Optimization for Master-UAV-powered Auxiliary-Aerial-IRS-assisted IoT Networks: An Option-based Multi-agent Hierarchical Deep Reinforcement Learning Approach

Jingren Xu, Xin Kang, *Senior Member, IEEE*, Ronghaixiang Zhang, and Ying-Chang Liang, *Fellow, IEEE*

***Abstract*—This paper investigates a master unmanned aerial vehicle (MUAV)-powered Internet of Things (IoT) network, in which we propose using a rechargeable auxiliary UAV (AUAV) equipped with an intelligent reflecting surface (IRS) to enhance the communication signals from the MUAV and also leverage the MUAV as a recharging power source. Under the proposed model, we investigate the optimal collaboration strategy of these energy-limited UAVs to maximize the accumulated throughput of the IoT network. Depending on whether there is charging between the two UAVs, two optimization problems are formulated. To solve them, two multi-agent deep reinforcement learning (DRL) approaches are proposed, which are centralized training multi-agent deep deterministic policy gradient (CT-MADDPG) and multi-agent deep deterministic policy option critic (MADDPOC). It is shown that the CT-MADDPG can greatly reduce the requirement on the computing capability of the UAV hardware, and the proposed MADDPOC is able to support low-level multi-agent cooperative learning in the continuous action domains, which has great advantages over the existing option-based hierarchical DRL that only support single-agent learning and discrete actions.**

***Index Terms*—UAV, IRS, rechargeable, multi-agent deep reinforcement learning, hierarchical**

I. Introduction

The Internet of Things (IoT), aiming at providing a pervasive connection of digital devices, is a novel cutting edge technology for the future digital world [2], [3]. Its existence has affected various aspects of our life and supports the development of smart city, smart agriculture, smart farming, smart manufacturing, and etc., as a core technology [4]. With diverse use cases anticipated in IoT, the types of IoT devices are expected to be diversified, and the deployment of different IoT devices are also expected to vary significantly. Some types of IoT devices, such as sensors, are expected to deployed sparsely in certain scenarios such as smart agriculture. For these scenarios, the traditional base station (BS) based communication solutions may not be cost-effective or high-efficient due to the long-distance path loss.

In contrast to the traditional terrestrial BS, unmanned aerial vehicle (UAV) communications not only benefit from the considerable channel gain brought by higher probability of line-of-sight (LoS) channel, but also enjoy the coverage gain by taking advantage of high maneuverability [5]. Thus, UAVs have recently drawn significant attention in data collection tasks of IoT devices [6]. Related optimization problems of ~~location~~ deployment, trajectory design and resource allocation in UAV based wireless networks have been studied in [7]-[10]. Yang *et al.* [7] maximized the energy efficiency in a UAV-assisted backscatter communication network by jointly optimizing the devices' wake-up schedule, power control, and the UAV's trajectory. Ye *et al.* [8] proposed a dynamic time division multiple access (TDMA) structure for full-duplex UAV-enabled wireless powered network. In the multi-UAV cooperative scenario, [9] proposed an iterative algorithm based on block coordinate descent to solve mixed integer non-convex optimization problem, where users' communication scheduling, UAVs' trajectory, and power control were highly coupled. In 3-D space, a novel framework for full-duplex orthogonal frequency-division multiple access UAV-enabled wireless-powered IoT networks was proposed in [10], where a swarm of UAVs was deployed to simultaneously charge all devices and then ~~flied~~ to new locations to collect information from scheduled devices. However, limited battery capacity on board is a key factor affecting the performance of UAVs in practical, which was ignored in aforementioned work. Although [7] took it into account, there's no guarantee of successful return after task termination. Considering the downsides of traditional charging methods, e.g. suspending tasks or wasting time in landing, there are some works dedicated to prolonging the lifetime of UAV networks. In [11], sustainable solar-powered UAVs were applied in random access IoT networks, where UAVs can be recharged through solar radiation. This method is highly weather-dependent and cannot be used in bad weather areas. Another promising technique is wireless-powered UAV [12]. Such researches usually deploy static charger with high-voltage power lines on the ground [13], [14]. However, the bad channel conditions requires a demanding charging distance. As a result, UAV has to find a trade-off between charging and communications.

This work was supported in part by the National Natural Science Foundation of China under Grants U1801261, by the National Key R&D Program of China under Grant 2018YFB1801105, by the Key Areas of Research and Development Program of Guangdong Province, China, under Grant 2018B010114001, by the Science and Technology Development Fund, Macau SAR, under Grant 0009/2020/A1, by the Central Universities under Grant ZYGX2019Z022 and by the 111 Project under Grant B20064. Part of this work was presented in IEEE Global Communications Conference 2021 [1].

J. Xu, X. Kang, R. Zhang and Y.-C. Liang are with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China (e-mail: jingrenxu@163.com; kangxin83@gmail.com; fredrickzhang@yahoo.com; liangyc@ieee.org).

On the other hand, the intelligent reflecting surface (IRS), benefiting from the breakthrough on the fabrication of inexpensive programmable meta-material, can improve the channel environment and extend the coverage [15]. Existing researches on IRS-assisted UAV communications can mainly be categorized into two groups: 1) IRS is deployed at a terrestrial point, e.g. the surface of buildings, to enhance the signals from the UAV [16], [17]. In [18], the average worst-case secrecy rate was maximized in an IRS-aided UAV secure communication system. Under the circumstance, the IRS is supposed to be in sight of both the UAV and the users [19], which is highly dependent on the deployment environment. 2) IRS is deployed in the air to achieve 360° panoramic full-angle reflection and enjoy the LoS air-to-ground (A2G) channel [20]. In [21] and [22], an integrated UAV-IRS was implemented as a relay node to provide an enhanced link to the ground users. The beamforming as well as the UAV's position ~~were~~ optimized to achieve a better channel quality. Besides, the authors in [23] considered an IRS-assisted multi-layer UAVs network, where a quasi-stationary UAV equipped with an IRS was placed at higher altitudes to serve smaller UAVs below. Howerever, the works aforementioned treat UAVs as spot-deployable carriers and ignore the gains brought by their high mobility.

Deep reinforcement learning (DRL) methods can deal with non-convex optimization problems with high complexity and is widely used in various UAV-assisted communication scenarios [24]. [25] proposed a deep Q-learning based method for UAV trajectories design and backscatter device scheduling. One single UAV is inefficient especially when the ground devices are widely distributed. Thus in [26], a fully-distributed control solution based on the deep deterministic policy gradient (DDPG) algorithm [27] was designed to navigate a UAV swarm for fair communication coverage. However, the aforementioned works ignore the limited on-board battery capacity which ~~is~~ unpractical for time-consuming tasks [28]. This constraint was considered in [29] and a distance-related reward was designed to help the UAV learn to return to the charging station. However, the method used in ~~this article~~, i.e., overlaying a smaller reward with a larger one, leads to training instability. The authors in [30] proposed an option-based DRL methods with hierarchical control to deal with such problems. To be specific, each UAV makes a choice among options, such as communication, charging and task termination, according to upper-level policies, and execute specific actions based on low-level policies. But the low-level policies in [30], e.g. the trajectories of UAVs, were replaced by the artificial strategies for simplicity without optimization. In addition, fully-distributed training costs more time or needs more hardware resources on parameters optimization compared with the centralized method, which leads to unsatisfactory performance in certain cases [31].

To deal with the limitations of aforementioned works, in this paper, we propose using a rechargeable Auxiliary UAV (AUAV) equipped with an IRS to assist the data collection of Master UAV (MUAV)-powered IoT network. To be specific, the IRS-equipped rechargeable AUAV is used to enhance the communication signals from the MUAV, and it also leverages the MUAV as a flying recharging power source.

The main contributions are listed as follows:

- We propose a novel master-UAV-powered auxiliary-aerial-IRS-assisted IoT network, which is different from existing works as follows: a) Both the MUAV and the AUAV are no longer deployed at fixed points. Instead, both ~~of them~~ are allowed to move freely within the IoT network. In this way, our model can enjoy the gains brought by both UAVs' high mobility and the IRS's signal enhancement. b) A wireless charging mechanism is added, through which the AUAV can harvest energy from the MUAV during the task instead of suspending task to recharge.
- Under the proposed new model, two network throughput maximization problems are formulated, depending on whether there is wireless charging between the AUAV and the MUAV. We investigate the optimal collaboration strategies (including power allocation, UAV trajectories, task termination, and charging) of these UAVs to maximize the accumulated throughput of the IoT network. Especially, unlike existing works, we assume both the AUAV and the MUAV are battery-limited, and we request both of them to return the originating point when the data collection task terminates, which makes our optimization problem more practical but more challenging.
- To solve the formulated optimization problems, two new multi-agent DRL approaches are proposed: a) For the problem without charging between UAVs, we propose a centralized training multi-agent DRL approach, referred to as centralized training multi-agent deep deterministic policy gradient (CT-MADDPG), to deal with the learning of low-level policies. Unlike the traditional MADDPG, the CT-MADDPG removes many unnecessary parameters, which greatly reduces the requirement on the computing hardware. b). For the problem with charging between UAVs, we model the optimization problem as a multi-agent semi-markov decision process (SMDP), and we propose a hybrid hierarchical DRL approach , which is named as multi-agent deep deterministic policy option critic (MADDPOC). Unlike the existing option-based hierarchical DRL that only support single-agent learning and discrete actions, the proposed MADDPOC not only supports low-level multi-agent cooperative learning but also support continuous actions.

The remainder of this paper is organized as follows. Section II describes the system model. In Section III, the scenario task is elaborated and problems are formulated. Section IV shows the CT-MADDPG algorithm with centralized training method for the non-charging scenario. In Section V, we propose the MADDPOC algorithm as a hierarchical extension of the CT-MADDPG for the charging scenario. In Section~~,~~ VI we present simulation results for the proposed algorithms. Finally, we conclude the paper in Section VII.

Notations: A scalar x, a vector $\mathbf{x}$, and a matrix $\mathbf{X}$ are given by x, $\mathbf{x}$, and $\mathbf{X}$, respectively. $\mathbb{R}^{M\times N}$ and $\mathbb{C}^{M\times N}$ represent the space of a $M \times N$ matrix with real and complex entries, respectively. $\|\cdot\|$ refers to a vector norm. $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose of a matrix. $\otimes$ and $\odot$ donate the Kronecker

product and Hadamard product of two matrices , respectively. $\mathcal{CN}\left(0,\sigma^2\right)$ represents a circularly symmetric complex Gaussian (CSCG) vector's distribution with zero mean and noise variance σ^2, and $\sim$ means "distributed as". $diag\left(x_1,\cdots,x_k\right)$ represents a diagonal matrix with the elements given by $\{x_1,\cdots,x_k\}$. The operator $\mod$ returns the remainder after division.

II. System Model

We consider a dual-UAV-enabled downlink TDMA IoT network serving K single-antenna GNs with fixed locations assisted by an aerial IRS as shown in Fig. 1. The MUAV carries an uniform plane array (UPA) with $S_{\mathrm{M}}=S_{\mathrm{M}_x}\times S_{\mathrm{M}_y}$ antennas for both information and wireless power transmission. Meanwhile, the AUAV is equipped with an IRS which consists of $S_{\mathrm{R}}=S_{\mathrm{R}_x}\times S_{\mathrm{R}_y}$ passive reflecting units for signal enhancement, as well as an UPA with $S_{\mathrm{A}}=S_{\mathrm{A}_x}\times S_{\mathrm{A}_y}$ antennas deployed on the top for energy harvesting. The orientation of the IRS can be adjusted flexibly to make sure it's always perpendicular to the y axis. Assumed that all antennas and IRS units share a single ratio frequency (RF) chain, but each of them has an individual phase shift. In addition, a high-capacity battery with energy E_{Mmax} is provided to the MUAV, making it a small mobile charging station, while the AUAV is equipped with a lightweight battery with low capacity E_{Amax} in consideration of the on-board IRS's weight and UAV's restricted loading capacity, and needs to be charged wirelessly by the MUAV. The communication and charging processes are assumed to be simplex, i.e., only one is allowed per slot, and the request for charging is available at any time [32]. Besides, two UAVs operate at the fixed altitudes of $H_{\mathrm{M}},H_{\mathrm{A}}$, where $H_{\mathrm{M}}>H_{\mathrm{A}}$, respectively. The total serving time T is divided into N_τ equal-length slots with duration time δ_τ (s), i.e., $T=N_\tau\delta_\tau$, where N_τ depends on the specific policy. Given the communication fairness, the MUAV will automatically change its linked GN per N_u slots starting with the GN closest to the initial UAV location, i.e. $k=(k+\lfloor\frac{n+1}{N_u}\rfloor)\mod(K+1)$, where $k\in\{1,\cdots,K\}$ labels the index of currently linked GN.

