Deep Reinforcement Learning for Energy-Efficient Data Dissemination Through UAV Networks

Abubakar S. Ali[1], ***Student Member, IEEE,*** Ahmed A. Al-Habob[2], ***Member, IEEE,*** Shimaa Naser[1], ***Member, IEEE,*** Lina Bariah[3], ***Senior Member, IEEE,*** Octavia A. Dobre[2], ***Fellow, IEEE,*** and Sami Muhaidat[1, 4], ***Senior Member, IEEE***

[1]KU 6G Research Center, Department of Computer and Communication Engineering, Khalifa University, Abu Dhabi, United Arab Emirates [2]Department of Electrical and Computer Engineering, Memorial University, St. John's, NL A1C 5S7, Canada
[3]Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE.
[4]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

The work of A. A. Al-Habob and O. A. Dobre has been supported in part by the Canada Research Chairs Program CRC-2022-00187.

ABSTRACT The rise of the Internet of Things (IoT), marked by unprecedented growth in connected devices, has created an insatiable demand for supplementary computational and communication resources. The integration of Unmanned aerial vehicles (UAVs) within IoT ecosystems presents a promising avenue to surmount these obstacles, offering enhanced network coverage, agile deployment capabilities, and efficient data gathering from geographically challenging locales. UAVs have been recognized as a compelling solution, offering extended coverage, flexibility, and reachability for IoT networks. Despite these benefits, UAV technology faces significant challenges, including limited energy resources, the necessity for adaptive responses to dynamic environments, and the imperative for autonomous operation to fulfill the evolving demands of IoT networks. In light of this, we introduce an innovative UAV-assisted data dissemination framework that aims to minimize the total energy expenditure, considering both the UAV and all spatially-distributed IoT devices. Our framework addresses three interconnected subproblems: device classification, device association, and path planning. For device classification, we employ two distinct types of deep reinforcement learning (DRL) agents—Double Deep Q-Network (DDQN) and Proximal Policy Optimization (PPO)—to classify devices into two tiers. To tackle device association, we propose an approach based on the nearest-neighbor heuristic to associate Tier 2 devices with a Tier 1 device. For path planning, we propose an approach that utilizes the Lin-Kernighan heuristic to plan the UAV's path among the Tier 1 devices. We compare our method with three baseline approaches and demonstrate through simulation results that our approach significantly reduces energy consumption and offers a near-optimal solution in a fraction of the time required by brute force methods and ant colony heuristics. Consequently, our framework presents an efficient and practical alternative for energy-efficient data dissemination in UAV-assisted IoT networks.

INDEX TERMS Data dissemination, Deep learning, Internet-of-things (IoT), Reinforcement Learning (RL), Unmanned Aerial Vehicle (UAV).

I. Introduction

THe exponential growth of connected devices, due to the emergence of Internet-of-Things (IoT), and the growing number of deployed sensors in smart cities call for enabling new traffic services to enhance user's experience [1]. This exponential growth of connected smart devices poses new challenges in terms of network capacity and massive connectivity [2]. Therefore, the task of data dissemination/aggregation has become much more challenging, motivating the need for revolutionary solutions that can reduce the dependency on the network infrastructure [3]. Recently, research efforts on unmanned aerial vehicles

(UAVs), also known as drone-based communication systems, have been growing in industry and academia alike, targeting strict requirements, particularly ultra-low latency and unprecedented communication reliability. UAVs constitute the basic building block of aerial networks, whose inherent features, such as flexibility and mobility, enable several new disruptive applications [4]. The recent advancements in UAVs have made considerable reduction in their production cost, making them affordable for numerous public and civil applications, such as border surveillance, traffic monitoring, and data dissemination/aggregation, to name a few [5].

On the other hand, reinforcement learning (RL) is a machine learning paradigm, where an intelligent agent learns from the interaction with an environment [6]. In particular, the agent learns to map states to actions to maximize numerical reward. The field of RL has become one of the most active research areas in machine learning, neural networks, and artificial intelligence. In particular, researchers have adopted RL algorithms to solve complex optimization problems that are difficult to tackle.

A. Motivation

An IoT network comprises a massive number of connected devices, which create a huge amount of data traffic that results in network congestion. On the other hand, energy consumption is a major concern in the context of UAV-assisted data dissemination, in which UAVs are used to disseminate data to multiple IoT devices. As a result, an energy-efficient solution that reduces the UAV's overall travel distance and communication energy, as well as the devices' energy consumption, should be proposed. This work tries to answer the question of whether a method exists that can perform joint classification, association, and path planning. Conventional methods, such as genetic algorithms, ant colony, simulated annealing, as well as other optimization methods, were investigated in [7]–[11]. In these methods, the "lifetime" behavior of many non-learning agents is evaluated, each employing a different policy when interacting with their environments, and eventually, those with the most rewards are selected. These methods can be effective if the space of policies is sufficiently small, there is ample time for searching, and the environment can be structured so that good policies are easy to find. Additionally, conventional methods are more attractive when the state of the environment cannot be accurately sensed by the learning agent [12].

In this paper, we consider the application of deep reinforcement learning (DRL) to minimize the energy expenditure of the underlying system model. To the best of the authors' knowledge, no report of such a framework has recently appeared in open literature. The motivation behind employing DRL for device classification in our UAV-assisted data dissemination framework hinges on several key advantages over traditional clustering or deep learning methods. Firstly, DRL is inherently suited for dynamic environments where interaction with the environment is crucial for learning optimal strategies. Unlike static clustering algorithms or supervised deep learning approaches, DRL agents learn by continuously interacting with the environment, making them ideal for scenarios where the state of the system can change over time. Secondly, DRL can handle complex, sequential decision-making processes more effectively. The device classification problem in UAV networks involves making a series of decisions that are dependent on each other, a scenario where DRL's ability to consider long-term outcomes can be particularly beneficial. Lastly, DRL's model-free nature allows it to learn optimal policies directly from high-dimensional sensory input, eliminating the need for manual feature selection or engineering that traditional methods might require. Therefore, DRL's adaptability, decision-making capabilities, and direct learning from data make it a compelling choice for the device classification problem in dynamic UAV-assisted IoT networks.

The main contributions of this work are summarized as follows:

- Formulate a UAV-assisted data dissemination problem as a Markov decision problem (MDP) with the objective of minimizing the UAV's energy consumption and the energy consumed by the IoT devices.
- We proposed a DRL-based solution, where an agent interacts with an environment to solve the classification sub-problem of the joint optimization problem.
- Two variants of association algorithm based on the nearest neighbor heuristic, in order to obtain a near optimum devices association, are proposed.
- Developing an efficient algorithm for solving the UAV path planning sub-problem, based on the Lin-Kernighan heuristic (LKH) algorithm, to obtain an optimum or near optimum UAV tour.
- Analyzing the computational complexity of the baseline, the brute force, and the proposed DRL approaches.

B. Organization of the Paper

The structure of this paper is as follows: Section II presents the background and related works. The system model is described in Section III. Section IV introduces the baseline approaches. Our proposed DRL-based UAV-assisted data dissemination methodology is detailed in Section B. Section VI outlines the performance evaluation metrics utilized. Results and their ensuing discussions are showcased in Section VII, and finally, conclusions are drawn in Section VIII.

II. Background and Related Works

This study is anchored in a specialized application of UAV networks, situated amidst a rapidly advancing landscape where DRL significantly enhances UAV network functionalities. The advent of DRL has sparked notable advancements in areas such as autonomous UAV navigation, dynamic resource allocation, and real-time decision-making processes, each contributing to the development of UAV systems that are not only more efficient but also highly adaptable and intelligent [13]. The application of DRL within UAV-assisted IoT networks, in particular, plays a crucial role in overcoming intricate challenges related to energy management, optimizing flight paths, and devising effective device-to-device communication strategies [10]. Our contribution to this evolving domain leverages DRL to refine data dissemination processes in UAV-assisted IoT networks, specifically targeting the enhancement of device classification, association, and path planning mechanisms. This focus on applying advanced AI techniques to improve UAV network operations in support of IoT ecosystems marks a significant step forward in making these networks more functional and efficient. The subject of data dissemination, which stands as a core component of contemporary communication networks, has received considerable attention in recent studies, including [7]–[11], [13]. Among these, [7] and [8] specifically explore the optimization of UAV flight states in three dimensions. These research efforts employ methods such as alternating descent techniques and concave-convex procedures, aiming to maximize the efficiency of data dissemination by optimizing energy consumption and ensuring equitable data distribution across devices.

Further advancements in UAV-assisted strategies are highlighted in [9], where a cognitive UAV approach is proposed to enhance data dissemination in IoT devices. The primary goal here is to maximize the minimum received bits by the devices, tackled through a mixed-integer non-linear program. Meanwhile, [10] investigates UAV trajectory optimization to maximize the sum rate of edge users. This complex problem is approached through an iterative algorithm that addresses the mixed-integer non-convex challenges. In contrast, [11] proposes a two-tiered framework focusing on minimizing total energy consumption in UAV-assisted tasks, employing ant colony optimization for solving the joint optimization challenges.

The authors in [13] present another perspective by employing UAVs for data dissemination with the aim of maximizing throughput while minimizing transmission delay. The study introduces the recursive least squares algorithm for efficient and accurate vehicle mobility prediction. This theme of efficiency and optimization is further explored through the lens of RL in subsequent research.

In the context of RL, [14] marks a significant contribution with its focus on UAV-mounted mobile edge computing. The study formulates a Markov Decision Process (MDP) to enable dynamic UAV trajectory planning, considering mobile terminal user locations. A similar approach is seen in [15], where an MDP in a UAV-assisted communication scenario is explored. The use of Deep Reinforcement Learning (DRL) in these studies aims to optimize UAV strategies for efficient data collection, delivery, and energy management.

