Intelligent Reflecting Surface-Assisted Fresh Data Collection in UAV Communications

Hongli Huang, Juan Liu(✉), and Lingfu Xie

Ningbo University, Ningbo 315211, China
{2011082102,liujuan1,xielingfu}@nbu.edu.cn

Abstract. This paper considers employing Intelligent reflecting Surface (IRS) in unmanned aerial vehicle (UAV) assisted wireless communications to ensure the freshness of the collected data in Internet of Things (IoT). We aim to minimize the average Age of Information (AoI) of the sensor nodes (SNs) by jointly optimizing the UAV flight trajectory, the SNs' scheduling and the IRS phase shift matrix. It is modeled as a Markovian Decision Process (MDP) problem. A deep reinforcement learning algorithm based on a Twin Delayed Deep Deterministic Policy Gradient (TD3) is proposed to learn and find the optimal UAV trajectory and scheduling of the SNs. For a scheduled transmission, the IRS is used based on the channel information to align the signal phase shifts. Simulation results show that IRS-assisted UAV data collection can significantly reduce the AoI of the SNs.

Keywords: Unmanned aerial vehicle · intelligent reflecting surface · deep reinforcement learning · age of information

1 Introduction

Reliable and timely sensory information by ground sensor nodes (SNs) is critical to applications in Internet of Things (IoT). It is generally challenging for the SNs with limited battery capacity to communicate reliably over long distances. In recent years, unmanned aerial vehicles (UAVs) are routinely used as mobile data collectors in IoT due to their high mobility and easy deployment. The age of information (AoI) is a measure to SNs' information freshness. In [1], AoI was defined as the time elapsed from the generation of the latest packet by a source node to its reception by a target node. For AoI-oriented UAV data collection, [2] designed an online flight trajectories of the UAV based on deep reinforcement learning (DRL) method to minimize the SNs' weighted sum of AoIs, and [3] studied the influence of SNs' sampling and the queueing process on the SNs' average AoI. The above works only account for the Line of Sight (LoS) channel between the UAV and the SN.

In the urban environment, however, the LoS link between the UAV and the SN is likely to be blocked by obstructions like tall buildings. Intelligent reflecting Surfaces (IRS) is one of the technologies that have great potential in future

Q. Liang et al. (Eds.): CSPS 2022, LNEE 874, pp. 189–197, 2023.
https://doi.org/10.1007/978-981-99-2362-5_24

wireless networks. It is is a plane composed of a large number of low-cost passive reflective elements, each of which can independently adjust the phase of an incoming signal. This allows for intelligently reconstructing the wireless propagation environment and improving the channel quality [4]. IRS is also amenable to installation. As a result, IRS can be used to overcome the channel blockage between the UAV and the SNs. Moreover, with IRS, other aspects of the communication systems, e.g., the UAV energy consumption [5] and the network throughput [6], can be improved.

In contrast to the above works, we consider the deployment of an IRS for UAV-assisted data collection in the IoT. We assume the energy of the UAV is limited and there is no charging station. For either periodic or random sampling of the SNs, we aim to minimize the SNs' average AoI by jointly optimizing the UAV flight trajectory, the SNs' scheduling and IRS phase shift matrix. This is modeled as a finite Markov decision process (MDP). To solve this problem, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm [7] in DRL is proposed to learn and find the optimal policy for the flight trajectory and node scheduling. For a scheduled transmission from a SN to the UAV, the IRS, based on the channel information, is operated in such a way that the phase shifts of the signals are aligned. Simulation results demonstrate that with IRS and the learned optimal policy, both the average AOI and the transmission power of the SNs can be significantly reduced.

