# MAPPO-Based Cooperative UAV Trajectory Design with Long-Range Emergency Communications in Disaster Areas

Kai Li is also an organizing committee member for The 1st IEEE Workshop on Wireless outdoor, Long-Range and Low-Power Networks (WOLOLO).

**Yue Guan**

**Sai Zou**

**Bochun Wu**

**Kai Li***

**Wei Ni**

*CISTER Research Centre

CISTER-TR-230401

2023/06/12

# MAPPO-Based Cooperative UAV Trajectory Design with Long-Range Emergency Communications in Disaster Areas

Yue Guan, Sai Zou, Bochun Wu, Kai Li*, Wei Ni

*CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail: yueguanee@foxmail.com, dr-zousai@foxmail.com, wubochun@fudan.edu.cn, kai@isep.ipp.pt, Wei.Ni@data61.csiro.au

https://www.cister-labs.pt

## Abstract

This paper investigates the cooperative real-time trajectory design issue for multiple unmanned aerial vehicles (UAVs) to support long-range emergency communications in disaster areas. To quickly restore the communication links between mobile users (MUs) and base stations, UAVs equipped with a radio frequency (RF) module and a free space optical (FSO) module serve as relay nodes. Given the difficulty of setting up a central controller for UAVs and the urgency of emergency communication, the UAV trajectory design issue is formulated as a distributed cooperative optimization problem. A collaborative multi-UAV trajectory design method based on multi-agent proximal policy optimization (MAPPO) is adopted to improve RF/FSO channel throughput. Compared with the state-of-the-art DRL methods, MAPPO can achieve higher RF allocation efficiency and increase the FSO communication backhaul capacity.

# MAPPO-Based Cooperative UAV Trajectory Design with Long-Range Emergency Communications in Disaster Areas

Yue Guan[1], Sai Zou[1,*], Bochun Wu[2], Kai Li[3], Wei Ni[4]
[1]College of Big Data and Information Engineering, Guizhou University, China ({yueguanee, dr-zousai}@foxmail.com)
[2]Fudan University, China (wubochun@fudan.edu.cn)
[3]CISTER Research Centre, Portugal (kaili@ieee.org)
[4]CSIRO, Australia (Wei.Ni@data61.csiro.au)

***Abstract*—This paper investigates the cooperative real-time trajectory design issue for multiple unmanned aerial vehicles (UAVs) to support long-range emergency communications in disaster areas. To quickly restore the communication links between mobile users (MUs) and base stations, UAVs equipped with a radio frequency (RF) module and a free space optical (FSO) module serve as relay nodes. Given the difficulty of setting up a central controller for UAVs and the urgency of emergency communication, the UAV trajectory design issue is formulated as a distributed cooperative optimization problem. A collaborative multi-UAV trajectory design method based on multi-agent proximal policy optimization (MAPPO) is adopted to improve RF/FSO channel throughput. Compared with the state-of-the-art DRL methods, MAPPO can achieve higher RF allocation efficiency and increase the FSO communication backhaul capacity.**

***Index Terms*—unmanned aerial vehicle, trajectory optimization, long-range, free space optical communication, radio frequency, multi-agent proximal policy optimization**

## I. Introduction

When earthquakes, tsunamis, flash floods, and other disasters occur, communication facilities may be damaged, and then emergency communication becomes dominant. The research on emergency communication has attracted the joint attention of academia and industry [1]–[3]. Considering the destructiveness of ground roads and the urgent emergency communication guarantees when disasters occur, long-distance communication for irregular mobile users (MUs) in disaster areas has become difficult. Free space optical (FSO) communication uses laser light waves as the carrier wave and the atmosphere as the transmission medium without laying optical fibers, providing large communication capacity and high-speed, long-distance transmission. Unmanned aerial vehicles (UAVs) are flexible and can fly quickly and accurately to a given location. How to use the UAV equipped with an FSO module (connecting with the ground base station) and a radio frequency (RF) module (access in a disaster area) as an emergency communication tool in a disaster area is of great significance [4], [5].

The introduction of UAVs in emergency communications has enormous advantages. An optimal location deployment method for UAVs is proposed to optimize unevenly distributed access and area coverage in disaster areas [6]. However, autonomous rescues in disaster areas tend to move erratically while search and rescue personnel are organized, and these become highly spatio-temporal dynamic characteristics of access [7]. The cooperative pursuit of multiple UAVs in a dynamic environment is a complex problem.

As a branch of artificial intelligence, deep reinforcement learning (DRL) is often used for decision-making in complex situations. It also has many contributions to the field of UAV deployment. DRL has many advantages, which can make multi-step decisions to obtain optimal rewards in dynamic environments [8]. However, single DRL is mainly used for dynamic environments in low-dimensional action and state spaces [9]. Multi-agent DRL can work together in global and high-dimensional state space [10], especially multi-agent proximal policy optimization (MAPPO) algorithm [11]. Can the MAPPO algorithm be introduced to alleviate the dimensional explosion of the state and action space of UAVs' cooperative work in emergency communication?

