Multi-Agent Learning Based Packet Routing in Multi-Hop UAV Relay Network

Jiawei Chen*, Ruijin Ding*, Wen Wu†, Jun Liu‡, Feifei Gao*, and Xuemin (Sherman) Shen§
*Department of Automation, Tsinghua University, Beijing, China
†Frontier Research Center, Peng Cheng Laboratory, Shenzhen, China
‡Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China
§Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada
Email: {chenjiaw20, drj17}@mails.tsinghua.edu.cn, {juneliu, feifeigao}@tsinghua.edu.cn, {w77wu, sshen}@uwaterloo.ca

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9882288

***Abstract*—In this paper, we investigate the packet routing problem in a multi-hop unmanned aerial vehicles (UAV) relay network, where multiple UAVs serve as relays between a base station (BS) and remote ground users (GUs) for enhancing network throughput. The challenges are that the dynamic network topology due to UAV mobility leads to volatile wireless connection. Moreover, there exists strong interference among UAVs due to line-of-sight communication links. Towards this end, we propose a novel multi-agent deep reinforcement learning based algorithm, named as MAQMIX for: 1) designing proper UAVs' trajectories to provide reliable network connection between the BS and remote GUs; 2) properly allocating frequency resource among UAVs to alleviate interference; 3) choosing a proper next-hop UAV for each data packet. Specifically, two training mechanisms are incorporated in the MAQMIX, i.e., intra UAV and inter UAV training mechanisms, which can tackle large action space issue and coordinate the training among UAVs, respectively. Simulation results show that the MAQMIX can significantly outperform baseline schemes in terms of transmission time and network throughput.**

I. Introduction

Recently, unmanned aerial vehicle (UAV)-assisted communication has attracted increasing attention from both industry and academia due to its multi-fold advantages [1]. First, UAVs can provide on-demand communication services due to their high manoeuvrability. For example, UAVs can be deployed to provide emergency communication services for natural disasters. Second, UAVs can provide line-of-sight (LOS) communication links to ground users (GUs), which can enhance network throughput [2]. Third, UAVs can act as data relays to connect remote users, thereby significantly enhancing network coverage in remote areas [3].

In this paper, we aim to facilitate communication services in remote areas, in which BSs are sparsely deployed and a number of remote GUs cannot connect to the BSs directly. As such, UAVs can be deployed as multiple relays for establishing network connection between these remote GUs and BSs to deliver data packets, i.e., multi-hop UAV relay network. However, the mobility of UAVs leads to the dynamic change of network topology. Due to the limited communication range, the connectivity between UAVs can be unreliable. As a result, the UAV trajectories and the next hop selection need to be well designed. In addition, since the LOS links among UAVs bring serious interference, frequency resource allocation should be considered to mitigate the interference. We investigate the packet routing problem to minimize the transmission time of the data packets and enhance network throughput through three types of decisions, i.e., UAV trajectory design, frequency resource allocation, and the next hop selection. This problem is quite challenging for its complexity.

Deep reinforcement learning (DRL) is a potential solution for such complex problem. DRL has been widely used in UAV-assisted communications. In [4], Liu *et al.* leverage DRL and propose an energy-efficient UAV control policy for fair communication coverage. In [5], an air-ground integrated network is studied, where the multi-user access control policy and UAVs' trajectories are jointly optimized using a multi-agent DRL (MADRL) framework for maximizing total throughput while guaranteeing the fairness among GUs. However, the above packet routing problem cannot be directly solved by traditional DRL algorithms due to the following reasons. First, each UAV needs to make the three types of decisions at the same time, leading to extremely large action space. Second, the objective of traditional DRL or MADRL is for each agent, while in this paper, the objective is to minimize the transmission time for each packet, which is determined by all UAVs. It is difficult to coordinate the UAVs to achieve this objective only through reward design.

In this paper, in order to tackle the large action space and coordinate the training among UAVs, we propose a novel MADRL algorithm, named as multi-agent QMIX (MAQMIX), which has two novel advantages. First, the MAQMIX has an intra UAV training mechanism where each UAV agent is decomposed into multiple subagents to tackle the large action space issue. Second, the MAQMIX has a novel inter UAV training mechanism, which coordinate the training among UAVs to deal with the inconsistency between the objective of minimizing transmission time of data packets and maximizing cumulative reward of each UAV agent. The simulation results have shown that MAQMIX can enhance the network throughput and reduce the transmission time significantly.

