Unmanned Systems, Vol. 14, No. 2 (2026) 267–280

DOI: 10.1142/S2301385026300015

UNMANNED SYSTEMS
https://www.worldscientific.com/worldscinet/us

Application of Reinforcement Learning in UAV Tasks: A Survey

Jiahao Fu [*,†,§], Feng Yang [†,‡,¶]

[*]National Elite Institute of Engineering,
Northwestern Polytechnical University, Xi'an 710072, P. R. China

[†]Key Laboratory of Information Fusion Technology,
Ministry of Education, Xi'an 710072, P. R. China

[‡]School of Automation, Northwestern Polytechnical University,
Xi'an 710072, P. R. China

The burgeoning demand for unmanned aerial vehicles (UAVs) across diverse domains can be attributed to their high flexibility, ease of deployment and low operational costs. Concomitantly, rapid advancements in reinforcement learning have emerged as a viable avenue for augmenting the autonomy of UAVs. This paper provides a comprehensive overview of the foundational concepts and methodologies of reinforcement learning and taxonomizes its applications in UAV decision-making into three primary categories: fundamental tasks encompassing obstacle avoidance and path planning, advanced tasks involving cooperative control, and complex tasks requiring adversarial decision-making. Additionally, the challenges associated with implementing reinforcement learning in UAV applications are critically examined. In the final section, we envision future research directions and provide a comprehensive summary of the study. This will assist practitioners and researchers in selecting appropriate reinforcement learning algorithms for their drone mission applications.

Keywords: Reinforcement learning; UAV; collaborative task execution; adversarial decision-making.

1. Introduction

In recent years, with the rapid advancement of technology, unmanned aerial vehicles (UAVs) have become an integral part of the modern technological landscape [1]. Whether in civilian domains (such as logistics transportation, power line inspection and agricultural irrigation) or military applications (including enemy reconnaissance and intelligent air combat), UAVs play a vital role [2]. Figure 1 provides examples of various tasks performed by drones. To offer a comprehensive overview of task types, this paper focuses on military UAVs.

The effectiveness of drone task execution largely depends on the decision-making techniques employed. Researchers have been investigating various decision-making methods to improve drone performance. For instance, the authors of [3–5] utilized game theory to facilitate drone decision-making. With the rise of heuristic algorithms, approaches such as genetic algorithms [6–8], ant colony optimization [9, 10], and particle swarm optimization [11, 12] have been frequently applied to tasks such as path planning and obstacle avoidance. Knowledge-based methods integrate extensive expert experience and have also been used as a foundation for drone decision-making. However, a significant limitation of these methods is the challenge of representing human expertise through formal mathematical rules. Nowadays, as the demand for drones to became more autonomous and intelligent increased, reinforcement learning has emerged as a promising machine learning approach, providing new solutions for autonomous decision-making and behavior optimization in UAVs.

Reinforcement Learning (RL) is a method that enables an intelligent agent to learn optimal strategies through continuous interaction with its environment [13]. In RL, an UAV

Received 30 October 2024; Revised 26 December 2024; Accepted 26 December 2024; Published 6 March 2025. This paper was recommended for publication in its revised form by editorial board member, Shiyu Zhao.
Email Addresses: [§]*fjh@mail.nwpu.edu.cn*, [¶]*yangfeng@nwpu.edu.cn*
[¶]Corresponding author.

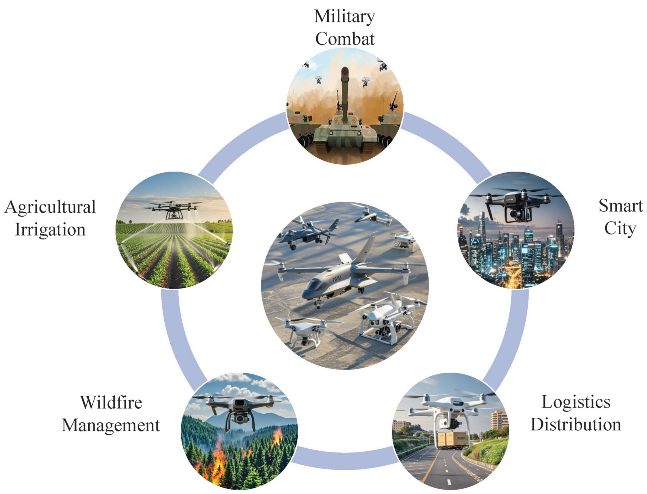


Fig. 1. UAV use case.

(Agent) obtains the environmental state [14] through various onboard sensors and performs actions (Actions) based on the environmental state (State). It then receives feedback (Reward) from the environment, thereby learning how to maximize cumulative rewards over the long term [15]. This learning approach has inherent advantages in the UAV domain because UAVs need to continuously interact with the environment and make adaptive decisions based on environmental changes while performing tasks [16].

In the review of decision-making for drones based on reinforcement learning, reference [17] outlines the application of artificial intelligence methods to the path planning problem for drone swarms, categorizing them into four main types: reinforcement learning, evolutionary computation, swarm intelligence and graph neural networks. However, the discussion of reinforcement learning is limited to the aspect of path planning. The authors of [18] focused on the formation problem of drone swarms, with similarly limited exploration of reinforcement learning applications. The authors of [19] primarily introduced the use of various drone simulators under reinforcement learning frameworks, with a stronger emphasis on the design of simulators and simulation environments. Reference [20] provides an overview of reinforcement learning control in fixed-wing drones, highlighting aspects such as attitude control, course control, swarming and landing glide. However, these pertain only to basic functionalities and lack coverage of more complex task scenarios.

Unlike the aforementioned literature, this paper not only encompasses the techniques discussed in those works but also delves into the fundamental nature of decision-making. It systematically transitions from simple, single-task scenarios to complex, multi-task environments, providing a comprehensive and systematic overview of various reinforcement learning algorithms applied in different UAV tasks.

2. Fundamentals of Reinforcement Learning and Decision-Making for UAVs

Regardless of the form of the reinforcement learning algorithm, its fundamental theoretical framework can be explained using the Markov Decision Process (MDP) [21]. The MDP process can be represented in the form of the following five-tuple:

$$(S, A, R, T, \gamma) \tag{1}$$

In the MDP framework, S represents the set of environmental states that the agent can be in, describing the agent's situation at a given time; A represents the set of actions that the agent can take, defining all possible actions the agent can execute; R denotes the reward provided by the environment, which is the feedback or reward signal the agent receives after interacting with the environment, with the reward magnitude typically related to the agent's goal; T is the state transition function, representing the probability distribution of the agent transitioning from state s_t to state s_{t+1} after taking action a_t, describing the impact of the agent's behavior on the environmental state; and γ is the discount factor, typically taking values between 0 and 1, which serves to reduce the influence of future rewards, emphasizing the importance of current rewards and preventing the agent from focusing too much on long-term rewards at the expense of short-term ones [22]. Through repeated interactions, the agent can continuously acquire new states and receive rewards to learn and update its model [23], as illustrated in Fig. 2.

The sequence of actions taken by the agent from the initial state to the terminal state constitutes an episode. The goal of MDP is to learn an optimal policy that maximizes the cumulative reward over the entire episode. In other words, MDP aims to find a policy that maximizes the expected sum of reward obtained by the agent from the initial state to the terminal state:

$$G_t = E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\right]. \tag{2}$$

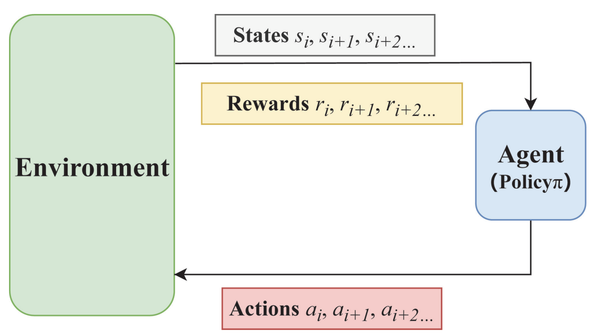


Fig. 2. Fundamental framework of reinforcement learning.

G_t represents the total reward, E denotes the expectation and R represents the immediate reward after each action. To achieve reward maximization, two primary categories of reinforcement learning methodologies are commonly employed [24]: value-based reinforcement learning methods [25–30], which estimate action-value functions [22, 31, 32] and policy-based reinforcement learning approaches [33–38], which directly optimize policy parameters.

