AoI-Energy-Aware UAV-Assisted Data Collection for IoT Networks: A Deep Reinforcement Learning Method

Mengying Sun, *Student Member, IEEE*, Xiaodong Xu, *Senior Member, IEEE*, Xiaoqi Qin, *Member, IEEE*, and Ping Zhang, *Fellow, IEEE*

***Abstract*—Thanks to the inherent characteristics of flexible mobility and autonomous operation, unmanned aerial vehicles (UAVs) will inevitably be integrated into 5G/B5G cellular networks to assist remote sensing for real-time assessment and monitoring applications. Most existing UAV-assisted data collection schemes focus on optimizing energy consumption and data collection throughput, which overlook the temporal value of collected data. In this article, we employ Age of Information (AoI) as a performance metric to quantify the temporal correlation among data packets consecutively sampled by the Internet of Things (IoT) devices, and investigate an AoI-energy-aware data collection scheme for UAV-assisted IoT networks. We aim to minimize the weighted sum of expected average AoI, propulsion energy of UAV, and the transmission energy at IoT devices, by jointly optimizing the UAV flight speed, hovering locations, and bandwidth allocation for data collection. Considering the system dynamics, the optimization problem is modeled as a Markov decision process. To cope with the multidimensional action space, we develop a twin-delayed deep deterministic (TD3) policy gradient-based UAV trajectory planning algorithm (TD3-AUTP) by introducing the deep neural network (DNN) for feature extraction. Through simulation results, we demonstrate that our proposed scheme outperforms the deep Q-network and actor–critic-based algorithms in terms of achievable AoI and energy efficiency.**

***Index Terms*—Age of Information (AoI), data collection, deep reinforcement learning (RL), energy efficiency, unmanned aerial vehicle (UAV) trajectory planning.**

I. Introduction

INTERNET OF THINGS (IoT) is one of the key enablers for smart city, which continuously monitors the ambient environment and serves as precious data source for the digitization of urban space [1], [2]. This rush to digitization has resulted in a significant growth in the scale of IoT devices (IoTDs), whose number is projected to reach 500 billion by 2025 [3], generating data at an exponential rate and ever unprecedented. A massive amount of IoTDs are deployed to monitor a physical process and report the system status periodically, such as temperature, humidity, and light intensity [4]. The collected data are exploited to support real-time management applications, such as traffic control, autonomous driving [5], industrial control, etc. Under such scenarios, the freshness of collected data is of critical importance to the quality of informed decisions. Stale information could be harmful or even deleterious to the system performance. To quantify the freshness of collected data, we employ the concept of Age of Information (AoI) as the performance metric, which is defined as the elapsed time since generating the most recently received status update[6], [7]. Compared with traditional performance metrics, such as throughput and transmission delay, AoI captures the temporal correlations among consecutively sampled data packets, where a newly sampled packet is featured by a smaller value of age. Considering the limited battery capacity at IoTDs, it is of critical importance to optimize the transmission energy so that more data packets can be sampled for fresh updates [8], [9].

Nowadays, AoI has been investigated widely as a new metric to evaluate the timeliness of the collected data in the IoT networks [10]–[15], which places emphasis on the "freshness" of the data in the destination node. Rovira-Sugranes and Razi [10] investigated the AoI minimization for IoT networks and obtained the closed-form sampling rate for IoTDs. Gu *et al.* [11] analyzed the average peak AoI for IoTDs for both overlay and underlay schemes, respectively. Also, they derived simple asymptotic expressions of the average peak AoI, which could determine the superiority of underlay schemes and overlay schemes. Gu *et al.* [12] studied the status update of the IoTDs, which kept transmitting the current status update repeatedly, and they minimized the average AoI by optimizing the transmission power of IoTDs and the maximum allowable transmission times. Xu *et al.* [13] modeled the procedure of computing and transmission as a tandem queue in IoT networks and proposed a derivative-free algorithm to find the optimal updating frequency with

Manuscript received December 12, 2020; revised March 5, 2021 and April 11, 2021; accepted May 3, 2021. Date of publication May 10, 2021; date of current version December 7, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2020YFB1806905; in part by the National Natural Science Foundation of China under Grant 61871045 and Grant 61801045; in part by the BUPT Excellent Ph.D. Students Foundation under Grant CX2020213; and in part by the China Scholarship Council. *(Corresponding author: Xiaodong Xu.)*

Mengying Sun, Xiaoqi Qin, and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: smy_bupt@bupt.edu.cn; xiaoqiqin@bupt.edu.cn; pzhang@bupt.edu.cn).

Xiaodong Xu is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (email: xuxiaodong@bupt.edu.cn).

Digital Object Identifier 10.1109/JIOT.2021.3078701

minimum average peak AoI. Zhou and Saad [14], [15] optimized data sampling, device scheduling, and updating process jointly with the target of minimizing the average AoI. Some of these works analyzed the AoI performance according to the queue disciplines in theory, and some works applied the traditional methods to minimize the AoI by optimizing the resource allocation.

To combat the constrained transmission power at IoTDs and the unavailable uplink channels, unmanned aerial vehicle (UAV)-assisted remote sensing is considered as a promising solution to improve the energy efficiency of data collection [16]–[18]. Thanks to the flexible deployment of UAV in 3-D space, it can provide strong Line-of-Sight (LoS) links and small path-loss exponent to the ground IoTDs [19]–[21]. Zhang *et al.* [22] and Hu *et al.* [23] characterized the tradeoff between energy consumption and age performance for UAV-assisted data collection. A higher flight speed reduces the traveling time of UAV and, therefore, yields a smaller age, at the cost of higher propulsion energy consumption. Moreover, the selection of hovering spots of UAV determines the uplink transmission time and transmission power consumption at IoTDs. Therefore, the joint optimization of power consumption and age performance is necessary to improve the efficiency and quality of UAV-assisted data collection.

A. Related Works

In this section, we review state-of-the-art studies on AoI-energy-aware UAV-assisted data collection that is relevant to this research. These related works can be classified into two areas: 1) AoI-aware UAV-assisted data collection and 2) energy-aware UAV-assisted data collection.

1) AoI-Aware UAV-Assisted Data Collection: More and more works focused on the UAV-assisted data collection with the target of minimizing the average AoI or peak AoI of IoTDs [23]–[31]. Hu *et al.* [23] investigated that multi-UAVs executed sensing tasks through cooperative sensing and transmission to minimize the AoI. To solve the problem, they developed a compound-action actor–critic algorithm (CA2C) based on deep reinforcement learning (RL). Abd-Elmagid and Dhillon [24] investigated the UAV trajectory planning to minimize the average peak AoI by relaying the information data of IoT networks. Li *et al.* [25] considered general and heterogeneous sampling behaviors among source IoTDs, and developed a new near-optimal low-complexity scheduling algorithm to minimize the AoI. Zhou *et al.* [26] investigated AoI-based UAV trajectory planning for fresh data collection with the target of minimizing the expected AoI and developed the deep Q-network (DQN) algorithm to solve the formulated problem. Liu *et al.* [27] optimized a joint sensor node association and trajectory planning strategy jointly to strike a balance between the sensor nodes' uploading time and the UAV's flight time in various scenarios. The required time for energy harvesting, data collection, and UAV's flight time for each IoTD was jointly considered to minimize the average AoI [28], and the dynamic programming (DP) and ant colony (AC) heuristic algorithms were designed to achieve the optimal solution. Zhang *et al.* [29] formulated a joint sensing time, transmission time, UAV trajectory, and task scheduling optimization problem to minimize the AoI. The problem was decomposed into two subproblems, and an iterative algorithm was proposed to deal with the sensing time, transmission time, and UAV trajectory optimizing subproblem. In [30], based on the deep RL algorithm, the authors optimized the flight trajectory of the UAV and scheduling of status update packets jointly to achieve the minimum weighted sum of AoI of different processes. Cheung [31] minimized the AoI of all the nodes by considering the BSs' load and the UAV trajectory. The UAV-dependent cost was introduced as the sum of the selected base station's (BSs) AoI and the handover penalty. Jia *et al.* [32] proposed an AoI-aware UAV trajectory scheme by considering the energy consumption of IoTDs, and employed the DP approach to achieve the optimized strategy. Li *et al.* [33] proposed a UAV trajectory planning model for data collection with minimizing expired data packets in the whole sensor system, and then relaxed the obscure original problem into a min–max-AoI-optimal path problem, which was solved by the RL method.

2) Energy-Aware UAV-Assisted Data Collection: For UAV-assisted data collection, most researchers focus on the energy consumption of IoTDs, the propulsion energy of the UAV, data rate, and energy efficiency. Zeng *et al.* [34] proposed an efficient algorithm to optimize the hovering locations and durations, as well as the flying trajectory connecting these hovering locations. The target of this work was to minimize the UAV energy consumption, including propulsion energy and communication energy, satisfying the throughput requirement for each ground node. Zhan and Zeng [9] and Yang [35] investigated the UAV-enabled data collection system for ground nodes to achieve the tradeoff between the propulsion energy consumption and the transmission energy of the ground nodes. Li *et al.* [36] maximized the energy efficiency of the UAV by optimizing the UAV trajectory, device transmit power, and computation data scheduling, and they applied the alternating direction method of multipliers algorithm to obtain the optimal solution. However, it is not always practical to use the ground controllers to adjust the UAV flight status or preconfigure the trajectory since the network environment might be complex and changes unpredictably. Therefore, it will be an alternative way that the UAV learns to fly and collect data fully autonomously according to the current environment state.

Some researchers employed RL methods to investigate the UAV trajectory planning for data collection. Liu *et al.* [37] considered the energy efficiency and fairness among the ground nodes in the UAV-assisted data collection process, and proposed a deep RL algorithm based on deep deterministic policy gradient (DDPG) to solve the formulated problem. Liu *et al.* [38] studied the UAV-assisted data collection by considering the limited energy of UAV and coverage range to maximize the total amount of collected data. The proposed method integrated the convolutional neural network (CNN) for feature extraction and then made the decision under the guidance of the multiagent DDPG method in a fully distributed manner. Liu *et al.* [39] jointly designed the UAV trajectory and power control to maximize the instantaneous sum transmit rate

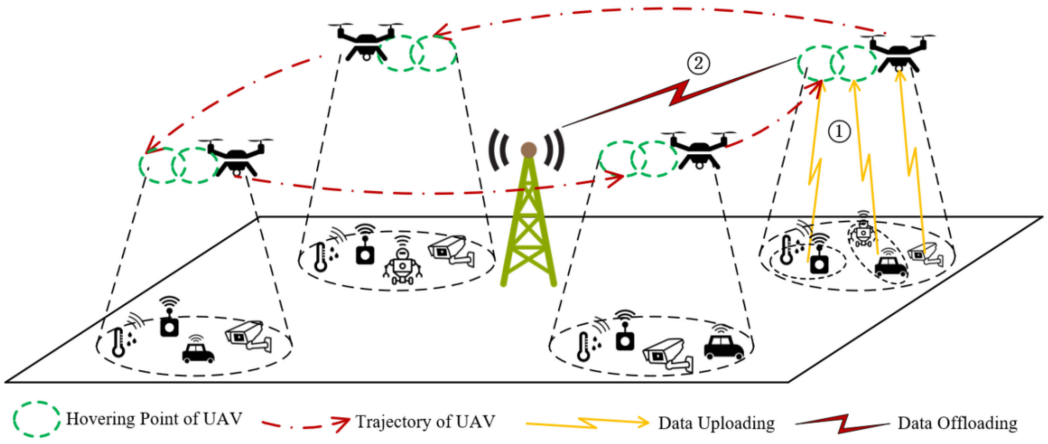


Fig. 1. UAV-assisted data collection in the IoT network.

by satisfying the rate requirement in a multi-UAV scenario. Besides, a multiagent Q-learning-based method was proposed for predicting the future UAV location based on the users' movement.

