







Packet Routing in Dynamic Multi-Hop UAV Relay Network: A Multi-Agent Learning Approach

Ruijin Ding , *Student Member, IEEE*, Jiawei Chen , *Student Member, IEEE*, Wen Wu , *Senior Member, IEEE*, Jun Liu , *Member, IEEE*, Feifei Gao , *Fellow, IEEE*, and Xuemin Shen , *Fellow, IEEE*

Abstract— The multi-hop unmanned aerial vehicle (UAV) network can serve as data relays where ground users (GUs) do not have reliable direct connections to the base station (BS). Existing works mainly focus on simple dual-hop system. In this paper, we investigate the packet routing problem in a multi-hop UAV relay network to minimize the data transmission time and enhance the network throughput. However, the dynamic network topology due to UAV mobility makes the packet routing challenging since the limited communication range of each UAV leads to volatile wireless connection. Moreover, the line-of-sight communication links may cause strong interference among UAVs. Towards this end, we propose a novel multi-agent deep reinforcement learning based algorithm, named as multi-agent QMIX (MAQMIX) to: 1) design proper UAVs' trajectories to serve the moving GUs while maintaining the network connection; 2) allocate frequency resource properly among UAVs to alleviate the impact of interference; and 3) choose a proper next hop UAV for each data packet to reduce the transmission time and probability of network congestion. The proposed MAQMIX has two novel training mechanisms, i.e., intra-UAV and inter-UAV training mechanisms, which can tackle the large action space issue and coordinate the training among UAVs in the multi-hop UAV relay network. Simulation results demonstrate that the MAQMIX outperforms baseline schemes in terms of the network congestion avoidance, throughput, and transmission time.

Index Terms—Multi-agent reinforcement learning, multi-hop UAV relay network, resource allocation, trajectory design, interference mitigation.

Manuscript received 30 December 2021; revised 22 April 2022; accepted 5 June 2022. Date of publication 14 June 2022; date of current version 19 September 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102401, in part by the National Natural Science Foundation of China under Grant 61902214, in part by Tsinghua University Initiative Scientific Research Program, in part by Guoqiang Institute, Tsinghua University, and in part by Beijing Municipal Natural Science Foundation under Grant 4212002. Part of this work has been accepted by IEEE International Conference on Communications (ICC) 2022 [1]. The review of this article was coordinated by Dr. Phone Lin. (*Corresponding author: Jun Liu.*)

Ruijin Ding, Jiawei Chen, and Feifei Gao are with the Institute for Artificial Intelligence Tsinghua University (THUAI), State Key Lab of Intelligent Technologies and Systems, Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China (e-mail: drj17@mails.tsinghua.edu.cn; chenjiaw20@mails.tsinghua.edu.cn; feifeigao@ieee.org).

Wen Wu is with the Frontier Research Center, Peng Cheng Laboratory, Shenzhen, Guangdong 518055, China (e-mail: wuw02@pcl.ac.cn).

Jun Liu is with the Institute of Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China, and also with the Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: junliu@tsinghua.edu.cn).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Digital Object Identifier 10.1109/TVT.2022.3182335

I. INTRODUCTION

RECENTLY, unmanned aerial vehicle (UAV)-assisted communication has attracted increasing attention from both industry and academia due to its multi-fold advantages [2]–[5]. UAVs can be deployed flexibly for on-demand wireless communication systems due to their high manoeuvrability [6]–[8]. UAVs can also provide line-of-sight (LOS) links to ground users (GUs), which can improve the performance of communication systems [9]–[11]. In addition, the UAVs can act as the complement of terrestrial networks to serve the heavy traffic loads in hotspot areas, such as large-scale sport or musical events [12]–[14].

Existing researches on UAV-assisted communication mainly focus on two research directions. *First*, UAVs can serve as aerial base stations (BSs) to provide efficient and reliable temporary wireless communication services for GUs in emergency situations [15]–[19]. In [15] and [16], Zeng *et al.* derive mathematical models on the propulsion power of UAVs and then optimize UAVs' trajectories to facilitate energy-efficient communication services. In [17], the orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) schemes in the UAV communication system are investigated. The minimum average rate among GUs is maximized through joint UAV trajectory design and resource allocation. In [18], an air-ground coordinated communication system is studied, where the multi-user access control policy and UAVs' trajectories are jointly optimized to improve the overall throughput while guaranteeing the fairness among GUs. *Second*, UAVs can be deployed as relays between GUs and BSs [20]–[22] or between two user groups [23]–[25]. Compared with fixed relay station, the UAV-aided relay network can provide flexible relay services for mobile GUs due to UAV manoeuvrability. In [20], the UAV acts as an amplify-and-forward relay between a BS and GUs, and the trajectory as well as the transmit power of the UAV are optimized to minimize the outage probability of the relay network. In [21], Zhang *et al.* consider a UAV-assisted decode-and-forward relay communication system in a downlink maritime communication, and optimize the placement of UAV to improve the capacity of wireless backhaul links between the BS and the UAV. The UAV's trajectory and the transmit powers of UAV and BS are jointly optimized to maximize energy efficiency [22]. In [23], the UAV trajectory and time scheduling are jointly optimized to maximize the minimum secrecy rate in the UAV-relayed wireless networks. In [24], a UAV is employed as a full-duplex relay to assist the underlaid

device-to-device (D2D) communications. The sum throughput is maximized under the transmit power budget and UAV trajectory constraints. In [25], the signal-to-interference-ratio (SIR) of the system is maximized to enhance the throughput between a transmitter and receiver pair through UAV position planning. However, the existing works on UAV-aided relay network focus on the simple dual-hop system with only one UAV [20]–[24], i.e., the transmitter-to-UAV hop and UAV-to-receiver hop.

Our work belongs to the second research direction. In this paper, we aim to facilitate communication services in the areas where GUs do not have reliable direct connections to the BS. To serve these mobile GUs, UAVs can be deployed as data relays for establishing network connections between the GUs and the BS. We investigate the packet routing problem in the multi-hop UAV relay network, in which the GUs generate data packets and then multiple UAVs transmit them to a faraway BS hop by hop. However, the packet routing in the multi-hop UAV relay network faces some challenges. The mobility of UAVs leads to the dynamic network topology. Due to the limited communication range, the connection between UAVs can be unreliable, and the candidate next hop UAVs keep changing. Also, the movement of UAVs can affect the subsequent topology of the multi-hop UAV relay network. As a result, the UAV trajectories and the next hop selection need to be well designed. In addition, since there is strong interference among UAVs due to the LOS links, frequency resource allocation should be considered to mitigate the interference. The packet routing problem in the multi-hop UAV relay network is a sequential decision-making problem, since the trajectories of UAVs, frequency resource allocation, and next hop selection should be properly designed sequentially during the flight. There are many decision variables to optimize, which makes the problem intractable.

Deep reinforcement learning (DRL) including multi-agent DRL (MADRL) is a potential solution for such sequential decision-making problem [26], [27]. The learning agent can choose actions based on the current state, and then the network environment turns into the next state and returns a reward to the agent [28]. The objective of DRL is to maximize the cumulative reward [29], which means, the subsequent effects of the current action can be considered. Due to its superior performance, DRL has been widely used in UAV-assisted communications, such as UAV trajectory design [30], [31], and the resource allocation for UAV networks [32].

In this paper, we study the packet routing problem in a multi-hop UAV relay network, where multiple UAVs transmit the data packets generated from GUs to a faraway BS hop by hop. The objective is to minimize the transmission time of the data packets while maintaining the network connection and avoiding network congestion through determining three types of decisions, i.e., UAV trajectory design, frequency resource allocation, and the next hop selection. However, there are some issues when applying MADRL to the multi-hop UAV relay network. First, each UAV needs to make the three types of decisions at the same time, leading to large action space. The total action space size is the product of all the candidates of the three types of decisions. When the network scale increases, the action space will be extremely large and thus difficult to tackle. Second, the objective of the packet routing problem is

to minimize the transmission time for each packet, which is determined by all the UAVs along the packet transmission path. However, the objective of traditional DRL or MADRL is to maximize the cumulative reward of each agent. In other words, the optimization targets of packet routing problem and traditional DRL are different. Such objective inconsistency cannot be solved by simple reward shaping mechanism. In order to tackle these issues, we propose a new MADRL algorithm, named as multi-agent QMIX (MAQMIX), which has two novel training mechanisms, intra-UAV and inter-UAV training mechanisms, to tackle the large action space issue and coordinate the training among UAVs, respectively.

The main contributions of this paper are summarized as follows:

- We formulate the packet routing problem in a multi-hop UAV relay network, in which the objective is to design UAV trajectories, allocate frequency resource, and select the next hop UAV to minimize the transmission time of the data packets while maintaining the network connection and avoiding network congestion through determining three types of decisions.
- We design an intra-UAV training mechanism where each UAV is decomposed into three subagents responsible for UAV trajectory design, frequency resource allocation, and the next hop selection, respectively. Each UAV runs the multi-agent DRL algorithm, QMIX [33], to train the three subagents. The decomposition of UAV agent can reduce the action space significantly, thereby simplifying the training process.
- We further design a novel inter-UAV training mechanism, where each UAV agent uses the state-action value of the next hop UAV to update its own value function, such that the value function of each UAV agent can estimate the remaining transmission time required for each packet to be transmitted to the destination. With the inter-UAV training mechanism, the training among UAVs can be coordinated.