This paper studies the trajectory optimization and resource management of UAVs in disaster areas emergency communication, where UAVs equipped with RF/FSO serve MUs as relay base stations. To address the complexity, high-dimensionality and urgency of emergency rescue, we introduce a multi-DRL method, i.e., MAPPO, for multi-UAV cooperative emergency communication.

## II. Related WORK

The use of UAVs to assist ground base stations (GBSs) in covering on-signal areas has gained significant attention globally in recent years, as noted in several studies [12], [13]. Additionally, free-space optical (FSO) communication for high-speed point-to-point communication has also received tremendous attention, as demonstrated in [14], which explored the use of FSO in quickly establishing backhaul between UAVs and MUs in disaster areas.

[*] Corresponding author

This work is supported by the Social Science Foundation of China (21VSZ126), Natural Science Foundation of Guizhou (ZK[2022]-162), and Natural Science Foundation of Guizhou University (X2021167).

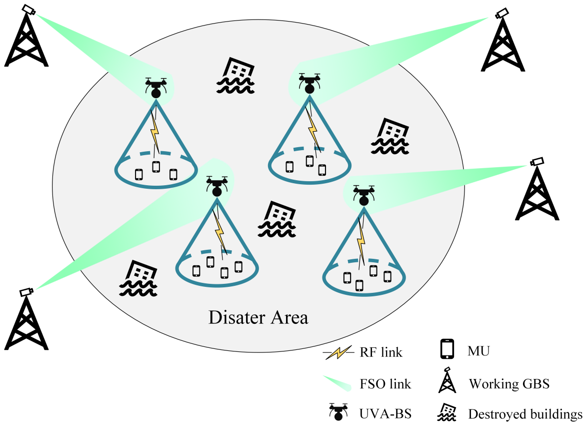

Fig. 1. Emergency communications in urban environments, where multiple UAVs cooperate to connect GBSs and MUs.

DRL is a popular ML technique used in solving the decentralized partially observable Markov decision process (Dec-POMDP) that arises in optimizing UAV trajectories for providing communication to MUs. This approach was applied in several studies [15]–[17], such as Qin et al.'s multi-agent DRL method for UAV trajectory design that aims to maximize throughput and ensure fair communication service to MUs [16]. In another study, UAV trajectory optimization is performed using a multi-agent reinforcement learning algorithm that considers the data that UAVs can cache [17].

Although several studies have focused on UAV trajectory optimization, none of them have addressed the cooperation problem of FSO/RF hybrid communication UAVs. This paper employs a multi-agent DRL method to design a trajectory optimization problem for a hybrid FSO/RF communication UAV facing dynamic MUs.

## III. SYSTEM MODEL

In this section, the system model (including the emergency communication scenario, RF channel model, MU mobility model, and FSO channel model) is introduced.

### A. Scenario Description

The scene of multi-UAV cooperation in disaster area emergency rescue is shown in Fig. 1. This system consists of $I$ UAVs, $J$ MUs, and $K$ GBSs, collected in sets $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{K}$, respectively. The GBSs are distributed around the disaster area. The MUs are randomly located with Poisson distribution. The UAVs provide RF communication services to the MUs. The UAVs can communicate with distant GBSs through an FSO channel.

In this study, we consider a slotted system with $T$ identical slots, as indexed by $t \in \mathcal{T} = \{1, 2, \cdots, T\}$. At the beginning of each time slot $t$, each MU $j \in \mathcal{J}$ moves randomly and requests radio resources. We assume a quasi-static environment where network state information remains unchanged during each time slot. We also assume disaster areas in the scenario of urban areas, as proposed by ITU-R [18].

### B. RF Channel Model

The RF channel between a UAV and an MU is a probabilistic line-of-sight (LoS) channel [19]. Between UAV $i$ and MU $j$, let $L_{i,j}^{\mathrm{LoS}}$ and $L_{i,j}^{\mathrm{NLoS}}$ denote the path losses in LoS and NLoS scenarios, respectively. They can be expressed as

$$L_{i,j}^{\mathrm{LoS}}\left(d_{i,j}\right) = B^{\mathrm{LoS}} + \gamma^{\mathrm{LoS}} \log d_{i,j} + G \tag{1}$$

and

$$L_{i,j}^{\mathrm{NLoS}}\left(d_{i,j}\right) = B^{\mathrm{NLoS}} + \gamma^{\mathrm{NLoS}} \log d_{i,j} + G, \tag{2}$$

where $d_{i,j}$ denotes the distance between UAV $i$ and MU $j$; $B^{\mathrm{LoS}}$ and $B^{\mathrm{NLoS}}$ denote the path losses at *reference distance* $d_{i,j} = 1$; $\gamma^{\mathrm{LoS}}$ and $\gamma^{\mathrm{NLoS}}$ denote the path loss exponents of LoS and NLoS transmissions, respectively; $G$ is a standard Gaussian random variable, i.e., $G \sim N(0, 1)$.