The remainder of this paper is organized as follows. Section II introduces the system model and formulates the optimization problem. Section III presents the proposed MAQMIX algorithm. Section IV provides simulation results, and the paper is concluded in Section V.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.

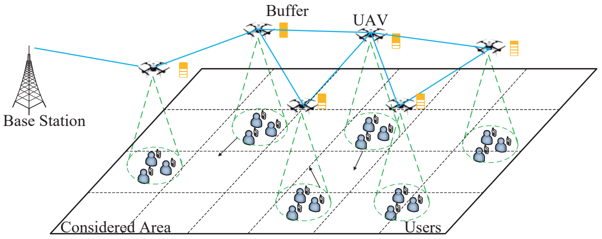

Fig. 1. Multi-hop UAV relay network.

II. System Model and Problem Formulation

A. Network Model

As shown in Fig. 1, we consider a multi-hop UAV relay network, in which multiple UAVs act as the data relays between remote GUs and a BS. The multi-hop UAV network consists of M UAVs, whose index set is denoted by $\mathcal{M} = \{1, 2, \cdots, M\}$. We consider a square area with a side length of L 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 in 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 $\boldsymbol{w}_m^u(t) \in \mathbb{R}^{2\times1}$, $0 \le t \le T$. There are K GUs in the network, whose index set is denoted by $\mathcal{K} = \{1, 2, \cdots, K\}$, which move with speed V^g and generate data packets in a Poisson distribution. The horizontal location of GU k is denoted by $\boldsymbol{w}_k^g(t) \in \mathbb{R}^{2\times1}$. The system works in a time slotted manner, i.e., $\mathcal{T} = \{1, 2, \cdots, T\}$. And the duration of each time slot is sufficiently short such that the moving directions of UAVs and GUs keep unchanged. The movement of the UAVs and GUs can be expressed via

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

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

where $\boldsymbol{e}_m^u(t)$ and $\boldsymbol{e}_k^g(t)$ are directions of UAV m and GU k.

B. Transmission Model between GU and UAV

At the beginning of each time slot, each GU associates the UAV with the strongest received signal strength (RSS). The associated UAV for GU k is denoted by $\chi_k^g(t)$. We introduce a binary variable $\mu_{k,m}^g(t)$ to indicate whether GU k is associated to UAV m to access at time slot t, $\sum_{m\in\mathcal{M}}\mu_{k,m}^g(t) = 1$. The number of GUs served by UAV m is $N_m^g(t) = \sum_{k\in\mathcal{K}}\mu_{k,m}^g$.

We adopt the probabilistic pathloss model for the communication link [6]. The probability of the LOS link between GU k and UAV m at time slot t can be approximated by

$$Pr_{k,m}^L(t) = \frac{1}{1+\eta_1\exp\left(-\eta_2\left(\arcsin\left(\frac{H}{d_{k,m}(t)}\right)-\eta_1\right)\right)}, \tag{2}$$

where η_1 and η_2 are two environment-related parameters, and $d_{k,m}(t)$ is the distance between GU k and UAV m. Then, the average air-to-ground (A2G) pathloss between GU k and UAV m at time slot t can be modeled as

$$PL_{k,m}(t) = PL_{FS}(t) + Pr_{k,m}^L(t)\times\eta_L + (1-Pr_{k,m}^L(t))\times\eta_N. \tag{3}$$

Here, $PL_{FS}(t) = 20\log d_{k,m}(t) + 20\log f_c + 20\log\left(4\pi/V^l\right)$, in which f_c and V^l denote the carrier frequency and the speed of light, respectively. η_L and η_N 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. 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

$$\alpha_{k,m}^g(t) = \frac{P^g}{10^{PL_{k,m}(t)/10}n_0\frac{B^g}{N_m^g(t)}}, \tag{4}$$

where n_0 is the noise power spectral density. Then the achievable transmission rate between GU k and UAV m at time slot t is

$$\zeta_{k,m}^g(t) = \begin{cases} \frac{\mu_{k,m}^g(t)B^g}{N_m^g(t)}\log_2(1+\alpha_{k,m}^g(t)), & \text{if } N_m^g(t) > 0, \\ 0, & \text{otherwise}. \end{cases} \tag{5}$$

At time slot t, GU k generates a data packet with a data size of $\sigma_k^g(t)$ and transmits it to UAV $\chi_k^g(t)$. The transmission time of the data packet is $\tau_k^g = \frac{\sigma_k^g(t)}{\zeta_{k,\chi_k^g(t)}^g(t)}$.