However, since the UAV propulsion energy and the transmission energy of IoTDs are also considerable aspects, few works jointly minimize the expected AoI, the energy consumption of UAV, and IoTDs for UAV-assisted data collection process. Most works applied complex optimization solutions, evolutionary algorithms, or heuristic approaches to the trajectory optimization problem, but the disadvantage is that the location and channel information are supposed to be known. This may not be suitable for the dynamic and unpredictable environment, as well as the continuous optimizing process. Nowadays, due to the new research tendency of using artificial intelligence (AI), precisely, RL came out as a new strategy that can grant the UAV sufficient intelligence to make local decisions to accomplish missions. RL can be utilized to train autonomous agents in a semisupervised manner to operate independently as a bottom-up alternative to the centralized systems. With the RL algorithm and training model, the UAV automatically and rapidly optimizes the trajectory by interacting with the environment.

B. Contributions

In this article, we investigate a UAV-assisted AoI-energy-aware data collection problem for a group of IoTDs deployed at a given geographic area. We aim to minimize the weighted sum of expected AoI, propulsion energy consumption of UAV, and transmission energy consumption at IoTDs. The optimization problem is a formulated subject to constraints of the energy consumption, transmission channel conditions, as well as the tolerable maximum AoI of IoTDs. Considering the unpredictable system dynamics, we reformulate our problem as a Markov decision process (MDP) and propose a deep RL-based algorithm to find the optimal solution for flight speed and hovering spot for UAV, and transmission scheduling for IoTDs to minimize the weighted sum of the expected AoI and energy consumption of IoTDs and UAV.

The main contributions of this article are summarized as follows.

1) We formulate an optimization problem to minimize the weighted sum of AoI and energy consumption of UAV as well as IoTDs by jointly considering the UAV speed, flight distance, as well as direction, the scheduled IoTDs, and bandwidth allocation. The formulated problem is subject to the maximum UAV speed, total bandwidth, and the maximum number of IoTDs scheduled by the UAV at one time.
2) To efficiently solve the minimization problem, we transform the proposed problem as an MDP. Due to the huge state space and high-dimensional continuous action in the problem, we leverage a deep neural network (DNN) to characterize the state as well as action spaces and propose a TD3-based AoI-energy-aware UAV trajectory planning algorithm (TD3-AUTP) based on the DDPG framework to obtain the optimal solutions for the flight speed, distance, and direction of the UAV, the IoTD scheduling, and channel allocation for IoTDs.
3) The proposed scheme is evaluated by comparing it with other existed works. Through simulation results, we verify that the proposed scheme can decrease the expected AoI, energy consumption, and average steps in one round compared with the existed works. It also demonstrates the necessity of optimizing the UAV trajectory, IoTD scheduling, and channel allocation jointly to enhance the performance of the UAV-assisted data collection.

The remainder of this article is organized as follows. We present the system model and assumptions in Section II. The problem reformulation and the proposed AoI-energy-aware UAV-assisted data collection scheme with the TD3 algorithm are provided in Section III. The simulation results and analyses are given in Section IV. Finally, we conclude this article in Section V.

II. System Model

As illustrated in Fig. 1, in a UAV-assisted IoT network, IoTDs need to upload the sampling data to the network continually; however, IoTDs cannot transmit the data to the remote BS directly due to the limited energy and unavailable channel

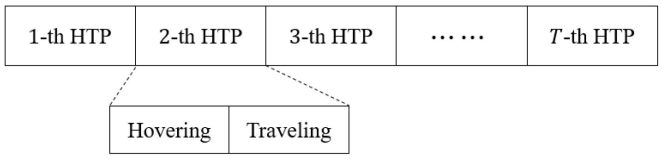


Fig. 2. Framework structure for UAV trajectory.

conditions. A UAV is dispatched as a mobile data collector to gather the data of IoTDs and then offload the data to the remote BS. Let $\mathcal{M} = \{1, 2, \ldots, M\}$ be the set of M IoTDs. The data size of IoTD $m \in \mathcal{M}$ is denoted as χ_m. The UAV works in the half duplex mode, which means that each node cannot transmit and receive data at the same time slot. Thus, the data collection (transmission from IoTD to UAV) and data forwarding (transmission from UAV to BS) do not coexist in the coherent interval.

To guarantee the stability of data reception for the UAV, the UAV hovers in the air when collecting data. As shown in Figs. 1 and 2, there exist two stages for the UAV in each hovering–traveling process (HTP), i.e., *h*overing and *t*raveling. In the first stage, the scheduled IoTDs upload the collected data to the UAV and the UAV is in the hovering status (marked as ① in Fig. 1). In the second stage, the UAV offloads the collected data to the BS before arriving at the next hovering spot, and the UAV is in the traveling status (marked as ② in Fig. 1). For energy saving, IoTDs keep monitoring the signal from the UAV until receiving the signal, and then IoTDs upload the data to the UAV. At the start of the hovering status, the UAV informs IoTDs to sample and upload the data to the UAV. Note that the data sampling time is so small that can be ignored.

A. Communication Model

We assume a time-slotted system and there are T unequal time slots, which are regarded as T HTPs in the time horizontal, and one HTP includes two parts, i.e., hovering duration and traveling duration. Let $\Gamma_m^t = [x_m^t, y_m^t] \in \mathbb{R}^{1\times 2}$ be the horizontal location of IoTD $m \in \mathcal{M}$ in the ground in the tth HTP, where $t \in \{1, 2, \ldots, T\}$. Assume that the UAV flies at a fixed altitude H, and the horizontal location of the UAV in the tth HTP is denoted as $\mathbf{q}^t = [x^t, y^t] \in \mathbb{R}^{1\times 2}$. In the hovering statues of the tth HTP, the distance between the UAV and IoTD m in the ground is denoted as $d_m^t = \sqrt{H^2 + \|\mathbf{q}^t - \Gamma_m^t\|^2}$. Let h_m^t be the channel coefficient between the UAV and IoTD m

$$h_m^t = \sqrt{\beta_m^t}\tilde{h}_m^t \tag{1}$$

where β_m^t means the large-scale fading effect, such as path loss and shadowing, and $\tilde{h}_m^t$ accounts for the small-scale fading, which is a complex-variable random variable satisfying the condition $\mathbb{E}\{|\tilde{h}_m^t|^2\} = 1$.

Note that in such practical scenarios, the UAV may not have any additional information about the heights, exact locations, number of the obstacles, as well as the channel conditions. Therefore, one must consider the randomness associated with the LoS and non-LoS (NLoS) links while designing the UAV-based communication system. For ground-to-air communications, each IoTD will typically have a LoS view toward a specific UAV with a given probability. This LoS probability depends on the environment, location of the device and the UAV, as well as the elevation angle. We introduce a common approach to model the LOS probability between the UAV and IoTD m in the tth HTP, denoted as

$$P_{m,\mathrm{LoS}}^t = \frac{1}{1 + C\exp\left(-D\left[\theta_m^t - C\right]\right)} \tag{2}$$

where C and D mean the propagation parameters and constant values, which depend on the carrier frequency and type of environment, such as rural, urban, or dense urban. $\theta_m^t = (180/\pi)\sin^{-1}((H/d_m^t))$ represents the elevation angle in degree. Naturally, the NLoS probability is given by $P_{m,\mathrm{NLoS}}^t = 1 - P_{m,\mathrm{LoS}}^t$. According to (2), we can see that the increasing elevation angle brings the larger LoS probability. Furthermore, we introduce the UAV-device channel gain β_m^t as [40], [41]

$$\beta_m^t = \begin{cases} \frac{1}{\eta_1}\left(\frac{4\pi f_c d_m^t}{c}\right)^{-\alpha}, & \mathrm{LoS} \\ \frac{1}{\eta_2}\left(\frac{4\pi f_c d_m^t}{c}\right)^{-\alpha}, & \mathrm{NLoS} \end{cases} \tag{3}$$

where f_c is the carrier frequency, α is the path-loss exponent, and η_1 and η_2 ($\eta_2 > \eta_1 > 1$) are the excessive path-loss coefficients in LoS and NLoS cases, respectively. c is the speed of light. Furthermore, the channel gain β_m^t depending on the LoS and NLoS can be reformulated as

$$\begin{aligned}\beta_m^t &= P_{m,\mathrm{LoS}}^t \frac{1}{\eta_1}\beta_0\left(d_m^t\right)^{-\alpha} + \left(1 - P_{m,\mathrm{LoS}}^t\right)\frac{1}{\eta_2}\beta_0\left(d_m^t\right)^{-\alpha} \\ &= \hat{P}_{m,\mathrm{LoS}}^t \beta_0\left(d_m^t\right)^{-\alpha}\end{aligned} \tag{4}$$

where $\hat{P}_{m,\mathrm{LoS}}^t = P_{m,\mathrm{LoS}}^t(1/\eta_1) + (1 - P_{m,\mathrm{LoS}}^t)(1/\eta_2)$ is recognized as the regularized LoS probability. $\beta_0 = ([4\pi f_c]/c)^{-\alpha}$ means the channel gain when the distance is set as 1 m.

