IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 May 2024, accepted 4 July 2024, date of publication 9 July 2024, date of current version 14 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3425497

SURVEY

Deep Reinforcement Learning for AoI Minimization in UAV-Aided Data Collection for WSN and IoT Applications: A Survey

OLUWATOSIN AHMED AMODU[1], CHEDIA JARRAY[2], RAJA AZLINA RAJA MAHMOOD[3], HUDA ALTHUMALI[4], (Graduate Student Member, IEEE), UMAR ALI BUKAR[5], ROSDIADEE NORDIN[6], NOR FADZILAH ABDULLAH[1], (Member, IEEE), AND NGUYEN CONG LUONG[7]

[1]Department of Electrical, Electronics and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia
[2]Become: Technology, Science, AI and Automation Laboratory, 75013 Paris, France
[3]Department of Communication Technology and Network, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM), Serdang, Selangor 43400, Malaysia
[4]Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, Jubail 31961, Saudi Arabia
[5]Centre for Intelligent Cloud Computing (CICC), Faculty of Information Science and Technology, Multimedia University, Bukit Beruang, Melaka 75450, Malaysia
[6]Department of Engineering, School of Engineering and Technology, Sunway University, Bandar Sunway, Selangor Darul Ehsan 47500, Malaysia
[7]Faculty of Computer Science, Phenikaa University, Hanoi 12116, Vietnam

Corresponding authors: Oluwatosin Ahmed Amodu (oluwatosin.amodu@ukm.edu.my) and Nor Fadzilah Abdullah (fadzilah.abdullah@ukm.edu.my)

This work was supported by Universiti Kebangsaan Malaysia through Dana Impak Perdana 2.0 under Grant DIP-2022-020.

ABSTRACT Deep reinforcement learning (DRL) has emerged as a promising technique for optimizing the deployment of unmanned aerial vehicles (UAVs) for data collection in wireless sensor networks (WSNs) and Internet of Things (IoT) applications. With DRL, UAV trajectory can be optimized, optimal data collection points can be determined, sensor node transmissions can be scheduled efficiently, and irregular traffic patterns can be learned effectively. In view of the significance of DRL for UAV-assisted IoT research in general and, more specifically, its use for time-critical applications, this paper presents a review of the existing literature on UAV-aided data collection for WSN and IoT applications related to the application of DRL to minimize the Age of Information (AoI), a recent metric used to measure the degree of freshness of transmitted information collected in data-gathering applications. This review aims to provide insights into the state-of-the-art techniques, challenges, and opportunities in this domain through an extensive analysis of a sizable range of related research papers in this domain. It discusses application areas of UAV-assisted IoT, such as environmental monitoring, infrastructure inspection, and disaster response. Then, the paper focuses on the proposed works, their optimization objectives, architectures, simulation libraries and complexities of the various DRL-based approaches used. Thereafter discussion, challenges, and some opportunities for future work are provided. The findings of this review serve as a valuable resource for researchers and practitioners, guiding further advancements and innovations in the field of DRL for UAV-aided data collection in WSN and IoT applications.

INDEX TERMS
Age of information (AoI), data acquisition, deep reinforcement learning (DRL), drones, energy-efficiency, Internet of Things (IoT), scheduling, trajectory, unmanned aerial vehicles (UAVs), wireless sensor networks (WSN).

The associate editor coordinating the review of this manuscript and approving it for publication was A. Taufiq Asyhari.

I. INTRODUCTION

The utilization of unmanned aerial vehicles (UAVs) for data collection in wireless sensor networks (WSNs) and IoT applications has gained significant attention due to their rapid mobility and maneuverability. In both rural and urban monitoring scenarios, UAVs deliver sensed data from remote areas to designated destinations, such as base stations. These data-gathering applications are vital in time-critical applications such as medical supply delivery and disaster scenarios where life could be at stake. Ensuring timely data delivery is crucial in such applications. This necessitates a clear definition of the concept of timely data delivery. The Age of Information (AoI) metric has recently been proposed to define timeliness in wireless communication applications, particularly, those that involve the transfer of data from source to destination.

A. AGE OF INFORMATION

AoI can be defined as the elapsed time since the most recent update packet generated at the source node was received at its destination [1]. AoI and its variants (average AoI, peak AoI, sum AoI, weighted sum AoI etc) have thus become popular in recent times as they could help to quantify the freshness of collected information in data-gathering applications, including data sensed by ground-based devices (such as sensors and internet of things (IoT) devices) and transported by UAVs. The evolution of AoI largely depends on the underlying system assumptions and several examples can be found in literature, for instance, [2], [3], [4], [5]. Similarly, AoI characterization has gained popularity and is scattered in different portions of literature. As a result of the strict AoI requirement for real-time status updates and applications, researchers have studied AoI-aware UAV-assisted wireless transmission in the IoTs [5]. AoI applications extends beyond UAVs and involves other real-world applications such as caching, energy harvesting networks, networked monitoring and cyber physical systems, as well as data-analytics applications, information-oriented systems and the IoTs [6]. Particularly, IoT applications constitutes one of the most popular sensing applications which have experienced unprecedented growth in the last couple of years. Many of these applications require data either in real-time or in a very fresh state. An industrial robot, for instance, is an example of IoT application that requires data sensing and timely delivery of information. Artificial intelligence may be used to analyze the data collected by these robots or sensor nodes deployed within the factory to optimize the production process in a timely fashion [7].

B. THE RISE OF DRL FOR AOI MINIMIZATION IN UAV-ASSISTED IOT

Researchers have focused on minimizing AoI in UAV-assisted IoT applications using machine learning-based methods. Particularly, most researchers formulate complex optimization problems with diverse constraints that cannot be effectively solved by traditional optimization algorithms especially due to the dimension of the problems such as the UAV trajectory and the large number of sensor node transmissions to be scheduled. Thus, most authors reformulate the problem as a Markov Decision Process (MDP) which is a prerequisite to using reinforcement learning-based algorithms to solve the formulated problems.

However, because many of these problems are multi-dimensional with large discrete or continuous spaces, different classes of deep reinforcement learning algorithms have been adopted. This is a very technical subject, and previous surveys [8], [9], [10] have not captured this technicality, even though they have provided a large understanding of the nature of problems solved and a general classification of aspects and objectives. One of the findings in these works is the predominance of deep reinforcement learning methods for solving the associated problems, leaving a gap to thoroughly review and study these works from a DRL perspective, especially since it is the heart of the technicality of the proposed solutions and the most fundamental background that new researchers should understand and acquire. This review aims to fill that gap by providing a broad understanding of the proposed solutions using DRL while identifying the classes of algorithms, their objectives, the MDP formulation, and their algorithms, as well as discussing some of the interesting insights that could be derived from the problem formulation and DRL algorithms used in the discussed works.

This survey provides an overview of DRL techniques, their applications, and the challenges associated with their implementations. The key concepts and advancements of various DRL-based implementations of a wide range of research papers have been reviewed and summarized. This survey aims to provide an overview of recent advancements, challenges, and applications of DRL in UAV-aided data collection. By examining the recent techniques, this survey aims to contribute to the development of effective DRL-based solutions for data collection in UAV-assisted systems towards improving trajectory optimization, energy management, scheduling planning, and other important aspects, and objective functions.

The survey covers various aspects of DRL, including an introduction to reinforcement learning and the role of neural networks, classes of RL algorithms, Bellman's Equation and common model-free RL algorithms. Subsequently, the proposals using DRL in both single and multi-agent environments for AoI minimization in UAV-assisted wireless communication are presented. Particularly, in this paper, we include papers that have not considered a terrestrial BS or data center [5], [11], [12], [13], [14], [15] as opposed to the prior two surveys on this subject [8], [9] that exclude these works. Next, the MDP formulation for the DRL framework in these works was explained in details including relevant equations and mathematical representations for the states, actions, and reward functions. Then a comprehensive overview of lessons learned from these works from diverse technical perspectives are explicitly discussed. Moreover,

the challenges associated with training DRL agents, such as sample inefficiency, exploration in high-dimensional spaces, stability of learning, and safety concerns, are presented.

II. RELATED SURVEYS, MOTIVATION AND CONTRIBUTIONS

To the best of our knowledge, this paper stands as the only survey targeted at the utilization of DRL for AoI minimization in UAV-assisted IoT applications, classifying the algorithms and providing details on different aspects such as target objectives, MDP representations, simulation libraries, algorithm complexities, and parameter settings, as well as challenges and future considerations. A discussion of some of the most relevant reviews related to AoI minimization and the use of DRL for UAVs are provided in what follows.

A. PRIOR SURVEYS ON AOI

Several researchers have reviewed different portions of AoI literature different from the goal in this paper, for instance, Yates et al. [16], provides a summary of contributions on AoI research for low-latency cyber physical systems and applications requiring time-stamped status updates. This includes methods of analysis and evaluations and scenarios involving single-hop and multi-hop networks. Abbas et al. [17] provides an overview of AoI and its variants in massive (large-scale) IoT networks focusing on queuing policy, scheduling, stochastic modeling and multiple access schemes. Similarly, Yu et al. [18] provided an overview of AoI in cellular internet of things providing its requirements, problem solving methods and challenges in addition to a proposed prediction-based scheme for status updates. Wang et al. [19] also presented a survey on AoI-optimal sampling policies as well as packet management strategies with a focus on resource and energy-constrained nodes. Amodu et al. [8], [9] have presented surveys on AoI minimization in UAV-assisted IoT where DRL was identified to be one of the most common trends in performance optimization [8], and the importance of design aspects such as trajectory optimization, scheduling and energy management were emphasized [8]. None of these works have explored the use of DRL for AoI minimization in UAV-IoT despite researchers have invested huge amounts of time on this subject and very significant achievements in this domain.

B. PRIOR SURVEYS ON THE APPLICATIONS OF DRL

Several related surveys have been conducted to explore the application of deep reinforcement learning in the context of unmanned aerial vehicles (UAVs). Table 2 provides a summary of some of reviews on DRL applications in various fields. These surveys provide comprehensive overviews of DRL techniques, algorithms, and applications specific to UAVs.

The authors [20] presented a comprehensive review on the applications of DRL with respect to wireless communication and networks. These include autonomous and decentralized wireless network applications such as UAV and IoT in which local decisions are required for optimal network performance within an uncertain environment. In such networks, RL has been deployed to obtain optimal policies for decision-making, especially when there is a finite and small state and action space, whereas DRL has been deployed for large-scale networks in situations with a larger state and action space. The authors provide a tutorial on both fundamental and advanced concepts and models relating to DRL. The variants and modifications aimed at solving communication and networking problems include data offloading, network security, preserving connectivity, dynamic network access, wireless caching, and data rate control. In addition, applications of DRL for resource sharing and data collection are discussed, with an exposition on challenges faced and open issues.

Although UAVs have gained popularity in a lot of civilian and military applications such as traffic patrol, surveillance, remote sensing, rescue operations, infrastructure inspection, environmental monitoring, etc., the autonomous UAV operation itself poses a major challenge, especially in unplanned circumstances. This has motivated the proposition of DRL for guiding, navigating, and controlling the UAV amongst other artificial intelligence algorithms. Thus the authors [21] focus on a detailed description of DRL-based techniques and identify their limitations for autonomous UAV control. According to the authors, most works have focused on the use of DRL for UAV control in simulation as opposed to real field-test scenarios.

There has been an increasing demand for drones in applications where UAVs are used autonomously to perform tasks and avoid obstacles using reinforcement learning algorithms. Choosing the right RL algorithm to tackle navigation problems is thus essential, which motivates the authors to identify the UAV navigation applications and tasks and discuss the frameworks and simulation software used for UAV navigation. The authors [22] classify RL algorithms using certain characteristics such as features and use-cases for different navigation problems to help technical experts select the most suitable RL algorithms for their peculiar problems and use-cases. Furthermore, gaps and opportunities for UAV navigation research were identified.

Advancement in cooperative multi-agent systems for carrying out complex tasks in a coordinative manner has increased the popularity of UAV applications in recent years. This motivates the authors [23] to study multi-UAV scenarios and classify them into five groups based on their unique tasks (coverage, communication, target-driven navigation, computational offloading, adversarial search and game). The authors systematically selected the works using DRL for scalable and cooperative multi-UAV communication while critically discussing some of their peculiarities and providing future research directions via a critique of the current assumptions and constraints in DRL-based collaborative multi-UAV research.

Modern cellular networks are largely characterized by high inter-cell interference, particularly because of the universal frequency reuse approach for maximizing spectral efficiency. This is even more challenging as UAVs are introduced into the cellular architecture, as the LOS links also contribute to the level of interference, thus motivating the need for interference management schemes. DRL-based interference management was proposed [24] due to the challenges of traditional solutions in which a priori knowledge of channel information of interfering signals is required. Furthermore, the authors discuss novel approaches to scale and decentralize algorithms via multi-agent reinforcement learning.

C. CONTRIBUTIONS

Through this survey, we aim to provide researchers, practitioners, and enthusiasts with a thorough understanding of the current state-of-the-art and developments in DRL for AoI minimization in UAV-assisted IoT. By highlighting the key techniques, applications, and challenges, we intend to inspire further research and advancements in this exciting and rapidly evolving field. Undoubtedly, DRL has many potential applications and benefits for various domains due to its ability to handle high-dimensional state and action spaces, as well as solve very complex real-world problems.

In this paper, we have used both Google Scholar and Scopus databases for selecting the reviewed papers. For the Scopus search, on June 29, 2023, we used the keywords *uav OR drone OR unmanned AND "age of information" OR "information freshness" OR "data freshness" AND "deep reinforcement learning" OR drl AND iot OR wsn* while excluding results on Edge computing to obtain the papers reviewed in this study. Different from all the prior surveys in Section II, this paper presents the following contributions:

- An overview of UAV-assisted WSN/IoT applications with a view to emphasizing the time criticality of some of these applications.
- A summary of the overall landscape and proposals using DRL for AoI minimization in UAV-assisted IoT applications with their RL architectural setup.
- A classification of the proposals using DRL within the framework of AoI-minimization for UAV-assisted IoT into three categories: policy-based, value-based, and actor-critic based on the DRL algorithms used for problem-solving.
- A summary of literature based on the target objective (trajectory, energy efficiency, scheduling) with the benchmark algorithms for each proposal.
- A summary of algorithm complexities, simulation libraries, and some derived lessons are provided.
- An exposition into challenges and future research considerations within the framework of DRL for AoI-minimization in UAV-assisted IoT.

TABLE 1. List of abbreviations.

Acronym	Meaning
AoI	Age of Information
AP	Affinity Propagation
A2C	Advantage Actor critic
A3C	Asynchronous Advantage Actor critic
BS	Base station
CH	Cluster head
CP	Clustering point
CSI	Channel State Information
D3QN	Deep double Q network
DC	Data center
DDPG	Deep Deterministic Policy Gradient
DPG	Deterministic Policy Gradient
DQN	Deep Q Network
DRL	Deep Reinforcement Learning
EH	Energy Harvesting
ESA	Expected sum AoI
FDRL	Federated Deep Reinforcement Learning
GSDRL	Guided Search DRL
IIoT	Industrial Internet of Things
IoT	Internet of Things
IoTD	Internet of Things Devices
LOS	Line-of-Sight
LPWAN	Low Power Wide Area Network
MDP	Markov Decision Process
MADRL	Multi-Agent DRL
POMDP	Partially Observable Markov Decision Process
ML	Machine learning
NWAoI	Normalized Weighted sum of AoI
QoS	Quality of Service
PPO	Proximal Policy Optimization
RF	Radio Frequency
RIS	Re-configurable Intelligent Surface
RL	Reinforcement Learning
SAC	Soft Actor Critic
SARSA	State–action–reward–state–action
SNR	Signal to noise ratio
TD3	Twin-delayed deterministic policy gradient
TRPO	Trust Region Policy Optimization
UAV	Unmanned Aerial Vehicle
VDN	Value Decomposition Networks
VPO	Variational Policy Optimization
WPT	Wireless Power Transfer
WSN	Wireless Sensor Networks

D. PAPER ORGANIZATION

This review follows a general to specific organization style where the paper is first positioned with respect to existing literature, background technical information is first provided, then the reviewed papers are summarized and details are provided on their problem formulation, algorithms proposed to solved the problems, simulation, complexity as well as challenges and open research opportunities. Specifically, this survey paper has the following structure: Section II-B presents several surveys that explore the DRL applications in UAVs environments, Section III highlights several common UAV-assisted IoT applications, Section IV outlines the common DRL algorithms, Sections V and VI presents recent research works related to AoI minimization using DRL in single and multi-agents architectural environments respectively, Section VII provides summary of critical issues including target objectives, algorithm complexities and parameters

TABLE 2. Summary of review papers on Deep Reinforcement Learning (DRL) implementation in various domains.

Ref.	Focus	Description
[25]	Survey on DRL and its verification methods.	This paper focuses on DRL's achievement of complex task mastery through agent training, highlighting the challenges of robustness, safety, and verification methods in real-world contexts, and the emergence of a new DRL subfield to tackle these challenges.
[26]	Survey on blockchain-ML applications in Industrial IoT (IIoT).	This research assesses the use of blockchain and machine learning (ML) in IIoT, focusing on consensus, storage, and communication, identifying security and privacy risks from a ML standpoint. The findings present practical IIoT blockchain solutions and point out areas for future research.
[27]	Survey on DRL-based practical solutions for electric power system control.	This paper reviews DRL-based research in electric power system control with emphasis on electric power system operating states and control levels addressing issues like cyber-security, data analysis, and load forecasting.
[28]	Survey on DRL algorithms for task offloading in MEC-based AANs.	The paper investigates the use of AANs for addressing IoT computation offloading challenges in areas lacking traditional computing infrastructure. It emphasizes DRL for efficient resource management in MEC services and task offloading in MEC-based AANs.
[29]	Survey on RL and DRL methods within wireless IoT environments.	The paper discusses challenges in integrating IoT devices into networks and highlights the use of RL and DRL for dynamic decision-making. It explores these applications in wireless IoT for tasks such as routing and energy management.
[30]	Survey on effective Blockchain-AI integration in ETS.	The article explores the integration of Blockchain (BC) and artificial intelligence in energy trading systems (ETS). It discusses the transaction security provided by the BC's consensus algorithms and forecasting and data analytics abilities by the AI's DRL algorithms.
[31]	Survey on DRL-based methods for smart robotics advancement.	The paper examines motion-planning policies for mobile robots, emphasizing DRL implementation in unstructured environments. It categorizes traditional DRL approaches, explores their theories and applications, discussesrecent DRL developments, and proposes research directions to enhance mobile robot motion-planning algorithms.
[32]	Survey on DRL-based on Job-Shop scheduling problems.	The paper discusses complex optimization scheduling problems in manufacturing and transport, often addressed through Job-Shop scheduling (JSS) and evolutionary algorithms. It presents a novel DRL architecture for JSS optimization and highlights its AI-transformative potential.
[33]	Survey on zero-shot generalization in DRL.	This survey focuses on creating RL algorithms that adapt to new scenarios at deployment, for real-world applications with diverse and unpredictable environments.
[34]	Survey on FDRL technique-based on vehicular networks.	The paper discusses 5G vehicular networks for intelligent transportation, addressing dynamic QoS needs and safety-critical systems. It suggests DRL as a solution for resource allocation.

setting in the discussed works, Section VIII presents some of the challenges identified from the studies and proposes opportunities in this research area. Finally, Section IX concludes the paper. Acronyms and their meanings are provided in Table 1 while a brief overview of the paper structure is provided in Figure 1 for ease of navigation.

III. UAV-ASSISTED IOT APPLICATIONS

In this section, we highlight several UAV-assisted IoT applications published in the literature. This is categorized into monitoring, including industrial and structural as well as environmental monitoring, smart city, data gathering, security, health, agriculture and disaster management applications. Fig. 2 provides a visual summary of some of these applications.

A. MONITORING APPLICATIONS

Integrated systems based on wireless sensor networks (WSNs) and unmanned aerial vehicles (UAVs) with electric propulsion are emerging as state-of-the-art solutions for large scale monitoring [35]. Application references are highlighted in various domains such as environmental, agriculture, emergency situations and homeland security. UAV-assisted IoT-based monitoring has been extensively studied in the literature [36] especially for large-scale data collection applications [37] and many more. In [38], multi-UAVs network architecture has also been considered.

1) INDUSTRIAL AND STRUCTURAL MONITORING

UAVs have huge potential as sensing tools in the industry because they can be used to proactively address many problems. For instance, they can be used to facilitate decision making and quantify production. They provide a very consistent technological solution for event monitoring and cost-saving data collection [39]. UAV assisted IoT has diverse applications in industrial and structural monitoring. In this respect, the authors in [40] deployed a pre-programmed drone as a surveillance gadget to monitor illegal electrical connections and terminate detected lines.

The authors in [39] propose a smart monitoring and control system which integrates UAV into an industrial control mechanism via an IoT gateway. Photos taken by the UAV are computed in the cloud instantaneously and in a systematic manner. In other words, the UAV performs visual supervision and the service is integrated into the control loop in an industrial concrete plant. The results indicate that it is feasible for efficient and reliable system operations useful for reducing waste and improving quality.

UAVs and WSNs are advantageous in bridge health monitoring. Bridges are exposed to different forms of damage after construction. Thus, it is important to perform qualitative bridge maintenance to improve the lifetime of bridges and their serviceability (thereby saving lives). In this regard, bridge inspection is fundamental [41]. WSNs have been identified as a very good alternative to visual bridge inspection, however, it has its associated challenges such as

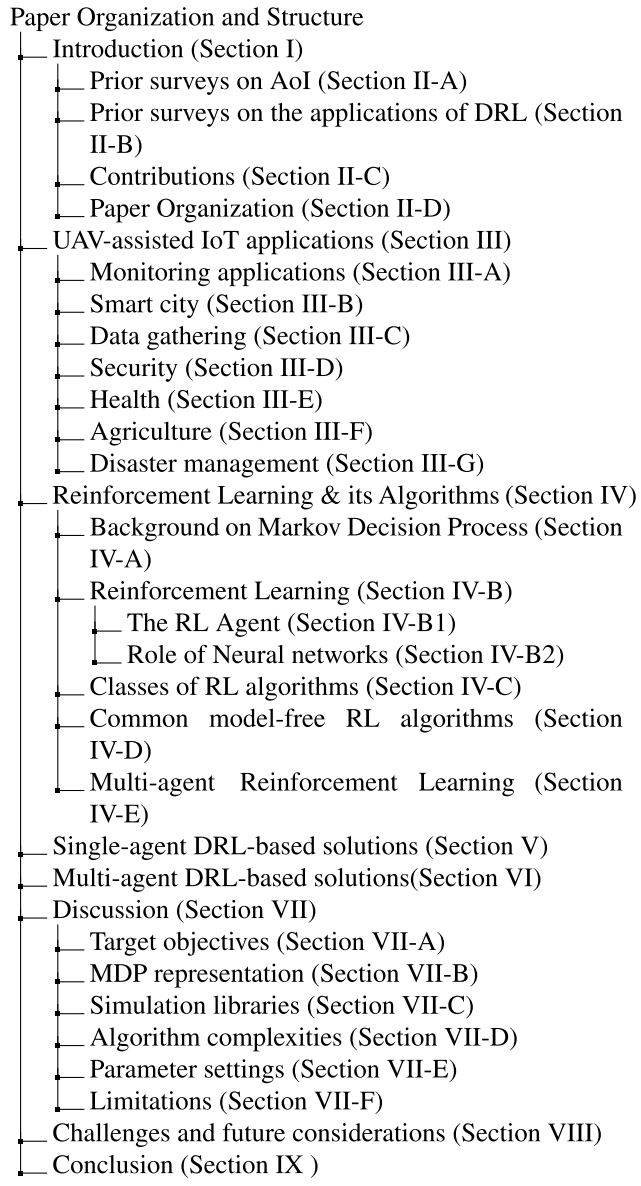

FIGURE 1. Organization of this paper.

battery failure. Hence, using UAVs can help to overcome these challenges for bridge health monitoring [41].

2) ENVIRONMENTAL MONITORING

Massive IoT networks facilitate a wide variety of real-time applications requiring local decision making and remote monitoring and control via sensors [42]. IoT devices are deployed to observe physical phenomena, such as temperature, pollution and humidity levels [2]. Air pollution detection is a typical environmental monitoring application of UAV-IoT which affects both our natural environment and our health [43]. Sensors can be placed in strategic areas to perform such monitoring tasks in smart cities effectively. In [44], the data acquisition system is incorporated in UAVs within the edge computing and IoT framework for early forest fire detection. An integrated self-configured UAV-WSN architecture can also be used to facilitate the a scale acquisition of environmental data [45]. UAVs can also be temporarily used to monitor marine environments rather than using close-range base stations or satellites which are costly in terms of infrastructure [46].

Furthermore, UAV-assisted IoT is promising for capturing greenhouse gas emissions [47]. Drone-enabled IoT relays facilitate high speed data collection for remote environmental monitoring [48]. UAVs can also be used for air quality monitoring. For instance, toxic gas detection sensor array can be mounted on UAVs [49] for accurate localization and better data monitoring. Environmental pollution and air quality monitoring is not only needed in towns and urban areas but also in remote and rural areas. Such solutions have been proven effective and economical in [50] in which Long Range (LoRA) mounted drone has been deployed that allows the operator to control the position of the drone to take measurements with the results are displayed in real-time on a web application. Such air monitoring solutions can also be infused into the smart city framework whereby processed data is displayed on a mobile device or a computer as seen in [43].

Another aspect of the environment that can be effectively monitored using UAV-assisted IoT is the detection of wildfires. Particularly, the occurrences of a wildfire has become frequent in certain part of the world and are usually severe such as bushfire incidents in Australia. This makes wildfire management and detection receive much attention. There are several ways by which wildfire can be detected but many of them have their shortcomings. For instance, the use of satellite imaging and remote-cameras result in late detection as well as poor reliability [51]. However, UAV-assisted IoT wildfire detection solution can be used to improve and optimize the detection probability even when there are constraints in terms of cost. IoT devices detect fires and report to UAVs which offer more reliable wildfire detection as compared to satellite imaging.

B. SMART CITY

UAV plays a major role in the development of IoT for smart environment and smart city applications [44]. The use of UAV-assisted IoT in smart city can take various forms. For instance, in an IoT Low Power Wide Area Network (LPWAN) deployment for smart city applications, UAVs can be deployed to collect data on residential's energy consumption which is useful in remote and rural areas [52]. As mentioned earlier, air quality monitoring can also be fused into smart city framework which include real-time monitoring and aerial-ground sensing, such as the one deployed in Peking and Xidian University in China [53].

