# A Survey on Reinforcement Learning Methods for UAV Systems

Hengsheng Chen[1], Yuanguo Lin[1], Mingjian Fu[2], Lina Yao[2], and Quan Z Sheng[2]

[1]Co-first authors
[2]Affiliation not available

January 28, 2025

# A Survey on Reinforcement Learning Methods for UAV Systems

HENGSHENG CHEN*, Fuzhou University, China
YUANGUO LIN*, Jimei University, China
MINGJIAN FU†, Fuzhou University, China
LINA YAO, The University of New South Wales, Australia
QUAN Z. SHENG, Macquarie University, Australia

In recent years, Unmanned Aerial Vehicles (UAVs) have attracted a lot of attention due to their flexibility and mobility. However, due to the increasingly complex environments faced by UAVs and the rising demands on UAV systems, traditional UAV control methods can no longer efficiently control the UAV under multi-constraint situations. Reinforcement Learning (RL), as an emerging robot control technology, is well suited to the needs of UAV systems in terms of its ability to interact with and learn from the environment. Therefore, RL-based UAV systems are gradually becoming a new trend in research. Nonetheless, as a new research field, it faces some challenges. To fully grasp the landscape of RL-based UAV systems, it is paramount to provide a comprehensive overview and analysis of the existing specific RL methods applied to UAV systems. In this survey, we first provide a comprehensive overview and summary of the application of RL in different UAV scenarios based on the classification of RL methods. After that, based on the existing relevant literature, we conduct a systematic analysis of the challenges and recent advancements when applying RL to UAV systems. Finally, we discuss the potential research directions for RL-based UAV systems.

CCS Concepts: • **Computing methodologies** → **Reinforcement learning**; • **Networks** → **Mobile networks**.

Additional Key Words and Phrases: Unmanned aerial vehicle, reinforcement learning, trajectory planning, resource scheduling

## 1 Introduction

Unmanned Aerial Vehicle (UAV) is a type of vehicle that is either controlled by a remote-controlled device or programmed to fly autonomously. Due to their flexibility and high maneuverability, UAV technology has gained significant attention and is employed in various military and civilian applications, such as search and rescue [5], mapping [132], transportation [117], and precision agriculture [106]. With the growing demand, a single UAV frequently falls short of fulfilling mission requirements in large-scale complex scenarios. As a result, researchers have begun exploring multi-UAV systems in different scenarios.

In complex environments, human control of UAV's action greatly reduces the efficiency of systems and even puts UAVs in danger due to operator errors. Therefore, autonomous control of UAVs has been studied by researchers. With the deepening of research, some researchers use the Proportional Integral Derivative (PID) method to control the behavior of UAV [10], but it relies on parameter adjustment and cannot adapt well to dynamic environments. Some researchers also

*Both authors contributed equally to this research.
†Corresponding author.

Authors' Contact Information: Hengsheng Chen, Fuzhou University, Fuzhou, China, 231027050@fzu.edu.cn; Yuanguo Lin, Jimei University, Xiamen, China, xdlyg@jmu.edu.cn; Mingjian Fu, Fuzhou University, Fuzhou, China, sinceway@fzu.edu.cn; Lina Yao, The University of New South Wales, Sydney, Australia, lina.yao@unsw.edu.au; Quan Z. Sheng, Macquarie University, Sydney, Australia, michael.sheng@mq.edu.au.

use Model Predictive Control (MPC) to optimize the UAV's decision-making [80], [126], which can accomplish some complex flight tasks, but also cannot help the UAV adapt to variable environments. The rise of artificial intelligence has provided a new track for autonomous control of UAV, and some researchers have used supervised and unsupervised learning to train models that can provide accurate predictions for UAV [62], [63]. However, both supervised and unsupervised learning require massive amounts of data to train the model, which is difficult to achieve in complex environments.

Reinforcement Learning (RL) [57], a significant branch of machine learning, focuses on learning and optimizing strategies through continuous interaction with the environment. This property enables UAVs to adjust their flight strategies based on feedback from the complex and dynamic environment. Additionally, the goal-directed nature of RL can direct UAVs to efficiently accomplish tasks, such as navigation [128] and target tracking [131], while optimizing overall performance. With advancements in deep learning, Deep Reinforcement Learning (DRL) [7] is gradually being investigated. Instead of storing state-action pairs with $Q$-tables as in traditional RL, DRL directly utilizes neural networks to generate decisions, making it easier to handle high-dimensional spaces.

**Related Work.** In recent years, using RL methods to realize autonomous control of UAVs has become a research trend. To understand the application of RL in UAV systems and to promote the field, AlMahamid et al. [4] focus on autonomous navigation of UAVs based on RL methods and discuss navigation tasks, frameworks, simulation software, challenges, and opportunities in the field. The authors in [98] detailed the computing offloading problem in aerial edge computing and discussed how RL algorithms can address the dynamics and heterogeneity challenges of the aerial edge computing environment. Bai et al. [8] investigate RL-based multi-UAVs from several different application scenarios and propose promising research directions, but neglect to discuss RL-based single UAV systems. In [89], the authors reviewed the application of DRL in single UAV-assisted communication networks, which does not fully reflect the complexity of RL for UAV systems.

**Our Contribution.** While the current works provide a detailed survey of RL-based UAV systems, some aspects are still not taken into account. This article aims to offer a thorough survey to highlight the powerful applications of RL in UAV systems. The key contributions of this survey are as follows:

- Contrast to existing surveys, we not only focus on articles examining multiple UAV systems, but also collect and analyze works investigating single UAV systems.
- We summarize and analyze the related work in terms of three classifications (i.e., value-based method, policy-based method, and Actor-Critic (AC) method) of RL algorithms for different UAV scenarios. This approach provides a more intuitive understanding of RL algorithm applications in UAV systems.
- We systematically analyze the challenges and recent advancements for RL applications in UAV systems, including high-dimensional space, limited observation, dynamic environment, and reward function definition.
- We analyze the issues of RL-based UAV systems and provide some potential future research directions, including data sampling in large-scale environments, sparse rewards in RL, cooperative control between UAVs, interpretability of RL, simulation to reality, security and privacy, and LLM for RL-based UAV systems.

As shown in Fig. 1, the rest of this article is structured as follows. Section 2 introduces the background, definition, and basis of RL, and lists some of its commonly used methods. Section 3 provides a comprehensive review of RL algorithms applied for different UAV systems scenarios. Section 4 explores the challenges and recent advancements when RL is applied to UAV systems. Section 5 highlights the problems that still exist in RL-based UAV systems and provides possible future directions. Finally, Section 6 summarizes this survey.

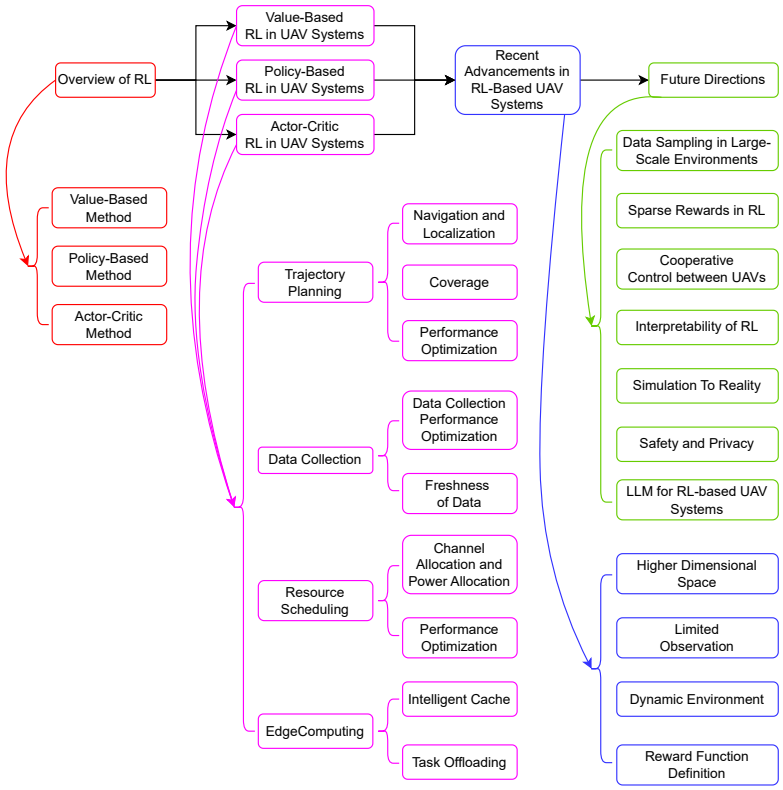


Fig. 1. Taxonomy of reinforcement learning-based UAV systems in this survey.

## 2 OVERVIEW OF REINFORCEMENT LEARNING

RL is a goal-directed learning algorithm that the agent learns to maximize a numerical reward through updating its policy [79]. At each time step, the agent selects and executes an action based on its observations and policy. This action results in a change in the environment, which then provides the agent with corresponding feedback according to those changes. After that, the agent updates its policy in response to this feedback. The agent repeats the process until the optimal policy is reached. RL, in practice, relies on a mathematical model called Markov Decision Process (MDP) [122] to operate. This model provides a structured approach to modeling and analyzing the entire learning and decision-making process of reinforcement learning.

RL is often employed to tackle various intricate problems. In practical applications, it is necessary to clarify the definition of the RL problem, including the core elements such as agent, state space, action space, and reward function. In an RL-based UAV framework, the UAV acts as an agent, and the state space can be the position of the UAV and the information observed by the UAV. The action space can be direction and speed. The reward function can include collision penalty and energy consumption reward. According to different learning objectives and strategies, RL algorithms can be categorized into three types, i.e., value-based method, policy-based method, and AC method.

### 2.1 Value-Based Method

In value-based method, the agent maintains a value table or function and uses it to select the most valuable action. This method aims to obtain an optimal value function that determines the optimal action in each state to maximize the expected return.

Q-learning algorithm [141] is a typical representative of value-based method. Q-learning can be utilized to address the control problem faced by UAVs in simple scenarios. Due to the nature of Q-learning, the actions of the UAV are usually discretized. Q-learning performs action selection by constructing a $Q$-value table, where each element of the $Q$-value table measures the maximum expected cumulative payoff that the UAV will gain when a given action is taken in a given state $s$. The $Q$-value table is continuously updated through iterative updating of values, and the $Q$-value iteration process is represented by the following equation:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma max Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)], \tag{1}$$

where $\alpha$ is the learning rate. However, Q-learning requires to maintain the Q-table continuously, making Q-learning unsuitable for complex tasks. The emergence of deep networks provided a direction for the improvement of Q-learning algorithms. With the deepening of research, algorithms such as Deep Q-Network (DQN) [95], double DQN [45], and dueling DQN [140] have been proposed. These algorithms use deep networks to process high-dimensional information and output corresponding actions, thus overcoming the limitations of traditional Q-learning algorithms.

### 2.2 Policy-Based Method

Unlike value-based method, policy-based method learns the policy by directly optimizing the parameter $\theta$ of the policy function $\pi_\theta$. Typically, we use Policy Gradient (PG) method [123] to optimize this parameter. PG method can be understood as a Monte Carlo (MC) combined with a neural network approach. PG method aims to maximize the expected cumulative return over the entire trajectory of the agent's interaction with the environment. Denoting this optimization objective as $J(\pi_\theta)$, the corresponding gradient of the function is as follows:

$$\nabla_\theta J(\theta) \propto \sum_{s \in S} \mu^\pi(s) \sum_{a \in A} q^\pi(s, a) \nabla_\theta \pi_\theta(a \mid s) = \mathbb{E}\left[\sum_a q^\pi(s_t, a) \nabla_\pi (a \mid s_t, \theta)\right], \tag{2}$$

where the symbol $\propto$ denotes a positive relationship and $\mu$ is on-policy distribution. Then, the parameter $\theta$ can be updated through a gradient ascent procedure, detailed as follows:

$$\theta = \theta + \beta \nabla J(\theta), \tag{3}$$

where $\beta$ is the step factor, which can be adjusted to realize the scaling function.

However, in an RL problem, slight parameter adjustments can lead to significant changes in the policy function. To improve the stability of the policy function, Trust Region Policy Optimization (TRPO) algorithm [111] and Proximal Policy Optimization (PPO) algorithm [112] are proposed. TRPO algorithm uses Kullback-Leibler Divergence to measure the proximity of two policies which is limited by setting a threshold. Unlike TPRO, PPO algorithm directly adds Kullback-Leibler Divergence to the original objective function, reducing the computational complexity.

### 2.3 Actor-Critic Method

AC method combines the advantages of both value-based method and policy-based method. It contains the actor network and the critic network. Specifically, the actor network optimizes its strategy based on the feedback of the value function provided by the critic network, while the critic network is responsible for training the value function and updating its evaluation with the help of the Temporal Difference (TD) error.

Within the AC method, the most well-known algorithm is Asynchronous Advantage Actor-Critic (A3C) [94]. A3C algorithm starts with a global network and creates multiple parallel environments, where each agent copies the parameters of the global network before learning. Next, each agent interacts with the environment and computes its respective gradient. Finally, the gradients are sent back to the global network. In the A3C algorithm, each agent uses different policies and updates its policy asynchronously and independently, which may lead to instability. In addition, there are many improved AC algorithms such as Advantage Actor-Critic (A2C), Soft Actor-Critic (SAC) [44], Deterministic Policy Gradient (DPG) [119], and Deep Deterministic Policy Gradient (DDPG) [78]. These algorithms can utilize the samples more efficiently and improve learning efficiency. Since the AC method involves two neural networks, it may exhibit instability during the training process.

## 3 RL-BASED UAV SYSTEMS

In this section, we provide a summary of the application of RL in different UAV scenarios following value-based method, policy-based method, and AC method, respectively.

### 3.1 Value-Based UAV Systems

This section further divides the related work to value-based UAV systems into four key components: the problem of UAV trajectory planning, the task of data collection in UAV systems, the resource allocation problem for UAV systems, and UAV-assisted edge computing networks. Fig. 2 illustrates the framework of value-based UAV systems.

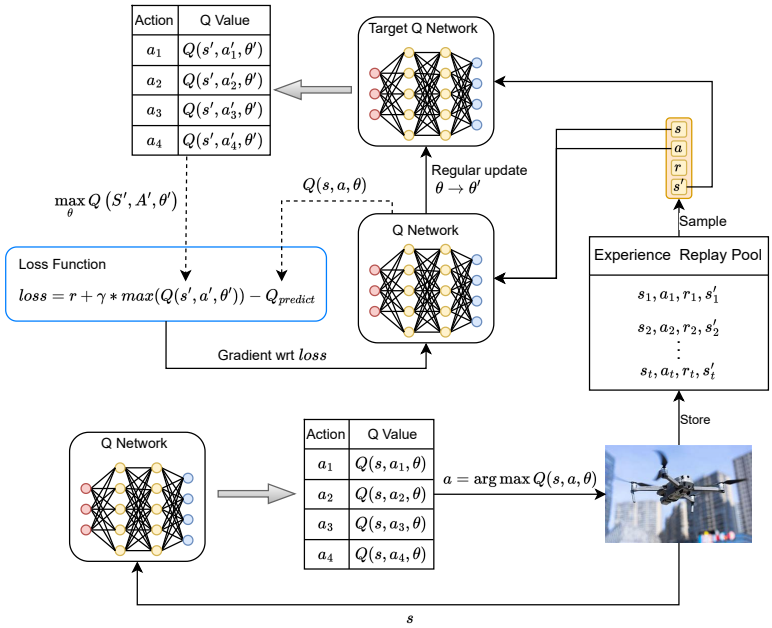


Fig. 2. Framework of value-based UAV system model.

*3.1.1 Trajectory Planning.* UAV trajectory planning entails employing specific algorithms to determine an flight trajectory for a UAV, thus ensuring its successful accomplishment of assigned tasks. Through the value-based method, UAVs can autonomously find the optimal trajectory to complete tasks. For example, [136] explores the trajectory planning problem when UAVs collect data from distributed IoT nodes in a non-cooperative situation. The authors used some practical

constraints to ensure UAV's safety while maximizing the amount of data collected from multiple IoT nodes. The authors chose the Dueling Double Deep Q-network (D3QN) framework for UAVs to learn strategies and accomplish tasks without any information provided by the environment and other UAVs. Specifically, the UAV can sense other UAVs within a certain range through the sensors it is equipped with and maximize the amount of data collected through a reward function.

The value-based UAV trajectory planning can be further divided into three key aspects: first focusing on navigation and localization [26], followed by an in-depth study of UAV coverage of Ground Devices (GDs) or specific areas [162], and finally, a study of UAV trajectory optimization to achieve system performance improvement [156].

**Navigation and Localization.** UAVs are often used for transportation and targeting due to their flexibility and mobility. However, complexity in the environment poses challenges for self-contained UAV flight. To investigate the navigation and obstacle avoidance capabilities of UAVs, the authors in [49] designed a multi-UAV transportation system. In this system, each UAV is required to accomplish two tasks. The UAVs need to transport objects to their destinations via forward trajectories and collect data from ground-based IoT devices via backward trajectories. Notably, UAVs on forward and backward tracks fly at different altitudes, thus Q-learning is used to train UAVs to avoid collisions with UAVs performing the same task.

In large-scale scenarios, it is difficult for UAVs to select targets that require communication or to maintain communication during their flight. To address the limitations imposed by sensors, the agent in [51] makes decisions based on the strength of the received signal, instead of the state information obtained by sensors. Specifically, UAV-ground links are regarded as agents that utilize massive Multiple-Input-Multiple-Output (MIMO) to receive and enhance the signal and employ DQN to optimize the policies. On the other hand, in [17], the authors addressed the problem of continuous communication in large-scale scenarios. Specifically, the UAV maintains communication by switching the objects it communicates with. However, such an approach brings additional overhead. To reduce the frequency of switches and energy consumption, the authors utilized Q-learning algorithm to optimize the policy of the UAV.

