Maximizing Age-Energy Efficiency in Wireless Powered Industrial IoE Networks: A Dual-Layer DQN-Based Approach

Haina Zheng, *Member, IEEE*, Ke Xiong, *Member, IEEE*, Mengying Sun, *Member, IEEE*, Huaqing Wu, *Member, IEEE*, Zhangdui Zhong, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*— This paper investigates the age of information (AoI) and energy efficiency of wireless powered industrial Internet of Everything (IIoE) network, where multiple low-power IIoE devices (IIoEDs) are wirelessly charged by a hybrid access point (HAP) to transmit their sensing information to the control nodes. To enhance the system's information timeliness with high energy efficiency, we define a novel performance metric, i.e., age-energy efficiency (AEE), which depicts the achievable AoI gain per unit energy consumption. Then, an optimization problem is formulated to maximize the system long-term AEE by jointly optimizing the IIoEDs scheduling and the HAP's transmit power. Due to the non-convexity of the formulated problem and the intractable challenges with discrete binary variables, we first model the problem as a two-stage discrete-time Markov decision process (MDP) with carefully designed state spaces, action spaces, and reward functions. We then propose a deep reinforcement learning (DRL)-based approach to find the effective scheduling strategy and transmit power. To improve the accuracy of the learned policy, we design a dual-layer deep Q-network (DLDQN) algorithm with fast convergence. Simulation results show that our proposed DLDQN algorithm can improve the AEE by at least 25% when the number of IIoEDs exceeds 50 compared with benchmarks. Moreover, with the proposed DLDQN algorithm, the system long-term AEE can be improved with the increase of the number of IIoEDs.**

***Index Terms*— Industrial Internet of Everything (IIoE), wireless power communication network (WPCN), age of information (AoI), energy efficiency (EE), deep reinforcement learning (DRL), dual-layer deep Q-network.**

I. Introduction

WITH the rapid evolution of industrial technologies and the beyond fifth-generation (B5G)/sixth-generation (6G) networks, a tremendous number of small-form industrial sensing devices are expected to be connected to construct industrial Internet of Everything (IIoE) [1]. Since IIoE devices (IIoEDs) usually are required to frequently sample [2] and transmit a lot of real-time update information for supporting the emerging real-time applications, the limited battery capacity becomes a challenging issue to the wide deployment of IIoEDs [3].

Thanks to the advance of wireless charging technologies, especially the radio frequency (RF)-based wireless power transfer (WPT) [4], IIoEDs are able to prolong their battery lifetimes with sustainable energy harvest (EH) [5]. On the other hand, with the explosive growth of IIoEDs, the energy consumption in IIoE networks increases dramatically, which is inconsistent with the green communication requirements of B5G/6G networks [6], [7]. To address this problem, energy efficiency (EE) in low-power IIoE network design has been widely investigated in both industry and academia [8]. Additionally, many emerging intelligent industrial applications (e.g., smart factory, smart plant and smart supply chain [9]) supported by IIoEDs commonly have high requirements on information timeliness [10], [11]. In such applications, outdated information may seriously degrade their performance, and even cause control decision errors or serious malfunctions [12]. To depict the information freshness, age of information (AoI) has been proposed as a novel fundamental performance index for emerging status update applications [13], [14].

Recently, AoI and EE have been widely studied for various wireless networks, including the wireless powered communication networks (WPCNs) [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25]. Specifically, in [17], the theoretical bounds of the average AoI for practical WPT networks were analyzed and derived as functions of the capacitor's size.

Manuscript received 25 March 2022; revised 12 August 2022, 5 December 2022, and 20 March 2023; accepted 4 June 2023. Date of publication 26 June 2023; date of current version 13 February 2024. This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2022JBQY004 and in part by the National Natural Science Foundation of China under Grant 62071033 and Grant 62221001. The associate editor coordinating the review of this article and approving it for publication was L. Duan. *(Corresponding author: Ke Xiong.)*

Haina Zheng and Ke Xiong are with the Engineering Research Center of Network Management Technology for High Speed Railway of Ministry of Education, School of Computer and Information Technology, the Collaborative Innovation Center of Railway Traffic Safety, the National Engineering Research Center of Advanced Network Technologies, and the Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China (e-mail: kxiong@bjtu.edu.cn).

Mengying Sun is 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).

Huaqing Wu is with the Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: huaqing.wu1@ucalgary.ca).

Zhangdui Zhong is with the State Key Laboratory of Rail Traffic Control and Safety and the Beijing Engineering Research Center of High-Speed Railway Broadband Mobile Communications, Beijing Jiaotong University, Beijing 100044, China (e-mail: zhdzhong@bjtu.edu.cn).

Xuemin Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: sshen@uwaterloo.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2023.3287885.

Digital Object Identifier 10.1109/TWC.2023.3287885

In [18], the authors studied the impact of the data packet size on the system average AoI in WPCNs, and a closed-form expression of the average AoI was derived. In [19], a sampling and updating policy to minimize the average AoI of WPCNs was investigated for a single source-destination pair. In [20], the average AoI of wireless powered relay aided networks with multiple sources was studied. In [21], the authors proposed an AoI-constrained scheduling strategy in WPCNs, and the impact of the packet arrival time on the system AoI was revealed. In [22], an AoI-based system utility was maximized in WPCNs with the energy regarded as a cost. The peak AoI minimization problem was studied in [23] for wireless powered cooperative networks, revealing the dominant impact of power-limited relays on the system performance. In [24], the EE of wireless powered sensor networks was maximized with AoI constraints, where both AoI and EE were modeled to find the optimal solution. In [25], the age-energy tradeoff was studied for WPCNs, in which a weighted AoI and energy consumption model was structured and the age-energy region was explored.

In general, wireless network design problems, such as power allocation and device scheduling, may be computational tractable by model-based approaches for small-scale networks. To realize real-time decision-making with low computational complexity, efficient model-free methods, such as deep reinforcement learning (DRL), have been utilized for large-scale wireless network designs [26], [27], [28]. Specifically, with deep neural networks (DNNs), DRL is able to learn from empirical data and make the effective decision rapidly, which has been regarded as an effective tool in dealing with mathematically intractable problems. On the other hand, for the traditional optimization method, only the static/quasi-static wireless environments are considered. The time-varying features of wireless channels make it challenging to design the power allocation and device scheduling schemes, as they have to be solved within the channel coherence time [29]. Therefore, DRL-based approaches have been widely applied in various wireless network designs.

The DRL-based approach has also been applied for real-time decision-making in WPCNs. For example, in [30], [31], and [32], the DRL-based adaptive power control strategies were investigated for WPCNs, where the EE was maximized for multi-node networks and relay-assisted networks, respectively. In [33], an online computation offloading strategy was studied in WPCNs via designing a DRL-based framework, in which an online deep Q-network (DQN) algorithm was implemented as a scalable solution. The authors in [34] proposed a data collection strategy to minimize the average AoI in RF-powered communication systems, where a DQN-based scheme was designed to find the near optimal solution. Similarly, in [35], an AoI-minimization problem was investigated with the energy involved as a weighted cost. The aforementioned works show that DRL has great potential in dealing with mathematically intractable problems for large-scale network design.

Different from existing works where AoI and EE were often separately investigated, in future wireless networks, lower AoI and higher EE are required at the same time. Therefore, how to improve AoI and EE simultaneously becomes a new challenge. In this work, we focus on *how to enhance the system's information timeliness with high EE, which will be a fundamental problem for future IIoE network design.* However, high information freshness (i.e., small AoI) depends on frequent sensing and fast transmitting, and requires excessive energy consumption [36]. This inherent tradeoff between AoI and energy consumption makes it non-trivial to solve our focused problem [37].

In this paper, we consider a wireless powered IIoE network, where multiple low-power IIoEDs are wirelessly charged by a hybrid access point (HAP) to transmit their sensing information to the control nodes (CNs). To enhance AoI with high EE, a novel metric, i.e., age-energy efficiency (AEE), is defined to evaluate the contribution of per-unit energy consumption to the achieved AoI gain. Based on the newly defined efficient performance metric, the long-term AEE is maximized by jointly optimizing the IIoEDs scheduling and the HAP's transmit power.

The main contributions of this work are summarized as follows.

- ***First,*** a novel performance metric, i.e., AEE, is defined in this work to characterize the achievable AoI gain per unit energy consumption. AEE takes both AoI and EE into account to evaluate the information timeliness and energy utilization performance of wireless networks.
- ***Second,*** to maximize the long-term AEE of the wireless powered IIoE network, we formulate an optimization problem by jointly optimizing the scheduling strategy and the HAP's transmit power, subject to the battery capacity constraint, the signal-to-noise-ratio (SNR) threshold constraint, the aging threshold constraint, and the transmit power threshold constraint.
- ***Third,*** to efficiently solve the formulated non-convex problem with discrete and continuous variables, we model it as a two-stage discrete-time Markov decision process (MDP) with carefully designed state spaces (including the normalized AoI, the IIoED's remaining energy, and the HAP's energy consumption), action spaces (including the charging or scheduling process, and the HAP's transmit power selection), and reward functions for state-action pairs. Then, we propose a DRL-based approach, where a dual-layer DQN (DLDQN) algorithm is designed to improve the accuracy of the learned policy. Specifically, the outer DQN is designed to find the effective scheduling strategy, and the inner one is designed to determine the optimal transmit power.
- ***Fourth,*** we theoretically prove that our proposed DLDQN algorithm has fast convergence speed compared with conventional optimization methods. This implies that our proposed algorithm is suitable for enhancing the system long-term AEE of real-time status update networks, especially in the highly dynamic wireless environment.

Simulation results show that, our proposed DLDQN-based algorithm can improve the AEE by at least 25% when the number of IIoEDs exceeds 50 compared with the benchmarks, such as the fixed power algorithm, the round robin algorithm, and the random algorithm. Moreover, with the proposed DLDQN

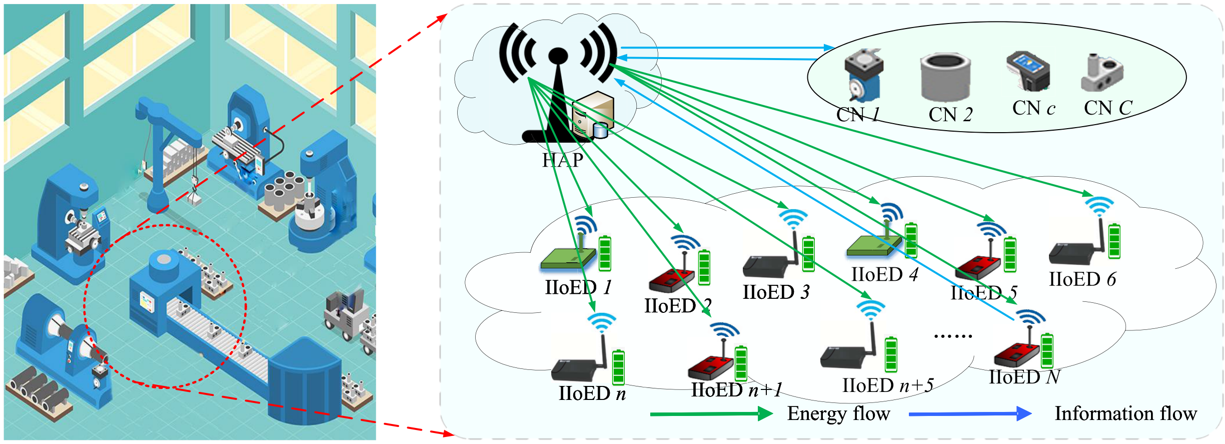


Fig. 1. Illustration of a wireless powered IIoE network.

algorithm, the system long-term AEE can be improved with the increase of the number of IIoEDs. In addition, the impacts of different system parameters (e.g., the distance between HAP and IIoEDs, the HAP's transmit power threshold, the update packet length, the energy buffer size, and the number of epoches) on the long-term AEE are also discussed and evaluated.

The remaining of this paper is organized as follows. In Section II and Section III, we describe the system model and the problem formulation, respectively. In Section IV, the modeled two-stage MDP formulation is presented, and in Section V, the proposed DLDQN algorithm framework is presented. In Section VI, simulation results are provided, followed by the conclusion in Section VII.

II. System Model

A. Network Model

As shown in Fig. 1, we consider a wireless powered IIoE network consisting of a HAP, a set of IoEDs, and a set of CNs, where the IIoEDs are denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$ with n representing the n-th IIoED, and the CNs are denoted by $\mathcal{C} = \{1, 2, \ldots, C\}$ with c representing the c-th CN. The IIoEDs are randomly deployed to sense the surrounding environment information (such as image, temperature, humidity and pressure), and the CNs are deployed as execution nodes to serve certain applications by making real-time industrial control decisions (such as fault warning, device scheduling and smart monitoring) based on the received update sensed by IIoEDs.

