# Multi-Objective Optimization for UAV-Assisted Wireless Powered IoT Networks Based on Extended DDPG Algorithm

Yu Yu, Jie Tang, *Senior Member*, *IEEE*, Jiayi Huang, Xiuyin Zhang, *Senior Member*, *IEEE*, Daniel Ka Chun So, *Senior Member*, *IEEE*, and Kai-Kit Wong, *Fellow*, *IEEE*

***Abstract*—This paper studies an unmanned aerial vehicle (UAV)-assisted wireless powered IoT network, where a rotary-wing UAV adopts fly-hover-communicate protocol to successively visit IoT devices in demand. During the hovering periods, the UAV works on full-duplex mode to simultaneously collect data from the target device and charge other devices within its coverage. Practical propulsion power consumption model and non-linear energy harvesting model are taken into account. We formulate a multi-objective optimization problem to jointly optimize three objectives: maximization of sum data rate, maximization of total harvested energy and minimization of UAV's energy consumption over a particular mission period. These three objectives are in conflict with each other partly and weight parameters are given to describe associated importance. Since IoT devices keep gathering information from the physical surrounding environment and their requirements to upload data change dynamically, online path planning of the UAV is required. In this paper, we apply deep reinforcement learning algorithm to achieve online decision. An extended deep deterministic policy gradient (DDPG) algorithm is proposed to learn control policies of UAV over multiple objectives. While training, the agent learns to produce optimal policies under given weights conditions on the basis of achieving timely data collection according to the requirement priority and avoiding devices' data overflow. The verification results show that the proposed MODDPG (multi-objective DDPG) algorithm achieves joint optimization of three objectives and optimal policies can be adjusted according to weight parameters among optimization objectives.**

***Index Terms*—Internet of Things (IoT), wireless power transfer (WPT), unmanned aerial vehicle (UAV), multi-objective optimization (MOO), deep deterministic policy gradient (DDPG).**

This work has been supported in part by Nation Key Research and Development Project under Grant 2019YFB1804100, in part by the National Natural Science Foundation of China under Grant 61971194, in part by Key Research and Development Project of Guangdong Province under Grant 2019B010156003, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515011607, in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2019D06), in part by the Fundamental Research Funds for the Central Universities under Grant 2019JQ08, and in part by the Research Fund Program of Guangdong Key Laboratory of Aerospace Communication and Networking Technology under Grant 2018B030322004. (*Corresponding author: Jie Tang.*)

Y. Yu, J. Huang, and X. Zhang are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: eeyu_yu@mail.scut.edu.cn; eejiayihuang@mail.scut.edu.cn; zhangxiuyin@scut.edu.cn).

J. Tang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: eejtang@scut.edu.cn).

D. K. C. So is with the School of Electrical and Electronic Engineering, University of Manchester, Manchester M13 9PL, U.K. (e-mail: d.so@manchester.ac.uk).

K.-K Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: kai-kit.wong@ucl.ac.uk).

## I. Introduction

Internet of Things (IoT), which can achieve the transparent and seamless integration of a large number of different and heterogeneous terminal systems at any time, any place, and anything inter-networking paradigm, has been regarded as a crucial and up-and-coming technology for future network and an important part of the new generation of information technology [1] [2]. In recent years, IoT technology has been widely used in residential life, industry, public management and other fields and has altered people's production patterns and life style. With the large-scale promotion and application of the IoT, the number of terminal devices grows explosively. As forecasted in [3] that about 500 billion devices will be equipped with sensors and connected to the Internet by 2030. The explosive growth of IoT devices puts forward higher requirements for communication system, including increased data rate and better coverage. In particular, timely transmission of gathering information is essential in typical sensing and monitoring scenario, e.g., power grid monitoring [4] and temperature and humidity monitoring [5]. For these applications, timely data collection (DC), on one hand, is crucial to the accuracy and reliability of derived decisions [6]. On the other hand, it is essential to avoid data overflows for the limited capacity of IoT devices. Besides, since IoT devices are power-limited typically, energy supply at mass low-power wide-area devices is another challenging issue in wireless IoT networks.

The development of 5G and even 6G is achieving ubiquitous connectivity at high-speed, low-latency, reliable and secure mobile broadband [7] [8], which has provided key technical support for ubiquitous deployment of the IoT technology. In particular, wireless power transfer (WPT) technology that based on radio frequency (RF) signal has been considered as a promising solution for energy supply problem of massive IoT devices [9]. Superior to obtaining energy from renewable sources, WPT can prolong the battery lifetime of widespread devices with stable as well as continuous energy over wireless link. In addition, it has a great advantage of low maintenance cost and high flexibility [10]. Since RF signal carries both energy and information, WPT is combined with wireless information transfer (WIT) technology to achieve simultaneous transmission of energy and information, so as to make the

best use of RF spectrum. In [11], the authors studied on the integration design of WPT and WIT. The design schemes were divided into three types, which were simultaneous wireless information and power transfer, wirelessly powered communication networks and wirelessly powered backscatter communication. How to meet the demands of data transmission and energy harvesting (EH) at the same time is the main challenge of WPT-based communications system. Besides, thanks to the high mobility, excellent maneuverability and low deployment cost, Unmanned aerial vehicle (UAV) has been applied to the wireless network to improve communication coverage, system capacity and deployment efficiency [12]–[14]. Combined with WPT, UAVs can execute DC and energy transfer for wildly distributed IoT devices. It has become a key component of IoT network [15] [16].

### A. Related Work

There have been many research works studied on the optimization of UAV-assisted wireless powered IoT networks in recent year. In [17]–[19], DC and energy transfer were processed using harvest-then-transmit protocol. It was on the basis of time division multiple access scheme where IoT devices harvested energy in downlink, and then used the energy to upload their gathered information in uplink. In [20] [21], UAV was assumed to be equipped with a full-duplex hybrid access point (HAP) and was able for simultaneous uplink WIT and downlink WPT. The optimization objectives of existing studies were various. In [17] [18] and [20], the objectives were to maximize the uplink throughput of all IoT devices. In [19], the authors aimed to maximize the minimum throughput of ground terminals. Besides, the authors of [21] considered of sum-throughput maximization, total-time minimization and total-energy minimization respectively. In some applications where environment changes rapidly, it is very important to ensure the real-time performance of data. To this end, in [22] and [23], the authors concerned about the freshness of the collected data and aimed to minimize the age of information of sensing data. Since UAV is energy-limited, energy consumption is an essential problem in UAV-assisted communication system. In [24], the authors derived a mathematical propulsion energy model of rotary-wing UAV and aimed to minimize the total energy consumption of UAV under the constraints of data rate of all the ground nodes. The energy consumption model was adopted in [25] and the tradeoff between energy consumption and mission completion time was revealed.

The mobility of UAV and the randomness and dynamics of IoT system pose great challenges to the optimization of UAV-assisted wireless IoT networks. Facing the complex and dynamic IoT network environments, UAV is required to equip with the ability to sense surrounding and the ability of real-time decision. Traditional optimization methods rapidly become unmanageable for these sophisticated network optimizations. Recently, artificial intelligence has been considered as the major innovative technique for UAV-assisted IoT system [26] [27]. Particularly, deep reinforcement learning (DRL), the integration of reinforcement learning (RL) [28] and deep learning [29], has become an emerging and promising technology and has attracted extensive attention. Taking full advantage of DRL algorithm, UAV can learn to build knowledge about the massive IoT environment without knowing the complete network information through iterative interaction, and then modifies its action strategy accordingly. In [30], a DRL-based multi-UAVs control strategy was proposed to achieve effective and fair communication coverage for ground Point-of-Interests in a target region. In particular, deep deterministic policy gradient (DDPG) algorithm [31] was leveraged for the continuous control task. This work was further studied in [32] and a multi-agent distributed solution was proposed. In [33], the authors considered the different priorities of required data and leveraged DRL technique to design the UAV's cruise route for data collection in the sensing region. A similar scenario was studied in [34] with multiple UAVs and multiple charging stations. In [30] and [32], the authors did not think of various priorities of users, which has been included in [33] and [34]. However, the priority requirements of data were assumed to be certain during the mission. Since in most practical application scenarios, massive ground nodes are deployed to observe real-time update of physical processes, dynamic priority requirements of sensing data should not be neglected. In [35] and [36], the authors proposed the data generation model to describe the real-time data update process, and developed online path planning algorithms for UAV based on Deep Q-Network (DQN) algorithm. As for the design of state space in environmental model, the information of all users and UAVs were included in state in [30] [32] and [34]. In [33] as well as [35], the observation of agent was set as a map and taken as input of convolutional neural network. These designs required a great deal of information about the environment.

