工程科学学报
Chinese Journal of Engineering

未知环境下无人机集群智能协同探索路径规划

王伟伦 尤明 孙磊 张秀云 宗群

Intelligent cooperative exploration path planning for UAV swarm in an unknown environment

WANG Weilun, YOU Ming, SUN Lei, ZHANG Xiuyun, ZONG Qun

引用本文:
王伟伦, 尤明, 孙磊, 张秀云, 宗群. 未知环境下无人机集群智能协同探索路径规划[J]. 北科大：工程科学学报, 2024, 46(7): 1197–1206. doi: 10.13374/j.issn2095–9389.2023.10.15.002
WANG Weilun, YOU Ming, SUN Lei, ZHANG Xiuyun, ZONG Qun. Intelligent cooperative exploration path planning for UAV swarm in an unknown environment[J]. *Chinese Journal of Engineering*, 2024, 46(7): 1197–1206. doi: 10.13374/j.issn2095–9389.2023.10.15.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2023.10.15.002

工程科学学报, 第 46 卷, 第 7 期: 1197−1206, 2024 年 7 月
Chinese Journal of Engineering, Vol. 46, No. 7: 1197−1206, July 2024
https://doi.org/10.13374/j.issn2095-9389.2023.10.15.002; http://cje.ustb.edu.cn

未知环境下无人机集群智能协同探索路径规划

王伟伦[1)], 尤　明[2)], 孙　磊[3)], 张秀云[1)✉], 宗　群[1)]

1) 天津大学电气自动化与信息工程学院, 天津 300072　2) 沈阳飞机设计研究所, 沈阳 110035　3) 陆军航空兵研究所, 北京 101100
✉通信作者, E-mail: zxy_11@tju.edu.cn

摘　要　随着无人机执行任务复杂性与环境种类多样性的不断提高, 多无人机集群系统逐渐得到国内外的广泛关注, 无人机路径规划成为当前研究热点. 考虑到传统路径规划算法一般需要先验地图信息, 在搜索救援等环境未知场景中难以满足, 本文提出了一种基于强化学习的未知环境下的无人机集群协同探索路径规划方法. 首先, 考虑无人机集群协同探索任务特点及动力学、避碰避障等约束条件, 基于马尔可夫决策过程, 建立无人机集群协同探索博弈模型与评价准则. 其次, 提出基于强化学习方法的无人机集群协同探索方法, 建立基于策略−评判网络的双网络架构, 并利用随机地图增强探索方法面对未知环境的泛化能力. 每架无人机在探索过程中不断收集地图信息, 并基于环境信息和个体间的共享信息调整自身策略, 通过迭代训练实现未知环境下的集群协同探索. 最后, 基于 Unity 搭建无人机集群协同探索虚拟仿真平台, 并与非合作的单智能体算法进行对比试验, 验证了本文所提算法在任务成功率、任务完成效率和回合奖励等方面均具有优势.

关键词　无人机; 集群; 深度强化学习; 自主探索; 路径规划
分类号　TP391.41

Intelligent cooperative exploration path planning for UAV swarm in an unknown environment

WANG Weilun[1)], *YOU Ming*[2)], *SUN Lei*[3)], *ZHANG Xiuyun*[1)✉], *ZONG Qun*[1)]

1) School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
2) Shenyang Aircraft Design and Research Institute, Shenyang 110035, China
3) Institute of Army Aviation, Beijing 101100, China
✉Corresponding author, E-mail: zxy_11@tju.edu.cn