**Coverage.** Area coverage plays an indispensable role in practical tasks such as data collection and monitoring. Through area coverage, GDs can receive timely services and upload generated data. As UAV technology advances, UAVs are gradually being used to tackle the challenge of area coverage. However, due to the complexity of the environment, it isn't easy to realize autonomous area coverage control of UAVs with traditional UAV methods. Therefore, researchers try to employ the RL method to realize the autonomous flight of UAVs for area coverage. For example, in both [162] and [164], UAVs are regarded as agents and adjust their locations according to the positions of GDs to fulfill the coverage requirements. To achieve effective UAV coverage, the authors of both works introduced the RL algorithm to assist the UAV find the optimal deployment location.

UAV systems can be used to survey complex terrain, for example, [97] researches the problem of covering irregular 3D terrain. The authors proposed a geometric method to map irregular 3D terrain surfaces to many weighted 2D planes. Then, a two-level hierarchical UAV swarm architecture, including Leader UAVs (LUAVs) and Follower UAVs (FUAVs), is proposed to cover the terrain. The LUAVs are trained by a swarm DQN algorithm to select patches to be covered, and the FUAVs perform specific coverage within patches according to a designed trajectory algorithm.

**Performance Optimization.** In RL-based UAV systems, the system performance is usually enhanced by RL algorithms to achieve the objectives more efficiently. The authors in [61] used fairness as a metric to train UAVs. In this system, the UAVs broadcast power to Energy Receivers (ERs) on the ground and ensure that all ERs receive approximately the same amount of power. Note that the location of the ERs is agnostic during UAV training. Therefore, the authors used Q-learning algorithm to update the behavioral policy of the UAV.

UAVs can also be used to provide services to users, [84] aims to maximize the total mean opinion score of Ground Users (GUs) in user stationary scenarios and user mobile scenarios, respectively. In this study, each UAV needs to find the optimal deployment location. First, a method, called GAK-means, is used to group users. After that, Q-learning algorithm is used to train the UAVs to traverse these groups to collect data efficiently.

Similar to [84], in [35], UAVs are utilized to provide services to GDs in post-disaster areas. Specifically, the GDs can offload tasks into the UAVs, after which the UAVs transfer the tasks to another set of GDs. Considering the complex environment of the post-disaster area, the authors modeled the trajectory and communication scheduling of the UAV as a joint optimization problem. Then a multi-step D3QN is proposed to solve the optimization problem. In the D3QN method, the authors used multi-step bootstrapping technique to obtain the return $R_{n:n+\wp_1}$ of the future $\wp_1$ steps and improve the loss function, the improved loss function is defined as

$$L = \left( R_{t:t+\wp_1} + \gamma^{\wp_1} \hat{Q}(s_{t+\wp_1}, \arg\max_{\hat{a}} Q(s_{t+\wp_1}, \hat{a}|\theta)|\hat{\theta}) - Q(s_t, a_t|\theta) \right)^2, \tag{4}$$

where $\hat{Q}$ is the $Q$ function value of the target network and $\hat{a}$ denotes all future actions. $\theta$ is the evaluation network parameter updated by utilizing the gradient descent method based on Eq. (4).

*3.1.2 Data Collection.* The high flexibility of UAVs enables them to collect data in wireless communication networks. UAV-assisted networks not only improve the efficiency of data collection but also avoid energy depletion of GDs by communicating between UAVs and GDs. However, due to limited energy and the requirement of fairness, the UAV must make reasonable decisions about its data collection and flight trajectory. RL algorithms can be well suited to guide UAVs in their decision-making to learn strategies adequately.

The value-based UAV data collection system can be further categorized into two aspects: data collection performance optimization [151] and data freshness [82]. The former aims to analyze how RL algorithms can help UAVs in data collection tasks, while the latter describes how RL algorithms can improve data freshness.

**Data Collection Performance Optimization.** To enhance the efficiency of UAVs in collecting data from GDs, [33] presents a Q-learning-based approach to design the UAV trajectory to collect data. Specifically, the deployment area is split into several grids. Each grid contains several sensors, and the center is configured with a wireless charging device for charging the UAV. It is worth noting that data collection is only possible when the UAV hovers over the devices. To improve the performance of the system, the authors proposed a Q-Learning-Based Energy Efficient Data Collection by the UAV (QEDU) algorithm and a Q-Learning-Based Throughput-Maximizing Data Collection by the UAV (QTDU) to find the corresponding UAV routes to optimize the UAV's strategy.

Since GDs are typically constrained by limited energy which restricts the sustainability of the services provided by UAVs and GDs. To that end, Kai Li et al. [68] proposed an onboard deep Q-network to enable UAVs to schedule online MPT and data collection to prevent battery exhaustion and data buffer overflow of the devices. This method minimizes the total packet loss of the GDs by optimally deciding the GD selection, modulation scheme, and instantaneous patrolling speed of the UAV. In addition, [149] investigates the data transmission scheme for Massive Machine Type Communication (mMTC) using UAVs equipped with a Simultaneous Wireless Information and Power Transfer (SWIPT) device. In this scheme, the UAV can charge GDs and collect data from GDs, while the UAV can also charge itself by receiving energy from the environment and Base Station (BS). To achieve maximum long-term utility, DQN algorithm is used to adjust the strategy for UAV data collection, transmission, and energy reception.

**Freshness of Data.** The average age of information (AOI) can effectively assess the freshness of data, defined as the elapsed time from the moment data is generated until it is received. In UAV data collection tasks, AOI is often used as an evaluation metric to indicate the performance of the system. In [21], the authors investigated wireless power networks for UAVs. This study considers dynamic time-varying channels and builds the corresponding channel model. To minimize the AOI under dynamic channel conditions, a DQN-based approach is proposed to find the optimal execution strategy for UAVs according to channel variations and under energy constraints.

In [1], UAVs are used as mobile BSs for data collection to optimize UAV trajectories under the constraints of AoI. The optimization problem is proved to be NP-hard. Thus, the authors employed a DQN-based algorithm to solve it. Moreover, the experience replay technique is utilized to enhance the stability of the training process. The authors in [113] investigated data collection scenarios in UAV swarms containing many sensors to collect them efficiently. Compared to [1], this work further considers the energy requirements of UAVs working continuously for a long period. Due to energy constraints, the UAVs need to travel to a designated area for recharging. To coordinate the data collection and charging strategies of the UAVs, a double DQN-based approach is proposed where the reward function combines an energy consumption and collision penalty to train the UAV to find more energy-efficient and safer strategies.

*3.1.3 Resource Scheduling.* UAVs usually have limited battery life and computing resources, and according to different mission requirements, UAVs need to allocate their resources to achieve efficient mission execution reasonably. Since UAV resource allocation is usually a dynamic process, decisions need to be made based on the real-time state. RL can learn the optimal decision-making strategy through continuous trial and error, making it an effective method for handling complex resource allocation problems. For example, in [19], the authors proposed a feasible communication scheme by combining Non-Orthogonal Multiple Access (NOMA), clustering, and RL. Specifically, the communication is divided into three phases. First, the authors introduced a Load-Balancing Fuzzy C-Means (LB-FCM) algorithm to categorize UAV swarms and select Cluster Heads (CHs), while ensuring balanced cluster sizes. The second phase is data aggregation, where members of each cluster establish communication with their CHs via U2U links. In this phase, a DQN-based RL framework, called MAARL, is applied to optimize resource allocation for each CH to maximize the data aggregation rate. The last phase is data offloading, where CHs transmit collected information to GUs using an RL-based algorithm to maximize the data offloading rate.

Focusing on the value-based resource allocation method for UAVs, we further investigate how the method performs task allocation [90] and how it improves the system's performance [65], [150].

**Channel Allocation and Power Allocation.** Cognitive Radio (CR) technology enables devices to dynamically access idle spectrum, and its integration into UAV systems is seen as an effective solution to address the spectrum allocation problem. In [71], the authors proposed a co-design strategy for the Cognitive UAV (CUAV) assisted networks aiming at energy-efficient traffic offloading. By embedding cognitive radios into CUAVs, the UAVs can identify and access the idle spectrum and build backhaul links by identifying and accessing the idle spectrum. To maximize the energy efficiency, the authors used the DDQN algorithm to update the control strategy of the CUAV based on the reward function defined as

$$r_t = \frac{\min\left\{W_{\mathrm{c}}^t, W_{\mathrm{tr}}^t\right\}}{\sigma \cdot \left(E_{\mathrm{pro}}^t + \sum_{k=1}^{K} b_{k,t} E_{\mathrm{com}}^{k,t}\right)}, \tag{5}$$

where $W_{\mathrm{c}}^t$ denotes the amount of data collected at time $t$ and $W_{\mathrm{tr}}^t$ denotes the data volume transmitted by the CUAV. The balance between $W_{\mathrm{c}}^t$ and $W_{\mathrm{tr}}^t$ depends on the time allocation of the offloading

phase, transmission power, and band selection. $\sigma$ is a bias factor used to weigh offloaded traffic and energy consumption of the CUAV. $E^t_{\text{pro}}$ and $E^{k,t}_{\text{com}}$ denote the propulsion energy consumption and communication energy consumption of CUAV, respectively, and both metrics depend on the trajectory length and time allocation. $K$ is the number of available bands and $b_{k,t} \in \{0, 1\}$ denotes the idleness of the k-th band. Therefore, to maximize the reward $r_t$, it is necessary to undertake a joint optimization of the trajectory, time allocation, transmission power, and band selection.

To guarantee the Quality of Service (QoS) of users in heavy-traffic communication networks, [42] proposes a digital twin dynamic resource allocation method. This method introduces airborne UAVs and Device-to-Device (D2D) communication, which enables efficient data transmission even for devices with limited communication capabilities. The authors proposed a joint optimization problem for relay node selection, bandwidth allocation, and trajectory assignment of UAVs. To address this problem, the authors introduced the DDQN algorithm for policy exploration. Similarly, [67] investigates a dynamic UAV efficient communication network in which UAVs fly in a given flight path and provide continuous communication services to GUs. Considering the dynamic nature of communication networks, a DQN-based approach is developed to ensure coverage continuity by adjusting UAVs' channel and power allocation strategies. Simulation results show that the DQN-based method exhibits good performance under different flight paths and coverage requirements.

**Performance Optimization.** To enhance the quality of air-to-ground communication, the authors in [60] proposed a new RL framework for maximizing the total system throughput while minimizing device drops. This RL framework can be divided into two parts: internal RL and external RL. Internal RL considers UAVs and terrestrial BSs as agents to find the optimal power transfer policy through Q-learning algorithm. External RL only treats the UAV as an agent and determines its optimal deployment location.

To coordinate collaboration among UAVs and achieve efficient resource allocation, [27] investigates the UAV trajectory planning problem in mobile edge computing. Assuming an urban area where GDs are randomly distributed and move with a certain strategy, UAVs need to move to the target location and respond to the requests from GDs while avoiding collisions. Due to the complexity of the environment and the user's changing requirements, a DQN-based approach is proposed. In this approach, the reward function combines collision penalties, user requirements, and energy penalties to plan a safe and energy-efficient trajectory and allocate resources efficiently.

*3.1.4 Edge Computing.* UAV-assisted edge computing utilizes UAVs for data transmission, processing, and control. Compared to traditional edge computing networks, UAV-assisted edge computing is more efficient and flexible. In UAV-assisted edge computing, the UAV is often used for task offloading and intelligent caching. Task offloading can offload excessive tasks from GDs to the UAV or the cloud to efficiently execute tasks. Intelligent caching reduces the demands of centralized servers by caching commonly used data or resources on edge devices. However, the associated task offloading decisions and content caching decisions are usually NP-hard problems. As a result, RL-based UAV-assisted edge computing has emerged.

In related work to value-based UAV-assisted edge computing, we focus on two main aspects: intelligent caching [6] and task offloading [52].

**Intelligent Cache.** The growth of multimedia applications has led to a surge in network traffic, resulting in a huge amount of computation. To fulfill GDs' need for low latency, [15] considers a wireless caching network where UAVs cache popular content and deliver it to GDs on demand. To provide better services to GDs, UAVs must update their location in real-time to track the corresponding mobile users. The authors proposed a Q-learning-based Cooperative Multi-Agent RL (CMARL) algorithm for training UAVs to adapt to the frequently changing environments. In this method, UAVs can help each other learn by sharing information.

Similar to [15], in [160], the authors also used a UAV as an airborne BS for caching popular content. To improve the QoS for terrestrial users, the authors used the NOMA technique for UAV cellular networks and proposed a joint optimization problem for UAV caching content updating and scheduling and content delivery latency. Then, A Q-learning-based RL algorithm that employs a soft greedy strategy to select actions is proposed to handle the optimization problem. Meanwhile, considering the limitations of the $Q$-table, a function approximation method is proposed, which employs Stochastic Gradient Descent (SGD) to search for actions and uses DNNs to establish a mapping between output states and actions.

**Task Offloading.** With today's growing demand for user traffic, terrestrial communication technologies are challenged with limited computing resources. The use of UAVs for computing offloading has started to gain attention to reducing the amount of computation at terrestrial BSs [124]. To tackle the challenges of increasing computing requirements in hotspots, [12] construct a novel UAV system by introducing wireless power supply technology and low-power scattering communication technology. In this system, the UAV can supply power to the hotspot and collect users' data by backhauling. Considering the stochastic nature of the task and the dynamic variations in channel conditions, a DQN-based approach is proposed to maximize the long-term utility of the hotspot. In addition, the method incentivizes UAVs to share resources by introducing Lagrange multipliers to design an optimal contract for each UAV.

In emergency scenarios, the system's computational efficiency is of great concern, as achieving higher efficiency enables better fulfillment of real-time demands. However, the system efficiency is limited by the computing resources and energy of the UAV, to this end, [158] proposes a DQN-based approach to maximize the system efficiency. In this work, the authors used a single UAV to serve the GU and train the UAV according to the proposed approach to find the best trajectory and offloading decision. However, in real emergency missions, the freshness of the extracted information plays a crucial role, affecting the detection accuracy. In [14], the AOI is further considered and constitutes a joint optimization problem with the calculation of unloading energy consumption and the UAV trajectory planning. In this case, the authors proposed a D3QN-based algorithm to help the UAV make decisions, including the flight direction and the task offloading sequence.

Table 1 lists related works that apply value-based mehod. It outlines the application scenarios, RL algorithms, and performance metrics in these works.

### 3.2 Policy-Based UAV Systems

In UAV systems, value-based method indirectly optimizes policies by evaluating the value of states, making it suitable for simple tasks. However, value-based method is usually less applicable when the UAV is required to achieve fine control. In contrast, policy-based method directly learns the policies, making it perform superiorly when dealing with complex policy requirements. Furthermore, this method is capable of learning stochastic policies, enabling UAVs to continuously experiment with new flight trajectories and policies during flight, thereby enhancing their exploration capabilities. The framework of policy-based UAV systems is shown in Fig. 3.

In this section, we outline the related work on UAV systems using policy-based method, which are developed from four application scenarios, i.e., trajectory planning, data collection, resource allocation and scheduling, and edge computing.

*3.2.1 Trajectory Planning.* In the policy-based UAV trajectory planning scenario, we focus on autonomous navigation and localization, coverage and overall performance optimization [110].

**Navigation and Localization.** During the actual flight, the environment is casually variable, which brings challenges to the flight of UAVs. To allow UAVs to quickly adapt to new environments and learn excellent flight strategies, [72] presents an approach based on the PPO algorithm.

Table 1. Overview of Value-based Method for UAV Systems.

| RL Classification | Scenario | Work | RL Algorithm | Performance Indicator | Year |
|---|---|---|---|---|---|
| Value-based | Trajectory Planning | [136] | D3QN | success rate, data collection rate, collision rate, data collection, success rate | 2022 |
| | | [49] | Q-Learning | average trajectory length | 2020 |
| | | [51] | DQN | coverage performance, convergence performance | 2019 |
| | | [17] | Q-Learning | number of handoffs, disconnectivity percentage, battery consumption | 2023 |
| | | [26] | Q-Learning | average localization error | 2020 |
| | | [162] | DQN | downlink capacity | 2020 |
| | | [164] | Q-Learning | energy efficiency | 2021 |
| | | [97] | DQN | area coverage | 2021 |
| | | [61] | Q-Learning | total harvested energy | 2019 |
| | | [156] | DQN | accumulated reward, average data buffer length, average residual battery level, average number of devices with data overflow, average number of devices out of battery | 2020 |
| | | [84] | Q-Learning | QoS | 2019 |
| | | [35] | D3QN | the whole operation time, 3D trajectory results, GD scheduling results, the robustness of the algorithm | 2024 |
| | Data Collection | [33] | Q-Learning | average throughput, delay, energy efficiency | 2021 |
| | | [151] | DQN | reward, completion time, | 2022 |
| | | [68] | DQN | network cost, packet loss rate, patrolling velocity | 2019 |
| | | [149] | DQN | average long-term expected utility, average delay, relay energy efficiency | 2020 |
| | | [82] | DQN | average AoI, energy efficiency, convergence performance | 2021 |
| | | [21] | DQN | AoI | 2022 |
| | | [1] | DQN | average cumulative reward, average AoI, average energy efficiency, average bandwidth efficiency, average UAV resource utilization | 2020 |
| | | [113] | DDQN | the cumulative reward and the fitness score | 2024 |
| | Resource Scheduling | [19] | Double DQN, DDPG | training reward, network sum rate, fairness of UAVs, priority of GS | 2023 |
| | | [90] | DQN | convergence performance, total reward, achievable sum rate | 2021 |
| | | [71] | DDQN | accumulated energy efficiency, trajectory result | 2022 |
| | | [42] | DQN | average delay, packet loss rate, communication overhead | 2023 |
| | | [67] | DQN | the spectral efficiency, the rate variance of ABS | 2023 |
| | | [65] | Q-Learning | average reward, accumulated average energy efficiency, accumulated average throughput, accumulated average number of outage users | 2021 |
| | | [60] | Q-Learning | fairness, accumulated reward, throughput, the number of outage devices | 2024 |
| | | [150] | Dueling DQN | energy efficiency, computation time | 2023 |
| | | [27] | Q-Learning | average path length, QoS, average risk, total time elapse | 2024 |
| | Edge Computing | [15] | Q-Learning | network throughput, throughput ratio | 2021 |
| | | [160] | Q-Learning | content delivery delay, cache hit radio | 2020 |
| | | [6] | Dueling DQN | convergence performance, QoS satisfaction, average minimum transmit power, total transmit power, cache resource utilization | 2021 |
| | | [124] | Double Q-Learning | package loss, network throughput, packet delay | 2021 |
| | | [52] | DQN | sum cost, task offloading decision accuracy | 2022 |
| | | [158] | DQN | rewards | 2021 |
| | | [14] | D3QN | energy consumption, averaged age of update, trajectory result | 2022 |
| | | [12] | DQN | the learning reward, task drop rate, energy storage, energy fluctuation variance, feasibility, efficiency | 2024 |

Specifically, the UAV is considered as an agent that needs to start from an initial position and track the user moving on a fixed path, while the agent is trained on different tasks during the training process. The variety of tasks enables the UAV to swiftly adapt to different environments. Notably, the UAV operates at a constant altitude, which means that UAVs are equivalent to moving on a flat surface, reducing the complexity of the training.