2 System Model

As shown in Fig. 1, we consider an IoT, where I SNs are distributed in the rectangular area with side lengths x_{max}^U and y_{max}^U to sample the environment and a rotary-wing UAV acts as a mobile base station to collect status-update information. The horizontal location of the i-th SN is expressed as $q_i = [x_i, y_i] \in \mathbb{R}^2 \, (i \in \mathcal{I} = \{1, ..., I\})$. However, obstructions such as tall buildings and trees in congested cities cause severe path loss and high attenuation to the air-to-ground channels between the UAV and SNs. In this case, we deploy an IRS on a high-rise building with height H_R to improve channel quality by reflecting signals controllably. The horizontal location of the IRS is defined as $q_R = [x_R, y_R] \in \mathbb{R}^2$. For simplicity, we assume a time-slotted system where the length of each time slot is T_{ts} seconds. The UAV flies at height H_U over the rectangular target area. The horizontal location of the UAV at the time slot t can be defined by $q_t^U = \left[x_t^U, y_t^U\right] \in \mathbb{R}^2 \, (t \in \mathcal{T} = \{1, \ldots, T\})$. Furthermore, $T \in \mathbb{N}$ depends on the UAV's maximum carried energy e_{max} and the service process.

2.1 Channel Model

The ground-to-air communication channel between each SN and the UAV includes two links: the direct link from the SN to the UAV, and the indirect link reflected by the IRS. The distance between the UAV and SN i at slot t is given by $d_t^{i,U} = \sqrt{||q_t^U - q_i||^2 + \left(H_U\right)^2}$. Similarly, the distances between the SN

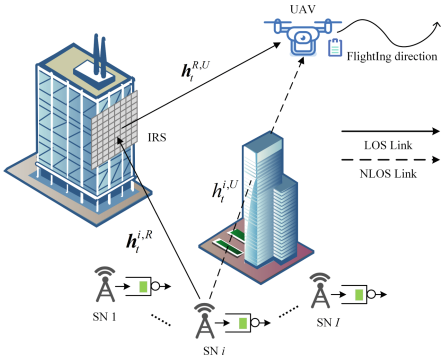


Fig. 1. IRS-assisted UAV data collection

i and IRS and between the IRS and UAV are denoted as $d_{i,R}$ and $d_t^{R,U}$, respectively. Assume that the IRS is composed of an $M \times M$ uniform planar array (UPA) with M^2 reflective elements. The set of reflective elements is defined as $\mathcal{M} = \{1, ..., M^2\}$.

1) Direct link: According to the channel model in [8], the features of LoS and non-line-of-sight (NLoS) links are preserved and appear with a certain probability, respectively. Then, the channel gain between the UAV and SN i at slot t is given by

$$h_t^{i,U} = \begin{cases} \sqrt{\frac{\beta_0}{\left(d_t^{i,U}\right)^{\alpha_1}}}\widetilde{h}_{i,t}, & \text{LoS link} \\ \sqrt{\nu\frac{\beta_0}{\left(d_t^{i,U}\right)^{\alpha_1}}}\widetilde{h}_{i,t}, & \text{NLoS link} \end{cases}, \tag{1}$$

where β_0 is the pathloss at the reference distance of 1 m, α_1 is the path loss exponent for the direct link, $\nu < 1$ denotes the attenuation factor due to NLoS, and $\widetilde{h}_{i,t}$ is the small-scale fading that follows the complex Gaussian distribution with mean 0 and variance 1.