The integration of DRL continues to reshape the UAV-assisted communication landscape, as evidenced in [16]. This study introduces a Deep Q-Network (DQN) based flight resource allocation scheme to optimize flight cruise and data collection schedules. Such advancements are pivotal in addressing the complex demands of modern communication networks.

Finally, the studies in [17] and [18] demonstrate the versatility of DRL in solving intricate problems like joint beamforming, power control, and interference coordination. Similarly, [12] applies DRL to manage resources in maritime networks within UAV-assisted edge computing environments. [19] innovates further by combining Lyapunov stochastic optimization with DRL, focusing on energy-efficient and low-latency solutions at the wireless edge. The culmination of these efforts is seen in [20] and [21], where DRL schemes are used to optimize various aspects of networking, underlining the critical role of UAVs and DRL in the evolution of wireless communication systems.

Our work introduces a pioneering approach to UAV-assisted data dissemination in IoT environments by employing a DRL framework. This marks a significant departure from traditional optimization methods such as genetic algorithms, ant colony optimization, and simulated annealing, which have been predominantly explored in existing studies. The novelty of our approach lies in its dynamic adaptability and ability to learn optimal strategies through direct interaction with the environment, a feature not inherently available in the deterministic or heuristic-based methods previously applied. By formulating the data dissemination problem as an MDP, our DRL framework is capable of continuously learning and improving, ensuring optimal decisions under varying network conditions. This not only enhances the efficiency of the UAV's path planning and data dissemination but also significantly reduces the system's overall energy consumption. Unlike conventional methods, our DRL approach offers a scalable and flexible solution capable of adapting to the ever-evolving landscape of IoT networks, thereby setting a new benchmark in UAV-assisted communication strategies.

III. System Model and Problem Formulation

In this section, we describe the system, communication, and power models of the considered UAV-assisted data dissemination problem.

A. System Model Description

In this work, we consider N IoT devices, a single UAV, and a library of F files. Further, we assume that the UAV hovers over a group of IoT devices to disseminate data. The IoT devices are classified into m Tier 1 devices, $\tau_{\mathbf{1}}$, which receive data directly from the UAV and L Tier 2 devices, $\tau_{\mathbf{2}}$, that receive data from devices in $\tau_{\mathbf{1}}$. The UAV moves from a docking location and hovers over $\tau_{\mathbf{1_1}}$ device at a fixed altitude h and disseminates the required files f_i by the $\tau_{\mathbf{1_1}}$ device, and then moves to $\tau_{\mathbf{1_2}}$ device. After the last Tier 1 device, $\tau_{\mathbf{1_m}}$ the UAV returns to its docking location. Fig. 1 presents the system model. In our proposed system model, the hierarchical communication framework allows for a scalable and energy-efficient approach to data dissemination in UAV-assisted IoT networks, ensuring that data reaches a wide array of devices in diverse and challenging environments.

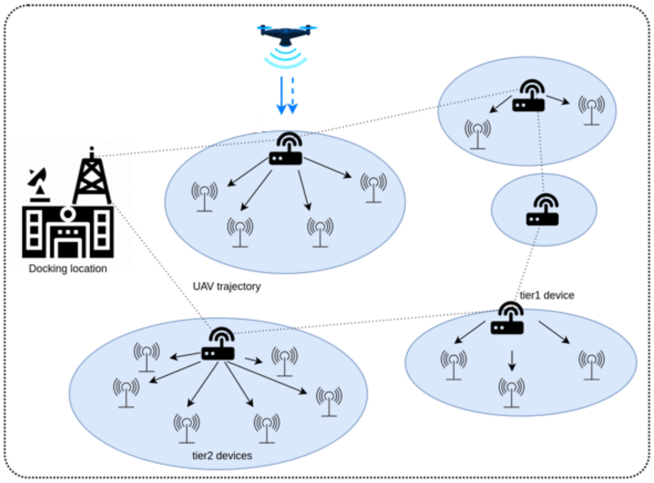

FIGURE 1. UAV-assisted data dissemination system model.

Furthermore, in our framework, we assume that each device is interested in downloading certain files from the file library. This is represented by a matrix $\mathbf{W}$ of size $N \times F$, whose elements w_{ij}, are given by

$$w_{ij} = \begin{cases} 1 & \text{if } n_i \text{ requires file } f_j \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The IoT devices are further assumed to be spatially distributed and have a fixed communication range, r_i. A matrix $\mathbf{C}$ of size $N \times N$ is introduced to represent the communication link between these devices. The device i has a communication link with device k if the distance l_{ik} between them is less than or equal to the communication range r_i. Therefore, the elements of $\mathbf{C}$ matrix are given by

$$c_{ik} = \begin{cases} 1 & \text{if } l_{ik} \leq r_i \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The distance to be traveled by a UAV should be reduced to minimize the overall energy expenditure. This can be achieved if there are fewer Tier 1 devices that the UAV has to visit and hover over. However, due to constraints in devices' communication links, and energy consumption when downloading files, there is a need to strategically classify the IoT devices. The classification is represented by a matrix $\boldsymbol{T}$ of size $N \times 2$ whose elements T_{iz} are described as

$$T_{iz} = \begin{cases} 1 & \text{if } z = 1 \text{ and } n_i \in \tau_1 \text{ or } z = 2 \text{ and } n_i \in \tau_2 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

After the optimal classification, there is also a need to optimize the trajectory which the UAV takes starting from the initial docking location, to all the Tier 1 devices and back to its docking location. This can be formulated as a travelling salesman problem (TSP), in which the matrix $\bar{\psi}$ of size $(m+2) \times 3$ depicts the tour taken by the UAV. Furthermore, there is a need to associate devices, where each Tier 2 device is associated with a Tier 1 device. This association is represented by a matrix $\boldsymbol{\mu}$ of size $N \times N$ and entries

$$\mu_{ik} = \begin{cases} 1 & \text{if } n_k \in \tau_2 \text{ and gets data from } n_i \in \tau_1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The strategic classification of IoT devices into Tier 1 and Tier 2 categories serves as a foundational element of our UAV-assisted data dissemination framework, aiming to optimize both the efficiency of data delivery and the system's overall energy expenditure. This classification is pivotal for the following reasons:

1) Efficiency in Data Dissemination: By classifying devices into tiers, our approach ensures that UAVs prioritize data dissemination to Tier 1 devices, which are then responsible for further distributing the data to Tier 2 devices. This hierarchical distribution model leverages the UAV's mobility and the IoT network's density to enhance data dissemination reach while minimizing UAV energy consumption.
2) Optimization of Energy Consumption: The classification facilitates a reduction in the UAV's travel distance, as it needs to hover only over Tier 1 devices. This directly correlates to a decrease in energy consumption, which is a critical factor in the operational efficiency of UAV-assisted networks. Additionally, by leveraging Tier 2 devices for further data dissemination, the system efficiently utilizes the existing communication links within the IoT network, further conserving energy.
3) Adaptability to Network Dynamics: Our classification strategy introduces a level of adaptability that allows the system to respond efficiently to dynamic network conditions, including changes in device density, energy availability, and data dissemination requirements. By continuously analyzing and potentially reclassifying devices based on current conditions, the UAV can adjust its path and dissemination strategy in real-time, ensuring optimal performance.

B. Problem Formulation

To obtain an energy-efficient system, the overall energy expenditure has to be minimized. Therefore, in this work, we formulate an optimization problem such that the total energy consumption is quantified by the sum of the total energy consumed by the UAV during the tour time, the UAV due to data dissemination, and the IoT devices to download their required data. Accordingly, the total expended energy by the UAV is given by [11]

$$
\begin{aligned}
E_{\text{UAV}}\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) = &\sum_{l=1}^{K+1} \left(P_{\text{hov}} + P_{\text{mov}}\right) \frac{\left\|\bar{\psi}_l - \bar{\psi}_{l+1}\right\|}{v} \\
&+ \sum_{i=1}^{N} \sum_{j=1}^{F} T_{i1} d_j^{(i)} \frac{\left(P_{\text{hov}} + P_T\right) L_j}{R_i},
\end{aligned}
\tag{5}
$$

where $\bar{\boldsymbol{\psi}}$ denotes the UAV tour matrix outlining the path the UAV follows to disseminate data, $d_j^i = 1$ if n_i or any other device with $\mu_{ik} = 1$ requires to download the file f_j, and $d_j^i = 0$ otherwise. Lj is the size of file f_j in bits, $R_i = B\log_2\left(1 + \frac{P_T}{\varphi_i N_0}\right)$ is the average data rate of the communication channel between the UAV and device n_i, with B denoting the channel bandwidth, P_T representing the UAV transmit power, φ_i as the average channel path-loss, and N_0 denoting the additive white Gaussian noise power. The norm of the distance between the current and next stopping point is represented by $\left\|\bar{\psi}_l - \bar{\psi}_{l+1}\right\|$. The power consumed by the UAV when it hovers at a fixed position is expressed as $P_{\text{hov}} = \sqrt{\frac{(Mg)^3}{2\pi r^2 p \vartheta}}$ with M denoting the UAV's mass, p representing the number of propellers, r the radius of the propeller, ϑ is the air density, and g denoting the earth gravity. Furthermore, the power consumed when the UAV moves from one position to another equals to $P_{\text{mov}} = \frac{v}{v_{\text{max}}}\left(P_{\text{max}} - P_{\text{stop}}\right) - P_{\text{stop}}$, where v and v_{max} are the traveling and maximum speed of the UAV, respectively, and P_{max} and P_{stop} are the power consumption of the UAV's hardware at full speed and idle states, respectively. Following [11], the energy expenditure of device n_i when downloading data from the UAV is given by

$$
\begin{aligned}
E_i\left(\boldsymbol{T}, \boldsymbol{\mu}\right) = T_{i1} &\left(\sum_{j=1}^{F} d_j^{(i)} \frac{P_r L_j}{R} + \sum_{k=1}^{N} \sum_{j=1}^{F} \mu_{ik} w_{kj}\right) \\
&+ T_{i2} \left(\sum_{k=1}^{N} \sum_{j=1}^{F} \mu_{ki} w_{ij} \frac{P_r L_j}{R_{ki}}\right),
\end{aligned}
\tag{6}
$$

where P_r is the power consumed by the receiver, and $R_{ki} = B\log_2\left(1 + \frac{P_{\tau_i}}{\varphi_{ki}\sigma^2}\right)$ is the average data rate for the terrestrial communication link between device n_i and device n_k with φ_{ki} representing the average channel path-loss.

