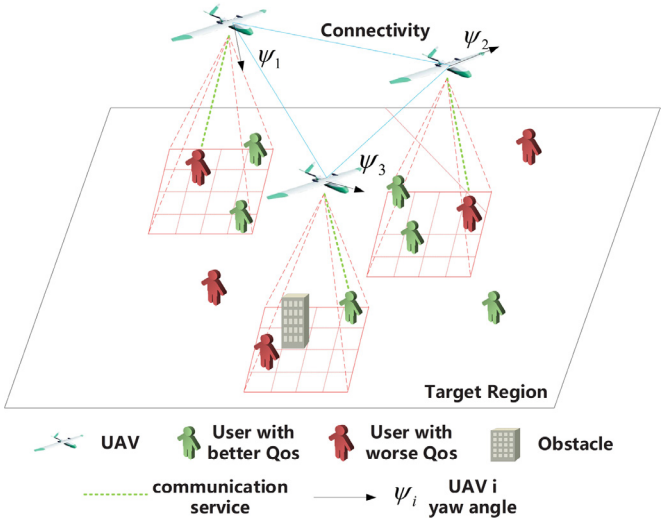


Fig. 3. Description of UAV–user communication service association strategy.

To reflect the correspondence between UAV i and user k, we define a binary variable $\alpha_{k,i}$, indicating whether UAV i is communicating with user k. We assume that at each time step t, each UAV can serve at most one user, and each user can be served by at most one UAV. Therefore, the constraints on $\alpha_{k,i}^t$ are as follows:

$$\sum_{k=1}^{M} \alpha_{k,i}^t \le 1, \quad \sum_{i=1}^{N} \alpha_{k,i}^t \le 1, \quad \forall k, i \tag{6}$$

Then, the achievable transmission rate of user k over time period T is given by

$$R_k^T = \frac{\Delta t}{T} \sum_{t=0}^{T} \sum_{i=1}^{N} \alpha_{k,i}^t \log_2\left(1 + \gamma_{k,i}\right) \tag{7}$$

where Δt is the time step. We define the communication QoS of user k, Q_k, within task time T as

$$Q_k = T * R_k^T \tag{8}$$

To simplify the UAV–user association problem, we assume that when the UAVs provide communication services at each time step, the user served by the UAV is one with the worst communication QoS in its service area and is not obstructed by obstacles. If multiple UAVs serve the same area simultaneously, the UAV that can provide the best service quality will first select the user to serve, and then the other UAVs select the other users to serve in turn. Therefore, the value of $\alpha_{k,i}$ can be expressed as follows:

$$\alpha_{k,i}^{t+1} = 1 \ \text{ where } \ k = \arg\min_k R_k^t \ \text{ and } \ i = \arg\max_i \gamma_{k,i}^{t+1} \tag{9}$$

The UAV–user association strategy is shown in Fig. 3.

3.5. Problem formulation

We consider the communication QoS of the users through two aspects:

(1) Increase the average value of the communication QoS of all users over a period of time T;
(2) Reduce the variance of the communication QoS of all users over a period of time T.

Based on the above two considerations, the communication QoS of the users is ensured by maximizing that of the user with the worst communication QoS. The problem can be described as

$$\begin{aligned} \max_{\psi_{ic}} \quad & \min Q_k, i = 1, 2, \ldots N, k = 1, 2, \ldots M \\ \text{s.t.} \quad & \left|\dot{\psi}_i\right| \le n_{\max} g / V_i \\ & \mathbf{p}_i \in T_R \\ & \left\|\mathbf{p}_i, \mathbf{p}_j\right\| \le d_{\text{com}} \end{aligned} \tag{10}$$

where the first constraint is the UAV's dynamic constraint, the second constraint means that the UAV cannot go out of the target region, and the third constraint means that the UAV cannot leave the network. Our objective is to determine the optimal control policy ψ_{ic} for each UAV to provide users with a fair and efficient communication service under the conditions that satisfy the above constraints.

4. Learning-based resilience guarantee design

Considering the complexity of the above optimization problem caused by the dynamic changes of the environment, users, and UAVs, traditional optimization methods cannot be applied. In this section, we formulate this problem as an RL problem in the context of a POMDP and solve it using a DRL method.

4.1. Reinforcement learning

The optimization problem defined in Section 3 can be formulated as a POMDP that is solved with reinforcement learning. In this RL framework, each UAV learns the control policy by interacting with the environment. Typically, at time t, each UAV independently obtains observation $\mathbf{o}_t$ from the environment and takes action $\mathbf{a}_t$ according to its policy $\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ parameterized by θ. Then, the environment changes and the UAV receives reward r_{t+1} that evaluates its selected action. The objective of each UAV is to learn an optimal policy that maps observations to actions by maximizing a long-term cumulated reward

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{11}$$

where $\gamma\,(0 < \gamma < 1)$ is the discount factor.

We use a DRL method to implement the control policy for each UAV, where a deep neural network parameterized by θ is used to approximate the UAV's control policy and a policy gradient algorithm to train this deep neural network for determining the optimal parameter θ^* that satisfies the constraints of (10). Next, the ingredients of DRL are introduced.

4.2. Observation space

At time t, the observation of UAV i consists of three parts, i.e., $\mathbf{o}_i^t = [\mathbf{o}_{i,1}^t, \mathbf{o}_{i,2}^t, \mathbf{o}_{i,3}^t]$, which is shown in Fig. 4.

(1) Observation $\mathbf{o}_{i,1}^t$ denotes the information of the users in service area C_A^i obtained by UAV i at time t, including the user's position relative to the UAV and the user's current communication QoS, i.e., $\mathbf{o}_{i,1}^t = [\mathbf{o}_{i,1}^{t,k_1}, \ldots, \mathbf{o}_{i,1}^{t,k_j}, \ldots]$, $\mathbf{o}_{i,1}^{t,k_j} = [d_{i,k_j}^t, \phi_{i,k_j}^t, Q_{i,k_j}^t]$, $k_j \in C_A^i$. The position of user k relative to UAV i is expressed in the form of polar coordinates in the velocity heading coordinate system of UAV i with distance $d_{i,k}^t$ and angle $\phi_{i,k}^t$. Observation $\mathbf{o}_{i,1}^t$ contains information of all users within the service area C_A^i of UAV i, so the dimension of $\mathbf{o}_{i,1}^t$ is $3 * M_i$ ($M_i = \left|C_A^i\right|$), where M_i denotes the total number of users in service area C_A^i, i.e., $\mathbf{o}_{i,1}^t \in \mathbf{R}^{3*\mathrm{M}_i}$.

(2) Observation $\mathbf{o}_{i,2}^t$ denotes the state information of other UAV j obtained by UAV i through the communication network at time t, including the position of UAV j relative to UAV i and the position and the communication QoS of the user with the worst communication QoS in the service area of UAV j. At time t, the position of UAV j relative to UAV i is expressed in the form of polar coordinates in the velocity heading coordinate system of UAV i with distance $d_{j,i}^t$ and angle $\phi_{j,i}^t$. The communication QoS of the user with the worst QoS within the service area of UAV j is expressed as $Q_j^{\min}(Q_j^{\min} = \min Q_k, k \in C_A^j)$, and the user's position is expressed in the form of polar coordinates in the velocity heading coordinate system of UAV i with distance $d_j^{Q_{\min}}$ and angle $\phi_j^{Q_{\min}}$. Thus, observation $\mathbf{o}_{i,2}^t$ can be represented as $\mathbf{o}_{i,2}^t = [\mathbf{o}_{i,2}^{t,1}, \ldots, \mathbf{o}_{i,2}^{t,j}, \ldots, \mathbf{o}_{i,2}^{t,N}]$, $\mathbf{o}_{i,2}^{t,j} = [d_{j,i}^t, \phi_{j,i}^t, Q_j^{\min}, d_j^{Q_{\min}}, \phi_j^{Q_{\min}}]$, $j = 1, 2, \ldots N, j \ne i$, and the dimension of $\mathbf{o}_{i,2}^t$ is $5 * (N-1)$, i.e., $\mathbf{o}_{i,2}^t \in \mathbf{R}^{5*(N-1)}$.

