Toward Autonomous Multi-UAV Wireless Network: A Survey of Reinforcement Learning-Based Approaches

Yu Bai, Hui Zhao, *Graduate Student Member, IEEE*, Xin Zhang, Zheng Chang, *Senior Member, IEEE*, Riku Jäntti, *Senior Member, IEEE*, and Kun Yang, *Fellow, IEEE*

***Abstract*—Unmanned aerial vehicle (UAV)-based wireless networks have received increasing research interest in recent years and are gradually being utilized in various aspects of our society. The growing complexity of UAV applications such as disaster management, plant protection, and environment monitoring, has resulted in escalating and stringent requirements for UAV networks that a single UAV cannot fulfill. To address this, multi-UAV wireless networks (MUWNs) have emerged, offering enhanced resource-carrying capacity and enabling collaborative mission completion by multiple UAVs. However, the effective operation of MUWNs necessitates a higher level of autonomy and intelligence, particularly in decision-making and multi-objective optimization under diverse environmental conditions. Reinforcement Learning (RL), an intelligent and goal-oriented decision-making approach, has emerged as a promising solution for addressing the intricate tasks associated with MUWNs. As one may notice, the literature still lacks a comprehensive survey of recent advancements in RL-based MUWNs. Thus, this paper aims to bridge this gap by providing a comprehensive review of RL-based approaches in the context of autonomous MUWNs. We present an informative overview of RL and demonstrate its application within the framework of MUWNs. Specifically, we summarize various applications of RL in MUWNs, including data access, sensing and collection, resource allocation for wireless connectivity, UAV-assisted mobile edge computing, localization, trajectory planning, and network security. Furthermore, we identify and discuss several open challenges based on the insights gained from our review.**

Manuscript received 22 January 2023; revised 24 June 2023 and 25 August 2023; accepted 4 October 2023. Date of publication 12 October 2023; date of current version 22 November 2023. This work was supported in part by NSFC under Grant 62071105, and in part by Sichuan NSF under Grant 2022NSFSC0544. *(Corresponding author: Zheng Chang.)*

Yu Bai is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland (e-mail: yu.bai@aalto.fi).

Hui Zhao and Xin Zhang are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: zhaohui2022@std.uestc.edu.cn; 202121080639@std.uestc.edu.cn).

Zheng Chang is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Faculty of Information Technology, University of Jyväskylä, 40014 Jyväskylä, Finland (e-mail: zheng.chang@jyu.fi).

Riku Jäntti is with the Department of Information and Communications Engineering, Aalto University, 02150 Espoo, Finland (e-mail: riku.jantti@aalto.fi).

Kun Yang is with the School of Computer Sciences and Electrical Engineering, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: kunyang@essex.ac.uk).

Digital Object Identifier 10.1109/COMST.2023.3323344

***Index Terms*—Unmanned aerial vehicle (UAV), multi-UAV wireless network, reinforcement learning, UAV-assisted communication network, UAV-assisted mobile computing.**

I. INTRODUCTION

UNMANNED Aerial Vehicles (UAVs), commonly referred to as drones, have steadily grown into pivotal elements within various professional fields, marking a paradigm shift in numerous practices and operations. Their broad application spectrum spans several civilian spheres such as aerial photography, precision agriculture, environmental monitoring, and search and rescue operations, among others [1], [2], [3], [4]. These uses highlight UAVs' transformative potential in different industries and their expanding role in contemporary society. This burgeoning relevance of UAVs is substantiated by their market growth dynamics. According to a comprehensive report by UAV Industry Insights, the global UAV market, valued at approximately U.S.$30.6 billion in 2022, is expected to nearly double, reaching an estimated worth of U.S.$55.8 billion by 2030. This projection represents a Compound Annual Growth Rate (CAGR) of 7.8%, underscoring the persistent rise of the UAV industry in the coming years [5].

The inherent flexibility and cost-effective deployment of UAVs act as catalysts for the advancement of multifaceted applications. However, the capabilities of a single UAV-based platform often fall short of meeting the growing demand for flexible and autonomous applications. The coverage and functionality offered by a solitary UAV are limited, posing challenges in addressing increasingly intricate missions. Consequently, there is a rising interest in the development of multi-UAV-based frameworks, aimed at augmenting the scale of existing single UAV-based applications and introducing enhanced degrees of freedom and autonomy [6], [7].

Multi-UAV Wireless Networks (MUWNs) offer superior coverage and expanded service resources, thus facilitating coordinated and cooperative efforts among multiple UAVs to tackle large-scale, multi-objective tasks. These networks sustain the collective functionality of multiple UAVs while retaining the autonomy of individual units, thereby aiding the completion of applications involving more complex tasks. Innovative scenarios such as MUWN-assisted Internet of Things (IoT) [8], MUWN-assisted cellular communications [9], and MUWN-assisted edge computing have been proposed [10], indicating the versatile potential of MUWNs. MUWNs also act as vital components in the upcoming integration of 6G wireless technologies, holding the promise of significantly enhancing the reach, reliability, and resilience of 6G systems [11]. Leveraging MUWNs, 6G systems can harness their capabilities to ensure uninterrupted, high-speed, and low-latency communication, especially in challenging terrains and dense urban areas. These complex applications of UAVs further emphasize the need for advanced orchestration methods and smarter algorithms in MUWNs, including task allocation, trajectory planning, and resource management. It also requires an optimization approach to balance various objectives such as energy efficiency, communications coverage, and user satisfaction. Moreover, the inherent dynamics and unpredictability of MUWN operating environments, influenced by factors like flight-path obstacles and fluctuating user demands, necessitate a high degree of adaptability and intelligence within MUWNs. Traditional deterministic or heuristic algorithms may not fully accommodate these needs due to their inherent limitations in adaptability and scalability, especially considering the network's dynamic and distributed nature.

To address these challenges, researchers have been exploring various machine learning techniques, among which Reinforcement Learning (RL) has shown promising results. Over the past decade, RL has demonstrated remarkable success in various fields such as game playing [12], autonomous vehicles [13], and robotic control [14]. RL-based algorithms possess the capability to learn and adapt based on environmental cues, offering an optimal solution for handling dynamic conditions and identifying optimal strategies [15], [16]. RL's unique capability of learning from interaction and optimizing complex control policies makes it a prime candidate for the coordination and operation of MUWNs. RL has already demonstrated its efficacy in enhancing MUWNs by optimizing key factors like strategic deployment and latency in the context of 6G technology [17], [18].

Furthermore, recent advancements in deep learning have fostered the development of Deep Reinforcement Learning (DRL), wherein deep neural networks are employed to estimate the value function and policy in RL [19]. This enables the RL algorithms can handle high-dimensional state and action spaces [12]. Leveraging the function approximation capability of deep learning, DRL can effectively address scalability issues in large-scale MUWNs. Consequently, the utilization of DRL in MUWNs is considered to hold substantial potential, facilitating the development of adaptable, efficient, and intelligent MUWNs.

Given the increasing importance of RL and DRL in MUWNs, it is necessary to provide a comprehensive survey on this topic, summarizing the current advanced applications, identifying the challenges, and discussing potential future directions. This paper aims to serve this need, providing valuable insights for researchers and practitioners interested in the intersection of RL and MUWNs.

A. Related Surveys

Though several existing reviews have discussed the applications and advantages of RL or UAVs from different perspectives, there is still a lack of a comprehensive survey on RL-enabled MUWNs. In [20], the authors present a tutorial detailing the benefits, applications, and challenges of UAVs in wireless communication. The authors of [21] provide an analysis of game-theoretic solutions that can potentially mitigate issues in UAV-assisted networks. In [9], the authors examine UAV cellular communications from multiple perspectives, including practical aspects, standardization advancements, regulations, and security challenges. In [22], the authors emphasize the potential of single-agent RL algorithms in various applications but neglect the discussion of multi-agent RL algorithms and MUWNs. In [23], the authors explore security-related challenges and present emerging technologies in UAV networks. In [24], the authors narrowly focus on RL algorithms without extensively considering possible applications. In [25], the authors offer a comprehensive review of wireless networks for future aerial communications, shedding light on connectivity requirements, wireless communication technologies, and network architectures. In [26], the authors survey the applications of multi-agent RL in the future Internet, with a specific section dedicated to the trajectory design of UAV networks. However, the authors overlook single-agent RL applications as well as a multitude of other MUWN applications. The authors of [27] and [28] focus on the application of RL in autonomous navigation and path planning in MUWNs, leaving many other MUWN applications unexplored. In [29], the authors provide a survey on the system architecture and networking design of aerospace-integrated networks. Table I provides a comparative summary of the surveys mentioned, highlighting that even the most recent ones do not offer a comprehensive overview of the applications of RL-enabled MUWNs, which also indicates there is a gap in the existing literature.

B. Scope and Contribution of Our Survey

This paper aims to provide a comprehensive survey and an in-depth analysis of RL-enabled MUWNs. To ensure a clear understanding for a broad spectrum of readers, we include an instructional guide to RL techniques in the second section. Then, we summarize the cutting-edge applications of RL-enabled MUWNs. These applications have been divided into six categories, with each further subdivided according to important tasks or targets pertaining to the specific application.

The first four categories are dedicated to advanced areas within RL-enabled MUWNs, as illustrated in Fig. 1.

- *Data Access, Sensing, and Collection:* MUWNs can enhance real-time data sensing and transmission, a significant benefit for IoT nodes with limited uplink capabilities [30]. As depicted in Fig. 1, UAVs gather data from IoT devices and then transmit this data to the Data Center. By adjusting factors such as UAV positioning, transmission power, and other related actions, RL can optimize data access, sensing, and collection performance. Within this section, we elaborate on three pivotal performance aspects: maximization of data collection, enhancement of data collection efficiency, and optimization of data freshness.
- *Resource Allocation for Wireless Connectivity:* MUWNs enhance existing terrestrial networks by offering

TABLE I
EXISTING SURVEYS ON RL, MUWN, AND RELATED AREAS

Works	Topic	Scope		
		MUWNs	RL	Diverse Applications of RL-Enabled MUWNs
[20]	A tutorial of benefits, applications, and challenges of UAVs in wireless communication	✔	✗	✗
[21]	A survey of game theory in UAVs communication	✔	✗	✗
[9]	A survey on UAV cellular communications	✔	✗	✗
[22]	A survey on applications of DRL in communications and networking	✗	✔	✗
[23]	A survey on security issues and emerging solutions for UAVs	✔	✗	✗
[24]	A tutorial of DRL for AI-enabled wireless networks	✗	✔	✗
[25]	A survey of wireless networks for future aerial communications	✔	✗	✗
[26]	A survey on applications of multi-agent RL in future internet	✔	✔	✗
[27]	A survey on autonomous UAV navigation using RL	✔	✔	✗
[28]	A review of artificial intelligence applied to path planning in UAV swarms	✔	✔	✗
[29]	A survey on aerospace integrated networks innovation for empowering 6G	✔	✗	✗
Our survey	A comprehensive survey of the applications of RL in MUWNs	✔	✔	✔

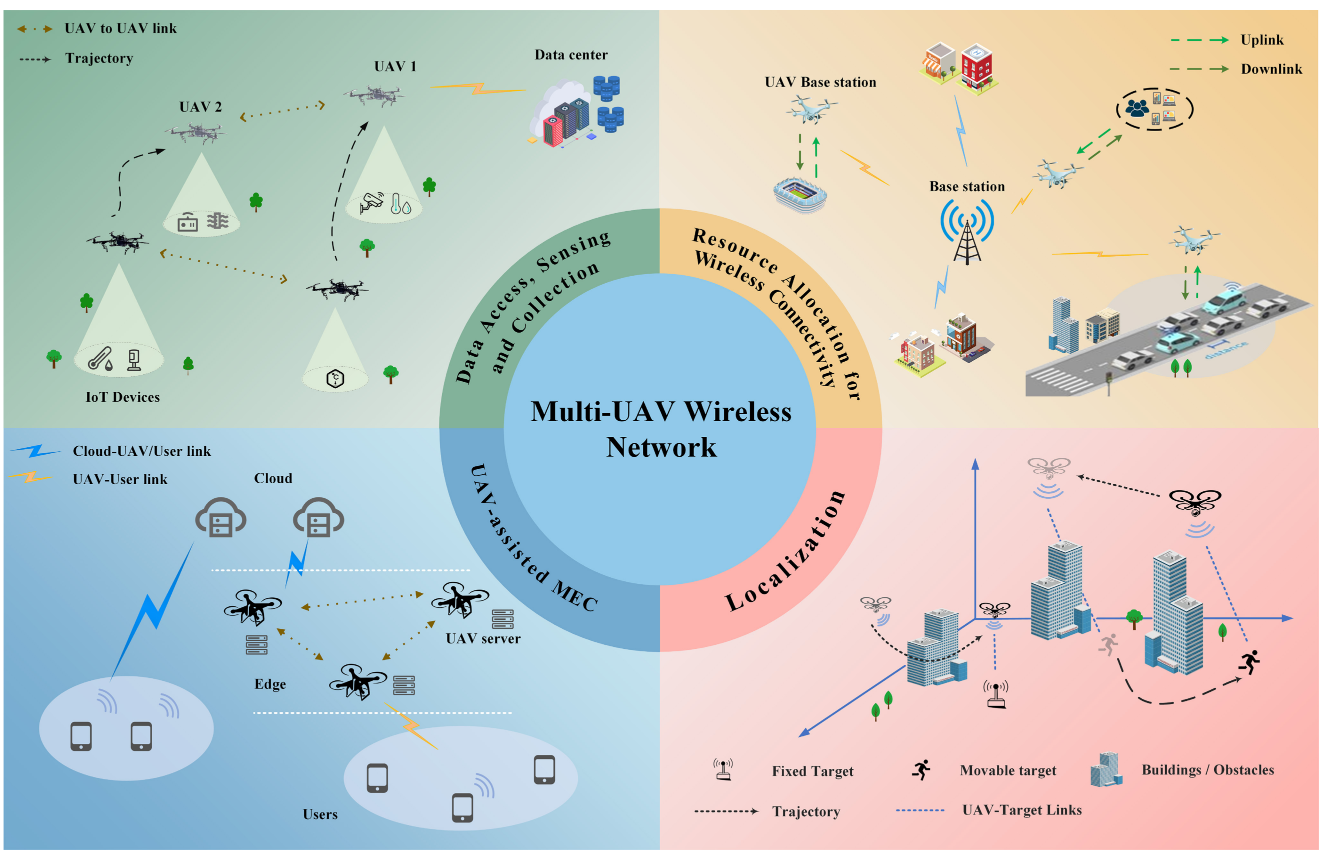


Fig. 1. MUWNs in different applications.

adaptive coverage, flexible deployment, and streamlined management [20], [31]. As illustrated in Fig. 1, UAVs can provide communication services to diverse terminals, such as specific highway segments and dense ground user populations. Consequently, the performance of these networks heavily relies on judicious resource allocation and sophisticated optimization technique. RL equips MUWNs with the capability to adjust strategies in unpredictable and ever-changing environments. This section predominantly discusses spectrum allocation and power control, wireless power transfer, and caching strategies.

- *UAV-Assisted Mobile Edge Computing (MEC):* The advent of MEC, a new architectural paradigm, allows UAVs equipped with computing and communication platforms to provide offloading services for latency and energy-sensitive applications [32]. As depicted in Fig. 1, UAVs function as edge servers, delivering real-time services to users. RL-based applications and algorithms can enhance UAV-assisted MEC systems by aiding decision-making processes such as offloading decisions and computation resource allocation. We discuss four key performance areas in MEC systems: latency-related, energy-related optimization and design, resource management and computation offloading, and optimization of other objectives.
- *Localization:* UAVs play a crucial role in scenarios like search and rescue, surveillance, target tracking, and positioning [3], [33]. As illustrated in Fig. 1, UAVs have the capability to localize both static and movable targets within intricate environments. RL can be instrumental in enhancing the precision of UAV-based localization. This is achieved by optimizing the received signal from targets or refining the distance metrics to targets via meticulous adjustments to the UAV's movement and power.

The final two application categories delve into trajectory planning and network security, which are both critical considerations in MUWNs.

- *Trajectory Planning:* Effective trajectory planning is a critical aspect of UAVs in MUWNs, as it not only influences the movement of UAVs but also plays a significant role in achieving optimization objectives [28], [34], [35]. We highlight the importance of effective trajectory planning and flexibility for RL-enabled MUWNs to excel across various applications. This section focuses on navigation and provides typical examples showcasing how trajectory planning can optimize MUWN performance.
- *Network Security:* Owing to their open and multi-connective nature, MUWNs are susceptible to attacks [9]. RL can help optimize the location and transmission power of MUWNs to avoid eavesdropping or interference.

Fig. 2 illustrates the structure of these applications. We hope that this survey can shed light on autonomous and intelligent MUWNs, and provide a clear vision for researchers and engineers in related fields.

The primary contributions of this survey are outlined as follows:

- We offer an instructional guide to RL techniques and outline a clear procedure for utilizing RL in MUWNs.
- We provide a comprehensive overview of the applications of RL in MUWNs. To facilitate understanding, we categorize recent research works into six groups and provide detailed discussions for each category.
- We highlight the open challenges in the field to guide readers towards future research directions concerning the applications of RL-enabled MUWNs.

The remainder of this paper is given as follows. Section II presents basic knowledge and popular algorithms of RL. Section III examines the application of RL for data access, sensing, and collection in MUWNs. Section IV analyzes the RL-based resource allocation schemes for wireless connectivity in MUWNs. Section V demonstrates RL applications in UAV-assisted MEC. Section VI presents the localization-related applications. In Section VII, the related works on RL-based trajectory planning for MUWNs are reviewed. Applying RL for MUWN security is discussed in Section VIII. Section IX highlights open issues and challenges. Finally, we conclude this survey in Section X. To facilitate reading, the abbreviations used in this paper are listed in Table II.

TABLE II
ABBREVIATIONS

Abbreviation	Description
A2C	Advantage Actor-Critic
A3C	Asynchronous Advantage Actor-Critical
AC	Actor-Critic
AoI	Age of Information
BS	Base Station
COMA	Counterfactual Multi-Agent Policy Gradients
CSI	Channel State Information
CTDE	Centralized Training and Decentralized Execution
D2D	Device to Device
D3QN	Dueling Double DQN
DCA	Difference of Convex Algorithm
DDPG	Deep Deterministic Policy Gradient
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
HAP	High Altitude Platform
IL	Independent Learning
IoT	Internet of Things
LoS	Line-of-Sight
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MEC	Mobile Edge Computing
MUWNs	Multi-UAV Wireless Networks
NOMA	Non-Orthogonal Multiple Access
POMDP	Partially Observable Markov Decision Process
PPO	Proximal Policy Optimization
QoE	Quality of Experience
QoS	Quality of Service
RF	Radio Frequency
RIS	Reconfigurable Intelligent Surface
RL	Reinforcement Learning
RSS	Received Signal Strength
SAC	Soft Actor-Critic
SARL	Single-Agent Reinforcement Learning
SINR	Signal-to-Interference Plus Noise Ratio
TD3	Twin-Delayed DDPG
TRPO	Trust Region Policy Optimization
UAV	Unmanned Aerial Vehicle
UE	User Equipment
VDN	Value-Decomposition Networks
WPT	Wireless Power Transfer

II. PRELIMINARY ON REINFORCEMENT LEARNING

This section provides some basics about RL. Initially, we introduce the fundamentals of RL and its mathematical representation. Subsequently, we provide a summary of the classic RL algorithms. Lastly, we delineate a general procedure for employing RL in MUWNs, enabling readers to understand how RL can be leveraged to improve MUWNs.