Given altitude $h_i$ for UAV $i$, we can define the *elevation angle* between UAV $i$ and MU $j$ as $\varphi_{i,j} = \sin^{-1}(\frac{h_k}{d_{i,j}})$. The LoS probability [20] is given by

$$\rho_{i,j}^{\mathrm{LoS}}\left(\varphi_{i,j}\right) = \frac{1}{1 + \eta \exp[-\beta\left(\varphi_{i,j} - \eta\right)]}, \tag{3}$$

where $\eta$ and $\beta$ are environmental parameters w.r.t. the urban environmental deployment model. On the other hand, the NLoS probability is given by

$$\rho_{i,j}^{\mathrm{NLoS}}\left(\varphi_{i,j}\right) = 1 - \rho_{i,j}^{\mathrm{LoS}}\left(\varphi_{i,j}\right). \tag{4}$$

To sum up, the statistical path loss jointly with LoS and NLoS probabilities can be calculated as

$$L_{i,j}^{\mathrm{avg}} = L_{i,j}^{\mathrm{LoS}} \rho_{i,j}^{\mathrm{LoS}} + L_{i,j}^{\mathrm{NLoS}} \rho_{i,j}^{\mathrm{NLoS}}. \tag{5}$$

The signal-to-interference-plus-noise ratio (SINR) for the RF channel between UAV $i$ and MU $j$ is given by

$$\Gamma_{i,j} = \frac{P_{i,j}^{r} 10^{-\frac{L_{i,j}^{\mathrm{avg}}}{10}}}{\sigma^2 + \sum_{m=1, m\neq i}^{N_{\mathrm{uav}}} P_{m,j}^{r} 10^{-\frac{L_{m,j}^{\mathrm{avg}}}{10}}}, \tag{6}$$

where $\sigma^2$ is the additive white Gaussian noise (AWGN) power, $N_{\mathrm{uav}}$ is the number of UAVs, and $P_{i,j}^{r}$ is the transmission power of the UAV, $\sum_{m=1, m\neq i}^{N_{\mathrm{uav}}} P_{m,k,j}^{r} 10^{-\frac{L_{m,j}^{\mathrm{avg}}}{10}}$ is the aggregate interference power. The UAV communicates with MUs to ensure a high SINR. In case an MU connects each UAV with SINR lower than threshold $\delta$, the MU is considered as *off-state* [21]. The constraint on SINR is given by

$$SINR(\,\Gamma_j) < \delta\,. \tag{7}$$

Thus, the achievable rate of MU $j$ can be written as

$$S_{i,j} = \frac{B_i}{N_{\mathrm{mu},i}} \log_2 \left(1 + \frac{P_{i,j}^{r} 10^{-\frac{L_{i,j}^{\mathrm{avg}}}{10}}}{\sigma^2 + \sum_{m=1, m\neq i}^{N_{\mathrm{uav}}} P_{m,j}^{r} 10^{-\frac{L_{m,j}^{\mathrm{avg}}}{10}}}\right), \tag{8}$$

where $B_i$ is the channel bandwidth of UAV $i$, $N_{\mathrm{mu},i}$ is the number of MUs connected to UAV $i$, and $N_{\mathrm{uav}}$ is the number of UAV. Here, we assume that MUs evenly divide the channel bandwidth.

### C. MU Mobility Model

We assume that MUs walk on a continuous plane while having position $(x(t), y(t))$ and velocity $(V_x(t), V_y(t))$ at time slot $t$. The MU moves freely within the disaster area while position and velocity are statistically independent. The introduction of fluid-dynamics to describe the crowd flow has been widely used [22]. Here Maxwell-Boltzmann distribution is introduced to describe the velocity of MU. The probability density function, $P(V_x)$, for a single velocity component, $V_x$, of MUs' walking velocity is given by

$$P\left(V_x\right) = \frac{1}{\sqrt{2\pi} v_{r.m.s.}} \exp\left(-\frac{V_x^2}{2v_{r.m.s.}^2}\right), \tag{9}$$

where $v_{r.m.s.}$ is the root-mean-square of the velocity $v \equiv |V|$. $V_y$ and $V_x$ have the same representation, and their combined velocity probability density function is given by

$$P\left(V\right) = \frac{1}{2\pi v_{r.m.s.}^2} \exp\left(-\frac{V^2}{2v_{r.m.s.}^2}\right). \tag{10}$$