Let p_m^t be the transmission power of IoTD m in the tth HTP, and the data rate between the UAV and IoTD m can be expressed as

$$\mathcal{R}_m^t = b_m^t c_m^t = b_m^t \log_2\left(1 + \frac{p_m^t\left|h_m^t\right|^2}{\sigma^2}\right) \tag{5}$$

where b_m^t and c_m^t mean the assigned bandwidth for IoTD m and the achievable spectral efficiency, respectively. σ^2 denotes the additive white Gaussian noise power at the receiver. Similarly, the received data rate of the BS from the UAV can be expressed as

$$\mathcal{R}_{\mathrm{uav}}^t = b_{\mathrm{uav}} \log_2\left(1 + \frac{p_{\mathrm{uav}}^t\left|h_{\mathrm{uav}}^t\right|^2}{\sigma^2}\right) \tag{6}$$

where b_{uav} and p_{uav}^t mean the bandwidth and transmit power of the link between UAV and BS, respectively. Then, we can obtain the latency of data transmission from IoTD m to the UAV as $D_m^{\mathrm{up},t} = (\chi_m/\mathcal{R}_m^t)$, and the uploading latency can be expressed as

$$D^{\mathrm{up},t} = \max_{m\in\mathcal{M}}\left\{I_m^t \frac{\chi_m}{\mathcal{R}_m^t}\right\} \tag{7}$$

where binary I_m^t indicates whether IoTD m is scheduled in the tth HTP, which is denoted as

$$I_m^t = \begin{cases} 1, & \text{if IoTD } m \text{ is scheduled in the } t\text{th HTP.} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

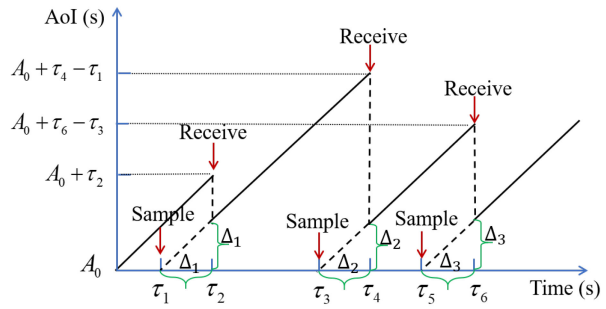


Fig. 3. Achieved AoI as the time goes by.

The latency of offloading collected data to the BS from the UAV can be expressed as

$$D^{\text{off},t} = \frac{\sum_{m\in\mathcal{M}} \chi_m}{\mathcal{R}^t_{\text{uav}}}. \tag{9}$$

B. Age of Information

We introduce the AoI as the key performance metric to characterize the timeliness of the sampling information updates, which is defined as the time elapsed since the most recently received update was generated [42]. At any time τ, let $o(\tau)$ be the generating time of the sampling data, which has been received by the BS; thus, the AoI can be defined as $A^\tau = \tau - o(\tau)$. As illustrated in Fig. 3, the AoI starts from the initial age A_0, and increases with time at rate 1 until an update packet is received. At time τ_1, the data are sampled by each IoTD, at time τ_2, the sampled data are received by BS, and the previous data are replaced. Thus, the AoI in the BS is updated as $\Delta_1 = \tau_2 - \tau_1$. The process repeats.

We evaluate the AoI of IoTD m in the end of the tth HTP, which means the time when the UAV arrives at the next hovering spot exactly

$$A_{m,t} = \begin{cases} D^{\text{up},t} + D^{\text{fly},t}, & \text{if } I^t_m = 1 \\ A_{m,t-1} + D^{\text{up},t} + D^{\text{fly},t}, & \text{otherwise} \end{cases} \tag{10}$$

where $D^{\text{up},t}$ means the latency for data collection and $D^{\text{fly},t}$ indicates the flight latency of the UAV from the last hovering spot to the next one. It is noted that the collected data of the UAV are assumed to be offloaded to the BS before the UAV arrives at the next hovering spot, i.e., $D^{\text{off},t} \le D^{\text{fly},t}$, which can be guaranteed by adjusting the speed of the UAV. In practice, the offloading latency is lower than the traveling latency. The achieved average AoI of all IoTDs at the end in the tth HTP is expressed as

$$A^t = \frac{1}{M}\sum_{m=1}^{M} A_{m,t}. \tag{11}$$

C. Energy Consumption

The energy consumption consists of two parts: 1) the transmission energy consumption of IoTDs and 2) the energy consumption of the UAV.

1) The energy consumption of IoTD m can be expressed as

$$E^t_m = p^t_m \frac{\chi^t_m}{\mathcal{R}^t_m} = p^t_m \frac{\chi^t_m}{b^t_m \log_2\left(1 + \frac{p^t_m |h^t_m|^2}{\sigma^2_{\text{uav}}}\right)} \tag{12}$$

and the total energy consumption of IoTDs is expressed as

$$E^t_{\text{IoT}} = \sum_{m=1}^{M} E^t_m. \tag{13}$$

2) The energy consumption of the UAV can be expressed as [34]

$$E^{t,\text{tra}}_{\text{UAV}} = \frac{l^t}{V}\left(P_0\left(1+\frac{3V^2}{U^2_{\text{tip}}}\right) + \frac{P_i \nu_0}{V} + \frac{1}{2} d_0 \rho s \Lambda V^3\right) \tag{14}$$

$$E^{t,\text{hov}}_{\text{UAV}} = P_h D^{t,\text{hov}} \tag{15}$$

where $E^{t,\text{tra}}_{\text{UAV}}$ and $E^{t,\text{hov}}_{\text{UAV}}$ indicate the traveling energy consumption and hovering energy consumption. V is the flight speed of the UAV and (l^t/V) means the traveling time.

U_{tip} indicates the tip speed of the rotor blade, and ν_0 is known as the mean rotor induced velocity in the hovering status. d_0 and s are the fuselage drag ratio and rotor solidity, respectively. ρ and Λ mean the air density and rotor disc area, respectively. P_0, P_i, and P_h are constants representing the blade profile power, induced power in the hovering status, and hovering energy consumption, respectively. $D^{t,\text{hov}}$ is the hovering duration in the tth HTP. Refer to [34] for more detailed information. For the tth HTP, considering that the UAV traveling energy is generated in the last HTP, the energy consumption of the UAV can be expressed as $E^t_{\text{UAV}} = E^{t-1,\text{tra}}_{\text{UAV}} + E^{t,\text{hov}}_{\text{UAV}}$.

D. Problem Formulation

Let tuple $[l^t, \theta^t, V^t]$ be the traveling decision of the UAV in the tth HTP, where l^t, θ^t, and V^t denote the flight distance, flight direction, and flight speed, respectively. Since our work aims to minimize the weighted sum of AoI and the energy consumption, the network utility can be expressed as

$$U = \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\left[-A^t\right] + \zeta \mathrm{E}\left[-E^t_{IoTD}\right] + \omega \mathrm{E}\left[-E^t_{\text{UAV}}\right] \tag{16}$$

where ζ and ω represent the weight factors about transmission energy of IoTDs and the UAV propulsion energy, respectively. E[·] indicates the expected operation over the finite number of HTPs.

We determine the weight factors by evaluating the proportions of the estimated achievable AoI, energy of IoTDs, and propulsion energy of the UAV. We find the upper bound of each weight factor when one of the three aspects is completely emphasized and the other factors are negligible. Then, we adjust the weight factors based on the upper bounds and adopt one group of the weight factors according to the test experiences as well as the preferences on the three aspects.

Our problem is formulated as

$$\begin{aligned}\mathbf{P}: \max_{\{l^t,\theta^t,V^t,b_m^t\}} \quad & U \\ \text{s.t.} \quad & c1: V^t \le v^{\max} \\ & c2: x_{\min} \le x^t \le x_{\max}, y_{\min} \le y^t \le y_{\max}, \\ & c3: \sum_{m=1}^{M} b_m^t \le B \\ & c4: \sum_{m=1}^{M} I_m^t \le \delta \end{aligned} \tag{17}$$

where $c1$ means the speed of the UAV is limited. $c2$ represents the trajectory area of the UAV is bounded. The total bandwidth allocated to IoTDs is B shown in $c3$. $c4$ means that the number of IoTDs scheduled in parallel is no more than δ.

E. Problem Analysis

Lemma 1: To achieve the minimized expected AoI of all IoTDs, the UAV needs to serve all IoTDs in one flight round and the UAV flies for data collection in the given geographic area continuously.

Proof: Since the time of a flight round is larger than the uplink transmission time, an IoTD will wait for a long time to be served again if it is not be scheduled in this flight round; therefore, the AoI of this IoTD will be accumulated. One flight round indicates the process from that the first IoTD is scheduled to the condition that all IoTDs are served. When a flight round ends, the next flight round starts and all IoTDs are in unserved status. Thus, our problem can be transformed to a trajectory planning problem in one flight round. ■

In one flight round, the hovering spot is optimized since the number of IoTDs that the UAV can schedule is limited. Therefore, two conditions should be considered. The first condition is that the UAV flies close to IoTDs because this can provide a high uplink transmission rate, but there is a disadvantage that the number of IoTDs served simultaneously can be limited. The second condition is that the UAV selects a hovering spot where the UAV can serve as most IoTDs as possible and the uplink transmission rates are optimized under this condition. However, the disadvantages of this condition are that the uplink transmission energy consumption for the scheduled IoTDs is relatively high and the uploading latency for each IoTD increases. Besides, the flight distance and speed should be taken into consideration. Based on Lemma 1, the sensing area of UAV is divided into several subareas and all subareas can be served one by one. The UAV flies out a subarea only when all IoTDs in this subarea are scheduled from the moment it flies in. The problem can be transformed to optimize the utility of each subarea. This way can effectively avoid that the UAV flies back to serve IoTDs, which were scheduled just now and improve training efficiency and performance.

In our work, the UAV spot optimization and IoTD scheduling affect the energy consumption. When the hovering spot and the scheduled IoTDs are determined, the AoI of all IoTDs will be stressed. Since the transmission latency in any HTP can directly affect the AoI of all IoTDs, we optimize the bandwidth allocation to achieve the minimized transmission latency.

Lemma 2: When the transmission latencies of IoTDs are equal, the minimum AoI will be obtained. The optimal bandwidth allocation for IoTD i can be achieved as

$$b_i^t = \frac{B}{1 + \sum_{i,j\in\mathbb{I}^t, j\neq i} \frac{c_i^t \chi_j^t}{\chi_i^t c_j^t}} \ \forall i \in \mathbb{I}^t \tag{18}$$

where $\mathbb{I}^t$ means the set of IoTDs that are scheduled in the tth HTP.

Proof: For the bandwidth allocation, the problem is reformulated as

$$\min_{\{b_i^t\}} \ \max_{i\in\mathbb{I}^t} \frac{\chi_i^t}{b_i^t c_i^t} \tag{19}$$

$$\text{s.t.} \ \sum_{i\in\mathbb{I}^t} b_i^t \le B. \tag{20}$$

From problem (19), we can infer that the optimal solution lies on the condition that the achievable transmission latencies of the scheduled IoTDs are equal. Assume that the latencies of two IoTDs are different, i.e., $[\chi_i^t/(b_i^t c_i^t)] < [\chi_j^t/(b_j^t c_j^t)]$, the achievable solution is $[\chi_j^t/(b_j^t c_j^t)]$. However, by adjusting the bandwidth b_i^t and b_j^t, we have $[\chi_i^t/((b_i^t - \Delta)c_i^t)] = [\chi_j^t/((b_j^t + \Delta)c_j^t)] < [\chi_j^t/(b_j^t c_j^t)]$, and obtain that $[\chi_j^t/(b_j^t c_j^t)]$ is not the optimal solution. Therefore, the optimal solution is achieved when the transmission latencies are equal.

