Received 27 December 2023; revised 17 February 2024; accepted 6 March 2024. Date of publication 18 March 2024; date of current version 17 April 2024.

Digital Object Identifier 10.1109/OJCOMS.2024.3377706

Meta Reinforcement Learning for UAV-Assisted Energy Harvesting IoT Devices in Disaster-Affected Areas

MARWAN DHUHEIR, AIMAN ERBAD (Senior Member, IEEE), ALA AL-FUQAHA (Senior Member, IEEE), AND ABEGAZ MOHAMMED SEID (Member, IEEE)

Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

CORRESPONDING AUTHOR: M. DHUHEIR (e-mail: aldhuheir2011@gmail.com)

This work was supported by the NPRP-Standard (NPRP-S) 13th Cycle Grant from Qatar National Research Fund under Grant NPRP13S-0205-200265.

Open Access funding provided by the Qatar National Library.

ABSTRACT Over the past decade, Unmanned Aerial Vehicles (UAVs) have attracted significant attention due to their potential applications in emergency-response applications, including wireless power transfer (WPT) and data collection from Internet of Things (IoT) devices in disaster-affected areas. UAVs are more attractive than traditional techniques due to their maneuverability, flexibility, and low deployment costs. However, using UAVs for such critical tasks comes with challenges, including limited resources, energy constraints, and the need to complete missions within strict time frames. IoT devices in disaster areas have limited resources (e.g., computation, energy), so they depend on the UAVs' resources to accomplish vital missions. To address these resource problems in a disaster scenario, we propose a meta-reinforcement learning (RL)-based energy harvesting (EH) framework. Our system model considers a swarm of UAVs that navigate an area, providing wireless power and collecting data from IoT devices on the ground. The primary objective is to enhance the quality of service for strategic locations while allowing UAVs to dynamically join and leave the swarm (e.g., for recharging). In this context, we formulate the problem as a non-linear programming (NLP) optimization problem aimed at maximizing the total EH IoT devices and determining the optimal trajectory paths for UAVs while adhering to the constraints related to the maximum time duration, the UAVs' maximum energy consumption, and the minimum data rate to achieve a reliable transmission. Due to the complexity of the problem, the combinatorial nature of the formulated problem, and the difficulty of obtaining the optimal solution using conventional optimization problems, we propose a lightweight meta-RL solution capable of solving the problem by learning the system dynamics. We conducted extensive simulations and compared our approach with two state-of-the-art models using traditional RL algorithms represented by a deep Q-network algorithm, a Particle Swarm Optimization (PSO) algorithm, and one greedy solution. Our simulation results show that the proposed Meta-RL algorithm can outperform the IoT EH of the DQN, PSO algorithm, and the greedy solution by 25%, 32%, and 45%, respectively. The results of our simulations also demonstrate that our proposed approach outperforms the competitive solutions in terms of efficiently covering strategic locations with a high satisfaction rate and high accuracy.

INDEX TERMS Energy harvesting, UAVs positions, energy consumption, meta-reinforcement learning, UAVs, strategic locations.

I. INTRODUCTION

WIRELESS power transfer (WPT) holds immense potential in revolutionizing the Internet of Things (IoT) landscape by addressing key challenges associated with powering and maintaining numerous connected devices in a plethora of critical applications such as post-disaster

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

scenarios, search and rescue operations, etc [1], [2], [3]. The IoT devices, crucial for data collection, monitoring, and communication in post-disaster scenarios, often grapple with depleted energy reserves, intensifying the need for innovative solutions. Compounding this challenge is the impracticality of relying on traditional powering systems, which may be compromised or inaccessible in disaster-stricken areas. WPT emerges as a transformative solution, offering the potential to overcome the limitations of conventional power sources by providing cable-free charging for energy-constrained IoT devices [4]. Specifically, WPT mechanisms that operate on radio frequency (RF) signals are considered an alternative solution to traditional power supply to address the energy supply challenges faced by many IoT devices [5].

Utilizing ground chargers for transmitting power to IoT devices has been investigated in many research studies [6], [7]. Nonetheless, this technique poses several limitations, causing the efficiency of WPT to decrease due to poor line of sight (LoS) and long distances, particularly in post-disaster scenarios where traditional power infrastructure may be compromised. To address this challenge, an unmanned aerial vehicle (UAV) equipped with WPT capabilities can serve as a dynamic and adaptable platform for delivering power to remote or inaccessible areas affected by disasters [8]. The advantages of utilizing the UAV include effective deployment cost, flexibility, maneuverability, scalability, and direct LoS channel. However, utilizing one UAV to cover the entire area with high efficiency is insufficient due to the limitations of energy and the UAVs' flight duration. UAV swarm can provide a cost-effective and reliable solution to collect data and provide services to dispatched IoT devices over a wide geographical area with terrestrial infrastructure impacted by natural or man-made disasters [1]. The deployment of UAVs for WPT in disaster-stricken areas necessitates strategic path planning to maximize energy harvesting (EH) and efficiently reach critical locations, hereinafter referred to as strategic locations. The goal is to find the UAVs' trajectories, ensuring they cover strategic locations within the affected area while concurrently harvesting energy to collect data from the ground IoT devices. Path planning algorithms need to consider the energy-harvesting capabilities of the UAVs, the distribution of IoT sensors' energy-demanding devices, and the post-disaster dynamic environmental conditions.

Deep reinforcement learning (DRL) has recently appeared as a promising solution for UAVs and their autonomous movements. The agent is trained to learn the optimal control policy and make autonomous decisions through interactions with their environment. In particular, DRL techniques solve several challenges in state-of-the-art techniques. First, they offer real-time and online-based solutions to most of the complicated problems of using UAVs that navigate an area and learn how to interact with environments, allowing them to be used as intelligent machines in places that humans cannot reach, such as volcanoes. Second, DRL solves and provides efficient solutions to complex problems that traditional optimization techniques cannot solve. These exceptional features enable the DRL solutions to be an excellent choice for most path planning and UAV missions. Most of the UAV's environments are dynamic, and a prompt interaction is required to be taken, which poses a challenge for conventional RL techniques [9]. Meta-reinforcement learning (Meta-RL) techniques have emerged as a potential solution to address this challenge, in which the agent is trained to learn the optimal policies in environments with similar constructions quickly [10].

Nevertheless, most of the existing studies investigate single aspects of utilizing UAVs for WPT and wireless information transfer (WIT), such as considering a single UAV or ignoring the practical constraints of UAVs, including UAVs' energy consumption and UAVs' flight duration. This study addresses a critical gap by concentrating on the optimization of WPT and data collection services specifically tailored for strategic locations. This investigation delves into the intricate problem of WPT and data collection while meticulously considering practical constraints inherent in UAV operations, such as energy consumption, limited flight duration, and the minimum data rate for efficient data transmission. The objective is to strike a balance between maximizing EH, ensuring efficient data collection, and adhering to the UAV's operational limitations. The proposed study contributes to the advancement of UAV-based disaster response capabilities by providing a comprehensive analysis of the intricate interplay between WPT, data collection, and the inherent challenges associated with practical UAV constraints. Our contributions can be summarized as follows:

- We formulate a system model involving a UAV swarm that navigates a designated area, providing WPT to scattered IoT devices and enabling WIT from ground-distributed IoT devices. Unlike previous research work, this model considers essential constraints, including UAVs' energy consumption, minimum data rate, and maximum flight duration, focusing on covering strategic locations such as post-disaster-stricken areas for better UAV services.
- We delineate our approach as an optimization problem aimed at maximizing the total EH of IoT devices through downlink channels from UAVs. The harvested energy in the downlink channels enables the IoT devices to transmit their data to the UAVs via uplink channels. The optimization problem seeks maximum EH by evaluating the trajectory paths of UAVs through the covered area, specifically focusing on strategic locations to enhance service delivery.
- The formulated optimization problem is a non-linear programming (NLP) problem, and employing conventional optimization techniques for its solution is challenging and time-consuming. To address this, we opt for a real-time solution leveraging deep learning techniques. Given the dynamic nature of post-disaster environments, traditional reinforcement learning (RL) struggles to adapt to continuous changes. Consequently,

we adopt a lightweight and online solution by employing meta-RL.
- We investigate the proposed system model through extensive simulations to prove its performance by testing it on various parameters and comparing our adopted meta-RL solution with two state-of-the-art algorithms: RL with DQN algorithm and particle swarm optimization (PSO), and a greedy solution. We also draw the maximum threshold on EH and data rate as a benchmark for our solutions. We demonstrate that our adopted meta-RL algorithm outperforms the competitive algorithms.

The rest of this article is organized as follows: Section II presents the related work, and Section III presents the description of our system model. In Section IV, we delineate the problem formulation. Section V introduces the adopted meta-RL model. Section VI explains the implementation results of the proposed approach. At the end, Section VII concludes and discusses the future research directions.

II. RELATED WORK

Various optimization objectives of WPT are investigated in existing studies. In [2], UAVs are explored for WPT in unknown environments. The optimization problem jointly optimizes UAVs' search effectiveness, energy harvesting, and energy consumption. Another study, [16], optimizes the allocation of time slots for different tasks to maximize the offloaded data rate to the base station. In [11], [17], researchers focus on joint optimization of UAV trajectory planning and analog beamforming to enhance wireless power transmission. The work in [12] explores the maximization of EH by finding the optimal position of the UAV. While these studies tackle the WPT challenge from different angles, critical parameters such as using a single UAV for the mission and UAV's time duration and energy consumption are not considered in the optimization formulation. To overcome the problem, the studies in [13], [14] investigate WPT with multiple UAVs to complete the mission while minimizing the completion time and data rate, respectively. Nevertheless, these studies explore the problem of WPT while ignoring crucial parameters that affect the delivery of wireless power, such as the constraints related to energy harvesting from UAVs. To tackle these problem, we investigate the problem of minimizing EH of IoT devices while considering crucial parameters that affect WPT and data collection.

Over the past decades, extensive research has delved into optimizing UAV path planning for WPT to IoT devices. Dynamic solutions, particularly DRL, have garnered significant attention in UAV-enabled WPT systems. In [3], the approach involves a single UAV navigating an area to provide power to ground IoT devices. However, relying on a single UAV proves insufficient to cover and adequately serve the area. Moreover, utilizing DRL techniques for real-time path planning solutions is also investigated in [9], [15], [18], [19]. In [9], meta-RL is employed for UAV trajectory planning, leveraging dynamic meta-RL attributes to maximize the ground coverage of users in dynamic environments. Meanwhile, Xu et al. in [18] focuse on multi-UAV trajectory planning for data collection, emphasizing the minimization of mission completion time. Another study, [19], uses DRL techniques for UAV position control to track a ground-based object, and Wang et al. in [15] utilize DRL to maximize the computation efficiency while optimizing UAVs and mobile devices positions. Bezziane et al. in [20] investigated the communication protocols between UAVs to ensure efficient communication and efficient UAVs trajectory planning. Notably, none of the previously mentioned studies [9], [15], [18], [19], [20] investigates estimating the UAV's position specifically for providing WPT and WIT, with a focus on strategic locations in the covered area, which is deeply investigated in this paper.

For efficient use of UAVs for data collection, UAV-based approaches need to strike a balance among trajectory paths, energy consumption, and completion time [21], [22], [23]. In [21], the authors advocate for joint optimization of UAV positions and transmit power to ensure reliable information transmission and swift data collection from ground users. Meanwhile, [22] introduces joint position and travel path optimization to minimize energy consumption during data collection, focusing on optimizing trajectory paths to collect more user data within energy constraints. In [23], DRL is employed for UAV trajectory planning optimization, emphasizing finding optimal paths under time constraints and ensuring timely delivery of collected data to a central station. These studies [21], [22], [23] primarily concentrate on optimizing UAV path planning to minimize energy consumption while maximizing coverage for extensive data collection. However, none of these studies investigates UAV path planning for efficient data collection by utilizing UAVs for energy harvesting while considering critical constraints for improving WPT and WIT, explicitly for disaster-affected areas, with a focus on vulnerable spots to ensure strategic trajectory planning for data collection from the most affected ground IoT devices. Our approach addresses these challenges by utilizing UAVs for WPT and WIT by estimating the optimal UAVs paths. In particular, UAVs navigate the entire area and strategically focus on trajectories in critical locations (e.g., post-earthquake, flood-prone areas), ensuring a minimum data rate for reliable data transmission to the UAVs in the swarm. Additionally, as UAVs traverse the area, they remotely provide power to IoT devices, ensuring successful data collection before uploading it to the remote-control station.

