Chinese Journal of Electronics
Vol.32, No.6, Nov. 2023

Multi-Objective Coordinated Optimization for UAV Charging Scheduling in Intelligent Aerial-Ground Perception Networks

ZHOU Yi[1,3], CHENG Xiang[1,3], SHI Huaguang[1,2], JIN Zhanqi[1,3], NING Nianwen[1,3], and LIU Fuqiang[4]

(1. *School of Artificial Intelligence, Henan University, Zhengzhou 450046, China*)

(2. *Henan Engineering Research Center for Industrial Internet of Things, Zhengzhou 450046, China*)

(3. *International Joint Research Laboratory for Cooperative Vehicular Networking, Zhengzhou 450046, China*)

(4. *College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China*)

Abstract — The unmanned aerial vehicles (UAVs)-assisted intelligent traffic perception system can provide effective situation awareness. However, UAVs are required to be recharged before the energy is exhausted, which may cause task interruption. To address this concern, the charging UAV (CUAV) is employed to provide wireless charging for the mission UAVs (MUAVs). This paper studies the charging scheduling problem of the CUAV under the premise of optimizing the MUAVs deployment. We first model the MUAVs deployment problem considering the energy consumption and data transmission and establish the CUAV charging model. Then, the above problem is formulated as a multi-objective multi-agent stochastic game process to simplify the decisions-making of MUAVs and CUAV, based on which we propose the utility-based Pareto optimal deployment and charging algorithm, which reduces the computing complexity by equivalent utility of the MUAVs while using Kullback-Leibler divergence to constrain solutions. Next, to ensure the effectiveness of policy update, the multi-agent communication protocol is adopted to improve policy exploration efficiency. Simulation results show that the proposed algorithm outperforms existing works in terms of energy efficiency and charging by comparing with the Pareto front of different methods, endurance anxiety of the MUAVs, and charging utilization under different task modes.

Key words — Wireless charging, UAVs deployment, Multi-objective optimization, Multi-objective reinforcement learning, Pareto optimal.

I. Introduction

Nowadays, traffic perception mainly relies on fixed sensor equipment. However, the quality of collected data is vulnerable to the blind spots of cameras and the bad weather, and thus cause the wrong scheduling judgments of control center [1]. With the popularization of 6G network for the space-air-ground integration, unmanned aerial vehicles (UAVs) characterized by high mobility can be rapidly deployed to cover roads and conduct tasks by equipping with cameras [2]. Mission UAVs (MUAVs) transmit the data to the base station by continuous visual field coverage [3], [4]. However, the limited battery capacity of MUAVs cannot support durable flight [5]. It is worth noting that the energy constraint problem cannot be solved well by merely optimizing the energy of MUAVs [6]. Thus, the wireless power transmission (WPT) technology for UAVs has emerged to provide MUAVs with sustainable energy. As a stable energy source, the charging UAV (CUAV) is employed to provide wireless charging for MUAVs to ensure their endurance [7].

In the above scenario, the deployment of MUAVs and charging of CUAV are two main problems. Most studies regard UAVs task energy efficiency as the main optimization objective [8]–[11]. The researches on UAV charging include single objective optimization [12] that considers charging and multi-objective optimization (MOO) that optimizes both charging and deployment [13], [14]. The former only focuses on charging and ignores UAVs tasks while the latter ignores the coupling relationship between charging and deployment.

Manuscript Received Nov. 16, 2022; Accepted Feb. 14, 2023. This work was supported by the National Natural Science Foundation of China (62176088), the Program for Science & Technology Development of Henan Province (212102210412, 222102210067, 222102210022), and the Young Elite Scientist Sponsorship Program by Henan Association for Science and Technology (2022HYTP013).

For the deployment and charging problems of UAVs, MOO methods have attracted extensive attention [15], [16]. However, they usually transform the MOO problems into single objective by weighted sum, which is difficult to find reasonable weight and may lead to local optimal solution [17]. To address these issues, Pareto optimal solutions have emerged [18]. However, because the solution search space will increase when the number of objectives increase, the solution process will consume huge computing resources.

Reward driven deep reinforcement learning (DRL) [19] can make up for the complex solution of nonlinear problems. In this paper, a multi-agent deep reinforcement learning (MADRL) [20] method combining with the Pareto optimal solution is used to obtain optimal strategies. The main contributions are as follows.

- The MOO problem of UAVs deployment and charging is modelled as a multi-agent multi-objective stochastic game. In addition, the utility-based Pareto optimal deployment and charging (UPDC) algorithm is proposed by considering the MADRL and MOO to guarantee the tradeoff of solutions.
- To ensure rapid update of joint policies, the CUAV is regard as altruistic agent, and combine equivalent utility with the advantage function to improve learning efficiency of the UPDC algorithm. In addition, the KL divergence is used as the constraint to ensure Pareto optimal policies of deployment and charging are solved within the trust region.
- To ensure the effectiveness of policy update, the temporal message control (TMC) protocol is leveraged based on vectorized multi-objective agent network and message transmission process is filtered to improve the exploration efficiency of optimal strategies.

The remainder of this paper is organized as follows. Section II discusses the related work. In Section III, we first describe the deployment of MUAVs and the charging scheduling of CUAV, respectively, and then introduce the multi-objective optimization problem. Section IV describes the implementation process of the proposed UPDC algorithm. The results and analyses of the experiments are presented in Section V. We conclude our paper and discuss future research work in Section VI.

II. Related Work

This section will review the related work from the following two aspects: 1) UAVs deployment and energy optimization; 2) UAVs deployment and charging.

1. UAVs deployment & energy optimization

For UAVs deployment, one of the objectives is to ensure that MUAVs complete tasks within the energy constraints. For the UAV-aided communication, Li *et al.* [21] considered deploying MUAVs to provide expandable connection for users, in which the network energy efficiency is maximized by jointly optimizing the UAV association, location and resource allocation. For deployment and energy optimization in mobile edge computing, Chen *et al.* [22] formulated the information freshness-aware task offloading as a stochastic game, which aims to maximize long-term payoff of mobile users. Pang *et al.* [23] presented an IRS-assisted UAV network, which aims to guarantee high energy-efficiency communication between the UAV and users.

To provide dynamic services for IoT devices, Mozaffari *et al.* [24] studied UAVs-assisted data collection, and proposed a new framework to optimize space layout and movement to minimize energy consumption of MUAVs. Liu *et al.* [25] proposed an efficient DRL-based method to maximize the energy efficiency considering communication coverage and fairness. Samir *et al.* [26] designed a trajectory optimization method for MUAVs to maximize vehicle coverage and minimize the energy consumption. And Zhang *et al.* [27] proposed an energy-saving deployment algorithm by balancing the flight distance and final service altitude of heterogeneous MUAVs to maximize residual energy. Huang *et al.* [28] studied the UAV pair-supported relaying problem in IoT systems, which receiver is used as relay between transmitter and destination, and proposed dueling DDQN method to solve the optimization problem. Wu *et al.* [29] proposed a multi-UAV-based cooperative framework to balance the accuracy and efficiency of UAVs searching and localization, and studied two algorithms to decide UAVs flight direction and estimate the position of interference source.

The above researches [21]–[29] mainly focus on optimizing the deployment. However, MUAVs cannot provide long-term services due to the limited energy. For the UAV-assisted intelligent traffic perception task considered in this paper, the uninterrupted tasks of the MUAVs are crucial. Therefore, it is necessary to optimize the charging scheduling scheme in dynamic deployment process to maintain continuous tasks.

2. UAVs deployment & charging

To maintain the continuous task of MUAVs, the CUAV is employed to provide charging. Zhu *et al.* [30] proposed a WPT scheme for UAVs data collection and designed a trajectory scheduling algorithms for the CUAV to provide charging. Similarly, Fu *et al.* [31] studied the problem of data collection based on wireless charging, and employed Q-learning to find the optimal strategy. Xiong *et al.* [32] solved the MOO problem to obtain the best strategy for MUAV collect and transmit data while maximizing the system long-term effectiveness. Fu *et al.* [33] optimized the UAVs trajectory while maximizing residual energy of the CUAVs. Li *et al.* [7] established a uninterruptible charging model, and formalized charging problem as an optimization problem to reduce the charging waste. However, the work aforementioned [7], [31]–[32] focus on charging for the MUAVs under full deployement, but ignores the coupling relationship between charging and deployment. Finally, we summarize the critical aspects and difference for the existing literature and our work in Table 1.

Table 1. Summary of relevant papers in critical aspects and difference

Optimization objective	References	Critical aspects	Difference from our work
UAVs Energy Efficient	[6], [8]–[11], [21]–[29]	1) Energy-aware deployment; 2) Energy efficient trajectory; 3) Cooperative deployment.	Single objective optimization of UAVs task energy efficiency
UAVs Charging Scheduling	[5], [7], [12]–[14]	1) Charge scheduling optimization; 2) Nondisruptive UAV charging; 3) Seamless and long-term service.	Single objective optimization of wireless power transmission, charging coverage and scheduling
UAVs Energy Efficient and Charging Scheduling	[15], [17], [30]–[33]	1) Joint deployment and charging; 2) Multi-objective optimization; 3) Far field wireless charging.	Multi-objective optimization of energy efficiency and charging is converted into a single objective problem by weighted sum

This paper mainly focuses on how to obtain tradeoff solutions among coupling-relationship MOO. Some researches have explored novel methods to achieve a balance between two conflicting optimization objectives. For example, Reymond *et al.* [34] proposed the Pareto-DQN algorithm to estimate the Pareto front with a high-dimensional state-space and could obtain the approximately real Pareto front. Wang *et al.* [35] proposed the Pareto-optimal actor-critic method to obtain optimal policies by optimizing the coupling objectives, which was not affected by the concavity and convexity of the Pareto front. For convenience of reading, the important symbols are listed in Table 2 with the corresponding descriptions.