In the modeling of UAV control based on Markov processes, we summarize the state space, action space and reward settings according to the literature. The state space may include: the UAV's position, velocity, orientation and ammunition information; the enemy's position, velocity, and orientation; the distance between the UAV and the enemy; and the target location information. To reduce computational complexity, the action space is often discretized, such as dividing the UAVs movement directions into different angles, acceleration actions, deceleration actions, and so on. In various tasks, the reward design for UAVs typically includes: a high reward for completing the target objective; rewards for avoiding obstacles or enemies during flight; negative rewards (i.e. penalties) for each step taken or time spent; and negative rewards for energy consumption or ammunition usage [39–45].

When applying reinforcement learning methods to UAV decision-making, the classical exploration-exploitation dilemma is inevitable. Since UAVs cannot obtain complete environmental information during actual mission execution, we encourage them to perform diverse actions to receive different reward feedback, thereby avoiding local optima. However, UAVs cannot take an infinite number of actions; they need to reasonably balance exploration and exploitation to maximize rewards. The ratio of exploration to exploitation can be determined using various strategies such as ϵ-greedy, Upper Confidence Bound (UCB), and Gradient Bandit algorithms.

To further enhance UAV decision-making capabilities, researchers have integrated neural networks with reinforcement learning methodologies in this domain, known as Deep Reinforcement Learning (DRL). For instance, during UAV navigation, environmental state images can be acquired through a monocular camera, processed by convolutional neural networks (CNNs), and subsequently utilized within a reinforcement learning framework to output optimal actions while maximizing discounted returns [46].

3. Applications of Reinforcement Learning in UAV

Different reinforcement learning algorithms have been employed to address various UAV mission challenges. Currently, civilian UAV applications primarily include inspection and surveillance, agricultural operations, meteorological data collection, exploration and mapping. In military contexts, UAVs are mainly utilized for reconnaissance, strike missions, swarm operations and communication relays. Regardless of whether UAVs are used for civilian or military purposes, whether operating individually or as part of a swarm, obstacle avoidance, and path planning remain fundamental tasks. The remaining tasks are categorized in this paper into collaborative cooperation and adversarial decision-making, as illustrated in Fig. 3.

3.1. *UAV obstacle avoidance and path planning*

3.1.1. *Obstacle avoidance*

The obstacle avoidance capability of UAVs is a fundamental guarantee for their safe flight in executing all tasks. This can be achieved by leveraging various onboard sensors such as cameras and radar [47]. The data provided by these devices are crucial for training reinforcement learning algorithms, which in turn enables neural network models to learn how to effectively identify and avoid obstacles in complex spatial environments.

The research by ZHAO *et al.* highlights the potential of Reinforcement Learning (RL) in UAV missions. They designed a UAV obstacle avoidance method based on the Q-learning algorithm, which allows the UAV to interact with the environment and effectively evade obstacles during flight, thus enhancing safety [48]. Aleksandra Faust proposed an RL strategy based on minimum residual oscillation trajectories, ensuring that autonomous aerial cargo UAVs can navigate quickly and safely in environments with static obstacles [49]. Combining deep neural networks with reinforcement learning, such as the Deep Q-learning algorithm, has also been used to explore optimal strategies for UAV obstacle avoidance [50]. Besides using specific obstacle coordinate information as input data for obstacle avoidance simulation tests, image information during UAV navigation can also serve as environmental states for DRL methods. Ou *et al.* [51] proposed a novel learning-based framework for achieving autonomous obstacle avoidance in monocular vision-based quadrotor UAVs. This framework employs a two-stage architecture, comprising a perception module

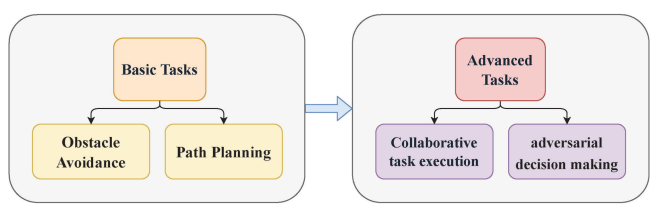


Fig. 3. Task relationship graph.

and a decision-making module. The perception module extracts depth information from onboard camera images using an unsupervised learning approach, which serves as input to the decision-making module. The decision-making module leverages Dueling Double Deep Recurrent Q-learning (D3RQN) to mitigate the limitations of the onboard monocular camera's observational capabilities and select practical obstacle avoidance actions. In addition to depth information images serving as inputs for reinforcement learning algorithms, other image modalities, such as RGB images and event-based imaging types, are also employed. Ma *et al.* [52] designed a saliency detection algorithm using a deep CNN to extract monocular visual cues. This algorithm mimics the human visual detection system, accurately estimating the position of obstacles within the field of view (FOV). It then combines with RL-based action selection, enabling UAVs to learn to avoid obstacles during flight. AlMahamid *et al.* [53] asserted that RGB images lack depth information, impeding the accurate assessment of the distance between the UAV and surrounding objects, which can lead to unstable flight maneuvers. Depth information is crucial for constructing an effective reward function that penalizes UAV behaviors that bring it excessively close to obstacles. Consequently, some approaches employ both RGB images and depth maps as inputs to provide richer environmental information. In contrast, event-based images present brightness changes as one-dimensional temporal sequences. Although they do not include RGB channels, event-based images can efficiently capture rapid changes within a scene and are more storage-efficient compared to traditional 2D images.

These different types of image data outputs reflect the state of the UAV at various time points, and this state information is fed as input data into an RL agent. The RL agent makes decisions based on this input data to control the UAV's flight actions, thereby avoiding obstacles. The design of the RL agent's neural network structure depends on several key factors: first, the type of input data; second, the quantity of input data; and finally, the chosen algorithm. For instance, when using the DQN algorithm to process RGB images or depth maps, CNNs are typically employed to extract features, followed by fully connected layers for further processing. This is due to the excellent performance of CNNs in image processing. For event-based image data, spiking neural networks (SNNs) are utilized. This network structure is particularly suited for handling spatiotemporal data and can recognize spatiotemporal patterns within it [54].

In swarm mode, the obstacle avoidance problem for UAVs is more complex than that for a single UAV. It involves a higher-dimensional state space and action space, necessitating full consideration of the coordination between UAVs. Niu *et al.* [55] proposed a multi-strategy evolutionary algorithm based on an improved Artificial Ecosystem Optimizer and reinforcement learning (MEAEO-RL) to address the collaborative path planning challenges for multiple UCAVs in complex environments. This algorithm is designed to generate optimal candidate paths for each UCAV, ensuring that they can concurrently reach their destinations while accounting for temporal variables and obstacle avoidance. By constructing a brain-inspired learning framework and integrating swarm intelligence with human cognitive mechanisms, the intelligence of the swarm agents is significantly enhanced. Furthermore, the study implemented a multi-strategy database to replace the traditional single update method of the Artificial Ecosystem Optimizer (AEO) during the consumption phase. By utilizing the cumulative experience from reinforcement learning to select effective update strategies during this phase, the algorithm's computational complexity is reduced. Path planning simulation experiments conducted in a series of complex three-dimensional environments demonstrated that the algorithm excels in robustness, improved convergence accuracy, and the ability to plan cooperative paths for multiple UCAVs while adhering to various constraints.

Similarly, avoiding dynamic obstacles has always been a challenging problem. While numerous path planning algorithms have been proposed to help UAVs find effective paths between the source and the destination, these algorithms often fall short in addressing the issue of dynamic obstacles. To tackle this challenge, Sonny *et al.* [56] introduced a Q-learning-based algorithm that not only handles static obstacles but also adapts to dynamically changing environments, thereby achieving effective avoidance of dynamic obstacles. They incorporated a shortest-distance priority strategy to minimize the distance the UAV needs to travel to reach the target. Additionally, the proposed Q-learning algorithm employed a grid-based method to solve the path planning problem and maximized the reward by learning from the behavior of agents in the environment. Through comparative results, this method demonstrated improved performance in learning and path length compared to advanced path planning methods such as A*, Dijkstra and SARSA algorithms. This highlights that UAV obstacle avoidance and path planning are inseparable issues; path planning is not merely about determining a path from the start to the end but also ensuring safety and efficiency during the flight.