The remainder of this paper is organized as follows. Section II introduces the system model in detail and presents the formulated optimization problem. Section IV presents the details of the proposed MAQMIX algorithm. Simulation results are provided in Section V, followed by the conclusion in Section VI.

Notations: Scalars are denoted by lower-case letters, vectors are denoted by boldface letters, and matrixes are denoted by capital letters. $A_{i,j}$ denotes the element in the i -th row and j -th column of matrix A . The Euclidean norm of vector \mathbf{a} is denoted by $\|\mathbf{a}\|$. \mathbb{R}^M denotes the space of M -dimensional real vectors. $\mathbb{E}_\pi[\cdot]$ denotes the expectation of a random variable following policy π .

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we present the network model, the transmission model between a GU and a UAV, the transmission model between UAVs, and then formulate the problem.

A. Network Model

As shown in Fig. 1, we consider a multi-hop UAV relay network, in which GUs do not have reliable direct connections

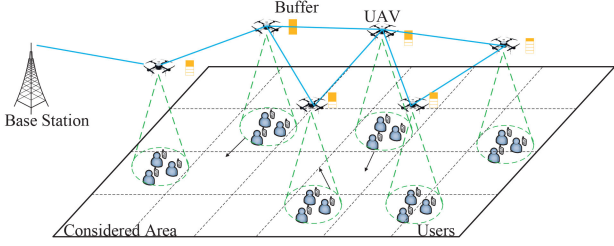


Fig. 1. A UAV relay network. Each GU transmits data packets to a selected UAV, and the UAVs formulate a multihop network to forward the data packets to the BS.

to a BS, and multiple UAVs act as the relays between GUs and the BS. The multi-hop UAV relay network consists of M UAVs, whose index set is denoted by $\mathcal{M} = \{1, 2, \dots, M\}$. We consider a square area with a side length of L_s km, and the BS is located at the edge of the considered area. The UAV M hovers near the BS and keeps connecting to the BS, which is the destination UAV of the multi-hop UAV relay network. All the UAVs fly at an altitude H with fixed speed V^u , and the horizontal location of UAV m at time t is denoted by $\mathbf{w}_m^u(t) \in \mathbb{R}^{2 \times 1}$, $0 \leq t \leq T$, where the superscript ‘ u ’ in this paper means the ‘UAV’ and T is the total service time for the multi-hop UAV relay network. There are K GUs in the network, whose index set is denoted by $\mathcal{K} = \{1, 2, \dots, K\}$, which move with speed V^g and generate data packets to be transmitted to the BS through UAV relays. The superscript ‘ g ’ means the ‘GU’. The horizontal location of GU k is denoted by $\mathbf{w}_k^g(t) \in \mathbb{R}^{2 \times 1}$. The system is assumed to work in a time slotted manner, i.e., $\mathcal{T} = \{1, 2, \dots, T\}$, and the duration of each time slot δ_t is sufficiently short such that the moving directions of UAVs and GUs keep unchanged within a time slot. The movement of the UAVs and GUs can be expressed via

$$\mathbf{w}_m^u(t+1) - \mathbf{w}_m^u(t) = V^u \delta_t \mathbf{e}_m^u(t), \forall m \in \mathcal{M}, \quad (1a)$$

$$\mathbf{w}_k^g(t+1) - \mathbf{w}_k^g(t) = V^g \delta_t \mathbf{e}_k^g(t), \forall k \in \mathcal{K}, \quad (1b)$$

where $\mathbf{e}_m^u(t)$ and $\mathbf{e}_k^g(t)$ are the directions of UAV m and GU k at time slot t . Since the locations and the directions are linear with each other, the optimization of UAV trajectories $\{\mathbf{w}_m^u(t)\}_{m \in \mathcal{M}, t \in \mathcal{T}}$ is equivalent to the optimization of UAV flying directions $\{\mathbf{e}_m^u(t)\}_{m \in \mathcal{M}, t \in \mathcal{T}}$.

B. Transmission Model Between GUs and UAVs

At the beginning of each time slot, each GU is associated to the UAV with the strongest received signal strength (RSS), which would transmit a newly generated packet. The associated UAV for GU k is denoted by $\chi_k^g(t)$ and is the source UAV of the newly generated packet in the multi-hop UAV relay network. We introduce a binary variable $\mu_{k,m}^g(t)$ to indicate whether GU k is associated to UAV m at time slot t , i.e., $\mu_{k,\chi_k^g(t)}^g(t) = 1$. Otherwise, $\mu_{k,m}^g(t) = 0, \forall m \neq \chi_k^g(t)$. Thus we have

$$\sum_{m \in \mathcal{M}} \mu_{k,m}^g(t) = 1, \quad (2)$$

and the number of GUs served by UAV m is represented by $N_m^g(t) = \sum_{k \in \mathcal{K}} \mu_{k,m}^g(t)$.

Due to the high probability of the line-of-sight (LOS) links, the air-to-ground (A2G) channel is different from the terrestrial channel. In addition to free space pathloss, the existence of buildings, trees, and other physical objects can bring excessive pathloss [34]. Taking the occurrence of LOS and non-LOS (NLOS) links into account, we adopt the probabilistic pathloss model for the communication link, where the probability of the LOS link between GU k and UAV m at time slot t can be approximated by a simple modified sigmoid function:

$$Pr_{k,m}^{LOS}(t) = \frac{1}{1 + \eta_1 \exp\left(-\eta_2 \left(\arcsin\left(\frac{H}{d_{k,m}(t)}\right) - \eta_1\right)\right)}. \quad (3)$$

Here, η_1 and η_2 are two parameters related to the communication scenario, and $d_{k,m}(t) = \sqrt{\|\mathbf{w}_m^u(t) - \mathbf{w}_k^g(t)\|^2 + H^2}$ is the distance between GU k and UAV m . Moreover, the probability of the NLOS links is given by

$$Pr_{k,m}^{NLOS}(t) = 1 - Pr_{k,m}^{LOS}(t). \quad (4)$$

Then, the average A2G pathloss between GU k and UAV m at time slot t can be modeled as

$$PL_{k,m}(t) = PL_{FS}(t) + Pr_{k,m}^{LOS}(t) \times \eta_{LOS} + Pr_{k,m}^{NLOS}(t) \times \eta_{NLOS}, \quad (5)$$

where $PL_{FS}(t) = 20 \log d_{k,m}(t) + 20 \log f_c + 20 \log (4\pi/V^l)$ is the free space pathloss, in which f_c and V^l denote the carrier frequency and the speed of light, respectively.¹ Furthermore, η_{LOS} and η_{NLOS} are the mean values of excessive pathloss in the LOS and NLOS links.

Each UAV serves its covered GUs in a frequency division multiple access (FDMA) mode, and the frequency bands of UAVs are orthogonal to each other. We assume that each UAV has the same amount of bandwidth B^g to serve GUs. Let P^g denote the transmission power of GUs, and the SNR at UAV m from GU k can be expressed as

$$\alpha_{k,m}^g(t) = \frac{P^g}{h_{k,m}^g(t) n_0 \frac{B^g}{N_m^g(t)}}, \quad (6)$$

where $h_{k,m}^g(t) = 10^{PL_{k,m}(t)/10}$, and n_0 is the noise power spectral density. Then the achievable transmission rate between GU k and UAV m at time slot t is

$$\zeta_{k,m}^g(t) = \begin{cases} \mu_{k,m}^g(t) \frac{B^g}{N_m^g(t)} \log_2(1 + \alpha_{k,m}^g(t)), & \text{if } N_m^g(t) > 0; \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

The data size of packet of each GU is generated from a Poisson distribution with parameter λ . At time slot t , GU k generates a packet with a data size of $\sigma_k^g(t)$ and then transmits it to UAV $\chi_k^g(t)$. The transmission time of the data packet is

$$\tau_k^g = \frac{\sigma_k^g(t)}{\zeta_{k,\chi_k^g(t)}^g(t)}. \quad (8)$$

¹The above pathloss expressions are in unit of dB.

Each UAV maintains a buffer to cache the received data packets, which follows the first-in-first-out rule. The remaining buffer size of UAV m is denoted by $l_m(t)$. When GU k chooses UAV m to transmit its newly generated data packet and $l_m(t) \geq \sigma_k^g(t)$, UAV m will immediately reserve the space in its buffer for the packet, i.e., $l_m(t+1) = l_m(t) - \sigma_k^g(t)$.² Otherwise, the packet will be dropped due to insufficient remaining buffer size triggering a *network congestion* event.

C. Transmission Model Between UAVs

Since the distance between a source UAV and the destination UAV M may exceed the direct communication range d^c , each packet would be transmitted through multiple hops until it reaches the destination UAV M . Let $\xi = \{m_0, m_1, \dots, m_{N_\sigma^u}\}$ denote the transmission path with N_σ^u hops connecting the source UAV and the destination UAV, where m_n represents the index of the n -th hop UAV on path ξ . Note that we have $m_{N_\sigma^u} = M$.