### B. Contribution

In this paper, we study an UAV-assisted wireless powered IoT network where the UAV is equipped with a full-duplex HAP. Different from the above research works which either optimized single objective or optimized several objectives separately, we consider the joint optimization of multiple objectives. Our aim is to maximize the sum data rate and the harvested energy while reducing the energy consumption of the UAV. A DRL-based framework is proposed to solve the multi-objective optimization (MOO) problem. Instead of taking the full information of the system as input of neural network, we extract small amounts of information that closely related to the flying decision of the UAV to make up the state vector. The contributions of our work can be summarized as follow:

- We propose the UAV-assisted data collection and energy transfer in wireless powered IoT network, where the requirements of IoT devices to upload data are updated in real time and UAV adopts the fly-hover-communicate protocol to successively visit IoT devices according to their requirements priorities.
- We investigate a MOO problem that aims to maximize sum data rate and total harvested energy and to minimize UAV's energy consumption simultaneously, where

a MODDPG algorithm to find optimal policies of UAV's flight decision is developed. To achieve MOO, we design the reward as a 4-dimensional vector, where three elements correspond to the three optimization objectives and another auxiliary element ensures the completion of the basic task, and extend the classical DDPG algorithm to multi-dimensional reward.
- Through the training results, we show that the optimal policy based on the proposed MODDPG algorithm is more flexible than traditional rule-based policies. By modifying the weight parameters, the optimal policies can be adjusted to achieve the coordination and optimization of multiple objectives under different priorities.

The remainder of this paper is organized as follows. The system model and the MOO problem are presented in Section II. In Section III, we give a brief introduction of RL. In Section IV, we propose the MODDPG algorithm for UAV-assisted data collection and energy transfer, including the construction of the environmental model and the design of algorithm framework. Simulation results are shown and analyzed in Section V and we generalize conclusions in Section VI.

## II. System Description and Problem Formulation

In this section, we first present a wireless powered IoT network where UAV executes data collection and energy transfer and then propose the MOO optimization problem.

### A. System Model

As Fig. 1 shows, we consider a wireless powered IoT network with a two-antenna UAV and $J$ single-antenna IoT devices. The IoT devices are randomly located in a limited geographic area. Considering that the UAV is energy-limited, each flying mission lasts for a specific period with duration $T > 0$. Fly-hover-communicate protocol is adopted in our work. As stipulated in the protocol, the UAV does not communicate with ground nodes whiling flying and only processes DC and energy transfer when hovering. The UAV is equipped with a HAP. When it is hovering at a corresponding location, it operates in full-duplex mode. It transmits energy to IoT devices in downlink with one antenna and collects data from IoT devices in uplink with the other antenna simultaneously.

*1) **IoT device**:* We use the $\mathcal{J} \triangleq \{j = 1, 2, \cdots, J\}$ to denote IoT devices. They are distributed randomly on the ground. $[x_j, y_j]$ denotes the location of device $j$. We consider the practical application scenario that IoT devices monitor a variety of physical processes online. The status update packets about their observed processes are gathered and stored in their data buffer in real time. $l_j(t)$ denotes data in a queue waiting to be uploaded at $t, 0 \leq t \leq T$. It is updated according to

$$l_j(t + \Delta t) = l_j(t) + \lambda_j(t)\Delta t, \tag{1}$$

where $\Delta t$ is the update interval, $\lambda_j(t)$ is the data generation rate of devices $j$ at $t$. We assume that $\lambda_j(t)$ obeys the Poisson distribution, and the parameter of Poisson distribution of different devices are different. The maximum of $l_j(t)$ is typically constrained by hardware limitation and it is assumed to be bounded by $[0, l_{\max}]$, where $l_{\max}$ is the storage capacity

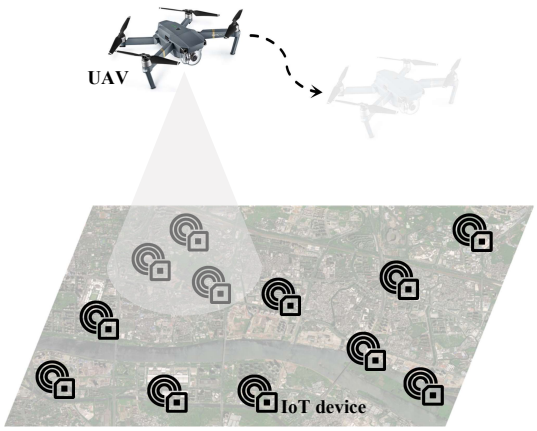


Fig. 1: System model

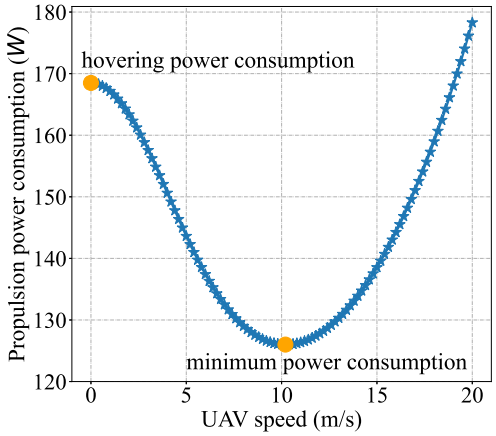


Fig. 2: Propulsion power consumption versus speed V

of the data buffer and it is assumed the same for all devices. If the data buffer is filled up with data, the older data may be overwritten by new data, or the new gathered data may be dropped, both of which will result in data loss. Thus it is of great importance for IoT devices to upload their gathered information to UAV in time. We assume that the time division multiple access protocol is applied for the information transmission of IoT devices and the uplink transmit power of devices are $P_u$. The transmission data size in bit corresponding to $l_{\max}$ is Q. The data to be transferred at $t$ is

$$Q_j(t) = \frac{l_j(t)}{l_{\max}} Q. \tag{2}$$

Since the length of data buffer and the data generation rate vary from device to device, their priorities to upload data is different. We denote $q_j^u(t)$ to represent the data upload priority of device $j$. It is given as

$$q_j^u(t) = \lambda_j(t) \frac{l_j(t)}{l_{\max}}. \tag{3}$$

The data transmission priority not only relies on the ratio of gathered data to the storage capacity, but is also affected by data generation rate. It contains the prediction of future priority.

*2) **UAV**:* As for UAV, we assume it flies at a fixed altitude $H > 0$. The horizontal location at time $t$ is denoted as $[x_u(t), y_u(t)]$, and the hovering altitude is omited here. The UAV determines its next action in real time and updates the position accordingly. The flight control of the UAV is described by flight speed $v(t)$ and yaw Angle $\theta(t)$, where $v(t)$ is limited by the maximum flying velocity $v_{\max} = 20$ m/s and $\theta(t) \in [-\pi, \pi]$. While flying, the propulsion power consumption with speed $V$ can be calculated by [24] as follows.

$$\begin{aligned} P(V) =& P_0 \left(1 + \frac{3V^2}{U_{tip}^2}\right) + P_i \left(\sqrt{1 + \frac{V^4}{4v_0^4}} - \frac{V^2}{2v_0^2}\right)^{1/2} \\ &+ \frac{1}{2} d_0 \rho s A V^3. \end{aligned} \tag{4}$$

The propulsion power consumption of UAV includes blade profile, induced power and parasite power, corresponding to the three parts of the above formula. $P_0$ is blade profile power in hovering and $U_{tip}$ is the tip speed of rotor blade. $P_i$ and $v_0$ denote induced power and the mean rotor induced velocity under the hover condition. As for parasite power, $d_0$, $\rho$, $s$, $A$ respectively denote the fuselage drag ratio, air density, rotor solidity and rotor disc area. The variation trend of propulsion power consumption with speed is shown in Fig. 2. We can find that the power decreases first and then increases in the acceleration. The speed corresponding to the lowest power consumption is mentioned as maximum-endurance (ME) speed $V_{\text{ME}}$. And the hovering power consumption $P_{hov} = P_0 + P_i$ can be calculated by setting $V = 0$.

We assume that the range of UAV's DC and energy transfer is limited. UAV only charges and collects data from IoT devices that fall within the coverage. This assumption is reasonable since communication is inefficient when IoT devices are too far away from UAV. We denote $D_{dc}$ and $D_{eh}$ to represent the maximum coverage radius of data collection and energy transfer. At each moment, UAV chooses an IoT device as the target device for data collection. Once the target device falls within $D_{dc}$, UAV will hover at the corresponding location to receive information and transmit energy to other devices within $D_{eh}$ at the same time until the target device completes its data upload. We denote $P_d$ to represent the downlink transmit power of UAV.

