

基于改进型 DQN 的无人机覆盖路径规划算法

王兴旺¹, 罗锦宽², 张紫洋², 迟玉璞², 蒋金佑², 于美铭³

(1. 吉林大学 计算机科学与技术学院, 长春 130012; 2. 吉林大学 软件学院, 长春 130012; 3. 吉林大学 公共计算机教学与研究中心, 长春 130012)

摘要: 针对无人机覆盖路径规划中传统 Deep Q-network (DQN) 存在的覆盖率低、路径冗余度高、收敛速度慢等问题, 提出了一种基于多机制融合的改进型 DQN 模型。首先, 将无人机的工作空间建模为二维栅格地图, 进行预处理并设计奖励函数。其次, 基于传统 DQN, 逐步融合 Double DQN、Dueling network 和 Priority experience replay 三种机制, 运用独热编码等技术搭建多尺度卷积神经网络, 构建了改进型 DQN 模型。最后, 设计可视化仿真实验, 并以栅格覆盖率等评价指标进行消融实验。实验结果表明: 改进型 DQN 模型在覆盖路径规划任务中表现出色, 其性能优于传统 DQN 等模型。相较于传统 DQN, 其栅格覆盖率提升了 8.28%, 栅格重复率、路径长度、碰撞障碍物次数、超出边界次数分别降低了 28.44%、45.14%、77.31%、90.14%, 且网络收敛速度显著加快。

关键词: 计算机应用; 深度强化学习; 覆盖路径规划; 无人机; 栅格地图

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2026)09-2444-12

DOI: 10.13229/j.cnki.jdxbgxb.20250237

Algorithm of coverage path planning for unmanned aerial vehicle based on improved DQN

WANG Xing-wang¹, LUO Jin-kuan², ZHANG Zi-yang², CHI Yu-pu², JIANG Jin-you², YU Mei-ming³

(1. College of Computer Science and Technology, Jilin University, Changchun 130012, China; 2. College of Software, Jilin University, Changchun 130012, China; 3. Public Computer Education and Research Center, Jilin University, Changchun 130012, China)

Abstract: Aiming at the problems of low coverage, high path redundancy and slow convergence speed of traditional DQN in UAV coverage path planning, an improved DQN model based on multi-mechanism fusion was proposed. Firstly, the workspace of the UAV was modeled as a two-dimensional grid map. It was preprocessed and the reward function was designed. Secondly, based on the traditional DQN, three mechanisms of Double DQN, Dueling Network and Priority Experience Replay were gradually integrated, and a multi-scale convolutional neural network was built using techniques such as one-hot encoding to construct an improved DQN model. Finally, a visual simulation experiment was designed, and an ablation experiment was performed using evaluation indicators such as grid coverage. The experimental results show

收稿日期: 2025-03-21.

基金项目: 国家重点研发计划项目 (2023YFB4502304); 吉林大学创新创业训练计划项目 (X202410183312).

作者简介: 王兴旺 (1990-), 男, 副教授, 博士. 研究方向: 深度神经网络优化. E-mail: xww@jlu.edu.cn

通信作者: 于美铭 (1995-), 女, 工程师. 研究方向: 物联网与深度神经网络. E-mail: yumm@jlu.edu.cn

that the improved DQN model performs well in the coverage path planning task, and its performance is better than that of traditional DQN and other models. Compared with traditional DQN, its grid coverage rate is increased by 8.28%. Also, the grid repetition rate, path length, number of collisions with obstacles, and number of times exceeding boundaries are reduced by 28.44%, 45.14%, 77.31%, and 90.14%. The network convergence speed is significantly accelerated as well.

Key words: computer applications; deep reinforcement learning; coverage path planning; unmanned aerial vehicle; grid maps

0 引言

无人机技术以其高效、灵活和低成本的优势,在林业^[1]、空气质量监测^[2]、灾害响应^[3]等多个领域得到应用。利用无人机实现智能化的巡视检查已成为当前研究的热点方向之一。覆盖路径规划(Coverage path planning, CPP)作为无人机巡检的关键技术,旨在为其规划最优轨迹以实现对目标区域的全覆盖访问。该技术还广泛应用于无人水面舰艇^[4]、智能扫地机器人^[5]、救援机器人^[6]等领域。

传统CPP方法主要基于经典图论算法,如Dijkstra算法^[7]通过迭代更新节点集合求解最短路径;Hart等^[8]提出的A*算法通过引入启发式函数提升搜索效率。而Xu等^[9]提出了多目标粒子群优化算法,以进行更稳定、效果更好的路径规划。唐颂等^[10]提出了一种改进的遗传算法,将任务分配与航迹规划信息统一编码,解决了多无人机协同规划中计算规模大、效率低的问题。

近年来,深度强化学习(Deep reinforcement learning, DRL)的突破性发展为CPP研究提供了新范式。Luo等^[11]通过优化神经网络激活函数简化了状态空间建模过程;屈博琛^[12]提出了基于深度学习自动微分的无人机路径规划算法,实现了无人机在多个障碍物下的最优路径规划问题求解;值得注意的是,DeepMind团队^[13]提出的深度Q网络(Deep Q-network, DQN)通过深度神经网络逼近Q值函数,开创了端到端决策的新模式。后续研究相继提出重要改进:Van Hasselt等^[14]设计的双网络架构(Double DQN, DDQN)有效解决了Q值过度估计问题;Wang等^[15]提出了竞争网络结构(Dueling network),增强了网络对状态价值和动作优势的分离评估能力;Schaul等^[16]引入了优先经验回放(Prioritized experience replay, PER),通过TD误差加权采样显著提升了训练效率。

尽管现有方法在特定维度改进了DQN性能,但单一改进策略对系统整体性能的提升有限。此外,现有研究尚缺乏对多种改进机制综合应用的系统性探讨。鉴于此,本文致力于在传统DQN的基础上,融合DDQN、Dueling network和PER三种机制,构建一种改进型DQN模型,以期在无人机覆盖路径规划任务中取得更优的性能表现。

1 工作空间建模与预处理

1.1 工作空间建模

为有效模拟覆盖路径规划任务,现作如下假设:

(1)将无人机的工作空间建模为 $H \times W$ 的二维栅格地图;每个栅格代表一定的区域面积,反映实际地理特征。

(2)每个栅格的大小近似为视觉传感器一次拍摄所覆盖的区域面积,如图1中的蓝色部分所示(彩图见电子版,下同)。

(3)不考虑无人机的体积,将其建模为直径等于栅格边长的黑色圆形智能体,移动方向设置为水平和垂直方向。

(4)当智能体到达栅格的中心位置时,认为该栅格已被覆盖。忽略智能体的转弯半径及能量消耗。

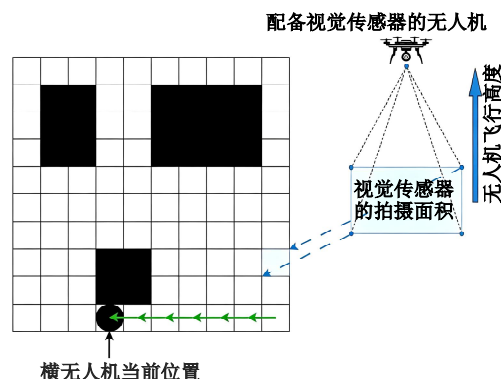


图1 栅格地图构建示意图

Fig. 1 Schematic diagram of grid maps construction

本文采用颜色编码与数值标识相结合的方式构建二维栅格地图,以区分不同类型的栅格单元。一些自然地貌特征如小山丘、岩石区、天然屏障,视为栅格地图中的障碍物,设定为黑色且数值标识为-1,表示无人机无法通行和覆盖此类栅格。空旷区域设为白色且数值标识为0,表示智能体可通行且尚未覆盖的栅格。已被覆盖的栅格数值标识为1,其存在绿色箭头图标,表明智能体的覆盖路径。

为便于强化学习算法的实现与测试,本文将同时存在黑、白两种区域的栅格单元皆设置为黑色栅格,即工作空间中出现的障碍物都是由若干正方形黑色栅格单元组成的,如图2所示。

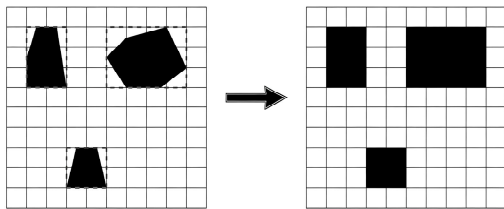


图2 障碍物规则化示意图

Fig. 2 Schematic diagram of obstacle regularization

1.2 奖励函数的设计

智能体应尽可能选择通行未覆盖的区域,同时避免重复覆盖,提高路径规划效率。为引导智能体高效完成覆盖路径规划任务,本文奖励函数设计为:

$$r = \begin{cases} 1, & G(i, j) = 0 \\ -1, & G(i, j) = 1 \\ 0.25, & \vec{G}(i, j) = \vec{G}(i', j') \\ -20, & G(i, j) = -1 \text{ or } i, j < 0 \text{ or } i, j > 10 \\ 20, & \text{done} \end{cases} \quad (1)$$

式中: $G(i, j)$ 表示智能体采取行动后在栅格地图中的位置信息,值为0表示其进入未被覆盖的栅格,给予奖励+1;值为1表示其进入已被覆盖的栅格,给予奖励-1;值为-1表示其碰撞障碍物,给予奖励-20; i 和 j 任意值小于0或大于10时,表示其飞越栅格地图边界,给予奖励-20;当智能体完成全覆盖时,获得奖励+20;无人机在直线方向上连续运动时,行驶同等路程用时最短, $\vec{G}(i, j)$ 表示智能体的运动方向,当其连续向同一方向运动时,给予奖励+0.25。

2 深度强化学习模型

2.1 基础模型

2.1.1 Q-Learning

强化学习是机器学习的一个重要分支,与监督学习不同,其不依赖于标注数据,而是通过智能体与环境的交互来学习最优策略,从而在给定任务中最大化长期回报。它的基本框架由智能体、环境状态、动作和奖励这几个要素组成。智能体执行动作后,环境将会更新状态,并根据任务完成情况给予即时奖惩反馈,随后智能体更新策略,并根据新的环境状态和反馈执行新的动作。

Watkins^[17]提出的Q-Learning是一种无模型(Model-Free)和离策略(Off-Policy)的强化学习算法,旨在通过与环境的交互,学习一个动作-价值Q函数以确定最优行为策略,最大化长期回报。

在Q-Learning中,贝尔曼方程描述了最优策略下的Q值递归关系,定义了最优Q值函数,在每个状态-动作对下,最大化预期回报的Q值。其确定了Q-Learning的优化目标,并为Q值的更新提供了理论依据。Q-Learning采用结合了蒙特卡罗方法和动态规划方法的时间差分学习算法,通过实时计算和调整TD误差以更新Q值,在与环境的交互过程中,逐步逼近最优Q值。Q值的更新公式为:

$$Q(s, a) \leftarrow Q(s, a) + \omega \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right) \quad (2)$$

式中: s 为当前状态; a 为当前选择的动作; r 为执行动作后获得的即时奖励; s' 为执行动作后的新环境状态; a' 为下一个可能采取的动作; ω 为学习率,控制新获得的信息对Q值更新的影响; γ 为折扣因子,衡量未来奖励对当前决策的影响; $Q(s, a)$ 表示当前状态 s 下选择动作 a 的Q值; $\max_{a'} Q(s', a')$ 表示在下一个状态 s' 中,所有可能动作中的最大Q值。

$Q(s, a)$ 本质上是一个存储所有状态-行动对 (s, a) 的Q值的二维表格,即Q值表(Q-Table)。每个条目 $Q(s, a)$ 表示在状态 s 下采取行动 a 时,智能体所能获得的期望回报。Q值表通常初始化为零或随机的小值,其随着智能体与环境的交互不断更新。

2.1.2 传统DQN模型

Q-Learning算法在离散和低维状态空间中表现较好,但随着状态空间维度的增加,Q值表的规模呈指数级增长导致存储和计算成本过高,难以处理高维或连续的状态-动作空间。

为解决这一问题,DeepMind团队^[13]提出了DQN,创新地将深度学习与Q-Learning相结合,使用深度神经网络逼近Q值函数,从而避免了Q值表在高维状态空间中的存储问题,能够处理更加复杂的状态空间。DQN中神经网络的输入是当前环境状态 s ,输出是所有可能动作 a 的Q值。网络的训练目标是 minimized Q值预测与目标Q值之间的误差,使网络学习到最优的动作-价值函数,准确地估计Q值。同时,DQN在Q-Learning的基础上引入了若干关键技术,主要包括经验回放和目标网络。