2) Indirect link: The channel gain between the IRS and SN i obeys the Rician fading at slot t, $\boldsymbol{h}_t^{i,R} \in \mathbb{C}^{M^2 \times 1}$, which can be expressed as $\boldsymbol{h}_t^{i,R} = \sqrt{\frac{\beta_0}{(d_{i,R})^{\alpha_2}}}\left(\sqrt{\frac{k}{1+k}}\boldsymbol{h}_{LoS}^{i,R} + \sqrt{\frac{1}{1+k}}\boldsymbol{h}_{t,NLoS}^{i,R}\right)$, where k is the Rician factor, α_2 is the path loss exponent between the SN and the IRS, $\boldsymbol{h}_{LoS}^{i,R}$ is the LOS component, and $\boldsymbol{h}_{t,NLoS}^{i,R}$ is the NLOS component modeled as a complex Gaussian variable with mean 0 and variance 1. Here, $\boldsymbol{h}_{LoS}^{i,R} = [1, \cdots, e^{-j\frac{2\pi d}{\lambda}(M-1)\sin\theta_{i,R}\cos\zeta_{i,R}}]^H \otimes [1, \cdots, e^{-j\frac{2\pi d}{\lambda}(M-1)\sin\theta_{i,R}\sin\zeta_{i,R}}]^H \in \mathbb{C}^{M^2\times 1}$, where d is the distance between the IRS elements, λ is the carrier length, $\theta_{i,R}$ and $\zeta_{i,R}$ represent the vertical and horizontal angle-of-departures (AoDs) from the SN i to IRS at slot t, respectively. In addition, the geographical relationships are $\sin\theta_{i,R} = \frac{H_R}{\|q_R - q_i\|}$, $\sin\zeta_{i,R} = \frac{|x_i - x_R|}{\|q_R - q_i\|}$ and $\cos\zeta_{i,R} = \frac{|y_i - y_R|}{\|q_R - q_i\|}$.

The channel between the UAV and the IRS is dominated by the LoS link. Similarly, the channel gain at slot t is expressed as $\boldsymbol{h}_t^{R,U} = \sqrt{\frac{\beta_0}{\left(d_t^{R,U}\right)^2}} \boldsymbol{h}_{t,LoS}^{R,U} \in \mathbb{C}^{M^2 \times 1}$, where $\boldsymbol{h}_{t,LoS}^{R,U}$ is the LOS component from the IRS to the UAV. The IRS phase shift matrix at slot t is defined as $\Theta_t = \text{diag}\left(e^{j\theta_t^1}, \ldots, e^{j\theta_t^{M^2}}\right)$, where $\theta_t^m \in [0, 2\pi)$ is the phase shift of the m-th element. Therefore, the signal-to-noise ratio (SNR) can be computed as $\eta_t^{i,U} = \frac{P\left|\left(\boldsymbol{h}_t^{R,U}\right)^H \Theta_t \boldsymbol{h}_t^{i,R} + h_t^{i,U}\right|^2}{\sigma^2}$, where P is the SN's transmit power and σ^2 is the noise power. If the SNR is less than a threshold η_{th}, i.e., $\eta_t^{i,U} < \eta_{th}$, the UAV cannot decode the received signal successfully.

2.2 Queuing Model

Each SN samples periodically or randomly the environmental information, referred to as fixed sampling and random sampling. The sensed information is packaged into an update packet of ω bits with a timestamp [3]. Then, the packet is stored in the buffer of the SN and waits for collection by the UAV. Let $g_t^i \in \{0, 1\}$ denotes the sampling action of SN i at slot t. Specifically, $g_t^i = 1$ denotes that SN i generates an update packet at slot t, and otherwise $g_t^i = 0$. Once an update packet arrives at SN i in the beginning of each slot t, its lifetime is recorded and updated as

$$U_t^i = \begin{cases} 0, & g_t^i = 1 \\ U_{t-1}^i + 1, & \text{otherwise} \end{cases}. \tag{2}$$

Let $c_t^i \in \{0, 1\}$ be the binary user scheduling variable. $c_t^i = 1$ indicates that the SN i is associated and ready to send one update packet to the UAV at slot t, and otherwise $c_t^i = 0$. To fully exploit the IRS, it is assumed that the UAV is associated with at most one SN in each time slot. If the update packet is successfully delivered to the UAV with the aid of IRS, we say that SN i is served at slot t. Accordingly, the service state of SN i is set $z_t^i = 1$. If the transmission is failed or no transmission takes place, the service state is set as $z_t^i = 0$. After a successful transmission, the AoI of this SN is updated according to the lifetime of the delivered update packet, and otherwise the AoI increases by one after a time slot. At the beginning of each slot t, the AoI is updated as

$$A_t^i = \begin{cases} U_{t-1}^i + 1, & z_{t-1}^i = 1 \\ A_{t-1}^i + 1, & \text{otherwise} \end{cases}. \tag{3}$$

The average AoI of all SNs at time slot t is given by $\overline{A_t} = \frac{1}{I} \sum_{i=1}^{I} A_t^i$.