*3) **Channel Model**:* We denote the downlink channel power gain and uplink channel power gain of wireless communication link between UAV and IoT device $j$ as $h_j(t)$ and $g_j(t)$, respectively. The practical air-to-ground channel model that combined with line-of-sight (LoS) link and non-line-of-sight (NLoS) link is considered. The mathematical description of corresponding pass loss is given as follows

$$L_j(t) = \begin{cases} \gamma_0 d_j^{-\tilde{\alpha}}, & \text{LoS link} \\ \mu^{\text{NLoS}} \gamma_0 d_j^{-\tilde{\alpha}}, & \text{NLoS link} \end{cases}, \tag{5}$$

where $\gamma_0 = (\frac{4\pi f_c}{c})^{-2}$ represents channel power gain at the reference distance of $d_0 = 1$ m, with $f_c$ denoting the carrier frequency and $c$ denoting the speed of light. $d_j^{-\tilde{\alpha}}$ is the propagation distance between UAV and IoT device $j$, where $\tilde{\alpha}$ stands for the path loss exponent. $\mu^{\text{NLoS}}$ is the additional attenuation coefficients of NLoS link. As for IoT device $j$, the LoS probability at time $t$ can be expressed as

$$P_j^{\text{LoS}}(\theta_j(t)) = \frac{1}{1 + a \exp(-b(\theta_j(t) - a))}. \tag{6}$$

The LoS probability of channel condition depends largely on the propagation environment. $a$ and $b$ are constant values that depend on the carrier frequency and the type of environment. It also influenced by the relative location of the communicating parties. $\theta_j(t)$ is the elevation angle of UAV and IoT device in degree. It is given as $\theta_j(t) = \frac{180}{\pi} \sin^{-1}\left(\frac{H}{d_j(t)}\right)$. $d_j(t) = \sqrt{H^2 + (x_u(t) - x_j)^2 + (y_u(t) - y_j)^2}$ is the distance between UAV and the IoT device $j$. The probability of the NLoS component then can be given by $P_j^{\text{NLoS}}(\theta_j(t)) = 1 - P_j^{\text{LoS}}(\theta_j(t))$. We assume the uplink and downlink channels are approximately equal. As a result, the channel power gain between UAV and IoT device $j$ is given as

$$\begin{aligned} h_j(t) &\approx g_j(t) \\ &= \left(P_j^{\text{LoS}}(\theta_j(t)) + \mu^{\text{NLoS}} P_j^{\text{NLoS}}(\theta_j(t))\right) \gamma_0 d_j(t)^{-\tilde{\alpha}}. \end{aligned} \tag{7}$$

*4) **Energy Harvesting Model**:* The UAV works in full duplex mode at its hovering stage. It keeps transmitting RF signals to devices with the constant transmit power $P_d$ when it receives information from the target device over uplink channel. The devices within it energy transfer coverage range excepts the target device will be charged. The received power at device $j$ is

$$P_j^r(t) = |h_j(t)|^2 P_d, \quad \forall \Delta d_j(t) \le D_{eh}. \tag{8}$$

It this paper, we apply the non-linear EH model [37]. Different from linear model, non-linear EH model considers the saturation limitation of the circuits and is more practical. Through RF-EH circuit, the harvested energy is described by

$$P_j^h(t) = \frac{P_{\text{limit}} e^{cd} - P_{\text{limit}} e^{-c(P_j^r(t) - d)}}{e^{cd}(1 + e^{-c(P_j^r(t) - d)})}, \tag{9}$$

where $P_{\text{limit}}$ is the maximum output DC power, $c$ and $d$ are constants that depend on related circuit characteristics of the EH system.

### *B. Problem Formulation*

In this work, we aim to maximize sum data rate and total harvested energy, and minimize energy consumption of UAV at the same time. The UAV is required for perception of IoT environment and implement real-time path planning. The decision of UAV flight trajectory and the choose of hovering position should consider quality of service of devices and energy consumption of UAV. Furthermore, the avoidance of data overflow of all IoT devices is of great importance. To this end, the UAV successively visits IoT devices according to their real-time requirements priority. For example, the IoT device $\hat{j} = \arg\max_j q_j^u(t)$ will be chosen as the target device of UAV at $t$. When UAV flies close to the target device enough, e.g., $d_j(t) \le D_{dc}$, it hovers at the corresponding location and starts collecting data in uplink and transmitting energy in downlink. Let $k, 0 < k \le K$ represent the $k^{th}$ hovering of UAV in a

mission, where $K \geq 0$ denotes the total number of times that UAV hovers to communicate with IoT devices. We denote the corresponding communication device of the $k^{th}$ hovering as $j^k$. Then the transmission data rate at the $k^{th}$ hovering is given as

$$R^k = W \log_2 \left(1 + \frac{P_u |g_{j^k}(t)|^2}{\sigma_n^2}\right), \tag{10}$$

where $W$ is the wireless communication bandwidth and $\sigma_n^2$ is the channel noise power at UAV. To upload all of the gathered data to UAV, the hovering time can be calculated by

$$t^k = \frac{Q_{j^k}(t)}{R^k}. \tag{11}$$

At the same time, the UAV keeps transmitting energy to devices within its energy transfer coverage expect device $j^k$ in downlink. According to (8)(9), the harvested power at device $j$ is given by

$$E_j = P_j^h(t) t^k, \quad \forall \Delta d_j(t) \leq D_{eh}, j \neq j^k. \tag{12}$$

Then the total harvested energy at the $k^{th}$ hovering is given as

$$E^k = \sum_{j, \forall d_j(t) \leq D_{eh}, j \neq j^k} E_j. \tag{13}$$

The sum data rate and the total harvested energy of all the hovering stages in a mission are given as following.

$$R_{sum} = \sum_{k=0}^{K} R^k, \tag{14}$$

$$E_{total}^h = \sum_{k=0}^{K} E^k. \tag{15}$$

And the total energy consumption for UAV's flying and hovering in the task duration is given as

$$E_{total}^c = \int_0^T P(v(t))\, dt. \tag{16}$$

It should be noted that energy consumption also includes communication energy. Since we assume the downlink transmit power $P_d$ a constant, this component is not included in the optimization objective. The MOO problem can be formulated as

$$\mathbf{P1}: \max_{v(t), \theta(t)} \ (R_{sum}, \ E_{total}^h, \ -E_{total}^c) \tag{17}$$

$$\text{s.t.} \quad v(t) \in [0, v_{\max}], \tag{18}$$

$$\theta(t) \in [-\pi, \pi]. \tag{19}$$

As for the sum data rate, its maximization depends on the amount of devices uploading data over the UAV mission period, that is, the total number of hoverings $K$ and the data rate in each hovering. It can be easily concluded that to maximize $R_{sum}$, on one hand, the UAV should fly at a higher speed so that it can visit more IoT devices. On the other hand, the hovering location should be close to the target device so as to improve the data rate as well as shorten the communication time of each hovering. From this aspect, hovering over the device is the best choice. As for the maximization of total harvested energy, besides the maximization of $K$, we hope that more devices fall within the coverage of UAV at each hovering. In addition, the smaller distance between UAV and charging devices, the better. It may conflict with the UAV's hovering directly over the target device to get the maximum data rate. As for the objective of UAV's energy consumption, it is clearly that $V_{\text{ME}}$ can achieve its minimization. However, it may be not fast enough to collect more data and charge more devices. What's more, the low flying speed may lead to IoT device's data overflow.

As we can see, these three objectives are in conflict with each other partly. Since the devices are randomly distributed and their data generation are dynamic, it is substantially complex and may impose considerable computational cost to find out an optimal hovering location and make flying decision. Furthermore, the environment is partially observed, traditional model-based methods like dynamic programming method are unable to fix this problem. Recently, DRL has shown excellent ability of solving complex problems and is regarded as one of the core technologies of artificial intelligence. As the integration of deep learning and RL, it owns the strong understanding ability and decision-making ability and thus can realize end-to-end learning. It has shown great potential in solving sophisticated network optimizations. DDPG, which is one of the classical DRL algorithms, has been proved that can learn effective polices in continuous action spaces using low-dimensional observations [31]. It is suitable for our proposed UAV's flight decision problem where flying speed and yaw Angle are chosen in continuous interval. Since the reward of original DDPG algorithm is scalar, we extend it to multi-dimensional reward for the MOO problem. A MODDPG algorithm is proposed for UAV-assisted data collection and energy transfer and weight parameters are introduced to describe the preferences of the objectives.

## III. Preliminaries

Here we give a brief introduction of reinforcement learning [28]. RL is one of the fields of machine learning. It emphasizes that agent learns through interaction with the environment directly, with on need for imitative monitoring signals or complete modeling of the surrounding environment. Based on the formal framework of the markov decision processes (MDP), RL problem can be described by a five-tuple $< \mathcal{S}, \mathcal{A}, r, p, \gamma >$. $\mathcal{S}$ and $\mathcal{A}$ are set of states and actions. $r$ is reward function. $p$ is transition function that indicates the probability of moving from one state to the next state. And $\gamma$ is discounting factor to exponentially discount the value of future rewards. The agent uses the discount factor to adjust the importance of rewards over time. $\gamma$ is always a positive real value less than one. Policy and value function are core elements of RL. A policy defines the agent's behavior in an environment. It typically represented by a function that determines the next action $a \in \mathcal{A}$ to take given a state $s \in \mathcal{S}$ and denoted as $\pi(a|s)$. The agent aims to learn a policy that maximizes the discounted return during an episode. The discounted return $G$ at time step $t$ can be calculated by

$$G_t = \sum\nolimits_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{20}$$

where $r$ is usually denoted by $r(s,a)$. It is the reward that agent obtains when it executes action $a$ at state $s$. During the learning process, the agent learns to optimize its policy toward reaching a better policy with the experience transition. And the optimal policy can be obtained as

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[G|\pi\right]. \tag{21}$$

Value function determines what is good for the agent in the long run, which helps agent to learn and find optimal policies. Particularly, action-value function, which is also referred to as Q-function, defines the value of action $a$ in state $s$ under a policy $\pi$. It is denoted as $Q^{\pi}(s,a)$ and given formally as

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[G_t|s_t = s, a_t = a]. \tag{22}$$

Original RL algorithms use table, or (non) linear function as approximator to estimate the Q-function. Table is used in Q-learning [38]. It is a classical off-policy algorithm where the actions that the Q-learning function learns from are outside the current policy. And the agent relies on the Q-table to select the best action. It is obvious that this method is not applicable to higher dimensional state and action. To this end, deep neural network is introduced as more powerful nonlinear function approximators. DQN uses neural network as the function approximator to approximate Q-function [39]. And the Q network $\theta^Q$ is optimized by minimizing the loss between Q-function and target value:

$$L\left(\theta^Q\right) = \mathbb{E}\left[y_t - \left(Q\left(s_t, a_t|\theta^Q\right)\right)^2\right]. \tag{23}$$

The target value $y_t$ is obtained by

$$y_t = r\left(s_t, a_t\right) + \gamma Q\left(s_{t+1}, a'|\theta^Q\right). \tag{24}$$

However, DQN can only make decisions on problem with discrete and low-dimensional action spaces and unable to solve continuous action control problem. For continuous control task, policy gradient algorithm was applied and an actor-critic approach has been proposed in [40]. It combines two types of RL algorithms that based on values (like Q-learning) and action probabilities (like Policy Gradients) and thus can effortlessly select the right action from the continuous action space. On the basis of the actor-critic framwork and learning from successful insights of DQN, DDPG was proposed in [31] and turned out to be efficient for robustly solving a complex problems with continuous action spaces from a variety of fields.