The HAP[1] is dedicatedly deployed by service provider to wirelessly charge the IIoEDs, schedule the IIoEDs according to CNs' requests, and broadcast the sensing update packets to CNs. In real IoE networks, as IIoEDs are generally with relatively small physics sizes, low costs, and large-scale deployments, we consider that each IIoED is equipped with a single antenna and employs the "harvest-then-transmit" protocol to avoid the mutual interference between energy transfer and information transmission. Moreover, similar to many existing works [33], [34], we employ the popularly deployed time division multiple access (TDMA) protocol in real IoE networks, where only one IIoED is scheduled in each epoch.[2]

B. Transmission Model

The transmission protocol is shown in Fig. 2. We consider one system operation time frame T, which is also the maximal tolerable response time for all CNs. The time period T is slotted into a sequence of epochs, i.e., $\mathcal{K} = \{1, 2, \ldots, K\}$, with equal time interval of $t = \frac{T}{K}$. At epoch k, either the WPT or the scheduling process of one IIoED is performed at the HAP. Note that t is set to be smaller than the wireless channel coherence time. Hence, at each epoch, the channel coefficient is regarded as unchanged, while from the current epoch to the next epoch, the channel coefficient varies independently. At epoch k, we denote the WPT index by $v(k) \in \{0, 1\}$, where $v(k) = 1$ means that the HAP charges IIoEDs via WPT at epoch k, and $v(k) = 0$ means that the WPT process does not occur. The scheduling index is denoted by $o_n(k) \in \{0, 1\}$. When the HAP schedules the sensing update packet generated by IIoED n at epoch k, $o_n(k) = 1$; otherwise, $o_n(k) = 0$. Note that at epoch k, it satisfies that

$$v(k) + \sum\nolimits_{n \in \mathcal{N}} o_n(k) = 1, \forall k \in \mathcal{K}. \tag{1}$$

[1] Since the CN desire to collect status update packets from different IIoEDs, the HAP is deployed and integrated with a central controller to coordinate the IIoED scheduling and also broadcast the sensing update packets to CNs. Specifically, it collects the global channel state information (CSI) and the transmission-related information of IIoED, and hence coordinates the WPT and the scheduling. To explore the system performance limit, perfect CSI is assumed in this paper. Compared to imperfect CSI, the performance achieved by our work can be regarded as an upper bound and used as a guideline to design the wireless networks.

[2] Although by deploying some advanced multiple access protocols, such as orthogonal frequency division multiple access (OFDMA), non-orthogonal multiple access (NOMA), and sub-carrier multiple access (SCMA), the better system performance may be achieved, such protocols require the IIoEDs to have high synchronization and decoding capabilities. For instance, OFDMA requires the IIoEDs to meet strict synchronization requirements. NOMA and SCMA require the IIoEDs to have strong decoding capabilities, which brings relatively high hardware requirements to the IIoEDs. Thus, in this work, we also consider the TDMA protocol due to its simplicity and popularity for deployment.

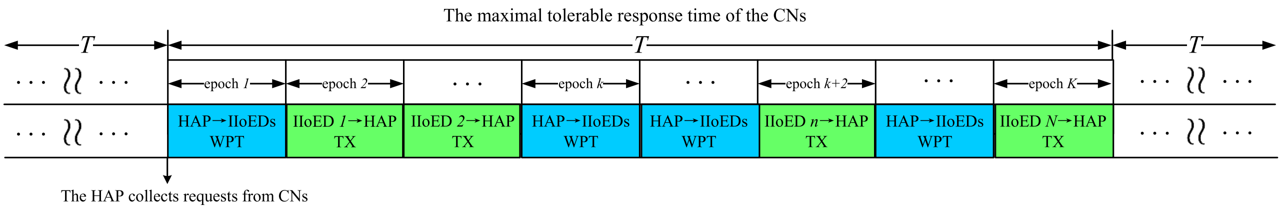


Fig. 2. Illustration of the transmission protocol.

C. EH Model

Consider the case where the HAP performs the WPT process by broadcasting RF signals at epoch k. Then, IIoEDs store and accumulate[3] the harvested energy from RF-signal in their energy buffer. Denote the HAP's transmit power and the wireless channel coefficient[4] between the HAP and the IIoEDs by $P_{\text{HAP}}(k)$ and $h_n(k)$, respectively. Via RF-based WPT, the stored energy in IIoED n's energy buffer at epoch k is

$$E_n^{\text{e}}(k) = \min\left\{\eta_{\text{eh}} P_{\text{HAP}}(k)|h_n(k)|^2 t, M_{\text{eh}} t\right\}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{2}$$

where the non-linear EH model [41] is employed. $\eta_{\text{eh}} \in (0,1]$ and M_{eh} are the energy conservation efficiency and the saturation threshold of IIoEDs' EH circuits, respectively. Thus, at epoch k, the energy accumulated in IIoED n's energy buffer is given by

$$E_n^{\text{s}}(k) = \min\left\{E_n^{\text{r}}(k) + v(k)E_n^{\text{e}}(k), B_n\right\}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{3}$$

where $E_n^{\text{r}}(k)$ is the remaining energy stored in IIoED n's energy buffer at the beginning of k, and B_n is the capacity size of IIoED n's energy buffer. Generally, IIoED is equipped with a small energy buffer, such as capacitor, due to its low cost and small size. In practice, only when $E_n^{\text{s}}(k)$ reaches IIoED n's trigger threshold, i.e., energy capacity threshold[5] B_n, IIoED n is able to be triggered to stop harvesting energy and wait to be scheduled to generate and deliver its sensing update packet.[6]

D. Information Delivery Model

Suppose at epoch k, IIoED n is scheduled, and it will generate and deliver its sensing update packet to the HAP within epoch k. The received information signal at the HAP from IIoED n is expressed as

$$y_n(k) = \sqrt{P_n(k)} h_n(k) x_n(k) + n(k), \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{4}$$

where $x_n(k) \in \mathbb{C}$ with $\mathbb{E}\left\{|x_n(k)|^2\right\} = 1$ is the information symbol sent by IIoED n at epoch k. $P_n(k)$ is the IIoED n's transmit power at epoch k. $n(k) \sim \mathcal{CN}(0, \sigma^2)$ represents the Additive White Gaussian Noise (AWGN) at the HAP with σ^2 being the receiver noise power. Accordingly, at epoch k, the received signal-to-noise-ratio (SNR) at the HAP from IIoED n is given by

$$\gamma_n(k) = \frac{P_n(k)|h_n(k)|^2}{\sigma^2}, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{5}$$

To ensure the success of information decoding, when the HAP schedules the sensing update packet generated by IIoED n at epoch k, its received SNR has to surpass the pre-specified threshold γ_{th}. Hence, it satisfies that

$$\gamma_n(k) \geq \gamma_{\text{th}}, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{6}$$

Correspondingly, at epoch k, the achievable information rate at the HAP from IIoED n is given by

$$R_n(k) = W \log_2\left(1 + \gamma_n(k)\right), \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{7}$$

where W is the system bandwidth. Let L_n denote the length of the update packet generated by IIoED n. In order to successfully deliver the update packet to the HAP from IIoED n at epoch k, it satisfies that

$$R_n(k) t \geq L_n, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{8}$$

Following (5), (7) and (8), the minimal required energy to successfully deliver an update packet for IIoED n at epoch k is given by[7]

$$E_n^{\text{t}}(k) = \frac{\sigma^2 t\left(2^{\frac{L_n}{tW}} - 1\right)}{|h_n(k)|^2}, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{9}$$

Then, the remaining energy stored in IIoED n's energy buffer at the end of k reduces to

$$E_n^{\text{r}}(k) = B_n - E_n^{\text{t}}(k), \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{10}$$

and IIoED n is triggered to restart the energy accumulation.

[3] For the WPT-based system, if it does not accumulate energy, the system will not have enough energy to transmit. Therefore, accumulating energy to enable transmission is the first thing needed to be solved.

[4] The channel is considered to be reciprocal in the downlink energy transfer and uplink information delivery [38], [39], [40].

[5] Although when the accumulated energy reaches the transmit energy, triggering the scheduling is also a feasible way to run the system, in practice, whether the transmit energy is enough for delivering a packet is hard to determine. Moreover, due to the dynamic wireless channel status, deciding how to transmit based on the channel status at each time epoch will bring excessive additional overhead of accurate perception of the channel estimation. In our studied IoE scenarios, the IIoED is with low power consumption and small size, so they may not have sufficient channel perception capability. To this end, we employ a simple fixed buffer threshold as the trigger condition to trigger the scheduling, similar to many existing works [33], [34] in low-power IoE networks.

[6] For many passive IoE devices, their sensing energy consumption usually is much less than the transmission energy consumption. For the case that the sensing energy consumption cannot be ignored, similar to [42] and [43], we define the sensing energy consumption as a constant E^{a}, where E^{a} is a static circuit maintenance constant. Then, by putting the constant E^{a} into the construction of the system model and the formulation of the problem, our proposed approach is still applicable.

[7] Note that we consider that the minimal required energy E_n^{t} for transmitting one update packet is generally less than the pre-determined IIoED n's energy capacity threshold B_n, so each transmission is guaranteed to be successful finished.

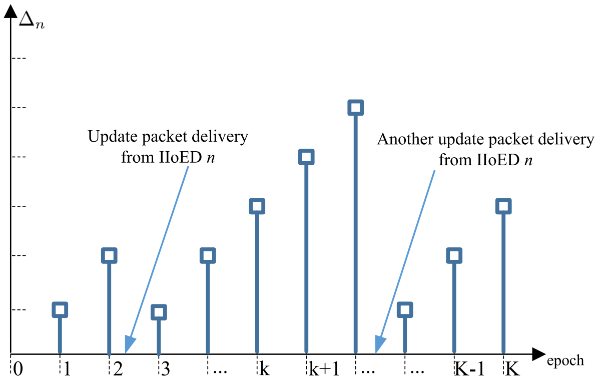


Fig. 3. Illustration of the AoI for IIoED n.

E. AoI Model

To capture the timeliness of the received sensing update packet, the AoI metric is employed to depict how fresh the information is. At epoch k, the AoI of the current update packet generated by IIoED n at epoch k is given by

$$\Delta_n(k) = k - g_n + 1, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{11}$$

where g_n $(g_n \leq k)$ is the generation and delivery time of the current update packet of IIoED n. For example, at epoch k, when $g_n < k$, it means that the current update packet of IIoED n is generated and delivered at epoch g_n, and the value of $\Delta_n(k)$ increases to $(k - g_n + 1)$. When $g_n = k$, it means that the current update packet of IIoED n is generated and delivered at epoch k, and the value of $\Delta_n(k)$ reduces to 1. As shown in Fig. 3, the value of $\Delta_n(k)$ increases discretely until the next sensing update packet delivered to the HAP.

To ensure that the information from IIoED n is fresh, it satisfies that

$$\Delta_n(k) \leq \widehat{\Delta}_n, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{12}$$

where $\widehat{\Delta}_n$ represents the tolerable information aging threshold from IIoED n. It means that before the AoI of the update packet generated by IIoED n grows to $\widehat{\Delta}_n$, the corresponding packet must be scheduled by the HAP.

We denote the number of requests for IIoED n's sensing update packets from all of CNs by M_n. Based on [44], [45], and [46], M_n is regarded as the weight factor of IIoED n's AoI, which means the importance of the IIoED n's AoI. Hence, the weighted sum of all IIoEDs' AoI at epoch k is given by

$$\Delta(k) = \frac{\sum_{n\in\mathcal{N}} M_n \Delta_n(k)}{\sum_{n\in\mathcal{N}} M_n}, \forall k \in \mathcal{K}. \tag{13}$$

F. AEE Model

Similar to the EE defined as the ratio between the achievable information rate and the required energy, we define the AEE as the ratio between the achievable AoI gain and its required energy consumption,[8] i.e.,

$$\Phi = \frac{\widehat{\Delta} - \Delta}{E}, \tag{14}$$

where $\widehat{\Delta}$ is the maximum tolerable information aging of the considered system, Δ is the system achievable AoI, and their difference is the achievable AoI gain of the system. E is the total required energy consumption of the system. As shown in (14), the system long-term AEE can be improved by increasing the AoI gain and reducing total energy consumption. For IIoED n, its long-term average AEE over whole K epochs is expressed as

$$\Phi_n(\boldsymbol{P}_{\text{HAP}}, \boldsymbol{v}, \boldsymbol{O}) = \frac{\sum_{k\in\mathcal{K}} \left(\widehat{\Delta}_n - \Delta_n(k)\right)}{K \sum_{k\in\mathcal{K}} o_n(k) E_n^{\text{t}}(k)}, \forall n \in \mathcal{N}. \tag{15}$$

For the system, its long-term average AEE over whole K epochs is expressed as

$$\Phi_{\text{HAP}}(\boldsymbol{P}_{\text{HAP}}, \boldsymbol{v}, \boldsymbol{O}) = \frac{\sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} M_n \left(\widehat{\Delta}_n - \Delta_n(k)\right)}{K \sum_{n\in\mathcal{N}} M_n \sum_{k\in\mathcal{K}} v(k) P_{\text{HAP}}(k) t}. \tag{16}$$