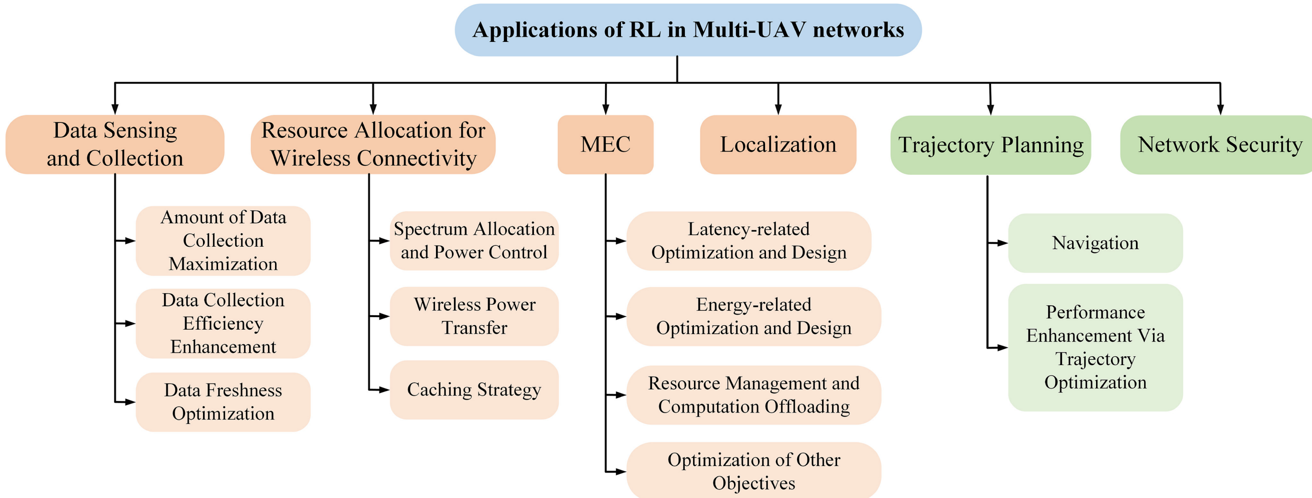


Fig. 2. A classification of the applications of RL in MUWNs.

A. Fundamentals of RL

RL is a goal-oriented learning and autonomous decision-making method. The core of RL is a decision-maker, commonly referred to as an agent, which learns to make beneficial decisions through interactions with its environment. The crux of this interaction is a reward system where actions leading to favorable outcomes are positively reinforced, thus guiding the agent to strive towards maximizing cumulative rewards over time. The action-guiding process within the RL is an iterative refinement of the agent's strategy, known as policy. The agent employs an initial policy to interact with the environment, observes the subsequent outcomes, and fine-tunes the policy in line with the received feedback. This iterative learning process is maintained until an optimal or near-optimal policy is obtained.

RL can address a wide range of complex problems that necessitate intelligent decision-making and adaptive behavior, such as game playing, robotics, navigation, resource management, and task scheduling. In essence, RL can be effectively applied to any problem that can be conceptualized as a sequence of decisions targeting a specific objective.

In practical applications, RL is operationalized via its mathematical representation known as a Markov Decision Process (MDP) [36]. The MDP offers a systematic mechanism to model and analyze the complete process of learning and decision-making in RL. A detailed discussion on MDP is presented in the subsequent section.

B. MDP

A MDP can be mathematically represented as a tuple (S, A, P, R, γ), where S and A denote the sets of the agent's states and actions respectively. The state transition probability set is expressed as $P : S \times A \times S \rightarrow [0, 1]$. $R : S \times A \times S \rightarrow \mathbb{R}$ signifies the set of rewards $r(s_t, a_t, s_{t+1})$ received by the agent from the environment. Commonly, we consider a finite MDP with times steps T. The agent-environment interaction concludes naturally at the final step T, such as upon the completion of a task. Therefore, the interaction within a finite MDP can be recorded as a sequence $\{s_0, a_0, s_1, a_1, \ldots, s_T, a_T\}$, commonly referred to as an episode. Fig. 3 illustrates the MDP. The agent chooses an action a_t based on the current state s_t. In response to the agent's action, the environment reciprocates with an updated state s_{t+1} and offers a reward $r(s_t, a_t, s_{t+1})$ to the agent. It is crucial to note that the action is to maximize the discount cumulative return $G_t = \sum_{k=0}^{T} \gamma^k r(s_{t+k}, a_{t+k}, s_{t+1+k})$, rather than the immediate reward $r(s_t, a_t, s_{t+1})$. Here, $\gamma \in [0, 1]$ serves as a discount rate employed to strike a balance between immediate and future rewards. As mentioned, RL is to find the optimal policy. The policy of the agent is recorded as $\pi(a_t|s_t)$, representing the probability of choosing action a_t at state s_t.

In practical applications, agents often lack complete environmental information. For example, a UAV gathers only partial information about the environment through its sensors. In such cases, the RL problem is modeled as a Partially Observable MDP (POMDP) [37]. A POMDP can be described as a 7-tuple $(S, A, P, R, \gamma, O, Z)$, where S, A, P, R, and γ are defined as in MDP. Here, O represents the set of possible observations and $Z : S \times A \times O \rightarrow [0, 1]$ describes the probability distribution over observations.

C. Single-Agent Reinforcement Learning

Having presented the fundamental concepts of RL and MDP, we can now turn our attention to a particular subset of RL problems known as Single-Agent Reinforcement Learning (SARL). SARL pertains to scenarios in which a single agent is involved in the learning and decision-making process. This section provides an overview of the milestone SARL algorithms, helping the readers gain a comprehensive understanding of the SARL landscape. As we continue to build upon these concepts, keep in mind that there are also scenarios involving multiple agents. These are referred to as Multi-Agent Reinforcement Learning (MARL) problems, which will be discussed in the subsequent section.

SARL algorithms can be categorized into two distinct types, namely value-based and policy-based SARL, depending on the methodologies adopted for the derivation of optimal policies.

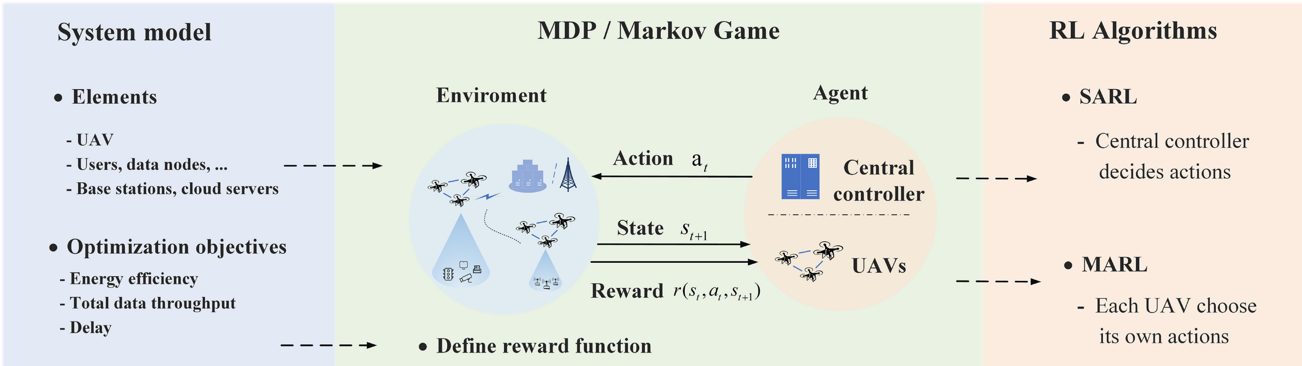


Fig. 3. A general procedure for applying RL to MUWNs.

1) Value-Based SARL Algorithms: Value-based SARL algorithms operate by computing the value function that provides the expected cumulative reward for each state or state-action pair. The aim is to identify the optimal value function, which offers the highest expected cumulative reward for each state or state-action pair. This optimal value function subsequently determines the optimal policy, wherein the agent selects the action that results in the state with the highest expected value at every step.

A significant advancement in RL was the advent of the Q-learning algorithm [38]. The Q-learning algorithm utilizes an action-state function, $Q(s_t, a_t)$, commonly referred to as the Q-function, to compute the expected cumulative reward for each state-action pair. Through iterative computation, this algorithm converges to the optimal Q-value, as represented by the following equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r(s_t, a_t, s_{t+1}) + \gamma max_a Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (1)$$

where α is the learning rate. In *Q*-learning, a Q-table is used to record different Q-values corresponding to various states and actions. However, a Q-table is constrained by its ability to record only a limited number of actions and states.

To overcome this limitation, the Deep Q-Network (DQN) [12] employs a deep neural network to map states and actions to Q-values, referred to as the Q-network. DQN updates its parameters using gradient descent. A key feature of DQN is the implementation of a target network that mirrors the original Q-network in terms of parameters. This target network serves a critical function in stabilizing the learning process. While the primary Q-network's parameters are updated frequently, the target network's parameters are kept fixed over a certain number of steps. This reduces the correlations between the target and predicted Q-values and aids in mitigating the risk of feedback loops. Upon reaching the specified number of steps, the parameters of the target network are synchronized with those of the Q-network, thereby ensuring a continual, gradual learning process. Both Q-learning and DQN are instances of temporal-difference learning algorithms [39]. These algorithms update their estimates following each state transition, irrespective of waiting for the final result of an episode.

Following the advent of DQN, researchers have developed various enhancements to refine and extend the utility of DQN. Double DQN [40] resolves the issue of overestimation, where the learned Q-value often exceeds the actual Q-value. Furthermore, the method of Prioritized Experience Replay [41] introduces a mechanism that assigns weights to data samples based on their significance. This allows the algorithm to prioritize more informative data during the learning process, thereby reducing the variance and enhancing learning efficiency. Dueling DQN [42] employs the concept of an advantage function, which serves to evaluate the relative quality of each action taken in a given state s. In other words, it quantifies the advantage of choosing a particular action.

The Dueling Double DQN (D3QN) incorporates the strengths of both the Double DQN and the Dueling DQN methodologies. It amalgamates the overestimation rectification of Double DQN with the differentiated action valuation provided by the advantage function in Dueling DQN, thus creating a more refined and effective learning algorithm.

2) Policy-Based RL Algorithms: Policy-based RL algorithms aim to find the optimal policy directly without needing to learn a value function. In these methods, the policy is parameterized (often using neural networks in complex environments), and the parameters are iteratively updated to improve the policy. The principle underlying policy-based methods is known as Policy Gradient [43]. In simple terms, Policy Gradient methods aim to maximize the expected return by adjusting the policy parameters in the direction that maximally improves the performance. This is typically achieved using methods of gradient ascent. In this way, Policy Gradient methods iteratively enhance the policy until an optimal policy is obtained.

One of the fundamental algorithms of Policy Gradient methods is the REINFORCE algorithm [44]. It utilizes a Monte Carlo method, generating data of an episode based on the current policy, to estimate the gradient for updates. For policy gradient algorithms, the size of the gradient update step profoundly influences the performance. Techniques such as Trust Region Policy Optimization (TRPO) [45] update the policy under constraints measured by the Kullback-Leibler

Divergence. Proximal Policy Optimization (PPO) [46], a more user-friendly approach than TRPO, incorporates the penalty term of Kullback-Leibler Divergence directly into the objective function. Currently, PPO is the most widely adopted policy gradient method.

However, policy gradient methods encounter a significant issue: they exhibit high variance when learning from raw episode data. This variance leads to inefficient learning and negatively impacts the training process. To address this, Actor-Critic (AC) algorithms were developed. The AC methods utilize a 'critic' to estimate the value function, while the 'actor' estimates the policy. By estimating the value function, the critic aids in reducing the variance associated with policy gradient methods, resulting in more efficient learning. With the advent of neural networks, these roles are often assigned to separate networks within the model: the actor network and the critic network.

The Asynchronous Advantage Actor-Critic (A3C) method introduced a parallel form of AC algorithms [47]. A3C uses a global network and multiple worker nodes that interact with the environment to gather data and update parameters. The worker nodes update their parameters asynchronously with the global network. However, it was found that the strength of A3C comes more from the parallel training aspect rather than asynchronous updates. Following this observation, the Advantage Actor-Critic (A2C) method was developed, which updates parameters synchronously to ensure uniformity among worker nodes. Today, A2C is generally considered more effective than A3C. The main purpose of both A3C and A2C is to reduce correlation in the training data by using multiple workers to gather data.

These algorithms, however, are primarily designed MDPs with discrete action spaces. For MDPs with continuous action spaces, the Deep Deterministic Policy Gradient (DDPG) method is employed [48]. DDPG combines elements of DQN and AC algorithms, creating a deterministic policy where the relationship between states and actions is deterministic [49]. Notably, DDPG employs target networks for both the actor and critic networks, leading to a total of four networks in the DDPG structure. A subsequent improvement of DDPG, the Twin-Delayed Deep Deterministic Policy Gradient (TD3) [50], introduced additional enhancements including clipped double-Q learning, delayed updates, and target policy smoothing.

The Soft Actor-Critic (SAC) algorithm [51], another variant of AC, introduces entropy regularization to the optimization process. This approach encourages the policy to explore more options and reduces the likelihood of premature convergence to sub-optimal policies.

D. Multi-Agent Reinforcement Learning

While SARL focuses on decision-making processes involving a single learning agent interacting with its environment, real-world scenarios often encompass multiple interacting agents. Examples range from a single robot control to multiple robots coordinating to complete a task. The actions of each agent can significantly impact the others. Hence, MARL becomes a necessary paradigm, accounting for these complex multi-agent interactions.

This section first introduces the Markov Games, the multi-agent counterpart to the MDP used in SARL. Then, we explore various approaches employed to handle interactions between agents, a critical aspect of transitioning from SARL to MARL. Finally, we will summarize some of the classical MARL algorithms in use today.

1) Markov Games: Markov games [52] are multi-agent extensions of MDPs. A partially observable Markov game for N agents can be denoted as a tuple $(N, S, \{A_i\}_{i\in N}, \{O_i\}_{i\in N}, P, \{R_i\}_{i\in N}, \gamma)$. N denotes the number of agents. S represents the state of all agents. $\{A_i\}_{i\in N}$ and $\{O_i\}_{i\in N}$ are the sets of actions and observations for each agent. To choose actions, each agent i uses a stochastic policy $\pi_{\theta_i} : O_i \times A_i \rightarrow [0, 1]$, which produces the next state according to the state transition function $P : S \times A_1 \times \cdots \times A_N \rightarrow S^2$. $\{R_i\}_{i\in N}$ is a set of rewards for each agent.

2) Training Scheme: The primary distinction between MARL and SARL lies in the interaction among agents. In the work presented in [53], multi-agent systems are categorized into three types based on the relationship between agents: fully cooperative games, fully competitive games, and mixed games. Hence, a primary challenge in MARL is the training of multiple agents to cooperate or compete with each other.

A rudimentary approach to MARL training is Independent Learning (IL), which does not take into account the influence of other agents. This approach essentially overlooks both cooperation and competition between agents. Although IL may produce adequate results in certain practical scenarios, it may lead to non-stationarity problems, which can negatively impact convergence [54].

A more commonly utilized framework for MARL training is Centralized Training with Decentralized Execution (CTDE). In this setup, each agent has access to the information of all other agents during the training phase. However, during the execution phase, agents make decisions based solely on their local information. This CTDE framework offers a practical solution for MARL and is widely used in recent research.

3) MARL Algorithms: Coming to specific MARL algorithms, Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [55] represents a landmark algorithm. MADDPG extends the DDPG algorithm to multi-agent environments. In MADDPG, each agent has its own independent critic network, actor network, and reward function, allowing the algorithm to tackle multi-agent problems in cooperative, competitive, and mixed environments. During training, each agent can view the local observations and actions of all other agents. Hence, centralized training is required despite each agent having an independent critic network.

Another advanced algorithm is Counterfactual Multi-Agent Policy Gradients (COMA) [56], designed to tackle the multi-agent credit assignment problem in Decentralized POMDP (Dec-POMDP) [57]. In COMA, the agents share a joint critic network based on the local observations and actions of all agents, while each agent's actor network is independent and based only on local observations. Since all agents in Dec-POMDP cooperate fully and share the same global reward, each agent may not know the impact of its actions on the global reward. COMA addresses this issue by computing each agent's individual advantage function.

Further, value decomposition-based algorithms can also be used to solve the Dec-POMDP problem. Such methods decompose the joint Q value, Q_{total}, into a combination of each agent's Q value, Q_i. Value-Decomposition Networks (VDN) [58], for example, assume that the Q-value of each agent can be summed to yield a joint action-value function. This addition in VDN implies a linearity assumption. QMIX [59], on the other hand, combines Q_i to Q_{total} using a hybrid network, and employs global state information to improve algorithm performance. At the same time, to ensure global maximization, QMIX assumes that the derivative of Q_{total} with respect to Q_i is non-negative.

E. Adapting RL for MUWNs

This section outlines the general procedure for applying RL in MUWNs.

As shown in Fig. 3, the system model of practical MUWN applications encompasses various elements such as UAVs and users. It also includes optimization objectives for MUWNs, such as energy efficiency and total throughput.

Central to the application of RL in MUWNs is the formulation of the system model into a MDP or a Markov Game. This formulation entails two critical steps. Firstly, defining the elements as the agent and the environment. As Fig. 3 depicts, the agent is typically defined as the central controller or an individual UAV. In most SARL scenarios, the central controller, acting as the agent, is tasked with allocating tasks or target locations to UAVs. Conversely, when individual UAVs are defined as agents, their actions could be specified as speed, direction, or designation of their own target user.

The second step involves designing the reward function in accordance with the optimization objectives. In instances where there are multiple optimization objectives, it is possible to incorporate multiple reward or penalty terms into the reward function.

Depending on the agent defined earlier, either SARL or MARL could be leveraged for MUWNs. Given that neural networks are frequently employed in RL as a parameterized model for policy or value, these networks can be tailored to suit the specific scenario.

III. Data Access, Sensing and Collection

In IoT applications, transmitting data to nearby Base Stations (BSs) can be challenging for wireless nodes with limited energy. To overcome this challenge, UAVs can be utilized to facilitate data access, sensing, and collection for nodes that are unable to access the network infrastructure. Fig. 4 provides a visual representation of a typical data collection scenario, wherein UAVs collect data from various devices and transfer it to the central data center.

During these operations, MUWNs confront complex decision-making tasks including the allocation of data nodes, optimization of flight trajectories, and determination of appropriate transmission power levels. RL-based agents can take on the role of decision-makers in this context, helping to address these challenges and enhance the overall performance of data access, sensing, and collection tasks.