To enable UAVs to reach their destinations safely, [77] proposes a PPO-based framework for achieving autonomous collision avoidance of UAVs in limited airspace. Specifically, this framework considers a region containing $N$ UAVs and $M$ obstacles, and each UAV needs to reach its destination without encountering obstacles at the same time as the other UAVs. The UAVs observe part of the information through sensors, and the observation space of the $i$-th UAV is noted as $o_t$ =

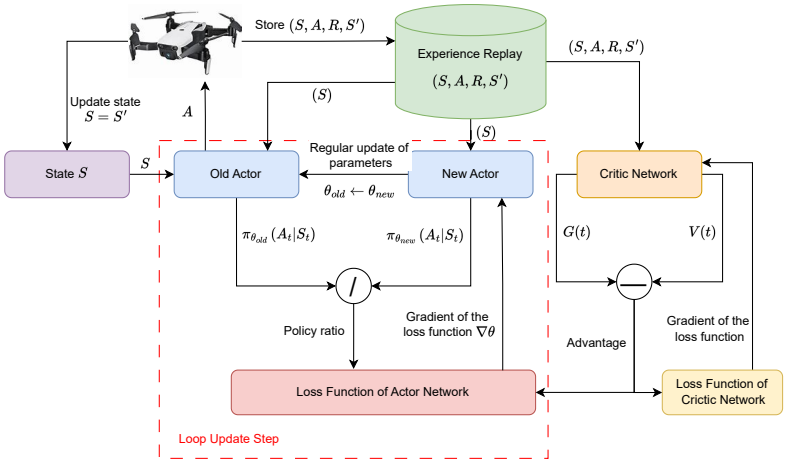


Fig. 3. Framework of policy-based UAV system model.

$\sum_i^{N_i} \left\langle {}^i o_t^{nbr}, {}^i o_t^g, {}^i o_t^{vw} \right\rangle + {}^i o_t^{prevw}$, where ${}^i o_t^{nbr}$ is the neighbor position, ${}^i o_t^g$ is target position, ${}^i o_t^{vw}$ denotes the velocity and ${}^i o_t^{prevw}$ is the ith UAV's predefined speed. To achieve autonomous collision avoidance for the UAVs, the authors introduced the PPO algorithm to train the UAVs. Considering that the PPO algorithm suffers from the defect of steady-state error, a Generalized Integrator Compensator (GIC) is used to preprocess the observation space set $s_o^t$ of all UAVs at time $t$, and the result is denoted as $s_g^t$, which is defined as

$$s_g^t = s_o^t + \sum_{i=1}^{t} \varsigma^{N_t+1-i} s_o^i, \tag{6}$$

where $\varsigma$ is an integral compensation coefficient. $\varsigma^{N_t+1-i}$ denotes the weight coefficient at time $i$, which indicates that more recent historical states will receive more attention. In addition, the authors added LSTM to the actor network and critical network to enhance the algorithm's efficiency.

**Coverage.** Using multiple UAVs to rapidly establish communications in disaster areas and provide coverage to as many GDs as possible is a major challenge. To help UAVs better collaborate and cover as many GUs as possible, Liang et al. [56] et al. proposed a UAV path planning method that combines Graph Aggregator (GAT) and PPO. Specifically, the method first utilizes GAT to aggregate the observation information of UAVs and the information of neighboring UAVs, where the information of neighboring UAVs is selected by a relational topology graph. After that, the feature vectors output from the GAT are treated as the state at the current moment and input into the PPO. In this way, it allows the PPO algorithm to make full use of its own and neighbor's observation information for better collaboration. However, covering too many GUs may result in insufficient UAV resources to satisfy all GUs. To this end, Guan et al. [40] introduced SINR constraints and UAV resource limitation constraints, aiming to satisfy high-quality services while covering more GUs. In this work, the UAV communicates through RF modules and FSO modules worn on its body. To cope with different channel conditions and the mobility of GUs, a MAPPO-based UAV trajectory planning method is proposed to schedule UAVs. In addition, To enhance the efficiency of the training process, the authors introduced an enhanced K-Means algorithm that reduces the dimensionality of the state space by grouping the GUs.

**Performance Optimization.** In UAV communication scenarios, UAVs must constantly adjust their positions to provide communication for GDs. To this end, the authors in [31] proposed a

method based on the improved Multi-Agent Proximal Policy Optimization (MAPPO) algorithm for optimizing the UAV trajectory. This method introduces a policy trimming and average evaluation mechanism to address the bias estimation and local optimum convergence problems of the MAPPO algorithm. Specifically, the policy tailoring restricts action choices to speed up learning, while the average evaluation mechanism increases evaluation accuracy using multiple networks. Considering the inefficiency of traditional Downlink-Uplink (DL-UL) coupling, the authors in [18] introduced the idea of DL-UL Decoupling (DUDe) , which allows users to associate different BSs. Due to the intractable non-convexity problem arising from in-band full-duplex communication and DUDe, the authors formulated the joint problem of decoupled DL-UL association and trajectory planning as a Partially Observed Markov Decision Process (POMDP). To address the POMDP, an improved clip and count-based PPO algorithm is developed.

In UAV listening scenarios, excellent listening capabilities can help UAVs do their job better. In [41], the authors considered a scenario where legitimate UAVs are used to listen in on suspicious UAVs. In this case, the legitimate UAVs improve their listening capability by transmitting jamming signals to the suspicious UAVs and adjusting their flight direction. Due to environmental uncertainties, the authors formulated the sequential decision-making problem as an MDP. To solve the MDP, a MAPPO-based method is developed to optimize the legitimate UAVs' policies of trajectory planning and power control. The method first obtains the interference power allocation policy in the current state by a constrained linear programming method. Then, based on the interference power allocation policy, the MAPPO algorithm is employed to learn the movement policy of each UAV to improve the listening performance.

*3.2.2 Data Collection.* In the policy-based UAV data collection scenario, We further categorize the purpose into two categories: data collection optimization and data freshness [108].

**Data Collection Optimization.** With flexibility in collecting measurement data, UAVs have great potential to improve the accuracy of spectrum mapping (SC). [69] investigate the scenario of using UAVs in SC, and the main goal is to use UAVs to efficiently collect RSS measurements to maximize SC accuracy and energy efficiency. In this system, the transmitter source is dynamic and unknown, making it difficult to obtain a priori knowledge. To this end, this article proposes a PPO-based trajectory optimization algorithm that introduces a backtracking method to construct a backtracking dominance function to solve the sparse feedback problem.

[165] develops a new UAV-IoT system for collecting data from CHs. In this system, the UAV knows the locations of all the CHs in advance. Therefore, the UAV needs to optimize its trajectory to access each CH while reducing energy consumption. This optimization problem is treated as a generalized traveler's problem and then solved by a DRL approach based on sequence-to-sequence neural networks. Specifically, the system trains the Seq2Seq neural network model by using the REINFORCE algorithm [145] to make it output the optimal CH selection policy. However, the location and number of GDs are often unknown because the sensing devices are at the periphery or cannot communicate due to accidents. This creates a challenge for data collection in UAV systems. To this end, the authors in [83] proposed a PPO-based UAV trajectory planning method for data collection tasks in IoT networks. First, the PPO algorithm optimizes the hover point location of the UAV by using the historical location of the GD to maximize the number of communications. After that, the simulated annealing method is employed to traverse all the hover points to reduce the energy costs of the UAV. Finally, a Cluster-Head Searching Algorithm with Autonomous Exploration Pattern (CHSA-AEP) is proposed to ensure the UAV collects data efficiently.

**Freshness of Data.** Improving data freshness can effectively meet the low-latency requirements in delay-sensitive applications. However, in UAV systems, the complex environment makes traditional methods ineffective in improving data freshness. [20] investigates the data collection

problem of UAV crowd sensing in delay-sensitive applications, and proposes a PPO-based sensing framework for collecting data efficiently. The framework includes a synchronous computational architecture that integrates the information of neighboring time slots through GRUs to generate more energy-efficient UAV flight strategies. To achieve multi-objective optimization, the authors proposed a trade-off reward function for UAV training, combining multiple constraints.

An RL-based mean field resource allocation method is proposed in [28] to improve the freshness of data. In this work, The problem of optimizing both trajectory planning and resource allocation for UAVs is formulated as a Mean Field Game (MFG) problem. Since the MFG problem is difficult to solve with traditional optimization problems, the authors proposed a new Mean Field Hybrid Proximal Policy Optimization (MF-HPPO) method to solve it. In this method, the Fokker-Planck-Kolmogorov (FPK) equation is used to adjust the speed of UAVs, and LSTM is used to predict the network state. Both FPK and LSTM improve the UAVs' data acquisition performance.

*3.2.3 Resource Scheduling.* Policy-based UAV resource scheduling approaches can be further categorized into two key parts: channel allocation and power allocation [47], which focuses on resource management, and performance optimization [37], aiming at improving overall efficiency.

**Channel Allocation and Power Allocation.** RIS technology provides a new direction for UAV wireless communication in multi-obstacle scenarios. With RIS technology, the signal can be transmitted to overcome the bottleneck of building obstruction. [100] studies the downlink of communication-impaired UAVs and UEs and proposes a Reconfigurable Intelligent Surface (RIS)-assisted multi-UAV wireless network in conjunction with RIS technology. To maximize the system's energy efficiency, the authors presented the optimization problem combining the transmit power of the UAVs and the phase shift of the RIS and proposed two DRL-based methods based on DDPG and PPO to solve it, respectively. Based on [100], the authors in [101] further considered the mobility of UAVs and the limited energy of IoT devices. In this system, UAVs need to fly autonomously for communication and energy transfer. In this case, the authors proposed two DRL-based methods to jointly optimize the UAV's control policy to complete the flight task.

[32] proposes a game theory and DRL-based approach aimed at jointly optimizing resource allocation and UAV deployment to improve the system's throughput. To achieve interference management among UAVs, the method employs a game-theory-based price mechanism that employs pricing factors to adjust the resource allocation and reduce the dimensionality of the optimization variables. Then, a Price-Based Proximal Policy Optimization (3PO) algorithm is proposed to find the optimal policy for allocating resources and deploying the UAV.

**Performance Optimization.** In emergency scenarios, it is vital to improve the efficiency of the system. To this end, both [137] and [66] have investigated the resource allocation strategies of UAVs in emergency scenarios. Specifically, [137] addresses the application of UAVs in wearable networks for medical emergency response. The authors proposed an improved K-Means method to group wearable devices, and an algorithm based on the collaborative decision-making of multiple agents to determine the order in which UAVs access these groups. Moreover, a PPO-based approach is proposed to determine the resource allocation policy for UAVs to maximize the total throughput and efficiency of the network. [66] investigates an emergency communication system that consists of a UAV and multiple Simultaneously Transmitting and Reflecting Reconfigurable Intelligent Surfaces (STAR-RISs) and multiple GUs. In this system, communication between the UAV and GUs is facilitated through the STAR-RIS. Considering the special requirements of the disaster area scenario, the authors proposed a joint optimization problem for the UAV trajectory, resource allocation, and passive beamforming of STAR-RIS. To solve this problem, the authors proposed an algorithm that combines Lagrangian relaxation and PPO to help establish excellent communication between UAVs and GUs.

*3.2.4 Edge Computing.* This section focuses on UAV-assisted edge computing where a policy-based method is used. We further explore UAV-assisted edge computing by categorizing them into two main groups: one is intelligent caching [138] and another is task offloading.

**Intelligent Caching.** By caching content, it is possible to deliver content faster to devices that need it, thereby improving the QoS. For example, [3] investigates a scenario where the UAV provides services to cars on a highway. In this case, the UAV partially caches services from a content repository and can provide content services to passing vehicles. The authors formulated the cache replacement, trajectory planning, and resource allocation of the UAV together as an optimization problem. To tackle this problem and further reduce the energy cost, the PPO-clip algorithm is introduced to adjust the control policy of the UAV. The reward consists of two parts, a positive reward is provided when the UAV provides sufficient service to the vehicle and a negative reward is provided when the UAV movement consumes energy. Building on [3], [2] considers a collaborative caching system to provide content services to vehicles. Specifically, the system architecture consists of a Roadside Unit (RSU), a UAV, and a set of vehicles. In this architecture, the RSU and the UAV collaborate to provide complete content to the vehicles and the vehicles can upload the received content to the UAV, thereby refreshing the UAV's cache. In addition, the authors proposed a Dual-Task DRL (DTDRL) method to facilitate collaboration between RSUs and UAVs.

Content caching is also commonly used in multimedia applications to improve user satisfaction by delivering content to users instantaneously. In [54], the authors studied the multimedia content delivery problem for UAV-assisted cellular networks using edge caching techniques. To minimize the delay for users to access the content, the authors modeled cache placement, multiuser associations, flight trajectory planning, and power allocation as an optimization problem. Then, a Double Clip Proximal Policy Optimization (DC-PPO) algorithm is proposed to solve this problem. In addition, the authors used a Beyond the Boundary of Explored Regions (BeBold)-based exploration criterion to encourage the UAV to explore uncharted territory.

**Computing Offloading.** Communication in disaster areas relies on the timeliness of task completion, but limited computing resources constrain the time for task completion. To this end, [58] employs UAVs to assist the ground BS in the computation. Considering that the heterogeneous QoS requirements of computing tasks cannot be met by only auxiliary computation via UAVs, the authors introduced a hierarchical aerial computation system for High Altitude Platforms (HAPs). Specifically, UAVs can collect tasks from GUs and offload some of them to the HAP through their own policies, reducing their tasks to meet the latency requirements better. To find the best policies for UAVs, the authors developed a MAPPO-based algorithm to deal with a joint optimization problem combining resource allocation, and collision avoidance constraints.

Similar to [58], [121] also studied a system using UAVs to assist in computation, aiming to minimize the total delay and energy consumption of the system. Specifically, the authors proposed an improved Evolutionary Multi-Objective Reinforcement Learning (EMORL) to optimize conflicting objectives simultaneously. The method is divided into a warm-up phase and an evolutionary phase. In the warm-up phase, a set of initial policies is obtained by optimizing randomly generated tasks using Multi-Task Multi-Objective Proximal Policy Optimization (MMPPO). Then, the evolutionary phase optimizes these initial strategies by generating a set of new tasks based on the objective weights and repeats this step until a predefined number of evolutionary generations is reached.

Table 2 summarizes the related work on UAV systems based on policy-based method. It gives the specific RL algorithms used in the related work and the performance metrics of the experiments.

### 3.3 Actor-Critic UAV Systems

The advantages of the AC method are that it performs well in balancing multiple requirements and can effectively make real-time adjustments. Therefore, the AC method is particularly suitable

Table 2. Overview of policy-based method for UAV systems.

| RL Classification | Scenario | Work | RL Algorithm | Performance Indicator | Year |
|---|---|---|---|---|---|
| Policy-based | Trajectory Planning | [72] | PPO | cumulative average reward value, success rate, average time, success rate of collision avoidance, average minimum path, average path, cumulative average minimum step value | 2020 |
| | | [77] | PPO | the convergence of the reward curve, the coverage of PoIs, energy consumption, a comprehensive metric that combines the PoI coverage and energy efficiency | 2023 |
| | | [56] | PPO | average sum rate | 2023 |
| | | [40] | MAPPO | the number of interrupted MUs, the total throughput between the UAVs and the serving MUs, the backhaul throughput of FSO communication | 2024 |
| | | [110] | PPO | average flow throughput, empirical distribution | 2019 |
| | | [31] | MAPPO | mean episode reward | 2023 |
| | | [41] | decentralized MAPPO | eavesdropping rate, success rate | 2023 |
| | | [18] | PPO | reward | 2022 |
| | Data Collection | [69] | PPO | the normalized absolute error, energy consumption | 2023 |
| | | [165] | REINFORCE | trajectory fit, energy consumption | 2021 |
| | | [83] | PPO | UAV energy consumption, time utilization ratio | 2022 |
| | | [20] | PPO | data collection ratio, average system delay, geographical fairness, energy consumption, energy efficiency | 2021 |
| | | [108] | PPO | average AoI | 2020 |
| | | [28] | PPO | AoI | 2024 |
| | Resource Scheduling | [100] | DDPG, Clip-PPO | energy efficiency | 2021 |
| | | [101] | DDPG, Clip-PPO | the total network sum-rate | 2022 |
| | | [47] | TRPO, DDPG | convergence performance, energy efficiency | 2021 |
| | | [32] | PPO | cumulative rewards, system throughput, system energy efficiency | 2023 |
| | | [137] | PPO | convergence performance, service fairness rate, energy consumption | 2024 |
| | | [66] | PPO | average rates, cumulative energy consumption, throughput | 2023 |
| | | [37] | PPO | sum rate, energy efficiency, spectral efficiency, Jain's fairness index | 2024 |
| | Edge Computing | [3] | PPO-Clip | convergence, energy efficiency level, amount served to vehicles, energy consumption | 2020 |
| | | [2] | PPO | service rates of vehicles | 2021 |
| | | [54] | PPO | cumulative reward, content acquisition delays | 2022 |
| | | [138] | PPO | effective AoI | 2023 |
| | | [58] | MAPPO | rewards, amount of computed tasks, average satisfaction ratio | 2023 |
| | | [121] | MAPPO | the inverted generational distance, hyper volume, comprehensive objective indicator, Friedman test | 2022 |

for UAV wireless communication networks that need to consider multiple factors and cope with dynamic changes. The system based on the AC algorithm is demonstrated in Fig. 4.