2.3 Problem Description

The UAV consumes energy on flight and hovering. The energy consumption for receiving and decoding the update packets is relatively small and can be omitted.

The UAV hovers and collects data during the transmission interval T_s and then flies to the next location during the remaining time $T_{ts} - T_s$. If no data needs to be collected, the UAV flies across the entire time slot. Therefore, the energy consumption at slot t can be expressed as

$$e_t^{co} = \begin{cases} P_t^f T_{ts}, & \sum_{i=1}^{I} c_t^i = 0 \\ P_{ho} T_s + P_t^f \left(T_{ts} - T_s\right), & \text{otherwise} \end{cases}, \tag{4}$$

where P_t^f is the UAV flight power as a function of flight speed v_t^U and P_{ho} is the hovering power, which can be obtained from Eq. (11) in [5]. Then, the remaining energy of the UAV at slot t can be computed as $e_t^{re} = e_{t-1}^{re} - e_t^{co}$.

The objective is to minimize the weighted sum of the average AoI of all SNs and the UAV's energy consumption by jointly optimizing the UAV flight trajectory $\boldsymbol{Q} = \left[q_t^U, \forall t \in \mathcal{T}\right]$, SN scheduling $\boldsymbol{C} = \left[c_t^i, \forall i \in I, \forall t \in \mathcal{T}\right]$, and IRS's phase shift matrix $\boldsymbol{\Phi} = [\Theta_t, \forall t \in \mathcal{T}]$. The optimization problem can be formulated as

$$\min_{\boldsymbol{Q},\boldsymbol{C},\Phi} \quad \frac{1}{T} \sum_{t=1}^{T} \left(\overline{A_t} + \delta e_t^{co}\right) \tag{5}$$

$$\begin{aligned} s.t.\ & C1: \sum_{i=1}^{I} c_t^i \le 1,\ c_t^i \in \{0,1\},\ \forall t \in \mathcal{T}, \\ & C2: 0 \le \theta_t^m < 2\pi,\ \forall m \in \mathcal{M}, \\ & C3: 0 \le x_t^U \le x_{max}^U,\ 0 \le y_t^U \le y_{max}^U,\ \forall t \in \mathcal{T}, \\ & C4: 0 < \left\| q_t^U - q_{t-1}^U \right\| < v_{max}^U T_{ts},\ \forall t \in \mathcal{T}, \end{aligned}$$

where v_{max}^U is the maximum flying speed of the UAV and δ is the relative importance factor. It is quite difficult to solve the above mixed integer non-convex problem. Hence, we propose the TD3-based algorithm for UAV-enabled data collection which is able to make the best decision quickly and accurately even when the scale of the problem is very large.

3 TD3-Based UAV Data Collection Method

Then, a TD3 algorithm is proposed for the UAV-enabled data collection to find the optimal UAV trajectory and SN scheduling policy efficinetly. During each packet transmission, the IRS's phase shift matrix is optimized based on the perfectly estimated channel state to maximize the received SNR at the UAV.

3.1 Optimization of IRS Phase Shift Matrix

Given the SN scheduling and UAV's location, the received SNR is maximized by optimizing the phase shifts of IRS, which is equivalent to the following problem:

$$\min_{\Theta_t} \left| \left(\boldsymbol{h}_t^{R,U}\right)^H \Theta_t \boldsymbol{h}_t^{i,R} + h_t^{i,U} \right|^2 \tag{6}$$

$$s.t.\ 0 \leq \theta_t^m < 2\pi, \forall m \in \mathcal{M},$$

From [9], the optimal phase shifts of IRS can be obtained by aligning the phases of the direct and indirect links between the UAV and the associated SN. In particular, when scheduling SN i, the optimal phase shift of the m-th element of IRS can be expressed as $\theta_t^{m,*} = \phi_t^{i,U} - \left(\omega_{t,m}^{i,R} + \omega_{t,m}^{R,U}\right), \forall m \in \mathcal{M}$, where $\phi_t^{i,U}$, $\omega_{t,m}^{i,R}$ and $\omega_{t,m}^{R,U}$ are the phases of the direct SN-UAV link, and the SN-IRS and IRS-UAV links via element m, respectively.