Based on the above, the total energy expenditure can be written as

$$
E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) = E_{UAV}\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) + \sum_{i=1}^{N} E_i\left(\boldsymbol{T}, \boldsymbol{\mu}\right). \tag{7}
$$

Consequently, we formulate the following energy expenditure minimization problem:

$$
\begin{aligned}
\min_{\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}} \quad & E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) \\
\text{s.t.} \quad & T_{i1} + T_{i2} = 1, \quad \forall\, 1 \le i \le N, \\
& \sum_{i=1}^{N} \mu_{ik} T_{i1} = T_{k2}, \quad \forall\, 1 \le k \le N, \\
& \mu_{ik} c_{ik} = \mu_{ik}, \quad \forall\, 1 \le i, k \le N, \\
& T_{i1}, T_{i2}, \mu_{ik} \in \{0, 1\}, \quad \forall\, 1 \le i, k \le N.
\end{aligned}
\tag{8}
$$

The constraints in (8) are set to ensure that all devices are classified in Tier 1 or Tier 2. Additionally, they guarantee that each Tier 2 device has been associated with a Tier 1 device, and that a communication link exists between each device in Tier 2 and a device in Tier 1, respectively.

While our current problem formulation presented in (8) primarily focuses on minimizing energy consumption within UAV-assisted IoT networks, we acknowledge the existence of additional complexities, including network dynamics, diverse device capabilities, and environmental factors. These elements indeed play a critical role in the practical deployment and efficiency of such networks. We intend to address these additional challenges in subsequent works, aiming to incrementally refine our understanding and solutions for the multifaceted nature of UAV-assisted IoT networks.

IV. Baseline Approaches For UAV-Assisted Data Dissemination

In this section, we will present the baseline approaches based on which we compare our proposed DRL-based solutions to. These are summarized in Table 1.

A. Naive Approach

The naive baseline approach is used as a benchmark to analyze the performance of the proposed approach. In this approach, all devices are assumed to be in Tier 1. This indicates that no strategic classification is required, no association is needed, and the UAV has to follow a tour that stops and disseminates data at every IoT device in the order of their deployment.

B. Brute Force Algorithm (BFA) Approach

The optimum solution of the objective function in (8) can be solved iteratively using the brute force algorithm (BFA). To obtain the solution using BFA, we need to iterate over all possible combinations of potential classifications of devices. For each classification combination, we will iterate over all possible permutations of device association, and finally solve the path planning problem in the form of TSP iteratively. This is a very expensive method that will guarantee an optimal solution.

TABLE 1. Summary of Baseline Approaches

Approach	Description	Strengths	Weaknesses	Algorithmic Basis
Naive Approach	Assumes all devices are Tier 1, requiring no device classification or association, and the UAV disseminates data to each device directly.	Simplistic and straightforward, providing a clear benchmark for comparison.	Lacks efficiency in energy use and does not leverage device-to-device communication capabilities.	N/A
Brute Force Algorithm (BFA)	Iteratively explores all possible combinations of device classifications, associations, and UAV path planning to find the optimal solution.	Guarantees finding the optimal solution, providing a solid benchmark for optimality.	Computational complexity increases factorially with the number of devices, making it impractical for larger networks.	Detailed in the manuscript.
Ant-Colony Optimization (ACO) Approach	Inspired by the foraging behavior of ants, this heuristic algorithm finds near-optimal paths for data dissemination.	Finds near-optimal solutions efficiently for larger networks. Good balance between performance and computational demand.	May not always find the absolute best solution. Performance depends on specific parameters and heuristic rules.	The comprehensive algorithm is succinctly encapsulated in Algorithm 1.

Algorithm 1: ACO Algorithm for Data Dissemination Using a UAV

```
Input: $N, \psi_0, \psi_h, h, W, C, A, I, \alpha, \beta$
Initialize: $\tau_i, \tau_{0i}, \tau_{ik}, \tau'_{ik} \forall 1 \leq i, k \leq N$; $O \leftarrow \infty$;
for Iteration = 1 to I do
    $O_1 \leftarrow \infty$; $O_2 \leftarrow \infty$;
    for $a = 1$ to $A$ do
        Obtain $T^{(a)}, \mu^{(a)}$, and $\psi^{(a)}$ using (9), (10), and (12);
        Evaluate $O(T^{(a)}, \mu^{(a)}, \psi^{(a)})$ using (8);
        if $O > O(T^{(a)}, \mu^{(a)}, \psi^{(a)})$ then
            $T^* \leftarrow T^{(a)}$; $\mu^* \leftarrow \mu^{(a)}$; $\psi^* \leftarrow \psi^{(a)}$;
        end
        if $O_1 > O(T^{(a_1)}, \mu^{(a_1)}, \psi^{(a_1)})$ then
            $T^{(a_1)} \leftarrow T^{(a)}$; $\mu^{(a_1)} \leftarrow \mu^{(a)}$;
             $\psi^{(a_1)} \leftarrow \psi^{(a)}$;
        end
        else if $O_2 > O(T^{(a_2)}, \mu^{(a_2)}, \psi^{(a_2)})$ then
            $T^{(a_2)} \leftarrow T^{(a)}$; $\mu^{(a_2)} \leftarrow \mu^{(a)}$;
             $\psi^{(a_2)} \leftarrow \psi^{(a)}$;
        end
        Deposit pheromone for $a_1$ and $a_2$ using (14);
    end
end
return $T^*, \mu^*, \psi^*$;
```

C. Ant Colony Optimization Approach

The Ant Colony Optimization (ACO) paradigm, a bio-inspired, stochastic search algorithm, employs a collection of agent-based models, referred to as 'ants', which cooperatively navigate the solution space of an optimization problem through the exchange of pheromone-like information. Emulating the path-finding mechanisms of their biological counterparts, these agents incrementally construct solutions, informed by pheromone trails that are both strengthened by usage and diminished via evaporation. Over time, this leads to a convergence on shorter, more optimal paths. Analogous to reinforcement learning, the ACO algorithm assigns a higher weight to more favorable solutions, thus enabling a progressive improvement over iterations.

In the context of optimizing UAV network data dissemination, the authors in [11], propose an approach where a colony of ants A engages in a sequential, three-part procedure to address the optimization problem defined in (8). Each ant $a \in A$ executes a tour comprising device classification, device association, and UAV trip determination to derive a solution iteratively.

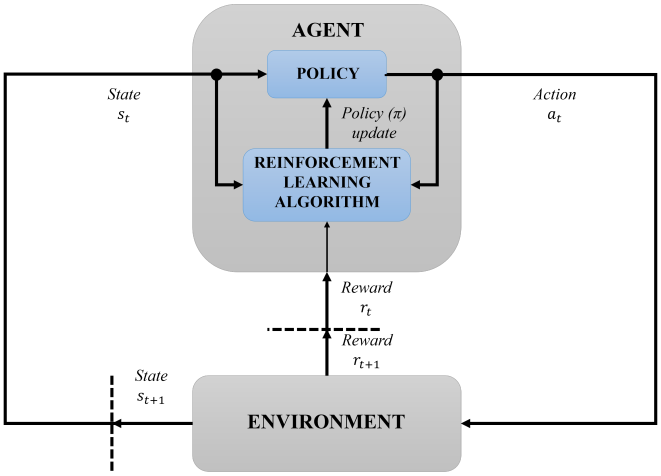

FIGURE 2. Elements of RL.

V. DRL-Based UAV-assisted Data Dissemination

In this section, we introduce a comprehensive DRL-based solution designed to address the energy-efficient d ata dissemination problem in UAV networks as defined in (8). Our approach harnesses the capabilities of a DRL framework to autonomously guide a UAV in making data dissemination decisions that optimize energy usage across the network. Central to this solution is the use of a Double Deep Q-Network (DDQN) algorithm, or a proximal policy optimisation (PPO) algorithm, where the UAV acts as a single-agent, dynamically interacting with the network environment.

Through iterative training, the UAV effectively learns to classify IoT devices into tiers, associate Tier 2 devices with Tier 1 devices, and intricately plan its flight path to minimize energy consumption while ensuring timely data delivery. By encapsulating the state space, action space, and the reward mechanism within a structured learning paradigm, our framework demonstrates a novel application of DRL in the context of UAV-assisted IoT networks. The following subsections detail the components of our proposed solution, including the MDP formulation, the DDQN/PPO agent approach, device association and path planning algorithms, and performance metrics critical to evaluating the efficiency o f o ur strategy.