(3) Observation $\mathbf{o}_{i,3}^t$ denotes the displacement from UAV i to the boundary of the target region T_A at time t, expressed by four variables, namely $d_i^U = Y_{\max} - y_i$, $d_i^D = y_i - Y_{\min}$, $d_i^L = x_i - X_{\min}$, and $d_i^R = X_{\max} - x_i$, where $Y_{\max}$, $Y_{\min}$, $X_{\max}$, and $X_{\min}$ are the boundary values of the target region T_A; that is, the target region can be expressed as $\mathbf{p}_i \in T_A,\ \mathbf{p}_i = (x_i, y_i),\ X_{\min} \le x_i \le X_{\max},\ Y_{\min} \le y_i \le Y_{\max}$. Thus, observation $\mathbf{o}_{i,3}^t$ can be represented as $\mathbf{o}_{i,3}^t = [d_i^U, d_i^D, d_i^L, d_i^R]$.

4.3. Action space

We discretize the action space of UAV i as $\mathbf{a}_i \in \{-30, -20, -10, 0, 10, 20, 30\}$ deg, which represents the UAV heading angle that needs to be changed. The heading angle command of UAV i is calculated by

$$\psi_{ic} = \psi_i + \Delta\psi_i \tag{12}$$

where $\Delta\psi_i$ is sampled from $\mathbf{a}_i$ according to the selection probabilities calculated by the deep neural network.

4.4. Network architecture

We use a deep neural network to implement the control policy of UAV i, which inputs observation $\mathbf{o}_i^t$ and outputs the selection probabilities of actions $\mathrm{P}(\mathbf{a}_i^t|\mathbf{o}_i^t)$. All UAVs have the same control policy and share the same deep neural network, whose architecture is shown in Fig. 5.

As shown in Fig. 5, to deal with the variable length observation $\mathbf{o}_{i,1}^t$, we limit its maximum length to 5; that is, observation $\mathbf{o}_{i,1}^t$ processed by the network is the information of maximum 5 users. When the number of users is less than 5, we fill the remaining input positions with 0. When the number of the users in the UAV's service area is greater than 5, these 5 users are those with the worst communication QoS in service area C_A^i at time t; they are sorted from the least to the largest communication QoS. When the number of users in the UAV's service area is less than 5, these 5 users are those discovered latest by UAVs; that is, UAVs have a memory of the states of the 5 latest users. They are sorted from the latest to the oldest time of discovery. Similarly, to process the variable length observation $\mathbf{o}_{i,2}^t$, we limit its maximum length to 2; that is, observation $\mathbf{o}_{i,2}^t$ processed by the network is the states of maximum 2 UAVs. When the number of other UAVs is less than 2, we fill the remaining input positions with 0. These two UAVs are the farthest from UAV i at time t; they are sorted from the farthest to the nearest distance from UAV i. The cropped observations $\mathbf{o}_{i,1}^t$ and $\mathbf{o}_{i,2}^t$ and observations $\mathbf{o}_{i,3}^t$ are merged and processed by three fully connected layers, and the network finally outputs the probability values $\mathrm{P}(\mathbf{a}_i^t|\mathbf{o}_i^t)$. In the above network structure, the last fully

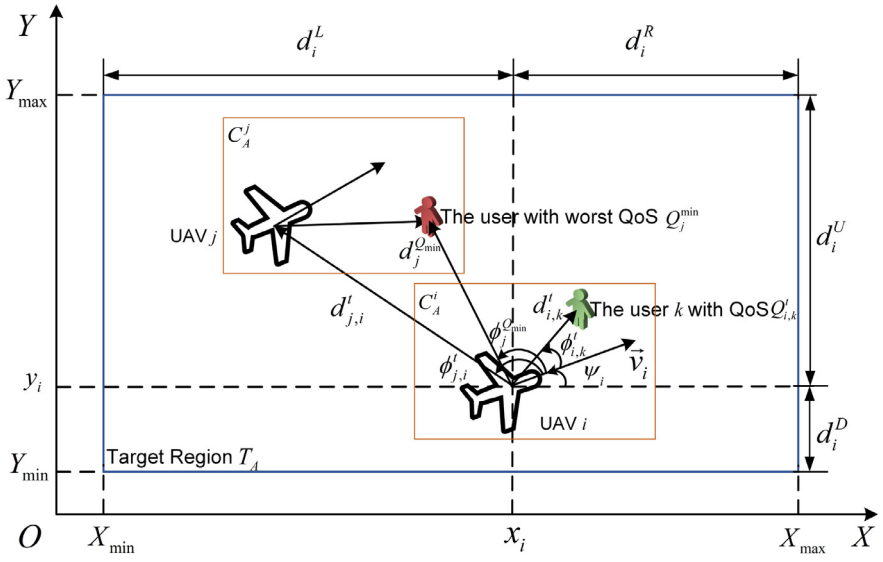


Fig. 4. Observation of UAV i.

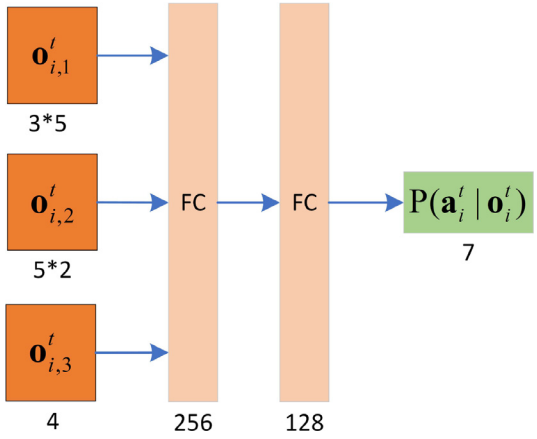


Fig. 5. Deep neural network architecture.

connected layer uses the sigmoid function as the activation function, limiting the output to (0,1), and the remaining layers use the ReLU function as the activation function. The final selected action ψ_{ic} is obtained by sampling the network's output $P(\mathbf{a}_i^t|\mathbf{o}_i^t)$.

4.5. Reward function

Our objective is to provide users with a fair and efficient communication service by maximizing the communication QoS of the user with the worst QoS. At the same time, the UAV is not allowed to leave the communication network and fly out of the target region. To achieve the above goals, the reward obtained by UAV i at time t is designed as

$$r_i^t = r_{i,Q}^t + r_{i,net}^t + r_{i,ser}^t \tag{13}$$

where $r_{i,Q}^t$ is the reward related to the users' communication QoS, $r_{i,net}^t$ is the reward related to the UAVs' communication network, and $r_{i,ser}^t$ is the reward related to the UAV's position relative to the target region.