为打破样本之间存在的时间相关性,DQN引入了经验回放池,将智能体与环境交互产生的所有四元组经验 (s_t, a_t, r_t, s_{t+1}) 都存储在该经验池中。训练过程中,DQN从经验池中随机采样小批量数据进行网络更新,有效降低了数据之间的时间相关性,提高了训练的效率和稳定性。

由于Q值函数本身是由神经网络近似的,故直接使用主网络计算和更新Q值具有不稳定性。因此,DQN引入了与当前主网络结构相同的目标网络,其参数定期从主网络复制,更新滞后于主网络。主网络用于选择动作和估计Q值,目标网络用于计算更新时的目标Q值,从而降低了训练的不稳定性,减少了训练过程中的震荡和发散风险。

目标Q值计算公式为:

$$y_m = r + \gamma \max_{a'} Q(s', a'; \theta^-) \quad (3)$$

式中: y_m 为DQN的目标Q值; $\max_{a'} Q(s', a'; \theta^-)$ 为下一个状态 s' 下所有可能动作中的最大Q值,由目标网络估计; θ^- 为目标网络的参数。

DQN通过最小化预测Q值与目标Q值之间的均方误差来训练网络,损失函数定义为:

$$L(\theta) = E_{(s,a,r,s') \sim D} \left[\left(y_m - Q(s, a; \theta) \right)^2 \right] \quad (4)$$

式中: D 为经验回放池中存储的经验样本所定义的经验分布; $Q(s, a; \theta)$ 为主网络对状态 s 和动作 a 的估计Q值; θ 为主网络的参数。

通过不断更新网络参数,DQN能够学习到一个近似最优的Q值函数,实现高效的决策。

传统DQN的框架如图3所示。

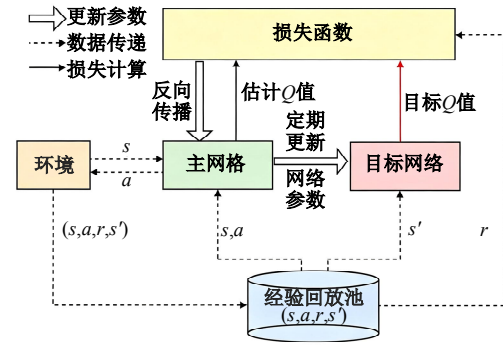


图3 DQN的框架示意图

Fig. 3 Schematic diagram of DQN architecture

2.2 对传统DQN单一策略的改进

DQN的出现标志着深度强化学习的重大突破,为解决复杂环境中的决策问题提供了一个有效的工具,但其存在的Q值过度估计、动作价值估计不够准确和样本采样效率低等问题仍需进一步研究改进。

2.2.1 对Q值过度估计问题的改进

传统DQN在Q值更新时,使用同一个网络选择和评估动作,这可能导致Q值过度估计的问题,使智能体偏向选择Q值被高估的动作,影响策略的稳定性和模型性能。Van Hasselt等^[14]提出了DDQN,其核心思想在于将动作选择与动作评估分离,利用两个独立的网络分别执行这两个任务,从而减少了单一网络对动作价值的过度估计。

传统DQN原有两个网络,故无须再构建新网络。现由主网络根据当前状态 s' 选择具有最大Q值的动作 a' ,再通过目标网络据此计算该动作对应的Q值 $Q(s', a'; \theta^-)$,成功减少了Q值过度估计的现象。

DDQN的目标Q值计算公式为:

$$y_n = r + \gamma Q(s', \arg \max_{a'} Q(s', a'; \theta); \theta^-) \quad (5)$$

式中: y_n 为DDQN机制的目标Q值; $\arg \max_{a'} Q(s', a'; \theta)$ 指对下一状态 s' 的所有可能动作 a' 计算 $Q(s', a'; \theta)$,并取令该值最大的动作 a' ; $Q(s', \arg \max_{a'} Q(s', a'; \theta); \theta^-)$ 指使用目标网络的参数 θ^- 评估该动作在状态 s' 中的Q值。

DDQN的损失函数定义为:

$$L(\theta) = E_{(s,a,r,s') \sim D} \left[\left(y_n - Q(s, a; \theta) \right)^2 \right] \quad (6)$$

2.2.2 对网络架构的改进

传统 DQN 计算动作 Q 值时未考虑不同动作在不同状态下的实际价值差异。这未充分利用状态本身的价值信息,不能体现每个动作的相对重要性,导致在某些状态下动作价值的估计不够准确。Wang 等^[15]提出了 Dueling network 架构,通过将 Q 函数分解为状态价值函数和动作优势函数的加权组合,显著提升了 Q 值估计的效率和准确性:

$$Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + \left(A(s, a; \theta, \alpha) - \frac{1}{|K|} \sum_{a'} A(s, a'; \theta, \alpha) \right) \quad (7)$$

式中: θ 为共享的网络参数; α 和 β 分别为动作优势函数和状态价值函数的参数; $|K|$ 为可选动作的总数; $V(s; \theta, \beta)$ 为状态价值函数,表示无论采取何种动作,智能体能够获得的期望累积奖励; $A(s, a; \theta, \alpha)$ 为动作优势函数,表示在状态 s 下,动作 a 相对于该状态下所有可能动作的平均优势; $\sum_{a'} A(s, a'; \theta, \alpha) / |K|$ 为在状态 s 下,所有动作优势函数的平均值; $Q(s, a; \theta, \alpha, \beta)$ 为主网络对状态 s 和动作 a 的估计 Q 值,此时动作优势函数和状态价值函数的参数分别为 α 和 β 。

网络结构实现如图 4 所示,原网络用于提取状态共享特征的前几层不变,但在输出层之前添加了两个独立的分支,分别计算状态价值和动作优势,最终加权组合得到 Q 值输出。这种分解方式使网络能够分别估计状态的整体价值和各动作相对于该状态的优势。在不显著增加计算复杂度的情况下,提升了 Q 值估计的效率和准确性,使训练过程趋于稳定。

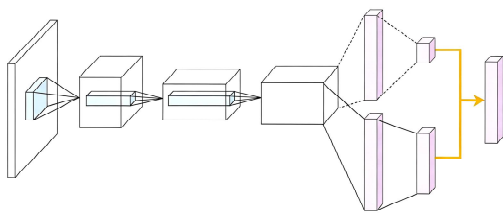


图 4 Dueling network 示意图

Fig. 4 Schematic diagram of Dueling network

2.2.3 对经验回放机制的改进

传统 DQN 的经验回放机制对所有存储的经验样本给予相同的采样概率,从经验池中随机抽取样本进行训练。然而,其未能充分考虑不同样本的重要性,导致对网络训练和策略更新较为关