3.1.2. *Path planning*

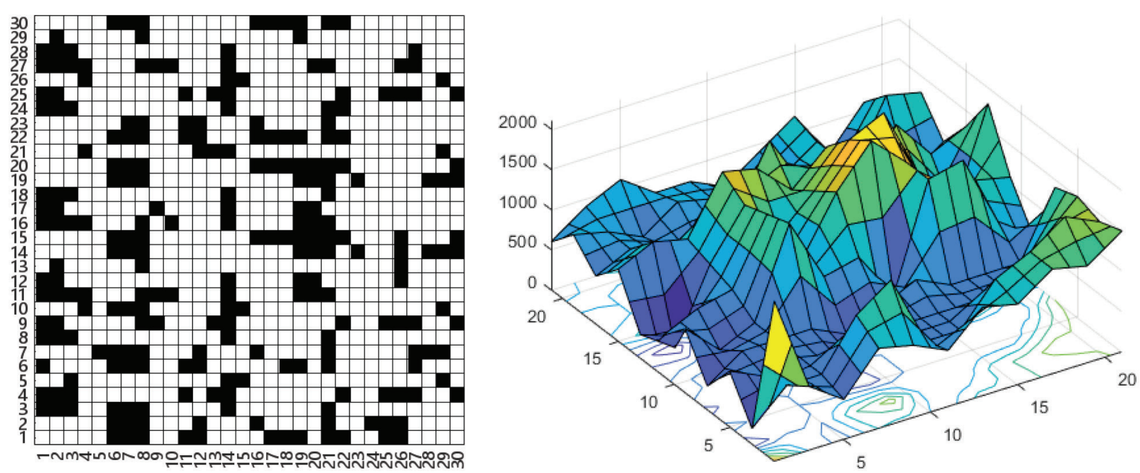
UAV path planning can be categorized into global path planning and local path planning. Global path planning involves devising routes based on pre-existing environmental information to navigate around obstacles within the environment. In contrast, local path planning relies on the

UAVs onboard sensors to perceive information in the immediate vicinity, enabling real-time route adjustments. Path planning for UAVs is typically conducted within two-dimensional or three-dimensional spatial frameworks, as illustrated in Fig. 4.

Path planning can be addressed using various techniques and algorithms. In this work, we focus on reinforcement learning technologies for solving both global and local path planning problems. In this approach, the reinforcement learning agent receives information from the environment and outputs the optimal path points based on a reward function.

Zhang *et al.* proposed a geometric reinforcement learning method for UAV path planning learning. This method uses a reward matrix to represent the optimal path from the current point to the target point, with the reward matrix updated based on geometric distance and risk information [57]. The Q-learning algorithm, as one of the classic reinforcement learning algorithms, is also an effective approach for solving UAV path planning problems. For instance, Tianze Zhang *et al.* divided the quadrotor UAVs actions into forward, backward, left, right and hover, establishing a corresponding Q-table to explore the maximum reward, i.e. the optimal path, in a grid map simulation environment [58]. Jones, Djahel and Welsh introduced an innovative method called Multiple Q-table Path Planning (MQTPP), specifically designed for UAV path planning in urban environments. Compared to traditional Q-learning algorithms, MQTPP adaptively re-plans paths within a single learning stage, significantly enhancing computational efficiency during flight while nearly achieving optimal path planning results [59]. In addressing the path planning problem for UAV swarms in obstacle-laden environments, Puente-Castro *et al.* [60] proposed a Q-learning-based system that leverages artificial neural networks (ANN) for self-adjustment, optimizing the path by learning from both failures and successes. The study demonstrated that as the number of UAVs increases, the system is capable of achieving fewer actions across maps of varying sizes. Furthermore, it was shown that employing a two-layer ANN to implement the Q-learning algorithm is sufficient to operate in maps with obstacles, thereby providing an effective solution for the autonomous control of UAV swarms. Wang *et al.* [61] addressed the problem of time-constrained path planning for UAV operations in complex unknown environments by introducing an Improved Exploration Mechanism (IEM). The IEM combines Prioritized Experience Replay (PER) and curiosity-driven exploration to enhance sampling efficiency and encourage exploration, thereby overcoming local optimum issues in sparse reward environments. However, increased exploration raises computational complexity. In addition to value-based methods, policy-based methods are also quite effective. Silvirianti *et al.* [62] proposed Quantum-Based Deep Deterministic Policy Gradient (Q-DDPG) and Quantum-Based Recurrent DDPG (Q-RDDPG) schemes for time-series optimization in UAV communications. This research leverages quantum models to reduce computational complexity and training loss, particularly optimizing action selection within continuous state and action spaces. The Q-DDPG and Q-RDDPG approaches exhibited superior performance in trajectory optimization and dynamic resource allocation for UAV communications, achieving higher rewards and lower training losses compared to traditional DDPG methods.

In exploring the auxiliary role of UAVs in smart cities and the Internet of Things (IoT), Xi *et al.* proposed a lightweight real-time path planning method based on reinforcement learning, known as the Adaptive Soft Actor-Critic (ASAC)



(a) Two-Dimensional Space (b) Three-Dimensional Space

Fig. 4. Various spatial simulation environments.

algorithm. This approach significantly enhances the UAVs path planning capabilities in complex urban environments by optimizing the training process, network architecture, and algorithm model [63].

In the field of real-time path planning, Lee *et al.* [64] proposed a novel UAV control approach based on Goal-Conditioned Reinforcement Learning (GCRL). This method enhances the flexibility of UAV agents in performing diverse tasks by training them to be guided by subgoals. Compared to traditional path planning methods, this algorithm enables UAVs to receive user-defined subgoals in real-time in unknown environments and execute complex maneuvers such as high-altitude flight, low-altitude flight, penetration, and avoidance. Experimental results show that UAV agents, even pre-trained in simple environments, are capable of completing challenging tasks in various complex scenarios, demonstrating the potential applicability of this approach across different environments.

With the foundational capabilities of obstacle avoidance and path planning ensured, UAVs can adapt to more complex task requirements. The following discussion will categorize tasks into two types: cooperative collaboration and adversarial decision-making.

3.2. *UAV collaborative cooperation*

In recent years, significant advancements have been made in the study of cooperative collaboration among UAV swarms. By integrating multiple UAV systems, these swarms can perform tasks that are more complex and diverse than those a single UAV can handle. Cooperative collaboration in UAV swarms involves multiple UAVs operating as a unified system, communicating and coordinating with each other to achieve specific tasks or goals. This collaborative approach shows tremendous potential for applications in various fields, including military operations, search and rescue missions, environmental monitoring and agricultural spraying.

The key technologies for cooperative collaboration among UAVs include the following:

(i) Communication Protocols: Ensuring effective communication between UAVs and the ground control station.
(ii) Path Planning: Designing optimal paths for each UAV to avoid collisions and overlap.
(iii) Task Allocation: Assigning tasks rationally based on the UAVs' capabilities and mission requirements.
(iv) Cooperative Control Algorithms: Facilitating coordinated actions and decision-making among UAVs.

Peng *et al.* considered a communication system between a UAV swarm and a base station, which is interfered with by multiple interference sources. They proposed a multi-dimensional anti-interference reinforcement learning algorithm that incorporates energy constraints and tunes antenna beams, effectively improving decision-making and enhancing overall UAV communication quality [65]. Xu *et al.* used the DQN algorithm to address the UAV formation control problem under group communication constraints. Their trained model can complete target waypoint searches while maintaining general communication quality within the group [66]. Zhao *et al.* [67] explored the optimization of task offloading in UAV-assisted mobile edge computing environments using multi-agent deep reinforcement learning (MADRL). They proposed a cooperative MADRL framework and specifically designed the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm to address the challenges posed by high-dimensional continuous action spaces. This work not only takes into account the limited computational and energy capacities of UAVs but also optimizes UAV trajectories, computational task allocation, and communication resource management to minimize the total execution delay and energy consumption. Through simulation results, they demonstrated that the proposed multi-UAV multi-edge cloud task offloading method outperforms other optimization approaches in terms of performance.

Liu and Bai [68] proposed an intent communication-based cooperative traffic monitoring method for multiple UAVs. This approach utilizes multi-agent reinforcement learning (MARL) neural networks to generate intent information and optimizes flight control and task allocation strategies by incorporating intent data from other UAVs. The method effectively enhances the collaboration between UAVs when managing large numbers of random traffic monitoring tasks, significantly improving vehicle navigation accuracy and achieving a more balanced workload distribution in autonomous vehicles. Therefore, it is evident that utilizing reinforcement learning methods is an effective approach to solving homogeneous UAV cooperative tasks. UAV swarm operations have made significant contributions across various tasks.

In certain missions, different types of UAVs are often required to handle various functional modules, making the cooperative operation of heterogeneous UAVs a hot topic in the UAV field. Zhan *et al.* conducted an in-depth study on UAV swarms in complex multi-agent collaboration scenarios by proposing a Multi-Agent Proximal Policy Optimization (MAPPO) algorithm within the Ray framework. This research achieved effective collaboration among heterogeneous UAVs by incorporating global information in the Critic network and using local information in the Actor network. Furthermore, the use of curriculum learning for phased training enhanced the algorithm's generalization performance. Experimental results in a Unity3D collaborative combat environment demonstrated that the MAPPO algorithm, upon convergence, completed target tasks with the fewest steps and achieved the highest average cumulative reward compared to other algorithms, showcasing exceptional performance [69].

