Computer Science and Application 计算机科学与应用, 2026, 16(5), 95-109
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165168

Hans汉斯

基于强化学习的三维无人机路径规划综述

张楚寒[1]，姜福宏[1]，王梦嫒[2*]，程　超[3]

[1]魏桥国科(滨州)科技有限公司，山东　滨州
[2]魏桥国科(北京)科技有限公司，北京
[3]滨州魏桥国科高等技术研究院，山东　滨州

收稿日期：2026年4月15日；录用日期：2026年5月13日；发布日期：2026年5月22日

摘　要

传统无人机路径规划算法依赖精确环境模型，在复杂动态三维环境中存在适应性差、实时性差等明显局限性。文章首先阐述了传统路径规划算法的不足，引入深度强化学习作为解决该问题的全新技术路径。其次全面梳理值函数、策略梯度及值－策略混合三大类典型强化学习算法，深入探讨各类算法的核心原理，从单机与多机两个维度，系统总结了强化学习在三维无人机路径规划中的改进成果与应用进展。最后，聚焦无人机实际飞行场景的独特挑战，明确强化学习的应用瓶颈并总结未来发展方向，为该领域的理论研究与工程实践提供系统性参考。

关键词

无人机，路径规划，强化学习，策略梯度

A Review of 3D UAV Path Planning Based on Reinforcement Learning

Chuhan Zhang[1], Fuhong Jiang[1], Mengbi Wang[2*], Chao Cheng[3]

[1]Weiqiao Guoke (Binzhou) Technology Co., Ltd., Binzhou Shandong
[2]Weiqiao Guoke (Beijing) Technology Co., Ltd., Beijing
[3]Weiqiao Guoke High-Tech Research Institute, Binzhou Shandong

Received: April 15, 2026; accepted: May 13, 2026; published: May 22, 2026

Abstract

Traditional unmanned aerial vehicle (UAV) path planning algorithms rely heavily on accurate environmental models and exhibit significant limitations, including poor adaptability and inadequate

*通讯作者。

文章引用: 张楚寒, 姜福宏, 王梦嫒, 程超. 基于强化学习的三维无人机路径规划综述[J]. 计算机科学与应用, 2026, 16(5): 95-109. DOI: 10.12677/csa.2026.165168

real-time performance in complex dynamic three-dimensional (3D) environments. This thesis first elaborates on the inherent limitations of traditional path planning algorithms and introduces deep reinforcement learning as an innovative approach to address these challenges. Subsequently, it comprehensively reviews three major categories of reinforcement learning algorithms—namely value-based, policy gradient-based, and value-policy hybrid methods, delving into their core principles and systematically synthesizing their advancements and application outcomes in 3D UAV path planning from both single-UAV and multi-UAV perspectives. Finally, focusing on the unique challenges of actual UAV flight scenarios, the thesis clarifies the application bottlenecks of reinforcement learning and summarizes the future development directions, providing systematic references for theoretical research and engineering practice in this field.

Keywords

UAV, Path Planning, Reinforcement Learning, Policy Gradient

1. 引言

无人机具有垂直起降能力和高机动性[1]，已被广泛应用于军事侦察、灾害救援等多个领域。在此背景下，高效的路径规划技术成为提升无人机自主作业能力的关键。然而，三维环境中的路径规划不仅要考虑传统二维规划中的避障问题，还需应对复杂地形约束、动态障碍物、多机协同等多重挑战，特别是在森林、城市、洞穴等复杂环境中，如图 1 所示。

Figure 1. 3D path planning diagram for UAVs in complex environments
图 1. 复杂环境下无人机三维路径规划图

传统无人机路径规划算法主要分为三类[2]：1) 图搜索算法：常见的有 A*算法[3]和 Dijkstra 算法[4]，将环境离散成网格，通过启发式函数或代价累加在图中寻找全局最优路径，仅适用于静态、完全已知场景，且计算量大。2) 随机采样算法：如快速遍历随机树算法(Rapidly-Exploring Random Tree, RRT) [5]与概率路线图(Probabilistic Road Map Method, PRM) [6]，通过概率采样生成可行轨迹，但存在稳定性差等不足。3) 数值优化：通过构建目标函数，结合无人机动力学等约束条件，采用数学优化方法求解最优解，以实现路径规划的动态响应与约束满足，例如模型预测控制(Model Predictive Control, MPC) [7]。传统算法均依赖精确环境模型，将路径规划视为已知模型下的优化问题。然而，无人机实际飞行环境具有部分可观测、非线性、强耦合等特点，传统方法难以满足实时性与适应性要求。强化学习为突破上述困境提供了全新的解决方法。作为数据驱动的决策方法，强化学习无需依赖精确的环境模型，而是通过智能体与环境之间的互动来进行自我优化。这种方法具有高动态环境适应性、泛化能力强和灵活性高等特点[8]-[10]。

尽管现有文献已对无人机路径规划算法进行了广泛梳理，但专门针对强化学习在该领域应用的综述仍鲜有涉及，也缺乏对不同任务场景下算法适配性的系统分析。为弥补这一不足，本文聚焦于三维无人机路径规划，梳理了强化学习算法的演进历程，重点剖析其在单机与多机场景中的改进策略与应用进展，明确各类算法的适用边界，为该领域的后续研究提供结构化参考。

2. 路径规划定义与强化学习基础框架

2.1. 三维无人机路径规划定义

无人机路径规划是在满足动力学、避障、能耗、通信等约束下，为无人机集群规划最优三维路径技术，具备部分可观测、多智能体交互、多目标优化、动态不确定等特点，该问题可建模为部分可观测马尔可夫决策过程，其数学表达如下：

$$P_{ss'}^{a} = E\left(S_{t+1} = s' \mid S_t = s, A_t = a\right) \tag{1}$$

其中，状态空间 S 包含环境与无人机状态，动作空间 A 为无人机机动指令集。

2.2. 强化学习基本框架

强化学习是一种面向序贯决策问题的机器学习方法，智能体通过与环境交互学习动作策略以最大化累积奖励[11]。强化学习的框架由智能体、环境、状态空间、动作空间、奖励函数及策略函数构成。

2.2.1. 状态空间

在无人机路径规划的强化学习模型中，状态空间 A_t 是智能体感知环境并做出决策的输入依据，主要包含三个方面：

1) 无人机本体状态：位置、姿态、速度、加速度、剩余能量等；
2) 环境感知状态：地形、固定障碍物的坐标与轮廓、禁飞区域边界、航点信息等；
3) 任务和目标状态：主要涵盖目标点位置、任务优先级、执行进度、通信链路状态等。

2.2.2. 动作空间

动作选择是强化学习中的关键要素，无人机依据不同策略来选择下一步动作，根据即时奖励持续优化策略，以最大化长期累积回报。无人机采用 27 个方向的动作，包含当前网格平面及上、下共三个平面的运动方向，如图 2 所示。