By considering the average, $\bar{Q}_t$, and the standard deviation, Q_t^{std}, of the communication QoS of all users at time t, the reward can be designed as

$$r_{i,Q}^t = 10{*}N * ((\bar{Q}_t - \bar{Q}_{t-1}) - (Q_t^{std} - Q_{t-1}^{std})) \tag{14}$$

where $\bar{Q}_t = \frac{1}{M}\sum_{k=1}^{M} Q_k^t$ and $Q_t^{std} = \sqrt{\frac{1}{M}\sum_{k=1}^{M}(Q_k^t - \bar{Q}_t)}$.

When UAV i leaves the UAVs' communication network, the reward it receives is

$$r_{i,net}^t = \frac{d_{com} - \max(d_{i,j}^t)}{\sqrt{(X_{max} - X_{min})^2 + (Y_{max} - Y_{min})^2}}, \quad j = 1, 2, \ldots, N, i \neq j \tag{15}$$

When UAV i flies out of the target area, the reward it receives is

$$r_{i,ser}^t = \frac{\min(0, d_i^U) + \min(0, d_i^D) + \min(0, d_i^L) + \min(0, d_i^R)}{\sqrt{(X_{max} - X_{min})^2 + (Y_{max} - Y_{min})^2}} \tag{16}$$

4.6. Training algorithm

A large number of DRL algorithms have been developed and have found a wide range of applications [42,43]. In this study, PPO, a policy-based DRL algorithm, is used to train the deep neural network owing to its benefits of optimizing control policies with guaranteed monotonic improvement and high sampling efficiency. In the training process, a framework of centralized training and decentralized execution is employed, where each UAV independently observes environment information and executes actions, and then the experience collected by all the UAVs is used to train the network.

As shown in Algorithm 1 , network training comprises two processes: the experience collecting process and policy updating process. In the first process, each UAV uses the shared policy to select the action and collects experience until it reaches the maximum time step T. In the second process, the policy network π_θ is optimized E_π times with loss function $L^{CLIP}(\theta)$ and the state value network V_ϕ is optimized E_V times with loss function $L^V(\phi)$ on the same minibatch data sampled from the collected experience. The network structure of the state value network V_ϕ is the same as the policy network π_θ except that there is only one output value in the last layer. The optimization tool used is Adam optimizer [44].

5. Simulation results and analysis

In this section, we first describe the setup and the parameters used in the simulation performed to demonstrate the generalization capability and robustness of the learned policy in different

Algorithm 1: PPO with multiple UAVs providing communication services.

Initialize policy network π_θ and value function V_ϕ, let $\pi_{\theta'} = \pi_\theta$
for *episode = 1,2,...,* **do**
  for *step = 1,2,...,T* **do**
    for *UAV i = 1, 2, N* **do**
      run policy $\pi_{\theta'}$, collecting experience $\{\mathbf{o}_i^t, r_i^t, \mathbf{a}_i^t\}$
      Estimate advantages using
      $\hat{A}_i^t = \delta_i^t + (\gamma\lambda)\delta_i^{t+1} + \ldots + \ldots + (\gamma\lambda)^{T-t+1}\delta_i^{T-1}$
      where $\delta_i^t = r_i^t + \gamma V(\mathbf{o}_i^{t+1}) - V(\mathbf{o}_i^t)$
    end
  end
  for *j = 1,2,...,E_π* **do**
    $L^{CLIP}(\theta) =$
    $-\hat{\mathbb{E}}_t\left[\min\left(r_i^t(\theta)\hat{A}_i^t, \text{clip}\left(r_i^t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i^t\right)\right]$
    $r_i^t(\theta) = \frac{\pi_\theta(\mathbf{a}_i^t|\mathbf{o}_i^t)}{\pi_{\theta'}(\mathbf{a}_i^t|\mathbf{o}_i^t)}$
    Optimize surrogate $L^{CLIP}(\theta)$ wrt θ, with E_π epochs, minibatch size $B \le NT$ and the learning rate $l_{r\theta}$.
    (Note: clip$(x, x_{\min}, x_{\max})$ limits the value of x between $x_{\min}$ and $x_{\max}$)
  end
  $\theta' \leftarrow \theta$
  for *k = 1,2,...,E_V* **do**
    $L^V(\phi) = \sum_{i=1}^{N}\sum_{t=1}^{T}\left(\sum_{t'\ge t}\gamma^{t'-t}(r_i^{t'} - \bar{r}) - V_\phi(\mathbf{o}_i^t)\right)^2$
    Optimize surrogate $L^V(\phi)$ wrt ϕ, with E_V epochs, minibatch size $B \le NT$ and the learning rate $l_{r\phi}$
  end
end

Table 1
Training parameters in Algorithm 1.

Parameters	Values
T	500
N	2,3,4,5
γ	0.99
λ	0.95
E_π	10
ϵ	0.1
B	64
$l_{r\theta}$	2e−5
E_V	10
$l_{r\phi}$	1e−3

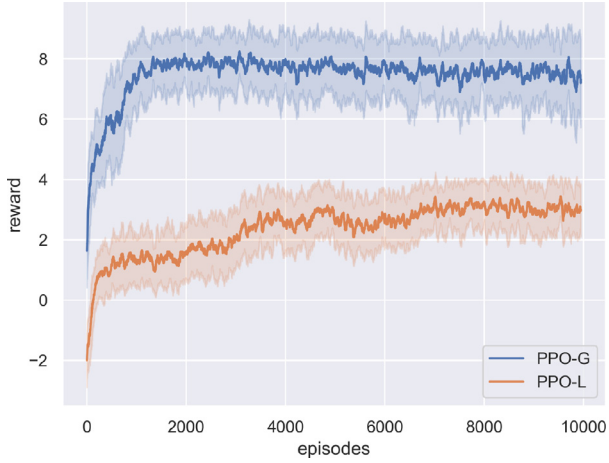


Fig. 6. Curves of the average and variance of the rewards obtained from the training episodes. PPO-G represents the policy with global service capability and PPO-L represents the policy with local service capability.

scenarios. Finally, we compare our policy with other methods in several scenarios.

5.1. Simulation setup and training results

In the training process, we consider an environment of size 500 m × 500 m × 150 m with a communication range d_{com} of 200 m. The UAV's communication service area C_A is set as a square of size 100 m × 100 m. Each UAV has the same speed, i.e., V_i = 10 m/s ($i = 1, 2, \ldots N$). For avoiding collisions, each UAV flies at different altitudes ranging 100–150 m. The maximum lateral overload limit of the UAV is set to $n_{\max} = 1$. All the users have the same speed, which is set to $V_{user}^k = 1$ m/s. For communication channel parameters, we set $\beta_0 = -60$ dB, $\sigma_0^2 = -110$ dB m, and $P_{\max} = 0.1$ W. The simulation step is set to $\Delta t = 0.5$ s.

The deep neural networks are implemented with Pytorch [45] in Python. The parameters in Algorithm 1 are listed in Table 1.

We train two different policies, one with global service capability and the other with local service capability. The results with global communication capabilities are provided as a baseline comparison for the results with local communication capabilities. On the one hand, it shows the results that each method can achieve when global communication service capabilities are available; on the other hand, it reflects the impact on user communication QoS when only local communication service capabilities are available by comparing with results with global communication capabilities. When the policy has global service capability, the communication distance between UAVs is not limited, and each UAV can obtain state information of all users in the environment. When the policy has local service capability, the communication distance between UAVs is set to 200 m, and each UAV can only obtain state information of all users within its service area. Both policies are trained using 10,000 episodes. At the beginning of each training episode, the numbers of UAVs, users, and obstacles are randomly selected from [2,5], [5, 20], and [0, 5], respectively. In addition, the positions of the UAVs, obstacles, and users are randomly reset. The obstacles are modeled as cylinders with a height ranging 20–80 m and a constant radius of 15 m. We record the average and variance of the reward every 50 episodes, where the reward is computed as the sum of the rewards received in each episode. The results shown in Fig. 6 indicate that the cumulative rewards of both policy training keep increasing and converge at the end of the training, implying that both trained control policies enable the UAVs to provide better communication services for users in the training environments and receive stable rewards. As we expected, the UAVs with global service capabilities can receive more rewards, indicating that they can provide better services to users because they can determine which users have poor communication QoS and thus provide communication services to those users.