A multitude of research studies have delved into UAVs path planning, with a predominant focus on ensuring transmission reliability [24], [25], [26], [27]. Particularly, the authors in [24] investigated UAVs' role as relay devices in reliability systems, emphasizing the transmission of control information. In a related context, [25] investigated the required data rate for successful control data delivery between the base station and UAVs while maintaining Quality of Service (QoS) reliability standards. Meanwhile,

TABLE 1. Summary of relevant related works.

Reference	Optimization Objective(s)	Optimization Parameter(s)	Optimization Constraint(s)	Data Collection	UAVs type	Strategic location	Optimization Technique(s)
[2]	Maximize UAV's search efficiency, total harvested energy, and UAV's energy utilization efficiency. Minimize UAV's flight energy consumption.	clustering sensor network, UAVs motion control, energy utilization efficiency.	UAV's energy consumption	NO	single UAV	NO	UAV Motion Control (UMC) algorithm and Dynamic Genetic Clustering (DGC)
[3]	Maximize sum-rate, EH, and minimize UAV's energy consumption.	UAV's flight speed, and elevation angle	UAV's energy consumption	YES	single UAV	NO	Deep learning with extended deep deterministic policy gradient (DDPG).
[11]	Maximize EH of ground users	Transmit power, analog beamforming, UAVs positions	UAV's energy consumption	NO	single UAV	NO	Cosine-based approximation algorithm
[12]	Maximize EH of Ground Nodes (GN)	UAV position	UAV's energy consumption	NO	single UAV	NO	V-shaped and Inverted Trapezoidal WPT,
[13]	Minimize total completion time and UAVs' energy consumption.	UAVs' path planning	Flying time, coverage region, QoS requirement, and number of UAVs.	YES	multi UAVs	NO	Peer-to-peer UAV-IoT sensing networks and clustering UAV-IoT sensing networks
[14]	Minimize the maximum data rate.	UAVs' path planning, and time allocation.	UAVs energy consumption, and UAVs flight duration	YES	multi UAVs	NO	Successive convex approximation method
[15]	Maximize the computation efficiency.	Allocate charging time, schedule computation tasks, UAVs' path planning, mobile devices optimal positions.	UAVs and mobile devices energy consumption.	NO	multi UAVs	NO	Multi-task Deep Reinforcement Learning (DRL) and a heuristic algorithm.
our work	Maximize EH of IoT devices	UAVs' path planning	UAVs' energy consumption, UAVs' flight duration, and minimum data rate for data transmission.	YES	multi UAVs	YES	Meta-Reinforcement Learning (Meta-RL)

the work by She et al. in [26], delved into optimizing UAV positions to achieve reliable transmissions. Despite their contributions to maximizing data rate for transmission reliability, these studies did not consider the crucial aspect of minimum energy consumption by UAVs, which is the primary focus of our work in this paper. Our research in [27] studied the reliability of transmission. However, in this paper, we consider a simple scenario where the UAVs navigate an area to find the optimal positions of UAVs that minimize the completion time. We did not investigate delivering WPT to ground IoT devices to improve the services that UAVs were intended to offer, which is the main focus of this work. Furthermore, we introduce several dynamic solutions, including two state-of-the-art algorithms based on RL and PSO and one greedy algorithm, to enrich the implementation of our proposed approach. We present a comprehensive summary of some closely related works in TABLE 1.

This paper tackles the intricacies associated with employing a UAV swarm for WPT to navigate specific strategic locations to enhance service provision. The primary goal is to determine optimal UAV positions that maximize the efficiency of EH of IoT devices. This optimization accounts for various factors, including the cumulative energy consumption of UAVs, ensuring a minimum data rate for reliable transmission, and adhering to the maximum time duration constraints of UAV operations. To address this complex challenge, we formulate the problem as an NLP problem, a computationally demanding task for conventional optimization techniques. Consequently, we propose an innovative online solution employing a deep learning technique, specifically meta-RL, to acquire knowledge about the optimal dynamics of the control policy for real-time decision-making. This approach aims to overcome the inherent complexities of the optimization problem, providing a dynamic and adaptive solution to the challenges associated with UAV swarm-enabled WPT in strategically important areas.

III. SYSTEM MODEL

Figure 1(a) depicts our system model. The covered area is partitioned into equal-sized C cells, $\bar{C} = \{1, 2, \ldots, C\}$, with each cell $c \in \bar{C}$. According to this approach, U UAVs, represented as $\bar{U} = \{1, 2, \ldots, U\}$, navigate to cover C cells, with a focus on some strategic locations denoted by q cells, where $q \in \bar{C}$ cells, representing areas most impacted by natural disasters like earthquakes or hurricanes. Each UAV at one cell is denoted by u^c is tasked with remotely delivering

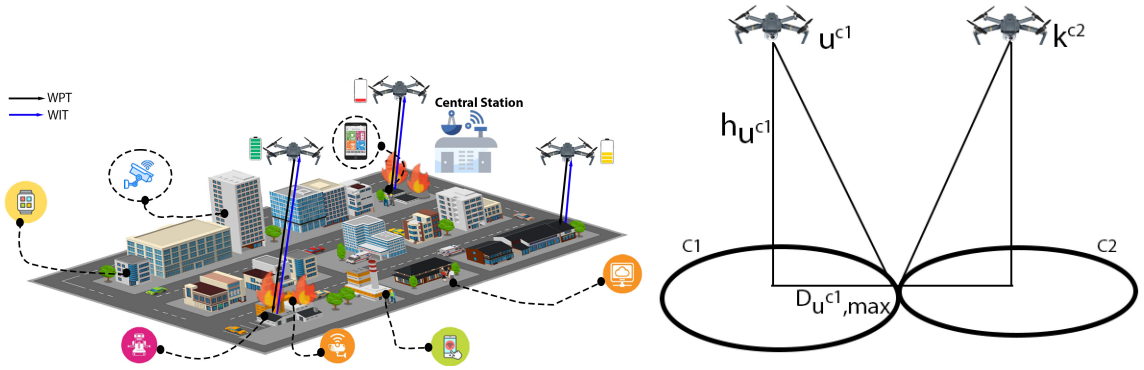


FIGURE 1. System model for multi-UAVs covering an area with strategic locations. The mission of UAVs is to deliver wireless power and collect data from distributed IoT devices.

power and collecting data from randomly distributed IoT devices, $N \gg 1$, where $\bar{N} = \{1, 2, \ldots, N\}$. The positions of ground IoT devices are denoted by i, $Q_i = [x_i, y_i, 0]$, with $i \in \bar{N}$, are assumed to be known using global positioning systems (GPS). Each UAV is equipped with two antennas, one for WPT in the downlink channel, facilitated by UAVs onboard batteries, and the other for collecting data from ground IoT devices [3]. The movements and directions of UAVs are optimized using a central station (control center) denoted as B_c, as illustrated in Figure 1(a). The UAVs depart from this central station and return to it to upload the collected data, pointing to completing one round of the UAVs. Furthermore, Figure 1(b) illustrates the maximum distances between UAVs, emphasizing the need for each UAV to maintain a safe distance to prevent collisions.

Let us consider that the UAVs operate within a time frame T, where $T > 0$ in seconds (s). For simplicity, we assume that each UAV maintains a fixed altitude h_{u^c}, complying with regulatory standards and safety requirements, with $u^c \in \bar{U}$ representing the UAV index in the set $\bar{U}$. Additionally, we assume that the time frame T is divided into M equal time slots, characterized by the set $\bar{M} = \{1, \ldots, M\}$ representing the set of time slots within one-time frame [28]. Each time slot t indicates a single movement of the UAVs and is defined by the time step duration $\mu = \frac{T}{M}$, chosen to ensure a stable time interval for determining the 3D location of the UAV. Consequently, based on time slot t, the 3D position of the UAVs can be expressed as $Q_{u^c}[t] = [x_{u^c}[t], y_{u^c}[t], h_{u^c}[t]]$, where $t \in \bar{M}$ denotes the time slot index in the set $\bar{M}$. UAVs navigate the area in each time slot t, providing WPT and collecting data. The collected data from ground IoT devices is uploaded to the ground central station B_c after the UAVs complete a full-time frame T. The UAVs provide WPT in the downlink channel from batteries carried onboard UAVs.

A. WIRELESS CHANNEL MODEL

In our approach and for practical scenarios, the obstacle information, including their number, height, and locations, might not be known; hence, the randomness of the availability of LoS and non-line-of-sight (NLoS) channels of the air-to-ground link between UAVs and IoT devices are considered. Note that the LoS and NLoS depend on the type of environment (e.g., rural, urban, suburban, etc.), the location of UAVs and IoT devices, and the altitude of the flying UAVs. Hence, the probability of the expression of LoS is given by [29]:

$$P_{i,u^c}^{LoS} = \frac{1}{1 + \omega_1 \exp\left(-\omega_2\left[\theta_{i,u^c} - \omega_1\right]\right)}, \tag{1}$$

where ω_1 and ω_2 are constant parameters, and their values are specified based on the type of the environment, θ_{i,u^c} represents the elevation angle between the UAV u^c and the IoT device i. In particular, $\theta_{i,u^c} = \frac{180}{\pi} \times \sin^{-1}(\frac{h_{u^c}}{d_{i,u^c}})$, where $d_{i,u^c} = \sqrt{(x_{u^c} - x_i)^2 + (y_{u^c} - y_i)^2 + h_{u^c}^2}$ is the distance between the UAV u^c and the IoT device i. The probability of NLoS is given by $P_{i,u^c}^{NLoS} = 1 - P_{i,u^c}^{LoS}$.

The path loss models of LoS and NLoS are expressed as follows [29]:

$$L_{LoS}^{i,u^c} = \psi_{LoS}\left(\frac{4\pi f_c d_{i,u^c}}{c}\right)^2 \tag{2}$$

$$L_{NLoS}^{i,u^c} = \psi_{NLoS}\left(\frac{4\pi f_c d_{i,u^c}}{c}\right)^2, \tag{3}$$

where f_c is denoted as the carrier frequency and c is the speed of light. ψ_{LoS} and ψ_{NLoS} are denoted to the excessive path loss related to the loss of free space propagation for LoS and NLoS, respectively. The total average path of the communication link between the UAVs and the IoT devices is given by:

$$\bar{L}_{i,u^c} = P_{i,u^c}^{LoS} L_{LoS}^{i,u^c} + P_{i,u^c}^{NLoS} L_{NLoS}^{i,u^c} \tag{4}$$

Moreover, the average channel gain of the communication link between IoT devices and UAVs is $\bar{g}_{i,u^c} = (1/\bar{L}_{i,u^c})$. The data rate for data transmission between IoT devices and UAVs is expressed as:

$$\rho_{i,u^c} = B_{i,u^c}\, log_2\left(1 + \frac{P_r}{\sigma^2}\right), \tag{5}$$

where B_{i,u^c} is the transmission bandwidth between the UAV u^c and IoT device i, P_r is the received power at the UAV u^c, and $P_r = P_i \times \bar{g}_{i,u^c}$, P_i is denoted to the transmit power at the device i, and σ^2 represents the thermal noise power. Hence, to achieve reliable data collection at UAVs from IoT devices at each time slot t, the following condition needs to be satisfied:

$$\rho^t_{i,u^c} \geq \rho^{th} \tag{6}$$

B. DEVICE-TO-DEVICE (D2D) TIME DELAY MODEL

D2D time delay is the time required to complete the data collection and send it to the central station. The completion time contains the time required to collect the data from IoT devices at time slot T^t_{data} and the time required by the UAVs to move between two successive time-slots t and $t+1$ and denoted by $T^t_{u^c}$, and they can be expressed by the following expressions:

$$T^{i,u^c}_{data}[t] = \sum_{i \in N} \frac{K_i}{\rho^t_{i,u^c}} \tag{7}$$

$$T_{u^c}[t] = \frac{\|Q_{u^c}[t+1] - Q_{u^c}[t]\|}{V}, \ t = 1, \ldots, M, \tag{8}$$

where K_i is the IoT's data packets and V is the average speed of u^c-th UAV traveling between two consecutive locations. Then, the total completion time of u^c-th UAV mission is given by:

$$T^t_{tot} = |T_{max} - \left(T^{i,u^c}_{data}[t] + T_{u^c}[t]\right)| \tag{9}$$

In each cell, the varying number of IoT devices necessitates different visitation frequencies for data upload. We consider the time UAVs spend traversing through IoT devices and the time to collect the data. During each time slot, the maximum flight time of UAVs is reduced by the time required for traversal and data collection. To adhere to completion time constraints and encourage UAVs to prioritize strategic locations with higher demand services, thereby enhancing the total EH of IoT devices, we impose the completion time constraint as follows:

$$T^t_{tot} \leq T_{max}, \tag{10}$$

where T_{max} is the maximum time for UAVs to complete their mission, which is updated at each movement of UAVs.