The size of a UAV swarm directly affects the complexity of task planning, with large-scale UAV clusters facing the problem of dimensionality explosion. Zuo *et al.* [70] proposed a fast and robust algorithm that combines the Simple Attention Model (SAM) with Dynamic Information Reinforcement Learning (DIRL) to address the task planning challenges for large-scale UAV clusters. Dai *et al.* [71] introduced a UAV deployment and resource allocation method based on Multi-Agent Cooperative Environment Learning (MACEL), aimed at optimizing the dynamic positioning and resource allocation in UAV communication networks through a distributed architecture. This method leverages the DQN algorithm, enabling each UAV to act as an independent agent that learns the mapping from partially observed state information to positioning and resource selection decisions. However, this approach lacks the ability to make decisions based on global information.

3.3. *UAV adversarial decision-making*

With the rapid advancement of manufacturing and artificial intelligence technologies, UAVs have made significant strides not only in the civilian sector but also in modern aerial combat due to their superior performance. This aligns with the other category of tasks discussed in this article: adversarial decision-making.

Intelligent decision-making technology is essential for UAVs to perform various tasks — such as reconnaissance, early warning, and electronic warfare — while maintaining high survival rates [72–75]. As shown in Table 1, numerous related projects have been initiated worldwide over the past two decades [76].

In academia, numerous scholars have extensively researched UAV adversarial decision-making, recognizing the superiority of reinforcement learning-based maneuvering decisions. Chen *et al.* [77] designed and developed an autonomous UAV maneuvering decision-making human-machine air combat system based on a deep Q-learning network. This air combat system has been validated to effectively simulate real air combat, confirming the effectiveness of autonomous UAV maneuvering decisions. Vlahov *et al.* [78] used the A3C algorithm to train one-on-one UAV air combat strategies and transitioned the learned strategies from simulation to real-world flight testing, demonstrating that the application of reinforcement learning in UAV adversarial decision-making is not limited to simulations. Li *et al.* proposed a UAV air combat decision-making method based on parallel self-play reinforcement learning, enhancing the autonomous decision-making capabilities of UAVs in one-on-one air combat. This study established an air combat model, defined missile attack zones, and employed the Soft-Actor-Critic (SAC) algorithm to construct the decision model, achieving maneuvering processes. Additionally, the study analyzed the stability of the closed-loop air combat decision system controlled by neural networks using the Lyapunov function and verified through simulation results the effectiveness of the proposed algorithm in improving the generalization performance of UAV control decisions [79]. Zhang *et al.* [80] developed a multi-UAV cooperative air combat decision-making model based on multi-agent reinforcement learning. They employed bidirectional recurrent neural networks to enable effective communication between UAVs and integrated target allocation and situational assessment to generate cooperative tactical maneuver strategies. These strategies guide UAVs to achieve an overall situational advantage and

Table 1. Important air combat projects in the last two decades.

Country or region	Time	Content
France	2003	The “Horizon” Program aims to develop unmanned combat drone prototypes to ensure battlefield dominance and defense requirements.
Russia	2010	The “National Weapons Program 2011–2020” plans the development of UAV systems and forces.
USA	2016	The “Gremlin” Program aims to explore technologies such as aerial launch and recovery of small drone swarms, low-cost airframe design and complex digital flight control systems to develop new combat methods.
USA	2017	The OFFSET Program aims to achieve breakthroughs in the operational capabilities of autonomous drone swarm systems by combining existing autonomous swarms with new manned-unmanned teaming technologies.
USA	2019	The Air Combat Evolution project aims to develop intelligent air combat algorithms to achieve cooperative operations between MAVs and UAVs.
USA	2020	The AlphaDogfight Program aims to test AI algorithms for air combat and to realize joint operational capabilities between MAV and UAV.

effectively defeat opponents. This work provides new perspectives and methodological guidance for autonomous air combat decision-making in high-tech local warfare. Jiang *et al.* [81] designed a multi-agent transformer network structure to leverage the homogeneity of UAV state information for the structured processing of complex situational data. Specifically, they utilized a self-attention mechanism to calculate each UAV's local situational information, incorporating virtual objects to assist in computing the weight distribution of local situations. This approach results in a more effective global situational representation, providing a more accurate foundation for maneuver decision-making in short-range air combat scenarios involving UAV swarms.

Besides focusing on situational information as the basis for maneuvering decisions, Wang *et al.* [82] proposed an enhanced multi-UAV air combat decision-making method based on hierarchical reinforcement learning from the perspective of action types. By designing a hierarchical decision network based on tactical action types, they simplified the complexity of the maneuvering decision space and improved the number of valuable experiences by decomposing high-quality combat experiences obtained during training, thereby reducing the complexity of strategy learning. Experimental results using the advanced UAV simulation platform JSBSim showed that the proposed method demonstrated superior performance in both balanced and disadvantageous air combat environments compared to various baseline algorithms.

In adversarial decision-making, implementing preemptive strategies often yield significant results. Liu *et al.* [75] studied adversarial games in air combat environments with multiple UAVs and proposed a novel Graph Attention Multi-Agent Soft Actor-Critic Reinforcement Learning and Target Prediction Network (GA-MASAC-TP Net) to address multi-UAV adversarial issues. This method enhances graph attention neural networks and group target prediction networks, enabling each agent to learn how to handle teammate information and predict the future trajectories of targets, thereby improving its strategies during adversarial processes. The research results indicate that compared to traditional multi-agent reinforcement learning methods, GA-MASAC-TP Net shows higher mission success rates and lower health losses in simulated environments, effectively forming an efficient multi-UAV adversarial strategy [83].

4. Challenges of Reinforcement Learning in UAV Decision-Making

In the development of autonomous decision-making for UAVs, several challenges have emerged. This paper outlines two of the most significant issues: the challenge of establishing accurate UAV motion models, and the inherent challenges of reinforcement learning in training and precision control.

4.1. *Establishing UAV motion models*

Establishing a motion model for UAV is a complex process involving several technical challenges, primarily including the following aspects:

Dynamic Complexity: The motion of UAVs is influenced by multiple factors such as airflow, wind, gravity and inertia. The interaction of these factors results in highly complex dynamic behavior.

Nonlinear Characteristics: The equations governing UAV motion are typically nonlinear, posing significant challenges for modeling and control. Nonlinear systems can exhibit unpredictable behavior, necessitating advanced mathematical tools and techniques for effective handling.

Dimensionality Issue: UAVs move in three-dimensional space, requiring consideration of multiple degrees of freedom such as roll, pitch and yaw. This increases the complexity of modeling. Additionally, the motion models of fixed-wing UAVs and multi-rotor UAVs differ significantly, as shown in Figs. 5 and 6.

Model Validation and Testing: The developed motion model needs to be validated through experiments, which may involve flight testing. This process can be costly and carries inherent risks.

Safety and Robustness: The UAV motion model must be capable of handling unexpected situations and model uncertainties to ensure flight safety and system robustness.

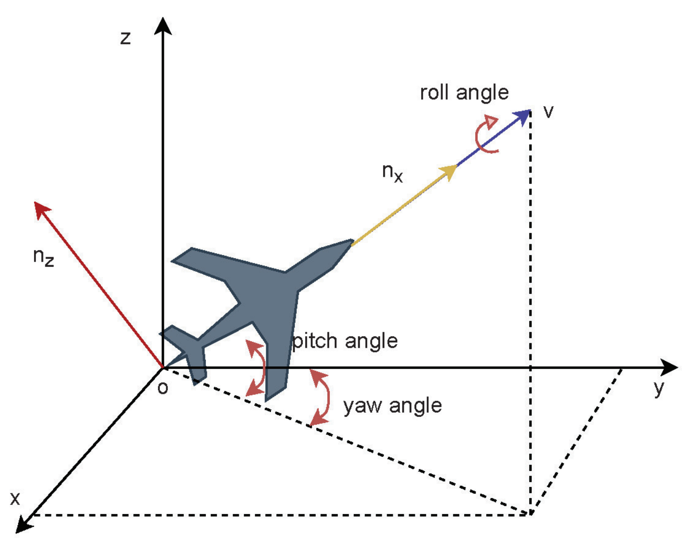


Fig. 5. Fixed-wing UAV dynamics model.