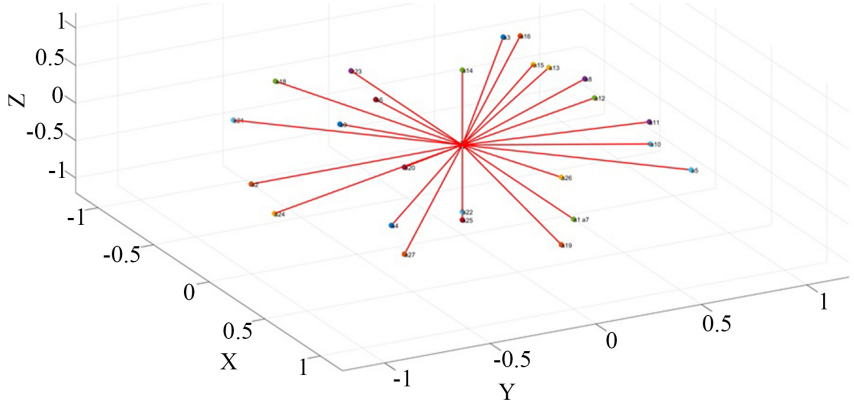

Figure 2. Action selection diagram corresponding to the plane where the drone is located
图 2. 无人机所在平面对应的动作选择示意图

2.2.3. 奖励函数

奖励函数是强化学习中的核心组成部分，决定了智能体的行为目标和学习方向。奖励函数 R 负责引导无人机的行为，使其尽可能高效地达成最终目标，同时避免发生碰撞。对于每架无人机 i 而言奖励可以定义为以下公式：

$$R_t^i = r_{target}^i - r_{collision}^i - r_{effort}^i \tag{2}$$

r_{target}^i 是无人机缩短目标距离给予的正向奖励，与 $\left\| p_t^i - p_{target}^i \right\|$ 成正比；$r_{collision}^i$ 是与障碍物或其他无人机发生碰撞时的惩罚，r_{effort}^i 抑制不必要的能量消耗。

2.2.4. 策略函数

策略函数决定了智能体在特定状态下的动作选择。在无人机路径规划中，策略函数 $\pi(a|s)$ 的优化目标使无人机能够在动态和不确定的环境中学习到最优的路径规划方案。通过不断与环境交互，智能体根据奖励函数的反馈来更新策略函数，逐步提高其决策的准确性和适应性，如下式所列：

$$\pi^*(a|s) = \arg\max_{\pi} E_{\pi}\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{3}$$

其中，r_t 是在时间步长 t 获得的奖励；γ 是折扣因子 $(0 \le \gamma \le 1)$，用于平衡即时奖励和未来奖励的重要性。

3. 无人机路径规划中的强化学习典型算法

本章针对性梳理应用在三维无人机路径规划中的强化学习算法，并根据学习机制分为值函数、策略梯度和值–策略混合三大类进行详细介绍。图 3 为强化学习算法的发展脉络图，清晰呈现了各类强化学习算法的演进历程与分类关联，其中以蓝色标注的算法为本文重点介绍，其也是适用于无人机三维路径规划场景的核心算法。

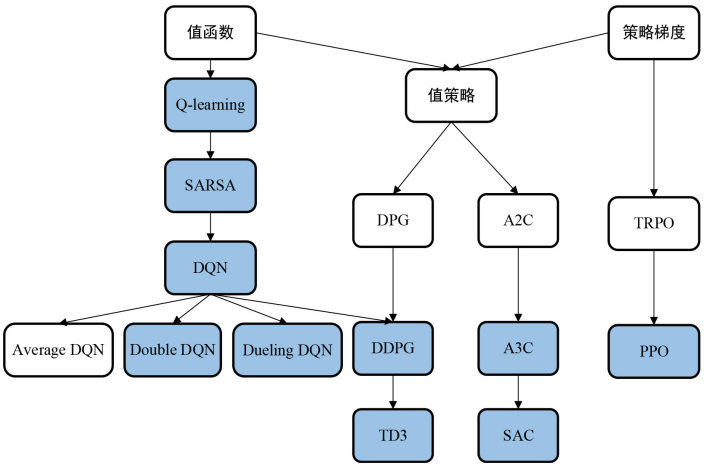

Figure 3. Evolutionary path of reinforcement learning algorithms
图 3. 强化学习算法演进脉络

3.1. 基于值函数的方法

在强化学习的早期应用中值函数方法是主流，其核心思想是根据价值函数的大小推导出最优策略，进而选择动作，典型算法包括 SARSA (State-Action-Reward-State-Action) [12]、DQN (Deep Q-Network) [13] 等。

3.1.1. Q-Learning

Q-Learning [14]是一种经典的基于价值函数的算法，核心思想并非优化已有策略，而是构建全新策略，采用表格形式(Q 表)存储 Q 值，该 Q 表每个状态对应一行，每个动作对应一列，单元格存储相应状态－动作对的估计 Q 值。其输入是状态和动作，输出是各个状态和动作的价值，通过不断更新表格来学习如何在一个给定的环境中执行最佳动作，最终得到一个训练较好的 Q 表。更新公式如式(4)所示。算法具体流程见图 4。

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r+\gamma\max_{a_{t+1}}Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\right] \tag{4}$$

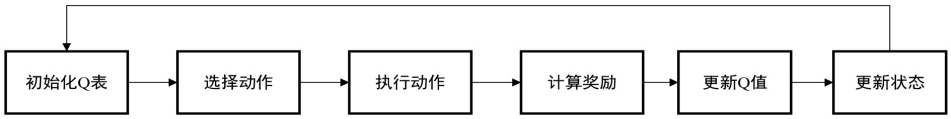

Figure 4. Flowchart of the Q-Learning algorithm
图 4. Q-Learning 算法流程图

然而，Q-Learning 存在以下局限：首先，更新 Q 值时采用下一时刻最优动作的 Q 值，易产生估计偏差，导致 Q 值过高估计；其次，状态空间较大时，Q 表的存储与检索开销显著增加。此外，状态维度增加使状态－动作对呈指数增长，导致可用于学习的样本变得稀疏，进而影响算法的学习性能。

3.1.2. SARSA

SARSA 是一种在线策略型强化学习方法，以时序交互序列为学习依据，凭借其优异的泛化性能得到广泛应用[15]-[17]。SARSA 依据无人机当前真实执行的飞行动作进行价值评估，不会盲目追求高奖励而选择危险动作，在路径规划中体现出天然的安全保守特性，更适合对安全性要求较高的自主飞行场景。如图 5 给出了 SARSA 算法的示意图，智能体在状态 s 下采用当前的行为策略产生一个新动作 a，智能体此时并不执行该动作，而是通过动作价值函数得到后一个状态动作对 (s',a') 的价值，利用这个新的价值和即时奖励 r 来更新前一个状态动作对 (s',a') 的价值[18]，更新公式如下：

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_{t+1}+\gamma Q(s_{t+1},a_{t+1})-Q(s_t,a_t)\right] \tag{5}$$

虽然 SARSA 的在线策略更安全、碰撞风险低；但仅支持离散动作，收敛慢、易陷入局部最优。

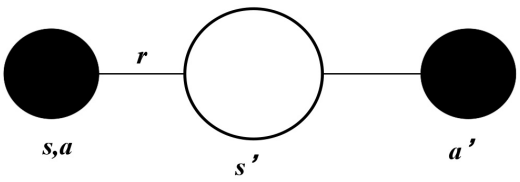

Figure 5. The structure diagram of the SARSA algorithm
图 5. SARSA 算法结构图

3.1.3. DQN

2015 年 DeepMind 团队将 DQN 首次应用于 Atari 游戏并获得突破[19]，该算法以 Q-learning 为基础，融合值函数与深度神经网络，以神经网络替代 Q 表，引入经验回放和固定 Q 目标网络解决训练不稳定的问题，通过迭代更新 Q 网络学习最优策略，其更新公式和损失函数如式(6)、式(7)所示，结构图如图 6 所示。