Based on the above analysis, the optimal solution lies on the condition

$$\frac{\chi_i^t}{b_i^t c_i^t} = \frac{\chi_j^t}{b_j^t c_j^t} \Rightarrow b_j^t = \frac{c_i^t \chi_j^t}{\chi_i^t c_j^t} b_i^t. \tag{21}$$

Aligning with condition (20), we can obtain the optimal solution as (18). ■

III. TD3-Based UAV-Assisted Data Collection Algorithm

In this section, the formulated minimization problem about the AoI-energy-aware UAV-assisted data collection is reformulated as a discrete-time MDP with an infinite horizon discounted criterion, and then we discuss the general solution.

A. MDP Formulation for UAV-Assisted Data Collection

The problem ($\mathbf{P}$) is reformulated as an MDP with the target of maximizing the utility related to the expected AoI and energy consumption. Generally, the MDP is marked as the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, P \rangle$, which models an agent's sequential decision-making process and includes four elements. $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, and P represent the state set, action set, output reward, and state transition probability, respectively, where $\mathcal{R}$ is defined as $\mathcal{S} \times \mathcal{A} \to \mathcal{R}$ and the state transition probability is defined as $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to P$. The UAV is regarded as the agent of the formulated MDP. The state set, action set, and reward function are provided as follows.

1) *State:* During the tth HTP, the local state of the UAV can be characterized as $s^t = \langle \mathbf{q}^t, \mathbf{A}^t, \mathbf{C} \rangle$, where the first part $\mathbf{q}^t = [x^t, y^t]^T$ means the location of the UAV, $\mathbf{A}^t = \{A_1^t, A_2^t, \ldots, A_M^t\}$ means the current AoI of all

IoTDs, and $\mathbf{C}^t = \{C_1^t, C_2^t, \ldots, C_M^t\}$ indicates the coverage condition indicator. Binary variable $C_m^t = 1$ means IoTD m has been served in the tth HTP of the current flight round, *vice versa*.

2) *Action:* During the tth HTP, the action of the UAV is defined as $a^t = \langle l^t, \theta^t, V^t \rangle$, where l^t means the flight distance with guaranteeing the condition $c2$, and $\theta^t \in (0, 2\pi]$ indicates the UAV flight direction.
3) *Reward:* During the tth HTP, the reward is related to the average AoI, energy consumption of IoTDs, and the UAV propulsion energy, which is expressed specifically as

$$r^t = -A^t - \zeta * E_{IoTD}^t - \omega * E_{\text{UAV}}^t - \hbar^{\text{penalty}} + \hbar^{\text{award}} \tag{22}$$

where positive constant $\hbar^{\text{penalty}}$ indicates that the UAV will receive a penalty when the UAV cannot serve IoTDs which have not been scheduled in this flight round, and the positive constant $\hbar^{\text{award}}$ means that the UAV will receive an award when the UAV covers all IoTDs of one subarea. Both parameters are enabled to improve the performance of the learning process by avoiding the UAV to hover in the spot with no IoTDs to serve and stimulating the UAV to serve all IoTDs within fewer steps, respectively.
4) *State Transition Probability:* The state transition probability is defined as the probability of next state s^{t+1} when the agent task action a^t in state s^t, which is marked as $P(s^{t+1}|s^t, a^t.)$.

Let $\pi = (\pi_l, \pi_\theta, \pi_V)$ be the stationary policy of the UAV traveling, where π_l, π_θ, and π_V denote the UAV traveling distance policy, traveling direction policy, and traveling speed policy, respectively. When deploying π^t, the UAV observes state s^t at in the tth HTP, and make decisions for traveling distance, direction, and speed, that is, $\pi^t(s^t) = (\pi_l(s^t), \pi_\theta(s^t), \pi_V(s^t))$.

Based on the assumptions on the UAV mobility and the AoI evolutions, the randomness lying in a sequence of the states over the time horizon $\{s^t : t \in \mathbb{N}^+\}$ can be easily verified as Markovian with the following controlled state transition probability:

$$\mathbb{P}\left(s^{t+1}\middle|s^t, \pi\left(s^t\right)\right) = \mathbb{P}\left(x^{t+1}, y^{t+1}\middle|x^t, y^t\right) \cdot \mathbb{P}\left(A^{t+1}\middle|x^{t+1}, y^{t+1}\right) \tag{23}$$

where $\mathbb{P}$ denotes the probability of an event occurrence. With the given stationary control policy π and the initial state $s^1 \in \mathcal{S}$, the expected long-term discounted utility function can be expressed as

$$\mathbb{V}(s, \pi) = E_\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r^t\left(s^t, a^t\right)\middle|s^1 = s\right] \tag{24}$$

where $\gamma \in (0, 1]$ means the discount factor and γ^{t-1} denotes the discount factor for the former $(t-1)$ state transmission processes. Following the control policy π over the HTPs, $E_\pi[\cdot]$ denotes the expectation over different decision makings under different states following a control policy π across the HTPs. Equation (24) can be considered as the optimization function for the UAV. The AoI-energy-aware UAV data collection problem can be reformulated as a single-agent MDP, and (24) can be rewritten as

$$\mathbb{V}(s, \pi) = E_\pi[r(s, a)] + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s'|s, \pi(s)\right)\mathbb{V}\left(s', \pi\right). \tag{25}$$

The optimal policy can be derived based on the Bellman optimally equation

$$\mathbb{V}^*(s, \pi) = \mathbb{V}\left(s, \pi^*\right) = \max_{a \in \mathcal{A}}\left\{\mathrm{E}_\pi[r(s, a)] + \gamma \sum_{s' \in \mathcal{S}} P\left(s'|s, \pi(s)\right)\mathbb{V}\left(s', \pi\right)\right\}. \tag{26}$$

Furthermore, we introduce the definition of the right-side hand of (26)

$$Q^*(s, a) = Q^{\pi^*}(s, a) = E_\pi[r(s, a)] + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s'|s, \pi(s)\right)\mathbb{V}^*\left(s', \pi\right). \tag{27}$$

Then, we have

$$\mathbb{V}^*(s, \pi) = \max_{a \in \mathcal{A}} Q^*(s, a). \tag{28}$$

The optimal value function $\mathbb{V}^*$ can be obtained from $Q^*(s, a)$, which is expressed as

$$Q^*(s, a) = E_\pi[r(s, a)] + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}\left(s'|s, \pi(s)\right) \max_{b \in \mathcal{A}} Q^*\left(s', b\right) \tag{29}$$

where s and a are the current state and action, respectively. s' means the produced state when taking action a. Based on information tuple $\langle s, a, r, s' \rangle$, optimal Q function $Q^*(s, a)$ can be achieved in a recursive way. The updates of the Q function can be expressed as

$$Q^{t+1}(s, a) = (1 - \alpha)Q^t(s, a) + \alpha\left[r^t + \gamma \max_{b \in \mathcal{A}} Q^t\left(s', b\right)\right] \tag{30}$$

where $\alpha \in (0, 1]$ denotes the learning rate. It has been proven that the Q-learning process converges and eventually finds the optimal policy, and three conditions should be satisfied: 1) the state transition probability under the optimal stationary control policy is stationary; 2) $\sum_{t=1}^{\infty} \alpha^t$ is infinite and $\sum_{t=1}^{\infty} (\alpha^t)^2$ is finite; and 3) the state–action pairs are visited infinitely often.

B. TD3 Algorithm Based on DDPG Framework

In this work, we consider a model-free deep RL algorithm since the agent cannot make decisions according to the uncertain environment. Moreover, the environment state space $\mathcal{S}$ and action space $\mathcal{A}$ are infinite. Thus, we should adopt the policy-based and model-free deep RL algorithms, such as policy gradient, actor–critic, deterministic policy gradient (DPG), and DDPG [43]. DDPG is an improved actor–critic algorithm, which combines the advantages of policy gradient and DQN algorithms. The DQN algorithm is developed from the Q-learning algorithm by introducing the DNN. However, a disadvantage exists in the DDPG algorithm, which is that

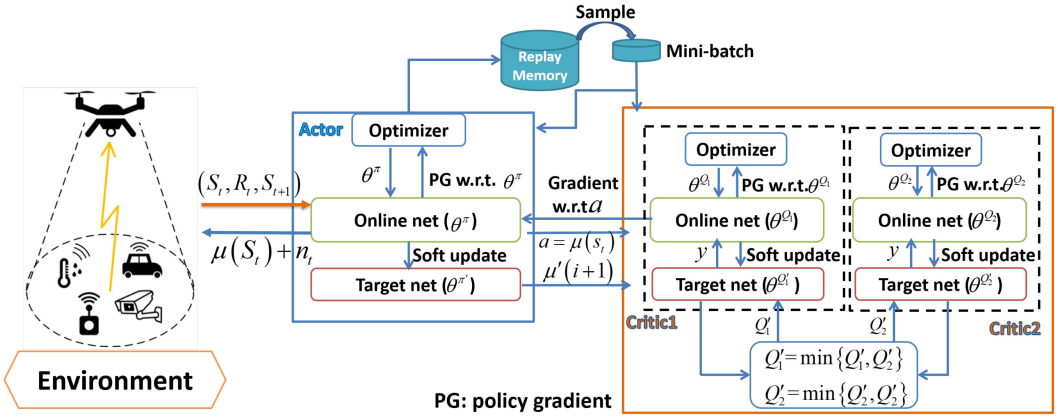


Fig. 4. Structure of TD3 with AoI-energy-aware UAV-assisted date collection.

the Q-value is always overestimated so that the learned policy is invalid. The TD3 can effectively solve this issue in the DDPG by introducing clipped double critic networks as shown in Fig. 4.

In the TD3 algorithm, the action space is formulated by setting the bounds for each action. In every decision process, the actions are selected within the bounds. In the DQN algorithm, one action is selected from a fixed action space, while the TD3 algorithm is good at dealing with the problem with multidimensional variables and the multiple actions are bounded together for optimization. Based on the current state, we can derive a probability distribution with multiple actions by the minibatch sampling and neural networks estimation. According to the largest probability, the solution with actions can be obtained.