A. MDP Formulation

In this section, we define the RL elements, i.e., state space, action space and long-term cumulative reward within the context of the proposed framework, respectively. The elements of RL are summarized in Fig. 2. The MDP in our UAV-assisted data dissemination framework is defined by the state space S, action space A, and the reward function R. Specifically:

- State Space ($\mathcal{S}$): The system state at time-step t, defined as s_t, consists of the following four parts: 1) The classification matrix ($\boldsymbol{T}$); 2) The association matrix ($\boldsymbol{\mu}$); 3) The UAV tour matrix ($\bar{\boldsymbol{\psi}}$); and 4) The total energy expenditure ($E(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}})$). Hence, the state space of the system can be characterized by $\mathcal{S} = \left\{\boldsymbol{T}; \boldsymbol{\mu}; \bar{\boldsymbol{\psi}}; E(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}})\right\}$ of size $2N+N^2+3(N+2)+1$. At the initial system state, s_{t_0} is a tuple consisting of $[\boldsymbol{T_0}; \boldsymbol{\mu_0}; \bar{\boldsymbol{\psi}}_{\boldsymbol{0}}; E_0(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}})]$, where $\boldsymbol{T_0}$ is obtained when all devices are set as Tier 1, $\boldsymbol{\mu}$ is a matrix with elements $\mu_{ik} = 0$ indicating no association, $\bar{\boldsymbol{\psi}}_{\boldsymbol{0}}$ is the obtained tour when the UAV visits all devices, and $E_0(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}})$ is the energy obtained for this initial state setup. The final or terminal state s_{t_f} is the state when convergence is met.
- Action Space ($\mathcal{A}$): The UAV is assumed to perform a tour and hover over Tier 1 devices to disseminate data. These Tier 1 devices depicted by the $\mathbf{T}$ matrix can be represented by a decimal number. The agent will select an action $a_t \in \mathcal{A} = \left\{1, \cdots, 2^N - 1\right\}$ at each time-step, where $\mathcal{A}$ is the action space of size 2^N, that will cause the transition of the system from s_t to s_{t+1}. For example, for $N = 4$, $\mathcal{A} = \{1, \cdots, 15\}$. If $a_t = 10 \equiv 1001_2$, then $T_{11}, T_{14} = 1$, while $T_{22}, T_{23} = 0$.
- Reward Function ($\mathcal{R}$): When action a_t is taken under state s_t, the reward r_t which is a function of the obtained $E(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}})$ can be given by

$$r_t = \begin{cases} -10 & \text{if } E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) = \infty \\ -1 & \text{if } E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) < E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right)_{\text{best}} \\ 0 & \text{if } E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) = E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right)_{\text{best}} \\ 10 & \text{if } E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right) > E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right)_{\text{best}}, \end{cases} \tag{9}$$

where $E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right)_{\text{best}} = \arg\min\left(E\left(\boldsymbol{T}, \boldsymbol{\mu}, \bar{\boldsymbol{\psi}}\right)\right)$ obtained in previous timesteps. The objective of the proposed system is to find an optimal policy $\pi^*(\cdot)$, typically a function approximator with tunable parameters, that maps a_t given s_t to maximize the cumulative reward.

The necessity of employing DRL over conventional optimization methods arises from the dynamic and complex nature of the problem. The UAV-assisted data dissemination framework involves a continuously changing environment, with varying network conditions and device distributions. Moreover, the application of DRL enables the system to leverage trial-and-error learning to discover innovative strategies for device classification, association, and path planning that minimize energy consumption while maintaining effective data dissemination.

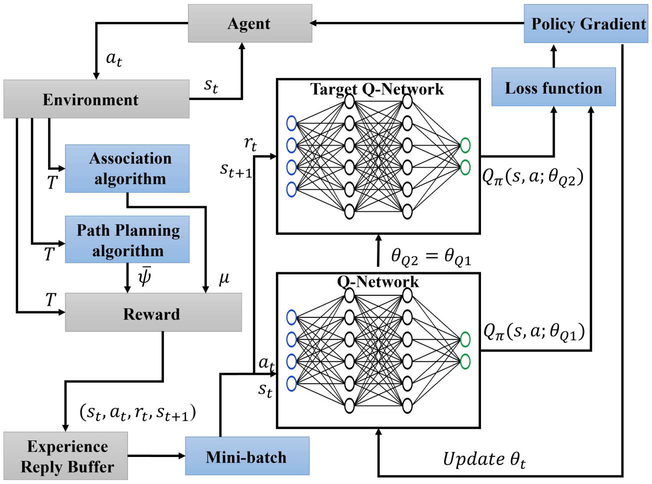

FIGURE 3. Proposed DRL-based UAV-assisted data dissemination solution framework.

B. Proposed Approach

Although dynamic programming techniques can be employed to solve MDP and find the optimal policy $\overset{*}{\pi}(\cdot)$, a high computational complexity arises for large-scale MDPs. In recent years, RL algorithms have been increasingly employed by researchers to tackle the issue of dimensionality for large-scale MDPs. RL is a goal-oriented computational approach where an agent interacts with an unknown dynamic environment to learn how to perform a task. The objective of the agent is to maximize the cumulative reward for the task without being explicitly programmed and without human intervention [6].

In the proposed approach, the environment is modeled as described in Section III. The environment is considered dynamic and unknown at the start of every training episode. The setup, however, remains unchanged throughout the time steps of each episode.In the context of our UAV-assisted data dissemination system, depicted in Fig. 1, the UAV serves as the central entity responsible for collecting and processing network data essential for the DRL-based approach. It compiles comprehensive information regarding IoT device locations, energy states, and data demands. This rich dataset is critical for the UAV to comprehend the prevailing network conditions and to facilitate informed decision-making for optimizing data dissemination pathways.

TABLE 2. DDQN Q-Network Dimension.

Layer	Type	Dimension
Input layer	Feature input	$\vert\mathcal{S}\vert$ features
Hidden layer 1	Fully connected	$\vert\mathcal{S}\vert \times 256$
Hidden layer 2	Fully connected	256×256
Output layer	Fully connected	$256 \times \vert\mathcal{A}\vert$

1) Devices Classification

- **DDQN Agent Approach**

  In the proposed framework depicted in Fig. 3, we adopt a double DQN (DDQN) as the agent learning algorithm, to solve the UAV-assisted data dissemination problem. The DQN algorithm is a model-free, online, off-policy RL method in which a value-based RL agent is employed to train a critic that estimates and returns future rewards [22], [23]. However, one limitation of DQN is its tendency to overestimate action values for large-scale function approximation. This limitation is addressed by the utilization of DDQN, which provides generalized problems with large-scale function approximation, while offering improved performance [24]. The selection of this type of agent type is motivated by the fact that our observation space is continuous and our action space is discrete. Our DDQN algorithm implementation is presented in Algorithm 2.

  The implemented DDQN agent has two function approximators in the form of neural networks, whose weights θ_{Q1} and θ_{Q2} are updated with every iteration to obtain $\overset{*}{\pi}(\cdot)$. However, $\overset{*}{\pi}(\cdot)$ can only be achieved as $t \rightarrow \infty$, which implies that the neural networks can only approximate, and hence, the name function approximators. In DDQN, a target critic network is used to generate an action, i.e., the device classification in decimal representation, and a critic network is used to determine the Q-value of the action. The critic network is made up of four layers described in Table 2, and a ReLU activation function $f(x) = \max(0, x)$ is chosen [25]. The experience reply buffer $\mathcal{D}$ stores the experience of the agent, which is the transition pair at time-step t and is defined as (s_t, a_t, r_t, s_{t+1}).

  This shows that at a state s_t applying an action a_t on the environment yields a reward r_t and takes the environment to state s_{t+1}. The mini-batch is an experience pair of size K, that is sampled from $\mathcal{D}$ and used as input to the critic network. The approximation of the Q-value should be able to approach the state-action value function estimated by the DDQN. This state-action value function is given by the Bellman equation as

$$Q_{\overset{*}{\pi}}\left(s_t, a_t\right) = \mathbb{E}_{s_t}\left[r_{s_t, s_{t+1}, a_t} + \gamma \max_{a_{t+1}} Q_{\overset{*}{\pi}}(s_{t+1}, a_{t+1}) \mid s_t, a_t\right], \tag{10}$$

  where $0 < \gamma < 1$ is the discount factor that determines how important the prediction of future rewards is. Therefore, at the end of every time step during training, the DDQN updates the weights to minimize the mean-squared error loss,

$$\min_{\theta_t} L_t\left(\theta_t\right), \tag{11}$$

  where $L_t(\theta_t) = \mathbb{E}_{s_t, a_t}[(y_t - Q_\pi(s_t, a_t; \theta_t))^2]$, with y_t denoting the estimated function value at time-step t when s_t is the current state, and a_t is the taken action, which is given by

$$y_t = \mathbb{E}_{s_t}\left[r_{s_t, s_{t+1}, a_t} + \gamma \max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1}; \theta_{t-1}) \mid s_t, a_t\right]. \tag{12}$$

  The stochastic gradient descent (SGD) algorithm [26] is used during training to update the weights θ_t at every time-step t. In SGD, the weights are initialized randomly and iteratively updated based on a learning rate η. The weight update formula is given as

$$\theta_t = \theta_t - \eta \nabla L_t(\theta_t). \tag{13}$$

  Building upon this framework, the process flow of our DRL-based solution for UAV-assisted data dissemination is further elaborated in Fig. 3. This figure demonstrates the intricate interactions between the UAV, functioning as the DDQN agent, and the dynamic environment it operates within. During the training phase, the UAV amasses real-time network data, encompassing the locations of IoT devices, their data needs, and prevailing energy levels. Such information drives the refinement of the DDQN's policy gradient and loss function, ultimately enhancing the action-value function iteratively. The association and path planning algorithms, integral components of the environment,