The communication links between UAVs can be assumed to be LOS links, since there is negligible blockage in the sky. The channel gain between UAV m and UAV m' follows the free space pathloss model:

$$\|h_{m,m'}^u(t)\|^2 = \frac{\rho_0}{\|\mathbf{w}_m^u(t) - \mathbf{w}_{m'}^u(t)\|^2}, \quad (9)$$

where ρ_0 denotes the channel gain at a reference distance of 1 m. All UAVs have the same transmission power P^u , and the received power of UAV m' from UAV m is represented by

$$P_{m,m'}^r(t) = P^u \|h_{m,m'}^u(t)\|^2. \quad (10)$$

Each UAV will choose the next hop UAV and the transmission band for every packet cached in its buffer. There are N_B orthogonal frequency bands with equal amount of bandwidth, whose index set is denoted by $\mathcal{B} = \{1, \dots, b, \dots, N_B\}$, for UAVs to choose to forward data packets. The bandwidth of each band is B^u . The index of the chosen frequency band of UAV m is denoted by $b_m(t)$, and we introduce a binary indicator $\mu_{m,b}^B(t)$ to indicate whether UAV m chooses frequency band b . Thus $\mu_{m,b_m(t)}^B(t) = 1$, and $\mu_{m,b}^B(t) = 0$ for $b \neq b_m(t)$. Since the UAVs working in the same frequency band can interfere with each other within the communication range, the signal-to-interference-plus-noise ratio (SINR) at UAV m_{n+1} from UAV m_n can be expressed as (11)

$$\begin{aligned} & \alpha_{m_n, m_{n+1}}^u(t) \\ &= \frac{P_{m_n, m_{n+1}}^r(t)}{\sum_{i \in \mathcal{M} \setminus \{m_{n+1}\}} \mu_{i, b_{m_n}}^B(t) \mu_{i, m_{n+1}}^c(t) P_{i, m_{n+1}}^r(t) + n_0 B^u} \end{aligned} \quad (11)$$

where $\mu_{i, m_{n+1}}^c(t) \in \{0, 1\}$ is the communication range indicator. Specifically, $\mu_{i, m}^c(t) = 1$ indicates that UAV i is within the communication range of UAV m , otherwise $\mu_{i, m}^c(t) = 0$. Then the transmission rate from UAV m_n to UAV m_{n+1} is

$$R_{m_n, m_{n+1}}^u = B^u \log_2(1 + \alpha_{m_n, m_{n+1}}^u(t)). \quad (12)$$

²If one UAV receives the data packets from multiple GUs at the same time, we have $l_m(t+1) = l_m(t) - \sum_{k \in \mathcal{K}} \mu_{k, m}^g(t) \sigma_k^g(t)$.

TABLE I
NOTATIONS OF THE TRANSMISSION MODEL

Notation	Meaning
χ_k^g	The associated UAV for GU k
$\mu_{k, m}^g$	The association indicator of UAV m and GU k
N_σ^g	The number of GUs served by UAV m
B^g	Frequency bandwidth of each UAV used to serve GUs
$\alpha_{k, m}^g$	SNR at UAV m from GU k
$\zeta_{k, m}^g$	Transmission rate between GU k and UAV m
ξ	Transmission path of a data packet
m_n	The index of the n -th hop UAV on path ξ
B^u	Frequency bandwidth of each band among UAVs
b_m	The index of chosen band of UAV m
$\mu_{m, b}^B$	The chosen band indicator
$\alpha_{m_n, m_{n+1}}^u$	The SINR at UAV m_{n+1} from UAV m_n
$\mu_{i, m}^c$	The communication range indicator
σ^u	The size of data packet among UAVs

At each time slot, each UAV forwards the first data packet cached in its buffer list. Taking a data packet of size σ^u at UAV m_n as an example,³ UAV m_{n+1} will reserve the corresponding buffer size for this packet, i.e., $l_{m_{n+1}}(t+1) = l_{m_n}(t) - \sigma^u$.⁴ The network congestion occurs when $l_{m_{n+1}}(t) < \sigma^u$. The transmission time $\tau_{m_n, m_{n+1}}^u$ for the packet in this hop is calculated as

$$\tau_{m_n, m_{n+1}}^u = \frac{\sigma^u}{R_{m_n, m_{n+1}}^u}. \quad (13)$$

And the overall transmission time along the path is

$$t = \sum_{n=0}^{n=N_\sigma^u-1} \tau_{m_n, m_{n+1}}^u. \quad (14)$$

The notations of the transmission model are listed in Table I.

D. Problem Formulation

We aim to minimize the transmission time of each packet and avoid the network congestion via optimizing UAV trajectories $\{\mathbf{w}_m^u(t)\}_{m \in \mathcal{M}, t \in \mathcal{T}}$, frequency resource allocation $\{b_m(t)\}_{m \in \mathcal{M}, t \in \mathcal{T}}$, and the transmission path ξ . Thus the problem can be formulated as:

$$\min_{\mathbf{w}_m^u, b_m, \xi} \sum_{n=0}^{n=N_\sigma^u-1} \tau_{m_n, m_{n+1}}^u \quad (15a)$$

$$\text{s.t. } \|\mathbf{w}_m^u(t+1) - \mathbf{w}_m^u(t)\| = V^u \delta_t, \forall m \in \mathcal{M}, \quad (15b)$$

$$\|\mathbf{w}_m^u(t) - \mathbf{w}_{m_{n+1}}^u(t)\| \leq d^c, \forall m_n \in \xi, \quad (15c)$$

$$l_{m_{n+1}}(t) \geq \sigma^u, \quad (15d)$$

$$l_m(t) \geq \sigma_k^g(t), \quad (15e)$$

³There is no subscript and suffix σ^u because once the packet is transmitted to the multi-hop UAV relay network, it does not matter that when was the data packet generated by which user.

⁴If UAV m_{n+1} is chosen as the next hop UAV for multiple packets from multiple UAVs, then it will immediately reserve the corresponding buffer for all the incoming packets.

$$\mathbf{w}_m^u(0) = \mathbf{w}_m^0, \quad (15f)$$

$$\psi_l \leq \mathbf{w}_m^u(t) \leq \psi_u, \quad \forall m \in \mathcal{M}, \quad (15g)$$

where \mathbf{w}_m^0 is the initial location of UAV m , and ψ_l and ψ_u denote the lower and upper boundary coordinates of the considered area, respectively. The constraint (15b) is the UAV speed constraint, and (15c) means that the next hop UAV should be within the communication range of the current UAV, otherwise a service outage event occurs. The constraints (15d) and (15e) indicate that the UAV should have sufficient remaining buffer size for the incoming packets. The UAVs are restricted in the considered area, as indicated by (15g).

According to (1a), since $\mathbf{w}_m^u(t)$ and $\mathbf{e}_m^u(t)$ have a one-to-one correspondence, the trajectory variables $\mathbf{w}_m^u(t)$ in (15a) can be replaced by flying direction variables $\mathbf{e}_m^u(t)$. This replacement makes the constraint (15b) negligible, which simplifies the optimization problem. Hence, problem (15) can be rewritten as:

$$\min_{\mathbf{e}_m^u, b_m, \xi} \sum_{n=0}^{n=N_\sigma^u-1} \tau_{m_n, m_{n+1}}^u \quad (16a)$$

$$\text{s.t. } \|\mathbf{w}_{m_n}^u(t) - \mathbf{w}_{m_{n+1}}^u(t)\| \leq d^c, \quad \forall m_n \in \xi, \quad (16b)$$

$$\mathbf{w}_m^u(0) = \mathbf{w}_m^0, \quad (16c)$$

$$l_{m_{n+1}}(t) \geq \sigma^u, \quad (16d)$$

$$l_m(t) \geq \sigma_k^g(t), \quad (16e)$$

$$\psi_l \leq \mathbf{w}_m^u(t) \leq \psi_u, \quad \forall m \in \mathcal{M}. \quad (16f)$$

However, problem (16) is still challenging to be solved. The reason is four-fold: (1) This is a sequential decision-making problem; (2) UAVs need to make decisions on their trajectories, frequency resource allocation and the next hop UAV at the same time. Hence the action space of each UAV is very large such that it is difficult to search an optimal policy; (3) These three types of decisions are coupled each other, which further complicates the joint decision making process. For example, the flying directions determine the next available hop UAVs, since the next hop UAV should be in the communication range of the current UAV; (4) In addition, the objective of problem (16) is dedicated for each data packet, which is determined by all the UAVs along the packet transmission path. While the objective of traditional DRL is to maximize the cumulative reward of each agent. Since the decision-making agents are the UAVs, this inconsistency cannot be solved by simple reward shaping, and maximizing the cumulative reward can not solve the problem (16) directly.

III. PRELIMINARIES

In a single-agent DRL algorithm [35], the agent observes the environment state $s(t)$, executes action $a(t)$, and then receives reward $r(t)$. The goal of DRL is to find a policy $\pi(a|s)$ that maps a state to an action for maximizing the expected discounted

cumulative reward $\mathbb{E}_\pi[\sum_i^T \gamma^i r(t)]$, where $\gamma \in (0, 1)$ is the discount factor. The state-action value function is defined as

$$Q(s(t), a(t)) = \mathbb{E}_\pi \left[\sum_{i=0}^T \gamma^i r(t+i+1) \middle| s(t), a(t) \right], \quad (17)$$

which is the expectation of the discounted cumulative reward. Then the greedy policy is $\pi(a|s) = \arg \max_a Q(s, a)$.

Take the deep Q network (DQN) [36] as an example. The DQN uses a deep neural network (DNN) to approximate the state-action value function, i.e., $Q(s(t), a(t); \theta)$, where θ is the weight of the DNN. However, DNN can cause instability and divergence issues when applied in DRL. Experience replay and target network are often used to tackle the issues. Specifically, the experience replay buffer \mathcal{C} stores the state transition samples $(s(t), a(t), r(t), s(t+1))$, and DRL samples a mini-batch to break the correlation between sequential samples. The target network θ' has the same architecture as the origin DNN θ , and copies the weights of θ every N_u steps. The network weights remain unchanged between two updates to break the correlation between training target and the original DNNs. The DNN is trained via minimizing the loss function, i.e.,

$$L(\theta) = \frac{1}{N_s} \sum_i [y(i) - Q(s(i), a(i); \theta)]^2, \quad (18)$$

where

$$y(i) = r(i) + \gamma Q'(s(i+1), \pi(a|s(t+1)); \theta'), \quad (19)$$

is the training target, and N_s is mini-batch size.