To reduce the packet loss probability, each UAV maintains a buffer to cache the received data packets, which follows the first-in-first-out rule. The buffer status of UAV m is denoted by $l_m(t)$. Once $l_m(t) < \sigma_k^g(t)$, the packet will be dropped due to insufficient buffer size, triggering a *network congestion* event.

C. Transmission Model between UAVs

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

$$\|h_{m,m'}^u(t)\|^2 = \frac{\rho_0}{\|\boldsymbol{w}_m^u(t) - \boldsymbol{w}_{m'}^u(t)\|^2}, \tag{6}$$

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

There are N_B orthogonal frequency bands with equal amount of bandwidth B^u, whose index set is denoted by $\mathcal{B} = \{1, \cdots, b, \cdots, N_B\}$. Each UAV should choose a proper frequency band to avoid interference. 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 band b. The communication range of a UAV is denoted by d^c. The signal-to-interference-plus-noise ratio (SINR) at UAV m' from UAV m can be expressed as

$$\alpha_{m,m'}^u(t) = \frac{P_{m,m'}^r(t)}{\sum_{i\in\mathcal{M}\backslash m'}\mu_{i,b_m(t)}^B(t)\mu_{i,m'}^c(t)P_{i,m'}^r(t) + n_0B^u}, \tag{7}$$

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.

where $\mu^c_{i,m'}(t) \in \{0,1\}$ is the communication range indicator with $\mu^c_{i,m'}(t) = 1$ indicating that UAV i is within the communication range of UAV m', and $\mu^c_{i,m'}(t) = 0$ otherwise. Then the transmission rate from UAV m to UAV m' is

$$R^u_{m,m'} = B^u \log_2(1 + \alpha^u_{m,m'}(t)). \tag{8}$$

The data size of the data packet from UAV m to UAV m' is σ^u, and the transmission time can be expressed as $\tau^u_{m,m'} = \sigma^u / R^u_{m,m'}$. The packet is transmitted in the multi-hop UAV relay network until it reaches UAV M (i.e., the destination). The number of hops traversed by the packet is denoted by N^u_σ, and the corresponding path is $\xi = \{m_0, m_1, \cdots, m_{N^u_\sigma}\}$, where $m_{N^u_\sigma} = M$ represents the destination. The overall transmission time along the path is $t = \sum_{n=0}^{N^u_\sigma - 1} \tau^u_{m_n,m_{n+1}}$.

D. Problem Formulation

We aim to minimize the transmission time of each packet and avoid the network congestion via optimizing UAVs' trajectories $\{\boldsymbol{w}^u_m(t)\}_{t\in\mathcal{T}}$, frequency resource allocation $b_m(t)$, and the next-hop UAV selection m'. Thus the problem can be formulated as

$$\begin{aligned}
\min_{\boldsymbol{e}^u_m, b_m, m_n \in \mathcal{M}} \quad & \sum_{n=0}^{N^u_\sigma - 1} \tau^u_{m_n,m_{n+1}} && \text{(9a)}\\
\text{s.t.} \quad & \|\boldsymbol{w}^u_m(t) - \boldsymbol{w}^u_{m'}(t)\| \le d^c, \forall t \in \mathcal{T} && \text{(9b)}\\
& \boldsymbol{w}^u_m(0) = \boldsymbol{w}^0_m, && \text{(9c)}\\
& l_{m'}(t) \ge \sigma^u, && \text{(9d)}\\
& l_m(t) \ge \sigma^g_k(t), && \text{(9e)}\\
& \boldsymbol{\psi}_l \le \boldsymbol{w}^u_m(t) \le \boldsymbol{\psi}_u, \forall t \in \mathcal{T}. && \text{(9f)}
\end{aligned}$$

Problem (9) is challenging to be solved since 1) UAVs should make decisions on their trajectories, frequency resource allocation and the next-hop UAV at the same time, and hence the action space of each UAV is very large; 2) the three types of actions are coupled each other which further complicates the joint decision making process; and 3) the objective function of problem (9) is for each data packet, while the decision-making agents are the UAVs, thus maximizing the cumulative reward can not solve the problem (9).