ABSTRACT　Owing to the increasing complexity of task execution and a wide range of variability in environmental conditions, a single unmanned aerial vehicle (UAV) is insufficient to meet practical mission requirements. Multi-UAV systems have vast potential for applications in areas such as search and rescue. During search and rescue missions, UAVs acquire the location of the target to be rescued and subsequently plan a path that circumvents obstacles and leads to the target. Traditional path-planning algorithms require prior knowledge of obstacle distribution on the map, which may be difficult to obtain in real-world missions. To address the issue of traditional path-planning algorithms that rely on prior map information, this paper proposes a reinforcement learning-based approach for the collaborative exploration of multiple UAVs in unknown environments. First, a Markov decision process is employed to establish a game model and task objectives for the UAV cluster, considering the characteristics of collaborative exploration tasks and various constraints of UAV clusters. To maximize the search and rescue success rate, UAVs must satisfy dynamic and obstacle-avoidance constraints during mission execution. Second, a reinforcement learning-based method for the collaborative exploration of multiple UAVs is proposed. The multiagent soft actor–critic (MASAC) algorithm is used to iteratively train the UAVs' collaborative exploration

收稿日期: 2023−10−15
基金项目: 国家自然科学基金资助项目 (62373273, 62073234, 62373268, 62022060)

strategies. The actor network generates UAV actions, while the critic network evaluates the quality of these strategies. To enhance the algorithm's generalization capability, training is conducted in randomly generated map environments. To avoid UAVs being obstructed by concave obstacles, a breadth-first search algorithm is used to calculate rewards based on the path distance between the UAVs and targets rather than the linear distance. During the exploration process, each UAV continuously collects and shares the map information with all other UAVs. They make individual action decisions based on the environment and information obtained from other UAVs, and the mission is considered successful if multiple UAVs hover above the target. Finally, a virtual simulation platform for algorithm validation is developed using the Unity game engine. The proposed algorithm is implemented using PyTorch, and bidirectional interaction between the Unity environment and Python algorithm is achieved through the ML-Agents (Machine learning agents) framework. Comparative experiments are conducted on the virtual simulation platform to compare the proposed algorithm with a non-cooperative single-agent SAC algorithm. The proposed method exhibits advantages in terms of task success rate, task completion efficiency, and episode rewards, validating the feasibility and effectiveness of the proposed approach.

KEY WORDS unmanned aerial vehicle; swarm; deep reinforcement learning; automatic exploration; path planning

无人机(UAV)指无人驾驶航空器, 与有人机相比, 无人机具有成本低、尺寸小、机动性高、隐蔽性好以及生存能力较强等优势, 被广泛应用于战场侦察、物资运输、气象监测和资源勘探等军用和民用领域. 随着任务环境的复杂性和任务种类的多样性不断提高, 单架无人机由于载荷、续航能力有限, 难以完成大规模的复杂任务. 为了解决单架无人机存在的种种问题, 需要对无人机集群系统的任务分配[1]、目标检测[2]、航路规划等关键技术加强研究, 使未来无人机的应用方式向智能化、集群化发展[3].

作为无人机集群控制的关键技术之一, 航路规划指根据各无人机的特定任务, 考虑障碍物等环境因素以及编队避碰和无人机自身飞行性能限制, 在满足多约束条件的前提下为多无人机规划出从起始点到目标点的可行航路, 实现指定性能指标最优[4]. 航路规划问题的求解方法包括以迪杰斯特拉算法(Dijkstra)和 A 星算法(A Star)为代表的路径搜索算法[5−6]、以人工势场法(APF)为代表的势场方法[7−8]、以粒子群(PSO)和差分进化(DE)为代表的启发式算法[9−13] 和以模型预测控制(MPC)为代表的最优控制方法[14−16]. 航路规划问题是一个非确定性多项式(NP)问题, 当问题求解规模不断增大时, 难以保证求解计算的实时性, 无法适应动态航路规划的任务需求[17]. 此外, 无人机探索任务作为一类特殊的航路规划问题, 不仅具有一般航路规划问题的难点, 还面临着环境信息不确定的挑战, 无人机在探索过程中不断获得关于环境的新信息, 这就要求算法具有对环境变化做出快速响应的在线动态规划能力.