C. DATA GATHERING

Data gathering is the most studied application of UAV-assisted WSN/IoT based architecture. Most of the related literature on age minimization had focused on this application

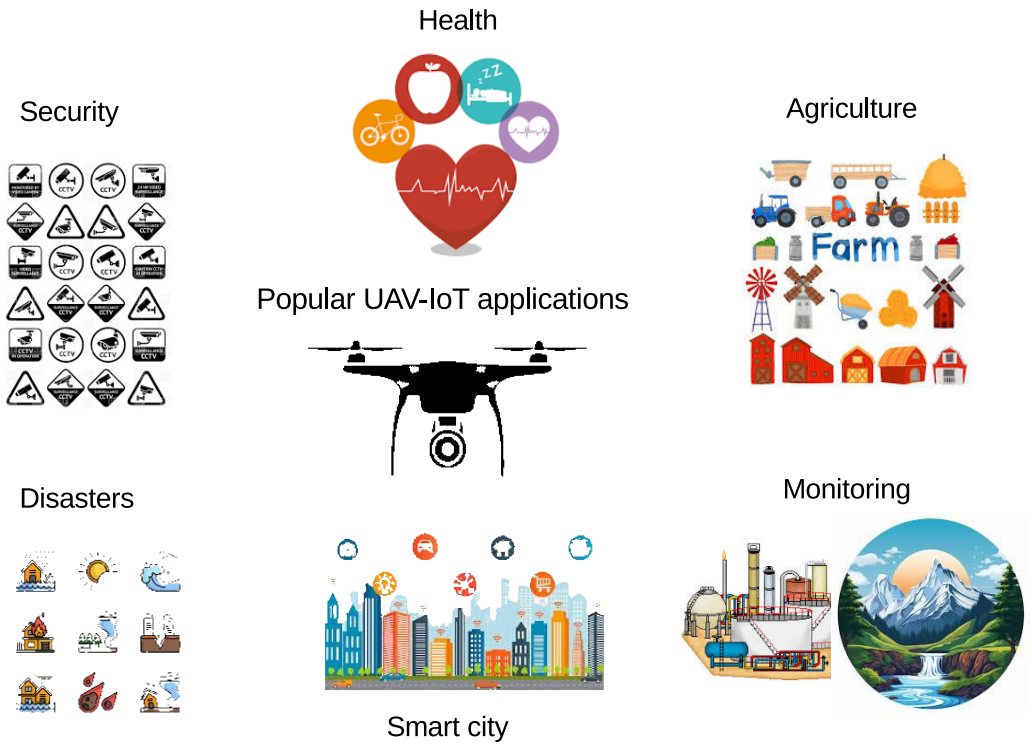

FIGURE 2. Popular UAV-assisted IoT applications.

and identified timely and safe data collection as the crucial requirements of UAV-WSN operation. There have been several studies on the deployment of UAV-WSN for data gathering with their unique peculiarities [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68]. In order to further motivate the applications of UAV-IoT-based data gathering, it is important to note that some of these works are aimed at addressing unique problems. For instance, [59] aims to reduce cost of multi-hop WSN transmissions via a mobile data collector and UAV. Reference [63] aims at addressing problems related to lack of real-time data, [62], [65] aim at reducing energy consumption while [67] aims to reduce congestion of several SN concurrent transmissions to UAV. The work by [62] has shown that data gathering in WSNs using UAVs provides an efficient solution in regards to energy considerations.

D. SECURITY

UAV-assisted WSN/IoT also plays a significant role in security. For example, drones have been extensively studied as surveillance devices in smart cities [69]. Particularly, quadcopter drones can be used as for border surveillance [70]. Hence, public safety applications would largely benefit from drones [71]. The use of UAVs has been studied and implemented as a cost-effective security system in IoT. Similarly, an IoT-based drone surveillance for industrial security applications has been developed [72]. A swarm of UAVs can also be deployed to perform search missions [73]. UAVs can also be deployed in autonomous indoor flight operations [74], [75], [76]. They have also proven useful for crowd surveillance [77] as well as IoT emergency communications [78].

E. HEALTH

One major use of drones in recent years has been in the area of health related. Specifically, the COVID-19 pandemic had increased researchers interest in drones, specifically for delivering drugs, medical consumables and equipments. Drones could also have been used to disinfect surroundings and verifty the conformance of the public with social distancing regulations. In [79], the role of UAV-assisted IoTs in managing the COVID-19 impact was extensively studied.

F. AGRICULTURE

Agriculture is indeed another popular application of UAVs. UAV-assisted IoT has a huge potential in agricultural applications and has been widely studied [80], [81], [82], [83], [84], [85], in particular for spraying pesticides [86], as a management platform [87] and as an intelligent framework for precision farming [88]. Moreover, the IoT-based edge UAV swarms have been used for distributed aerial processing in smart farming [89]. This architecture has few unique advantages including cost savings [90], [91]. In the context of UAV-assisted IoT, several critical issues have also been studied such as UAV path optimization [92], and the use of remote sensing drones as mobile gateways in precision

agriculture [93]. Also, the integration of IoT and drones for monitoring orchards from pests [94] and smart irrigation and real-time field monitoring [95]. Drones-assisted IoT was not only used to monitor crops [96] but also in animal [97] as well as pest and diseases [98] monitoring. In summary, UAV-IoT is promising in agricultural monitoring applications [99] and this includes many sub aspects of agriculture including soil monitoring [100] as well as wildlife [101] and water quality monitoring [102].

G. DISASTER MANAGEMENT

The widespread deployment of UAVs in ad hoc environments is attributed to their ability to provide dynamic solutions, such as in search and rescue operations [103]. UAV-assisted WSN plays a significant role in large-scale monitoring applications including disaster management [37]. Communication in pre-disaster scenarios is easy since infrastructure is fully operational [104]. In disaster management, IoT has become an integral part of data exchange [105]. It plays a vital role in the effective detection and management of natural disasters as multiple SNs are deployed in a large area to observe physical processes [42]. The role of SNs in disaster management include to sense, collect, process, and send data to server via sinks [105]. Sensors can be used in several applications, especially in hard-to-reach places, and disaster and post-disaster scenarios. For instance, sensors can be used for searching, locating, and rescuing survivors in disasters [106]. The use of UAV-assisted IoT has been extensively studied for disasters (see [107], [108]), rescue operations [109], [110] and emergency services [111].

IV. REINFORCEMENT LEARNING AND ITS ALGORITHMS

A. BACKGROUND ON MARKOV DECISION PROCESS

Markov decision process (MDP) is a mathematical framework used for modeling decision-making problems in which the outcomes are partly random and controllable. Thus, any decision-making problems can leverage on MDPs, including those adopting reinforcement learning-based solutions since the outcomes are under the control of an agent that makes the decision and are partly random in nature [20]. Optimization problems which can be solved via dynamic programming and reinforcement learning find MDPs framework very useful.

MDP framework is defined by a set of states, S, action, A, transition probability, p (from state s to next state s' after the execution of action, a), and the immediate reward, r, obtained after the action is performed. For short it is represented by a tuple representing (S, A, p, r) [20].

In several cases, state measurements are partially observable, which motivates the use of partially observable MDPs(POMDPs), a generalized MDP. For POMDPs-based problems, the agents do not fully observe the states and hence, they maintain a probability distribution of states based on observations and observation probabilities [112]. Thus, POMDPs-based problems involve two additional components in comparison to MDPs, i.e., a set of observations, and observation probabilities [112]. In the next section, RL is explained.[1,2]

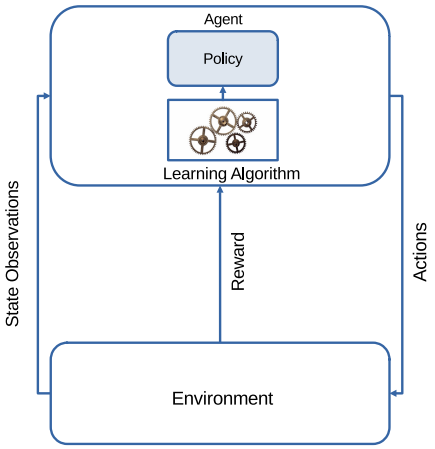

FIGURE 3. Typical RL architecture.

B. REINFORCEMENT LEARNING

Reinforcement learning (RL) is a machine learning training method in which an agent makes decisions and takes actions, observes the results of the actions, then adjusts the actions to achieve optimal results. RL basically consists of an agent and an environment. Since the agent acts on the environment, it requires some feedback on how well it is acting which is achieved via either a reward or penalty provided by the environment (refer Fig. 3). Actions have effects on the environment and make the environment change states. Based on the feedback from the environment, either reward as positive reinforcement or penalty as negative reinforcement, the agent makes intelligent decisions via this learning or training process.

With the aforementioned, an agent learns its environment independently by taking random actions and observing the reward or penalty over a long period of time to determine which action provides it with the most reward. However, since it needs to explore all the possible states, its actions can lead to exploring unnecessary areas (that is areas with low reward). Hence the agent's learning process become less efficient and time consuming. In situations whereby the model of the environment is not known beforehand, these model-free RL algorithms learn directly from experience or trial-and-error and use the feedback they receive to update their internal policies or value functions. Model-free RL algorithms are known useful for solving complex problems. On the other hand, if some information about the environment or underlying process is known beforehand, model-based

[1] https://www.mathworks.com/videos/reinforcement-learning-part-2-understanding-the-environment-and-rewards-1551976590603.html

[2] https://www.mathworks.com/videos/reinforcement-learning-part-3-policies-and-learning-algorithms-1554395009678.html

RL approach is preferred. Model-based RL algorithms learn the dynamics of the environment from experience and use the learned model to predict the outcomes of actions. This approach takes a shorter time since not all areas of the state space have to be explored unnecessarily.

1) THE RL AGENT

An agent is mainly composed of two interdependent components namely policy and a learning algorithm. To choose an effective policy, the nature of the underlying environment should be taken into consideration. The policy function takes the state as input and outputs the actions which can be represented by a state-to-action mapping (via Q-function), usually in tabular form with states as rows and actions as columns, also known as the Q-table. Using this table, the quality (Q) of the agent's decisions can be activated based on the desired objectives.

In many cases, there may be a very large number of possible actions taken by the agent, which may be difficult to capture in a table. In such cases, a continuous function is used instead. However, this makes it particularly difficult for us humans to learn the right parameters to predict the nature of this function accurately ahead of time, especially for highly complex or high-degree systems. Thus, a global function approximator that can handle continuous state-action spaces without needing a priori knowledge of structure is required. This motivates the use of neural networks, especially when it is impractical to use tables for a huge state-action space.

2) ROLE OF NEURAL NETWORKS

Neural networks constitute the basic "deep" component of deep reinforcement learning. A neural network is a group of nodes, or artificial neurons, connected in a way that allows them to be functional as a universal function approximator. Neural networks are used to represent policy in the agent. Given the right combination of nodes and connections (and weights), neural networks can mimic any input-output relationship. We want a neural function that can approximate complex functions that are difficult to solve; thus, it is important to make choices to ensure the neural network accurately captures the complexity of the problem at hand without making it overly complex, thus making training difficult. Since the neural network represents the brain of the intelligent agent, it is important to identify the classes of RL i.e., policy function-based, value function-based, and actor-critic.

C. CLASSES OF RL ALGORITHMS

1) VALUE FUNCTION-BASED ALGORITHMS

In value function-based learning, a function takes a state and a possible action as input and outputs the value of taking that action. This value is the some of the total discounted rewards from that state, so the policy (looks into the future), checks the value of every action and chooses the action with the highest value. In other words, the function criticizes the choices of the agent as it looks at its possible action (critic). Value in this case is beyond the instant reward from an action, it is the maximum expected return in the future.

The agent learns these values as it takes random actions, gets into a new state and collects the reward. Then the value of the action from that state (quality) is updated based on the reward using Bellman's Equation.

Bellman's Equation: Bellman's Equation helps the agent to solve the Q-table over time since it breaks up the problem into simpler steps rather than solving the value of the state-action pair in one step via dynamic programming. In Bellman's Equation, the value of a state-action pair is compared to what is in the Q-table to obtain the error (inaccuracy in prediction). The error is multiplied by the learning rate and the resulting * value is added to the old estimate. If the agent finds itself in the same state at a different time it would update the value and tweak it when it chooses the same action. This it done repeatedly until the true value of every state-action pair is sufficiently determined to exploit the optimal path.

Value functions can handle continuous state space without a lookup table (i.e. using a neural network). In this case, the state observation and action are provided as input and the neural network returns a value. However, for an infinite action space, the use of the policy to check all possible actions will be impossible. Using a neural network initial values can be random and then the learning algorithm uses the Bellman's Equation (or its variant) to determine the new value and update weights and biases in the network correspondingly. If enough state space has been explored by the agent, it can approximate the value function sufficiently well and select optimal action at any given state. For the on-policy value functions, the Bellman equations are represented as[3]

$$V^{\pi}(s) = \underset{\substack{a\sim\pi \\ s'\sim P}}{E} r(s,a) + \gamma V^{\pi}(s'), \tag{1}$$

$$Q^{\pi}(s,a) = \underset{s'\sim P}{E} r(s,a) + \gamma \underset{a'\sim\pi}{E} Q^{\pi}(s',a'), \tag{2}$$

in which $s' \sim P$ is used to represent $s' \sim P(\cdot|s,a)$, showing the next state s' is sampled from the transition rules of the environment; $a \sim \pi(\cdot|s)$ is shortened as $a \sim \pi$; while $a' \sim \pi(\cdot|s')$ is represented as $a' \sim \pi$ in short form.

As for the optimal value function, the Bellman equations are represented as[4]

$$V^{*}(s) = \max_{a} \underset{s'\sim P}{E} r(s,a) + \gamma V^{*}(s'), \tag{3}$$

$$Q^{*}(s,a) = r(s,a) + \gamma \max_{\substack{a' \\ s'\sim P}} Q^{*}(s',a') \tag{4}$$

One major difference between the Bellman equation for both on-policy and optimal value functions is the presence or absence of max over the actions. Thus, when the agent gets to choose its action is influenced by the inclusion or exclusion of the max. When included, it implies that the agent has to

[3]https://spinningup.openai.com/en/latest/spinningup/rl_intro.html
[4]https://spinningup.openai.com/en/latest/spinningup/rl_intro.html

choose the action that leads to the highest value whenever it chooses its action to act in an optimal fashion.

2) POLICY FUNCTION-BASED ALGORITHMS

Policy function-based neural network algorithms determine the agent's action. They can work with a stochastic policy in which the policy outputs a probability of taking a decision with exploration and exploitation probabilities factored in. The probabilities are tuned towards a direction that produces more reward over time. So the agent acts a certain way, collects reward along the way and updates the network to increase the probability of actions that yields the best reward. However, using this method, the obtained result may lead to a local maximum as it uses the policy gradient method.

3) ACTOR-CRITIC FUNCTION-BASED ALGORITHMS

Since action space needs to be small to appreciate the value policy-based method, a combination of the actor (network tries to take what it thinks it is the best action at a current state) and critic then estimates the value of the state and action (taken by the actor), we can develop a solution for continuous action space since only a single action (taken by the actor) needs to be evaluated rather than trying to find the best action by evaluating all actions.

An actor chooses an action (policy function-based) that is applied to the environment. The critic estimates what it thinks is the value of the state and action pair and then uses the reward to determine how accurate its prediction was by determining the difference between the new estimated value of the previous state and the old values of the previous state from the critic network. The new estimated value is based on the received reward and discounted value of the current state. This is used to know whether things went better than expected or not. This error is used by the critic to update itself in the manner that the value function would, to make a better prediction. So the actor updates itself based on the response from the critic and error term to improve its probability of taking actions in the future. This way action and critic networks are combined together to learn the optimal behaviour. The actor learns the right action using the critic feedback. The critic learns the value function from the reward to properly criticise the action of the actor. Using the actor-critic method the best features of both policy and value function algorithms can be taken advantage of and both continuous state and action spaces can be handled by the actor-critic and the learning process is faster when the returned reward has high variance.

D. COMMON MODEL-FREE RL ALGORITHMS

In this section, a brief description of common model-free RL algorithms is provided[5] (refer Fig. 4).

1) VPO

Policy gradient algorithm functions by updating policy parameters using stochastic gradient ascent on the policy

[5]https://spinningup.openai.com/en/latest/spinningup/rl_intro2.html

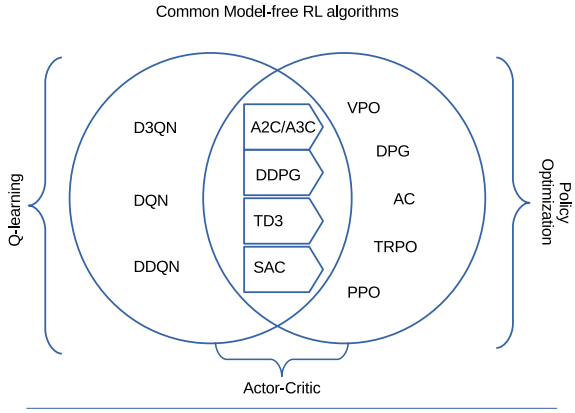

FIGURE 4. Common model-free RL algorithms using Q-learning, policy and actor-critic functions.

performance (see Equation 5 below)

$$\theta_{k+1} = \theta_k + \alpha \nabla_\theta J(\pi_{\theta_k}) \tag{5}$$

They compute advantage function estimates using infinite-horizon discounted return while using finite-horizon discounted policy gradient formula. Exploration involves taking random actions depending on the initial conditions and training procedure and as time goes by the policy begins to exploit rewards that it has found. However, the policy may still get trapped in a local optimum.

2) TRPO

TRPO updates policies by taking the biggest possible step for improving performance while satisfying the closeness constraint on the old and new policies (i.e. how close they are allowed to be). This constraint is expressed in a form that can be related to (but not exactly) the distance between probability distributions. TRPO avoids the problem in vanilla policy gradients by which new and old policies are close within the parameter space and thus large step sizes pose risks.

3) DDPG

DDPG is an off-policy algorithm that can be used strictly in environments characterized by continuous action spaces. It basically involves learning a Q-function and learning a policy. It uses Bellman's equation to describe the optimal action-value function, i.e., as a function approximator for the Q-function.

4) TD3

Although DDPG does a good job, sometimes it fails when the learned Q-function begins to overestimate the Q-values which breaks the policy as a result of exploiting errors in the Q-function. Twin Delayed DDPG can address this problem by learning two Q-functions instead of one. The smaller of the two Q-values from the target in Bellman's error loss function. It also ensures the policy is not updated as frequently, i.e., once every two Q-function updates. Finally, it performs target

policy smoothing, in which it adds noise to the target action to make it more difficult for the policy to exploit errors in the Q-functions by smoothing out Q over changes in action.

5) PPO

In PPO, the motivation is similar to that of TRPO in which they are concerned about taking the largest possible step for policy improvement with the available data while preventing an accidental performance breakdown. As opposed to TRPO which uses a complex second-order method, PPO uses first-order methods while using some techniques to ensure the new policies are close enough to the old ones. They are easier to implement and can perform as well as TRPO. They can be used in environments with discrete or continuous action spaces

6) SAC

Soft Actor Critic (SAC) optimizes a stochastic policy in an off-policy manner thus integrating both stochastic policy optimization and DDPG-type approaches. The policy is trained to obtain the best trade-off between expected return and randomness in the policy (entropy), thus, mimicking the exploration-exploitation trade-off since when the entropy is increased, more exploration takes place and learning is accelerated. Similarly, it can prevent premature convergence.

E. MULTI-AGENT REINFORCEMENT LEARNING

This section provides a simple but intuitive background on multi-agent RL systems.[6] Multi-agent system involves multiple agents sharing a common environment (refer Fig. 5) such as a swarm of UAVs that operate in a formation or several autonomous vehicles driving through the same intersection. They could also be distributed controllers that aim at accomplishing a common goal such as several smart homes trying to schedule power. All these are examples of cooperative agents since they work together to achieve a common goal. In other cases, agents could aim at maximizing their own personal goals (benefits) while minimizing those of other agents (adversarial). It is possible to have systems with both cooperative and adversarial multi-agent systems.

In designing multi-agent systems, it is important to clearly define how agents should perform actions or coordinate themselves. Sometimes it is better for agents to learn some of its behaviour on their own without a shared reward function, while in other cases, it is preferred that there is a common reward function for all agents. This decision has its own trade-offs, collaboration is facilitated by a shared reward, while agents might be lazy (to learn and earn more rewards) as compared to an unshared reward on the other hand a localized reward breeds or intensifies competition (for the limited) at the expense of the potential to earn more rewards (by the agent) [113].

[6]https://www.mathworks.com/videos/an-introduction-to-multi-agent-reinforcement-learning-1657699091457.html

For single-agent RL, the goal is to update the agents' policy as time goes by to maximize the reward. However, for multi-agent RL, multiple agents interact with an environment and each of those agents uses an RL mechanism to update their policy over time.

Introducing multiple agents that learn and interact with each other brings new challenges. For a decentralized learning architecture in which each agent is trained independently from the others, each agent tries to accumulate the highest reward regardless of the actions of other agents. New information is shared between agents which is advantageous with respect to the minimum communication overhead involved and the simplicity of the system.

However, since the agent does not know how much of the overall objective has been achieved by other agents they can learn how to avoid repeating those actions that are not required (due to the actions of other agents). Decentralized architecture involves a trade-off between performance and complexity.

For performance, it would be much better if agents shared their experience (relative to the final objective). However, since all agents are still in the learning process their policies change which makes it difficult for other agents to sufficiently track/understand the dynamics of the environment as the environment becomes non-stationary, thus making the MDP time variant. RL algorithm expects a stationary environment. Thus, the agents may not converge to a solution as each agent continuously changes its policy. Although it is possible that after a large number of episodes, agents can learn to work together. However, such learning is not fully complete and also depends on their prior states especially if the agents have different characteristics i.e. not identical.

Centralized architecture (refer Fig. 6) requires some higher level process that collects agents' experiences and a policy is learnt using the pool of collected information The policy is learnt is then distributed back to the agents. This is meritorious, especially in the case whereby each agent is identical (i.e. they have the same observations in actions as a single optimal policy would suffice each of them) i.e. that will control their actions to achieve an overall aim. This also reduces the amount of learning that takes place as each agent learns from the experience of the other. A stationary environment has also been created for the agent, as all agents are considered as larger entities and know about the policies of others.

V. SINGLE AGENT DRL-BASED SOLUTIONS

In this section, a list of summaries of various studies in single UAV scenarios using model-free DRL algorithms to minimize AoI is presented (refer Table 3). The organization of this section is provided in Fig. 7.

In the next sections, works using various model-free DRL algorithms in single UAV environments (as well as single-agent) of value-based, policy-based and actor critic-based classes have been identified and discussed. Table 4 provides the list summary of the works using these DRL algorithms.

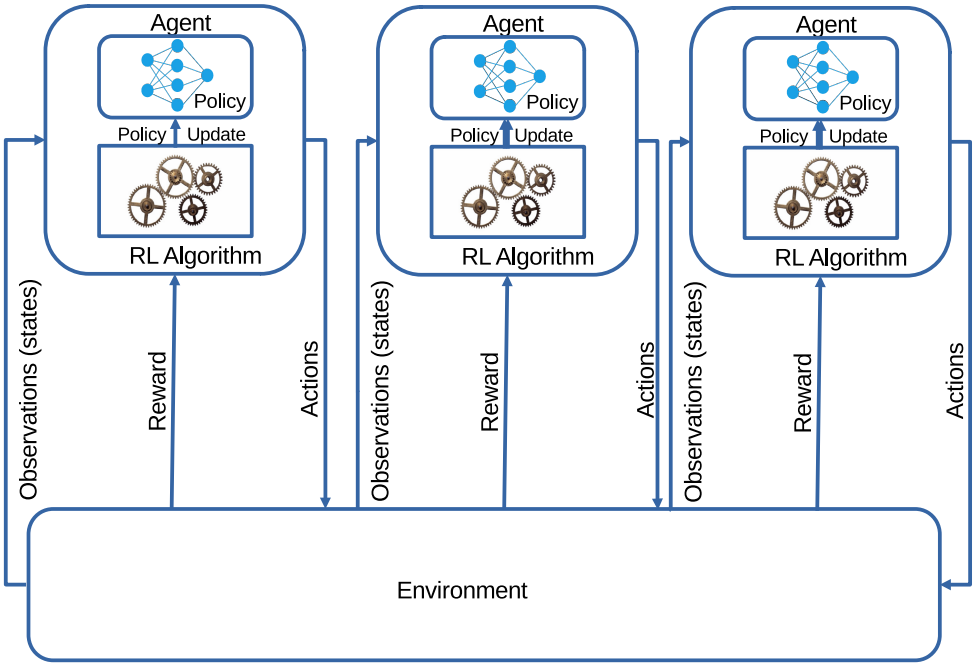

FIGURE 5. Typical MADRL architecture with 3 agents.

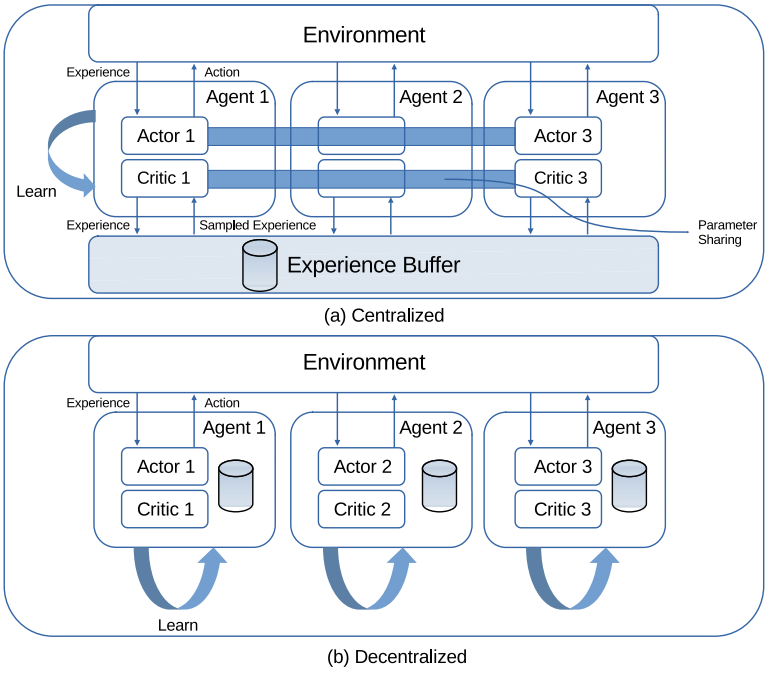

FIGURE 6. Centralized and Decentralized MADRL architectures.