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\left(r_t+\gamma\max\left(Q(s_{t+1},a_{t+1})\right)-Q(s_t,a_t)\right) \tag{6}$$

$$L(\theta)=E\left[\left(r_t+\gamma^{*}\max\left(Q\left(s_{t+1},a_{t+1},\theta^{-}\right)\right)-Q\left(s_t,a_t,\theta\right)\right)^{2}\right] \tag{7}$$

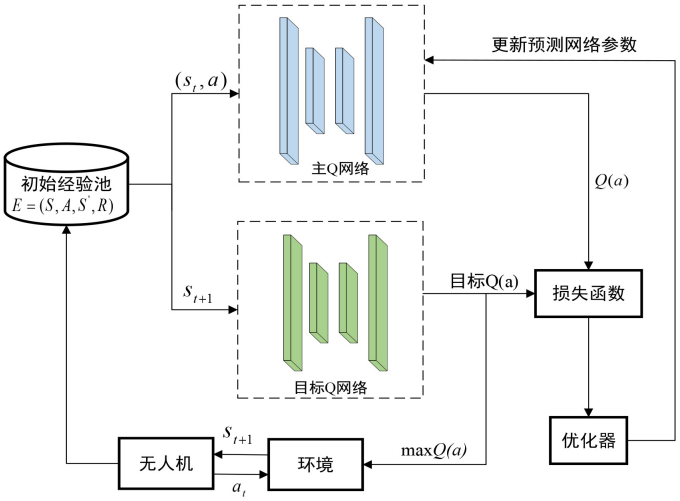

Figure 6. The structure diagram of DQN algorithm
图 6. DQN 算法结构图

DQN 凭借深度网络、经验回放和目标网络首次突破高维应用局限，为无人机三维路径规划奠定基础，然而，其固有的 Q 值过估计及策略评估效率不足等问题，限制了在复杂动态场景下的直接应用。为此，研究者提出 DDQN、Dueling DQN 等改进算法，下文将对此进行具体分析。

3.1.4. DDQN

DDQN (Double Deep Q-Network)由 Van Hasselt 等人于 2016 年提出[20] [21]，其核心思路是将动作选择与价值评估分离，通过两个独立的神经网络分别实现这两个功能，可有效缓解传统 DQN 在无人机路径规划中普遍存在的 Q 值过估计问题，避免因高估危险飞行动作价值而引发碰撞风险，但网络解耦设计复杂。DDQN 网络结构与 DQN 网络结构相同，表达式如下。

$$a_{t+1}=\arg\max_{a_{t+1}}Q\left(s_{t+1},a_{t+1},\theta\right) \tag{8}$$

$$y_t^{DDQN}=r+\lambda Q^{-}\left(s_{t+1},\arg\max_{a+1}Q\left(s_{t+1},a_{t+1},\theta\right),\theta^{-}\right) \tag{9}$$

3.1.5. Dueling DQN

Dueling DQN [22]在网络结构上采用对抗网络实现状态值与动作优势的解耦[23]，其输入与 DQN 相同，均为状态信息，但对路径决策的价值输出方式更贴合无人机自主飞行特性，二者结构对比如图 7 所示。Dueling DQN 将输出拆分为表征状态重要程度的状态值 Value [24]，以及对应各动作优劣 Advantage。Value 可以用于表征无人机当前所处空域的安全程度、可飞行性等环境基础价值；另一部分 Advantage 用于评估不同飞行动作，如前进、转向、爬升、下降等对路径优劣的影响程度。最终通过公式(10)将两部分融合得到完整 Q 值，使无人机更关注空域本身的安全性，同时合理区分机动动作的优劣，提升复杂环境下路径决策的效率与稳定性。

$$Q\left(s,a;\theta,\alpha,\beta\right)=V\left(s;\theta,\beta\right)+\left(A\left(s,a;\theta,\alpha\right)-\max_{a_{t+1}\in|A|}A\left(s,a_{t+1};\theta,\alpha\right)\right) \tag{10}$$

α 和 β 分别表示 Advantage 和 Value 函数全连接层的参数。与 DQN 和 DDQN 相比，Dueling DQN 的策略评估更高效，但网络结构实现较复杂。

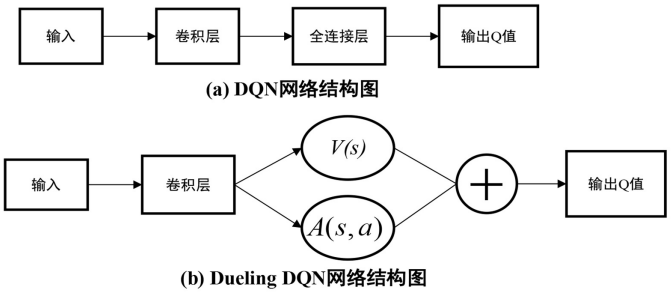

Figure 7. The structure diagram of Dueling DQN algorithm
图 7. Dueling DQN 算法结构图

3.2. 基于策略梯度的方法

值函数方法适合离散动作空间，但无人机的真实飞行是连续的，而基于策略梯度的方法则直接输出动作概率分布，为无人机提供了精细化操作的可能，常见算法代表有 PPO (Proximal Policy Optimization) [25]算法等。

PPO：

在采用策略梯度算法训练智能体时，易因生成劣质路径用于训练导致系统性能骤降、规划出现死锁碰撞，且样本无法复用、利用率低。为解决上述问题，Schulman 等人[26]提出了 PPO 算法(图 8)，通过限制策略更新幅度保障训练稳定高效；引入剪切目标函数替换原目标函数，避免策略过度更新引发的不稳定性，同时结合优势估计提升样本效率[27]，其剪切目标函数如下。

$$L^{PPO}(\phi)=E_t\left[\min\left(r_t(\theta)\hat{A}_t, f_{clip}\left(r_t(\theta),1-\varepsilon,1+\varepsilon\right)\hat{A}_t\right)\right] \tag{11}$$

$$r_t(\theta)=\frac{\pi_\theta(a_t\mid s_t)}{\pi_{old}(a_t\mid s_t)} \tag{12}$$

f_{clip} 为数值限幅函数，超参数 ε 控制剪切幅度，$r_t(\theta)$ 为新旧策略在状态 t 下的动作概率比；A_t 为优势函数，E_t 为时间步长期望。传统 PPO 使用固定的剪切超参数 ε，难以适应动态环境的变化。可设计自适应剪切系数，根据策略更新前后的优势函数方差动态调整 ε 值，在策略波动剧烈时收紧剪切范围，在稳定时适当放宽，从而平衡探索与稳定性。