键的样本利用不充分,而其他对模型训练贡献较小的样本则可能被反复采样。这减缓了神经网络的收敛速度,降低了计算资源利用率。

Schaul 等^[16]提出了 PER,基于每个样本的时序差分(TD)误差大小确定其优先级并调整其被抽取的概率,使智能体能够更多地学习具有较大 TD 误差的样本,有效地加速学习过程的收敛。

TD 误差反映了某个经验样本的预测 Q 值与实际观测奖励之间的差距,因此,较大的 TD 误差表明该样本对于更新 Q 值函数有关键作用。该改进不仅提高了有效样本的利用率,还加速了智能体对关键经验的学习,提升了模型的训练效率和性能。

每个经验样本 i 的优先级由其 TD 误差的绝对值决定,TD 误差的计算公式为:

$$\delta_i = r + \gamma Q\left(s', \arg \max_{a'} Q(s', a'; \theta); \theta^-\right) - Q(s, a; \theta) \quad (8)$$

式中: δ_i 是经验样本 i 的 TD 误差。

优先经验回放机制采用比例优先级的方法和 SumTree 数据结构以实现对经验样本的高效采样和优先级更新。

比例优先级是 PER 机制的关键部分,它通过对每个样本的优先级进行加权来确定样本被抽取的概率。对于每个经验样本 i 的优先级 p_i ,定义为 TD 误差的绝对值加上一个小的常数项,并通过参数进行缩放,如下所示:

$$p_i = (|\delta_i| + \epsilon)^\mu \quad (9)$$

式中: ϵ 为一个小常数,用于确保优先级 p_i 不为零,避免该样本被完全忽略; μ 为一个超参数,用于控制优先级对采样概率的影响程度。

样本 i 的采样概率 $P(i)$ 则与其优先级成正比:

$$P(i) = \frac{p_i^\sigma}{\sum_j p_j^\sigma} \quad (10)$$

式中: $\sum_j p_j^\sigma$ 为对所有经验样本的优先级进行加权求和,用于归一化。

为避免高优先级经验样本的过度采样对学习过程产生偏差,需要对样本进行重要性采样校正。样本的权重 w_i 计算公式为:

$$w_i = \left(\frac{1}{N \cdot P(i)} \right)^\sigma \quad (11)$$

式中: N 为经验池的大小; σ 为一个超参数,用于控

制重要性采样的调整程度。

SumTree是一种二叉树结构,其叶结点存储每个样本的优先级,而内部结点存储子树结点的优先级总和。这种特殊的结构使得优先级的更新和采样操作可以在 $O(\log N)$ 时间复杂度内完成,从而显著提高了大规模经验回放的效率。SumTree的树形结构如图5所示,其只用于存储经验样本的优先级,而经验样本的数据块仍存放在经验回放池中。

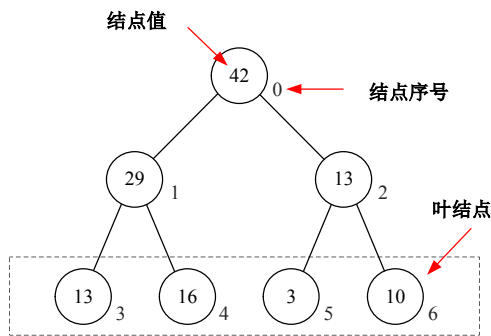


Fig. 5 Structure of SumTree

2.3 改进型DQN模型

单一改进策略对传统DQN整体性能的提升有限,本文通过结合上文所述的DDQN、Dueling network和PER三种机制,分别在不同层面改进传统DQN,使改进型DQN的整体性能得到了显著提升。改进型DQN模型的工作流程为:

(1)初始化阶段。首先,构建强化学习环境并获取初始状态。随后,分别初始化采用了Dueling Network架构的主网络与目标网络,其中主网络负责在线学习,目标网络则用于提供稳定的目标估计。此外,还生成了一个PER缓冲区,用于存储智能体与环境交互过程中获得的经验样本,并为每条样本分配初始优先级。

(2)数据采集阶段。智能体在该阶段采用epsilon-greedy策略与环境交互,根据预设概率随机选择动作以保证充分的探索,大多数情况下仍采取主网络选择的最优动作。执行选定动作后,环境返回相应反馈信息,智能体将获得的新经验样本及其初始优先级存入优先经验回放缓冲区中。

(3)核心训练阶段。当缓冲区中存储的经验样本达到一定数量后,训练过程进入核心阶段。首先,根据样本的优先级从缓冲区中抽取小批量数据,并为每条样本计算重要性采样权重,以校正非均匀采样可能引入的数据分布偏差。接着,主

网络利用Dueling network架构分别估计当前状态的价值和各动作的优势,结合二者生成加权组合Q值,选取Q值最高的动作。随后,目标网络利用DDQN机制,对该动作进行价值评估,计算目标反馈与当前估计之间的误差,并结合PER重要性采样权重形成加权损失,为后续梯度下降更新提供依据。完成梯度下降更新后,将基于更新后的预测误差调整缓冲区中相应样本的优先级,确保预测误差较大的样本在后续训练中有更高的采样概率。此外,训练过程中会定期将主网络的最新参数同步复制给目标网络,以维持目标估计的稳定性。

(4)训练循环。整个流程不断循环,直至达到设定的训练帧数或满足性能收敛条件。

本文完成了对传统DQN的三部分改进工作,逐步融合DDQN、Dueling network和PER三种机制,构建了改进型DQN模型,其框架如图6所示。这3种改进策略分属不同维度,它们互不冲突、相辅相成,有效地提升了传统DQN的整体性能。融合3种机制后,改进型DQN的损失函数为:

$$L(\theta) = E_{i \sim p} [w_i (y_n - Q(s, a; \theta, \alpha, \beta))^2] \quad (12)$$