When considering *multi-agent* scenarios with N_a agents, each agent has its own state-action value function $Q_i(o_i, a_i; \theta_i)$.⁶ QMIX [33] proposes a joint action-value function $Q^{tot}(\mathbf{o}, \mathbf{a})$, where \mathbf{o} and \mathbf{a} are the joint observations and actions of all agents, i.e., the joint action-value function ensures that a global argmax operation performed on Q^{tot} yields the same result as a set of individual argmax operations performed on each Q_i , i.e.,

$$\arg \max_{\mathbf{a}} Q^{tot}(\mathbf{o}, \mathbf{a}) = \left(\begin{array}{c} \arg \max_{a_1} Q_1(o_1, a_1) \\ \vdots \\ \arg \max_{a_{N_a}} Q_{N_a}(o_{N_a}, a_{N_a}) \end{array} \right), \quad (20)$$

which means the choice of greedy action for each agent in a decentralized way can lead to the greedy joint actions. In order to satisfy (20), monotonicity should be enforced through a constraint on the relationship between Q^{tot} and each Q_i [33]:

$$\frac{\partial Q^{tot}}{\partial Q_i} \geq 0, \quad \forall i \in \{1, 2, \dots, N_a\}. \quad (21)$$

In order to guarantee the monotonicity, QMIX uses a mixing network that inputs the state-action value of each agent Q_i and outputs Q^{tot} . The weights (excluding the bias) of the mixing network are produced by separate hypernetworks [37] and are restricted to be non-negative. The nonnegativity can be obtained by absolute activation function.

⁵The inequality symbols between two vectors indicate that each element at the corresponding position in the two vectors satisfies the inequality relationship.

⁶In multi agent scenarios, the agents usually cannot observe the entire environment state. Each agent can only have local observation based on the state $o_i(t) = f_i(s(t))$, where $f_i(\cdot)$ is the observation function.

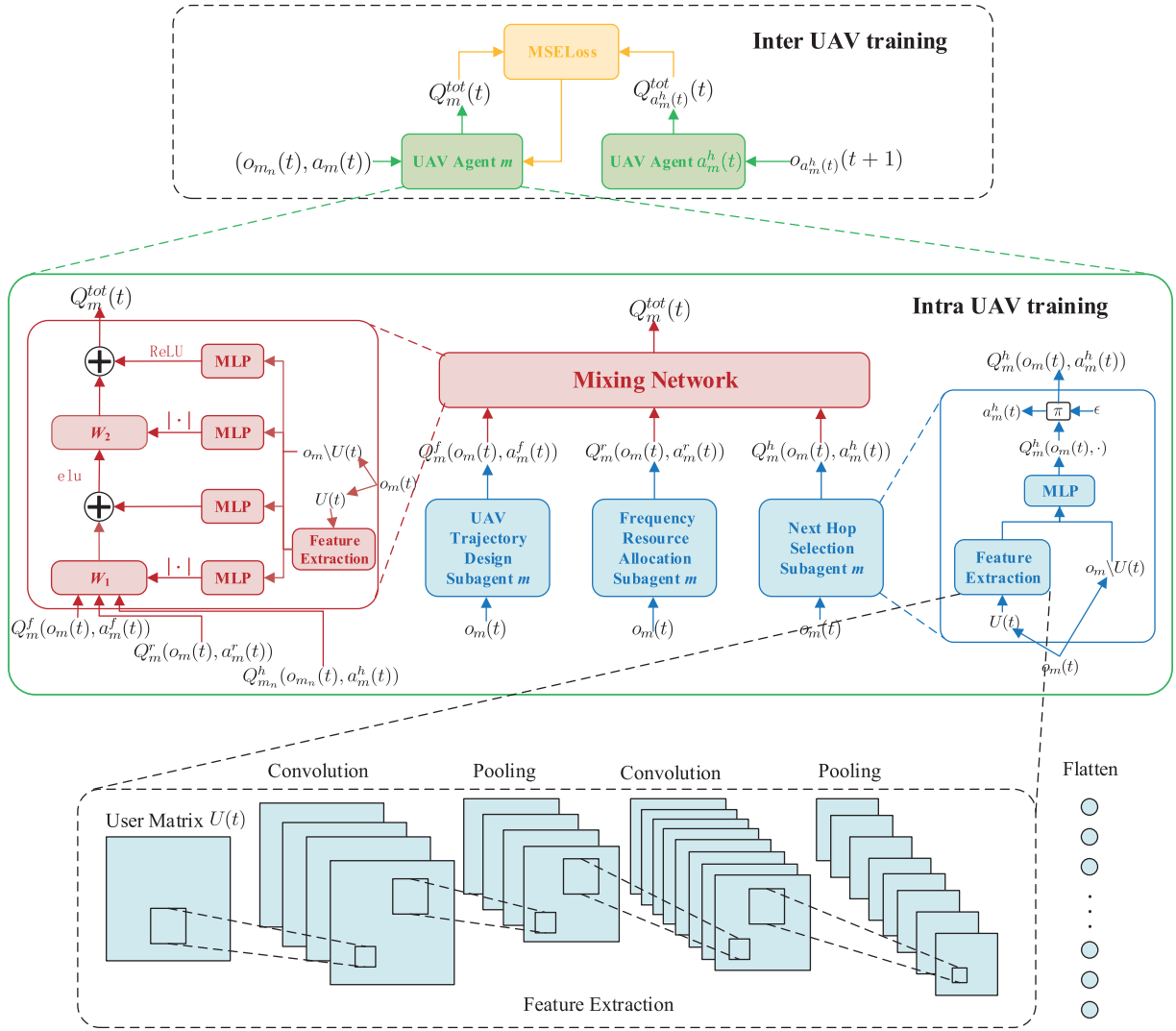


Fig. 2. Illustration of the MAQMIX algorithm. Each agent runs the multi-agent algorithm QMIX to tackle large action space issue, while the Q network is trained by the designed inter-UAV training mechanism.

IV. MAQMIX ALGORITHM

In this section, we present the proposed MAQMIX. As shown in Fig. 2, the MAQMIX has two designed mechanisms: an intra-UAV training mechanism and an inter-UAV training mechanism. Moreover, the convolutional neural networks (CNNs) are applied to extract feature of the user distribution. In the intra-UAV training part, each UAV agent is decomposed into three subagents which are responsible for trajectory design, frequency resource allocation, and the next hop selection, respectively. While in the inter-UAV training part, the training among UAVs are coordinated. Specifically, each UAV uses the state-action value of other agent to update its own state-action value, such that the state-action value can estimate the remaining transmission time for each packet.

The MAQMIX includes two phases: training phase and implementation phase. In the training phase, the DNN is trained offline, and the exploration is needed to search the optimal policy. While in the implementation phase, DNN only needs

forward propagation, which consumes much less resources than training. In addition, there is no need for exploration in the implementation phase.

A. Key Elements of the Proposed DRL Algorithm

We first introduce the designed observation, action and reward in the proposed algorithm.

1) *Observation*: The observation space of UAV agent m $o_m(t)$ consists of six components:

- The locations of all GUs $\{\mathbf{w}_k^g(t)\}_{k \in \mathcal{K}}$: According to Section II-A, the GUs choose the UAV with the strongest RSS to associate. Hence, the GUs distribution can affect UAV trajectory design. For example, more UAVs are needed to offload the data traffic where GUs are denser. In order to better extract the distribution characteristics of GUs, we map the GU locations $\{\mathbf{w}_k^g(t)\}_{k \in \mathcal{K}}$ to a GU distribution matrix $\mathbf{W}^g(t) \in \mathbb{R}^{N_c \times N_c}$ such that the CNNs

can be applied. Specifically, the considered area is equally divided into 10×10 grids, and $\mathbf{W}_{i,j}^g(t)$ represents the number of GUs in the corresponding grid;

- The channel gain between UAV m and other UAVs $\{\|h_{m,m'}^u(t)\|^2\}_{m' \in \mathcal{M} \setminus \{m\}}$;
- The UAVs' locations $\{\mathbf{w}_m^u(t)\}_{m \in \mathcal{M}}$;
- The remaining buffer size of all UAVs $\{L_m(t)\}_{m \in \mathcal{M}}$;
- The number of UAVs that select each frequency band $\{N_b^f\}_{b \in \mathcal{B}}$;
- The data size of the current data packet for transmission σ^u .

2) *Action*: The action of each UAV $a_m(t)$ consists of three components $a_m(t) = \{a_m^f(t), a_m^r(t), a_m^h(t)\}$, which are responsible for UAV trajectory design, frequency resource allocation, and the next hop selection, respectively.

- The flying direction $a_m^f(t) = e_m^u(t)$: For simplicity, the flying direction is discretized into five directions: left, right, forward, backward, and hover, i.e., $\{(-1, 0), (1, 0), (0, 1), (0, -1), (0, 0)\}$;
- The frequency resource allocation $a_m^r(t) = b_m(t)$: Each UAV can choose its transmission frequency band $b_m(t)$ to reduce the interference among UAVs;
- The next hop UAV $a_m^h(t) = m_{n+1}$: In our problem, the UAV chooses proper next hop to minimize the transmission time.