III. MAQMIX Algorithm

A. Preliminaries

We first give a brief introduction to DRL and QMIX [7]. In a *single agent* DRL algorithm, 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

$$Q(s(t), a(t)) = \mathbb{E}_\pi \left[\sum_{i=0}^{T} \gamma^i r(t+i+1) \middle| s(t), a(t) \right], \tag{10}$$

which is the expectation of the discounted cumulative reward.

Take the deep Q network (DQN) [8] 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. To tackle instability issues caused by applying DNN in DRL, experience replay and target network are used in DQN. The DNN is trained via minimizing the loss function, i.e.,

$$L(\theta) = \frac{1}{N_s} \sum_i [y(i) - Q(s(i), a(i); \theta)]^2, \tag{11}$$

where

$$y(i) = r(i) + \gamma Q'(s(i+1), \pi(a|s(t+1)); \theta'), \tag{12}$$

is the training target.

In *multi-agent* scenarios with N_a agents, each agent has its own state-action value function $Q_i(o_i, a_i; \theta_i)$. The joint action-value function ensures that global $\arg\max$ performed on Q^{tot} yields the same result as a set of individual $\arg\max$ operations performed on each Q_i, i.e.,

$$\arg\max_{\boldsymbol{a}} Q^{tot}(\boldsymbol{o}, \boldsymbol{a}) = \begin{pmatrix} \arg\max_{a_1} & Q_1(o_1, a_1) \\ & \vdots \\ \arg\max_{a_{N_a}} & Q_{N_a}(o_{N_a}, a_{N_a}) \end{pmatrix}, \tag{13}$$

which means the choice of greedy action for each agent in a decentralized way can lead to greedy joint actions. In order to satisfy (13), QMIX uses a mixing network with non-negative weights to input the state-action value of each agent Q_i and output Q^{tot} such that QMIX can have the monotonicity:

$$\frac{\partial Q^{tot}}{\partial Q_i} \ge 0, \forall i \in \{1, 2, \cdots, N_a\}. \tag{14}$$

B. Key Elements of the Proposed DRL Algorithm

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

- The locations of all GUs $\{\boldsymbol{w}^g_k(t)\}_{k\in\mathcal{K}}$: We map the GU locations $\{\boldsymbol{w}^g_k(t)\}_{k\in\mathcal{K}}$ to a GU distribution matrix $\boldsymbol{W}^g(t) \in \mathbb{R}^{N_c \times N_c}$ in which $\boldsymbol{W}^g_{i,j}$ represents the number of GUs in the corresponding grid;
- The channel gain between UAV m and other UAVs $\{\|h^u_{m,m'}(t)\|^2\}_{m'\in\mathcal{M}\setminus\{m\}}$;
- The UAVs' locations $\{\boldsymbol{w}^u_m(t)\}_{m\in\mathcal{M}}$;
- The buffer status of all UAVs $\{L_m(t)\}_{m\in\mathcal{M}}$;
- The number of UAVs that select each frequency band $\{N^f_b\}_{b\in\mathcal{B}}$;
- The data size of the current data packet σ^u.

2) Action: The action space of each UAV $a_m(t)$ consists of three parts:

- The flying direction $\boldsymbol{e}^u_m(t)$, including left, right, forward, backward, and hover, i.e., $\{(-1,0), (1,0), (0,1), (0,-1), (0,0)\}$;
- The frequency resource allocation $b_m(t)$;
- The next-hop UAV selection m'.

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

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.

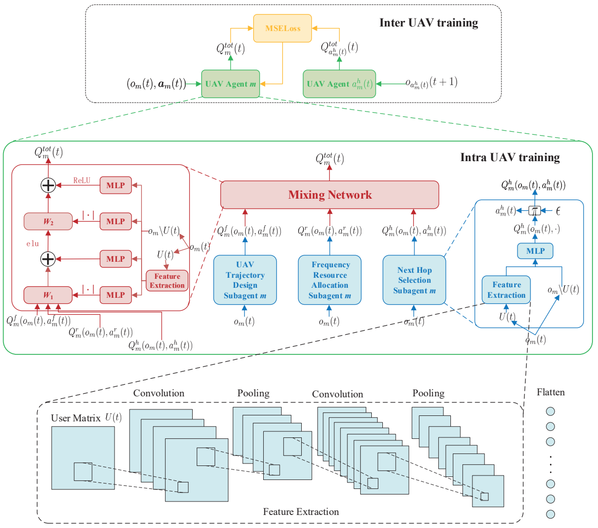

Fig. 2. The structure of MAQMIX.