3.2 TD3 Algorithm Design

MDP Problem Formulation: The optimization problem can be modeled as a finite MDP. In the sequel, we define the state space, action space and reward function, respectively.

1) State space: The system state at slot t is defined as $s_t = \left[q_t^U, e_t^{co}, \boldsymbol{A}_t, \boldsymbol{U}_t\right]$, where $\boldsymbol{A}_t = \left[A_t^i, \forall i \in I\right]$ and $\boldsymbol{U}_t = \left[U_t^i, \forall i \in I\right]$.

2) Action space: The system action at slot t is defined as $a_t = \left[\mu_t^U, d_t^U, \boldsymbol{C}_t\right]$, which includes the UAV's flight angle $\mu_t^U \in (0, 2\pi]$ and distance $d_t^U \in [0, d_{i,max}]$. Therefore, the horizontal position of the UAV at slot $t+1$ is updated as $q_{t+1}^U = \left[x_t^U + d_t^U \cos\mu_t^U, y_t^U + d_t^U \sin\mu_t^U\right]$.

3) Reward function: Given the state s_t and action a_t, the reward function at time slot t can be defined as $r_t(s_t, a_t) = -\left(\overline{A_t} + \delta e_t^{co}\right) + p_t$, where p_t is the penalty at slot t that gives punishment for an invalid action. For example, if the current action a_t causes the UAV to fly out of the target area, we set $p_t < 0$ and otherwise $p_t = 0$.

The objective is to find the optimal policy π^* to minimize the long-term return function $C_\pi = \mathbb{E}_\pi\left[\sum_{t=1}^{T}(\gamma)^{t-1} r_t(s_t, a_t)|s_1\right]$, where $\mathbb{E}_\pi$ is the expectation under policy π, $\gamma \in [0, 1]$ is the discount factor, and s_1 is the initial state.

TD3 Algorithm: The TD3 algorithm is based on an Actor-Critic framework consisting of deep neural networks (DNN) [7] to find the optimal policy, which has one Actor network that obtains the deterministic policy $\pi_\vartheta(s)$, and two Critic networks that obtains the value function $Q_\varphi(a, s)$. In addition, there are two target Critic networks with function $Q_{\varphi'}(a, s)$ and one target Actor network with function $\pi_{\vartheta'}(s)$. The Actor network can randomly extract mini-batches of samples from the replay buffer to train the network parameters. The policy gradient is $\nabla_\vartheta J(\vartheta) = N^{-1}\sum \nabla_a Q_{\varphi_1}(s, a)|_{a=\pi_\vartheta(s)} \nabla_\vartheta \pi_\vartheta(s)$, where N is the mini-batch size. The target Actor network copys the Actor network parameters periodically to stabilize the training process, and the target Critic network is the same. The smaller Q value in the two target Critic networks is selected as the target value: $y = r + \gamma \min_{l=1,2} Q_{\varphi_l'}\left(s', \pi_{\vartheta'}\left(s'\right) + \varepsilon\right)$, where $\varepsilon \sim \text{clip}\left(N(0, \sigma), -c, c\right)$ denotes the noise trimmed according to the normal distribution,which can avoid the overestimation problem. Then, the loss function is used to train the two Critic networks, which is expressed as $L(\varphi_i) = N^{-1}\sum\left(y - Q_{\varphi_l}(s, a)\right)^2$. The details of TD3-based UAV Data Collection are shown in Algorithm 1.