Table 5 provides comparison among different DQN-based works using single agent based on the optimization objectives, metrics used as well as benchmarks.

A. DQN-BASED ALGORITHMS AND THEIR TARGET OBJECTIVES

1) DQN FOR OPTIMIZING TRAJECTORY

Considering the traffic patterns of IoT devices could be diverse, such as fire alarm sensor as compared to temperature sensor, the authors [3] study a scenario whereby UAV-assists ground nodes with unknown traffic generation patterns. The fact that IoT devices were assumed to transmit data with different traffic patterns makes AoI minimization more challenging, thus to maintain information freshness, the authors formulated the online AoI-optimized UAV trajectory planning problem as an MDP due to the complex association and interaction pattern between the UAV and IoT devices to find the optimal policy for efficient trajectory planning.

TABLE 3. Summary of DRL-based algorithms in single UAV scenarios and their functions.

Ref	Name	Function
[1]	NCRL using DQN	Neural combinatorial-based DRL using DQN for obtaining optimal UAV scheduling policy for status updates
[12]	SAC-AO-RIS	DRL-based soft actor-critic algorithm with prioritized recent experience replay used to learn a stable policy for UAV trajectory optimization and IoTD scheduling
[13]	DQN-based with experience reply memory	Deep Q-network-based trajectory design method with *state space of* energy efficiency, rest energy, and AoI and the efficient reward function for EE optimization
[15]	DQN-scheme	For minimizing average AoI of ground nodes via a joint optimization of UAV trajectory, information transmission scheduling and EH at ground nodes
[114]	DQN-scheme	For finding an asymptotically optimal policy for optimizing trajectory of UAV and minimize both SN average age of information and packet drop rates
[115]	D3QN	DRL to solve a formulated problem for energy-efficient fresh data collection in rechargeable UAV-assisted IoT networks
[116]	Double DQN	Jointly optimizing UAV flight trajectory and transmission scheduling sequence of sensors
[117]	GSDRL	Guided DRL to help UAV independently complete data collection and forwarding from different initial locations in a rapid fashion
[118]	PPO	PPO was used to solve the formulated mixed-integer convex optimization problem for optimizing UAV altitude, communication schedule and RIS phase shift in the absence of a priori knowledge of activation pattern of IoTD
[119]	TD3-AUTP	TD3-AUTP algorithm in addition to the introduction of DNN for feature extraction for jointly optimizing the UAV flight speed and hovering location as well as bandwidth allocation for data collection to minimize average weighted sum of expected average AoI, UAV propulsion energy and IoTD transmission energy
[120]	A3C	For real-time decision making within the DRL framework to achieve low UAV energy consumption and minimize the AoI
[121]	DQN-scheme	For achieving optimal UAV flight trajectory and transmissions scheduling of SNs for minimizing the weighted sum AoI DRL
[122]	DRL	DRL for obtaining the optimal policy for scheduling status update packets while optimizing UAV trajectory to minimize weighted sum AoI
[3]	A-TP	DRL-based solution for solving the complex association interaction pattern between the UAV and IoT device for fresh data collection

TABLE 4. DRL-based algorithms for single UAV scenarios.

Model Free RL	Algorithm	References
Value-based	DQN-based	[1] [13] [15] [114] [122] [121] [123]
	DDQN	[116]
	D3QN	[115]
	AoI Trajectory Planning	[3]
Policy-based	PPO	[124] [125] [118]
Actor critic	SAC	[12]
	TD3	[119] [126]
	A3C	[120]
	DQN+Policy Gradient	[117]

The AoI was assumed to increase when the UAV does not visit a SN and it increases when the SN nodes are visited. This evolution of AoI for IoTDs can be represented as follows

$$A_{k,n} = \begin{cases} k\tau - \sum_{j=1}^{i_{k,n}} \delta[j], & C_{k,n} = 1 \\ A_{k-1,n} + \tau, & \text{otherwise.} \end{cases} \tag{6}$$

in which $\delta[j]$ is a random variable, $i_{k,n}$ is the index of the newest packet at IoT device n as of time slot k, $\mathcal{C}_k$ is the set of IoT devices covered by the UAV within time slot k.

They devised a DRL-based A-TP (AoI-based Trajectory Planning) algorithm which converges rapidly due to the adoption of a randomized policy for pre-training the deployed deep neural networks. Via extensive simulation, the authors show that the proposed algorithm can significantly reduce the AoI of the data collected by IoT devices and is robust even in the dynamic environment considered.

2) DQN-BASED ALGORITHMS FOR OPTIMIZING TRAJECTORY, ENERGY AND SCHEDULING PLANNING

The authors [15] study UAV-assisted wireless networks in which a dispatched UAV wirelessly charges multiple ground nodes using RF energy transfer, then the ground nodes deploy the harvested energy for uploading sensed data to the UAV. The authors formulate an optimisation problem to minimise the average AoI of UAVs via the joint optimisation of UAV trajectory, ground node information transmission, and energy harvesting scheduling. The formulated problem is a combinatorial optimisation problem with a set of binary variables, which makes solving it difficult. Thus, the authors reformulate the problem as a Markov process with a large state space, and a DQN was deployed to find a near-optimal solution using DRL. The authors constructed two networks, one for evaluating the reward accrued by an action performed in a current state and the other to predict realistic actions. The authors also study the impact of energy punishment in the reward function to save energy, however a trade-off was observed with respect to AoI. Similarly, the authors show the

- Single agent DRL-based solutions (Section V)
  - DQN-based algorithms and their target objectives (Section V-A)
    - DQN for optimizing trajectory (Section V-A1)
    - DQN-based algorithms for optimizing trajectory, energy and scheduling planning (Section V-A2)
    - DQN-based algorithms for optimizing trajectory and energy (Section V-A3)
    - DQN-based algorithms for optimizing trajectory and scheduling planning (Section V-A4)
    - DQN-based algorithm for optimizing trajectory (Section V-A5)
  - Policy-based DRL algorithms using single agent (Section V-B)
    - PPO-based algorithm for scheduling planning (Section V-B1)
  - Actor critic-based DRL algorithms using single agent (Section V-C)
    - Actor critic-based algorithms for optimizing trajectory and scheduling planning (Section V-C1)
    - Actor critic-based algorithm for optimizing other objective functions (Section V-C2)
  - Lessons learned (Section V-D)

FIGURE 7. Organization of Section V.

effect of packet size, transmit power, and GN distribution area on the ground node's AoI. The DQN algorithm was shown to converge, and an improvement in AoI was observed compared to other studied or baseline schemes.

The authors [115] investigate energy-efficient fresh data collection in rechargeable rotary wing UAV-assisted IoT in which a UAV leaves its initial location (with full energy) to collect data from SNs in IoT and should reach its final destination under a given time constraint. Particularly, a UAV is dispatched from a depot (with full energy from its initial location), flies over it to collect packets, and lands in its final position. While it is flying, the UAV (with an assumed circular coverage area) can harvest energy from charging stations with a maximum recharging distance in order to keep its battery level above the required threshold. The goal is to minimise the weighted sum of the average AoI and the average recharging price by jointly optimising the UAV trajectory, scheduling, and energy recharging while ensuring the remaining energy of the UAV should not be less than a threshold. The formulated optimisation problem was shown to be a constrained nonlinear integer programming problem that is difficult to computationally solve. Thus, they formulated the problem as a finite-horizon MDP, which was solved by a proposed duelling double deep q network, i.e., the d3qn algorithm for the UAV to learn its trajectory, and for scheduling, and energy recharging at each time slot. Simulation results show that, compared to baseline policies, the algorithm can reduce the weighted sum of AoI and average recharging price significantly.

TABLE 5. Comparison on optimization objectives (T=trajectory, E=energy, S = scheduling), metrics and benchmark algorithms using value-based DRL algorithms in single UAV scenarios.

Ref.	T	E	S	Metric	Studied Algorithms
[15]	✓	✓	✓	Average AoI	1.Proposed DQN-scheme 2.Random Walk 3.Energy-based 4.Greedy
[115]	✓	✓	✓	Average AoI	1.Proposed D3QN 2. Distance-based 3.AoI-based
[13]	✓	✓	✘	Average AoI	1.Proposed DQN with experience-reply memory 2. DQN without experience-reply memory
[123]	✓	✓	✘	Average AoI	1.Proposed DQN with reply memory 2.Greedy 3.DQN
[1]	✓	✘	✓	Average AoI	1.Proposed NCRL 2. Proposed LSTM-autoencoder 3.Discretized DQN 4.Weight-based
[121]	✓	✘	✓	Average AoI	1.Proposed DQN-scheme 2.AoI-based 3.Distance-based
[116]	✓	✘	✓	Average AoI	1.Proposed DDQN 2. DQN 3.Distance-based
[122]	✓	✘	✓	Weighted Sum AoI	1.Proposed DQN-based policy 2.Distance-based 3.Random walk policy
[3]	✓	✘	✘	AoI CDF	1.Proposed AoI-based Trajectory Planning 2. Round Robin scheduler 3. Age-based MaxWeight scheduling
[114]	✓	✘	✘	Average AoI & packet drop	1.Proposed DQN-scheme 2.Greedy 3.SARSA

3) DQN-BASED ALGORITHMS FOR OPTIMIZING TRAJECTORY AND ENERGY

DRL approach was deployed to achieve both data freshness and energy-efficiency optimization for multi UAV navigation in [123], in which the authors consider multiple UAV-BS for providing connectivity to IoT devices for improving information freshness. The authors formulate an energy-efficient trajectory optimization problem for maximizing energy efficiency via an optimal UAV-BS trajectory policy. To ensure data freshness at the ground BS, the authors incorporated energy and AoI constraints and propose an agile DRL with an experience replay model to solve the formulated problem. The state space is extremely large which makes the proposed solution appealing as finding the best trajectory policy is too complex for the UAV-BS. The trained model (using the proposed solution) is applied to achieve an effective real-time trajectory policy for the UAV-BS to capture network states over time. The proposed approach proves to be more energy efficient than the

baseline algorithm, greedy algorithm, and Deep Q network approaches.

Learning-assisted UAV data collection for achieving information freshness in IoT is the focus in [13]. Particularly, this is because guaranteeing information freshness of energy-limited UAVs can be very challenging. Thus, the authors study the trajectory design of a multi-UAV system with a large number of ground sensor devices that send information to the UAV-BS with constraints on information freshness. Thus, the paper considers the trajectory design of a single-antenna multi-UAV-enabled communication network which collects data from single-antenna IoT devices for maximising energy efficiency with constraints in the AoI. In the studied architecture, each ground device sends its information to its associated UAV. The authors first formulated an energy-efficiency maximization problem under safety distance, rest energy, and AoI constraints. The non-convex nature of the objective function and unknown dynamic environment (unknown space of the UAV trajectory) motivated the authors to propose the Deep Q network for efficient trajectory design using the architecture in Fig. 8. Here, the state space includes energy efficiency, AoI, and rest energy efficiency. Similarly, an efficient reward function for energy-efficiency maximization is adopted. The system is validated and simulation results show that the proposed scheme achieves a better energy efficiency performance compared with the benchmark scheme.

4) DQN-BASED ALGORITHMS FOR OPTIMIZING TRAJECTORY AND SCHEDULING PLANNING

The authors in [1] study a wireless network topology by which battery-limited UAVs flies around to collect status update packets from ground nodes which are observing some physical phenomenon. The authors formulate the problem of minimizing the normalized weighted sum AoI associated with the observed physical processes in which the AoI metric is described as

$$A_m(t) = A_m^{min} + t - t_{i-1,m}, \forall t \in [t_{i-1,m}, t_{i,m}) \quad i \in \{1, \ldots, n_m\} \tag{7}$$

in which A_m^{min}: The minimum value of AoI, $t_{i,m}$: The time instant at which node m transmit an update packet.

This problem includes two major components which are the optimizing the flight trajectory of the UAV as well as how update packet transmissions are scheduled. The problem was initially formulated as a mixed-integer programming problem. Then, for a given scheduling policy, the authors develop a convex optimization-based solution to determine the optimal trajectory and update. The formulated NWAoI minimization is provided as follows:

$$\bar{G}(\boldsymbol{t}_1, \ldots, \boldsymbol{t}_M) \triangleq \frac{1}{\tau^2} \sum_{m=1}^{M} \lambda_m \sum_{i=1}^{n_m+1} (t_{i,m} - t_{i-1,m})^2. \tag{8}$$

However, given the combinatorial nature of this problem, a finite horizon MDP with finite state and action spaces was formulated as a complement to the convex optimization. Finite-horizon dynamic programming could not be used in this case due to computational practicality and thus a neural combinatorial-based DRL is employed in view of the large state space of the MDP. The deep RL architecture is as shown in Fig. 9. This aims to obtain the optimal scheduling policy given the derived equations and constraints related to velocity, location and energy. For large-scale scenarios with several nodes, DQN cannot learn optimal scheduling, thus the authors proposed the use of a long short-term memory (LSTM)-based autoencoder to map the state space to a fixed-size vector representation. The proposed neural combinatorial based DRL significantly performs better than the baseline polices such as discretized state DQN and weight-based policies with regards to the achievable normalized weighted sum AoI per process.

The paper [121] focuses on addressing the challenge of balancing information freshness and energy consumption during UAV-based data collection from IoT devices. To achieve this the authors introduce an energy constraint taking into consideration both the energy consumption for communication and propulsion. Thus, the power consumption of the UAV is modelled as

$$\tilde{P}(V_t) = \gamma + P_1 \left(\sqrt{1 + \frac{V_t^4}{4v_0^4}} - \frac{V_t^2}{2v_0^2} \right)^{\frac{1}{2}} + \frac{1}{2} d_0 \rho s_0 A V_t^3, \tag{9}$$

where $\gamma = P_0 \left(1 + \frac{3V_t^2}{U_{tip}^2}\right) +$, P_0, P_1 is the blade profile power and derived power of the UAV in the hovering state, V_t is the velocity of the UAV at slot t, U_{tip} is the tip speed of the rotor blade of the UAV, v_0 is the mean rotor induced velocity in the hovering state, d_0 is the fuselage drag ratio, ρ is the density of air, s_0 is the rotor solidity, and A is the area of the rotor disk. The authors formulated the problem as MDP and deployed a DRL-based solution to minimize the weighted sum of AoI to identify the best trajectory route for the UAV as well as the optimal scheduling policy for SNs. Results show the superiority of the proposed DQN-based solution for UAV-assisted data collection as compared to the baseline schemes.

The authors in [116] deployed DDQN for achieving information freshness in UAV-assisted IoT. In the studied architecture, the BS sends the UAV to fly to SNs to collect data while ensuring the UAV energy is more than enough. Whenever UAV energy is low, it returns to the destination. Data sampled by SNs is stored in a buffer such that new packets replicate old ones before the UAV collects them. In this case, the AoI is determined by data sampling of each SN, the queuing waiting time, and the UAV-aided transmission process. Thus, the sampling and queuing process of SN impacts the UAV's AoI optima trajectory significantly, which has rarely been captured in prior studies. The authors modelled the problem as MDP and defined the state space, action space, and reward function. The authors jointly optimize the flight trajectory of the UAV and the transmission scheduling sequence of SNs. To overcome

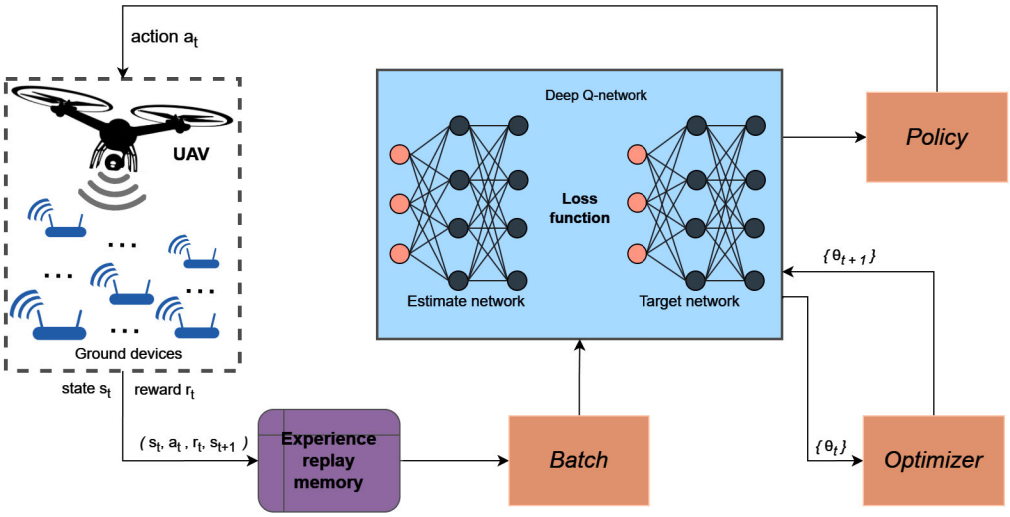

FIGURE 8. DRL Framework for UAV's Trajectory [13].

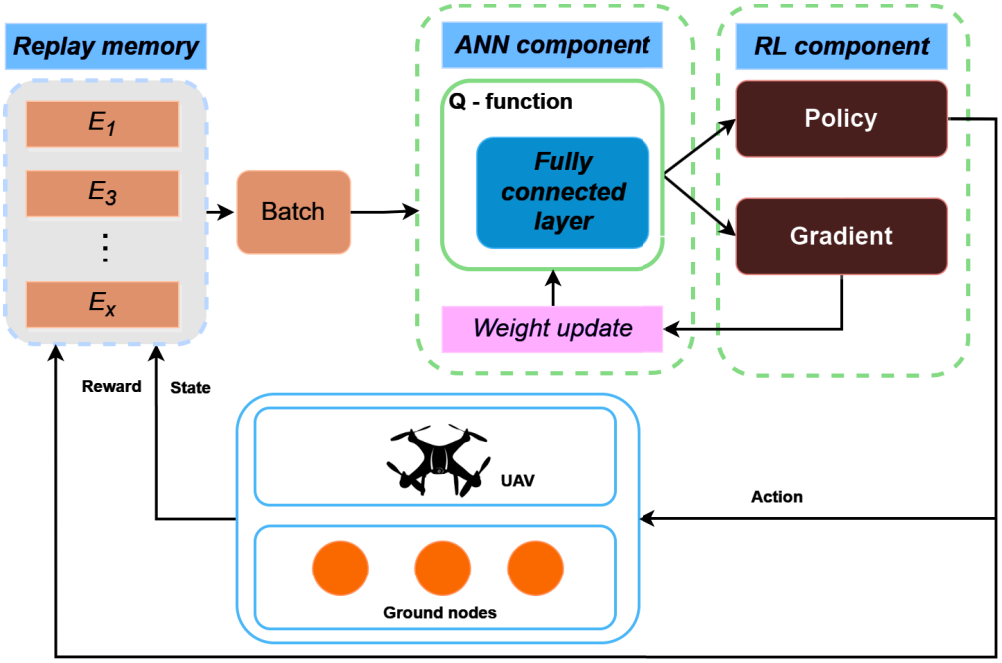

FIGURE 9. Typical Example of the DRL Architecture [1], [122].

the dimension disaster, the authors propose a node data collection algorithm based on double deep Q learning. The proposed algorithm was compared to other algorithms, and the authors show the system's performance under different sampling strategies. Via a large number of simulation experiments, the authors show that the proposed DDQN that the proposed algorithm can improve UAV AoI compared to the baseline schemes and reduce the packet loss rate of SNs.

The authors [122] focus on jointly optimizing UAV trajectory and the efficient scheduling of energy constrained ground nodes' status update packets with the objective of minimizing the weighted sum AoI. The authors formulate the problem as a MDP with finite state and action spaces then developed a DRL algorithm used to overcome the problem of dimensionality associated with the formulated problem. The algorithm consists of ANN for state space dimension reduction and the RL component for optimization of the policy. Abstracting different physical processes, results show that the DRL-based approach significantly performs better than the baseline policies such as random walk and distance-based policies.

5) DQN-BASED ALGORITHM FOR OPTIMIZING TRAJECTORY

The authors [114] study the age-optimal data collection problem for energy-constrained-UAV-aided IoT in which data is sampled either in a random or fixed manner. Thus, in the studied scenario, the authors jointly consider data sampling, queueing, and UAV-aided relaying. This paper assumes a sample and replace policy to update packets in the buffer (i.e., new data replaces old ones). To obtain the age-optimal trajectory, the authors formulate the problem as a finite-horizon MDP taking buffer management into consideration with the objective of minimising the weighted sum AoI, packet drop rate, and UAV energy consumption. Due to the ''curse of dimensionality'', the authors propose a DRL (in this case, DQN-based learning) algorithm to design the age-optimal trajectory for the UAV while also keeping the packet drop rate to the bearest minimum. Via simulation, they show that the proposed algorithm successfully reduces AoI and packet drop rate (better than SARSA policy) based on its experience having learned the network topology and SN sampling status.

B. POLICY-BASED DRL ALGORITHMS USING SINGLE AGENT

Table 6 provides comparison among different policy-based works using single agent based on the optimization objectives, metrics used as well as benchmarks.

1) PPO-BASED ALGORITHM FOR SCHEDULING PLANNING

DRL was deployed for achieving energy and AoI-efficient data collection in rechargeable UAV-aided IoT in [125]. The authors focus on optimizing the AoI in UAV-RIS assisted IoT network in which RIS mounted on UAVs are deployed for increasing the throughput capacity of the network by functioning as relays between the BS and IoTD. The signal transmission to IoTD will be affected by uncoordinated UAVs or RIS phase shift elements. The authors developed two model-free DRL approaches to minimize the average sum of AoI in the network. Particularly, off-policy DQN and on-policy PPO were used to solve the problem by optimizing the RIS phase shift, UAV-RIS location, and IoTD scheduling jointly. The stability and convergence of the two algorithms were evaluated and the result showed that the on-policy approach, i.e., PPO performed better than DQN (off-policy) with respect to stability, and convergence speed under different environment settings.

The authors in [124] focus on scheduling IoT devices and dynamic UAV altitude control. The overall objective was to minimize the expected weighted sum AoI of IoT devices' sampled data, which depends on the channel conditions. The studied network model thus incorporates the reliability of the wireless channel and provides analytical characterization. The problem is formulated as a mixed integer non-linear programming problem; thus, linear and dynamic programming cannot address this. Thus, the problem is formulated as an MDP and the authors deploy an agent on the UAV to make decisions at each time slot based on the environmental dynamics it learns for obtaining the optimal altitude. Particularly, higher altitude improves the LoS links at the expense of weaker received signals due to higher path loss. The hybrid discrete-continuous action space and tight coupling of altitude and scheduling make the problem challenging. Thus the authors deploy online DRL with PPO to effectively solve the formulated problem

A solution towards the integration of UAV and RIS focuses on optimising the phase shift of RIS elements to improve different performance metrics [118]. Thus, the authors focus on studying learning-based IRS-assisted age-aware data collection in UAV-assisted IoT. The authors study the potential of RIS-assisted UAV-enabled data collection from IoT with its objective is to minimize the expected sum AoI by optimizing UAV altitude, communication schedule, and RIS phase shift. The network involves IoTD with limited transmission capabilities that collect data (a stochastic process), and sampled data is processed by the BS. The authors study a single RIS deployed as a passive relay to forward sampled data to the BS, considering the different IoTD activation patterns. If the SNR exceeds the required threshold, the data is sent to the BS while considering AoI constraint. The authors formulated the framework as an optimisation problem considering the SNR constraints, UAV altitude constraints, and IoTD scheduling constraints with the objective of minimising the expected sum AoI. Because the optimisation problem is quite challenging due to the unknown activation pattern of IoTD, the authors chose to apply PPO for solving the mixed integer non-convex optimization problem for learning the randomness of the IoTD activation pattern and controlling the UAV altitude, RIS element phase shift, and communication scheduling for minimising ESA. The authors developed two baseline policies to evaluate the effectiveness of the proposed algorithm: 1) random walk policy in which the IoTD is randomly selected to relay status update information along with the adjustment of the RIS phase to ensure a reflected signal can be added at the selected IoTD constructively while randomly changing the UAV altitude. Also, the other policy is hovering with a greedy policy, by which the UAV searches for the best height that satisfies the reliability constraint for most of the IoTD. The UAV then selects the IoTD with the maximum current AoI. The policies compared with the baseline are adequate, as the former policy exploits all the possible actions, which may result in selecting actions that decrease the AoI, while the latter policy always selects IoTD with a higher AoI to relay their status updates. The authors observe that the proposed algorithm can minimise ESA for a lower number of IoTD since each IoTD enjoys more frequent scheduling. However, for a large number of IoTDs, ESA increases as more scheduling is needed to decrease ESA. Hovering with a greedy policy performs better than the random walk policy since it selects IoTDs with the highest value. The proposed algorithm outperforms all the baseline algorithms as it can learn the activation pattern of IoTD and adjust UAV altitude.

TABLE 6. Comparison on optimization objectives (T=trajectory, E=energy, S = scheduling), metrics and benchmark algorithms using policy-based DRL algorithms in single UAV scenarios.

Ref.	T	E	S	Others	Metric	Studied Algorithms
[124]	✘	✘	✓	UAV altitude	Expected Weighted Sum AoI	1.Proposed PPO 2. Random Deployment with Random Scheduling (RDRS) 3.Heuristic Deployment with Greedy Scheduling (HDGS)
[118]	✘	✘	✓	UAV altitude, RIS phase shift	Expected sum AoI	1.Proposed PPO 2. Random Walk 3.Hovering with greedy
[125]	✘	✘	✓	UAV-RIS location, RIS phase shift	Average sum AoI	1.On-Policy PPO 2.Off-Policy DQN

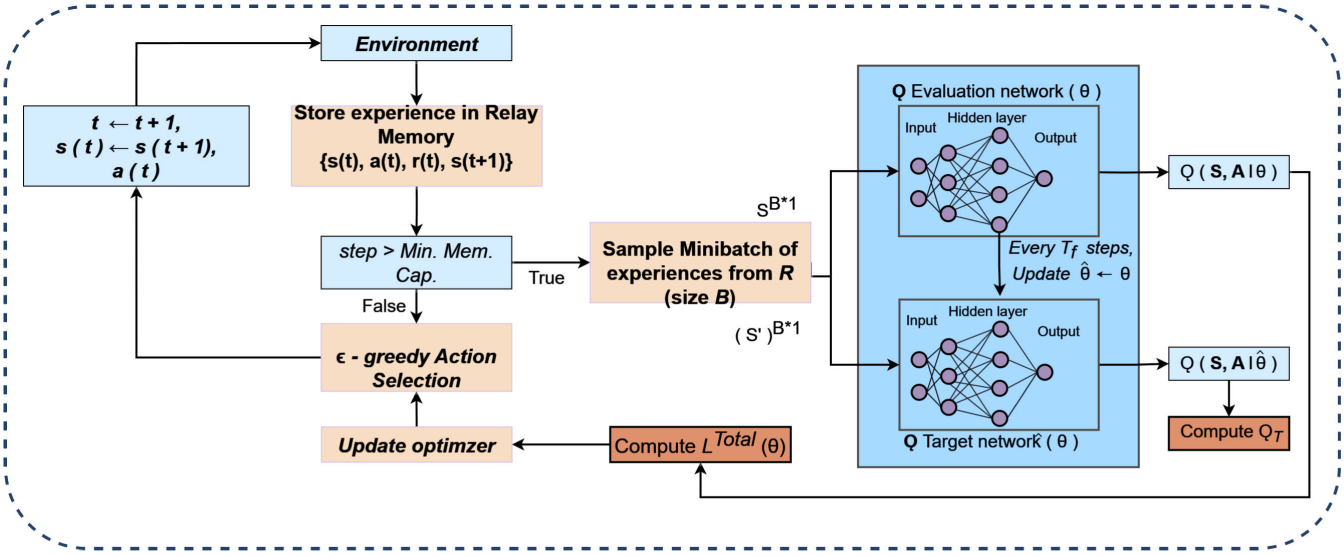

FIGURE 10. Typical representation of the DQN architecture in [125].