For the sake of convenience, we donate the time-varying horizontal locations $\mathbf{l}_{\mathrm{M}}[n]=[x_{\mathrm{M}}[n],y_{\mathrm{M}}[n]],\mathbf{l}_{\mathrm{A}}[n]=[x_{\mathrm{A}}[n]\,y_{\mathrm{A}}[n]],n\in\{0,\cdots,N_\tau\}$, for the MUAV and AUAV, respectively, and $\mathbf{l}_{\mathrm{G}_k}=[x_{\mathrm{G}_k},y_{\mathrm{G}_k}],\mathbf{l}_{\mathrm{C}}=[x_{\mathrm{C}},y_{\mathrm{C}}]$ for the stationary GN k and charge station. The distance between MUAV and AUAV, AUAV and GN k, MUAV and charge station, AUAV and charge station, are donated by $d_{\mathrm{MA}}[n],d_{\mathrm{AG_k}}[n],d_{\mathrm{MC}}[n],d_{\mathrm{AC}}[n]$, respectively.

A. Channel Model

The transmission links between the MUAV, AUAV and GNs can be regarded as A2G channels which suffer from the extra path loss brought by the buildings, trees and other obstacles with high altitude. Such channel is generally donated by a probabilistic-average-based LoS (PLoS) model, which is expressed as

$$\begin{aligned}\overline{PL}(d,\Delta H)=PL^{\mathrm{FS}}(d)+\Gamma^{\mathrm{LoS}}(d,\Delta H)\eta_{\mathrm{LoS}}\\+(1-\Gamma^{\mathrm{LoS}}(d,\Delta H))\eta_{\mathrm{NLoS}},\end{aligned}\tag{1}$$

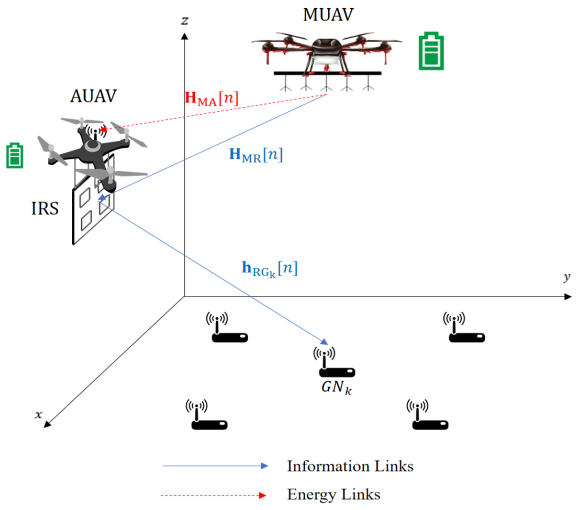


Fig. 1: System model.

where $PL^{\mathrm{FS}}(d)=20\log\frac{4\pi f_c d}{c}$ is the free space path loss related to the distance d, f_c donates the carrier frequency, c refers to the velocity of light, and $\eta_{\mathrm{LoS}},\eta_{\mathrm{NLoS}}$ are the excessive path loss component for LoS and NLoS links, respectively. Besides, Γ^{LoS} donates the probability of LoS connectivity, which is modified by a simple sigmoid function [33], given by

$$\Gamma^{\mathrm{LoS}}(d,\Delta H)=\frac{1}{1+\eta_a\exp\left(-\eta_b\left(\arcsin\left(\frac{\Delta H}{d}\right)-\eta_a\right)\right)},\tag{2}$$

where ΔH refers to the altitude difference, η_a and η_b are constants related to the type of propagation environment.

Therefore, the channel response vectors from MUAV to IRS, from MUAV to AUAV's UPA and from IRS to GN k at time slot n, donated by $\mathbf{H}_{\mathrm{MR}}[n]\in\mathbb{C}^{S_{\mathrm{R}}\times S_{\mathrm{M}}},\mathbf{H}_{\mathrm{MA}}[n]\in\mathbb{C}^{S_{\mathrm{A}}\times S_{\mathrm{M}}}$ and $\mathbf{h}_{\mathrm{RG_k}}[n]\in\mathbb{C}^{S_{\mathrm{R}}\times 1}$, respectively, can be modeled as

$$\begin{aligned}\mathbf{H}_{\mathrm{MR}}[n]=\sqrt{\alpha_{\mathrm{MR}}[n]}\mathbf{a}\left(\theta^{\mathrm{A}}_{\mathrm{MR}}[n],\xi^{\mathrm{A}}_{\mathrm{MR}}[n],S_{\mathrm{R}_x},S_{\mathrm{R}_y}\right)\\\otimes\mathbf{a}\left(\frac{\pi}{2}-\theta^{\mathrm{D}}_{\mathrm{MR}}[n],\xi^{\mathrm{D}}_{\mathrm{MR}}[n],S_{\mathrm{M}_x},S_{\mathrm{M}_y}\right)^H,\end{aligned}\tag{3}$$

$$\begin{aligned}\mathbf{H}_{\mathrm{MA}}[n]=\sqrt{\alpha_{\mathrm{MA}}[n]}\mathbf{a}\left(\theta^{\mathrm{A}}_{\mathrm{MA}}[n],\xi^{\mathrm{A}}_{\mathrm{MR}}[n],S_{\mathrm{A}_x},S_{\mathrm{A}_y}\right)\\\otimes\mathbf{a}\left(\theta^{\mathrm{D}}_{\mathrm{MA}}[n],\xi^{\mathrm{D}}_{\mathrm{MR}}[n],S_{\mathrm{M}_x},S_{\mathrm{M}_y}\right)^H,\text{and}\end{aligned}\tag{4}$$

$$\mathbf{h}_{\mathrm{RG_k}}[n]=\sqrt{\alpha_{\mathrm{RG_k}}[n]}\mathbf{a}\left(\theta^{\mathrm{D}}_{\mathrm{RG_k}}[n],\xi^{\mathrm{D}}_{\mathrm{RG_k}}[n],S_{\mathrm{R}_x},S_{\mathrm{R}_y}\right),\tag{5}$$

where $\alpha_{\mathrm{MR}}[n]=\alpha_{\mathrm{MA}}[n]=\frac{1}{\overline{PL}(d_{\mathrm{MA}}[n],H_M-H_A)}$, $\alpha_{\mathrm{RG_k}}[n]=\frac{1}{\overline{PL}(d_{\mathrm{AG_k}}[n],H_A)}$ are the corresponding path loss defined by the PLoS model mentioned above. Besides, $\theta^{\mathrm{A}}_{\mathrm{MR}}[n]$, $\xi^{\mathrm{A}}_{\mathrm{MR}}[n]$, $\theta^{\mathrm{D}}_{\mathrm{MR}}[n]$, $\xi^{\mathrm{D}}_{\mathrm{MR}}[n]$, $\theta^{\mathrm{A}}_{\mathrm{MA}}[n]$, $\theta^{\mathrm{D}}_{\mathrm{MA}}[n]$, $\theta^{\mathrm{D}}_{\mathrm{RG_k}}[n]$ and $\xi^{\mathrm{D}}_{\mathrm{RG_k}}[n]$ are the related vertical and horizontal AOAs/AODs shown in Fig 2. In addition, the steering vector $\mathbf{a}$ is expressed as:

$$\mathbf{a}(\theta,\xi,K_x,K_y)=f_{K_x}(\sin\theta cos\xi)\otimes f_{K_y}(\sin\theta sin\xi),\tag{6}$$

where $f_M(u)=\left[1,e^{-j\frac{2\pi\triangle d}{\lambda_c}u},\cdots,e^{-j\frac{2\pi\triangle d}{\lambda_c}(M-1)u}\right]^H$, λ_c denotes the carrier wavelength, and $\triangle d$ denotes the antenna or IRS unit spacing. For simplicity, we set $\Delta d/\lambda_c=1/2$.

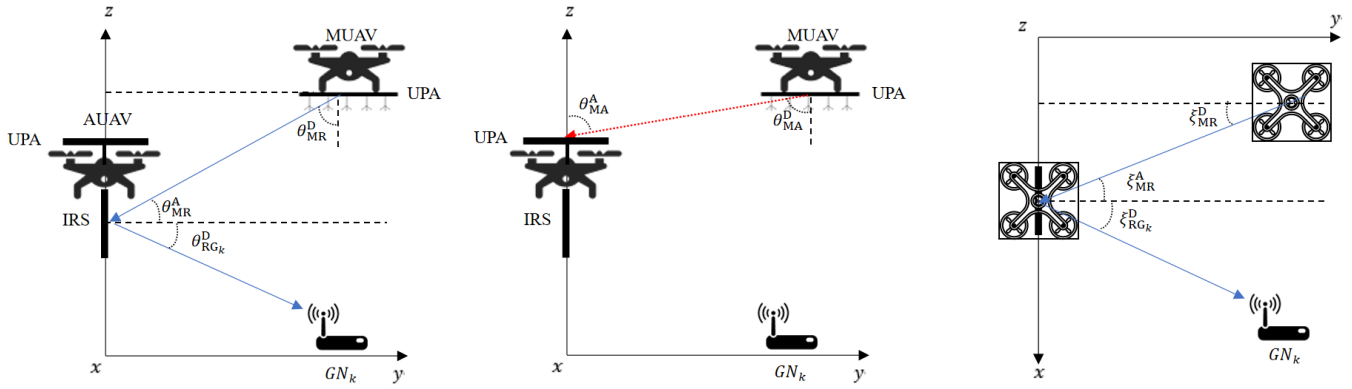


Fig. 2: (a) Vertical AoAs/AoDs between the MUAV, IRS and GN k. (b) Vertical AoAs/AoDs between the MUAV and AUAV's UPA when charging. (c) Horizontal AoAs/AoDs between the MUAV, AUAV and GN k.

B. Information Transmission Model

The direct transmission path from the MUAV to GN is ignored due to the high operation altitude. Thus, the end-to-end channel form the MUAV to GN k via the IRS carried by the AUAV at time slot n can be calculated as

$$\mathbf{h}_{\mathrm{MG}_k}[n]^H = \mathbf{h}_{\mathrm{RG}_k}[n]^H \mathbf{\Phi}[n] \mathbf{H}_{\mathrm{MR}}[n], \tag{7}$$

where $\mathbf{\Phi}[n] = \mathrm{diag}\left(\mathrm{e}^{\mathrm{j}\varphi_{1,1}[n]}, \cdots, \mathrm{e}^{\mathrm{j}\varphi_{\mathrm{S_{R_x}},\mathrm{S_{R_y}}}[n]}\right) \in \mathbb{C}^{\mathrm{S_R}\times\mathrm{S_R}}$ donates the phase shift matrix of the IRS and $\varphi_{S_{\mathrm{R}_x},S_{\mathrm{R}_y}}[n] \in [0, 2\pi)$ refers to the phase control introduced to the $\left(S_{\mathrm{R}_x}, S_{\mathrm{R}_y}\right)$-th reflecting element. Once the MUAV sends modulated symbol $x_k^{\mathrm{Tx}}[n]$ with zero mean and unit variance at time slot n, the received signal at GN k via reflected path can be written as

$$y_k^{\mathrm{Rx}}[n] = p_t[n]\mathbf{h}_{\mathrm{MG}_k}[n]^H \mathbf{w}_k[n] x_k^{\mathrm{Tx}}[n] + z_k[n], \tag{8}$$

where $\mathbf{w}_k[n] \in \mathbb{C}^{S_{\mathrm{M}}\times 1}$ is a unit-power beamformer adopted by the MUAV to serve GN k and $p_t[n]$ donates the information transmit power. The noise at the receivers follows the circular complex Gaussian distribution, i.e., $z_k[n] \sim \mathcal{CN}(0, \sigma^2)$, where σ^2 refers to noise power. For simplicity, we assume maximum ratio transmission (MRT) is adopted for power allocation at the transmitter, i.e. $\mathbf{w}_k[n] = \frac{1}{\sqrt{S_{\mathrm{M}}}}\mathbf{a}\left(\frac{\pi}{2} - \theta^{\mathrm{D}}_{\mathrm{MR}}[n], \xi^{\mathrm{D}}_{\mathrm{MR}}[n], S_{\mathrm{M_x}}, S_{\mathrm{M_y}}\right)$. Then the optimal phase shift control at the IRS can be directly derived in terms of AoAs/AoDs:

$$\begin{aligned}\varphi_{s_{\mathrm{R}_x},s_{\mathrm{R}_y}}[n] = \frac{2\pi\Delta d}{\lambda_c}\big[&\left(s_{\mathrm{R}_x}-1\right)\left(\sin\theta^{A}_{\mathrm{MR}}[n]\cos\xi^{A}_{\mathrm{MR}}[n]\right.\\ &+\sin\theta^{D}_{RG_k}[n]\cos\xi^{D}_{\mathrm{RG_k}}[n]\left.\right) - \left(s_{\mathrm{R}_y}-1\right)\left(\sin\theta^{A}_{\mathrm{MR}}[n]\right.\\ &\sin\xi^{A}_{\mathrm{MR}}[n] + \sin\theta^{D}_{RG_k}[n]\sin\xi^{D}_{\mathrm{RG_k}}[n]\left.\right)\big], \forall s_{\mathrm{A}_x}, s_{\mathrm{A}_y}, n.\end{aligned} \tag{9}$$