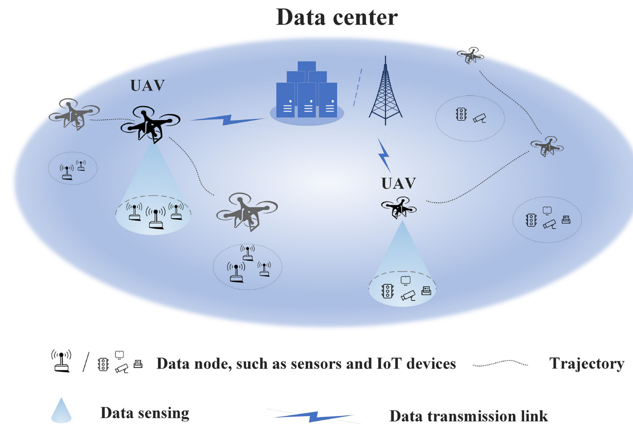


Fig. 4. Data access, sensing, and collection in MUWN.

We have categorized related works into three key aspects based on their main focus in data collection tasks: data collection maximization, data collection efficiency enhancement, and data freshness optimization.

A. Data Collection Maximization

A primary objective of MUWNs in data collection task is to optimize the collection of data from multiple terminals while operating within defined constraints of time and energy resources.

In [60], the authors introduce a mobile crowd-sensing system designed for urban areas with data nodes and multiple obstacles. UAVs are utilized for data collection and obstacle avoidance in this region. The system model divides the target area into multiple grids, disregarding the effect of UAV flight altitude. Consequently, the location information of UAVs, obstacles, and data nodes can be integrated into a 2-dimensional (2D) plane. The paper employs a Convolutional Neural Network (CNN), known for its significant advantages in image processing [61], to extract features from the 2D plane. To maximize the data collection within a specified number of time slots, while ensuring fairness among data nodes and UAVs' energy efficiency, the paper adopts a MADDPG-based model. Each UAV's actions consist of direction and distance traveled. The reward function incorporates data collection amounts, fairness degree, energy consumption, and obstacle collision penalty. A subsequent work [62] from the same research group employ a PPO-based model to address a similar crowd-sensing system. In [62], a BS acts as an agent, assigning tasks to UAVs and directing their actions. In addition to the three optimization objectives mentioned in [60], this study also incorporates data freshness as a consideration. To fully leverage the temporal information of adjacent timeslots in the training data, the paper incorporates Gated Recurrent Units (GRU) [63], a model based on Recurrent Neural Networks (RNNs), during the feature extraction step.

In [30], the authors investigate a UAV-aided wireless sensor network where UAVs carry out missions based on predetermined trajectories and can provide opportunistic assistance to terrestrial sensor networks. The work presents a real-time scheme for UAV scheduling, bandwidth allocation, and power control based on DQN and DDPG to optimize long-term data transmission. The actions of agents involve selecting the target data node, configuring the data node's transmit power, and allocating bandwidth for the UAV. The reward is the cumulative amount of transmitted data. The DQN-based model is adopted for discrete action space, in which transmit power and bandwidth are discretized into fixed numbers. The DDPG-based model is employed to solve continuous action space.

Some practical constraints are considered in [64]. These constraints include collision avoidance and kinematic limitations. The system employs a UAV that is dispatched to collect data within an area containing other moving UAVs and static obstacles. Communication or information exchange among the UAVs is not possible. A sensor on the dispatched UAV is capable of detecting nearby UAVs within a certain range. To enable the UAV to learn decision-making and task accomplishment without relying on additional environmental or UAV-provided information, a model based on D3QN is utilized. The action space of the UAV in this model encompasses rotation angle and speed, considering realistic constraints, while the height factor is disregarded. The reward function used in the model takes into account the quantity of data collected, task completion time, movement penalties, and collision penalties.

In [65], the authors incorporate an additional energy supply as an integral part of the data collection task. UAVs are categorized into two groups: data collectors and energy transmitters. Data collectors are responsible for gathering data from IoT devices, while energy transmitters provide wireless energy transfer to these devices. The primary objective of the research is to maximize the data throughput of IoT devices, ensure data freshness, and enhance the energy efficiency of UAVs. To achieve these goals, UAVs are trained using a model based on MADDPG. The actions available to the UAVs include horizontal flight direction, flying distance, and interval altitudes. The reward function for energy transmitters is based on the residual energy of the device and the number of upload attempts made by each device. For data collectors, the reward function considers data throughput and data freshness. Additionally, there are penalties incorporated within the reward functions to impose constraints on the actions taken by the UAVs.

B. Data Collection Efficiency Enhancement

This section shifts the focus towards enhancing the efficiency of data collection tasks, primarily by reducing the time required for data collection and minimizing the distances traversed by UAVs.

In [66], a multi-UAV-aided data collection system is presented within the context of backscatter communication. UAVs have the ability to activate backscatter devices to gather data from them. If the energy of UAVs is insufficient to complete the mission, they can recharge at a designated charging station. The primary objective of this research is to minimize the total flight time once UAVs have completed their tasks. To achieve this, the backscatter sensors are clustered using the Gaussian mixture model clustering method. Each UAV is then dispatched to collect data from a specific cluster, dividing the target area into multiple sub-areas. If UAVs are unable to cross the boundaries between sub-areas, each UAV only needs to consider the area it is responsible for. However, if UAVs can traverse the boundaries, collaboration between different UAVs becomes necessary. The authors refer to these two boundaries as the deterministic boundary and the ambiguous boundary, respectively. Accordingly, a DQN is adopted for the first situation, while a multi-agent DQN is used for the second situation. The action space for UAVs consists of three choices: moving to a backscatter node, charging, or collecting data. The reward function incorporates penalties for flight time, energy consumption, and flight distance.

The authors in [67] study a case in which UAVs collect data from a variety of IoT devices. Similar to [66], All IoT nodes are clustered first and then each UAV is responsible for a cluster. The clustering approach adopted in this work is K-means. The boundary between clusters is the same as the deterministic boundary in [66]. A multi-agent DQN is used to minimize mission time. The acceleration and velocity of a UAV constitute its action space.

In [68], UAVs are employed for crowd-sensing in an environment with obstacles. In contrast to the previous two works, UAVs in this work select their sensing objects on their own rather than being dispatched to a specific cluster. A MADDPG-based algorithm is used to implement mission selection and path planning for UAVs. A 3-dimensional (3D) simulation environment is built in this work. The action space of a UAV is 3D angular velocity. The authors consider obstacle avoidance, motion control, and the amount of collected data when developing the reward function. The final results show an improvement in both task completion speed and energy efficiency.

An integrated UAV and vehicle system for data collection in a smart city environment is considered in [69]. Vehicles are dispatched to collect most of the data through genetic algorithms. Then UAVs are employed to collect the remaining data. A DQN-based model is adopted in UAV trajectory planning to reduce the flying distance and reduce consumption.

C. Data Freshness Optimization

This section examines the scenario of continuous data collection, where data freshness emerges as a crucial consideration. In this context, UAVs have a dual role in gathering data and transmitting it to BSs or users. As mentioned earlier, the optimizations presented in [62] and [65] take into account the freshness of the collected data. To assess the level of freshness, the Age of Information (AoI) metric is employed [73]. A smaller AoI indicates greater freshness and vice versa. This investigation gives insight into the balance required between data collection and data freshness in a dynamic environment.

TABLE III
SUMMARY OF WORKS ON MUWN-ASSISTED DATA ACCESS, SENSING, AND COLLECTION

Issues	Works	Main Findings	Optimization Objects	RL Algorithms
Data Collection Maximization	[60]	Develops a DRL technique for trajectory design of UAVs in a MUWN-assisted mobile crowd sensing system	Fairness, energy efficiency minimization	Distributed DDPG
	[62]	Proposes a CTDE-based DRL for trajectory planing of UAV-based mobile crowd sensing tasks under time constraints	Fairness, energy efficiency, and data freshness minimization	PPO
	[30]	Presents a DRL approach for optimizing UAV scheduling and power control in UAV-assisted wireless sensor networks	Long-term data transmission maximization	DQN & DDPG
	[64]	Uses DRL to solve practical constraints such as collision avoidance and kinematic limitations in UAV-based data collection	Obstacle avoidance	D3QN
	[65]	Proposes a MADDPG method to help two teams of UAVs transfer power and collects data from a sensor network	Data freshness, energy efficiency minimization	MADDPG
Data Collection Efficiency Enhancement	[66]	Applies SARL and MARL for task allocation of a multi-UAV-aided data collection system in backscatter communication	Total flight time minimization	DQN & Multi agent DQN
	[67]	Utilizes DRL to optimize the UAV's speed, acceleration, and collision avoidance in data collection task	Total flight time minimization	Multi-agent DQN
	[68]	Creates a 3D simulation environment and employs DRL to efficiently control UAVs for data collection	Obstacle avoidance, task completion speed improvement, and energy efficiency minimization	MADDPG
	[69]	Applies DQN to enhance the coverage ratio and reduce collection costs in a vehicle- and UAV-assisted IoT	Flight distance of UAVs minimization	DQN
Data Freshness Optimization	[70]	Introduces a distributed sense-and-send protocol and a compound-action actor-critic algorithm to minimize AoI	AoI minimization	DQN & DDPG
	[71]	Applies DRL to optimize the task allocation and trajectory of UAV in a data collection system with both cellular links and UAV-to-device link	AoI minimization	DDPG
	[72]	Utilizes DRL to jointly optimize UAVs' trajectories and scheduling	AoI minimization	DDPG

In [70], a cooperative group of UAVs performs sensing tasks and transmits collected data to a central BS for a given time. A distributed sense-and-send protocol is proposed that clearly divides the entire task into different steps, including information change, decision-making, sensing and data collection, and data transmission. The authors propose a DDPG- and DQN-based model to minimize the accumulated AoI. The action space of the UAV agent contains its selected task and sensing location, in which the task is a discrete action and the location is continuous. To handle the hybrid action space, training methods of the DQN and DDPG algorithms are adopted. The reward function is the reduction of AoI after the task is completed. A similar case can be found in [71], while some transition tasks from UAVs to users are added. Specifically, data collected by UAVs can be transmitted to the BS via cellular links or directly to mobile devices through UAV-to-device communications. Unlike [70], a MADDPG-based model is utilized to train each UAV in [71]. The action space consists of two continuous actions: sensing location and transmission location. The reward is minus AoI. Both of the above works define the action as a change of position, and the UAV will fly to the specified position at a constant speed.

In [72], the authors propose the utilization of UAVs for data collection from sensors installed on vehicles. The UAV-assisted area is defined as a one-way road of a specific length. Upon a vehicle entering this road, the AoI of its data starts to accumulate. To achieve the minimum cumulative AoI within a designated time frame, the authors employ a model based on DDPG. The action space comprises three components: selected targets, flying distance, and flying direction. The reward functions incorporate penalties for high AoI, long distances, and specific flight restrictions.

In summary, RL-enabled MUWNs offer a promising solution to enhance data access, sensing, and collection tasks. The performance enhancements achieved are evident in the three subsections discussed: data collection maximization, efficiency enhancement, and data freshness optimization. To provide a concise overview of these studies, Table III presents the main findings, optimization objectives, and the specific RL methodologies employed in each work.

IV. RESOURCE ALLOCATION FOR WIRELESS CONNECTIVITY

MUWNs present a beneficial complement to existing terrestrial communication networks, offering flexible deployment, easy management, and adaptive coverage [97], [98]. The utilization of UAVs as aerial BSs, caches, or relays can significantly enhance the network capacity and Quality of Service

(QoS) of the overall network compared to traditional terrestrial networks [31], [74], [77], [79], [91], [99]. However, the performance of such UAV-based networks hinges significantly on the rationality of resource allocation and optimization strategies, such as optimization of power, bandwidth, and storage resources. While traditional techniques such as game theory [100] and convex optimization [101] have been used, RL offers a more suitable approach, empowering MUWNs to adjust their strategy in uncertain and dynamic environments, ultimately reaching an optimal solution. Furthermore, the RL-based agents model enables the optimization of multiple resource allocations simultaneously and maximizes multiple performance aspects of the system.

This section provides an overview of RL-based resource allocation applications for wireless connectivity in MUWNs. We categorize these applications into three main areas: spectrum allocation and power control, wireless power transfer, and cache strategies.

A. Spectrum Allocation and Power Control

MUWNs offer a solution for providing network access to ground users in terrestrial cellular communication or remote areas where terrestrial network access is unavailable. With their flexible deployment and high mobility, MUWNs can effectively address these connectivity challenges. However, the limited availability of frequency resources and onboard hardware on UAVs imposes constraints on radio resource utilization. Therefore, it becomes crucial to implement reasonable spectrum and power control schemes to ensure the efficient usage of MUWNs. Resource management schemes, aimed at maximizing system throughput or fairness, are commonly employed in various applications such as UAV-assisted terrestrial communications [102], IoT [78], and IoV [94]. These schemes play a pivotal role in optimizing spectrum and power allocation within MUWNs.

1) Spectrum Allocation: Aiming at the scenario of terrestrial communication assisted by quad-rotor UAV, the authors of [74] jointly consider 3D UAV trajectory design and frequency band allocation. The authors aim to maximize the defined fair throughput, so as to determine the UAV 3D location and the frequency band allocation strategies. A DDPG-based model is proposed to obtain the optimal strategies by offline training and implementation phrases with forward propagation. In the practical application of DRL, there exist dimension imbalance, gradient vanishing, and training oscillations. Dimension spread, pre-activation and softmax reference techniques are adopted in [74] to address these issues respectively. Simulation results show that compared with randomly generated strategies, the proposed method achieves much better performance in terms of fairness and total throughput.

The authors of [75] jointly consider UAVs' positions, UAV-user Equipment (UE) association, and transmit beamforming at the UAVs. Such a joint design requires accurate Channel State Information (CSI) estimation to obtain a better system performance. However, the rapidly changing position of the UAV makes it hard to obtain perfect CSI. Motivated by this, a Q-learning-based method in conjunction with a deterministic optimization technique called Difference of Convex Algorithm (DCA) is proposed. Different from the used models in [103] and [104] which the CSI availability is assumed without the location information of the UAVs, the authors propose a DQN method to obtain accurate CSI associated with the location of UAVs. High-quality transmit beamforming is produced and UAV-UE association is determined based on the calculated location of UAVs. Meanwhile, the reward is computed by DCA to construct the decision policy of Q-learning which can update the location of the UAV. With the aid of DCA, the Q-learning algorithm only needs to train the UAV with a smaller set of variables, which greatly reduces the state-action space. Experiment results demonstrate that the DCA can achieve higher sum achievable rates of UEs compared with traditional convex approximation-based solutions. In practical applications, due to insufficient communication resources for signaling, the UAVs are usually unable to acquire user location and channel parameters beforehand and have to make decisions based on their observations. In this case, the authors of [76] propose a distributed MARL approach based on DQN to jointly design trajectory and power control, which can optimize the cellular network downlink throughput. Decisions on user-channel assignment, transmit power allocation, and one-step movement are made by UAVs. In the distributed learning architecture, global model training is accomplished via local gradient trained by each UAV such that the whole training process can be carried out in parallel without observation data sharing. Experiment results show that ground users achieve almost equal individual throughput, indicating the proposed fairness indicator can optimize policy to achieve better fairness.

In some typical IoT scenarios, massive devices require high-quality and ubiquitous connectivity over a large scale area. With a High Altitude Platform (HAP) or BS, UAVs (or even a single UAV) can act as aerial BSs to improve terrestrial network throughput, expand network coverage, and offload compute-intensive tasks from IoT devices [105], [106]. The network coverage in remote and hard-to-reach areas is considered in [78]. The aerial access network consists of one HAP and a UAV swarm that supports the IoT devices by providing communication and computation services. In order to improve the energy efficiency of aerial computing servers and QoS of IoT devices located in underserved areas, the authors jointly address the problem of device-UAV association, task offloading, and bandwidth allocation with the goal of maximizing QoS as well as minimizing total energy consumption of IoT devices. Taking into account the non-convexity and complexity of the targeted problem as well as the dynamic nature of the network, the HAP and UAV swarm are regarded as multi-agents so the targeted issue can be transformed into Markov games. Furthermore, the authors propose a MARL algorithm based on DDPG. To ensure the stability and convergence of learning, a CTDE framework is adopted. It can be found that using UAV to provide edge computing services receive significant research interest, as it enables devices with computational-intensive tasks to offload them to the edge server [107] in a flexible way. The authors of [108] investigate the sum power minimization problem in a MEC system with multiple UAVs. Especially, spectrum allocation and transmission power control are jointly optimized. To acquire the optimal resource allocation strategy, a semi-distributed federated MARL algorithm with the integration of federated learning [109] and DRL is proposed. Through this proposed algorithm, the UEs can learn models quickly by training the local data. In addition, UAV can assist the development of a vehicular network which is a crucial use case for the upcoming 6G [110]. The vehicular network has created diversified novel applications with extremely multiple service requirements, including ultra-high reliability, delay sensitivity, high bandwidth requirements, and computing-intensive applications [111], [112]. In [77], the UAVs are acting as the aerial BSs to enhance the network resource allocation fairness for vehicles. The authors jointly consider the UAVs' flying range, communication range, and energy constraints to maximize the total throughput of vehicles. A DRL approach based on the SAC is proposed to achieve proportional bandwidth allocation and maximize the total throughput of all vehicles.

In the field of radio resource sharing, cognitive radio, and software-defined radio technologies provide substantial contributions. RL also plays a vital role, offering potential benefits to these technologically-enhanced MUWNs. One notable example of this application is discussed in [81], where the authors put forward a system that employs a cluster of cognitive UAVs operating in an overlay mode to access multiple orthogonal primary spectrum resources. The tasks allocated to these UAVs include cooperative area sensing, data backhaul operations, and cooperative spectrum sensing. Consequently, these operations enable cognitive UAVs to opportunistically exploit idle spectrum resources owned by primary users. To optimize the utility of primary user channels, the researchers incorporate a multi-agent DQN framework. Independent agent training is leveraged within this framework for primary user channel selection. The proposed algorithms, as evidenced by numerical simulations, have demonstrated the potential in improving both sensing accuracy and channel utilization. Another contribution in the same domain is presented in [82]. The authors present a novel framework that involves a UAV-assisted IoT network utilizing a UAV as a relay to alleviate the increasing data traffic. By incorporating cognitive radio technology, the UAV is capable of identifying and utilizing idle spectrums, thereby establishing a wireless backhaul link to facilitate efficient data transmission, particularly in scenarios where spectrum availability is limited. A DQN-based model is employed to enhance the energy efficiency of the traffic offloading system. The proposed strategy simultaneously optimizes four actions of UAV: the UAV's flying trajectory, the timing for data collection or transmission, the control of transmission power, and the band selection. By integrating cognitive radio technology with RL, the UAV achieves enhanced functionality, particularly in dynamic and complex environments. Similarly, the author of [83] introduces a cognitive satellite-UAV framework for IoT. In this design, satellites hold spectrum priority, while UAVs share the spectrum to serve ground users via Non-Orthogonal Multiple Access (NOMA) technology. The authors present a joint optimization problem aimed at minimizing ground users' transmission latency, while controlling UAVs' power allocation and trajectory. To address this issue, a MADDPG-based algorithm is proposed, demonstrating minimized transmission latency and thus proving promising for IoT services via cognitive satellite-aerial networks. Finally, the authors in [84] propose an integrated framework that combines software-defined radios to enable updates in UAV networking. They consider the constraints of power and computational resources inherent to UAVs. Ground stations equipped with SDR transceivers gather information related to UAV status and performance assessment. This information is subsequently processed through RL algorithms, which in turn guide the SDR transceivers to modify the communication configurations, synchronizing the updating of UAV networking. The authors illustrate the effectiveness of this approach with an example that involves adjusting the connection links between UAVs to maximize bandwidth and channels while minimizing power consumption. The integration of software-defined radio and RL, as proposed in this study, shows promising potential for resource allocation in MUWNs.