Table 2. Table of Important Symbols

Symbol	Description	Symbol	Description
m, M	The index and total number of MUAVs	$V_t^{m,k}$	Data transmission rate from MUAV$_m$ to MUAV$_k$
$U_{m,t}$	The position of MUAVs	ξ	Energy efficiency of MUAVs
E_m	Propusion energy consumption of MUAV$_m$	ρ	Charging capacity of the CUAV
E_{re_m}	The residual energy of MUAV$_m$	r_t	The reward value
E_0	Initial energy of MUAVs	s_t, a_t	State and action of all UAVs
$E_{\text{threshold}}$	The residual energy threshold of MUAVs	V^{π_i}	Utility of single policy
$E_{\min}, E_{\max}$	The minimum and maximum charging capacity	N	The number of intersections

III. System Model and Problem Formulation

Fig.1 shows the rechargeable UAVs-assisted traffic perception system, which consists of two kinds of UAVs. The MUAVs are deployed for perception, and one CUAV is employed for the charging. M MUAVs patrols among N target areas. T consecutive time slots with equal slot duration τ are considered. There are two patrol modes for MUAVs, which can serve different traffic conditions in an energy-saving way.

- Patrol mode 1: High-connectivity priority based cruise (HCPC). Each MUAV is responsible for covering one road section in time slot t $(t = 1, 3, \ldots, T)$ in the form of round-trip patrol between two intersections. Its main goal is to maintain stable connectivity of MUAVs.
- Patrol mode 2: Low-overhead priority based cruise (LOPC). Each MUAV patrol roads along a clockwise or counter-clockwise direction within square area. This patrol mode mainly maintains the minimum required number of MUAVs on the premise of keeping the basic connectivity to save the deployment overhead.

Since the goals of MUAVs and the CUAV are different, the change of MUAVs status will affect decisions of the CUAV, and CUAV's strategies will affect

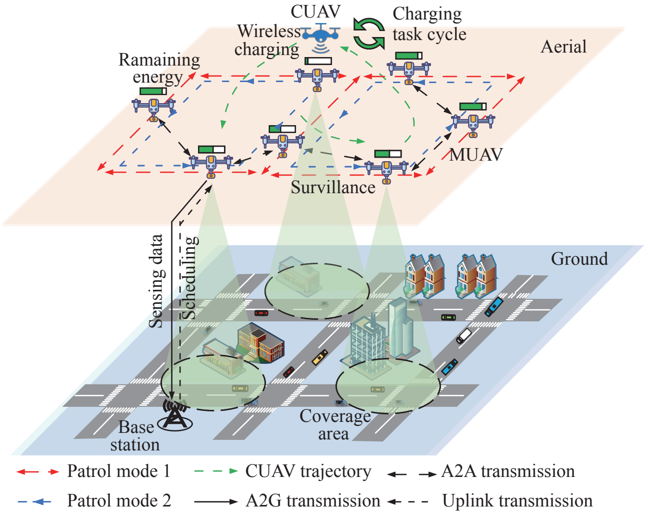

Fig. 1. Rechargeable UAVs-assisted traffic perception system.

the MUAVs task. In this section, we will describe the problem from three parts: MUAVs deployment, CUAV scheduling, and MOO of deployment and scheduling.

1. Deployment problem

This paper focuses on MUAVs' energy efficiency, which is related to energy consumption, network connectivity and patrol efficiency. The coordinate of MUAV$_m$ ($m=1,2,\dots,M$) in time slot t ($t=1,2,\dots,T$) is denoted as $U_{m,t}=[x_{m,t},\, y_{m,t},\, z_{m,t}]$, where $x_{m,t}\in[x_{\min}, x_{\max}]$, $y_{m,t}\in[y_{\min}, y_{\max}]$ and $z_{m,t}\in[z_{\min}, z_{\max}]$ are the horizontal coordinate, vertical coordinate, and the altitude of each UAV, respectively. The speed of MUAV$_m$ in time slot t is $v_{m,t}\in[v_{\min}, v_{\max}]$. In addition, for two patrol modes, the flight distance of MUAV$_m$ between slots t and $t+1$ is constrained as follows:

$$v_{\min}\tau \le \|U_{m,t+1}-U_{m,t}\|_2 \le v_{\max}\tau \tag{1}$$

Reliable data transmission needs to consider the network connectivity, data transmission rate, and the current residual energy comprehensively.

1) Energy consumption model and communication model of MUAVs

The energy consumption of MUAVs consists of two aspects: propulsion energy consumption [36] and communication energy consumption. The propulsion energy consumption of MUAV$_m$ is calculated as

$$E_m(t)=\int_0^t P(V_m(t))\,\mathrm{d}t \tag{2}$$

where $V_m(t)$ denotes the instantaneous speed of MUAV$_m$ in time slot t, and $P(V_m(t))$ is the propulsion power. Thus, the residual energy of MUAV$_m$ is

$$E_{\mathrm{re}_m}(t)=E_0-(E_m(t)+E_{\mathrm{c}_m}(t)) \tag{3}$$

where E_0 is the maximum battery capacity of MUAV$_m$, $E_{\mathrm{c}_m}(t)$ is communication energy consumption, which includes the energy consumption caused by the data reception and transmission. When an MUAV residual energy drops to the residual energy threshold, the MUAV has been selected as the priority service object by the CUAV. Whether the energy consumption is high or low during hovering, the MUAV will eventually get charging service to recover the task. Therefore, the hovering energy consumption of MUAVs is ignored.

The stable communication network should be established among MUAVs to transmit perception data. Data transmission is not only related to transmission rate, but related to its own residual energy. This is because that the MUAVs will suspend their current tasks and switch to the hovering state to wait for the charging from the CUAV when the residual energy is lower than the threshold. Thus, link interruption occurs and the real-time data transmission will fail.

Due to the high-quality line of sight links among MUAVs, the average path loss between MUAV$_m$ and MUAV$_k$ during data transmission is exspressed as [37]

$$L_t^{m,k}=20\log\left(\frac{4\pi f d_{m,k}}{c}\right)+\eta_{\mathrm{LoS}} \tag{4}$$

where $d_{m,k}$ indicates the distance between MUAV$_m$ and MUAV$_k$, c is velocity of light, and η_{LoS} is the average additional pass loss. Assuming that the transmission power ψ_{tx} and the average noise power ψ_n are fixed, the received signal-to-noise ratio is $\mathrm{snr}_t^{m,k}=\psi_{tx}-L_t^{m,k}-\psi_n$. The transmission rate from MUAV$_m$ to MUAV$_k$ can be expressed as $V_t^{m,k}=W\log(1+\mathrm{snr}_t^{m,k})$, where W is bandwidth.

When the residual energy of MUAV is lower than threshold value $E_{\mathrm{threshold}}$, the data transmission rate $V_t^{m,k}$ from MUAV$_m$ to MUAV$_k$ will be affected, i.e.,

$$V_t^{m,k}=\begin{cases} V, & E_{\mathrm{re}_m}>E_{\mathrm{threshold}}\\ \lambda V, & E_{\mathrm{re}_m}\le E_{\mathrm{threshold}}\end{cases} \tag{5}$$

where $E_{\mathrm{threshold}}$ is the residual energy threshold of MUAV, and $\lambda\in(0,\ 1)$ is the decay factor.

2) Task energy efficiency

To jointly consider energy consumption, network connectivity and patrol efficiency, the energy efficiency indicator ξ is designed as

$$\xi = \frac{1}{E_T}\left(\sum_{t=1}^{T}\sum_{m=1}^{M-1}\sum_{k=m+1}^{M}V_t^{m,k}\right)\sum_{m=1}^{M}\kappa_m \tag{6}$$

where $E_T=\sum_{t=1}^{T}\sum_{m=1}^{M}E_m(t)$ is MUAVs' total energy consumption. $\kappa_m=\frac{1}{T\tau+cT_w}\sum_{t=1}^{T}d_t^m$ indicates the patrol efficiency where T_w is the charging waiting time, c is the binary decision variable, it is 1 when the residual energy is greater than the threshold, and 0 otherwise. Due to the energy limitation, T_w will decrease the patrol efficiency. $\sum_{t=1}^{T}d_t^m$ is the movement distance of MUAV$_m$.

3) The MUAVs deployment problem

The maximization of energy efficiency is set as the optimization goal to ensure all MUAVs complete the perception task. The optimization Objective 1 is

$$f_1(X_1)=\max_{X_1}\xi \tag{7}$$

where X_1 is the continuous independent variable, i.e.,

$$\begin{aligned} X_1=&\left[\mathbb{X}^{1\times M},\mathbb{Y}^{1\times M},\mathbb{Z}^{1\times M},\mathbb{E}^{1\times M},\mathbb{V}^{1\times M}\right]\\ =&\left[x_1,x_2\ \dots,x_M,\ y_1,y_2,\dots,y_M,\ z_1,z_2,\dots,z_M,\right.\\ &\left.E_1,E_2,\dots,E_M,\ V_1,V_2,\dots,V_M\right] \end{aligned} \tag{8}$$

where $(\mathbb{X}^{1\times M},\mathbb{Y}^{1\times M},\mathbb{Z}^{1\times M})$ denotes the position of MUAVs, and $(\mathbb{E}^{1\times M},\mathbb{V}^{1\times M})$ denotes the residual energy and the transmission rate of MUAVs.

It is obvious that the residual energy of MUAVs is related to the energy efficiency. In this paper, the energy efficiency of MUAVs is a necessary indicator to measure the task. Because the reduction of the residual energy will change the status of MUAVs, i.e., from moving to hovering and waiting for charging, which will have a negative impact on the energy efficiency in terms of equation (6). Therefore, to maintain the endurance of MUAVs while ensuring higher energy efficiency, it is necessary to schedule the CUAV to provide active charging. Because the position of MUAVs is dynamic, and the energy efficiency changes with the change of residual energy of each MUAV, which is challenging for CUAV to adopt effective charging scheduling strategies for MUAVs that need charging services.