The derivation details are available in Appendix A, and the achievable data rate of GN k at time slot n can be calculated as

$$R_k[n] = \log_2\left(1 + \frac{p_t[n]\left\|\left(\mathbf{h}_{\mathrm{MG_k}}[n]\right)^H \mathbf{w}_k[n]\right\|^2}{\sigma^2}\right). \tag{10}$$

C. Energy Transmission Model

The received power at the AUAV at time slot n can be written as

$$p_r[n] = p_c\left\|\mathbf{H}_{\mathrm{MA}}[n]\mathbf{w}_{\mathrm{M}}[n]\right\|^2, \tag{11}$$

where the steering vector $\mathbf{w}_{\mathrm{M}}[n] \in \mathbb{C}^{S_{\mathrm{M}}\times 1}$ denotes the phase shift vector of UPA carried by the MUAV and constant p_c the charging power transmitted by the MUAV. Similarly, we also adopt MRT for power allocation, i.e. $\mathbf{w}_{\mathrm{M}}[n] = \frac{1}{\sqrt{S_{\mathrm{M}}}}\mathbf{a}\left(\theta^{\mathrm{D}}_{\mathrm{MA}}[n], \xi^{\mathrm{D}}_{\mathrm{MR}}[n], S_{\mathrm{M_x}}, S_{\mathrm{M_y}}\right)$. Then the received power can be simplified as

$$p_r[n] = S_{\mathrm{M}} S_{\mathrm{A}} p_c \alpha_{\mathrm{MA}}[n]. \tag{12}$$

D. Quad-Rotor UAV Energy Consumption Model

In this paper, we consider a quad-rotor UAV propulsion consumption model which contains three parts: the blade profile power, the parasite power and the induced power to overcome the profile drag of blades, fuselage drag, and induced drag, respectively. All of these are related to the velocity of UAV v as mentioned in [34]:

$$\begin{aligned}P_{\mathrm{UAV}}(v) = &\underbrace{P_0\left(1 + \frac{3v^2}{\Omega^2 R^2}\right)}_{\text{Blade profile}} + \underbrace{P_i\left(\sqrt{1 + \frac{v^4}{4v_0^4}} - \frac{v^2}{2v_0^2}\right)^{\frac{1}{2}}}_{\text{Induced}}\\ &+ \underbrace{\frac{1}{2}d_0\rho s A v^3}_{\text{Parasite}}.\end{aligned} \tag{13}$$

The physical meanings of the parameters in (13) are summarized in Table I. Besides, the power for policy computing and IRS controller is negligible compared with the propulsion consumption, which are not considered. Thus, we only take the energy consumption of propulsion, communication and charging into account.

TABLE I: Physical meaning of parameters in Quad-Rotor UAV Energy Consumption Model

Parameters	Physical meanings	Simulation values
Ω	Blade angel velocity	300 (radians/second)
R	Rotor radius	0.4 (meter)
ρ	Air density	1.225 (kg/m^2)
s	Rotor solidity	0.05 (m^3)
A	Rotor disc area	0.503 (m^2)
v_0	Induced velocity for rotor in forwarding flight	4.03 (meter/second)
d_0	Fuselage drag ratio	0.6
P_0	Blade profile power in hovering status	79.86 (watt)
P_i	Induced power in hovering status	88.63 (watt)

III. Task Description and Problem Formulation

A. Task Description

In our proposed scenario, two UAVs leave the initial charging station and provide communication services for GNs in sequence, and do not return until the task termination. Thus, the AUAV with low-capacity battery will power off before the MUAV. We consider two cases here: 1) Both UAVs terminate the task and return to the charging station. 2) The AUAV harvests energy from the MUAV to continue the task. In other words, not only do UAVs have to decide when to communicate, charge and terminate, i.e., over-option policy, but also how to execute them, i.e., intra-option policy. Given the low time proportion of the latter two processes, we replace their intra-option policies with fixed strategies in the sense of energy minimization, since the optimization make little sense. Specifically, the two UAVs will approach each other at the max-energy-efficiency velocity, i.e., $v_{mee} = \arg\max_{v}(\frac{v}{P_{\mathrm{UAV}}(v)})$, and hover when the horizontal positions overlap in the process of charging, while for the termination process, two UAVs will head straight to the charging station at v_{mee}. Note that the termination process performs multi-time-step execution dependent on $d_{\mathrm{MC}}[n]$ and $d_{\mathrm{AC}}[n]$, while the other two are executed step by step. For the sake of latter presentation, we introduce the concept of option to represent the process to be performed. Denote $\omega^n \in \{0, 1, 2\}$ as the current option, referring to the processes of charging, communication and task termination, respectively. Then the leftover energy of MUAV and AUAV at time slot n can be formulated as

$$E_{\mathrm{M}}[n] = E_{\mathrm{M}_{max}} - \delta_\tau \sum_{z=0}^{n-1}(P_{\mathrm{UAV}}(v_{\mathrm{M}}[z]) + (1-\xi_d)\omega^n p_t[z]) - \xi_d E_{\mathrm{Mr}}[n], \tag{14}$$

$$E_{\mathrm{A}}[n] = \min\big(E_{\mathrm{A}_{max}}, E_{\mathrm{A}_{max}} - \delta_\tau \sum_{z=0}^{n-1}(P_{\mathrm{UAV}}(v_{\mathrm{A}}[z]) - (1-\xi_d)(1-\omega^n)\alpha_c p_r[z]) - \xi_d E_{\mathrm{Ar}}[n]\big), \tag{15}$$

where $v_{\mathrm{M}}[z], v_{\mathrm{A}}[z]$ donate the time-varying velocities of the MUAV and AUAV, respectively, α_c donates the energy conversion efficiency and ξ_d is a binary variable masking the termination step, i.e. $\omega^{N_\tau} = 2$. Note that the remaining energy of AUAV after charging cannot exceed the battery capacity and $E_{ir}[n]$ donates the energy for return, given by

$$E_{ir}[n] = \delta_\tau P_{\mathrm{UAV}}(v_{mee})\frac{d_{i\mathrm{C}}[n]}{v_{mee}}, i = \mathrm{M}, \mathrm{A}. \tag{16}$$

Besides, we also denote

$$E_{il}[n] = E_i[n] - E_{ir}[n], i = \mathrm{M}, \mathrm{A}, \tag{17}$$

as the conditional leftover energy if return to the charging station at time slot n.

B. Problem Formulation

In this paper, our target is to maximize the accumulative throughput of GNs by jointly optimizing the over-option policy and intra-option policy, including the trajectories of two UAVs and the transmit power at the MUAV. Particularly, both the non-charging and charging scenarios are considered.

1) Non-charging Scenario: That is, $\omega^n = 1, \forall n \in \{0, \cdots, N_\tau - 1\}$. Besides, the moving range is a rectangular area with the x-dimensional limitation $[x_{min}, x_{max}]$ as well as the y-dimensional limitation $[y_{min}, y_{max}]$. Both UAVs can move freely in the specified area and adjust velocities dynamically. Similarly, the velocity and the transmit power for communication also have the bound v_{max}, p_{tmax}. Meanwhile, the UAV crash due to battery exhaustion is not expected at any time, even on the way back to the charging station. So the task ends when the constraint (18g) is about to be violated. In practice, signals cannot be reflected when the MUAV and GN k are at different sides of the IRS, so we set the same-side constraint (18h) to avoid it. Moreover, we also take into account the limited communication coverage capabilities of the MUAV in the constraint (18i). For the convenience of illustration, denote $\mathbf{L} = \{\mathbf{l}_{\mathrm{M}}[n], \mathbf{l}_{\mathrm{A}}[n], \forall n\}$ as the trajectory set of two UAVs and $\mathbf{P}_t = \{p_t[n], \forall n\}$ the communication power set allocated to the GNs. Then the problem (P1) can be mathematically formulated as

$$\begin{aligned}
(\mathrm{P1}): \max_{\mathbf{L},\mathbf{P}_t} \quad & \delta_\tau \sum_{n=0}^{N_\tau}(1-\xi_d)\omega^n R_k[n], && \text{(18a)}\\
\text{s.t.} \quad & x_{\mathrm{M}}[n] \in [x_{min}, x_{max}], x_{\mathrm{A}}[n] \in [x_{min}, x_{max}], && \text{(18b)}\\
& y_{\mathrm{M}}[n] \in [y_{min}, y_{max}], y_{\mathrm{A}}[n] \in [y_{min}, y_{max}], && \text{(18c)}\\
& v_{\mathrm{M}}[n] \in [0, v_{max}], v_{\mathrm{A}}[n] \in [0, v_{max}], && \text{(18d)}\\
& p_t[n] \in [0, p_{tmax}], && \text{(18e)}\\
& \mathbf{l}_{\mathrm{M}}[0] = \mathbf{l}_{\mathrm{A}}[0] = \mathbf{l}_{\mathrm{M}}[N_\tau] = \mathbf{l}_{\mathrm{A}}[N_\tau] = \mathbf{l}_{\mathrm{C}}, && \text{(18f)}\\
& E_{il}[n] > 0, i = \mathrm{M}, \mathrm{A}, && \text{(18g)}\\
& (y_{\mathrm{M}}[n] - y_{\mathrm{A}}[n]) \cdot (y_{G_k}[n] - y_{\mathrm{A}}[n]) > 0, && \text{(18h)}\\
& \|\mathbf{l}_{\mathrm{M}}[n] - \mathbf{l}_{\mathrm{G}_k}[n]\| \le l_{min}. && \text{(18i)}
\end{aligned}$$

2) Charging Scenario: Given the influence of IRS weight on the loading capacity, the battery capacity of AUAV is significantly reduced. A UPA is applied to the top of AUAV for energy harvesting from the MUAV. Besides the optimization variables mentioned in P1, we also focus on the option selection to ensure flight endurance and successful return. Note that the constraint (19b) is defined in terms of the current leftover energy instead of $E_{il}[n]$, since the charging mechanism is added. Besides, the constraints (19c) and (19d) only restrain the communication slots, i.e. $\omega^n = 1$. Denote

$\Omega_\omega = \{\omega^n, \forall n\}$ as the option set selected by UAVs. Then the problem (P2) can be formulated as

$$\text{(P2)}: \max_{\mathbf{L},\mathbf{P}_t,\Omega_\omega} \quad \delta_\tau \sum_{n=0}^{N_\tau} (1-\xi_d)\omega^n R_k[n], \tag{19a}$$

$$\text{s.t.} \quad (18b),(18c),(18d),(18e),(18f),$$

$$E_i[n] > 0, i = \mathrm{M}, \mathrm{A}, \tag{19b}$$

$$(1-\xi_d)\omega^n \cdot (y_{\mathrm{M}}[n] - y_{\mathrm{A}}[n]) \cdot (y_{G_k}[n] - y_{\mathrm{A}}[n]) \geq 0, \tag{19c}$$

$$(1-\xi_d)\omega^n \cdot \|\mathbf{l}_{\mathrm{M}}[n] - \mathbf{l}_{\mathrm{G}_k}[n]\| \leq l_{min}. \tag{19d}$$

IV. Multi-Agent DRL for P1

The optimization problem (P1) is challenging since it needs a joint design of the UAVs' trajectories and the transmit power, which are highly coupled due to the energy consumption. Thus, in this section, we propose an algorithm named CT-MADDPG to solve this problem.