Numerical results show that the proposed algorithm performs well with respect to AoI.

C. ACTOR CRITIC-BASED DRL ALGORITHMS USING SINGLE AGENT

Table 7 provides comparison among different actor critic-based works using single agent based on the optimization objectives, metrics used as well as benchmarks.

1) ACTOR CRITIC-BASED ALGORITHMS FOR OPTIMIZING TRAJECTORY AND SCHEDULING PLANNING

UAV and RIS are promising for improving the capacity, coverage and reliability of wireless communications [127]. As opposed to prior works which studied UAV-aided fresh data collection in 2D environments, the authors [12] leverage RIS to mitigate the impact of blockages due to buildings on the AoI performance in a 3D urban IoT scenario. The authors aim to minimise the AoI of all the IoTD by optimising UAV flight trajectory, IoTD transmission scheduling, and RIS discrete and shift. The problem is, however, challenging due to the high correlation of channel information, complex building distribution, and dynamic UAV trajectory. In such cases, it is not possible or applicable to apply traditional optimisation techniques, so the problem is reformulated as an MDP while accommodating the optimization of phase shift control of RIS and UAV trajectory as well as IoTDs transmission scheduling. A DRL-based algorithm (SAC-AO-RIS) combining soft actor-critic (SAC) [128] and alternating optimisation (AO) was developed. SAC algorithm with a high exploration ability is leveraged for learning the UAV trajectory and scheduling policy of IoTD. Also, a simple AO algorithm is used to effectively optimize the RIS phase shift. To ensure the training procedure is stable and converges well, the recent prioritized experience replay technique was exploited. The authors show via simulations that the proposed scheme can effectively reduce the average episodic AoI as compared to the baseline methods.

The authors in [119] focus on path learning for multiple battery-recharged UAVs to optimize the age of information

TABLE 7. Comparison on optimization objectives (T=trajectory, E=energy, S = scheduling), metrics and benchmark algorithms using actor critic-based DRL algorithms in single UAV scenarios.

Ref.	T	E	S	Others	Metric	Studied Algorithms
[12]	✓	✘	✓	RIS phase shift	Average AoI	1.Proposed SAC-AO-RIS 2. SAC-RA-RIS 3.SAC-NO-RIS 4. TP-AO-RIS 5.TP-RA-RIS 6.TP-NO-RIS 7.FTGS-NO-RIS
[119]	✓	✘	✓	-	Average AoI	1. Proposed TD3-AUTP 2. A-TP 3.CA2C 4. Greedy
[126]	✓	✘	✓	-	Average AoI	1. Proposed TD3 2. PPO Algo
[120]	✓	✘	✓	-	Average AoI	1.Proposed A3C 2. DQN
[117]	✘	✘	✘	data collection, forwarding strategy	Average AoI	1. Proposed GSDRL 2.DQN 3.DDQN

and power of IoT devices deployed in a geopgraphic area which continually uploads data. Particularly, UAVs have beeen deployed in this work due to the energy limitations and poor channel conditions which makes it impossible for IoTDs to transmit to the BS directly. Thus the mobile data collector (UAV) is dispatched to gather IoTD data and offload to BS. AoI-energy-aware data collection for UAV-assisted IoT studied in this work aims to minimize the weighted sum expected AoI, UAV propulsion energy, and IoTD transmission energy via the joint optimization of UAV flight speed, hovering location, and bandwidth allocation for data collection. Due to the nature of the system dynamics, the authors formulated the problem with respect to the maximum tolerable AoI of IoTDs, transmission channel constraints, and energy consumption. The problem was modelled as an MDP and a twin delayed deep deterministic policy gradient-based UAV trajectory planning algorithm was proposed based on the architecture in Fig. 12 to solve the problem by deploying the deep neural network for feature extraction. Via simulations, the author showed that the proposed scheme outperforms the deep Q network and actor-critic-based algorithms with respect to achievable AoI and energy efficiency.

In a RIS-assisted UAV wireless communication for achieving information freshness in IoT, [126] jointly optimize UAV flight trajectory, SN scheduling, and IRS phase shift matrix. The problem is modelled as an MDP and a DRL algorithm based on Twin delayed deep deterministic policy gradient was proposed for learning and finding optimal UAV trajectory and SN scheduling. For transmissions that have been scheduled, the IRS is used for aligning signals and shifts based on channel information. Via simulation, the authors show that IRS-assisted UAV data collection can reduce SNs AoI significantly.

In [120] the authors deploy A3C for UAV-assisted data collection in WSN with EH, and the objective is to minimize AoI while maximizing EH-powered UAV trajectory and transmission opportunities of SNs. The A3C algorithm is proposed to achieve real-time decision-making for the DRL framework. The algorithm works as follows: Initially, all workers, global and shared parameters, and the global shared counter is initialized. The following is done in all episodes: gradients are reset to zero, the initial energy is set to maximum energy and the state is set to the initial state, starting with the first slot to the last slot, the action is executed according to the policy if sufficient energy is available i.e energy is greater than the threshold energy. After the action is taken, the position of the UAV and its energy is updated, otherwise the UAV must remain in the air for energy harvesting from ground nodes. The AoI is updated and the cost is determined before it moves to the next state, whenever there are no more update packets, the actor adjusts and optimises the policy while the A3C deploys entropy for increasing exploratory actions. The global network parameters are then updated asynchronously based on the cumulative gradients and pull parameters from the global network.

2) ACTOR CRITIC-BASED ALGORITHM FOR OPTIMIZING OTHER OBJECTIVE FUNCTIONS

For timely data collection in UAV-based Internet of Things networks with SNs having different timeliness priorities, the [117] aim at helping UAVs with different initial positions to independently complete data collection tasks via a guided search deep reinforcement learning algorithm. Data collection is modelled as a sequential decision problem for minimizing the average AoI or maximizing the number of collected nodes for a specific environment. Using GSDRL

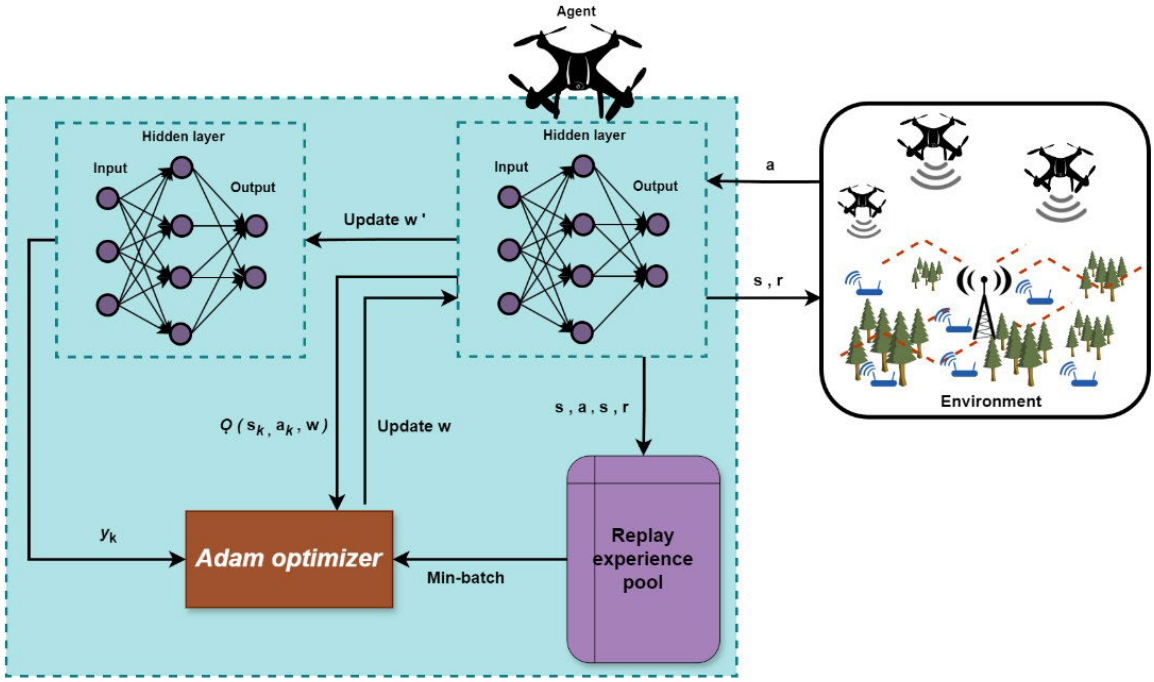

FIGURE 11. Illustration of the DRL Algorithm for UAV data collection and forwarding [117].

(see 11) the authors optimize the data collection strategy. After the network has been trained using GSDRL, the UAV can rapidly perform autonomous navigation and decision-making to complete complex tasks. The proposed GSDRL can effectively adapt to diverse environments and obtain a good data collection and forwarding strategy.

D. LESSONS LEARNED

Most of the studied works have deployed DQN algorithm applicable for discrete and continuous state space problems while D3QN has also been studied for complex setups involving trajectory scheduling and energy recharging. Value-based algorithms have mostly been deployed in literature especially when researchers began to deploy DRL for trajectory optimization. Trajectory optimization is the most common objective that has appeared in these works; either as a stand-alone objective or as one of the objectives. This is followed by communication scheduling and energy efficiency. After value-based algorithms, actor-critic algorithms have also been deployed in recent times for optimizing UAV trajectory and scheduling. The incorporation of reconfigurable intelligent surfaces and the optimization of their phase shift as well as data forwarding have also been studied with actor-critic algorithms. On the other hand, policy-based algorithms have mainly been deployed to optimize scheduling decisions and UAV altitude as well as RIS phase shift.

VI. MULTI-AGENT DRL-BASED SOLUTIONS

In this section, a list of summaries of various studies in multi-UAV scenarios (and also multi-agent-based) using model-free DRL algorithms to minimize AoI is presented (refer Table 8). The organization of this section is provided in Figure 15.

In addition, works using various model-free DRL algorithms in multi UAV environments of value-based, policy-based, and actor-critic-based classes have been identified and discussed. Table 9 provides the list summary of the works using these DRL algorithms.

A. DQN-BASED DRL ALGORITHMS USING MULTI-AGENTS

Table 10 provides comparison among different value-based works using multi agents based on the optimization objectives, metrics used as well as benchmarks.

1) DQN-BASED ALGORITHMS FOR OPTIMIZING TRAJECTORY AND ENERGY

AoI-aware DRL-based UAV trajectory planning in wireless-powered IoT networks was the focus in [5]. In other words, the authors utilise multiple UAVs to wirelessly charge low-power IoTDs and collect fresh information from them while considering the realistic energy constraints of low-power IoTDs and the dynamic channel conditions. UAVs are deployed to wirelessly charge IoT devices. IoT devices upload fresh information to UAVs based on harvested energy (using non-linear EH model) thus facilitating a sustainable IoT network. To avoid interference, devices are not allowed to harvest and transmit at the same time. UAVs make the decision as to where or which direction to fly and the device that should be visited next. This is achieved by considering the trade-off between data collection and energy transmission and the distance to the destination node.

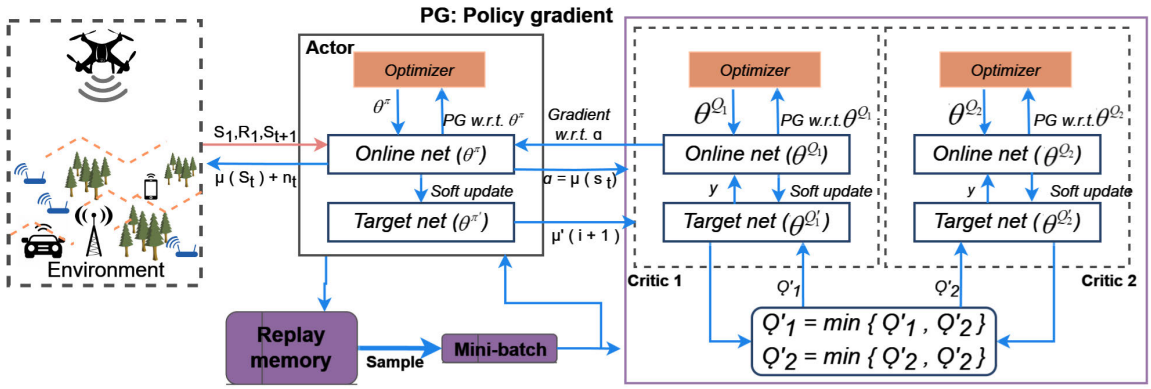

FIGURE 12. Combination of the TD3 structure and AoI-energy-aware UAV-assisted data collection (TD3 algorithm based on DDPG framework) [119].

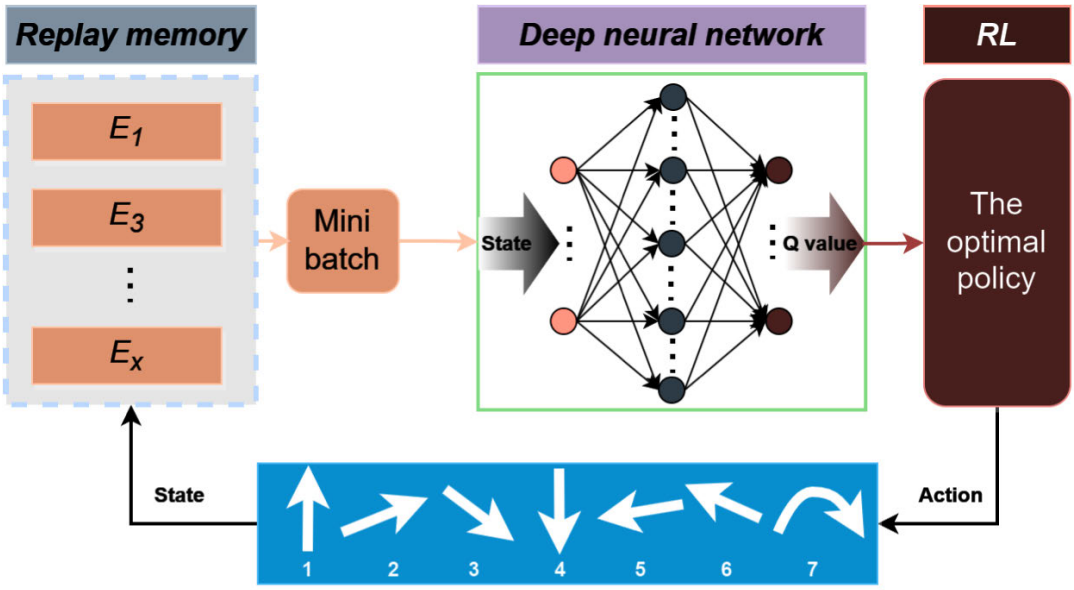

FIGURE 13. The schematic view of the deep Q-learning architecture [5].

The authors consider the blockage rate and density of the surrounding area captured by constraints. The UAV employs fixed transmit power to communicate with and interact with the devices. The authors establish a practical LOS/NLOS channel model for exploring the influence of the dynamic time-varying channel on AoI. The authors formulated the problem as an offline sequential decision-making problem in the presence of dynamic channel conditions. A novel DRL-based proactive UAV trajectory planning algorithm based on Fig. 13 was proposed to automatically adjust the UAV flight policy based on the channel conditions variations and the energy transmission vs data collection trade-off. The results show that the proposed UAV trajectory planning algorithm can reduce AoI significantly (by up to 20% to 65%) as compared to other trajectory planning algorithms.

The use of DRL for energy-efficient UAV trajectory design was considered in [131]. The authors consider multiple UAV-assisted IoT in which the UAVs function as relay nodes between SN and BS. Each UAV relays the information from IoTD to BS (located at the centre of the map). There exists charging spots at fixed positions, such as corners of the terrain. An optimization problem is formulated to jointly plan the trajectory of UAV while minimizing AoI of received messages and considering UAV energy consumption. To address this problem, the authors proposed the DRL algorithm to find the optimal policy for UAV trajectory with nine movement directions of the UAV at each instant. The deployed DRL-based algorithm is based on the architecture in Fig. 16. Deep-Q network functions as a function approximator for estimating the state-action value function. The proposed scheme converges first and yields a

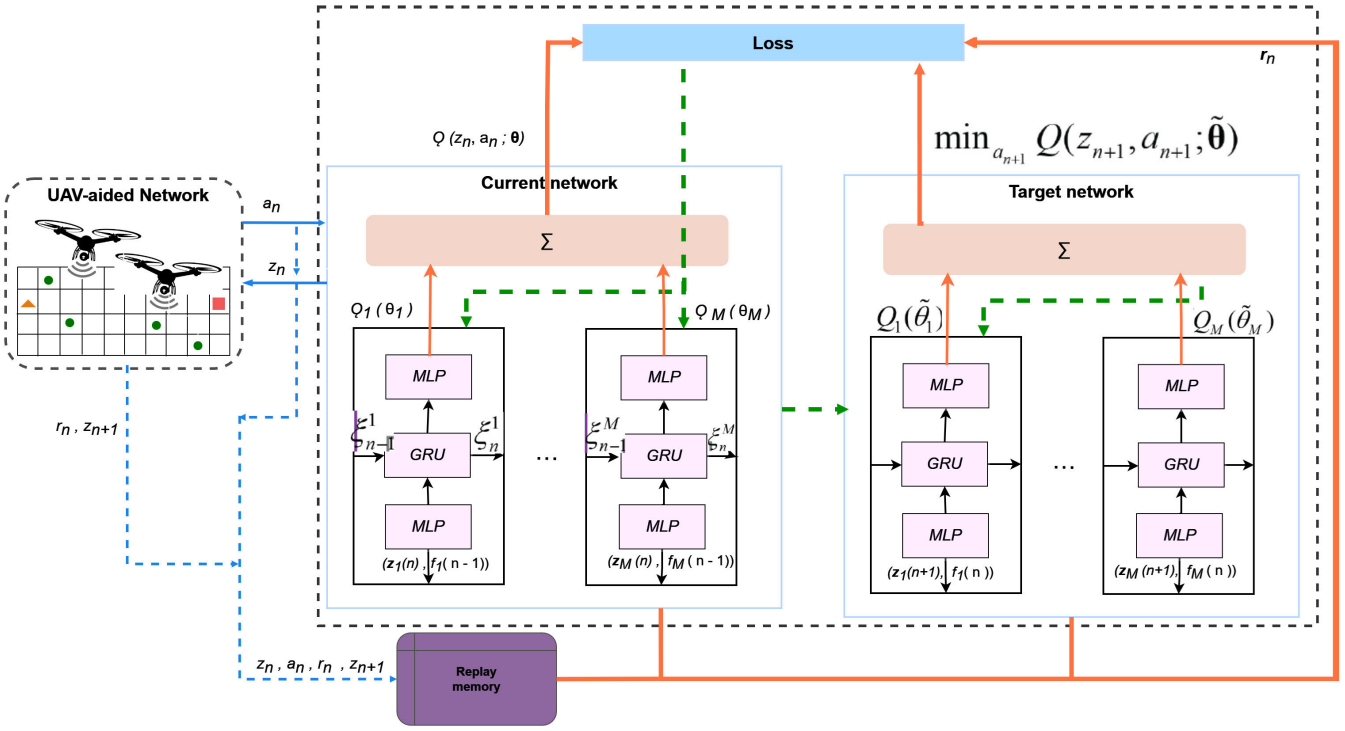

FIGURE 14. Schematic representation of the VDN architecture for multi-UAV-aided Data collection [129].

TABLE 8. Summary of DRL-based algorithms in multi-UAV scenarios and the functions.

Ref	Name	Function
[5]	PUTP	UAV trajectory planning for adjusting the flight policy based on variations in channel conditions and energy transmission and data collection trade-off
[14]	DRL-CTDE	To solve finite horizon Decentralized POMDP which was used to characterize the non-stationary environment in the studied multi-UAV data collection framework
[118]	PPO	To obtain the optimal online scheduling policy, control the dynamic UAV altitude and maintain the AoI at the BS over the dynamic and unreliable channel conditions
[125]	Off-Policy DQN and On-Policy PPO	Two model-free DRL for jointly optimizing RIS phase shift, location of UAV-RIS, and IoT transmission scheduling for large scale IoT
[130]	VDN	Find an optimal solution for POMDP-based WPT-powered multi-UAV assisted data collection via centralized training and distributed execution
[131]	DRL-based	For solving formulated UAV trajectory optimization problem to minimize the AoI of received messages
[132]	DRL-based	For optimizing the flying distance and average AoI under constraints in UAV energy
[133]	QMAUTP	MADRL was used for solving a formulated decentralized POMDP to minimize the AoI for sensing tasks
[134]	DRL-based	Multi-agent DRL-based solution for power control and trajectory design subtasks with independent state, action spaces and reward function
[129]	Sarsa and VDN-based	UAV flight trajectory design in the absence of knowledge of SN sampling modes
[135]	Deep Q-network	For estimating the state-action value function to optimize UAV trajectory, AoI and minimize energy consumption
[136]	soft actor-critic	DRL-based solution was used for optimal UAV positioning. A soft actor-critic algorithm was used to train UAV agents. They execute required actions for flight location selection
[137]	MADRL-"TEAM"	multi-agent DRL for optimizing the trajectories of two teams of UAVs for maximizing throughput, minimizing AoI, improving energy utilization and energy transfer

TABLE 9. DRL-based algorithms for multi UAV scenarios.

Model Free RL	Algorithm	References
Value-based	DQN-based	[5] [14] [130] [131] [133] [134] [129] [135] [138]
Actor critic	DRL-based	[132]
	MADRL-based	[137]

lower AoI and energy reduction than the random walk policy. Particularly, the average AoI is reduced by approximately 25% and requires half less energy when compared to the baseline scheme.

- Multi-agent DRL-based solutions (Section VI)
  - DQN-based algorithms using multi-agents (Sec VI-A)
    - DQN-based algorithms for optimizing trajectory and energy (Section VI-A1)
    - DQN-based algorithm for optimizing trajectory and scheduling planning (Section VI-A2)
    - DQN-based algorithms for optimizing energy (Section VI-A3)
  - AC-based DRL using multi-agents (Section VI-B)
    - Actor Critic-based DRL for optimizing trajectory and energy (Section VI-B)
  - Lessons learned (Section VI-C)

FIGURE 15. Organization of section V.

TABLE 10. Comparison on optimization objectives (T=trajectory, E=energy, S = scheduling), metrics and benchmark using value-based DRL algorithms in multiple UAVs scenarios.

Ref.	T	E	S	Metric	Studied Algorithms
[5]	✓	✓	-	Sum AoI	1.Proposed AoI-based (PUTP) 2.Random-based 3. Distance-based
[131]	✓	✓	-	Average AoI	1.Proposed DRL 2.Random Walk
[133]	✓	✓	-	AoI	1.Proposed QMAUTP(QMIX+GRU) 2.IUTP MADRL-based 3. VUTP MADRL-based
[134]	✓	✓	-	AoI	1.Proposed DQN-scheme 2.Q-learning scheme
[135]	✓	✓	-	AoI	1.Proposed DQN-based 2.Greedy algo. 3.Nearest neighbor 4.Random Walk
[138]	✓	✓	-	AoI	1.Proposed VDN
[14]	✓	-	✓	Total Average AoI	1.Proposed QMIX-based (DRL-CTDE) 2.Cluster-based 3.Nearest Scheduling 4.IDQN-based
[130]	-	✓	-	Average AoI	1.Proposed VDN 2.Greedy policy 3.Nearest neighbor
[129]	-	✓	-	Average AoI	1.Proposed Sarsa + VDN 2.DQN policy 3.Greedy policy 4.Nearest neighbor 5.Sarsa policy

A Multi-agent DRL approach was considered for cooperative UAVs with the aim of achieving AoI optimal trajectory planning in [133]. The authors consider multi-UAV-assisted IoT architecture in which UAVs perform their actions cooperatively to collect data packets generated by IoTDs to be transmitted to the BS while ensuring information freshness. The distributed cooperative multi-UAV dynamic trajectory planning problem was formulated as a decentralized partially observable MDP (Dec POMDP) where packet updates arrive in a stochastic manner and are thus unknown to the UAVs. A multi-agent DRL was devised to address this challenge due to the unknown environmental dynamics and high conflict collision constraints. The developed algorithm leverages QMIX and GRU techniques. The authors show by simulation that the proposed algorithm is effective for solving the formulated problem.

The authors [134] consider a UAV-enabled data collection system with single-antenna UAVs and IoTDs. The SNs, i.e., ground devices, transfer their data independently to UAV stations, in which users share sub-carriers to access the wireless network in an efficient manner. A UAV functions as a beacon for gathering information by flying in a specific area. In order to satisfy communication quality requirements and flight safety, each UAV has the same flight height and a maximum service radius. The paper adopts a probabilistic LoS channel model. IoTD using the same sub-carrier will cause interference to suppress the rate performance, and thus the authors explore power control for each sub-carrier to improve the rate performance of the UAV system. Considering the dire importance of information freshness, the authors optimise the trajectory of the UAV for the time-sensitive IoTDs. The joint power control and UAV trajectory design for achieving information freshness problems was formulated. The problem is non-convex and thus decomposed into two sub-components: power control and trajectory optimization with constraints on UAV maximum transmit power and minimum rest energy to achieve flight safety. Then, an MA DRL scheme is proposed to solve these sub-problems with independent state and action spaces as well as independent reward function. Via simulations, the authors show the proposed schemes yield better performance gains compared to the benchmark schemes.