2. Charging scheduling problem

For CUAV, this paper mainly focuses on two issues: 1) Which MUAV will be selected by CUAV to charge? 2) How much energy required for charging the MUAV to meet the current system endurance and ensure higher energy efficiency. In this subsection, the mobility model and charging model of the CUAV are established, and the optimization objective is introduced.

1) CUAV mobility model

The coordinate of the CUAV in time slot t $(t=1,2,\dots,T)$ is expressed as $C_t=[x_t,\ y_t,\ z_t]$. The speed of the CUAV is $v_{ct}\in[v_{\min},v_{\max}]$. CUAV will fly towards the target MUAV at the speed v_{ct} to provide charging when its central controller determines the current or next objective, and the flight time in the process is $T_{cs}=d_{cm}/v_{ct}$, which is a part of charging waiting time T_w, where $d_{cm}=\|U_{m,t}-C_t\|_2$ indicates the distance between CUAV and the MUAV$_m$.

For the charging, the CUAV needs to select the appropriate service object by evaluating their current states. This paper focuses on the impact of CUAV charging strategies on the overall task of MUAVs, the additional energy consumption of the CUAV scheduling and energy supplement are ignored.

2) CUAV charging model

For the CUAV charging, the charging efficiency is determined by the electricity-to-laser conversion efficiency of transmitter ϕ_{el}, transmission efficiency ϕ_{lt} and laser-to-electricity conversion efficiency of the receiver ϕ_{le} [38]. Assuming that the working power of the laser charging transmitter is P_t, the laser charging power P_c is

$$P_c=P_t\phi_{el}\phi_{lt}\phi_{le} \tag{9}$$

Thus, the CUAV charges one of MUAVs with $E_{\text{harv}}=P_c\cdot T_c$, where T_c is the duration of charging.

In terms of CUAV charging scheduling, Jain's fairness index [39] is used to guide fair charging decisions. For charging decision, it is considered from two aspects, namely, the objective of charging service and the degree of energy transmission. The construction process of the CUAV fair charging index is given as follows:

$$f_r=\frac{\left(\sum_{m=1}^{M}E_{\text{re}_m}^{t}\right)^2}{M\times\sum_{m=1}^{M}{E_{\text{re}_m}^{t}}^2} \tag{10}$$

$$f_c=\frac{\left(\sum_{m=1}^{M}C_m\right)^2}{M\times\sum_{m=1}^{M}{C_m}^2} \tag{11}$$

$$f_t=w_1f_r+w_2f_c \tag{12}$$

where f_r represents whether the CUAV charges MUAVs with low residual energy, C_m is the score of energy obtained by MUAV$_m$, $C_m=E_{\text{harv}}/E_m$, E_m denotes consumed energy of MUAV, f_c is used to measure whether MUAVs are charged in a relatively fair way, f_t indicates whether CUAV serves all MUAVs fairly, and $w_1\in[0,1]$ and $w_2\in[0,1]$ are two adjustable weights, $w_1+w_2=1$, which can be adjusted by the residual energy of MUAV$_m$ and charging capacity of the CUAV. The charging capacity of the CUAV can be expressed as $\rho=E_{\text{harv}}\times f_t$ by fairness optimization.

3) The CUAV charging scheduling problem

Due to the energy limitation, MUAVs will suspend the current task to hover and wait for charging service when the residual energy is insufficient, which will lead to the following problems: 1) It may cause the interruption of MUAVs network connectivity, which will affect the reliable data transmission to a certain extent; 2) It will affect the cooperation among MUAVs while degrading the overall task energy efficiency.

To address the above two challenges, the CUAV is scheduled to provide charging services for MUAVs, and the optimization Objective 2 considering charging fairness is given as follows:

$$f_2(X_2)=\max_{X_2}\rho \tag{13}$$

where X_2 represents the continuous independent variable of the objective function, i.e.,

$$\begin{aligned}X_2&=[\mathbb{E}_{\mathbb{C}}^{1\times M},\mathbb{A}^{1\times M}]\\&=[E_{C_1},E_{C_2}\dots,E_{C_M},A_1,A_2\dots,A_M]\end{aligned} \tag{14}$$

and $(\mathbb{E}_{\mathbb{C}}^{1\times M},\mathbb{A}^{1\times M})$ denotes the CUAV charging capacity and the variable of charging scheduling.

3. Multi-objective optimization problem

In this paper, the deployment of the MUAVs and charging of CUAV need to meet the above two objectives simultaneously, i.e., maximizing the energy efficiency of task and maximizing the charging amount. The MUAVs and CUAV have a correlation that charging decisions of the latter depend on the residual energy and mobility strategies of the former. The action state and network topology of MUAVs are in dynamic change, which brings challenges for the CUAV to plan rational charging decisions.

Therefore, to obtain trade-off solutions between objectives, we balance the charging and energy efficiency by solving a MOO problem. Fig.2 shows the decision-making of the MUAVs and CUAV and the relationship between the two objectives.

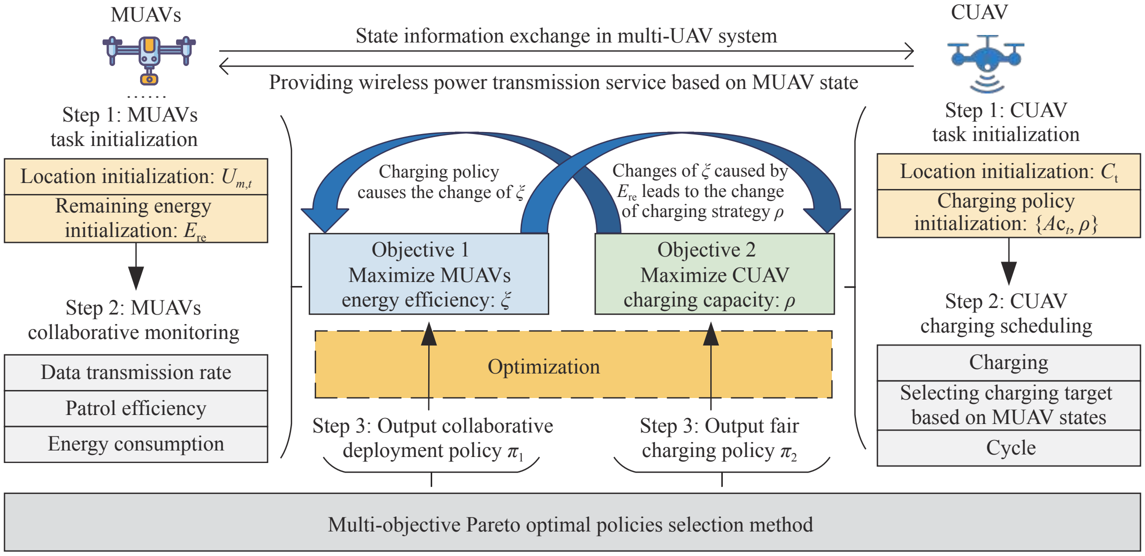
MUAVs
CUAV
State information exchange in multi-UAV system
Providing wireless power transmission service based on MUAV state
......
Step 1: MUAVs task initialization
Location initialization: $U_{m,t}$
Remaining energy initialization: E_{re}
Step 2: MUAVs collaborative monitoring
Data transmission rate
Patrol efficiency
Energy consumption
Charging policy causes the change of ξ
Changes of ξ caused by E_{re} leads to the change of charging strategy ρ
Objective 1 Maximize MUAVs energy efficiency: ξ
Objective 2 Maximize CUAV charging capacity: ρ
Optimization
Step 3: Output collaborative deployment policy π_1
Step 3: Output fair charging policy π_2
Step 1: CUAV task initialization
Location initialization: C_t
Charging policy initialization: $\{Ac_t, \rho\}$
Step 2: CUAV charging scheduling
Charging
Selecting charging target based on MUAV states
Cycle
Multi-objective Pareto optimal policies selection method

Fig. 2. Decision Framework of MUAVs and CUAV.

By determining the position and residual energy of each MUAV while considering the correlation among the MUAVs and between the CUAV and MUAVs, we formulate the MOO problem of UAVs deployment and charging scheduling. Combined with Section III.1 and Section III.2, we take the energy efficiency of the MUAVs and the charging amount of the CUAV as optimization objectives, and explore balanced solutions. The above MOO problem is expressed as follows:

$$\mathcal{P}\colon \max_{X_1 \cup X_2} F = \max_{X_1 \cup X_2} \{f_1,\ f_2\} \tag{15}$$

$$\begin{aligned}
\text{s.t.}\quad & \text{C1: } X_{\min} \leq x_{m,t} \leq X_{\max}, \forall m \in M, \forall t \in T,\\
& \text{C2: } Y_{\min} \leq y_{m,t} \leq Y_{\max},\\
& \text{C3: } Z_{\min} \leq z_{m,t} \leq Z_{\max},\\
& \text{C4: } E_{\text{remaining}} \geq E_{\text{threshold}},\\
& \text{C5: } E_{\min} \leq \rho_{m,t} \leq E_{\max},\\
& \text{C6: } Ac_t \in \{1, 2, \ldots, M\}, \forall t \in T.
\end{aligned}$$

where $X_{\min}$ and $X_{\max}$ are the minimum and maximum range covered by the MUAV in the field of view. $Y_{\min}$ and $Y_{\max}$ represent the minimum and maximum coverage of MUAV longitudinal vision. $Z_{\min}$ and $Z_{\max}$ are the lowest and highest vertical height. $E_{\min}$ and $E_{\max}$ are the minimum and maximum power that CUAV replenishes for MUAVs. Ac_t indicates the scheduling relationship between the CUAV and MUAVs, i.e., $Ac_t = m$ if the CUAV chooses to charge the m-th MUAV at tth time slot. C1–C3 are movement constrains of MUAVs, respectively. C4 constrains the residual energy of MUAVs. C5 constrains the charging capacity of CUAVs within an appropriate range. An appropriate charging strategy will increase the mission endurance of MUAVs. C6 indicates the charging scheduling of the CUAV, i.e., which MUAV is selected for charging service.