2) Power Allocation: The transmit power allocation and control are crucial for the design of MUWNs due to the hardware limitation and energy supplement of the UAVs. In [31], the problem of dynamic resource allocation on providing on-demand communication service for ground users in the MUWN is studied. To maximize the expected rewards, the authors formulate the long-term resource allocation problem as a stochastic game and propose a MARL framework based on Q-learning to obtain the optimal resource allocation strategy, including power allocation and UAV-device association. In [102], to minimize the interference caused by UAVs, the authors propose a downlink/uplink decoupled access scheme for cellular-enabled UAV communication systems. The control and data links of UAVs and the uplinks and downlinks of UEs are separated into different serving BSs and operating frequencies. A DQN-based approach is proposed to improve the energy efficiency of the system. The ground users are considered to be a process of continuous movement, so the UAV swarm that provides cellular offloading needs to be redeployed dynamically according to the mobility of the user. In an effort to solve this pertinent dynamic problem, the authors of [79] present a mutual DQN algorithm to determine the optimal 3D trajectory and power allocation of UAVs. The limited battery capacity determines the service duration of UAV-assisted terrestrial cellular networks. Optimizing the deployment location and transmission power of UAVs in the network to maximize energy efficiency is an effective way to improve the QoE of ground users. The authors of [80] develop a multi-agent DQN-based approach for controlling UAV deployment and power transmissions individually. The proposed algorithm can be implemented in practical UAVs with limited computation power due to the fact that distributed Q-learning doesn't require deep neural networks for function approximation. Signal enhancement and interference reduction can be accomplished by reflecting the received signal to the destination using an intelligently controlled reflection original, i.e., Reconfigurable Intelligent Surface (RIS). In scenarios with dense buildings, frequent shadow effects will cause serious channel fading. The RIS-assisted MUWNs have the capacity to facilitate network performance. In [113], the authors investigate RIS-aided MUWNs, and DRL method based on DDPG and PPO technique is utilized to jointly optimize the transmit power at the UAV and the phase-shift matrix at the RIS.

B. Wireless Power Transfer

1) WPT for IoT Devices: In IoT networks, batteries of low-power devices with small capacities need to be recharged or replaced periodically to ensure long-term operation. The UAVs can act as RF energy transmitters, which can charge devices through WPT to extend the lifetime of IoT networks effectively [114], [115]. Transmission distance from UAV to devices can be flexibly adjusted and Line-of-sight (LoS) channel can be achieved easily owing to the high maneuver ability of UAVs. As a consequence, WPT-enabled UAV-assisted IoT networks can enhance data collection lifecycle and RF energy transfer efficiency remarkably.

In UAV-assisted WPT scenarios, the coupling between UAV trajectory design and resource allocation, as well as the massive number of devices bring high complexity to the network design, which is not suitable for using traditional approaches including convex optimization [116] and dynamic programming [117]. DRL is very effective in this context to solve the mixed-integer nonconvex programming problem in UAV-assisted IoT systems. The authors of [86] jointly optimize UAV trajectory and wireless energy scheduling to minimize the average delay of the IoT devices. UAVs are dispatched to charge IoT devices by using WPT and IoT devices can transmit data in the uplink using received energy. The formulated problem is hard to solve due to its complex combinational optimization nature. Then a DRL method based on DQN constructed by two artificial neural networks is presented to find a near-optimal solution.

Enhancing the power supply of energy-constrained IoT devices is crucial to improve the data freshness in IoT networks. The authors of [87] propose an AoI-oriented UAV-enabled WPT approach under time-varying channel conditions to ensure sustainable energy supply for efficient data transmission. The RIS-assisted UAV communication is investigated in [90], where the authors jointly optimize the UAV trajectory and the power allocation of the UAV, the energy harvesting scheduling of IoT devices, and the phase-shift matrix of the RIS. Two DRL algorithms based on DDPG and PPO are proposed to obtain the optimal solution to maximize the network sum-rate. The authors of [118] investigate the fairness problem between energy transmitters (i.e., UAVs) and energy receivers (i.e., IoT devices). In order to maximize the minimum harvested energy, the author uses a SARL-based scheme to optimize the trajectory of UAV.

In the UAV-assisted wireless powered communication system, service requests of IoT devices are in general time-varying and uncertain. The authors of [88] propose a DQN-based scheme for UAV trajectory planning to minimize the average data buffer length and maximize the remaining charge of the battery of the UAVs. The authors of [85] consider joint optimization of multiple objectives with a goal of maximizing the sum data rate and the harvested energy while reducing the energy consumption of the UAV. A multiple-objective DDPG algorithm is developed to find the optimal trajectory of the UAVs. Battery drain issues and buffer overflow of IoT devices may bring mistakes in uplink data transmission. Resource management problem is large-scale and time-varying in MUWNs, which induces trickiness in finding the optimal solution within an acceptable time. Motivated by these issues, the authors of [89] present a novel data-driven DRL framework based on DDPG to train resource allocation of the UAVs, so that the data packet loss is minimized.

2) WPT for UAVs: The lifecycle of a UAV determines the duration of providing aerial services (e.g., goods delivery, data collection and wireless connectivity for ground users). Extending the lifetime of MUWNs to improve the QoS is an urgent problem to be solved [119]. However, terrestrial charge stations pre-installed in fixed locations are not conducive to the UAVs providing intelligent services in the time-varying environment. One feasible solution is to provide power during the mission of the UAV by using far-field WPT technology. For instance, it is possible to empower onboard batteries of UAVs wirelessly by flying energy sources [120]. The authors of [120] propose a DDPG-based scheme to jointly optimize the sum-energy received by the UAV swarm, the energy loading process of terrestrial charge stations and the most energy-efficient trajectories of power supply UAVs. To improve the service duration of UAVs, the authors of [121] investigate a typical scenario, in which the UAVs are categorized into charging UAV and mission UAV. The trajectory and recharging process of charging UAVs are designed by using DDPG to minimize mission time. The authors of [122] study the UAV charging scheduling strategy according to the location distribution of UAVs. Furthermore, in [123], the authors study the application of UAV swarm for data collection with wireless charging. A Q-learning-based algorithm is presented to find the optimal trajectory of UAVs with consideration of energy consumption for hovering and flying of UAVs and delay due to wireless charging. Similarly, the authors of [124] investigate the optimal trajectory to minimize the flying distance and service duration of UAVs by using Q-learning.

Owing to both the IoT devices and UAVs being energy constraints, ensuring the energy supply of UAVs and IoT devices to provide continuous service is difficult to be addressed. The UAV-assisted data collection and delivery scheme in IoT is investigated in [125], where the UAVs are equipped with WPT infrastructure to charge devices, as well as receiving energy from BS to charge UAV batteries. However, in some specific remote areas, the UAV swarm cannot receive energy from BS. Under this circumstance, joint consideration of energy supply for IoT devices and UAVs is still an under-investigation problem.

C. Caching Strategy

The rapid development of mobile communication has brought exponential growth in data traffic, of which more than half is used for content transmission [126] and leads to backhaul network congestion [127]. One strategy for mitigating this issue involves moving network resources closer to users, such as caching popular content at BSs or edge nodes. This approach enhances network QoS and Quality of Experience (QoE) [128], [129]. However, distributing content in high-demand hotspots exclusively via BSs or edge nodes with fixed capacity results in lower data rates due to user mobility and fixed server locations. This setup may lead to two potential issues: (a) the inability to timely distribute cached content to users, and (b) the lack of cached content requested by new users, both of which generate additional network latency and bandwidth consumption.

In hotspots with temporary high user density, adding fixed terrestrial cache nodes is often cost-prohibitive. As a solution, MUWNs assist terrestrial cellular networks by leveraging UAVs as cache-enabled aerial BSs, thereby enhancing access capacity and reducing network delay. Previous works have utilized traditional methods such as heuristic algorithms [130] and Lyapunov optimization [131] to design cached content placement and distribution. However, finding the optimal solution within polynomial time complexity using traditional methods is challenging due to the time-varying nature of the environment, including channel states and user locations.

Caching content at UAVs has been demonstrated to be effective in reducing backhaul traffic for applications with intensive content requests, such as augmented reality and multimedia. The authors of [91] investigate cache-enabling UAV NOMA networks to assist terrestrial cellular communication. The user association, power allocation of NOMA, deployment and caching placement of UAVs are jointly optimized to minimize the content delivery delay. A DDPG-based DRL algorithm is proposed to find the optimal power allocation, UAV deployment and Caching Placement with low complexity. To tackle the long-term caching placement and resource allocation, the authors of [132] adopt a Q-learning-based DRL algorithm, in which the UAV as the agent learns and selects action with soft ϵ-greedy strategy. In [92], the authors present a generic algorithm to integrate the use of human-centric features and random waypoint user mobility. To maximize QoE satisfaction and reduce the transmit power of the UAVs, a DRL method based on dueling DQN is proposed to predict the location of UAVs and contents to cache at the UAVs. In [93], the authors investigate the content transmission problem, in which multiple cache-enabled UAVs are evolved to offload data traffic in a heavy-crowded cellular network. In this novel optimization problem, multiuser association, cache placement, UAV trajectory and transmission power are jointly considered. The macro BS and UAVs act as agents interacting with the environment and a dual-clip PPO algorithm is designed to achieve minimizing the sum content acquisition delay of user devices.

Device-to-device (D2D) caching is an effective approach to improve network throughput and alleviate backhaul burden. In [95], the authors consider the cache placement optimization problem in a D2D-enabled MUWN. To improve the QoE of the requesting devices, a cache placement strategy optimization problem is investigated in order to minimize the file access latency for all users. The authors present a DDPG-based optimization algorithm to determine the concrete cached content and location. Furthermore, the authors

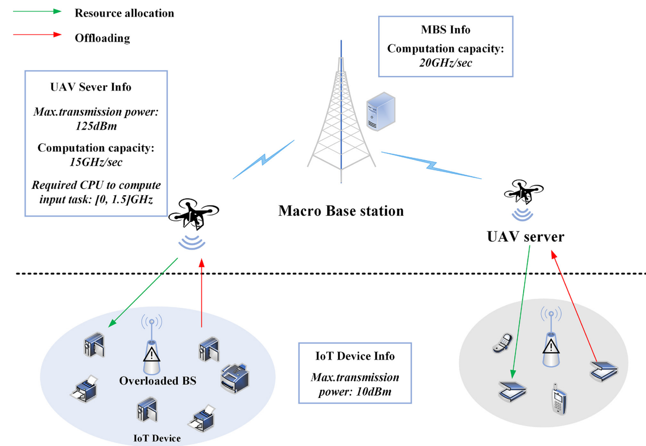


Fig. 5. A scenario of MUWN-assisted MEC.

of [96] investigate a dynamic 3D trajectory design of multiple UAVs in a wireless D2D caching network. Cooperative MARL is elaborately designed to cope with time-varying network topology and the coexistence of aerial and terrestrial caching nodes. Simulation results demonstrated the proposed method can significantly improve the network throughput. In vehicular networks, infotainment will become one of the crucial ways to enhance the vehicle user experience. MUWNs can be deployed to provide content delivery for vehicular networks. In [94], the authors consider the problem of content delivery to vehicles on road segments with either overloaded or no available communication infrastructure. Specifically, caching decisions, UAV trajectory, and radio resource allocation are jointly designed. Due to the dynamic environment, a PPO-based algorithm is implemented to find the caching strategy, and mobility of UAV, which can maximize the energy efficiency of the UAV.

In conclusion, RL offers a compelling solution for resource allocation in MUWNs. The strategy enhancements for various resource allocations have been explored, including spectrum allocation and power control, wireless power transfer, and cache strategies. To succinctly summarize these studies, Table IV provides a consolidated overview of the main findings, optimization objectives, and RL algorithms utilized across a variety of MUWN applications.

V. UAV-Assisted Mobile Edge Computing

MEC emerges as a novel architecture that significantly enhances the user experience on edge devices. By equipping UAVs with computing and communication platforms, they offer offloading services to fulfill the requirements of latency-sensitive and energy-sensitive applications. Fig. 5 illustrates a scenario of UAV-assisted MEC, where IoT devices can access the computing resources offered by UAV servers. Detailed information about this setup, such as the computing capacity and transmission power of the UAV, the capacity of the macro base station, and the transfer power of IoT devices, is provided based on the specifications described in [145]. To enhance the performance of MUWNs-based MEC systems, it is crucial to make informed decisions regarding offloading, computation resource allocation, and power management.

TABLE IV
SUMMARY OF WORKS ON RESOURCE ALLOCATION OF MUWNS

Issue	Works	Main Findings	Optimization Objects	RL Algorithms
Spectrum Allocation and Power Control	[74]	Develops a DRL-based method for 3D UAV trajectory design and frequency band allocation	Fairness throughput maximization	DDPG
	[75]	Uses a Q-learning approach to manage UAVs' positions and a difference of convex algorithm for beamforming and UAV-UE association	Sum-rate maximization	Q-Learning
	[76]	Introduces a MARL approach for resource allocation and trajectory design in a decentralized UAV-aided cellular network	Overall throughput, fairness throughput maximization	DQN
	[77]	Proposes a DRL method for optimal UAV positioning and resource allocation in vehicular communications	Total throughput maximization	SAC
	[78]	Presents a MARL approach for optimal resource allocation and offloading in aerial access IoT networks	Number of completed tasks maximization, energy consumption minimization	DDPG
	[79]	Introduces a mutual DQN algorithm for efficient trajectory design and power allocation in UAV networks	Throughput maximization	DQN
	[80]	Proposes a multiagent Q-learning algorithm for optimal deployment and power control in UAV-based networks	Energy efficiency maximization, number of outage users minimization	Q-Learning
	[31]	Develops a MARL framework for dynamic resource allocation in UAV networks	Throughput and power consumption trade-off	Q-Learning
	[81]	Uses a hybrid MARL approach for spectrum sensing and channel access in cognitive UAV networks	Utility of primary user channels maximization	Multi-agent DQN
	[82]	Uses DRL to optimize UAV paths and control in an uncertain network environment	Energy efficiency maximization	DQN
	[83]	Proposes a MARL method to control UAVs in a cognitive satellite-UAV network	Users' transmission latency minimization	MADDPG
	[84]	Proposes an integrated framework that combines software defined radios to enable updates in UAV networking	Bandwidth and channels maximization, power consumption minimization	Q-learning
WPT	[85]	Applies extended DDPG for optimizing UAV path planning considering data rate, energy harvest, and UAV energy consumption	Sum-rate, harvested energy maximization, energy consumption minimization	DDPG
	[86]	Uses DQN to optimize UAV trajectory and scheduling in UAV-assisted WPT network	Average AoI minimization	DQN
	[87]	Proposes DRL-based UAV trajectory planning to optimize system-level AoI in IoT networks	Sum of AoI minimization	DQN
	[88]	Applies DQN for UAV flight control to balance data buffer length and battery life in a MUWN-assisted WPT system	Average data buffer length minimization, residual battery level maximization	DQN
	[89]	Develops an on-board DQN for transmit power allocation, trajectory design, and transmission schedule in UAV-assisted WPT and data collectection system	Overall data package loss minimization	DQN
	[90]	Proposes DRL algorithms to maximize the total network sum-rate in RIS-assisted UAV communications	Sum-rate maximization	DDPG & PPO
Cache Strategy	[91]	Proposes DRL for joint optimization of resources, UAV deployment, and caching to minimize delivery delay in UAV NOMA networks	Content delivery delay minimization	DDPG
	[92]	Uses dueling DQN to optimize UAV location and content caching	QoE satisfaction maximization	Dueling DQN
	[93]	Utilizes PPO to optimize user association, cache placement, and UAV trajectory in a cellular network	Sum content acquisition delay minimization	PPO
	[94]	Applies PPO for optimization of caching decisions, UAV trajectory, and radio resource allocation in vehicle content delivery scenarios	Energy efficiency maximization	PPO
	[95]	Proposes a DDPG-based caching placement strategy in UAV-relaying networks	File access delay minimization	DDPG
	[96]	Presents a MARL framework for dynamic 3D trajectory design of cache-enabled UAVs	Network throughput maximization	Q-Learning

RL can be a valuable approach in tackling this challenge and improving system performance. By employing RL techniques, agents can autonomously learn and make optimal decisions related to offloading tasks, allocation of computation resources, and even the design of UAV edge server trajectories. RL enables the system to adapt and optimize its operations based on the observed environment, performance metrics, and predefined reward functions. This empowers the system to dynamically adjust its strategies and ultimately enhance the overall performance of the MUWNs-based MEC system.

This section is structured into four subsections, each highlighting a crucial aspect of the MEC system. Firstly, we present an overview of the efforts made to optimize latency and energy-related aspects in UAV MEC systems. Next, we delve into resource management and computation offloading within the context of UAV MEC systems. Lastly, we examine the system from the perspective of other objective optimization, considering objectives beyond latency and energy optimization, such as packet loss rate and fairness.

A. Latency-Related Optimization and Design

This section is dedicated to exploring latency-related optimization and design in UAV-assisted MEC systems. Latency is a critical factor in ensuring efficient and responsive communication between IoT devices and UAV servers. By minimizing latency, the system can achieve improved real-time performance and user experience.

In [133], a group of UAVs formed a network and is used to provide MEC services and complicated computing tasks are separated into various task streams. Considering the dynamic network state and the energy constraints of the UAV, the average mission response time minimization problem is formulated and modeled by a MDP. A MARL-based algorithm is then proposed to address the formulated problem. Besides, the authors extend the proposed on-policy Multi-agent AC-based computation offloading algorithm to an off-policy algorithm. In [134], considering the unloading delay and energy efficiency of users, the authors formulate a mixed-integer nonlinear problem in which the interaction between multi-users and multi-UAV is modeled by game theory. Furthermore, a MADDPG-based method is used to optimize the trajectory of the UAV and achieve obstacle avoidance. Space-air-ground integrated networks have recently attracted a lot of research interest. In [135], considering the mobility of UAVs and dynamic network traffic, the authors propose a double Q-learning algorithm with an improved delay-sensitive replay memory algorithm to allow UAVs to learn based on local and neighbor history information and make decisions about offloading policies. The proposed algorithm greatly reduces the packet loss rate and transmission delay.