Multi-UAV path learning was deployed for AoI and power optimization in IoT networks with battery-charged UAVs [135]. A set of deployed IoTDs with multiple UAVs relay data sensed by SNs to BS. An optimization problem is formulated which jointly optimizes the UAV trajectory while reducing the energy consumption and AoI of received messages. To ensure UAV operates in an energy-efficient manner, the UAV can get recharged in charging stations/depots. The complex optimization problem is solved via DRL. The authors use a Deep Q network to estimate the state-action value function. The proposed scheme converges quickly and achieves lower ergodic energy consumption and ergodic rate compared with benchmark algorithms such as the greedy algorithm, nearest neighbor algorithm, and random walk.

The authors in [138] focus on using DRL approach for timely data collection for UAV-assisted IoT. The authors study a UAV-assisted wireless powered IoT data collection scenario when UAV function as mobile charging stations for wireless recharging of SNs. To design the trajectory and mobile path of mobile charging stations as well as jointly optimize AoI, the authors formulate the problem as a partially observed MDP with large state and action spaces.

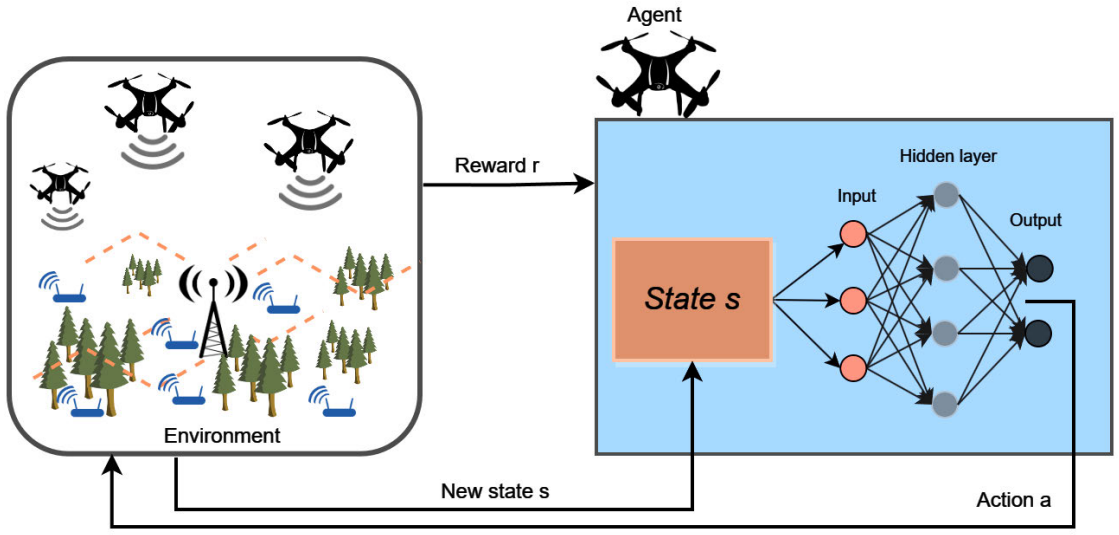

FIGURE 16. Illustration of the DQN architecture [131].

They deploy a multi-agent DRL based on VDN for making real-time decisions based on the partial observation of the environment. The authors show, through simulations, that the proposed algorithm is effective.

2) DQN-BASED ALGORITHM FOR OPTIMIZING TRAJECTORY AND SCHEDULING PLANNING

A learning-based multi-UAV cooperative data collection for information freshness in IoT was studied in [14]. UAVs with limited onboard energy leave their initial location to collect data from sensors in cooperation with other UAVs. UAV must stop at the destination while ensuring non-negative residual energy to ensure it completes its mission successfully. To design the UAV trajectory and the SN transmission scheduling policies for information freshness, various constraints such as collision avoidance, trajectory, kinematic, etc are considered as constraints. The multi-UAV data collection problem is modelled as a decentralized partially observable MDP Dec-POMDP as each UAV is not aware of the environmental dynamics and can observe only a part of the SNs. To address this challenge, the authors proposed the use of a multi-agent DRL-based algorithm with centralized learning and decentralized execution. The agent and mixing network's centralized training is done offline and the training process is stabilized via two sets of neural networks as shown in Fig. 17. Action masks were used to filter invalid actions and ensure the constraints were not violated. Via simulations, the authors show that the proposed algorithm can reduce total average AoI, and the action mask method improves convergence speed.

3) DQN-BASED ALGORITHMS FOR OPTIMIZING ENERGY

Ref [130] consider ML solutions for multi-UAV-assisted data collection in wireless-powered IoT in which the UAVs acts as mobile data collector for charging sensor nodes. SNs are charged using the RF transfer from UAVs and the harvested energy from the UAV is used to upload data to the UAV. SNs sample environments at fixed or random intervals based on the sampling mode and update packets of sampled information are delivered by multiple UAVs. UAVs cooperate to collect data based on their partial observations, e.g., locations and energy levels, and the AoI status and ttl of the most recent update packet of each SN within its coverage. The authors assume random sampling, with its update packet arrival is based on a Poisson distribution. The objective is to improve the SN service time and freshness of collected information. The authors modelled the problem as a partially observed MDP with a large observation action space in which the UAV intelligently learns the environment (as an agent) and makes intelligent decisions. They employ a multi-agent VDN-based MARL algorithm in which each UAV acts as an agent that takes independent decisions on flight and data collection based on partial and time-varying observations to obtain the optimal strategy via the multi-agent DRL framework. Results indicate the SN average AoI increases with an increase in sampling interval, and it is higher with random sampling compared to fixed sampling. Then VDN performs better than the non-learning policies, i.e., greedy and nearest neighbour policies, to achieve a smaller average AoI.

The authors of [129] study the multi-UAV-aided data collection problem when the sampling mode of SN is unknown to the UAV. They deploy state-of-the-art RL methods for designing UAV flight trajectories. SNs randomly or periodically samples data packets and multiple energy-constrained UAVs are dispatched to collect update packets from SNs while the UAV is flying over them. The trajectory planning problem is formulated as MDP with the objective of minimizing the weighted sum average AoI of SN under collision avoidance and energy capacity constraints. The

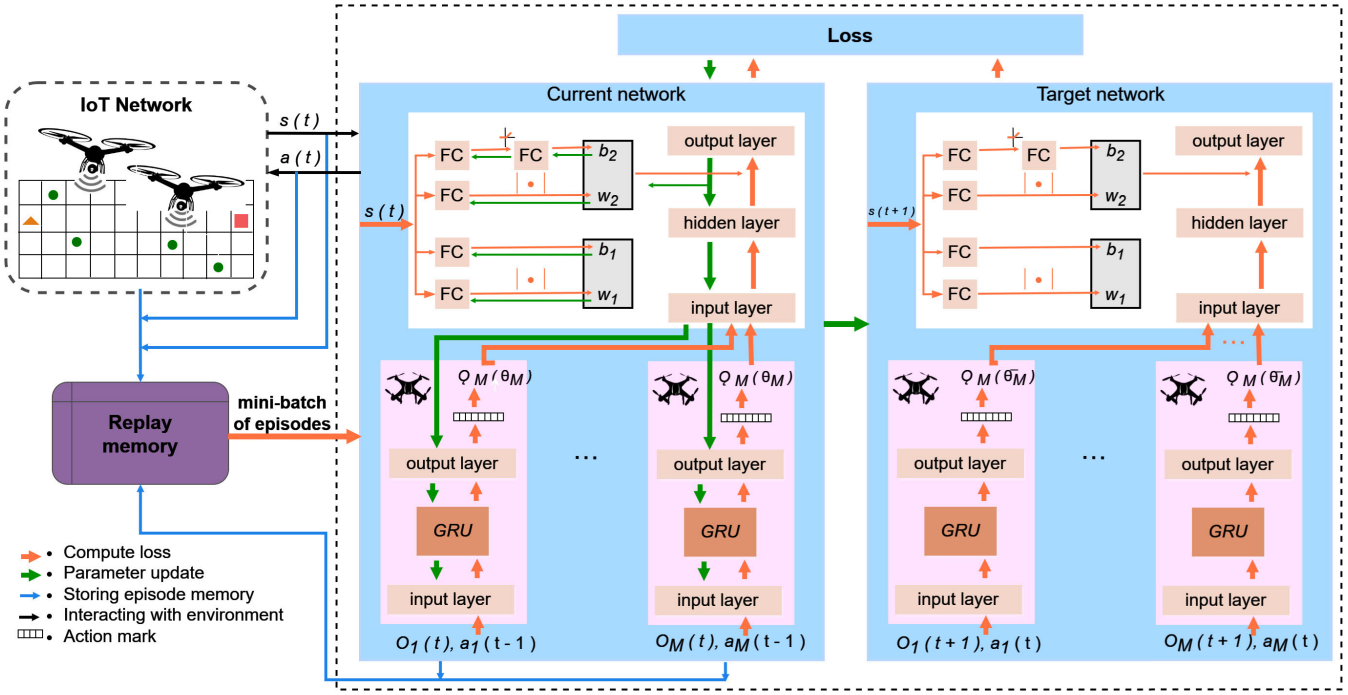

FIGURE 17. Illustration of the QMIX algorithm for the training phase of the neural network parameters [14].

TABLE 11. Comparison on optimization objectives (T=trajectory, E=energy, S = scheduling), metrics and benchmark algorithms using actor critic-based DRL algorithms in multiple UAVs scenarios.

Ref.	T	E	S	Metric	Studied Algorithms
[137]	✓	✓	-	Average AoI	1.Proposed TEAM DDPG-based 2.MADQN 3. DDQN
[136]	✓	-	-	Sum AoI	1.Proposed SAC 2.DQN
[132]	✓	-	-	Average AoI	1.Proposed DRL-model 2.GA 3.Distance DRL

authors propose two learning-based algorithms based on SARSA (for optimal policy) and VDN (based on the architecture in Fig. 14) to effectively carry out data collection tasks of SN. SARSA-based algorithms can facilitate optimal policy asymptotically when some conditions are satisfied. The VDN, one of the most popular MADRL, makes UAVs take decisions independently on their flight and data collection using the partially observed network information. Via simulations, the effectiveness of the two learning-based algorithms over traditional policies was shown.

B. ACTOR CRITIC-BASED DRL USING MULTI-AGENTS

Table 11 provides a comparison among different actor critic-based works using multi agents based on the optimization objectives, metrics used as well as benchmarks.

Actor Critic-Based DRL for Optimizing Trajectory And Energy: The authors of [137] focus on synchronizing a team of UAVs for the timely collection of data and energy transfer via DRL. For data collection, UAVs should first engage in wireless energy transfer to supply IoTDs with the required energy in the downlink. IoTD performs wireless information transfer to UAVs in the uplink using the energy harvested. However, whenever the same UAV performs WIT and WET, its energy usage and data collection time are largely sacrificed and it is even difficult to coordinate between UAVs for improving WE and WIT performance. With UAVs divided into two teams to act as data collectors and energy transfer respectively, a MADRL is used to optimize both teams trajectory, maximum throughput, and minimum expected AoI as well as energy utilization of UAV and improve energy transfer. Simulation results show that the proposed scheme effectively synchronizes UAV teams and adapts their trajectory for large-scale dynamic IoT environments.

The authors [132] explore UAVs with heterogeneous energy capabilities. A set of SNs is distributed in an area to sense environmental data while the UAV swarm collects data from each SN and uploads it to the BS or DC where the UAVs are initially located. In order to prevent collisions, the authors assume that UAVs can fly at different altitudes. UAVs send signals for data acquisition to the sensors when they hover above them. Such UAVs can collect data from multiple SNs (via linear path) while ensuring they have sufficient energy to return to the data center or they just return to the data center also via a linear flying path (from origin to destination). In this case, the linear distance corresponds to the energy consumed during UAV flight. In this context, the authors focus on path planning for the UAV swarm for optimizing the

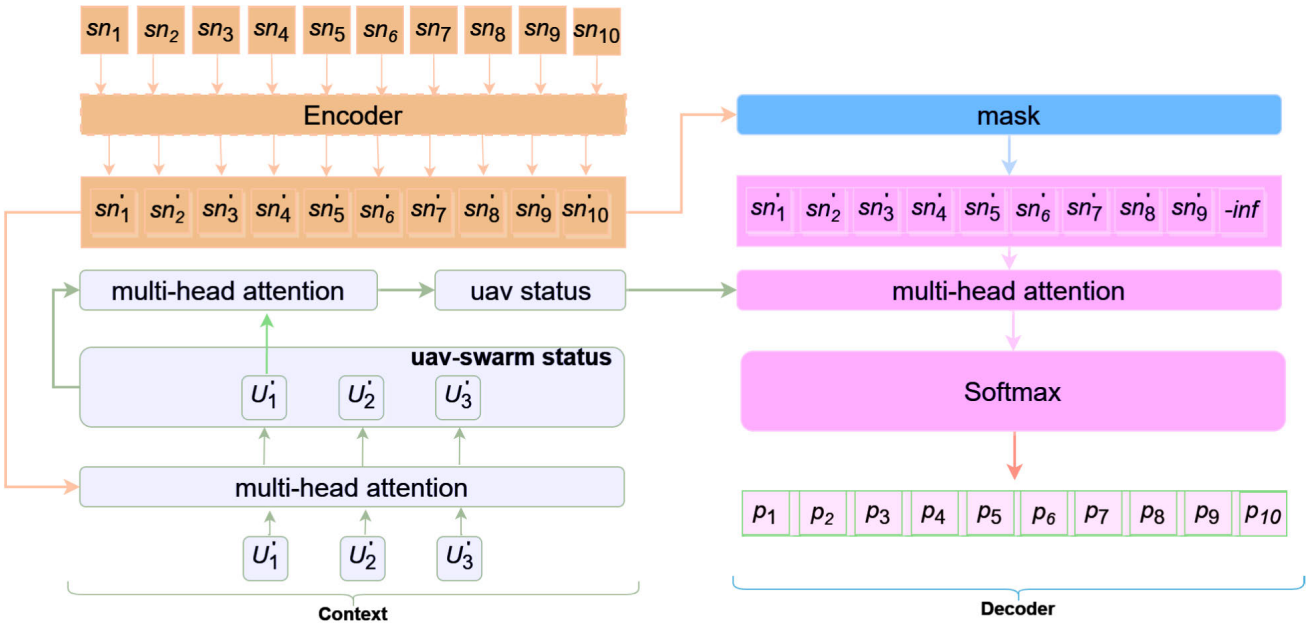

FIGURE 18. Attention-based model for computing encoder output from UAV-swarm status [132].

AoI with UAV cooperation. They also designed an attention-based DRL algorithm using an encode-decoder model (refer Fig. 18) for heterogeneous UAV path planning for optimizing the AoI considering UAV's energy constraints. They conduct simulations to prove the fast convergence of the proposed algorithm, high optimisation capability, and reliability for use in heterogeneous UAV cooperative swarm with AoI constraints.

DRL and device matching were deployed to achieve fresh and energy-efficient data collection in satellite-controlled rechargeable UAV-assisted IoT networks in [136]. UAV flies to collect data from IoT devices that sense information and the collected data is transferred to the satellite. To improve the information freshness, AoI is minimized via UAV trajectory design. A satellite is used as the central controller for collecting the minimum AoI values between UAV and IoT device which is obtained by training the UAV via DRL. The training helps to minimize AoI with optimal UAV positioning. The satellite builds a preferred list of UAVs and IoTDs based on the AoI values and finishes the matching via Gale-Shapley algorithm. Simulation results emphasize the merit of the proposed approaches over others.

C. LESSONS LEARNED

Multiple UAVs could collaborate to avoid collisions along their trajectories while aiming to achieve other objectives such as minimizing energy efficiency, AoI, or optimizing scheduling decisions. Such multi-agent systems are cooperative.

In cooperative systems, the agents collaborate to maximize the long-term return. Multi-agent setups could also be fully competitive, where the return of all agents adds up to zero, while mixed settings combine both the cooperative and competitive, for instance, some agents cooperate with teammates while competing with other teams. Some of the challenges of multi-agent schemes include non-stationarity, varying learning speeds, and issues in scalability [139].

Most of the studies in this category mainly deploy the DQN algorithm. Similarly, multi-agent DRL-based solutions have mainly been deployed to optimize the UAV trajectory, and energy consumption minimization while scheduling planning was less studied. In addition, average AoI is one of the most common metrics studied using multi-agent DRL schemes.

VII. DISCUSSION

A. TARGET OBJECTIVES OF DRL-BASED SCHEMES

Figures 19 and 20 shows a list of some of the DRL-based algorithms and modifications that have been deployed in the reviewed papers. D3QN was used to solve the MDP to determine UAV trajectory, SN scheduling and energy recharging in [115]. DDQN was used to overcome the disaster of dimension in [116], while in [14] because each UAV is unaware of the environmental dynamics a multi-agent DRL with centralized learning and decentralized execution was proposed. The complex optimization problem formulated for optimizing UAV trajectory and minimizing device energy consumption and AoI was solved using a deep Q network. DQN and PPO were proposed to minimize the average sum AoI by jointly optimizing phase shift, location, and IoTD transmission scheduling.

In [5], DRL was used to achieve optimal system level AoI under dynamic channel conditions which adjust the UAV flight policy based on the channel variations and trade-off between energy transmission and data collection. TD3 lends

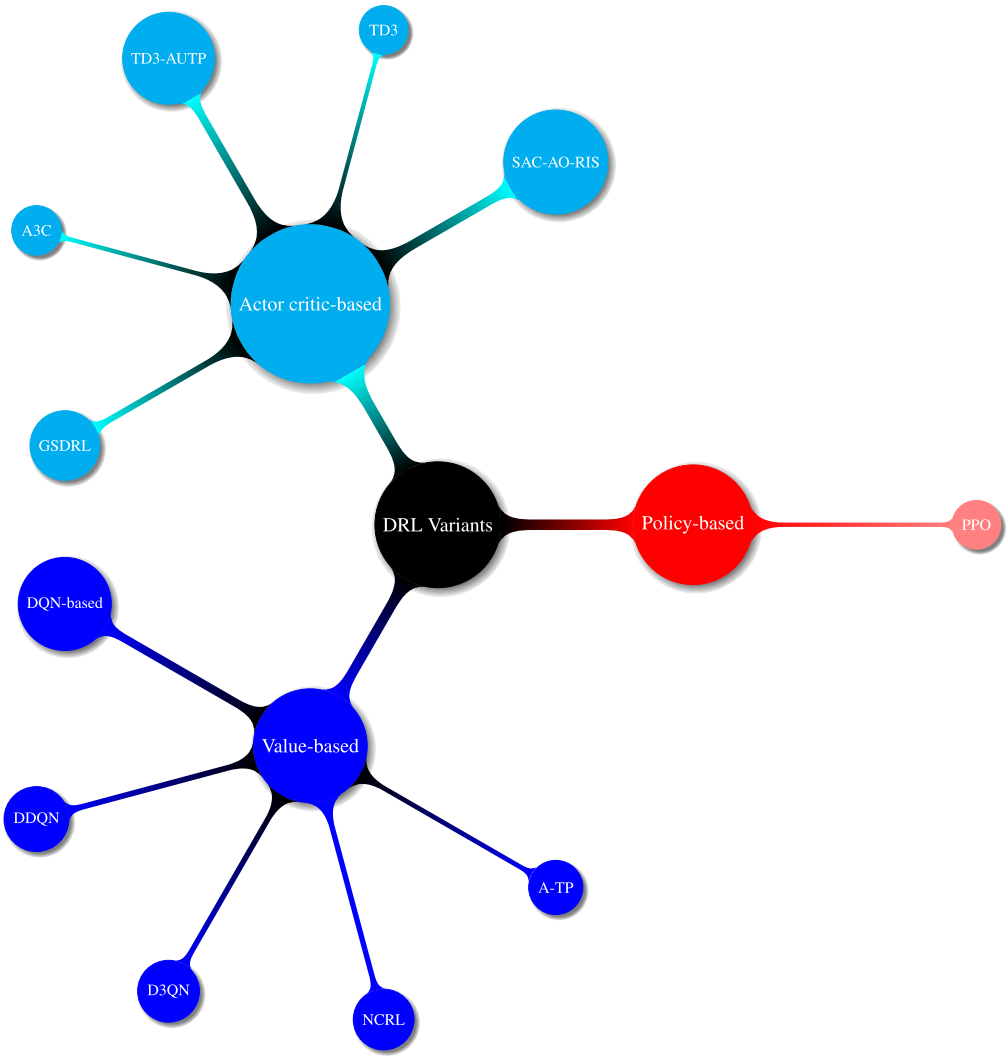

FIGURE 19. A list of some of the variants on DRL studied for single UAV scenarios.

itself well for learning and finding the optimal UAV trajectory and SN scheduling in [126]. Agile DRL with experience replay was used to solve the energy efficiency maximization problem with contextual constraints (e.g. AoI at ground BS) in [123] while DRL-based satellite control for collecting minimum AoIs between UAVs and IoTDs were captured in [136]. DRL was used for trajectory design with the state space considers residual energy, energy efficiency, and AoI in [13]. VDN was used in [138] to make real-time decisions according to the partial observations of the environment.

MADRL was used to address issues due to unknown environmental dynamics and conflict collision constraints in [133]. VDN in MADRL framework was used to find the optimal strategy for UAV to learn its environment and make independent decisions in [130]. The problem in [119] was formulated as MDP due to the system dynamics and TD3 policy gradient-based trajectory planning algorithm was used to deal with the multi-dimensional action space. MADRL scheme was used to handle power control and trajectory design sub-problem in [134]. PPO is used to solve mixed integer non-convex optimization problems regarding altitude optimization, communication scheduling and phase shift optimization in [118]. A-TP using DRL is used to handle traffic generation pattern with unknown topology in [3] in which MDP was used to model the complex interaction between UAV and IoT.

DRL is used to design the UAV flight trajectory without knowing the sampling mode of each SN in [129]. Therein SARSA and VDN were deployed to meet the data collection requirements. In [131], DRL was used to solve the complex optimization problem associated with multi-UAVs functioning as relays using a DQN for estimating state-action value functions. In [117], GSDRL helps to tackle the problems of UAVs with different initial positions which are to complete data collection and forwarding independently. Online free DRL to solve the problem where of dynamic altitude control and optimal scheduling policy for UAVs under unreliable channels [124]. In [132], DRL based on attention mechanism

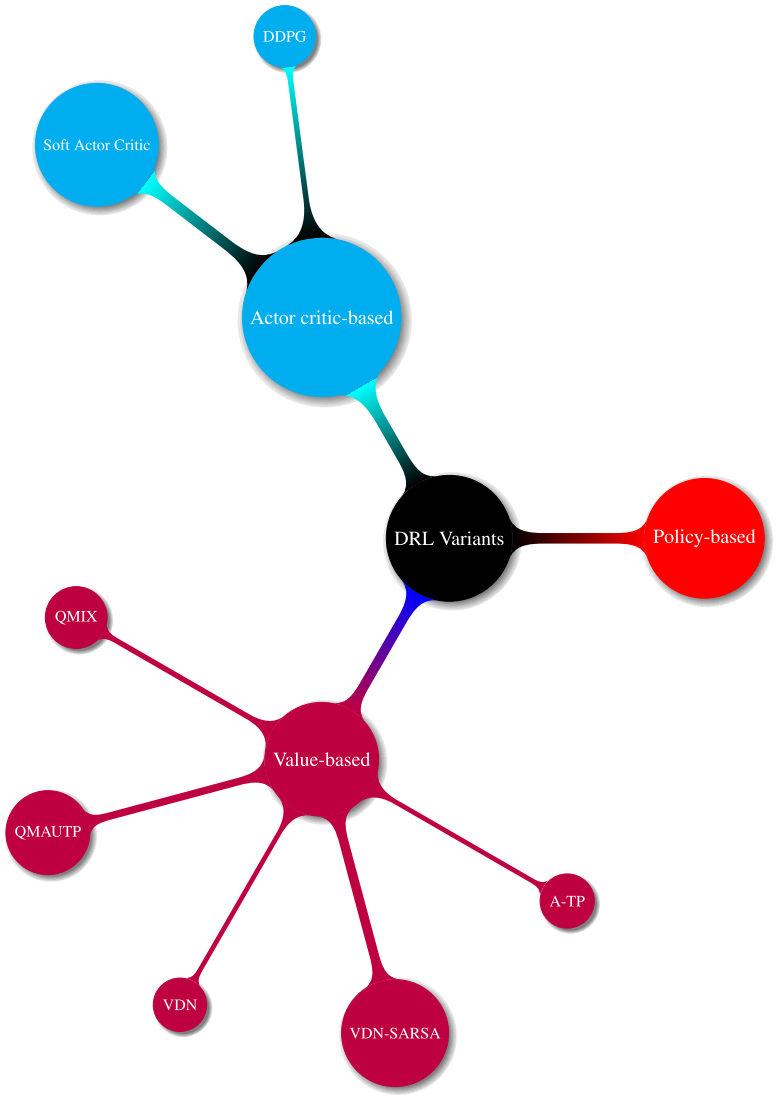

FIGURE 20. A list of some of the DRL variants used in multi-UAV scenarios.

were used to optimize the AoI under energy constraints. UAV-assisted data collection leverages on DRL algorithm for flight trajectory optimization and transmission schedule to overcome the curse of dimensionality in [121]. Also, MADRL was used to jointly optimize a team of UAVs to minimize the expected AoI, and maximize throughput and energy utilization of UAVs in [137].

B. MDP REPRESENTATIONS

Markov decision process plays a vital role in modeling problems to be solved by reinforcement learning. A summary of the MDP representations for some of the reviewed papers is provided in Tables 12 and 13 with observations summarized in Table 14. State representations play an important role in the optimization of UAV-assisted IoT research. State representations within the studied frameworks in this paper involved (horizontal or 3D) UAV current location, AoI of IoTDs, coverage indicator, updated packet lifetime, UAV residual energy, ratio between AoI and planning range of IoTD, UAV rest energy, difference between remaining time and minimum time needed for UAV to reach the destination, energy stored in the battery of SN, difference in UAV battery status, SN location information, indicators to indicate whether data is collected or not, and node energy after update.

As for actions, UAV movement is the most common abstraction of the action of the agent under a maximum of nine cardinals: North, East, West, North, Southeast, Southwest, Northeast, Northwest, Hovering. Similarly, energy harvesting related parameters, scheduling, and flight speed can all be considered as actions for the UAV. As for the reward it could be defined as negative sum of AoI, or summation of penalty for AoI violation from IoT nodes, or new AoI for optimal scheduling, or sum of AoI and energy consumption amongst others. Refer to Tables

TABLE 12. DRL-based algorithms and their MDP structure.