In the structure of the TD3, there are two types of networks, including one actor network and two critic networks. Each actor net and critic net contain two subnets: 1) online net and 2) target net. These six neural networks consist of various layers, and all layers contain their corresponding parameters. Actor net parameters $\boldsymbol{\vartheta} = \{\vartheta^{\pi}, \vartheta^{\pi'}\}$ denote the online parameter and target net parameter, respectively. Critic net parameters $\boldsymbol{\theta} = \{\theta^{Q_1}, \theta^{Q_2}, \theta^{Q'_1}, \theta^{Q'_2}\}$ mean the online net and target net parameters of critic net 1, as well as the online net and target net parameters of critic net 2. The TD3 uses a reply memory $\mathbb{D}$ to store the experiences generated from the training process. Besides, the random experience selection breaks the correlation among the experiences in minibatch $\mathbb{I}$. If the memory is already full, the oldest experiences will be removed to make room for the latest one. On the basis of the TD3, we show the three advantages of TD3 algorithm as follows.

1) *Clipped Double-Q Learning:* Compared with DDPG, there are two critic nets. By choosing the minimum Q-value of the target nets, the target network can be well constructed.
2) *Delayed Policy Updates:* The updating of the parameter in the target net is slower than that of the online net.
3) *Target Policy Smoothing:* Some noise generated in the algorithm is added to the action, which makes the Q function changes with all the actions smoothly. Then, the policy is not easy to be suffered from the error of the Q function.

C. Deep Critic-Networks Training and Update

In the TD3 algorithm, the critic network aims to simulate the real Q-table without the constrains of dimensionality. As aforementioned, DQN is a variant of Q-learning, which applies a DNN as the function approximator, and DDPG takes the advantage of DQN. The critic net is trained by minimizing the loss function as follows:

$$L\left(\theta^{Q_i}\right) = E\left(y^t - Q_i\left(s^t, a^t \middle| \theta^{Q_i}\right)\right)^2, i = 1, 2 \tag{31}$$

where $Q_i(s^t, a^t|\theta^{Q_i}.) \approx Q_i(s^t, a^t)$, and θ^{Q_i} is parameter of the DNN in the online net of the ith critic net and y^t is target value, which can be estimated by

$$y^t = r^t\left(s^t, a^t\right) + \gamma \min_{i=1,2} Q_i'\left(s^t, a^t \middle| \theta^{Q_i'}\right) \tag{32}$$

where $Q_i'(s^t, a^t|\theta^{Q_i'}.)$ means the Q-function in the target net of the critic net, which has the same structure of the online net.

The experience reply mechanism is the characteristic of the TD3 algorithm, where a reply memory buffer $\mathbb{D}$ is utilized to store experience tuple $\langle s^t, a^t, r^t, s^{t+1}\rangle$. In the training process, minibatch $\mathbb{I}$ is sampled from reply memory $\mathbb{D}$ to update the parameters of the nets. Based on the policy gradient methods, θ^{Q_i} is updated as

$$\begin{aligned}\nabla_{\theta^{Q_i}} L\left(\theta^{Q_i}\right) &= -E_{s^t}\Big[\left(y_j - Q_i\left(s_j, a_j \middle| \theta^{Q_i}\right)\right) \\ &\quad \times \nabla_{\theta^{Q_i}} Q_i\left(s_j, a_j \middle| \theta^{Q_i}\right)\Big] \\ &= -\frac{1}{|\mathbb{I}|}\sum\nolimits_{j\in\mathbb{I}}\left(y_j - Q_i\left(s_j, a_j \middle| \theta^{Q_i}\right)\right) \\ &\quad \times \nabla_{\theta^{Q_i}} Q_i\left(s_j, a_j \middle| \theta^{Q_i}\right) \\ &\quad \forall i = 1, 2 \end{aligned} \tag{33}$$

$$\theta^{Q_i} := \theta^{Q_i} - \alpha_c \nabla_{\theta^{Q_i}} L\left(\theta^{Q_i}\right), \forall i = 1, 2 \tag{34}$$

where α_c is a positive constant reflecting the learning rate for the value function evaluation.

Algorithm 1 TD3-AUTP

1: **Initialization:** discount factor γ, update rate υ, penalty $\hbar^{\text{penalty}}$, award $\hbar^{\text{award}}$. For the UAV agent, the online net parameters of Actor net and two Critic nets: ϑ^{π}, θ^{Q_1} and θ^{Q_2}; the target net parameters is the copy of corresponding online nets: $\vartheta^{\pi'}$, $\theta^{Q'_1}$ and $\theta^{Q'_2}$; replay memory $\mathbb{D}$, minibatch $\mathbb{I}$, $\mathbb{I} \subseteq \mathbb{D}$; the maximum number of episodes $E_{\max}$; the maximum steps in each episode $T_{\max}$.
2: **for** episode = 1 to $E_{\max}$ **do**
3: reset the state s^0, the reward is reset as 0.
4: **for** step = 1 to $T_{\max}$ **do**
5: Choose action a_t based on the policy $a^t = \mu(s^t) + n_t$, and obtain the reward $r^t\left(s^t, a^t\right)$ and the next observation s^{t+1}. Store the tuple $\left\langle s^t, a^t, r^t, s^{t+1}\right\rangle$ into the replay memory $\mathbb{D}$.
6: Sample the Mini-batch $\mathbb{I}$ from the experience replay memory $\mathbb{D}$ based on the sample strategy. $\tilde{a} = \mu\left(s^{t+1}\right) + n_t, n_t \sim clip(\mathcal{N}(0, \sigma), -c, c)$, $y = r + \gamma \min_{i=1,2} \theta^{Q'_i}\left(s^{t+1}, \tilde{a}\right)$.
7: Update Critics by minimizing the loss, $\theta^{Q_i} = \arg\min_{\theta^{Q_i}} L\left(\theta^{Q_i}\right) = \frac{1}{|\mathbb{I}|}\sum_{j\in\mathbb{I}}\left(y_j - Q_i\left(s_j, a_j \middle| \theta^{Q_i}\right)\right)^2, i = 1, 2$
8: Update the actor policy by the sampled policy gradient according to (35),
9: Every Z steps, update the target net parameters of Actor and Critic network:
10: $\vartheta^{\pi'} \leftarrow \upsilon\vartheta_t^{\pi} + (1-\upsilon)\vartheta_t^{\pi'}$
11: $\theta^{Q'_i} \leftarrow \upsilon\theta_t^{Q_i} + (1-\upsilon)\theta_t^{Q'_i}, i = 1, 2$
12: **end for**
13: **end for**

D. Deep Actor-Networks Training and Update

The actor network is trained to generate deterministic policy $\mu(s^t)$. The policy-gradient algorithm is applied to iteratively evaluate and improve the policy according to the gradient.

Considering the goal of the agent, the actor updates the parameters to make $\mu(s^t)$ approximate the optimal UAV trajectory policy to achieve the maximum $E\{Q_1(s^t, a^t|\theta^{Q_1})\}$. Note that we take the parameters of first critic network to update the actor network. Thus, the policy objective function, used to evaluate policy performance under a given parameter ϑ^{π}, can be defined as $J(\vartheta^{\pi}) = E\{Q_1(a^t)\}$.

With minibatch $\mathbb{I}$, the net parameter ϑ^{π} is updated by applying the policy gradient method

$$\nabla_{\vartheta^{\pi}} J\left(\vartheta^{\pi}\right) \approx \frac{1}{|\mathbb{I}|}\sum\nolimits_{j\in\mathbb{I}} \nabla_a Q_1\left(s^t, a^t\middle|\theta^{Q_1}\right)\Big|_{a=\mu\left(s_j|\vartheta^{\pi}\right)} \times \nabla_{\vartheta^{\pi}}\mu\left(s_j\middle|\vartheta^{\pi}\right). \tag{35}$$

The update for parameter ϑ^{π} of policy gradient can be expressed as

$$\vartheta^{\pi} := \vartheta^{\pi} - \alpha_a \nabla_{\vartheta^{\pi}} J\left(\vartheta^{\pi}\right) \tag{36}$$

where $\alpha_a > 0$ is a positive constant and represents the learning rate of the actor net.

E. Complete Algorithm

Summarizing all the components, we present the pseudocode of the proposed TD3-based AoI-energy-aware UAV trajectory planning algorithm in Algorithm 1.

Our solution works as follows. At the beginning, we initialize the discount factor γ, update rate υ, penalty $\hbar^{\text{penalty}}$, and award $\hbar^{\text{award}}$. For the UAV agent, we initialize online net parameters for actor net and two critic nets ϑ^{π}, θ^{Q_1}, and θ^{Q_2}, and the corresponding target net parameters $\vartheta^{\pi'}$, $\theta^{Q'_1}$, and $\theta^{Q'_2}$. The UAV is equipped with a reply buffer, which is also initialized to store the experiences. In each episode, we reset the configurations of network environment $s^0 \in \mathcal{S}$, including the initialized AoI of all IoTDs, the initialized position of the UAV, and the initialized coverage state for all IoTDs. At the beginning of each step, the UAV agent selects action a^t according to the current state s^t. Then, the UAV executes selected action a^t, and obtains new state s^{t+1}, and receives corresponding reward r^t.

In the learning process, based on the sample strategy, minibatch $\mathbb{I}$ including sample $\langle s^t, a^t, r^t, s^{t+1}\rangle$ is sampled from the memory buffer $\mathbb{D}$. To explore more effective actions and achieve the more accurate Q function, random noise n_t is added to the derived action based on the strategy and current state s^{t+1}, and we have new action $\tilde{a}$, which is shown in line 6 of Algorithm 1, and c is the boundary of action $\tilde{a}$. Based on $\tilde{a}$, we obtain target value y according to (32). Then, we update the parameters of online net of the critic network by minimizing the loss function according to (31). By the sampled policy gradient method in (35) and (36), the parameters of the online net of the actor network can be updated. Since the updating period of target net is larger than that of the online net, the parameters of the target networks will be updated every Z steps based on update rate υ shown in lines 10 and 11.

F. Discussions

Based on the above discussions, we can see that the training algorithm includes the relay buffer and six neural networks. The test validation of the training algorithm is comprised of the online actor net. Comapared with the training algorithm, the latency of test validation can be ignored. Then, we mainly analyze the complexity of the training algorithm.

The complexity of the proposed TD3-AUTP algorithm mainly depends on the number of training steps. The complexity of the training algorithm can be derived with regard to the floating point operations per second (FLOPS). In the TD3-AUTP algorithm, all neural networks are the fully connected layer networks. For the dot product of a P vector and a $P \times Q$ matrix, the FLOPS is calculated as $(2P-1)Q$. For each column in the matrix, the operations are composed of P multiplications and $P-1$ additions. Besides, a calculation of the activation layer in the neural networks is needed. For a single FLOP calculation, the multiplication, addition, subtraction, division, exponentiation, etc., are counted. Therefore, the functions of different activation layers correspond to different FLOPS. Hence, the computations are Q with Q inputs for Relu layers, $4 \times Q$ for sigmoid layers, and $6 \times Q$ for tanh layers.