For the above two objectives, if one objective is excessively optimized, it is difficult to meet the requirements of maximizing the energy efficiency. Once the charging waiting or data transmission are impaired, the charging scheduling fails. In summary, the MOO problem of UAVs deployment and charging in this paper is NP hard [40]. Hence, we employ the multi-objective MADRL method to solve the problem. In addition, we transform it into a multi-objective decision making problem [41] while combining Pareto optimal to meet the policy requirements of different optimization goals.

IV. Multi-Objective Optimization with Pareto-Optimal MADRL

We design a multi-UAV deployment and charging scheduling algorithm based on multi-objective MADRL. First, we describe the MOO problem $\mathcal{P}$ as a multi-objective multi-agent (MOMA) decision-making problem, and then propose the UPDC algorithm to solve it.

1. Multi-objective Markov decision process

To denote the MOMA decision-making problem, we introduce multi-objective stochastic game (MOSG) [42], i.e., multi-objective Markov decision process (MOMDP). Therefore, the multi-objective problem is described as a multi-objective stochastic game tuple (S, A, T, γ, R), which is used to model the decision-making process. The tuple consists of five parts, i.e., state space S, joint action set A, state transition probability $P \in [0,\ 1]$, reward discount factor $\gamma \in (0,\ 1)$, and reward function $R = R_1 \times R_2 \times \cdots \times R_i \times \cdots \times R_{M+1}$, where R_i denotes the vector reward function of agent i $(i = 1, 2, \ldots, M+1)$ for each objective. The state space, action space and rewards are defined as follows.

State Space S State $s_t \in S$ is the set of agent observations in time slot t, and $s_t = (U_{m,t},\ E_r,\ L_p)$.

- $U_{m,t} = [x_{m,t},\ y_{m,t},\ z_{m,t}]$ denotes the position of MUAVs m $(m = 1,\ 2,\ ...,\ M)$ in time slot t.
- $E_r = \{E_{\text{re}_m} | m = 1, 2, \ldots, M\}$ denotes the residual energy set of MUAVs in time slot t.
- $L_p = [x_p,\ y_p]$ denotes location of target points.

Action Space A The action space consists of all possible actions taken by agents (i.e., the MUAVs and CUAV) during the task. $a_t \in A$, and both agents take actions in the same action space. $a_t = (\alpha_t,\ d_t,\ \rho_{m,t})$ consists of the following three parts:

- $\alpha_t \in [0,\ 2\pi]$ denotes flight direction of MUAVs and CUAVs in time slot t.
- $d_t \in [0, d_{\max}]$ denotes flight distance of MUAVs and CUAVs in time slot t.
- $\rho_{m,t} \in [0,\ E_{\max}]$ denotes the amount of charge that CUAV charges MUAV m in time slot t.

Reward R The task environment is partially observable and non-stationary for the MUAVs and CUAVs. They can evaluate current actions and understand the environment according to the reward received. Agents can learn the efficient control strategy of MOO problems through the reward function. In this paper, the reward is designed as a two-dimensional vector:

$$R = \{r_t\} = \{[r_{te}(t),\ r_c(t)]\} \tag{16}$$

where $r_{te}(t)$ and $r_c(t)$ correspond to two optimization objectives: maximization of energy efficiency and charging capacity. $r_{te}(t)$ and $r_c(t)$ are given as

$$r_{te}(t) = \begin{cases} 10^4 \times \xi, & \forall E_{\text{re}_m} > E_{\text{threshold}} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

$$r_c(t) = \begin{cases} \rho + r_c\ , & \text{if } T_w = 0 \\ 0\ , & \text{otherwise} \end{cases} \tag{18}$$

where the former is the joint reward maintained by the MUAVs, and the latter is the the CUAV independent reward. r_c is a positive reward compensation value to encourage the CUAV to actively conduct charging.

Reasonable reward setting is useful to guide agents to explore strategies. For optimization goal, maximizing the energy efficiency of the MUAVs is based on reasonable charging. Excessive charging of a MUAV by the CUAV may make other MUAVs switch to the state of charging waiting, which will cause the degradation of the energy efficiency. Thus, appropriate punishments are considered to ensure fair charging. The punishment to the CUAV is denoted as P_m and is given as

$$P_m = \begin{cases} p_1, & \exists \text{ MUAV } m, E_{\text{re}_m} < E_{\text{threshold}} \\ p_2, & \text{if } \xi < \xi_{\min} \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

where P_1 is 10ρ, and P_2 is $\rho + r_c$, which means the CUAV will be punished when the residual energy of the MUAVs is lower than $E_{\text{threshold}}$, and the total energy efficiency of the MUAVs is lower than the minimum energy efficiency required. P_m is 0 when the remaining energy is sufficient to execute current task.

2. UPDC algorithm

As a typical policy based the MADRL algorithm, HATRPO [43] is suitable for agents to learn strategies in the continuous action space under complex UAV deployment and charging task. The reason is that the theorem that monotonic improvement of policies is proved and the trust region update is introduced to keep efficient policies update step. Therefore, the HATRPO algorithm is used as the basic algorithm in our work. We combine the solution of multi-objective pareto optimal solution set with the HATRPO algorithm to obtain the tradeoff of two optimization problems solutions by taking the vector reward as bridge.

Fig.3 shows the architecture of the UPDC algorithm. It adopts the framework of centralized training and distributed execution [44]. In the training phase, the utility of MUAVs is equated as a part of the advantage function to reduce the complexity of solutions, and the trust region update is applied in the utility-based Pareto optimal advantage function gradient update. The TMC protocol is also introduced based on the multi-objective agent network. Next, we will describe the algorithm from three aspects: 1) Equivalence of Pareto optimal based on utility; 2) Trust region update of Pareto ptimal policies; 3) Efficient policies exploration for policy update.

1) Equivalence of Pareto optimal based on utility

For MUAVs, each agent reacts depend on the policy π, and the optimization of policy π means maximizing the expected discounted long-term reward

$$V^{\pi_i} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, a_t, s_{t+1}) | \pi\right] \tag{20}$$

where V^{π_i} indicates the utility about single policy π_i of

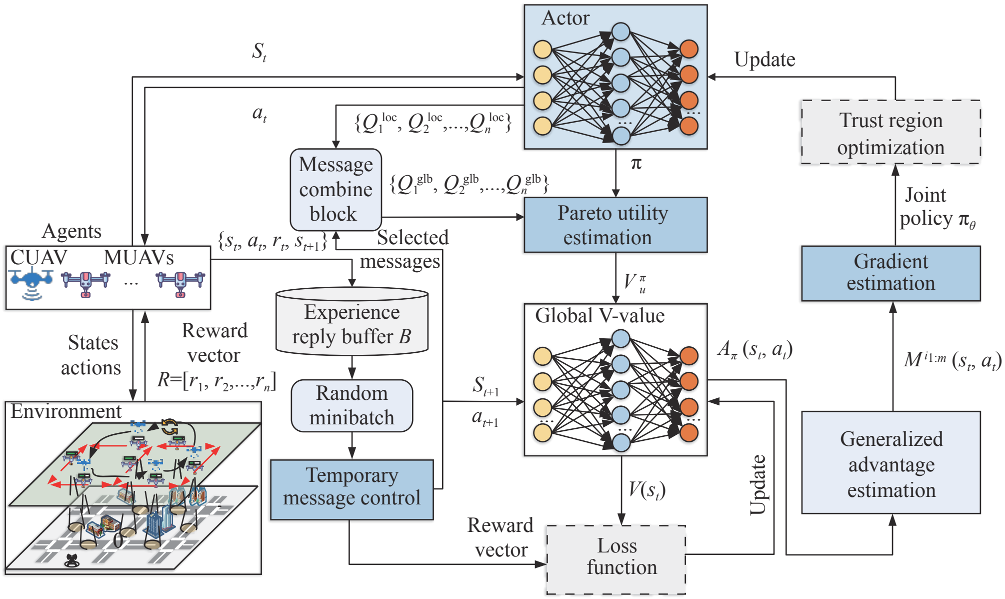
Actor
Update
S_t
a_t
$\{Q_1^{\mathrm{loc}}, Q_2^{\mathrm{loc}},\dots,Q_n^{\mathrm{loc}}\}$
Trust region optimization
Message combine block
$\{Q_1^{\mathrm{glb}}, Q_2^{\mathrm{glb}},\dots,Q_n^{\mathrm{glb}}\}$
π
Pareto utility estimation
Joint policy π_θ
Agents
CUAV
MUAVs
…
$\{s_t, a_t, r_t, s_{t+1}\}$
Selected messages
Gradient estimation
V_u^π
Experience reply buffer B
Global V-value
$A_\pi(s_t, a_t)$
$M^{i1:m}(s_t, a_t)$
States actions
Reward vector
$R=[r_1, r_2,\dots,r_n]$
Random minibatch
S_{t+1}
a_{t+1}
Environment
Generalized advantage estimation
Temporary message control
$V(s_t)$
Update
Reward vector
Loss function

Fig. 3. The architecture of UPDC Algorithm.

each agent and $\pi=(\pi_1,\pi_2,\dots,\pi_n)$ represents joint strategy. $R_i(s_t,a_t,s_{t+1})$ is the rewards obtained by agent i for an action $a_t\in A$ under state $s_t\in S$.

However, the computing of the utility function of each agent will consume huge computing resources. Therefore, for MUAVs, this paper considers the team-based utility function, namely obtaining joint utility function V_u^π by the agents joint policies

$$V_u^\pi = u\left(\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R_i(s_t,a_t,s_{t+1})|\pi\right]\right) \tag{21}$$

The CUAV makes decisions with the state of other MUAVs as input. We regard CUAV as an altruistic agent and take the utility of other agents as the goal, and combine the joint advantage function $A_\pi(s,a)\triangleq Q_\pi(s,a)-V_u^\pi$ in HATRPO to update gradient. For different sets 1 and 2 of agents, the multi-agent advantage function [43] is given as follows:

$$A_\pi^1(s,a^1,a^2)\triangleq Q_\pi^{1,2}(s,a^1,a^2)-Q_\pi^2(s,a^2) \tag{22}$$

Since the joint utility function of MUAVs is the result of cooperation, we equate the team-based utility function with the multi-agent state action value function to improve the computing ability of algorithm.