5.2. Results with trained policy in different scenarios

In this subsection, we use the trained control policies in different scenarios to verify their performance. We conduct 100 random experiments for each test case and calculate the average values of the average and variance of the communication QoS of all users. The results obtained using the trained policies are shown in Fig. 7. Fig. 7(a) shows the results with varying number of users. We divide the test experiment into six test cases (three each for policies with global and local service capabilities), where the number of UAVs is fixed at 3 and the number of users is 5, 10, and 20, respectively. The horizontal axis shows the test cases, and the

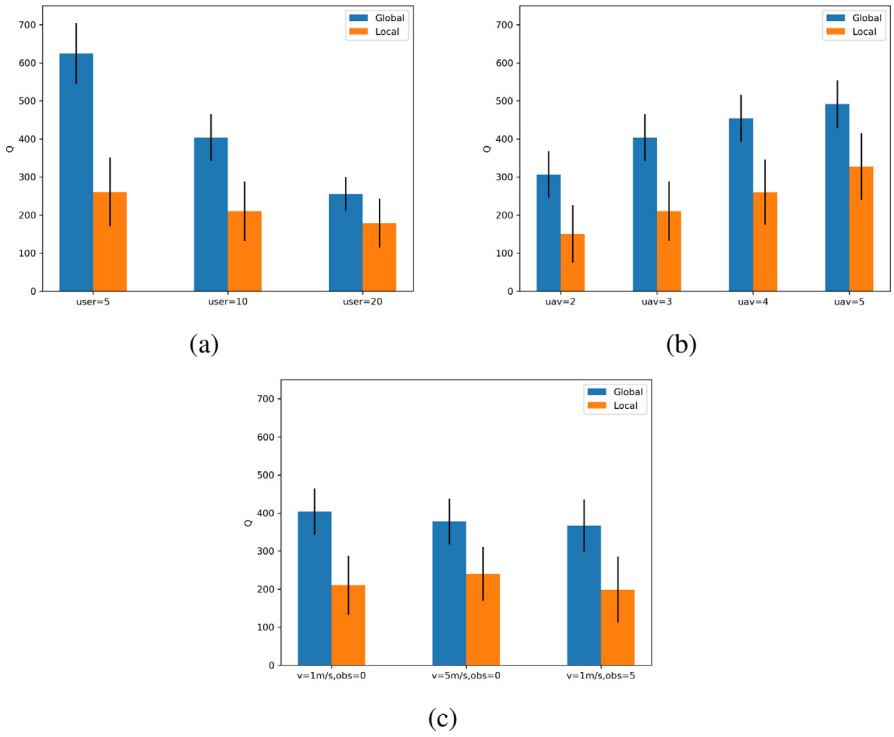


Fig. 7. The communication QoS of users in different scenarios when the trained policies are used. (a) The results with varying number of users where the number of UAVs is fixed at 3. (b) The results with varying number of UAVs where the number of users is fixed at 10. (c) The results with varying environment, i.e., the speed of users' movement increases and the number of obstacles increases.

vertical axis shows the users' communication QoS. All subsequent results are displayed in the same way. Fig. 7(a) shows that when the number of users increases, the average communication QoS of users decreases, which is in line with our expectations. In addition, the average communication QoS of users with local service capability is significantly less than that with global communication service capability, and the variance of the communication QoS of users with local service capability is larger than that with global communication service capability. This is because the UAVs have limited communication range and local state information of users and cannot comprehensively consider the communication QoS of all users. This phenomenon can also be seen in Fig. 7(b) and (c).

Fig. 7 (b) shows the results with varying number of UAVs. We divide the test experiment into eight test cases (four each for policies with global and local service capabilities), where the number of users is fixed at 10 and the number of UAVs is 2, 3, 4, and 5, respectively. First, we observe that as the number of UAVs increases, the average communication QoS of users gradually improves, which is inevitable. It is noteworthy that as the number of UAVs increases, the rate at which users' communication QoS improves decreases. This is because of the following two reasons. First, multiple UAVs can already cover all users and meet the needs of users; further increasing the number of UAVs does not result in more improvement. Second, by contrast, increasing number of UAVs will lead to increased communication interference between them, which reduces the service quality of the UAVs for users. Moreover, comparison of the results with global and local service capabilities indicates that when the number of UAVs is 2 and they have global service capability, the communication QoS of the users is the same as when the number of UAVs is 5 and they have local service capability. This suggests that increasing the number of UAVs can compensate for the lack of communication service capabilities of UAVs.

Fig. 7 (c) shows the results with varying environment. We consider two environment changes: (1) the speed of users' movement increases, and (2) obstacles exist in the environment that may obstruct the communication between UAVs and users. In the first case, the user movement speed is increased to 5 m/s. In the second case, five obstacles are considered in the environment. As shown in Fig. 7(c), in the six test cases (three each for policies with global and local service capabilities), the users' communication QoS is similar, implying that the trained control policies can adapt well to changes in the environment. The results also show that the trained control policies enable the UAVs with local service capability to adapt to changes in the environment and provide users with reliable communication services.

The trajectories of UAVs and the communication QoS of users in a test case are shown in Figs. 8 and 9, respectively. In this test case, the number of UAVs is 3, the number of users is 10, the user movement speed is 1 m/s, and no obstacles are present. Fig. 8 shows the results with global service capability. We can see that UAVs can serve different users separately, while giving priority to users with poor communication QoS, thus ensuring the simultaneous improvement of the users' communication QoS. Fig. 9 shows the results with local service capability. We can observe the following three phenomena:

(1) When the UAVs serve users, due to the limitation of the communication range, the distance between the UAVs must be maintained within the communication range, which reduces the UAVs' exploration abilities and makes it difficult for them to find new users.
(2) UAVs have a certain ability to explore new users, which can be seen from the trajectories of the UAVs at $t = 50$ s, 100 s, and 150 s; this allows the UAVs to explore new users even when some users have been discovered, ensuring that other users in the environment are also served.
(3) When there are more concentrated users in a certain area, the UAVs will provide more services in this area, which can be seen by the trajectories of the UAVs near users U4, U5, and U7.