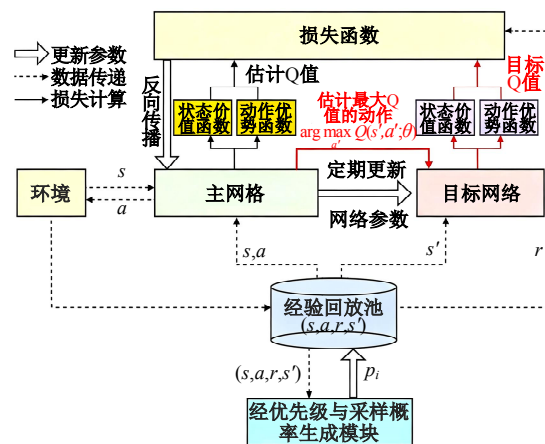


图6 改进型DQN的框架

Fig. 6 Architecture of improved DQN

2.4 神经网络结构

现搭建改进型DQN的深度学习神经网络结构,其主要由独热编码机制、空间注意力机制^[18]、通道注意力机制^[19]、多尺度卷积层、深度特征提取、价值流和优势流等模块构成,如图7所示。

独热编码机制是一种将分类数据转换为二进制格式的数据编码方法,广泛应用于机器学习、路径规划等任务。本文利用独热编码技术,在网络

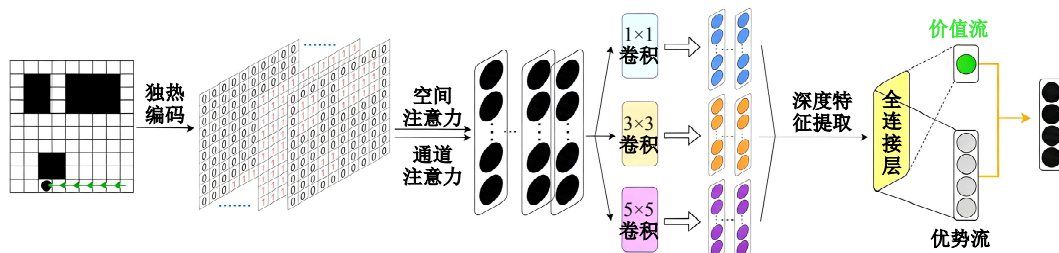


图7 改进型DQN的神经网络结构

Fig. 7 Neural network architecture of Improved DQN

的输入层将二维栅格地图信息转换为障碍物区域、未覆盖栅格、已覆盖栅格和智能体当前位置共4个通道信息,以便智能体更高效地学习与处理。

空间注意力机制通过强调特征图中重要的空间位置来增强模型的特征表示能力。该模块由一个卷积层和一个Sigmoid激活函数组成。卷积层的输入通道数为4,输出通道数为1,卷积核大小为 7×7 ,填充为3,将生成一个与输入特征图大小相同的注意力权重图,对每个空间位置的特征进行加权,突出重要的空间区域。Sigmoid激活函数将输出值归一化到 $[0, 1]$ 范围内,对输入特征图进行加权。

通道注意力机制旨在对各通道的贡献进行加权,以提升模型的特征选择能力。首先,对输入特征图应用自适应池化操作,分别进行全局平均池化和全局最大池化。池化后得到的特征通过两个输入输出维度均为4的全连接层进行处理,第一个全连接层和第二个全连接层分别后接ReLU激活函数、Sigmoid激活函数,生成通道注意力权重,并将其与输入特征图的每个通道逐元素相乘,从而增强关键特征。

多尺度卷积层包含3个并行的卷积分支,分别使用 1×1 、 3×3 和 5×5 的卷积核提取不同尺度的特征,以提高模型的泛化能力。其中, 1×1 卷积分支的输入通道数为4,输出通道数为32; 3×3 卷积分支的输入通道数为4,输出通道数为64,填充为1; 5×5 卷积分支的输入通道数为4,输出通道数为32,填充为2;每个卷积层后接批归一化和ReLU激活函数。3个卷积分支提取的特征图通过拼接操作整合成一个大小为 $128 \times H \times W$ 的特征图。

在多尺度特征提取后,模型通过深度特征提取模块进行处理,进一步细化特征表示。该模块包含3个卷积层,卷积核大小均为 3×3 ,填充为1,通道数从128逐渐减少到64。每个卷积层后接批

归一化和ReLU激活函数。

价值流的输入大小为 $128 \times H \times W$,首先将其展平为 $128 \times H \times W$ 大小的一维向量,接着输入到第一个全连接层,输出大小为512,激活函数为ReLU;第二个全连接层的输出大小为256;第三个全连接层的输出大小为1,对应状态价值。优势流的结构与价值流相同,展平后的向量输入到第一个全连接层,输出大小为512,激活函数为ReLU;第二个全连接层的输出大小为256;第三个全连接层的输出大小为4,分别对应智能体动作空间中4个动作的优势。最终Q值通过价值流和优势流的加权组合得出。

3 可视化仿真实验分析

本文设计可视化仿真实验,对DQN、DQN+①、DQN+①+②、DQN+①+②+③(改进型DQN)共4种算法模型的覆盖性能进行评估对比。其中,①为DDQN机制,②为Dueling Network机制,③为PER机制。实验结果表明,改进型DQN相比于其他模型,其综合覆盖性能更好且收敛速度更快。

3.1 实验环境与参数配置

本实验的硬件环境为Intel(R) Xeon(R) Platinum 8481C+RTX4090D(24 GB),软件环境为Ubuntu 20.04+Python 3.8+CUDA 11.3+Torch 1.10.0+Gym 0.26.2。结合算法训练效果调整相关参数,仿真实验参数设置如表1所示。

3.2 参数敏感性分析与实验验证

当同一栅格被智能体覆盖两次及以上,则该栅格被视为重复覆盖。设需要覆盖的栅格数量为 N_a ,已覆盖的栅格数量 N_c ,重复访问栅格的次数为 R_c ,该回合中智能体运动的路径长度为 R_l 。

栅格覆盖率 C_{rate} 的计算公式为:

$$C_{rate} = \frac{N_c}{N_a} \quad (13)$$

表 1 仿真实验参数设置

Table 1 Parameter settings for simulation experiments

参数	含义	值
ω	网络学习率	1e-4
γ	折扣因子	0.99
ϵ	见式(9)	1e-5
μ	见式(9)	0.6
σ	见式(11)	0.4
H	二维栅格地图高度	10
W	二维栅格地图宽度	10
frame_max	最大训练帧数	1 000 000
frame_start	训练开始帧数	10 000
$ D $	回放经验池最大样本容量	100 000
$ K $	可选动作的总数	4
Batch_size	每次训练选取的样本数	32
epsilon_max	最大贪婪系数	1.0
epsilon_min	最小贪婪系数	0.01
epsilon_decay	至最小贪婪系数的帧数	30 000
max_steps	每回合最大步数	300
agent_num	智能体数量	1
update_freq	网络参数更新间隔/帧	1 000
obstacles_per	障碍物栅格的占比/%	16