The action space has a cardinality of $5 \times N_B \times (M - 1)$, which is extremely large and thus difficult to tackle as the number of UAVs increases.

3) *Reward*: The reward of each UAV consists of the following five components:

- The transmission time of the current hop $\tau_{m_n, m_{n+1}}^u$, which follows the objective function (16a). Since the objective of DRL is to minimize the transmission time, the reward can be defined as the opposite value of the transmission time;
- The constant penalty for flying out of the considered area is denoted by r_m^s . The constant penalty r_m^s restricts the UAV to the considered area, which reflects the constraint (16f);
- The constant penalty r_m^e for the network congestion that occurs at the next hop (16d);
- The congestion penalty for the congestion caused by the packets from GUs, which means the constraint (16e) is violated. If the packet generated from GU k exceeds the remaining buffer size of UAV m , i.e., UAV m is overloaded, the UAVs within the distance d^ϕ from GU k will receive a penalty $\{r_i^\phi\}_{i \in \mathcal{M} \setminus \{m\}}$, whose value is based on the distance between UAVs and GU k , while the UAVs beyond d^ϕ will not be punished. The UAVs within the communication range get penalty because the congestion caused by the packet from GU means that areas near this GU need more UAVs to serve. UAVs outside the range are not punished because they are too far away from the current area with high user density, and there is no need for these UAVs to serve this area. Note that $\{r_i^\phi\}_{i \in \mathcal{M} \setminus \{m\}}$ is for UAVs other than m , and can be specifically defined as

$$r_i^\phi = r^\beta - \mu_{i,k}^d \kappa_d \|\mathbf{w}_i^u - \mathbf{w}_k^g\|. \quad (22)$$

Here, r^β is the bonus for encouraging UAVs to fly close to the overloaded UAV to offload data traffic, and κ_d is a positive weight for balancing reward and distance. $\mu_{i,k}^d$ is the range indicator, with $\mu_{i,k}^d = 1$ indicating that the distance between UAV i and GU k is within d^ϕ , and $\mu_{i,k}^d = 0$ otherwise;

- The outage penalty for the selected UAV exceeding the communication range r_m^c . The outage penalty r_m^c prevents the UAV from violating the constraint (16b).

The overall reward of UAV m is defined as

$$r_m = -\kappa_t \tau_{m,m'}^u + \mu_m^s r_m^s + \mu_m^e r_m^e + r_m^\phi + \mu_m^c r_m^c, \quad (23)$$

where κ_t is a positive adjusting weight, while μ_m^s , μ_m^e and μ_m^c indicate whether UAV m flies out of the considered area, chooses the next hop UAV that occurs congestion, and chooses the next hop UAV out of the communication range, respectively. Note that r_m^ϕ comes from other agents, whose remaining buffer size is insufficient for receiving the packets from GUs.

B. Intra-UAV Training Mechanism

To tackle large action space issue, each UAV agent runs the MADRL algorithm QMIX. Specifically, we assume that each UAV consists of three subagents, i.e., trajectory design subagent, frequency resource allocation subagent, and the next hop selection subagent, being responsible for three actions defined in Section IV-A2. As such, the large action space is decomposed into three small action space. The decomposition reduces the action space from $5 \times N_B \times (M - 1)$ to $5 + N_B + (M - 1)$, which significantly reduces the training complexity. The greedy policy of UAV agent m is

$$\pi_m(a_m | o_m) = \left\{ \underset{a_m^f}{\operatorname{argmax}} Q_m^f(o_m, a_m^f), \underset{a_m^r}{\operatorname{argmax}} Q_m^r(o_m, a_m^r), \underset{a_m^h}{\operatorname{argmax}} Q_m^h(o_m, a_m^h) \right\}, \quad (24)$$

where $Q_m^f(\cdot)$, $Q_m^r(\cdot)$, and $Q_m^h(\cdot)$ are the state-action value functions of the subagents for trajectory design, frequency resource allocation, and the next hop selection, respectively. According to (20), we have

$$\underset{a}{\operatorname{argmax}} Q_m^{\text{tot}}(o_m, a) = \pi_m(a_m | o_m). \quad (25)$$

C. Inter-UAV Training Mechanism

Although each UAV is decomposed into three subagents, the mixing network can still output a state-action value Q_m^{tot} . Hence, we can treat each UAV as one agent in the inter-UAV training process.

Let us modify the training target defined in (19) and use the state-action value of the next hop UAV to calculate the training target, which is easy to obtain through information exchange among UAVs [38]. The training target defined in (19) is converted to

$$y_m(i) = r_m(i) + \gamma Q_{m'}^{\text{tot}}(o_{m'}(i), \pi_{m'}(a_{m'} | o_{m'}); \theta_{m'}^Q). \quad (26)$$

Using the state-action value from other agents to calculate the training target can break the correlation between the target and the origin DNN. As a result, the separate target network is not required in the MAQMIX.

Note that after sufficient training, the agents hardly receive the penalties defined in Section IV-A3. The state-action value of each agent Q_m^{tot} can be regarded as the estimated opposite value of the subsequent transmission time of the current data packet. For example, if UAV m_n connects to the destination UAV, the reward of UAV m_n is the opposite value of the transmission time between UAV m_n and the destination UAV, and then the state-action value Q_m^{tot} can be regarded as the estimation of the opposite value of the transmission time from UAV m_n to the destination UAV. UAV m_{n-1} chooses UAV m_n as the next hop UAV, and then the training target in (26) becomes $r_{m_{n-1}} + Q_{m_n}^{tot}$, where $r_{m_{n-1}}$ is the opposite value of the transmission time between UAV m_{n-1} and UAV m_n , and $Q_{m_n}^{tot}$ is the estimation of the opposite value of the transmission time from UAV m_n to the destination UAV. Thus, $Q_{m_{n-1}}^{tot}$ can be regarded as the estimation of the opposite value of transmission time from UAV m_{n-1} to the destination UAV. Similarly, we can infer that the state-action value Q_m^{tot} can be regarded as the estimated opposite value of the transmission time from UAV m to the destination UAV. Maximizing the state-action value is equal to minimizing the transmission time.

D. Network Architecture

As shown in Fig. 2, each UAV agent has three subagents. In each subagent, the GU distribution matrix in the observation is first input into the feature extraction. Then the extracted features and the remaining observations are input to a multilayer perceptron (MLP) that consists of multiple fully connected layers and *ReLU* activation layers. The MLP outputs the state-action values of all possible actions. The subagents choose actions using ϵ -greedy policy. Specifically, the subagent chooses a random action with probability ϵ or chooses the action with the maximum value with probability $1 - \epsilon$.

The mixing network inputs the three chosen state-action value and outputs the total value Q_m^{tot} . The weights of the mixing network are produced by the hypernetwork. Similarly, the hypernetwork first extracts features of the GU distribution matrix, and then the features as well as other observations are input into the MLPs. The outputs of the MLPs are activated by the absolute activation function and are considered as the weights of fully connected layers of the mixing network.

E. Training Algorithm

The proposed MAQMIX algorithm is episodic with episode length T . At the beginning of each episode, the UAVs and GUs are randomly distributed in the considered area.

The GUs generate data packets and send them to the UAVs with the strongest RSS. After the last data transmission, the UAV m gets the observation, inputs $o_m(t)$ to three subagents, and outputs corresponding state-action values. In the training phase, the subagents choose actions based on ϵ -greedy policy, which encourage the exploration to prevent the agents

Algorithm 1: Training Phase of the MAQMIX.

- 1: Architecture of each UAV state-action value network that includes three subagents and a mixing network, UAV to GU frequency band B^g , UAV frequency bands B^u , experience replay buffer \mathcal{C} , episode length T , batch size N_s , UAV-BS speed V , the number of UAVs M and the number of GUs K ;
 - 2: Well-trained network parameters of all UAVs;
 - 3: The networks of all UAV agents are all randomly initialized. The weights of the mixing network is produced by separate hypernetworks and restricted to be non-negative;
 - 4: Initialize the experience replay buffer;
 - 5: **for** each episode **do**
 - 6: Initialize locations of UAVs and GUs;
 - 7: **for** each time slot t **do**
 - 8: The GUs generate data packets and transmit them to the corresponding UAVs;
 - 9: **for** each UAV m **do**
 - 10: The agent m gets the observation $o_m(t)$;
 - 11: The three subagents take the observation as input, and output the flying direction, frequency resource allocation, and the next hop UAV via ϵ -greedy policy;
 - 12: **end for**
 - 13: All agents take actions;
 - 14: The agents obtain the rewards $\{r_m(t)\}_{m \in \mathcal{M}}$ and the next observations $\{o_m(t+1)\}_{m \in \mathcal{M}}$;
 - 15: Store $(\{o_m(t)\}_{m \in \mathcal{M}}, \{a_m(t)\}_{m \in \mathcal{M}}, \{r_m(t)\}_{m \in \mathcal{M}}, \{o_m(t+1)\}_{m \in \mathcal{M}})$ in experience replay buffer;
 - 16: **end for**
 - 17: Sample several random minibatches of N_s experience tuples from replay buffer;
 - 18: **for** each agent m **do**
 - 19: Calculate the target according to (19);
 - 20: Update the network by minimizing the loss (18);
 - 21: **end for**
 - 22: **end for**
-

from local optimum. In the implementation phase, the exploration is not required, since the DNNs have been well-trained. Then the environment proceeds with agents taking the selected actions $\{a_m(t)\}_{m \in \mathcal{M}}$, and transitions to the next state. The agents obtain their rewards $\{r_m(t)\}_{m \in \mathcal{M}}$ and the next observations $\{o_m(t)\}_{m \in \mathcal{M}}$. The experience replay buffer stores the experience tuple $(\{o_m(t)\}_{m \in \mathcal{M}}, \{a_m(t)\}_{m \in \mathcal{M}}, \{r_m(t)\}_{m \in \mathcal{M}}, \{o_m(t+1)\}_{m \in \mathcal{M}})$. When the buffer is full, batches of experience tuples are randomly sampled and the DNNs of agents are trained via minimizing the loss (18). The proposed MAQMIX algorithm is summarized in Algorithm 1.