Algorithm 1. TD3-based UAV Data Collection

Initialize Critic and actor networks $Q_{\varphi_1}, Q_{\varphi_2}$, π_ϑ with random φ_1, φ_2 and ϑ;
Initialize target networks $\varphi_1^{'} \leftarrow \varphi_1$,$\varphi_2^{'} \leftarrow \varphi_2$ and $\vartheta^{'} \leftarrow \vartheta$;
Initialize replay buffer and learning rate α;
for $episode = 1 : episode_{max}$ **do**
  Set $t = 1$, $e_t^{co} = e_{max}$, observe the initial state s_t;
  repeat
    Select action with exploration noise $a_t \sim \pi_\vartheta(s_t) + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$;
    Execute the action a_t, calculate the optimal IRS phase shift matrix Θ_t^*, update the UAV position q_{t+1}^U, energy e_t^{re}, and AoI $A_t^i\,(i = 1, 2, \ldots, I)$;
    Obtain the reward r_t and the new state s_{t+1}, store experience (s_t, a_t, r_t, s_{t+1}) in replay buffer;
    Sample a mini-batch of N transitions from replay buffer;
    Update the Critic networks $\varphi_l = \underset{\varphi_l}{\operatorname{argmin}} L(\varphi_l), l = 1, 2$;
    if $t \bmod t_{update}$ **then**
      Update the Actor network $\vartheta = \vartheta - \alpha \nabla_\vartheta J(\vartheta)$;
      Update target networks: $\varphi_l^{'} \leftarrow \tau\varphi_l + (1-\tau)\varphi_l^{'}, l = 1, 2$, $\vartheta^{'} \leftarrow \tau\vartheta + (1-\tau)\vartheta^{'}$;
    end if
  until $e_t^{re} < e_{th}$;
end for

4 Simulation Results

We consider a 300 m × 400 m rectangular target area, and set the lower left corner of the area as the coordinate origin. The coordinate of IRS is set as $[0, 150, 30]$, and the horizontal coordinate of three SNs are set as $[10, 180]$, $[85, 350]$, $[225, 50]$. The random sampling process is modeled as a Poisson process. The system simulation parameters are set as follows: $T_{ts} = 1\,s$, $T_s = 0.5\,s$, $H_U = 60\,\text{m}$, $\beta_0 = -45\,\text{dB}$,$\alpha_1 = 3.1$, $\alpha_2 = 2.3$, $\sigma^2 = -110\,\text{dBm}$, $\omega = 110\,\text{KB}$, $\delta = 0.001$, $\eta_{th} = 0.77$, $v_{max}^U = 40\,\text{m/s}$, $e_{max} = 1.2\,e5\text{J}$, $e_{th} = 8\,e3\text{J}$. If there is no specific explanation, the reinforcement learning parameters are shown in Table 1.

Table 1. Learning parameters

Parameter	Value	Parameter	Value
Learning rate for actor	1e-4	Learning rate for critic	1e-3
Exploration noise	0.1	Policy noise	0.15
Software update factor	0.004	Reward discount	0.98
Total number of training episodes	6e4	Batch size	128

Figure 2 shows the convergence curves of the proposed TD3 and PPO algorithms, when fixed sampling with rate 0.2 is applied and the RIS is 15×15. It is observed that the TD3 algorithm converges faster and more stably, and is more

suitable for our studied problem. This is because the PPO algorithm tends to have insufficient explorations and usually find a suboptimal policy.

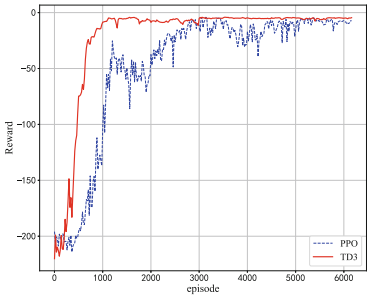
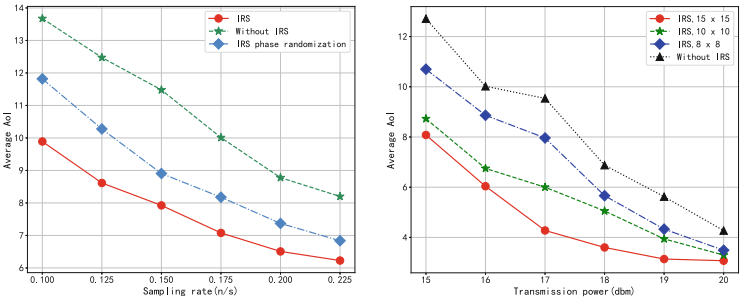


Fig. 2. Convergence comparison between TD3 and PPO algorithms.