3) Reward: The reward of each UAV consists of five parts:

- The transmission time of the current hop $\tau^u_{m,m'}$, which follows the objective function;
- The constant penalty reward for flying out of the considered area r^s_m;
- The constant penalty reward r^e_m for the network congestion that occurs in the next hop;
- The congestion penalty for the congestion caused by the packets from GUs (9e): If the packet generated from GU k exceeds the buffer size of UAV m, the UAVs within the distance d^ϕ from GU k will receive a penalty $\{r^\phi_i\}_{i \in \mathcal{M} \backslash m}$, whose value is based on the distance between UAVs and GU k, i.e.,

$$r^\phi_i = r^\beta - \mu^d_{i,k} \kappa_d \|\boldsymbol{w}^u_i - \boldsymbol{w}^g_k\|. \tag{15}$$

  Here r^β is the bonus for encouraging UAVs to fly close to the congested UAV to offload the traffic, κ_d is a positive weight for balancing reward and distance and $\mu^d_{i,k}$ is the range indicator;
- The outage penalty reward for the selected UAV exceeding the communication range r^c_m.

The overall reward of UAV m is defined as

$$r_m = -\kappa_t \tau^u_{m,m'} + \mu^s_m r^s_m + \mu^e_m r^e_m + r^\phi_m + \mu^c_m r^c_m, \tag{16}$$

where κ_t is a positive adjusting weight like κ_d, while μ^s_m, μ^e_m and μ^c_m are the service region, UAV hop congestion, and communication range out indicators respectively.

C. Intra UAV Training Mechanism

As shown in Fig. 2, to tackle large action space issue, each UAV agent runs the QMIX algorithm. Specifically, each UAV consists of three subagents, i.e., trajectory design subagent, resource allocation subagent, and the next hop selection subagent. As such, the large action space is decomposed into three small action spaces, which simplifies the training complexity. The greedy policy of UAV agent m is

$$\begin{aligned}\pi_m(a_m|o_m) =& \{\arg\max_a Q^f_m(o_m, a), \arg\max_a Q^r_m(o_m, a),\\ & \arg\max_a Q^h_m(o_m, a)\}\end{aligned} \tag{17}$$

where $Q^f_m(\cdot)$, $Q^r_m(\cdot)$, and $Q^h_m(\cdot)$ are the state-action value functions of the subagents for trajectory deisgn, frequency resource allocation, and the next-hop UAV selection, respectively. According to (13), we can get

$$\arg\max_a Q^{tot}_m(o_m, a) = \pi_m(a_m|o_m). \tag{18}$$

D. Inter UAV Training Mechanism

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

We use the value of the next-hop UAV to calculate the training target, which is easy to obtain through information exchange among UAVs [9]. The training target defined in (12) is converted to

$$y_m(i) = r_m(i) + \gamma Q^{tot}_{m'}(o_{m'}(i), \pi_{m'}(a_{m'}|o_{m'}); \theta^Q_{m'}). \tag{19}$$

Note that after sufficient training, the agents hardly receive the penalties defined in Section III-B3. The state-action value of each agent Q^{tot}_m can be regarded as an estimate of the transmission time. Then (19) can be regarded as the summation of transmission time of the current hop and the subsequent required transmission time estimated by the next-hop UAV. Thus we can use the estimated value of the next-hop UAV to calculate the training target of the current UAV.

E. Network Architecture

As shown in Fig. 2, each UAV agent has three subagents. In each subagent, the GU distribution matrix is first input into the feature extraction. A multilayer perceptron (MLP) inputs the extracted features and the remaining observations and outputs the state-action values of all possible actions. The subagents choose actions using ϵ-greedy policy.

The mixing network inputs the three chosen state-action values and outputs the total value Q^{tot}_m. The weights of the mixing network are produced by the hypernetwork. Similarly, the hypernetwork takes the observations as input into the MLPs. The outputs of the MLPs are activated by the absolute activation function and are considered as the weights of the mixing network.