A. Markov Game for Multi-UAV Cooperation

In a multi-agent environment, the decision making by an agent is influenced by others'. We use the Markov Game as a multi-agent extension of Markov decision process (MDP) [35] to model it. A Markov Game can be represented by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{R}$ is the reward space, $\mathcal{P}$ is the transition probability, and $\gamma \in [0,1]$ is the reward discount factor. At time slot n, agent i observes the current state $\mathbf{s}_i^n \in \mathcal{S}$ and takes an action $\mathbf{a}_i^n \in \mathcal{A}$ based on its own policy. Then the environment transits into the next state $\mathbf{s}_i^{n+1}$ with the transition probability $\mathcal{P}\left(\mathbf{s}_i^{n+1} | \mathbf{s}_i^n, \mathbf{a}_1^n, \cdots, \mathbf{a}_L^n\right)$ and generates a feedback in term of timely reward r_i^n. Each agent is aimed at maximizing the accumulative discounted rewards, given by

$$R_i^{acc} = \sum_{n=0}^{N_\tau} \sum_{m=0}^{N_\tau - n} (\gamma)^m r_i^{n+m}, \tag{20}$$

where N_τ is the time horizon aforementioned.

B. CT-MADDPG Approach

MADDPG is a multi-agent DRL method applied in multi-agent continuous action space within actor-critic framework. Its main idea is adopting the deep neutral network (DNN), also known as function approximator, to approximate the deterministic policy $\mu : \mathcal{S} \to \mathcal{A}$ and critic the action executed in a certain state by a fitting function Q for each agent [24]. To wherever possible reduce the number of network parameters that need to be optimized, as shown in Fig 3, we adopt the mode of centralized training, in which all agents share a single reward r^n and a global critic network Q is used to evaluate all executed actions in a global observation $\mathbf{o}^n$ containing all environment information. We call the MADDPG algorithm with the centralized training method CT-MADDPG. Note that $\mathbf{o}^n = \mathbf{s}_i^n$ in general except in partially observable environments [36], where agents can only observe local states caused by the limited communication capabilities. Besides, each agent

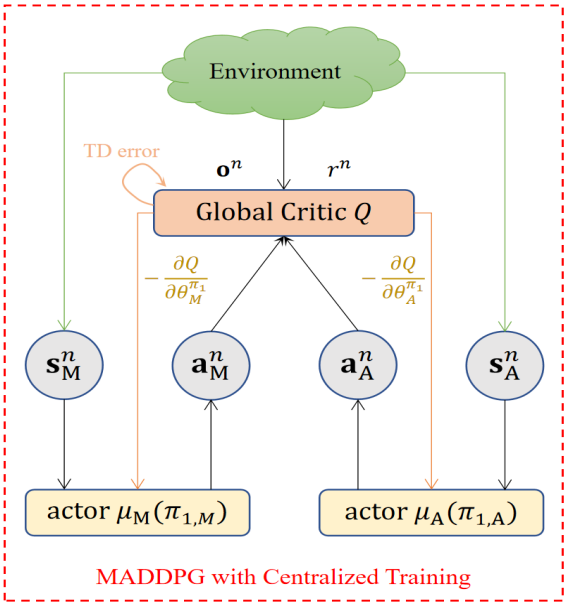


Fig. 3: The architecture of CT-MADDPG.

is assigned an actor network μ_i and executes the action in a distributed way.

Due to the divergence caused by the non-linear function approximator, two techniques are usually used to resolve this issue: the experience replay and the target network. In our scenario, CT-MADDPG stores both agents' experience transitions $\left\{\left\langle \mathbf{o}^n, \mathbf{s}_i^n, \mathbf{a}_i^n, r^n, \mathbf{o}^{n+1}, \mathbf{s}_i^{n+1}\right\rangle, i = \mathrm{M}, \mathrm{A}\right\}$ into a replay buffer in the exploration stage and randomly samples a mini-batch from it during the training phase. Moreover, the target networks of actor i and global critic, μ_i' and Q', which are used to compute the updated target, have the same architectures as the evaluation networks μ_i and Q. Specifically, the global critic network can be trained by minimizing the temporal difference (TD) error:

$$L\left(\theta^Q\right) = \frac{1}{2}\left(y - Q\left(\mathbf{o}^n, \mathbf{a}^n; \theta^Q\right)\right)^2, \tag{21}$$

where

$$y = r^n + \gamma Q'\left(\mathbf{o}^{n+1}, \mathbf{a}^{n+1}; \theta^{Q'}\right) * (1-\xi_d) \tag{22}$$

is the update target. Note that $\mathbf{a}^n = \{\mathbf{a}_\mathrm{M}^n, \mathbf{a}_\mathrm{A}^n\}$ donates the actions executed by two agents, $\mathbf{o}^{n+1}$ donates the next global observation and $\mathbf{a}^{n+1} = \left\{\mu_i'(\mathbf{s}_i^{n+1}; \theta_i^{\mu'}), i = \mathrm{M}, \mathrm{A}\right\}$ is given by the target actor networks of both agents. Besides, θ^Q and $\theta^{Q'}$ refer to the parameters of evaluation and target critic networks as well as the evaluation and target actor networks are weighted by θ_i^μ and $\theta_i^{\mu'}$, respectively. Moreover, the actor network of agent i is trained by minimizing the Q value:

$$L\left(\theta_i^\mu\right) = -Q\left(\mathbf{o}^n, \mu_i\left(\mathbf{s}_i^n; \theta_i^\mu\right); \theta^Q\right). \tag{23}$$

In addition, the target networks are updated by slowly tracking the learned networks:

$$\theta_i^{\mu'} = (1-\tau)\,\theta_i^{\mu'} + \tau\theta_i^\mu, \tag{24}$$

$$\theta^{Q'} = (1-\tau)\,\theta^{Q'} + \tau\theta^Q, \tag{25}$$

where $\tau \ll 1$ donates the soft replace rate. The other two tricks of CT-MADDPG are Gaussian action noise $\mathcal{N}_\mathcal{A}$ addition to

ensure the exploratory of agents and noise attenuation in the late training period to achieve better convergence.

C. Space Design

In this subsection, we give a rational design including the spaces of state, action and reward for (P1).

State: The states observed by the MUAV and the AUAV at time slot n are the same one-dimensional vectors with nine components:

- $\mathbf{l}_{\mathrm{M}}[n], \mathbf{l}_{\mathrm{A}}[n]$: the horizontal positions of two UAVs.
- $H_{\mathrm{M}}, H_{\mathrm{A}}$: the operation altitudes of two UAVs.
- $\mathbf{l}_{\mathrm{G}_k}$: the horizontal position of GN k.
- $E_{\mathrm{M}}[n], E_{\mathrm{A}}[n]$: the leftover energy of two UAVs.
- $E_{\mathrm{M}l}[n], E_{\mathrm{A}l}[n]$: the conditional leftover energy defined in (17).

Note that the location information can be observed through additional GPS devices, and the electricity information can be shared with simple signal communications, which are readily available in practice. Hence, the full state can be formulated as

$$\mathbf{s}_i^n = \big\{\mathbf{l}_{\mathrm{M}}[n], H_{\mathrm{M}}, \mathbf{l}_{\mathrm{A}}[n], H_{\mathrm{A}}, \mathbf{l}_{\mathrm{G}_k}, E_{\mathrm{M}}[n], E_{\mathrm{A}}[n], E_{\mathrm{M}l}[n], E_{\mathrm{A}l}[n]\big\}, i = \mathrm{M}, \mathrm{A}, \quad (26)$$

and has 12 dimensions. Moreover, the global observation $\mathbf{o}^n = \mathbf{s}_i^n$ specifically in our scenario.

Action: The actions for the MUAV and the AUAV at time slot n donated by $\mathbf{a}_{\mathrm{M}}^n$ and $\mathbf{a}_{\mathrm{A}}^n$ are the outputs of their actor networks that contain the scalar elements shown in the following:

- $v_{\mathrm{M}}[n], v_{\mathrm{A}}[n]$: The MUAV's and the AUAV's velocity.
- $a_{\mathrm{M}}[n], a_{\mathrm{A}}[n] \in [-\pi, \pi]$: The azimuthal angles.
- $p_t[n]$: The MUAV's transmit power.

According to the above definitions, the action sets of the MUAV and the AUAV can be written as

$$\mathbf{a}_{\mathrm{M}}^n = [v_{\mathrm{M}}[n], a_{\mathrm{M}}[n], p_t[n]], \quad (27)$$

$$\mathbf{a}_{\mathrm{A}}^n = [v_{\mathrm{A}}[n], a_{\mathrm{A}}[n]]. \quad (28)$$

Reward: In the centralized training mode, there is a global reward for agents sharing, which depends on the current global state and all executed actions. In our scenario, the reward design includes two parts:

- Single-step throughput reward: First, we need to motivate both UAVs by including a single-step throughput reward in a shared manner in order to solve (P1). The reward is designed as

$$r_{th}^n = \xi_{th}[n]\omega^n \left(R_k[n]\right)^{\kappa_{th}}, \quad (29)$$

  where κ_{th} is a factor to adjust the distribution of the reward. $\xi_{th}[n] = 0$ if violating the constraint (18h) or (18i), otherwise, $\xi_{th}[n] = 1$.
- Cross-border penalty: To prevent the UAVs from leaving the operation area, we set the cross-border penalty as

$$r_{cb}^n = \kappa_{cb}\left(\xi_{Mcb}[n] + \xi_{Acb}[n]\right), \quad (30)$$

  where $\xi_{Mcb}[n]$ and $\xi_{Acb}[n]$ are binary indicators implying whether the MUAV and the AUAV are outside the square area or not, respectively. Besides, κ_{cb} is a negative constant.

Thus, the overall global reward at time slot n is formulated as

$$r^n = r_{cb}^n + r_{th}^n. \quad (31)$$

V. Option-based Multi-Agent HDRL for P2

The CT-MADDPG method aforementioned only performs as a surrogate of the intra-option policy, while in the charging scenario, an extra higher-level surrogate of the over-option policy is needed. To deal with this issue, we combine the CT-MADDPG with the options framework [37] to enable the multi-agent hierarchical control.

A. Semi-Markov Game for Multi-UAV Options Learning

Options: The options framework allows a reinforcement learning agent to represent, learn and plan with temporally extended action. The temporally extended action, also known as Markovian option ω, is usually represented by a triplet $\langle I_\omega, \pi_\omega, \beta_\omega\rangle$ in which $I_\omega \subseteq \mathcal{S}$ is an initial set, π_ω is an intra-option policy, and $\beta_\omega : \mathcal{S} \to [0, 1]$ is a termination function. If $\mathbf{s}^n \in I_\omega$, then the option ω^n is available at state $\mathbf{s}^n$. Once an option is selected, a series of corresponding intra-option actions will be executed based on π_ω until it terminates with $\beta_\omega : \mathcal{S} \to \mathcal{A}$, then the over-option policy $\pi_\Omega : \mathcal{S} \to \Omega$ will make a new choice, and repeats the aforementioned procedure.

SMDP: Distinguished from the mode of single-step action execution in traditional MDP, the states may change several times between two adjacent decisions in SMDP, while the intermediate states may be irrelevant to the agent. Moreover, the intra-option policy π_ω can be even replaced with the artificial policy served as a sequence of pre-specified actions. Hence, the SMDP can be regarded as a special application of MDP in temporal abstract scenario, where the time step of each policy execution is a random variable.