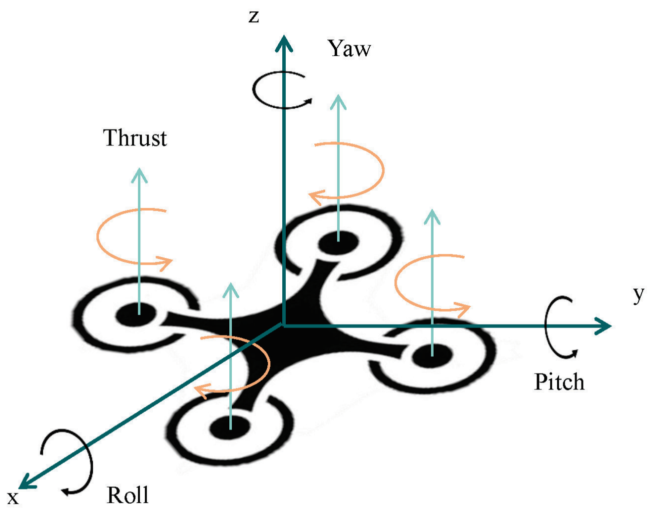


Fig. 6. Multi-rotor UAV dynamics model.

Özbek *et al.* [84] and Din *et al.* [85] proposed the Harfang3D Dog-Fight Sandbox, a semi-realistic flight simulation environment that comprehensively considers various factors to simulate UAV dynamics, environmental states and aerodynamics. This environment allows users to customize different task scenarios, including one-on-one dogfights and multi-UAV adversarial engagements. By using DRL techniques, an intelligent decision-making system was established to train the agents.

Bøhn *et al.* [86] utilized DRL agents to interact with controlled systems capable of handling complex nonlinear dynamics, exploring optimal control laws. This enabled attitude control of fixed-wing aircraft operating on original nonlinear dynamics, and the trained controller models were deployed and tested on real UAVs.

When solving the dimensionality issues of UAV motion models, the action space of UAVs is typically divided into discrete and continuous action spaces. Some researchers (e.g. Hu *et al.* [87, 88]; Yang *et al.* [89]; Wu *et al.* [90]; Wang *et al.* [91]) have developed intelligent maneuver decision models based on DRL and constructed corresponding maneuver libraries, providing UAVs with a variety of maneuver options. These models employ the DQN algorithm and utilize discretized maneuver actions to help UAVs obtain effective strategies in close or long-range combat, either to defeat opponents or evade attacks. Specifically, Hu *et al.* [87] adopted 15 discrete typical maneuver actions to replace the continuous action space of UAVs, and experimental results confirmed the effectiveness of this method in simulating UAV spatial movements and achieving intelligent decision-making. To address the mismatch between discrete maneuvers and the continuous maneuvers of real fighter jets, other researchers (e.g. Yue *et al.* [92]; Li *et al.* [93]; Zhang *et al.* [94]; Yang *et al.* [95]; Yoo *et al.* [96]) began using continuous control inputs to replace discrete actions such as climbing and left turns. They applied algorithms such as DDPG and PPO, enabling UAVs to learn and make intelligent decisions in dynamic environments. Li *et al.* [93] developed a UAV dynamics model that considers continuous control variables and designed an intelligent decision-making framework based on different aerial combat intents, such as attack, evade, pursuit and energy conservation. Experimental results show that under the guidance of MAV intent, the intelligence level of UAVs has been enhanced.

4.2. *Reward function design*

RL is a machine learning method that allows agents to learn how to make decisions through interactions with the environment. In reinforcement learning, the reward function is one of the core components, defining the immediate reward an agent receives when performing an action or reaching a state. The design of the reward function is crucial to the learning performance of the agent [97, 98]. Below are some key points to consider when designing a reward function:

Define clear objectives: The reward function should explicitly reflect the agent's ultimate goal. For instance, if the goal is for the agent to complete a task as quickly as possible, the reward function should incentivize swift actions.

Immediate and long-term rewards: When designing the reward function, it's important to balance immediate rewards with long-term rewards. Immediate rewards can rapidly guide the agent's actions, while long-term rewards assist the agent in learning deeper strategies.

Sparse versus dense rewards: Sparse rewards mean the agent only receives rewards upon reaching specific goals, whereas dense rewards provide feedback at every step or state. Sparse rewards can reduce computational overhead but may slow down the learning process, while dense rewards can speed up learning but might require more complex strategies to avoid local optima.

Positive versus negative feedback: The reward function can be positive (i.e. encouraging certain behaviors) or negative (i.e. punishing certain behaviors). In design, it's important to consider how to guide the agent's behavior through positive and negative feedback.

Exploration versus exploitation: The reward function should encourage the agent to strike a balance between exploring new actions and exploiting known effective behaviors. This can be achieved by designing the reward function, for example, by increasing rewards for exploratory behaviors.

Multi-task learning: If the agent is required to accomplish multiple tasks simultaneously, the reward function may need to be designed as multi-dimensional to reflect the relative importance of each task.

Safety and constraints: In some applications, such as autonomous driving or robotic operations, the reward function may need to include safety constraints to ensure the agent's behavior does not violate safety rules.

Timeliness of feedback: Rewards should be given to the agent as quickly as possible to enable faster learning.

Interpretability: In certain applications, the design of the reward function should account for interpretability so that humans can understand and verify the agent's behavior.

The design of a reward function is a process that requires customization based on the specific application scenario and goals. In practical applications, the reward function may need to be optimized through experimentation and adjustments to achieve the best learning outcomes.

Wang *et al.* [16] particularly emphasized the design of the reward function, adopting a sparse reward scheme, where the UAV is only rewarded upon successfully completing a navigation task. This approach simplifies the reward mechanism but is only suitable for tasks with simple objectives. Hou *et al.* studied the Subtask-Masked Curriculum Learning for Reinforcement Learning (SubMas-RL) method, specifically addressing the problem of UAV maneuver decision-making in sparse reward tasks. They constructed a sparse reward structure based on key events, setting unique rewards for each situation. This design avoids the difficulty of designing dense reward functions due to insufficient expert experience [99].

Wu *et al.* [100] adopted a dense reward structure to guide UAV decision-making. In their research, the reward consisted of four components: a reward for quickly reaching the destination, a constant negative reward for each UAV action, a reward for encouraging safe exploration, and a constraint reward based on the distance between the UAV and obstacles. This reward design provides both long-term goal rewards and immediate feedback for each action, while also encouraging the agent to explore more of the space.

In designing the reward function for the joint optimization problem concerning UAV trajectory and resource allocation, Silvirianti *et al.* [101] not only considered the UAVs flight trajectory, user grouping, and power allocation, but also quantified the UAVs energy efficiency performance using cumulative rewards, effectively making energy efficiency an optimization objective to find the optimal UAV operational strategy.

In the study by Jinsong Hu *et al.*, they proposed the Twin-Delayed Deep Deterministic Policy Gradient Aided Covert Transmission Algorithm (TD3-CT), which introduces a reward-shaping mechanism. This mechanism not only considers the improvement of effective throughput but also encourages the UAV to optimize its flight path and transmission strategy while maintaining stealth [102]. Chansuparp and Jitkajornwanich, in their research, proposed the Augmentative Backward Reward (ABR) [103], a design that breaks through the limitations of traditional reward mechanisms. Assigning rewards proportional to each transition's contribution toward achieving the goal, mitigates the bias that may be introduced by goal-oriented reward mechanisms.

Li *et al.* [104] addressed the issue of insufficient hard constraints during UAV path planning, which makes it difficult to ensure safety during the learning and execution phases. They proposed the Shield-DDPG algorithm, where the Shield acts as a protective mechanism to ensure that the UAVs state trajectory satisfies a safety specification. This method enhances the safety of reinforcement learning in solving UAV decision-making problems.

5. Future Prospects

Based on the existing research foundation and technological development trends, future research can further expand in the following directions:

(i) Algorithm innovation and integration: Future research may focus on developing new reinforcement learning algorithms capable of better handling continuous action spaces and high-dimensional state spaces. At the same time, it will explore how to integrate current RL algorithms with other machine learning techniques (such as deep learning and imitation learning) to enhance decision-making efficiency and accuracy.

(ii) Real-time decision-making and responsiveness: The application of reinforcement learning in UAV decision-making will increasingly emphasize real-time performance. Researchers will focus on optimizing algorithms to reduce computational resource consumption and enhance UAVs' ability to respond quickly in dynamic environments.

(iii) Environmental awareness and learning: UAVs will require deeper environmental perception capabilities to make better decisions. Future research may explore how to leverage RL algorithms to enable UAVs to learn environmental characteristics from experience and utilize this knowledge for more accurate decision-making.