The detailed MAQMIX is listed in Algorithm 1.

IV. Performance Evaluation

A. Simulation Settings

We set the the episode length as $T = 1,000$ and the side length of the square area as $L_s = 1$ km. UAVs fly at a altitude $H = 300$ m. There are $M = 10$ UAVs and $K = 50$ GUs in the network. The speed of UAVs and GUs are $V^u = 10$ m/s, $V^g = 5$ m/s, respectively. The data size of the generated data

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Training Phase of MAQMIX

```
Randomly initialize the networks of all UAV agents;
Initialize the experience replay buffer;
for each episode do
    for each time slot $t$ do
        The GUs generate data packets and transmit them to the corresponding UAVs;
        for each UAV $m$ do
            The agent $m$ gets the observation $o_m(t)$;
            The three subagents take the observation as input, and output the corresponding actions via $\epsilon$-greedy;
        end for
        All agents take actions;
        The agents obtain the rewards $\{r_m(t)\}_{m\in\mathcal{M}}$ and the next observations $\{o_m(t+1)\}_{m\in\mathcal{M}}$;
        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;
    end for
    Sample $N_s$ experience tuples randomly;
    for each agent $m$ do
        Update the network by minimizing the loss (11);
    end for
end for
```

TABLE I
COMMUNICATION RELATED PARAMETERS

Notation	Meaning	Value
n_0	Noise power spectral density	10^{-17} W/Hz
P^g	Transmission power of GUs	0.01 W [10]
P^u	Transmission power of UAVs	0.03 W
η_1	Parameter of the channel model	4.88
η_2	Parameter of the channel model	0.43
η_{LOS}	Parameter of the channel model	0.1 dB
η_{NLOS}	Parameter of the channel model	21 dB
B^g	Bandwidth for UAV to serve GUs	2×10^5 Hz
B^u	Bandwidth among UAVs	1×10^6 Hz

packet follows a Poisson distribution with parameter $\lambda = 20$. The communication related system parameters are summarized in Table I.

In the proposed MAQMIX algorithm, we divide the service area into 10×10 girds whose side length is $L_g = 100$ m, and utilize 2 CNN layers followed by pooling layers to extract the features of GUs' distribution. For each subagent, we adopt a fully-connected network with 5 layers, and each hidden layer is activated by $ReLU$ functions. The mixing network consists of hypernetwork layers with embedding dimension of 64, the filters and bias of which are generated from the states through a single linear layer followed by an absolute activation function. The parameters of reward in (16) are listed in TABLE II. We set the random exploration possibility $\epsilon = 0.1$, and the learning rate is set to $L_r = 0.001$. For each learning step, each agent randomly samples $N_s = 32$ experiences for training. The capacity of the experience replay buffer is set to $|\mathcal{C}| = 15,000$.

TABLE II
REWARD PARAMETERS

Notation	Value
κ_t	100
κ_d	0.01
r^ϕ	5
r_m^s	-50
r_m^c	-50
r_m^e	-50
d^ϕ	800

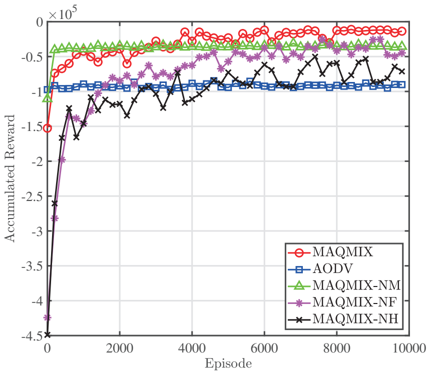

Fig. 3. Cumulative reward performance of all agents in terms of episodes.

B. Simulation Results

We compare the proposed MAQMIX algorithm with the baseline AODV algorithm [11]. In the AODV, the locations of all UAVs are fixed, and the frequency resources are randomly allocated. Besides, in order to verify the effectiveness of each subagent, we propose three other benchmarks which remove one of the three subagents respectively and only optimize the other two subagents:

- MAQMIX-NM: The locations of all UAVs are fixed and the other two 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. As shown in Fig. 3, 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 AODV. At the beginning of the training process, the cumulative rewards of MAQMIX-NF and MAQMIX-NH are much lower than that of 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 disconnectivity of the chosen next hop, leading to the penalty. MAQMIX-NM converges much faster than others. The reason is that the frequency resource allocation

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.