This section outlines related works that employ AC method in four UAV scenarios (i.e., trajectory planning, data collection, resource allocation, and edge computing).

*3.3.1 Trajectory Planning.* We further divide the AC-based UAV systems for trajectory planning into three primary parts: navigation and localization [129], [152], coverage, and performance optimization [133]. The first two sections describe how RL can help UAVs with navigation tasks, and the last section illustrates the effectiveness of RL applied to UAV systems.

**Navigation and Localization.** Due to the uncertainty of the environment and the limited perception of the UAV, the authors in [128] proposed a DRL-based approach to address the sparse rewards challenge for UAV navigation tasks. In the training process, the UAV completes tasks of different difficulties according to a prior policy in the nonexpert assistant and learns a new policy using the A3C algorithm. Denoting the learning objective of the UAV as $\eta(\pi_b)$, the gradient of the learning objective is defined as

$$\nabla_\theta \eta(\pi_b) = \mathbb{E}_{s\sim d_{\pi_b}(s)}\left[\nabla_\theta \log \pi(a|s) A_{\pi_b}(s,a)\right], \tag{7}$$

where $d_{\pi_b}$ denotes the distribution of states under the behavior policy $\pi_b$. The behavior policy $\pi_b$ combines the learning policy $\pi(a|s)$ and the prior control policy $\pi_h(a|s)$, which is achieved

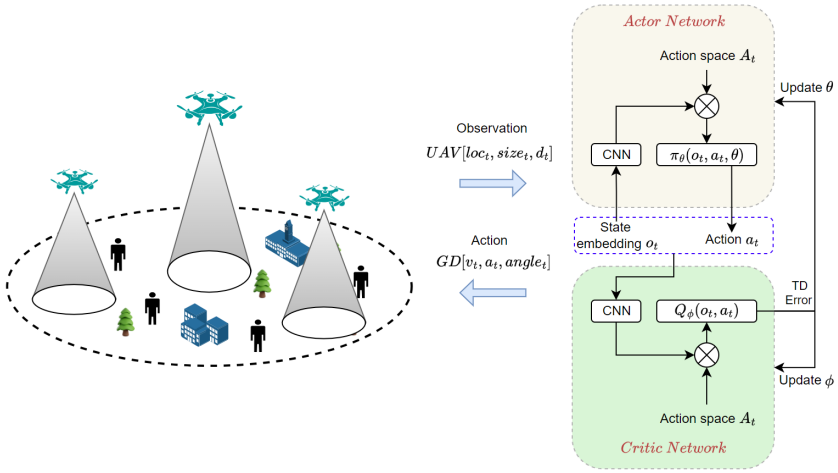


Fig. 4. Framework of AC-based UAV system model.

by $\pi_b(a|s) \propto \pi(a|s) \cdot \pi_h(a|s)$. As training progresses, the influence of the prior control policy on the policy learned by the UAV gradually decreases. Finally, the UAV can independently learn high-performance flight control strategies.

The increasing study of UAVs for target tracking highlights the necessity of ensuring safe tracking in practical applications. To this end, a two-phase method proposed in [131], which combines expert experience and DRL, helps the UAV accomplish its tasks safely. In the first phase, the UAV is trained to learn flight policy based on expert experience obtained through a sample generator combining artificial potential fields and the PID method. In the second phase, the UAV further optimizes the flight policy by screening for superior experiences and employing the Twin-Delayed Deep Deterministic Policy Gradient (TD3) algorithm [34]. Similarly, to achieve safe and efficient tracking of the UAV, the authors in [114] designed a timing controller to enable the UAV to approach the target quickly. Considering the tracking errors caused by environmental disturbances, a novel approach based on the DDPG algorithm is developed to train the UAV to improve its environmental adaptation and obstacle avoidance capability. In addition, the authors introduced a training paradigm that uses old strategies to make decisions based on the current state to improve the training efficiency.

**Coverage.** Full communications coverage allows GDs to be serviced on time and improves the QoS. In [109], UAVs are deployed to provide vehicle coverage in dynamic environments. Considering the limited energy, UAVs can return to charging stations for recharging in this system. To reduce the number of deployed UAVs while improving vehicle coverage rate and energy efficiency, the DDPG algorithm is introduced to find the optimal UAV deployment strategy.

In large-scale scenarios, the coverage task can be well accomplished using multi-UAV communication. For example, [147] establishes a stochastic communication model, in which coverage and position information is transmitted between UAVs and the information map is updated by an information fusion method. Through this model, UAVs can collaborate to efficiently accomplish the area coverage task. In addition, the authors in [154] configured cameras on UAVs for effective surveillance of industries in a smart city. Specifically, the leader UAV can collect information from each UAV and distribute the collected information to the remaining UAVs.

**Performance Optimization.** To improve the QoS of users in an ultra-dense city, [99] proposes a UAV deployment and trajectory design method that combines deep Echo-State Network (ESN) and AC method. For user mobility, the authors used ESN to find the user's movement model and output the coordinates of the next user. After that, a Fast Global K-Means (FGKM) method is applied

to cluster the users into different groups and determine the center of mass of each group as the initial placement location of the UAV. Finally, the UAV is trained using the AC method to find the optimal deployment location and the best trajectory.

The cooperative control of UAVs and unmanned ground vehicles (UGVs) enhances monitoring system efficiency by leveraging their complementary strengths. [74] investigates a method to collaboratively control UAVs and UGVs to monitor an unknown spatiotemporal field, and propose a cooperative model. Specifically, each UAV transmits monitoring information to the UGV, which utilizes this information for trajectory planning. The MPC-based approach is proposed to ensure that UAVs fly within the communication range of the UGV. Considering the problem of information transmission delay and cumulative information constraints, the authors proposed a method based on Multi-Agent Twin-Delayed Deep Deterministic Policy Gradient (MATD3) that allows the UGV to find the optimal trajectory using delay measurements.

To improve the anti-jamming capability of UAVs, [50] introduces a joint optimization problem for the trajectory planning of UAVs and RIS configuration, i.e., maximizing the received data rate of UAVs in the presence of jamming. To achieve this optimization problem, the authors proposed a DRL model based on DDPG and TD3 that enables the UAV to continuously optimize its trajectory and RIS configuration based only on the received data rate.

*3.3.2 Data Collection.* We further categorize the works on AC-based UAV systems for data collection into two parts. The first part focuses on data collection optimization [142]. The second part explores how AC method can improve the freshness of data.

**Data Collection Optimization.** UAVs have been used in data collection scenarios in smart cities and emergency scenarios. However, the presence of obstacles creates a challenge to achieve high-quality data collection. Therefore, researchers employ RL methods to address this challenge. In [81], the authors investigated how to provide high-quality services to GDs in cities through UAV communication. Specifically, the authors divided the deployment environment into a $M \times N$ grid, in which multiple obstacles and mobile data nodes are distributed, and the UAVs need to avoid the obstacles while moving to the data nodes and collecting data. To collect as much data as possible under fairness and energy constraints, a DRL-based method is proposed. The method extracts features through DNN and then helps the UAV to update its strategy through DDPG.

The issue of energy limitations presents significant challenges for UAVs in efficiently performing data collection tasks. To address this, [92] proposes a system based on multi-UAV and pairs of UGVs. In this system, UAVs are responsible for collecting data and charging IoT devices, while UGVs are tasked with recharging the UAVs. The authors proposed a MADDPG-based approach to plan the trajectories of UAVs and UGVs, aiming to minimize the AoI and the total energy consumption.To improve the efficiency of data collection by UAVs, the authors in [159] proposed a DDPG-based method to learn data collection strategies for UAVs. To make the proposed method more relevant, the authors assumed that the number and duration of time slots are not fixed. According to this assumption, the authors proposed a task completion time minimization problem.

**Freshness of Data.** The system in [107] provides a scenario for the use of UAVs in an intelligent transportation system, especially in terms of keeping information fresh. Specifically, this system collects timely data generated by ground vehicles by deploying UAVs on roadways without ground BSs. Because of the ever-changing environment, the authors proposed a DDPG-based UAV control method to optimize the UAV's flight trajectory and data collection strategy to minimize the AoI.

To enhance energy efficiency and AOI, [102] divides UAVs into two teams for collecting data and charging IoT devices. By dividing the work differently, the problem of switching the operation of a single UAV for data collection and power transmission is avoided. A DDPG-based approach is proposed to achieve collaborative control of the two teams, enabling them to charge IoT devices

and collect generated data promptly. The action space of a UAV consists of three components, i.e., the direction of flight of the UAV, the flight distance, and the interval height. The reward function ensures that there is no collision between UAVs by increasing the safe distance penalty.

[38] utilizes UAVs to assist IoT devices in adapting to different environments. Specifically, UAVs learn through a knowledge base and need to sequentially access the IoT device in various environments to adjust its policy. Meanwhile, an AC method is proposed to utilize the energy of the UAV efficiently. This method strikes a balance between AoI and energy consumption by optimizing the allocation of IoT resources. Simulation results validate that the proposed method significantly improves the performance of IoT devices and reduces UAV energy consumption.

*3.3.3 Resource Scheduling.* Due to the existing resource scheduling methods for UAVs still facing lots of drawbacks, the following works focus on how the AC method can perform reasonable resource allocation and achieve higher performance of the systems.

**Channel Allocation and Power Allocation.** The computing resources that UAVs can carry are limited, and it is crucial to allocate them rationally and utilize them efficiently. In [86], the authors focused on how to utilize the remaining resources to access the ground communication network when UAVs perform specific tasks. Considering the power of sensors, the authors used data-uploading clusters, where cluster members periodically summarize the data to the cluster leader, who then uploads them. The members of each cluster are determined using the K-Means method. To maximize data transmission, the authors developed a DQN-based method and a DDPG-based method to update the resource allocation strategies of UAVs promptly.

Considering the limited spectrum resources in IoT networks, [43] introduces a cognitive satellite-over-the-air network. The network shares spectrum resources with UAVs and satellites while giving satellites the primary right to utilize spectrum resources. In addition, the authors introduced the NOMA technique to improve spectrum utilization. To guarantee the QoS for terrestrial users, a MADDPG-based method is proposed to minimize the total delay by optimizing flight trajectories and resource allocation strategies of UAVs. Aiming at the problem of limited resources of GDs in smart agriculture, [59] investigates a UAV-assisted edge computing network. Specifically, GDs can forward part of the monitored and collected data to the associated UAVs. After that, the UAVs choose to perform the tasks locally or offload them to the MEC hosts according to their computing capabilities. Considering the dependency between energy requirement and delay requirement, a method combining Graph Convolutional Neural Network (GCN) and AC method is proposed to find the optimal resource allocation policy.

**Performance Optimization.** UAVs are often used as mobile BSs to provide services to GUs. In this case, QoS is an important metric for measuring the quality of network services. An aerial platform is considered in [64] to provide communication and computing services to remote or disaster areas. Specifically, the airborne platform consists of an aerial platform and UAVs, which allow GDs to offload tasks effectively. The authors treated IoT correlation, task offloading, and resource allocation as a joint optimization problem, aiming to maximize QoS and energy efficiency. Considering the non-convex and complex nature of this optimization problem, the authors proposed a MADDPG-based approach, where high-altitude platforms and UAVs are treated as agents to find the optimal policy. Similarly, the authors in [157] proposed a DDPG-based approach to maximize the average QoS. Specifically, the approach finds the optimal control strategy for the UAV by jointly optimizing the UAV's trajectory, resource allocation, and task offloading strategy. In addition, the loss is minimized by a criterion network and a target network to achieve smoother learning.

[29] studies a downlink transmission network that contains a BS, a RIS-equipped UAV, and multiple GUs. The BS employs the RSMA technique to handle the interference and send out signals, while the UAV amplifies and reflects the signals to the GUs via the RIS. Considering the limited

energy, the authors formulated the resource management problem of the UAV as an optimization problem aiming to maximize the total energy efficiency. For the proposed optimization problem, a SAC-based algorithm is proposed to help UAVs find the optimal resource allocation policy.

*3.3.4 Edge Computing.* We further investigate two application scenarios in AC-based UAV-assisted edge computing systems: intelligent caching [161] and computing offloading [104].

**Intelligent Caching.** The use of UAVs for content caching can provide high-quality services to areas where some communication equipment is lacking, thus meeting the demand for services from ground equipment. In [76], cached UAVs are utilized to provide communication services to GDs in areas lacking infrastructure. To satisfy the service requirements of each type of device while minimizing the system latency, the authors proposed a MADDPG-based approach to jointly optimize the trajectory planning, access sequence, and resource allocation of UAVs. In addition, the method introduces a particle swarm optimization approach to reduce the agent's action space, thus improving the convergence performance. Similarly, [87] meets the service demands of different users by setting multiple service routes. In addition, this work considers the high QoS requirements in VR scenarios. Consequently, the caching policy and offloading policy of the UAV need to be optimized to minimize latency and energy consumption. Considering the constraints of limited delay, energy consumption, and caching, an approach combining Federated Learning (FL) [70] and AC method is proposed to solve the optimization problem.

D2D has gained widespread attention due to its low cost and low latency, and network performance can be effectively improved by utilizing D2D technology. In [130], the authors studied a UAV network that utilizes D2D, in which the user devices are divided into Devices for Requesting (DoRs) and Devices for Caching (DoCs). DoRs send content requests to the UAV or DoCs, and the DoCs cache some of the content. Since some of the requested content is not cached by the UAV or DoCs, it is necessary to obtain and select the cache location from the BS. To address this problem, a DDPG-based approach is proposed to optimize the cache placement policy of the UAV and DoCs, aiming to minimize the file access latency and improve the QoS.

**Computing Offloading.** When monitoring remote areas or disaster relief, the inability of the 5G network to provide full coverage of the area allows IoT devices in the area to perform only a limited number of computations. These IoT devices can't work well when the computing requirements increase. To this end, [16] uses the Space-Air-Ground Integrated Network (SAGIN) technique to provide computation offloading services for IoT devices in remote areas. IoT devices can assign tasks to UAVs and cloud servers, or process these tasks locally. The authors used Virtual Machines (VMs) to represent UAV-assisted edge computing, which requires an efficient resource allocation and task scheduling strategy due to the limited computing resources of UAVs. In the resource allocation of the VM, the authors removed some tasks with the most demanding latency requirements and ensured that the sum of resources obtained by each task did not exceed the limit. In the task allocation process, the authors used the AC method to train IoT devices to minimize the total system cost concerning task latency.

Since there is a bias between the offline environment and the real environment, models trained offline may not work in the real environment well. To minimize this bias, the authors in [73] introduced the GAN technique into the mission offloading scenario for UAVs. Specifically, the GAN is first used to sample data from the environment and generate a simulated environment that closely approximates the real environment. After that, the MATD3 algorithm is employed to train UAVs and GUs within the simulated environment. In this approach, UAVs and GUs are regarded as different agents, where UAVs need to find the optimal flight paths and task assignment strategies and GUs need to determine the proportion of task offloading.

Table 3. Overview of actor-critic algorithms for UAV systems.

| RL Classification | Scenario | Work | RL Algorithm | Performance Indicator | Year |
|---|---|---|---|---|---|
| Actor-Critic | Trajectory Planning | [129] | DDPG | success rate, extra time, extra distance, average speed | 2020 |
| | | [128] | A3C | mean success rate | 2020 |
| | | [152] | MADDPG | average reward, success rate, trapped rate, collision rate, path efficiency | 2023 |
| | | [131] | TD3 | episode reward, episode length, tracking distance, flight speed | 2023 |
| | | [114] | DDPG | reward, relative distances between UAVs and obstacles | 2023 |
| | | [109] | DDPG | average coverage, reward | 2020 |
| | | [147] | SAC | cumulative coverage | 2024 |
| | | [154] | A2C | reward convergence, surveillance, comparison schemes, trained behaviors, computation cost comparison | 2022 |
| | | [99] | AC | average reward, trajectory result | 2022 |
| | | [133] | DDPG | convergence performance, time consumed, energy consumption | 2021 |
| | | [74] | MATD3 | reward, final time, sums of measurements, observability constants, tracking errors | 2024 |
| | | [50] | DDPG, TD3 | received data rate | 2023 |
| | Data Collection | [81] | MADDPG | data collection ratio, energy consumption ratio, geographical fairness, energy efficiency | 2019 |
| | | [142] | MADDPG | speed, data completion percentage, PoI completion percentage, energy efficiency, energy consumption, geographical fairness | 2022 |
| | | [92] | MADDPG | AoI, energy consumption, fairness, throughput, number of active UAVs and UGVs | 2024 |
| | | [159] | DDPG | completion time | 2023 |
| | | [107] | DDPG | expected weighted sum AoI, CDF of the average age | 2020 |
| | | [102] | MADDPG | accumulated reward, average AoI of devices, average throughput, energy efficiency | 2022 |
| | | [38] | AC | average AOI, average reward, energy consumption | 2024 |
| | Resource Scheduling | [86] | DQN, DDPG | cumulative reward, loss, total data transmission | 2022 |
| | | [43] | MADDPG | total transmission latency, convergence properties | 2022 |
| | | [59] | AC | energy consumption, average delay, energy-time cost | 2024 |
| | | [64] | MADDPG | convergence performance, total reward, delay satisfaction ratio, total energy consumption | 2022 |
| | | [157] | DDPG | convergence, average QoS | 2022 |
| | | [29] | SAC | average system energy efficiency | 2024 |
| | Edge Computing | [76] | MADDPG | completion time | 2024 |
| | | [87] | MAAC | delay, energy consumption | 2023 |
| | | [161] | DDPG | average content delivery delay | 2021 |
| | | [130] | DDPG | accumulated reward, total file access latency | 2021 |
| | | [16] | AC, PG | average total delay, convergence, total cost, energy consumption, weighted delay | 2019 |
| | | [104] | AC | average energy consumption, convergence performance | 2023 |
| | | [73] | MATD3 | average task latency, energy consumption | 2024 |

Table 3 presents an overview of the related work that incorporates the AC method, offering more detailed information as the specific RL methods employed and the evaluation metrics selected.