栅格重复率 R_{rate} 的计算公式为:

$$R_{\text{rate}} = \frac{R_c}{R_l} \tag{14}$$

首先以 1e-4 的超参数学习率、0.99 的超参数折扣因子和 $\vec{G}(i,j)=\vec{G}(i',j')$ 时 +0.25 的奖励机制为部分基准参数,训练出一组模型数据作为对照组。随后,通过控制变量法分别调整单一参数,训练 6 组数据。每次训练时,智能体都将在具有随机障碍物和随机初始位置的环境中执行覆盖路径规划任务。共记录 3 000 个回合的数据,最后各项指标均取所有测试回合的平均值,实验结果如表 2 所示。

表 2 参数敏感性分析实验结果

Table 2 Experimental results of parameter sensitivity analysis

改进型 DQN 模型相关参数	栅格覆盖 率/%	栅格重复 率/%	路径长度	碰撞障碍物 次数	超出边界 次数
$\omega=1\text{e}-4, \gamma=0.99, \vec{G}(i,j)=\vec{G}(i',j')$ 时奖励值 +0.25	97.90	35.83	154.41	5.46	0.34
$\omega=1\text{e}-3$	98.37	31.33	144.91	3.97	0.51
$\omega=1\text{e}-5$	83.89	71.03	291.75	13.11	1.97
$\gamma=0.8$	96.91	43.36	173.43	3.40	2.44
$\gamma=0.9$	97.85	39.97	164.51	4.41	1.54
$\vec{G}(i,j)=\vec{G}(i',j')$ 时奖励值 +0	92.41	53.94	230.00	15.79	2.01
$\vec{G}(i,j)=\vec{G}(i',j')$ 时奖励值 +0.5	98.04	35.67	158.67	9.03	0.27

对于本文覆盖路径规划任务,当学习率 ω 设置为 1e-3 时,模型训练初期收敛速度快,栅格覆盖率迅速上升,但后期由于梯度更新步幅过大导致出现震荡现象,训练不稳定;当学习率 ω 设置为 1e-4 时,能平衡收敛速度和稳定性;当学习率 ω 设置为 1e-5 时,收敛速度显著降低,在相同的训练帧数下,整体覆盖性能一般。

当折扣因子 γ 设置为 0.99 时,鼓励智能体进行长期规划,覆盖尽可能多的区域,探索更优的全局路径,降低路径长度和冗余;当折扣因子 γ 分别设置为 0.9 和 0.8 时,容易导致智能体更多地关注短期奖励,陷入局部最优。

当 $\vec{G}(i,j)=\vec{G}(i',j')$ 时给予奖励值 +0,即智能体在同一方向上连续运动时不给予正奖励。缺乏直线运动的引导,智能体可能会更多地采取折返或曲线路径,在局部区域迂回,转向频率提高,导致栅格重复率增加和路径冗余;当 $\vec{G}(i,j)=\vec{G}(i',j')$ 时给予奖励值 +0.25,能有效地引导智能体更多地探索覆盖新的区域,提高栅格覆盖率,优化全局路径;当 $\vec{G}(i,j)=\vec{G}(i',j')$ 时给予奖励值 +0.5,总体覆盖效果与奖励值 +0.25 相近,但引导直线方向的奖励值过高可能会增加智能体碰撞障碍物和超出边界的次数。

3.3 模型可视化仿真结果

训练过程中,智能体的每个动作及相应的环境都会产生一个样本,存放在经验回放池中。当其存储的经验样本达到一定数量后,神经网络开始训练;若智能体完成全覆盖任务或路径长度大于设定的回合最大步数,则该回合结束,进入新一轮回合;当训练次数达到设定的最大训练帧数时,结束训练。

网络训练时,使用 epsilon-greedy 策略,通过动态调整 ϵ 贪婪系数的值,平衡探索环境和利用网络策略的权重。训练过程中, ϵ 从 epsilon_max 开始,经过一定帧数训练达到 epsilon_min。本文对 4 种算法在训练过程中的损失曲线和奖励曲线进行对比。

如图 8 所示,横坐标表示训练次数,纵坐标表示损失值。由图可见,传统 DQN 损失偏大,收敛最慢;Improved DQN 相较其他算法更稳定,且收敛速度最快。

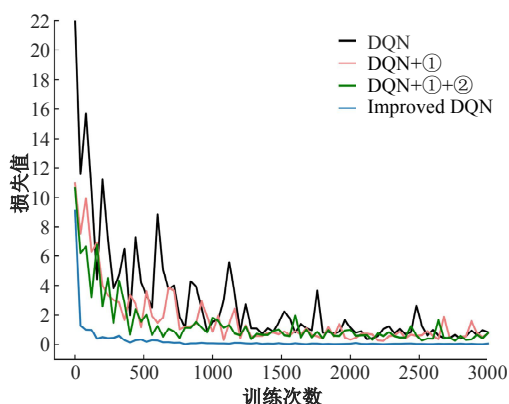


图 8 网络训练损失曲线

Fig. 8 Network training loss curve

如图 9 所示,横坐标表示训练回合,纵坐标表示奖励值。由图可见,传统 DQN 的平均奖励值最小且波动较大;Improved DQN 相较其他算法,收敛后奖励值最高、更稳定且波动较小。