## IV. MODDPG ALGORITHM FOR UAV-ASSISTED DATA COLLECTION AND ENERGY TRANSFER

In this section, we first build the environmental model which maps the system model to the interaction environment of MDP. And then propose an UAV-assisted data collection and energy transfer algorithm for the proposed MOO problem.

### A. Environmental Model

It is of great importance to cast the optimization problem into the MDP in a right way. The agent depends on the interaction with the environment to adjust its behavior and learn optimal policies. Here we give detail description of the design of state space, action space and reward in our model.

*1) **State Space**:* Collecting the real-time service requirements of all the IoT devices relies on frequent information exchange between the UAV and IoT devices. It will occupy a large amount of wireless resources and cause delay, greatly reducing the efficiency of the system. To be more practical, we assume that the UAV can only observe its own state and partial network information. To be specific, UAV can observe its own location, the cumulative number of flights out of the restricted area, the location of the target device, and the number of devices with data loss. And then the state space is defined symbolically as

$$\mathcal{S} \triangleq \{\mathbf{s}_t\} = \{[d^x_{\hat{j}}(t), d^y_{\hat{j}}(t), x_u(t), y_u(t), N_f(t), N_d(t)]\}, \tag{25}$$

where $[d^x_{\hat{j}}(t), d^y_{\hat{j}}(t)]$ is the distance between the target device and the UAV under the cartesian coordinates. Once the UAV has finished the data collection of the target device, a new one will be selected according to the status of the system at the time. This element helps to guide the UAV to get the target devices into its data collection coverage. $N_f(t)$ records the cumulative number of times that the UAV has continuously exceeded the restricted area by the time $t$. Combining with UAV's absolute position $[x_u(t), y_u(t)]$, it helps to keep the UAV from flying out of the designated area that causes unnecessary waste of resources. And the number of devices with data loss $N_d(t)$ will drive the UAV to service the high-demand devices timely. In practical scenarios, the global network information is incapable to obtain and the real-time knowledge about each device is unknowable at the UAV. Besides, most of the information is not necessary for decision-making. In our setting, we extract a small amount of necessary information to represent the state of the environment. These elements of state space will enable the UAV to have a good overall perception of the environment. Furthermore, it overcomes the lack of network information which is common problem that exists in massive uncertain IoT system.

*2) **Action Space**:* Observing the state, the UAV makes action decision in real time. The action space is defined as

$$\mathcal{A} \triangleq \{\mathbf{a}_t\} = \{[v(t)cos(\theta(t)), v(t)sin(\theta(t))]\}. \tag{26}$$

We use $[cos(\theta(t)), sin(\theta(t))]$ to represent yaw and then the network will learn a normalized two-dimensional vector. The flying speed $v(t)$ and the yaw Angle $\theta(t)$ are assumed to be continuous value in the interval $[0, v_{\max}]$ and $[-\pi, \pi]$ respectively. It enlarges the control freedom of the UAV as well as improves the efficiency of the control scheme comparing to discrete action space.

*3) **Reward**:* Since the environment is partially observed, the UAV depends on the reward to evaluate its decision, infer the distribution of states and learn and know the environment. Besides, the agent relies on the well-designed reward function to learn effective control policy for the proposed MOO problem. According to our optimization problem, the reward is designed as a 4-dimensional vector.

$$\mathcal{R} \triangleq \{\mathbf{r}_t\} = \{[r_{dc}(t), r_{eh}(t), r_{ec}(t), r_{aux}(t)]\}, \tag{27}$$

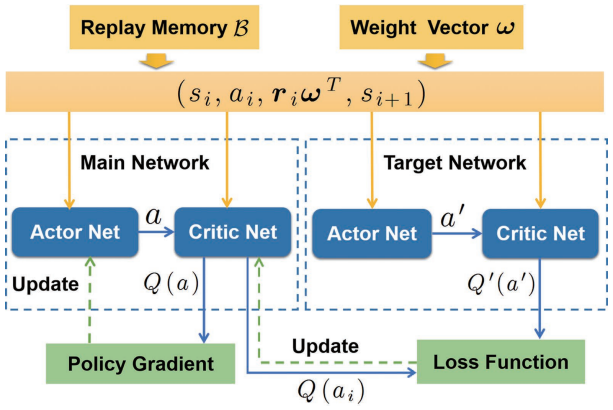


Fig. 3: Framework of MODDPG algorithm

where $r_{dc}(t)$, $r_{eh}(t)$, $r_{ec}(t)$ correspond to the three optimization objectives: maximization of sum data rate, maximization of total harvested energy and minimization of UAV's energy consumption. They are designed as following.

$$r_{dc}(t) = \begin{cases} 100 \times R^k, & \text{at UAV's } k^{th} \text{ hovering} \\ 0, & \text{otherwise} \end{cases}. \tag{28}$$

$$r_{eh}(t) = \begin{cases} 100 \times (E^k + \sum_{0}^{J} \mathbb{I}_{d_j(t) \le D_{eh}}), & \\ & \text{at UAV's } k^{th} \text{ hovering} \\ 0, & \text{otherwise} \end{cases}. \tag{29}$$

$$r_{ec}(t) = \begin{cases} -P_{hov}, & \text{if UAV is hovering} \\ -P(v(t)), & \text{otherwise} \end{cases}. \tag{30}$$

Once the target device falls within the data collection coverage radius of UAV, the UAV will hover to process data collection and energy transfer. Otherwise the UAV is in flying stage. We give more rewards to the agent for its higher data rate, more harvested energy at more IoT devices in hovering, and punish it for its higher energy consumption at both flying and hovering stages. $w_{dc}$, $w_{eh}$ and $w_{ec}$ are priority weights associated with each attribute. In addition, there is an auxiliary reward $r_{aux}(t)$ that given as

$$r_{aux}(t) = -d_j^x(t) - d_j^y(t) - N_f(t) - N_d(t). \tag{31}$$

It can be seen that $r_{aux}(t)$ includes the distance between UAV and the target device. It will be small if the UAV is far away from the target device, which helps the UAV recognize the location of the target device so as to get close to it. Besides, if the UAV tries flying out of the restricted area or leads to IoT devices' data overflow due to the failure of timely data collection, it will get negative reward. We inflict punishment on UAV's bad flight decisions to drive the UAV to learn to finish the basic tasks no matters the preferences of the optimization objectives. The corresponding weight $w_{aux}$ is set as 1 all the time.

### *B. MODDPG Algorithm*

The algorithm framework of MODDPG is presented in Fig.3. Based on DDPG architecture, we maintain an actor network $\mu(s|\theta^\mu)$ to specify the main policy that builds a mapping from states to actions and a critic network $Q(s, a|\theta^Q)$ to estimate the action value. $\theta^\mu$ and $\theta^Q$ are parameters of two networks. The weights of both the actor network and critic network are initialized from a truncated normal distribution centered on 0 with standard deviation $\sqrt{2/f}$ , where $f$ is the number of input units in the weight tensor. The biases are all initialized as 0.001. Besides, target network is applied to the actor-critic architecture to calculate the target values. Specifically, a target actor network $\mu'(s, a|\theta^{\mu'})$ and a target critic network $Q'(s, a|\theta^{Q'})$ are created by copying the parameters of mian network in the initialization phase.