C. ENERGY CONSUMPTION MODEL

UAVs have limited energy capacity as a result of their constrained onboard batteries. The battery lifespan is influenced by various factors, such as the UAV's energy source, type, weight, and speed. Usually, the UAV's energy usage can be categorized into three main components: propulsion energy, communication energy, and WPT energy. Communication energy is the energy required to configure the communication link to the IoT devices and data collection through the uplink and the energy of the dissemination to the central station that is taken at each time step of UAV's movement. The communication energy is smaller than the propulsion and WPT energy [30]. To model the propulsion energy, we utilize the propulsion-power model designed for rotary-wing UAVs as in [30]:

$$\varepsilon^t_{prop,u^c} = \underbrace{\eta_i \sqrt{\left(\sqrt{\left(1 + \frac{v^4_{u^c}}{4v^4_0}\right)} - \frac{v^2_{u^c}}{2v^2_0}\right)}}_{\text{Induced Power}} + \underbrace{\eta_b \left(1 + \frac{3v^2_{u^c}}{v^2_{tip}}\right)}_{\text{Blade Power}} + \underbrace{\frac{1}{2} f_0 \varphi r D_a v^3_{u^c}}_{\text{Parasite Power}}, \tag{11}$$

where η_i refers to the blade profile power and η_b refers to the induced power, v^2_{tip} indicates the speed of the UAV's rotor blade, v_0 is the rotor induced velocity, f_0 refers to fuselage drag ratio, r refers to the rotor solidity, φ refers to the air density, and D_a is the area of the rotor disc. To calculate the hovering power consumption, equation (11) is used with zero speed of the UAV, i.e., $v_{u^c} = 0$, as follows:

$$\varepsilon^t_{(hov,u^c)} = \eta_i + \eta_b. \tag{12}$$

Therefore, the propulsion energy consumption of UAV u^c at time slot t is obtained as follows:

$$\varepsilon^t_{(u^c,prop)} = \begin{cases} \varepsilon_{(prop,u^c)} \times t & \text{if } v_{u^c} > 0 \\ \varepsilon_{(hov,u^c)} \times t & \text{if } v_{u^c} = 0. \end{cases} \tag{13}$$

Moreover, the communication energy consumption is the total energy consumed by each UAV to collect the data from the IoT devices, and the energy required to disseminate the updated state with the central station B_c. If the distance between a UAV and an IoT device or the distance between the UAV and the central station is large, more communication energy is consumed, and hence, the energy consumption will be depleted quickly, and a UAV with sufficient power traverses this cell to provide services. Then, the communication energy consumed by all the UAVs (in Joules) in the swarm is expressed as:

$$\varepsilon^t_{(u^c,comm)} = \sum_{u \in U} P_{(u^c,comm)} \left(T^{i,u^c}_{data}[t] + T^{B_c,u^c}_{data}[t]\right), \tag{14}$$

where $P_{(u^c,comm)}$ is the UAV's communication power, $T^{B_c,u^c}_{data} = \frac{S_u}{\rho^t_{B_c,u^c}}$ is the time required to transmit the UAVs's state to the central station, S_u is the UAV's state to be shared with the central station, and $\rho^t_{B_c,u^c}$ is the data rate required to transmit the UAV's state to the central station B_c. This energy is designed to calculate the energy consumed by the UAV to collect data, and disseminate UAV's state to the central station while focusing on strategic locations. At each time slot t, at least one strategic location is visited by one UAV.

The UAVs navigate the area and work in a full duplex scenario. The UAVs transmit wireless power to the IoT devices with constant transmit power P_{wpt} in the downlink channel and collect information in the uplink channel from the IoT devices that exist within the UAVs' maximum

coverage power, i.e., all of the IoT devices within the coverage of the UAV receives wireless power from that UAV $\forall \Delta d_i^t \leq D_{max}$, where D_{max} is the maximum radius coverage of the UAV; hence, the received power at the IoT device i at time slot t can be expressed as follows [3]:

$$P_r^t = P_{wpt} \, |\bar{g}_{i,u^c}^t|^2, \forall \Delta d_i^t \leq D_{max}. \tag{15}$$

Thus, the UAVs energy consumption for WPT can be calculated as [3]:

$$\varepsilon_{(u^c,wpt)}^t = \sum_{u^c \in U} P_r^t \mu_{wpt}, \tag{16}$$

where μ_{wpt} is the time required to deliver wireless power to the IoT devices. Hence, the total UAVs' energy consumption at time slot t can be calculated as follows:

$$\varepsilon_{u^c,tot}^t = \varepsilon_{(u^c,prop)} + \varepsilon_{(u^c,comm)} + \varepsilon_{(u^c,wpt)}. \tag{17}$$

Due to their limited battery lifespan, UAVs need to have sufficient energy to accomplish their mission; hence, we add a constraint to ensure sufficient energy is available for the UAVs during their mission. The battery status $\Omega_{u^c}^t$ at each time slot t can be obtained as follows:

$$\Omega_{u^c}^t = \Omega_{u^c}^{t-1} - \varepsilon_{u^c,tot}^t, \tag{18}$$

where Ω_u^{t-1} is the battery level at the end of $t-1$. Let $\Omega_{u^c}^0$ denote the battery capacity before the mission starts, in which $\Omega_{u^c}^0 = \Omega_{u^c}^{init} + \Omega_{u^c}^{min}$, where $\Omega_{u^c}^{init}$ is the battery capacity of the UAV that is assigned for the mission, and $\Omega_{u^c}^{min}$ is the minimum battery level for the UAV to return to its central station, therefore, $\Omega_{u^c}^t \in [\Omega_{u^c}^{min}, \Omega_{u^c}^0]$.

For the UAVs to satisfy the constraint of the maximum energy consumption in the one-time frame, the total energy consumption of UAVs in time slot t should be greater than the minimum energy conceptions of the UAVs, i.e., UAVs need to have sufficient energy to complete the mission as expressed in the following constraint:

$$\Omega_{u^c}^t \geq \Omega_{u^c}^{min}, \quad \forall u^c \in U, \forall t \in \bar{M}. \tag{19}$$

D. ENERGY HARVESTING MODEL

For a practical scenario of the EH model, we consider the non-linear EH model [31]. Compared to a linear model, a non-linear model considers the practical limitations of the circuits. The EH of the RF-EH circuit can be expressed by [3]:

$$P_{r,i}^{harv} = \frac{P_{lim} e^{ab} - P_{lim} e^{-a(P_r^t - b)}}{e^{ab} 1 + e^{-a(P_r^t - b)}}, \tag{20}$$

where P_{lim} is the threshold of the output DC power, the parameters a and b refer to the characteristics of the EH circuits.

The total EH of IoT devices in one time frame T that is received from the swarm of UAVs can be expressed as:

$$P_{r,tot}^{harv} = \sum_{i \in N} \delta_{u^c,i,q} \, . \, P_{r,i}^{harv}, \forall u^c \in U, \forall q \in C, \tag{21}$$

where $\delta_{u,i,q}$ is a binary constraint to encourage UAVs to pass through strategic locations and exploit their energy to collect data from the IoT devices, and it is defined as:

$$\delta_{u^c,i,q} = \begin{cases} 1, & \text{if UAV } u^c \text{ is collecting data from IoT } i \text{ that} \\ & \text{is in strategic location } q, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

IV. PROBLEM FORMULATION

The main objective is to maximize the total EH of the ground IoT devices presented in equation (21) by finding the optimal positions of UAVs while respecting the constraints of the maximum completion time, minimum UAVs' energy consumption, and minimum achievable data rate for reliable data collection. The problem formulation is expressed as follows:

$$\boldsymbol{P} = \max_{Q} \; P_{r,tot}^{harv} \tag{23}$$

subject to:

$$\text{C1: } \rho_{i,u^c}^t \geq \rho^{th}, \tag{23a}$$
$$\text{C2: } T_{tot}^t \leq T_{max}, \tag{23b}$$
$$\text{C3: } \Omega_{u^c}^t \geq \Omega_{u^c}^{min}, \quad \forall u^c \in \bar{U}, \forall t \in \bar{M}, \tag{23c}$$
$$\text{C4: } d_{u^c,k^v} \geq 2D_{max}, \; \forall u^c, k^v \in \bar{U}, \tag{23d}$$
$$\text{C5: } \sum_{i=1}^{N} \delta_{u^c,i,q} \geq 1, \forall u^c \in \bar{U}, \forall q \in \bar{C}, \tag{23e}$$
$$\text{C6: } Q(0) = Q(M), \tag{23f}$$

$$\delta_{u^c,i,q} \in \{0, 1\} \tag{24}$$

The objective function presented in Equation (23) is designed to maximize the total EH of the distributed IoT devices. It achieves this by determining optimal UAV paths that encompass navigating to cover the area, delivering wireless power, and collecting data from IoT devices, with a specific focus on visiting strategic locations. The mission needs to adhere to several constraints, including the maximum completion time, the minimum data rate, and the maximum energy consumption of UAVs to ensure reliable data transmission from IoT devices to airborne UAVs. The optimization problem presented in Equation (23) is classified as an NLP problem due to the non-linearity inherent in Equations (23), (23a), and (23c).

The constraint defined in Equation (23a) is imposed to guarantee that the transmission data rate surpasses a specified threshold, ensuring the reliable transmission of data packets, which mainly depends on the distance between the UAV and the IoT device. When the distance is very large, more EH to IoT device is required, causing quick consumption of UAV's energy and channel gain between the UAV and the IoT device are affected, causing degradation in data rate as indicated in Equation (5), hence a violation of constraint (23a). In case of violating the constraint in Equation (6), a UAV with sufficient energy and time visits this cell to provide services to the IoT device in which it takes more time and

TABLE 2. Symbols list.

Symbol	Details	Symbol	Details
N	The number of IoT devices	U	The number of UAVs
Q_i	The position of i IoT device	Q_{u^c}	The positions of UAV u^c that is in cell c.
T	The number of time frame	M	The number of time slots in one time frame T
d_{i,u^c}	The distance between UAV u^c at cell c and IoT device i	f_c	The carrier frequency
c	The speed of light	ρ_{i,u^c}	The data rate between UAV u^c and IoT device i
P_i	The transmit power at device i	P_r	The received power at the UAV u^c
B_{i,u^c}	The transmission bandwidth between the UAV u^c and IoT device i	ω_1 and ω_2	Constant parameters and their values are specified based on the type of the environment
V	The average speed of u^{th} UAV traveling between two consecutive locations	$\tau_{u^c,t}$	The u^{th} UAV's travelling time between two successive time-slots t and $t+1$
T_{data}	The time required to transfer the data from IoT devices	T_{com}	The time required by the UAVs to accomplish the mission
K_i	The IoT's data packets	$\Omega_{u^c}^{min}$	The minimum safe energy return of UAV u^c.
$P_{r,i}^{harv}$	The EH of IoT device i	σ^2	Thermal noise power
$\varepsilon_{u^c,tot}^t$	The total energy consumption of UAV u^c at time slot t.	$\bar{g}_{i,u^c}$	Channel gain between UAV u^c and IoT device i
T_{max}	The maximum time for UAVs to complete their mission	$\delta_{u^c,i,q}$	A binary constraint to encourage UAVs pass through strategic locations
P	The UAVs transitions probability	π	meta-RL policy
γ	Discount factor	α	Learning rate
$v^\pi(s)$	The value of applying an action to environment state s	θ	Neural network weights
ϵ	Exploration rate	A	The vector set of actions

energy than closer IoT devices to upload its data to the UAV successfully. Simultaneously, the constraint specified in Equation (23b) is established to ensure that UAVs operate within their maximum duration before returning to the central station. The energy consumption of the UAV is bounded, leading to the formulation of constraint (23c), which ensures that the UAV's energy consumption remains sufficient for its mission, encompassing the delivery of WPT and data collection. Additionally, constraint (23d) is introduced to maintain a safe distance between UAVs, mitigating the risk of collisions within the UAV swarm. The constraint presented in Equation (23e) is devised to guarantee that, at each time step or time slot *t*, at least one UAV visits one of the strategic locations. Lastly, The constraint in Equation (23f) indicates that the starting point of the UAVs and the endpoint is the same.