We can derive the computing complexity of Ξ^a layers of the actor fully connected neural network and Ξ^c layers of the critic fully connected neural network as

$$l_{\text{act}} l_i + 2\sum_{j=0}^{\Xi^a - 1} l_{\text{actor},j} l_{\text{actor},j+1} + 4\sum_{k=0}^{\Xi^c - 1} l_{\text{critic},k} l_{\text{critic},k+1}$$

$$= \mathcal{O}\left(\sum_{j=0}^{\Xi^a-1} l_{\text{actor},j} l_{\text{actor},j+1} + \sum_{k=0}^{\Xi^c-1} l_{\text{critic},k} l_{\text{critic},k+1}\right) \quad (37)$$

where l_{act} indicates the corresponding parameters determined by the type of the activation layer. l_i means the unit number in the ith layer, and l_0 is the input size. $l_{actor,j}$ and $l_{critic,k}$ mean the unit numbers in the jth layer of actor network and the kth layer of the critic network, respectively.

Next, we will discuss the convergence performance. First, our problem is an MDP, where the state set, actor set, as well as reward set are mapping uniquely. In the training process, the UAV is encouraged to fully explore the environment of training in the early stage with larger noise n_t. The policy gradient algorithm is applied to guide the optimizing process for obtaining the higher reward; thus, the experiences with better performance are emphasized and accumulated as the episode goes by. Therefore, the expected reward increases until the proposed TD3-AUTP algorithm achieves the stable and optimal solution. In Section IV, we will show the convergence performance by comparing it with the existed schemes.

The optimality of the proposed TD3-AUTP cannot be directly derived since our problem is a continuous optimization problem with multidimensional actions and states. The DQN is a value-based algorithm, the action space of which is discrete and finite. In each action decision process, a single action is selected from the action space. When increasing the dimension of the action space, the action space will exponentially grow. Besides, it is difficult to map multidimensional actions to the action space. However, the TD3 algorithm developed from the policy-based DDPG algorithm [43]–[45] by introducing clipped double critic networks. Same as the DDPG algorithm, the TD3 algorithm is an actor–critic and model-free algorithm, the action space of which is continuous. The TD3 algorithm can directly output a deterministic result for each learning agent according to the trained policy, so the output dimension of TD3 is fixed, which is the number of the optimizing variables. Even if the number of the variables increases, the output dimension of TD3 will not grow on a large scale. Therefore, compared with the DQN, the TD3 can effectively reduce the action size as well as the quantization error, and solve problems with the multidimensional actions. The objective function in our problem is not convex about the optimizing variables. The optimizing variables, such as UAV speed, flight direction, and flight distance, are tightly coupled together. In the training process, we cannot guarantee that the optimal solution can be reached. However, if we enlarge the reply memory and the size of the minibatch, the achieved result will approximate to the optimal solution.

IV. Simulation and Analysis

In this section, simulation results are carried out to illustrate the efficiency of the proposed TD3-AUTP algorithm. The initial location of the UAV is set as (-100 m, -100 m), and 40 IoTDs are deployed randomly in the area of 200×200 m^2. IoTDs are allocated with subchannels to send the sampled data to the UAV when they are covered by the UAV. The maximum number of IoTDs that can be scheduled in parallel is set as 3. The maximum horizontal distance that the UAV can serve each IoTD is set as 20 m. The simulation platform is presented as: CPU Intel i7-9750H and GPU Nvidia GTX-1660Ti.

TABLE I
Simulation Parameters

Symbol	Value
the propagation D	0.14
the propagation C	11.95
the path loss exponent α	2
the carrier frequency f_c	2 GHz
Noise Power σ^2	-130 dBm
the maximum bandwidth for UAV	2 MHz
the height of the UAV H	50 meters [27]
the discount factor for future rewards γ	0.99
Additional path loss to free space for Los η_1	3 dB [40]
Additional path loss to free space for NLos η_2	23 dB [40]
Learning Rate for Actor α_a	0.0001
Learning Rate for Critic α_c	0.001
Decay the Action Randomness	0.99
Discount factor γ	0.95
Soft Replacement Value	0.01

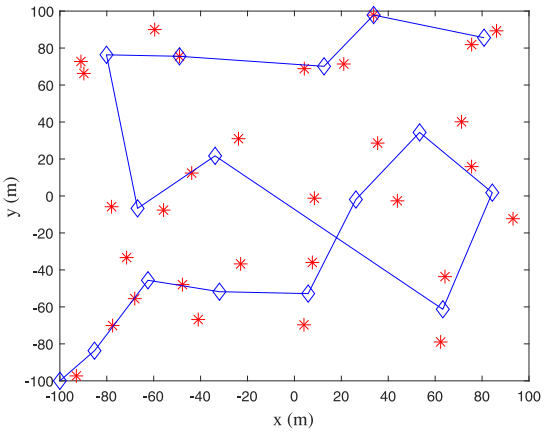


Fig. 5. Example of the UAV trajectory planning for data collection with distance-aware greedy algorithm.

The other parameters are listed in Table I.

For performance comparisons, we provide three related methods are as follows.

1) AoI-aware trajectory planning (A-TP) algorithm with DRL technique [26].
2) A-TP based on CA2C algorithm [23].
3) The distance-based greedy method (Greedy). The UAV decides the next hovering position based on the achieved the minimum AoI [28].

It is noted that the advantage of the greedy method is that the location information of all IoTDs is known to the UAV, and the UAV will not interact with the environment. In the greedy method, the UAV just takes the action according to the location and AoI of each IoTD with the target of achieving the largest utility. In the A-TP method, the DQN algorithm is adopted, where the actions are trained separately due to the characteristic of the DQN algorithm that the action space is limited and action is discrete. In the CA2C method, the actor–critic architecture is applied; however, it is difficult to guarantee the convergence of this method.

Fig. 5 shows the UAV trajectory that the UAV serves the nearest IoTDs with the greedy algorithm. The start point of

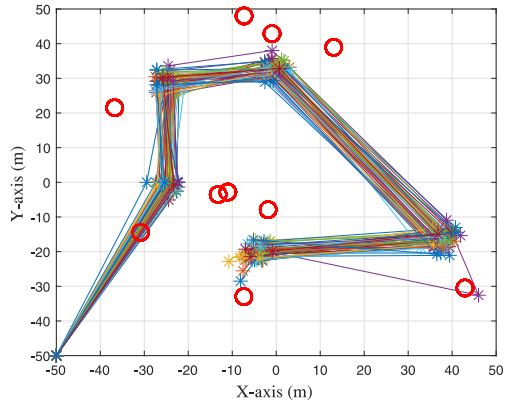


Fig. 6. Examples of the UAV trajectory planning for data collection with TD3-AUTP algorithm for the first subarea.

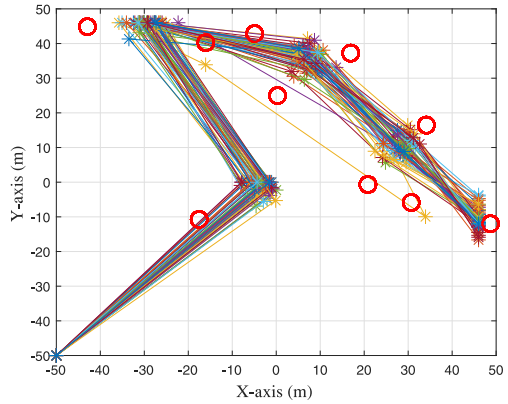


Fig. 7. Examples of the UAV trajectory planning for data collection with TD3-AUTP algorithm for the second subarea.

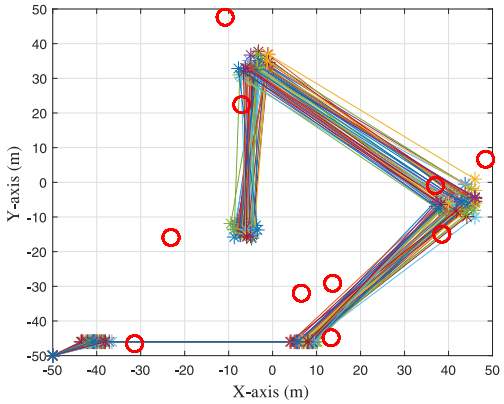


Fig. 8. Examples of the UAV trajectory planning for data collection with TD3-AUTP algorithm for the third subarea.

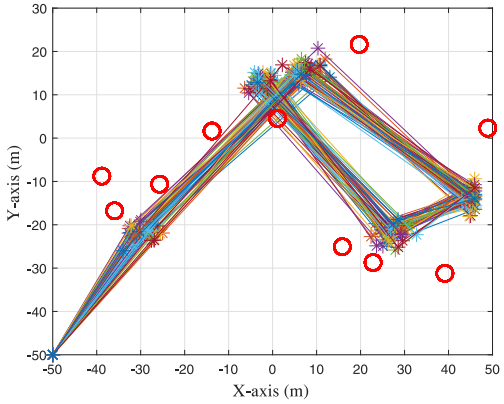


Fig. 9. Examples of the UAV trajectory planning for data collection with TD3-AUTP algorithm for the fourth subarea.

the UAV is (−100 m, −100 m), the blue rhombus symbols indicate the hovering spots of the UAV, and the red asterisk symbols denote the locations of IoTDs. The UAV covers the nearest IoTDs until all IoTDs are scheduled. This method has a disadvantage that the UAV may fly a relatively longer distance to cover the uncovered IoTDs. The trajectory route may not be formulated as the closed loop, which will cause that the first covered IoTD suffers a relatively large duration to be scheduled again.

Figs. 6–9 show the UAV trajectory with the proposed TD3-AUTP algorithm. The coverage area is divided into four subareas, the subarea is set as 100×100 m^2, and the maximum speed of the UAV is set as 30 m/s. In these figures, the red hollow circle symbols present the locations of IoTDs and the asterisk symbols denote the hovering spots of the UAV. For each subarea, we obtain the training model by the TD3-AUTP algorithm, and test the model with 50 times. These 50 test routes are superimposed on a single figure. On the one hand, the UAV can minimize distance for energy saving and flight latency reduction; on the other hand, more IoTDs are expected to be scheduled for AoI reduction.

The proposed TD3-AUTP algorithm is implemented based on Tensorflow, which is an open-source machine learning library. We use a fully connected DNN that has two hidden layers of 300 neurons and use the activation function of ReLU and tanh. The experience memory buffer size is set as 500 000, and the minibatch is set as 2000.