随着以神经网络为核心的人工智能技术不断发展, 多种基于学习的算法被应用于无人机航路规划[18]. Wu 等[19] 提出了一种有效的深度强化学习训练方法, 通过设计惰性训练方法和将冗余经验移除操作, 可以减少大部分训练时间而不会损失过多的准确性. Li 等[20] 构建了一个通用的探索框架, 将探索问题分解为决策、规划和建图子过程, 增强系统的模块化程度, 提出一种将深度强化学习与和即时定位与地图构建(SLAM)相结合的决策算法, 该算法使用深度神经网络从局部地图中学习探索策略. Lu 等[21] 提出 MGRL(Markov graph reinforcement learning) 算法, 将图神经网络与强化学习结合, 对现实环境建立图网络模型, 通过强化学习训练模型参数, 实现视觉导航. Sonny 等[22] 提出一种基于网格图的路径规划方法, 优先考虑最短无人机路径与路径障碍物, 采用 Q 学习方法实现动态避障与路径重规划. 现有研究大多只关注单架无人机在探索任务中的路径规划, 未考虑无人机集群内部多无人机间的信息共享与协作.

基于此, 本文提出了一种基于强化学习的未知环境下多无人机协同探索路径规划方法. 首先, 基于马尔可夫决策过程建立无人机集群协同探索博弈模型, 考虑无人机集群协同探索任务特点设置约束条件. 其次, 提出基于强化学习方法的无人机集群协同探索方法, 通过模型建立—迭代训练—在线决策的过程优化无人机集群路径规划策略, 基于无人机集群内部信息共享, 实现未知环境下的协同探索. 最后, 基于 Unity 搭建无人机集群协同探索虚拟仿真平台, 并与非合作的单智能体算法进行对比试验, 验证了本文所提方法的有效性.

1 无人机集群探索问题描述与模型建立

1.1 无人机集群探索问题描述

1.1.1 任务场景定义

考虑到无人机集群在不同环境下执行探索任

务的条件存在较大差异, 对本文研究的任务场景做出以下说明和限定:

(1)假设某任务区域内存在若干静态障碍物、5 架己方无人机与 1 个待救援目标, 己方无人机以避开障碍物并到达待救援目标地点为任务目标.

(2)无人机需要探索的环境为四周存在边界的长方形城市环境, 环境中分布着建筑物, 无人机飞行过程中不能与建筑物和其他无人机发生碰撞. 无人机飞行高度为 40 m 且只在固定高度的二维平面内飞行, 将高度大于等于飞行平面的建筑物视为障碍物.

(3)在任务开始时, 数架无人机被随机抛洒到环境的不同区域, 可以保证初始状态不会与建筑物发生撞击.

(4)每架无人机都装有传感器和通信装置, 可以感知自身周围的局部环境障碍物信息, 并可以在无人机集群内部进行信息共享.

(5)在环境中存在一个目标物体, 目标的位置在初始时已知, 且不会随时间改变.

(6)若在规定时间内, 有 3 架及以上无人机到达待救援目标地点并同时在地点上空悬停 3 个时间步以上, 那么任务成功, 否则时间耗尽后任务失败.

(7)当任务结束时, 到达目标地点所花费时间越少, 收益越大.

综上, 无人机集群探索任务目标可以描述为: 考虑己方无人机飞行速度、避障避碰、边界等约束条件, 利用各无人机的自身观测信息和无人机集群共享信息, 在规定时间内到达待救援目标地点.

1.1.2 任务约束条件

在执行任务的过程中, 受自身设备及安全限制, 无人机需满足各项约束条件, 包含速度约束、避障约束、避碰约束及边界约束等:

(1)速度约束.

在探索任务中, 受自身机动能力的限制, 无人机存在最大速度限制, 即

$$\|v_i\| \leqslant v_{\max} \tag{1}$$

其中, i为无人机的编号, $i \in [1,N]$, N为无人机的数量; v_i为无人机当前时刻的速度, $v_{\max}$为无人机的最大速度.