V. PERFORMANCE EVALUATION

In this section, we conduct extensive simulations to evaluate the proposed MAQMIX algorithm.

TABLE II
COMMUNICATION RELATED PARAMETERS

Notation	Meaning	Value
n_0	Noise power spectral density	10^{-17} W/Hz [15]
P^g	Transmission power of GUs	0.01 W [15]
P^u	Transmission power of UAVs	0.03 W
η_1	The parameter of the probabilistic pathloss model	4.88 [39]
η_2	The parameter of the probabilistic pathloss model	0.43 [39]
η_{LOS}	The parameter of the probabilistic pathloss model	0.1 dB [39]
η_{NLOS}	The parameter of the probabilistic pathloss model	21 dB [39]
B^g	The bandwidth used by each UAV to serve the GUs	2×10^5 Hz
B^u	The bandwidth of each orthogonal band in UAV to UAV transmission	1×10^6 Hz [15]

TABLE III
DETAILED MLP ARCHITECTURES OF THE THREE SUBAGENTS

	UAV trajectory design	Frequency resource allocation	Next hop selection
Input layer	54	54	54
hidden layer 1	64	64	64
hidden layer 2	64	64	64
hidden layer 3	64	64	64
hidden layer 4	64	64	64
Output layer	5	4	9

A. Simulation Settings

We set the side length of the square area as $L_s = 1$ km, where all GUs and UAVs are restricted to the considered area. There are $K = 50$ GUs in this area. The episode length is $T = 1,000$. UAVs fly at an altitude $H = 300$ m. The communication range of each UAV is $d^c = 400$ m. Taking the coverage of the considered simulation area and the communication range of UAVs into account, the number of UAVs is set to $M = 10$, such that the UAVs can cover the entire area even in some extreme cases. The speed of UAVs and GUs are $V^u = 10$ m/s, $V^g = 5$ m/s, respectively. The data size of the generated packet is sampled from a Poisson distribution with parameter $\lambda = 20$. The communication related system parameters are summarized in Table II.

In the proposed MAQMIX algorithm, we divide the considered area into 10×10 grids whose side length is $L_g = 100$ m. Each subagent has a CNN with two convolutional layers to extract the feature of user distribution. The input channel and output channel of each layer is set to 1. The kernel size, stride, and padding of both layers are set to 3, 2, and 1, respectively. Each subagent also has a MLP to output the state-action values, and each hidden layer is activated by *ReLU* function. The detailed architectures are shown in Table III. The mixing network has one convolutional layer to extract the features of user distribution and then has a MLP to output the total state-action value. The MLP has one hidden layer with 64 neurons. The input size is 3, which represents the state-action value of the three subagents. Note that the weights of the mixing network are produced by the hypernetworks. The hypernetworks take the state as input, and output weights of the mixing network, which is activated by absolute function.

TABLE IV
FLOPS OF THE THREE SUBAGENTS

	UAV trajectory design	Frequency resource allocation	Next hop selection
FLOPs	34367	34240	34875

TABLE V
REWARD PARAMETERS

Notation	Meaning	Value
κ_t	Weight between reward and transmission time	100
κ_d	Weight between penalty and distance	0.01
r^β	Bonus in (22)	5
r_m^s	Penalty for flying out of the considered area	-50
r_m^c	Penalty for the selected UAV exceeding communication range	-50
r_m^e	Penalty for the congestion	-50
d^ϕ	The distance threshold about the reward defined in (22)	800

The parameters of reward in (23) are listed in Table V. We set the random exploration possibility $\epsilon = 0.1$, and adopt the ADAM optimizer [40] with the starting learning rate $L_r = 0.001$. For each learning step, each agent randomly samples $N_s = 32$ experiences from the experience replay buffer for training. The capacity of the experience replay buffer is set to 15,000.

B. Complexity Analysis

The computational complexity has been analyzed in terms of floating operations (FLOPs) [41]. For convolutional layers, FLOPs can be calculated as

$$\text{FLOPs} = 2H_{in}W_{in}(C_{in}K^2 + 1)C_{out}, \quad (27)$$

where H_{in} , W_{in} and C_{in} are the height, width and the number of channels of the input feature map, K is the kernel size, and C_{out} is the number of output channels.

For fully-connected layers, FLOPs can be calculated as

$$\text{FLOPs} = (2N_{in} - 1)N_{out}, \quad (28)$$

where N_{in} is the number of input neurons and N_{out} is the number of output neurons.

Considering the neural network architecture of each subagent, the FLOPs of the three subagents are calculated and listed in Table IV. The difference in FLOPs among the three subagents is caused by the difference of the output layers. Note that the three subagents can make decisions in parallel.

C. Simulation Results

We compare the proposed MAQMIX algorithm with the baseline algorithm AODV [42]. AODV is a well-known ad hoc routing algorithm, which is based on the shortest path principle. In the AODV algorithm, the locations of all UAVs are fixed, and frequency resources are randomly allocated. In addition, in order to verify the effectiveness of each subagent, we propose three other benchmarks which remove one of the three subagents

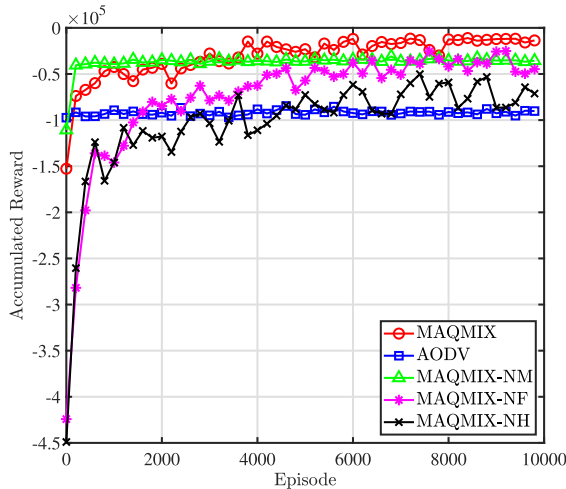


Fig. 3. Cumulative reward performance of all agents with respect to episodes.

respectively and only optimize the other two subagents. These benchmarks are detailed as follows:

- MAQMIX-NM: In the MAQMIX-NM, the trajectory design subagent is ignored, and the locations of all UAVs are fixed. The other two subagents, frequency resource allocation and the next hop selection subagents, are trained by MAQMIX.
- MAQMIX-NF: The frequency resources are randomly allocated, while the other two subagents are trained by MAQMIX.
- MAQMIX-NH: Each UAV chooses the shortest path as the alternative of next hop selection subagent, while the other two subagents are trained by MAQMIX.

We first investigate the cumulative reward of all agents in each episode to verify the effectiveness of the proposed algorithm. As shown in Fig. 3, the cumulative reward shows the tendency of increment with the training process, which reveals the effectiveness of the MAQMIX. With all the three subagents, the MAQMIX algorithm achieves the best performance by the end of training, and all the three benchmarks simplified from MAQMIX outperform the baseline AODV. At the beginning of the training process, the cumulative rewards of the MAQMIX-NF and MAQMIX-NH are much lower than that of the MAQMIX-NM. This is because at the beginning, the UAVs fly randomly due to the trajectory design subagent, which results in volatile connections among UAVs. The packets are frequently dropped due to the disconnection of the chosen next hop UAV, leading to the penalty. The MAQMIX-NM converges much faster than others and the cumulative reward is much more stable after convergence. The reason is that the frequency resource allocation and the next hop selection subagents are much easier to train compared with trajectory design subagent. This is also the reason why the cumulative reward of the MAQMIX in the first 1,000 episodes rises rapidly despite the trajectory design subagent. Moreover, the lack of trajectory subagent in the MAQMIX-NM leads to the stability, since the network topology is static. The cumulative reward of the MAQMIX-NH is lower than that of the MAQMIX-NM and MAQMIX-NF after

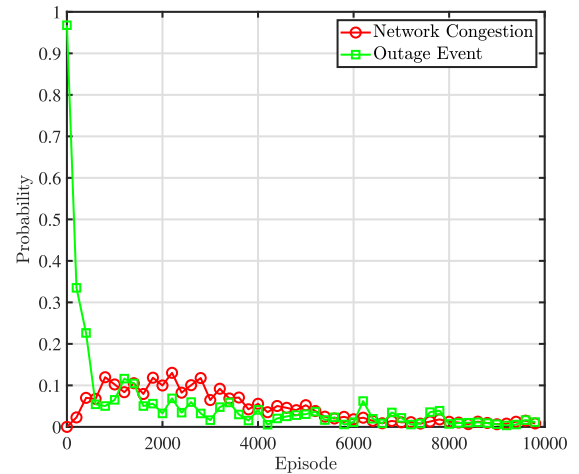


Fig. 4. Probability of network congestion and outage of the proposed algorithm.

convergence, because the dynamic network topology caused by trajectory design subagent deteriorates the performance of the shortest path routing protocol.