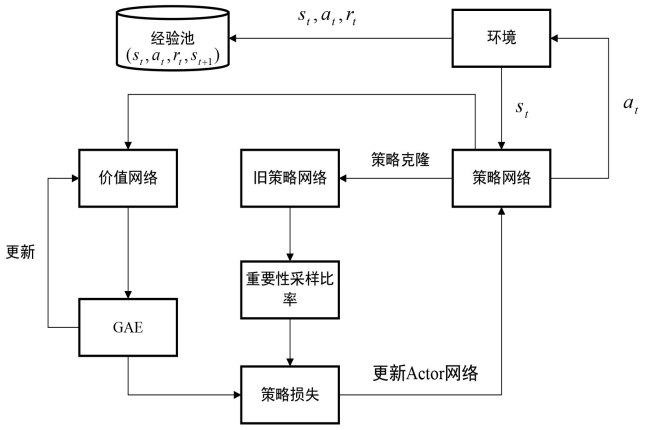

Figure 8. The structure diagram of the PPO algorithm
图 8. PPO 算法结构图

3.3. 基于值－策略混合的方法

价值函数与策略类强化学习方法各有优劣：前者依赖预定义任务价值，需平衡价值与智能体数量；后者受策略空间复杂度与搜索效率制约，易陷局部最优且计算开销大[28]。为融合二者优势，研究者提出 Actor-Critic 算法[29]来更高效地处理连续动作与高维状态空间问题。

3.3.1. DDPG

DDPG (Deep Deterministic Policy Gradient)算法[30]结合确定性策略梯度(DPG) [31]与深度学习，用于解决连续控制问题。它基于 AC 架构并采用 DQN 学习策略，是面向无人机连续动作空间设计的无模型深度强化学习方法，本质为 DNN + DPG [32]，算法框架见图 9。在无人机路径规划中，DDPG 借助经验回放打破样本时序相关性，提升飞行数据利用效率；通过价值网络迭代逼近最优动作价值，并以时序差分误差完成网络更新；同时引入独立目标网络，有效避免价值过估计问题，保障无人机飞行控制策略更新稳定、轨迹平滑安全。在此基础上不断优化策略网络参数，使无人机能够输出连续、平滑的机动指令，实现复杂环境下的精准避障与稳定跟踪。

$$\begin{cases} \nabla_\theta q\left(s_j, \mu\left(s_j;\theta\right);\omega\right) = \nabla_\theta \mu\left(s_j;\theta\right)\cdot\nabla_a q\left(s_j, \hat{a}_j;\omega\right) \\ \theta \leftarrow \theta + \beta\cdot\nabla_\theta \mu\left(s_j;\theta\right)\cdot\nabla_a q\left(s_j, \hat{a}_j;\omega\right) \end{cases} \tag{13}$$

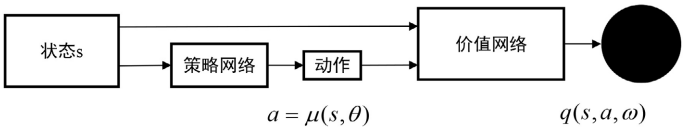

Figure 9. The schematic diagram of the DDPG algorithm
图 9. DDPG 算法示意图

3.3.2. TD3

TD3 (Twin Delayed Deep Deterministic Policy Gradient)算法[33]作为 DDPG 的改进版本，使用了 6 个神经网络，适用于连续动作空间，已广泛应用于无人机追击、近距空战反追击等场景。结构如图 10 所示。

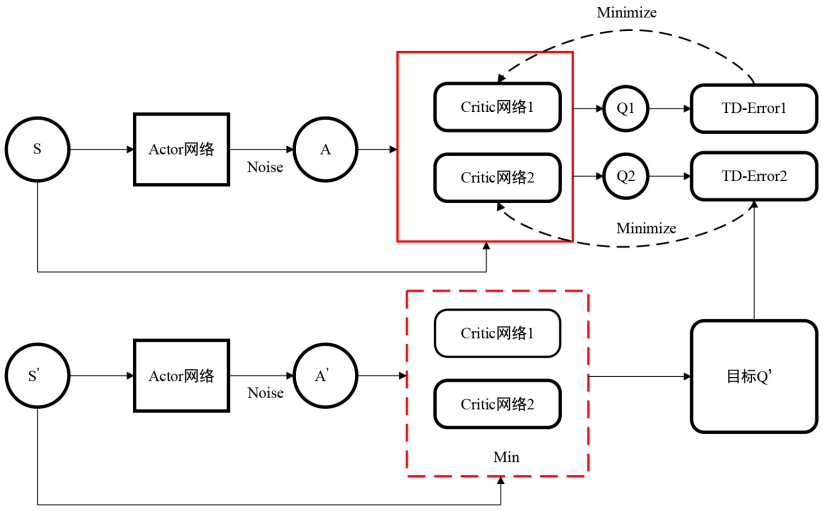

Figure 10. The structure diagram of the TD3 algorithm
图 10. TD3 算法结构图

TD3 算法提出三项关键改进，首先利用双 Q 网络估计价值，选取较小值以削减正向偏差；其次，对策略网络及目标网络实施延迟更新，使其更新步调慢于 Q 网络，从而稳定训练过程；最后，在目标策略中注入噪声，提升动作探索的多样性。其目标 Q 值的计算公式如式(14)所示。这些改进显著增强了 TD3 在高维连续任务中的性能与收敛可靠性。与 DDPG 相比，TD3 稳定性极强，但网络数量多、计算量大，策略偏保守。

$$y = r + \gamma \min_{i=1,2} Q_i\left(s_{t+1}, \mu\left(s_{t+1} \mid \theta^{\mu}\right) \mid \theta^{Q_i}\right) \tag{14}$$

3.3.3. A3C

A3C (Asynchronous Advantage Actor-Critic) [34]是一种通过同时运行多个局域网络异步更新全局网络的算法，尤其适用于无人机这类需要连续、实时、多场景并行探索的自主飞行任务。Actor 网络根据无人机当前飞行状态输出动作概率分布，决策无人机的飞行方向、速度与转弯策略；Critic 网络对当前飞行状态与动作的价值进行评估，用于指导 Actor 网络更新策略。多个 Actor-Critic 代理独立工作，与全局网络交换学习结果，整个过程是异步完成的，使无人机在复杂三维环境中能够更快、更稳定地学习到安全、平滑的最优飞行路径，如图 11 所示。

$$A\left(a_t, s_t; \theta^{-}, \theta_v^{-}\right) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V\left(s_{t+k}; \theta_v^{-}\right) - V\left(s_t; \theta_v^{-}\right) \tag{15}$$

θ 是策略参数，θ_v 是状态值函数的参数，不同状态对应不同的 k，k 取值的上界为 t，V 为状态函数。A3C 算法样本利用率高、训练稳定性强，但其在多机协同的场景中容易出现适配不足、协同机制缺失等问题。