### D. FSO Channel Model

Since the FSO transmitting power of the GBS is much higher than that of the UAV, the main consideration here is the limitation of the backhaul rate of the UAV to the GBS on the FSO channel [23]. The backhaul rate $\kappa$ of the FSO connection between a UAV and a GBS can be expressed respectively as

$$\kappa = \frac{P_t \xi_t \xi_{atm} \phi^2}{\pi(\varkappa_t/2)^2 d_{i,k}^2 \vartheta_p Z_b} \tag{11}$$

and

$$\xi_{atm} = 10^{\frac{-\varrho d_{i,k}}{10}}, \tag{12}$$

where $P_t$ is the FSO transmit power of the UAV; $\xi_t$ denotes the optical efficiencies of the transmitter and receiver; $\xi_{atm}$ denotes the value of the atmospheric transmission at the laser transmitter wavelength; $\phi$ stands for the UAV receiver FSO beam diameter; $\varkappa_t$ is the transmitter divergence; $\vartheta_p = \hbar c/\lambda$ is the photon energy. While $\hbar$ is Planck's constant, $c$ is the speed of light, $\lambda$ stands for the transmission wavelength; and $Z_b$ is the mean receiver sensitivity in the number of photons$/b$. The distance between UAV $i$ and GBS $k$, $d_{i,k}$, and the atmospheric attenuation coefficient (dB/km), $\varrho$, are given by

$$d_{i,k} = \sqrt{(d_{i,k}^{\mathrm{h}})^2 + \left(h_i - h_k\right)^2} \tag{13}$$

and

$$\varrho = \frac{3.91}{\nu}\left(\frac{\lambda}{550nm}\right)^{-Q}, \tag{14}$$

where $d_{i,k}^{\mathrm{h}}$ is the horizontal distence between UVA $i$ and GBS $k$; $h_k$ is the altitude of the GBS; $\nu$ represents the visibility in kilometers; $Q$ is the environmental quality function [24]; and $\lambda$ is the wavelength of the transmitted signal. The relationship between different degrees of visibility and $Q$ is given by

$$Q = \begin{cases} 1.6, & \text{if} \quad \nu > 50 \text{ km}, \\ 1.3, & \text{if} \quad 6 \text{ km} < \nu < 50 \text{ km}, \\ 0.16\nu + 0.34, & \text{if} \quad 1 \text{ km} < \nu < 6 \text{ km}, \\ \nu - 0.5, & \text{if} \quad 0.5 \text{ km} < \nu < 1 \text{ km}, \\ 0, & \text{if} \quad \nu < 0.5 \text{ km}. \end{cases} \tag{15}$$

As a result, the backhaul rates $R$ of the UAVs to the GBSs under different visibility levels can be calculated.

## IV. MAPPO-BASED UAV TRAJECTORY DESIGN

### A. Problem Transformation

Considering the dynamic characteristics of the environment, the distribution of the UAVs, and the locality of the UAVs' observations, we formulate the UAVs trajectory optimisation problem as a Dec-POMDP problem. Dec-POMDP is a mathematical framework for multi-agent decision-making problems. In emergency communication scenarios, UAVs are regarded as distributed agents which execute flight strategies based on their local observations. The observations, states, actions, and rewards of this Dec-POMDP at time $t$ are defined as $O$, $S$, $A$, and $R$. The detailed definitions of Dec-POMDP elements are given as follows.

**Observation** $O(t)$**:** At time slot $t$, the agent collects the environmental information within the observation range, which includes the position and speed of the UAV, the position and speed of the MUs, and the position information of the GBSs in time slot $t$, so the observation $O(t)$ is defined as

$$\begin{aligned} O(t) = \{&u(t), m_1(t), m_2(t), ..., m_j(t), g_1, g_1, ..., g_k, \\ &u^v(t), m_1^v(t), m_2^v(t), ..., m_j^v(t)\}, \end{aligned} \tag{16}$$

where $u(t) = (x_{\mathrm{uav}}^t, y_{\mathrm{uav}}^t, h_{\mathrm{uav}}^t)$ and $m(t) = (x_{\mathrm{mu}}^t, y_{\mathrm{mu}}^t)$ represents the position of UAV and MUs in time slot $t$, respectively. Because the position of the GBS is constant, $g = (x_{gbs}, y_{gbs})$ defines the position of the GBS. In addition, $u^v(t)$ and $m^v(t)$ represent the dynamic speed of the UAV and the MU, respectively.

**State** $S(t)$**:** The system state is composed of the states of all UAVs, MUs, and GBSs in the environment, and the positions of the UAVs and MUs can change over time. Thus, the state space $S(t)$ at slot $t$ can be defined as

$$\begin{aligned} S(t) = \{&u_1(t), u_2(t), ..., u_{N_{\mathrm{uav}}}(t), m_1(t), m_2(t), ..., m_{N_{\mathrm{mu}}}(t), \\ &g_1, g_1, ..., g_{N_{gbs}}, u_1^v\left(t\right), u_2^v\left(t\right), ..., u_{N_{\mathrm{uav}}}^v(t), \\ &m_1^v(t), m_2^v(t), ..., m_{N_{\mathrm{mu}}}^v(t)\}. \end{aligned} \tag{17}$$

**Action** $A(t)$**:** The action defines the value that guides the agent to act according to the policy function in the MDP. The action space of UAVs is defined as

$$A(t) = \{\varpi_i, \psi_i \mid \varpi \in [0,1], \psi_i \in [-180°, 180°]\} \; i \in I, \tag{18}$$

where $\varpi$ defines the flight radius of the UAV, $\psi$ defines the flight angle of the UAV. Each UAV maintains a fixed altitude for a time slot $t$ and flies within a continuous radius and angle.