Ref	State	Action	Cost/Reward	Objective
[130]	State update process: the lifetimes of the update packets and AoI values of the SNs in its coverage.	The flight orientation of each UAV during each time slot, such as north, south, west, or east.	A weighted combination of the average AoI for SN, the average energy consumption of the UAV, and a possible penalty.	Find the optimal scheduling policy.
[1]	The state of the environment contains: 1) the initial battery levels and the time instant of operation, and 2) the node energy levels after updates with timestamps.	Node m scheduled to transmit; $a_n = 0$ ends policy, no further updates.	The NWAoI value's decrease.	Discover the optimal control policy that minimizes the ESA, governing both the UAV's altitude and scheduling decisions in the presence of an unknown activation pattern.
[118]	The state of the system in time-slot n is characterized by the SNR at the BS, the altitude of the UAV, and the AoI.	The UAV modifies its altitude, either moving up or down, and hovers during time slot n, determining which IoTD to send its status update to the BS.	A sum of negative AoI values.	AoI minimization.
[136]	UAV position at time t.	The UAV's distance traveled (d_t) and its heading or orientation (θ_t) at time t.	The AoI with a negative value at time t.	AoI minimization and UAV trajectory optimization.
[13]	UAV position, energy efficiency, UAV rest energy, AoI of current state.	The UAV's flight orientation.	Energy efficiency.	In a multi-UAV setup, the goal is to optimize energy efficiency.
[115]	At time slot t, the UAV's location on the ground, the AoI values for all sensors, the time difference between the time left and the minimum time required to reach the final destination, and the remaining energy of the UAV.	At time slot t, the UAV's motion, the SNs' scheduling, and an indicator signaling whether the UAV is recharging during that time slot.	The combined weight-adjusted total of the average AoI and the average cost of recharging.	Minimize the combined weighted total of the average AoI and the average cost of recharging.
[134]	The transmit rate of each sub-carrier. The rate of the link, the 3D coordinates of the UAV.	The proportion of the highest transmission power to the size of the operational area. The UAV's flight maneuvers at time t.	Sum of AoI.	Minimize the AoI of collected data under UAV flying energy constraints.
[15]	Energy stored in the battery of ground node and AoI current slot. Horizontal coordinates of UAV and time needed to reach final destination.	Movement of UAV, data collection, scheduling of energy harvesting for ground node.	AoI of the packets collected by the UAV and remaining time to reach final destination.	Optimizing trajectory, energy and scheduling.
[119]	In the t-th HTP, s^t serves as a characterization of the UAV's local state, encompassing key information including the UAV's position, the current AoI for all IoTDs, and a coverage condition indicator.	During t-th HTP, flight distance of UAV guarantee a condition c_2.	The average AoI linked to the energy usage of IoT devices, UAV propulsion energy, and a specified penalty.	Optimizing energy consumption and data collection throughput.
[140]	The state of UAVs and the target region at time slot t represented by the Location of UAV, the flying time of target region.	The UAV action is characterized by the flying directions, the flying distances, the processing decision for data in sub-regions that they cover.	The freshness of the whole traget region, the negative form of energy consumed by each UAV.	Optimize real-time global situational awareness amid dynamic environmental changes with minimal energy consumption.
[121]	At time slot t, the relevant factors include the UAV's ground projection, the AoI for all connected SN, the disparity between the remaining time and the minimum time needed to reach the final destination, and the variation between the remaining energy of the UAV and the energy necessary for the UAV to reach the final destination.	UAV move and schedule of SN.	The minimum weighted average AoI of all nodes while adhering to constraints. An extra reward is given when the UAV reaches the destination with non-negative energy, but violations lead to imposed punishments.	Design the optimal flight path for the UAV and schedule sensor transmissions to minimize the weighted sum of the AoI.
[123]	The trajectory position for navigation of the UAV-BS.	The learning agent's trajectory planning is determined by the current state, and the available options are drawn from a predefined set, representing various UAV-BS navigation configurations.	The energy efficiency of UAV-BS.	Maximize energy efficiency through the optimization of the trajectory policy for the UAV-BS.

C. SIMULATION LIBRARIES

Pytorch and Tensorflow are the most common libraries used in the research on DRL-assisted AoI minimization for UAV-IoT. For instance, [1], [3], [115], [121], and [122] consider Tensorflow-Agents library, while [132] deployed Python 3.9 and Pytorch. Similarly, [14], [124], and [131] used Pytorch library while [13] deployed Tensorflow 1.14, on Python3.7 (Intel i7 CPU). The authors in [135] also used Pytorch on the NVIDIA Tesla V100 GPU. Oubbati et. al [137] also used Tensorflow 1.14.0 on Python 3.6.9 while [15] deployed Tensorflow 1.5.0 on Python 3.6. As we can observe not all the reviewed works have provided details on the

TABLE 13. DRL-based algorithms and their MDP structure (contd).

Ref	State	Action	Cost/Reward	Objective
[124]	A combination of AoI at a particular time slot, the status of the virtual queue, achievable rate, and status update size that could be sent to the BS.	During its operation, the UAV adjusts its altitude and decides which IoT device to send status updates to or when to schedule transmissions from its virtual queue to the base station.	The reward, R, is the sum of normalized factors, including an IoT network penalty for high AoI status updates and a network penalty for UAV altitude violations.	Acquire knowledge of environmental dynamics for managing UAV altitude and scheduling policies.
[122]	The status of UAV in time slot n is defined by both its geographical position and the variance between the time remaining and the time needed to reach its ultimate destination.	The UAV choose to stay or move to an adjacent cell, if UAV is at a boundary cell and $v(n)$ leads outside the region in the next slot, it remains in the current location. UAV selects a ground node for update packet reception.	Total expected cost of system.	Attain the optimal policy that reduces the weighted sum of Age of Information (AoI).
[130]	Regarding UAV m, updates are made to the lifetimes of update packets and the AoI values for the SNs within its coverage area. Concurrently, the position of UAV m is revised as a part of this process.	UAV hovers to transmit energy, recieve data and flies to one of its adjacent grid.	Weighted summation of the SN' average AoI, UAV's average energy consumption and potential penalty.	Determine the optimal approach within the framework of multi-agent DRL.
[132]	UAV state contain position, remaining power. The state of the SN encompasses details regarding its geographical position and whether data collection has occurred or not.	UAV flight to certain target node then collecting data and flying to the data center to transmit data back.	The reward function R has two components: the information age reward (r^{α}) decreases with lower average AoI, aiming to minimize cumulative information age, while the collection reward (r^{c}) is associated with the number of sensors collected by the UAV swarm.	Leveraging end-to-end training for the purpose of optimizing the average age while taking into account UAV energy constraints.
[131]	The location of each UAV, the AoI for all Sensor Nodes (SNs) covered by the UAVs, the disparity between the battery levels of each UAV and the cumulative energy required to reach the nearest charging depot, and the energy consumption due to packet relays in a worst-case scenario where UAVs relay packets during every time slot t.	UAV move and schedule policy.	The weighted sum of AoI of all devices.	Determining the optimal policy for UAV trajectories involving nine movement directions at each time instant.
[133]	The state of the m-th agent includes the AoIs at IDs and the BS, the current position and the energy state of the m-th UAV.	The UAV has the option to hover or move in the north, south, east, or west direction.	The reward in the study has two parts: the cost is the weighted average AoI at the BS per time slot, and a penalty r_p is imposed if the joint action $A(t)$ breaches the distance safety constraint.	Devise cooperative trajectories for a group of UAVs to collect and continuously transmit data packets generated at specific IDs to the base station.
[5]	Within the time slot n, the state consists of two parts: the index of the hexagonal cell where the UAV is currently positioned during time slot n, the ratio of the current AoI to the planning range N for all IoT devices during time slot n.	The UAV decides whether to remain stationary above the current hexagon or transition to an adjacent hexagon.	The negative sum of AoI.	Automatic UAV flight policy adjustments considering channel variations and the energy transmission-data collection trade-off.
[114]	The AoIs of the K SNs, The lifetime of the SNs' update packets, The UAV's horizontal location and the UAV's residual energy e.	The UAV's flight action at each timeslot is chosen from the set of four flight directions: North, South, West, East, The AoIs of SNs are updated, the UAV's residual energy $e(n)$ gradually update over time.	Weighted sum of the SNs' average AoI.	Design an age-optimized trajectory for the UAV while simultaneously minimizing the packet drop rate.

simulation libraries and software versions. Similarly some of the information available on the computing capacity of the CPU or GPU indicates the need for huge computational power for such simulations.

D. ALGORITHM COMPLEXITIES

The algorithm complexities of several of the algorithms proposed in literature are provided here. It is worthy of note that some papers do not provide the complexity analysis with respect to their Big-O notation. The complexity of the PPO-based algorithm in [124] is $\mathcal{O}\left((P-1)\cdot n_p^2\right) \sim \mathcal{O}\left(n_p^2\right)$ where n_p is the number of neural units in the p^{th} hidden layer, P is the transmission power. The complexity of the QMIX-based algorithm in [14] is $M\mathcal{O}^a\left((N+1)(N_1+1)(N_2+1)\right)+\mathcal{O}^m$ where M is the set of UAV, N the set of SN, N1 and N2 are positive integers. The complexity of the algorithm for cooperative UAV teams in [137] is $\mathcal{O}\left((N+M)\times T\left(\sum_{i=1}^{\Re-1}\iota_{Ac,i}\iota_{Ac,i+1}+\sum_{j=1}^{\Im-1}\iota_{Cr,i}\iota_{Cr,i+1}\right)\right)$, where N is the number of UAV-ETs, M is the number of UAV-DCs, T is the number of time slots, $\iota_{Cr,i}$ is the

TABLE 14. Observations in MDP frameworks for DRL-based solutions.

Ref	Observation
[118]	A vector with M elements representing the SNR during time-slot n, achieved through coherent combination of signals from distinct paths using RIS element phases, At any given time-slot n, the UAV altitude $H_U[n] \in (Hmin, Hmax)$, and the evolution of $A_i[n]$ of IoTD i
[141]	Network environment (including the BS, sensing target, and mobile devices) is regarded as the state of UAVs in the system before the $n-th$ frame $s_i^{(n)}$ contain the frame index n, current/sensing and transmission location, the remaining data size, the AoI, the cycle index, the stage indicator
[119]	The energy consumption of IoTDs, the UAV flight distance, flight direction and flight speed and the AoI values
[123]	The observation space for the learning agent (located at the ground BS) is defined at each time step, encompassing UAV-BS current location $p^u_{current}$, target position p_{end}, energy efficiency μ involves UAV-BSs, and average age for the navigation optimization δ.
[124]	Network status: the AoI of all the IoT devices, status of the virtual queue, the time elapsed at the UAV virtual queue associated with all IoT devices, the achievable rate, the status-update size
[122]	The condition of a ground node m at time slot n is described as the state $s_m(n)$ which is represented by Battery level, AoI at the UAV, while the state of the UAV is given by the location of UAV
[3]	the horizontal and the vertical coordinate of the UAV within the time slot k x_k, the current AoI of all IoT devices within the time slot k
[130]	The observation of UAV m in time slot n can be described by AoI values, lifetimes of the update packets, position and remaining energy of the UAV
[133]	agent m's observation at time t captures ID-related AoI; each agent can only observe the AoI at one ID only if it visits that ID

TABLE 15. DRL-based algorithms and their complexities.

Ref	Algorithm	Complexity
[124]	PPO-based	$\mathcal{O}\left((P-1)\cdot n_p^2\right) \sim \mathcal{O}\left(n_p^2\right)$
[15]	DQN-based	$\mathcal{O}\left(N + b\sum_{p=1}^{P} n_{p-1}\cdot n_p\right)$
[14]	QMIX-based	$M\mathcal{O}^a\left((N+1)(N_1+1)(N_2+1)\right) + \mathcal{O}^m$
[137]	TEAM	$\mathcal{O}\left((N+M)\times T\left(\sum_{i=1}^{\Re-1} \imath_{Ac,i}\imath_{Ac,i+1} + \sum_{j=1}^{\Im-1} \imath_{Cr,i}\imath_{Cr,i+1}\right)\right)$
[117]	GSDRL	$\mathcal{O}\left(2^{l_0}(L/d)^{2l_0}\right)$
[12]	SAC-AO-RIS	$\mathcal{O}\left(B\times E\times T\left(2(K+2)h_1 + 2h_1h_2 + (K+3)h_2\right) + 8M\right)$
[129]	SARSA-based	$\mathcal{O}\left(\left(3M + 2K + \left\|\hat{\nu}\right\|^M\right)\cdot N\cdot N_e\right)$
[118]	PPO-based	$\mathcal{O}\left(\sum_{q=1}^{Q-1} n_q\cdot n_{q-1}\right) \sim \mathcal{O}\left(n_q^2\right)$
[142]	DP-based	$\mathcal{O}\left(2^M M^2\right)$
[141]	DDPG-based	$\mathcal{O}\left(q^2\right)$
[143]	DRL-based	$\mathcal{O}\left(\sum_{p=1}^{P} n_p\cdot n_{p-1}\right)$
[144]	CA2C	$\mathcal{O}\left(N_b\cdot N^2\cdot M\right)$

critic network, $\imath_{Ac,i}$ is the number of the neurons in the i^{th} layer of the actor network, $\Re$, $\Im$ is the number of fully-connected layers in the actor and critic networks. The GSDRL algorithm in [117] has a complexity of $\mathcal{O}\left(2^{l_0}(L/d)^{2l_0}\right)$, where $l_0 = m_0 + N_0$ is the number of actions the UAV takes, m_0 is the number of SN, N_0 is the number of groups, L is the length of rectangle, d is the length of each grid. The complexity of the SAC-AO-RIS algorithm in [12] is $\mathcal{O}\left(B\times E\times T\left(2(K+2)h_1 + 2h_1h_2 + (K+3)h_2\right) + 8M\right)$ where k is the number of IoTDs, M is the number of RIS elements, h_1, h_2 are the numbers of two hidden fully-connected layers. The SARSA-based algorithm in [129] has a complexity of $\mathcal{O}\left(\left(3M + 2K + \left|\hat{\nu}\right|^M\right)\cdot N\cdot N_e\right)$ where k, M are the number of SNs and number of UAV, N is the number of steps, N_e is the number of episodes, $\hat{\nu}$ is the number of flight actions of each UAV. The PPO algorithm in [118] has a complexity of $\mathcal{O}\left(\sum_{q=1}^{Q-1} n_q\cdot n_{q-1}\right) \sim \mathcal{O}\left(n_q^2\right)$ where n_q is the number of neural units in the q^{th} hidden layer. Also, the DP-based algorithm in [142] has a complexity of $\mathcal{O}\left(2^M M^2\right)$ where M is the number of ground SNs. Similarly, the DDPG-based algorithm in [141] has a complexity of $\mathcal{O}\left(q^2\right)$ where q is the number of neurons. The DRL-based algorithm in [143] is characterized by a complexity of $\mathcal{O}\left(\sum_{p=1}^{P} n_p\cdot n_{p-1}\right)$, where n_p is the number of neural units in fully-connected layer p. The CA2C algorithm in [144] is characterized by a complexity of $\mathcal{O}\left(N_b\cdot N^2\cdot M\right)$ where N is the number of tasks, M is the number of UAVs, and N_b: sampling Batch size. The complexities of all the aforementioned algorithms are summarized in Table 15. In many of the works reviewed some of the algorithms can not be expressed as a formal Big-O notation due to the complex nature of the system.

E. PARAMETER SETTINGS OF DRL-BASED ALGORITHMS

This section discusses the parameter settings of DRL-based algorithms. The setting of the main DRL parameters are summarized in Table 16. The key parameter in DRL-based algorithm is the learning rate which has been set to different values in the reviewed algorithms. As shown in Table 16, the learning rate values range from 0.00008 [120] to 0.4 [13], [114]. By exploring the results of AoI in the reviewed algorithms, it can be concluded that the learning rate setting has an impact on the achieved AoI. For example, the authors in [136] examined the overall AoI of SAC-based algorithm in 6000 episodes for two different values of learning rates equal to 0.1 and 0.01 respectively. The results show that the higher value of learning rate (0.1) has achieved less AoI

which indicates that as the learning rate increases, the AoI is decreases. However, this relation can not be generalized since other algorithms such as T3D, PPO, Sarsa and DQN have achieved an AoI less than SAC with learning rates less than 0.01 [118], [119], [121], [129]. Thus, studying the impact of learning rate settings on AoI can be considered an open issue to be investigated deeply. Table 16 shows other parameters setting such as discount factor, number of hidden layers, number of neurons, activation function, exploring probability and optimizer techniques that have been used in the reviewed algorithms. These parameters are summarized here to provide the reader with a full vision of the possible settings for each parameter. This allows the researchers to open up new research questions on studying the impact of these settings on the results of the AoI parameter.

F. LIMITATIONS

The algorithms identified in Figures 19 and 20 generally fall under the different categories of DRL algorithms. In this section, we provide a general discussion of DRL algorithms worthy of consideration in different applications, including UAV-assisted IoT applications.

DRL algorithms have limitations such as poor training stability and overestimation especially during the cold start phase as agents need to interact continuously with the environment to learn the best strategies. Such agents could learn incorrect experiences at the early stages of training which could lead to failure in training [145]. Similarly, traditional DRL algorithms usually achieve a good result for the target optimization objective and cannot be generalized to objectives beyond the reward function [145]. The following are the limitations of the DRL algorithms commonly deployed in literature [146]:

- Although Q-learning is easy to use and straightforward, it is characterized by slow convergence and unsuitable for continuous action space.
- SARSA manages stochastic environments and policies, however, it is slow and difficult for continuous action space.
- Although DQN can handle high dimensional state space, it can potentially overestimate Q-values.
- Policy gradient algorithms can handle continuous action space and learn stochastic policies while optimizing non-differentiable objective functions. However, they are characterized by high variance in gradients and are hyper-parameter-sensitive.
- Actor-critic algorithms combine the advantage of both policy gradient and value-based schemes, thus, they can handle both discrete and continuous action space problems. However, it is difficult to balance exploration and exploitation and they have high variance to gradients.
- DDPG can handle continuous action space and high dimensional state space problems. Similarly, it can learn deterministic policies, however it is unstable and characterized by overestimation bias.

TABLE 16. Parameters setting of DRL-based algorithms.

Parameter	Value	Reference
Learning rate α	0.4	[114], [13]
	0.008	[130]
	0.002	[121]
	0.001	[119], [118], [124]
	0.0008	[120]
	0.0005	[133]
	0.0001	[119]
	0.00008	[120]
Discount Factor γ	0.99	[133]
	0.95	[119]
	0.90	[120], [118], [124], [5], [13], [130]
No. of hidden layers	100	[123]
	4	[133]
	3	[13], [124], [136]
	2	[119], [118]
No. of neurons	300	[119], [130]
	256	[121], [133]
	200	[121]
	128	[133], [136]
	64	[13], [124], [136], [133]
Activation Function	tanh	[119], [13]
	ReLU	[119], [121], [3], [130]
	Softmax	[13]
Number of episodes	200 000	[5]
	100 000	[130]
	20 000	[13]
	10 000	[120]
	6000	[124]
	100	[123]
	4	[124]
Exploring Probability ϵ	from 1 to 0.0005	[133]
	from 1 to 0.05	[133]
	from 0.9999 to 0.1	[120]
	from 0.9999 to 0.05	[130]
	from 0.9 to 0	[121]
Reply memory size	50 000	[130]
	40 000	[13], [120]
	10 000	[133]
	200	[123]
Mini-batch size	200	[121]
	128	[3]
	64	[13], [133]
Update policy length	300	[121]
	240	[118], [124]
	200	[130], [133], [120]
Greedy decrement	0.05	[13]
	0.0001	[121]
Clip fraction	0.2	[120], [124]
Optimizer technique	Adam	[121], [118], [124]
Loss coefficients	0.5 and 0.01	[118], [124]

- TD3 solves the overestimation bias problem in DDPG. However, it is unstable and requires careful hyperparameter tuning

VIII. CHALLENGES AND FUTURE CONSIDERATIONS

DRL has made significant contributions to UAV control from various points of view motivating researchers to go the extra mile to use these algorithms which require more time than classical control algorithms [21]. In this section, we discuss challenges and future considerations of research. Before providing further details, some of these

issues are summarized below.[7] Particularly setting up all the parameters and components of the MDP and choosing a DRL algorithm to optimize an important design objective is not as straightforward as it might seem even everything is well set up, defined and the algorithm converges well. A number of important considerations regarding how learned policy can be adjusted if the result is not satisfactory and the level of confidence of the designer that the deployed policy will work on real hardware is quite challenging.

The DRL is a process that is difficult to understand fine-grained details of how it works for solving a particular problem with respect to neurons, weights and biases, which makes it difficult to predict its performance in a new environment or when a particular parameter is changed. Similarly, a redesign of the entire MDP framework and perhaps retraining might be required to see the expected results in a different environment which is time consuming. Recent efforts have been working towards explainable DRL (see [147], [148], [149]) but this is yet to be fully explored within UAV-IoT. Another issue is how very realistic models can be designed especially in view of the huge dynamics in real settings such as noise and disturbances. Thus physical tests are important for UAVs especially for verification of models and tuning of reward functions. It is also important to ensure effective error handling to prevent unplanned occurence during practical deployments by creating a safe mode.

A. CHOICE OF ALGORITHMS

The choice of algorithms for solving specific problems largely determines the quality of obtained solutions with respect to several important factors such as stability, convergence and the amount of time required for learning. Different classes of algorithms have unique trade-offs and choosing the most suitable algorithm is not a trivial task. Some algorithms require huge computational time to run which may limit their applications for energy and computationally limited drones as well as reliability of the solution for the target application. For instance, DQN has a longer convergence time while PPO is characterized by faster convergence. Some algorithms are sensitive to changes such as PPO while others are more unstable such as DQN. Some other algorithms do very well with continuous action spaces e.g. TD3. It is thus important to design and deploy the most suitable algorithms in view of the several performance measures and considerations such as training time, convergence time, energy savings, amount of computation required, reliability, ability to handle continuous space and generalizability of solutions. For instance, applying SARSA is challenging for large networks where a much greater number of SNs and UAVs are participating [129]. Particularly, for the following reasons: 1) The value of the Q-table becomes intractable computationally as the state-action pairs increase significantly due to the large state-action space. 2) Communication overhead increases significantly since BS would have to collect information about the network from more UAVs and SN via over-control channels. 3) Due to the coverage limitation, UAVs might not be able to cover all SNs due to the limited coverage range as the network gets larger, which makes efficient data gathering from SNs very difficult. It would also be difficult for BS to perfectly know the sampling state of the lifetime of update packets, especially when SN operates in random sampling mode, which leads to poor performance loss.

The study of which algorithms are most suitable for various forms of UAV navigation such as propulsion, hovering, cruising and landing for more accurate prediction of the UAV behaviour and quality of communication links is imperative particularly because that would facilitate the practical deployment of these algorithms in several real-life applications

B. TRAINING

Training data is important for training DNN. However, it has its inherent challenges [13]. For instance, how to obtain a sufficient number of training data in a time-changing environment is challenging, despite being crucial for optimising the neural network. According to empirical evidence, independent training data can enhance stability and improve neural network convergence. As such, obtaining independent training data is another challenge in optimising the neural network. Thus, the experience replay and random sampling methods are adopted [13].

Most authors do not mention how long it took to train their models and the computational power of the simulation engines. Without a doubt, training DRL models could take a considerable amount of time particularly, since millions of iterations might be required to obtain meaningful results. This is especially due to the deep learning components of the algorithm [112] and the back propagation procedure. Techniques to improve the training procedure are thus required to make DRL solutions more attractive for both research and practice. Imitation learning is one potential approach that can be leveraged to shorten the training time by leveraging on the structural knowledge of efficient heuristic algorithms [150]. Both inadequate training and overtraining are not desirable for best results [151]. However, training is not as straightforward because of the need for hyperparameter tuning. It is also important to ensure training is tractable which can be achieved by reducing the action space (without compromising on model accuracy). How to effectively achieve this is thus important. Although training a single agent is time consuming, it is much more so for multi-agent systems [152]. Thus, handling multi-agent systems with reasonable training time and complexity is quite important. Besides the aforementioned it is also important to be able to make generalizations from training experiences to adapt to environmental changes [153] especially since a lot

[7] https://www.mathworks.com/videos/reinforcement-learning-part-5-overcoming-the-practical-challenges-of-reinforcement-learning-1558604830037.html

of uncertainties are bound to occur within the operational environment of the UAV. Finally, training can sometimes be frustrating if the reward function or action space is not properly designed, leading to failure in learning or overly long learning time.

C. CONVERGENCE

DRL suffers from slow convergence and higher search cost, especially in path planning problems in complex environments [154]. In addition to its slow convergence, it is also faced with the problem of overfitting and poor exploration in such environments. Particularly, off policy algorithms experiences longer convergence time due to the use of the epsilon greedy method and the variance between Q-values [155]. Also, in some cases, convergence to local optima occurs especially if the exploration is not sufficiently diverse in high dimensional space [153]. Similarly, uncertatinty estimation relies on function approximation and in situations with high dimensional state action, the estimation errors might not converge easily [156]. Convergence problem can also be as a result of unstable training and hence, a proper choice of algorithm is required for this, e.g. DQN converges slowly although its use of NN to approximately replace the Q function in order to reduce the dimension of input data [157].

D. CHOICE OF PARAMETERS AND HYPERPARAMETERS

It is essential to optimise the parameters of the neural network using a suitable loss function to obtain the optimal Q-function [13]. Some optimizers, such as the gradient descent algorithm and the Adam algorithm, can be used to obtain optimal neural network parameters based on the loss function and training data set [13].

DRL involves a huge amount of parameters which requires a large amount of data to train with huge computing resources. This makes it challenging to deploy in real-time, AoI sensitive applications especially for the resource limited resources particularly when they are expected to learn themselves. It is possible to use heuristic algorithms for some of the problems which DRL are deployed for, however, DRL performs much better overall than many heuristic algorithms, in fact, in all the results so far, DRL has yielded the best performance. For this reason, DRL can be deployed as a benchmark or an avenue to understand the weakness of existing heuristic algorithms [150] so that they can be improved for use in UAV based data gathering missions. In addition, DRL can be targeted at large scale and multi-dimensional optimization problems.