Solving the objective function outlined in Equation (23) poses a challenge when employing conventional optimization techniques, primarily due to the non-linearity inherent in constraints (6) and (19). Consequently, we opt for a real-time-based solution, employing meta-RL to address both the objective function and its associated constraints. The subsequent sections will elaborate on the application and methodology of this real-time solution.

V. META-REINFORCEMENT LEARNING FOR EFFICIENT ENERGY HARVESTING AND UAVS PATH PLANNING

In this section, we present the system model of our adopted solution to address the optimization problem (23) utilizing a deep learning technique. Conventional RL algorithms require a specific number of episodes to learn the optimal policy and converge into the maximum expected reward. With any change in the environment, the agent has difficulties converging so quickly to the maximum rewards. Due to the dynamic environment of our adopted scenario, in which UAVs join and disconnect many times due to the continuous time of working, hence, we adopt a lightweight solution using meta-RL to solve the problem of dynamicity in the environment, namely meta-RL. Meta-RL proves to be an effective technique for handling dynamic environments and learning optimal policies with fewer episodes compared to traditional deep learning methods [10]. Given the dynamic nature of the UAV swarm environment in post-disaster scenarios, where UAVs can depart and join the swarm intermittently (e.g., for recharging), our adopted approach needs to exhibit flexibility in accommodating these unforeseen changes while adhering to some critical constraints during the process of delivering wireless power and collecting data from IoT devices. Due to the complexity of the problem (23), conventional RL algorithms are slow to converge again to respect the constraints of path planning. Consequently, meta-learning is adopted to quickly cope with new environments efficiently across related tasks. Specifically, a model is trained to acquire the capability to learn new tasks effectively and converge rapidly, demonstrating superior efficiency in handling a variety of related tasks compared to focusing solely on a single task.

Meta-RL applies meta-learning to RL and aims to make the agent learns the general policy of related tasks. A task consists of a set of states and, actions and rewards [10]. In particular, meta-RL agent does not aim to learn the optimal policy of a specific task; instead, they aim to learn the general policy that can be applied to new environments with the same family of tasks. The benefit is that it enables the agent

to quickly reach the optimal policy of new environments with a minimum number of episodes.

In our approach, each episode encompasses a series of time steps, representing a single time frame T, during which the UAVs in the swarm depart from the central station and complete one round. These time steps correspond to the time slot t, and at each step, the algorithm has the task of choosing the best positions for UAVs. The decision-making process, aimed at selecting the optimal positions for UAVs within the swarm, is guided by multiple considerations. These include the imperative to maximize total EH while adhering to constraints related to UAVs' energy consumption and ensuring successful data transmission from ground IoT devices to UAVs within the allotted time duration of the UAVs. Our approach is conceptualized as a Markov Decision Process (MDP), denoted by (S, A, P, R, γ). In this framework, S encapsulates the environment state, A signifies the action vector representing the path planning of the UAVs, P characterizes the probability of possible transitions, R denotes the rewards associated with each action taken by the agent, and γ is dedicated to the discount learning factor. This MDP formulation provides a structured framework for decision-making within the dynamic environment of UAV swarm navigation.

A. ENVIRONMENT MODELING

The environment in our approach is represented by a swarm of UAVs navigating a designated area, with a specific emphasis on strategic locations. Let us designate π as the optimal stochastic policy that the agent endeavors to learn, where $\pi : S \times A \rightarrow [0, 1]$. The meta-RL algorithm receives information from a centralized agent that interacts with the environment, selecting action A, receiving either a reward or penalty for that action, and incrementally refining the optimal policy π^* based on the accumulated reward R at each time step t. The algorithm aims to attain the optimal policy π^* with maximum $v^{\pi^*}(s)$ for all parameters $s \in S$. The values within the set $v^{\pi}(s)$ reflect the feedback from the reward subsequent to executing action a in state s and can be elucidated as:

$$v^{\pi}(s) = \mathbf{E}_{a_t, s_{t+1}}\left(\sum_{k=1}^{\infty} \gamma^{k-t} R_k | S_t = s\right), \tag{25}$$

where $\mathbf{E}$ represents the expected value.

B. STATES AND ACTIONS

The RL agent requires informative cues about the environment to enhance the system's performance. Our approach employs a state vector encompassing pivotal parameters: the positions of UAVs, strategic locations, the record of visited cells in the grid, the number of IoT devices in each cell, the index of the UAVs, and the energy consumption of all next paths. The agent's action involves determining trajectory directions for UAVs, with a specific emphasis on strategic locations. These actions are chosen to maximize the objective function defined in Equation (23). This function seeks to optimize total EH, considering critical parameters like UAV energy consumption, maximum duration time, and minimum data rate for reliable packet transmission. Figure 2 illustrates the asynchronous advantage actor-critic framework. The adopted system is a centralized framework where the agent receives the environment, state, and action for processing them, facilitating fast convergence and rapid learning of the optimal policy.

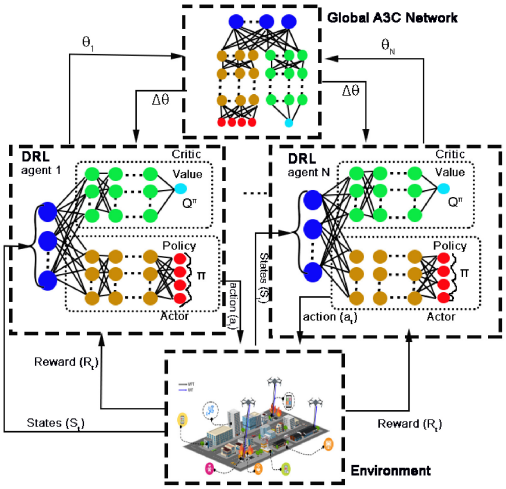


FIGURE 2. System Model for A3C algorithm for a swarm of UAVs covering an area with strategic locations. The UAVs mission is WPT in the downlink channel and WIT in the uplink channel.

The state S contains the details that affect the movements of UAVs in the grid and help the agent learns the optimal policy by choosing the correct action. Therefore, at each time step, the agent receives the information that contains the following details:

- $Q_{u_u^c} = \{Q_{u_1^c}, \ldots, Q_{u_u}^C\}$: the position of UAVs in the grid,
- $P_{u_1} = \{P_{u_1}, \ldots, P_{u_u}\}$: the UAVs paths in each time frame T,
- $q = \{q_1, \ldots, q_n\}$: the position of strategic location in the grid.
- $N_i = \{N_1, \ldots, N_C\}$: the number of IoT devices at each cell in the grid.
- $\varepsilon_{tot} = \{\varepsilon_1, \ldots, \varepsilon_8\}$: the energy consumption that the UAV will consume in all neighboring cells,

where q_n indicates the index of the strategic locations in the grid.

Let us define the action taken by the agent to learn the optimal policy of actions $A^*(t) = \{a_1^*, \ldots, a_8^*\}$. The action of UAV path planning at a time slot or time step t, which represents the directions of UAVs that can take in the neighboring cells, can be written as $A(t) = \{a_1, \ldots, a_8\}$, which are up, down, left, right, up-right, up-left, down-left, down-right. The selection of the optimal policy can be expressed as:

$$A^*(t) = \text{argmax}_{a_i} A(t) \tag{26}$$

C. REWARD FUNCTION

The reward function plays a pivotal role in guiding the algorithm toward learning the optimal policy. In our algorithm, the agent receives a positive reward when the UAV's direction is selected to maximize EH and an additional reward when visiting one of the strategic locations. The algorithm focuses on finding UAV paths that maximize EH while adhering to the constraints outlined in Section IV. Specifically, a positive reward is granted when all constraints outlined in Equation (27) are respected. Conversely, the agent incurs a negative reward if any of the constraints specified in Equation (27) are not respected. The imperative for UAVs is to traverse through strategic locations and provide better services. The instantaneous reward function, encapsulating the total reward for adhering to the constraints of our adopted solution, can be expressed as follows:

$$R_t = C1 \times C2 \times C3 \times C4 \times C5 \tag{27}$$

Equation (27) outlines the total rewards that the algorithm aims to maximize by selecting the appropriate action for the UAV swarm. The algorithm receives a positive reward when all constraints are adhered to, and in case of any violations, penalties are incurred. C1 represents the objective function in equation (23), incentivizing the algorithm to choose the direction that maximizes EH among all available paths. C2 corresponds to the constraint in equation (23a), ensuring the minimum data rate for transmitting data from the IoT device to the UAV. C3 pertains to the constraint in equation (23b), indicating the maximum duration of UAVs in a one-time frame T. Therefore, if all maximum time T_{max} is consumed, i.e., violating this constraint, the UAV returns back to the central station even if it has not completed its maximum steps for the episode by the minimum energy for safe return. The agent considers this violation, and a penalty is received as feedback of this episode. The UAV, with enough time, needs to traverse through this cell and provide services to the IoT devices. C4 refers to the constraint established in equation (23c), ensuring that UAVs operate within their energy consumption capabilities, guaranteeing they have sufficient energy to accomplish their mission. Therefore, if any UAV does not have sufficient energy to accomplish the mission, i.e., constraint violation, the UAV returns to the central station even if it has not completed its maximum steps for the episode, and the agent receives a penalty pointing to energy constraint violation. The agent will learn from this path that it needs more energy and causes violation; hence, another UAV with strong energy traverses through this cell and provides services. C5 addresses the constraint in equation (23d), ensuring that UAVs maintain a safe distance during navigation to prevent collisions. When all these constraints are observed, the agent receives a positive reward, signifying that a commendable action was taken, encouraging UAVs to navigate within their capabilities and aligning with practical considerations for flying UAVs.

The UAV swarm needs to traverse strategic locations, deliver wireless power, and collect data from IoT devices. Each strategic location carries various service demands based on the number of IoT devices within the cell and its importance. This demand service factor represents the QoS in these strategic locations and signifies how satisfied strategic locations are with the services coming from the UAV swarm. As UAVs traverse a strategic location, they fulfill a portion of the demand service, denoted by ϕ_q. Consequently, we define the rewards for satisfying the number of visits to these strategic locations, as done in [32], by:

$$R_q = \frac{1}{1 + \sum_{i=1}^{q} \phi_q(t)}, \tag{28}$$

where q is the set of strategic locations.

D. RL AGENT MODELING

The agent aims to maximize their rewards by experimenting with multiple actions in the environment. Through the interaction of actions and the environment, the agent learns the optimal policy π^* from their surroundings. The optimal policy is achieved by formulating a strategy encompassing the optimal sets of action-state values $Q(S_t, A_t)$. These values assist the agent in understanding how to anticipate future rewards by selecting the best action from the available optimal set of actions A_t. The action-value $Q(S_t, A_t)$ of the agent signifies the performance metrics of the chosen action that the RL agent should take in the subsequent same states. It can be represented as:

$$Q(s_t, a_t) = \mathbf{E}_{a_t, s_{t+1}}\left(\sum_{k=1}^{N} \gamma^k R_{t+k} | S_t = s, A_t = a \right), \tag{29}$$

where γ denotes the discount factor and $\gamma \in [0, 1]$ which indicates the connection of the immediate reward to the long-term rewards. The transitions between states-actions can be calculated by [19]:

$$Q(s, r) \leftarrow Q(s, r) + \alpha\left(R_t + \gamma \max_a \hat{Q}(s, a) - Q(s, a) \right), \tag{30}$$

where α indicates the learning factor.