### 3.4 Summary

In this section, we provide an overview of different application scenarios of UAV systems through three different types of RL approaches. From the literature reviewed, it is clear that value-based algorithms are mainly applied in simpler scenarios, such as simple flights within limited areas. In contrast, policy-based algorithms are more effective in handling complex scenarios that involve continuous action spaces. Moreover, AC-based algorithms integrate the advantages of the value-based method and policy-based method, which can quickly adapt to changes in the environment. Overall, RL can offer practical guidance for a wide range of UAV application scenarios.

## 4 RECENT ADVANCEMENTS IN RL-BASED UAV SYSTEMS

RL offers an excellent solution for realizing UAV autonomous control. However, due to the complexity of UAV systems, RL faces many challenges in practical applications. To address these challenges,

many related researchers have proposed various solutions. This section summarizes the related works from four aspects: higher dimensional space, limited observation, dynamic environment, and reward function definition.

### 4.1 Higher Dimensional Space

When UAVs operate in large and intricate environments, they face multidimensional challenges posed by environmental information, decision complexity, and mission requirements. Excessive dimensionality can cause the agent's training process to consume huge computing resources or even become infeasible. In this case, it is crucial to optimize the agent's state and action space.

To cope with the intelligent jamming attacks of UAVs in high-dimensional space, Li et al. [75] embedded domain knowledge directly into the training process of the UAV. This knowledge constrains the exploration of both jammers and UAVs, significantly reducing the state space. Moreover, it accelerates the training process by engaging the UAV to adopt more efficient behavior.

Shurrab et al. [118] proposed an RL model with state space reduction techniques for target localization. Instead of directly adopting the position of the UAV as a state input, the authors utilized the readings from the UAV-equipped sensors to represent the state. Moreover, the authors optimized the dimensionality of the state space by assuming that similar readings are considered to be in the same state. In addition, the UAV's movements are discretized to reduce the dimensionality of the action space. To improve the convergence of RL algorithms, Hosseinzadeh et al. [48] used the QRF intelligent filtering algorithm to filter unnecessary state information. Specifically, the authors assumed that the communication range of UAVs is a spherical region, and UAVs in the region can exchange position information. According to the spherical coordinates of the target UAV, the authors employed the QRF algorithm to narrow down the state space to the UAVs within the sector where the target UAV is located, thus greatly reducing the complexity.

### 4.2 Limited Observation

When performing tasks, such as tracking or surveying, UAVs rely on sensors to observe their surroundings. Due to the limitation of sensor ranging and the influence of environmental obstacles, UAVs usually only observe incomplete information, which leads to instability in the RL training process. Therefore, how to utilize the UAV's local observation information is the key to the decision-making process. Chen et al. [13] assumed that the GUs are equipped with GPS positioning devices to provide accurate position information to UAVs, overcoming the limitations of sensors. In addition, a probabilistic-based LoS model is used to account for dynamic communication links between UAVs and GUs to construct the optimization problem.

Communication between UAVs is another solution for addressing the problem of limited perception. Through establishing communications and sharing information, UAVs can obtain information that cannot be surveyed by themselves, leading to better decision-making. Zhang et al. [163] designed a Graph Vision and Communication (GVis&Comm) framework based on Heterogeneous Graphical Neural Network (HGNN). The graph vision module increases the weight by self-attention to distinguish the number of users that change during processing. The communication module adopts an encoder-decoder structure to handle discrete information exchange between UAVs, thus reducing backpropagation overhead. Specifically, the framework first generates state information through the GVis layer, and then exchanges and aggregates the information with the neighbors through the Comm layer. Finally, the MARL algorithm selects the actions based on the obtained information, and then, updates the policy via the loss function designed by

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{|\mathcal{B}| \cdot N_{\text{UAV}}} \sum_{b=1}^{|\mathcal{B}|} \sum_{i=1}^{N_{\text{UAV}}} \left( \delta_{i,b}^{t} - Q(o_{i,b}^{t+1}, a_{i,b}^{t+1}; \boldsymbol{\theta}) \right)^2, \tag{8}$$

where $\delta_{i,b}^{t}$ denotes the TD error, and $b$ is the sample indices. This collaboration between HGNN and MARL effectively addresses the problem of limited observation of individual UAVs, improving overall system efficiency. Additionally, radio technology can be used to enable communication between UAVs. In [127], UAVs can share their information over a radio interface within a limited communication range to coordinate the exploration of an area and achieve efficient surveillance. Considering the security issue of information, Wang et al. [139] proposed a secure FL framework to enhance the quality and security of shared sensory data among UAVs.

### 4.3 Dynamic Environment

The mobility of UAVs results in UAVs always operating in a dynamic environment, which increases the computational complexity of RL algorithms. Therefore, effectively addressing these challenges is essential to ensure the stability, reliability, and efficiency of the UAV's training process.

*4.3.1 Network Topology Changes.* Network topology changes in UAV systems refer to variations in the connections and communication links between the nodes of the UAV network, usually caused by changes in the state or location of the UAV or mobile terminal. Such changes increase the difficulty for UAVs to adapt to the environment, necessitating the design of flexible and adaptive RL methods.

To address the challenge posed by variations in the number and location of GUs, Jiang et al. [55] developed two architectures: a DNN-based architecture and a DRL-based architecture. The former employs incremental learning to adapt to changes and applies an entropy-checking mechanism to optimize the sample generator globally. The latter is used to train DNNs and introduces additional action refinement to help mobile nodes make real-time decisions.

In [116], the UAV must continuously adjust its position to offer computing offloading services for different GDs, leading to constant changes in the communication links. In this case, the authors defined these network topology changes as part of the state space and guided the UAV's training process by designing an adaptive reward function. In addition, Prioritized Experience Replay (PER) is introduced to enable the UAV to adapt more effectively to changes in the network topology. Similarly, [103] incorporates the location information of mobile users into the state space, allowing the UAV to make decisions based directly on the GUs' location.

*4.3.2 Environmental Condition Changes.* In UAV systems, environmental alterations primarily encompass meteorological variations and shifts in obstacle positioning. These dynamic changes of such external factors complicate the state space and dynamic space of UAVs. Therefore, considering environmental alterations in the optimization process of UAV decision-making enhances UAVs' adaptability to the dynamic environment.

To enable UAVs to avoid flying obstacles successfully, Ma et al. [91] proposed a saliency-based RL method to train UAVs. This method combines saliency techniques with CNNs to extract obstacle features from image data and estimate their locations accurately. Subsequently, the AC method was used to process this obstacle position information and find the optimal obstacle avoidance strategy.

[105] investigates coordinated communication and packet forwarding for UAVs in scenarios with obstacles and wind. The study builds a model of wind direction and speed to help UAVs make more energy-efficient decisions under the influence of wind. Additionally, an obstacle-aware model is proposed to enable UAVs to select unobstructed routes for forwarding data packets. In short, this approach first generates a strategy for UAVs under wind and obstacle interference. Subsequently, Q-Learning is used to evaluate the strategy and optimize it based on a reward function defined as

$$R^{t}\left(s^{t}, a^{t}\right) = -\alpha \cdot \underbrace{\epsilon\left(f^{t}, \hat{f}^{t}\right)}_{\text{residualenergy}} - \beta \cdot \underbrace{\tau\left(cost\left(a^{t}\right)\right)}_{\text{obstacle}}, \tag{9}$$

where the state $s^t$ includes the wind model. $\epsilon\left(f^t, \hat{f}^t\right)$ represents the cost of flight energy and $cost\left(a^t\right)$ is the obstacle-awareness recovery time.

### 4.4 Reward Function Definition

The reward function is a key part of RL to evaluate the performance of actions taken by an agent in a given state. An effective reward function can enhance the overall efficiency and performance of the system. Generally, different reward functions need to be designed based on different scenarios to meet the system requirements. For instance, to shorten the trajectory length in UAV trajectory planning, Jarray et al. [53] defined a reward function that combines static and dynamic rewards. The static reward is determined by the attributes of the current UAV state to help the UAV maintain a certain direction. The dynamic reward is determined by the distance between the UAV and the target to guide the UAV toward the target. In the data collection scenario, Sherman et al. [115] introduced the AoI into the reward function for maximizing the freshness of information. In [155], the authors designed a three-part reward function. Each part of that function acts as a constraint on the UAV's decision-making to optimize the UAV's communication object selection, resource scheduling, and flight trajectory. Additionally, Liu et al. [85] considered communication delay in the assessment of system performance. They reduce the communication delay by introducing a penalty reward in the designed reward function.

In practical communication scenarios, ensuring the security of communications is critical. Wen et al. [144] designed competitive reward functions to maximize the summed secrecy rate of legitimate UAVs and the eavesdropping rate of listening UAVs. The reward functions for the legitimate UAV and the eavesdropping UAV, denoted as $r_1(t)$ and $r_2(t)$ respectively, are defined as follows:

$$r_1(t) = \sum_{k=1}^{K} \lambda_k(t) \left[R_{C,k}(t) - R_{E,k}(t)\right], \tag{10}$$

$$r_2(t) = \sum_{k=1}^{K} \lambda_k(t) R_{E,k}(t), \tag{11}$$

where $\lambda_k(t)$ is a binary variable denoting the communication situation. $R_{C,k}(t)$ and $R_{E,k}(t)$ denote the communication rate and the eavesdropping rate, respectively. Legitimate UAVs and eavesdropping UAVs continuously optimize their strategies through this competitive reward function. The authors in [22] designed two reward functions to maximize the average secrecy rate and minimize the secrecy interruption duration. Based on these reward functions, the authors developed two PPO-based algorithms to optimize the corresponding policies.

### 4.5 Summary

Overall, many studies address the challenges encountered in applying RL algorithms to UAV systems. One group of researchers design clever reward functions to help UAVs adapt to complex environments. Meanwhile, others address the problems encountered in RL-based UAV systems by introducing techniques (e.g., graph neural networks and K-mean method). Table 4 summarizes the recent advancements in applying RL to UAV systems. Although these approaches have yielded effective results, there are still some challenges and unresolved parts of RL-based UAV systems, which we will explore and analyze in the next section.

## 5 FUTURE DIRECTIONS

Although significant progress has been made in RL-based UAV systems, many challenges and opportunities remain. This section provides an overview of potential directions in this area.

Table 4. Overview of recent advancements in RL-based UAV systems

| Challenge | Work | Limitation | Model | Year |
|---|---|---|---|---|
| Higher Dimensional Space | [75] | state-space explosion in anti-jamming UAV systems | Knowledge-Based RL | 2021 |
| | [118] | lack of autonomy, dimensional catastrophe and limited generalization capability in existing UAV target localization methods | DQN | 2023 |
| | [48] | the performance optimization problem of routing algorithms in FANET | Q-Learning+QRF | 2023 |
| Limited Observation | [13] | multi-UAV trajectory design and user association during data transmission in complex environments | CFG+MADRL | 2022 |
| | [163] | UAV perception of limited and dynamic environment for communication process between multiple UAVs and GDs in large-scale complex scenarios | GNN+MARL | 2022 |
| | [127] | collision avoidance in target detection and monitoring of UAV swarms | distributed DQN | 2021 |
| | [139] | privacy and data misuse issues when integrating UAV sensory data | Blockchain + LDP + RL | 2020 |
| Dynamic Environment | [55] | MEC architecture for real-time decision-making in large-scale dynamic scenarios | DNN + DQN | 2020 |
| | [116] | sparse reward, high-dimensional space in dynamic, stochastic and time-varying environments for UAV task offloading | MATD3 + PER | 2024 |
| | [103] | optimal location deployment of UAVs in dynamic environments | AC + DQN | 2023 |
| | [91] | UAV obstacle avoidance under real-time obstacle movement | Saliency Detection + AC | 2018 |
| | [105] | wind and obstacles affecting UAV communications in disaster areas | Heuristic Greedy + A3C | 2023 |
| Reward Function Definition | [53] | learning inefficiency due to the use of fixed reward values during UAV training | Q-Learning | 2023 |
| | [115] | the performance optimization problem for IoT in dynamic environments | Off-Policy DQN + On-Policy PPO | 2023 |
| | [155] | multi-objective optimization problem for UAVs in emergency communication scenarios | CO + SAC | 2023 |
| | [85] | latency and limited resource issues in UAV networks for maritime communications | DRL | 2022 |
| | [144] | information theft by eavesdroppers in UAV communication networks | TDMA + MADDPG | 2022 |
| | [22] | information theft by eavesdroppers in UAV communication networks | SCA + PPO | 2024 |

**Data Sampling in Large-Scale Environments.** RL models usually require many samples to train the agent to adapt to the environment. Consequently, the quality and quantity of data directly affect the policy that the agent learns. In large-scale complex environments, the sampling of UAVs may be restricted due to obstacles and complex terrain, resulting in insufficient samples. To solve this problem, we can adopt a hybrid framework of offline RL and diffusion model [120]. This framework generates a large-scale dataset using the diffusion model. In addition, employing Generative Adversarial Networks (GAN) [39] to generate samples is also an effective scheme.

**Sparse Rewards in RL.** In RL-based UAV systems, UAVs learn optimal strategies by receiving rewards from their surroundings. However, due to the complexity of the environment, UAVs may only receive feedback on an important event or the completion of a task, making it difficult for UAVs to learn the strategy quickly. To solve this problem, additional rewards are usually introduced to guide the UAV's behavior. However, designing a multi-constrained and reliable reward function remains challenging and requires further exploration and research. To this end, we can embed curriculum learning into RL [135],[9],[153]. Based on course learning, the UAV first learns policy using the RL method in simple environments. Then, the UAV utilizes the learned policy to explore complex environments and update the policy. Subsequently, the UAV explores more complex environments until it can adapt to the eventual complex and dynamic environments.

**Cooperative Control between UAVs.** The emergence of MARL provides a feasible solution for UAV swarm collaborative control, and the CTDE framework is the most widely used. The core idea of CTDE is to centralize the information of all agents for strategy learning, but the agents rely only on local information for decision-making. However, the limitation of lack of consensus in the CTDE framework may lead to inconsistent collaborative decision-making of the agents, which may affect the system's performance. Hierarchical consensus-based multi-agent RL (HC-MARL) [30] addresses this limitation by introducing the multi-head attention mechanism, providing a solution to realize cooperative control of UAV swarms. We can also introduce game theory [24] in MARL to analyze the behaviors among agents and design reasonable payoff functions to realize effective collaboration among agents. In addition, developing an efficient and reliable communication mechanism can also allow UAVs to utilize more information for decision-making, thus facilitating cooperation.

**Interpretability of RL.** The goal of explainable RL is to elucidate the decision-making processes of RL agents in sequential decision-making settings [93]. In traditional RL methods, the strategies that agents have learned often suffer from the black-box problem. In certain sensitive domains, such as healthcare and the military, the interpretability and reliability of decisions are crucial. Similarly, in RL-based UAV systems, achieving model interpretability can make agent decisions more transparent and increase user confidence. He et al. [46] used the SHAP [88] value estimation method to explain UAV movements using the DRL model. However, the SHAP method is still superficial and suffers from shortcomings such as a lack of local accuracy. To this end, we can develop a principled interpretable DRL from an early stage, without directly adding existing interpretable modules.

**Simulation to Reality.** Most current research mainly conducts experiments in simulated environments. However, there is usually a significant difference between real and simulated environments. Some studies train agents by introducing randomness in simulated environments, such as domain randomization [125]. Other studies consider hybrid training approaches that combine simulated and real environments, such as transfer learning [143]. We can adopt these approaches to enhance UAVs' generalization capabilities to achieve autonomous control in the real world. In conclusion, as UAV technology evolves, the simulation-to-real problem remains an intriguing direction for further exploration.

**Safety and Privacy.** Currently, few studies have considered UAV systems' security and privacy issues. When UAV systems use wireless communication, they may be subjected to external interference or attack, affecting the system's security and even leaking personal privacy. In addition, outside interferences and attacks may lead to data poisoning, affecting RL's decision-making ability and increasing the risk of malicious exploitation. To solve this problem, a type of RL that incorporates risk-awareness, known as safe RL [36], has been proposed. The safe RL guides the learning process by introducing the risk factor and considers the external factors to prevent abnormal actions of the agent. To protect individual privacy, we can use FL and RL to co-train the model. FL does not require centralized data but builds models based on datasets on multiple distributed devices. RL method can also employ blockchain technology [134], [96] to improve system safety by encrypting the data, ensuring it remains untampered during collection or exchange. In addition, setting a suitable privacy budget for the training process through differential privacy techniques [25], [148] can also be used to limit the attacker's access to the data. Nevertheless, the security and privacy issues of UAV systems issues still need further exploration and research.