While updating the network parameters, a random mini-batch of experience tuples are sampled uniformly from replay memory. Different from the original DDPG, which is single-objective MDP with scalar reward signal, the reward in the experience tuples is a vector. Since the value of the action depends on the preferences among competing objectives, we use the linear weighting method to calculate the weighted sum

**Algorithm 1** MODDPG algorithm for UAV-assisted data collection and energy transfer

**Input:** a weight vector $\boldsymbol{w} = [w_{dc}, w_{eh}, w_{ec}, w_{aux}]$.

Initialize main network and target network;
Initialize replay memory $\mathcal{B}$, Initialize $\sigma^2 = 2.0$, $\epsilon = 0.9999$ for action exploration;
**for** episode $:= 1, \cdots, \mathrm{M}$ **do**
  **for** step $t := 1, \cdots, T$ **do**
    Update the environment status and observe the current state $s_t$;
    Select action according to $a_t \sim \mathcal{N}(\mu(s_t|\theta^\mu), \epsilon\sigma^2)$;
    Execute action $a_t$ and limit UAV in designated area, observe reward $\boldsymbol{r}_t$, transit to the next state $s_{t+1}$;
    Store the experience tuple $(s_t, a_t, \boldsymbol{r}_t, s_{t+1})$ into replay memory $\mathcal{B}$;
    **if** update **then**
      Randomly sample a mini-batch transitions from $\mathcal{B}$.
      Compute $y_i$

$$\begin{cases} \boldsymbol{r}_i\boldsymbol{w}^{\mathrm{T}}, & \text{for teminal } s_{i+1} \\ \boldsymbol{r}_i\boldsymbol{w}^{\mathrm{T}} + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{\boldsymbol{Q}'}\right), & \text{otherwise.} \end{cases}$$

      Update critic network by minimizing the critic loss (35);
      Update actor network by maximizing the actor loss (36);
      Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \tag{32}$$

$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}. \tag{33}$$

      Decay the action randomness: $\sigma^2 \leftarrow \sigma^2\epsilon$.
    **end if**
  **end for**
**end for**

TABLE I: Simulation parameters

| | |
|---|---|
| Bandwidth ($B$) | 1MHz |
| Noise power($\sigma_n^2$) | -90dBm |
| Reference channel power gain ($\gamma_0$) | -30dB |
| Attenuation coefficients of NLoS link ($\mu$) | 0.2 |
| Path loss exponent ($\tilde{\alpha}$) | 2.3 |
| parameters of LoS probability($a$, $b$) | 10, 0.6 |
| blade profile power ($P_0$) | 79.86 |
| induced power ($P_i$) | 88.63 |
| Tip speed of rotor blade ($U_{tip}$) | 120m/s |
| Mean rotor induced velocity in hover ($v_0$) | 4.03 |
| Fuselage drag ratio ($d_0$) | 0.6 |
| air density ($\rho$) | $1.225 km/m^3$ |
| Rotor solidity ($s$) | 0.05 |
| rotor disc area ($A$) | 0.503 $m^2$ |
| maximum output DC power ($P_{limit}$) | 9.079$\mu$W |
| parameters of EH model($c$, $d$) | 47083, 2.9 $\mu$W |

TABLE II: Network configurations

| **Parameters** | **Values** |
|---|---|
| Network structure for actor | [400,300] |
| Network structure for critic | [400,300] |
| Number of training episodes | 1600 |
| Learning rate for actor | $10^{-3}$ |
| Learning rate for critic | $10^{-3}$ |
| Reward discount | 0.9 |
| Replay memory size | 8000 |
| Batch size | 64 |
| Initial exploration variance | 2.0 |
| Final exploration variance | 0.1 |
| Soft target updates parameter | 0.001 |

of elements of the reward vector with the given weights, which is given as $r = \boldsymbol{r}\boldsymbol{w}^{\mathrm{T}}$, where $\boldsymbol{w} = [w_{dc}, w_{eh}, w_{ec}, w_{aux}]$. Then the reward vector is transformed into scalar form. It should be noted that through this design, the MODDPG algorithm is suitable for MOO problem with arbitrary number of objectives. And it also supports single objective optimization (SOO) problem. In our setting, all of the weight parameters are chosen in the interval $[0.0, 1.0]$ according to the importance preference of each attribute. With the target network, the target value $y_i$ is calculated as following.

$$y_i = \boldsymbol{r}\boldsymbol{w}^{\mathrm{T}} + \gamma Q'\left(s_{i+1}, \mu'\left(s_{i+1}|\theta^{\mu'}\right)|\theta^{Q'}\right). \tag{34}$$

To optimize the main critic network, we calculate the difference between target value and Q-function given by the main critic network. Then the main critic network is trained by using the gradient descent method to minimize the loss function, which is defined as the mean square error of the difference.

$$L\left(\theta^Q\right) = \mathbb{E}\left[\left(Q\left(s_i, a_i|\theta^Q\right) - y_i\right)^2\right]. \tag{35}$$

The loss function of actor network is simply obtained by calculating the sum of Q-function for the states. We use main critic network and pass action computed by main actor network to compute the Q-function. The loss function of actor network is

$$L\left(\theta^\mu\right) = \mathbb{E}\left[Q\left(s_i, \mu\left(s_i|\theta^\mu\right)|\theta^Q\right)\right]. \tag{36}$$

The chain rule is applied to update actor network weights by maximizing $L\left(\theta^\mu\right)$. And the parameters of two target networks will update during the training using "soft" target update.

To ensure adequate exploration of the continuous action spaces, An exploration policy is applied to the actor policy. In detail, at each decision step, the action is selected from a random process that follow a Gaussian distribution with expectation $\mu\left(s_t, |\theta_t^\mu\right)$ and variance $\epsilon\sigma^2$, where $\epsilon$ is an adjustable parameter to decay the action randomness of the training process. The complete algorithm is presented in **Algorithm 1**.

## V. SIMULATION RESULTS AND DISCUSSION

In this section, we present numerical results to evaluate the performance of MODDPG algorithm. We set the number of

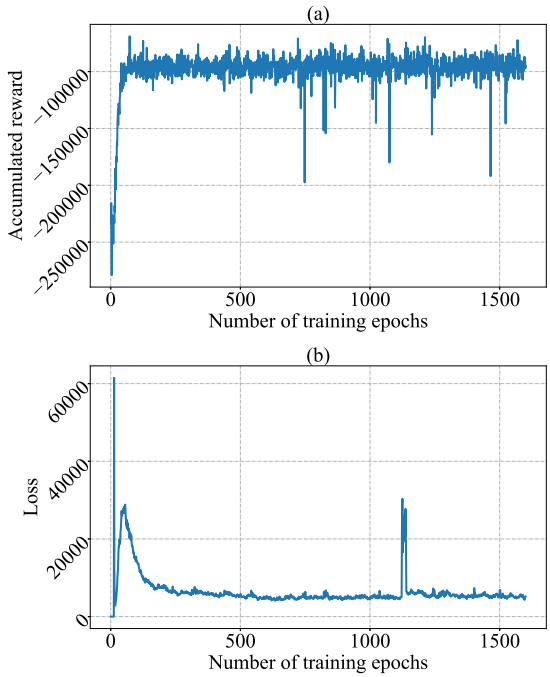


Fig. 4: Training curves of the network: (a) Accumulated reward; (b) Loss.

IoT devices to 100, the mission period to 10 minutes. IoT devices are randomly distributed in a square area with the range of 400 m by 400 m. At the beginning of each task, the UAV begins its mission at a random position in the designated area. It flies at an altitude of 10 m and the maximum flying speed $v_{\max} = 20$ m/s [22]. The radius of UAV's coverage are set to $D_{dc} = 10$ m and $D_{eh} = 30$ m. The transmit power of the UAV and IoT devices are set to $P_d = 40$ dBm and $P_u = -20$ dBm respectively [19]. The data cache of IoT devices are updated per second. Their expectations of the Poisson process of data accumulation are randomly assigned from the set $\{4, 8, 15, 20\}$. The capacity of data buffer $l_{\max}^d$ is 5000 packets and the corresponding transmission data size is set to $Q = 10$ Mbits. Other system parameters are listed in Table I, where the parameters are set by referring to [19] [24]. The structure and parameters of the MODDPG network are

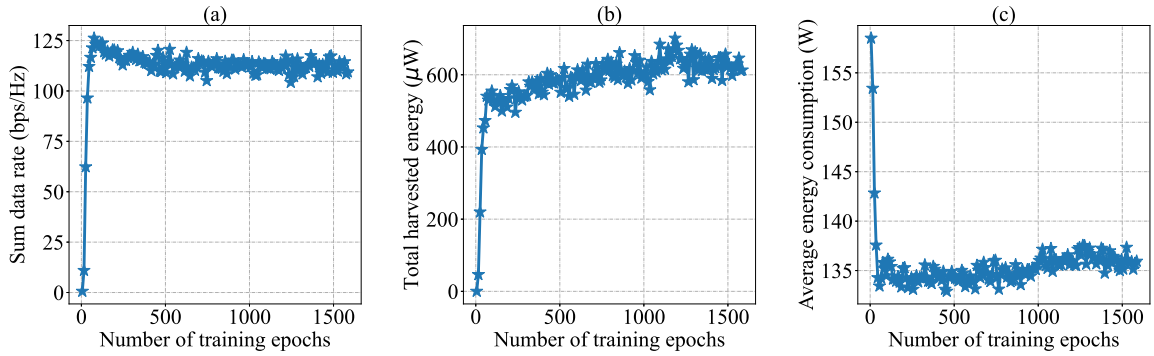


Fig. 5: Training curves tracking optimization objectives: (a) Sum data rate; (b) Total harvested energy; (c) Average energy consumption.