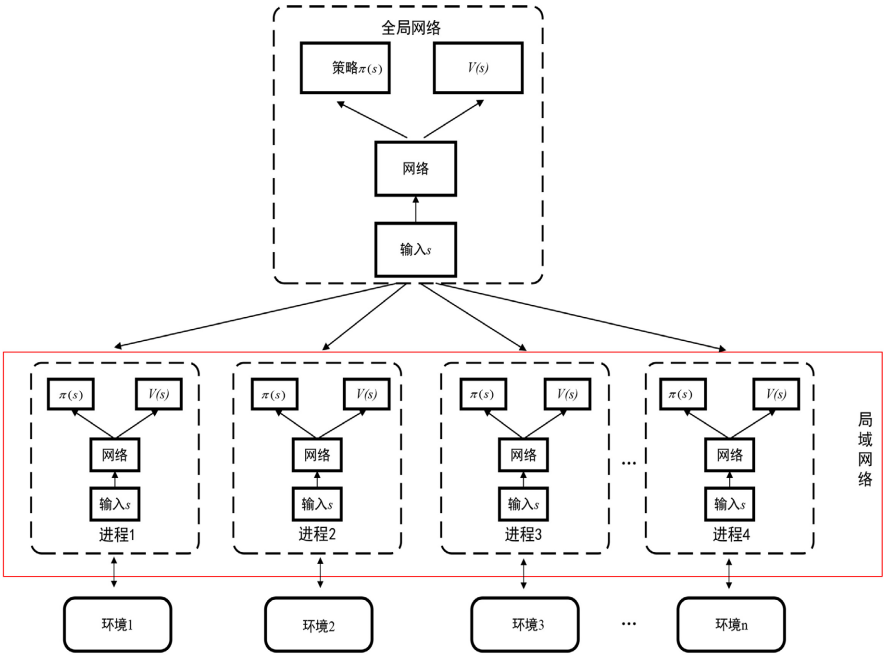

Figure 11. The asynchronous training framework of the A3C algorithm
图 11. A3C 算法异步训练框架

3.3.4. SAC

SAC (Soft Actor-Critic)算法[35]通过引入熵正则化与软化更新机制增强智能体在动态环境中的自适应能力，并借鉴 TD3 双 Q 网络截断策略，通过双独立 Q 网络取最小值抑制价值过估计，算法结构如图

12 所示。SAC 算法的优化目标是在最大化期望累积奖励的同时最大化策略熵，以保持策略的随机性，其目标函数如式(16)所示。

$$J(\pi)=\sum_{t=0}^{T}E_{(s_t,a_t)\sim\rho^{\pi}}\left[r(s_t,a_t)+\alpha H\left((\pi\cdot|s_t)\right)\right] \tag{16}$$

ρ^{π} 表示在策略 π 下 $r(s_t,a_t)$ 的分布，$H(\cdot)$表示熵值，α 表示超参数。SAC 算法在无人机实际飞行任务中表现出极强的稳定性，在障碍物分布、气象条件存在差异的多变空域环境中，仍能保持一致且优异的规划性能；且算法本身对超参数不敏感，在不同的交互环境中设置同样的超参数依然能够获得优秀的性能[36]，缺点是收敛耗时较长。

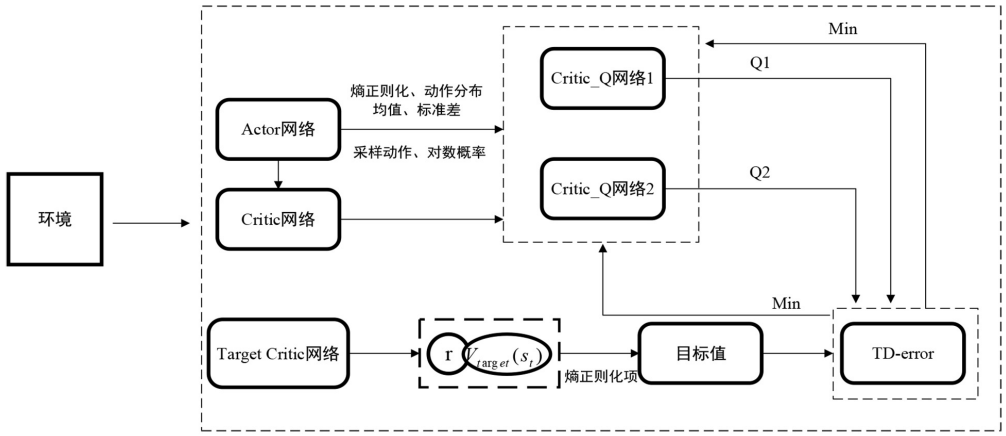

Figure 12. The framework of the SAC algorithm
图 12. SAC 算法结构图

3.4. 强化学习典型算法比较

强化学习算法选型直接决定无人机路径规划的效率、稳定性。为明确优缺点和应用适配性，本节对比归纳了典型算法的性能优缺点与适用场景，见表 1。

Table 1. Comparison of typical reinforcement learning algorithms and applicable scenarios
表 1. 强化学习典型算法比较与适用场景

理论框架	算法	优点	缺点	适用场景
值函数	Q-Learning	原理简单、易于实现	易过估计	状态空间简单场景
	SARSA	逐步更新策略	只适用于离散动作	在线学习场景、安全要求高的环境
	DQN	具有一定泛化能力	易过估计，效率低	静态全局路径规划
	DDQN	缓解过估计	解耦效果依赖网络设计	障碍物密集、灾害现场等安全敏感场景
	Dueling DQN	高效策略评估	实现较复杂	农业植保、物流配送等开阔空域场景
策略梯度	PPO	训练稳定性好	易过拟合	动态环境局部避障
值–策略混合	DDPG	收敛速度快	超参数敏感	连续动作空间
	TD3	稳定性高	策略保守，参数多	分布式决策、并行探索的复杂场景
	A3C	异步训练快	资源消耗大	围捕、覆盖搜索等分布式决策场景
	SAC	高维动作支持	初期探索效率低，收敛耗时较长	静动态障碍共存的复杂三维环境

4. 单机与多机场景下强化学习算法的改进与应用

4.1. 单无人机路径规划

单无人机路径规划研究以自主避障、目标追踪与路径优化为核心。针对未知环境感知延迟、动态场景收敛慢等问题，不少研究者对各类强化学习算法进行改进。Wu [37]提出了一种自适应转换速度 Q-Learning 算法(ACSQL)，并将其分为两个阶段，第一阶段通过自适应速度调整提升探索效率并结合传感器探测信息初始化 Q 表，解决训练收敛慢与高不确定性问题；第二阶段采用子域搜索算法获取安全短路径。通过仿真实验证明 ACSQL 在不同场景下路径长度短于 DDPG 和 IDWA 法。文献[38]将遗传算法(GA)与 Q-Learning 集成，GA 生成广泛候选路径，Q-Learning 结合无人机当前速度及与静态障碍物的距离进行理性决策，实现路径规划决策的实时优化与动态调整，提升任务能效与避碰能力的同时兼顾无人机速度与障碍物距离约束。通过实验证明 GA/QL 算法在能效上优于经典 GA 方法，性能提升 57.14%。针对传统 SARSA 算法依赖 ε-greedy 策略导致的收敛速度慢、易陷入局部最优两大问题，文献[39]将模拟退火(SA)全局优化策略引入 SARSA 算法，以 SA 替代传统 ε-greedy 策略，但未改进 SARSA 在线策略、单步更新的固有特性，易引发训练波动，复杂任务中学习效率受限。Chao 等人[40]提出了 E-DQN 仿生路径规划算法，基于 DQN 神经网络拟合高维状态价值的能力，引入大脑多巴胺奖惩机制，提升模型训练的稳定性与收敛性。Zhu 等人[41]提出了 ROLSM-DDQN 算法，借助 DDQN 动作选择与价值评估分离的结构优势，有效抑制 Q 值过估计，并将风扰动融入能量消耗模型与奖励函数，实现了无人机任务飞行中的最小阻力路径规划，相较于 DQN 和 DDQN 算法，ROLSM-DDQN 在步数、成本和路径长度上分别降低了 5.41%~6.25%、7.58%~8.60%和 10.23%。然而，该算法依赖全局信息预测进行离线规划，难以满足动态环境的实时性需求。Jiang 等[42]将 DDQN 拓展至灾害救援实际场景，提出了 DDQN-SSQN 算法，通过状态拆分与 RSSI 信号辅助决策，降低了无人机搜索被困人员的决策难度，提升了复杂任务场景下的收敛速度与任务完成率。