**LLM for RL-based UAV Systems.** RL can be essentially viewed as a sequential decision-making process, where the agent takes the appropriate action based on the state at each moment. However, since UAVs operate in large-scale environments, the application of RL algorithms frequently encounters the challenge of sparse rewards. To address this problem, many researchers have begun to combine the Large Language Model (LLM) with RL. The LLM can learn without massive samples, making LLM usable for reward design in RL. In [146], the authors proposed a framework that

combines RL and LLM for application in the field of air combat. In this framework, LLM utilizes a knowledge base to learn to understand complex environments and the rules of the game. After that, LLM predicts the future state of the enemy aircraft based on the input environment information and thus guides the RL model. In addition, the LLM's strong contextual understanding and analytical capabilities enable it to provide valuable suggestions for agent decision-making, e.g., [23], [11]. We can use LLM to generate decisions for UAVs, thereby reducing non-essential exploration. Nowadays, the capabilities of LLM have not been fully embodied in RL-based UAV systems. We believe that integrating LLM with RL can provide a promising direction for the autonomous control of UAVs.

## 6 CONCLUSION

UAV systems serve as an effective technology for accomplishing diverse tasks in complex environments, such as real-time monitoring and data collection. In recent years, RL-based UAV systems have gradually attracted the interest of researchers. RL-based UAV systems can help UAVs autonomously optimize their flight and control strategies, and therefore, they can perform various tasks more accurately than other UAV manipulation methods. In this survey, we present a thorough review of RL-based UAV systems for different scenarios according to the three main categories of RL (i.e., value-based method, policy-based method, and AC method). We analyze the applications of RL in UAV systems in four parts: trajectory planning, data collection, resource scheduling, and edge computing. In addition, We analyze the challenges and recent advancements in applying RL to UAV systems, including high-dimensional space, limited observation, dynamic environment, and reward function definition. Finally, we provide an overview of the issues still existing in the field and suggest possible future directions for its development.

## References


[1] Sarder Fakhrul Abedin, Md. Shirajum Munir, Nguyen H. Tran, Zhu Han, and Choong Seon Hong. 2021. Data Freshness and Energy-Efficient UAV Navigation Optimization: A Deep Reinforcement Learning Approach. *IEEE Transactions on Intelligent Transportation Systems* 22, 9 (2021), 5994–6006. https://doi.org/10.1109/TITS.2020.3039617

[2] Ahmed Al-Hilo, Moataz Samir, Chadi Assi, Sanaa Sharafeddine, and Dariush Ebrahimi. 2021. A cooperative approach for content caching and delivery in UAV-assisted vehicular networks. *Vehicular Communications* 32 (2021), 100391. https://doi.org/10.1016/j.vehcom.2021.100391

[3] Ahmed Al-Hilo, Moataz Samir, Chadi Assi, Sanaa Sharafeddine, and Dariush Ebrahimi. 2021. UAV-Assisted Content Delivery in Intelligent Transportation Systems-Joint Trajectory Planning and Cache Management. *IEEE Transactions on Intelligent Transportation Systems* 22, 8 (2021), 5155–5167. https://doi.org/10.1109/TITS.2020.3020220

[4] Fadi AlMahamid and Katarina Grolinger. 2022. Autonomous Unmanned Aerial Vehicle navigation using Reinforcement Learning: A systematic review. *Engineering Applications of Artificial Intelligence* 115 (2022), 105321. https://doi.org/10.1016/j.engappai.2022.105321

[5] Ebtehal Turki Alotaibi, Shahad Saleh Alqefari, and Anis Koubaa. 2019. LSAR: Multi-UAV Collaboration for Search and Rescue Missions. *IEEE Access* 7 (2019), 55817–55832. https://doi.org/10.1109/ACCESS.2019.2912306

[6] Stephen Anokye, Daniel Ayepah-Mensah, Abegaz Mohammed Seid, Gordon Owusu Boateng, and Guolin Sun. 2022. Deep Reinforcement Learning-Based Mobility-Aware UAV Content Caching and Placement in Mobile Edge Networks. *IEEE Systems Journal* 16, 1 (2022), 275–286. https://doi.org/10.1109/JSYST.2021.3082837

[7] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine* 34, 6 (2017), 26–38. https://doi.org/10.1109/MSP.2017.2743240

[8] Yu Bai, Hui Zhao, Xin Zhang, Zheng Chang, Riku Jäntti, and Kun Yang. 2023. Toward Autonomous Multi-UAV Wireless Network: A Survey of Reinforcement Learning-Based Approaches. *IEEE Communications Surveys & Tutorials* 25, 4 (2023), 3038–3067. https://doi.org/10.1109/COMST.2023.3323344

[9] Zhenshan Bing, Hongkuan Zhou, Rui Li, Xiaojie Su, Fabrice O. Morin, Kai Huang, and Alois Knoll. 2023. Solving Robotic Manipulation With Sparse Reward Reinforcement Learning Via Graph-Based Diversity and Proximity. *IEEE Transactions on Industrial Electronics* 70, 3 (2023), 2759–2769. https://doi.org/10.1109/TIE.2022.3172754

[10] S. Bouabdallah and R. Siegwart. 2005. Backstepping and Sliding-mode Techniques Applied to an Indoor Micro Quadrotor. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. 2247–2252. https:

//doi.org/10.1109/ROBOT.2005.1570447

[11] Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 150, 38 pages.

[12] Che Chen, Shimin Gong, Wenjie Zhang, Yifeng Zheng, and Yeo Chai Kiat. 2024. DRL-Based Contract Incentive for Wireless-Powered and UAV-Assisted Backscattering MEC System. *IEEE Transactions on Cloud Computing* 12, 1 (2024), 264–276. https://doi.org/10.1109/TCC.2024.3360443

[13] Gong Chen, Xiangping Bryce Zhai, and Congduan Li. 2023. Joint Optimization of Trajectory and User Association via Reinforcement Learning for UAV-Aided Data Collection in Wireless Networks. *IEEE Transactions on Wireless Communications* 22, 5 (2023), 3128–3143. https://doi.org/10.1109/TWC.2022.3216049

[14] Hao Chen, Xiaoqi Qin, Yixuan Li, and Nan Ma. 2022. Energy-aware Path Planning for Obtaining Fresh Updates in UAV-IoT MEC systems. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. 1791–1796. https://doi.org/10.1109/WCNC51071.2022.9771867

[15] Yu-Jia Chen, Kai-Min Liao, Meng-Lin Ku, Fung Po Tso, and Guan-Yi Chen. 2021. Multi-Agent Reinforcement Learning Based 3D Trajectory Design in Aerial-Terrestrial Wireless Caching Networks. *IEEE Transactions on Vehicular Technology* 70, 8 (2021), 8201–8215. https://doi.org/10.1109/TVT.2021.3094273

[16] Nan Cheng, Feng Lyu, Wei Quan, Conghao Zhou, Hongli He, Weisen Shi, and Xuemin Shen. 2019. Space/Aerial-Assisted Computing Offloading for IoT Applications: A Learning-Based Approach. *IEEE Journal on Selected Areas in Communications* 37, 5 (2019), 1117–1129. https://doi.org/10.1109/JSAC.2019.2906789

[17] Nesrine Cherif, Wael Jaafar, Halim Yanikomeroglu, and Abbas Yongacoglu. 2024. RL-Based Cargo-UAV Trajectory Planning and Cell Association for Minimum Handoffs, Disconnectivity, and Energy Consumption. *IEEE Transactions on Vehicular Technology* 73, 5 (2024), 7304–7309. https://doi.org/10.1109/TVT.2023.3340177

[18] Chen Dai, Kun Zhu, and Ekram Hossain. 2023. Multi-Agent Deep Reinforcement Learning for Joint Decoupled User Association and Trajectory Design in Full-Duplex Multi-UAV Networks. *IEEE Transactions on Mobile Computing* 22, 10 (2023), 6056–6070. https://doi.org/10.1109/TMC.2022.3188473

[19] Xunhua Dai, Zhiyu Lu, Xuehan Chen, Xinyi Xu, and Fengxiao Tang. 2024. Multiagent RL-Based Joint Trajectory Scheduling and Resource Allocation in NOMA-Assisted UAV Swarm Network. *IEEE Internet of Things Journal* 11, 8 (2024), 14153–14167. https://doi.org/10.1109/JIOT.2023.3340669

[20] Zipeng Dai, Chi Harold Liu, Rui Han, Guoren Wang, Kin K. Leung, and Jian Tang. 2023. Delay-Sensitive Energy-Efficient UAV Crowdsensing by Deep Reinforcement Learning. *IEEE Transactions on Mobile Computing* 22, 4 (2023), 2038–2052. https://doi.org/10.1109/TMC.2021.3113052

[21] Qi Dang, Qimei Cui, Zhenzhen Gong, Xuefei Zhang, Xueqing Huang, and Xiaofeng Tao. 2022. AoI Oriented UAV Trajectory Planning in Wireless Powered IoT Networks. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. 884–889. https://doi.org/10.1109/WCNC51071.2022.9771588

[22] Runze Dong, Buhong Wang, Kunrui Cao, Jiwei Tian, and Tianhao Cheng. 2024. Secure Transmission Design of RIS Enabled UAV Communication Networks Exploiting Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 73, 6 (2024), 8404–8419. https://doi.org/10.1109/TVT.2024.3357821

[23] Yuqing Du, Olivia Watkins, Zihan Wang, Cédric Colas, Trevor Darrell, Pieter Abbeel, Abhishek Gupta, and Jacob Andreas. 2023. Guiding pretraining in reinforcement learning with large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*. JMLR.org, Article 346, 21 pages.

[24] Martin Dufwenberg. 2011. Game theory. *WIREs Cognitive Science* 2, 2 (2011), 167–173. https://doi.org/10.1002/wcs.119 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcs.119

[25] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12. https://doi.org/10.1007/11787006_1

[26] Dariush Ebrahimi, Sanaa Sharafeddine, Pin-Han Ho, and Chadi Assi. 2021. Autonomous UAV Trajectory for Localizing Ground Objects: A Reinforcement Learning Approach. *IEEE Transactions on Mobile Computing* 20, 4 (2021), 1312–1324. https://doi.org/10.1109/TMC.2020.2966989

[27] Muhammad Ejaz, Jinsong Gui, Muhammad Asim, Mohammed A. El-Affendi, Carol Fung, and Ahmed A. Abd El-Latif. 2024. RL-Planner: Reinforcement Learning-Enabled Efficient Path Planning in Multi-UAV MEC Systems. *IEEE Transactions on Network and Service Management* 21, 3 (2024), 3317–3329. https://doi.org/10.1109/TNSM.2024.3378677

[28] Yousef Emami, Hao Gao, Kai Li, Luis Almeida, Eduardo Tovar, and Zhu Han. 2024. Age of Information Minimization Using Multi-Agent UAVs Based on AI-Enhanced Mean Field Resource Allocation. *IEEE Transactions on Vehicular Technology* 73, 9 (2024), 13368–13380. https://doi.org/10.1109/TVT.2024.3394235

[29] Sajad Faramarzi, Sepideh Javadi, Farshad Zeinali, Hosein Zarini, Mohammad Robat Mili, Mehdi Bennis, Yonghui Li, and Kai-Kit Wong. 2024. Meta Reinforcement Learning for Resource Allocation in Aerial Active-RIS-Assisted Networks With Rate-Splitting Multiple Access. *IEEE Internet of Things Journal* 11, 15 (2024), 26366–26383. https://doi.org/10.1109/JIOT.2024.3397007

[30] Pu Feng, Junkang Liang, Size Wang, Xin Yu, Xin Ji, Yiting Chen, Kui Zhang, Rongye Shi, and Wenjun Wu. 2024. Hierarchical Consensus-Based Multi-Agent Reinforcement Learning for Multi-Robot Cooperation Tasks. arXiv:2407.08164 [cs.AI] https://arxiv.org/abs/2407.08164

[31] Zikai Feng, Mengxing Huang, Di Wu, Edmond Q. Wu, and Chau Yuen. 2023. Multi-Agent Reinforcement Learning With Policy Clipping and Average Evaluation for UAV-Assisted Communication Markov Game. *IEEE Transactions on Intelligent Transportation Systems* 24, 12 (2023), 14281–14293. https://doi.org/10.1109/TITS.2023.3296769

[32] Shu Fu, Xue Feng, Ajmery Sultana, and Lian Zhao. 2024. Joint Power Allocation and 3D Deployment for UAV-BSs: A Game Theory Based Deep Reinforcement Learning Approach. *IEEE Transactions on Wireless Communications* 23, 1 (2024), 736–748. https://doi.org/10.1109/TWC.2023.3281812

[33] Shu Fu, Yujie Tang, Yuan Wu, Ning Zhang, Huaxi Gu, Chen Chen, and Min Liu. 2021. Energy-Efficient UAV-Enabled Data Collection via Wireless Charging: A Reinforcement Learning Approach. *IEEE Internet of Things Journal* 8, 12 (2021), 10209–10219. https://doi.org/10.1109/JIOT.2021.3051370

[34] Scott Fujimoto, Herke van Hoof, and David Meger. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 1587–1596.

[35] Yunfei Gao, Xiaopeng Yuan, Dingcheng Yang, Yulin Hu, Yue Cao, and Anke Schmeink. 2024. UAV-Assisted MEC System With Mobile Ground Terminals: DRL-Based Joint Terminal Scheduling and UAV 3D Trajectory Design. *IEEE Transactions on Vehicular Technology* 73, 7 (2024), 10164–10180. https://doi.org/10.1109/TVT.2024.3367624

[36] Javier García and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* 16, 1 (Jan. 2015), 1437–1480.

[37] Benmeziane Imad-Ddine Ghomri, Mohammed Yassine Bendimerad, and Fethi Tarik Bendimerad. 2024. DRL-Driven Optimization for Energy Efficiency and Fairness in NOMA-UAV Networks. *IEEE Communications Letters* 28, 5 (2024), 1048–1052. https://doi.org/10.1109/LCOMM.2024.3377005

[38] Zhenzhen Gong, Omar Hashash, Yingze Wang, Qimei Cui, Wei Ni, Walid Saad, and Kei Sakaguchi. 2024. UAV-Aided Lifelong Learning for AoI and Energy Optimization in Non-Stationary IoT Networks. *IEEE Internet of Things Journal* (2024), 1–1. https://doi.org/10.1109/JIOT.2024.3406220

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (Oct. 2020), 139–144. https://doi.org/10.1145/3422622

[40] Yue Guan, Sai Zou, Haixia Peng, Wei Ni, Yanglong Sun, and Hongfeng Gao. 2024. Cooperative UAV Trajectory Design for Disaster Area Emergency Communications: A Multiagent PPO Method. *IEEE Internet of Things Journal* 11, 5 (2024), 8848–8859. https://doi.org/10.1109/JIOT.2023.3320796

[41] Delin Guo, Lan Tang, Xinggan Zhang, and Ying-Chang Liang. 2024. Joint Optimization of Trajectory and Jamming Power for Multiple UAV-Aided Proactive Eavesdropping. *IEEE Transactions on Mobile Computing* 23, 5 (2024), 5770–5785. https://doi.org/10.1109/TMC.2023.3311484

[42] Qi Guo, Fengxiao Tang, and Nei Kato. 2023. Resource Allocation for Aerial Assisted Digital Twin Edge Mobile Network. *IEEE Journal on Selected Areas in Communications* 41, 10 (2023), 3070–3079. https://doi.org/10.1109/JSAC.2023.3310065

[43] Shaoai Guo and Xiaohui Zhao. 2023. Multi-Agent Deep Reinforcement Learning Based Transmission Latency Minimization for Delay-Sensitive Cognitive Satellite-UAV Networks. *IEEE Transactions on Communications* 71, 1 (2023), 131–144. https://doi.org/10.1109/TCOMM.2022.3222460

[44] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. arXiv:1801.01290 [cs.LG] https://arxiv.org/abs/1801.01290

[45] Hado van Hasselt, Arthur Guez, and David Silver. 2016. Deep reinforcement learning with double Q-Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. AAAI Press, 2094–2100.