(iv) Human-machine interaction: Future UAV systems will place more emphasis on the human-machine interaction experience. Reinforcement learning can be used to develop UAVs capable of understanding and predicting human intentions, thereby achieving more natural and effective human-machine collaboration.

(v) Interpretability and transparency: With growing demands for transparency and interpretability in algorithmic decision-making, future research will explore ways to make RL algorithms' decision-making processes more transparent, enabling users and regulatory bodies to understand and trust UAV decisions.

(vi) Hardware and software co-development: Advances in hardware, such as more efficient computing chips and more advanced sensors, will evolve in tandem with RL algorithm development, enhancing the overall performance of UAV decision-making systems.

6. Conclusion

This paper explores the application of reinforcement learning in the field of UAV decision-making. Reinforcement learning algorithms control UAVs to move in the target direction by using agents to interact with the environment states, and this method is expected to further enhance UAV autonomy and intelligence. In this paper, we first review the fundamental concepts and theories of reinforcement learning; then, UAV tasks are classified into three categories: obstacle avoidance and path planning, cooperative collaboration and adversarial decision-making. We progressively summarize the contributions of current researchers, starting from basic functional tasks. Next, we extract the key challenges and difficulties reinforcement learning faces in UAV decision-making, along with proposed solutions. Finally, we provide an outlook on future research directions.

Acknowledgment

This work has no foundation support.

ORCID

Jiahao Fu https://orcid.org/0009-0008-8916-5413
Feng Yang https://orcid.org/0000-0001-6991-1628

References

[1] Y. Zeng, R. Zhang and T. J. Lim, Wireless communications with unmanned aerial vehicles: Opportunities and challenges, *IEEE Commun. Mag.* **54**(5) (2016) 36–42.

[2] M. A. Bedford, Unmanned aircraft system (UAS) service demand 2015–2035: Literature review projections of future usage, Technical Report 13-01, DOT-VNTSC-DoD (2013).

[3] H. Park, B.-Y. Lee, M.-J. Tahk and D.-W. Yoo, Differential game based air combat maneuver generation using scoring function matrix, *Int. J. Aeronaut. Space Sci.* **17**(2) (2016) 204–213.

[4] P. Runyan, W. Heping, W. Zhengping and L. Yu, Decision-making of aircraft optimum configuration utilizing multi-dimensional game theory, *Chin. J. Aeronaut.* **23**(2) (2010) 194–197.

[5] R. Vidal, O. Shakernia, H. J. Kim, D. H. Shim and S. Sastry, Probabilistic pursuit-evasion games: Theory, implementation, and experimental evaluation, *IEEE Trans. Robot. Autom.* **18**(5) (2002) 662–669.

[6] F. C. J. Allaire, M. Tarbouchi, G. Labonté and G. Fusina, FPGA implementation of genetic algorithm for UAV real-time path planning, *J. Intell. Robot. Syst.* **54**(1) (2009) 495–510.

[7] A. A. Najm and I. K. Ibraheem, Nonlinear PID controller design for a 6-dof UAV quadrotor system, *Eng. Sci. Technol. Int. J.* **22**(4) (2019) 1087–1097.

[8] F. Yan, J. Chu, J. Hu and X. Zhu, Cooperative task allocation with simultaneous arrival and resource constraint for multi-UAV using a genetic algorithm, *Expert Syst. Appl.* **245** (2024) 123023.

[9] S. Gao, J. Wu and J. Ai, Multi-UAV reconnaissance task allocation for heterogeneous targets using grouping ant colony optimization algorithm, *Soft Comput.* **25** (2021) 7155–7167.

[10] Z. Zhen, D. Xing and C. Gao, Cooperative search-attack mission planning for multi-uav based on intelligent self-organized algorithm, *Aerosp. Sci. Technol.* **76** (2018) 402–411.

[11] S. Shao, Y. Peng, C. He and Y. Du, Efficient path planning for UAV formation via comprehensively improved particle swarm optimization, *ISA Trans.* **97** (2020) 415–430.

[12] C. Huang, X. Zhou, X. Ran, J. Wang, H. Chen and W. Deng, Adaptive cylinder vector particle swarm optimization with differential evolution for UAV path planning, *Eng. Appl. Artif. Intell.* **121** (2023) 105942.

[13] L. P. Kaelbling, M. L. Littman and A. W. Moore, Reinforcement learning: A survey, *J. Artif. Intell. Res.* **4** (1996) 237–285.

[14] X. Liu, Y. Liu and Y. Chen, Reinforcement learning in multiple-UAV networks: Deployment and movement design, *IEEE Trans. Veh. Technol.* **68**(8) (2019) 8036–8049.

[15] C. J. C. H. Watkins and P. Dayan, Q-learning, *Mach. Learn.* **8**(3) (1992) 279–292.

[16] C. Wang, J. Wang, J. Wang and X. Zhang, Deep-reinforcement-learning-based autonomous UAV navigation with sparse rewards, *IEEE Internet Things J.* **7**(7) (2020) 6180–6190.

[17] A. Puente-Castro, D. Rivero, A. Pazos and E. Fernandez-Blanco, A review of artificial intelligence applied to path planning in UAV swarms, *Neural Comput. Appl.* **34**(1) (2022) 153–170.

[18] Y. Bu, Y. Yan and Y. Yang, Advancement challenges in UAV swarm formation control: A comprehensive review, *Drones* **8**(7) (2024) 320.

[19] J. H. Chan, K. Liu, Y. Chen, A. S. M. S. Sagar and Y.-G. Kim, Reinforcement learning-based drone simulators: Survey, practice, and challenge, *Artif. Intell. Rev.* **57**(10) (2024) 281.

[20] D. J. Richter, R. A. Calix and K. Kim, A review of reinforcement learning for fixed-wing aircraft control tasks, *IEEE Access* **12** (2024) 103026–103048.

[21] M. L. Puterman, Markov decision processes, in *Handbooks in Operations Research and Management Science*, Vol. 2 (1990), pp. 331–434.

[22] R. S. Sutton, Learning to predict by the methods of temporal differences, *Mach. Learn.* **3**(1) (1988) 9–44.

[23] C. Liu, X. Xu and D. Hu, Multiobjective reinforcement learning: A comprehensive overview, *IEEE Trans. Syst. Man Cybern.: Syst.* **45**(3) (2015) 385–398.

[24] J. Jia and W. Wang, Review of reinforcement learning research, in *2020 35th Youth Academic Annual Conf. Chinese Association of Automation (YAC)* (IEEE, 2020), pp. 186–191.

[25] A. Konar, I. Goswami Chakraborty, S. J. Singh, L. C. Jain and A. K. Nagar, A deterministic improved q-learning for path planning of a mobile robot, *IEEE Trans. Syst. Man Cybern.: Syst.* **43**(5) (2013) 1141–1153.

[26] X. Jiang, K. Shi, Y. Liu and Y. Liu, Intelligent control strategy of urban rail train based on sarsa(λ) algorithm, in *2024 4th Int. Conf. Neural Networks, Information and Communication Engineering (NNICE)* (IEEE, 2024), pp. 1615–1618.

[27] Y. Qin, X. Liu and Y. Mao, Multilabel decoupling diagnosis for compound faults by improved deep q-network, *IEEE Trans. Instrum. Meas.* **73** (2024) 1–10.

[28] S. Shahnooshi, J. Ebrahimi and A. Bakhshai, A reinforcement learning controller based on double DQN for DC microgrids with constant power loads, in *2024 IEEE Canadian Conf. Electrical and Computer Engineering (CCECE)* (IEEE, 2024), pp. 376–380.

[29] X. Tao and A. S. Hafid, Deepsensing: A novel mobile crowdsensing framework with double deep q-network and prioritized experience replay, *IEEE Internet Things J.* **7**(12) (2020) 11547–11558.

[30] M. Tang and V. W. Wong, Deep reinforcement learning for task offloading in mobile edge computing systems, *IEEE Trans. Mob. Comput.* **21**(6) (2022) 1985–1997.

[31] R. Bellman, Dynamic programming and lagrange multipliers, *Proc. Natl. Acad. Sci.* **42**(10) (1956) 767–769.

[32] R. Coulom, Efficient selectivity and backup operators in Monte-Carlo tree search, in *Int. Conf. Computers and Games* (Springer, Berlin, Heidelberg, 2006), pp. 72–83.

[33] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Mach. Learn.* **8**(3) (1992) 229–256.

[34] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross and V. Goel, Self-critical sequence training for image captioning, in *2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), pp. 1179–1195.

[35] Q. Lou, Y. Zhou and X. Li, Multi-UAV optimal formation control via actor-critic reinforcement learning algorithm, in *2022 4th Int. Conf. Industrial Artificial Intelligence (IAI)* (IEEE, 2022), pp. 1–6.