PPO 凭借稳定的策略更新特性，成为单机局部避障的优选算法，Qi 等[43]提出了频率分解 PPO 算法(FD-PPO)，充分发挥 PPO 策略更新稳定、不易发散的优势，设计启发式奖励函数来解决无人机路径规划问题，对比实验表明，FD-PPO 算法在平均奖励、路径长度和成功率方面都优于 PPO 算法。Tian 等人[44]提出了 TSEB-DDPG 算法，利用 HL-PSO 生成经验路径；再将经验轨迹、碰撞信息、实时转移数据存入三个经验缓冲区，相较传统 DDPG 等方法，收敛最快、精度最高，在真实场景中路径长度、规划耗时与成功率综合表现最优。文献[45]提出了 PP-CMNTD3 算法，借鉴人工势场设计密集奖励，实现避障、目标趋近与能耗平衡。Zhao 等[46]将 YOLOv7 作为感知前端与 TD3 结合，使算法在无障/有障环境成功率达 93%、92%，且可移植性强，但该成果未在真实无人机平台部署验证。Zhou 等人[47]面向无人机三维在线轨迹规划需求，利用 SAC 算法最大熵框架带来的优异探索能力与随机策略优势，在 SAC 的 Actor 网络引入自注意力机制强化高维状态关键特征提取，融合改进人工势场法设计多维度奖励函数以解决奖励稀疏问题，基于数字高程模型(DEM)设计状态空间适配复杂地形与突发威胁，显著提升轨迹规划成功率与平滑度。

4.2. 多无人机路径规划

多无人机路径规划以集群协同决策、全局路径优化与机间避碰为核心，相较于单机规划，还需解决无人机间的相互躲避、航线交叉与动作冲突等问题，对算法全局决策、多智能体协调与实时响应能力要求更高。赵天隆等[48]将引入优先经验回放的 DDQN 算法应用于无人机编队长机全局路径规划，结合改进人工势场法实现僚机的高效协同，凭借 DDQN 训练稳定、泛化性较好的特性，使僚机路径平滑度提升

82.6%~90.9%，充分验证了 DDQN 算法在多机编队全局优化中的应用价值。Wang 等人[49]在 Dueling DQN 算法基础上提出 IM-Dueling DQN 算法，将应用场景扩展至三维静动态混合障碍物空间，设计兼顾多目标的综合奖励函数和改进双经验池策略，使无人机群在复杂环境中仍能规划出安全、高效的飞行路径。Zhang 等人[50]利用 TD3 训练稳定、探索均匀的特点，将遗传算法与之结合，引入最大均值差方法增强策略多样性，并通过改进突变算子与设计混合奖励函数提升算法稳定性与收敛速度，任务成功率达 95%。

A3C 算法凭借异步更新特性提升了样本利用率与训练稳定性，但其在多机集群场景中暴露出多任务适配不足、协同机制缺失、易出现动作冲突等问题。针对上述问题，Qiao 等人[51]结合 A3C 异步训练效率高与 PPO 更新稳定的双重优势，设计四种战略行为专属奖励函数强化任务导向，并引入经验回放机制，一定程度上提升了算法的场景适配性，但改进后的算法计算开销大幅增加，难以满足机载端的计算资源约束。现有多无人机协同大多依赖简单跟随与基础碰撞惩罚，缺乏精细化实时协调与冲突预判，且以中心化控制为主，未经过大机群、复杂地形等实际场景验证，难以应对动态需求与突发状况。而强化学习算法仍停留在仿真阶段，尚未从安全约束、硬件特性、场景复杂性及实时性需求出发开展系统性设计，这正是从仿真走向真实应用的核心突破口。

4.3. 现有研究的共性问题与局限性分析

当前，基于强化学习的无人机三维路径规划研究存在以仿真为主，缺乏真机部署、模型假设过于理想化、评估指标单一、缺少多维度综合评价等共性问题，成为算法从仿真走向实际应用的主要障碍。表 2 汇总了上述文献在实验验证、模型假设、评估体系三方面的共性与局限性。

Table 2. Summary and comparison of deep reinforcement learning algorithms in path planning applications
表 2. 深度强化学习算法在路径规划应用总结与对比

算法	实验设计	模型假设	评估指标
ACSQL [37]	仅在仿真环境中验证	传感器探测信息无噪，环境完全可知	仅评估路径长度，未对比计算耗时或成功率
GA + Q-Learning [38]	静态环境仿真	无人机速度与障碍物距离可精确获取	能效提升 57.14%
SARSA + SA [39]	低复杂度场景仿真	状态空间离散且规模有限	仅评估收敛速度，未给出路径长度或避障成功率
E-DQN [40]	AirSim 仿真	多巴胺奖惩机制线性映射到无人机控制	路径规划运行时间 41.364 s，优于 DQN
ROLSM-DDQN [41]	全局信息离线规划	风扰动模型已知且可预先建模	与 DDQN 相比步数下降 5.41%，能耗下降 7.58%，路径长度下降 10.23%
DDQN-SSQN [42]	灾害救援特定场景仿真	RSSI 信号强度与距离呈理想单调关系	追踪成功率达到 99.5%
FD-PPO [43]	仿真环境验证	启发式奖励函数可泛化至不同地形	与 PPO 相比，FD-PPO 的路径更短，避障成功率更高
TSEB-DDPG [44]	应急通信与救援任务场景	城市 3D 环境含不同高度建筑等静态障碍	成功路径长度 38.76 m，成功率 87.6%
PP-CMNTD3 [45]	仿真实验，未部署真机	假设目标无人机运动轨迹可预测	成功率 73.6%，碰撞率 11.4%，规划成功平均消耗电量较高
YOLOv7 + TD3 [46]	仿真实验，未部署真机	人工势场设计的密集奖励可迁移至未知环境	有/无障碍环境下路径规划成功率 93%/92%
SAC + 自注意力 [47]	基于 DEM 仿真，未验证真实地形	假设数字高程模型可精确表征复杂地形	成功率与平滑度提升

续表

DDQN + 人工势场 [48]	长机–僚机编队仿真	假设长机全局路径规划结果可被僚机跟随	所有僚机的路径平滑度都有提升，分别达到 90.9%与 82.6%
IM-DuelingDQN [49]	静动态混合障碍物场景，但未说明动态障碍物运动规律	双经验池策略可无冲突地协调多机决策	平均奖励、收敛速度、稳定性、任务成功率、避障能力全部优于 Q-Learning、DQN、DDQN 算法
GM-TD3 [50]	仅验证单追逃场景，未扩展至多机协同避障	假设目标运动模型已知	GM-TD3 成功率约 95%
PPO-A3C-PER [51]	无人机空战仿真环境	四种战略行为可覆盖所有空战场景	机动决策成功率 94.7%，收敛速度提升约 42%