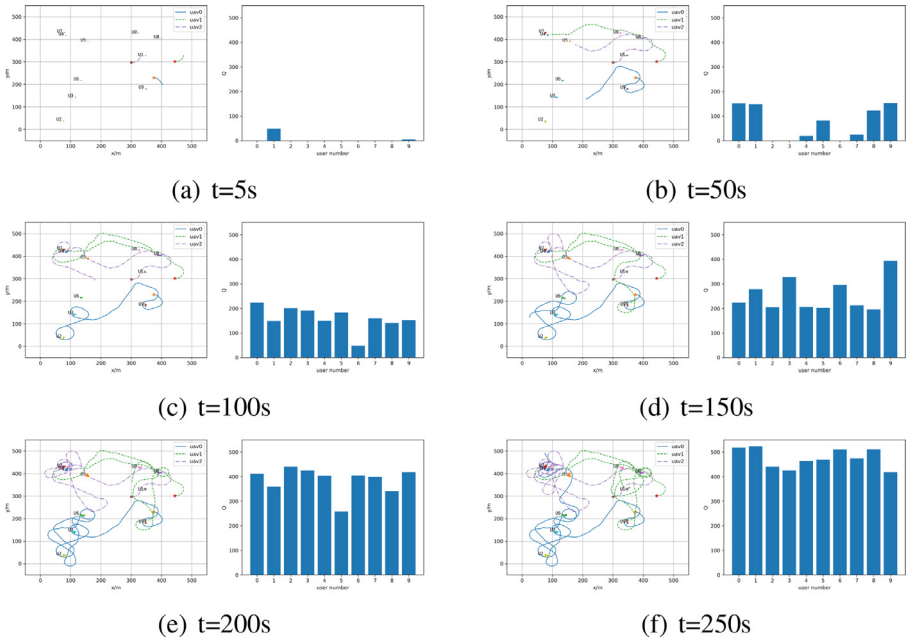


Fig. 8. Trajectories of UAVs and the communication QoS of users with global service capability.

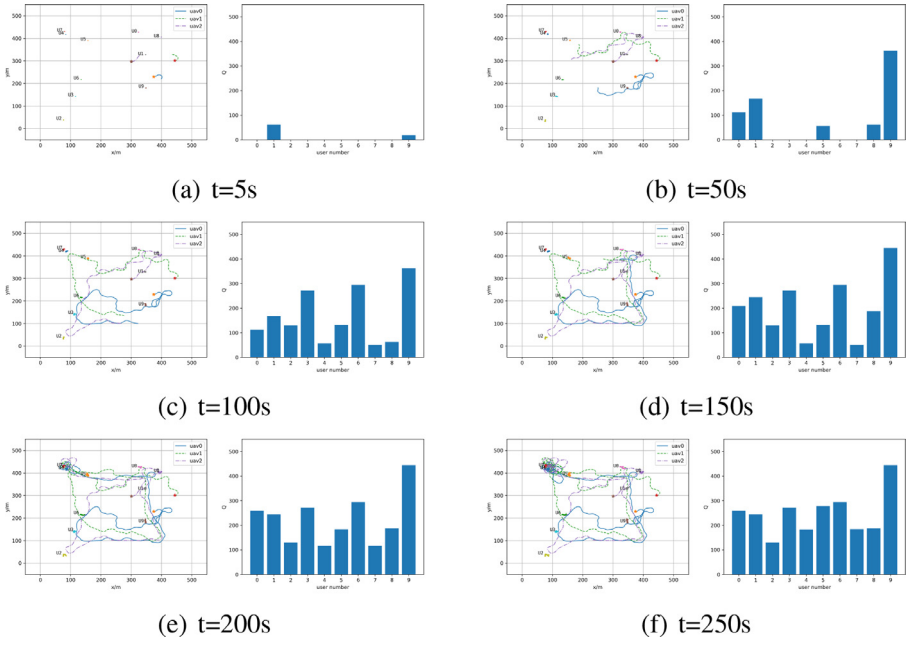


Fig. 9. Trajectories of UAVs and the communication QoS of users with local service capability.

The above results demonstrate that our trained control policies can adapt to different scenarios, has good generalization ability and robustness, and responds resiliently to changes in the environment, providing guaranteed communication services to users.

5.3. Comparison results with other algorithms

In this subsection, we compare our trained control policies with three different algorithms, namely the deep *Q*-learning network (DQN) [18], which is a value-based deep reinforcement learning, a consensus-based bundle algorithm (CBBA) [46], and the distributed Hungarian method (DHA) [47]. The parameters of DQN are listed in Table 2. Firstly, the comparison experiments are conducted in eight different test scenarios with global and local service capabilities. Global communication service capabilities are provided as a benchmark for the results with local communication service capabilities. In each test scenario, we conducted 100 random experiments and calculated the average values of the average and vari-

Table 2
Training parameters of the DQN.

Parameter	Value
Batch size	64
Replay memory size	100,000
Discount factor	0.99
Learning rate	5e−4
Number of UAVs	2,3,4,5
Max step	500

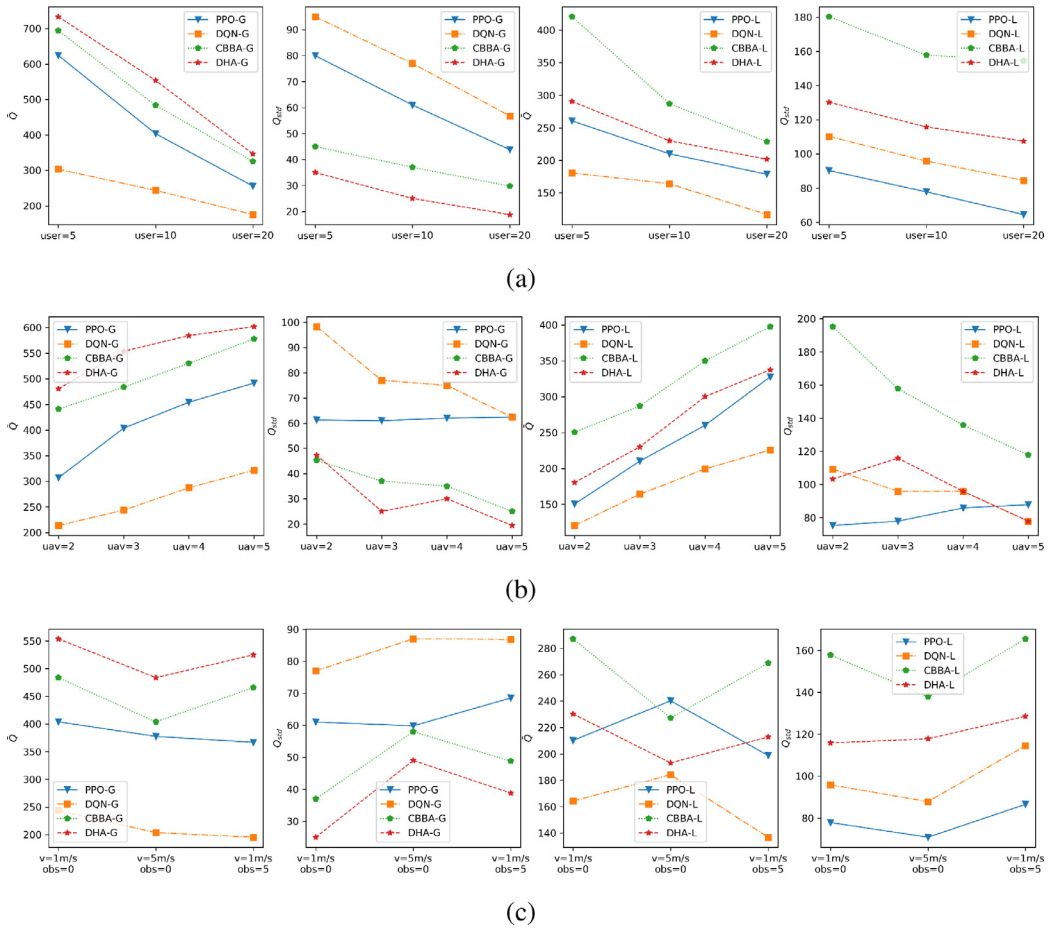


Fig. 10. Comparison results between PPO, DQN, CBBA and DHA in different scenarios. (a) The results with varying number of users where the number of UAVs is fixed at 3. (b) The results with varying number of UAVs where the number of users is fixed at 10. (c) The results with varying environment, i.e., the speed of users' movement increases and the number of obstacles increases. $\bar{Q}$ represents the average communication QoS of all users and Q_{std} represents the variance of the communication QoS of all users.

ance of all users' communication QoS. The comparison results are shown in Fig. 10.