As we introduce the options framework to the system, the set $\Omega_\omega = \{\omega^n, \forall n\}$ should also be added to the Markov game tuple. Hence, we can use semi-Markov game, i.e. $\langle \mathcal{S}, \Omega_\omega, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma\rangle$, to model (P2). Note that the option serves as global decisions for both agents specifically in our scenario, since whether the communication or chargeing process can not be finished by a single UAV, and also there is no guarantee that both UAVs will make the same choice simultaneously.

B. MADDPOC

We focus on the global options learning for both agents in this subsection. To achieve this goal, the CT-MADDPG algorithm is combined with the options framework by adding two sets of independent differentiable parameterized function approximators. As shown in Fig 4, $\pi_{\omega,i}$ denotes the agent i's intra-option policy of option ω parameterized by $\theta_i^{\pi_\omega}$, β the termination function of options parameterized by θ^β, and Q_Ω the global state-option value function parameterized by θ^{Q_Ω}. Besides, donate β', Q'_Ω weighted by $\theta^{\beta'}$, $\theta^{Q'_\Omega}$ for the target networks of β and Q_Ω, respectively.

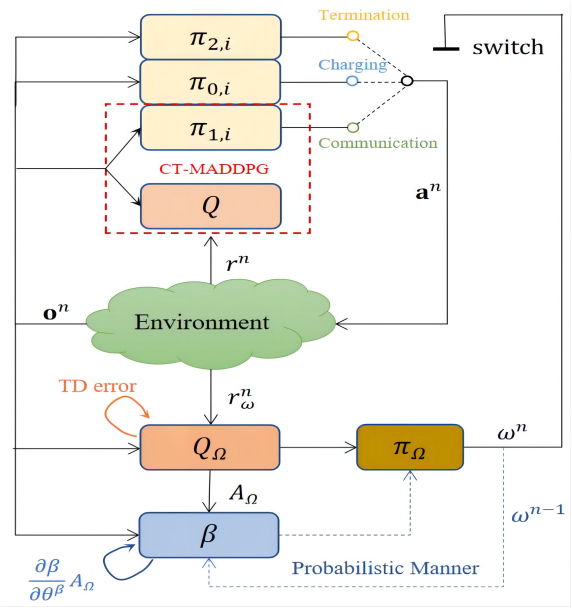


Fig. 4: The architecture of MADDPOC.

Just like the critic network evaluates actions, the function Q_Ω is used to evaluate how good or bad different choices are under a particular state. Thus, its outputs can be written as $Q_\Omega\left(\mathbf{o}^n;\theta^{Q_\Omega}\right)\in\mathcal{R}^{1\times n_\omega}$, where n_ω is the total number of available options. And $\beta\left(\mathbf{o}^n;\theta^{\beta}\right)\in\mathcal{R}^{1\times n_\omega}$, the outputs of β network, donates the termination probability of each option. Similarly, let $Q'_\Omega(\mathbf{o}^{n+1};\theta^{Q'_\Omega})$ and $\beta'(\mathbf{o}^{n+1};\theta^{\beta'})$ denote the outputs of their target networks, respectively. For the sake of distinction, we define a function as

$$\mathcal{F}(\mathbf{o},\omega)\triangleq\sum\mathcal{F}(\mathbf{o};\theta)\odot\mathbf{e}_\omega, \tag{32}$$

where $\mathcal{F}\in\{Q_\Omega,\beta,Q'_\Omega,\beta'\}$ donates some kind of network weighted by $\theta\in\{\theta^{Q_\Omega},\theta^{\beta},\theta^{Q'_\Omega},\theta^{\beta'}\}$, which takes the observation infromation o as input. Note that $\mathcal{F}(\mathbf{o};\theta)$ donates the output vector and $\mathcal{F}(\mathbf{o},\omega)$ the value corresponding to the option ω, which can be formulated as a product sum form of the option's one-hot code $\mathbf{e}_\omega$ and the network output vector.

As mentioned above that the last option ω^{n-1} will continue to perform util it terminates, and then a new choice ω^n will be made based on the over-option policy π_Ω. Two methods are available here:

$$\omega^n=\arg\max_{\omega}\left(Q_\Omega\left(\mathbf{o}^n;\theta^{Q_\Omega}\right)\right), \tag{34}$$

$$\omega^n=\begin{cases}\omega^{n-1},\text{with probability }1-\beta(\mathbf{o}^n,\omega^{n-1})\\ \arg\max_{\omega}\left(Q_\Omega\left(\mathbf{o}^n;\theta^{Q_\Omega}\right)\right),\text{otherwise}\end{cases} \tag{35}$$

for greedy policy and probabilistic policy, respectively. Besides, the related update target of Q_Ω takes a mixed form of probability:

$$\begin{aligned}y_\Omega=r^n_\omega+\gamma\big(&\left(1-\beta'(\mathbf{o}^{n+1},\omega^n)\right)Q'_\Omega(\mathbf{o}^{n+1},\omega^n)\\&+\beta'(\mathbf{o}^{n+1},\omega^n)Q'_{\Omega new}\big)*(1-\xi_d),\end{aligned} \tag{36}$$

where $\beta'(\mathbf{o}^{n+1},\omega^n),Q'_\Omega(\mathbf{o}^{n+1},\omega^n)$ donate the termination probability and expected advantage value of option ω^n related to the next state, respectively. Note that r^n_ω refers to the timely reward feedback given by the environment after the option execution. Besides, the expected advantage of new option is usually in the form of maximum advantage estimates, i.e.

Algorithm 1 Training process for MADDPOC

Initialize all the networks, including $Q_\Omega,\beta,Q,\pi_{1,i}$ weighted by $\theta^{Q_\Omega},\theta^{\beta},\theta^{Q},\theta^{\pi_1}_i$, as well as their target nets $Q'_\Omega,\beta',Q',\pi'_{1,i}$. Set target weights: $\theta^{Q'_\Omega}=\theta^{Q_\Omega},\theta^{\beta'}=\theta^{\beta},\theta^{Q'}=\theta^{Q},\theta^{\pi'_1}_i=\theta^{\pi_1}_i$. Note that $i=M,A$ denotes the subscript of the MUAV and AUAV, respectively.
Initialize the artificial intra-option policies $\pi_{0,i},\pi_{2,i}$ weighted by untrainable parameters $\theta^{\pi_0}_i$, $\theta^{\pi_2}_i$.
Initialize the queues $D_\Omega,D_\mathcal{A}$ with the same capability C as experience replay buffers.
for episode = 1 to $\mathcal{M}$ **do**
  Initialize the primal option $w^{-1}=1$.
  while not ξ_d **do**
    Get global observation $\mathbf{o}^n$ from the environment.
    if D_Ω is not full **then**
      Make choices and execute actions randomly.
    else
      Select current option according to (35).
      Get actions based on the intra-option policy: $\mathbf{a}^n=\{\pi_{\omega^n,i}\left(\mathbf{o}^n;\theta^{\pi_{\omega^n}}_i\right)+(1-\xi_d)\omega^n\mathcal{N}_\mathcal{A},i=M,A\}$, where $\mathcal{N}_\mathcal{A}\sim\mathcal{N}(0,\sigma^2_\mathcal{A})$ donates the action noise and $\sigma^2_\mathcal{A}$ donates the noise power.
    end if
    if $\mathbf{a}^n_i\in\mathbf{a}^n$ violates the constraint (18e) or (18d) **then**
      Clip $\mathbf{a}^n_i$ by the action bounds.
    end if
    Execute the clipped actions and get feedbacks in terms of the next state $\mathbf{o}^{n+1}$ and rewards r^n,r^n_ω.
    if $\mathbf{o}^{n+1}$ violates the constraint (18b) or (18c) **then**
      Clip $\mathbf{o}^{n+1}$ by the area bounds.
    end if
    Update the state $\mathbf{o}^n=\mathbf{o}^{n+1}$.
    Store $\left\langle\mathbf{o}^n,r^n_\omega,\mathbf{o}^{n+1},\omega^n,\omega^{n-1},\xi_d\right\rangle$ into D_Ω.
    if $\omega^n=1$ **then**
      Store $\left\langle\mathbf{o}^n,\mathbf{a}^n,r^n,\mathbf{o}^{n+1},\xi_d\right\rangle$ into $D_\mathcal{A}$.
    end if
    if D_Ω is full **then**
      Sample a random minibatch of $\mathcal{B}$ transitions from $D_\Omega,D_\mathcal{A}$ and train the networks $Q_\Omega,\beta,Q,\pi_{1,i}$ by minimizing the losses (38), (39), (21) and (23), respectively.
      Update the target networks using the soft replace methods (41), (42), (24) and (25).
    end if
  end while
  if episode$>\mathcal{M}_d$ **then**
    Decay the action noise in an exponential manner:

$$\sigma_\mathcal{A}=\sigma_{\mathcal{A}_0}\gamma_\mathcal{A}^{\frac{\text{episode}-\mathcal{M}_d}{\mathcal{M}_f}}, \tag{33}$$

    where $\sigma_{\mathcal{A}_0}$ donates the initial standard deviation of $\mathcal{N}_\mathcal{A}$, $\gamma_\mathcal{A}$ denotes the decay rate and $\mathcal{M}_f$ refers to the decay frequency.
  end if
end for

$Q'_{\Omega new} = \max\left(Q'_{\Omega}(\mathbf{o}^{n+1};\theta^{Q'_{\Omega}})\right)$. To avoid the drawbacks of overestimation, we use the double-Q technique [38] instead:

$$Q'_{\Omega new} = Q'_{\Omega}\left(\mathbf{o}^{n+1}, \arg\max_{\omega}\left(Q_{\Omega}(\mathbf{o}^{n};\theta^{Q_{\Omega}})\right)\right). \quad (37)$$

Moreover, the network Q_Ω can be trained in the same way as the global critic Q by minimizing the TD error:

$$L\left(\theta^{Q_\Omega}\right) = \frac{1}{2}\left(y_\Omega - Q_\Omega\left(\mathbf{o}^n, \omega^n\right)\right)^2. \quad (38)$$

In addition, the β net is updated by minimizing the loss masked by ξ_d:

$$L\left(\theta^{\beta}\right) = \beta(\mathbf{o}^n, \omega^{n-1}) A_\Omega * (1-\xi_d), \quad (39)$$

where $A_\Omega = \left(Q_\Omega\left(\mathbf{o}^n, \omega^n\right) - V_\Omega\right)$ is a constant, donating the termination advantage value of option ω^n. When the option choice is suboptimal with respect to the expected value V_Ω over all options, the value A_Ω is negative and it drives the gradient corrections up, which increases the odds of terminating. Note that V_Ω takes the form of max advantage under the greedy policy, i.e. $V_\Omega = \max\left(Q_\Omega(\mathbf{o}^n;\theta^{Q_\Omega})\right)$, or is given in a probabilistic manner:

$$\begin{aligned} V_\Omega = &\left(1-\beta(\mathbf{o}^n,\omega^{n-1})\right) Q_\Omega(\mathbf{o}^n,\omega^n) \\ &+ \beta(\mathbf{o}^n,\omega^{n-1}) \max\left(Q_\Omega\left(\mathbf{o}^n;\theta^{Q_\Omega}\right)\right) \end{aligned} \quad (40)$$

Also, the target networks are updated by applying soft replace, which is the same as (24) and (25):

$$\theta^{Q'_\Omega} = (1-\tau)\,\theta^{Q'_\Omega} + \tau\theta^{Q_\Omega}, \quad (41)$$

$$\theta^{\beta'} = (1-\tau)\,\theta^{\beta'} + \tau\theta^{\beta}. \quad (42)$$

Different from the policy gradient (PG) method used in [37], we no longer design additional networks for intra-option policy learning which can be directly replaced by the CT-MADDPG base, i.e. $\pi_{\omega=1,i} = \mu_i$. In other words, the networks Q_Ω, β and $Q, \pi_{1,i}$ maintain two independent experience buffers for the purposes of options learning and intra-option policies optimization, respectively. Besides, the complete training process of MADDPOC is shown in Algorithm 1.

C. Space Design

Detailed design on the spaces of state, option, action and reward for MADDPOC is elaborated in this subsection.