III. Problem Formulation

For the considered system, our goal is to maximize its long-term average AEE. We formulate an optimization problem by jointly optimizing the IIoED scheduling and the HAP's transmit power. Mathematically, the AEE maximization problem is expressed by

$$\mathbf{P}_1 : \max_{\boldsymbol{P}_{\text{HAP}}, \boldsymbol{v}, \boldsymbol{O}} \Phi_{\text{HAP}}\left(\boldsymbol{P}_{\text{HAP}}, \boldsymbol{v}, \boldsymbol{O}\right) \tag{17}$$

$$\text{s.t.} \quad E_n^{\text{s}}(k) \leq B_n, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17a}$$

$$o_n(k) E_n^{\text{t}}(k) \leq B_n, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17b}$$

$$o_n(k) \gamma_n(k) \geq \gamma_{\text{th}}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17c}$$

$$\sum\nolimits_{n\in\mathcal{N}} o_n(k) + v(k) = 1, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17d}$$

$$o_n(k) \in \{0, 1\}, v(k) \in \{0, 1\}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17e}$$

$$\Delta_n(k) \leq \widehat{\Delta}_n, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{17f}$$

$$0 \leq v(k) P_{\text{HAP}}(k) \leq P_{\text{HAP}}^{\max}, \forall k \in \mathcal{K}, \tag{17g}$$

where $\mathbf{P}_{\text{HAP}} = [P_{\text{HAP}}(1), \ldots, P_{\text{HAP}}(K)]$ is the HAP's transmit power during T, $\boldsymbol{v} = [v(1), \ldots, v(K)]$ is the WPT process vector at the HAP during T, and $\mathbf{O} = [\mathbf{o}(1), \ldots, \mathbf{o}(K)]$ is the IIoEDs scheduling process vector during T with $\mathbf{o}(k) = [\mathbf{o}_1(k), \ldots, \mathbf{o}_N(k)]$ being the scheduling index vector of IIoEDs at epoch k. Constraint (17a) is the energy capacity threshold constraint, which means that the energy accumulated in IIoED n's energy buffer should not exceed its capacity threshold. Constraint (17b) is the energy causality constraint, which means that for IIoED n, the energy used for transmitting

[8] Although the multi-objective optimization may also pursue the AoI gain maximization and the energy consumption minimization simultaneously, it cannot effectively and clearly characterize the contribution of per Joule energy to the AoI gain in a network system. As a matter of fact, reducing AoI usually requires consuming more energy. Since the wireless system is energy-constrained, the benefits brought by consuming the same energy under different system states are also different. By defining AEE, this difference and changing laws of AoI gain versus energy could be better described.

update packet should not exceed its stored energy in energy buffer. Constraints (17c), (17f), and (17g) are the received SNR from IIoED n, the tolerable information aging for IIoED n, and the HAP's transmit power threshold, respectively. Constraints (17d) and (17e) indicate that the HAP performs either the WPT process to charge the IIoEDs or the IIoEDs scheduling process to schedule only one IIoED to deliver its generated sensing update packet to the HAP at epoch k.

Problem $\mathbf{P}_1$ is a non-convex mixed integer programming problem with high complexity. Specifically, one needs to search among the $(N+1)^K$ possible choices to find the best WPT or scheduling strategy $\{\boldsymbol{v}^*, \boldsymbol{O}^*\}$, and search among the optional values to find the optimal HAP's transmit power $\mathbf{P}^*_{\text{HAP}}$. Due to the large search space, quickly solving such a combinatorial optimization problem is difficult to achieve via conventional methods, especially when the network expands. In view of the fact that the system long-term AEE usually depends on the time-varying wireless channel, we model Problem $\mathbf{P}_1$ as an MDP in the next section.

IV. MDP Formulation

In this section, we reformulate Problem $\mathbf{P}_1$ as a MDP. The state spaces, action spaces, and reward functions of the formulated MDP are designed as follows.

A. State Spaces

1) State Space for Stage 1: In stage 1, the HAP performs the WPT or IIoEDs scheduling process action. The state space includes the state of all IIoEDs, CNs, and the HAP, which is denoted by four elements. The first element of state space is the number of requests from CNs for the sensing update packet generated by IIoED n at epoch k, i.e., $\{M_n(k)\}_{n\in\mathcal{N},k\in\mathcal{K}}$. The second element of state space is the normalized AoI of the sensing update packet generated by IIoED n at epoch k, i.e., $\{\tilde{\Delta}_n(k)\}_{n\in\mathcal{N},k\in\mathcal{K}}$, where $\tilde{\Delta}_n(k)$ is given by

$$\tilde{\Delta}_n(k) = \frac{M_n\Delta_n(k)}{\sum_{n\in\mathcal{N}} M_n\Delta_n(k)}, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{18}$$

The normalized AoI is considered because only the rank of AoI from different IIoEDs instead of the value of them affects the IIoEDs scheduling order. The third element of state space is the remaining energy stored in IIoED n's energy buffer at epoch k, i.e., $\{E^{\text{r}}_n(k)\}_{n\in\mathcal{N},k\in\mathcal{K}}$. The fourth element of state space is the HAP's total energy consumption at epoch k, i.e., $\{E^{\text{t}}_{\text{HAP}}(k)\}_{k\in\mathcal{K}}$.

In a word, the state of all IIoEDs, CNs and the HAP is denoted by

$$\boldsymbol{S}^{\text{stg1}} = \{\boldsymbol{S}(1), \boldsymbol{S}(2), \ldots \boldsymbol{S}(k), \ldots \boldsymbol{S}(K)\}, \tag{19}$$

where the state matrix $\boldsymbol{S}(k)$ at epoch k is given by

$$\boldsymbol{S}(k) = \{\boldsymbol{s}_1(k), \ldots, \boldsymbol{s}_n(k), \ldots, \boldsymbol{s}_N(k), s_{\text{HAP}}(k)\}, \forall k \in \mathcal{K}, \tag{20}$$

in which $\boldsymbol{s}_n(k) = \{M_n(k), \tilde{\Delta}_n(k), E^{\text{r}}_n(k)\}_{n\in\mathcal{N},k\in\mathcal{K}}$ and $s_{\text{HAP}}(k) = \{E^{\text{t}}_{\text{HAP}}(k)\}_{k\in\mathcal{K}}$.

At the beginning of each epoch, the HAP has the exact knowledge of all IIoEDs' and CNs' state information in this epoch.

2) State Space for Stage 2: When the HAP selects the WPT process action in stage 1, the transmit power selecting action needs to be performed in stage 2. The power selection affects the levels of IIoEDs' energy buffers at the current state and the normalized AoIs of the sensing update packets generated by IIoEDs at the next state. Therefore, in stage 2, the state space at epoch k includes the normalized AoIs of the sensing update packet generated by all IIoEDs and the levels of all IIoEDs' energy buffers, which is expressed as

$$\boldsymbol{S}^{\text{stg2}} = \left\{\{\tilde{\Delta}_n(k)\}, \{I^{\text{full}}_n(k)\}\right\}, \forall n \in \mathcal{N}, k \in \mathcal{K}, \tag{21}$$

where $I^{\text{full}}_n(k)$ indicates whether the IIoED n's energy buffer is full at epoch k, and is expressed as

$$I^{\text{full}}_n(k) = \begin{cases} 1, \text{if } E^{\text{r}}_n(k) = B_n, \forall n \in \mathcal{N}, k \in \mathcal{K} \\ 0, \text{otherwise}, \end{cases} \tag{22}$$

where $I^{\text{full}}_n(k) = 1$ means that the energy stored in IIoED n's energy buffer is full at epoch k. When $I^{\text{full}}_n(k) = 0$, IIoED n cannot be triggered to generate and be scheduled to deliver an update packet.

B. Action Spaces

1) Action Space for Stage 1: The action space for stage 1 includes the HAP's decisions to perform the WPT or IIoEDs scheduling process. Specifically, the HAP's action at epoch k is expressed as

$$a^{\text{stg1}}(k) \in \{0, \mathcal{N}\}, \forall k \in \mathcal{K}. \tag{23}$$

The space size of $a^{\text{stg1}}(k)$ at epoch k is $N+1$. $a^{\text{stg1}}(k) \in \{\mathcal{N}\}$ means that the HAP performs the IIoEDs scheduling process action to schedule one IIoED to deliver its sensing update packet at epoch k, and $a^{\text{stg1}}(k) = 0$ means that the HAP performs the WPT process action.

Based on (18), the normalized AoI of the sensing information packet generated by IIoED n at the next state based on the current action is given by

$$\tilde{\Delta}_n(k+1) = \frac{M_n\left(\Delta_n(k) \times \mathcal{I}(n, a^{\text{stg1}}(k)) + 1\right)}{\sum_{n\in\mathcal{N}} M_n\left(\Delta_n(k) \times \mathcal{I}(n, a^{\text{stg1}}(k)) + 1\right)}, \tag{24}$$

$\forall n \in \mathcal{N}$, $k \in \mathcal{K}$, where $\mathcal{I}(\cdot)$ is a WPT or IIoEDs scheduling indicator function. When variable n equals to the value of $a^{\text{stg1}}(k)$ at epoch k, $\mathcal{I}(n, a^{\text{stg1}}(k)) = 0$, and it indicates that the HAP schedules IIoED n; Otherwise, $\mathcal{I}(n, a^{\text{stg1}}(k)) = 1$, and the WPT or one of the rest of $n-1$ IIoEDs scheduling process is performed by the HAP. The remaining energy in IIoED n's energy buffer at epoch $k+1$ based on the current action is

$$E^{\text{r}}_n(k+1) = \begin{cases} \min\{E^{\text{r}}_n(k) + E^{\text{e}}_n(k), B_n\}, \\ \qquad \text{if } a^{\text{stg1}}(k) = 0, \forall k \in \mathcal{K}, \\ B_n - E^{\text{t}}_n(k), \\ \qquad \text{if } a^{\text{stg1}}(k) = n, \forall n \in \mathcal{N}, k \in \mathcal{K}. \end{cases} \tag{25}$$

The HAP's energy consumption at epoch $k+1$ based on the current action is given by

$$E^{\text{t}}_{\text{HAP}}(k+1) = \begin{cases} E^{\text{t}}_{\text{HAP}}(k) + P_{\text{HAP}}(k)t, \\ \qquad \text{if } a^{\text{stg1}}(k) = 0, \forall k \in \mathcal{K}, \\ E^{\text{t}}_{\text{HAP}}(k), \\ \qquad \text{if } a^{\text{stg1}}(k) = n, \forall n \in \mathcal{N}, k \in \mathcal{K}, \end{cases} \tag{26}$$

where $E_{\text{HAP}}^{\text{t}}(k)$ is the HAP's transmission energy consumption at epoch k.

2) Action Space for Stage 2*:* The action space for stage 2, i.e., the HAP's transmit power selection, at epoch k is

$$a^{\text{stg2}}(k) \in \boldsymbol{P}_{\text{HAP}}, \tag{27}$$

where $\boldsymbol{P}_{\text{HAP}}$ is the power selection space with its size being D. Here, D means that we equally divide the HAP's maximum transmit power into D levels. Hence, the available transmit power selecting action for charging IIoEDs is given by

$$\boldsymbol{P}_{\text{HAP}} = \left[\frac{P_{\text{HAP}}^{\max}}{D}, \frac{2P_{\text{HAP}}^{\max}}{D}, \ldots, P_{\text{HAP}}^{\max}\right]. \tag{28}$$

C. Reward Functions

1) Reward Function for Stage 1*:* The reward function for stage 1 is affected by two factors: (1) the normalized AoI of the sensing update packet collected by HAP, and (2) the energy consumption of the HAP and the IIoEDs. Thus, the corresponding reward function at epoch K is designed as

$$R^{\text{stg1}}(K) = \frac{1}{K}\sum\nolimits_{k\in\mathcal{K}} \frac{\sum_{n\in\mathcal{N}} M_n\left(\widetilde{\widehat{\widetilde{\Delta}}}_n - \widetilde{\Delta}_n(k)\right)}{v(k)P_{\text{HAP}}(k)t\sum_{n\in\mathcal{N}} M_n}. \tag{29}$$

2) Reward Function for Stage 2*:* The reward function for stage 2 is designed by considering the tradeoff between AoI and energy consumption.[9] With a larger HAP's transmit power, more IIoEDs can be fully charged to be scheduled at the next epoch, leading to a lower AoI. However, this may also cause more potential energy wastage due to the limited IIoED energy buffer capacities. Therefore, the reward function at epoch k for the transmit power selection is designed as

$$R^{\text{stg2}}(k) = \sum\nolimits_{n\in\mathcal{N}}\left(I_n^{\text{full}}(k)\cdot\tilde{\Delta}_n(k) - \mu E_n^{\text{w}}(k)\right), \forall k \in \mathcal{K}, \tag{30}$$

where μ is the weight factor,[10] and $E_n^{\text{w}}(k)$ is the wasted energy of IIoED n at epoch k, which is given by

$$E_n^{\text{w}}(k) = \max\left\{P_{\text{HAP}}(k)t - E_n^{\text{r}}(k+1), 0\right\}, \forall n \in \mathcal{N}, k \in \mathcal{K}. \tag{31}$$

D. Transition Probability

The transition probability is the probability that the system transits from the current state $\left\{\boldsymbol{S}^{\text{stg1}}(k), \boldsymbol{S}^{\text{stg2}}(k)\right\}$ to the next state $\left\{\boldsymbol{S}^{\text{stg1}}(k+1), \boldsymbol{S}^{\text{stg2}}(k+1)\right\}$ with the HAP performing action $\left\{a^{\text{stg1}}(k), a^{\text{stg2}}(k)\right\}$ at epoch k. The transition probability is denoted by $\Pr\left(\left\{\boldsymbol{S}^{\text{stg1}}(k+1), \boldsymbol{S}^{\text{stg2}}(k+1)\right\}|\left\{\boldsymbol{S}^{\text{stg1}}(k), \boldsymbol{S}^{\text{stg2}}(k)\right\}, \left\{a^{\text{stg1}}(k), a^{\text{stg2}}(k)\right\}\right)$. As the transition probability is difficult to acquire in practical implementation, in the next section, we propose a DRL-based approach to find the effective scheduling strategy and transmit power for Problem $\mathbf{P}_1$.