In [136], a UAV-assisted MEC system for the aquatic environment is considered. The authors construct a two-layer UAV marine communication network, which includes a centralized upper-layer UAV and a group of distributed lower-layer UAVs, and MEC is implemented in the upper layer. The authors present a delay minimization problem considering both communication and computation delays and transform this problem into a MDP. The DQN-based algorithm and DDPG-based algorithm are then proposed to optimize the trajectory of the upper UAV. As we can observe, in the scenario of UAV-assisted MEC, many RL-based solutions can greatly reduce the time for UAVs to complete tasks. However, due to the memory requirements of RL, RL-based schemes may face some difficulties in implementation. The method proposed in [137] reduces the required information and training time on the premise of ensuring the learning process. In [138], the authors propose a task offloading scheme with federated learning based on DQN, which can optimize the time for sensing task of UAV by reducing the energy consumption of computing, while maximizing the task completion rate. The proposed scheme can also improve offloading performance while ensuring the privacy of UAV data. Additionally, the authors propose lightweight agnostic defense mechanisms to combat backdoors in multi-UAV settings.

B. Energy-Related Optimization and Design

Energy consumption optimization is another critical challenge in UAV-assisted MEC systems, particularly in remote areas where the energy supply for UAVs is often limited. This section reviews relevant research and works focused on energy optimization in such systems.

In [139], UAVs can serve as aerial BSs to provide edge services to vehicles on the road. The cloud computing center server can predict road traffic conditions in real-time, and assign the UAVs to different mission areas. In the formed trajectory optimization problem, the goal is to reduce the flying and turning energy of the UAV. In addition, the proposed energy-saving deployment scheme based on RL can obtain the optimal hovering position of the UAV. In [140], the energy consumption of all users is minimized by jointly optimizing user association, resource allocation, and UAV trajectories. Firstly, the proposed trajectory optimization problem is solved by the convex optimization method. In addition, for the dynamic environment where UAVs take off from different locations, an AC-based trajectory optimization method is proposed, which can make real-time decisions. In [141], the authors investigate the problem of minimizing energy consumption in MEC systems. The goal of the problem is to minimize the energy consumption of devices considering the optimization of the UAV trajectory and the AoI of the state update package. The authors propose an RL scheme for decision-making using the D3QN-based model.

C. Resource Management and Computation Offloading

This section focuses on resource management and computation offloading decisions in UAV-assisted MEC systems. The reasonable allocation of spectrum, time, computing, and other resources plays a crucial role in maximizing their utilization and improving system performance.

In [142], the authors propose a blockchain-MEC model that includes a cloud-based server and various UAVs, where the server is used to perform blockchain tasks, and the UAVs with MEC units are used to collect data from local devices. In the MEC system, many ground BS and UAVs can transmit tasks to the cloud-based server and also run some blockchain tasks. The authors model resource management and pricing in such a MEC system as a Stackelberg game, and propose an unsupervised hierarchical deep learning algorithm based on deep Q-learning and Bayesian deep learning. In [143], a network framework for a collaborative UAV-assisted MEC system is presented, in which UAVs can help each other to perform computing tasks. Considering the interference mitigation from UAVs to devices, the authors simultaneously optimize the offloading decision and resource management strategy to maximize the long-term utility. The problem is defined as a semi-Markov process considering the random needs of users and the time-varying communication channel and then the proposed RL-based algorithms are implemented in both centralized and distributed manners.

In [144], the multi-dimensional resource management problem of UAV -vehicular MEC system is studied. To efficiently provide on-demand resource allocation, both the macro eNodeB and the UAV are installed with multi-access MEC servers, and collaborate to make association decisions and resource allocation strategies for the vehicles. A distributed optimization problem is formulated to maximize the number of offloaded tasks while satisfying heterogeneous QoS requirements, and then a MADDPG-based solution is proposed. Through offline centralized training of the MADDPG model, the MEC server can quickly make vehicle association and resource allocation decisions during the online execution phase.

In [145], multiple UAVs are used to provide computing offloading and resource allocation services for IoT devices. Considering the QoS requirements of IoT devices, the authors present a computing cost minimization problem in terms of energy and latency, and extend it to a Markov game. In [146], the authors investigate a multi-UAV-assisted hierarchical MEC network. To maximize the average QoE overall time slots, the allocation of bandwidth and computing resources and the UAV's trajectory are jointly optimized. In [147], the authors investigate computing offloading in space-air-ground integrated networks. A resource allocation and task scheduling method is proposed to efficiently allocate computing resources to different virtual machines. The authors define the offloading decision problem as a MDP and utilize policy gradient and AC methods to improve system performance.

D. Optimization of Other Objectives

This section takes a comprehensive view of the UAV-assisted MEC system by considering objectives beyond latency and energy optimization, packet loss rate and fairness.

In a UAV-assisted MEC system, multiple performance metrics should be able to be jointly optimized in a dynamic manner to guarantee continuous high-quality service. The authors of [148] propose a distributed architecture that uses multi-agent AC to dynamically offload tasks from UAVs to the edge cloud. By learning the best actions from the environment, the total delay in receiving the message from the user and the power of UAV are minimized together. In [149], the authors investigate a system for coordinated task offloading between multiple UAVs and multiple edge nodes. The objective of this problem is to minimize execution latency and energy consumption by jointly optimizing the trajectory of the UAV, task allocation, and radio resource. The authors convert it into a MDP and propose an algorithm based on multi-agent TD3 to solve it efficiently. In [150], the authors consider a heterogeneous wireless network model that includes multiple devices, multiple MEC servers, and multiple UAVs, where both UAVs and servers can harvest energy from renewable resources. The proposed computing offloading problem is to minimize the cost sum of latency and energy consumption. In order to achieve the above goals, the authors propose two DQN-based methods to find the optimal offloading strategy.

In [151], the authors study network slicing with UAVs and MEC devices. Since the computing unit on the UAV consumes considerable energy and affects the flight process, the authors design a system controller that can turn off the computing unit of the UAV, and can offload computing tasks to other UAVs. The main objective is to maximize power consumption, task loss, and incurred delay. The system model is extended to a Markov process so that the authors use an RL-based scheme to solve this multi-objective problem. In [152], the authors investigate the problem of UAV trajectory design in the UAV-assisted MEC system. The goal of the work is to optimize geographic fairness for all users, fairness for UAV load balancing and total energy consumption. A MARL algorithm based on MADDPG is proposed to independently design each UAV trajectory. After obtaining the trajectory of the UAV, a low-complexity method is introduced to optimize the offloading decision of users.

Most of the existing research has focused on computing offloading and UAV trajectory optimization problems, while it is difficult to handle dynamic environments where the locations of UAVs and devices are constantly changing. Despite many RL-based methods, it is difficult to deal with situations with multiple UAVs and a large number of devices. In [153], the authors propose a DDPG-based scheme, which can obtain real-time scheduling strategies in dynamic environments for large-scale UAV-assisted MEC. Furthermore, the proposed method can be more scalable through hierarchical RL and improves learning efficiency, computing efficiency and average task latency. As we can see, effectively utilizing SARL or MARL in the UAV-assisted MEC system networks also faces many challenges. With a large number of different types of nodes and UAVs, the state and action spaces can grow exponentially. This leads to a decrease in the convergence performance of the RL-based algorithm. In addition, in many papers, computation offloading and resource allocation are often performed in a single workflow, that is, they are performed sequentially. In order to further improve the system performance, multiple workflows can be considered to arrange and schedule different tasks to simplify the implementation of RL.

In conclusion, integrating RL in MUWN shows promise for optimizing UAV-assisted MEC systems. The four subsections in this section highlight the potential of RL in optimizing various aspects of the system. Table V provides a summarized overview of the main findings, optimization objectives, and RL algorithms used in different MUWN applications within UAV-assisted MEC systems.

VI. Localization

UAVs possess substantial potential in applications such as search and rescue, surveillance, target tracking, and positioning. Search and rescue operations, for instance, often occur in contexts where natural disasters have resulted in unavailable public services and disrupted infrastructures. In such challenging conditions, MUWNs can serve as crucial auxiliary tools for tasks like target search and rescue and positioning. By leveraging a range of sensors installed on UAVs, including cameras, radars, and signal receivers, MUWNs can acquire comprehensive environmental awareness capabilities. For instance, airborne cameras paired with computer vision

TABLE V
SUMMARY OF WORKS ON UAV-ASSISTED MEC USING RL

Issue	Works	Main Findings	Optimization Objects	RL Algorithms
Latency-related Optimization and Design	[133]	Proposes a MUWN-assisted computation offloading system using MARL to make offloading decisions and allocate bandwidth	Average mission response time minimization	Multi-agent AC
	[134]	Develops an approach combining potential game and MADDPG for optimizing UAVs' trajectory, service assignment and data offloading	Offloading delay minimization, energy efficiency maxmization	MADDPG
	[135]	Designs a RL-based traffic offloading method for Space-Air-Ground Integrated Networks	Packet drop rate,and transmission delay minimization	Q-learning
	[136]	Introduces DQN and DDPG algorithms for optimizing the trajectory of UAV and configuration of virtual machines in a maritime UAV communication network	Latency minimization	DQN & DDPG
	[137]	Proposes a simplified DQN-based formalization for UAV to cloud task offloading	Total delay minimization	DQN
	[138]	Proposes a federated DRL-based task offloading and a triggerless backdoor attack scheme, and corresponding lightweight defense mechanisms	mission time and task completion rate maximization	DQN
Energy-related Optimization and Design	[139]	Designs a pre-dispatch UAV-assisted system that uses DRL for optimal UAV flight trajectory and deployment in vehicular MEC	Energy cost of UAVs minimization	DQN
	[140]	Proposes two trajectory control algorithms for MUWN-assisted MEC to optimize energy consumption and user association	UEs' energy consumption minimization	AC
	[141]	Proposes a DRL solution for UAV path planning and energy efficient computation offloading in a UAV-IoT MEC system	Devices' energy consumption minimization	D3QN
Resource Management and Allocation	[142]	Introduces a hierarchical RL algorithm for resource management and pricing in a MEC IoT system with UAVs	Payoffs of BSs and peers maximization	Deep Q-learning
	[143]	Develops cooperative offloading and resource management schemes in a UAV-enabled MEC network for power IoT system using DRL	Long-term utility maximization	Distributed DRL
	[144]	Proposes a MADDPG for resource management in MEC- and UAV-assisted vehicular networks	Number of offloaded tasks maximization	MADDPG
	[145]	Introduces a MADRL approach to minimize computation cost and ensure QoS in a multi-UAV-enabled IoT edge network	Overall network computation cost minimization	MADDPG
	[146]	Proposes a DRL algorithm for joint resource allocation and path planning in UAV-assisted edge computing	Average total QoE maximization	DDPG
	[147]	Proposes a learning-based system for optimizing computation offloading in space-air-ground integrated networks	Sum delay minimization	Policy gradient
Other objective Optimization	[148]	Utilizes MARL to create a dynamic task offloading system in UAV networks	Improve the energy efficiency, and task completion time	Multi-agent AC
	[149]	Introduces a MARL system for collaborative task offloading in UAV-assisted MEC systems	Sum of execution, delays and energy minimization	Multi-agent TD3
	[150]	Develops a DRL scheme for distributed computation offloading in heterogeneous wireless networks with multi-access MEC	Weighted average cost minimization	DQN
	[151]	Creates a 5G network slice extension system managed via RL, integrating UAVs equipped with multi-access MEC facilities	The defined objective with power consumption, job loss, and incurred delay maximization	Q-Learning
	[152]	Uses MARL for trajectory planning in multi-UAV assisted MEC, enhancing UAV management and offloading decisions	Geographical fairness, fairness of each UAV' UE-load, overall energy consumption of UEs maximization	MADDPG
	[153]	Proposes a scalable scheduling approach for large-scale UAV-assisted MEC using hierarchical RL	Computing efficiency maximization average task latency minimization	DDPG

technology enable the detection and location of ground targets. Furthermore, the location of these targets can be verified through the measurement of Received Signal Strength (RSS) by onboard RF sensors. RL-based agents thus can optimize the RSS or distance to ensure accurate target localization by the UAV. Fig. 6 illustrates a typical localization scenario.

This section presents a review of research works on target tracking, surveillance, and localization utilizing RL-enabled MUWNs.

In [33], the authors introduce a target tracking system that involves a single RF target with a fixed position and multiple UAVs equipped with omnidirectional RSS sensors. In this system, the reward function of the RL model is set as the current RSS value minus the average RSS over a certain period of time in the past. Therefore, the RL model can instruct the UAV to fly in the direction of increasing RSS value. According to prior knowledge, the greater the RSS, the closer the UAV is to the target, and the path loss between them is the smallest.

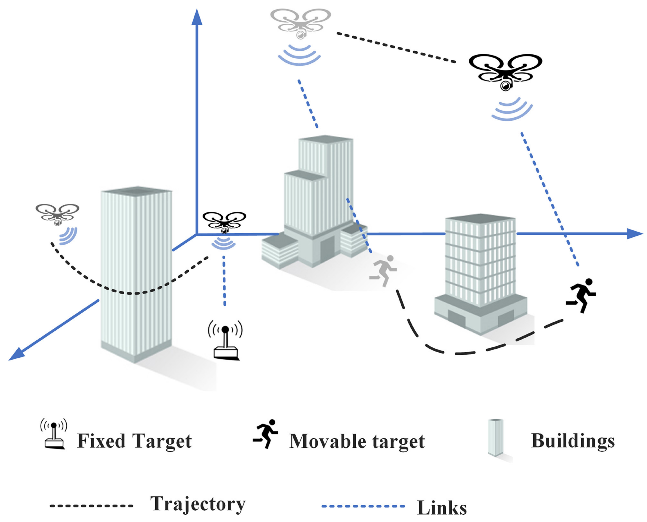


Fig. 6. A scenario of MUWN-assisted Localization.

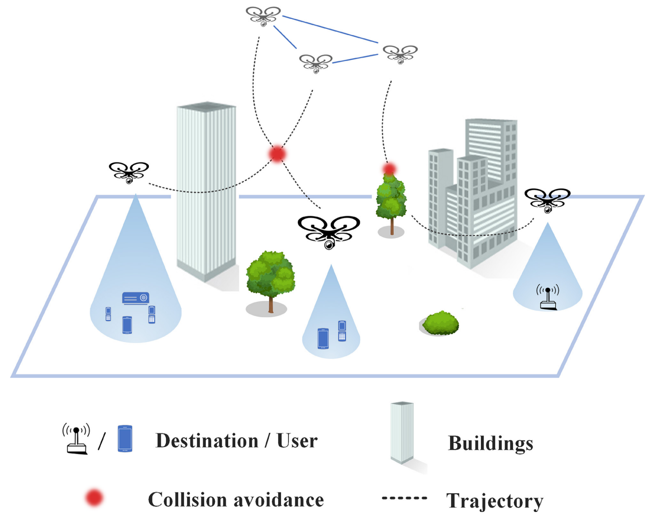


Fig. 7. A scenario of trajectory planning in MUWN.

Moreover, this work measures multiple pairs of RSS values as well as the distance between the UAV and the target under each RSS value. Then uses supervised learning to find the mapping relationship between RSS and distance. Specifically, UAVs are guided to fly to the target position by a multi-agent Q-learning algorithm, which maximizes the RSS value. Gaussian process regression algorithm is used to obtain the mapping relationship between RSS and distance.

A monitoring MUWN is proposed by [154] in which UAVs are used to monitor users in a specific area. The purpose of this study is to optimize the energy consumption and trajectory of UAVs so that they are able to cover a larger area. Specifically, the authors select one of the UAVs as a leader, responsible for communicating with other UAVs and collecting observation information from them. An AC-based model is proposed to implement the optimization objectives. Although the authors set all UAVs as agents, the program actually combines the observed states of all agents and calculates and selects actions collectively, which is similar to a single-agent system.

In [156], multiple UAVs are used to detect and track multiple static objects within a two-dimensional square grid environment. The size of the square grid is $M \times M$, and the observable range of each UAV is a square area of $F \times F$ around itself, while $F < M$. Using distributed DQN training, each UAV acts as an agent that explores the environment, avoids static obstacles, and locates the target. Similarly to [156], multiple UAVs are dispatched to track targets in a fixed area in [157]. Even though the target is only a single object, the environment simulation and UAV motion are more complex. There are six degrees of freedom available to UAVs. The authors adopt a MADDPG-based algorithm, which promotes UAV cooperation more effectively than distributed DQN in [156].

A system in which multiple UAVs can track dynamic targets is described in [158]. The tracked target is also a UAV. This study assumes that UAV swarms are able to acquire the location of target UAVs through sensors. A multi-agent SAC model is used for training the UAV swarm. Additionally, this work considers the power consumption of the UAV during tracking and proposes an energy-saving strategy

In conclusion, this section presents a compilation of literature on MUWN-based target localization applications. To enhance clarity, Table VI is provided, offering a summary of the targets, sensors or sensing information utilized, and the optimization objectives addressed in the discussed literature.

VII. Trajectory Planning

Trajectory planning plays a crucial role not only in the movement of MUWNs but also in achieving their optimization objectives. Fig. 7 illustrates a scenario depicting trajectory planning in MUWNs. Unlike traditional path planning approaches, RL-based trajectory design focuses on defining actions for the agent, such as the direction of movement, speed, or target location of UAVs. By utilizing a reward function, incentives are provided to guide UAVs towards discovering optimal trajectories. To ensure obstacle avoidance for MUWNs, collision penalty items can be incorporated into the reward function. By considering energy and user constraints within the reward function, trajectories with improved efficiency and service quality for MUWNs can be obtained.

This section further categorizes the works on trajectory planning into two parts. The first part focuses on UAV navigation, which includes collision avoidance, flocking behavior, and coverage. The second part explores task performance enhancement via trajectory optimization.

A. Navigation

This section focuses on the direct movement control of UAVs based on RL, with a specific emphasis on addressing topics such as collision avoidance, flocking behavior, and coverage. The main objective of this line of research is to enable effective navigation and coordination among multiple UAVs within the MUWN framework. By leveraging RL algorithms, optimal movement strategies can be developed to enhance the overall performance and efficiency of the UAVs in various scenarios.

TABLE VI
SUMMARY OF WORKS ON MUWN-ASSISTED LOCALIZATION

Issues	Works	Targets	Sensors or Sensing Info	Optimization objective	RL Algorithms
Target tracking	[33]	A fixed-position RF target	RSS sensors	Receiving RSS maximization	Q-learning
Surveillance	[154]	Ground users	Cameras	Cover a large area, overlapping and shadow areas minimization	A2C
Search and rescue	[155]	Multiple moving targets with radio transmitters	Radio receivers	Positioning error minimization	DQN
Target detect and surveillance	[156]	Multiple static objects	Square observable area	Surveillance performance	DQN
UAVs and targets pursuit-evasion	[157]	A moving target	Game environment	Tracking performance	MADDPG
Target tracking	[158]	A single moving UAV	Location of the target	Tracking performance and energy-saving	Multi-agent SAC

The avoidance of obstacles during UAV flight is an important aspect of trajectory planning. Some literature has explored the RL-based obstacle avoidance strategy of single UAV [159], [160], [161]. In MUWNs, mutual avoidance is a challenging research problem. RL can be used to assist UAVs in autonomously learning avoidance strategies and achieving intelligent flight.