Algorithm 2: DQN-Based Framework for UAV-assisted Data Dissemination Problem.

```
Initialize $\theta_{Q1}, \theta_{Q2}, \epsilon_t = 1, \delta, i = j = 0,$ and $K$;
while $j < |\mathcal{E}|$ do
    set $s_t = s_{t_0} = [T_0, \mu_0, \bar{\psi}_0, E_0]$;
    while $t < |\mathcal{T}|$ do
        $X_t \sim U\,(0, 1)$;
        if $X_t < \epsilon_t$ then
            $a_t = \text{random}(1, \cdots, 2^N - 1)$;
        else
            $a_t = \arg\max_{a_t} Q(s_t, at \mid \theta_{Q1})$;
        end
        $a_t \mapsto \mathbf{T}$;
        Obtain $\mu$ matrix from Algorithm 2 or 3;
        Obtain $\bar{\psi}$ from Algorithm 4;
        Evaluate $E(\mathbf{T}, \mu, \bar{\psi})$;
        Obtain $r_t$ and $s_{t+1}$;
        Store the experience $[s_t, a_t, r_t, s_{t+1}]$ in $\mathcal{D}$;
        Sample a random mini-batch of $K$
          experiences from $\mathcal{D}$;
        if $s_t == s_{t_f}$ then
            $y_t = r_t$;
        else
            $a_{t_{\max}} = \arg\max_{a_{t+1}} Q(s_{t+1}, a_{t+1} \mid \theta_{Q1})$;
            $y_t = \mathbb{E}_{s_t, a_t}[r_{s_t, s_{t+1}, a_t} + \gamma Q_\pi(s_{t+1}, a_{t_{\max}} \mid$
              $\theta_{Q2}) \mid s_t, a_t]$;
        end
        Update Q-network parameters
          $\theta_{Q1} = \theta_t - \eta \nabla L_t(\theta_t)$;
        where $L_t(\theta_t) = \mathbb{E}_{s_t, a_t}[(y_t - Q_\pi(s_t, a_t; \theta_t))^2]$;
        Update the target Q-network parameters $\theta_{Q2}$;
        Update the exploration rate $\epsilon_{t+1} = \epsilon_t - \delta$;
        Set $s_t = s_{t+1}$ $t = t + 1$;
    end
    $j = j + 1$;
end
Output optimal policy $\overset{*}{\pi}$;
```

deliver feedback in the form of reward signals, thereby enabling a robust learning loop that underpins the UAV's operational decisions.

- **PPO Agent Approach** Within our UAV-assisted data dissemination framework, as illustrated in Fig. 3, we integrate the Proximal Policy Optimization (PPO) algorithm as an alternative agent learning strategy. PPO, a policy gradient method for reinforcement learning, offers several advantages, particularly for problems with continuous action spaces or when seeking to balance exploration with exploitation in complex environments [27]. Unlike value-based methods, PPO operates directly on the policy space, enabling it to manage the stochasticity of actions in a more nuanced manner.

Algorithm 3: PPO-Based Framework for UAV-assisted Data Dissemination Problem.

```
Initialize policy parameters $\theta$, value function
  parameters $\phi$, and environment $\mathcal{E}$;
for each iteration do
    Collect set of trajectories $\mathcal{D}_\theta$ by running policy
      $\pi_\theta$ in the environment;
    Compute rewards-to-go $\hat{R}_t$ and advantage
      estimates $\hat{A}_t$ based on the current value function
      $V_\phi$;
    Update the policy by maximizing the PPO-Clip
      objective via SGD:
      $$\theta \leftarrow \arg\max_\theta \frac{1}{|\mathcal{D}_\theta| T} \sum_{\tau \in \mathcal{D}_\theta} \sum_{t=0}^{T} \min\left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip}\left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right),$$
      ;
    Update the value function by minimizing the loss
      via SGD:
      $$\phi \leftarrow \arg\min_\phi \frac{1}{|\mathcal{D}_\theta| T} \sum_{\tau \in \mathcal{D}_\theta} \sum_{t=0}^{T} \left( V_\phi(s_t) - \hat{R}_t \right)^2$$
      ;
end
Output optimized policy parameters $\theta^*$;
```

The core of the PPO algorithm lies in its objective function, which modifies the policy gradient for improved training stability and efficiency. It uses a clipped surrogate objective to prevent overly large updates, thereby ensuring a smooth policy evolution. This characteristic is crucial in our setting, where sudden changes in UAV flight patterns or data dissemination strategies could lead to suboptimal outcomes or instability. Our PPO algorithm implementation is presented in Algorithm 3. Our implementation of PPO employs separate neural networks for the policy (actor) and value (critic) functions, denoted as π_θ and V_ϕ, respectively, where θ and ϕ are the network parameters. The actor-network is responsible for selecting actions based on the current state, while the critic estimates the value of being in a given state, helping to gauge the advantage of performed actions.

Training involves iteratively updating the policy based on the clipped objective and refining the value function estimate to reduce the variance of policy updates. The dual nature of this approach—optimizing both the policy and value estimates—enables our framework to adaptively improve UAV control and data dissemination strategies over time. Algorithm adjustments, such as the entropy bonus to encourage exploration or the specific form of the advantage function, are fine-tuned based on empirical performance within our simulated environment.

Algorithm 4: Association Algorithm.

```
Initialize $i = z = 0$;
Given $a_t$, devices location $(x_k, y_k)$;
while $i < L$ do
    Extract $\tau_{2_i}$ device location $d_{2_i} \leftarrow (x_k, y_k)$;
    while $z < m$ do
        Extract $\tau_{1_z}$ device location $d_{1_z} \leftarrow (x_k, y_k)$;
        $d_z = D_e(d_{2_i}, d_{1_z})$;
        $z = z + 1$;
    end
    $\tau_{1_i} = \text{arg} \min_{d_z}(\tau_1(d_z))$;
    $\tau_{2_i} \in \tau_{1_i}$;
    $i = i + 1$;
end
Obtain $\boldsymbol{\mu}$ matrix based on the association;
```

The adaptability of PPO, combined with its effectiveness in continuous action domains, makes it an ideal choice for our UAV-assisted framework, enabling efficient, scalable, and robust data dissemination strategies in complex and dynamic environments.

2) Devices Association

In our framework, we develop an algorithm that enables devices association. The input of the association algorithm is the current action a_t applied to the environment by the agent. The action a_t is translated into the classification matrix $\boldsymbol{T}$ and then this is used to associate the devices. We employ the nearest neighbor heuristic [28] to develop an the association algorithm. In our proposed association algorithm, each τ_2 device is associated with a τ_1 device that is nearest to it in physical distance. Here we use devices' locations as input and employ the Euclidean distance $D_e(\cdot, \cdot)$ to realize the association, as described in Algorithm 4.

3) Path Planning

To obtain the optimum tour $\bar{\boldsymbol{\psi}}$, we propose a path planning algorithm. Similar to the association algorithm, the action a_t applied to the environment by the agent is translated into the classification matrix $\boldsymbol{T}$, and then used to obtain the physical locations of the τ_1 devices. In our framework, we employ the LKH algorithm, which is an effective method for obtaining an optimum and near-optimum solution for the symmetric TSP [29]. For the route mapping, we employ the greedy nearest neighbor heuristic, while for route-improvement, we employ the 2-opt heuristic of the LKH algorithm [30]. The implementation of the DRL-based framework is shown in Algorithm 5. In the algorithm, the UAV starts from an initial docking location ψ_0, takes a tour $\bar{\boldsymbol{\psi}}$ and returns to the docking location ψ_0.

Algorithm 5: UAV Path Planning Algorithm.

```
Initialize $i = z = 0$;
Given $a_t$, devices location $(x_k, y_k)$, $\psi_0$;
uavTour$_0$ = $[\psi_0]$;
while $z < m$ do
    Extract $\tau_{1_z}$ device location $d_{1_z} \leftarrow (x_k, y_k)$;
    $d_z = D_e(\text{uavTour}_0, d_{1z})$;
    $\psi_z = \text{arg} \min_{d_z}(\tau_{1_z}(d_z))$;
    uavTour$_0$ = $[\psi_0; \psi_z]$;
    $z = z + 1$;
end
Set $q_{\text{min}} = 0$;
while $i < m - 2$ do
    $d_z = D_e(d_{i+1}, d_{i+2})$;
    while $z < m$ do
        $d_{z2} = D_e(d_{i+1}, d_{i+3})$;
        $d_{z3} = D_e(d_{i+3}, d_{i+4})$;
        $d_{z4} = D_e(d_{i+2}, d_{i+2})$;
        $q = (d_{z2} - d_{z3}) + (d_{z4} - d_{z1})$;
        if $q < q_{\text{min}}$ then
            swap$(d_{z1}, d_{z2})$;
            swap$(d_{z3}, d_{z4})$;
        end
        uavTour$_z$ = [uavTour$_{z-1}$; $\psi_j$];
        $z = z + 1$;
    end
    $i = i + 1$;
end
Obtain uavTour$_{\text{best}}$;
```

To address the complexity and dynamic nature of UAV-assisted IoT networks, our methodology integrates the DDQN algorithm alongside targeted heuristics. The DDQN algorithm was chosen for its proven efficacy in reducing the overestimation bias present in traditional Q-learning methods, thereby providing more stable and reliable learning outcomes in environments with highly variable and uncertain dynamics, such as those encountered in UAV-assisted data dissemination tasks [23]. This choice is underpinned by DDQN's ability to decouple the selection of actions from the evaluation of their potential rewards, an advancement that significantly enhances decision-making quality in complex scenarios. Furthermore, the incorporation of specific heuristics, such as the nearest-neighbor and Link Kernighan heuristics for device association and path planning, respectively, is motivated by their computational efficiency and robust performance in solving optimization problems that are NP-hard [29]. The combination of DDQN with these heuristics represents a deliberate strategy to harness complementary strengths: DDQN's advanced learning capabilities and the heuristics' efficiency and proven applicability in related optimization challenges.

VI. Performance Metrics

In this section, we discuss the performance metrics employed to quantify the efficiency of our proposed DRL-based framework for the UAV-assisted data dissemination problem.

A. Computational Complexity

Here, we present the mathematical expression of the computational complexity of the baseline solution, the brute force solution, and the proposed DRL-based solution.

1) Baseline Approach

The baseline approach is used as a benchmark to analyze the performance of the proposed approach. In this approach, all devices are assumed to be in Tier 1. This indicates that no strategic classification is required, no association is needed, and the UAV has to follow a tour that stops and disseminates data at every IoT device in the order of their deployment. The overall computational complexity of this approach based on the device classification, device association, and UAV path planning is presented in Table 3.