V. Proposed DLDQN Algorithm

In this section, we present our proposed DRL-based approach, i.e., a DLDQN algorithm, based on the two-stage MDP to solve Problem $\mathbf{P}_1$.

A. Framework of Our Proposed DLDQN

For the considered system, the HAP is regarded as an agent, and the wireless power IIoE network scenario is regarded as the environment. We propose a DLDQN algorithm with a dual-layer[11] framework as shown in Fig. 4, where DQNs are employed for both the WPT or IIoED scheduling process and the HAP's transmit power selection. Specifically, in the outer layer, we utilize a DQN to train the IIoEDs scheduling strategy, where two DNNs, i.e., the online network and the target network, are trained with the same structure but different parameters. The online network is trained by updating the Q-value function that evaluates the expected cumulative reward after performing the current scheduling action in the current state. The target network is trained to predict realistic scheduling action. In the inner layer, we train another DQN for the HAP's transmit power selection, where two DNNs are utilized in a similar way to that in the outer-layer DQN. For both two DQNs, a mini-batch method is used to sample the historical experience stored in the Experience Memory (EM) for updating network parameters. Finally, the WPT or IIoEDs scheduling process action and HAP's transmit power selecting action are made to converge with all constraints listed in (17a)-(17g) of Problem $\mathbf{P}_1$ being satisfied. The designs of the two DQNs are detailed as follows.

B. Outer-Layer DQN for WPT or IIoEDs Scheduling

The outer-layer DQN is composed of two phases, i.e., the WPT or IIoEDs scheduling process action generation and the IIoEDs scheduling strategy update. Specifically, in phase 1, the generation of the WPT or IIoEDs scheduling process action relies on the Q-value function learned by a DNN. At epoch k, the Q-value function in the outer-layer DQN is

$$\begin{aligned} Q^{\text{ou}}(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k)|\theta^{\text{ou}}(k)) = \mathbb{E}\left[V^{\text{ou}}(k)|\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k),\right.\\ \left.\{\boldsymbol{v}^*(k), \boldsymbol{O}^*(k)\}\right], \forall k \in \mathcal{K}, \end{aligned} \tag{32}$$

where $\theta^{\text{ou}}(k)$ is the parameters of the outer-layer DQN at epoch k, e.g., the weights of connecting the hidden neurons. $V^{\text{ou}}(k)$ is the discounted cumulative reward of the scheduling process at epoch k, which is given by

$$V^{\text{ou}}(k) = \sum\nolimits_{i\in\mathbb{N}} \zeta^i R^{\text{stg1}}(k+i+1), \forall k \in \mathcal{K}, \tag{33}$$

[9] According to [47], a designer reward function is used to evaluate the agent, while a separate agent reward function is used to guide agent behavior, which guarantees that the agent learns to achieve the expected goal. In our proposed DLDQN algorithm, the reward function for stage 2 is designed to guide HAP's behavior directly, which guarantees that the HAP learns to achieve the maximal system long-term AEE.

[10] Actually, the weight factor has impact on the tradeoff between the normalized AoI and the wasted energy. When more attention is paid to the information freshness, the weight factor should be a small value; otherwise, when more attention is paid to the energy consumption, the weight factor should be a large value.

[11] Although single-layer DQN may also solve our considered problem, as the scheduling strategy and power selection are in a single action space, the action space is with high dimension and the influence relationship between the scheduling strategy and power selection is difficult to learn. Hence, we adopt the double-layer DQN to reduce the dimension of the action space and enhance the accuracy of the learned policy.

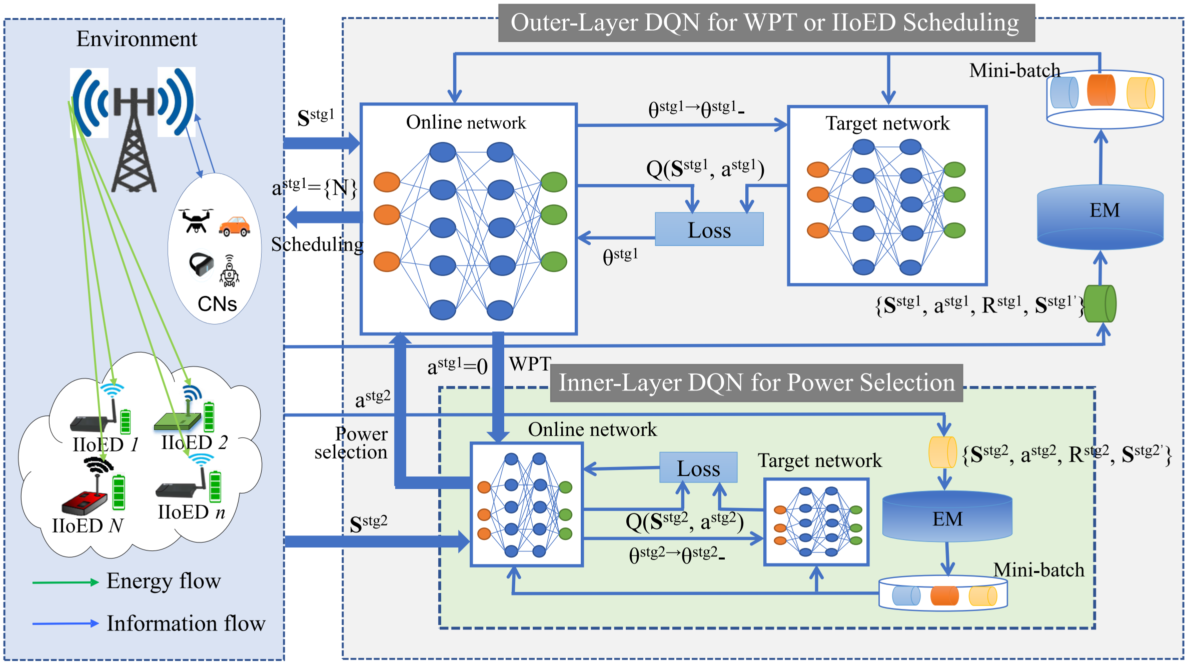

Outer-Layer DQN for WPT or IIoED Scheduling
Environment
$\mathbf{S}^{\text{stg1}}$
Online network
$\theta^{\text{stg1}}\rightarrow\theta^{\text{stg1}}_{-}$
Target network
Mini-batch
$a^{\text{stg1}}=\{N\}$
Scheduling
$Q(\mathbf{S}^{\text{stg1}}, a^{\text{stg1}})$
Loss
θ^{stg1}
EM
CNs
$\{\mathbf{S}^{\text{stg1}}, a^{\text{stg1}}, R^{\text{stg1}}, \mathbf{S}^{\text{stg1}'}\}$
$a^{\text{stg1}}=0$
WPT
Inner-Layer DQN for Power Selection
a^{stg2}
Power selection
Online network
Loss
Target network
$\{\mathbf{S}^{\text{stg2}}, a^{\text{stg2}}, R^{\text{stg2}}, \mathbf{S}^{\text{stg2}'}\}$
EM
$\mathbf{S}^{\text{stg2}}$
$Q(\mathbf{S}^{\text{stg2}}, a^{\text{stg2}})$
$\theta^{\text{stg2}}\rightarrow\theta^{\text{stg2}}_{-}$
Mini-batch
IIoED 1
IIoED 2
IIoED N
IIoED n
Energy flow
Information flow

Fig. 4. Illustration of the framework of our proposed DLDQN algorithm.

where ζ is the discount factor,[12] and i is the count indicator. Besides, $\{\boldsymbol{v}^*(k), \boldsymbol{O}^*(k)\}$ is the best IIoEDs scheduling strategy at epoch k, which is obtained by maximizing $V^{\text{ou}}(k)$ at the current epoch, i.e.,

$$\{\boldsymbol{v}^*(k), \boldsymbol{O}^*(k)\} = \arg \max_{\{\boldsymbol{v}(k), \boldsymbol{O}(k)\}} \mathbb{E}\left[V^{\text{ou}}(k)|\{\boldsymbol{v}(k), \boldsymbol{O}(k)\}\right], \forall k \in \mathcal{K}. \tag{34}$$

Based on the *Bellman Optimality Equality* in [48], the Q-value function is updated by

$$Q^{\text{ou}}(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k)|\theta^{\text{ou}}(k)) = R^{\text{stg1}}(k+1) + \zeta \max_{a^{\text{stg1}}} Q(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}), \forall k \in \mathcal{K}. \tag{35}$$

Then, the best IIoED scheduling process action at epoch k is

$$a^{\text{stg1}*} = \arg \max_{a^{\text{stg1}}} Q(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}), \forall k \in \mathcal{K}. \tag{36}$$

Subsequently, in phase 2, a batch of sampled historical experiences, i.e., $\{\boldsymbol{S}^{\text{stg1}}, a^{\text{stg1}}, \boldsymbol{S}^{\text{stg1}}(k+1), R^{\text{stg1}}(k+1)\}$, are drawn from the EM to train the two DNNs. The parameters of the online network and those of the target network are updated from $\theta^{\text{stg1}}(k)$ and $\theta^{\text{stg1}}_{-}(k)$ to $\theta^{\text{stg1}}(k+1)$ and $\theta^{\text{stg1}}_{-}(k+1)$, respectively, by minimizing the loss function. The loss function at epoch k is given by

$$L(\theta^{\text{stg1}}(k+1)) = \mathbb{E}\left[R^{\text{stg1}}(k) + \zeta \max_{a^{\text{stg1}}} Q(\boldsymbol{S}^{\text{stg1}}(k+1), a^{\text{stg1}}| \theta^{\text{stg1}}(k)) - Q(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k)|\theta^{\text{stg1}}(k))\right]^2, \forall k \in \mathcal{K}. \tag{37}$$

By adopting the *stochastic gradient descent approach*, the gradient of the parameter θ^{stg1}'s update at epoch $k+1$ is expressed as

$$\nabla L(\theta^{\text{stg1}}(k+1)) = \mathbb{E}\left[R^{\text{stg1}}(k) + \zeta \max_{a^{\text{stg1}}} Q(\boldsymbol{S}^{\text{stg1}}(k+1), a^{\text{stg1}}|\theta^{\text{stg1}}(k)) - Q(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k)|\theta^{\text{stg1}}(k+1))\right] \times \nabla Q(\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k)|\theta^{\text{stg1}}(k+1)), \forall k \in \mathcal{K}. \tag{38}$$

C. Inner-Layer DQN for Transmit Power Selection

Similar to the outer-layer DQN, the inner-layer DQN is also composed of two phases, i.e., the transmit power selecting action generation and the transmit power selection update. Specifically, at epoch k, the Q-value function in phase 1 is

$$Q^{\text{in}}(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k)|\theta^{\text{in}}(k)) = \mathbb{E}\left[V^{\text{in}}(k)|\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k), P^*_{\text{HAP}}(k)\right], \forall k \in \mathcal{K}, \tag{39}$$

where $\theta^{\text{in}}(k)$, $V^{\text{in}}(k)$, and $P^*_{\text{HAP}}(k)$ are the inner-layer DQN's parameters, the discounted cumulative reward of the transmit power selection, and the optimal HAP's transmit power at epoch k, respectively. $V^{\text{in}}(k)$ and $P^*_{\text{HAP}}(k)$ are obtained by maximizing $V^{\text{in}}(k)$ at epoch k and are given by

$$V^{\text{in}}(k) = \sum\nolimits_{i \in \mathbb{N}} \zeta^i R^{\text{stg2}}(k+i+1), \forall k \in \mathcal{K}, \tag{40}$$

and

$$P^*_{\text{HAP}}(k) = \arg \max_{P_{\text{HAP}}(k)} \mathbb{E}\left[V^{\text{in}}(k)|P_{\text{HAP}}(k)\right], \forall k \in \mathcal{K}, \tag{41}$$

respectively. Then, the inner DQN's Q-value function is

$$Q^{\text{in}}(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k)|\theta^{\text{in}}(k)) = R^{\text{stg2}}(k+1) + \zeta \max_{a^{\text{stg2}}} Q(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}), \forall k \in \mathcal{K}, \tag{42}$$

[12] The discount factor is an intrinsic parameter of DQN, which reflects the influence weight of the previous experience on the current learning as the earlier experience has a smaller influence on the current learning. The smaller the value of the discount factor ζ, the smaller the influence of the earlier experience on the current learning.