(2)避障约束.

由于区域内存在若干静态障碍物, 出于安全考虑, 无人机在飞行过程中不能与障碍物碰撞, 即两者的相对距离须保持在安全范围, 即

$$\Delta d_{ik} > d_{\min} \tag{2}$$

其中, $\Delta d_{ik} = \|p_i - p_k\|$, 表示第$i$个无人机相对其周围第$k$个障碍物之间的距离, p_i和p_k分别表示第i个无人机的位置以及第k个障碍物中心点的位置. $d_{\min}$表示无人机的最大安全半径.

(3)避碰约束.

为安全考虑, 己方无人机之间不能互相碰撞, 两者的相对距离也需要保持在安全范围内, 即

$$\Delta d_{ij} > d_{\min} \tag{3}$$

其中, $\Delta d_{ij} = \|p_i - p_j\|$表示第$i$个无人机相对第$j$个无人机的距离, $i, j \in [1,N]$, $i \neq j$.

(4)边界约束.

为了探索任务的顺利进行, 无人机在飞行过程中不能超过给定任务区域, 即

$$0 \leqslant p_{i,\xi} \leqslant d_{\text{bound},\xi} \tag{4}$$

其中, $\xi \in [1, \xi_{\max}]$表示无人机的运动维度, 对于本文研究的二维探索环境, $\xi_{\max} = 2$. $d_{\text{bound},\xi}$表示ξ维度上的区域边界. $p_{i,\xi}$表示ξ维度上无人机的位置.

1.2 无人机集群探索模型建立

强化学习最主要的部分是智能体和环境. 在智能体与环境交互的每一步中, 智能体获取到环境的状态信息, 然后根据自身的策略决定要采取的行动. 当智能体采取行动后, 环境会发生变化达到新的状态, 同时给予智能体奖励来告知其当前环境状态的好坏.

根据上述过程, 可以将强化学习问题抽象为一个马尔可夫决策过程(MDP), MDP 由五元组$<\boldsymbol{S},\boldsymbol{A},P,R,\gamma>$构成. 其中, $\boldsymbol{S}$为所有有效状态的集合; $\boldsymbol{A}$为所有有效动作的集合; $R:\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\to\mathbb{R}$为奖励函数, $r=R(s,a,s')$代表了从状态s采取动作a后, 转移到新状态s'时获得的奖励; $P:\boldsymbol{S}\times\boldsymbol{A}\to P(\boldsymbol{S})$为状态转移概率函数, $P(s'|s,a)$代表了从状态s采取动作a后, 转移到新状态s'的概率; γ为衰减因子, 它的值越大, 表示越关注未来的奖励.

下面给出无人机集群探索任务的马尔可夫决策过程模型.

1.2.1 状态集S和观测集O

状态(State)和观察(Observation)都是无人机在与环境交互过程中所获取的对于环境的描述, 不同之处在于状态是对环境整体的完备描述, 而观察只是对环境的部分描述, 可能遗漏部分信息. 然而, 单架无人机很难获取到环境的全部状态, 因此对于多无人机的协同探索任务有必要对状态和观测加以区分.

无人机集群协同探索环境状态信息如表 1 所

示, 其中包括二维矩阵和一维向量表示的数据. 二维矩阵主要用来表示与地图障碍物有关的信息, 一维向量主要用来表示位置、速度等与地图无关的信息. 环境状态信息是最全面的信息, 包括全部的地图障碍物信息和所有无人机的详细信息.

无人机观测信息如表 2 所示. 与状态信息不同的是, 观测信息的设置尽可能贴近实际任务的执行情况, 相比全局状态信息, 观测信息只包含单架无人机自身传感器的信息和最低限度的无人机集群共享信息.