**Reward** $R(t)$**:** The reward is what the agent gets in return after interacting with the environment. In the problem of UAV communication rescue in disaster-stricken areas, each UAV has the same purpose: to maximise the MU throughput, the backhaul rate to GBS, and the number of connected MUs. Based on (7), (8), and (11), the rewards for all UAVs at time slot $t$ are expressed as

$$R(t) = \left( \sum_{i=1}^{N_{\text{uav}}} \sum_{j=1}^{N_{\text{mu},i,j}} S_{i,j} + \sum_{i=1}^{N_{\text{uav}}} \chi\kappa \right) \exp(\frac{\Upsilon}{\Theta_{\text{mu}}}), \quad (19)$$

where $\Theta_{\text{mu}}$ represents the number of disconnected MUs, and there is a formula $\Theta_{\text{mu}} = N_{\text{mu}} - N_{\text{mu},j}$ to calculate, $N_{\text{mu},i,j}$ is the number of MUs provided with communication services by UAV $i$; $\chi$ and $\Upsilon$ represent the impact factors applied to balance rewards, where $\chi$ is fixed, and $\Upsilon$ is positively correlated with the number of MUs.

### B. MAPPO

To solve the Dec-POMDP constructed above, a centralised value function and a distributed policy function are adopted in MAPPO, which can achieve distributed execution while obtaining the maximum reward.

Maximizing the expectation of rewards is necessary to solve the trajectory optimization problem for UAVs. The Actor selects actions according to a policy $\pi$, and another component Critic evaluates the value of the selected actions. The global UAV rewards are accumulated as

$$r_t(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T_n} R(\tau^i) \pi_\theta(\alpha_t^i | s_t^i), \quad (20)$$

where $\tau$ is the sequence of experiences of the UAV interacting with the environment, which is stored in the experience replay area $\mathcal{D}_k = \tau^i$; $\sum_{t=1}^{T_n} \pi_\theta(\alpha_t^i, s_t^i)$ is the probability that a sequence $\tau^i$ occur. The recent reward is more important in the dynamic interaction between the UAV and the environment, so there is a discounted reward $R(\tau) = \sum_{t=1}^{T_n} \mu^t r$, and $\mu$ represents the discount factor. To maximise the agent's cumulative reward, the policy network parameter $\theta$ is updated by gradient, which can be expressed as

$$\nabla \hat{r}_\theta = E_{\tau \sim \pi_\theta(\tau)} [A^\theta(s_t, \alpha_t) \nabla \log \pi_\theta(\alpha_t^i | s_t^i)], \quad (21)$$

where $A^\theta = \pi_\theta(s_t, a_t) - V_\phi(s_t)$ is the advantage function that is used to replace $R(\tau)$, and $V_\phi(*)$ represents the value function. The advantage function ensures that $a_t$ selected by the $\pi_\theta(s_t, a_t)$ under $s_t$ is superior to the other possible actions.

Considering that the policy of DRL is on-policy, i.e., the trained agent is the same as the agent that sampled the data. To make the adoption data reproducible, applying the importance sampling method is introduced, (20) and (21) are rewritten as

$$\nabla \hat{r}_\theta = E_{\tau \sim \pi_{\theta'}(\tau)} \left[ \frac{\nabla \pi_\theta(\alpha_t^i | s_t^i)}{\pi_{\theta'}(\alpha_t^i | s_t^i)} A^{\theta'}(s_t, \alpha_t) \right] \quad (22)$$

and

$$J^{\theta'}(\theta) = E_{\tau \sim \pi_{\theta'}(\tau)} \left[ \frac{\pi_\theta(\alpha_t | s_t)}{\pi_{\theta'}(\alpha_t | s_t)} A^{\theta'}(s_t, \alpha_t) \right], \quad (23)$$

where the $\pi_{\theta'}$ is the policy that generates the training data. Too large a variance between $\pi_\theta$ and $\pi_{\theta'}$ can lead to reduced sample efficiency and cause the training process to become extremely unstable [25]. Therefore, Kullback-Leibler (KL) divergence used to describe the measure of variability between two probability distributions is given by

$$D_{\text{KL}}(\pi_\theta || \pi_{\theta'}) = \int_{-\infty}^{+\infty} \pi_\theta(a_t, s_t) \log \frac{\pi_\theta(a_t, s_t)}{\pi_{\theta'}(a_t, s_t)} d(a_t, s_t). \quad (24)$$