TABLE 3. Baseline Approach Complexity.

Problem	Complexity
Classification	$\mathcal{O}(1)$
Association	$\mathcal{O}(1)$
Path Planning	$\mathcal{O}(N^2)$
Overall	$\mathcal{O}(1+1+N^2) \approx \mathcal{O}(N^2)$

2) Brute Force Algorithm (BFA) Approach

The optimum solution of the objective function in (8) can be solved iteratively using the brute force algorithm (BFA). To obtain the solution using BFA, we need to iterate over all possible combinations of potential classifications of devices. For each classification combination, we will iterate over all possible permutations of device association, and finally solve the path planning problem in the form of TSP iteratively. This is a very expensive method that will guarantee an optimal solution. The overall computational complexity of this approach based on device classification, device association, and UAV path planning for N devices is presented in Table 4. This approach is impractical for a number of IoT devices $N > 10$ due to the factorial expression.

3) Ant-Colony Based Approach

In the ant-colony optimization framework, an individual ant a executes $\mathcal{O}(N)$, $\mathcal{O}(N^2F)$, and $\mathcal{O}(N^2)$ computations to derive $\boldsymbol{T}^{(a)}$, $\boldsymbol{\mu}^{(a)}$, and $\bar{\psi}^{(a)}$ respectively. The process of assessing the objective function is accomplished through $\mathcal{O}(N^2F^2 + NF^2)$ operations, which simplifies to $\mathcal{O}(N^2F^2)$. Thus, the total computational burden of the ACO method is $\mathcal{O}(N^4F^3AI + N^2I)$, simplifying to $\mathcal{O}(N^4F^3AI)$. This is substantially more efficient than the computationally intensive exhaustive search, which has a complexity of $\mathcal{O}(2^N[N(N-1)/2 + N!]N^2F^2)$, equivalent to $\mathcal{O}(2^N N! N^2 F^2)$.

TABLE 4. BFA Approach Complexity.

Problem	Complexity
Classification	$\mathcal{O}(2^N)$
Association	$\mathcal{O}(N^2)$
Path Planning	$\mathcal{O}(N!)$
Overall	$\mathcal{O}(2^N N^2 N!)$

4) Proposed DRL-Based Approach

The overall-complexity of the proposed solution is evaluated as the sum of the three sub-problems' complexity, i.e., classification, association, and path-planning sub-problems. For a given N, the overall computational complexity of this approach during training of the agent and during run-time for N devices is presented in Table 5.

TABLE 5. DRL-based Approach Complexity.

Process	Complexity
Association algorithm [21]	$\mathcal{O}(N^2)$
Path planning algorithm [25]	$\mathcal{O}(N^3\text{log}N)$
Size of action space $\lvert\mathcal{A}\rvert$	$\mathcal{O}(2^N)$
Size of state space $\lvert\mathcal{S}\rvert$	$\mathcal{O}(N^2)$
Number of episodes $\lvert\mathcal{E}\rvert$	$\mathcal{O}(E)$
Number of time-steps $\lvert\mathcal{T}\rvert$	$\mathcal{O}(T)$
Number of hidden layers $\lvert\mathcal{H}\rvert$	$\mathcal{O}(H)$
Hidden layer dimension $\lvert\mathcal{L}\rvert$	$\mathcal{O}(L)$
Convergence-Overall	$\mathcal{O}(ETL^{2H} \cdot 2^N N^2)$
Run-time-Overall	$\mathcal{O}(2^N N^2)$

B. Convergence

Convergence is defined in terms of the number of episodes at which the target average total reward is achieved. It is expected that as the number of IoT devices N increases, the convergence time also increases. This is due to an increase in the state space input size, as well as the action space output size. From Table 4, it can be seen that the action space input increases exponentially with an increase in N which results in a longer convergence time. The obtained expression has no factorial function which makes it more practical than the BFA.

C. Average Reward

Average reward reflects the cumulative efficiency of the agent's decisions over time, with higher (less negative) average rewards corresponding to more optimal policies.

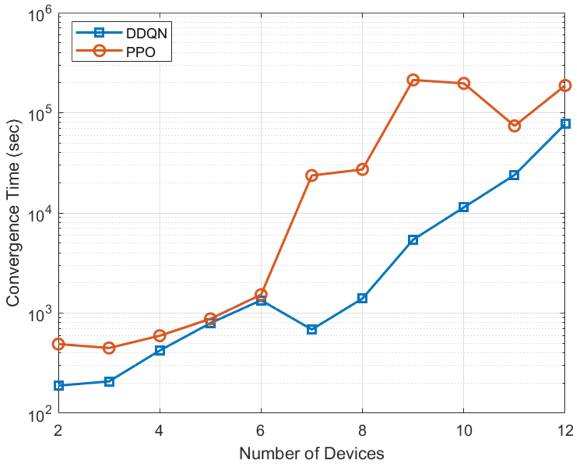

FIGURE 4. Convergence time of the DRL agents as a function of the number of IoT devices N.

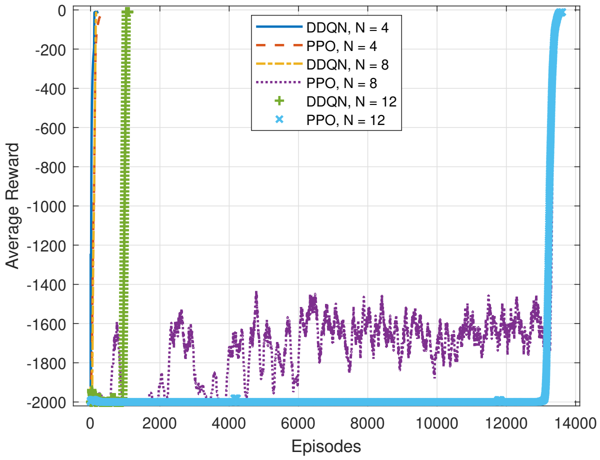

FIGURE 5. Average reward of the DRL agents during training as a function of the number of episodes $|E|$.

The average reward thus serves as an indication of the algorithm's ability to balance immediate costs with long-term gains, which in the case of UAV networks, translates to the judicious management of resources like battery life and bandwidth while fulfilling the network's data requirements.

D. Run-time

The exhaustive search was set as the upper bound and the calculation of the run-time complexity is possible and given as $O(2^N N^2 N!)$, where N is the number of IoT devices in the system. In the proposed approach however, $|\mathcal{H}|$ and $|\mathcal{L}|$ remain fixed after the agent is being trained, leading to a significantly reduced run-time computational complexity, as indicated in Table 4. This is a consequence of the action space and state space dominating the overall complexity for high values of N.

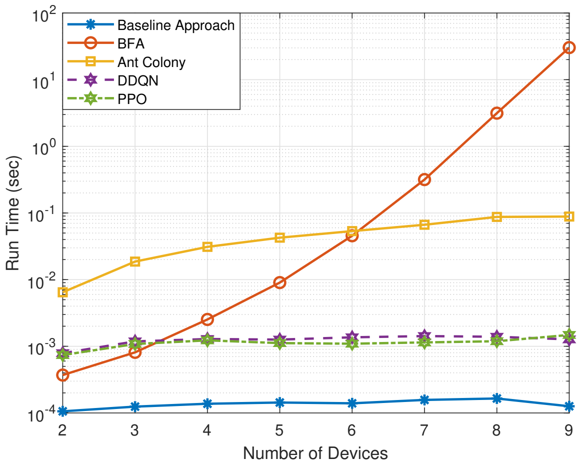

FIGURE 6. Run-time as a function of the number of IoT devices N.

E. Energy Expenditure

Using the energy expression given in (7), we obtained the energy expenditure with variable system model parameters to demonstrate the generalization of the policy $\pi(\cdot)$ obtained by the DDQN agent.

VII. Results and Discussions

In this section, the efficiency of the proposed DRL-based solution is evaluated in terms of the performance metrics described in Section V. We follow the simulation setup in [11] to design the system model in Section III. The default system parameters, unless otherwise stated, are presented in Table 6. The RL hyper-parameters of the proposed scheme are presented in Table 7. These hyper-parameters are tuned during training to achieve good policy for the agent. The effect of some of the main parameters on the total energy expenditure of the proposed framework is studied and the results are compared with the benchmark solutions. The results are obtained as average of 100 runs.

1) Convergence

The relationship between the training convergence time and the number of IoT devices, N, is captured in Fig. 4. Contrary to typical expectations in DRL implementations, we observe that DDQN uniformly outperforms PPO across the range of N. Specifically, as N escalates from 2 to 12, the DDQN algorithm maintains a consistently lower convergence time in comparison to PPO. This inverse relationship between convergence time and network size highlights the effectiveness of DDQN in our UAV-assisted data dissemination context, underscoring its computational efficiency and scalability. This performance is consistent with the computational complexity analysis presented in Table 4, and reinforces DDQN's suitability for larger scale networks where expedient policy learning is crucial.

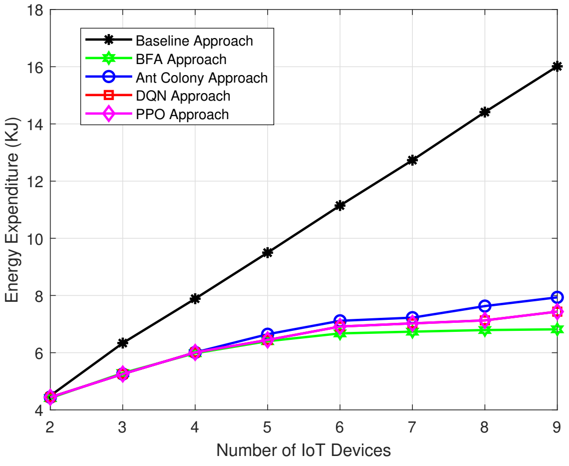

FIGURE 7. Energy expenditure as a function of the number of IoT devices N, with $r_i = 500$ m, $L_j = 5$ kb, $f_i = 20$ files, and $F = 70$ files.