As shown in Fig. 10, the following two conclusions can be drawn:

(1) As shown in the first and second columns in Fig. 10(a)–(c), in the case of the UAVs with global service capabilities, the communication QoS using the policy trained with the DHA-G is the best, followed by the CBBA-G and PPO-G. The policy trained with the DQN-G has the worst communication QoS. The policy trained with the DHA-G has the highest average communication QoS and the minimum variance of communication QoS. The results of DQN-G are opposite to those of DHA-G, and the results of CBBA-G and PPO-G are in between those of DHA-G and DQN-G. This is because, given the global information, the DHA-G and CBBA-G can accurately place the UAVs near the user who needs communication services and they do not use the process of randomly searching for users; thus, they achieve similar results. However, because the policies trained with the PPO-G and DQN-G use deep neural networks, there is certain randomness in their outputs and in the process of providing communication services to users; they randomly search for users, which reduces their communication QoS. In addition, the results indicate that the performance of the policy trained with PPO-G is better than that of the policy trained with the DQN-G.

(2) As shown in the third and fourth columns in Fig. 10(a)–(c), in the case of the UAVs with local service capabilities, the communication QoS using the policy trained with PPO-L and the DHA-L is similar. Both being better than the QoS obtained using the policy trained with the other two methods. In most cases, the average of the users' communication QoS using the policy trained with the DHA-L is greater than using the policy trained with PPO-L. However, the variance of users' communication QoS using the policy trained with the DHA-L is also greater than using the policy trained with PPO-L. The main reason is that the policy trained with PPO-L learns during training to strike a direct balance between exploiting known information and exploring unknown information, which gives the policy a strategy to find unknown users, allowing it to serve unexplored users in the target area and provide fair communication services to all users. In particular, as shown in the third column in Fig. 10(c), when the user's speed is 5m/s, the policy trained with PPO-L has the highest average communication QoS and the minimum variance of communication QoS compared with other methods, which shows the policy trained with PPO-L can adapt to environments where the user's location changes rapidly, demonstrating its robustness to dynamic changes in the environment. The average of the users' communication QoS using the policy trained with the CBBA-L is the largest among all the compared methods; however, the variance of users' communication QoS is also the largest. This is because the CBBA-L does not have a good mechanism for randomly searching for users and can only provide communication services to users within the local communication service area's scope. This will result in some users being well served while others not being

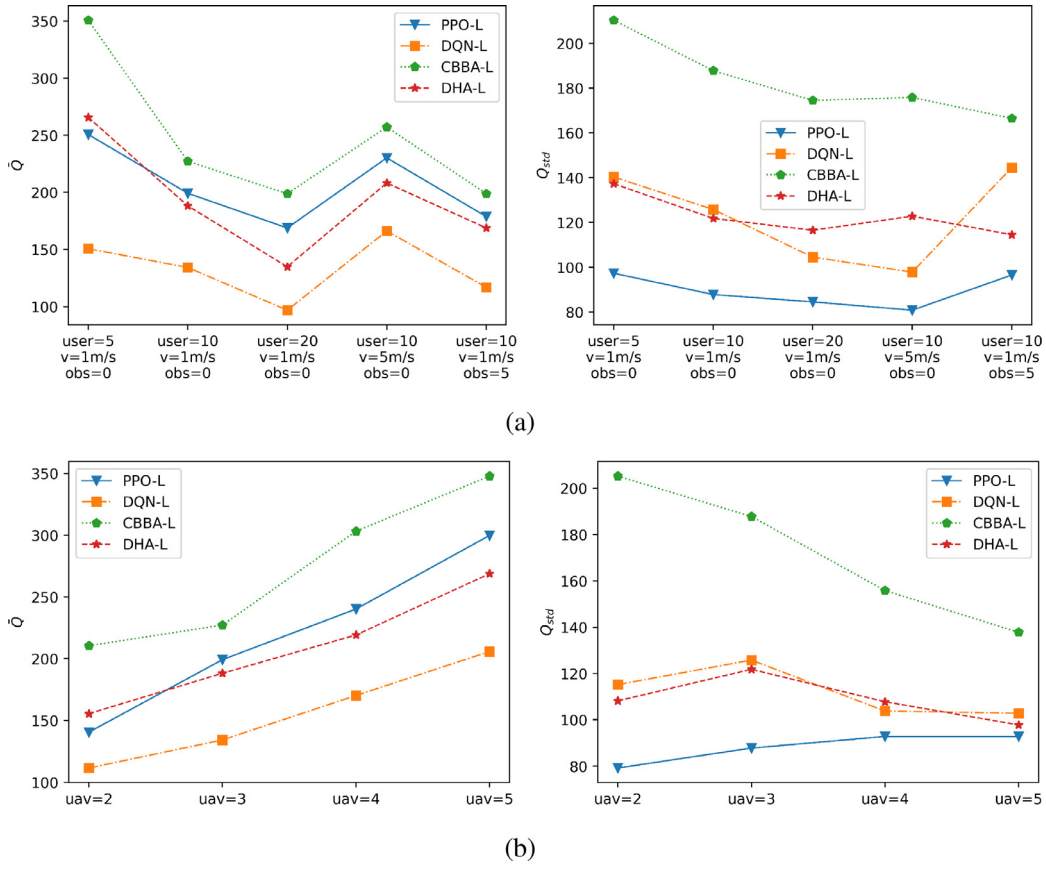


Fig. 11. Comparison results between PPO, DQN, CBBA, and DHA in an environment with environmental disturbances. The UAVs have local communication service capability. (a) The results with different environments where the number of UAVs is fixed at 3. (b) The results with varying UAVs where the number of users is fixed at 10, the user's speed is 1m/s. $\bar{Q}$ represents the average communication QoS of all users and Q_{std} represents the variance of all users' communication QoS.

served. The communication QoS using the policy trained with the DQN-L is the worst. The results indicate that a stochastic policy has certain advantages compared with a deterministic policy on the problem of multiple UAVs providing users with communication services, regardless of UAVs' communication service capabilities.

Besides, UAVs usually have only local communication service capabilities in practical applications due to energy consumption and economic cost. At the same time, the UAVs will be disturbed by various environmental disturbances, such as unstable communication, malfunctioning sensing system and positioning system, which will lead to a large error in the UAV's estimation of the user's and other UAV's states, which will cause an impact on the user's communication service. To measure the robustness of the above methods to such environmental disturbances, we simulate such environmental disturbances by adding a small period of disturbance error to each UAV's observations at regular intervals. Specifically, we add Gaussian noise of duration 10 s to the UAV's observations $\mathbf{o}_{i,1}^t$ and $\mathbf{o}_{i,2}^t$ at $t = 50$ s, $t = 100$ s, $t = 150$ s, and $t = 200$ s, respectively, i.e., $\mathbf{o}_{i,1}^t = \mathbf{o}_{i,1}^t + 0.5\mathcal{N}(0,1)$, $\mathbf{o}_{i,2}^t = \mathbf{o}_{i,2}^t + 0.5\mathcal{N}(0,1)$. (Note: The UAV's observations are normalized to be in the interval $[-1,1]$) The comparison results are shown in Fig. 11.