[46] Lei He, Nabil Aouf, and Bifeng Song. 2021. Explainable Deep Reinforcement Learning for UAV autonomous path planning. *Aerospace Science and Technology* 118 (2021), 107052. https://doi.org/10.1016/j.ast.2021.107052

[47] Tai Manh Ho, Kim-Khoa Nguyen, and Mohamed Cheriet. 2021. UAV Control for Wireless Service Provisioning in Critical Demand Areas: A Deep Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology* 70, 7 (2021), 7138–7152. https://doi.org/10.1109/TVT.2021.3088129

[48] Mehdi Hosseinzadeh, Saqib Ali, Liliana Ionescu-Feleaga, Bogdan-Stefan Ionescu, Mohammad Sadegh Yousefpoor, Efat Yousefpoor, Omed Hassan Ahmed, Amir Masoud Rahmani, and Asif Mehmood. 2023. A novel Q-learning-based routing scheme using an intelligent filtering algorithm for flying ad hoc networks (FANETs). *Journal of King Saud University - Computer and Information Sciences* 35, 10 (2023), 101817. https://doi.org/10.1016/j.jksuci.2023.101817

[49] Yu-Hsin Hsu and Rung-Hung Gau. 2022. Reinforcement Learning-Based Collision Avoidance and Optimal Trajectory Planning in UAV Communication Networks. *IEEE Transactions on Mobile Computing* 21, 1 (2022), 306–320. https://doi.org/10.1109/TMC.2020.3003639

[50] Shuyan Hu, Xin Yuan, Wei Ni, Xin Wang, and Abbas Jamalipour. 2023. RIS-Assisted Jamming Rejection and Path Planning for UAV-Borne IoT Platform: A New Deep Reinforcement Learning Framework. *IEEE Internet of Things Journal* 10, 22 (2023), 20162–20173. https://doi.org/10.1109/JIOT.2023.3283502

[51] Hongji Huang, Yuchun Yang, Hong Wang, Zhiguo Ding, Hikmet Sari, and Fumiyuki Adachi. 2020. Deep Reinforcement Learning for UAV Navigation Through Massive MIMO Technique. *IEEE Transactions on Vehicular Technology* 69, 1 (2020), 1117–1121. https://doi.org/10.1109/TVT.2019.2952549

[52] Shafkat Islam, Shahriar Badsha, Ibrahim Khalil, Mohammed Atiquzzaman, and Charalambos Konstantinou. 2023. A Triggerless Backdoor Attack and Defense Mechanism for Intelligent Task Offloading in Multi-UAV Systems. *IEEE Internet of Things Journal* 10, 7 (2023), 5719–5732. https://doi.org/10.1109/JIOT.2022.3172936

[53] Raja Jarray and Soufiene Bouallègue. 2023. Reinforcement Learning-Based Path Planning Approach for Unmanned Arial Vehicles. In *2023 IEEE International Conference on Artificial Intelligence & Green Energy (ICAIGE)*. 1–6. https://doi.org/10.1109/ICAIGE58321.2023.10346406

[54] Jiequ Ji, Kun Zhu, and Lin Cai. 2023. Trajectory and Communication Design for Cache- Enabled UAVs in Cellular Networks: A Deep Reinforcement Learning Approach. *IEEE Transactions on Mobile Computing* 22, 10 (2023), 6190–6204. https://doi.org/10.1109/TMC.2022.3181308

[55] Feibo Jiang, Kezhi Wang, Li Dong, Cunhua Pan, Wei Xu, and Kun Yang. 2021. AI Driven Heterogeneous MEC System with UAV Assistance for Dynamic Environment: Challenges and Solutions. *IEEE Network* 35, 1 (2021), 400–408. https://doi.org/10.1109/MNET.011.2000440

[56] Zhiling Jiang, Yining Chen, Ke Wang, Bowei Yang, and Guanghua Song. 2023. A Graph-Based PPO Approach in Multi-UAV Navigation for Communication Coverage. *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL* 18, 6 (DEC 2023). https://doi.org/10.15837/ijccc.2023.6.5505

[57] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W. Moore. 1996. Reinforcement Learning: A Survey. *J. Artif. Intell. Res.* 4 (1996), 237–285.

[58] Hongyue Kang, Xiaolin Chang, Jelena Mišić, Vojislav B. Mišić, Junchao Fan, and Yating Liu. 2023. Cooperative UAV Resource Allocation and Task Offloading in Hierarchical Aerial Computing Systems: A MAPPO-Based Approach. *IEEE Internet of Things Journal* 10, 12 (2023), 10497–10509. https://doi.org/10.1109/JIOT.2023.3240173

[59] Fahime Khoramnejad, Aisha Syed, W. Sean Kennedy, and Melike Erol-Kantarci. 2024. Energy and Delay Aware General Task Dependent Offloading in UAV-Aided Smart Farms. *IEEE Transactions on Network and Service Management* 21, 5 (2024), 5033–5048. https://doi.org/10.1109/TNSM.2024.3391664

[60] Eunjin Kim, Junsu Kim, Jae-Hyun Kim, and Howon Lee. 2024. HiMAQ: Hierarchical multi-agent Q-learning-based throughput and fairness improvement for UAV-Aided IoT networks. *Journal of Network and Computer Applications* 223 (2024), 103813. https://doi.org/10.1016/j.jnca.2023.103813

[61] Sungmo Ku, Sangwon Jung, and Chungyoung Lee. 2019. UAV Trajectory Design Based on Reinforcement Learning for Wireless Power Transfer. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. 1–3. https://doi.org/10.1109/ITC-CSCC.2019.8793294

[62] Harrison Kurunathan, Hailong Huang, Kai Li, Wei Ni, and Ekram Hossain. 2024. Machine Learning-Aided Operations and Communications of Unmanned Aerial Vehicles: A Contemporary Survey. *IEEE Communications Surveys & Tutorials* 26, 1 (2024), 496–533. https://doi.org/10.1109/COMST.2023.3312221

[63] Mohamed-Amine Lahmeri, Mustafa A. Kishk, and Mohamed-Slim Alouini. 2021. Artificial Intelligence for UAV-Enabled Wireless Networks: A Survey. *IEEE Open Journal of the Communications Society* 2 (2021), 1015–1040. https://doi.org/10.1109/OJCOMS.2021.3075201

[64] Demeke Shumeye Lakew, Anh-Tien Tran, Nhu-Ngoc Dao, and Sungrae Cho. 2023. Intelligent Offloading and Resource Allocation in Heterogeneous Aerial Access IoT Networks. *IEEE Internet of Things Journal* 10, 7 (2023), 5704–5718. https://doi.org/10.1109/JIOT.2022.3161571

[65] Seungmin Lee, Heejung Yu, and Howon Lee. 2022. Multiagent Q-Learning-Based Multi-UAV Wireless Networks for Maximizing Energy Efficiency: Deployment and Power Control Strategy Design. *IEEE Internet of Things Journal* 9, 9 (2022), 6434–6442. https://doi.org/10.1109/JIOT.2021.3113128

[66] Jiayi Lei, Tiankui Zhang, Xidong Mu, and Yuanwei Liu. 2024. NOMA for STAR-RIS Assisted UAV Networks. *IEEE Transactions on Communications* 72, 3 (2024), 1732–1745. https://doi.org/10.1109/TCOMM.2023.3333880

[67] Jiandong Li, Chengyi Zhou, Junyu Liu, Min Sheng, Nan Zhao, and Yu Su. 2024. Reinforcement Learning-Based Resource Allocation for Coverage Continuity in High Dynamic UAV Communication Networks. *IEEE Transactions on Wireless Communications* 23, 2 (2024), 848–860. https://doi.org/10.1109/TWC.2023.3282909

[68] Kai Li, Wei Ni, Eduardo Tovar, and Abbas Jamalipour. 2019. On-Board Deep Q-Network for UAV-Assisted Online Power Transfer and Data Collection. *IEEE Transactions on Vehicular Technology* 68, 12 (2019), 12215–12226. https://doi.org/10.1109/TVT.2019.2945037

[69] Li Li, Wei Li, Jun Wang, Xiaonan Chen, Qihang Peng, and Wei Huang. 2023. UAV Trajectory Optimization for Spectrum Cartography: A PPO Approach. *IEEE Communications Letters* 27, 6 (2023), 1575–1579. https://doi.org/10.

1109/LCOMM.2023.3265214
[70] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine* 37, 3 (2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749
[71] Xuanheng Li, Sike Cheng, Haichuan Ding, Miao Pan, and Nan Zhao. 2023. When UAVs Meet Cognitive Radio: Offloading Traffic Under Uncertain Spectrum Environment via Deep Reinforcement Learning. *IEEE Transactions on Wireless Communications* 22, 2 (2023), 824–838. https://doi.org/10.1109/TWC.2022.3198665
[72] Xuan Li, Qiang Wang, Jie Liu, and Wenqi Zhang. 2020. Trajectory Design and Generalization for UAV Enabled Networks:A Deep Reinforcement Learning Approach. In *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. 1–6. https://doi.org/10.1109/WCNC45663.2020.9120668
[73] Yangyang Li, Lei Feng, Yang Yang, and Wenjing Li. 2024. GAN-powered heterogeneous multi-agent reinforcement learning for UAV-assisted task offloading. *Ad Hoc Networks* 153 (2024), 103341. https://doi.org/10.1016/j.adhoc.2023.103341
[74] Zhuo Li, Yunlong Guo, Gang Wang, Jian Sun, and Keyou You. 2024. Informative Trajectory Planning for Air-Ground Cooperative Monitoring of Spatiotemporal Fields. *IEEE Transactions on Automation Science and Engineering* (2024), 1–12. https://doi.org/10.1109/TASE.2024.3382730
[75] Zhiwei Li, Yu Lu, Xi Li, Zengguang Wang, Wenxin Qiao, and Yicen Liu. 2021. UAV Networks Against Multiple Maneuvering Smart Jamming With Knowledge-Based Reinforcement Learning. *IEEE Internet of Things Journal* 8, 15 (2021), 12289–12310. https://doi.org/10.1109/JIOT.2021.3062659
[76] Zewu Li, Chen Xu, Zhanpeng Zhang, and Runze Wu. 2024. Deep reinforcement learning based trajectory design and resource allocation for task-aware multi-UAV enabled MEC networks. *Computer Communications* 213 (2024), 88–98. https://doi.org/10.1016/j.comcom.2023.11.006
[77] Chengqing Liang, Lei Liu, and Chen Liu. 2023. Multi-UAV autonomous collision avoidance based on PPO-GIC algorithm with CNN–LSTM fusion network. *Neural Networks* 162 (2023), 21–33. https://doi.org/10.1016/j.neunet.2023.02.027
[78] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2019. Continuous control with deep reinforcement learning. (2019). arXiv:1509.02971 [cs.LG] https://arxiv.org/abs/1509.02971
[79] Yuanguo Lin, Yong Liu, Fan Lin, Lixin Zou, Pengcheng Wu, Wenhua Zeng, Huanhuan Chen, and Chunyan Miao. 2024. A Survey on Reinforcement Learning for Recommender Systems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 10 (2024), 13164–13184. https://doi.org/10.1109/TNNLS.2023.3280161
[80] Björn Lindqvist, Sina Sharif Mansouri, Ali-akbar Agha-mohammadi, and George Nikolakopoulos. 2020. Nonlinear MPC for Collision Avoidance and Control of UAVs With Dynamic Obstacles. *IEEE Robotics and Automation Letters* 5, 4 (2020), 6001–6008. https://doi.org/10.1109/LRA.2020.3010730
[81] Chi Harold Liu, Zheyu Chen, and Yufeng Zhan. 2019. Energy-Efficient Distributed Mobile Crowd Sensing: A Deep Learning Approach. *IEEE Journal on Selected Areas in Communications* 37, 6 (2019), 1262–1276. https://doi.org/10.1109/JSAC.2019.2904353
[82] Lingshan Liu, Ke Xiong, Jie Cao, Yang Lu, Pingyi Fan, and Khaled Ben Letaief. 2022. Average AoI Minimization in UAV-Assisted Data Collection With RF Wireless Power Transfer: A Deep Reinforcement Learning Scheme. *IEEE Internet of Things Journal* 9, 7 (2022), 5216–5228. https://doi.org/10.1109/JIOT.2021.3110138
[83] Run Liu, Zhenzhe Qu, Guosheng Huang, Mianxiong Dong, Tian Wang, Shaobo Zhang, and Anfeng Liu. 2023. DRL-UTPS: DRL-Based Trajectory Planning for Unmanned Aerial Vehicles for Data Collection in Dynamic IoT Network. *IEEE Transactions on Intelligent Vehicles* 8, 2 (2023), 1204–1218. https://doi.org/10.1109/TIV.2022.3213703
[84] Xiao Liu, Yuanwei Liu, and Yue Chen. 2019. Reinforcement Learning in Multiple-UAV Networks: Deployment and Movement Design. *IEEE Transactions on Vehicular Technology* 68, 8 (2019), 8036–8049. https://doi.org/10.1109/TVT.2019.2922849
[85] Ying Liu, Junjie Yan, and Xiaohui Zhao. 2022. Deep Reinforcement Learning Based Latency Minimization for Mobile Edge Computing With Virtualization in Maritime UAV Communication Network. *IEEE Transactions on Vehicular Technology* 71, 4 (2022), 4225–4236. https://doi.org/10.1109/TVT.2022.3141799
[86] Yitong Liu, Junjie Yan, and Xiaohui Zhao. 2022. Deep-Reinforcement-Learning-Based Optimal Transmission Policies for Opportunistic UAV-Aided Wireless Sensor Network. *IEEE Internet of Things Journal* 9, 15 (2022), 13823–13836. https://doi.org/10.1109/JIOT.2022.3142269
[87] Zhikai Liu, Navneet Garg, and Tharmalingam Ratnarajah. 2024. Multi-Agent Federated Reinforcement Learning Strategy for Mobile Virtual Reality Delivery Networks. *IEEE Transactions on Network Science and Engineering* 11, 1 (2024), 100–114. https://doi.org/10.1109/TNSE.2023.3292570
[88] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. (2017), 4768–4777.
[89] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. 2019. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Communications Surveys & Tutorials* 21, 4 (2019), 3133–3174. https://doi.org/10.1109/COMST.2019.2916583

[90] Phuong Luong, François Gagnon, Le-Nam Tran, and Fabrice Labeau. 2021. Deep Reinforcement Learning-Based Resource Allocation in Cooperative UAV-Assisted Wireless Networks. *IEEE Transactions on Wireless Communications* 20, 11 (2021), 7610–7625. https://doi.org/10.1109/TWC.2021.3086503

[91] Zhaowei Ma, Chang Wang, Yifeng Niu, Xiangke Wang, and Lincheng Shen. 2018. A saliency-based reinforcement learning approach for a UAV to avoid flying obstacles. *Robotics and Autonomous Systems* 100 (2018), 108–118. https://doi.org/10.1016/j.robot.2017.10.009

[92] Kaddour Messaoudi, Abdullah Baz, Omar Sami Oubbati, Abderrezak Rachedi, Tahar Bendouma, and Mohammed Atiquzzaman. 2024. UGV Charging Stations for UAV-Assisted AoI-Aware Data Collection. *IEEE Transactions on Cognitive Communications and Networking* (2024), 1–1. https://doi.org/10.1109/TCCN.2024.3394859

[93] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2024. Explainable Reinforcement Learning: A Survey and Comparative Review. *ACM Comput. Surv.* 56, 7, Article 168 (April 2024), 36 pages. https://doi.org/10.1145/3616864

[94] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 1928–1937.

[95] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (Feb. 2015), 529–533. https://doi.org/10.1038/nature14236

[96] Abegaz Mohammed Seid, Aiman Erbad, Hayla Nahom Abishu, Abdullatif Albaseer, Mohamed Abdallah, and Mohsen Guizani. 2023. Blockchain-Empowered Resource Allocation in Multi-UAV-Enabled 5G-RAN: A Multi-Agent Deep Reinforcement Learning Approach. *IEEE Transactions on Cognitive Communications and Networking* 9, 4 (2023), 991–1011. https://doi.org/10.1109/TCCN.2023.3262242

[97] Zhiyu Mou, Yu Zhang, Feifei Gao, Huangang Wang, Tao Zhang, and Zhu Han. 2021. Deep Reinforcement Learning Based Three-Dimensional Area Coverage With UAV Swarm. *IEEE Journal on Selected Areas in Communications* 39, 10 (2021), 3160–3176. https://doi.org/10.1109/JSAC.2021.3088718

[98] Ahmadun Nabi, Tanmay Baidya, and Sangman Moh. 2024. Comprehensive survey on reinforcement learning-based task offloading techniques in aerial edge computing. *Internet of Things* 28 (2024), 101342. https://doi.org/10.1016/j.iot.2024.101342

[99] Maedeh Nasr-Azadani, Jamshid Abouei, and Konstantinos N. Plataniotis. 2022. Single- and Multiagent Actor–Critic for Initial UAV's Deployment and 3-D Trajectory Design. *IEEE Internet of Things Journal* 9, 16 (2022), 15372–15389. https://doi.org/10.1109/JIOT.2022.3150184

[100] Khoi Khac Nguyen, Saeed R. Khosravirad, Daniel Benevides da Costa, Long D. Nguyen, and Trung Q. Duong. 2022. Reconfigurable Intelligent Surface-Assisted Multi-UAV Networks: Efficient Resource Allocation With Deep Reinforcement Learning. *IEEE Journal of Selected Topics in Signal Processing* 16, 3 (2022), 358–368. https://doi.org/10.1109/JSTSP.2021.3134162

[101] Khoi Khac Nguyen, Antonino Masaracchia, Vishal Sharma, H. Vincent Poor, and Trung Q. Duong. 2022. RIS-Assisted UAV Communications for IoT With Wireless Power Transfer Using Deep Reinforcement Learning. *IEEE Journal of Selected Topics in Signal Processing* 16, 5 (2022), 1086–1096. https://doi.org/10.1109/JSTSP.2022.3172587

[102] Omar Sami Oubbati, Mohammed Atiquzzaman, Hyotaek Lim, Abderrezak Rachedi, and Abderrahmane Lakas. 2022. Synchronizing UAV Teams for Timely Data Collection and Energy Transfer by Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 71, 6 (2022), 6682–6697. https://doi.org/10.1109/TVT.2022.3165227

[103] Nahid Parvaresh and Burak Kantarci. 2023. 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 Journal of the Communications Society* 4 (2023), 700–712. https://doi.org/10.1109/OJCOMS.2023.3251297

[104] Peng Qin, Shuo Wang, Zhou Lu, Yuanbo Xie, and Xiongwen Zhao. 2023. Deep Reinforcement Learning-Based Energy Minimization Task Offloading and Resource Allocation for Air Ground Integrated Heterogeneous Networks. *IEEE Systems Journal* 17, 3 (2023), 4958–4968. https://doi.org/10.1109/JSYST.2023.3266769

[105] Chengyi Qu, Francesco Betti Sorbelli, Rounak Singh, Prasad Calyam, and Sajal K. Das. 2023. Environmentally-Aware and Energy-Efficient Multi-Drone Coordination and Networking for Disaster Response. *IEEE Transactions on Network and Service Management* 20, 2 (2023), 1093–1109. https://doi.org/10.1109/TNSM.2023.3243543

[106] Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Thomas Lagkas, and Ioannis Moscholios. 2020. A compilation of UAV applications for precision agriculture. *Computer Networks* 172 (2020), 107148. https://doi.org/10.1016/j.comnet.2020.107148

[107] Moataz Samir, Chadi Assi, Sanaa Sharafeddine, Dariush Ebrahimi, and Ali Ghrayeb. 2020. Age of Information Aware Trajectory Planning of UAVs in Intelligent Transportation Systems: A Deep Learning Approach. *IEEE Transactions on Vehicular Technology* 69, 11 (2020), 12382–12395. https://doi.org/10.1109/TVT.2020.3023861

[108] Moataz Samir, Chadi Assi, Sanaa Sharafeddine, and Ali Ghrayeb. 2022. Online Altitude Control and Scheduling Policy for Minimizing AoI in UAV-Assisted IoT Wireless Networks. *IEEE Transactions on Mobile Computing* 21, 7 (2022), 2493–2505. https://doi.org/10.1109/TMC.2020.3042925

[109] Moataz Samir, Dariush Ebrahimi, Chadi Assi, Sanaa Sharafeddine, and Ali Ghrayeb. 2021. Leveraging UAVs for Coverage in Cell-Free Vehicular Networks: A Deep Reinforcement Learning Approach. *IEEE Transactions on Mobile Computing* 20, 9 (2021), 2835–2847. https://doi.org/10.1109/TMC.2020.2991326

[110] Vidit Saxena, Joakim Jaldén, and Henrik Klessig. 2019. Optimal UAV Base Station Trajectories Using Flow-Level Models for Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking* 5, 4 (2019), 1101–1112. https://doi.org/10.1109/TCCN.2019.2948324

[111] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15)*. JMLR.org, 1889–1897.