5. 挑战与未来发展

5.1. 现存关键挑战

1) 建模与动态性挑战：实际飞行中地形、动态障碍等易导致状态空间膨胀，精确建模困难，使算法收敛慢、易陷局部最优，环境突发变化还会干扰规划稳定性[52]；

2) 多目标适配不足：强化学习奖励函数多为固定权重且无统一标准，难以兼顾路径优化、能耗控制、避障安全等多重目标，泛化能力弱；

3) 场景迁移性差：不同作业场景环境特征差异大，单场景训练模型难以直接迁移至陌生场景，需大量重训，增加部署与时间成本，难以投入实际应用。

5.2. 未来发展方向

1) 构建边缘云–机载端的协同架构，边缘云完成模型预训练与轻量化优化，机载端部署精简模型，结合低延迟通信实现快速决策闭环；依托边缘分布式算力，提升复杂场景下的实时响应与安全可靠性；

2) 针对奖励函数权重固定、策略泛化性弱的问题，研发自适应的多目标奖励函数以平衡规划冲突；引入模仿学习等机制迁移单场景经验，降低重训成本、提升模型多场景适配性；

3) 发展基于世界模型的强化学习方法，借鉴 Dreamer 算法让无人机学习环境模型预测状态与奖励，在虚拟模型中规划学习，大幅减少真实交互数据依赖，适配真机低试错需求；

4) 借助大语言模型的语义解析与知识迁移能力，将自然语言任务转化为强化学习可识别的目标与约束，注入先验知识减少无效探索，同时通过实时交互反馈适配动态任务调整。

6. 总结

本文围绕无人机三维路径规划问题，深入探讨了强化学习在解决该问题中的理论优势与应用潜力。结合无人机高动态、强约束、资源有限等特性，重点分析对比了基于值函数、策略梯度及值策略混合的三类强化学习算法在路径规划中的改进策略及典型成果；总结未来发展方向，旨在为相关领域的研究与实践提供有益的参考与借鉴。

基金项目

本研究得到地球深部探测与矿产资源勘查国家科技重大专项支持(编号：2024ZD1002600)。

参考文献

[1] Yang, L., Qi, J.T., Xiao, J. and Yong, X. (2014) A Literature Review of UAV 3D Path Planning. *Proceeding of the* 11*th World Congress on Intelligent Control and Automation*, Shenyang, 29 June-4 July 2014, 2376-2381.

https://doi.org/10.1109/wcica.2014.7053093

[2] 聂虹宇, 张广玉, 李德才, 等. 多旋翼无人机的环境感知与运动规划方法综述[J]. 信息与控制, 2025, 54(3): 353-371.

[3] 李晓辉, 苗苗, 冉保健, 等. 基于改进 A*算法的无人机避障路径规划[J]. 计算机系统应用, 2021, 30(2): 255-259.

[4] 李亚飞, 赵瑞. 城市复杂环境下多目标无人机路径规划研究[J]. 南京航空航天大学学报, 2024, 56(6): 1002-1012.

[5] Huang, Y., Li, H., Dai, Y., Lu, G. and Duan, M. (2024) A 3D Path Planning Algorithm for UAVs Based on an Improved Artificial Potential Field and Bidirectional RRT. *Drones*, **8**, Article 760. https://doi.org/10.3390/drones8120760

[6] 曾国奇, 赵民强, 刘方圆, 等. 基于网格 PRM 的无人机多约束航路规划[J]. 系统工程与电子技术, 2016, 38(10): 2310-2316.

[7] Tripicchio, P., Unetti, M., D'Avella, S. and Avizzano, C.A. (2023) Smooth Coverage Path Planning for UAVs with Model Predictive Control Trajectory Tracking. *Electronics*, **12**, Article 2310. https://doi.org/10.3390/electronics12102310

[8] Azar, A.T., Koubaa, A., Ali Mohamed, N., Ibrahim, H.A., Ibrahim, Z.F., Kazim, M., *et al*. (2021) Drone Deep Reinforcement Learning: A Review. *Electronics*, **10**, Article 999. https://doi.org/10.3390/electronics10090999

[9] Sun, H., Zhang, W., Yu, R. and Zhang, Y. (2021) Motion Planning for Mobile Robots—Focusing on Deep Reinforcement Learning: A Systematic Review. *IEEE Access*, **9**, 69061-69081. https://doi.org/10.1109/access.2021.3076530

[10] Zhu, K. and Zhang, T. (2021) Deep Reinforcement Learning Based Mobile Robot Navigation: A Review. *Tsinghua Science and Technology*, **26**, 674-691. https://doi.org/10.26599/tst.2021.9010012

[11] 熊斯, 李逸琛, 欧阳权, 等. 基于强化学习的无人机集群航迹规划研究综述[J]. 空间电子技术, 2025, 22(6): 1-8, 123.

[12] Tanimoto, Y. and Fukumizu, K. (2024) State-Separated SARSA: A Practical Sequential Decision-Making Algorithm with Recovering Rewards. arXiv: 2403.11520.

[13] 许振阳, 陈谋, 韩增亮, 等. 复杂环境下基于 TCPDQN 算法的低空飞行器动态航路规划[J]. 机器人, 2025, 47(3): 383-393.

[14] Watkins, C.J. and Watkins, P. (1989) Learning from Delayed Rewards. Ph.D. Thesis, King's College.

[15] 张泽华, 杨波, 傅广, 等. 基于 SARSA 的动态蜂群算法求解作业车间调度问题[J]. 组合机床与自动化加工技术, 2023(6): 188-192.

[16] 陈一波, 赵知劲. 基于 SARSA 学习的跳频系统智能抗干扰决策算法[J]. 现代电子技术, 2023, 46(1): 31-35.

[17] 司彦娜, 普杰信, 于晓升, 等. 基于径向基神经网络的多步 SARSA 控制算法[J]. 控制与决策, 2023, 38(4): 944-950.

[18] 黄鑫, 张志佳, 孙平, 等. 基于深度强化学习的路径规划算法综述[J]. 机器人, 2026, 48(1): 196-216.

[19] 于天浩, 周航, 贾鑫悦, 等. 基于改进 DQN 算法的无人机路径规划算法研究[J]. 航空计算技术, 2025, 55(6): 59-63, 79.

[20] 王艺霖, 张烈平, 尹亚梦, 等. 基于改进 DDQN 的移动机器人路径规划算法[J]. 桂林航天工业学院学报, 2025, 30(5): 770-783.

[21] Van Hasselt, H., Guez, A. and Silver, D. (2016) Deep Reinforcement Learning with Double Q-Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**, 2094-2100. https://doi.org/10.1609/aaai.v30i1.10295

[22] Wang, Z., Schaul, T., Hessel, M., *et al*. (2016) Dueling Network Architectures for Deep Reinforcement Learning. *International Conference on Machine Learning*, New York, 19-24 June 2016, 1995-2003.