As shown in Fig. 11(a) and (b), in most cases, the policy trained with PPO-L has a higher average communication QoS than that trained with DHA-L and the minimum variance of communication QoS compared with other policies. The policies trained with CBBA-L and DQN-L have obvious shortcomings with the maximum variance of communication QoS and the minimum average communication QoS, respectively. This demonstrates that the policy trained with PPO-L has strong robustness to environmental disturbances and can provide guaranteed resilient communication services to users. An important reason is that the disturbance of the observations will have a large impact on the planning results of the DHA-L, making the planning results deviate from the actual situation. Besides, since the policy trained with PPO-L is a deep neural network trained in various environments, it has better robustness to disturbing observations. Also, as shown in the first column in Fig. 11(a), when the number of the users is 5, the user's speed is 1m/s and there is no obstacle in environment, the policy trained with the DHA-L has a higher average communication QoS than that trained with PPO-L. This is because that when the number of users is small, the interference added to the users' states has less impact on the communication service of the UAV. Similarly, the first column of Fig. 11(b) shows similar results. When the number of UAVs is 2, the environmental disturbances have less impact on the UAV communication service, so the policy trained with the DHA-L has a higher average communication QoS than PPO-L.

5.4. Discussions

Although UAVs with global communication service capability can provide high-quality communication services, equipping them with better communication equipment will lead to increased weight of the UAV and consume more energy during the communication process. At the same time, equipping communication equipment with global communication service capability will increase the cost of UAVs. Therefore, in practice, the selection of UAVs with different communication service capabilities requires a trade-off between communication service performance and cost. The use of

UAVs will depend on the lesser of the cost due to the increased number of UAVs or the cost due to the use of global communication equipment.

In practical applications, the most common case is to use UAVs with local communication service capability to provide communication services to users over a large area. In such cases, the policy trained with PPO-L has a significant advantage over other methods (DQN-L, CBBA-L, DHA-L). The policy trained with DQN-L obviously performs less well than PPO-L with a lower average communication QoS and a higher communication QoS variance. Besides, the policy trained with CBBA-L has the highest average communication QoS with the highest communication QoS variance, leading to the fairness of user communication QoS is difficult to guarantee. Compared with the policy trained with PPO-L, the DHA-L policy has a higher average communication QoS. Still, it also has a higher communication QoS variance, so user communication QoS fairness is lower. Moreover, the policy trained with PPO-L has a higher average communication QoS and the minimum communication QoS variance compared to other policies when UAVs are disturbed by interference factors in the environment. So it is more robust to dynamic interference factors in the environment. For these reasons, it is worthwhile to use the policy trained with PPO-L when the UAV has local communication service capability to provide resilient and guaranteed communication QoS to users.

6. Conclusion

In this paper, we investigated the optimization control problem of multiple UAVs for providing users with resilient and guaranteed communication services. First, we modeled the problem as a multi-UAV optimal control problem with the objective of maximizing the communication QoS for the user with the worst QoS. Then, considering the complexity of this optimization problem owing to the dynamic changes in the environment, we formulated the control of each UAV as a POMDP and solved it using a policy-based DRL method, PPO, where each UAV shares the same control policy and takes action independently based on its observed information. In the training process, the centralized training and decentralized execution paradigm was used, where the experience collected by all UAVs was used to train the shared control policy. In addition, to provide efficient and fair communication services to users, the average and variance of the communication QoS of all users were considered in the design of the reward function. We conducted extensive simulations to evaluate the performance of the proposed algorithm. Simulation results show that the policies trained with PPO can adapt to different scenarios and provide resilient and guaranteed communication QoS to users. Moreover, the results show better performance than the policies trained with the DQN, CBBA, and DHA when the UAVs have local communication service capability. In the future, we will extend this work to heterogeneous UAVs and study how UAVs with different capabilities can collaboratively provide communication services to users.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, A. Sheth, Machine learning for internet of things data analysis: a survey, Digit. Commun. Netw. 4 (2018) 161–175.

[2] G. Nguyen, S. Dlugolinsky, M. Bobk, V. Tran, Ã. Garca, I. Heredia, P. Malk, L. Hluch, Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey, Artif. Intell. Rev. 52 (2019) 77–124.

[3] H. Yan, Coclustering of multidimensional big data—A useful tool for genomic, financial, and other data analysis, IEEE Syst. Man Cybern. Mag. (2017) 23–30.

[4] L. Bai, L. Cui, Y. Wang, Y. Jiao, E. Hancock, A quantum-inspired entropic kernel for multiple financial time series analysis (2020) 4453–4460.

[5] Y. Huang, G. Kou, A kernel entropy manifold learning approach for financial data analysis, Decis. Support Syst. 64 (2014) 31–42.

[6] L. Cui, L. Bai, L. Rossi, z. Zhang, Y. Jiao, E. Hancock, A preliminary survey of analyzing dynamic time-varying financial networks using graph kernels, Struct., Syntactic, Stat. Pattern Recognit. (2018) 237–247.

[7] L. Cui, L. Bai, Y. Wang, X. Jin, E. Hancock, Internet financing credit risk evaluation using multiple structural interacting elastic net feature selection, Pattern Recognit. 114 (2021) 107835.

[8] K. Stockinger, N. Bundi, J. Heitz, W. Breymann, Scalable architecture for big data financial analytics: user-defined functions vs. SQL, J. Big Data 6 (2019) 46.

[9] C. Jabbour, A. Jabbour, J. Sarkis, M. Filho, Unlocking the circular economy through new business models based on large-scale data: an integrative framework and research agenda, Technol. Forecast. Soc. Change 144 (2019) 546–552.

[10] N. Fikri, M. Rida, N. Abghour, K. Moussaid, A. Omri, An adaptive and real-time based architecture for financial data integration, J. Big Data 6 (2019) 97.

[11] H. Shakhatreh, A.H. Sawalmeh, A. Al-Fuqaha, Z. Dou, E. Almaita, I. Khalil, N.S. Othman, A. Khreishah, M. Guizani, Unmanned aerial vehicles (UAVs): a survey on civil applications and key research challenges, IEEE Access 7 (2019) 48572–48634.

[12] A. Sargolzaei, A. Abbaspour, C.D. Crane, Control of cooperative unmanned aerial vehicles: review of applications, challenges, and algorithms, in: Optimization, Learning, and Control for Interdependent Complex Networks, Springer, 2020, pp. 229–255.

[13] R. Santos, X.M. Pardo, X.R. Fdez-Vidal, Scene wireframes sketching for unmanned aerial vehicles, Pattern Recognit. 86 (2019) 354–367.

[14] J. Ren, X. Jiang, A three-step classification framework to handle complex data distribution for radar UAV detection, Pattern Recognit. 111 (2020) 107709.

[15] J. Ren, X. Jiang, Regularized 2-D complex-log spectral analysis and subspace reliability analysis of micro-doppler signature for UAV detection, Pattern Recognit. 69 (2017) 225–237.

[16] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F.R. Yu, K.-K. Wong, UAV-assisted emergency networks in disasters, IEEE Wirel. Commun. 26 (1) (2019) 45–51.

[17] C.H. Liu, Z. Chen, J. Tang, J. Xu, C. Piao, Energy-efficient UAV control for effective and fair communication coverage: a deep reinforcement learning approach, IEEE J. Sel. Areas Commun. 36 (9) (2018) 2059–2070.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

[19] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.

[20] W. Sihang, W. Jiapeng, M. Weihong, J. Lianwen, Precise detection of Chinese characters in historical documents with deep reinforcement learning, Pattern Recognit. 107 (2020) 107503.