In Fig. 3(a) and Fig. 3(b), we illustrate the average AoI performance of the proposed TD3-based method for different random sampling rates and transmission powers, respectively. As shown in Fig. 3(a), given any IRS phase control policy, it is observed that a higher sampling rate leads to the smaller average AoI, since the sensing data can be collected more frequently. By optimal phase alignment for IRS, our proposed scheme achieves the minimum AoI value for any sampling rate, which indicates that the IRS-aided UAV data collection scheme can significantly improve the information freshness. In Fig. 3(b), as either the number of IRS reflecting elements or the transmission power of the SN increases, the average AoI can be greatly reduced. In both subfigures, the transmission scheme without IRS leads to the highest AoI of SNs.



(a) IRS phase control (b) The number of IRS elements

Fig. 3. The average AoI performance of the proposed TD3 method

5 Conclusion

This paper investigated the efficient IRS-assisted UAV data collection problem for IoT. The IRS was deployed on a tall building to improve the channel quality of the UAV and SNs, and the optimal IRS reflection coefficient was obtained by phase alignment. The problem was modeled as a MDP problem, and we proposed the TD3 algorithm in DRL to find to find the optimal UAV trajectory and SN scheduling policy to minimizes the average AoI. Simulation results showed that integrating IRS into UAV data collection can effectively reduce the average AoI regardless of whether the SNs periodically or randomly sample environmental information, and the TD3 algorithm outperforms the PPO algorithm in terms of convergence speed and stability after convergence for the problems in this paper. The larger the IRS the lower the transmission power of the SN with guaranteed average AoI.

References

1. Kaul, S., Yates, R., Gruteser, M.: Real-time status: how often should one update? In: IEEE INFOCOM, pp. 2731–2735 (2012)
2. Abd-Elmagid, M.A., Ferdowsi, A., Dhillon, H.S., Saad, W.: Deep reinforcement learning for minimizing age-of-information in UAV-assisted networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2019)
3. Tong, P., Liu, J., Wang, X., Bai, B., Dai, H.: Deep reinforcement learning for efficient data collection in UAV-aided Internet of Things. In: IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–6 (2020)
4. Wu, Q., Zhang, S., Zheng, B., You, C., Zhang, R.: Intelligent reflecting surface-aided wireless communications: a tutorial. IEEE Trans. Commun. **69**(5), 3313–3351 (2021)
5. Cai, Y., Wei, Z., Hu, S., Ng, D., W. K., Yuan, J.: Resource allocation for power-efficient IRS-assisted UAV communications. In: IEEE International Conference on Communications Workshops (ICC Workshops), pp. 1–7 (2020)
6. Nguyen, K.K., Masaracchia, A., Sharma, V., Poor, H.V., Duong, T.Q.: RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning. IEEE J. Sel. Top. Sig. Process. **16**(5), 1086–1096 (2022)
7. Fujimoto, S., Hoof, H.V., Meger, D.: Addressing function approximation error in actor-critic methods. arXiv:1802.09477v3 (2018)
8. Al-Hourani, A., Kandeepan, S., Jamalipour, A.: Modeling air-to-ground path loss for low altitude platforms in urban environments. In: IEEE Global Communications Conference, pp. 2898–2904 (2014)
9. Wu, Q., Zhang, R.: Intelligent reflecting surface enhanced wireless network: joint active and passive beamforming design. In: IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2014)