2) Trust region update of Pareto optimal policies

After obtaining the equivalent advantage function, we compare the multi-objective Pareto non-dominated set achieved by the vector reward. When the target value is no longer improved, the Pareto optimal solution set is obtained. Since the monotone improvement property of the joint policy update in the trust region policy update [45], we combine the iterative process of obtaining the Pareto optimal policies with the update of trust region strategy to ensure near optimal solutions.

In the process of updating, the KL divergence constraint is used to limit the update step of approximate solution. The agents target function is given as follows:

$$L_\pi^1(\bar{\pi},\hat{\pi})\triangleq\mathbb{E}_{s,a^1,a^2}\left[A_\pi^1(s,a^1,a^2)\right] \tag{23}$$

where $\bar{\pi}$ and $\hat{\pi}$ denote other joint policies of different agents sets 1 and 2, respectively. For joint policies $\bar{\pi}$,

$$\begin{aligned}J(\bar{\pi})\geq J(\pi)+\sum_{m=1}^{n}\Big[&L_\pi^{i_{1:m}}\left(\bar{\pi}^{i_{1:m-1}},\bar{\pi}^{i_m}\right)\\&-CD_{\mathrm{KL}}^{\max}\left(\pi^{i_m},\bar{\pi}^{i_m}\right)\Big]\end{aligned} \tag{24}$$

where $J(\pi)\triangleq\mathbb{E}_{s,a}\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$ is the expected total reward. C is the coefficient, and KL divergence constraint is $\mathbb{E}\left[D_{\mathrm{KL}}(\pi^{i_m},\bar{\pi}^{i_m})\right]\leq\delta$, δ is threshold hyperparameter, which limits the step size of policy update and ensures it occurs within trust range.

For the selection of policies, two principles should be followed: 1) It must converge to the real Pareto front as much as possible; 2) The solution should be as dense as possible. Thus, hypervolume metric [46] and sparsity metric [47] are introduced to evaluate the quality of the solution to find the approximate Pareto front.

3) Efficient policies exploration for policy update

The update and selection of policies for maximum energy efficiency and charging capacity depend on exploration of policies. We utilize TMC protocol [48] to design multi-objective agent network that consists of three networks to control agents messages transmission, and improve the exploration efficiency of policies by reducing redundant data exchange.

The multi-objective agent network structure is shown in Fig.4. It consists of four parts: 1) Joint action generator, 2) Message encoder, 3) Message buffer, and 4) Message combination block. The network uses the joint action generator to obtain the vector local Q-value $\left\{Q_1^{\text{loc}}, Q_2^{\text{loc}}, \ldots, Q_n^{\text{loc}}\right\}$, and then combine it with the received message to obtain the vector global Q-value $\left\{Q_1^{\text{glb}}, Q_2^{\text{glb}}, \ldots, Q_n^{\text{glb}}\right\}$. Received message buffer receives messages while updating with a new message, and then selects the message with the valid bit $\text{val}(n) = 1$.

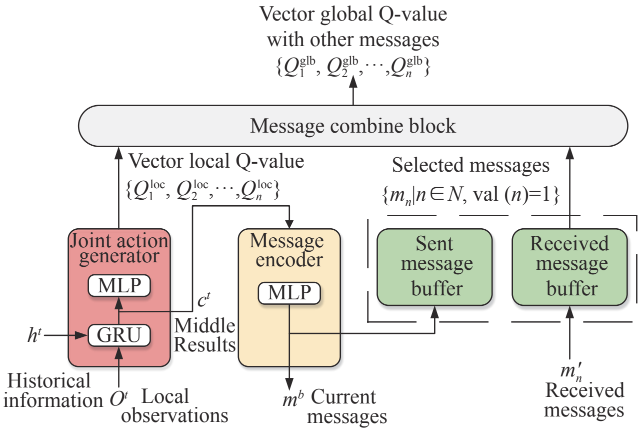

Fig. 4. Multi-objective agent network structure.

The communication protocol works between the sent message buffer and the received message buffer. The communication protocol is given as follows.

On the sender side,

$$m_s = \begin{cases} f_{\text{msg}}(c^t), & \text{if } \left\| f_{msg}(c^t) - m^b \right\| \geq \delta \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

where m_s denotes sent message, $f_{\text{msg}}(c^t)$ and m^b denote message generated by message encoder and saved message in the sent message buffer, respectively. t and t^{last} denote the current timestep and the last timestep at which agent broadcasts messages to the other agents, respectively. δ and ω_s are Euclidean distance threshold and the smoothing window size.

On the receiver side,

$$val\ (n) = \begin{cases} 0, & \text{if } t^{\text{updated}} > \omega_s \\ 1, & \text{otherwise} \end{cases} \tag{26}$$

where t^{updated} is time period, and valid bit is 0 when a message is expired.

The UPDC algorithm (Algorithm 1) is designed to address the MOO problem of MUAVs deployment and CUAV charging, and its principle is given as follows.

Parameter initialization (line 1) Initializing the experience replay buffer Bc, the actor network and Global V-value network, and the corresponding two message buffer during the training phrase.

Experience sampling (lines 3–6) The agents execute actions at each episode, and transfer the collected state transitions into reply buffer as policies learning basis, then filter redundant information by TMC protocol.

Equivalent advantage function generation (lines 7–10) Agents estimate joint utility of MUAVs by current joint policy π, and equate it with the state action value function for estimating advantage function $A_\pi(s_t, a_t)$ during the training phrase.

Policy update (lines 12–19) After determining the joint advantage function, the optimal strategies are obtained by calculating the gradient, then determining the update direction, and planning the update range.

$$g_k^{i_m} = \frac{1}{B} \sum_{b=1}^{B} \sum_{t=1}^{T} \nabla_{\theta_k^{i_m}} \log \pi_{\theta_k^{i_m}}^{i_m}(a_t^{i_m} | s_t^{i_m}) M^{i_{1:m}}(s_t, a_t) \tag{27}$$

where $M^{i_{1:m}}(s_t, a_t)$ is equivalent advantage function, and then determining the update direction by

$$d_k^{i_m} \approx g_k^{i_m} / H_k^{i_m} \tag{28}$$

where $H_k^{i_m}$ is the Hessian of the expected KL divergence, and then planning the update range of policy

$$\mathbb{E}\left[D_{\text{KL}}(\pi^{i_m}, \bar{\pi}^{i_m})\right] \leq \delta \tag{29}$$

lastly, the Global network is updated by loss function

$$\phi_{k+1} = \arg\min \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=0}^{T} \left(V(s_t) - R_t\right)^2 \tag{30}$$

where $V(s_t)$ represents the state value function of V-value network.

Algorithm 1 UPDC Algorithm

```
Input: Number of agents $n$, episodes $K$, coordinate of
  MUAV $U_{m,t}$, and residual energy $E_r$.
Output: The charging policies of CUAV.
//Parameter initialization
1: Initialize: Actor network, Global V-value network, experience
   reply buffer B, two message buffer and message encoder;
2: for $k=0,1,\ldots,K-1$ do
//Experience sampling
3:     Collect a set of experiences by performing MUAVs
       and CUAV joint policy $\pi$;
4:     Transfer transitions into experience reply buffer B;
5:     Sample a minibatch of $B$ transitions from B randomly;
6:     Filtering similar message by TMC;
//Equivalent advantage function generation
7:     Estimate utility function $V_u^\pi$ based on joint policy $\pi$;
8:     Equate $V_u^\pi$ with the state action value function;
9:     Estimate advantage function $A_\pi(s_t, a_t)$ based on glob-
```

```
al V-value network with GAE according to (22);
Set equivalent $M^{i_1}(s_t, a_t) = A_\pi(s_t, a_t)$;
//Policy update
for agent $i_m = i_1, i_2, ..., i_M$ do
    Estimate the gradient of the agent maximization objective by (27);
    Compute the update direction of gradient by (28);
    Update jointly the Pareto-optimal approximate solution by the KL-constraint by (29);
    Select Pareto-optimal solution by Hypervolume and Sparsity metric;
    Update the policy of agent $i_m$
end for
Update Global V-value network by loss function (30)
end for
```

3. Complexity analysis

This paper uses the complexity to measure the performance of the UPDC algorithm. Suppose the Actor network contains I fully connected layers and the Global V-value network contains J fully connected layers, the time complexity $O(N)$ can be calculated as follows:

$$O(N) = O\left(\sum_{i=1}^{I-1} u_{a,i} u_{a,i+1} + \sum_{j=1}^{J-1} u_{g,j} u_{g,j+1}\right) \tag{31}$$

where $u_{a,i}$ and $u_{g,j}$ are the neuron number in the i-th and j-th layer of the actor and global V-value network.

There is an $F \times H$ matrix for the fully connected layer. Therefore, the fully connected neural networks need the number of storage unit $(F+2) \times H$, and the space complexity is $O(N)$. It is essential for the sent message and received message buffer to distribute storage unit so as to store more experience, and space complexity is $O(S_m)$ and $O(R_m)$. Hence, the space complexity of the UPDC algorithm is $O(N) + O(S_m) + O(R_m)$.

V. Performance Evaluation

In this section, simulation results are performed to state the effectiveness of the UPDC algorithm. As shown in Fig.5, task area consists of 9 intersections is taken as simulation scenes. MUAVs perform tasks within road sections. Meanwhile, CUAV keep in standby while observing the MUAVs states. We set the task range to a square area of size 800 m × 800 m. The simulation parameter settings are listed in Table 3.