The agent is the combination of actor-critic algorithm in which the actor learns the optimal policy through dynamic interaction with the environment and the critic learns to evaluate those actions to improve the performance. The central agent exists at the central station that shares information with the UAVs about the positions of other UAVs, remaining energy and time, and the positions of the IoT devices in the grid, leading to efficient performance making sure that the UAVs work together to improve the system's performance.

Algorithm 1 delineates the steps of the meta-RL algorithm to learn the optimal policy and respect the constraints related to the objective function in (23). Each episode represents a one-time frame T, which indicates the paths of all UAVs until returning to the central station to upload the collected data to the central station. Also, each time slot represents a one-time step for the UAVs during one round of their mission. The algorithm aims to select the action that maximizes the

Algorithm 1 Meta-RL With Asynchronous Advantage Actor-Critic Algorithm

1: **Initialize:** define empty set: E_{harv}
2: **Initialize:** Q-network parameters θ and θ_v.
3: **Initialize:** T_{max} initial Time for the mission.
4: **Initialize:** old Q-network parameters θ' and θ'_v
5: **Output:** UAVs trajectory paths
6: **for** Each $r \in R$ episodes **do**
7: Gradient initialization: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$
8: Q-network initialization: $\theta' = \theta$ and $\theta'_v = \theta_v$
9: **for** Each time slot $t \in M$ **do**
10: **for** Each UAV $u^c \in \{1, \ldots, U\}$ **do**
11: $S(u) = \{Q_{u^c}[t], P_u[t], q_{u^c}, N_i^C, \varepsilon_{tot}^C\}$
12: choose action A_i based on ϵ
13: **if** constraints in Equation (27) **then**
14: $R_t += 1$
15: **else**
16: $R_t += 0$
17: **if** $S_{u^c}[t] \in q_u^C$ **then**
18: $E_{harv} += P_{r,tot}^{harv}$, Equation (23)
19: $R_t = R_t + R_q$, Equation (28)
20: UAV's consumed time Equation (9)
21: UAV's consumed energy Equation (18)
22: Update T_{max} for the UAV u^c
23: Observe S_{i+1}
24: Observe R_t
25: System reset with new UAVs positions
26: Save (S_i, A_i, r_i, S_{i+1}) in replay memory
27: Taking a mini-batch of (S_i, A_i, r_i, S_{i+1})
28: Gradient accumulation wrt θ' : $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i; \theta')(R - V(s_i; \theta'_v))$
29: Gradient accumulation wrt θ'_v : $d\theta \leftarrow d\theta + \partial(R - V(s_i; \theta'_v))^2/\partial\theta'_v$
30: Asynchronous update of θ using $d\theta$ and θ_v using $d\theta_v$

total reward with the help of the details provided in the state vector. If the agent chooses the right action that satisfies all the constraints designed in equation (27), it receives a reward of 1; otherwise if any constraint is not respected, a penalty of 0 is added to the total reward function [lines 13-16]. Each strategic location has a different demand service, indicating the number of IoT devices inside it that need WPT to upload their data to the UAVs. Hence, each UAV satisfies part of this demand service each time it visits a strategic location as indicated by equation (28), and the UAVs collaborate to satisfy their demand services. If the agent traverses through one of the strategic locations, the demand service it satisfies is added as a reward to the total reward function, and the total EH of the strategic location is calculated [lines 17-19]. Then, the remaining time and energy of the UAVs are calculated to be checked in the next movements as to whether the UAVs have sufficient energy and time to deliver WPT and collect data [lines 20-21]. The max time for UAV to provide services is updated to respect the constraint of maximum duration of UAVs [line: 22]. Then, the reward, action, and state are observed as feedback to the agent to learn the optimal policy that leads to the maximum reward. We emphasize that in traditional RL algorithms, the agent receives the reward just only as a reward at the end of each time step, while meta-RL receives the action, state, and outcome reward, so it learns faster and adapts itself to the new environments from the same type quicker than traditional RL algorithms.

VI. SIMULATIONS AND ANALYSIS

This section details the simulations conducted to test and evaluate the performance of the proposed system model. We compute the maximum thresholds for rewards, EH, and data rates as benchmarks to compare against the adopted algorithm, enriching the comparative analysis. Assuming all constraints are respected, i.e., the UAV swarm has sufficient energy, time, and data rate to deliver wireless power and collect data. The maximum threshold measures the maximum EH, reward, and data rate that all UAVs $\sum_{u^c=1}^{U} u^c$ can achieve in a one-time frame T. In particular, the maximum threshold calculations are based on the number of UAVs and IoT devices and the maximum values for rewards, EH, and data rates. Our proposed meta-RL solution is compared with one state-of-the-art RL-based algorithm represented by the DQN algorithm. The DQN algorithm leverages a replay buffer to determine the optimal policy from all past experiences [33]. We also compare our adopted Meta-RL algorithm with PSO and Greedy algorithms, outlined in Algorithms 2 and 3, respectively. In the PSO solution, the number of particles corresponds to the UAVs' paths within the grid, and it assesses paths that yield the best EH among available solutions, as outlined in Algorithm 2. Consequently, the PSO algorithm needs to examine all positions and follow the position with higher EH to attain the optimal solution. However, acquiring the optimal solution through PSO is challenging and time-consuming, given that UAVs depart from the central station via different particles, covering the grid in 8 different directions. Therefore, the number of expected paths they need to check for the one-time frame is $8^C \times U$ possibilities, where C is the number of cells, and U is the number of UAVs, resulting in an extensive number of possibilities. To simplify, we considered 1000 different particle positions within the covered area for PSO implementation. Conversely, the greedy solution examines the best EH among neighboring cells as UAVs traverse the grid, selecting the direction that yields the optimal result, as depicted in algorithm 3 [lines: 8-13].

The simulation analysis commences by presenting the total EH output of IoT devices from our adopted meta-RL solution, compared with the three alternative solutions. Subsequently, the total energy consumption of UAVs is analyzed, and the percentage of satisfaction of demand services is presented. The algorithm's adaptability to changes in the number of UAVs during training is also tested.

Algorithm 2 PSO Algorithm

Initialize: $vel, Pos, iteration, p_{best}, g_{best}, best_{fitness}$
Initialize: total EH E_{tot}
Initialize: vector contains UAVs positions as U_{u^c}
Initialize: vector contains strategic location positions as q
Output: UAVs paths with highest EH
Generate random particles (P) with cell numbers
for Each $itr \in 1000$ **do**
  for Each particle (i) **do**
    Calculate fitness of Equation (23)
    Update p_{best}, g_{best}
  for Each time slot $t \in M$ **do**
    for Each UAV $u^c \in \{1, \ldots, U\}$ **do**
      Update vel, Pos
      if $Pos > limit$ **then**
        $limit = pos$
        Calculate fitness of Equation (23, $best_{fitness}$)
        Update p_{best}, g_{best}
      $U_{u^c}[t] = index(p_{best})$
      if $U_{u^c}[t] \in q$ **then**
        $E_{tot} += best_{fitness}$
      UAV's consumed time Equation (9)
      UAV's consumed energy Equation (18)
      Update T_{max} for the UAV u^c
    system reset with new UAVs positions

Algorithm 3 Greedy Algorithm

Initialize: empty list of neighboring cells as N_{ne}
Initialize: total EH E_{tot}
Initialize: vector contains UAVs positions as U_{u^c}
Initialize: vector contains strategic location positions as q
for Each $itr \in 1000$ **do**
  for Each time slot $t \in M$ **do**
    for Each UAV $u^c \in \{1, \ldots, U\}$ **do**
      N_{ne} = EH of neighboring cells of $u_c[t]$
      $max_{EH} = 0$
      for $i \in N_{ne}$ **do**
        if $i \geq max_{EH}$ **then**
          $max_{EH} = i$
      $U_{u^c}[t] = index(max_{EH})$
      if $U_{u^c}[t] \in q$ **then**
        $E_{tot} += max_{EH}$
      UAV's consumed time Equation (9)
      UAV's consumed energy Equation (18)
      Update T_{max} for the UAV u^c
    system reset with central station UAVs positions

A. SIMULATION SETUP

In our approach, we conducted simulations on an area measuring 480 m × 480 m, divided into 36 equal-sized cells, to test the proposed system model. Within these 36 cells, three strategic locations were positioned at different places on the grid. In the simulation configurations, unless otherwise specified, we employed 3 UAVs and 400 IoT devices distributed randomly across the covered area. To ensure a fair comparison of the algorithms used in the simulations, we maintained consistent input configurations, including the same number of episodes, UAVs, and hardware for conducting experiments. The key parameters are summarized in Table 3. Each experiment was conducted over 50,000 episodes and averaged over 50 episodes for robust evaluation.

TABLE 3. Simulation parameters.

Parameters	Description	Value
V	Average flight speed	10 m/s
B_{i,u^c}	Bandwidth	1 MHz
σ^2	Noise power	-95 dBm
K_i	IoT devices data packet	2 KB
M	Time slots in one time frame T	18
γ	Discount factor	0.85
α	Learning rate	0.0001
T_{max}	Maximum flight time	180 s
$P_{u^c}^{comm}$	Communication power	5W
P_{wpt}	constant transmit power for WPT	300W
ρ^{th}	threshold for data rate	2 Mbps
h_{u^c}	the UAV's altitude	30 m
f_c	Carrier frequency	5 MHz
P_{lim}	Threshold of the output DC power	9.079e-6 μ W
a and b	EH circuits characteristics	47083e-6, 2.9e-6 μ W
ω_1 and ω_2	Parameters of urban areas	11.95, 0.14
ψ_{LoS} and ψ_{NLoS}	The excessive path loss of free space propagation for LoS and NLoS	3, 23 dB

B. NUMERICAL RESULTS AND SIMULATION ANALYSIS

Figure 3 illustrates the frequency of visits to strategic and non-strategic locations within a single time frame T for various algorithms. All approaches successfully encourage UAVs to visit strategic locations more frequently, meeting service requirements in those crucial areas. The meta-RL solution in Figure 3(a) notably outperforms other solutions by prioritizing visits to strategic locations. Conventional RL solutions, represented by the DQN algorithm in figure 3(b), achieve comparable results to the PSO algorithm presented in figure 3(c). Both DQN and PSO algorithms show comparable and superior results to the greedy solution in figure 3(d). The key distinction lies in the fact that deep learning-based solutions endow the agent with knowledge of the entire paths, maximizing EH and thereby enhancing rewards and overall system performance. The PSO solution requires the number of participants to equal the number of cells; each participant has eight paths, resulting in a challenging implementation with $8^C \times U$ possibilities, where C is the number of cells, and U is the number of UAVs. Therefore, we considered 1000 movements of C participants, showing results better

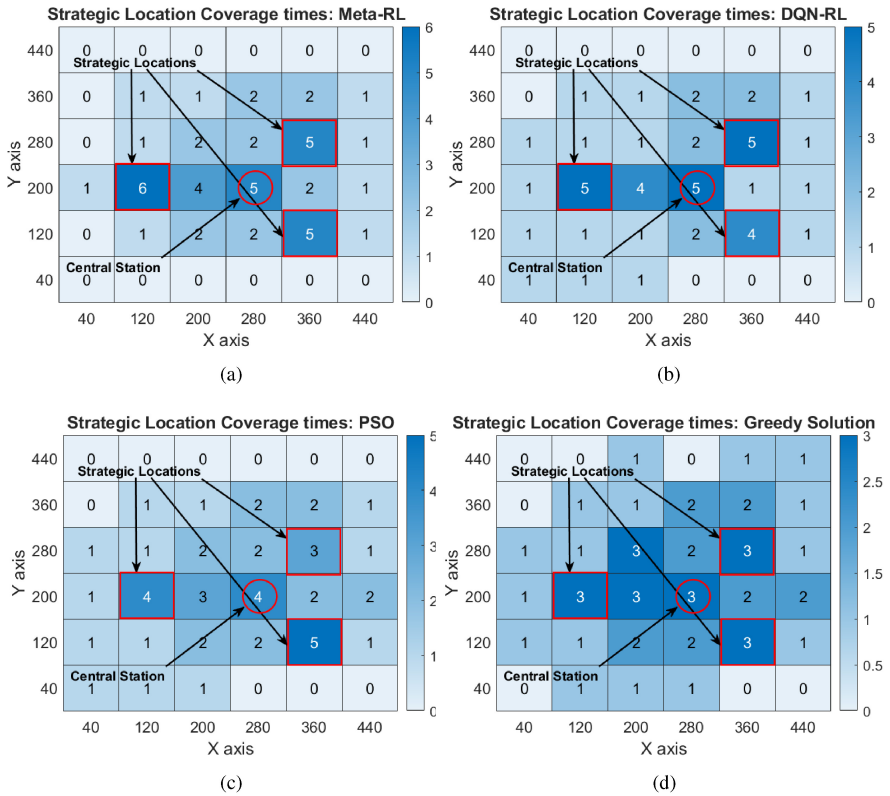


FIGURE 3. Comparing the number of visits to strategic locations with nonstrategic locations in one time frame T for different algorithms. All the algorithms pass through the strategic locations, providing better services than nonstrategic locations.

than the greedy solution. In contrast, the greedy solution only possesses information about the next step, resulting in inferior performance compared to competitive solutions.