[112] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:1707.06347 [cs.LG] https://arxiv.org/abs/1707.06347

[113] Mincheol Seong, Ohyun Jo, and Kyungseop Shin. 2024. Age of information minimization in UAV-assisted data harvesting networks by multi-agent deep reinforcement curriculum learning. *Expert Systems with Applications* 255 (2024), 124379. https://doi.org/10.1016/j.eswa.2024.124379

[114] Xingling Shao, Yi Xia, Zewei Mei, and Wendong Zhang. 2023. Model-guided Reinforcement Learning Enclosing for UAVs with Collision-free and Reinforced Tracking Capability. *Aerospace Science and Technology* 142 (2023), 108609. https://doi.org/10.1016/j.ast.2023.108609

[115] Michelle Sherman, Sihua Shao, Xiang Sun, and Jun Zheng. 2023. Optimizing AoI in UAV-RIS-Assisted IoT Networks: Off Policy Versus On Policy. *IEEE Internet of Things Journal* 10, 14 (2023), 12401–12415. https://doi.org/10.1109/JIOT.2023.3246925

[116] Huaguang Shi, Yuxiang Tian, Hengji Li, Jian Huang, Lei Shi, and Yi Zhou. 2024. Task offloading and trajectory scheduling for UAV-enabled MEC networks: An MADRL algorithm with prioritized experience replay. *Ad Hoc Networks* 154 (2024), 103371. https://doi.org/10.1016/j.adhoc.2023.103371

[117] Behzad Shirani, Majdeddin Najafi, and Iman Izadi. 2019. Cooperative load transportation using multiple UAVs. *Aerospace Science and Technology* 84 (2019), 158–169. https://doi.org/10.1016/j.ast.2018.10.027

[118] Mohammed Shurrab, Rabeb Mizouni, Shakti Singh, and Hadi Otrok. 2023. Reinforcement learning framework for UAV-based target localization applications. *Internet of Things* 23 (2023), 100867. https://doi.org/10.1016/j.iot.2023.100867

[119] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*. JMLR.org, I–387–I–395.

[120] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*. PMLR, Lille, France, 2256–2265.

[121] Fuhong Song, Huanlai Xing, Xinhan Wang, Shouxi Luo, Penglin Dai, Zhiwen Xiao, and Bowen Zhao. 2023. Evolutionary Multi-Objective Reinforcement Learning Based Trajectory Control and Task Offloading in UAV-Assisted Mobile Edge Computing. *IEEE Transactions on Mobile Computing* 22, 12 (2023), 7387–7405. https://doi.org/10.1109/TMC.2022.3208457

[122] Richard S. Sutton and Andrew G. Barto. 1998. Reinforcement Learning: An Introduction. *IEEE Trans. Neural Networks* 9 (1998), 1054–1054.

[123] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. (1999), 1057–1063.

[124] Fengxiao Tang, Hans Hofner, Nei Kato, Kazuma Kaneko, Yasutaka Yamashita, and Masatake Hangai. 2022. A Deep Reinforcement Learning-Based Dynamic Traffic Offloading in Space-Air-Ground Integrated Networks (SAGIN). *IEEE Journal on Selected Areas in Communications* 40, 1 (2022), 276–289. https://doi.org/10.1109/JSAC.2021.3126073

[125] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 23–30. https://doi.org/10.1109/IROS.2017.8202133

[126] Hoang Duong Tuan, Ali Arshad Nasir, Andrey V. Savkin, H. Vincent Poor, and Eryk Dutkiewicz. 2021. MPC-Based UAV Navigation for Simultaneous Solar-Energy Harvesting and Two-Way Communications. *IEEE Journal on Selected Areas in Communications* 39, 11 (2021), 3459–3474. https://doi.org/10.1109/JSAC.2021.3088633

[127] Federico Venturini, Federico Mason, Francesco Pase, Federico Chiariotti, Alberto Testolin, Andrea Zanella, and Michele Zorzi. 2021. Distributed Reinforcement Learning for Flexible and Efficient UAV Swarm Control. *IEEE Transactions on Cognitive Communications and Networking* 7, 3 (2021), 955–969. https://doi.org/10.1109/TCCN.2021.3063170

[128] Chao Wang, Jian Wang, Jingjing Wang, and Xudong Zhang. 2020. Deep-Reinforcement-Learning-Based Autonomous UAV Navigation With Sparse Rewards. *IEEE Internet of Things Journal* 7, 7 (2020), 6180–6190. https://doi.org/10.1109/JIOT.2020.2973193

[129] Dawei Wang, Tingxiang Fan, Tao Han, and Jia Pan. 2020. A Two-Stage Reinforcement Learning Approach for Multi-UAV Collision Avoidance Under Imperfect Sensing. *IEEE Robotics and Automation Letters* 5, 2 (2020), 3098–3105. https://doi.org/10.1109/LRA.2020.2974648

[130] Di Wang, Qianqian Liu, Jie Tian, Yuan Zhi, Jingping Qiao, and Ji Bian. 2021. Deep Reinforcement Learning for Caching in D2D-Enabled UAV-Relaying Networks. In *2021 IEEE/CIC International Conference on Communications in China (ICCC)*. 635–640. https://doi.org/10.1109/ICCC52777.2021.9580299

[131] Jiahua Wang, Ping Zhang, and Yang Wang. 2023. Autonomous target tracking of multi-UAV: A two-stage deep reinforcement learning approach with expert experience. *Applied Soft Computing* 145 (2023), 110604. https://doi.org/10.1016/j.asoc.2023.110604

[132] Kaiwen Wang, Lammmert Kooistra, Ruoxi Pan, Wensheng Wang, and João Valente. 2024. UAV-based simultaneous localization and mapping in outdoor environments: A systematic scoping review. *Journal of Field Robotics* 41, 5 (2024), 1617–1642. https://doi.org/10.1002/rob.22325 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/rob.22325

[133] Liang Wang, Kezhi Wang, Cunhua Pan, Wei Xu, Nauman Aslam, and Arumugam Nallanathan. 2022. Deep Reinforcement Learning Based Dynamic Trajectory Control for UAV-Assisted Mobile Edge Computing. *IEEE Transactions on Mobile Computing* 21, 10 (2022), 3536–3550. https://doi.org/10.1109/TMC.2021.3059691

[134] Taotao Wang, Soung Chang Liew, and Shengli Zhang. 2021. When blockchain meets AI: Optimal mining strategy achieved by machine learning. *International Journal of Intelligent Systems* 36, 5 (2021), 2183–2207. https://doi.org/10.1002/int.22375 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.22375

[135] Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2022), 4555–4576. https://doi.org/10.1109/TPAMI.2021.3069908

[136] Xueyuan Wang, M. Cenk Gursoy, Tugba Erpek, and Yalin E. Sagduyu. 2022. Learning-Based UAV Path Planning for Data Collection With Integrated Collision Avoidance. *IEEE Internet of Things Journal* 9, 17 (2022), 16663–16676. https://doi.org/10.1109/JIOT.2022.3153585

[137] Xiaojie Wang, Xin Wan, Hongjing Ji, Hao Hu, Ziqiang Chen, and Yulong Xiao. 2024. Joint UAV Deployment and User Scheduling for Wireless Powered Wearable Networks. *IEEE Internet of Things Journal* 11, 12 (2024), 21299–21311. https://doi.org/10.1109/JIOT.2024.3360078

[138] Yun Wang, Shu Fu, Changhua Yao, Haijun Zhang, and Fei Richard Yu. 2023. Caching Placement Optimization in UAV-Assisted Cellular Networks: A Deep Reinforcement Learning-Based Framework. *IEEE Wireless Communications Letters* 12, 8 (2023), 1359–1363. https://doi.org/10.1109/LWC.2023.3274535

[139] Yuntao Wang, Zhou Su, Ning Zhang, and Abderrahim Benslimane. 2021. Learning in the Air: Secure Federated Learning for UAV-Assisted Crowdsensing. *IEEE Transactions on Network Science and Engineering* 8, 2 (2021), 1055–1069. https://doi.org/10.1109/TNSE.2020.3014385

[140] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. 2016. Dueling network architectures for deep reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 1995–2003.

[141] Christopher Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8 (1992), 279–292.

[142] Kaimin Wei, Kai Huang, Yongdong Wu, Zhetao Li, Hongliang He, Jilian Zhang, Jinpeng Chen, and Song Guo. 2022. High-Performance UAV Crowdsensing: A Deep Reinforcement Learning Approach. *IEEE Internet of Things Journal* 9, 19 (2022), 18487–18499. https://doi.org/10.1109/JIOT.2022.3160887

[143] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3 (2016), 1–40. https://doi.org/10.1186/s40537-016-0043-6

[144] Chaoyang Wen, Yuan Fang, and Ling Qiu. 2022. Securing UAV Communication Based on Multi-Agent Deep Reinforcement Learning in the Presence of Smart UAV Eavesdropper. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. 1164–1169. https://doi.org/10.1109/WCNC51071.2022.9771555

[145] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8 (1992), 229–256. https://doi.org/10.1007/BF00992696

[146] Xiancai Xiang, Jian Xue, Lin Zhao, Yuan Lei, Chao Yue, and Ke Lu. 2024. Real-time Integration of Fine-tuned Large Language Model for Improved Decision-Making in Reinforcement Learning. In *2024 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN60899.2024.10650538

[147] Jian Xiao, Guohui Yuan, Yuxi Xue, Jinhui He, Yaoting Wang, Yuanjiang Zou, and Zhuoran Wang. 2024. A deep reinforcement learning based distributed multi-UAV dynamic area coverage algorithm for complex environment. *Neurocomputing* 595 (2024), 127904. https://doi.org/10.1016/j.neucom.2024.127904

[148] Yilin Xiao, Liang Xiao, Xiaozhen Lu, Hailu Zhang, Shui Yu, and H. Vincent Poor. 2021. Deep-Reinforcement-Learning-Based User Profile Perturbation for Privacy-Aware Recommendation. *IEEE Internet of Things Journal* 8, 6 (2021),

4560–4568. https://doi.org/10.1109/JIOT.2020.3027586

[149] Zehui Xiong, Yang Zhang, Wei Yang Bryan Lim, Jiawen Kang, Dusit Niyato, Cyril Leung, and Chunyan Miao. 2021. UAV-Assisted Wireless Energy and Data Transfer With Deep Reinforcement Learning. *IEEE Transactions on Cognitive Communications and Networking* 7, 1 (2021), 85–99. https://doi.org/10.1109/TCCN.2020.3027696

[150] Jinyong Xu. 2024. Efficient trajectory optimization and resource allocation in UAV 5G networks using dueling-Deep-Q-Networks. *Wireless Networks* (2024), 6687–6697. https://doi.org/10.1007/s11276-023-03488-1

[151] Shu Xu, Xiangyu Zhang, Chunguo Li, Dongming Wang, and Luxi Yang. 2022. Deep Reinforcement Learning Approach for Joint Trajectory Design in Multi-UAV IoT Networks. *IEEE Transactions on Vehicular Technology* 71, 3 (2022), 3389–3394. https://doi.org/10.1109/TVT.2022.3144277

[152] Yuntao Xue and Weisheng Chen. 2024. Multi-Agent Deep Reinforcement Learning for UAVs Navigation in Unknown Complex Environment. *IEEE Transactions on Intelligent Vehicles* 9, 1 (2024), 2290–2303. https://doi.org/10.1109/TIV.2023.3298292

[153] Chao Yan, Chang Wang, Xiaojia Xiang, Kin Huat Low, Xiangke Wang, Xin Xu, and Lincheng Shen. 2024. Collision-Avoiding Flocking With Multiple Fixed-Wing UAVs in Obstacle-Cluttered Environments: A Task-Specific Curriculum-Based MADRL Approach. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (2024), 10894–10908. https://doi.org/10.1109/TNNLS.2023.3245124

[154] Won Joon Yun, Soohyun Park, Joongheon Kim, MyungJae Shin, Soyi Jung, David A. Mohaisen, and Jae-Hyun Kim. 2022. Cooperative Multiagent Deep Reinforcement Learning for Reliable Surveillance via Autonomous Multi-UAV Control. *IEEE Transactions on Industrial Informatics* 18, 10 (2022), 7086–7096. https://doi.org/10.1109/TII.2022.3143175

[155] Chiya Zhang, Xinjie Li, Chunlong He, Xingquan Li, and Dongping Lin. 2023. Trajectory optimization for UAV-enabled relaying with reinforcement learning. *Digital Communications and Networks* (2023). https://doi.org/10.1016/j.dcan.2023.07.006

[156] Jidong Zhang, Yu Yu, Zhigang Wang, Shaopeng Ao, Jie Tang, Xiuyin Zhang, and Kai-Kit Wong. 2020. Trajectory Planning of UAV in Wireless Powered IoT System Based on Deep Reinforcement Learning. In *2020 IEEE/CIC International Conference on Communications in China (ICCC)*. 645–650. https://doi.org/10.1109/ICCC49849.2020.9238842

[157] Liang Zhang, Bijan Jabbari, and Nirwan Ansari. 2022. Machine Learning Driven UAV-assisted Edge Computing. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. 2220–2225. https://doi.org/10.1109/WCNC51071.2022.9771769

[158] Lu Zhang, Zi-Yan Zhang, Luo Min, Chao Tang, Hong-Ying Zhang, Ya-Hong Wang, and Peng Cai. 2021. Task Offloading and Trajectory Control for UAV-Assisted Mobile Edge Computing Using Deep Reinforcement Learning. *IEEE Access* 9 (2021), 53708–53719. https://doi.org/10.1109/ACCESS.2021.3070908

[159] Shuai Zhang, Weiqi Liu, and Nirwan Ansari. 2023. Completion Time Minimization for Data Collection in a UAV-enabled IoT Network: A Deep Reinforcement Learning Approach. *IEEE Transactions on Vehicular Technology* 72, 11 (2023), 14734–14742. https://doi.org/10.1109/TVT.2023.3280848

[160] Tiankui Zhang, Ziduan Wang, Yuanwei Liu, Wenjun Xu, and Arumugam Nallanathan. 2020. Caching Placement and Resource Allocation for Cache-Enabling UAV NOMA Networks. *IEEE Transactions on Vehicular Technology* 69, 11 (2020), 12897–12911. https://doi.org/10.1109/TVT.2020.3015578

[161] Tiankui Zhang, Ziduan Wang, Yuanwei Liu, Wenjun Xu, and Arumugam Nallanathan. 2022. Joint Resource, Deployment, and Caching Optimization for AR Applications in Dynamic UAV NOMA Networks. *IEEE Transactions on Wireless Communications* 21, 5 (2022), 3409–3422. https://doi.org/10.1109/TWC.2021.3121584

[162] Wenqi Zhang, Qiang Wang, Xiao Liu, Yuanwei Liu, and Yue Chen. 2021. Three-Dimension Trajectory Design for Multi-UAV Wireless Network With Deep Reinforcement Learning. *IEEE Transactions on Vehicular Technology* 70, 1 (2021), 600–612. https://doi.org/10.1109/TVT.2020.3047800

[163] Xiaochen Zhang, Haitao Zhao, Jibo Wei, Chao Yan, Jun Xiong, and Xiaoran Liu. 2023. Cooperative Trajectory Design of Multiple UAV Base Stations With Heterogeneous Graph Neural Networks. *IEEE Transactions on Wireless Communications* 22, 3 (2023), 1495–1509. https://doi.org/10.1109/TWC.2022.3204794

[164] Chenxi Zhao, Junyu Liu, Min Sheng, Wei Teng, Yang Zheng, and Jiandong Li. 2021. Multi-UAV Trajectory Planning for Energy-Efficient Content Coverage: A Decentralized Learning-Based Approach. *IEEE Journal on Selected Areas in Communications* 39, 10 (2021), 3193–3207. https://doi.org/10.1109/JSAC.2021.3088669

[165] Botao Zhu, Ebrahim Bedeer, Ha H. Nguyen, Robert Barton, and Jerome Henry. 2022. Joint Cluster Head Selection and Trajectory Planning in UAV-Aided IoT Networks by Reinforcement Learning With Sequential Model. *IEEE Internet of Things Journal* 9, 14 (2022), 12071–12084. https://doi.org/10.1109/JIOT.2021.3133278