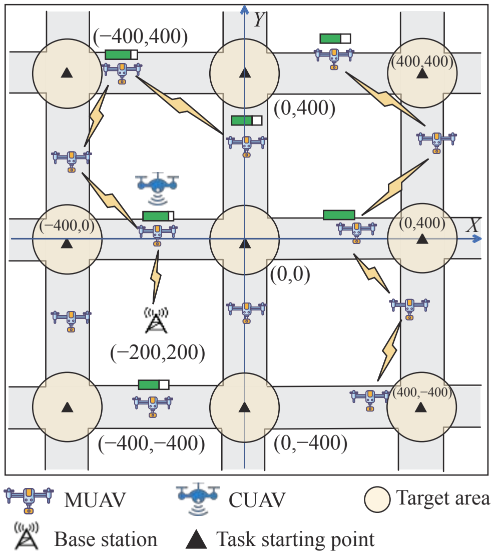

Fig. 5. Simulation scenario.

Table 3. Simulation parameter settings

Parameters	Value
Reward discount factor	0.99
Number of MUAVs (M)	3–12
Number of CUAV (C)	1
Number of target points (N)	4–9
Coverage radius of UAVs (r)	8 m
Width of road segment (g)	14 m
Length of road segment (d)	400 m
Horizontal coordinate range of MUAVs	$[-400, 400]$
Vertical coordinate range of MUAVs	$[-400, 400]$
Minimum speed of UAVs ($E_{\min}$)	2 m/s
Maximum speed of UAVs ($E_{\max}$)	10 m/s
Initial energy of UAVs (E_0)	2000 kJ
Flying energy consumption (E_m)	0.6 kJ/m
Communication energy consumption (Ec_m)	0.001 kJ/s
Charging rate of the CUAV	10 kJ/s

Due to the limited battery capacity of MUAVs, they will exhaust energy soon. Considering MUAVs patrolling sections 5 times and CUAV charges all MUAVs as a cycle. Experiments are conducted under above condition to verify the algorithm performance.

For deployment and charging, taking the number of crossroads N as variable. We explore the strategies of MUAV and CUAV under different tasks.

For patrol mode 1 HCPC, one MUAV is deployed in one road section, and $M \geq N$. Each MUAV operates on road section in the form of round-trip patrol between two intersections. Three task cases with different number of MUAVs M and the number of intersections N are considered to evaluate the performance of algorithm about addressing complex MOMA continuous decisions-making problems. Three task cases are given as follows:

- Case 1: $N = 4$, $M = 4$;
- Case 2: $N = 6$, $M = 7$;
- Case 3: $N = 9$, $M = 12$.

For patrol mode 2 LOPC, each MUAV patrols sections along a clockwise or counter-clockwise direction. Since this mode focuses on low MUAVs number deployment overhead, and thus a small number of MUAVs that can maintain basic connectivity are deployed in a square area. In the mode, $M < N$, and three different situations are considered to evaluate the continuous decision-making performance of the proposed algorithm. Three task cases are given as follows:

- Case 1: $N = 4$, $M = 3$;
- Case 2: $N = 6$, $M = 5$;
- Case 3: $N = 9$, $M = 8$.

1. Performance and analysis

Fig.6 shows that the vector rewards convergence curves of the UPDC algorithm in different cases.

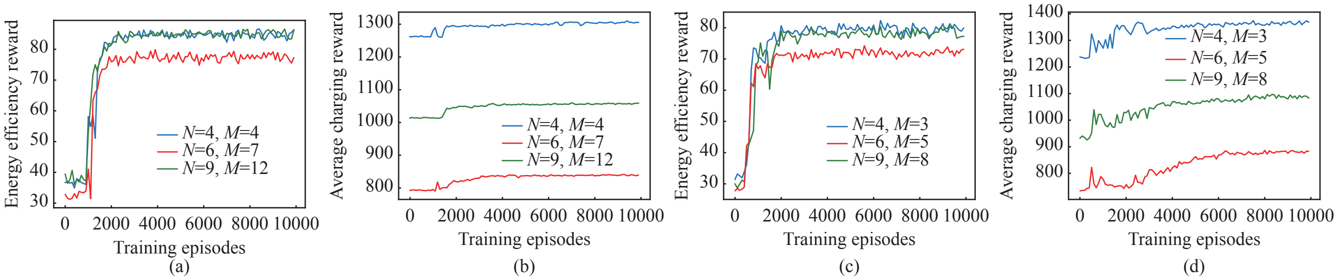

Fig. 6. Training episodes and vector reward of the UPDC algorithm in three cases with two patrol modes. For patrol mode 1 HCPC: (a) MUAVs energy efficiency reward; (b) CUAV average charging reward. For patrol mode 2 LOPC: (c) MUAVs energy efficiency reward; (d) CUAV average charging reward.

As shown in Fig.6, for patrol mode 1 HCPC, MUAVs energy efficiency and CUAV charging capacity have changed when the task scale extends from 2 × 2 and 2 × 3 intersections to 3 × 3 intersections by combining with Fig.6(a) and Fig.6(b). The converged energy efficiency reward has reduced from 87 to 76, and the CUAV charging capacity has also reduced from 1300 to 850. The reason is that the number of MUAVs increases with the expansion of the task. To ensure all MUAVs can be served, the policies with energy efficiency rewards as utility guarantee deployment is not affected at the cost of reducing the charging capacity.

For patrol mode 2 LOPC, to save deployment resources, the number of MUAVs has reduced. For the three cases, it is obvious that the charging capacity received by MUAVs increases, because the number of MUAVs decreases, which means that the CUAV is able to take care of all MUAVs and maintain their endurance. Through the fair charging scheduling strategy, each MUAV can obtain more charging times. However, its energy efficiency reward has declined, i.e., after convergence, case 1 and case 2 have reached 80 at most, and case 3 has only reached 70, which has declined compared with patrol mode 1. The reason is that the decrease of MUAVs number causes that the connectivity cannot always be maintained, which leads to the decline in transmission rate but saves deployment resources and improves the endurance of MUAVs.

It is worth noting that the MUAVs energy efficiency can also reach the threshold energy efficiency at the initial stage of task execution, which owes to the joint advantage function of the global policies updates in trust region. In addition, due to the TMC protocol, the vector reward can quickly converge from 1800 to 2000 episodes, which shows that the improved agent network successfully filters out redundant messages.

2. Performance comparisons with three related schemes

In this section, the UPDC algorithm is compared with POAC, PDQN, and OU-MADDPG.

- PDQN: Reference [34] proposed the Pareto-DQN algorithm, which can estimate Pareto front with a complex high-dimensional multi-objective state space.
- POAC: Reference [35] proposed the Pareto-optimal actor-critic approach, which is independent with objectives preference, and not affected by the concavity and convexity of the Pareto front.
- OU-MADDPG: Reference [30] proposed the UAV charging scheduling and trajectory planning algorithm based on MADDPG for MUAVs tasks.

As shown in Fig.7, for MUAV patrol mode 1, in Fig.7(a) and Fig.7(b), the Pareto Front trend is close by the selection of hypervolume and sparsity indicators. The reason is that the number of MUAVs need charging is within the acceptable range of CUAV. When $M = 4$, the UPDC algorithm can supplement nearly 65% power for MUAVs while keeping energy efficiency without loss. Similarly, when $M = 6$, with the increase of MUAVs number, to ensure MUAVs can get timely service and are not disturbed by charging waiting, CUAV reduces the maximum amount of charging.

From Fig.7(c), the expansion of the task means that the burden of MUAVs and CUAV increases when $M = 12$. For MUAVs, energy efficiency is more vulnerable to the negative impact of charging waiting and communication interruption. For CUAV, it is not only need to focus on MUAVs status in advance, but need to adjust the charging strategy in time. It can be seen that the energy efficiency of MUAVs decreases when the charging amount exceeds 1000, and when the charging

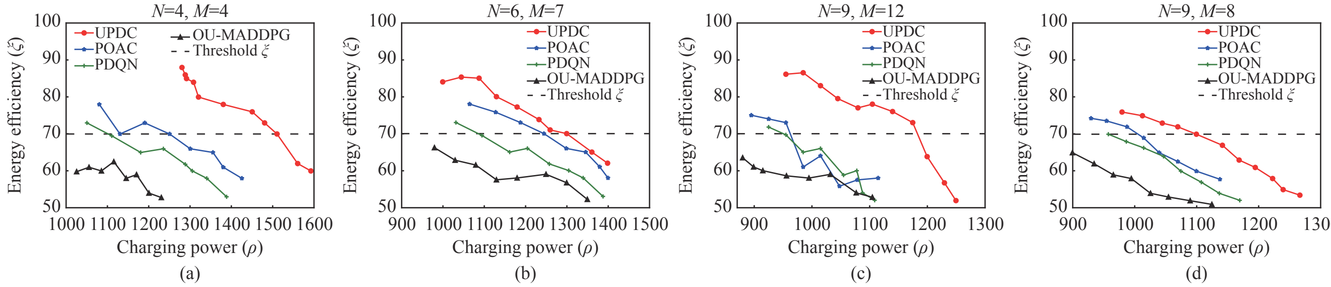

Fig. 7. Multi-objective policies performance of the UPDC, POAC, PDQN, and OU-MADDPG algorithms in different cases with two patrol modes. HCPC: (a) N=4, M=4. (b) N=6, M=7. (c) N=9, M=12, LOPC: (d) N=9, M=8.

amount reaches 1250, the energy efficiency drops to the lowest, namely, some MUAVs do not receive services. However, its sparsity increases, but the remaining optimal strategy points still dominate other methods.

For patrol mode 2, from Fig.7(d), 8 MUAVs are deployed in the area with 9 intersections and 12 sections. Since the number of MUAVs is lower than that of case 3 in patrol mode 1, the energy efficiency is affected by connectivity. Compared with case 3 in patrol mode 1, the average charging amount accepted by MUAVs has increased by about 3%–5%. However, each MUAV with low residual energy can obtain more services, although the reduction of connectivity leads to the decline of ξ, the number of MUAV M decrease while the mission sustainability is easier to achieve.