Figure 4 compares the performance of different algorithms in training and converging to the maximum demand service for the deep learning-based solutions. Demand service satisfaction evaluates how effectively the algorithms can meet the diverse demands of strategic locations for visits and service provision. As depicted in the figure, the meta-RL algorithm demonstrates superior convergence compared to the conventional RL algorithm. This outcome assesses the agent's ability to learn the optimal policy accurately, maximizing rewards. Initially, the agent explores various experiments and subsequently commences learning the optimal policy until converging to the maximum expected reward.

Figure 5 presents the outcomes of the total EH achieved by different algorithms. The figure illustrates the results of the objective function formulated in equation (23). During the exploration phase, the agent explores different actions and searches for actions that maximize the EH of IoT devices. Subsequently, they converge into the optimal policy, meaning UAVs have learned the optimal positions that maximize EH.

Figures 6 and 7 present a comparison of different algorithms in terms of EH in strategic and nonstrategic locations while varying the number of UAVs in the swarm. In both cases, as the number of UAVs increases, EH at IoT devices also increases. The maximum expected EH occurs

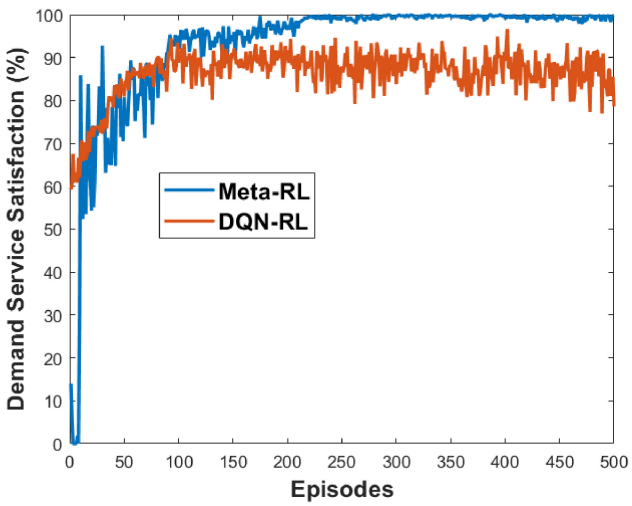


FIGURE 4. Demand service satisfaction comparison of different algorithms. The convergence of the algorithms to the maximum expected value tests the ability of the agent to learn the optimal policy of respecting the constraints. Meta-RL converges with higher accuracy than the DQN algorithm.

when all IoT devices receive sufficient WPT from UAVs, enabling them to transmit all their data to the UAVs. Meta-RL demonstrates performance closer to the maximum EH compared to competitive algorithms, aligning with the primary objective of the problem. Specifically, Meta-RL tends to harvest more energy than DQN, PSO, and the greedy solution in strategic locations and less EH than

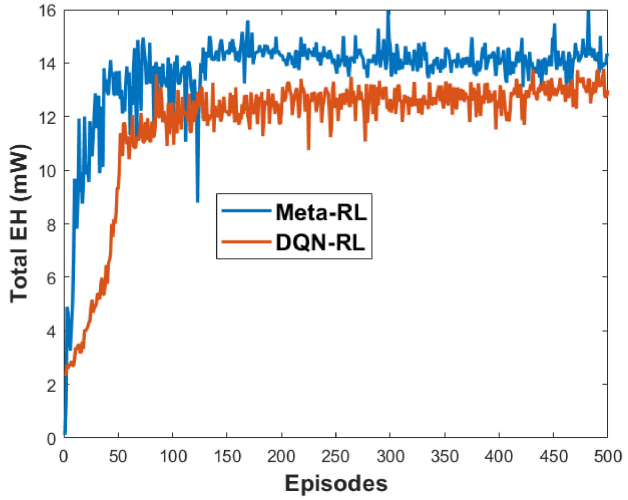


FIGURE 5. Average total EH in strategic locations of different algorithms. The agent spends a few episodes exploring different actions and then learn the optimal policy, which is the maximum EH of IoT devices, which is the main objective of the optimization in Equation (23).

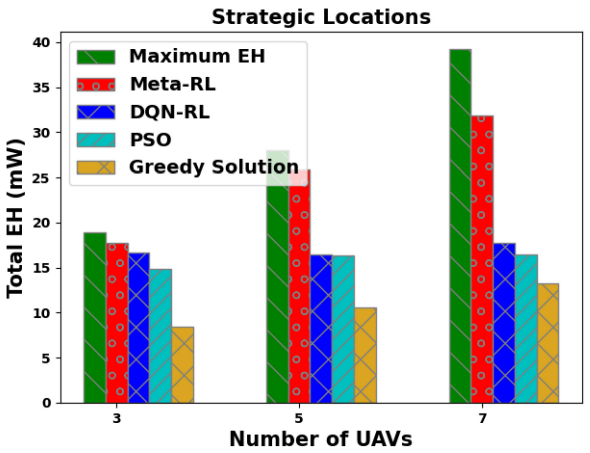


FIGURE 6. Average total EH comparison of different algorithms in strategic locations in terms of changing the number of UAVs. The meta-RL algorithm provides better results regarding the IoT EH which is the highest among the competitive algorithms, and closest to maximum EH. Also, more IoT EH is achieved when the number of UAVs increases.

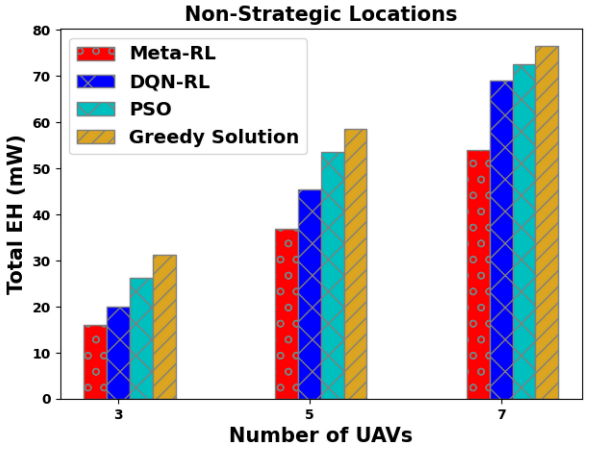


FIGURE 7. Average total EH comparison of different algorithms in nonstrategic locations in terms of changing the number of UAVs. The meta-RL algorithm provides better results regarding the IoT EH, the lowest among the other algorithms. Also, more IoT EH is achieved when the number of UAVs increases.

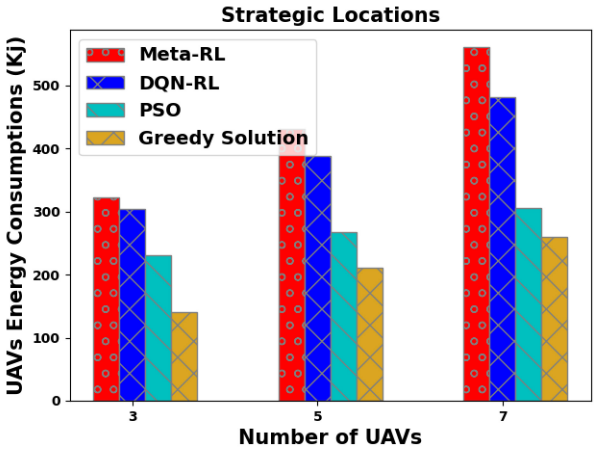


FIGURE 8. UAVs' energy consumption comparison of different algorithms in strategic locations in terms of changing the number of UAVs. The meta-RL algorithm provides better results regarding the UAVs' energy consumption, which is the highest among the other algorithms, i.e., it is successfully exploiting their energy in strategic locations. Also, more UAVs energy consumption is consumed when the number of UAVs increases.

in nonstrategic locations. The DQN algorithm provides comparable EH results to the PSO algorithm, outperforming the greedy solution.

Similarly, figures 8 and 9 compare the meta-RL solution, traditional RL algorithm, PSO, and greedy solution in terms of UAVs' energy consumption in strategic and nonstrategic locations as the number of UAVs varies in the swarm. As depicted in the figures, meta-RL tends to expend more energy than others in strategic locations and demonstrates lower energy consumption in nonstrategic locations. This behavior aligns with its priority to serve strategic locations, which is the main objective of the problem.

Figure 10 illustrates the average total rewards obtained by the deep learning-based algorithms. The average total rewards indicate the convergence of the agent to learn the optimal policy, achieved when all the constraints outlined in equation (27) are satisfied. The agent explores different states initially, progressively learning the optimal policy that yields higher rewards. Our adopted meta-RL algorithm demonstrates comparable convergence to the conventional RL algorithm.

Figure 11 compares different algorithms in terms of changing the UAV's transmit power. As the UAV's transmit power increases, the EH of IoT devices also increases proportionally. Higher UAV transmit power enhances the resources available, leading to increased EH at IoT devices. The meta-RL algorithm achieves results closer to the maximum EH and outperforms all competitive solutions. The conventional RL solution also outperforms the PSO algorithm and the greedy solution, as the agent learns the optimal policy of UAV trajectory paths that maximize EH. The key distinction lies in the fact that the RL agent and the PSO participant have knowledge about all paths that

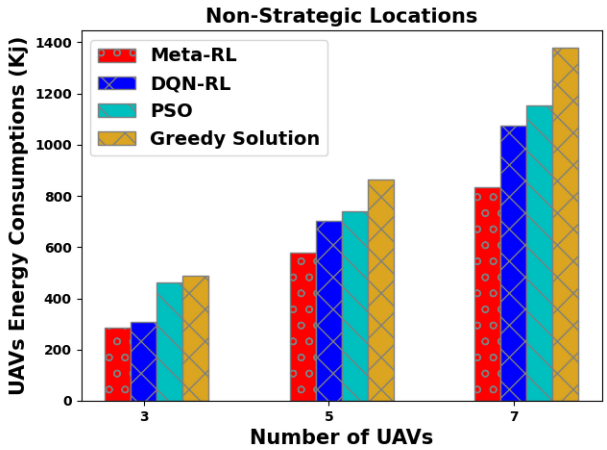


FIGURE 9. UAVs' energy consumption comparison of different algorithms in nonstrategic locations in terms of changing the number of UAVs. The meta-RL algorithm provides better results regarding the UAVs' energy consumption, which is the lowest among the other algorithms, i.e., it is successfully exploiting their energy in strategic locations. Also, more UAVs energy consumption is consumed when the number of UAVs increases.

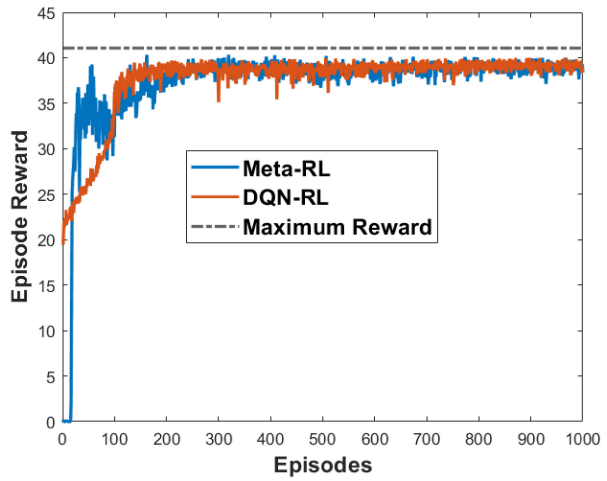


FIGURE 10. Episode reward comparison of different algorithms. The agent starts exploring different actions until learning the optimal policy that increases the reward designed in Equation (27).