[36] H.-C. Jang, Y.-C. Huang and H.-A. Chiu, A study on the effectiveness of A2C and A3C reinforcement learning in parking space search in urban areas problem, in *2020 Int. Conf. Information and Communication Technology Convergence (ICTC)* (IEEE, 2020), pp. 567–571.

[37] H. Lei, H. Ran, I. S. Ansari, K.-H. Park, G. Pan and M.-S. Alouini, DDPG-based aerial secure data collection, *IEEE Trans. Commun.* **72**(8) (2024) 5179–5193.

[38] H. Wang, Z. Zhou, J. Jiang, W. Deng and X. Chen, Autonomous air combat maneuver decision-making based on PPO-BWDA, *IEEE Access* **12** (2024) 119116–119132.

[39] Q. Gu, Z. An, L. Chen and K. Wang, Path planning of UAV using step-wise deep q-learning algorithm, in *2022 8th Annual Int. Conf. Network and Information Systems for Computers (ICNISC)* (IEEE, 2022), pp. 295–297.

[40] Q. Jin, X. Gao, Z. Guo and Z. Hou, Autonomous maneuver decision of UAV in air combat based on scenario-transfer deep reinforcement learning, in *Proc. 2021 Int. Conf. Autonomous Unmanned Systems (ICAUS 2021)*, eds. M. Wu, Y. Niu, M. Gu and J. Cheng (Springer, 2022), pp. 2616–2625.

[41] W. Su, M. Gao, X. Gao, X. Zhu and D. Fang, Adaptive decision-making with deep q-network for heterogeneous unmanned aerial vehicle swarms in dynamic environments, *Comput. Electr. Eng* **119** (2024) 109621.

[42] M. Sun, Z. Yang, X. Dai, X. Nian, H. Wang and H. Xiong, Deep reinforcement learning based on curriculum learning for drone swarm area defense, in *Proc. 2022 Int. Conf. Autonomous Unmanned Systems (ICAUS 2022)*, eds. W. Fu, M. Gu and Y. Niu (Springer Nature, 2023), pp. 1119–1128.

[43] J. Tang, Y. Liang and K. Li, Dynamic scene path planning of UAVs based on deep reinforcement learning, *Drones* **8**(2) (2024) 60.

[44] X. Yu, X. Zhou and Y. Zhang, Collision-free trajectory generation for UAVs using Markov decision process, in *2017 Int. Conf. Unmanned Aircraft Systems (ICUAS)* (IEEE, 2017), pp. 56–61.

[45] M. Zhang, C. Yan, W. Dai, X. Xiang and K. H. Low, Tactical conflict resolution in urban airspace for unmanned aerial vehicles operations using attention-based deep reinforcement learning, *Green Energy Intell. Transp.* **2**(4) (2023) 100107.

[46] N. Salvatore, S. Mian, C. Abidi and A. D. George, A neuro-inspired approach to intelligent collision avoidance and navigation, in *2020 AIAA/IEEE 39th Digital Avionics Systems Conf. (DASC)* (IEEE, 2020), pp. 1–9.

[47] Y. Lu, Z. Xue, G.-S. Xia and L. Zhang, A survey on vision-based UAV navigation, *Geo-Spat. Inf. Sci.* **21**(1) (2018) 21–32.

[48] Z. Yijing, Z. Zheng, Z. Xiaoyi and L. Yang, Q learning algorithm based UAV path learning and obstacle avoidence approach, in *2017 36th Chinese Control Conf. (CCC)* (Dalian, China, 2017), pp. 3397–3402.

[49] A. Faust, I. Palunko, P. Cruz, R. Fierro and L. Tapia, Automated aerial suspended cargo delivery through reinforcement learning, *Artif. Intell.* **247** (2017) 381–398, Special Issue on AI and Robotics.

[50] Y. Li, M. Li, A. Sanyal, Y. Wang and Q. Qiu, Autonomous UAV with learned trajectory generation and control, in *2019 IEEE Int. Workshop on Signal Processing Systems (SiPS)* (Nanjing, China, 2019), pp. 115–120.

[51] J. Ou, X. Guo, M. Zhu and W. Lou, Autonomous quadrotor obstacle avoidance based on dueling double deep recurrent q-learning with monocular vision, *Neurocomputing* **441** (2021) 300–310.

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

[53] F. AlMahamid and K. Grolinger, Autonomous unmanned aerial vehicle navigation using reinforcement learning: A systematic review, *Eng. Appl. Artif. Intell.* **115** (2022) 105321.

[54] N. Salvatore, S. Mian, C. Abidi and A. D. George, A neuro-inspired approach to intelligent collision avoidance and navigation, in *IEEE Digital Avionics Systems Conf.* (IEEE, 2020), pp. 1–9.

[55] Y. Niu, X. Yan, Y. Wang and Y. Niu, Three-dimensional collaborative path planning for multiple UCAVs based on improved artificial ecosystem optimizer and reinforcement learning, *Knowl.-Based Syst.* **276** (2023) 110782.

[56] A. Sonny, S. R. Yeduri and L. R. Cenkeramaddi, Q-learning-based unmanned aerial vehicle path planning with dynamic obstacle avoidance, *Appl. Soft Comput.* **147** (2023) 110773.

[57] B. Zhang, Z. Mao, W. Liu and J. Liu, Geometric reinforcement learning for path planning of UAVs, *J. Intell. Robot. Syst.* **77**(2) (2015) 391–409.

[58] T. Zhang, X. Huo, S. Chen, B. Yang and G. Zhang, Hybrid path planning of a quadrotor UAV based on q-learning algorithm, in *2018 37th Chinese Control Conf. (CCC)* (Wuhan, China, 2018), pp. 5415–5419.

[59] M. R. Jones, S. Djahel and K. Welsh, An efficient and rapidly adaptable lightweight multi-destination urban path planning approach for UAVs using q-learning, *IEEE Trans. Intell. Veh.* (2024) 1–12.

[60] A. Puente-Castro *et al.*, Q-learning based system for path planning with unmanned aerial vehicles swarms in obstacle environments, *Expert Syst. Appl.* **235** (2024) 121240.

[61] Z. Wang, W. Gao, G. Li, Z. Wang and M. Gong, Path planning for unmanned aerial vehicle via off-policy reinforcement learning with enhanced exploration, *IEEE Trans. Emerg. Top. Comput. Intell.* **8**(3) (2024) 2625–2644.

[62] B. Silvirianti, B. Narottama and S.-Y. Shin, UAV coverage path planning with quantum-based recurrent deep deterministic policy gradient, *IEEE Trans. Veh. Technol.* **73**(5) (2024) 7424–7429.

[63] M. Xi, H. Dai, J. He, W. Li, J. Wen, S. Xiao and J. Yang, A lightweight reinforcement-learning-based real-time path-planning method for unmanned aerial vehicles, *IEEE Internet Things J.* **11**(12) (2024) 21061–21072.

[64] G. Lee, K. Kim and J. Jang, Real-time path planning of controllable UAV by subgoals using goal-conditioned reinforcement learning, *Appl. Soft Comput.* **146** (2023) 110660.

[65] J. Peng, Z. Zhang, Q. Wu and B. Zhang, Anti-jamming communications in UAV swarms: A reinforcement learning approach, *IEEE Access* **7** (2019) 180532–180543.

[66] C. Xu, K. Zhang and H. Song, UAV swarm communication aware formation control via deep q network, in *2020 IEEE 39th Int. Performance Computing and Communications Conf. (IPCCC)* (Austin, TX, USA, 2020), pp. 1–2.

[67] N. Zhao, Z. Ye, Y. Pei, Y.-C. Liang and D. Niyato, Multi-agent deep reinforcement learning for task offloading in UAV-assisted mobile edge computing, *IEEE Trans. Wirel. Commun.* **21**(9) (2022) 6949–6960.

[68] S. Liu and Y. Bai, Multiple UAVs collaborative traffic monitoring with intention-based communication, *Comput. Commun.* **210** (2023) 116–129.

[69] G. Zhan, X. Zhang, Z. Li, L. Xu, D. Zhou and Z. Yang, Multiple-UAV reinforcement learning algorithm based on improved PPO in ray framework, *Drones* **6**(7) (2022) 166.

[70] L. Zuo, S. Gao, Y. Li, L. Li, M. Li and X. Lu, A fast and robust algorithm with reinforcement learning for large UAV cluster mission planning, *Remote Sens.* **14**(6) (2022) 1304.