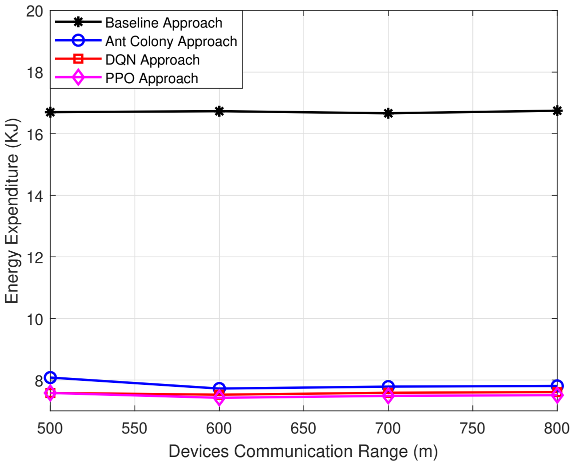

FIGURE 8. Energy expenditure as a function of the devices' communication range r_i, with $N = 12$, $L_j = 5$ kb, $f_i = 20$ files, and $F = 70$ files.

2) Average Reward

Fig. 5 presents the progression of average rewards per episode for the DDQN and PPO algorithms across different network sizes (N = 4, 8, 12). Initially, both algorithms exhibit similar reward trajectories, indicating a comparable exploration phase. As episodes progress, the DDQN algorithm converges more consistently to higher reward values, suggesting a stable learning and optimization process. In contrast, the PPO algorithm experiences more pronounced fluctuations, especially as the network size increases. This is indicative of the greater challenge PPO faces in scaling with the complexity of the task. For N = 12, DDQN demonstrates a robust performance with less variance in rewards, while PPO's performance is characterized by significant oscillations, reflecting a less stable policy learning trajectory. The observed trends underscore DDQN's potential for achieving more efficient policy learning in UAV-assisted data dissemination environments, particularly in larger networks.

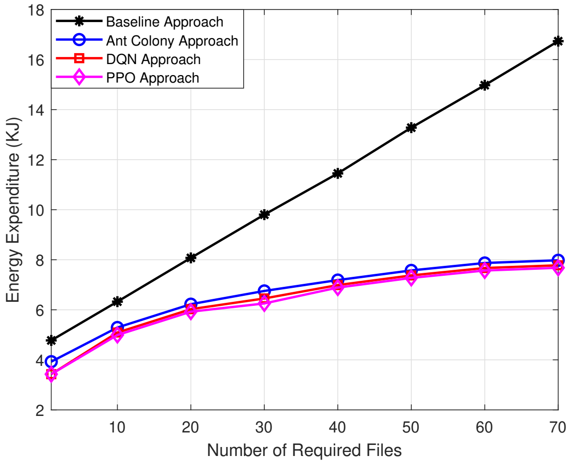

FIGURE 9. Energy expenditure as a function of the number of required files f_i, with $N = 12$, $r_i = 500$ m, $L_j = 5$ kb, and $F = 70$ files.

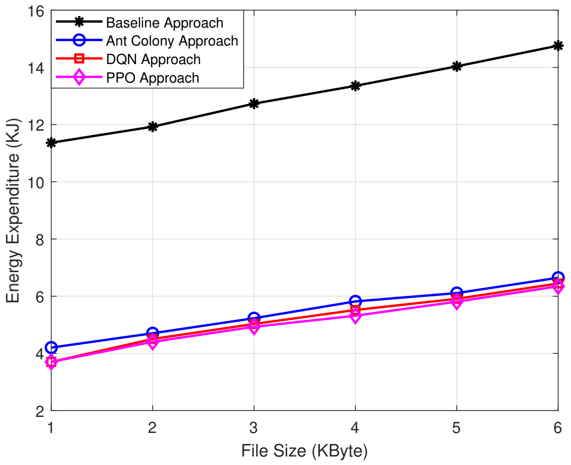

FIGURE 10. Energy expenditure as a function of the file size L_j, with $N = 12$, $r_i = 500$ m, $f_i = 20$ files, and $F = 70$ files.

3) Run-time

The execution time performance of various benchmark approaches, alongside the proposed DRL-based algorithms (DDQN and PPO), as a function of the number of IoT devices, N, is depicted in Fig. 5. For the sake of an equitable comparison, GPU acceleration has been excluded. We observe that the run-time of the baseline approach remains consistently negligible across different values of N. Conversely, the run-time of the Brute-Force Algorithm (BFA) escalates significantly with the increase in N, becoming computationally impractical for $N > 10$.

Interestingly, both the ant-colony algorithm and the proposed DRL-based approaches, including DDQN and PPO,

TABLE 6. System Model Parameters.

Parameter	Value	Parameter	Value
P_T	20 dBm	M	0.5 kg
L_j	5 kb	α	1
γ_0	-60 dB	E	10^3
P_{max}	5 W	A	100
h	100 m	B	20 kHz
p	4	ϵ	3
l	10^3	P_r	0.0126 W
σ	10	v_{max}	12 m/s
P_{T_i}	20 dBm	r	20 cm
σ^2	-110 dBm	β	1
r_i	500 m	$\lvert\zeta_{ik}\rvert^2$	$\sim \text{Exp}\,(1)$
P_{stop}	0 W	I	10^3

TABLE 7. RL Hyper-parameters.

Parameter	Value
Number of episodes $\lvert\mathcal{E}\rvert$	10000
Number of time-steps $\lvert\mathcal{T}\rvert$	100
Discount factor γ	0.99
Initial exploration rate ϵ	1
Exploration decay δ	0.005
Minimum exploration rate ϵ_{min}	0.01
Target critic smooth factor τ	1e-3
Target critic update frequency	1
Experience buffer size $\mathcal{D}$	10000
Minimum batch size K	64
Averaging window size	100
Target average reward	0

maintain relatively stable run-times across the spectrum of N, with only a marginal upward trend. It is notable that for $N < 7$, the BFA outpaces both the ant-colony and DRL-based methods. However, as N surpasses 7, the DRL-based methods demonstrate superior time efficiency compared to the BFA. Additionally, the proposed DRL-based algorithms consistently outperform the ant-colony algorithm in computational time, offering a more expedient approach to finding near-optimal solutions for all considered network sizes.

The relatively flat run-time curve of the proposed DRL-based approaches, particularly PPO, suggests an advantage in scalability and computational efficiency, positioning them as favorable strategies for larger IoT networks where quick adaptation to dynamic environments is paramount.

4) Energy Expenditure

The energy expenditure obtained by the proposed DRL-based framework as a function of the number of IoT devices N, devices communication range r_i, number of required files by the IoT devices f_i and file size L_j is presented in Figs. 6-9, respectively. Fig. 7 presents the energy expenditure as a function of the number of IoT devices N for various optimization approaches. As observed in the figure, there is a clear trend of increasing energy expenditure with an increasing number of IoT devices across all methods. Notably, the brute-force algorithm (BFA) achieves the lowest energy expenditure, demonstrating its ability to exhaustively search for the global minimum. However, it is crucial to consider the practicality of the BFA, especially as N grows beyond 9, where the computational complexity becomes prohibitive for real-world applications due to the factorial growth in the solution space. In contrast, the proposed DRL-based approach achieves energy expenditure levels close to those of the BFA, indicating its effectiveness in optimization tasks. Moreover, the DRL-based approach significantly outperforms the baseline approach while remaining computationally feasible for larger problem scales. Remarkably, for scenarios with more than 4 IoT devices, both variants of the proposed DRL-based approach achieve lower energy expenditure compared to the ant-colony optimization method, highlighting the superior efficiency and effectiveness of DRL-based optimization in UAV-assisted IoT networks. These findings emphasize the potential of the proposed DRL-based framework to provide efficient and scalable solutions for energy expenditure optimization in real-world UAV-assisted IoT scenarios, where computational efficiency and effectiveness are critical considerations.

Fig. 8 illustrates the impact of device communication range, r_i, on the energy expenditure for various optimization strategies. The baseline approach remains invariant with changes in r_i, affirming its non-adaptive trajectory planning which is not influenced by communication range. Conversely, the energy expenditure for the proposed DRL-based methods, encapsulated by the DDQN and PPO algorithms, exhibits a marginal decrement with the increase in r_i. This demonstrates the DRL approaches' proficiency in harnessing the broader communication capabilities to optimize the UAV's flight plan, consequently reducing energy usage. The Ant Colony approach, however, shows a gradual uptick in energy expenditure, suggesting a less effective use of increasing r_i in path optimization. These trends underscore the nuanced behaviors of different optimization strategies in response to variable device communication ranges within UAV-assisted networks.

As depicted in Fig. 9, the energy expenditure trends upward with the increase in the number of required files for all optimization strategies. The baseline approach evidences a proportional increase, reflecting its lack of adaptive efficiency in managing larger file distributions. In contrast, the proposed DRL-based methods, DDQN and PPO, along with the Ant Colony optimization, manifest markedly improved energy efficiency. Notably, the DRL-based algorithms exhibit a marginal increase in energy usage, maintaining a near-parallel trajectory irrespective of the growing number of files. This is indicative of their sophisticated optimization capabilities, which mitigate the impact of larger file counts on energy consumption. The Ant Colony approach, albeit superior to the baseline, does not match the efficiency of the DRL strategies, underscoring their effectiveness in energy-limited UAV network environments. The similarity in the performance of DDQN and PPO approaches suggests comparable optimization proficiencies in the domain of data dissemination.