1.2.2 动作集 A

由于障碍物是将连续的环境离散化为栅格地图形式表示的, 无人机的动作空间也可以从空间中的连续移动离散化为栅格地图的格子间移动. 很自然地, 可以将无人机的动作设置为向前、后、左、右移动一格, 然而这种方法也存在一些问题: 首先, 这种移动方式没有考虑到无人机的方向. 人类在现实中操控无人机时, 是以无人机前后左右的朝向为基准, 而不是以地理的东南西北为基准. 其次, 这种移动方式没有考虑到无人机的速度. 当栅格较密集时, 如果无人机仍每次只移动一格, 会导致决策效率过低; 如果将每个方向的移动细分为移动不同距离, 则动作空间的维度会成倍增加. 改进的动作集如表 3 所示.

表 1 环境状态信息

Table 1 State information of the environment

Data type	State name	Size	Description
Two-dimensional matrix	Obstacle area	(L_1, L_2)	0 for passable area, 1 for obstacle
	Discovered path	(L_1, L_2)	Initially set to 0, changed to 1 when the grid cell is detected and is a passable area.
	Explored area	(L_1, L_2)	Initially set to 0, changed to 1 when the grid cell is detected.
One-dimensional vector	Target position	$(2,)$	The position of the target in the grid map
	UAV positions	$(2N,)$	The positions of all UAVs in the grid map
	UAV directions	$(N,)$	The directions of all UAVs
	UAV speeds	$(N,)$	The speeds of all UAVs

Notes: L_1 and L_2 are the length and width of the grid map.

表 2 无人机观测信息

Table 2 Observation information of UAV

Data type	Name	Size	Description
Two-dimensional matrix	Discovered path	(L_1, L_2)	Initially set to 0, changed to 1 when the grid cell is detected and is a passable area.
	Explored area	(L_1, L_2)	Initially set to 0, changed to 1 when the grid cell is detected.
One-dimensional vector	Target position	$(2,)$	The position of the target in the grid map
	UAV position	$(2,)$	The position of this UAV in the grid map
	UAV direction	$(1,)$	The direction of this UAV
	UAV speed	$(1,)$	The speed of this UAV
	Nearest obstacles	$(4,)$	The distances to the nearest obstacle in the four directions
	Other UAV Positions	$(2\times(N-1),)$	The positions of other UAVs in the grid map

表 3 无人机动作集

Table 3 Action set of UAV

Action name	Description
Turn left	The UAV's position remains unchanged, and its direction turns 90° to the left.
Turn right	The UAV's position remains unchanged, and its direction turns 90° to the right.
Decelerate and move forward	The UAV's speed decreases by 1, its direction remains unchanged, and it moves forward based on its speed.
Maintain speed and move forward	The UAV's speed remains unchanged, its direction remains unchanged, and it moves forward based on its speed.
Accelerate and move forward	The UAV's speed increases by 1, its direction remains unchanged, and it moves forward based on its speed.
emergency stop	The drone sets its speed to 0, and its direction remains unchanged.

1.2.3 奖励函数R

以规定时间内点亮地图与到达目标地点为最终目标, 建立奖惩机制, 确定单步决策获得的收益值. 对于单架无人机, 收益值主要包括两部分:

(1)探索奖励.

无人机到达救援目标地点之前, 因为地图障碍物信息是未知的, 需要先对地图进行探索才能找到一条通往目标地点的通路. 但是, 探索的过程要以目标为导向, 不能进行无意义的探索. 因此可以采用一种考虑到障碍物分布的更加复杂的奖励函数, 即通过广度优先搜索[23]的方式, 计算到达目标的最短路径, 并根据无人机是否按照最短路径移动, 对无人机的行为进行奖励或惩罚, 从而使得智能体学习到路径规划的能力.

广度优先搜索(BFS)是较为简便的图的搜索算法之一, 其属于盲目搜寻法, 目的是系统地展开并检查图中的所有节点. 计算栅格地图中各点到目标的最短移动距离, 其流程如算法 1 所述.