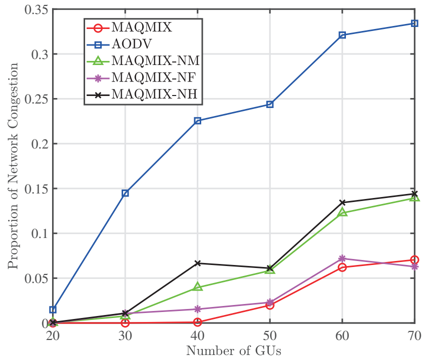

Fig. 4. Proportion of network congestion versus the number of GUs.

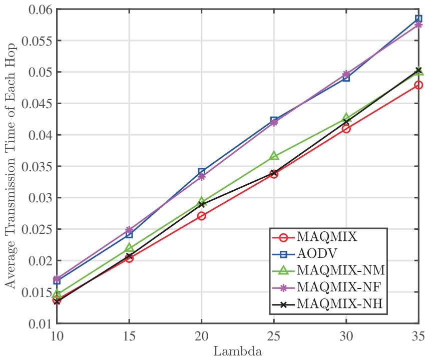

Fig. 5. Average transmission time of each hop versus the value of λ.

and the next hop selection subagents are much easier to train compared with trajectory design subagent. The cumulative reward of MAQMIX-NH is lower than that of MAQMIX-NM and MAQMIX-NF after convergence, because the dynamic network topology caused by trajectory design subagent deteriorates the performance of the shortest path routing protocol.

Figure 4 illustrates the impact of the number of GUs on the proportion of network congestion. It can be observed that in MAQMIX algorithm, the proportion 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, MAQMIX can reduce the probability of network congestion by 91.88% compared with AODV when the number of GUs is $K = 50$. This is because AODV is a routing protocol using the shortest path criterion, while MAQMIX can learn from the historical experience to avoid network congestion by choosing the transmission path properly. We can also observe that MAQMIX-NF achieves similar congestion reduction as MAQMIX. This is because the network congestion is mainly affected by the next hop selection and the trajectory design rather than the frequency resource allocation.

In Fig. 5, it can be seen that the average transmission time of each hop increases monotonically with the increment of λ. This is because the transmission time of each hop increases linearly with the packet size, while the transmission rate between UAVs is barely affected by packet size. In MAQMIX algorithm, the average transmission time reduces by 20.76% compared with AODV.

V. Conclusion

In this paper, we have investigated the packet routing problem in the multi-hop UAV relay network to minimize the transmission time and enhance the network throughput. We have proposed a novel MAQMIX algorithm to enhance the network throughput and shorten the transmission time by leveraging the intra training mechanism to tackle the large action space issue and the inter training mechanism to coordinate the training among UAVs. The proposed MAQMIX can be used to solve the problems with large action space and that the optimization objective is jointly determined by all agents. For future work, we will investigate the store-carry-forward routing scheme in the multi-hop UAV relay network.

Acknowledgement

This work is supported by National Key Research and Development Program of China (2018AAA0102401), and Natural Sciences and Engineering Research Council of Canada (NSERC). *(Corresponding author: Jun Liu)*

References

[1] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, “AI-assisted network-slicing based next-generation wireless networks,” *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 45–66, Jan. 2020.

[2] 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.

[3] S. Fu, Y. Tang, Y. Wu, N. Zhang, H. Gu, C. Chen, and M. Liu, “Energy-efficient uav-enabled data collection via wireless charging: A reinforcement learning approach,” *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10 209–10 219, 2021.

[4] 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.

[5] 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.*, 2021, doi:10.1109/JIOT.2021.3062091.

[6] 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.

[7] 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. ICML*, 2018, pp. 4295–4304.

[8] 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*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[9] 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. ICNC*, 2020, pp. 932–937.

[10] 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.

[11] C. Perkins, E. Belding-Royer, and S. Das, “RFC3561: Ad hoc on-demand distance vector (AODV) routing,” United States, 2003.

Authorized licensed use limited to: BEIHANG UNIVERSITY. Downloaded on May 31,2023 at 09:01:12 UTC from IEEE Xplore. Restrictions apply.