Algorithm 1 Our Proposed DLDQN Algorithm

```
Input: Initial $W$, $T$, $K$, $N$, $C$, $D$, $B$, $L$, $X$, $\eta_{\text{eh}}$, $M_{\text{eh}}$, $\gamma_{\text{th}}$, $P_{\text{HAP}}^{\max}$, $\zeta$, $\varepsilon$, $\boldsymbol{S}^{\text{stg1}}$, $\boldsymbol{S}^{\text{stg2}}$, $\theta^{\text{stg1}}$, $\theta_{-}^{\text{stg1}}$, $\theta^{\text{stg2}}$, $\theta_{-}^{\text{stg2}}$, Episode;
1  for $episode = 1 : Episode$ do
2      for $k = 1 : K$ do
3          Observe state $\boldsymbol{S}^{\text{stg1}}(k)$, and generate action $a^{\text{stg1}}(k)$;
4          if $a^{stg1}(k) \neq 0$ then
5              Execute action $a^{\text{stg1}}(k)$, obtain reward $R^{\text{stg1}}(k)$, and go to the next state $\boldsymbol{S}^{\text{stg1}}(k+1)$;
6              Store experience $\{\boldsymbol{S}^{\text{stg1}}(k), a^{\text{stg1}}(k), R^{\text{stg1}}(k), \boldsymbol{S}^{\text{stg1}}(k+1)\}$ to the EM;
7              Sample a mini-batch $H$ from the EM;
8              Update parameter $\theta^{\text{stg1}}(k+1)$ according to (38);
9              Update state $\boldsymbol{S}^{\text{stg2}}(k)$ based on state $\boldsymbol{S}^{\text{stg1}}(k)$;
10         else
11             Observe state $\boldsymbol{S}^{\text{stg2}}(k)$, and generate action $a^{\text{stg2}}(k)$;
12             Execute action $a^{\text{stg2}}(k)$, obtain reward $R^{\text{stg2}}(k)$, and go to the next stats $\boldsymbol{S}^{\text{stg2}}(k+1)$;
13             Store experience $\{\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k), R^{\text{stg2}}(k), \boldsymbol{S}^{\text{stg2}}(k+1)\}$ to the EM;
14             Sample a mini-batch $H$ from the EM;
15             Update parameter $\theta^{\text{stg2}}(k+1)$ according to (45);
16             Update state $\boldsymbol{S}^{\text{stg2}}(k)$ based on state $\boldsymbol{S}^{\text{stg2}}(k)$;
17         if mod(k, X)==0 then
18             Update the target DNN of both the outer-layer DQN and the inner-layer DQN by $\theta^{\text{stg1}} \to \theta_{-}^{\text{stg1}}$, $\theta^{\text{stg2}} \to \theta_{-}^{\text{stg2}}$.
19     Reset state $\boldsymbol{S}^{\text{stg1}}$ and state $\boldsymbol{S}^{\text{stg2}}$;
Output: $\theta^{\text{stg1}}$, $\theta_{-}^{\text{stg1}}$, $\theta^{\text{stg2}}$, $\theta_{-}^{\text{stg2}}$.
```

where the optimal power selecting action at epoch k is

$$a^{\text{stg2}*} = \arg\max_{a^{\text{stg2}}} Q(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}), \forall k \in \mathcal{K}. \tag{43}$$

Subsequently, in phase 2, the parameters of the online network and those of the target network are updated from $\theta^{\text{stg2}}(k)$ and $\theta_{-}^{\text{stg2}}(k)$ to $\theta^{\text{stg2}}(k+1)$ and $\theta_{-}^{\text{stg2}}(k+1)$, respectively. The loss function and the update gradient of the parameter θ^{stg2} at epoch k and epoch $k+1$ are given by

$$L(\theta^{\text{stg2}}(k+1)) = \mathbb{E}\Big[R^{\text{stg2}}(k) + \zeta \max_{a^{\text{stg2}}} Q(\boldsymbol{S}^{\text{stg2}}(k+1), a^{\text{stg2}}| \theta^{\text{stg2}}(k)) - Q(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k)|\theta^{\text{stg2}}(k))\Big]^2, \quad \forall k \in \mathcal{K}, \tag{44}$$

and

$$\nabla L(\theta^{\text{stg2}}(k+1)) = \mathbb{E}\Big[R^{\text{stg2}}(k) + \zeta \max_{a^{\text{stg2}}} Q(\boldsymbol{S}^{\text{stg2}}(k+1), a^{\text{stg2}}|\theta^{\text{stg2}}(k)) - Q(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k)|\theta^{\text{stg2}}(k+1))\Big] \times \nabla Q(\boldsymbol{S}^{\text{stg2}}(k), a^{\text{stg2}}(k)|\theta^{\text{stg2}}(k+1)), \forall k \in \mathcal{K}, \tag{45}$$

respectively.

D. Computational Complexity

To analyze the algorithm complexity, we first summarize our proposed DLDQN algorithm in Algorithm 1. Particularly, the training procedure begins with initializing the system parameters and the network parameters. Then, in lines 1-9 of Algorithm 1, we observe the current state in stage 1 and use a ε-greedy approach to explore the environment at the current state, in which case the HAP generates a random WPT or IIoEDs scheduling process action with probability ε, whereas an optimized action is selected with probability $1-\varepsilon$. In lines 10-16, we observe the current state in stage 2, and then the HAP performs transmit power selecting action as in stage 1. As the training goes on, the value of the probability ε is reduced to make the HAP learn the best action. For the current state, when the corresponding action is performed, its reward is obtained and the system transits to the next state. Meanwhile, the experiences are stored in the EM, and a mini-batch of historical experiences, i.e., H, are sampled from EM to train the online network and the target network of both the outer-layer DQN and the inner-layer DQN. In lines 17-18, the target network of both the outer-layer DQN and the inner-layer DQN are updated in every X episodes.

Then, we discuss the computational complexity of our proposed DLDQN algorithm. For the outer DQN, its training complexity contains two loops. The outer loop performs one training episode, and the inner loop performs K scheduling strategy and updates the parameters of the outer-layer DNNs. Particularly, the computational complexity of K scheduling strategy is $\mathcal{O}(K)$, and that of updating DNN's parameters is $\mathcal{O}(\sum_{p=1}^{P} n_{p-1} \cdot n_p)$, in which n_p is the neural node number at the layer p of a neural network [49]. With the mini-batch of experiences being H, the computational complexity of training the DNNs is $\mathcal{O}(H \cdot \sum_{p=1}^{P} n_{p-1} \cdot n_p)$. Then, the computational complexity of the outer-layer DQN is given by $\mathcal{O}(K + H \cdot \sum_{p=1}^{P} n_{p-1} \cdot n_p)$. Similar to the outer-layer DQN, the computational complexity of the inner-layer DQN includes performing K power selection and training the inner-layer DNNs. Therefore, the total computational complexity of our proposed DLDQN algorithm is summarized as $\mathcal{O}(K + H \cdot \sum_{p=1}^{P} n_{p-1} \cdot n_p)$.

VI. Simulation Results

In this section, we provide extensive simulation results to evaluate the effectiveness of our proposed DLDQN algorithm in terms of the long-term AEE in our considered wireless powered IIoE network.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
The number of epochs in one time frame	1000
The time period of an epoch	0.1 ms
The number of IIoEDs	50
The number of CNs	10
The number of requests from CNs	[2, 3, 4]
The HAP's transmit power threshold	5 W
The system bandwidth	1 MHz
The system noise power	10^{-9} mW
The size of the IIoED's energy buffer	0.1 μJ
The efficiency of EH circuits	0.8
The saturation threshold of EH circuits	24 mW
The update packet length	1000 bits
The tolerable information aging threshold	100 ms
The random action probability	0.1
The learning rate	0.0004
The training interval	20
The size of EM	10000
The size of mini-batch	256
The number of episodes	400
The update interval of the target network	100

A. Simulation Setup

The simulated network scenario is shown in Fig. 1. The simulation is conducted over a two-dimensional coordinate plane, in which the HAP is positioned at the original point (i.e., (0, 0)) and covers a circle of radius 10 m. The IIoEDs and CNs are randomly positioned at arbitrary points on the plane within 20 m. The wireless channel coefficients are considered to follow the Rician fading model, where the power loss factor is set to be 2 and the Rician factor is set to be 3.5. The rest of the simulation parameters are listed in Table I. The simulation experiments are run on Python 3.8 and Tensorflow 1.6.0. To implement our proposed DLDQN algorithm, the Q-network contains two hidden layers with 256 and 128 neurons, respectively. All simulation parameters described above do not change unless otherwise specified.

B. Our Proposed Design Vs. Benchmark Designs

To verify the effectiveness of our proposed DLDQN algorithm, three benchmarks are considered:

- **DQN Algorithm**: Compared with our proposed DLDQN algorithm, this algorithm considers performing the WPT or IIoEDs scheduling process and the transmit power selection at the common layer. The other settings are the same as that of our proposed DLDQN algorithm.
- **Round Robin Algorithm**: Compared with our proposed DLDQN algorithm, this algorithm only considers optimizing the power selection, but does not consider optimizing IIoEDs scheduling. The IIoEDs scheduling strategy is designed as a one-by-one sequential scheduling strategy. Its transmit power selection is the same as that of our proposed DLDQN algorithm.

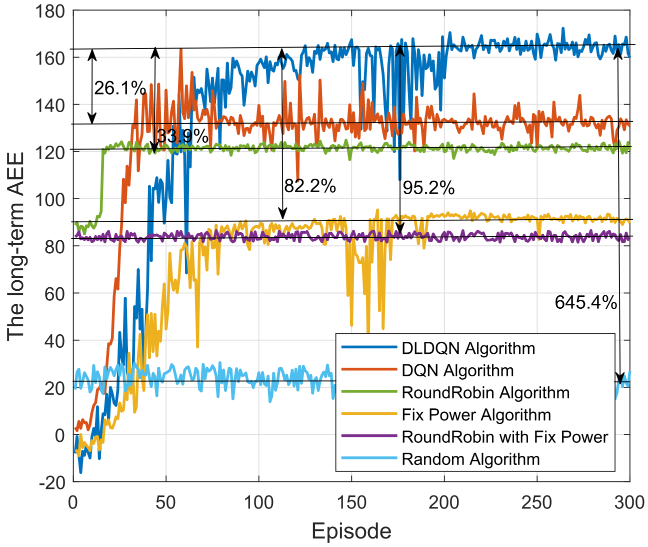


Fig. 5. The long-term AEE achieved by different algorithms.

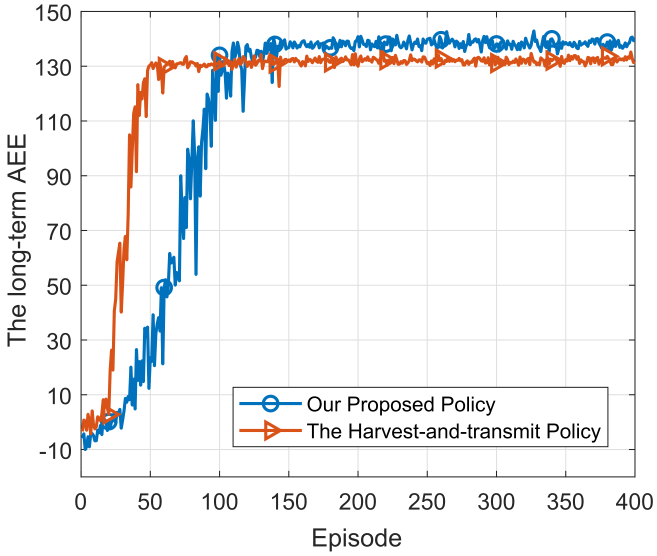
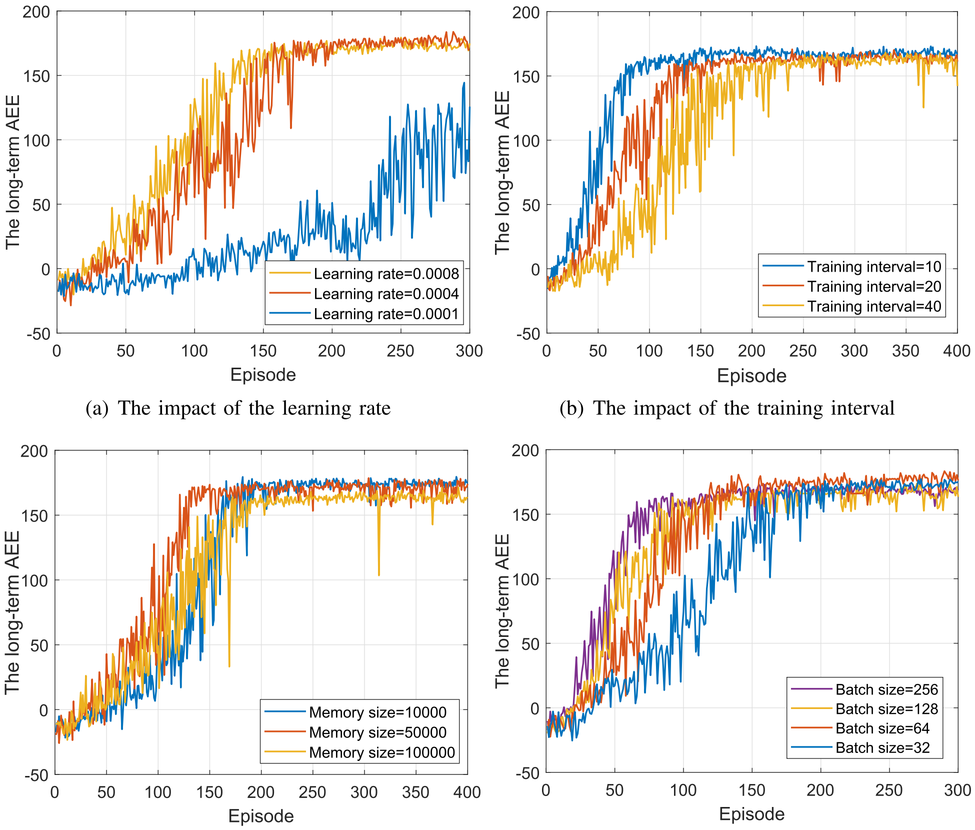


Fig. 6. The long-term AEE achieved by different policies.