For the fluctuation of ξ in Fig.7, the reason is that the change of ξ depends on the change of energy state caused by charging strategy. In the process of centralized training, the charging strategy adopted by the CUAV may only consider partial MUAVs and ignores others, which will lead to MUAVs enter the charging waiting state. Once residual energy of a certain MUAV is insufficient, and ξ will decrease. When the charging strategy changes, i.e., the average charging amount ρ increases, the number of MUAVs that do not receive charging services decreases, andξ will increase again.

Fig.8 shows that the charging utilization of the four methods under different cases. Taking two cycles as the MUAVs task goal and the CUAV can charge eight times at most, we verified the effectiveness of the charging strategy by comparing the energy efficiency of MUAVs before and after charging.

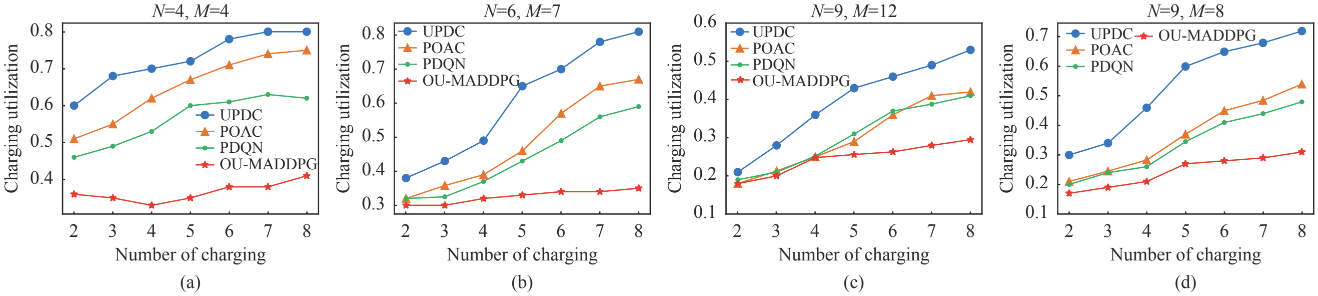

Fig. 8. The charging utilization of the UPDC, POAC, PDQN, and OU-MADDPG algorithms under different cases with two patrol modes. HCPC: (a) N=4, M=4, (b) N=6, M=7, (c) N=9, M=12, LOPC: (d) N=9, M=8.

Figs.8(a), (b), and (c) represent three cases of the patrol mode 1 and Fig.8(d) is the same as Fig.7(d). We compare the method performance based on case 3 in patrol mode 2.

For MUAVs patrol mode 1 HCPC:

1) In Fig.8(a), due to the small number of UAVs, the energy efficiency of the three methods tends to increase with the increase of charging times. The UPDC algorithm can obtain the highest charging utilization, and the CUAV's fair charging scheduling can competent for the cycle task of 4 MUAVs and execute charging service for MUAVs in advance.

2) From Fig.8(b), as the number of MUAVs increases, the space for updating of strategies also increases. However, our method can still guarantee reasonable solution. In the first round of cycle, MUAVs patrol between 4 round-trip and 5 round-trip, MUAVs with lower residual energy will appear successively. It is obvious that it cannot meet the task needs of case 2 when charging 2 to 4 times, and need to charge 5 to 7 times to obtain high charging utilization while ensuring there are no mission interruption.

3) As shown in Fig.8(c), for case 3, other methods have fallen into local optimal solution after charging 7 to 8 times while our method can still maintain high charging utilization. The reason is the UPDC al-

gorithm can schedule CUAV in advance while keeping MUAVs reasonable deployment.

For MUAVs patrol mode 2 LOPC:

4) Fig.8(d) shows the charging utilization comparison in case 3. For the UPDC algorithm, as the number of MUAVs decreases, the charging utilization increases with the increase of charging times. Moreover, the turning point of significant increase appears at the 4 to 5 charging. The reason is that agents focus on the penalties, which associates with the coupling constraints and will make agents are easier to find approximately balanced policies, i.e., the more charging times, the more durable tasks the MUAVs can perform.

For the mission completion time and endurance anxiety, since MUAVs in patrol mode 2 can obtain more energy without maintaining full connectivity and can quickly complete tasks without endurance anxiety. Therefore, we consider comparing performance of methods about two indicators in complex patrol mode 1. In Fig.9, our method can converge to a lower completion time. For case 2, because of fast optimal policy exploration and effective policy update, the charging schedule can ensure MUAVs are charged before residual energy reaches threshold. The lower time steps mean the better joint policies, namely there are fewer transmission interruptions and less charging waiting time. For Fig.10, after the training phrase, we compare the relationship between charging times and the number of charging waiting times under the three cases. Obviously, our method has the least endurance anxiety, which shows that it can not only schedule CUAV well, but enable MUAVs to perform tasks in the form of energy-saving.

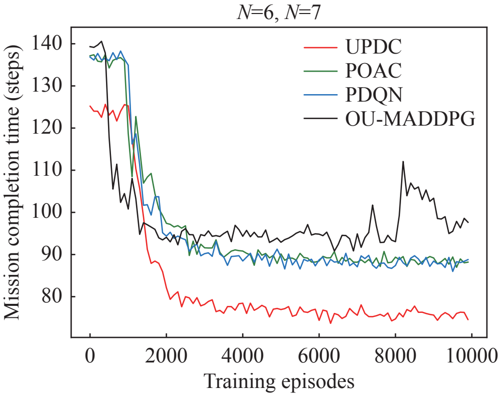

Fig. 9. The mission completion time of the UPDC, POAC, PDQN, and OU-MADDPG algorithms under case 2 with patrol mode 1.

VI. Conclusions and Future Work

In this paper, we studied the charging scheduling problem of CUAV considering the deployment of MUAVs. First, the MUAVs deployment problem is modelled considering the energy consumption and data transmission, then the CUAV charging model is established based on fair charging. The deployment and charging problem is formulated as a MOSG process to balance energy efficiency and charging. In addition, the UPDC algorithm is proposed to solve the MOO problem, which can ensure the credible update and rational selection of the Pareto Optimal policies. Finaly, our method has yield faster exploration of policies and higher MUAVs energy efficiency and CUAV charging capacity than other benchmark methods. However, the paper only solves the MOO problem with one CUAV serves MUAVs. As to more practical issue, such as urban level UAV-assisted perception, which also needs to consider multi-task allocation. Our method also cannot solve the multi-task problem in MOO problem. At present, we are also conducting research on multi-task problem for scheduling multiple CUAVs in complex scenarios and plan to solve it by combining distributed deployment methods in our future work.

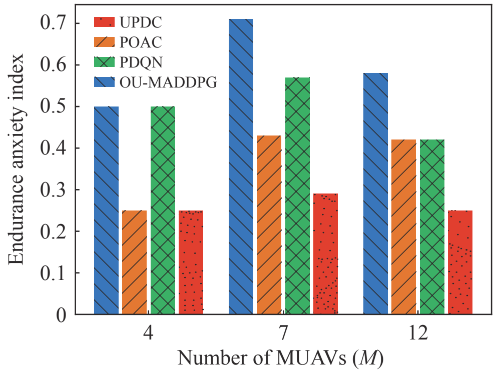

Fig. 10. The endurance anxiety index of the UPDC, POAC, PDQN, and OU-MADDPG algorithms under case 2 with patrol mode 1.

References

[1] G. X. Liu, H. Shi, K. Abbas, *et al.*, "Smart traffic monitoring system using computer vision and edge computing," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.8, pp.12027–12038, 2022.

[2] A. V. Savkin and H. L. Huang, "Navigation of a UAV network for optimal surveillance of a group of ground targets moving along a road," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.7, pp.9281–9285, 2022.

[3] N. Dilshad, J. Hwang, J. Song, *et al.*, "Applications and challenges in video surveillance via drone: A brief survey," in *Proceedings of 2020 International Conference on Information and Communication Technology Convergence*, Jeju, Korea (South), pp.728–732, 2020.

[4] Z. Liu, C. Zhan, Y. Cui, *et al.*, "Robust edge computing in UAV systems via scalable computing and cooperative computing," *IEEE Wireless Communications*, vol.28, no.5, pp.36–42, 2021.

[5] X. W. Li, H. P. Yao, J. J. Wang, *et al.*, "Rechargeable multi-UAV aided seamless coverage for QoS-guaranteed IoT networks," *IEEE Internet of Things Journal*, vol.6, no.6,

pp.10902–10914, 2019.

[6] Q. Chen, H. Zhu, L. Yang, *et al.*, "Edge computing assisted autonomous flight for UAV: Synergies between vision and communications," *IEEE Communications Magazine*, vol.59, no.1, pp.28–33, 2021.

[7] M. Q. Li, L. Liu, Y. Gu, *et al.*, "Minimizing energy consumption in wireless rechargeable UAV networks," *IEEE Internet of Things Journal*, vol.9, no.5, pp.3522–3532, 2022.

[8] Z. H. Yang, W. Xu, and M. Shikh-Bahaei, "Energy efficient UAV communication with energy harvesting," *IEEE Transactions on Vehicular Technology*, vol.69, no.2, pp.1913–1927, 2020.

[9] Y. W. Nie, J. H. Zhao, J. Liu, *et al.*, "Energy-efficient UAV trajectory design for backscatter communication: a deep reinforcement learning approach," *China Communications*, vol.17, no.10, pp.129–141, 2020.

[10] X. Zhang, X. H. Wang, X. P. Xu, *et al.*, "Demand learning and cooperative deployment of UAV networks," *Chinese Journal of Electronics*, vol.31, no.3, pp.408–415, 2022.

[11] C. W. Wang, Y. L. Cui, D. H. Deng, *et al.*, "Trajectory optimization and power allocation scheme based on DRL in energy efficient UAV-aided communication networks," *Chinese Journal of Electronics*, vol.31, no.3, pp.397–407, 2022.