We want to constrain the magnitude of each $\theta$ update. The objective function can be clipped by clipping the parameter $epsilon$, which can be expressed as

$$J_{\text{clip}}^\theta(\theta) = E_t[\min(r_t(\theta) A^\theta, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon) A^\theta)]. \quad (25)$$

In the presence of a multi-agent environment, centralised training and distributed execution can effectively address the Dec-POMDP constructed above. In the centralised training phase, the value network with access to the global state changes part of Dec-POMDP into MDP. At this phase, the primary optimisation objective of the value network for parameter $\phi$ is given by

$$V_\phi^i(\phi) = \max \frac{1}{T} \sum_{\mathcal{D}_i^s} \sum_{t=0}^{T} (V_\phi^i(s_t) - \hat{r}_t)^2, \quad (26)$$

where $\mathcal{D}_i^s$ is the global experience trajectory buffer storing trajectory $\tau_s$, $\tau_s$ is defined by $(s_t, s_{t+1}, a_t, \hat{r}_t, ...)$ at time slot $t$. To train the value network stably, the value function needs to be normalised by the mean and standard deviation.

In addition, the main optimisation objectives of the policy network are given by

$$J^{\theta'}(\theta) = E_{\tau_i \sim \pi_{\theta'}(\tau_i)} \left[ \frac{\pi_\theta(\alpha_t^i | o_t^i)}{\pi_{\theta'}(\alpha_t^i | o_t^i)} A^{\theta'}(o_t^i, \alpha_t^i) \right], \quad (27)$$

where $\tau_i$ is the sequence generated by UAV $i$ using local observation to interact with the environment. Each UAV updates the parameters $\theta_i$ according to $\tau_i$.

In the execution phase, UAV $i$ can execute action $a_i$ under policy $\pi_\theta(\alpha^i | o^i)$ based on local observation $o_i$. During this process, no additional information is required for UAV $i$. Thus, UAVs with distributed execution can serve the maximum number of MUs and total throughput.

## V. SIMULA RESULTS

### A. Simulation Parameters

This section evaluates the proposed MAPPO-based UAV trajectory optimisation scheme. Specially, we utilise the gym to build a UAV flight simulation environment and use the PPO and multi-agent deep deterministic policy gradient (MADDPG) algorithms [26] as benchmarks. The three methods employ the same model, where we build a six-layer fully-connected neural network with the same structure for the policy network and the value network to compare convergence performance

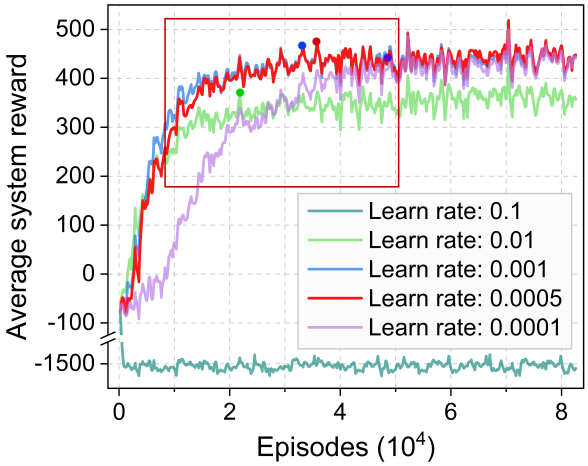

Fig. 2. Convergence speed of average system reward under different learning rates.

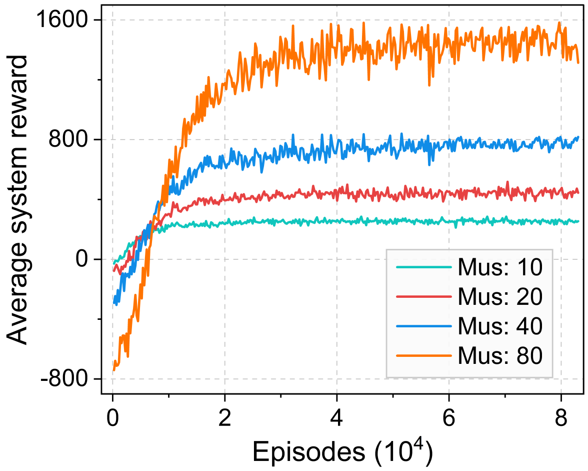

Fig. 3. Convergence performance under different MUs.