- **Fixed Power Algorithm**: Compared with our proposed DLDQN algorithm, this algorithm only considers optimizing IIoEDs scheduling, but does not consider optimizing the power selection. That is, the HAP employs a fixed transmit power to charge the IIoEDs. Its IIoEDs scheduling strategy is the same as that of our proposed DLDQN algorithm.
- **Round Robin Algorithm with Fixed Power**: Compared with our proposed DLDQN algorithm, this algorithm designs the IIoEDs scheduling as a one-by-one sequential scheduling strategy, and the HAP employs a fixed transmit power to charge the IIoEDs.
- **Random Algorithm**: Compared with our proposed DLDQN algorithm, this algorithm designs the IIoEDs scheduling and transmit power selection strategies as random strategies.

Figure 5 shows the achievable long-term AEE versus episode of our proposed DLDQN algorithm with five bench-



(a) The impact of the learning rate

(b) The impact of the training interval

(c) The impact of the memory size

(d) The impact of the batch size

Fig. 7. The convergence performance of our proposed DLDQN algorithm with different network parameters.

marks. With the increase of episode, the achieved system long-term AEEs of the four algorithms, i.e., our proposed DLDQN algorithm, the DQN algorithm, the round robin algorithm, and the fixed power algorithm, first increase significantly, and then keep stable with small fluctuations. On the other hand, the achieved system long-term AEEs of rest two algorithms, i.e., the round robin algorithm with fixed power and the random algorithm, fluctuate without increasing. The long-term AEE obtained by our proposed DLDQN algorithm outperforms that of the other five benchmarks. For example, it can be improved by about 6 times compared with the random algorithm and can be improved by about 80% compared with the fixed power algorithm. Besides, it can be improved by about 25% compared with the DQN algorithm. The reason is that the single-layer DQN makes joint decisions on all the tightly coupled variables, and the resulting large decision space makes it hard to improve the policy accuracy. Moreover, the convergence speed of our proposed DLDQN algorithm is comparable with that of the DQN algorithm.

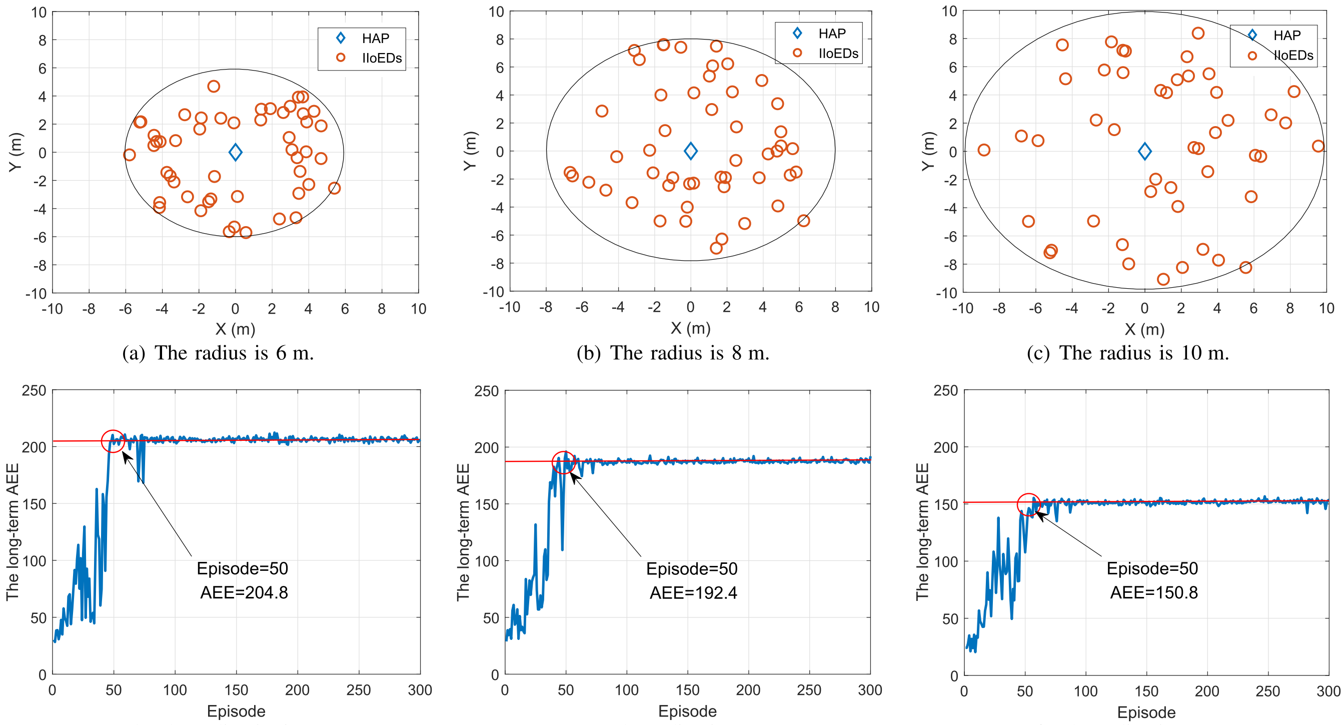
Figure 6 shows the achievable long-term AEE versus different policies. From Fig. 6, the achieved system long-term AEE of our proposed policy outperforms that of the harvest-and-transmit policy, which exhausts the harvested energy for each transmission epoch. The reason may be that for the harvest-and-transmit policy, it consumes more energy to achieve the same AoI gain as our proposed policy, so the AEE performance is degraded.

C. Convergence Performance

Figure 7 shows the impacts of different network parameters on the convergence performance of our proposed DLDQN algorithm, including different learning rates, training intervals, memory sizes, and batch sizes. In Fig. 7(a), a small learning rate makes our proposed DLDQN algorithm converge with a relatively slow speed. When the learning rate is larger than 0.0004, the converge speed is not obviously improved. In Fig. 7(b), our proposed DLDQN algorithm converges faster with a shorter training interval. In Fig. 7(c), a relatively small memory, i.e., 10^4 in our simulations, brings a fast convergence performance for our proposed DLDQN algorithm, while a relatively large memory, i.e., 10^5 in our simulations, requires more training data to converge. In Fig. 7(d), a large batch size leads to a fast convergence speed of our proposed DLDQN algorithm, because a large size of mini-batch can take the advantage of more training data. However, a large batch size costs more time for training. In the following simulations, by jointly considering the AEE performance, convergence speed, and computation time, we set the learning rate, training interval, memory size, and batch size to be 0.0004, 10, 10^4, and 256, respectively.

D. The Impact of the Distance Between HAP and IIoEDs on the Long-Term AEE

Figure 8 shows the long-term AEE versus the distance between HAP and IIoEDs. To evaluate the impact of this



(a) The radius is 6 m. (b) The radius is 8 m. (c) The radius is 10 m.

(d) The radius is 6 m. (e) The radius is 8 m. (f) The radius is 10 m.

Fig. 8. The impact of distance between HAP and IIoEDs on the system long-term AEE.

distance on the system long-term AEE achieved by our proposed DLDQN algorithm, we simulate three different scenarios, where the distances between HAP and IIoEDs are set to 6 m, 8 m, and 10 m, respectively. In Fig. 8, the long-term AEE increases with the decrease of the distance between HAP and IIoEDs. The reason is that when the distance between HAP and IIoEDs increases, the energy transmission efficiency from the HAP to the IIoEDs is significantly reduced. Hence, the HAP requires more time to wirelessly charge the IIoEDs' energy buffer, which deteriorates the AoI performance and further degrades the system long-term AEE.

Figure 9 depicts the long-term AEE versus the distance between HAP and IIoEDs, where the number of IIoEDs are set to 30 and 40. In Fig. 9, the long-term AEE achieved by our proposed DLDQN algorithm can be improved by about 20%∼30% compared with the fixed power algorithm. Besides, with the decrease of the distance between HAP and IIoEDs, such as from 14 m to 6 m, the improving rate of the achievable long-term AEE gain with our proposed DLDQN algorithm increases by about 6%∼8%. It means that when the distance between HAP and IIoEDs is short, a higher system long-term AEE gain is achieved by optimizing the HAP's transmit power. Moreover, the system long-term AEE can be improved when the number of IIoEDs increases from 30 to 40.

E. The Impact of the Number of IIoEDs on the Long-Term AEE

Figure 10 shows the system long-term AEE versus the number of IIoEDs, where the distance between the HAP and the IIoEDs is set to 16 m. In Fig. 10, we simulate three scenarios where the numbers of IIoEDs are 20, 30, and 40, respectively.

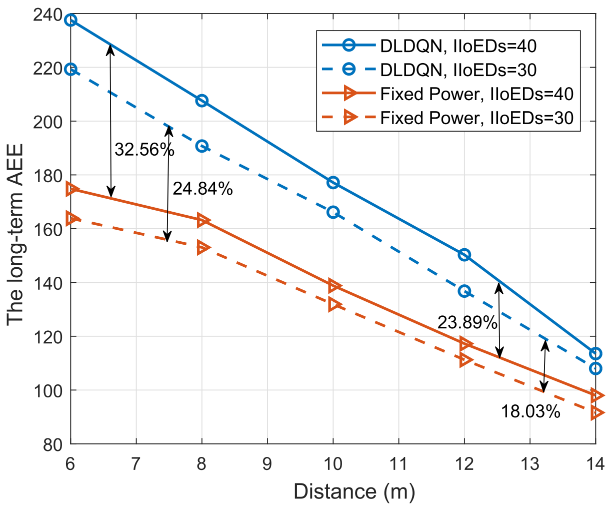
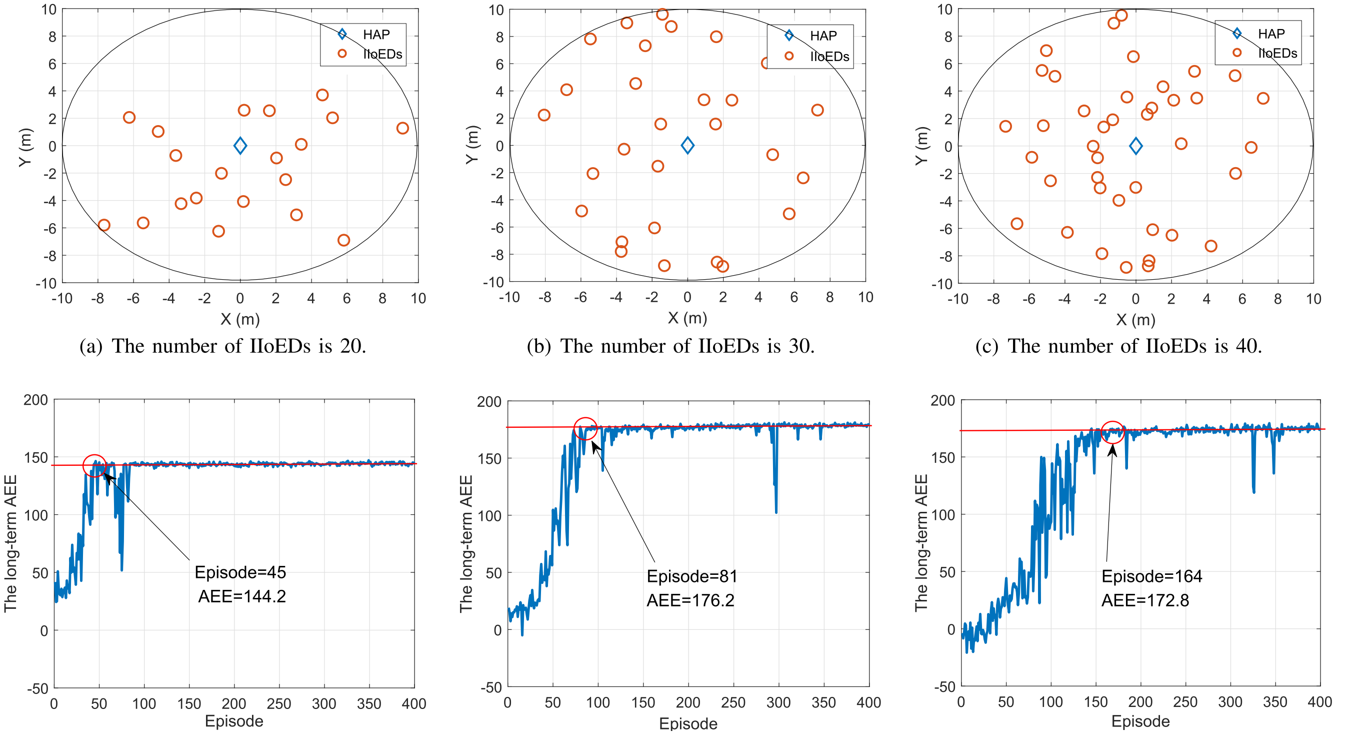


Fig. 9. The system long-term AEE versus the distance between HAP and IIoEDs.