[71] Z. Dai, Y. Zhang, W. Zhang, X. Luo and Z. He, A multi-agent collaborative environment learning method for UAV deployment and resource allocation, *IEEE Trans. Signal Inf. Process. Netw.* **8** (2022) 1–16.

[72] Y. Dong, J. Ai and J. Liu, Guidance and control for own aircraft in the autonomous air combat: A historical review and future prospects, *J. Aerosp. Eng.* **233** (2019) 5943–5991.

[73] A. T. Azar *et al.*, Drone deep reinforcement learning: A review, *Electronics* **10**(9) (2021) 999.

[74] A. Din, I. Mir and S. Faiza, Development of reinforced learning based non-linear controller for unmanned aerial vehicle, *J. Ambient Intell. Humaniz. Comput.* **14** (2022) 4005–4022.

[75] R. Chai, A. Tsourdos, A. Savvaris and S. Chai, Review of advanced guidance and control algorithms for space/aerospace vehicles, *Prog. Aerosp. Sci.* **122** (2021) 100696.

[76] X. Wang, Y. Wang, X. Su, L. Wang, C. Lu, H. Peng and J. Liu, Deep reinforcement learning-based air combat maneuver decision-making: Literature review, implementation tutorial and future direction, *Artif. Intell. Rev.* **57**(1) (2024).

[77] Y. Chen, J. Zhang, Q. Yang, Y. Zhou, G. Shi and Y. Wu, Design and verification of UAV maneuver decision simulation system based on deep q-learning network, in *2020 16th Int. Conf. Control, Automation, Robotics and Vision (ICARCV)* (Shenzhen, China, 2020), pp. 817–823.

[78] B. Vlahov, E. Squires, L. Strickland and C. Pippin, On developing a UAV pursuit-evasion policy using reinforcement learning, in *2018 17th IEEE Int. Conf. Machine Learning and Applications (ICMLA)* (Orlando, FL, USA, 2018), pp. 859–864.

[79] B. Li, J. Huang, S. Bai, Z. Gan, S. Liang, N. Evgeny and S. Yao, Autonomous air combat decision-making of UAV based on parallel self-play reinforcement learning, *CAAI Trans. Intell. Technol.* **8**(1) (2023) 64–81.

[80] J. Zhang, Q. Yang, G. Shi, Y. Lu and Y. Wu, UAV cooperative air combat maneuver decision based on multi-agent reinforcement learning, *J. Syst. Eng. Electron.* **32**(6) (2021) 1421–1438.

[81] F. Jiang, M. Xu, Y. Li, H. Cui and R. Wang, Short-range air combat maneuver decision of UAV swarm based on multi-agent transformer introducing virtual objects, *Eng. Appl. Artif. Intell.* **123** (2023) 106358.

[82] H. Wang and J. Wang, Enhancing multi-UAV air combat decision making via hierarchical reinforcement learning, *Sci. Rep.* **14**(1) (2024) 4458.

[83] D. Liu, Q. Zong, X. Zhang, R. Zhang, L. Dou and B. Tian, Game of drones: Intelligent online decision making of multi-UAV confrontation, *IEEE Trans. Emerg. Top. Comput. Intell.* **8**(2) (2024) 2086–2100.

[84] M. Özbek, S. Yıldırım, M. Aksoy, E. Kernin and E. Koyuncu, Harfang3d dog-fight sandbox: A reinforcement learning research platform for the customized control tasks of fighter aircrafts, preprint (2022), arXiv:2210.07282.

[85] A. Din, I. Mir, F. Gul and S. Mir, Non-linear intelligent control design for unconventional unmanned aerial vehicle, in *AIAA Scitech 2023 Forum*, National Harbor, MD, 2023.

[86] E. Bøhn, E. M. Coates, D. Reinhardt and T. A. Johansen, Data-efficient deep reinforcement learning for attitude control of fixed-wing UAVs: Field experiments, *IEEE Trans. Neural Netw. Learn. Syst.* **35**(3) (2024) 3168–3180.

[87] J. Hu, L. Wang, T. Hu, C. Guo and Y. Wang, Autonomous maneuver decision making of dual-UAV cooperative air combat based on deep reinforcement learning, *Electronics* **11**(3) (2022) 467.

[88] D. Hu, R. Yang, J. Zuo, Z. Zhang, J. Wu and Y. Wang, Application of deep reinforcement learning in maneuver planning of beyond-visual-range air combat, *IEEE Access* **9** (2021) 32282–32297.

[89] Q. Yang, J. Zhang, G. Shi, J. Hu and Y. Wu, Maneuver decision of UAV in short-range air combat based on deep reinforcement learning, *IEEE Access* **8** (2020) 363–378.

[90] L. Wu, C. Wang, P. Zhang and C. Wei, Deep reinforcement learning with corrective feedback for autonomous UAV landing on a mobile platform, *Drones* **6**(9) (2022) 238.

[91] X. Wang, H. Peng, J. Liu, X. Dong, X. Zhao and C. Lu, Optimal control based coordinated taxiing path planning and tracking for multiple carrier aircraft on flight deck, *Def. Technol.* **18**(2) (2022) 238–248.

[92] L. Yue, R. Yang, Y. Zhang, L. Yu and Z. Wang, Deep reinforcement learning for UAV intelligent mission planning, *Complexity* **2022** (2022) 1–13.

[93] B. Li, Z. Gan, D. Chen and S. Damjanovic, UAV maneuvering target tracking in uncertain environments based on deep reinforcement learning and meta-learning, *Remote Sens.* **12**(22) (2020) 3789.

[94] H. Zhang, H. Zhou, Y. Wei and C. Huang, Autonomous maneuver decision-making method based on reinforcement learning and Monte Carlo tree search, *Front. Neurorobot.* **16** (2022) 996412.

[95] K. Yang, W. Dong, M. Cai, S. Jia and R. Liu, UCAV air combat maneuver decisions based on a proximal policy optimization algorithm with situation reward shaping, *Electronics* **11**(16) (2022) 2602.

[96] J. Yoo, H. Seong, D. Shim, J. Bae and Y. Kim, Deep reinforcement learning-based intelligent agent for autonomous air combat, in *2022 IEEE/AIAA 41st Digital Avionics Systems Conf. (DASC)* (IEEE, 2022), pp. 1–9.

[97] Z. Wang, H. Li, H. Wu and Z. Wu, Improving maneuver strategy in air combat by alternate freeze games with a deep reinforcement learning algorithm, *Math. Probl. Eng.* **2020** (2020).

[98] N. Jiang, S. Jin and C. Zhang, Hierarchical automatic curriculum learning: Converting a sparse reward navigation task into dense reward, *Neurocomputing* **360** (2019) 265–278.

[99] Y. Hou, X. Liang, M. Lv, Q. Yang and Y. Li, Subtask-masked curriculum learning for reinforcement learning with application to UAV maneuver decision-making, *Eng. Appl. Artif. Intell.* **125** (2023) 106703.

[100] R. Wu, F. Gu and J. Huang, A multi-critic deep deterministic policy gradient UAV path planning, in *2020 16th Int. Conf. Computational Intelligence and Security (CIS)* (Guangxi, China, 2020), pp. 6–10.

[101] B. Silvirianti, B. Narottama and S.-Y. Shin, Layerwise quantum deep reinforcement learning for joint optimization of UAV trajectory and resource allocation, *IEEE Internet Things J.* **11**(1) (2024) 430–443.

[102] J. Hu, M. Guo, S. Yan, Y. Chen, X. Zhou and Z. Chen, Deep reinforcement learning enabled covert transmission with UAV, *IEEE Wirel. Commun. Lett.* **12**(5) (2023) 917–921.

[103] M. Chansuparp and K. Jitkajornwanich, A novel augmentative backward reward function with deep reinforcement learning for autonomous UAV navigation, *Appl. Artif. Intell.* **36**(1) (2022) 2084473.

[104] Y. Li, X. Zhang, Y. Zhu and Z. Gao, A UAV path planning method in three-dimensional urban airspace based on safe reinforcement learning, in *2023 IEEE/AIAA 42nd Digital Avionics Systems Conf. (DASC)* (Barcelona, Spain, 2023), pp. 1–7.

Jiahao Fu received M.S. degree from the Anhui University of Science and Technology, Huainan China in 2023. He is currently pursuing the Ph.D. degree with the National Elite Institute of Engineering, Northwestern Polytechnical University, Xian, China. His research interests include decision-making and reinforcement learning.

Feng Yang received his B.E., M.E. and Ph.D. degrees from the Northwestern Polytechnical University (NPU), Xian, China, in 2000, 2003 and 2007, respectively. He is currently Associate Professor at the School of Automation, NPU. His research interests include information fusion, target tracking, and radar data processing.