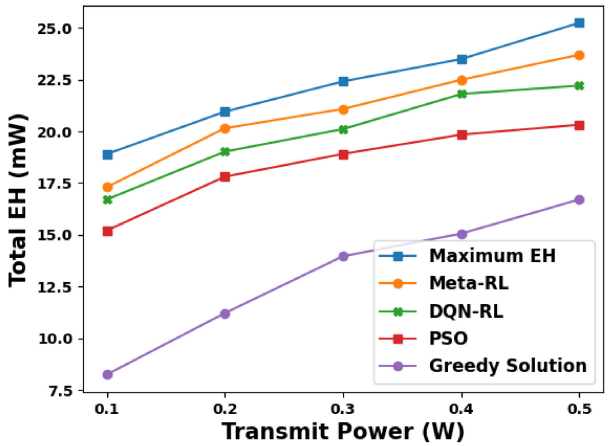


FIGURE 11. Average total EH comparison of different algorithms in terms of changing the UAV's transmit power. As the UAV's transmit power increases, the IoT EH increases simultaneously. Meta-RL solution achieves the highest EH of IoT among the competitive algorithms, and closest solution to the maximum EH of IoT devices.

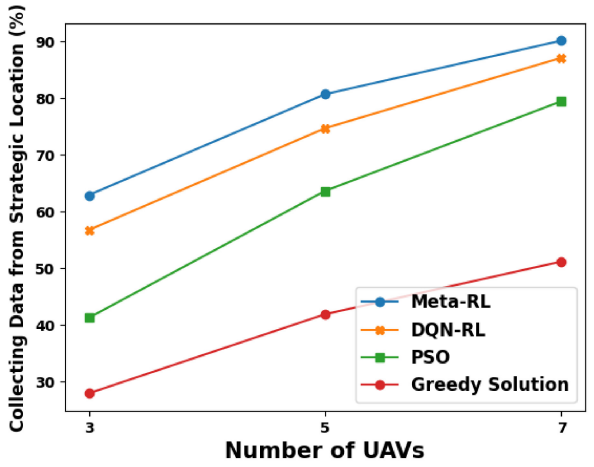


FIGURE 12. Comparing different algorithms based on the percentage of successfully collecting data from IoT devices as the number of UAVs increases in the swarm. The plot illustrates how well each algorithm performs in terms of data collection efficiency as the number of UAVs varies in the swarm.

maximize EH, while the greedy solution only has knowledge about the next step.

Figure 12 compares various algorithms regarding their success in collecting all data from IoT devices with varying numbers of UAVs in the swarm. As the number of UAVs increases, more data can be collected from IoT devices. Meta-RL excels in data collection, surpassing traditional RL, PSO, and the greedy solution. Specifically, with 7 UAVs, Meta-RL achieves approximately 90% data collection, outperforming DQN, PSO, and the greedy solution, which achieve 85%, 78%, and 50%, respectively. Figure 13 compares different algorithms in terms of the total sum rate achieved by each algorithm as the channel bandwidth increases. The plot illustrates how the total sum rate changes with varying channel bandwidths. As the channel bandwidth increases, the total sum rate also increases, indicating a stronger channel capable of collecting more data. For instance, with a 3 MHz channel bandwidth. The maximum threshold sum-data rate for 3 UAVs is 269.95 Mbpd. Moreover, the average sum rate for Meta-RL and DQN is 246.39 Mbps and 243.04 Mbps, compared to 164.89 Mbps and 111.94 Mbps for the PSO and greedy solution. Meta-RL achieves comparable results to DQN across all bandwidth scenarios. PSO consistently outperforms the greedy solution, as UAVs in the greedy solution only have knowledge about the next step of maximum total sum rate compared to deep learning-based solutions and PSO algorithms.

Figure 14 compares different algorithms in terms of total EH and total sum rate as the number of cell sizes increases. The plot illustrates how these metrics change with varying cell numbers in the grid. As the number of cells increases, both total EH and total sum rate decrease, indicating the need for more UAVs to enhance EH and data transmission

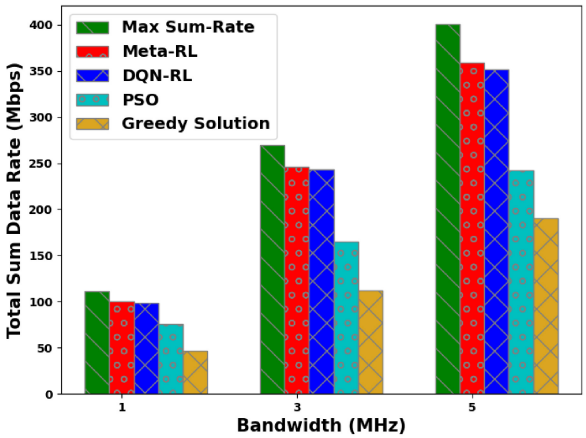


FIGURE 13. Comparing various algorithms based on the total sum rate of transmitting data to the UAVs as the channel bandwidth increases. The plot provides insights into the performance of each algorithm in terms of the total data rate during transmission.

rates. Meta-RL consistently outperforms DQN across all cell number scenarios in both total EH and data rates. DQN algorithm, in turn, surpasses PSO and the greedy solution. Furthermore, PSO outperforms the greedy solution in all scenarios. For instance, with a grid size of 7×7, the maximum threshold of EH and sum-rate for 3 UAVs are 14.83 mJ, 78.52 Mbps, respectively. Meta-RL achieves 12.79 mJ and 74.29 Mbps for total EH and total sum rate, respectively. In the same scenario, DQN achieves 11.12 mJ and 64.80 Mbps, PSO achieves 9.75 mJ and 50.84 Mbps, and the greedy solution achieves 6.42 mJ and 38.01 Mbps for total EH and total sum-rate, respectively.

Figure 15 demonstrates the results of the total EH of the different algorithms when changing the number of IoT devices in the grid. When the number of IoT devices increases, the total EH increases, pointing out that UAVs can deliver more wireless power to the increased IoT devices in the grid. In particular, the total EH improves from 7.85 mW to 19.05 mW when the number of IoT devices increases from 200 to 500 IoT devices in the grid for the maximum threshold when using 3 UAVs. Meta-RL outperforms the competitive algorithms and achieves closer results to the maximum threshold. More specifically, meta-RL outperforms traditional RL represented by DQN, where meta-RL achieved 17.59 mW compared to 16.05 mW for the DQN algorithm, 14.65 mW, and 12.32 mW for PSO and greedy solution when using 500 IoT devices. The DQN algorithm achieves better results than PSO and is a greedy solution. Moreover, the PSO algorithm achieves better results than the greedy solution.

We explore a practical scenario where the number of UAVs dynamically joins and disconnects from the swarm during the learning process. Figure 16 illustrates the corresponding implementation of average reward convergence. The simulation begins with a swarm of three UAVs. Notably, meta-RL exhibits rapid convergence to the maximum expected reward, achieving convergence in 900 episodes. Subsequently, two UAVs join the swarm, and meta-RL adeptly adapts to these changes in the environment, quickly converging to the maximum expected reward once again. Later in the learning process, three UAVs depart the swarm, possibly for recharging, and meta-RL demonstrates its adaptability by converging efficiently to the maximum expected reward. This capability showcases the resilience and flexibility of meta-RL in handling dynamic changes within the learning environment.

C. DISCUSSION AND LESSONS LEARNED

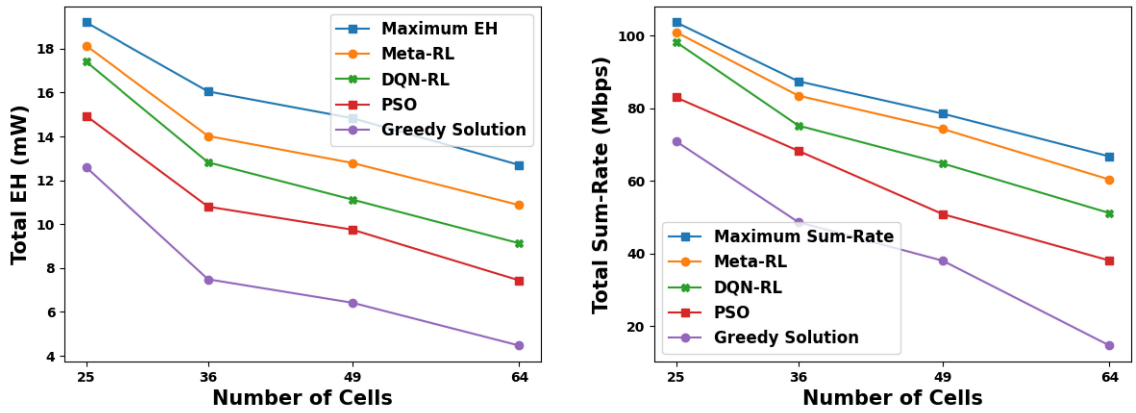
In this section, we delve into an insightful discussion and interpretation of the findings detailed in Section VI-B. The problem of using a swarm of UAVs to cover an area focusing on providing better services to strategic locations is investigated in this article. The strategic locations include an area affected by natural disasters like earthquakes, hurricanes, and floods in which the conventional communication system collapses, and dynamic and quick solutions are required to help people affected by the disaster. The services include working with a swarm of UAVs to provide WPT and WIT. We formulated the problem as an optimization problem that seeks to maximize IoT devices' EH in strategic locations. The formulated optimization problem is an NLP problem, which is challenging and time-consuming to solve using conventional optimization techniques. An online and real-time solution based on meta-RL is proposed using deep learning techniques. The adopted solution is compared with competitive solutions using conventional RL and also a greedy algorithm to enrich the results.

The results in Figures 3 show that the algorithms navigate the area, focusing on strategic locations that require better services. Our adopted meta-RL solution provides better results in which the strategic locations get more attention by around 60% compared to nonstrategic locations, then the traditional RL solution and PSO with around 50% more attention, then the worst case with the greedy solution.

Figure 4 shows the results of satisfying the demand services of strategic locations and makes a comparison with different algorithms. Each strategic location has a different demand service. When a UAV passes through one strategic location, it satisfies part of the demand service. It is also clear that meta-RL provides better demand satisfaction compared to conventional RL algorithms of the DQN algorithm.

The results in Figures 5 indicate the main objective function formulated in equation (23). In Figure 5, the IoT devices EH in strategic locations started low, and then the agent learned the optimal policy of maximizing it by choosing the best trajectory paths of UAVs.

The results in Figures 6, 7, 8, and 9 compare our adopted meta-RL algorithm with the competitive algorithms regarding IoT energy consumption and the total UAV energy consumption when the number of UAVs changes in the swarm in strategic and nonstrategic locations. In strategic locations, our adopted meta-RL algorithm provides IoT devices EH and UAVs energy consumptions higher than



(a) (b)

FIGURE 14. Comparing various algorithms based on the total EH to the IoT devices and the total sum-rate of the communication channel as the number of cells increases. The plot provides insights into the performance of each algorithm in terms of the total EH and total data rate during transmission.

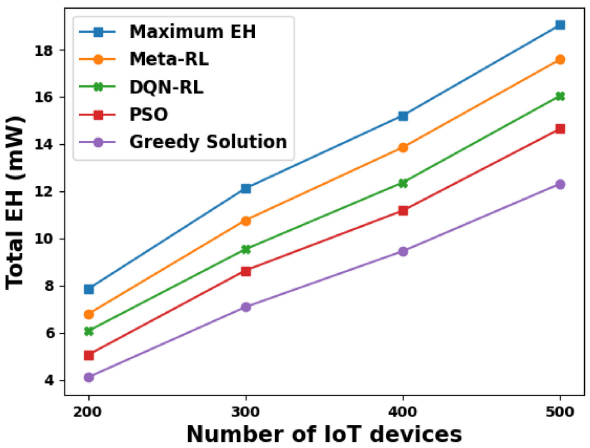


FIGURE 15. Comparing various algorithms based on the total EH as the number of IoT devices increases in the grid. The plot provides insights into the performance of each algorithm in terms of the total EH.