The system in [162] consists of multiple UAVs that attempt to avoid each other while flying from their starting points to their destinations. Each UAV is considered an agent in this study, and all agents employ the same avoidance strategy, which is trained using the DDPG algorithm. According to the reward function in this model, the UAV receives positive feedback when it is closer to the target point than it was the previous moment, and negative feedback when it is farther from the target point. A large reward will be awarded for reaching the target point, and a large penalty will be imposed for colliding with other UAVs. As UAVs cannot develop an effective strategy during the early stages of training, the data that a UAV successfully reached the target can only be collected after numerous attempts. In order to facilitate the convergence of the RL algorithm, this work divides the training process into two stages. In the first stage, optimal reciprocal collision avoidance [163] is used for supervised training of the UAV, in order to obtain an avoidance strategy as quickly as possible. To further optimize the policy, RL is employed in the second stage.

In [164], UAVs perform two tasks simultaneously. In the forward path, the UAV first flies from the starting point to the destination, performing the task of delivering the goods. In the backward path, the UAV returns to the starting point and collects data from IoT devices. During the flight, the UAVs are divided into different groups, and the flying heights of different groups of UAVs are different. By grouping, each UAV only needs to consider collisions between UAVs in the same group, which reduces the input dimensions of RL model. Additionally, the backward path considers a no-return traveling salesman problem that UAVs need to traverse multiple IoT devices. Q-learning is employed in this work.

In [165], the authors discuss a UAV flocking system where a UAV with missions and dependent control acts as the leader, while other UAVs, called followers, aim to flock with the leader. The followers in this study are trained as agents using Q-learning. The objective is to minimize the overall cost, which is determined by the distance and heading towards the leader. An integrated system of UAV flocking and obstacle avoidance is proposed in [166]. Similar to [165], the follower is still used as an agent. Followers compress their observations to a fixed length by encoding to cope with the possibly changing length of observations. Multi-agent D3QN is adopted to learn formation control and obstacle avoidance policy. To improve the training efficiency of RL, the authors employed a reference-point-based action selection strategy and an adaptive mechanism. Compared with [165], the authors take into account not only collision avoidance between UAVs, but also considered the avoidance of flocking UAVs from external obstacles in [167]. A system in which a swarm of UAVs travels through a complex of buildings to reach their destination is simulated. Additionally, the authors mention that the use of digital twin technology for highly simulating real-world environments may improve the possibility of using RL in a real-world environment. Collision avoidance is an essential aspect of UAV control in many applications. The issue of collision avoidance in data collection is addressed in [64]. An obstacle avoidance problem in the context of UAV chase and escape is discussed in [157].

In [168], the authors focus on utilizing UAV navigation in a massive multiple-input-multiple-output (MIMO) scenario. The authors employ a CNN to extract features from the system, which are then used for Q-learning training. Each UAV ground link is considered an agent, and the positioning of UAVs is based on the strength of the received signals.

In [169], the authors address the coverage problem of a three-dimensional irregular terrain surface using a UAV network. They project the 3D terrain surface into a series of weighted 2D patches. The UAVs are divided into two groups: low-level follower UAVs and high-level leader UAVs. Q-learning is applied to select patches for leader UAVs, and follower UAVs are dispatched to cover the patches based on a star communication topology.

B. Performance Enhancement via Trajectory Optimization

Trajectory planning plays a critical role in both the movement and optimization objectives of MUWNs. This section focuses on MUWN applications where performance can be enhanced through trajectory planning.

One such application involves the utilization of UAVs as aerial BSs to serve ground users. The inherent high mobility of UAVs enables them to dynamically plan trajectories and adjust their positions in real-time based on the current situation of ground users. This capability significantly improves the overall performance of the network, allowing for better coverage, improved signal quality, and enhanced user experience. A flow-level model based trajectory planning for MUWNs is proposed in [173]. The flow-level model in this paper defines a data flow as multiple Internet protocol packets belonging to an object such as a Web page or a video file. Data flows arrive randomly in time and space and are independent of each other. Due to the mobility of users, the locations of data flows may change over time. For given traffic density of users and pre-deployed UAV locations, a PPO-based model is utilized to plan the optimal UAV trajectory, thereby maximizing network throughput and reducing flow blocking probability. A simulation environment of a portion of downtown San Francisco is adopted. After planning with the PPO algorithm, the UAV-based aerial BSs achieved an approximately three-fold increase in average user throughput.

An issue of maximizing downlink capacity when the UAV-based aerial BSs can cover all terrestrial users is discussed in [34]. UAVs possess a dynamic three-dimensional motion, which is variable in speed, and the mobility of ground users is also taken into account in the literature. Each UAV acts as an independent agent capable of adjusting its real-time 3D position to track moving ground terminals. In this system, the UAV-based aerial BSs aim to provide high-quality wireless services to ground terminals as well as to ensure that all ground terminals are served.

In [174], a network is described where UAVs serve as aerial BSs to provide communication services to mobile users within a fixed area. The users are grouped using a genetic algorithm based on K-means clustering. Q-learning is then employed to train each UAV on deployment and maneuvering strategies. The optimization objective is to maximize the mean opinion score of the ground users. Positive rewards are given to UAVs when their actions improve the overall QoE sum, while negative rewards are assigned when the actions have a negative impact.

In [175], a network is presented where UAVs act as aerial BSs to provide communication services to vehicles on highways with weak cellular infrastructure. The authors utilize a DDPG model to optimize the trajectory of UAVs, aiming to maximize vehicle coverage with the fewest number of UAVs and minimum energy consumption. The reward function is designed to penalize uncovered vehicles, incentivize the deployment of additional UAVs, and encourage the UAVs to have surplus power for traveling to a charging station. Compared with [175], the fairness of ground users is considered in [176] on the basis of energy-saving coverage and throughput maximization. To achieve the above objectives, a MADDPG-based algorithm is used to optimize the UAV trajectory. In addition, the authors consider the issue of mutual collision avoidance among UAVs.

The purpose of [177] is to maximize the data downlink rate of users in a UAV-assisted communication network. A deep echo-state network is used to predict ground user movements. Then, a K-means-based method is employed to cluster users. Furthermore, the authors propose a single A2C algorithm that optimizes the deployment of UAVs. A multi-agent A2C algorithm is used to optimize the 3D trajectory of UAVs and improve the spectral efficiency of ground users. To achieve the above objectives, a CTDE-based multi-agent algorithm is used to optimize the UAV trajectory.

In [178], the authors proposes a description of the relationship between ground users and UAVs based on heterogeneous graphs. UAVs can communicate with each other and exchange graph information, allowing each UAV to obtain more observation information to better achieve cooperation. UAVs as agents, based on their own observation information and communication information with other UAVs, control their real-time position adjustments to maximize fair throughput and improve the QoS.

In [179], the authors explore the deployment of UAV as aerial BSs to offer downlink Internet connectivity following the incapacitation of ground BS due to natural disasters. Given the mobility of ground endpoints, the authors emphasize the need for real-time deployment of UAVs to enhance network performance. The author utilizes an actor-critic deep Q-learning model for decision-making, achieving continuous action control of the UAV. Simulation results indicate better performance in data rates and a decrease in the time required for optimal UAV deployment than deep Q-learning and Q-learning. Furthermore, the authors accentuate the potential of UAV networks in the context of 6G communications, especially in urban microcells. The real-time UAV deployment strategies discussed can offer a dense network of access points to facilitate ultra-reliable and low-latency communication.

Energy efficiency is another important performance metric in the MUWNs due to the energy limitation of UAVs. Effective trajectory planning can help to reduce the energy consumption of UAVs without compromising task completion. In [35], UAVs collect data from multiple IoT points and transmit it to a BS capable of multi-access edge computing via a backhaul communication link. The trajectory is planned using a DQN-based model that employs experience replay. The actions of agents are designated as target points for UAVs, which are numbered data collection points. Therefore, the trajectories of UAVs are the position transfer between different data nodes. Energy efficiency is maximized by adding an energy efficiency item to the reward function. The authors of [170] employ a MUWN to provide content coverage for users. In this environment, UAVs are able to go to charging piles to be charged. The authors simplify the model by dividing the target area into multiple grids, and the UAV moves from one grid center to another grid center. A decentralized multi-agent Q-learning model is proposed to maximize energy efficiency. In this model, each UAV is considered as two agents, one to control charging, called the energy learner, and the other to control task positioning called the cruise learner. In [172], energy efficiency and obstacle avoidance of UAVs are considered. Each UAV performs an assigned mission from a preset starting point to a set destination, while avoiding disturbances and obstacles in real-time by acquiring environmental information. TD3 is

TABLE VII
SUMMARY OF WORKS ON TRAJECTORY PLANNING FOR MUWNS

Issues	Works	Main Findings	Optimization Objectives	RL Algorithms
Navigation	[162]	Demonstrates a two-stage RL method for multi-UAV collision avoidance that plans a trajectory using noisy observations	Collision avoidance	DDPG
	[164]	Uses RL for collision avoidance and optimal trajectory planning in UAV networks, handling both object delivery and IoT data collection	Collision avoidance	Q-learning
	[166]	Develops a DRL framework to train UAVs in collision-free flocking, adapting to the number of followers and changing environment state	Flocking, collision avoidance	D3QN
	[167]	Proposes a Digital Twin-enabled DRL training framework to train multi-UAV systems for flocking motions, improving arrival rate and collision rate performance	Collision avoidance	DDPG
	[165]	Applies Q-Learning to manage flocking in a leader-follower topology among UAVs in a stochastic environment	Flocking	Q-learning
	[168]	Combines DRL with massive MIMO to optimize UAV navigation based on RSS	Navigation	Q-learning
	[169]	Presents a two-level hierarchical UAV swarm architecture to solve 3D irregular terrain surface coverage problems	Coverage	Q-learning
Performance Enhancement via Trajectory Optimization	[35]	Proposes a DRL model for UAV navigation in IoT networks	Energy efficiency, data freshness maximization	DQN
	[170]	Introduces a decentralized DRL for trajectory planning in multi-UAV scenarios	Energy efficiency maximization	Q-learning
	[171]	Presents a pointer network-A* DRL technique for UAV trajectory planning in wireless sensor networks	Energy efficiency maximization	A2C
	[172]	Modifies the TD3 model for online path planning of multiple UAVs	Energy efficiency maximization	TD3
	[173]	Uses a DRL approach with flow-level models to determine optimal UAVs' trajectories	Network throughput maximization, flow blocking probability minimization	PPO
	[34]	Presents a constrained DQN for optimizing the 3D trajectory design of multiple UAVs in a wireless network	Downlink capacity maximization	DQN
	[174]	Using Q-learning based method to optimize the deployment and dynamic movement of UAVs	Mean opinion score of ground users maximization	Q-learning
	[175]	Introduces a UAVs cell-free network utilizing DRL to optimize UAV trajectories in limited energy resources and insufficient environment knowledge	Vehicle coverage maximization, energy consumption minimization	DDPG
	[176]	Presents a MADDPG-absed method for optimizing the trajectory of UAVs in a dynamic communication system	Fairness of ground users, energy efficiency, throughput maximization	MADDPG
	[177]	Utilizes single-/multi-agent AC algorithms for initial UAV deployment and trajectory design	Data downlink rate of users maximization	A2C
	[178]	Using heterogeneous graphs to represent the relationship between UAVs and users, and enhance the cooperation of multi-agent through communication	Fair throughput maximization	Multi-agent DQN
	[179]	Leverage continuous AC deep Q-learning for optimal real-time UAV deployment	Sum data rate maximization	AC

used for path planning in this study. The UAVs employ the same strategy and do not cooperate with each other.

In conclusion, this section examines the role of RL in the movement of MUWNs. Table VII provides a summary of the research conducted on trajectory planning in RL-enabled MUWNs. It outlines the main findings, optimization objectives, and RL algorithms employed in these studies.

VIII. NETWORK SECURITY

MUWNs are susceptible to attacks due to their openness and multi-connectivity [9], [180], [181]. The presence of attackers or eavesdroppers in the environment poses a threat to data transmission and communication between UAVs and ground nodes. Fig. 8 provides an illustration of a network security scenario, depicting the presence of jammers and eavesdroppers. In this application, the action space of the RL agents typically includes the position or signal transmission power of UAVs. The reward function is designed to measure the corresponding security metric specific to the given environment. By employing RL, the location and power settings of the MUWN can be optimized to enhance security and mitigate eavesdropping or interference risks. This section presents a review of the existing literature on RL in the context of MUWN security.

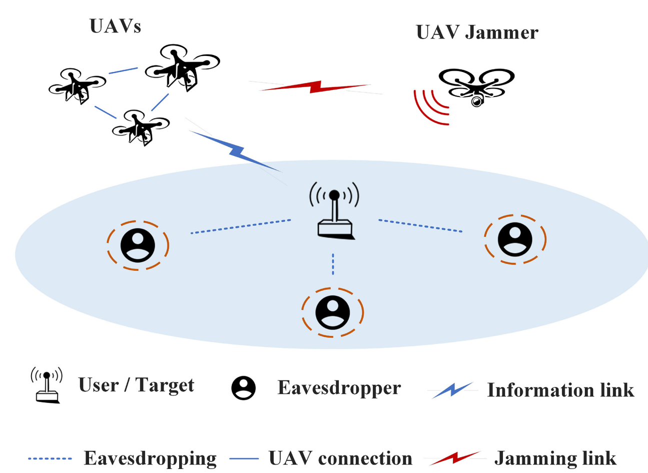


Fig. 8. A scenario of network security issues in MUWNs.

In [182], the authors introduce a framework for intrusion detection in MUWNs utilizing RL algorithms. The authors summarized three types of intrusion attacks that MUWNs might encounter, namely jamming attacks, impersonation attacks, and intrusion attacks. In jamming attacks, malicious users or jamming UAVs interfere with the target MUWN by

TABLE VIII
SUMMARY OF WORKS ON NETWORK SECURITY OF MUWNS

Issue	Works	Main Findings	Security risk sources	Optimization goals	RL algorithm
Network Security	[182]	Introduces a DRL approach for detecting malicious attacks in UAV aerial computing networks to enhance security services	Malware	Intrusion detection	DDPG
	[183]	Proposes a cooperative jamming approach utilizing MADDPG to enhance secure communications in UAV networks and defend against ground eavesdroppers	Ground eavesdroppers	Summed security rate of ground users maximization	MADDPG
	[184]	Presents a knowledge-based RL method to improve the convergence speed and performance of UAV networks against smart jammers with adaptive strategies	UAV jammers	Received SINR maximization	Q-learning
	[185]	Introduces a secure federated learning framework for UAV-assisted mobile crowdsensing, utilizing blockchain, privacy-preserving algorithms, and RL-based incentives to ensure privacy	Exchange of models	Pricing strategies of blockchain	Q-learning
	[186]	Proposes a UAV-enabled secure communication system, proposing a MADDPG-based algorithm to optimize UAV trajectory, power control, and node scheduling	UAV eavesdroppers	Total security rate maximization	MADDPG

sending false signals or noise. This prevents the MUWN from receiving information or obtaining inaccurate information. In impersonation attacks, the attacker pretends to be a UAV in the MUWN but provides incorrect services or maliciously steals data. In intrusion attacks, malicious programs are directly uploaded into the target MUWN. To demonstrate how RL can be used to detect intrusions, a simple DDPG-based case study is presented.

In [183], UAVs serve as aerial BSs for ground users. LoS is dominant in the aerial-to-ground channel link, facilitating easy wiretaps by ground eavesdroppers. In order to address this issue, the authors divide the UAVs into two groups. One group of UAVs acts as aerial BSs to send information to users. Another group of UAVs is jammers that transmit artificial noises to ground eavesdroppers. This system assumes that all UAVs are aware of the locations of ground users and ground eavesdroppers. Two groups of UAVs are trained using MADDPG to adjust their positions and powers in order to maximize the summed security rate of all ground users. The security rate for a ground user is the signal-to-noise ratio it receives minus the maximum signal-to-noise ratio that all ground eavesdroppers can accept. The authors also consider the case where the number of UAV jammers is less than that of ground eavesdroppers. UAV jammers are capable of dynamically adjusting their positions in order to jam all ground eavesdroppers.

In [184], a UAV swarm is dispatched to receive signals from a ground target. There are some UAV jammers near the target trying to jam the UAV swarm to reduce its received Signal-to-Interference plus Noise Ratio (SINR). In contrast to [183], [184] does not know the channel state or location of the jammers. The UAV swarm is viewed as a whole, and also as an agent trained by RL algorithm, while jammers adjust their transmit powers and positions depending on predefined strategies. The UAV swarm only uses RSS and SINR to make decisions. As the authors point out, although the jammer's location is unknown, they are able to predict the jammer's trajectory based on the knowledge that the RSS value changes with distance and the knowledge of the jammer's flight inertia. A Q-learning-based model is trained by using RSS, SINR and the predicted trajectories of jammers, which the authors call knowledge-based RL.

[186] present a scenario where a ground node sends confidential information to a legitimate UAV in the presence of a smart UAV eavesdropper. The eavesdropper can adjust its position for more effective eavesdropping. Similar to [184], the legitimate UAV is unable to obtain the trajectory of the eavesdropper. In particular, both the legitimate UAV and the eavesdropper are treated as agents in this study. They have completely opposing goals, with the legitimate UAV aiming to maximize the total security rate, and the eavesdropper looking to minimize it. The definition of security rate is the same as [183]. The authors reformulate the problem as a two-player zero-sum stochastic game problem. A MADDPG-based algorithm is used to train the two agents. Results indicate that the legitimate UAV selects a communication link away from the eavesdropper, whereas the eavesdropper does its best to listen in on confidential information.

A secure federated learning framework for UAV-assisted MCS is presented in [185]. The article uses blockchain technology to secure the exchange of local model updates and to verify the contributions. On the updated local model, a privacy-preserving algorithm is used to ensure that the privacy of the UAV is preserved. For participating UAVs and mission publishers, the exact parameters of the cost model and network model may not be available during federated learning. Utilize RL methods such as Q-learning when determining pricing strategies for UAVs and mission issuers when inaccurate system parameters are not available.

In conclusion, this section provides an overview of the security challenges faced by MUWNs and emphasizes the role of RL in addressing these problems. Table VIII presents specific details of the studies discussed in this section.

IX. OPEN CHALLENGES

In light of an analysis of the current work about RL in MUWNs in the above sections, we outline the following open challenges towards autonomous and intelligent MUWNs.

A. Multi-Objective Optimization

Multi-objective optimization remains a key and persisting research challenge within RL-based MUWNs. Present literature typically employs RL to solve multi-objective optimization by incorporating different optimization objectives as reward or penalty terms in the reward function. This process effectively facilitates the task of modeling complex multi-objective optimization problems, by endeavoring to maximize long-term rewards while concurrently optimizing each constituent of the reward function through iterative interactions. Despite the favorable outcomes demonstrated by this approach, several challenges persist. Firstly, there is no certainty that the structured reward function can accurately steer the RL model's training trajectory towards the desired direction. Secondly, determining the appropriate weight for each reward or penalty item, in order to maintain a balance among multiple objectives, remains problematic. Optimization objectives in MUWNs, such as energy efficiency and data throughput, are characterized by significant numerical variances. Thus, designing a model capable of effectively constraining and balancing numerous optimization goals is still a challenging task. Drawing upon the lessons learned from these challenges, future research should focus on devising more reliable methods for reward function design, in order to better guide the RL model's training. Additionally, methods that can effectively handle the numerical differences in optimization objectives and achieve a balanced optimization in MUWNs warrant further exploration and study. These learnings could potentially refine the way multi-objective optimization is handled within the realm of RL-based MUWNs.