[21] Z. Teng, B. Zhang, J. Fan, Three-step action search networks with deep Q-learning for real-time object tracking, Pattern Recognit. 101 (2020) 107188.

[22] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347(2017).

[23] H. Lu, X. Wei, H. Qian, M. Chen, A cost-efficient elastic UAV relay network construction method with guaranteed QoS, Ad Hoc Netw. 107 (2020) 102219.

[24] F. Lyu, P. Yang, W. Shi, H. Wu, W. Wu, N. Cheng, X.S. Shen, Online UAV scheduling towards throughput QoSguarantee for dynamic IoVs, in: ICC 2019-2019 IEEE International Conference on Communications (ICC), IEEE, 2019, pp. 1–6.

[25] W. Lin, Y. Zhou, H. Xu, J. Yan, M. Xu, J. Wu, Z. Liu, A tube-and-droplet-based approach for representing and analyzing motion trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 39 (8) (2016) 1489–1503.

[26] F. Cheng, S. Zhang, Z. Li, Y. Chen, N. Zhao, F.R. Yu, V.C.M. Leung, UAV trajectory optimization for data offloading at the edge of multiple cells, IEEE Trans. Veh. Technol. 67 (7) (2018) 6732–6736.

[27] M. Samir, M. Chraiti, C. Assi, A. Ghrayeb, Joint optimization of UAV trajectory and radio resource allocation for drive-thru vehicular networks, in: 2019 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2019, pp. 1–6.

[28] A. Bejaoui, K.-H. Park, M.-S. Alouini, A QoS-oriented trajectory optimization in swarming unmanned-aerial-vehicles communications, IEEE Wirel. Commun. Lett. 9 (6) (2020) 791–794.

[29] B. Perabathini, K. Tummuri, A. Agrawal, V.S. Varma, Efficient 3D placement of UAVs with QoS assurance in ad hoc wireless networks, in: 2019 28th International Conference on Computer Communication and Networks (ICCCN), IEEE, 2019, pp. 1–6.

[30] X. Li, H. Yao, J. Wang, S. Wu, C. Jiang, Y. Qian, Rechargeable multi-UAV aided seamless coverage for QoS-guaranteed iot networks, IEEE Internet Things J. 6 (6) (2019) 10902–10914.

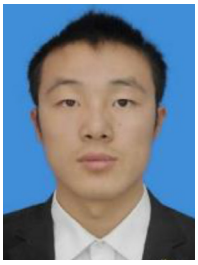
[31] S. Roth, A. Kariminezhad, A. Sezgin, Base-stations up in the air: multi-UAV trajectory control for min-rate maximization in uplink C-RAN, in: ICC 2019-2019 IEEE International Conference on Communications (ICC), IEEE, 2019, pp. 1–6.

[32] J. Hu, H. Zhang, L. Song, Z. Han, H.V. Poor, Reinforcement learning for a cellular internet of UAVs: protocol design, trajectory control, and resource management, IEEE Wirel. Commun. 27 (1) (2020) 116–123.

[33] A.M. Koushik, F. Hu, S. Kumar, Deep Q-learning-based node positioning for throughput-optimal communications in dynamic UAV swarm network, IEEE Trans. Cogn. Commun. Netw. 5 (3) (2019) 554–566.

[34] H. Bayerlein, P. De Kerret, D. Gesbert, Trajectory optimization for autonomous flying base station via reinforcement learning, in: 2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), IEEE, 2018, pp. 1–5.
[35] R. Ghanavi, E. Kalantari, M. Sabbaghian, H. Yanikomeroglu, A. Yongacoglu, Efficient 3D aerial base station placement considering users mobility by reinforcement learning, in: 2018 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2018, pp. 1–6.
[36] F. Wu, H. Zhang, J. Wu, L. Song, Trajectory design for overlay UAV-to-device communications by deep reinforcement learning, in: 2019 IEEE Global Communications Conference (GLOBECOM), IEEE, 2019, pp. 1–6.
[37] J. Hu, H. Zhang, L. Song, Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol, IEEE Internet Things J. 6 (4) (2018) 6177–6189.
[38] J. Cui, Y. Liu, A. Nallanathan, The application of multi-agent reinforcement learning in UAV networks, in: 2019 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2019, pp. 1–6.
[39] F. Wu, H. Zhang, J. Wu, L. Song, Cellular UAV-to-device communications: trajectory design and mode selection by multi-agent deep reinforcement learning, IEEE Trans. Commun. 68 (7) (2020) 4175–4189.
[40] S. Salehi, J. Hassan, A. Bokani, S.A. Hoseini, S.S. Kanhere, A QoS-aware, energy-efficient trajectory optimization for UAV base stations using Q-learning, in: 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), IEEE, 2020, pp. 329–330.
[41] H. Qiu, H. Duan, A multi-objective pigeon-inspired optimization approach to UAV distributed flocking among obstacles, Inf. Sci. 509 (2020) 515–529.
[42] S. Ivanov, A. D'yakonov, Modern deep reinforcement learning algorithms, arXiv preprint arXiv:1906.10025(2019).
[43] N.C. Luong, D.T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, D.I. Kim, Applications of deep reinforcement learning in communications and networking: a survey, IEEE Commun. Surv. Tutor. 21 (4) (2019) 3133–3174.
[44] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations (ICLR), 2015.
[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, 2019, pp. 8026–8037.
[46] H.-L. Choi, L. Brunet, J.P. How, Consensus-based decentralized auctions for robust task allocation, IEEE Trans. Robot. 25 (4) (2009) 912–926.
[47] S. Chopra, G. Notarstefano, M. Rice, M. Egerstedt, A distributed version of the hungarian method for multirobot assignment, IEEE Trans. Robot. 33 (4) (2017) 932–947.

Chengchao Bai was born in Zhangjiakou, China in 1990. He received the B.S. and Ph.D. degree in aerospace engineering from Harbin Institute of Technology, China, in 2013 and 2019, respectively. Now he is the Postdoc researcher in TU Delft. He has served as a member of Youth Editorial Board of the Journal Unmanned Systems Technology. He is a committee member of the IEEE RAS Technical Committee on Multi-Robot Systems, a committee member of the CSIG (China Society of Image and Graphing) technical committee on Machine Vision and the CICC (Chinese Institute of Command and Control) technical committee on Unmanned Systems His research interests include multi-robot learning, planetary exploration, intelligent sensing and large scale resilience cooperation. He is also interested in robot intelligence and its application.



Peng Yan was born in DingXi, China in 1996. He received the B.S. degree from Harbin Institute of Technology, China. He is pursuing Ph.D degree at Harbin Institute of Technology, China. His current interests include motion planning, decision making and behavior prediction.

Xiaoqiang Yu was born in Tongliao, China in 1994. He received the B.S. degree from Harbin Institute of Technology, China. He is pursuing Ph.D degree at Harbin Institute of Technology, China. His current interests include path planning, task assignment and mission planning in uncertain environment.

Jifeng Guo was born in Xi'an, China in 1977. He received the B.S., M.S. and Ph.D. degrees in aerospace engineering from Harbin Institute of Technology, China, in 2001, 2004 and 2007, respectively. From 2007 to 2004, he served as a lecturer and associate professor with Harbin Institute of Technology. Since 2015, he has been a Professor with School of Astronautics, Harbin Institute of Technology. He is the author of two books, more than 100 articles, and more than 30 inventions. His research interests include intelligent sensing, autonomous planning, on-orbit service and collaborative control. He is a member of editor board of the journal Unmanned Systems Technology and holds ten patents.