TABLE I
LIST OF MAIN SIMULATION PARAMETERS

| Parameters | Value |
|---|---|
| Additional path loss for LoS $(A_L)$ | 1 dB |
| Additional path loss for NLoS $(A_{NL})$ | 20 dB |
| Path loss exponent for LoS $(\gamma^{\mathrm{LoS}})$ | 2.09 |
| Path loss exponent for NLoS $(\gamma^{\mathrm{NLoS}})$ | 3.75 |
| Environment index $(\eta)$ | 9.6117 |
| Environment index $(\beta)$ | 0.1581 |
| Carrier frequency $(f_c)$ | 200 kHz |
| Channel bandwidth $(B)$ | 5 MHz |
| Noise power $(\sigma)$ | -95 dBm |
| UAV operating altitude $(h_{i,t})$ | 50 70 m |
| UAV maximum velocity $(v_{\max})$ | 20 m/s |
| UAV transmit power $(P^r_{i,j})$ | 30 dBm |
| FSO transmission power $(P_t)$ | 200 mW |
| Optical efficiencies $(\xi_t)$ | 0.8 |
| FSO receiver diameter $(\phi)$ | 0.06 m |
| FSO transmitter divergence $(\varkappa_t)$ | $2.07 \times 10^{-4}$ rad |
| FSO wavelength $(\lambda)$ | 1550 nm |
| Receiver sensitivity$(Z_b)$ | 100 photons/bit |

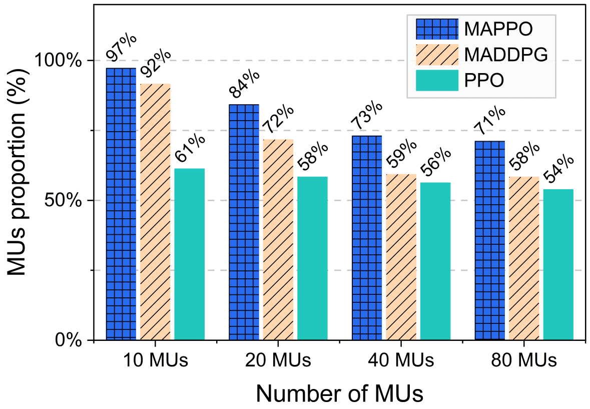

Fig. 4. Proportion of MUs serving different algorithms.

and number of the service MUs . For the proximal policy optimisation (PPO) algorithm [27], we assume that a central controller is installed on one UAV to control the actions of all UAVs.

Specifically, we simulate a disaster area with a 2 km2 km area. Remote GBSs were randomly distributed on the edge of the disaster area to establish FSO communication links with UAVs. MU has a Poisson distribution in the disaster area. Meanwhile, MUs velocity is distributed between 0.9 m/s and 2.2 m/s [28]. Initially, MUs with Poisson distribution are in the disaster area. The LoS and NLoS path loss exponents are $\gamma^{\mathrm{LoS}} = 2.09$ and $\gamma^{\mathrm{NLoS}} = 3.75$, respectively. More simulation parameters are shown in Table I.

### *B. Convergence Analysis*

Fig. 2 shows the trend of the average system reward, which is the accumulation of the average reward obtained by each UAV in an episode. Here, we evaluate the convergence speed of the MAPPO algorithm at various learning rates with 20 MUs, 3 UAVs, and 3 GBSs in the experiment. One can observe that when the learning rate is too large (e.g., the learning rate is set to 0.1), the algorithm will fail to converge. The red box indicates that the convergence speed will slow as the learning rate decreases. Furthermore, a too-low learning rate will only raise the system's average reward after convergence. When the learning rate is set to 0.005, the convergence speed balances well with the average value of the system reward.

Fig. 3 shows the trend of the average system reward changing with episodes based on MAPPO with different numbers of MUs. One can observe that the convergence speed of the system is related to the number of service MUs. In addition, when the number of MUs served by the system increases, the reward fluctuates after the system policy converges because the complexity and diversity of the environment increase as the number of MUs in the system increases.

### *C. Performance Evaluation*

Fig. 4 shows the impact of the four methods on the proportion of connected MUs. The distribution parameter of MUs is set to 10, 20, 40, and 80, and the number of UAVs is 3. The proportion of connected MUs is the ratio of disconnected MUs to the total number of MUs. It can be seen that the MU ratio of our proposed MAPPO algorithm is greater than the other algorithms. Also, MADDPG has better performance when the number of MUs is small. However, the performance decreases rapidly with the number of MUs increasing (reduced by 34%), which indicates that the MADDPG algorithm cannot be well applied in complex scenarios. In the ppo algorithm, the growth of MUs did not have a significant impact on the proportion of service MUs (remaining about 57%), because there has not been a good flight strategy for UAVs.

## VI. CONCLUSIONS

In this paper, considering a multi-UAV rescue system with hybrid FSO/RF communications under noise infection, we investigated the trajectory optimisation policies to maximise the number of served MUs and the total channel capacity. By transforming the problem into a Dec-POMDP, we proposed a MAPPO-based algorithm. The experimental results showed that MAPPO outperforms other DRL algorithms in turns of the number of served MUs and total channel capacity.

## REFERENCES

[1] D. Chen, Q. Qi, Z. Zhuang, J. Wang, J. Liao, and Z. Han, "Mean field deep reinforcement learning for fair and efficient UAV control," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 813–828, 2020.