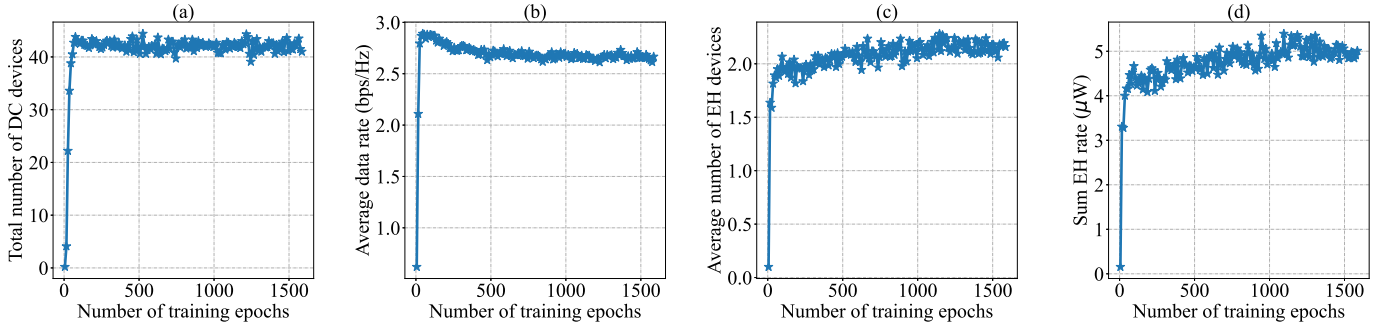


Fig. 6: Training curves tracking optimization results: (a) Total number of DC devices; (b) Average data rate; (c) Average number of EH devices; (d) Sum EH rate.

given in Table II. In our implementation, all the hidden layers are fully connected and ReLU function is used for activation. The final output layer of actor network is set to be a tahn layer to bound the actions.

First we show the effectiveness and convergence of the proposed MODDPG algorithm. The learning curves of the trained MODDPG agent are shown in Fig. 4. The agent's accumulated reward and the loss of critic network are given. The changing trend of three objectives as well as corresponding results during the training are also illustrated in Fig. 5 and Fig. 6. Here the weight parameters are set to $w_{dc} = w_{eh} = w_{ec} = 1.0$. We consider the jointly optimization of all three objectives. It can be seen in Fig. 4 that the agent quickly learns to obtain higher expected total rewards as training progresses. And then the accumulated reward converges steadily at a high level. At first about 10 epochs, the accumulated reward fluctuates at a very low level. It is because that the UAV is in complete experience stage. Without enough experience to learn from, the action is chosen randomly. At the same time, the loss of the network is 0 and all of the objectives are not optimized. When the replay memory is full, the UAV begins to sample the stored experience tuples to train networks. We can see that there is a major exploration and learning stage before about the 500th epochs. During this stage, the loss of network drops rapidly after a sharp rise. Within the same period, the sum data rate as well as total harvested energy increase rapidly while energy consumption is reduced. From Fig. 6, we can find that to improve sum data rate, the UAV learns to get more target devices into $D_{dc}$ to activate data collection. The average data rate reaches the maximum quickly. After that it declines slowly with the increase of average number of EH devices and sum EH rate, which represent average number of devices fall within $D_{eh}$ and sum of EH rate of all the EH devices at hovering stage respectively. In order to increase the total harvested energy, the agent makes a concession in the sum data rate. Besides, a sharp oscillation of training loss of critic network occurs at about 1100 epochs. After that the total harvested energy is further improved with higher energy consumption. It reveals that the agent further adjusts its control policy to find the trade-off between all the objectives. Finally, along with the oscillation convergence of the loss function, all optimization objectives are basically stable. It shows that our proposed algorithm can produce effective control policy of the UAV for the MOO problem.

For performance evaluation, we compare the policy produced by MODDPG algorithm, denoted as $P_{\text{MODDPG}}$, with two control policies that refer to [21] as

- $P_{V_{\max}}$: the UAV flies between the hovering position at the maximum speed $v = v_{\max} = 20$ m/s and hovers above the target device to collect data.
- $P_{V_{\text{ME}}}$: the UAV flies between the hovering position at the maximum-endurance speed $v = V_{\text{ME}} = 10.2$ m/s and hovers above the target device to collect data.

According to [21], $P_{V_{\max}}$ can achieve the maximization of the sum-throughput and $P_{V_{\text{ME}}}$ can achieve the minimization of the total-energy. Fig. 7 and Fig. 8 provide the comparison results on three optimal objectives and related indexes under different control policies. The coverage radius of data collection is set to 10 m, 15 m, 20 m, 25 m and 30 m. All the experimental

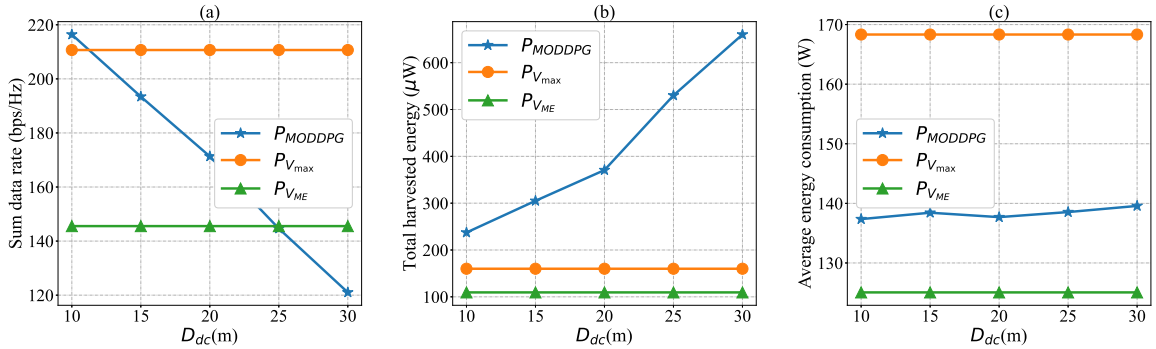


Fig. 7: Optimized objectives under different policies: (a) Sum data rate; (b) Total harvested energy; (c) Average energy consumption.

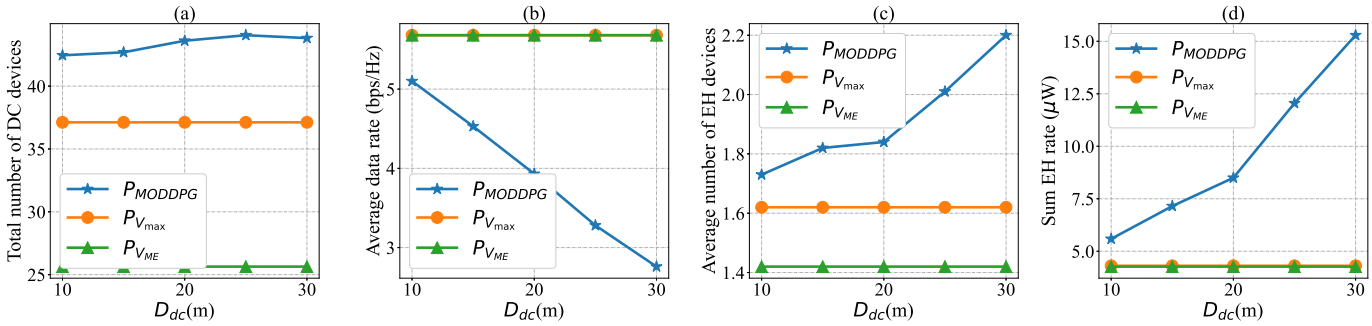


Fig. 8: Optimized results under different policies: (a) Total number of DC devices; (b) Average data rate; (c) Average number of EH devices; (d) Sum EH rate.

datas are average of 100 evaluation results. Under comparison policies $P_{V_{\max}}$ and $P_{V_{\text{ME}}}$, the UAV hovers above the target device to collect data, thus can achieve the best channel gain. As a result, their achieved average data rate reach the highest value with all $D_{\text{dc}}$. Besides, based on the adopted propulsion power consumption model, the average flying energy consumption under $P_{V_{\max}}$ and $P_{V_{\text{ME}}}$ reach the highest value and the lowest value respectively. From Fig. 7, we can find that the optimized sum data rate of our proposed $P_{\text{MODDPG}}$ policy outperforms the other two policies when $D_{dc} = 10$ m. It can be observed from Fig. 8 that $P_{\text{MODDPG}}$ achieves the highest total number of DC devices with all $D_{\text{dc}}$. Instead, the total number of DC devices under $P_{V_{\text{ME}}}$ is the least because of the low speed. Different from the compared policies that keep the same sum data rate in all cases of $D_{dc}$, the sum data rate under $P_{\text{MODDPG}}$ decreases with increasing $D_{dc}$. It is the result of the reduction of average data rate. Since we stipulate that the UAV will hover for data collection once the target device falls within $D_{dc}$, the distance between the UAV and the target device at hovering stage increases with $D_{dc}$.

In term of the performance of EH, the total harvested energy of $P_{\text{MODDPG}}$ is far higher than the other two policies with all $D_{dc}$. Besides, improvement becomes larger with increasing $D_{dc}$. As shown in Fig. 8, average number of EH devices and sum EH rate both increase with $D_{dc}$. The reason is that increasing $D_{dc}$ enhances the freedom of the control decisions of UAV. The UAV can choose its hovering position more flexibly as $D_{dc}$ increases. As a result, it can get more devices into its energy transmission coverage as well as shorten the power transmission distance to obtain higher total harvested energy of IoT devices. Instead, the information of the environment is unknown under $P_{V_{\max}}$ and $P_{V_{\text{ME}}}$ without feedback, so that these two compared policies are unable to adjust according to $D_{dc}$. Fig. 9 illustrates the UAV trajectory under $P_{\text{MODDPG}}$ and $P_{V_{\max}}$. We provide the flight path of the UAV between 5 target devices as an example to show the difference of two control policies. The uplink transmission data rate of each target device as well as sum EH rate of all the EH devices under the hover condition are shown in Fig. 9 (b) and (c). It can be observed that the UAV under $P_{\text{MODDPG}}$ chooses a hovering position that can achieve higher sum EH rate than $P_{V_{\max}}$ at the cost of decreased performance in data transmission. It reveals that the policy based on MODDPG algorithm has the advantage of achieving coordination and optimization of multiple objectives.