[23] 苏江玉. 基于深度强化学习的 USV 路径规划算法研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2023.

[24] 武曲, 张义, 郭坤, 等. 基于 DPES Dueling DQN 的路径规划方法研究[J]. 计算机应用与软件, 2023, 40(6): 147-153, 233.

[25] Xu, Y., Wei, Y., Wang, D., Jiang, K. and Deng, H. (2023) Multi-UAV Path Planning in GPS and Communication Denial Environment. *Sensors*, **23**, Article 2997. https://doi.org/10.3390/s23062997

[26] Schulman, J., Levine, S., Abbeel, P., *et al*. (2015) Trust Region Policy Optimization. *International Conference on Machine Learning*, Lille, 6-11 July 2015, 1889-1897.

[27] 万宇航, 朱子璐, 钟春富, 等. 基于改进 PPO 算法的机械臂动态路径规划[J]. 系统仿真学报, 2025, 37(6): 1462-1473.

[28] 程浩鹏, 朱涵, 杨高奇, 等. 深度强化学习及智能路径规划应用综述[J]. 现代计算机, 2022, 28(21): 1-10.

[29] Barto, A.G., Sutton, R.S. and Anderson, C.W. (1983) Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**, 834-846. https://doi.org/10.1109/tsmc.1983.6313077

[30] Lillicrap, T.P., Hunt, J.J., Pritzel, A., *et al*. (2016) Continuous Control with Deep Reinforcement Learning. *International Conference on Learning Representations*, San Juan, 2-4 May 2016.

[31] Silver, D., Lever, G., Heess, N., *et al*. (2014) Deterministic Policy Gradient Algorithms. *International Conference on Machine Learning*, Beijing, 21-26 June 2014, 387-395.

[32] 王树森. 深度强化学习[M]. 北京: 人民邮电出版社, 2022.

[33] Fujimoto, S., Hoof, H. and Meger, D. (2018) Addressing Function Approximation Error in Actor-Critic Methods. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 1587-1596.

[34] Mnih, V., Badia, A.P., Mirza, M., *et al*. (2016) Asynchronous Methods for Deep Reinforcement Learning. *International Conference on Machine Learning*, New York, 19-24 June 2016, 1928-1937.

[35] Haarnoja, T., Zhou, A., Abbeel, P., *et al*. (2018) Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 1861-1870.

[36] 周明鑫. 基于强化学习的多智能体自主任务分配[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工程大学, 2022.

[37] Wu, J., Sun, Y., Li, D., Shi, J., Li, X., Gao, L., *et al*. (2023) An Adaptive Conversion Speed Q-Learning Algorithm for Search and Rescue UAV Path Planning in Unknown Environments. *IEEE Transactions on Vehicular Technology*, **72**, 15391-15404. https://doi.org/10.1109/tvt.2023.3297837

[38] Saeed, R.A., Ali, E.S., Abdelhaq, M., Alsaqour, R., Ahmed, F.R.A. and Saad, A.M.E. (2024) Energy Efficient Path Planning Scheme for Unmanned Aerial Vehicle Using Hybrid Generic Algorithm-Based Q-Learning Optimization. *IEEE Access*, **12**, 13400-13417. https://doi.org/10.1109/access.2023.3344455

[39] 王现磊, 郝文宁, 陈刚, 等. 基于模拟退火策略的 SARSA 强化学习方法[J]. 计算机仿真, 2019, 36(4): 219-222, 228.

[40] Chao, Y., Dillmann, R., Roennau, A. and Xiong, Z. (2024) E-DQN-Based Path Planning Method for Drones in Airsim Simulator under Unknown Environment. *Biomimetics*, **9**, Article 238. https://doi.org/10.3390/biomimetics9040238

[41] Zhu, Y., Tan, Y., Chen, Y., Chen, L. and Lee, K.Y. (2024) UAV Path Planning Based on Random Obstacle Training and Linear Soft Update of DRL in Dense Urban Environment. *Energies*, **17**, Article 2762. https://doi.org/10.3390/en17112762

[42] Jiang, W., Bao, C., Xu, G. and Wang, Y. (2021) Research on Autonomous Obstacle Avoidance and Target Tracking of UAV Based on Improved Dueling DQN Algorithm. 2021 *China Automation Congress* (*CAC*), Beijing, 22-24 October 2021, 5110-5115. https://doi.org/10.1109/cac53003.2021.9728707

[43] Qi, C., Wu, C., Lei, L., Li, X. and Cong, P. (2022) UAV Path Planning Based on the Improved PPO Algorithm. 2022 *Asia Conference on Advanced Robotics, Automation, and Control Engineering* (*ARACE*), Qingdao, 26-28 August 2022, 193-199. https://doi.org/10.1109/arace56528.2022.00040

[44] Tian, S., Li, Y., Zhang, X., Zheng, L., Cheng, L., She, W., *et al*. (2024) Fast UAV Path Planning in Urban Environments Based on Three-Step Experience Buffer Sampling DDPG. *Digital Communications and Networks*, **10**, 813-826. https://doi.org/10.1016/j.dcan.2023.02.016

[45] 牟文心, 时宏伟. 基于改进 TD3 算法的无人机轨迹规划[J]. 计算机系统应用, 2024, 33(12): 197-209.

[46] Zhao, F.Y., Li, D.Y., Wang, Z.X., Mao, J.L. and Wang, N.Y. (2024) Autonomous Localized Path Planning Algorithm for UAVs Based on TD3 Strategy. *Scientific Reports*, **14**, Article No. 763. https://doi.org/10.1038/s41598-024-51349-4

[47] Zhou, Y., Shu, J., Hao, H., Song, H. and Lai, X. (2023) UAV 3D Online Track Planning Based on Improved SAC Algorithm. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, **46**, Article No. 12. https://doi.org/10.1007/s40430-023-04570-7

[48] 赵天隆, 陈龙胜, 张存富, 等. 融合强化学习与改进人工势场的无人机编队路径规划[J]. 航空兵器, 2025, 32(5): 54-63.

[49] Wang, W., Zhang, G., Da, Q. and Tian, Y. (2024) Path Planning with Improved Dueling DQN Algorithm for UAVs in Unknown Dynamic Environment. In: Li, S., Ed., *Computational and Experimental Simulations in Engineering*, Springer, 453-465. https://doi.org/10.1007/978-3-031-44947-5_36

[50] Zhang, Y., Ding, M., Yuan, Y., Zhang, J., Yang, Q., Shi, G., *et al*. (2024) Multi-UAV Cooperative Pursuit of a Fast-Moving Target UAV Based on the GM-TD3 Algorithm. *Drones*, **8**, Article 557. https://doi.org/10.3390/drones8100557

[51] Qiao, B., Jia, Z., Xiao, B. and Qian, H. (2025) Game Maneuver Decision-Making for Multi-UAV via PPO-A3C-PER Learning Method. In: Yan, L., Duan, H. and Deng, Y., Eds., *Advances in Guidance, Navigation and Control*, Springer, 72-81. https://doi.org/10.1007/978-981-96-2232-0_8

[52] 陈麒杰, 晋玉强, 韩露. 无人机路径规划算法研究综述[J]. 飞航导弹, 2020(5): 54-58.