DRL needs several of its hyperparameters to be set accurately before and during training to obtain the best results during training. Several of the reviewed papers have not emphasized how their hyperparameters have been tuned or the impact of the hyperparameters. Generalization framework for understanding the impact of change in hyperparameters are required to enhance the understanding of the performance gains that can be achieved. Particularly, DRL parameter optimization and hyperparameter choices go a long way in predicting the accuracy of DRL algorithm. Also, the most suitable architecture, learning rate, discount factor, etc are all required to obtain optimal performance. This issue with hyperparameters makes optimization challenging, particularly because DRL is sensitive to these parameters.

For instance, learning rate is a key parameter in DRL algorithms. It controls the speed and quality of learning [115]. When the learning rate is too low, the convergence rate of learning is also slow, and the learning falls into a local optimal solution. On the other hand, when the learning rate is high, the learning effect is quite sensitive, and thus convergence is not achieved [115].

E. STABILITY

High stability is required for DRL algorithms. However, unfortunately, not all algorithms are very stable. For instance, while DQN has yielded promising results in the minimization of AoI, trajectory and energy efficiency optimization, DQN suffer from issues such as poor convergence and instability. To address these challenges, it is important to prioritize algorithms that are robust to underestimation and overestimation as well as ensure stability. This calls for the need for proper engineering of different deployed algorithms to perfectly ensure stability [158]. For instance, the target network for DQN has to be updated at regular intervals whereas unstable jumps can be experienced during the updated intervals due to changes in parameter values during learning [151]. This makes it quite difficult to see a clear trend in the learning process. New updates should thus ensure policy learning maintains stability with low latency.

F. REAL DEPLOYMENTS

Implementation should be more application specific e.g. wildfire monitoring, rescue mission, exploration [21] with more realistic assumptions and very close-to-practice system models. For instance, UAVs would require precise navigation through obstacles to prevent collisions in many cases. Modeling real-world scenarios is quite complex and a lot of practical systems do not have separate training and evaluation environments [112]. As of writing, no real-implementation of DRL in a field setting for AoI-sensitive application has been reported in literature and as such all the reviewed works have run on simulation mode. It is thus difficult to predict some of the practical impairments towards successful implementation in real environments due to the huge gap between simulation and real deployments [159]. For instance, hardware resources and their uncertainties, robustness to faults, mismatched training and testing conditions, sample inefficiency when the learned policies are transferred to physical UAVs or learning policies directly on the hardware. A successful transfer of DRL agent from simulation to real environment in a safe, secure and efficient fail-safe and rigorous manner is quite challenging and needs to be validated thoroughly especially for critical AoI-sensitive applications. Although simulation

may provide experience that may be generalized in real world, it might not capture some of the details in real settings such as wind, etc. Environmental conditions include rain, wind speed, dust should be considered as part of the system disturbance [21]. It is thus important to ensure training experience can adapt to the uncertainties in the operation environment. This can be achieved via a huge number of trials. It is also possible to have drone monitoring in indoor environments too (with altitude control). The height control can be considered as one of the inputs of such models. Also, object avoidance is another important component. In a nutshell, researchers could explore real-life implementation of more advanced algorithms to facilitate smooth autonomous UAV navigation along with improved exploration for UAV control and navigation [21].

G. EXPLORATION

Another major challenge faced by DRL-based methods in wireless networks and UAV communications is the inherent exploration of RL algorithms. Exploration is necessary for learning, however, it can be quite time-consuming. Similarly, it is quite hard due to environmental challenges such as large state-action space, as its difficulty grows with increase in state-action space. For instance, real drones could require sensors for navigation and might also receive images in form of sensory input for navigation. At times the state space might have complex underlying structures which makes some states more difficult to access compared to other states with causal dependencies among states. Actions could also be combinatorial in nature or a hybrid of discrete and continuous thus making exploration more challenging. In multi-agent systems, exploration can also pose difficulty because local information is not sufficient to achieve coordinated exploration among agents. Global information is necessary but there exist inconsistencies between global and local perspectives which if not balanced can lead to redundant or inadequate exploration [156].

The common approach used in the literature is the epsilon-greedy approach [160], however, this method requires parameter tuning. Interestingly, there are several other unexplored approaches which have been reported in literature [156]: for instance using Bolzman exploration in which an agent draws action from Bolzman equation over its Q-values. The problem with this method however, is that it cannot be applied to continuous action state spaces. Another approach towards exploration is the use of Upper Confidence Bound (UCB) which measures the potential of each action by UCB of the reward expectation rather than using naive random exploration. Entropy regularization can also be used to promote exploration, especially for RL algorithms with a stochastic policy. The objective here is to encourage the agent to take diverse actions. However, regularization might deviate from the original optimization objective, which can be addressed by decreasing the influence of entropy regularization in the learning process. Bayesian optimization is another approach that can be employed for exploration to find the position of the global optimum. However, acquisition functions are required for efficient exploration to suit better selection.

Safe exploration is another important consideration, especially during the training process. Safety requirements are important in the evaluation, deployment, and sampling interaction. Finally, coordinated exploration in multi-agent systems for balancing local and global information is another promising area to explore. Research is required on multi-agent exploration, particularly with regards to partial observation, non-stationary high dimensional state-action space, and coordination.

H. COMPLEXITY

DRL can memorize specific sequence of actions or experiences which can lead to poor performance in certain unexpected circumstances. Techniques to improve this trades-off the solution such as regularization for higher complexity. DRL can be difficult to interpret and debug due to its black box nature. Models are complex and it is challenging to interpret due to the unknown internal workings of the learning process. This is more sensitive in safety-critical applications where agents cannot afford to make mistakes. Modeling real-world scenarios is complex and while some environments may be well understood, the complex and stochastic nature of the environmental dynamics makes it difficult to deploy the hand crafted solutions. For the case of DRL, a carefully designed reward function is required which might be very challenging in multi-objective non-linear problems especially those that requires assigning weights and considering trade-offs for complex problems. Other complex problems are problems where the state space has a complex structure. Exploration also adds to the complexity because of the combinatorial or hybrid discrete-continuous actions. The potential of using techniques such as entanglement quantum computing for achieving lower complexity can be explored, particularly, in view of the fact that classical computing is usually deployed despite DRL scales poorly with increase in problem size and complexity.

I. COMPUTATION

The common use of DRL is propelled by the existence of more affordable computational power e.g. Google Colab. UAVs are quite small compared to other vehicles and thus they have low memory and energy capacities thus they also have relatively low computational power. For this reason DRL may not be practical in certain use cases because implementation of optimization and learning-based AI require high computational power which is challenging to overcome. Thus it is important to reduce computational demands of the training in various neural network architectures, since significant computation power is required for training. Although GPU can be used, they are quite expensive despite Tensorflow and Pytorch can optimize between CPU and GPU, still there is an obvious trade-off between computation

and financial cost. This also applies to the case with multi-agent DRL-based schemes.

J. STATE-ACTION SPACE DIMENSION

Several problems are experienced due to high dimensional space and actions. For instance, most methods are or may not able to learn effectively in a large and complex action spaces as they only consider small discrete action space or low dimensional continuous action space. However, real world scenarios involve a large number of discrete or hybrid discrete and continuous action space which is quite complex [156]. Designing state, action and reward especially in high dimension state and high dimension action and sparse rewards are important [161]. Also, DRL suffers from convergence to local optima when operating in high dimensional space. Drones are faced with problems of high dimensions and continuous action nature especially for navigation, thus policy based DRL solutions are most suitable [162]. Nonetheless, both policy and value based schemes are commonly used.

DQN algorithm often output discrete actions because action selection is based on action-value function and are only suitable for discrete action space. Problems in this regard affects choice of network structure, space exploration, sample efficiency, and leads to overestimation of action-value function [161].

Whenever the transition probabilities in a model are not known, and the problem has a very large state-action space, traditional solutions for solving MDP, such as value and policy iteration becomes quite challenging [120]. In many large-scale scenarios with much larger node density (see [1] for example), DQN cannot learn optimal scheduling, thus other methods such as the use of a long short-term memory (LSTM)-based autoencoder are required to map the state space to a fixed-size vector representation to facilitate the AoI minimization.

Furthermore, it is hard for RL to learn with limited sample data. For instance, when using DRL, several millions of interactions are usually required for relatively simple problems, thus limiting its wide use in real-industry scenarios. Partially the question of how efficiently it explores the environment and collects information about information experiences could assist learning towards optimal policies, especially in complex environments with sparse rewards and noisy distractions.

Finding a control policy that effectively controls several closely related design components, such as the UAV altitude, the scheduling policy, and the RIS phase shift matrix can be quite challenging as shown in [118]. Such problems can be formulated as a hybrid discrete-continuous action space problem but it is quite difficult to combine these diverse actions into a single or one action space because of the large number of possible actions that should be considered. Efficient DRL algorithms would still be the best to solve large discrete action spaces as they increase the difficulty of learning, i.e., make learning hard [118].

IX. CONCLUSION

Reinforcement learning algorithms have become a revolutionary tool in several applications, permeating cutting edge applications in the last few years. Particularly, in modern research, they have become so popular and have provided solutions to extremely difficult and multi-dimensional problems in general sciences, medicine, engineering, agriculture to name a few. The use of reinforcement learning has also become a huge trend within the last few years. Particularly, it has proven to be successful for optimizing UAV navigation and control. In this paper, we consider the application of DRL for solving problems related to data and time-sensitive (albeit AoI-sensitive) applications which could span across several monitoring applications in IoT. Using DRL, the UAV can learn to perform actions that can optimize the designer's target objectives such as trajectory, scheduling of sensor nodes, learning traffic patterns, best hovering spots etc.

To provide an in-depth review of the use of DRL as a problem-solving tool in AoI-sensitive UAV-assisted IoT network architectures, the paper provides a background on several real applications of UAV-assisted IoT such as monitoring, smart city, data gathering, security, health, agriculture, and disaster management, some of which are very time sensitive. Then, a background on reinforcement learning and DRL is provided. Subsequently, the proposals relating to the studied works were classified, and briefly introduced.

Subsequently, we provide a discussion on target objectives, discuss the simulation libraries, algorithm complexity for several of the algorithms and most influential simulation parameters. Finally, we conclude with the challenges and future research directions that can be explored on this subject with emphasis on choice of algorithms, training, convergence, choice of parameters, stability, need for real testbeds, algorithm stability, exploration-exploitation dichotomy, complexity, computation, and dimension with respect to state-action space. In summary, RL is a robust framework that can be applied to achieve several vital objectives for the diverse applications of UAV-assisted IoT. As such aspects such as security and mobile edge computing, the use of non orthoganal multiple access, etc. As DRL is useful for UAVs, unmanned underwater vehicles can also benefit from DRL algorithms for planning their movement underwater and collecting data from underwater sensor networks.

ACKNOWLEDGMENT

The authors would like to thank everyone who provided valuable suggestions and support to improve the content, quality, and presentation of this article.

REFERENCES

[1] A. Ferdowsi, M. A. Abd-Elmagid, W. Saad, and H. S. Dhillon, "Neural combinatorial deep reinforcement learning for age-optimal joint trajectory and scheduling design in UAV-assisted networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1250–1265, May 2021.

[2] M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "On the role of age of information in the Internet of Things," *IEEE Commun. Mag.*, vol. 57, no. 12, pp. 72–77, Dec. 2019.

[3] C. Zhou, H. He, P. Yang, F. Lyu, W. Wu, N. Cheng, and X. Shen, "Deep RL-based trajectory planning for AoI minimization in UAV-assisted IoT," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[4] A. Munari and L. Badia, "The role of feedback in AoI optimization under limited transmission opportunities," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2022, pp. 1972–1977.

[5] Q. Dang, Q. Cui, Z. Gong, X. Zhang, X. Huang, and X. Tao, "AoI oriented UAV trajectory planning in wireless powered IoT networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 884–889.

[6] N. Pappas, M. A. Abd-Elmagid, B. Zhou, W. Saad, and H. S. Dhillon, *Age of Information: Foundations and Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2022.

[7] İ. Kahraman, A. Köse, M. Koca, and E. Anarim, "Age of information in Internet of Things: A survey," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 9896–9914, Mar. 2024.

[8] O. A. Amodu, U. A. Bukar, R. A. R. Mahmood, C. Jarray, and M. Othman, "Age of information minimization in UAV-aided data collection for WSN and IoT applications: A systematic review," *J. Netw. Comput. Appl.*, vol. 216, Jul. 2023, Art. no. 103652.

[9] O. Amodu, R. Nordin, C. Jarray, U. Bukar, R. R. Mahmood, and M. Othman, "A survey on the design aspects and opportunities in age-aware UAV-aided data collection for sensor networks and Internet of Things applications," *Drones*, vol. 7, no. 4, p. 260, Apr. 2023.

[10] U. A. Bukar, M. S. Sayeed, S. F. A. Razak, S. Yogarayan, and O. A. Amodu, "An exploratory bibliometric analysis of the literature on the age of information-aware unmanned aerial vehicles aided communication," *Informatica*, vol. 47, no. 7, pp. 91–114, Aug. 2023.

[11] J. Zhang, Y. Yu, Z. Wang, S. Ao, J. Tang, X. Zhang, and K.-K. Wong, "Trajectory planning of UAV in wireless powered IoT system based on deep reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jun. 2020, pp. 645–650.

[12] X. Fan, M. Liu, Y. Chen, S. Sun, Z. Li, and X. Guo, "RIS-assisted UAV for fresh data collection in 3D urban environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 632–647, Jan. 2023.

[13] X. Li, J. Li, and D. Liu, "Energy-efficient UAV trajectory design with information freshness constraint via deep reinforcement learning," *Mobile Inf. Syst.*, vol. 2021, pp. 1–9, Dec. 2021.

[14] X. Wang, M. Yi, J. Liu, Y. Zhang, M. Wang, and B. Bai, "Cooperative data collection with multiple UAVs for information freshness in the Internet of Things," *IEEE Trans. Commun.*, vol. 71, no. 5, pp. 2740–2755, May 2023.

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

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

[17] Q. Abbas, S. A. Hassan, H. K. Qureshi, K. Dev, and H. Jung, "A comprehensive survey on age of information in massive IoT networks," *Comput. Commun.*, vol. 197, pp. 199–213, Jan. 2023.

[18] B. Yu, X. Chen, and Y. Cai, "Age of information for the cellular Internet of Things: Challenges, key techniques, and future trends," *IEEE Commun. Mag.*, vol. 60, no. 12, pp. 20–26, Dec. 2022.

[19] S. Wang, Q. Sun, and H. Wang, "A survey on the optimisation of age of information in wireless networks," *Int. J. Web Grid Services*, vol. 19, no. 1, pp. 1–33, 2023.

[20] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[21] A. T. Azar, A. Koubaa, N. Ali Mohamed, H. A. Ibrahim, Z. F. Ibrahim, M. Kazim, A. Ammar, B. Benjdira, A. M. Khamis, I. A. Hameed, and G. Casalino, "Drone deep reinforcement learning: A review," *Electronics*, vol. 10, no. 9, p. 999, 2021.

[22] F. AlMahamid and K. Grolinger, "Autonomous unmanned aerial vehicle navigation using reinforcement learning: A systematic review," *Eng. Appl. Artif. Intell.*, vol. 115, Oct. 2022, Art. no. 105321.

[23] F. Frattolillo, D. Brunori, and L. Iocchi, "Scalable and cooperative deep reinforcement learning approaches for multi-UAV systems: A systematic review," *Drones*, vol. 7, no. 4, p. 236, Mar. 2023.

[24] M. Vaezi, X. Lin, H. Zhang, W. Saad, and H. V. Poor, "Deep reinforcement learning for interference management in UAV-based 3D networks: Potentials and challenges," 2023, *arXiv:2305.07069*.

[25] M. Landers and A. Doryab, "Deep reinforcement learning verification: A survey," *ACM Comput. Surveys*, vol. 55, no. 14s, pp. 1–31, Dec. 2023.

[26] Y. Wu, Z. Wang, Y. Ma, and V. C. M. Leung, "Deep reinforcement learning for blockchain in industrial IoT: A survey," *Comput. Netw.*, vol. 191, May 2021, Art. no. 108004.

[27] M. Glavic, "(Deep) reinforcement learning for electric power system control and related problems: A short review and perspectives," *Annu. Rev. Control*, vol. 48, pp. 22–35, 2019, doi: 10.1016/j.arcontrol.2019.09.008.

[28] T.-H. Nguyen and L. Park, "A survey on deep reinforcement learning-driven task offloading in aerial access networks," in *Proc. 13th Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2022, pp. 822–827.

[29] M. S. Frikha, S. M. Gammar, A. Lahmadi, and L. Andrey, "Reinforcement and deep reinforcement learning for wireless Internet of Things: A survey," *Comput. Commun.*, vol. 178, pp. 98–113, Oct. 2021.

[30] O. Jogunola, B. Adebisi, A. Ikpehai, S. I. Popoola, G. Gui, H. Gacanin, and S. Ci, "Consensus algorithms and deep reinforcement learning in energy market: A review," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4211–4227, Mar. 2021.

[31] H. Sun, W. Zhang, R. Yu, and Y. Zhang, "Motion planning for mobile robots—Focusing on deep reinforcement learning: A systematic review," *IEEE Access*, vol. 9, pp. 69061–69081, 2021.

[32] B. Cunha, A. M. Madureira, B. Fonseca, and D. Coelho, "Deep reinforcement learning as a job shop scheduling solver: A literature review," in *Proc. 18th Int. Conf. Hybrid Intell. Syst. (HIS)*, Porto, Portugal. Cham, Switzerland: Springer, Dec. 2020, pp. 350–359, doi: 10.1007/978-3-030-14347-3_34.

[33] R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel, "A survey of zero-shot generalisation in deep reinforcement learning," *J. Artif. Intell. Res.*, vol. 76, pp. 201–264, Jan. 2023.

[34] H. T. Nguyen, M. T. Nguyen, H. T. Do, H. T. Hua, and C. V. Nguyen, "DRL-based intelligent resource allocation for diverse QoS in 5G and toward 6G vehicular networks: A comprehensive survey," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–21, Aug. 2021.

[35] D. Popescu, F. Stoican, G. Stamatescu, O. Chenaru, and L. Ichim, "A survey of collaborative UAV–WSN systems for efficient monitoring," *Sensors*, vol. 19, no. 21, p. 4690, Oct. 2019.

[36] N. Baifeng and Z. Ke, "Design and implementation of Internet of Things+UAV flight monitoring and management system," in *Proc. IEEE 4th Int. Conf. Autom., Electron. Electr. Eng. (AUTEEE)*, Nov. 2021, pp. 404–408.

[37] M. A. Al-Mashhadani, M. M. Hamdi, and A. S. Mustafa, "Role and challenges of the use of UAV-aided WSN monitoring system in large-scale sectors," in *Proc. 3rd Int. Congr. Human-Computer Interact., Optim. Robotic Appl. (HORA)*, Jun. 2021, pp. 1–5.

[38] R. Macharia, K. Lang'at, and P. Kihato, "Interference management upon collaborative beamforming in a wireless sensor network monitoring system featuring multiple unmanned aerial vehicles," in *Proc. IEEE AFRICON*, Sep. 2021, pp. 1–6.

[39] M. Salhaoui, A. Guerrero-González, M. Arioua, F. Ortiz, A. El Oualkadi, and C. Torregrosa, "Smart industrial IoT monitoring and control system based on UAV and cloud computing applied to a concrete plant," *Sensors*, vol. 19, no. 15, p. 3316, Jul. 2019.

[40] R. Basak, I. Pal, A. Bandyopadhayay, and T. Guha, "IoT based drone operated monitoring of distribution transformers and terminating illegal power connections," in *Proc. 3rd Int. Conf. Electron., Mater. Eng. Nano-Technology (IEMENTech)*, Aug. 2019, pp. 1–5.

[41] H. Izadi Moud and M. Gheisari, "Coupling wireless sensor networks and unmanned aerial vehicles in bridge health monitoring systems," in *Proc. Int. Symp. Autom. Robot. Construction (IAARC)*, Jul. 2016, pp. 267–273.

[42] P. D. Mankar, Z. Chen, M. A. Abd-Elmagid, N. Pappas, and H. S. Dhillon, "Throughput and age of information in a cellular-based IoT network," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8248–8263, Dec. 2021.

[43] J.-I. Hernández-Vega, E. R. Varela, N. H. Romero, C. Hernández-Santos, J. L. S. Cuevas, and D. G. P. Gorham, "Internet of Things (IoT) for monitoring air pollutants with an unmanned aerial vehicle (UAV) in a smart city," in *Smart Technology*. Cham, Switzerland: Springer, 2018, pp. 108–120, doi: 10.1007/978-3-319-73323-4_11.

[44] N. Kalatzis, M. Avgeris, D. Dechouniotis, K. Papadakis-Vlachopapadopoulos, I. Roussaki, and S. Papavassiliou, "Edge computing in IoT ecosystems for UAV-enabled early fire detection," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2018, pp. 106–114.

[45] D. Popescu, C. Dragana, F. Stoican, L. Ichim, and G. Stamatescu, "A collaborative UAV-WSN network for monitoring large areas," *Sensors*, vol. 18, no. 12, p. 4202, Nov. 2018.

[46] C. Trasviña-Moreno, R. Blasco, Á. Marco, R. Casas, and A. Trasviña-Castro, "Unmanned aerial vehicle based wireless sensor network for marine-coastal environment monitoring," *Sensors*, vol. 17, no. 3, p. 460, Feb. 2017.

[47] B. Potter, G. Valentino, L. Yates, T. Benzing, and A. Salman, "Environmental monitoring using a drone-enabled wireless sensor network," in *Proc. Syst. Inf. Eng. Design Symp. (SIEDS)*, Apr. 2019, pp. 1–6.

[48] M. Zhang and X. Li, "Drone-enabled Internet-of-Things relay for environmental monitoring in remote areas without public networks," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7648–7662, Aug. 2020.

[49] S. Ghosh, K. Ghosh, S. Karamakar, S. Prasad, N. Debabhuti, P. Sharma, B. Tudu, N. Bhattacharyya, and R. Bandyopadhyay, "Development of an IoT based robust architecture for environmental monitoring using UAV," in *Proc. IEEE 16th India Council Int. Conf. (INDICON)*, Dec. 2019, pp. 1–4.

[50] L. Angrisani, A. Amodio, P. Arpaia, M. Asciolla, A. Bellizzi, F. Bonavolontà, R. Carbone, E. Caputo, G. Karamanolis, V. Martire, M. Marvaso, R. Peirce, A. Picardi, G. Termo, A. M. Toni, G. Viola, and A. Zimmaro, "An innovative air quality monitoring system based on drone and IoT enabling technologies," in *Proc. IEEE Int. Workshop Metrology Agricult. Forestry (MetroAgriFor)*, Oct. 2019, pp. 207–211.

[51] O. M. Bushnaq, A. Chaaban, and T. Y. Al-Naffouri, "The role of UAV-IoT networks in future wildfire detection," *IEEE Internet Things J.*, vol. 8, no. 23, pp. 16984–16999, Dec. 2021.

[52] G. B. Gaggero, M. Marchese, A. Moheddine, and F. Patrone, "A possible smart metering system evolution for rural and remote areas employing unmanned aerial vehicles and Internet of Things in smart grids," *Sensors*, vol. 21, no. 5, p. 1627, Feb. 2021.

[53] Z. Hu, Z. Bai, Y. Yang, Z. Zheng, K. Bian, and L. Song, "UAV aided aerial-ground IoT for air quality sensing in smart city: Architecture, technologies, and implementation," *IEEE Netw.*, vol. 33, no. 2, pp. 14–22, Mar. 2019.

[54] P. V. Pravija Raj, A. M. Khedr, and Z. Al Aghbari, "EDGO: UAV-based effective data gathering scheme for wireless sensor networks with obstacles," *Wireless Netw.*, vol. 28, no. 6, pp. 2499–2518, Aug. 2022.

[55] S. Singh, A. Malik, R. Kumar, and P. K. Singh, "A proficient data gathering technique for unmanned aerial vehicle-enabled heterogeneous wireless sensor networks," *Int. J. Commun. Syst.*, vol. 34, no. 16, p. e4956, 2021.

[56] G. P. Gupta, V. K. Chawra, and S. Dewangan, "Optimal path planning for UAV using NSGA-II based metaheuristic for sensor data gathering application in wireless sensor networks," in *Proc. IEEE Int. Conf. Adv. Netw. Telecommun. Syst. (ANTS)*, Dec. 2019, pp. 1–5.

[57] D. Ebrahimi, S. Sharafeddine, P.-H. Ho, and C. Assi, "UAV-aided projection-based compressive data gathering in wireless sensor networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1893–1905, Apr. 2019.

[58] I. Cardei, M. Cardei, and R. Papa, "UAV-enabled data gathering in wireless sensor networks," in *Proc. IEEE 37th Int. Perform. Comput. Commun. Conf. (IPCCC)*, Nov. 2018, pp. 1–8.

[59] H. Yomo, A. Asada, and M. Miyatake, "On-demand data gathering with a drone-based mobile sink in wireless sensor networks exploiting wake-up receivers," *IEICE Trans. Commun.*, vol. 101, no. 10, pp. 2094–2103, 2018.

[60] C. Y. Tazibt, M. Bekhti, T. Djamah, N. Achir, and K. Boussetta, "Wireless sensor network clustering for UAV-based data gathering," in *Proc. Wireless Days*, Mar. 2017, pp. 245–247.

[61] S. Say, M. E. Ernawan, and S. Shimamoto, "Cooperative path selection framework for effective data gathering in UAV-aided wireless sensor networks," *IEICE Trans. Commun.*, vol. 99, no. 10, pp. 2156–2167, 2016.

[62] A.-V. Vladuta, C. Grumazescu, I. Bica, and V. Patriciu, "Energy considerations for data gathering protocol in wireless sensor networks using unmanned aerial vehicles," in *Proc. Int. Conf. Commun. (COMM)*, Jun. 2016, pp. 237–240.

[63] S. Say, H. Inata, J. Liu, and S. Shimamoto, "Priority-based data gathering framework in UAV-assisted wireless sensor networks," *IEEE Sensors J.*, vol. 16, no. 14, pp. 5785–5794, Jul. 2016.

[64] S. Birtane and Ö. K. Sahingöz, "Data gathering from a large scale wireless sensor network with means of unmanned aerial vehicles," in *Proc. 24th Signal Process. Commun. Appl. Conf. (SIU)*, May 2016, pp. 2105–2108.

[65] S. Sotheara, K. Aso, N. Aomi, and S. Shimamoto, "Effective data gathering and energy efficient communication protocol in wireless sensor networks employing UAV," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 2342–2347.