As for flying energy consumption, it is observed from Fig. 7(c) that the optimized value of $P_{\text{MODDPG}}$ is far less than $P_{V_{\max}}$. Compared to $P_{V_{\text{ME}}}$, $P_{\text{MODDPG}}$ achieves higher sum data rate when $D_{dc}$ is smaller than 25 m. In addition, the total harvested energy is greatly improved under $P_{\text{MODDPG}}$ after more energy is sacrificed.

In the following experiments, we verify that the optimal policies are adjusted according to weight parameters. The parameters of comparison experiments are set as Table III. To compare the influence of different weight settings on the optimal policies, in "$opt_{DC}$", "$opt_{EH}$" and "$opt_{EC}$", we respectively set the weights associated with sum data rate, total harvested energy and energy consumption to 1.0, and weights of the other two objectives to 0.0. Different from $opt_{joint}$ that considers three objectives simultaneously, in

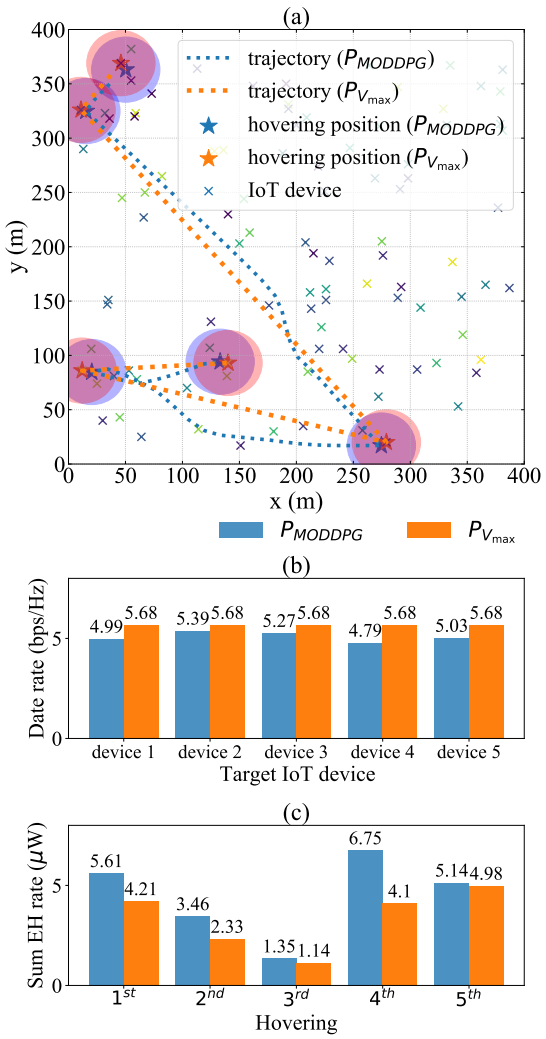


Fig. 9: Optimized results under different policies: (a) UAV trajectory; (b) Data rate; (c) Sum EH rate.

these three policies, we consider the optimization of three objectives separately. With this setting, we apply the MOO algorithm to solve SOO problem. The optimized results are given in Fig. 10 and Fig. 11. It is observed from Fig. 10 that the increasing $D_{dc}$ leads to reduced sum data rate and the improvement of total harvested energy under all polices. The reason can be found in Fig. 11. As the coverage radius of DC increases, the distance between the UAV and the target device becomes larger. On one hand, it enables the UAV to visit more devices in a mission mostly. On the other hand, it reduces the transmission data rate. In addition, the increasing $D_{dc}$ offers a wider choice scope of hovering positions. Thus more IoT devices can harvest energy from the UAV with better channel conditions.

Compared to three SOO policies, as shown in Fig. 10, the sum data rate achieved by "$opt_{joint}$" is higher than "$opt_{EH}$" and "$opt_{EC}$" in most $D_{dc}$ and only lower than "$opt_{DC}$" in all $D_{dc}$, which is consistent with the weights setting. In "$opt_{EH}$" and "$opt_{EC}$", with the weight $w_{dc}$ set to 0.0, the UAV makes flight decision without thinking about the sum data rate. Thus their data transmission performance are worse than "$opt_{joint}$",

TABLE III: Comparison experiment parameters

| Name | parameters |
|---|---|
| $opt_{joint}$ | $w_{dc} = 1.0,\ \ w_{eh} = 1.0,\ \ w_{ec} = 1.0$ |
| $opt_{DC}$ | $w_{dc} = 1.0,\ \ w_{eh} = 0.0,\ \ w_{ec} = 0.0$ |
| $opt_{EH}$ | $w_{dc} = 0.0,\ \ w_{eh} = 1.0,\ \ w_{ec} = 0.0$ |
| $opt_{EC}$ | $w_{dc} = 0.0,\ \ w_{eh} = 0.0,\ \ w_{ec} = 1.0$ |

the $w_{dc}$ of which is 1.0. In "$opt_{joint}$", with $w_{eh} = w_{ec} = 1.0$, the UAV also tries to optimize the total harvested energy and energy consumption while performing a task. As a result, the achieved sum data rate is lower than "$opt_{DC}$" that sets the sum data rate as the only goal. However, IoT devices can harvest more energy with less energy consumption of the UAV at the same time. Similarly, total harvested energy under "$opt_{joint}$" is higher than "$opt_{DC}$" and "$opt_{EC}$" in most $D_{dc}$. The performance of energy consumption under "$opt_{joint}$" is just worse than "$opt_{EC}$", far better than the other two policies. This proves that our proposed MODDPG algorithm successes to learn a control policy to simultaneously optimize multiple optimization aims.

Then we analyze the comparison results of three SOO policies. It can be observed from Fig. 10 that "$opt_{DC}$", "$opt_{EH}$" and "$opt_{EC}$" respectively achieve the highest sum data rate, the highest total harvest energy and the lowest average energy consumption with all $D_{dc}$. In "$opt_{DC}$" policy, the total number of DC devices as well as average data rate exceed other policies. The UAV learns to get more DC devices into its data collection coverage and get closer to the target devices to achieve better sum data rate. Consequently, total harvested energy of "$opt_{DC}$" is the lowest of all policies. It is because that trying to be close to the target device may have conflict with covering more EH devices under good channel gain. Since energy consumption is not concerned in this policy, it is much higher than the optimized value in "$opt_{EC}$". As for "$opt_{EH}$" policy, the total harvested energy promotes with the improvement of the number of EH devices and sum EH rate. The UAV not only learns to cover more EH devices at the hovering stage, it also hovers at a location with better channel gain. The optimization of total harvested energy also leads to more energy consumption than "$opt_{EC}$". The comparison results show that our proposed algorithm is able to produce optimal policies under different preference conditions.

## VI. Conclusion

In this paper, we investigated a MOO problem for UAV-assisted data collection and energy transfer in wireless powered IoT networks. Sum data rate, total harvested energy as well as energy consumption were optimized simultaneously. Since the IoT network was uncertain and dynamic, we proposed a MODDPG algorithm to achieve online control of the UAV. The reward function was designed as a multiple dimension vector corresponding to the optimization objectives. Thus the UAV learns to find joint optimization solution according weights parameters associated with objectives. Numerical results proved the validity of our algorithm and showed that the model can produce optimized policies under different

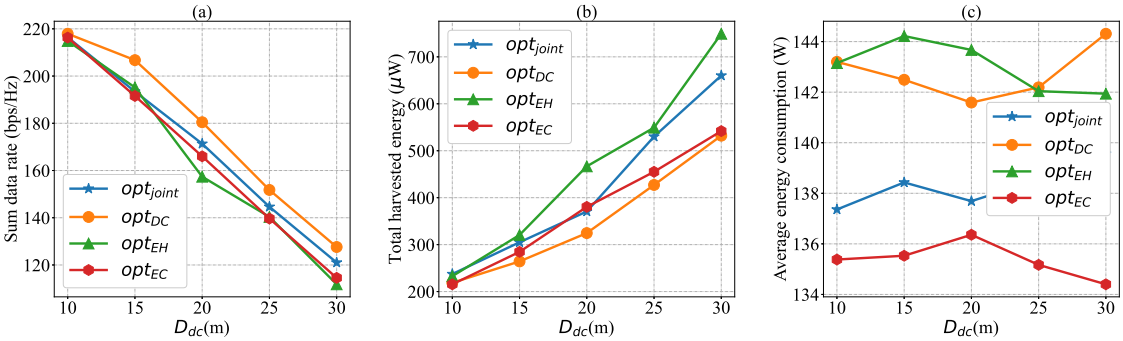


Fig. 10: Optimized objectives under different weight parameters: (a) Sum data rate; (b) Total harvested energy; (c) Average flying energy consumption.