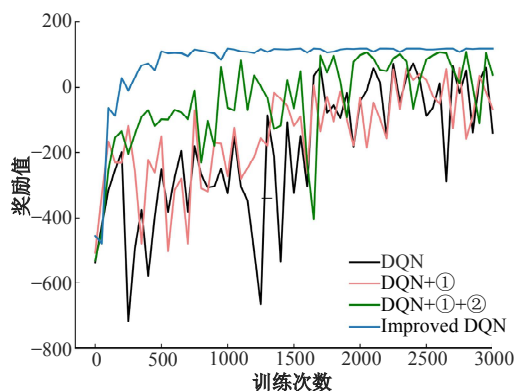


图 9 网络训练奖励曲线

Fig. 9 Network training reward curve

可视化仿真实验基于 OpenAI 平台的 Gym 库^[20]进行。4 种算法模型的仿真图如图 10 所示。

传统 DQN 模型在覆盖路径规划任务中表现欠佳。具体表现为路径冗余度高,方向转换频繁,碰撞障碍物以及超出边界的次数较多,且重复访

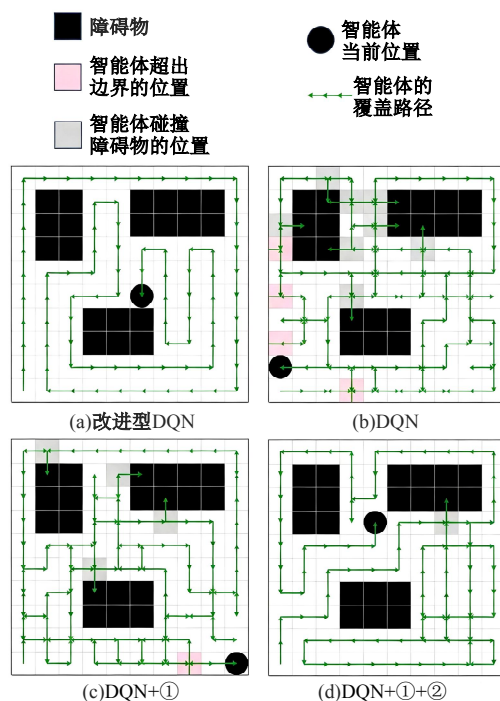


图 10 覆盖路径仿真图

Fig. 10 Coverage path simulation diagram

问栅格的频率较高,路径长度最长。这表明,传统 DQN 的原有机制在含有障碍物的二维栅格地图中,覆盖路径规划效果较差。

DQN+①模型相较于传统 DQN,在碰撞障碍物和超出边界的情况上有所改善,但总体提升效果一般。这表明,仅仅通过主网络选取评估值最高的动作,并利用目标网络对动作进行价值评估的策略,对传统 DQN 模型的改进有限,在覆盖路径规划效果上的提升幅度较小。

DQN+①+②模型的覆盖路径规划效果相比前两个模型有明显提升。该模型能够较好地保持在同一方向上运动,降低了栅格重复率,缩短了路径长度。这验证了 Dueling network 机制的有效性,这种 Q 函数分解方式能够有效区分环境状态的价值和各个动作的优势,从而使智能体能够根据不同环境状态选择合适的动作,减少路径迂回和冗余,碰撞障碍物和超出边界的动作得到有效控制。

改进型 DQN 模型的覆盖路径规划效果显著优于其他算法。其生成的路径平滑,转弯次数较少,栅格覆盖率最高,几乎不存在碰撞障碍物和超出边界的情况。这表明引入 PER 机制后,智能体能够充分学习 TD 误差较大的关键样本,有效提升了覆盖路径规划的性能。此外,智能体的转向

次数减少,成功学习到了同一方向连续移动时给予奖励的机制,进一步优化了其在覆盖路径规划任务中的表现。

3.4 消融实验

本消融实验旨在评估不同改进机制对 DQN 覆盖路径规划任务性能的影响。以传统 DQN 为基线模型,逐步引入 DDQN、Dueling network 和 PER 三种机制,并通过栅格覆盖率和栅格重复率这两个核心指标,以及路径长度、碰撞障碍物次数与超出边界次数 3 个辅助指标,综合评价模型的路径覆盖性能。对于每种模型,智能体将在具有随机障碍物和随机初始位置的环境中执行路径规划任务。实验的相关参数设置如表 1 所示,共记录 3 000 个回合的数据,最后各项指标均取所有测试回合的平均值。消融实验结果如表 3 所示。

作为基线模型,传统 DQN 的表现相对一般:栅格覆盖率为 89.62%,栅格重复率为 64.27%,表明其在探索方面有一定的局限,未能充分覆盖环境;路径长度为 281.5,路径冗余度较高,模型的路径规划能力有待提升;碰撞障碍物次数为 24.06,模型避障能力有待提高。

在单一策略的改进中,加入 PER 机制后,模型的表现提升最大。栅格覆盖率提升了 4.85%;栅格重复率降低了 13.09%;超出边界次数从 3.45 次减少至 2.62 次;路径长度缩短了 65.46 个栅格单位,减少了路径的迂回和频繁转向;碰撞障碍物次数降低为 11.69 次,模型的避障性能有所改善。这表明,PER 机制使模型更高效地利用高优先级的经验样本,显著提升了路径规划效率、决策精度和整体模型性能。

在双策略组合的改进中,加入 Dueling Network 机制和 PER 机制的模型性能最优。相比于

单一改进最优策略 PER,栅格覆盖率从 94.47% 提升至 96.26%;栅格重复率从 51.18% 降低至 44.76%;路径长度从 215.99 大幅缩短至 187.66 个栅格单位,减少了路径冗余;碰撞障碍物次数、超出边界次数分别减少了 0.93 次、1.85 次。这表明,Dueling network 机制通过区分状态的价值和动作的优势,进一步提高了模型在复杂环境下的决策能力和探索效率。

融合三种机制的改进型 DQN 模型展现了最佳的综合性能。相较于最优双策略组合改进,其栅格覆盖率提升了 1.64%,栅格重复率、路径长度、碰撞障碍物次数、超出边界次数分别降低了 8.93%、33.25 个栅格单位、5.14 次、0.43 次。这些显著的量化数据表明,所引入的 3 种机制在提高智能体的路径规划效率和可靠性方面发挥了关键作用。实验结果也证明了这 3 种机制在不同方面的有效性及良好的协同作用。同时,改进型 DQN 模型成功解决了无人机覆盖路径规划中,传统 DQN 模型存在的覆盖率低、路径冗余度高、收敛速度慢等问题。

表 3 中,仅 Dueling network 机制增加了模型的参数量(7.13 M)与计算开销(1.22 G),可部署在目前主流的嵌入式设备上(如 STM32F0 的 MIPS 为 60)。

3.5 泛化能力验证

本实验扩展实验场景,分别测试了不同模型在 $H=W=12$ 场景下和 $H=W=15$ 场景下的覆盖路径规划效果。由于栅格地图规模扩大,为确保模型能够充分探索,将参数 max_steps 分别设置为 max_steps=350 和 max_steps=400。实验共记录 3 000 个回合的数据,最后各项指标均取所有测试回合的平均值。扩展实验场景结果如表 4 所