[66] H. Zanjie, N. Hiroki, K. Nei, O. Fumie, M. Ryu, and Z. Baohua, "Resource allocation for data gathering in UAV-aided wireless sensor networks," in *Proc. 4th IEEE Int. Conf. Netw. Infrastructure Digit. Content*, Sep. 2014, pp. 11–16.

[67] M. Dong, K. Ota, M. Lin, Z. Tang, S. Du, and H. Zhu, "UAV-assisted data gathering in wireless sensor networks," *J. Supercomput.*, vol. 70, no. 3, pp. 1142–1155, Dec. 2014.

[68] S. V. Kashuba, V. I. Novikov, O. I. Lysenko, and I. V. Alekseeva, "Optimization of UAV path for wireless sensor network data gathering," in *Proc. IEEE Int. Conf. Actual Problems Unmanned Aerial Vehicles Develop. (APUAVD)*, Oct. 2015, pp. 280–283.

[69] A. Gohari, A. B. Ahmad, R. B. A. Rahim, A. S. M. Supa'at, S. Abd Razak, and M. S. M. Gismalla, "Involvement of surveillance drones in smart cities: A systematic review," *IEEE Access*, vol. 10, pp. 56611–56628, 2022.

[70] S. Berrahal, J.-H. Kim, S. Rekhis, N. Boudriga, D. Wilkins, and J. Acevedo, "Border surveillance monitoring using quadcopter UAV-aided wireless sensor networks," *J. Commun. Softw. Syst.*, vol. 12, no. 1, p. 67, Mar. 2016.

[71] D. Sikeridis, E. E. Tsiropoulou, M. Devetsikiotis, and S. Papavassiliou, "Wireless powered public safety IoT: A UAV-assisted adaptive-learning approach towards energy efficiency," *J. Netw. Comput. Appl.*, vol. 123, pp. 69–79, Dec. 2018.

[72] H. Ali, L. Y. Hang, T. Y. Suan, V. R. Polaiah, M. I. F. Aluwi, A. A. M. Zabidi, and M. Elshaikh, "Development of surveillance drone based Internet of Things (IoT) for industrial security applications," *J. Phys., Conf. Ser.*, vol. 2107, no. 1, Nov. 2021, Art. no. 012018.

[73] V. A. Dovgal, "A scheme of data analysis by sensors of a swarm of drones performing a search mission based on a fog architecture using the Internet of Things," in *Proc. Int. Conf. Ind. Eng., Appl. Manuf. (ICIEAM)*, May 2022, pp. 1073–1078.

[74] Y. Li, M. Scanavino, E. Capello, F. Dabbene, G. Guglieri, and A. Vilardi, "A novel distributed architecture for UAV indoor navigation," in *Proc. Int. Conf. Air Transp. (INAIR)*, vol. 35, 2018, pp. 13–22, doi: 10.1016/j.trpro.2018.12.003.

[75] T. T. Mac, C. Copot, R. D. Keyser, and C. M. Ionescu, "The development of an autonomous navigation system with optimal control of an UAV in partly unknown indoor environment," *Mechatronics*, vol. 49, pp. 187–196, Feb. 2018.

[76] J. M. S. Lagmay, L. J. C. Leyba, A. T. Santiago, L. B. Tumabotabo, W. J. R. Limjoco, and N. M. C. Tiglao, "Automated indoor drone flight with collision prevention," in *Proc. IEEE Region 10 Conf. (TENCON)*, Jun. 2018, pp. 1762–1767.

[77] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV-based IoT platform: A crowd surveillance use case," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 128–134, Feb. 2017.

[78] A. Bahr, M. A. Mehaseb, S. A. Doliel, S. El-Rabaie, and F. E. Abd El-Samie, "Power-aware 3D UAV placement for IoT emergency communications," in *Proc. 8th Int. Japan-Africa Conf. Electron., Commun., Computations (JAC-ECC)*, Dec. 2020, pp. 18–23.

[79] M. Rouault, W. Ejaz, M. Naeem, and R. Masroor, "The role of UAV-assisted IoT networks in managing the impact of the pandemic," *IEEE Commun. Standards Mag.*, vol. 5, no. 4, pp. 10–16, Dec. 2021.

[80] F. G. Costa, J. Ueyama, T. Braun, G. Pessin, F. S. Osório, and P. A. Vargas, "The use of unmanned aerial vehicles and wireless sensor network in agricultural applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2012, pp. 5045–5048.

[81] M. M. Vihari, U. Rani. Nelakuditi, and M. P. Teja, "IoT based unmanned aerial vehicle system for agriculture applications," in *Proc. Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Dec. 2018, pp. 26–28.

[82] T. Moribe, H. Okada, K. Kobayashl, and M. Katayama, "Combination of a wireless sensor network and drone using infrared thermometers for smart agriculture," in *Proc. 15th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2018, pp. 1–2.

[83] P. Radoglou-Grammatikis, P. Sarigiannidis, T. Lagkas, and I. Moscholios, "A compilation of UAV applications for precision agriculture," *Comput. Netw.*, vol. 172, May 2020, Art. no. 107148.
[84] N. Islam, M. M. Rashid, F. Pasandideh, B. Ray, S. Moore, and R. Kadel, "A review of applications and communication technologies for Internet of Things (IoT) and unmanned aerial vehicle (UAV) based sustainable smart farming," *Sustainability*, vol. 13, no. 4, p. 1821, Feb. 2021.
[85] S. Punia, H. Krishna, V. N. B, and A. Sajjad, "Agrosquad—An IoT based precision agriculture using UAV and low-power soil multi-sensor," in *Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT)*, Jul. 2021, pp. 1–6.
[86] B. S. Faiçal, F. G. Costa, G. Pessin, J. Ueyama, H. Freitas, A. Colombo, P. H. Fini, L. Villas, F. S. Osório, P. A. Vargas, and T. Braun, "The use of unmanned aerial vehicles and wireless sensor networks for spraying pesticides," *J. Syst. Archit.*, vol. 60, no. 4, pp. 393–404, Apr. 2014.
[87] A. Rao, H. Shao, and X. Yang, "The design and implementation of smart agricultural management platform based on UAV and wireless sensor network," in *Proc. IEEE 2nd Int. Conf. Electron. Technol. (ICET)*, May 2019, pp. 248–252.
[88] P. K. Singh and A. Sharma, "An intelligent WSN-UAV-based IoT framework for precision agriculture application," *Comput. Electr. Eng.*, vol. 100, May 2022, Art. no. 107912.
[89] A. Mukherjee, S. Misra, A. Sukrutha, and N. S. Raghuwanshi, "Distributed aerial processing for IoT-based edge UAV swarms in smart farming," *Comput. Netw.*, vol. 167, Feb. 2020, Art. no. 107038.
[90] J. Polo, G. Hornero, C. Duijneveld, A. García, and O. Casas, "Design of a low-cost wireless sensor network with UAV mobile node for agricultural applications," *Comput. Electron. Agricult.*, vol. 119, pp. 19–32, Nov. 2015.
[91] F. A. Almalki, B. O. Soufiene, S. H. Alsamhi, and H. Sakli, "A low-cost platform for environmental smart farming monitoring system based on IoT and UAVs," *Sustainability*, vol. 13, no. 11, p. 5908, May 2021.
[92] G. E. Just, M. E. Pellenz, L. A. D. P. Lima, B. S. Chang, R. Demo Souza, and S. Montejo-Sánchez, "UAV path optimization for precision agriculture wireless sensor networks," *Sensors*, vol. 20, no. 21, p. 6098, Oct. 2020.
[93] L. García, L. Parra, J. M. Jimenez, J. Lloret, P. V. Mauri, and P. Lorenz, "DronAway: A proposal on the use of remote sensing drones as mobile gateway for WSN in precision agriculture," *Appl. Sci.*, vol. 10, no. 19, p. 6668, Sep. 2020.
[94] G. Ristorto, A. Bojeri, G. Scarton, G. Giannotta, and G. Guglieri, "DroneONtrap project–integration of IoT technologies and drones for health status and pests monitoring of orchards," in *Proc. IEEE Int. Workshop Metrology Agricult. Forestry (MetroAgriFor)*, Nov. 2021, pp. 12–16.
[95] S. Jayanthy, G. Kiruthika, G. Lakshana, and M. Pragatheshwaran, "Early cotton plant disease detection using drone monitoring and deep learning," in *Proc. IEEE Int. Conf. Women Innov., Technol. Entrepreneurship (ICWITE)*, Feb. 2024, pp. 625–630.
[96] M. A. Uddin, A. Mansour, D. L. Jeune, M. Ayaz, and E.-H.-M. Aggoune, "UAV-assisted dynamic clustering of wireless sensor networks for crop health monitoring," *Sensors*, vol. 18, no. 2, p. 555, Feb. 2018.
[97] J. Xu, G. Solmaz, R. Rahmatizadeh, D. Turgut, and L. Bölöni, "Animal monitoring with unmanned aerial vehicle-aided wireless sensor networks," in *Proc. IEEE 40th Conf. Local Comput. Netw. (LCN)*, Oct. 2015, pp. 125–132.
[98] D. Gao, Q. Sun, B. Hu, and S. Zhang, "A framework for agricultural pest and disease monitoring based on Internet-of-Things and unmanned aerial vehicles," *Sensors*, vol. 20, no. 5, p. 1487, Mar. 2020.
[99] D. Popescu, F. Stoican, G. Stamatescu, L. Ichim, and C. Dragana, "Advanced UAV–WSN system for intelligent monitoring in precision agriculture," *Sensors*, vol. 20, no. 3, p. 817, Feb. 2020.
[100] S. Duangsuwan, C. Teekapakvisit, and M. M. Maw, "Development of soil moisture monitoring by using IoT and UAV-SC for smart farming application," *Adv. Sci., Technol. Eng. Syst. J.*, vol. 5, no. 4, pp. 381–387, 2020.
[101] A. Mitra, B. Bera, and A. K. Das, "Design and testbed experiments of public blockchain-based security framework for IoT-enabled drone-assisted wildlife monitoring," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2021, pp. 1–6.
[102] B. Etikasari, Husin, S. Kautsar, H. Y. Riskiawan, and D. P. S. Setyohadi, "Wireless sensor network development in unmanned aerial vehicle (uav) for water quality monitoring system," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 411, no. 1, Jan. 2020, Art. no. 012061.
[103] A. H. Wheeb, R. Nordin, A. A. Samah, and D. Kanellopoulos, "Performance evaluation of standard and modified OLSR protocols for uncoordinated UAV ad-hoc networks in search and rescue environments," *Electronics*, vol. 12, no. 6, p. 1334, Mar. 2023.
[104] M. Erdelj, M. Król, and E. Natalizio, "Wireless sensor networks and multi-UAV systems for natural disaster management," *Comput. Netw.*, vol. 124, pp. 72–86, Sep. 2017.
[105] A. S. Nandan, S. Singh, A. Malik, and R. Kumar, "A green data collection & transmission method for IoT-based WSN in disaster management," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25912–25921, Nov. 2021.
[106] Z. T. Alali and S. A. Alabady, "A survey of disaster management and SAR operations using sensors and supporting techniques," *Int. J. Disaster Risk Reduction*, vol. 82, Nov. 2022, Art. no. 103295.
[107] Z. Qadir, F. Ullah, H. S. Munawar, and F. Al-Turjman, "Addressing disasters in smart cities through UAVs path planning and 5G communications: A systematic review," *Comput. Commun.*, vol. 168, pp. 114–135, Feb. 2021.
[108] O. A. Saraereh, A. Alsaraira, I. Khan, and P. Uthansakul, "Performance evaluation of UAV-enabled LoRA networks for disaster management applications," *Sensors*, vol. 20, no. 8, p. 2396, Apr. 2020.
[109] T. Ahn, J. Seok, I. Lee, and J. Han, "Reliable flying IoT networks for UAV disaster rescue operations," *Mobile Inf. Syst.*, vol. 2018, pp. 1–12, Aug. 2018.
[110] H. Kim and K. Choi, "A modular wireless sensor network for architecture of autonomous UAV using dual platform for assisting rescue operation," in *Proc. IEEE SENSORS*, Oct. 2016, pp. 1–3.
[111] S. K. Datta, J.-L. Dugelay, and C. Bonnet, "IoT based UAV platform for emergency services," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, 2018, pp. 144–147.
[112] X. Xiang and S. Foo, "Recent advances in deep reinforcement learning applications for solving partially observable Markov decision processes (POMDP) problems: Part 1—Fundamentals and applications in games, robotics and natural language processing," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 3, pp. 554–581, Jul. 2021.
[113] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen, "Applications of multi-agent reinforcement learning in future internet: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1240–1279, 2nd Quart., 2022.
[114] P. Tong, J. Liu, X. Wang, B. Bai, and H. Dai, "Deep reinforcement learning for efficient data collection in UAV-aided Internet of Things," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Aug. 2020, pp. 1–6.
[115] M. Yi, X. Wang, J. Liu, Y. Zhang, and R. Hou, "Deep reinforcement learning for energy-efficient fresh data collection in rechargeable UAV-assisted IoT networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2023, pp. 1–6.
[116] J. Xu, X. Jia, and Z. Hao, "Research on information freshness of UAV-assisted IoT networks based on DDQN," in *Proc. IEEE 5th Adv. Inf. Manage., Communicates, Electron. Autom. Control Conf. (IMCEC)*, vol. 5, Dec. 2022, pp. 427–433.
[117] Y. Hu, Y. Liu, A. Kaushik, C. Masouros, and J. S. Thompson, "Timely data collection for UAV-based IoT networks: A deep reinforcement learning approach," *IEEE Sensors J.*, vol. 23, no. 11, pp. 12295–12308, Jun. 2023.
[118] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.
[119] M. Sun, X. Xu, X. Qin, and P. Zhang, "AoI-energy-aware UAV-assisted data collection for IoT networks: A deep reinforcement learning method," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17275–17289, Dec. 2021.
[120] N. Zhang, J. Liu, L. Xie, and P. Tong, "A deep reinforcement learning approach to energy-harvesting UAV-aided data collection," in *Proc. Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2020, pp. 93–98.
[121] M. Yi, X. Wang, J. Liu, Y. Zhang, and B. Bai, "Deep reinforcement learning for fresh data collection in UAV-assisted IoT networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, 2020, pp. 716–721.

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

[123] S. F. Abedin, Md. S. Munir, N. H. Tran, Z. Han, and C. S. Hong, "Data freshness and energy-efficient UAV navigation optimization: A deep reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5994–6006, Sep. 2021.

[124] M. Samir, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Online altitude control and scheduling policy for minimizing AoI in UAV-assisted IoT wireless networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 7, pp. 2493–2505, Jul. 2022.

[125] M. Sherman, S. Shao, X. Sun, and J. Zheng, "Optimizing AoI in UAV-RIS assisted IoT networks: Off policy vs. on policy," *IEEE Internet Things J.*, vol. 10, no. 14, pp. 12401–12415, Jul. 2023.

[126] H. Huang, J. Liu, and L. Xie, "Intelligent reflecting surface-assisted fresh data collection in UAV communications," in *Proc. Int. Conf. Commun., Signal Process., Syst.* Singapore: Springer, 2022, pp. 189–197, doi: 10.1007/978-981-99-2362-5_24.

[127] C. Y. Goh, C. Y. Leow, and R. Nordin, "Energy efficiency of unmanned aerial vehicle with reconfigurable intelligent surfaces: A comparative study," *Drones*, vol. 7, no. 2, p. 98, Jan. 2023.

[128] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1861–1870.

[129] Z. Li, P. Tong, J. Liu, X. Wang, L. Xie, and H. Dai, "Learning-based data gathering for information freshness in UAV-assisted IoT networks," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2557–2573, Feb. 2023.

[130] Z. Li, J. Liu, L. Xie, X. Wang, and M. Jin, "A deep reinforcement learning approach for multi-UAV-assisted data collection in wireless powered IoT networks," in *Proc. 14th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2022, pp. 44–49.

[131] E. Eldeeb, D. E. Pérez, J. M. de Souza Sant'Ana, M. Shehab, N. H. Mahmood, H. Alves, and M. Latva-Aho, "A learning-based trajectory planning of multiple UAVs for AOI minimization in IoT networks," in *Proc. Joint Eur. Conf. Netw. Commun. 6G Summit (EuCNC/6G Summit)*, 2022, pp. 172–177.

[132] L. Shi, X. Zhang, X. Xiang, Y. Zhou, and S. Sun, "Age of information optimization with heterogeneous UAVs based on deep reinforcement learning," in *Proc. 14th Int. Conf. Adv. Comput. Intell. (ICACI)*, Jul. 2022, pp. 239–245.

[133] K. Chi, F. Li, F. Zhang, M. Wu, and C. Xu, "AoI optimal trajectory planning for cooperative UAVs: A multi-agent deep reinforcement learning approach," in *Proc. IEEE 5th Int. Conf. Electron. Inf. Commun. Technol. (ICEICT)*, Aug. 2022, pp. 57–62.

[134] X. Li, B. Yin, J. Yan, X. Zhang, and R. Wei, "Joint power control and UAV trajectory design for information freshness via deep reinforcement learning," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC-Spring)*, 2022, pp. 1–5.

[135] E. Eldeeb, J. M. D. S. Sant'Ana, D. E. Pérez, M. Shehab, N. H. Mahmood, and H. Alves, "Multi-UAV path learning for age and power optimization in IoT with UAV battery recharge," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5356–5360, Apr. 2023.

[136] X. Wei, G. Zhang, and Z. Han, "Satellite-controlled UAV-assisted IoT information collection with deep reinforcement learning and device matching," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 1254–1259.

[137] O. S. Oubbati, M. Atiquzzaman, H. Lim, A. Rachedi, and A. Lakas, "Synchronizing UAV teams for timely data collection and energy transfer by deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6682–6697, Jun. 2022.

[138] F. Yang, J. Liu, and L. Xie, "UAV-assisted fresh data collection with MCS in wireless powered IoT," in *Proc. Int. Conf. Commun., Signal Process., Syst.*, in Lecture Notes in Electrical Engineering, vol. 874. Singapore: Springer, 2022, pp. 198–206, doi: 10.1007/978-981-99-2362-5_25.

[139] L. Canese, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, and M. Re, "Multi-agent reinforcement learning: A review of challenges and applications," *Appl. Sci.*, vol. 11, no. 11, p. 4948, May 2021.

[140] W. Fan, K. Luo, S. Yu, Z. Zhou, and X. Chen, "AoI-driven fresh situation awareness by UAV swarm: Collaborative DRL-based energy-efficient trajectory control and data processing," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, China, Aug. 2020, pp. 841–846.

[141] F. Wu, H. Zhang, J. Wu, Z. Han, H. V. Poor, and L. Song, "UAV-to-device underlay communications: Age of information minimization by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4461–4475, Jul. 2021.

[142] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. B. Letaief, "AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered IoT networks," *IEEE Internet Things J.*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021.

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

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

[145] C. Jia, H. He, J. Zhou, J. Li, Z. Wei, and K. Li, "A novel health-aware deep reinforcement learning energy management for fuel cell bus incorporating offline high-quality experience," *Energy*, vol. 282, Nov. 2023, Art. no. 128928.

[146] D. Han, B. Mulyana, V. Stankovic, and S. Cheng, "A survey on deep reinforcement learning algorithms for robotic manipulation," *Sensors*, vol. 23, no. 7, p. 3762, Apr. 2023.

[147] G. A. Vouros, "Explainable deep reinforcement learning: State of the art and challenges," *ACM Comput. Surveys*, vol. 55, no. 5, pp. 1–39, May 2023.

[148] L. He, N. Aouf, and B. Song, "Explainable deep reinforcement learning for UAV autonomous path planning," *Aerosp. Sci. Technol.*, vol. 118, Nov. 2021, Art. no. 107052.

[149] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning," *Knowledge-Based Syst.*, vol. 214, Feb. 2021, Art. no. 106685.

[150] R. N. Boute, J. Gijsbrechts, W. van Jaarsveld, and N. Vanvuchelen, "Deep reinforcement learning for inventory control: A roadmap," *Eur. J. Oper. Res.*, vol. 298, no. 2, pp. 401–412, Apr. 2022.

[151] N. Parvez Farazi, B. Zou, T. Ahamed, and L. Barua, "Deep reinforcement learning in transportation research: A review," *Transp. Res. Interdiscipl. Perspect.*, vol. 11, Sep. 2021, Art. no. 100425.

[152] A. Y. Majid, S. Saaybi, V. Francois-Lavet, R. V. Prasad, and C. Verhoeven, "Deep reinforcement learning versus evolution strategies: A comparative survey," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 2, 2023, doi: 10.1109/TNNLS.2023.3264540.

[153] E. Marchesini, D. Corsi, and A. Farinelli, "Genetic soft updates for policy evolution in deep reinforcement learning," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–15.

[154] F. Lotfi, O. Semiari, and W. Saad, "Semantic-aware collaborative deep reinforcement learning over wireless cellular networks," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 5256–5261.

[155] I. Jang, H. Kim, D. Lee, Y.-S. Son, and S. Kim, "Knowledge transfer for on-device deep reinforcement learning in resource constrained edge computing systems," *IEEE Access*, vol. 8, pp. 146588–146597, 2020.

[156] J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, and Z. Wang, "Exploration in deep reinforcement learning: From single-agent to multiagent domain," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jan. 19, 2023, doi: 10.1109/TNNLS.2023.3236361.

[157] Y. Yang, L. Juntao, and P. Lingling, "Multi-robot path planning based on a deep reinforcement learning DQN algorithm," *CAAI Trans. Intell. Technol.*, vol. 5, no. 3, pp. 177–183, Sep. 2020.

[158] V. Mai, "Reinforcement learning applied to the real world: Uncertainty, sample efficiency, and multi-agent coordination," Ph.D. thesis, Dept. Comput. Sci. Oper. Res., Université de Montréal, Dec. 2022.

[159] S. Zhang, Y. Li, and Q. Dong, "Autonomous navigation of UAV in multi-obstacle environments based on a deep reinforcement learning approach," *Appl. Soft Comput.*, vol. 115, Jan. 2022, Art. no. 108194.

[160] P. Cai, H. Wang, Y. Sun, and M. Liu, "DQ-GAT: Towards safe and efficient autonomous driving with deep Q-learning and graph attention networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21102–21112, Nov. 2022.

[161] J. Zhu, F. Wu, and J. Zhao, "An overview of the action space for deep reinforcement learning," in *Proc. 4th Int. Conf. Algorithms, Comput. Artif. Intell.*, 2021, pp. 1–10.

[162] N. Abouelenen, A. Alwarafy, and M. Abdallah, "Deep reinforcement learning for Internet of Drones networks: Issues and research directions," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 671–683, 2023.

OLUWATOSIN AHMED AMODU received the B.Tech. degree in electrical electronics engineering from The Federal University of Technology Akure, Nigeria, in 2012, and the master's degree in computer science with a specialization in distributed computing and the Ph.D. degree in wireless communication and network engineering from Universiti Putra Malaysia, in 2016 and 2021, respectively. He is currently a Postdoctoral Fellow with the Wireless Communication and Networks Laboratory, Universiti Kebangsaan Malaysia. He is also a Lecturer with Elizade University, Nigeria. His research interests include sensor networks, machine-type communications, device-to-device communication, stochastic geometry, reinforcement learning, unmanned aerial vehicles, and terahertz communications.

CHEDIA JARRAY received the B.Sc. and Ph.D. degrees in network and communications engineering from the National Engineering School of Gabes, University of Gabes, Tunisia, in 2013 and 2018, respectively. She is currently a Postdoctoral Researcher at Become: Technology, Science, AI, and Automation Laboratory in Paris, France, where she works remotely. Her research interests include 5G cellular networks, with a specific focus on machine-to-machine communications.

RAJA AZLINA RAJA MAHMOOD received the M.S. degree in software engineering from Universiti Teknologi Malaysia (UTM), in 2000. She is currently pursuing the Ph.D. degree in network security with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. She is also a member of the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia.

HUDA ALTHUMALI (Graduate Student Member, IEEE) received the bachelor's degree in computer science from Taif University, Saudi Arabia, in 2009, the master's degree in computer science from Universiti Putra Malaysia, in 2017, with a specialization in distributed computing, and the Ph.D. degree in computer networks, in 2022. She is currently an Assistant Professor with Imam Abdulrahman Bin Faisal University, Saudi Arabia. Her research interests include wireless networks, the IoT, cloud computing, and network security.

UMAR ALI BUKAR received the B.Sc. degree in business information technology (with a focus on e-commerce research and strategy) from Greenwich University, U.K., the M.Sc. degree in computer network management from Middlesex University, Dubai, and the Ph.D. degree from the Department of Software Engineering and Information Systems, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia. He is currently a Postdoctoral Research Fellow with the Centre for Intelligent Cloud Computing (CICC), Faculty of Information Science and Technology, Multimedia University, Melaka, Malaysia. His contributions have been published in prestigious peer-reviewed journals and international conferences. His IT career has included work on several niche projects, with responsibilities ranging from teaching to research and analysis. His research interests include crisis informatics, data analytics, text analysis, machine learning, and SLR.

ROSDIADEE NORDIN received the B.Eng. degree from Universiti Kebangsaan Malaysia (UKM), in 2001, and the Ph.D. degree from the University of Bristol, U.K., in 2011. He is currently a Professor with the Department of Engineering, School of Engineering and Technology, Sunway University. His research interests include beyond 5G wireless communications, specifically focusing on advanced wireless transmission techniques and channel modeling, aerial wireless communications, and wireless communications for the Internet of Things applications. He was a recipient of the Leadership in Innovation Fellowship, the Technopreneur Program under the Royal Academy of Engineering, U.K., in 2021, and the Top Research Scientists Malaysia (TRSM), a Prestigious Award under the Academy of Science Malaysia, in 2020.

NOR FADZILAH ABDULLAH (Member, IEEE) received the B.Sc. degree in electrical and electronics engineering from Universiti Teknologi Malaysia, in 2001, the M.Sc. degree (Hons.) in communications engineering from The University of Manchester, U.K., in 2003, and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, U.K., in 2012. She is currently an Associate Professor with Universiti Kebangsaan Malaysia, Selangor, Malaysia. Her research interests include 5G, millimeter wave, LTE-A, vehicular networks, massive MIMO, space-time coding, fountain codes, and channel propagation modeling and estimation.

NGUYEN CONG LUONG received the B.S. and M.S. degrees from the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology (HUST), Vietnam, and the Ph.D. degree from Institut Galilée, Université Sorbonne Paris Nord, France. His research interests include resource allocation and security management in the next-generation networks.