State: All networks take the observation information $\mathbf{o}^n$ aforementioned in (26) as inputs.

Option: Three options are available in our scenario as mentioned in III-A: $\omega^n = 0$ for charging, $\omega^n = 1$ for cooperative communication and $\omega^n = 2$ for task termination at time slot n. Given the serious problem of training sample truncation brought by the termination option, we make it unselectable in the early task. To be specific, we set an energy threshold E_{th} and reformulate the greedy method (34):

$$\omega^n = \arg\max_{\omega}\left(Q_\Omega\left(\mathbf{o}^n;\theta^{Q_\Omega}\right) - \mathbf{q}_{mask}\right), \quad (43)$$

where

$$\mathbf{q}_{mask} = \begin{cases} [0,0,q_m], & E_{\mathrm{M}} > E_{th}, \\ [0,0,0], & \text{otherwise}. \end{cases} \quad (44)$$

TABLE II: Environment and Training Settings

Parameters	Description	Value
$\eta_a, \eta_b, \eta_{\mathrm{LoS}}, \eta_{\mathrm{NLoS}}$	Dense Urban Environment	12.08, 0.11, 1.6, 23
σ^2	Background noise power	10^{-13}W/Hz
(x_{min}, x_{max})	Abscissa interval of the square area	(-100m, 100m)
(y_{min}, y_{max})	Ordinate interval of the square area	(-100m, 100m)
H_{M}, H_{A}	Flying heights of the MUAV and AUAV	100m, 98m
$\mathbf{l}_{\mathrm{C}}$	Location of the charging station	[-75m, 0m]
δ_τ	Slot duration	0.5s
N_u	Link switching frequency	30
$S_{\mathrm{M}}, S_{\mathrm{A}}$	Number of antennas on the UPA	128, 128
S_{R}	Number of IRS units	128
p_{tmax}	Maximum communication power	10W
p_c	Charge power transmitted by the MUAV	5000W
v_{max}	Maximum velocity of UAVs	20m/s
v_{mee}	Velocity with max efficiency	18.3m/s
l_{min}	Communication coverage distance	25m
α_c	Energy conversion efficiency	0.9
E_{th}	Leftover energy threshold	5000Joule
γ	Reward discount factor	0.95
τ	Soft replace rate	0.001
$\mathcal{B}$	Minibatch size	128
$\mathcal{C}$	Capacity of experience replay buffers	30000
$\mathcal{M}$	Total training epoch	1500
$\sigma_{\mathcal{A}_0}$	Initial action noise standard deviation	0.15 or 0.3
$\gamma_{\mathcal{A}}$	Decay rate of the action noise	0.99
$\mathcal{M}_f$	Decay frequency of the action noise	2
$\mathcal{M}_d$	Starting epoch of the noise decay	1000
$\kappa_{th}, \kappa_{cb}, \kappa_{dc}, \kappa_{po}$	Reward settings	1.2, -5, -5, -1000

Note that q_m is a big enough positive number, and the probabilistic method (35) can be rewritten in the same way.

Action: We have described the intra-option policies $\pi_{0,i}$ and $\pi_{2,i}$ as charging and termination processes, respectively, in the subsection III-A. In other words, when the agents choose either of the policies, their actions or action sequences to be executed are already artificially determined and no longer the outputs of their actor networks. Besides, CT-MADDPG is served as a surrogate of $\pi_{1,i}$. Thus, the action space of $\pi_{1,i}$ follows (27) and (28).

Reward: Two extra sparse rewards are added which helps to make options learning accurate and rapidly convergent:

- Deliberate cost: Charging can ensure the AUAV's endurance, but too much will impact the throughput performance. Also, an early return is a waste of energy. Therefore, we expect reasonable option switching only when necessary. Inspired by the deliberate cost proposed in [39], a penalty to control the switching frequency is set as

$$r^n_{dc} = \left(\xi_d + (1-\xi_d)\omega^{n-1}\left(1-\omega^n\right)\right)\kappa_{dc}, \quad (45)$$

  where κ_{dc} is a negative constant. Note that only the switch from communication to charging or termination is penalized since the former is what we encourage.
- Power-off penalty: To avoid the air crash midway due to battery exhaustion, we set a great penalty as

$$r^n_{po} = \xi_{po}[n]\kappa_{po}, \quad (46)$$

  where $\xi_{po}[n]$ donates a binary indicator representing whether the constraint (19b) is violated and κ_{po} is a negative constant.

In summary, the total reward to compute the update target of Q_Ω at time slot n is formulated as

$$r^n_\omega = r^n + r^n_{dc} + r^n_{po}. \quad (47)$$

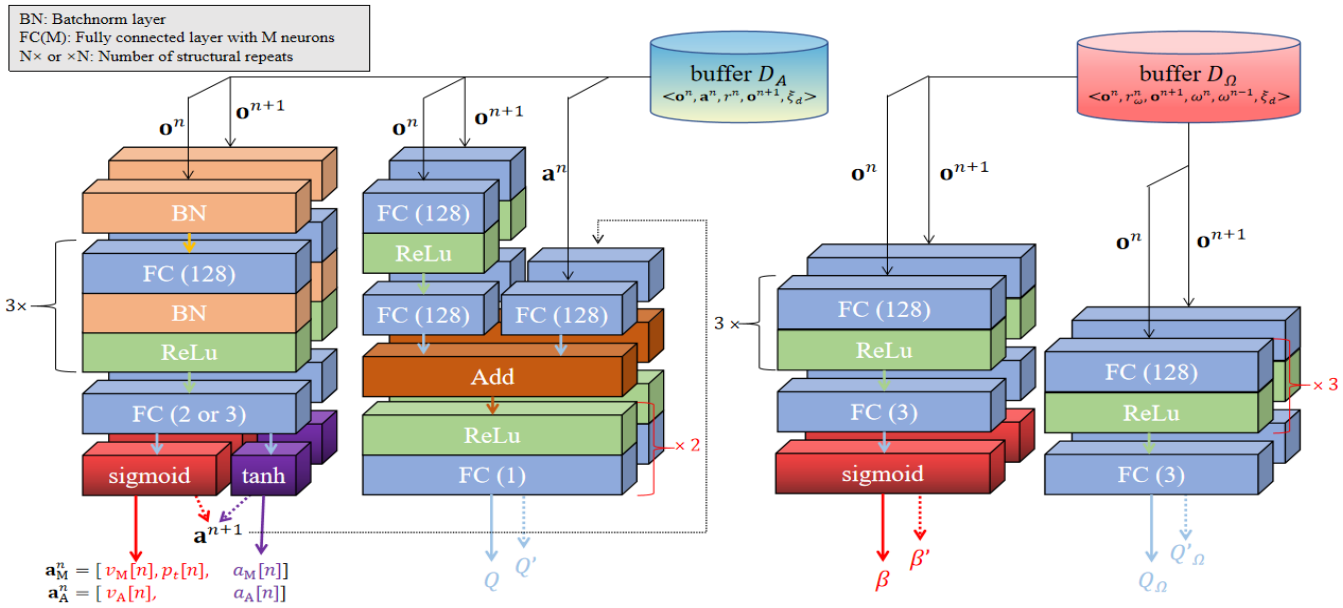


Fig. 5: Network structures of $\pi_{1,i}, Q, \beta, Q_\Omega$ and their target nets (from left to right).

VI. Numerical Results and Performance Evaluation

In this section, we present the simulation results as well as the performance analysis of the proposed algorithm.

A. Environment Settings

Related environment parameters are summarized in the front part of Table II. In order to minimize the length of time interval for the termination judgment, we set the threshold $E_{th} = 2p_c\sigma_\tau$, i.e., at least one more chance to charge, since the selection of termination before is a waste of energy and also makes the training samples truncated. Besides, the unspecified parameters, i.e. the number of GNs K, and the fully charged energy $E_{\mathrm{M}_{\max}}, E_{\mathrm{A}_{\max}}$, can be set manually.

B. Training Settings

We train MADDPOC using the Adam optimizer with learning rate 0.001. Specific structures of all networks, including $\pi_{1,i}, Q, \beta, Q_\Omega$ as well as their corresponding target nets, are shown in Fig 5. Given the inconsistent length of action normalization intervals, we make the initial derivation of action noise $\sigma_{\mathcal{A}_0}$ proportional to the normalized action range, e.g., $\sigma_{N_v} = 0.15, \sigma_{N_a} = 0.3, \sigma_{N_p} = 0.15$ for velocity, azimuthal angle and communication power, respectively. The rest parameters are shown in the lower half of Table II.

C. Performance and Analysis

We compare MADDPOC with other three baseline methods, MADDPOC without option masking (MADDPOC-NM), deep deterministic policy option critic (DDPOC) and option critic OC, while the CT-MADDPG algorithm proposed for (P1) is served as a contrastive benchmark in the non-charging scenario.

- MADDPOC-NM: No option masking is applied and all options are assumed to be selectable anytime in the framework of MADDPOC.
- DDPOC: A single-agent version of MADDPOC, where $\pi_{1,i}$ is replaced by a single actor network π_1 that outputs the actions of both UAVs.
- OC: The classical hierarchical architecture proposed in [37], where π_1 is trained by applying the policy gradient weighted by the advantage value A_Ω.

The latter two methods apply option masking by default. Unless specified otherwise, our default environment configuration is set to: $E_{\mathrm{M}_{\max}} = 35000$ Joule, $E_{\mathrm{A}_{\max}} = 15000$ Joule, and $K = 10$.

Fig. 6 presents the rewards convergence for different algorithms in the training phase. It can be seen that the MADDPOC's performance is the best before noise decay, while DDPOC suffers from high action dimensions and convergence a little bit slower. Besides, MADDPOC-NM is easily trapped into local optimal, i.e. prematurely terminate with surplus MUAV's energy. The reason lies in the hierarchical algorithm's inherent drawback of the mutual interference between upper and lower levels of decision-making. Specifically in our scenario, the over-option policy π_Ω cannot find appropriate charging points when the intra-option policy $\pi_{1,u}$ is not convergent, and thus related training samples are not generated because of the early termination. Moreover, accumulative rewards of all algorithms except OC have remarkable growth after noise attenuation, since the same side constraints (18h), (19c) are easily violated before. Note that the action deviations in OC architecture is served as trainable variables and artificial attenuation is impracticable, leading to the training instability and poor performance. Meanwhile in this stage, MADDPOC-NM steps out of learning stagnation and finds other appropriate charging points. In addition, CT-MADDPG cannot utilize the extra energy of MUAV and its throughput performance

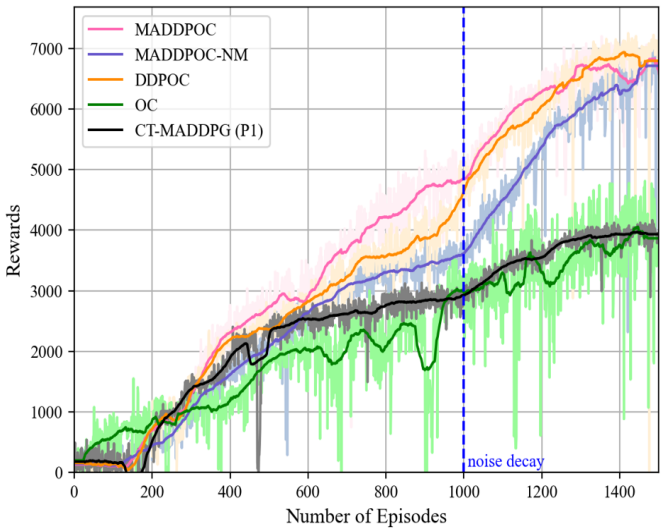


Fig. 6: Rewards convergence in training phase.