One of a major purpose is to prevent exceptional circumstances such as network congestion and outage event.⁷ As shown in Fig. 4, the probability of outage event decreases drastically in the first 1,000 episodes, which illustrates that the agents have learned how to choose UAVs within their communication range. However, the probability of network congestion shows a trend of rising at the beginning and then decreases after about 1,000 episodes. The reason is that whether the network congestion occurs can be counted only when the packets are transmitted to the UAV within the communication range. At the beginning, most UAVs cannot choose the next hop UAVs within their communication range properly, and thus there is almost no congestion. As the training process, more packets can be transmitted to the UAVs within the communication range, and the network starts to get congested. Then UAVs learn experience from congestion transitions via experience replay to avoid network congestion, such that the probability of congestions decreases to about 0 afterwards.

We plot trajectories of UAVs and GUs in Fig. 5. The GUs are divided into 3 different groups and move towards 3 different directions, and the destination UAV M hovers at coordinate of (950, 950). Although the trajectories of UAVs have some randomness, the UAVs exhibit the tendency of moving along with the GUs. The explanation of this phenomenon is that UAVs can achieve higher transmission rate due to the closer distance. In addition, UAVs hovering around cluster of GUs could provide more choices for GUs, and thus prevent the network congestion. For example, UAV₀ in Fig. 5 flies towards the cluster of GUs at the beginning. This is because all GUs gather inside the lower left grid and the UAVs should fly towards the GUs to offload the data traffic. Then UAV₀ moves along with a cluster of GUs, and hovers around the coordinate of (800, 400) at the end of episode.

⁷The outage event means the chosen next hop UAV that is out of its communication range.

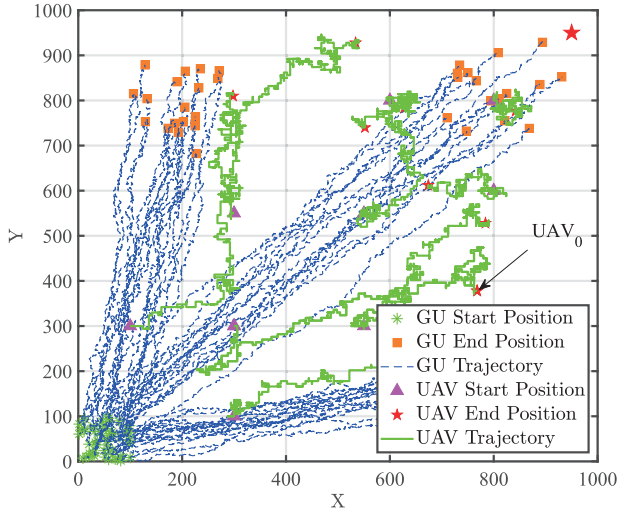


Fig. 5. Trajectory of UAVs. The UAVs exhibit the tendency of moving along with the GUs, and keep connected to each other.

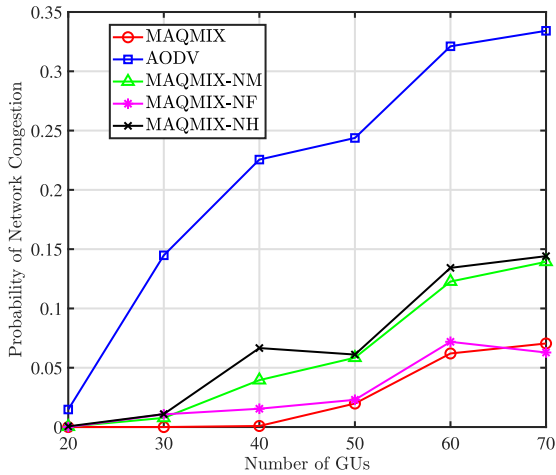


Fig. 6. Probability of network congestion versus the number of GUs.

For UAVs initialized around the destination, their cached data packets are mainly from other UAVs and do not need to provide service for GUs directly. Hence, these UAVs hover around the initial locations and have no obvious moving tendency.

Figure 6 illustrates the impact of the number of GUs on the probability of network congestion. It can be observed that in the MAQMIX algorithm, the probability of congestion increases monotonically from 0% to 7.05% as the number of GUs increases from 20 to 70. This is because more GUs generate more data packets, leading to higher network congestion probability. In addition, the MAQMIX reduces 91.88% network congestion compared with AODV when the number of GUs is $K = 50$, and all the other algorithms derived from the MAQMIX can also effectively reduce the congestion probability. This is because AODV is a routing protocol using the shortest path criterion, while the MAQMIX can learn from the historical experience to avoid network congestion by choosing the transmission path properly. We can also observe that the MAQMIX-NF achieves similar network congestion reduction as the MAQMIX. This is because the network congestion is mainly affected by the next

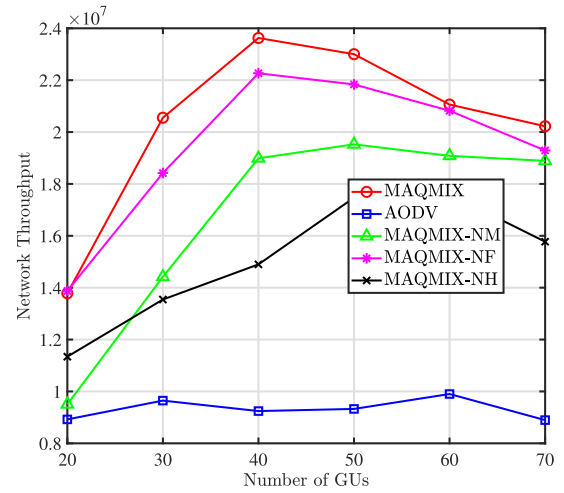


Fig. 7. Network throughput performance versus the number of GUs.

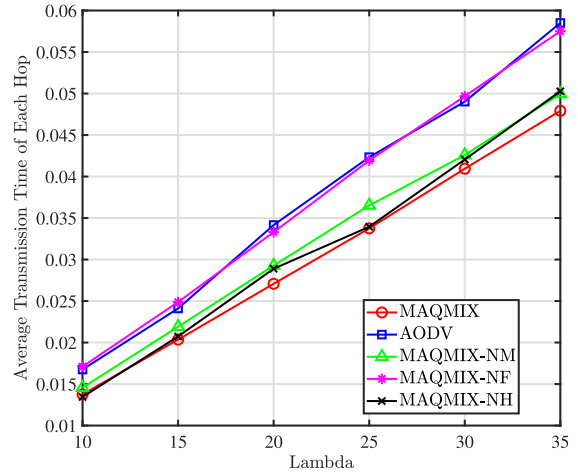


Fig. 8. Average transmission time of each hop versus the value of λ .

hop selection and trajectory design rather than the frequency resource allocation.

As shown in Fig. 7, the MAQMIX achieves the highest throughput compared with all benchmarks, and increases 60.87% total throughput of the whole system compared with AODV. We can observe that the throughput of the system first increases and then decreases as the number of GUs increases. The explanation is that when there are less GUs, the network is relatively idle. The increase in the number of GUs can result in the increase in network throughput. However, once the throughput reaches the network capacity, the increase in the number of data packets leads to a higher probability of network congestion, more data packets are dropped, and thus the network throughput decreases. Another notable phenomenon is that there is no significant difference between the throughput of the MAQMIX and that of the MAQMIX-NF algorithm. This can be explained by Fig. 6, since the two algorithms do not have significant difference in the probability of network congestion.

According to definition of Poisson distribution, the increment of parameter λ leads to the increment of average packet size. In Fig. 8, we can observe that the average transmission time of each hop increases monotonically with the increment of λ . In

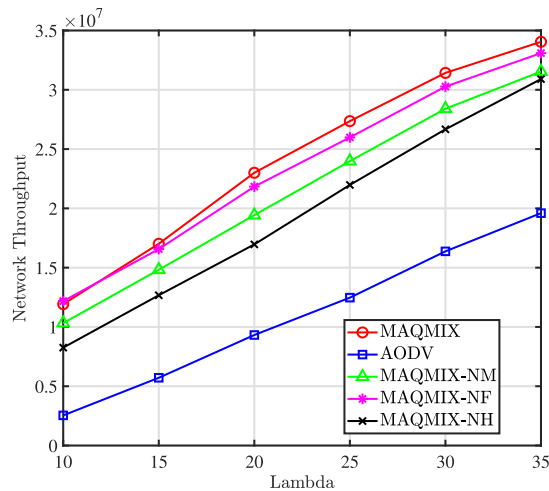


Fig. 9. Network throughput performance versus the value of λ .

the MAQMIX algorithm, the average transmission time reduces by 20.76% compared with AODV. We can observe an interesting phenomenon that the MAQMIX-NF has no significant reduction in average hop time compared with AODV, while the other two MAQMIX derived algorithms can get a relatively large reduction. This is because the frequency resource allocation subagent can reduce the interference among UAVs. According to (11), the reduction of interference leads to higher transmission rate, which in turn reduces the transmission time of each hop. Moreover, since the transmission rates among UAVs are also affected by UAV locations, the trajectory design subagent can also improve the transmission rate and bring the better performance of the MAQMIX compared with the MAQMIX-NM.