表 3 消融实验结果

Table 3 Ablation experiment results

DDQN	Dueing network	PER	栅格覆盖率/%	栅格重复率/%	路径长度	碰撞障碍物次数	超出边界次数	训练时间	#Param /M	#FLOPs /G
J	J	J	89.62	64.27	281.45	24.06	3.45	3:22:04	3.72	1.11
└	J	J	93.92	61.83	273.66	23.78	0.85	4:14:59	3.72	1.11
J	└	J	93.72	56.13	233.79	12.72	3.65	3:56:05	7.13	1.22
J	J	└	94.47	51.18	215.99	11.69	2.62	3:26:30	3.72	1.11
└	└	J	94.82	50.40	219.92	19.37	0.71	4:31:45	7.13	1.22
J	└	└	96.26	44.76	187.66	10.76	0.77	4:07:36	7.13	1.22
└	J	└	95.94	48.76	201.16	10.05	3.34	3:38:55	3.72	1.11
└	└	└	97.90	35.83	154.41	5.46	0.34	4:36:02	7.13	1.22

表 4 扩展实验场景结果
Table 4 Ablation experiment results

模型	H, W	max_steps	栅格覆盖率/%	栅格重复率/%	路径长度	碰撞障碍物次数	超出边界次数
DQN	10	300	89.62	64.27	281.45	24.06	3.45
DQN+①	10	300	93.92	61.83	273.66	23.78	0.85
DQN+②	10	300	93.72	56.13	233.79	12.72	3.65
DQN+③	10	300	94.47	51.18	215.99	11.69	2.62
改进型 DQN	10	300	97.90	35.83	154.41	5.46	0.34
DQN	12	350	84.68	63.46	349.70	20.36	26.14
DQN+①	12	350	86.14	64.84	348.76	13.07	43.30
DQN+②	12	350	85.16	61.47	345.89	24.45	31.46
DQN+③	12	350	89.80	58.61	329.08	18.76	12.15
改进型 DQN	12	350	94.13	57.00	320.95	13.00	5.46
DQN	15	400	80.77	53.10	399.98	29.49	19.75
DQN+①	15	400	84.64	50.86	400.00	30.84	30.76
DQN+②	15	400	83.14	51.68	399.99	30.51	35.20
DQN+③	15	400	87.60	52.56	399.94	18.17	25.78
改进型 DQN	15	400	92.70	49.66	391.99	14.29	4.96

示,验证了改进型 DQN 模型在不同地图规模下仍然可以很好地完成覆盖路径规划任务,泛化能力较强。

4 结束语

本文聚焦无人机覆盖路径规划任务,提出了一种融合多机制的改进型 DQN 模型。首先,进行工作空间的建模;之后,基于传统 DQN 模型,依次融合 DDQN、Dueling network 和 PER 三种机制,构建了改进型 DQN 模型;最终,在二维栅格地图上设计可视化覆盖路径规划仿真实验,以栅格覆盖率、栅格重复率、路径长度、碰撞障碍物次数、超出边界次数为评价指标进行消融实验。实验结果表明:本文改进型 DQN 模型的性能在多方面优于传统 DQN 等模型。同时,改进型 DQN 模型成功解决了无人机覆盖路径规划中,传统 DQN 模型存在的覆盖率低、路径冗余度高、收敛速度慢等问题。

参考文献:

- [1] Buchelt A, Adrowitzer A, Kieseberg P, et al. Exploring artificial intelligence for applications of drones in forest ecology and management[J]. Forest Ecology and Management, 2024, 551: No. 121530.
- [2] Bakirci M. Smart city air quality management through leveraging drones for precision monitoring[J]. Sustainable Cities and Society, 2024, 106: No. 105390.

- [3] Ibrahim Z T, He J. Slam technology on disaster response[J]. World Journal of Engineering and Technology, 2024, 12(3): 695-714.
- [4] Xing B, Wang X, Yang L, et al. An algorithm of complete coverage path planning for unmanned surface vehicle based on reinforcement learning[J]. Journal of Marine Science and Engineering, 2023, 11(3): No. 645.
- [5] Kim J. Autonomous robot vacuum system composed of a cleaner robot and a dust storage robot[J]. Journal of the Franklin Institute, 2024, 361(11): No. 106938.
- [6] Wan S, Chen Z, Dong J. An efficiency-based interactive dynamic technique with interval-valued hesitant fuzzy constraint cone for rescue route planning[J]. Expert Systems with Applications, 2023, 231: No. 120648.
- [7] Dijkstra E W. A note on two problems in connexion with graphs[J]. Numerische Mathematik, 1959, 1(1): 269-271.
- [8] Hart P E, Nilsson N J, Raphael B. A formal basis for the heuristic determination of minimum cost paths [J]. IEEE transactions on Systems Science and Cybernetics, 1968, 4(2): 100-107.
- [9] Xu R, Xia Y, Chen P. A Multi-Objective Particle Swarm Optimization Algorithm for Drone Path Planning in Forest Firefighting[C]//The 11th International Conference on Electrical and Electronics Engineering (ICEEE), Marmaris, Turkiye, 2024:523-527.
- [10] 唐颂, 吴建源. 基于改进遗传算法的协同航迹规划

- 方法[J]. 电光与控制, 2024, 31(7): 8-12, 26.
- Tang Song, Wu Jian-yuan. A Cooperative Trajectory Planning Method Based on Improved Genetic Algorithm[J]. Electronics Optics & Control, 2024, 31(7): 8-12, 26.
- [11] Luo J, Shun H, Hao Y. A path planning method for comprehensive area coverage utilizing neuron activity [C]//The 5th International Conference on Robotics, Intelligent Control and Artificial Intelligence(RICAI), Hangzhou, China, 2023: 379-383.
- [12] 屈博琛. 基于深度学习自动微分的无人机路径规划[J]. 自动化与仪表, 2025, 40(2): 66-72.
- Qu Bo-chen. Drone path planning based on deep learning with automatic differentiation[J]. Automation & Instrumentation, 2025, 40(2): 66-72.
- [13] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J/OL]. [2025-02-16]. <https://arxiv.org/abs/1312.5602>.
- [14] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 2016, 30(1): 2094-2100.
- [15] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning [C]//International Conference on Machine Learning, New York, USA, 2016: 1995-2003.
- [16] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[J/OL]. [2025-01-22]. <https://arxiv.org/abs/1511.05952>.
- [17] Watkins C J C H. Learning from delayed rewards [D]. Cambridge: King's College, University of Cambridge, 1989.
- [18] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28: 2017-2025.
- [19] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7132-7141.
- [20] Brockman G, Cheung V, Pettersson L, et al. Openai gym[J/OL]. [2025-01-26]. <https://arxiv.org/abs/1606.01540>.