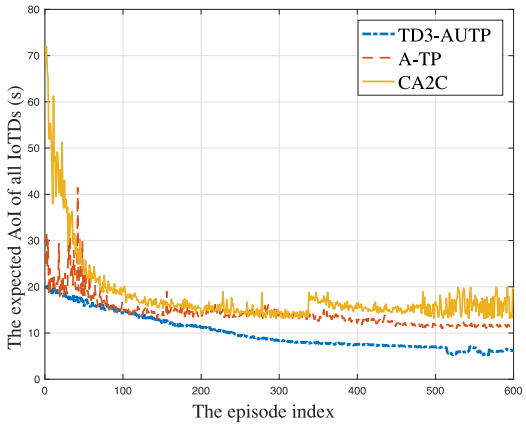


Fig. 10. Achieved expected AoI with the episodes.

Figs. 10 and 11 show the convergence performances of the TD3-AUTP, A-TP, and CA2C algorithms in the aspects of the expected AoI and number of steps (HTP processes) in one flight round. The CA2C algorithm has the poor convergence performance, and the TD3-AUTP and A-TP algorithms can converge to stable solutions with about 600 episodes. Compared with the CA2C and A-TP algorithms, the proposed

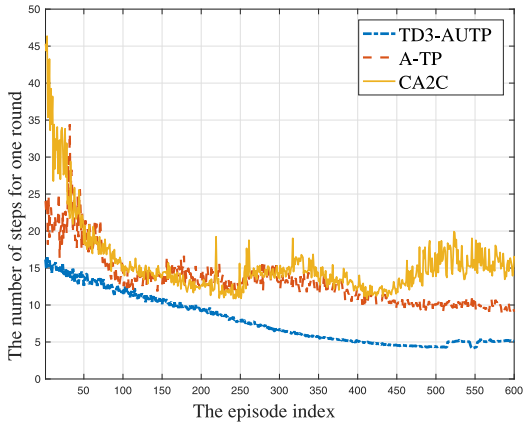


Fig. 11. Average number of steps with the episodes.

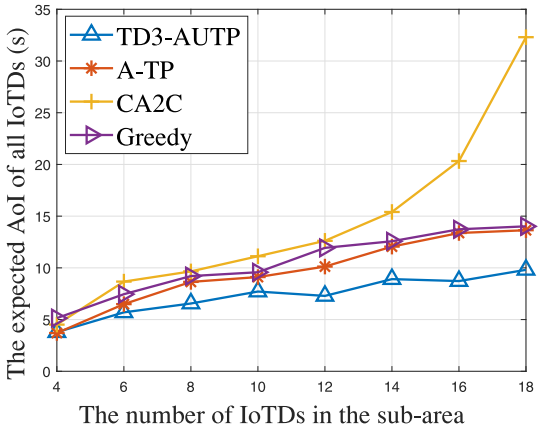


Fig. 12. Achieved expected AoI versus the number of IoTDs.

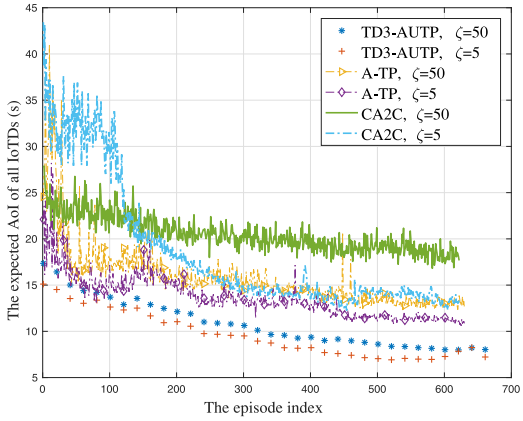


Fig. 13. Expected AoI of all IoTDs with different UAV energy weight factor.

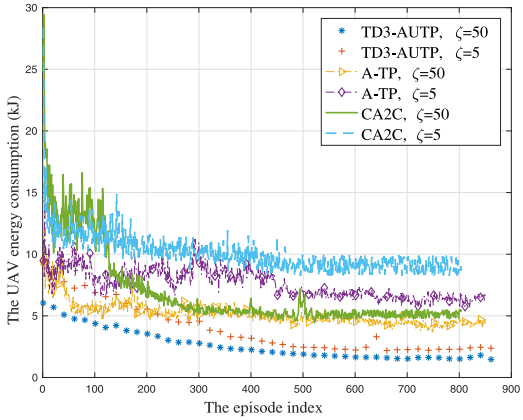


Fig. 14. UAV energy consumption for one round with different UAV energy weight factor.

TD3-AUTP algorithm can achieve the lower expected AoI with the less steps.

Fig. 12 shows the expected AoI performance under the different numbers of IoTDs with the subarea size set as 100×100 m^2. Note that the expected AoI indicates the average AoI of all IoTDs. For one IoTD drop, in the proposed scheme, as the number of IoTDs increases, the expected AoI increases with a little undulation, this is because the increased density of IoTDs makes it easier for the UAV to serve more IoTDs simultaneously. When the number of IoTDs gets larger, the performance of the CA2C algorithm deteriorates dramatically. When the number of IoTDs is set as 12, compared with the A-TP algorithm, CA2C algorithm, and greedy method, the proposed scheme can decrease the expected AoI by about 28%, 42.21%, and 38.91%, respectively. For the A-TP algorithm, since the action is discrete and the action space is limited, as well as the actions are trained in the distributed way, the performance is worse than that of the TD3-AUTP algorithm. As IoTDs increase and the state space becomes larger, in the TD3-AUTP algorithm, by increasing the memory buffer and minibatch size, the learning efficiency will be guaranteed; therefore, the proposed algorithm can achieve the better performance.

In Fig. 13, the weight factor of UAV energy consumption is set as 50 and 5, and the unit of the UAV energy is set as 10^3 J. The TD3-AUTP algorithm can achieve a relative lower expected AoI compared with the A-TP algorithm. When ζ is set as 5, the TD3-AUTP algorithm can achieve the lower expected AoI compared with that of the case $\zeta = 50$. The A-TP algorithm achieves a relatively larger expected AoI when ζ is set as 50. In Fig. 14, compared with the A-TP algorithm, the proposed TD3-AUTP algorithm brings about the higher UAV energy consumption, since TD3-AUTP obtains more utility from the expected AoI reduction with sacrificing more energy consumption.

We fix the number of IoTDs and increase the area to evaluate the learning efficiency of the TD3-AUTP algorithm. As shown in Fig. 15, as the extension of the coverage area, the expected AoI increases. On one hand, the flight distance increases and the flight time increases, on the other hand, as the exploration range increases, the accuracy of the three schemes will be destroyed under the same parameters. The proposed TD3-AUTP algorithm outperforms other algorithms in terms of the expected AoI. When the subarea is set as 140×140 m^2, compared with A-TP, CA2C, and greedy methods, the TD3-AUTP can decrease the expected AoI of all IoTDs about 46.94%, 65.34%, and 44.85%, respectively.

In Fig. 16, as the length on one side increases, the average number of steps for one round increases by applying these four methods. This is because that the flight distance is limited for

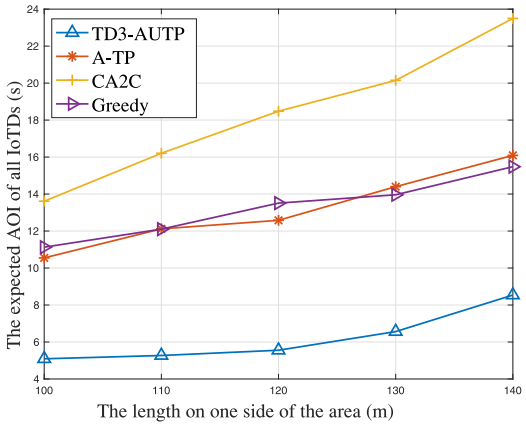


Fig. 15. Achieved expected AoI versus the length of each side of the coverage area.

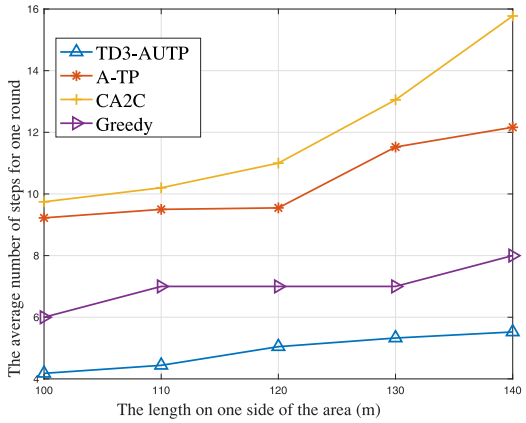


Fig. 16. Average number of steps versus the length of each side of the coverage area.

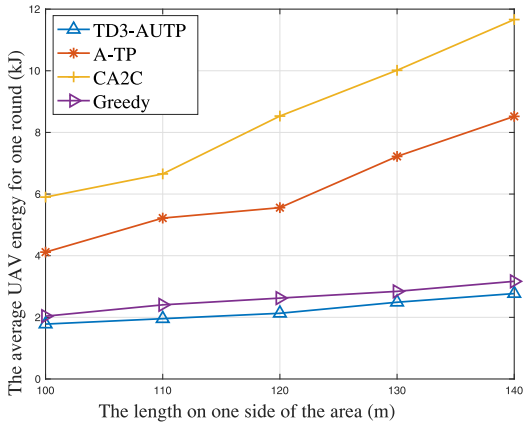


Fig. 17. Average UAV propulsion energy versus the length of each side of the coverage area.

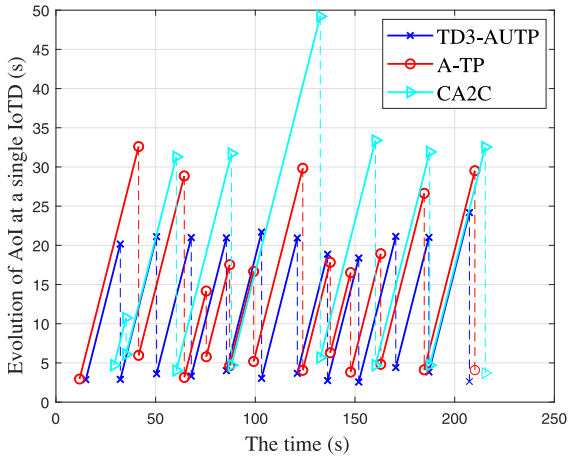


Fig. 18. AoI evolution for a single IoTD.

one HTP and the UAV prefers to provide a higher channel condition for each IoTD, thus less IoTDs will be severed in parallel due to the relatively longer distance. The proposed TD3-AUTP algorithm has better performance in the aspect of the number of steps for one round. With the length on each side of the area setting as 140 m, compared with A-TP, CA2C, and greedy methods, the TD3-AUTP algorithm can decrease the average number of steps by about 57.60%, 67.35%, and 30.97%, respectively.