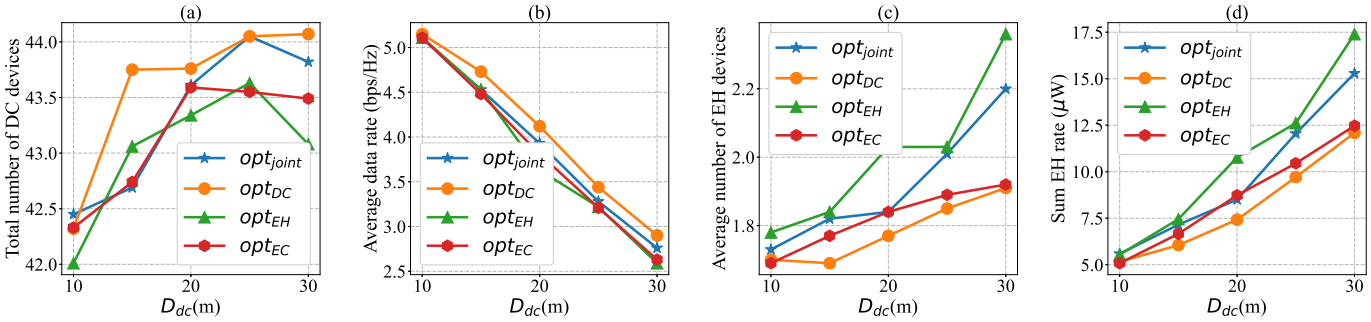


Fig. 11: Optimized results under different weight parameters: (a) Total number of DC devices; (b) Average data rate; (c) Average number of EH devices; (d) Sum EH rate.

weights. It should be noted that the proposed MOO algorithm was suitable for MOO problem with arbitrary number of objectives. In addition, it has been shown that UAV swarm has great advantages in completing complex tasks through the coordination of multiple UAVs. Therefore, it is worthy of studying the cooperative task and resource assignment, UAV-device pairing as well as multi-UAV path planning and collision avoidance for UAV swarm assisted wireless powered IoT network in the future.

## REFERENCES


[1] A. H. Mohd Aman, E. Yadegaridehkordi, Z. S. Attarbashi, R. Hassan, and Y. Park, "A Survey on Trend and Classification of Internet of Things Reviews," *IEEE Access*, vol. 8, pp. 111 763–111 782, 2020.

[2] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, and M. Mustaqim, "Internet of Things (IoT) for Next-Generation Smart Systems: A Review of Current Challenges, Future Trends and Prospects for Emerging 5G-IoT Scenarios," *IEEE Access*, vol. 8, pp. 23 022–23 040, 2020.

[3] "Cisco White Paper. Visual Networking Index: Forecast and Trends," Feb 2019. [Online]. Available: https://cyrekdigital.com/pl/blog/content-marketing-trendy-na-rok-2019/white-paper-c11-741490.pdf

[4] A. E. Kalør and P. Popovski, "Timely Monitoring of Dynamic Sources with Observations from Multiple Wireless Sensors," *arXiv e-prints*, p. arXiv:2012.12179, Dec. 2020.

[5] C. W. Tsai, C. F. Lai, M. C. Chiang, and L. T. Yang, "Data Mining for Internet of Things: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 77–97, 2014.

[6] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *2012 Proceedings IEEE INFOCOM*, 2012, pp. 2731–2735.

[7] G. Akpakwu, B. Silva, G. P. Hancke, and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," *IEEE Access*, vol. 6, no. 99, pp. 3619–3647, 2018.

[8] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When Machine Learning Meets Privacy in 6G: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2694–2724, 2020.

[9] O. L. A. Lpez, H. Alves, R. D. Souza, S. Montejo-Sanchez, E. M. G. Fernandez, and M. Latva-aho, "Massive Wireless Energy Transfer: Enabling Sustainable IoT Towards 6G Era," *IEEE Internet Things J.*, pp. 1–1, 2021.

[10] S. Bi, Y. Zeng, and R. Zhang, "Wireless powered communication networks: an overview," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 10–18, April 2016.

[11] B. Clerckx, R. Zhang, R. Schober, D. W. K. Ng, D. I. Kim, and H. V. Poor, "Fundamentals of Wireless Information and Power Transfer: From RF Energy Harvester Models to Signal and System Designs," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 4–33, 2019.

[12] C. T. Cicek, Z.-J. M. Shen, H. Gultekin, and B. Tavli, "3-D Dynamic UAV Base Station Location Problem," *arXiv preprint arXiv:2012.04909*, 2020.

[13] M. Bliss and N. Michelusi, "Adaptive Scheduling and Trajectory Design for Power-Constrained Wireless UAV Relays," *arXiv preprint arXiv:2007.01228*, 2020.

[14] H. Guo and J. Liu, "UAV-Enhanced Intelligent Offloading for Internet of Things at the Edge," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2737–2746, 2020.

[15] M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A Tutorial on UAVs for Wireless Networks: Applications, Challenges, and Open Problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 2019.

[16] B. Li, Z. Fei, and Y. Zhang, "UAV Communications for 5G and Beyond: Recent Advances and Future Trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, April 2019.

[17] S. Cho, K. Lee, B. Kang, K. Koo, and I. Joe, "Weighted Harvest-Then-Transmit: UAV-Enabled Wireless Powered Communication Networks," *IEEE Access*, vol. 6, pp. 72 212–72 224, 2018.

[18] L. Xie, J. Xu, and Y. Zeng, "Common throughput maximization for UAV-enabled interference channel with wireless powered communications," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3197–3212, 2020.

[19] J. Park, H. Lee, S. Eom, and I. Lee, "UAV-Aided Wireless Powered Communication Networks: Trajectory Optimization and Resource Allo-

cation for Minimum Throughput Maximization," *IEEE Access*, vol. 7, pp. 134 978–134 991, 2019.
[20] H.-T. Ye, X. Kang, J. Joung, and Y.-C. Liang, "Joint Uplink-and-Downlink Optimization of 3D UAV Swarm Deployment for Wireless-Powered NB-IoT Networks," *arXiv preprint arXiv:2008.02993*, 2020.
[21] H. Ye, X. Kang, J. Joung, and Y. Liang, "Optimization for Full-Duplex Rotary-Wing UAV-Enabled Wireless-Powered IoT Networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 5057–5072, 2020.
[22] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. B. Letaief, "AoI-Minimal Trajectory Planning and Data Collection in UAV-Assisted Wireless Powered IoT Networks," *IEEE Internet Things J.*, pp. 1–1, 2020.
[23] F. Wu, H. Zhang, J. Wu, L. Song, Z. Han, and H. V. Poor, "UAV-to-Device Underlay Communications: Age of Information Minimization by Multi-agent Deep Reinforcement Learning," *arXiv preprint arXiv:2003.05830*, 2020.
[24] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, 2019.
[25] F. Wu, D. Yang, L. Xiao, and L. Cuthbert, "Energy Consumption and Completion Time Tradeoff in Rotary-Wing UAV Enabled WPCN," *IEEE Access*, vol. 7, pp. 79 617–79 635, 2019.
[26] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, and A. L. Swindlehurst, "5G-and-Beyond Networks with UAVs: From Communications to Sensing and Intelligence," *arXiv preprint arXiv:2010.09317*, 2020.
[27] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, 2019.
[28] R. Sutton and A. Barto, *Reinforcement Learning:An Introduction*, 1998.
[29] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.
[30] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-Efficient UAV Control for Effective and Fair Communication Coverage: A Deep Reinforcement Learning Approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.
[31] 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," *arXiv preprint arXiv:1509.02971*, 2015.
[32] 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 Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1274–1285, 2020.
[33] B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-Based Energy-Efficient Data Collection by Unmanned Vehicles in Smart Cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1666–1676, April 2018.
[34] C. H. Liu, Z. Dai, Y. Zhao, J. Crowcroft, D. Wu, and K. K. Leung, "Distributed and Energy-Efficient Mobile Crowdsensing with Charging Stations by Deep Reinforcement Learning," *IEEE Trans. Mobile Comput.*, vol. 20, no. 1, pp. 130–146, 2021.
[35] S. Wan, J. Lu, P. Fan, and K. B. Letaief, "Towards Big data processing in IoT: Path Planning and Resource Management of UAV Base Stations in Mobile-Edge Computing System," *CoRR*, vol. abs/1906.05023, 2019. [Online]. Available: http://arxiv.org/abs/1906.05023
[36] J. Zhang, Y. Yu, Z. Wang, S. Ao, J. Tang, X. Zhang, and K. K. Wong, "Trajectory planning of uav in wireless powered iot system based on deep reinforcement learning," in *2020 IEEE/CIC International Conference on Communications in China (ICCC)*, 2020, pp. 645–650.
[37] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, 2015.
[38] C. Watkins, J. Christopher, and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
[39] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing Atari with Deep Reinforcement Learning," *CoRR*, vol. abs/1312.5602, 2013. [Online]. Available: http://arxiv.org/abs/1312.5602
[40] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic Policy Gradient Algorithms," *31st International Conference on Machine Learning, ICML 2014*, vol. 1, 06 2014.