[2] M. Matracia, M. A. Kishk, and M.-S. Alouini, "On the topological aspects of UAV-assisted post-disaster wireless communication networks," *IEEE Communications Magazine*, vol. 59, no. 11, pp. 59–64, 2021.

[3] Z. Yao, W. Cheng, W. Zhang, and H. Zhang, "Resource allocation for 5G-UAV-Based emergency wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3395–3410, 2021.

[4] J.-H. Lee, K.-H. Park, Y.-C. Ko, and M.-S. Alouini, "Throughput maximization of mixed FSO/RF UAV-aided mobile relaying with a buffer," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 683–694, 2020.

[5] L. Qu, G. Xu, Z. Zeng, N. Zhang, and Q. Zhang, "UAV-Assisted RF/FSO Relay System for Space-Air-Ground Integrated Network: A Performance Analysis," *IEEE Transactions on Wireless Communications*, 2022.

[6] D. Wu, X. Sun, and N. Ansari, "An FSO-based drone assisted mobile access network for emergency communications," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 3, pp. 1597–1606, 2019.

[7] G. Solmaz and D. Turgut, "Modeling pedestrian mobility in disaster areas," *Pervasive and Mobile Computing*, vol. 40, pp. 104–122, 2017.

[8] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[9] S. P. Gopi and M. Magarini, "Reinforcement learning aided uav base station location optimization for rate maximization," *Electronics*, vol. 10, no. 23, p. 2953, 2021.

[10] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC-and UAV-assisted vehicular networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 131–141, 2020.

[11] C. Yu, A. Velu, E. Vinitsky, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative, multi-agent games," *arXiv preprint arXiv:2103.01955*, 2021.

[12] S. Zhang, X. Sun, and N. Ansari, "Placing multiple drone base stations in hotspots," in *2018 IEEE 39th Sarnoff Symposium*. IEEE, 2018, pp. 1–6.

[13] H. Huang, A. V. Savkin, M. Ding, and M. A. Kaafar, "Optimized deployment of drone base station to improve user experience in cellular networks," *Journal of Network and Computer Applications*, vol. 144, pp. 49–58, 2019.

[14] W. Fawaz, C. Abou-Rjeily, and C. Assi, "UAV-aided cooperation for FSO communication systems," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 70–75, 2018.

[15] S. Lee, H. Yu, and H. Lee, "Multiagent Q-Learning-Based Multi-UAV Wireless Networks for Maximizing Energy Efficiency: Deployment and Power Control Strategy Design," *IEEE Internet of Things Journal*, vol. 9, no. 9, pp. 6434–6442, 2021.

[16] Z. Qin, Z. Liu, G. Han, C. Lin, L. Guo, and L. Xie, "Distributed UAV-BSs Trajectory Optimization for User-Level Fair Communication Service With Multi-Agent Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, pp. 12 290–12 301, 2021.

[17] Y.-J. Chen, K.-M. Liao, M.-L. Ku, F. P. Tso, and G.-Y. Chen, "Multi-agent reinforcement learning based 3D trajectory design in aerial-terrestrial wireless caching networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 8, pp. 8201–8215, 2021.

[18] M. Series, "Guidelines for evaluation of radio interface technologies for IMT-2020," *Report ITU*, pp. 2412–0, 2017.

[19] C. Liu, M. Ding, C. Ma, Q. Li, Z. Lin, and Y.-C. Liang, "Performance analysis for practical unmanned aerial vehicle networks with LoS/NLoS transmissions," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[20] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Further enhancements to LTE Time Division Duplex (TDD) for Downlink-Uplink (DL-UL) interference management and traffic adaptation," 3rd Generation Partnership Project (3GPP), Tech. Rep. 36.828, 06 2012, version 11.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2507

[21] M. Ding, P. Wang, D. López-Pérez, G. Mao, and Z. Lin, "Performance impact of LoS and NLoS transmissions in dense cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 2365–2380, 2015.

[22] L. Henderson, "The statistics of crowd fluids," *nature*, vol. 229, no. 5284, pp. 381–383, 1971.

[23] A. K. Majumdar, "Free-space laser communication performance in the atmospheric channel," *Journal of Optical and Fiber Communications Reports*, vol. 2, no. 4, pp. 345–396, 2005.

[24] I. I. Kim, B. McArthur, and E. J. Korevaar, "Comparison of laser beam propagation at 785 nm and 1550 nm in fog and haze for optical wireless communications," in *Optical wireless communications III*, vol. 4214. Spie, 2001, pp. 26–37.

[25] T. Xie, Y. Ma, and Y.-X. Wang, "Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[26] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in neural information processing systems*, vol. 30, 2017.

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[28] S. Chandra and A. K. Bharti, "Speed Distribution Curves for Pedestrians During Walking and Crossing," *Procedia - Social and Behavioral Sciences*, vol. 104, pp. 660–667, 2013, 2nd Conference of Transportation Research Group of India (2nd CTRG).