[12] M. Q. Li, L. Liu, J. Xi, *et al.*, "ECTSA: An efficient charging time scheduling algorithm for wireless rechargeable UAV network, " in *Proceedings of 2021 IFIP Networking Conference*, Espoo and Helsinki, Finland, pp.1–9, 2021.

[13] Y. Jin, Z. J. Qian, S. R. Gong, *et al.*, "Learning transferable driven and drone assisted sustainable and robust regional disease surveillance for smart healthcare," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.18, no.1, pp.114–125, 2021.

[14] J. W. Xu, K. Zhu, and R. Wang, "RF aerially charging scheduling for UAV fleet: A Q-learning approach, " in *Proceedings of the 15th International Conference on Mobile Ad-Hoc and Sensor Networks*, Shenzhen, China, pp.194–199, 2019.

[15] L. L. Liu, A. M. Wang, G. Sun, *et al.*, "Multiobjective optimization for improving throughput and energy efficiency in UAV-Enabled IoT," *IEEE Internet of Things Journal*, vol.9, no.20, pp.20763–20777, 2022.

[16] M. Mozaffari, W. Saad, M. Bennis, *et al.*, "A tutorial on UAVs for wireless networks: applications, challenges, and open problems," *IEEE Communications Surveys & Tutorials*, vol.21, no.3, pp.2334–2360, 2019.

[17] Y. Yu, J. Tang, J. Y. Huang, *et al.*, "Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm," *IEEE Transactions on Communications*, vol.69, no.9, pp.6361–6374, 2021.

[18] J. J. Wang, C. X. Jiang, H. J. Zhang, *et al.*, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol.22, no.3, pp.1472–1514, 2020.

[19] V. François-Lavet, P. Henderson, R. Islam, *et al.*, "An introduction to deep reinforcement learning," *Foundations and Trends*, no.3-4, pp.219–354, 2018.

[20] Y. D. Yang and J. Wang, "An overview of multi-agent reinforcement learning from game theoretical perspective," *arXiv preprint*, arXiv: 2011.00583, 2020.

[21] Y. Li, S. Y. Xu, Y. P. Wu, *et al.*, "Network energy-efficiency maximization in UAV-enabled air-ground-integrated deployment," *IEEE Internet of Things Journal*, vol.9, no.15, pp.13209–13222, 2022.

[22] X. F. Chen, C. Wu, T. Chen, *et al.*, "Information freshness-aware task offloading in air-ground integrated edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol.40, no.1, pp.243–258, 2022.

[23] X. W. Pang, N. Zhao, J. Tang, *et al.*, "IRS-assisted secure UAV transmission via joint trajectory and beamforming design," *IEEE Transactions on Communications*, vol.70, no.2, pp.1140–1152, 2022.

[24] M. Mozaffari, W. Saad, M. Bennis, *et al.*, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," *IEEE Transactions on Wireless Communications*, vol.16, no.11, pp.7574–7589, 2017.

[25] C. H. Liu, Z. Y. Chen, J. Tang, *et al.*, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol.36, no.9, pp.2059–2070, 2018.

[26] M. Samir, D. Ebrahimi, C. Assi, *et al.*, "Leveraging UAVs for coverage in cell-free vehicular networks: A deep reinforcement learning approach," *IEEE Transactions on Mobile Computing*, vol.20, no.9, pp.2835–2847, 2021.

[27] X. Zhang and L. J. Duan, "Energy-saving deployment algorithms of UAV swarm for sustainable wireless coverage," *IEEE Transactions on Vehicular Technology*, vol.69, no.9, pp.10320–10335, 2020.

[28] F. Huang, G. X. Li, H. C. Wang, *et al.*, "Navigation for UAV pair-supported relaying in unknown IoT systems with deep reinforcement learning," *Chinese Journal of Electronics*, vol.31, no.3, pp.416–429, 2022.

[29] G. Y. Wu and J. C. Gu, "Remote interference source localization: A multi-UAV-based cooperative framework," *Chinese Journal of Electronics*, vol.31, no.3, pp.442–455, 2022.

[30] K. Zhu, J. Yang, Y. Zhang, *et al.*, "Aerial refueling: Scheduling wireless energy charging for UAV enabled data collection," *IEEE Transactions on Green Communications and Networking*, vol.6, no.3, pp.1494–1510, 2022.

[31] S. Fu, Y. J. Tang, Y. Wu, *et al.*, "Energy-efficient UAV-enabled data collection via wireless charging: a reinforcement learning approach," *IEEE Internet of Things Journal*, vol.8, no.12, pp.10209–10219, 2021.

[32] Z. H. Xiong, Y. Zhang, W. Y. B. Lim, *et al.*, "UAV-assisted wireless energy and data transfer with deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol.7, no.1, pp.85–99, 2021.

[33] Y. J. Fu, H. B. Mei, K. Z. Wang, *et al.*, "Joint optimization of 3D trajectory and scheduling for solar-powered UAV systems," *IEEE Transactions on Vehicular Technology*, vol.70, no.4, pp.3972–3977, 2021.

[34] M. Reymond and A. Nowe, "Pareto-DQN: Approximating the Pareto front in complex multi-objective decision problems, " in *Proceedings of the Adaptive and Learning Agents Workshop 2019*, Montreal, Canada, 2019.

[35] T. H. Wang, Y. G. Luo, J. X. Liu, *et al.*, "Multi-objective end-to-end self-driving based on Pareto-optimal actor-critic approach, " in *Proceedings of 2021 IEEE International Intelligent Transportation Systems Conference*, Indianapolis, IN, USA, pp.473–478, 2021.

[36] 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.

[37] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments, " in *Proceedings of 2014 IEEE Global Communications Conference*, Austin, TX, USA, pp.2898–2904, 2014.

[38] Q. Q. Zhang, W. Fang, Q. W. Liu, *et al.*, "Distributed laser charging: A wireless power transfer approach," *IEEE Inter-*

net of Things Journal, vol.5, no.5, pp.3853–3864, 2018.

[39] A. B. Sediq, R. H. Gohary, R. Schoenen, *et al.*, "Optimal tradeoff between sum-rate efficiency and Jain's fairness index in resource allocation," *IEEE Transactions on Wireless Communications*, vol.12, no.7, pp.3496–3509, 2013.

[40] W. T. Wei, R. Y. Yang, H. X. Gu, *et al.*, "Multi-objective optimization for resource allocation in vehicular cloud computing networks," *IEEE Transactions on Intelligent Transportation Systems*, vol.23, no.12, pp.25536–25545, 2022.

[41] R. Rădulescu, P. Mannion, and D. M. Roijers, "Multi-objective multi-agent decision making: A utility-based analysis and survey," *Autonomous Agents and Multi-Agent Systems*, vol.34, no.1, article no.10, 2020.

[42] C. F. Hayes, R. Rădulescu, E. Bargiacchi, *et al.*, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol.36, no.1, article no.26, 2022.

[43] J. G. Kuba, R. Q. Chen, M. N. Wen, *et al.*, "Trust region policy optimisation in multi-agent reinforcement learning," *arXiv preprint,* arXiv: 2109.11251, 2021.

[44] M. Zhou, Z. Y. Wan, H. J. Wang, *et al.*, "MALib: A parallel framework for population-based multi-agent reinforcement learning," *arXiv preprint*, arXiv: 2106.07551, 2021.

[45] J. Schulman, S. Levine, P. Moritz, *et al.*, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, Lille, France, pp.1889–1897, 2015.

[46] W. J. Wang and M. Sebag, "Hypervolume indicator and dominance reward based multi-objective Monte-Carlo Tree Search," *Machine Learning*, vol.92, no.2, pp.403–429, 2013.

[47] J. Xu, Y. S. Tian, P. C. Ma, *et al.*, "Prediction-guided multi-objective reinforcement learning for continuous robot control," in *Proceedings of the 37th International Conference on Machine Learning*, Vienna, Austria, pp. 10607–10616, 2020.

[48] S. Q. Zhang, J. Y. Lin, and Q. Zhang, "Succinct and robust multi-agent communication with temporal message control," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 17271–17282, 2020.

ZHOU Yi received the B.S. degree in electronic engineering from the First Aeronautic Institute of Air Force, China, in 2002, and the Ph.D. degree in control system and theory from Tongji University, China, in 2011. He is currently a Full Professor and Deputy Dean with the School of Artificial Intelligence, Henan University, China. He is also the Director of International Joint Research Laboratory for Cooperative Vehicular Networks, Henan, China. His research interests include vehicular cyber-physical systems and multi-agent collaboration. (Email: zhouyi@henu.edu.cn)

CHENG Xiang is currently a postgraduate of Henan University, Zhengzhou, China. His research interests include multi-agent cooperation and multi-UAV cooperation deployment.
(Email: richard@henu.edu.cn)

SHI Huaguang (corresponding author) received the B.S. degree in electronic science and technology from Zhengzhou University, Zhengzhou, China, in 2014, and the Ph.D. degree in measurement technique and automation equipment from the University of Chinese Academy of Sciences, Beijing, China, in 2021. He is currently a Lecturer with the School of Artificial Intelligence, Henan University, Zhengzhou, China. His current research interests include industrial Internet of things, wireless networks, and multi-agent learning.
(Email: shihuaguang@henu.edu.cn)

JIN Zhanqi is currently a Postgraduate of Henan University, China. His current research interests include UAV-assisted communications and Intelligent reflective surface.
(Email: Jinzhanqi@henu.edu.cn)

NING Nianwen was born in 1991, Ph.D. He is currently a lecturer with the School of Artificial Intelligence, Henan University, Zhengzhou, China. His main research interests include intelligent traffic and graph neural network.
(Email: nnw@henu.edu.cn)

LIU Fuqiang was born in 1965, Ph.D. candidate, professor. He is winner of National Natural Science Foundation (key project), and his main research direction include Internet of Vehicles and Intelligent Transportation.
(Email: liufuqiang@tongji.edu.cn)