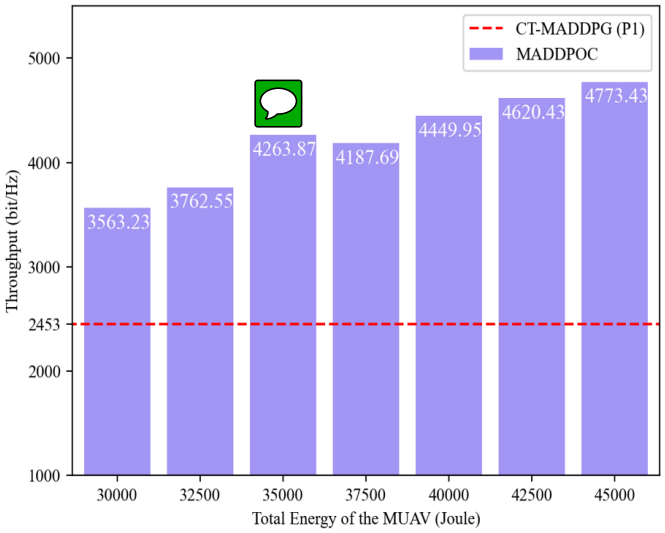


Fig. 8: Accumulated throughput vs MUAV's energy limit.

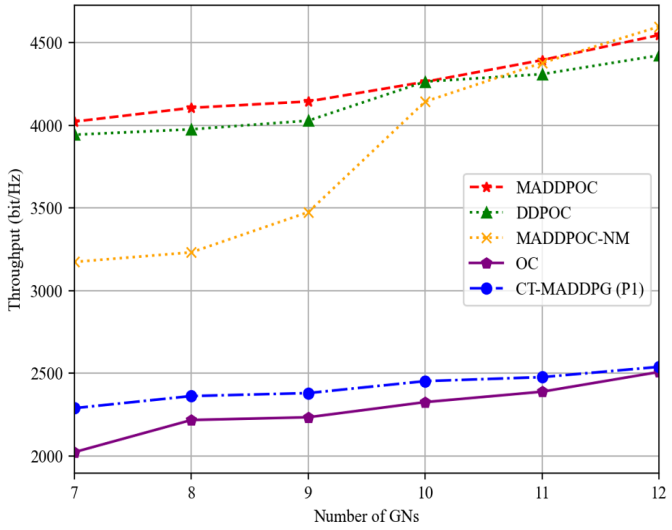


Fig. 7: Accumulated throughput vs number of GNs.

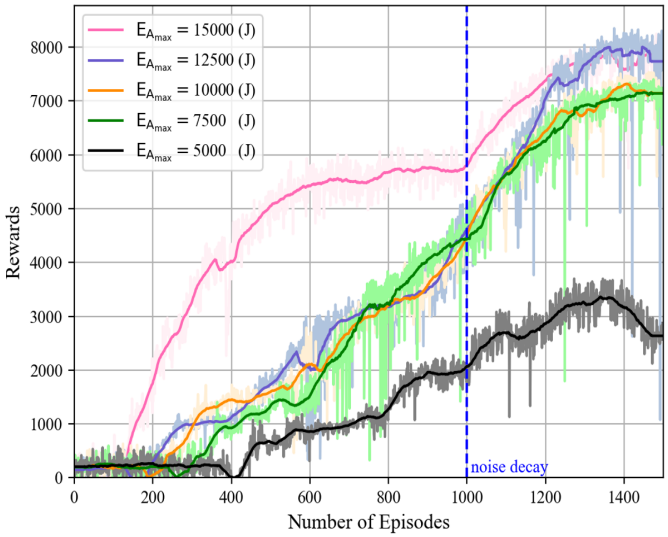


Fig. 9: Reward curves for different AUAV's energy limit.

depends on the AUAV's energy.

In Fig. 7, we compare the total throughput obtained by aforementioned methods varying with different numbers of GNs in deployment phase, where $\sigma_{\mathcal{A}} = 0$. We observe that MADDPOC has similar performance to DDPOC. Meanwhile, the total throughput has an upward trend, it is because that UAVs have higher probability of covering users when the number of GN increases. Besides, the performance of MADDPOC-NM is not stable, since it may not step out of local optimal, and unable to take full advantage of the charging mechanism. Moreover, the OC's performance is even below CT-MADDPG's. Because OC could not make reasonable choices in our scenario and degenerate into PG, which suffers from high dimensional actions and the constraint (19c).

Fig. 8 shows the impact of different MUAV's fully charged energy on throughput performance. An interesting thing we observe is that the throughput performance not always shows an increasing trend as the energy upper limit increases. There is even a decline when the upper limit increases from 35000 Joules to 37500 Joules, of which the reason is twofold. On the one hand, it can be observed from Fig 11(a) that both UAVs almost run out of energy simultaneously and terminate near the charging station in the former case. Agents tend to use similar trajectory policies if only the energy constraint changes. Thus, the MUAV has about 2500 Joules left which cannot be utilized in the latter case, since it's close to the energy required for a single charging. On the other hand, the over-option policy becomes torn between whether or not charge when the AUAV's battery is low in the case aforementioned, leading to the poor performance of the intra-option policy in the last few steps.

Similar situations can be seen in Fig. 9, where the rewards convergence of MADDPOC varying with different AUAV's energy limits is shown and $E_{\mathrm{M}_{\max}}$ is set to 45000 Joules. The throughput increases stepwise as the AUAV's energy increases, since the energy-underutilized phenomenon also exists in the cases of $E_{\mathrm{A}_{\max}} = 15000$ Joules and $E_{\mathrm{A}_{\max}} = 10000$ Joules. Another interesting thing is that the pink curve reaches a

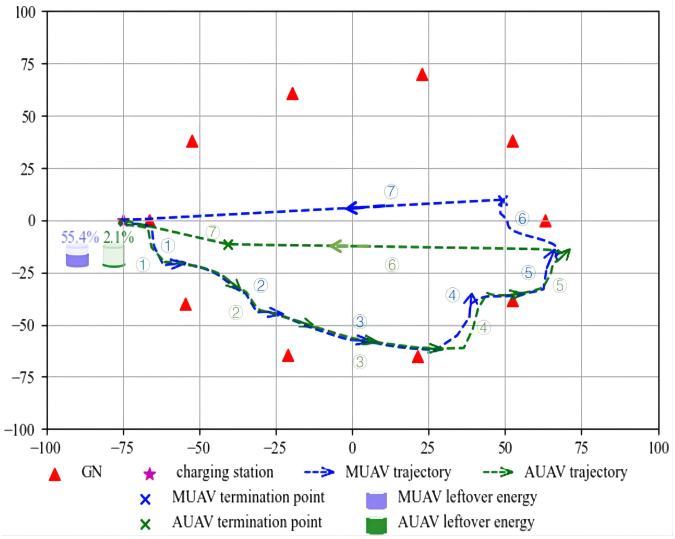


Fig. 10: The trajectory of non-charging scenario.

plateau at about 600 epochs, which is much faster than the others. Because the higher the AUAV's energy limit, the lower the probability of constraint (19b) violation. Then more positive samples will be provided within an epoch, making the training process more efficient. Besides, MADDPOC doesn't work when the AUAV's battery capacity becomes severe, e.g. $E_{\mathrm{A}_{\max}} = 5000$ Joules, since the high-frequency violation of the constraint (19b) makes the training process extremely difficult.

Fig. 10 plots the trajectory of non-charging scenario. The AUAV can only serve half of GNs and have to turn back midway. We observe that the AUAV will terminate somewhere close to the charging station, while the MUAV keeps quasi-stationary over the linked GN until its termination. On the one hand, there is a trade-off between termination and subsidiary communication for the AUAV, while the MUAV with high battery capacity only need to focus on providing high quality service. On the other hand, the range of MUAV's operation area is strictly restrained by (18i), so it cannot leave too far from the linked GN.

Fig. 11 shows the trajectories of charging scenario under different energy constraints. Note that the segment number pair, e.g. the ① in blue and green, represents the flight trajectories of two UAVs in the same period of time. The leftover energy of UAVs at each charging point as well as after termination is also labeled in terms of cylindrical icon. It can be observed that two trajectories almost coincide anytime, which depends on a high-degree cooperation between two actors $\pi_{1,\mathrm{M}}$ and $\pi_{1,\mathrm{A}}$. Because the data rate can be maximized when the IRS is close enough to the transmitter or the receiver. In this case, the constraint (18h) is easy to be violated even with a low action deviation, which supports the observation of Fig. 6. Besides, the leftover energy of both UAVs remains positive during the operation time, which guarantees the validity of trajectories. We also observe that MADDPOC always ends up with one or both ultimate energy of two UAVs since it's easy to judge from the positivity and negativity of the state information $E_{\mathrm{M}l}[n]$ and $E_{\mathrm{A}l}[n]$.

The numerical results under different energy constraints, including the items of total operation slots, charging count, leftover energy, accumulative throughput and energy efficiency are summarized in table III. It can be observed that MADDPOC is unable to take fully advantages of both UAVs' energy in most cases. Especially in the 4th case, working hours barely changes whether charge three or four times. Because it no longer makes sense to charge when the MUAV's leftover energy is just above the threshold for a single charging $p_c\delta\tau$ and the AUAV is running out of juice. The other key factor affecting energy efficiency is the termination location. The farther away UAVs are from the charging station at the termination step, the more energy UAVs wastes to return.

VII. Conclusions and Future Work

In this paper, we have investigated a MUAV-powered IoT network, in which we have proposed using a rechargeable AUAV equipped with an IRS to enhance the communication signals from the MUAV and also leverage the MUAV as a recharging power source. Under the proposed model, we have investigated the optimal collaboration strategy of these energy-limited UAVs to maximize the accumulated throughput of the IoT network. Depending on whether there is charging between the two UAVs, two optimization problems have been formulated. To solve them, we have proposed the CT-MADDPG algorithm to enable the UAVs' cooperative communication and the MADDPOC algorithm, which supports the multi-agent hierarchical control. The numerical results show that MADDPOC performs better than the benchmarks in terms of the convergence rate and the accumulative throughput. Furthermore, we also observed that the throughput performance not always shows an increasing trend as the energy limit increases because of the existence of energy-underutilization phenomenon.

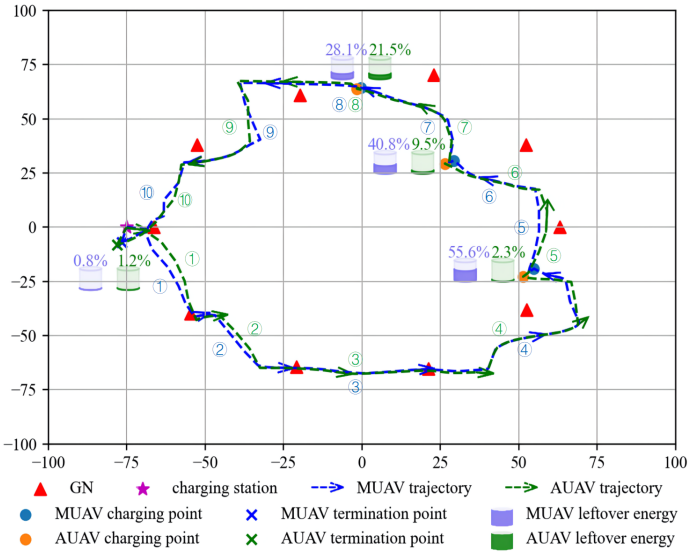
Although we deploy an IRS-equipped AUAV to help with communications, the phase optimization of IRS is replaced by MRT for simplicity. Given the dimensional differences among actions, it's not recommended to output the trajectories, transmit power and the phase of IRS simultaneously with a single network, since the phase with excessive dimension severely affects the performance of the other two low-dimensional actions. Thus, another phase-dedicated network is needed to be pre-trained before the optimization of trajectories and transmit power. In the case of perfect channel prediction, the channel matrix is formulated by the positions of UAV, IRS and linked GN, which can be effectively mapped to the optimized phase by deep transfer learning [40], while in the unperfect case, additional information from past channels is needed [41]. For the future work, we will investigate the trajectory design and phase optimization of multiple IRS-equipped AUAVs in a MUAV-powered IoT network.