When the number of IIoEDs increases from 20 to 30, the system long-term AEE achieved by our proposed DLDQN algorithm increases by about 22%, while when the number of IIoEDs increases from 30 to 40, the system long-term AEE decreases by about 2%. This phenomenon means that there exists an optimal number of IIoEDs to achieve the maximal long-term AEE with our proposed DLDQN algorithm, which provides an insightful guidance for practical deployment. With increasing number of IIoEDs, the EE of the system can be improved by fully utilizing the broadcast feature of wireless channels, while the AoI is degraded because more IIoEDs need to wait for a long time to be scheduled. Therefore, there exists



(a) The number of IIoEDs is 20.

(b) The number of IIoEDs is 30.

(c) The number of IIoEDs is 40.

(d) The number of IIoEDs is 20.

(e) The number of IIoEDs is 30.

(f) The number of IIoEDs is 40.

Fig. 10. The impact of the number of IIoEDs on the long-term AEE.

a tradeoff between the AoI and the minimal required energy. Before the number of IIoEDs increases to a certain threshold, i.e., 30 in our simulation, the EE has a dominant impact on the long-term AEE, while after the threshold, the AoI has a greater impact on the long-term AEE compared with EE.

F. The Impact of the Number of Epochs on the Long-Term AEE

Figure 11 shows the impact of the number of epochs in each episode on the long-term AEE. With more epochs in each episode, the convergence speed of our proposed DLDQN algorithm becomes faster, while its improving rate is deceasing. Besides, when the number of epochs in each episode is 1000, the system long-term AEE achieved by our proposed DLDQN algorithm increases by about 5% compared with the case with 500 epochs. Besides, when the number of epochs in each episode is larger than 1000, the system long-term AEE cannot be further improved.

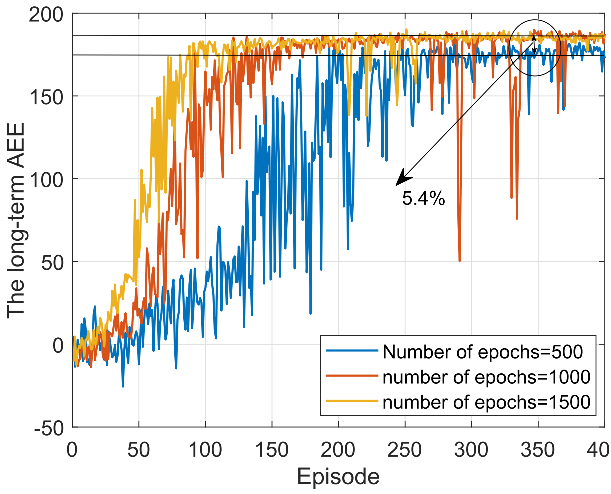


Fig. 11. The long-term AEE versus the number of epochs.

G. The Impact of the Weight Factor in Reward Function on the Long-Term AEE

Figure 12 shows the impact of the weight factor in reward function on the long-term AEE. With the increase of the weight factor μ, the system long-term AEE achieved by our proposed DLDQN algorithm first increases then decreases, which means that both AoI and energy consumption have impacts on the achievable system long-term AEE. When the weight factor is set to 1.3 in our simulations, it can realize the maximal system long-term AEE. Besides, when the weight factor is less than 0.8 or larger than 1.6, the system long-term AEE achieved by our proposed DLDQN algorithm decreases obviously.

H. The Impact of the Energy Buffer Capacity on the Long-Term AEE

Figure 13 shows the impact of the energy buffer capacity at the IIoEDs on the system long-term AEE. With different numbers of IIoEDs, the long-term AEE always increases with a larger energy buffer. Besides, when the number of IIoEDs is relatively small, i.e., 20 in our simulations, the system long-term AEE improvement achieved by increasing

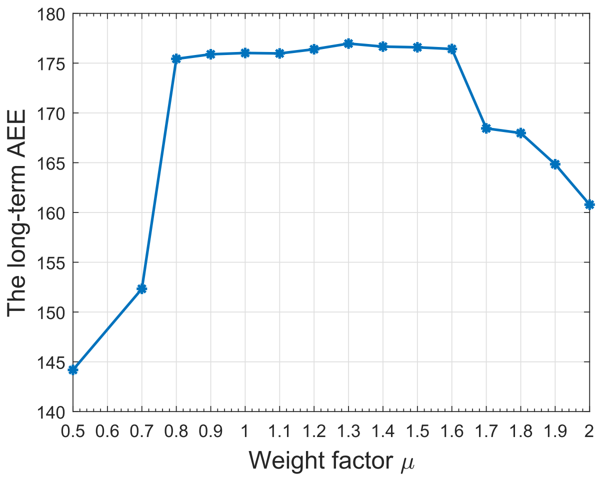


Fig. 12. The long-term AEE versus the weight factor in reward function.

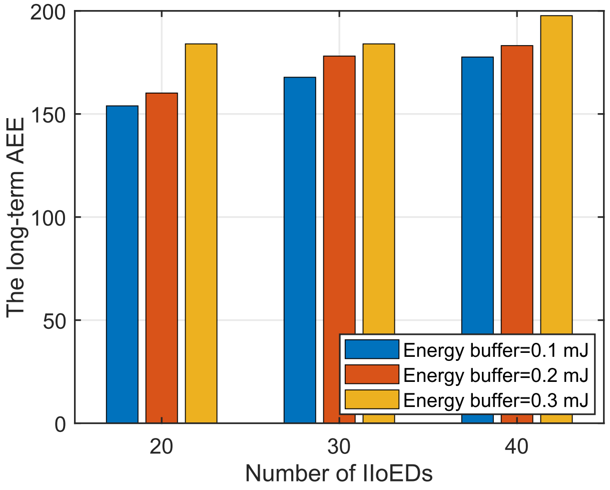


Fig. 13. The long-term AEE versus the energy buffer capacity.

the energy buffer is higher compared with the cases with 30 and 40 IIoEDs.

I. The Impact of the Status Update Packet Length on the Long-Term AEE

Figure 14 depicts the impact of the status update packet length, i.e., the information bits of a single status update packet, on the system long-term AEE, where the length of a single status update packet is set to 500 bits, 1000 bits, and 1500 bits, respectively. The long-term AEE significantly decreases with the increase of the packet length. For example, as shown in Fig. 14, when the packet length increases from 500 bits to 1000 bits with the number of IIoEDs being 30∼50, the system long-term AEE achieved by our proposed DLDQN algorithm is reduced by about 60%. When the packet length increases from 1000 bits to 1500 bits, the achieved system long-term AEE is reduced by about 70%∼80%. The reason is a long update packet requires more energy to transmit, which results in more time for energy transfer and a reduced AoI. Moreover, when the number of IIoEDs increases, the long-term AEE changes non-obviously, which means that the update packet length has a dominant impact on the system long-term AEE compared with the number of IIoEDs.

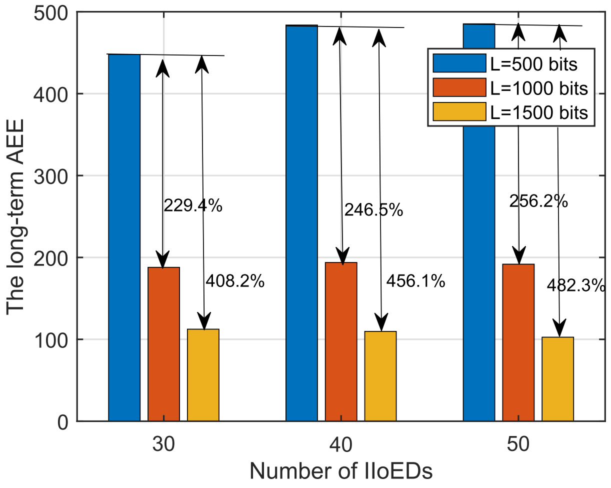


Fig. 14. The long-term AEE versus the update packet length.

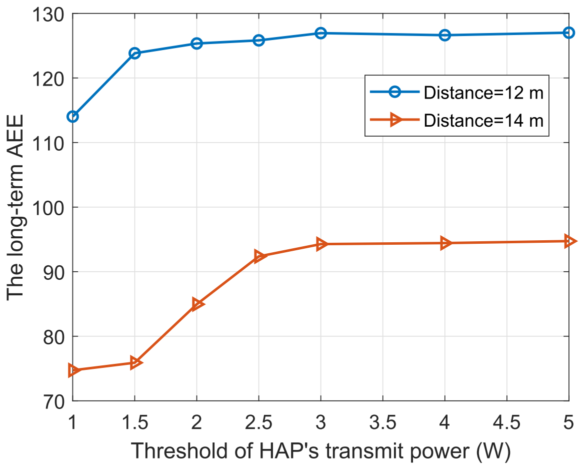


Fig. 15. The long-term AEE versus the threshold of HAP's transmit power.

J. The Impact of the Threshold of HAP's Transmit Power on the System Performance

Figure 15 shows the system long-term AEE versus the threshold of HAP's transmit power with different distances between the HAP and the IIoEDs, i.e., 12 m and 14 m, respectively. In Fig. 15, the system long-term AEE achieved by our proposed DLDQN algorithm first increases and then keeps stable, which shows a non-linear trend with the increase of the threshold of HAP's transmit power. The reason is that when the HAP's transmit power threshold is relatively small (i.e., 1.5 W when the distance is 12 m and 3 W when the distance is 14 m in our simulations), the limited HAP power resource constrains the system long-term AEE. When the threshold of HAP's transmit power increases, more power resources are available, allowing a better long-term AEE in the considered system.

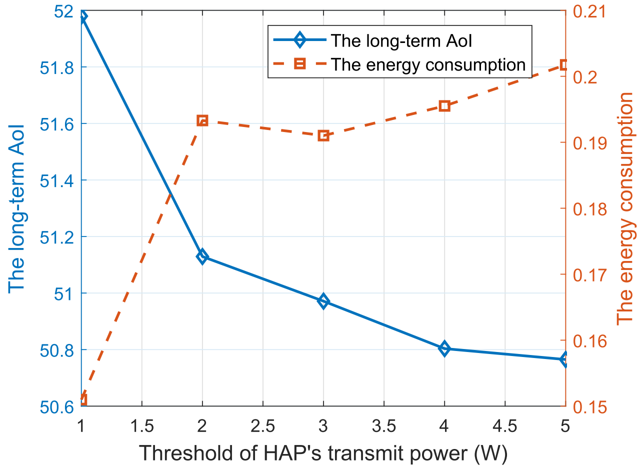


Fig. 16. The AoI and energy consumption versus the threshold of HAP's transmit power.

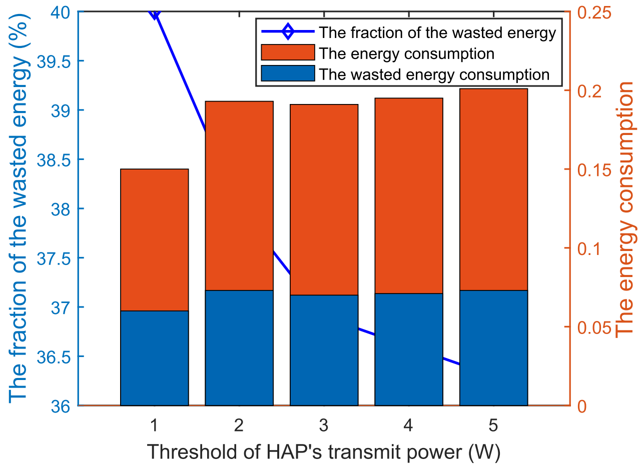


Fig. 17. The fraction of the wasted energy versus the threshold of HAP's transmit power.

Figure 16 shows the system long-term AoI and energy consumption versus the threshold of HAP's transmit power, where the distance between the HAP and the IIoEDs is 10 m. The long-term AoI and energy consumption achieved by our proposed DLDQN algorithm decreases and increases with the increase of the threshold of HAP's transmit power, respectively, which means that compared with reducing energy consumption, reducing AoI has a dominant impact on improving the system long-term AEE.

Figure 17 shows the fraction of the wasted energy versus the threshold of HAP's transmit power. It is shown that the fraction of the wasted energy in the total energy consumption decreases from about 40% to 36% with the increase of the threshold of HAP's transmit power, while the value of the wasted energy fluctuates around 0.075 J, which means that with the increase of the threshold of HAP's transmit power, the EE is improved by our proposed DLDQN algorithm, and the system long-term AEE is improved accordingly.

VII. Conclusion

In this paper, we have investigated a long-term AEE maximization problem in a wireless powered IIoE network. To solve the problem, we modeled it as a two-stage MDP and then proposed a DLDQN algorithm to find the effective scheduling strategy and transmit power. Simulation results demonstrated that our proposed algorithm can achieve significantly higher system long-term AEE compared with benchmarks. The proposed DLDQN algorithm is also scalable in improving the system long-term AEE when the number of IIoEDs increases. Moreover, we revealed that there exists a range of the number of IIoEDs, which achieves the maximal system long-term AEE, and can provide insightful guidance for the practical network deployment. In future work, we will investigate the AEE-related resource allocations for applications with integrated communications, sensing, and computing, in which the IIoEDs' sensed information requires integrated computational processing and communication services.

References

[1] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 1–30, 1st Quart., 2022.

[2] P. Yang, K. Guo, X. Xi, T. Q. S. Quek, X. Cao, and C. Liu, "Fresh, fair and energy-efficient content provision in a private and cache-enabled UAV network," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 97–112, Jan. 2022.

[3] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[4] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.

[5] K. Xiong, C. Chen, G. Qu, P. Fan, and K. B. Letaief, "Group cooperation with optimal resource allocation in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3840–3853, Jun. 2017.