B. Resource Allocation and Cooperation

Resource allocation and cooperation are fundamental elements of MUWNs, and their optimization remains an open problem. While Section IV reviews various works on resource allocation within RL-enabled MUWNs, such as spectrum, channel, energy, and cache allocations, issues persist in terms of UAV cooperation and resource allocation. Inefficient task distribution among UAVs can lead to unnecessary resource waste and performance deterioration. Some research employs SARL for resource allocation, which utilizes comprehensive information from all UAVs and users. However, given communication limitations, individual UAVs acting as agents are constrained in their decision-making capacity, as they lack a holistic view of the entire system. While MARL has been extensively explored in the context of MUWNs, it has not fully addressed the challenge UAV agents faced. These challenges underline the necessity of improving both resource allocation strategies and inter-UAV cooperation methods, with a particular emphasis on enhancing individual UAV's decision-making capabilities in the face of limited information. Such improvements are essential for fully realizing the potential of MARL in real-world MUWN applications.

C. Joint Trajectory Planning

Joint trajectory optimization for multi-UAVs constitutes a significant area requiring further exploration. As referenced in Section VII, trajectory planning is not just critical for the movement of MUWNs, but also instrumental in realizing their optimization goals. Existing research delves into UAV navigation via RL [168], and targeted optimization through trajectory design, such as maximizing data collection [60] and circumventing eavesdropping [186]. Nevertheless, the joint trajectory planning of multiple UAVs continues to pose challenges. Firstly, prevalent approaches often introduce strict constraints on UAV movements in the process of trajectory-based performance optimization. These may involve fixed altitude, speed, or designated task zones for UAVs to reduce design complexity. This highlights the need for more flexible trajectory planning methods. Secondly, joint trajectory planning necessitates consideration of the interplay between UAVs, underscoring the requirement for more efficient MARL algorithms. Therefore, the two primary lessons are the need for greater flexibility in trajectory planning, and the call for more advanced MARL algorithms to manage inter-UAV interactions effectively in joint trajectory planning.

D. Distributed DRL Framework

Developing a distributed MARL framework poses a significant challenge. The current state-of-the-art in MARL is the CTDE framework, wherein each agent can access the observations and actions of all others during training. A more fitting approach to intelligent MUWN development could be through distributed training among UAVs. Given that UAVs can communicate amongst themselves, they have the potential to leverage local information along with the data from other communicable UAVs to make decisions. It is worth noting that the exploration of multi-agent communication is an essential research focus for MARL algorithms. Therefore, the task of designing an efficient distributed DRL framework becomes both a daunting and rewarding endeavor. This task requires striking a balance between the practicality of distributed information access among UAVs and the technical requirements of maintaining effective learning algorithms.

E. Model Training and Implementation

There exists a notable gap between the existing literature on RL-enabled MUWNs and its real-world implementation. Most existing studies in this field primarily rely on simulated environments for their applications. RL models require an iterative process of interaction with the environment, often involving multiple failures before achieving a satisfactory model. Hence, it is crucial to devise strategies for training and implementing RL models in practical scenarios. Several approaches have been proposed to address this issue. Some methods utilize high-fidelity simulators to train agents and then employ transfer learning techniques to adapt the algorithm to real-world environments [187], [188]. Others incorporate pre-defined strategies or trajectories to limit unnecessary exploration and minimize errors during the learning process [189]. These efforts have provided valuable insights and lessons. Furthermore, they highlight the significance of balancing exploration and exploitation to ensure safe and efficient learning. However, the translation of RL techniques to real-world applications continues to pose significant challenges and presents exciting research opportunities.

F. Security and Privacy Issues

Security and privacy issues in MUWNs are rarely considered in current research. Due to their openness and multi-connections, MUWNs are vulnerable to unexpected attacks. While works in Section VII aim to prevent eavesdropping by adjusting the position and power of UAVs, there is a lack of research on MUWNs for the detection of security vulnerabilities, interference prevention, and intrusion in complex environments in real-life situations. While RL has been employed as an optimization technique in MUWNs, its potential in enhancing the security of UAVs remains an intriguing yet largely uncharted domain.

X. Conclusion

This paper presents a comprehensive survey on RL-based MUWNs. Firstly, we provide an introduction to the background of RL-enabled autonomous MUWNs. Then, we offer a tutorial on RL and review the recent advancements in RL algorithms. We also summarize the process of applying RL algorithms in MUWNs. Our analysis of RL applications in MUWNs is organized into six sections, covering data access, sensing, and collection; resource allocation for wireless connectivity; UAV-assisted MEC; localization; trajectory planning; and network security. Finally, open research directions are highlighted to shed light on the autonomous operation of multi-UAV network via RL approaches.

References

[1] S. Hayat, E. Yanmaz, and R. Muzaffar, "Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2624–2661, 4th Quart., 2016.

[2] D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis, "A review on UAV-based applications for precision agriculture," *Information*, vol. 10, no. 11, p. 349, 2019.

[3] M. Silvagni, A. Tonoli, E. Zenerino, and M. Chiaberge, "Multipurpose UAV for search and rescue operations in mountain avalanche events," *Geomatic. Nat. Hazards Risk*, vol. 8, no. 1, pp. 18–33, 2017.

[4] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of UAV-based target tracking and recognition systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3409–3422, Aug. 2020.

[5] E. Alvarado. "Drone market analysis 2022–2030." Sep. 2022. [Online]. Available: https://droneii.com/drone-market-analysis-2022-2030

[6] R. Shakeri et al., "Design challenges of multi-UAV systems in cyber-physical applications: A comprehensive survey and future directions," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3340–3385, 4th Quart., 2019.

[7] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2016.

[8] N. H. Motlagh, T. Taleb, and O. Arouk, "Low-altitude unmanned aerial vehicles-based Internet of Things services: Comprehensive survey and future perspectives," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 899–922, Dec. 2016.

[9] A. Fotouhi et al., "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart., 2019.

[10] M. Abrar, U. Ajmal, Z. M. Almohaimeed, X. Gui, R. Akram, and R. Masroor, "Energy efficient UAV-enabled mobile edge computing for IoT devices: A review," *IEEE Access*, vol. 9, pp. 127779–127798, 2021.

[11] G. Geraci et al., "What will the future of UAV cellular communications be? a flight from 5G to 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1304–1335, 3rd Quart., 2022.

[12] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.

[14] B. Kiumarsi, K. G. Vamvoudakis, H. Modares, and F. L. Lewis, "Optimal and autonomous control using reinforcement learning: A survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2042–2062, Jun. 2018.

[15] L. Lei, Y. Tan, K. Zheng, S. Liu, K. Zhang, and X. Shen, "Deep reinforcement learning for autonomous Internet of Things: Model, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1722–1760, 3rd Quart., 2020.

[16] Y. Xiao, J. Liu, J. Wu, and N. Ansari, "Leveraging deep reinforcement learning for traffic engineering: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2064–2097, 4th Quart., 2021.

[17] N. Parvaresh, M. Kulhandjian, H. Kulhandjian, C. D'Amours, and B. Kantarci, "A tutorial on AI-powered 3D deployment of drone base stations: State of the art, applications and challenges," *Veh. Commun.*, vol. 36, Aug. 2022, Art. no. 100474.

[18] Z. Cheng, M. Liwang, N. Chen, L. Huang, X. Du, and M. Guizani, "Deep reinforcement learning-based joint task and energy offloading in UAV-aided 6G intelligent edge networks," *Comput. Commun.*, vol. 192, pp. 234–244, Aug. 2022.

[19] Y. Li, "Deep reinforcement learning: An overview," 2017, *arXiv:1701.07274*.

[20] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart., 2019.

[21] M. E. Mkiramweni, C. Yang, J. Li, and W. Zhang, "A survey of game theory in unmanned aerial vehicles communications," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3386–3416, 4th Quart., 2019.

[22] N. C. Luong et al., "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.

[23] V. Hassija et al., "Fast, reliable, and secure drone communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2802–2832, 4th Quart., 2021.

[24] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 2nd Quart., 2021.

[25] A. Baltaci, E. Dinc, M. Ozger, A. Alabbasi, C. Cavdar, and D. Schupke, "A survey of wireless networks for future aerial communications (FACOM)," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2833–2884, 4th Quart., 2021.

[26] T. Li et al., "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.

[27] 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.

[28] A. Puente-Castro, D. Rivero, A. Pazos, and E. Fernandez-Blanco, "A review of artificial intelligence applied to path planning in UAV swarms," *Neural Comput. Appl.*, vol. 34, pp. 153–170, Jan. 2022.

[29] D. Zhou, M. Sheng, J. Li, and Z. Han, "Aerospace integrated networks innovation for empowering 6G: A survey and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 975–1019, 2nd Quart., 2023.

[30] Y. Liu, J. Yan, and X. Zhao, "Deep-reinforcement-learning-based optimal transmission policies for opportunistic UAVs-aided wireless sensor network," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13823–13836, Aug. 2022.

[31] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[32] S. A. Huda and S. Moh, "Survey on computation offloading in UAV-enabled mobile edge computing," *J. Netw. Comput. Appl.*, vol. 201, May 2022, Art. no. 103341.

[33] Y.-J. Chen, D.-K. Chang, and C. Zhang, “Autonomous tracking using a swarm of UAVs: A constrained multi-agent reinforcement learning approach,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13702–13717, Nov. 2020.
[34] W. Zhang, Q. Wang, X. Liu, Y. Liu, and Y. Chen, “Three-dimension trajectory design for multi-UAV wireless network with deep reinforcement learning,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 600–612, Jan. 2021.
[35] S. F. Abedin, M. 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.
[36] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 2014.
[37] G. E. Monahan, “State of the art—A survey of partially observable Markov decision processes: Theory, models, and algorithms,” *Manage. Sci.*, vol. 28, no. 1, pp. 1–16, 1982.
[38] C. J. Watkins and P. Dayan, “Q-Learning,” *Mach. Learn.*, vol. 8, no. 3, pp. 279–292, 1992.
[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[40] H. Van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-learning,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 30, 2016, pp. 2094–2100.
[41] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized experience replay,” 2015, *arXiv:1511.05952*.
[42] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, “Dueling network architectures for deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1995–2003.
[43] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 12, 1999, pp. 1057–1063.
[44] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Mach. Learn.*, vol. 8, no. 3, pp. 229–256, 1992.
[45] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.
[46] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017, *arXiv:1707.06347*.
[47] V. Mnih et al., “Asynchronous methods for deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.
[48] T. P. Lillicrap et al., “Continuous control with deep reinforcement learning,” 2015, *arXiv:1509.02971*.
[49] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.
[50] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1587–1596.
[51] 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.
[52] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *Proc. Mach. Learn.*, 1994, pp. 157–163.
[53] W. Du and S. Ding, “A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications,” *Artif. Intell. Rev.*, vol. 54, no. 5, pp. 3215–3238, 2021.
[54] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, “A survey of learning in multiagent environments: Dealing with non-stationarity,” 2017, *arXiv:1707.09183*.
[55] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6379–6390.
[56] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 2974–2982.
[57] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*. Cham, Switzerland: Springer, 2016.
[58] P. Sunehag et al., “Value-decomposition networks for cooperative multi-agent learning,” 2017, *arXiv:1706.05296*.
[59] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “Monotonic value function factorisation for deep multi-agent reinforcement learning,” *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 7234–7284, 2020.
[60] C. H. Liu, Z. Chen, and Y. Zhan, “Energy-efficient distributed mobile crowd sensing: A deep learning approach,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1262–1276, Jun. 2019.
[61] N. Aloysius and M. Geetha, “A review on deep convolutional neural networks,” in *Proc. Int. Conf. Commun. Signal Process. (ICCSP)*, 2017, pp. 588–592.
[62] Z. Dai, C. H. Liu, R. Han, G. Wang, K. Leung, and J. Tang, “Delay-sensitive energy-efficient UAV crowdsensing by deep reinforcement learning,” *IEEE Trans. Mobile Comput.*, vol. 22, no. 4, pp. 2038–2052, Apr. 2023.
[63] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, *arXiv:1412.3555*.
[64] X. Wang, M. C. Gursoy, T. Erpek, and Y. E. Sagduyu, “Learning-based UAV path planning for data collection with integrated collision avoidance,” *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16663–16676, Sep. 2022.
[65] 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.
[66] Y. Zhang, Z. Mou, F. Gao, L. Xing, J. Jiang, and Z. Han, “Hierarchical deep reinforcement learning for backscattering data collection with multiple UAVs,” *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3786–3800, Mar. 2021.
[67] S. Xu, X. Zhang, C. Li, D. Wang, and L. Yang, “Deep reinforcement learning approach for joint trajectory design in multi-UAV IoT networks,” *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3389–3394, Mar. 2022.
[68] K. Wei et al., “High-performance UAV crowdsensing: A deep reinforcement learning approach,” *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18487–18499, Oct. 2022.
[69] T. Li, W. Liu, Z. Zeng, and N. Xiong, “DRLR: A deep-reinforcement-learning-based recruitment scheme for massive data collections in 6G-based IoT networks,” *IEEE Internet Things J.*, vol. 9, no. 16, pp. 14595–14609, Aug. 2022.
[70] 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.
[71] 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.
[72] 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.
[73] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in *Proc. IEEE INFOCOM*, 2012, pp. 2731–2735.
[74] R. Ding, F. Gao, and X. S. Shen, “3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796–7809, Dec. 2020.
[75] P. Luong, F. Gagnon, L.-N. Tran, and F. Labeau, “Deep reinforcement learning-based resource allocation in cooperative UAV-assisted wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7610–7625, Nov. 2021.
[76] S. Yin and F. R. Yu, “Resource allocation and trajectory design in UAV-aided cellular networks based on multiagent reinforcement learning,” *IEEE Internet Things J.*, vol. 9, no. 4, pp. 2933–2943, Feb. 2022.
[77] T. Yuan, C. E. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, “Harnessing UAVs for fair 5G bandwidth allocation in vehicular communication via deep reinforcement learning,” *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4063–4074, Dec. 2021.
[78] D. S. Lakew, A.-T. Tran, N.-N. Dao, and S. Cho, “Intelligent offloading and resource allocation in heterogeneous aerial access IoT networks,” *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5704–5718, Apr. 2023.
[79] R. Zhong, X. Liu, Y. Liu, and Y. Chen, “Multi-agent reinforcement learning in NOMA-aided UAV networks for cellular offloading,” *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1498–1512, Mar. 2022.
[80] S. Lee, H. Yu, and H. Lee, “Multiagent Q-learning-based multi-UAV wireless networks for maximizing energy efficiency: Deployment and power control strategy design,” *IEEE Internet Things J.*, vol. 9, no. 9, pp. 6434–6442, May 2022.

[81] W. Jiang, W. Yu, W. Wang, and T. Huang, "Multi-agent reinforcement learning for joint cooperative spectrum sensing and channel access in cognitive UAV networks," *Sensors*, vol. 22, no. 4, p. 1651, 2022.

[82] X. Li, S. Cheng, H. Ding, M. Pan, and N. Zhao, "When UAVs meet cognitive radio: Offloading traffic under uncertain spectrum environment via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 824–838, Feb. 2023.

[83] S. Guo and X. Zhao, "Multi-agent deep reinforcement learning based transmission latency minimization for delay-sensitive cognitive satellite-UAV networks," *IEEE Trans. Commun.*, vol. 71, no. 1, pp. 131–144, Jan. 2023.

[84] J. Wang, Y. Liu, S. Niu, and H. Song, "Integration of software defined radios and software defined networking towards reinforcement learning enabled unmanned aerial vehicle networks," in *Proc. IEEE Int. Conf. Ind. Internet (ICII)*, 2019, pp. 44–49.

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

[86] 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.

[87] 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)*, 2022, pp. 884–889.

[88] J. Zhang et al., "Trajectory planning of UAV in wireless powered IoT system based on deep reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2020, pp. 645–650.

[89] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep Q-network for UAV-assisted online power transfer and data collection," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12215–12226, Dec. 2019.

[90] K. K. Nguyen, A. Masaracchia, V. Sharma, H. V. Poor, and T. Q. Duong, "RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1086–1096, Aug. 2022.

[91] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Joint resource, deployment, and caching optimization for AR applications in dynamic UAV NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3409–3422, May 2022.

[92] S. Anokye, D. Ayepah-Mensah, A. M. Seid, G. O. Boateng, and G. Sun, "Deep reinforcement learning-based mobility-aware UAV content caching and placement in mobile edge networks," *IEEE Syst. J.*, vol. 16, no. 1, pp. 275–286, Mar. 2022.

[93] J. Ji, K. Zhu, and L. Cai, "Trajectory and communication design for cache- enabled UAVs in cellular networks: A deep reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 22, no. 10, pp. 6190–6204, Oct. 2023.

[94] A. Al-Hilo, M. Samir, C. Assi, S. Sharafeddine, and D. Ebrahimi, "UAV-assisted content delivery in intelligent transportation systems-joint trajectory planning and cache management," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 8, pp. 5155–5167, Aug. 2021.

[95] D. Wang, Q. Liu, J. Tian, Y. Zhi, J. Qiao, and J. Bian, "Deep reinforcement learning for caching in D2D-enabled UAV-relaying networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, 2021, pp. 635–640.

[96] Y.-J. Chen, K.-M. Liao, M.-L. Ku, F. P. Tso, and G.-Y. Chen, "Multi-agent reinforcement learning based 3D trajectory design in aerial-terrestrial wireless caching networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8201–8215, Aug. 2021.

[97] H. Wu, X. Tao, N. Zhang, and X. Shen, "Cooperative UAV cluster-assisted terrestrial cellular networks for ubiquitous coverage," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2045–2058, Sep. 2018.

[98] C. Dong et al., "UAVs as an intelligent service: Boosting edge intelligence for air-ground integrated networks," *IEEE Netw.*, vol. 35, no. 4, pp. 167–175, Jul./Aug. 2021.

[99] Z. Sheng, H. D. Tuan, T. Q. Duong, and L. Hanzo, "UAV-aided two-way multi-user relaying," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 246–260, Jan. 2021.

[100] T. Fang, H. Tian, X. Zhang, X. Chen, X. Shao, and Y. Zhang, "Context-aware caching distribution and UAV deployment: A game-theoretic approach," *Appl. Sci.*, vol. 8, no. 10, p. 1959, 2018.

[101] Y. Li, H. Zhang, and K. Long, "Joint resource, trajectory, and artificial noise optimization in secure driven 3-D UAVs with NOMA and imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 11, pp. 3363–3377, Nov. 2021.

[102] Y. Shi, M. Q. Hamdan, E. Alsusa, K. A. Hamdi, and M. W. Baidas, "A decoupled access scheme with reinforcement learning power control for cellular-enabled UAVs," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17261–17274, Dec. 2021.

[103] M. Samir, S. Sharafeddine, C. M. Assi, T. M. Nguyen, and A. Ghrayeb, "UAV trajectory planning for data collection from time-constrained IoT devices," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 34–46, Jan. 2020.

[104] T. M. Nguyen, W. Ajib, and C. Assi, "A novel cooperative NOMA for designing UAV-assisted wireless backhaul networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2497–2507, Nov. 2018.