Fig. 9 reflects the impact of λ in terms of the network throughput. As illustrated in this figure, the MAQMIX-NF achieves the best performance in all three benchmark algorithms, since the frequency resource allocation action has negligible influence on the network throughput. The MAQMIX-NH has the worst performance in the three benchmark algorithms, and the reason is that the shortest path criterion under the dynamic network topology leads to higher possibility of outage event brought by the trajectory design subagent. This figure also illustrates that the throughput of all algorithms increases monotonically with the increment of λ . While the increment of the throughput in the MAQMIX tends to be flat when λ becomes large. The reason is that when the packets size is relatively small, the bottleneck of network throughput is the data packet size. Thus the increase of λ brings the linear increase of network throughput. However, the larger packet size also leads to the higher possibility of network congestion, which in turn reduces the network throughput.

VI. CONCLUSION

In this paper, we have investigated the packet routing problem in the multi-hop UAV relay network. We have proposed a novel MAQMIX algorithm, which leverages intra-UAV and inter-UAV training mechanisms to tackle the large action space issue and coordinate the training among UAVs, respectively. Simulation results have shown that the MAQMIX can enhance

the network throughput, reduce the network congestion probability, and shorten the transmission time. The intra-UAV training mechanism can be applied to tackle the problems with large action space, and the inter-UAV training mechanism can solve problems whose objectives are determined by multiple agents. For future work, we will investigate the store-carry-forward routing problem in the multi-hop UAV relay network.

REFERENCES

- [1] J. Chen, R. Ding, W. Wu, J. Liu, F. Gao, and X. Shen, "Multi-agent learning based packet routing in multi-hop UAV relay network," in *Proc. Int. Conf. Commun.*, 2022, pp. 1–5.
- [2] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surv. Tuts.*, vol. 24, no. 1, pp. 1–30, Jan.–Mar. 2021.
- [3] N. Kato, B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten challenges in advancing machine learning technologies toward 6G," *IEEE Wireless Commun.*, vol. 27, no. 3, pp. 96–103, Jun. 2020.
- [4] S. Chandrasekharan *et al.*, "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, May 2016.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] R. Ding, F. Gao, and X. Shen, "Deep reinforcement learning based 3D UAV trajectory design and frequency band allocation," in *Proc. IEEE Glob. Commun. Conf.*, 2020, pp. 1–6.
- [10] J. Zhao, F. Gao, L. Kuang, Q. Wu, and W. Jia, "Channel tracking with flight control system for UAV mmWave MIMO communications," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1224–1227, Jun. 2018.
- [11] W. Wu *et al.*, "Dynamic RAN slicing for service-oriented vehicular networks via constrained learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2076–2089, Jul. 2021.
- [12] F. Cheng *et al.*, "UAV trajectory optimization for data offloading at the edge of multiple cells," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6732–6736, Jul. 2018.
- [13] J. Zhao, F. Gao, G. Ding, T. Zhang, W. Jia, and A. Nallanathan, "Integrating communications and control for UAV systems: Opportunities and challenges," *IEEE Access*, vol. 6, pp. 67519–67527, 2018.
- [14] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao, and X. Shen, "Energy-efficient UAV-assisted mobile edge computing: Resource allocation and trajectory optimization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3424–3438, Mar. 2020.
- [15] 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.
- [16] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.
- [17] F. Cui, Y. Cai, Z. Qin, M. Zhao, and G. Y. Li, "Multiple access for mobile-UAV enabled networks: Joint trajectory design and resource allocation," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 4980–4994, Jul. 2019.
- [18] R. Ding, Y. Xu, F. Gao, and X. Shen, "Trajectory design and access control for air-ground coordinated communications system with multi-agent deep reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 8, pp. 5785–5798, Apr. 2021, doi: [10.1109/JIOT.2021.3062091](https://doi.org/10.1109/JIOT.2021.3062091).
- [19] W. Shi *et al.*, "Multi-drone 3-D trajectory planning and scheduling in drone-assisted radio access networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8145–8158, Aug. 2019.

- [20] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 161–164, Jan. 2018.
- [21] J. Zhang, F. Liang, B. Li, Z. Yang, Y. Wu, and H. Zhu, "Placement optimization of caching UAV-assisted mobile relay maritime communication," *China Commun.*, vol. 17, no. 8, pp. 209–219, Oct. 2020.
- [22] S. Ahmed, M. Z. Chowdhury, and Y. M. Jang, "Energy-efficient UAV relaying communications to serve ground nodes," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 849–852, Apr. 2020.
- [23] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-relaying-assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, May 2019.
- [24] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 1986–1999, Sep. 2018.
- [25] S. Hosseinalipour, A. Rahmati, and H. Dai, "Interference avoidance position planning in dual-hop and multi-hop UAV relay networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7033–7048, Nov. 2020.
- [26] N. Kato *et al.*, "Optimizing space-air-ground integrated networks by artificial intelligence," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 140–147, 2019.
- [27] X. Shen *et al.*, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 45–66, Jan. 2020.
- [28] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, pp. 1–14, 2016.
- [29] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5867–5876.
- [30] 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.
- [31] R. Ding, F. Gao, and X. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796–7809, 2020.
- [32] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.
- [33] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4295–4304.
- [34] 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.
- [35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
- [36] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.
- [37] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," in *Proc. Int. Conf. Learn. Representations*, pp. 1–18, 2017.
- [38] R. Ding, Y. Yang, J. Liu, H. Li, and F. Gao, "Packet routing against network congestion: A deep multi-agent reinforcement learning approach," in *Proc. Int. Conf. Comput., Netw. Commun.*, 2020, pp. 932–937.
- [39] 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. Int. Conf. Commun.*, 2016, pp. 1–5.
- [40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, pp. 1–15, 2015.
- [41] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.
- [42] C. Perkins, E. Belding-Royer, and S. Das, "RFC3561: Ad hoc on-demand distance vector (AODV) routing," United States, 2003.



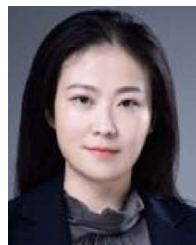
Ruijin Ding (Student Member, IEEE) received the B.Eng. degree in electrical and information engineering from the Dalian University of Technology, Dalian, China in 2017. He is currently working toward the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include optimization, UAV communications, and deep reinforcement learning.



Jiawei Chen (Student Member, IEEE) received the B.S. degree in 2020 from the Department of Automation, Tsinghua University, Beijing, China, where he is currently working toward the M.S. degree with the Department of Automation, Tsinghua University. His research interests include UAV-assisted communications, massive MIMO, and artificial intelligence assisted communications.



Wen Wu (Senior Member, IEEE) received the B.E. degree in information engineering from the South China University of Technology, Guangzhou, China, and the M.E. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2012 and 2015, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2019. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo. He is currently an Associate Researcher with Frontier Research Center, Peng Cheng Laboratory, Shenzhen, China. His research interests include 6G networks, pervasive network intelligence, digital twin, and network virtualization.



Jun Liu (Member, IEEE) received the Ph.D. degree from Northeastern University, Shenyang, China, in 2011. From 2011 to 2016, She was a Lecturer with the College of Information Science and Engineering, Northeastern University. From 2016 to 2018, she was a Postdoctoral Research Fellow with CECA Department, Peking University, Beijing, China. She is currently an Assistant Professor with the Institute of Network Sciences and Cyberspace, Tsinghua University, Beijing, China, and a Researcher with the Beijing National Research Center for Information Science and Technology, Beijing, China. Her research interests include wireless networks and space-air-terrestrial networks.



Feifei Gao (Fellow, IEEE) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada, in 2004, and the Ph.D. degree from the National University of Singapore, Singapore, in 2007. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an Associate Professor. He has authored or coauthored more than 180 refereed IEEE journal papers and more than 180 IEEE conference proceeding papers that are cited more than 12800

times in Google Scholar. His research interests include signal processing for communications, array signal processing, convex optimizations, and artificial intelligence assisted communications. Prof. Gao was the Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the Lead Guest Editor of IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, the Senior Editor of IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING and IEEE SIGNAL PROCESSING LETTERS, the Area Editor of IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, and *China Communications*. He was also the Symposium Co-chair of 2019 IEEE Conference on Communications (ICC), 2018 IEEE Vehicular Technology Conference Spring (VTC), 2015 IEEE Conference on Communications (ICC), 2014 IEEE Global Communications Conference (GLOBECOM), and 2014 IEEE Vehicular Technology Conference Fall (VTC), and also a Technical Committee Member of more than 50 IEEE conferences.



Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. He is also a registered Professional Engineer in Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian

Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Dr. Shen was the recipient of the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015, and Education Award in 2017 from the IEEE Communications Society (ComSoc), and Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013). He was also the recipient of the Excellent Graduate Supervision Award in 2006 from the University of Waterloo, and Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He was the Technical Program Committee Chair/Co-Chair of IEEE Globecom'16, IEEE Infocom'14, IEEE VTC'10 Fall, IEEE Globecom'07, and the Chair of the IEEE ComSoc Technical Committee on Wireless Communications. He is the President of the IEEE ComSoc. He was the Vice-President of Technical and Educational Activities, Vice-President of Publications, Member-at-Large on the Board of Governors, Chair of the Distinguished Lecturer Selection Committee, and Member of IEEE Fellow Selection Committee of the ComSoc. He was the Editor-in-Chief of the IEEE IoT JOURNAL, IEEE NETWORK, and *IET Communications*.