[6] M. Sheng, Y. Li, X. Wang, J. Li, and Y. Shi, "Energy efficiency and delay tradeoff in device-to-device communications underlaying cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 92–106, Jan. 2016.

[7] C. Han et al., "Green radio: Radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.

[8] Y. Lu, K. Xiong, P. Fan, Z. Ding, Z. Zhong, and K. B. Letaief, "Global energy efficiency in secure MISO SWIPT systems with non-linear power-splitting EH model," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 216–232, Jan. 2019.

[9] X. S. Shen et al., "Data management for future wireless networks: Architecture, privacy preservation, and regulation," *IEEE Netw.*, vol. 35, no. 1, pp. 8–15, Jan. 2021.

[10] M. Xie, J. Gong, and X. Ma, "Is the packetized transmission efficient? An age-energy perspective," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Jul. 2020, pp. 329–333.

[11] J. Zhong, R. D. Yates, and E. Soljanin, "Minimizing content staleness in dynamo-style replicated storage systems," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2018, pp. 361–366.

[12] Y. Tseng and Y. Hsu, "Online energy-efficient scheduling for timely information downloads in mobile networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1022–1026.

[13] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, "Minimizing age of information in vehicular networks," in *Proc. 8th Annu. IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw.*, Jun. 2011, pp. 350–358.

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

[15] M. Li, C. Chen, H. Wu, X. Guan, and X. Shen, "Age-of-information aware scheduling for edge-assisted industrial wireless networks," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5562–5571, Aug. 2021.

[16] Z. Fang, J. Wang, Y. Ren, Z. Han, H. V. Poor, and L. Hanzo, "Age of information in energy harvesting aided massive multiple access networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1441–1456, May 2022.

[17] I. Krikidis, "Average age of information in wireless powered sensor networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 628–631, Apr. 2019.

[18] H. Hu, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "$\alpha-\beta$ AoI penalty in wireless-powered status update networks," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 474–484, Jan. 2022.

[19] N. I. Miridakis, Z. Shi, T. A. Tsiftsis, and G. Yang, "Extreme age of information for wireless-powered communication systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 826–830, Apr. 2022.

[20] Y. Zheng, J. Hu, and K. Yang, "Average age of information in wireless powered relay aided communication network," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11311–11323, Jul. 2022.

[21] M. Moltafet, M. Leinonen, M. Codreanu, and N. Pappas, "Power minimization for age of information constrained dynamic control in wireless sensor networks," *IEEE Trans. Commun.*, vol. 70, no. 1, pp. 419–432, Jan. 2022.

[22] 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.

[23] Y. Khorsandmanesh, M. J. Emadi, and I. Krikidis, "Average peak age of information analysis for wireless powered cooperative networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 4, pp. 1291–1303, Dec. 2021.

[24] A. Valehi and A. Razi, "Maximizing energy efficiency of cognitive wireless sensor networks with constrained age of information," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 643–654, Dec. 2017.

[25] H. Zheng, K. Xiong, P. Fan, Z. Zhong, and K. Ben Letaief, "Age-energy region in wireless powered communication networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Jul. 2020, pp. 334–339.

[26] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, 4th Quart., 2019.

[27] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5438–5453, Oct. 2018.

[28] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multiuser cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[29] X. Shen et al., "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, 2020.

[30] T. Zhang and S. Mao, "Energy-efficient power control in wireless networks with spatial deep neural networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 1, pp. 111–124, Mar. 2020.

[31] R. Zhang, K. Xiong, Y. Lu, B. Gao, P. Fan, and K. B. Letaief, "Joint coordinated beamforming and power splitting ratio optimization in MU-MISO SWIPT-enabled HetNets: A multi-agent DDQN-based approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 677–693, Feb. 2022.

[32] H. Lee, D. Kim, and J. Lee, "Radio and energy resource management in renewable energy-powered wireless networks with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5435–5449, Jul. 2022.

[33] L. Huang, S. Bi, and Y. A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2581–2593, Nov. 2020.

[34] L. Liu, K. Xiong, J. Cao, Y. Lu, P. Fan, and K. B. Letaief, "Average AoI minimization in UAV-assisted data collection with RF wireless power transfer: A deep reinforcement learning scheme," *IEEE Internet Things J.*, vol. 9, no. 7, pp. 5216–5228, Apr. 2022.

[35] X. Wu, X. Li, J. Li, P. C. Ching, and H. V. Poor, "Deep reinforcement learning for loT networks: Age of information and energy cost tradeoff," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[36] X. Wu, J. Yang, and J. Wu, "Optimal status update for age of information minimization with an energy harvesting source," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 193–204, Mar. 2018.

[37] H. Huang, D. Qiao, and M. C. Gursoy, "Age-energy tradeoff optimization for packet delivery in fading channels," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 179–190, Jan. 2022.

[38] Y. Dong, M. J. Hossaini, J. Cheng, and V. C. M. Leung, "Robust energy efficient beamforming in MISOME-SWIPT systems with proportional secrecy rate," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 202–215, Jan. 2019.

[39] Y. Dong, A. El Shafie, M. J. Hossain, J. Cheng, N. Al-Dhahir, and V. C. M. Leung, "Secure beamforming in full-duplex MISO-SWIPT systems with multiple eavesdroppers," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6559–6574, Oct. 2018.

[40] A. Arafa, J. Yang, S. Ulukus, and H. V. Poor, "Age-minimal transmission for energy harvesting sensors with finite batteries: Online policies," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 534–556, Jan. 2020.

[41] Y. Dong, M. J. Hossain, and J. Cheng, "Performance of wireless powered amplify and forward relaying over Nakagami-m fading channels with nonlinear energy harvester," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 672–675, Apr. 2016.

[42] J. Gong, X. Chen, and X. Ma, "Energy-age tradeoff in status update communication systems with retransmission," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[43] J. Vincke, S. Kempf, N. Schnelle, C. Horch, and F. Schäfer, "A concept for an ultra-low power sensor network–detecting and monitoring disaster events in underground metro systems," in *Proc. 6th Int. Conf. Sensor Netw.*, 2017, pp. 150–155.

[44] M. Zhang, A. Arafa, J. Huang, and H. V. Poor, "Pricing fresh data," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1211–1225, May 2021.

[45] M. Sun, X. Xu, X. Qin, and P. Zhang, "AoI-energy-aware UAV-assisted data collection for IoT networks: A deep reinforcement learning method," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17275–17289, Dec. 2021.

[46] Y. Dong, H. Zhang, J. Li, F. R. Yu, S. Guo, and V. C. M. Leung, "An online zero-forcing precoder for weighted sum-rate maximization in green CoMP systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7566–7581, Sep. 2022.

[47] J. Sorg, *The Optimal Reward Problem: Designing Effective Reward for Bounded Agents*. Ann Arbor, MI, USA: Univ. Michigan, 2011.

[48] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[49] 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.

Haina Zheng (Member, IEEE) received the B.E. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University (BJTU), Beijing, China, in 2017. She is currently pursuing the Ph.D. degree with the School of Computer and Information Technology, BJTU. She was a Visiting Ph.D. Student with the BBCR Group, Department of Electrical and Computer Engineering, University of Waterloo, Canada, from February 2021 to March 2022. So far, she has contributed as the first author for four IEEE journal articles and five IEEE ComSoc flagship conference articles. Her current research interests include wireless powered networks, fog computing, and age of information. She received the Best Paper Award from IEEE ICC in 2020, the IEEE Technical Committee on Transmission Access and Optical Systems (TAOS) in 2020, and the 26th China Academic Annual Conference on Information Theory (CIEIT) in 2019.

Ke Xiong (Member, IEEE) received the B.S. and Ph.D. degrees from Beijing Jiaotong University (BJTU), Beijing, China, in 2004 and 2010, respectively.

From April 2010 to February 2013, he was a Post-Doctoral Research Fellow with the Department of Electrical Engineering, Tsinghua University, Beijing. Since March 2013, he has been a Lecturer and an Associate Professor with BJTU. From September 2015 to September 2016, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. He is currently a Full Professor and the Vice Dean of the School of Computer and Information Technology, BJTU. He has published more than 100 academic papers in refereed journals and conferences. His current research interests include wireless cooperative networks, wireless powered networks, and network information theory.

Dr. Xiong is a member of the China Computer Federation (CCF). He is also a Senior Member of the Chinese Institute of Electronics (CIE). He has received the Best Paper Award from IEEE ICSTSN 2023, the Best Paper Award from the 8th ICCCS 2023, the Best Paper Award from IEEE ICC 2020, the Best Paper Award from the IEEE TAOS Technical Committee in 2020, and the Best Paper Award of CIEIT Conference in 2018 and 2019. He also served as the Session Chair for IEEE GLOBECOM 2012, IET ICWMMN 2013, IEEE ICC 2013, and ACM MOMM 2014, the Publicity and Publication Chair for IEEE HMWC 2014, as well as the TPC Co-Chair for IET ICWMMN 2017 and IET ICWMMN 2019. He serves as the Associate Editor-in-Chief for the Chinese Journal *New Industrialization Strategy* and the Editor for *International Journal of Computer Engineering and Software*. In 2017, he served as the leading Editor for the Special Issue "Recent Advances in Wireless Powered Communication Networks" for *EURASIP Journal on Wireless Communications and Networking* and the Guest Editor for the Special Issue "Recent Advances in Cloud-Aware Mobile Fog Computing" for *Wireless Communications and Mobile Computing*. He also serves as a Reviewer for more than 15 international journals, including IEEE Transactions on Signal Processing, IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, IEEE Transactions on Vehicular Technology, IEEE Communication Letters, IEEE Signal Processing Letters, and IEEE Wireless Communication Letters.

Mengying Sun (Member, IEEE) received the bachelor's degree in communication engineering from the Beijing University of Chemical Technology (BUCT), Beijing, China, in 2016, and the Ph.D. degree in information and telecommunications engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, in 2022. From March 2021 to March 2022, she was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. She is currently a Post-Doctoral Researcher with BUPT. Her research interests include mobile edge computing, D2D communication, and semantic communications.

Huaqing Wu (Member, IEEE) received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2021. She was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, MacMaster University, from 2021 to 2022. She is currently an Assistant Professor with the Department of Electrical and Software Engineering, University of Calgary, Alberta, Canada. Her current research interests include B5G/6G, space-air-ground integrated networks, the Internet of Vehicles, mobile/edge computing/caching, and artificial intelligence (AI) for future networking. She received the Best Paper Award from IEEE GLOBECOM 2018, the *Chinese Journal on Internet of Things* 2020, and IEEE GLOBECOM 2022. She received the prestigious Natural Sciences and Engineering Research Council of Canada (NSERC) Post-Doctoral Fellowship Award in 2021.

Zhangdui Zhong (Fellow, IEEE) received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1988, respectively. He is currently a Professor and an Advisor of Ph.D. candidates with Beijing Jiaotong University, where he is also a Chief Scientist of the State Key Laboratory of Rail Traffic Control and Safety. He is also the Director of the Innovative Research Team of the Ministry of Education, Beijing, and a Chief Scientist of the Ministry of Railways, Beijing. His research has been widely used in railway engineering, such as the Qinghai-Xizang Railway, the Datong-Qinhuangdao Heavy Haul Railway, and many high-speed railway lines in China. He has authored or coauthored seven books, five invention patents, and over 200 scientific research papers in his research area. His research interests include wireless communications for railways, control theory and techniques for railways, and GSM-R systems. He is an Executive Council Member of the Radio Association of China, Beijing, and the Deputy Director of the Radio Association, Beijing. He has received the Best Paper Award from IEEE ICC 2020 and the Best Paper Award from the IEEE TAOS Technical Committee in 2020. He received the Maoyisheng Scientific Award of China, the Zhantianyou Railway Honorary Award of China, and the Top 10 Science/Technology Achievements Award of Chinese Universities.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, wireless network security, the Internet of Things, 5G and beyond, and vehicular networks.

Dr. Shen is a fellow of the Engineering Institute of Canada, the Canadian Academy of Engineering, and the Royal Society of Canada. He is a Foreign Member of the Chinese Academy of Engineering. He received the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, the R. A. Fessenden Award from IEEE, Canada in 2019, the Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, the James Evans Avant Garde Award from the IEEE Vehicular Technology Society in 2018, the Joseph LoCicero Award in 2015, the Education Award from the IEEE Communications Society in 2017, and the Technical Recognition Award from the Wireless Communications Technical Committee in 2019 and the AHSN Technical Committee in 2013. He has also received the Excellent Graduate Supervision Award from the University of Waterloo in 2006 and the Premier's Research Excellence Award (PREA) from the Province of Ontario, Canada, in 2003. He served as the Technical Program Committee Chair/Co-Chair for IEEE GLOBECOM 2016, IEEE Infocom 2014, IEEE VTC 2010 Fall, and IEEE GLOBECOM 2007, and the Chair for the IEEE Communications Society Technical Committee on Wireless Communications. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of the IEEE Fellow Selection Committee of the ComSoc. He is the President of the IEEE Communications Society. He served as the Editor-in-Chief for the IEEE Internet of Things Journal, *IEEE Network*, and *IET Communications*. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society and the IEEE Communications Society. He is a registered Professional Engineer of Ontario, Canada.