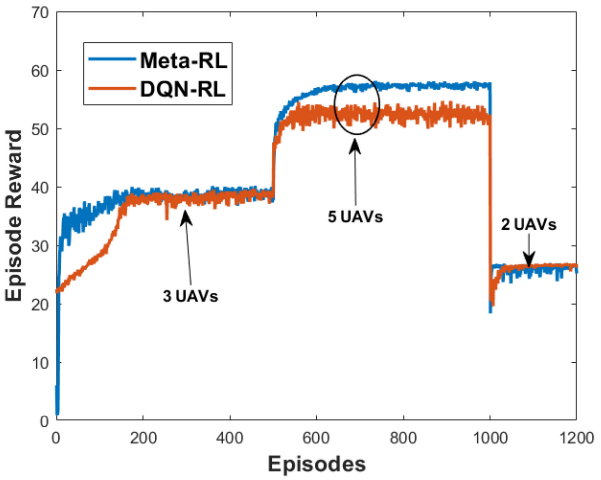


FIGURE 16. The adaptivity of Meta-RL and DQN algorithm to the environment changes of the learning. The algorithm started with three UAVs, then two more UAV joined the swarm, and after that, three UAVs left the swarm. Meta-RL algorithm learns the optimal policy quickly and converges to its maximum expected reward compared to DQN algorithm.

the other algorithms as the main focus is to harvest more energy and exploit the UAVs energy in strategic locations, which achieves higher IoT devices EH by around 18% than conventional RL algorithms and 32%. In contrast, meta-RL achieves less IoT devices and UAV energy consumption in nonstrategic locations as the main focus is to provide better services to strategic locations.

The result in Figure 10 indicates the ability of the agent to learn the optimal policy regarding respecting the constraints formulated in equation (27). Initially, the agent explores different actions and then learns the optimal policy to maximize the EH of IoT in strategic locations. It is also clear that our adopted meta-RL algorithm achieves better accuracy than the conventional RL algorithms with faster convergence. Figure 11 shows the effect of increasing the UAV's transmit power. When the UAV transmit power increases, the EH of the IoT also increases simultaneously. Moreover, our adopted meta-RL algorithm archives better IoT EH than the other algorithms.

Figure 12 provides a comparison of algorithms regarding their capability to collect data from IoT devices as the number of UAVs increases. Our adopted Meta-RL solution provides better results among competitive algorithms. Additionally, in Figure 13, we observe a comparison of the maximum threshold of the total sum-rate across all algorithms. Notably, Meta-RL achieves results closest to the maximum threshold of the total sum-rate, further highlighting its effectiveness in optimizing the communication channel's data transmission rate.

In Figure 14, the comparison unfolds as the number of cells in the covered area increases, impacting both the total EH and total sum-rate. Notably, as cell sizes increase, there is a simultaneous decrease in both total EH and total sum-rate. In Figure 16, the adaptability of the adopted Meta-RL solution to environmental changes is demonstrated. The solution exhibits dynamic convergence after abrupt changes

in the environment, showcasing its ability to adapt and stabilize efficiently.

In summary, the lessons concluded from the experiments, which answer the research questions presented at the beginning of this article, are summarized as follows:

- Our adopted meta-RL algorithm successfully satisfies the demand service of strategic locations better than the other algorithms. Furthermore, the meta-RL algorithm provides better results with higher accuracy compared to the other algorithms when satisfying the demand services of strategic locations. The convergence of the algorithms to the maximum expected value tests the ability of the agent to learn the optimal policy of respecting the constraints. Meta-RL converges with higher accuracy than DQN algorithms.
- The higher the number of UAVs, the higher the percentage of satisfaction of demand services in strategic locations. The meta-RL algorithm provides better results regarding the IoT EH, which is the highest among the other algorithms. Also, more IoT EH is achieved when the number of UAVs increases.
- Our adopted meta-RL algorithm achieves better results than the other algorithms regarding the number of visits to strategic locations, harvesting more energy, respecting the designed constraints, and exploiting the UAVs' energy consumption in strategic locations. The agent spends a few episodes exploring different actions and then learns the optimal policy, which is the maximum EH of IoT devices, which is the main objective of the optimization in equation (23).
- Increasing the UAV's transmit power leads to higher IoT device's EH as the UAVs become more robust and can increase the communication quality to deliver higher WPT, hence increasing the WIT.
- As the channel bandwidth increases, the total sum-rate of all algorithms increases, pointing that more data can be uploaded for the IoT devices to the UAVs. Meta-RL achieves a comparable result to the DQN algorithm and is better than PSO and the greedy algorithm.
- Increasing the number of cells leads to decrease the total EH and total sum-rate of data transmission.

VII. CONCLUSION AND FUTURE WORK

In this article, we investigated the problem of maximization of EH of IoT devices while respecting significant constraints, including the maximum UAV energy consumption, the maximum time duration, and the minimum data rate for successful data transmission. The UAVs navigate the area, transferring wireless power and collecting data from the distributed IoT devices on the ground. We formulated the problem as NLP, and due to the high complexity of the problem, we adopted a deep learning solution using meta-RL as a real-time-based solution. We compared the adopted solution with two state-of-the-art algorithms using RL and PSO and one greedy solution. We demonstrated that our solution outperforms the competitive solutions in terms of coverage and satisfying the demand requirements of the strategic locations.

In future work, we intend to enhance the resilience and efficiency of communication networks by exploring the impact of interference and the Doppler effect on communication links and UAVs' movements. Additionally, we plan to incorporate crucial parameters that can significantly affect the model's performance efficiency, such as mobile IoT devices and the integration of renewable energy sources like solar power for EH. This expansion will involve addressing more challenging parameters within the optimization problem.

REFERENCES

[1] L. Xie, X. Cao, J. Xu, and R. Zhang, "UAV-enabled wireless power transfer: A tutorial overview," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 2042–2064, Dec. 2021.

[2] J. Shi, P. Cong, L. Zhao, X. Wang, S. Wan, and M. Guizani, "A two-stage strategy for UAV-enabled wireless power transfer in unknown environments," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1785–1802, Feb. 2024.

[3] Y. Yu, J. Tang, J. Huang, X. Zhang, D. K. C. So, and K.-K. Wong, "Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6361–6374, Sep. 2021.

[4] X. Yuan, T. Yang, Y. Hu, J. Xu, and A. Schmeink, "Trajectory design for UAV-enabled multiuser wireless power transfer with nonlinear energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1105–1121, Feb. 2021.

[5] O. L. López, H. Alves, R. D. Souza, S. Montejo-Sánchez, E. M. G. Fernández, and M. Latva-Aho, "Massive wireless energy transfer: Enabling sustainable IoT toward 6G era," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 8816–8835, Jun. 2021.

[6] C. Sha, Y. Sun, and R. Malekian, "Research on cost-balanced mobile energy replenishment strategy for wireless rechargeable sensor networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3135–3150, Mar. 2020.

[7] Z. Wang, L. Duan, and R. Zhang, "Adaptively directional wireless power transfer for large-scale sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1785–1800, May 2016.

[8] A. M. Seid, J. Lu, H. N. Abishu, and T. A. Ayall, "Blockchain-enabled task offloading with energy harvesting in multi-UAV-assisted IoT networks: A multi-agent DRL approach," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 12, pp. 3517–3532, Dec. 2022.

[9] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3177–3192, Oct. 2021.

[10] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[11] X. Yuan, H. Jiang, Y. Hu, and A. Schmeink, "Joint analog beamforming and trajectory planning for energy-efficient UAV-enabled nonlinear wireless power transfer," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 10, pp. 2914–2929, Oct. 2022.

[12] M. Li, H. Li, P. Ma, and H. Wang, "Energy maximization for ground nodes in UAV-enabled wireless power transfer systems," *IEEE Internet Things J.*, vol. 10, no. 19, pp. 17096–17109, Oct. 2023.

[13] D. Van Huynh, T. Do-Duy, L. D. Nguyen, M.-T. Le, N.-S. Vo, and T. Q. Duong, "Real-time optimized path planning and energy consumption for data collection in unmanned ariel vehicles-aided intelligent wireless sensing," *IEEE Trans. Ind. Informat.*, vol. 18, no. 4, pp. 2753–2761, Apr. 2022.

[14] W. Luo, Y. Shen, B. Yang, S. Wang, and X. Guan, "Joint 3-D trajectory and resource optimization in multi-UAV-enabled IoT networks with wireless power transfer," *IEEE Internet Things J.*, vol. 8, no. 10, pp. 7833–7848, May 2021.

[15] X. Wang, J. Li, Z. Ning, Q. Song, L. Guo, and A. Jamalipour, "Wireless powered metaverse: Joint task scheduling and trajectory design for multi-devices and multi-UAVs," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 3, pp. 552–569, Mar. 2024.

[16] H. Yan, Y. Chen, and S.-H. Yang, "Time allocation and optimization in UAV-enabled wireless powered communication networks," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 2, pp. 951–964, Jun. 2022.

[17] S. Muy and J.-R. Lee, "Joint optimization of trajectory, beamforming, and power allocation in UAV-enabled WPT networks using DRL combined with water-filling algorithm," *Veh. Commun.*, vol. 43, Oct. 2023, Art. no. 100632.

[18] S. Xu, X. Zhang, C. Li, D. Wang, and L. Yang, "Deep reinforcement learning approach for joint trajectory design in multi-UAV IoT networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3389–3394, Mar. 2022.

[19] J. Moon, S. Papaioannou, C. Laoudias, P. Kolios, and S. Kim, "Deep reinforcement learning multi-UAV trajectory control for target tracking," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15441–15455, Oct. 2021.

[20] M. B. Bezziane, B. Brik, A. Messiaid, M. R. Kafi, A. Korichi, and A. B. Bezziane, "Impact of noise on data routing in flying Ad Hoc networks," *Opt. Quantum Electron.*, vol. 56, no. 4, p. 563, 2024.

[21] Z. Huang, C. Chen, and M. Pan, "Multiobjective UAV path planning for emergency information collection and transmission," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6993–7009, Aug. 2020.

[22] M. B. Ghorbel, D. Rodríguez-Duarte, H. Ghazzai, M. J. Hossain, and H. Menouar, "Joint position and travel path optimization for energy efficient wireless data gathering using unmanned aerial vehicles," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2165–2175, Mar. 2019.

[23] S. Wan, J. Lu, P. Fan, and K. B. Letaief, "Toward big data processing in IoT: Path planning and resource management of UAV base stations in mobile-edge computing system," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5995–6009, Jul. 2020.

[24] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 498–501, Jul. 2020.

[25] H. Ren, C. Pan, K. Wang, Y. Deng, M. Elkashlan, and A. Nallanathan, "Achievable data rate for URLLC-enabled UAV systems with 3-D channel model," *IEEE Wireless Commun. Lett.*, vol. 8, no. 6, pp. 1587–1590, Dec. 2019.

[26] C. She, C. Liu, T. Q. Quek, C. Yang, and Y. Li, "UAV-assisted uplink transmission for ultra-reliable and low-latency communications," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2018, pp. 1–6.

[27] M. Dhuheir, A. Erbad, and S. Sabeeh, "LLHR: Low latency and high reliability CNN distributed inference for resource-constrained UAV swarms," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2023, pp. 1–6.

[28] Y. Zeng, X. Xu, and R. Zhang, "Trajectory design for completion time minimization in UAV-enabled multicasting," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2233–2246, Apr. 2018.

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

[30] O. Ghdiri, W. Jaafar, S. Alfattani, J. B. Abderrazak, and H. Yanikomeroglu, "Offline and online UAV-enabled data collection in time-constrained IoT networks," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 4, pp. 1918–1933, Dec. 2021.

[31] E. Boshkovska, D. W. K. Ng, N. Zlatanov, and R. Schober, "Practical non-linear energy harvesting model and resource allocation for SWIPT systems," *IEEE Commun. Lett.*, vol. 19, no. 12, pp. 2082–2085, Dec. 2015.

[32] M. A. Dhuheir, E. Baccour, A. Erbad, S. S. Al-Obaidi, and M. Hamdi, "Deep reinforcement learning for trajectory path planning and distributed inference in resource-constrained UAV swarms," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 8185–8201, May 2023.

[33] W. Fedus et al., "Revisiting fundamentals of experience replay," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3061–3071.