The variation of energy expenditure with respect to file size is captured in Fig. 10. The Baseline approach exhibits a linearly increasing energy cost, indicative of its non-optimizing nature with respect to file size. In stark contrast, the Ant Colony and the proposed DRL-based algorithms, DDQN and PPO, present a significantly less pronounced rise in energy use as file size increases. This suggests an enhanced ability of these algorithms to handle larger data payloads without a corresponding linear increase in energy expenditure. It is particularly noteworthy that the DDQN and PPO algorithms are almost indistinguishable in their energy efficiency, revealing their comparable adeptness at optimizing energy consumption across different file sizes. This underscores the value of DRL approaches in energy-constrained environments, such as UAV networks, where efficient data management is paramount for operational success and longevity.

In recognizing the inherent challenges associated with the computational complexity of optimization algorithms, including our proposed DRL-based framework, it's important to note that as N increases, so too does the complexity of our algorithm. This is a common characteristic among many sophisticated optimization techniques, and while it presents challenges for scalability, it is crucial for ensuring the depth and thoroughness of the optimization process, particularly in complex, dynamic environments such as UAV-assisted IoT networks.

Our study's findings on enhancing energy efficiency in UAV-assisted IoT networks have profound practical implications across various sectors. In urban planning, energy-efficient UAV networks can support advanced monitoring and data collection for traffic management and infrastructure maintenance, facilitating smoother city operations and improved public safety. Within industrial operations, such deployments can enable real-time monitoring of vast industrial sites and logistics operations, ensuring timely maintenance and operational efficiency. Moreover, in the context of smart city development, energy-efficient UAV-assisted IoT networks are pivotal in deploying sensors and devices for environmental monitoring, public services optimization, and enhancing the quality of life for residents through intelligent data-driven decision-making. These applications underscore the significance of our research in contributing to the sustainable and efficient development of urban and industrial ecosystems.

VIII. Conclusions

In this paper, we have proposed a DRL scheme with the aim to solve a joint classification, association, and path planning optimization problem in a UAV-assisted data dissemination. The objective is to minimize the total energy expenditure while guaranteeing the delivery of required files to the IoT devices. The proposed system can effectively deal with dynamic environments as its execution only relies on the considered system model. The obtained results show that the DRL-based approach can reduce total energy expenditure as compared to the baseline approach for all performance metric parameters. Additionally, for all values of N, the DRL-based approach can achieve a near optimal solution within a shorter period of time compared to the ant-colony approach. However, it was observed that although the run-time of the proposed solution is very low and realistic for different values of N, the training convergence time increases with increasing N. As we look towards the future, enhancing the scalability of our DRL-based algorithm for larger IoT networks represents a pivotal area of our ongoing research. We aim to explore more computationally efficient DRL architectures, consider the potential of distributed DRL approaches, and investigate the integration of approximation algorithms. These strategies are intended to mitigate the impact of increased problem scale on computational complexity, ensuring that our framework remains both practical and effective for a wider range of network sizes.

Looking to the future, we are committed to further enhancing the scalability of our DRL-based framework to accommodate larger IoT networks. This includes investigating more computationally efficient DRL architectures that can manage the expanding action space without compromising on performance. In addition, we will consider the potential of distributed DRL approaches, which may offer a path toward parallelized learning and decision-making across multiple UAVs or network clusters. Another promising direction is the integration of approximation techniques aimed at reducing the complexity of action selection, thereby accelerating the convergence process.

Beyond these technical enhancements, we envisage our solution being deployed in real-world IoT networks, where its impact on energy efficiency could extend operational longevity and reduce costs. The adaptability of our DRL-based system positions it well for dynamic and unpredictable environments, which are characteristic of many IoT applications, from smart cities to precision agriculture. As IoT devices proliferate and UAV technologies advance, we anticipate that our framework will contribute to the evolution of smart, autonomous, and energy-efficient data dissemination methods that are both scalable and robust.

REFERENCES

[1] C. Y. Chong and S. P. Kumar, "Sensor Networks: Evolution, Opportunities, and Challenges," *Proceedings of the IEEE*, vol. 91, no. 8, pp. 1247–1256, Aug. 2003.

[2] 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.

[3] M. Ruiz, E. Álvarez, A. Serrano, and E. Garcia, "The Convergence between Wireless Sensor Networks and the Internet of Things; Challenges and Perspectives: a Survey," *IEEE Trans. Veh. Technol.*, vol. 14, no. 10, pp. 4249–4254, Oct. 2016.

[4] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile Internet of Things: Can UAVs Provide an Energy-Efficient Mobile Architecture?" in *Proc. IEEE Global Communications Conference (GLOBECOM)*, no. Iceca. IEEE, 2016, pp. 1–6.

[5] H. Ghazzai, M. B. Ghorbel, A. Kadri, M. J. Hossain, and H. Menouar, "Energy-Efficient Management of Unmanned Aerial Vehicles for Underlay Cognitive Radio Systems," *IEEE Trans. Green Commun. and Netw.*, vol. 1, no. 4, pp. 434–443, Dec. 2017.

[6] S. S. Sutton and G. A. Barto, *Reinforcement Learning: An Introduction, Second Edition*. MIT Press and The McGraw-Hill Companies Inc, 2018.

[7] Z. Xue, J. Wang, G. Ding, H. Zhou, and Q. Wu, "Maximization of Data Dissemination in UAV-Supported Internet of Things," *IEEE Wireless Commun. Lett.*, vol. 1, no. 1, pp. 185–188, Feb. 2019.

[8] ——, "Data Dissemination in IoT Using a Cognitive UAV," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 209–212, Feb. 2019.

[9] A. M. Almasoud and A. E. Kamal, "Data Dissemination in IoT Using a Cognitive UAV," *IEEE Trans. Cognitive Commun. Netw.*, vol. 5, no. 4, pp. 849–862, Dec. 2019.

[10] F. Cheng, S. Zhang, Z. Li, Y. Chen, N. Zhao, F. R. Yu, and V. C. M. Leung, "UAV Trajectory Optimization for Data Offloading at the Edge of Multiple Cells," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6732–6736, Jul. 2018.

[11] A. A. Al-Habob, O. A. Dobre, S. Muhaidat, and H. V. Poor, "Energy-Efficient Data Dissemination Using a UAV: An Ant Colony Approach," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 6732–6736, Jan. 2021.

[12] F. Xu, F. Yang, C. Zhao, and S. Wu, "Deep Reinforcement Learning Based Joint Edge Resource Management in Maritime Network," *China Commun.*, vol. 17, no. 5, pp. 211–222, May. 2020.

[13] F. Zeng, R. Zhang, X. Cheng, and L. Yang, "UAV-Assisted Data Dissemination Scheduling in VANETs," in *International Conference on Communications (ICC)*, no. Iceca. IEEE, 2018, pp. 1–6.

[14] Q. Liu, L. Shi, L. Sun, J. Li, M. Ding, and F. Shu, "Energy-Efficient Data Dissemination Using a UAV: An Ant Colony Approach," pp. 5723–5728, Jan. 2020.

[15] Z. Xiong, Y. Zhang, W. Y. B. Lim, J. Kang, D. Niyato, C. Leung, and C. Miao, "UAV-Assisted Wireless Energy and Data Transfer With Deep Reinforcement Learning," *IEEE Trans. Cognitive Commun. Netw.*, vol. 7, no. 1, pp. 85–99, Mar. 2021.

[16] K. Li, W. Ni, E. Tovar, and M. Guizani, "Joint Flight Cruise Control and Data Collection in UAV-Aided Internet of Things: An Onboard Deep Reinforcement Learning Approach," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9787–9799, Jun. 2021.

[17] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination," *IEEE Trans. Communun.*, vol. 68, no. 3, pp. 1581–1592, Dec. 2020.

[18] H. Qie, D. Shi, T. Shen, X. Xu, Y. Li, and L. Wang, "Joint Optimization of Multi-UAV Target Assignment and Path Planning Based on Multi-Agent Reinforcement Learning," *IEEE Access*, vol. 7, pp. 146 264–146 272, 2019.

[19] K. Stylianopoulos, M. Merluzzi, P. Di Lorenzo, and G. C. Alexandropoulos, "Lyapunov-driven deep reinforcement learning for edge inference empowered by reconfigurable intelligent surfaces," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49357.2023.10095112

[20] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *2020 IEEE Int. Conf. Commun. (ICC)*, 2020, pp. 1–6.

[21] Q. V. Do, Q.-V. Pham, and W.-J. Hwang, "Deep reinforcement learning for energy-efficient federated learning in uav-enabled wireless powered networks," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 99–103, 2022.

[22] V. Mnih, K. Kavukcuoglu, and D. Silver, "Human-level control through deep reinforcement learning," *Nature 518*, pp. 529–533, Feb. 2015.

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013.

[24] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," in *30th AAAI Conference on Artificial Intelligence*. AAAI, 2016, pp. 2094–2100.

[25] V. Nair and E. G. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *27th International Conference on Machine Learning Haifa, Israel 2010*, Jul. 2010, pp. 2094–2100.

[26] L. Bottou and O. Bousquet, "The Tradeoffs of Large Scale Learning," in *20th International Conference on Neural Information Processing Systems*, 2008, pp. 161–168.

[27] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[28] G. Kizilateş and F. Nuriyeva, "Deep Reinforcement Learning Based Joint Edge Resource Management in Maritime Network," *Adv. Intell. Syst. Comput.*, vol. 225, no. 1, pp. 111–118, Apr. 2013.

[29] Lin and Kernighen, "An effective heuristic algorithm for the traveling-salesman problem (k-opt)," 1973.

[30] A. Punnen, F. Margot, and S. Kabadi, "TSP heuristics: Domination analysis and complexity," *Algorithmica (New York)*, vol. 35, no. 2, pp. 111–127, Aug. 2003.