[105] Z. Jia, Q. Wu, C. Dong, C. Yuen, and Z. Han, "Hierarchical aerial computing for Internet of Things via cooperation of HAPs and UAVs," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5676–5688, Apr. 2023.

[106] H. Ahmadinejad and A. Falahati, "Forming a two-tier heterogeneous air-network via combination of high and low altitude platforms," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 1989–2001, Feb. 2022.

[107] F. Jiang, K. Wang, L. Dong, C. Pan, W. Xu, and K. Yang, "AI driven heterogeneous MEC system with UAV assistance for dynamic environment: Challenges and solutions," *IEEE Netw.*, vol. 35, no. 1, pp. 400–408, Jan./Feb. 2021.

[108] Y. Nie, J. Zhao, F. Gao, and F. R. Yu, "Semi-distributed resource management in UAV-aided MEC systems: A multi-agent federated reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13162–13173, Dec. 2021.

[109] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[110] Z. Mlika and S. Cherkaoui, "Network slicing with MEC and deep reinforcement learning for the Internet of Vehicles," *IEEE Netw.*, vol. 35, no. 3, pp. 132–138, May/Jun. 2021.

[111] R. Liu, A. Liu, Z. Qu, and N. N. Xiong, "An UAV-enabled intelligent connected transportation system with 6G communications for Internet of Vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 2, pp. 2045–2059, Feb. 2023.

[112] Z. Ning, J. Huang, X. Wang, J. J. P. C. Rodrigues, and L. Guo, "Mobile edge computing-enabled Internet of Vehicles: Toward energy-efficient scheduling," *IEEE Netw.*, vol. 33, no. 5, pp. 198–205, Sep./Oct. 2019.

[113] K. K. Nguyen, S. R. Khosravirad, D. B. da Costa, L. D. Nguyen, and T. Q. Duong, "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 3, pp. 358–368, Apr. 2022.

[114] L. Yang, Y. Zeng, and R. Zhang, "Wireless power transfer with hybrid beamforming: How many RF chains do we need?" *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6972–6984, Oct. 2018.

[115] B. Clerckx, R. Zhang, R. Schober, D. W. Kwan Ng, D. I. Kim, and H. V. Poor, "Fundamentals of wireless information and power transfer: From RF energy harvester models to signal and system designs," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 4–33, Jan. 2019.

[116] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.

[117] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[118] S. Ku, S. Jung, and C. Lee, "UAV trajectory design based on reinforcement learning for wireless power transfer," in *Proc. 34th Int. Tech. Conf. Circuits/Syst. Comput. Commun. (ITC-CSCC)*, 2019, pp. 1–3.

[119] Y. Liu, K. Xiong, Y. Lu, Q. Ni, P. Fan, and K. B. Letaief, "UAV-aided wireless power transfer and data collection in Rician fading," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3097–3113, Oct. 2021.

[120] O. S. Oubbati, A. Lakas, and M. Guizani, "Multiagent deep reinforcement learning for wireless-powered UAV networks," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16044–16059, Sep. 2022.

[121] K. Zhu et al., "Aerial refueling: Scheduling wireless energy charging for UAV enabled data collection," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 3, pp. 1494–1510, Sep. 2022.

[122] J. Xu, K. Zhu, and R. Wang, "RF aerially charging scheduling for UAV fleet: A Q-learning approach," in *Proc. 15th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, 2019, pp. 194–199.

[123] S. Fu et al., "Energy-efficient UAV-enabled data collection via wireless charging: A reinforcement learning approach," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 10209–10219, Jun. 2021.

[124] A. Merabet, A. Lakas, and A. N. Belkacem, "WPT-enabled UAV trajectory design for Healthcare delivery using reinforcement learning," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2022, pp. 271–277.

[125] Z. Xiong et al., "UAV-assisted wireless energy and data transfer with deep reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 85–99, Mar. 2021.

[126] C. Forecast, "Cisco visual networking index: Global mobile data traffic forecast update, pp. 2017–2022 White Paper," Cisco Public Inf., San Jose, CA, USA, 2019.

[127] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5G small cells: A radio resource management perspective," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 41–49, Oct. 2015.

[128] J. Yang, C. Ma, B. Jiang, G. Ding, G. Zheng, and H. Wang, "Joint optimization in cached-enabled heterogeneous network for efficient industrial IoT," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 5, pp. 831–844, May 2020.

[129] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.

[130] X. Xu, Y. Zeng, Y. L. Guan, and R. Zhang, "Overcoming endurance issue: UAV-enabled communications with proactive caching," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1231–1244, Jun. 2018.

[131] A. Asheralieva and D. Niyato, "Game theory and Lyapunov optimization for cloud-based content delivery networks with device-to-device and UAV-enabled caching," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 10094–10110, Oct. 2019.

[132] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Caching placement and resource allocation for cache-enabling UAV NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12897–12911, Nov. 2020.

[133] S. Zhu, L. Gui, D. Zhao, N. Cheng, Q. Zhang, and X. Lang, "Learning-based computation offloading approaches in UAVs-assisted edge computing," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 928–944, Jan. 2021.

[134] A. Gao, Q. Wang, W. Liang, and Z. Ding, "Game combined multi-agent reinforcement learning approach for UAV assisted offloading," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12888–12901, Dec. 2021.

[135] F. Tang, H. Hofner, N. Kato, K. Kaneko, Y. Yamashita, and M. Hangai, "A deep reinforcement learning-based dynamic traffic offloading in space-air-ground integrated networks (SAGIN)," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 276–289, Jan. 2022.

[136] Y. Liu, J. Yan, and X. Zhao, "Deep reinforcement learning based latency minimization for mobile edge computing with Virtualization in maritime UAV communication network," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4225–4236, Apr. 2022.

[137] A. Sacco, F. Esposito, G. Marchetto, and P. Montuschi, "A self-learning strategy for task offloading in UAV networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4301–4311, Apr. 2022.

[138] S. Islam, S. Badsha, I. Khalil, M. Atiquzzaman, and C. Konstantinou, "A triggerless backdoor attack and defense mechanism for intelligent task offloading in multi-UAV systems," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5719–5732, Apr. 2023.

[139] Z. Wu, Z. Yang, C. Yang, J. Lin, Y. Liu, and X. Chen, "Joint deployment and trajectory optimization in UAV-assisted vehicular edge computing networks," *J. Commun. Netw.*, vol. 24, no. 1, pp. 47–58, Feb. 2022.

[140] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and A. Nallanathan, "Deep reinforcement learning based dynamic trajectory control for UAV-assisted mobile edge computing," *IEEE Trans. Mobile Comput.*, vol. 21, no. 10, pp. 3536–3550, Oct. 2022.

[141] H. Chen, X. Qin, Y. Li, and N. Ma, "Energy-aware path planning for obtaining fresh updates in UAV-IoT MEC systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 1791–1796.

[142] A. Asheralieva and D. Niyato, "Distributed dynamic resource management and pricing in the IoT systems with blockchain-as-a-service and UAV-enabled mobile edge computing," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1974–1993, Mar. 2020.

[143] Y. Liu, S. Xie, and Y. Zhang, "Cooperative offloading and resource management for UAV-enabled mobile edge computing in power IoT system," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12229–12239, Oct. 2020.

[144] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC- and UAV-assisted vehicular networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 131–141, Jan. 2021.

[145] A. M. Seid, G. O. Boateng, B. Mareri, G. Sun, and W. Jiang, "Multi-agent DRL for task offloading and resource allocation in multi-UAV enabled IoT edge network," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 4, pp. 4531–4547, Dec. 2021.

[146] L. Zhang, B. Jabbari, and N. Ansari, "Machine learning driven UAV-assisted edge computing," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 2220–2225.

[147] N. Cheng et al., "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1117–1129, May 2019.

[148] A. Sacco, F. Esposito, G. Marchetto, and P. Montuschi, "Sustainable task offloading in UAV networks via multi-agent reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 5003–5015, May 2021.

[149] N. Zhao, Z. Ye, Y. Pei, Y.-C. Liang, and D. Niyato, "Multi-agent deep reinforcement learning for task offloading in UAV-assisted mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6949–6960, Sep. 2022.

[150] H. Ke, H. Wang, W. Sun, and H. Sun, "Adaptive computation offloading policy for multi-access edge computing in heterogeneous wireless networks," *IEEE Trans. Netw. Service Manag.*, vol. 19, no. 1, pp. 289–305, Mar. 2022.

[151] G. Faraci, C. Grasso, and G. Schembra, "Design of a 5G network slice extension with MEC UAVs managed with reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 10, pp. 2356–2371, Oct. 2020.

[152] L. Wang, K. Wang, C. Pan, W. Xu, N. Aslam, and L. Hanzo, "Multi-agent deep reinforcement learning-based trajectory planning for multi-UAV assisted mobile edge computing," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 1, pp. 73–84, Mar. 2021.

[153] T. Ren et al., "Enabling efficient scheduling in large-scale UAV-assisted mobile-edge computing via hierarchical reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7095–7109, May 2022.

[154] W. J. Yun et al., "Cooperative multiagent deep reinforcement learning for reliable surveillance via autonomous multi-UAV control," *IEEE Trans. Ind. Informat.*, vol. 18, no. 10, pp. 7086–7096, Oct. 2022.

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

[156] F. Venturini et al., "Distributed reinforcement learning for flexible and efficient UAV swarm control," *IEEE Trans. Cogn. Commun. Netw.*, vol. 7, no. 3, pp. 955–969, Sep. 2021.

[157] R. Zhang, Q. Zong, X. Zhang, L. Dou, and B. Tian, "Game of drones: Multi-UAV pursuit-evasion game with online motion planning by deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7900–7909, Oct. 2023.

[158] Z. Xia et al., "Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 931–945, Jan. 2022.

[159] C. Wang, J. Wang, J. Wang, and X. Zhang, "Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6180–6190, Jul. 2020.

[160] Z. Ma, C. Wang, Y. Niu, X. Wang, and L. Shen, "A saliency-based reinforcement learning approach for a UAV to avoid flying obstacles," *Robot. Auton. Syst.*, vol. 100, pp. 108–118, Feb. 2018.

[161] A. Singla, S. Padakandla, and S. Bhatnagar, "Memory-based deep reinforcement learning for obstacle avoidance in UAV with limited environment knowledge," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 1, pp. 107–118, Jan. 2021.

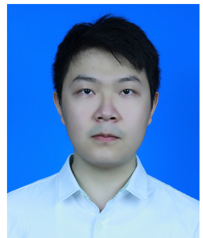
[162] D. Wang, T. Fan, T. Han, and J. Pan, "A two-stage reinforcement learning approach for multi-UAV collision avoidance under imperfect sensing," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3098–3105, Apr. 2020.

[163] J. v. den Berg, S. J. Guy, M. Lin, and D. Manocha, "Reciprocal n-body collision avoidance," in *Robotics Research*. Berlin, Germany: Springer, 2011, pp. 3–19.

[164] Y.-H. Hsu and R.-H. Gau, "Reinforcement learning-based collision avoidance and optimal trajectory planning in UAV communication networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 1, pp. 306–320, Jan. 2022.

[165] S.-M. Hung and S. N. Givigi, "A Q-learning approach to flocking with UAVs in a stochastic environment," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 186–197, Jan. 2017.

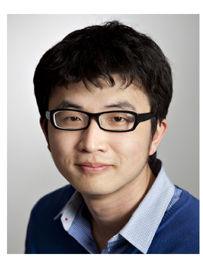
[166] C. Yan, C. Wang, X. Xiang, Z. Lan, and Y. Jiang, "Deep reinforcement learning of collision-free flocking policies for multiple fixed-wing UAVs using local situation maps," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1260–1270, Feb. 2022.
[167] G. Shen et al., "Deep reinforcement learning for flocking motion of multi-UAV systems: Learn from a digital twin," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 11141–11153, Jul. 2022.
[168] H. Huang, Y. Yang, H. Wang, Z. Ding, H. Sari, and F. Adachi, "Deep reinforcement learning for UAV navigation through massive MIMO technique," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1117–1121, Jan. 2020.
[169] Z. Mou, Y. Zhang, F. Gao, H. Wang, T. Zhang, and Z. Han, "Deep reinforcement learning based three-dimensional area coverage with UAV swarm," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3160–3176, Oct. 2021.
[170] C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li, "Multi-UAV trajectory planning for energy-efficient content coverage: A decentralized learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3193–3207, Oct. 2021.
[171] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and J. Henry, "UAV trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9540–9554, Sep. 2021.
[172] D. Hong, S. Lee, Y. H. Cho, D. Baek, J. Kim, and N. Chang, "Energy-efficient online path planning of multiple drones using reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 9725–9740, Oct. 2021.
[173] V. Saxena, J. Jaldén, and H. Klessig, "Optimal UAV base station trajectories using flow-level models for reinforcement learning," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 1101–1112, Dec. 2019.
[174] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.
[175] M. Samir, D. Ebrahimi, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Leveraging UAVs for coverage in cell-free vehicular networks: A deep reinforcement learning approach," *IEEE Trans. Mobile Comput.*, vol. 20, no. 9, pp. 2835–2847, Sep. 2021.
[176] Z. Qin, Z. Liu, G. Han, C. Lin, L. Guo, and L. Xie, "Distributed UAV-BSs trajectory optimization for user-level fair communication service with multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12290–12301, Dec. 2021.
[177] M. Nasr-Azadani, J. Abouei, and K. N. Plataniotis, "Single-and multi-agent actor-critic for initial UAV's deployment and 3D trajectory design," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 15372–15389, Aug. 2022.
[178] X. Zhang, H. Zhao, J. Wei, C. Yan, J. Xiong, and X. Liu, "Cooperative trajectory design of multiple UAV base stations with heterogeneous graph neural networks," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1495–1509, Mar. 2023.
[179] N. Parvaresh and B. Kantarci, "A continuous actor–critic deep Q-learning-enabled deployment of UAV base stations: Toward 6G small cells in the skies of smart cities," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 700–712, 2023.
[180] Y. Zhi, Z. Fu, X. Sun, and J. Yu, "Security and privacy issues of UAV: A survey," *Mobile Netw. Appl.*, vol. 25, no. 1, pp. 95–101, 2020.
[181] Y. Dang, C. Benzaïd, B. Yang, T. Taleb, and Y. Shen, "Deep-ensemble-learning-based GPS spoofing detection for cellular-connected UAVs," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25068–25085, Dec. 2022.
[182] J. Tao, T. Han, and R. Li, "Deep-reinforcement-learning-based intrusion detection in aerial computing networks," *IEEE Netw.*, vol. 35, no. 4, pp. 66–72, Jul./Aug. 2021.
[183] Y. Zhang, Z. Mou, F. Gao, J. Jiang, R. Ding, and Z. Han, "UAV-enabled secure communications by multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11599–11611, Oct. 2020.
[184] Z. Li, Y. Lu, X. Li, Z. Wang, W. Qiao, and Y. Liu, "UAV networks against multiple maneuvering smart jamming with knowledge-based reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 12289–12310, Aug. 2021.
[185] Y. Wang, Z. Su, N. Zhang, and A. Benslimane, "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1055–1069, Apr.–Jun. 2021.
[186] C. Wen, Y. Fang, and L. Qiu, "Securing UAV communication based on multi-agent deep reinforcement learning in the presence of smart UAV eavesdropper," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 1164–1169.
[187] J. Tan et al., "Sim-to-real: Learning agile locomotion for quadruped robots," 2018, *arXiv:1804.10332*.
[188] A. Anwar and A. Raychowdhury, "Autonomous navigation via deep reinforcement learning for resource constraint edge nodes using transfer learning," *IEEE Access*, vol. 8, pp. 26549–26560, 2020.
[189] A. Loquercio, A. I. Maqueda, C. R. Del-Blanco, and D. Scaramuzza, "DroNet: Learning to fly by driving," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 1088–1095, Apr. 2018.



Yu Bai received the M.S. degree from the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering, Aalto University, Espoo, Finland. His research interests include UAV networks and machine learning.

Hui Zhao (Graduate Student Member, IEEE) received the M.S. degree from the School of Computer Science, Beijing University of Technology, Beijing, China, in 2022. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include UAV wireless networks, Internet of Things, and deep reinforcement learning.

Xin Zhang is currently pursuing the M.S. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include UAV wireless networks, Internet of Things, and edge computing.



Zheng Chang (Senior Member, IEEE) received the Ph.D. degree from the University of Jyväskylä, Jyväskylä, Finland, in 2013. He has published over 140 papers in Journals and Conferences. His research interests include IoT, cloud/edge computing, security and privacy, and vehicular networks. He received Best Paper Awards from IEEE ICC in 2023, TCGCC, and APCC in 2017. He has been awarded as the 2018 IEEE Communications Society Best Young Researcher for Europe, Middle East, and Africa Region. He serves as an Editor of IEEE WIRELESS COMMUNICATIONS LETTERS, *Wireless Networks* (Springer), and *International Journal of Distributed Sensor Networks*, and a Guest Editor for IEEE NETWORK, IEEE WIRELESS COMMUNICATIONS, *IEEE Communications Magazine*, and IEEE INTERNET OF THINGS JOURNAL. He has participated in organizing workshops and special sessions in Globecom'19, WCNC'18-22, SPAWC'19, and ISWCS'18. He also serves as the Symposium Chair for ICC'20 and Globecom'23, and the Publicity Chair for INFOCOM'22.

Riku Jäntti (Senior Member, IEEE) received the M.Sc. degree (with Distinction) in electrical engineering and the D.Sc. degree (with Distinction) in automation and systems technology from the Helsinki University of Technology, in 1997 and 2001, respectively. He is a Full Professor of Communications Engineering with the Aalto University School of Electrical Engineering, Finland. Prior to joining Aalto in August 2006, he was a Professor Pro Tem with the Department of Computer Science, University of Vaasa. His research interests include machine type communications, disaggregated radio access networks, backscatter communications, quantum communications, and radio frequency inference. He is an Editorial Board Member of the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He has also been a Distinguished Lecturer of IEEE VTS (Class 2016).

Kun Yang (Fellow, IEEE) received the Ph.D. degree from the Department of Electronic and Electrical Engineering, University College London, U.K. He is currently a Chair Professor with the School of Computer Science and Electronic Engineering, University of Essex, U.K., leading the Network Convergence Laboratory. He is also an Affiliated Professor with UESTC, China. He has managed research projects funded by UK EPSRC, EU FP7/H2020, and industries. He has published 400+ papers and filed 30 patents. His main research interests include wireless networks and communications, future Internet, and edge computing. In particular he is interested in energy aspects of future communication systems, such as 6G, promoting energy self-sustainability via both energy efficiency (green communications and networks), and energy harvesting (wireless charging). He has been a Judge of GSMA GLOMO Award at World Mobile Congress–Barcelona since 2019. He serves on the editorial boards of a number of IEEE journals (e.g., IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *World Comparative Law*). He is a Deputy Editor-in-Chief of *IET Smart Cities*. He was a Distinguished Lecturer of IEEE ComSoc from 2020 to 2021. He is a Member of Academia Europaea, a Fellow of IET, and a Distinguished Member of ACM.