As shown in Fig. 17, the proposed TD3-AUTP algorithm can achieve the lower UAV energy consumption in one round than that of the A-TP and CA2C algorithms. As the coverage area increases, the advantage of the proposed TD3-AUTP algorithm is more obvious. When the length on each side of the subarea is set as 140 m, compared with the A-TP algorithm, CA2C algorithm, and greedy algorithm, the TD3-AUTP algorithm can decrease the energy consumption by about 67.47%, 76.23%, and 12.52%, respectively.

Fig. 18 shows the evolution of AoI at a single IoTD when the number of IoTDs is set as 12. The age performance is evaluated under the proposed TD3-AUTP, A-TP, and CA2C algorithms. Compared with A-TP and CA2C algorithms, the AoI evolution in the proposed TD3-AUTP algorithm keeps smoother and the achieved maximum AoI is relatively lower. The CA2C has the worst performance due to the poor convergence performance. On one hand, the TD3-AUTP algorithm can guarantee the effectiveness of collected data. On the other hand, it can help IoTDs monitor the signal of the UAV for data collection periodically.

V. Conclusion

In this article, we investigated an AoI-energy-aware UAV-assisted data collection scheme for IoT networks, where the UAV executes data collection tasks continuously in the given geographic area during a period. We formulated an optimization problem to minimize the weighted sum of the expected AoI and energy consumption by jointly optimizing the flight speed, direction, and distance of the UAV, the IoTD scheduling, and channel allocation for IoTDs. Due to the environment dynamics, the high-dimensional variables, and the continuously optimizing process, it is challenging to utilize traditional optimizing methods to achieve the long-term stable optimal solution. We reformulated the minimization problem as an MDP, and employed the RL method to characterize the optimizing process, and developed the TD3-AUPT algorithm based on the DDPG framework to achieve the long-term optimal solution. By adjusting the energy weight factors,

the proposed scheme can achieve the tradeoff among the expected AoI, the transmission energy for IoTDs, and the UAV propulsion energy. Compared with the existed schemes, the simulation results illustrate the efficiency of the TD3-AUPT algorithm in terms of the average AoI, the number of steps in one round, and the energy consumption. Based on this work, we will investigate the multi-UAV cooperative data collection scheme in future work.

REFERENCES

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[2] Q. Cui, J. Zhang, X. Zhang, K. Chen, X. Tao, and P. Zhang, "Online anticipatory proactive network association in mobile edge computing for IoT," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4519–4534, Jul. 2020.

[3] M. Samir, S. Sharafeddine, C. M. Assi, T. M. Nguyen, and A. Ghrayeb, "UAV trajectory planning for data collection from time-constrained IoT devices," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 34–46, Jan. 2020.

[4] G. Li *et al.*, "Energy efficient data collection in large-scale Internet of Things via computation offloading," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4176–4187, Jun. 2019.

[5] M. Samir, C. Assi, S. Sharafeddine, D. Ebrahimi, and A. Ghrayeb, "Age of information aware trajectory planning of UAVs in intelligent transportation systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12382–12395, Nov. 2020.

[6] A. Kosta, N. Pappas, and V. Angelakis, *Age of Information: A New Concept, Metric, and Tool*. Hanover, MA, USA: Now, 2017.

[7] H. Zheng, K. Xiong, P. Fan, Z. Zhong, and K. B. Letaief, "Age of information-based wireless powered communication networks with selfish charging nodes," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1393–1411, May 2021.

[8] Z. Wang, R. Liu, Q. Liu, J. S. Thompson, and M. Kadoch, "Energy-efficient data collection and device positioning in UAV-assisted IoT," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1122–1139, Feb. 2020.

[9] C. Zhan and Y. Zeng, "Aerial–ground cost tradeoff for multi-UAV-enabled data collection in wireless sensor networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1937–1950, Mar. 2020.

[10] A. Rovira-Sugranes and A. Razi, "Optimizing the age of information for blockchain technology with applications to IoT sensors," *IEEE Commun. Lett.*, vol. 24, no. 1, pp. 183–187, Jan. 2020.

[11] Y. Gu, H. Chen, C. Zhai, Y. Li, and B. Vucetic, "Minimizing age of information in cognitive radio-based IoT systems: Underlay or overlay?" *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10273–10288, Dec. 2019.

[12] Y. Gu, H. Chen, Y. Zhou, Y. Li, and B. Vucetic, "Timely status update in Internet of Things monitoring systems: An age-energy tradeoff," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5324–5335, Jun. 2019.

[13] C. Xu, H. H. Yang, X. Wang, and T. Q. S. Quek, "Optimizing information freshness in computing-enabled IoT networks," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 971–985, Feb. 2020.

[14] B. Zhou and W. Saad, "Minimum age of information in the Internet of Things with non-uniform status packet sizes," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1933–1947, Mar. 2020.

[15] B. Zhou and W. Saad, "Joint status sampling and updating for minimizing age of information in the Internet of Things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7468–7482, Nov. 2019.

[16] 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, 3rd Quart., 2019.

[17] S. Zhang, H. Zhang, and L. Song, "Beyond D2D: Full dimension UAV-to-everything communications in 6G," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6592–6602, Apr. 2020.

[18] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1346–1359, Jan. 2019.

[19] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5092–5106, Aug. 2018.

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

[21] Y. Liu, Z. Qin, Y. Cai, Y. Gao, G. Y. Li, and A. Nallanathan, "UAV communications based on non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 26, no. 1, pp. 52–57, Feb. 2019.

[22] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Wireless Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[23] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative Internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6807–6821, Nov. 2020.

[24] M. A. Abd-Elmagid and H. S. Dhillon, "Average peak age-of-information minimization in UAV-assisted IoT networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 2003–2008, Feb. 2019.

[25] C. Li, S. Li, Y. Chen, T. Hou, and W. Lou, "Minimizing age of information under general models for IoT data collection," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 2256–2270, Oct.–Dec. 2020.

[26] C. Zhou *et al.*, "Deep RL-based trajectory planning for AoI minimization in UAV-assisted IoT," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[27] J. Liu, P. Tong, X. Wang, B. Bai, and H. Dai, "UAV-Aided data collection for information freshness in wireless sensor networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2368–2382, Apr. 2021.

[28] 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.*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021.

[29] S. Zhang, H. Zhang, Z. Han, H. V. Poor, and L. Song, "Age of information in a cellular Internet of UAVs: Sensing and communication trade-off design," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6578–6592, Oct. 2020.

[30] M. A. Abd-Elmagid, A. Ferdowsi, H. S. Dhillon, and W. Saad, "Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[31] M. H. Cheung, "Age of information aware UAV network selection," in *Proc. WiOPT*, 2020, pp. 1–8.

[32] Z. Jia, X. Qin, Z. Wang, and B. Liu, "Age-based path planning and data acquisition in UAV-assisted IoT networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, 2019, pp. 1–6.

[33] W. Li, L. Wang, and A. Fei, "Minimizing packet expiration loss with path planning in UAV-assisted data sensing," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1520–1523, Dec. 2019.

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

[35] D. Yang, Q. Wu, Y. Zeng, and R. Zhang, "Energy tradeoff in ground-to-UAV communication via trajectory design," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6721–6726, Jul. 2018.

[36] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao, and X. Shen, "Energy-efficient UAV-assisted mobile edge computing: Resource allocation and trajectory optimization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3424–3438, Mar. 2020.

[37] 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, Jun. 2020.

[38] C. H. Liu, Z. Chen, and Y. Zhan, "Energy-efficient distributed mobile crowd sensing: A deep learning approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1262–1276, Jun. 2019.

[39] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, Aug. 2019.

[40] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.

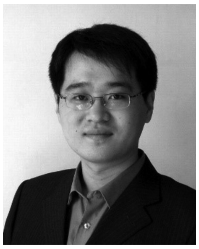
[41] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.

[42] Y. Sun, Y. Polyanskiy, and E. Uysal, "Sampling of the Wiener process for remote estimation over a channel with random delay," *IEEE Trans. Inf. Theory*, vol. 66, no. 2, pp. 1118–1135, Feb. 2020.

[43] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. ICLR*, 2016, pp. 1–14.
[44] G. Qiao, S. Leng, S. Maharjan, Y. Zhang, and N. Ansari, "Deep reinforcement learning for cooperative content caching in vehicular edge computing and networks," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 247–257, Jan. 2020.
[45] Y. Liu, X. Wang, J. Mei, G. Boudreau, H. Abou-Zeid, and A. B. Sediq, "Situation-aware resource allocation for multi-dimensional intelligent multiple access: A proactive deep learning framework," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 116–130, Jan. 2021.

Mengying Sun (Student Member, IEEE) received the B.S. degree in communication engineering from Beijing University of Chemical Technology, Beijing, China, in 2016. She is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications, Beijing.

Her research interests cover wireless communication, distributed computing, device-to-device communication, resource allocation, and mobile caching.



Xiaodong Xu (Senior Member, IEEE) received the B.S. degree in information and communication engineering and the master's degree in communication and information system from Shandong University, Jinan, China, in 2001 and 2004, respectively, and the Ph.D. degree of circuit and system from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2007.

He is currently a Professor with BUPT and a Research Fellow with Peng Cheng Laboratory, Shenzhen, China. He has coauthored nine books and more than 120 journal and conference papers. He is also the inventor or co-inventor of 39 granted patents. His research interests cover moving networks, D2D communications, and mobile-edge computing and caching.

Prof. Xu is an Associate Editor of IEEE ACCESS.

Xiaoqi Qin (Member, IEEE) received the Ph.D. degree from the Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA, in 2016.

She is a Lecturer with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research focuses on exploring new performance limits of next-generation wireless networks, and developing innovative resource sharing schemes based on value of information in intelligent Internet of things to support real-time applications.

Ping Zhang (Fellow, IEEE) received the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 1990.

He is currently a Professor with Beijing University of Posts and Telecommunications. He has published eight books and more than 400 papers, and he holds approximately 170 patents. His current research interests include mobile communications, ubiquitous networking, and service provisioning, especially in the key techniques of the 5G systems.

Prof. Zhang is an Executive Associate Editor-in-Chief on information sciences of *Chinese Science Bulletin*, a Member of next-generation broadband wireless communication network in National Science and Technology Major Project Committee, the 5th Advisory Committee of National Natural Science Foundation of China, The Ministry of Science and Technology (MOST) 863 Program Expert Team, and MOST IMT-Advanced 5G Expert Team, and the Chief Scientist of "973" National Basic Research Program of China.