Appendix A
Optimal Phase Shift Control at the IRS

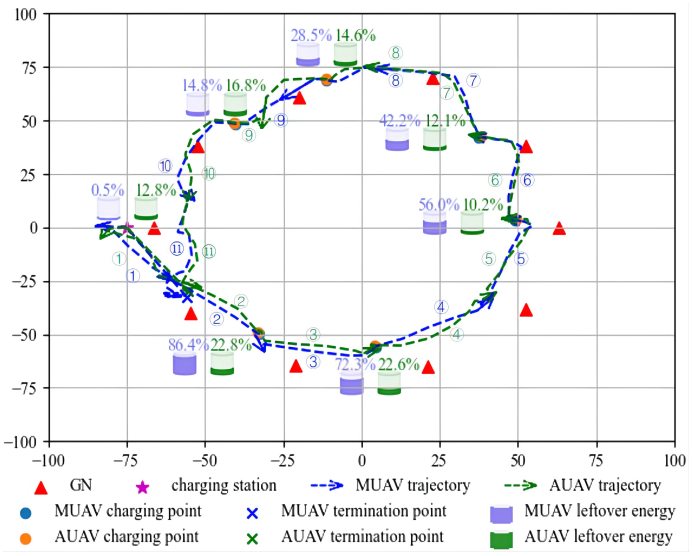
In the appendix, we derive the optimal phase control strategy at the IRS in our scenario. In the case of MRT, the signal

TABLE III: Numerical results under different energy constraints.

Index	$E_{\mathrm{M}_{\max}}$ (J)	$E_{\mathrm{A}_{\max}}$ (J)	N_τ	Charging count	$E_{\mathrm{M}}[N_\tau]$ (%)	$E_{\mathrm{A}}[N_\tau]$ (%)	Throughput (bit/Hz)	Energy efficiency (bit/Hz/Joule)
1	30000	15000	270	2	5.05%	$< 1\%$	3563	0.0792
2	32500		291	3	$< 1\%$	15.03%	3762	0.0792
3	35000		316	3	$< 1\%$	1.2%	4263	0.0853
4	37500		317	3	7.39%	1.40%	4187	0.0798
			318	4	$< 1\%$	17.37%	4151	0.0791
5	40000		340	4	$< 1\%$	8.57%	4449	0.0809
6	42500		358	4	3.67%	2.31%	4620	0.0803
7	45000		368	5	1.25%	13.53%	4773	0.0796
8		12500	365	5	$< 1\%$	$< 1\%$	4725	0.0822
9		10000	343	5	5.80%	$< 1\%$	4439	0.0807
10		7500	338	6	$< 1\%$	12.76%	4389	0.0836
11		5000	187	10-13	\	\	2299	0.0460



(a) $E_{\mathrm{M}_{\max}} = 35000$ Joule, $E_{\mathrm{A}_{\max}} = 15000$ Joule



(b) $E_{\mathrm{M}_{\max}} = 45000$ Joule, $E_{\mathrm{A}_{\max}} = 7500$ Joule

Fig. 11: The trajectories of charging scenario under different energy constraints.

noise ratio in (10) can be formulated as:

$$\begin{aligned}\gamma_s[n] &= \left\|\left(\mathbf{h}_{\mathrm{MG}_k}[n]\right)^H \mathbf{w}_k[n]\right\|^2 / \sigma^2 \\ &= \frac{S_{\mathrm{R}}^2 S_{\mathrm{M}}}{\sigma^2}\left\|\left(\mathbf{h}_{\mathrm{RG}_k}[n]\right)^H \mathbf{\Phi}[n]\,\mathbf{h}_{\mathrm{RM}}[n]\right\|^2 \\ &= \frac{S_{\mathrm{R}}^2 S_{\mathrm{M}}}{\sigma^2}\left\|\sum_{x=0}^{S_{\mathrm{R}_x}}\sum_{y=0}^{S_{\mathrm{R}_y}} e^{j\left(\varphi_{xy}^{\mathrm{RG}}[n]+\varphi_{x,y}[n]+\varphi_{xy}^{\mathrm{RM}}[n]\right)}\right\|^2. \end{aligned} \quad (48)$$

Note that $\mathbf{h}_{\mathrm{RM}}[n] = \mathbf{a}\left(\theta_{\mathrm{MR}}^{\mathrm{A}}[n], \xi_{\mathrm{MR}}^{\mathrm{A}}[n], S_{\mathrm{R}_x}, S_{\mathrm{R}_y}\right)$ according to (3), $\varphi_{xy}^{\mathrm{RG}}[n], \varphi_{xy}^{\mathrm{RM}}[n]$ denote the phases of xy-th elements of vectors $\left(\mathbf{h}_{\mathrm{RG}_k}[n]\right)^H, \mathbf{h}_{\mathrm{RM}}[n]$, respectively, and $\varphi_{x,y}[n]$ donates the phase of IRS's x, y-th unit. Let $\varphi_{xy}^{\mathrm{sum}}[n]$ represent the sum of the three phases. The formula (48) satisfies the Cauchy inequality:

$$\left\|\sum_{x=0}^{S_{\mathrm{R}_x}}\sum_{y=0}^{S_{\mathrm{R}_y}} e^{j\varphi_{xy}^{\mathrm{sum}}[n]}\right\| \le \sum_{x=0}^{S_{\mathrm{R}_x}}\sum_{y=0}^{S_{\mathrm{R}_y}}\left\|e^{j\varphi_{xy}^{\mathrm{sum}}[n]}\right\|, \quad (49)$$

equal if and only if $\varphi_{xy}^{\mathrm{sum}}[n] = \varphi, \forall x, y, n$, where $\varphi \in [0, 2\pi)$ is a constant. Let $\varphi = 0$, we can get $\varphi_{x,y}[n] = -(\varphi_{xy}^{\mathrm{RG}}[n] + \varphi_{xy}^{\mathrm{RM}}[n])$. Then the optimal phase in (9) can be obtained by substituting the specific expressions in (3) and (5).

REFERENCES

[1] J. Xu, X. Kang, R. Zhang, and Y.-C. Liang, "Joint power and trajectory optimization for IRS-aided master-auxiliary-UAV-powered IoT networks," in *Proc. IEEE GLOBECOM*, Madrid, Spain, Dec 2021, Accepted Paper.

[2] Y. Mehmood, F. Ahmad, I. Yaqoob, A. Adnane, M. Imran, and S. Guizani, "Internet-of-Things-based smart cities: Recent advances and challenges," *IEEE Communications Magazine*, vol. 55, no. 9, pp. 16–24, 2017.

[3] H. L. J. Ting, X. Kang, T. Li, H. Wang, and C.-K. Chu, "On the trust and trust modeling for the future fully-connected digital world: A comprehensive study," *IEEE Access*, vol. 9, pp. 106 743–106 783, 2021.

[4] S. Mohanty, U. Choppali, and E. Kouglanos, "Everything you wanted to know about smart cities: the Internet of Things is the backbone," *IEEE Consum. Electron. Mag.*, vol. 16, pp. 2162–2248, 2016.

[5] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016.

[6] G. Ding, Q. Wu, L. Zhang, Y. Lin, T. A. Tsiftsis, and Y.-D. Yao, "An amateur drone surveillance system based on the cognitive Internet of Things," *IEEE Communications Magazine*, vol. 56, no. 1, pp. 29–35, 2018.

[7] G. Yang, R. Dai, and Y.-C. Liang, "Energy-efficient UAV backscatter communication with joint trajectory design and resource optimization," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 926–941, 2021.

[8] H.-T. Ye, X. Kang, J. Joung, and Y.-C. Liang, "Optimization for full-duplex rotary-wing UAV-enabled wireless-powered IoT networks," *IEEE*

Transactions on Wireless Communications, vol. 19, no. 7, pp. 5057–5072, 2020.

[9] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, 2018.

[10] H.-T. Ye, X. Kang, J. Joung, and Y.-C. Liang, "Joint uplink-and-downlink optimization of 3D UAV swarm deployment for wireless-powered IoT networks," *IEEE Internet of Things Journal*, vol. PP, no. 99, pp. 1–1, 2021.

[11] S. Khairy, P. Balaprakash, L. X. Cai, and Y. Cheng, "Constrained deep reinforcement learning for energy sustainable multi-UAV based random access IoT networks with NOMA," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1101–1115, 2021.

[12] X. Zhai, H. Wang, J. Li, Z. Huang, and R. Gao, "A wireless charging method with lightweight pick-up structure for UAVs," *Electrical Engineering*, 2021.

[13] Y. Yan, W. Shi, and X. Zhang, "Design of UAV wireless power transmission system based on coupling coil structure optimization," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, 2020.

[14] M. Li, L. Liu, Y. Gu, Y. Ding, and L. Wang, "Minimizing energy consumption in wireless rechargeable UAV networks," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[15] Y.-C. Liang, R. Long, Q. Zhang, J. Chen, H. V. Cheng, and H. Guo, "Large intelligent surface/antennas (LISA): Making reflective radios smart," *Journal of Communications and Information Networks*, vol. 4, no. 2, pp. 40–50, 2019.

[16] Y. Cai, Z. Wei, S. Hu, D. W. K. Ng, and J. Yuan, "Resource allocation for power-efficient IRS-assisted UAV communications," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–7.

[17] Z. Wei, Y. Cai, Z. Sun, D. W. K. Ng, J. Yuan, M. Zhou, and L. Sun, "Sum-rate maximization for IRS-assisted UAV OFDMA communication systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2530–2550, 2021.

[18] S. Li, B. Duo, M. D. Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," *IEEE Transactions on Wireless Communications*, vol. 20, no. 10, pp. 6402–6417, 2021.

[19] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5218–5233, 2020.

[20] H. Lu, Y. Zeng, S. Jin, and R. Zhang, "Enabling panoramic full-angle reflection via aerial intelligent reflecting surface," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[21] T. Shafique, H. Tabassum, and E. Hossain, "Optimization of wireless relaying with flexible UAV-borne reflecting surfaces," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 309–325, 2021.

[22] S. Jiao, F. Fang, X. Zhou, and H. Zhang, "Joint beamforming and phase shift design in downlink UAV networks with IRS-assisted NOMA," *Journal of Communications and Information Networks*, vol. 5, no. 2, pp. 138–149, 2020.

[23] M. Al-Jarrah, A. Al-Dweik, E. Alsusa, Y. Iraqi, and M.-S. Alouini, "On the performance of IRS-assisted multi-layer UAV communications with imperfect phase compensation," *IEEE Transactions on Communications*, pp. 1–1, 2021.

[24] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multi-agent systems: A review of challenges, solutions, and applications," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[25] Y. Nie, J. Zhao, J. Liu, J. Jiang, and R. Ding, "Energy-efficient UAV trajectory design for backscatter communication: A deep reinforcement learning approach," *China Communications*, vol. 17, no. 10, pp. 129–141, 2020.

[26] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Transactions on Mobile Computing*, vol. 19, no. 6, pp. 1274–1285, 2020.

[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *Computer ence*, 2015.

[28] J. Baek, S. I. Han, and Y. Han, "Energy-efficient UAV routing for wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1741–1750, 2020.

[29] R. Ding, F. Gao, and X. S. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 19, no. 12, pp. 7796–7809, 2020.

[30] Y. Zhang, Z. Mou, F. Gao, L. Xing, J. Jiang, and Z. Han, "Hierarchical deep reinforcement learning for backscattering data collection with multiple UAVs," *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3786–3800, 2021.

[31] C. Yu, A. Velu, E. Vinitsky, Y. Wang, and Y. Wu, "The surprising effectiveness of MAPPO in cooperative, multi-agent games," 2021.

[32] X. Kang, Y.-C. Liang, and J. Yang, "Riding on the primary: A new spectrum sharing paradigm for wireless-powered iot devices," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6335–6347, 2018.

[33] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[34] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, 2019.

[35] P. Montague, "Reinforcement learning: An introduction, by Sutton, R.S. and Barto, A.G." *Trends in Cognitive Sciences*, vol. 3, no. 9, p. 360, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364661399013315

[36] Z. Mou, Y. Zhang, F. Gao, H. Wang, T. Zhang, and Z. Han, "Deep reinforcement learning based three-dimensional area coverage with UAV swarm," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3160–3176, 2021.

[37] P. L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," *AAAI*, pp. 1726–1734, 2017.

[38] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," *Computer ence*, 2015.

[39] J. Harb, P.-L. Bacon, M. Klissarov, and D. Precup, "When waiting is not an option : Learning options with a deliberation cost," 2017.

[40] Y. Ge and J. Fan, "Beamforming optimization for intelligent reflecting surface assisted MISO: A deep transfer learning approach," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 4, pp. 3902–3907, 2021.

[41] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, 2020.