算法1 广度优先搜索

输入:

二维栅格地图**map**, 目标坐标(i_{target}, j_{target})

输出:

存储最短移动距离的矩阵**res**

1: 初始化先进先出(FIFO)队列**queue**

2: 初始化用于存储最短移动距离的矩阵**dist**, 将矩阵**dist**的全部值置为−1

3: 将目标坐标加入队列**queue**, 将**dist**[i_{target}, j_{target}]的值置为0

4: 初始化距离计数器$c = 1$

5: while 队列**queue**不为空 do

6: 定义size =队列**queue**中当前元素个数

7: for 重复size次 do

8: 弹出队列**queue**的第一个元素, 坐标记为(i_{cur}, j_{cur})

9: for (i_{cur}, j_{cur})周围4个栅格 do

10: 该栅格坐标记为(i', j')

11: if (i', j')没有超出边界and **dist**[i', j'] == −1 and **map**[i', j'] == 1 then

12: 将(i', j')加入队列**queue**的末尾, **dist**[i', j'] = c

13: end if

14: end for

15: end for

16: 增加距离计数器$c = c + 1$

17: end while

BFS 算法的基本思想是, 从目标点开始, 将目标点加入队列, 然后依次将队列中的元素弹出, 并将其周围的元素加入队列, 直到队列为空. 通过一轮轮循环, 就可以访问所有的栅格并计算其到目标点的最短移动距离. BFS 算法返回的结果是一个矩阵, 若某一栅格值为−1, 则表示该栅格目前不能通过任何一条路径到达目标点, 否则表示该栅格到目标点的最短移动距离. 探索奖励为

$$r^{\text{explore}} = \mathbf{dist}[i, j] - \mathbf{dist}[i', j'] \tag{5}$$

其中, **dist**为 BFS 算法返回的二维矩阵, i, j为无人机原坐标, i', j'为无人机执行动作后的新坐标.

(2)目标奖励.

当无人机所在位置不能找到一条通往救援目标的通路时, 采用负的无人机与目标的距离作为奖励函数, 当无人机与目标距离较远时, 施加大的惩罚, 当无人机与目标距离较近时, 施加小的惩罚. 同时, 因为奖励总为负值, 为了最大化总回报, 这就要求无人机尽快完成任务, 结束该回合以避免再接受更多惩罚. 目标奖励为

$$r^{\text{target}} = -\sqrt{\left(i' - i_{\text{target}}\right)^2 + \left(j' - j_{\text{target}}\right)^2} \tag{6}$$

其中, i_{target}和j_{target}表示目标在栅格地图中的位置.

因此, 综合式 5) 和式 (6), 无人机的任务总奖励设计为

$$r = r^{\text{explore}} + r^{\text{target}} \tag{7}$$

基于上述的状态集、动作集与奖励函数, 建立了无人机集群协同探索博弈模型.

2 基于强化学习的无人机集群协同探索方法

2.1 无人机集群协同探索方法框架

无人机集群协同探索方法具体框架如图 1 所示.

步骤一: 模型建立. 根据任务约束条件和性能指标, 建立动态博弈模型; 在此基础上, 考虑任务目标和无人机搜索环境的约束条件, 以在有限时间内到达指定地点为目标, 根据人的专家经验, 建立奖惩机制, 确定每一时间步无人机立即收益值.

步骤二: 迭代训练. 采用中心式训练−分布式决策的强化学习算法, 建立决策网络和评价网络. 决策网络根据仿真平台所提供的无人机自身状态信息、探索过程中收集到的栅格地图信息和已知待救援目标位置信息, 决策无人机当前时刻的离散动作序号; 评价网络根据状态信息及决策信息评估决策结果的好坏, 并将新的交互经验补充到经验数据库中. 通过随机经验回放机制, 随机抽取数据库中的经验逐步训练评价网络和决策网络, 最终通过多次迭代的方式获取最优的无人机策略.

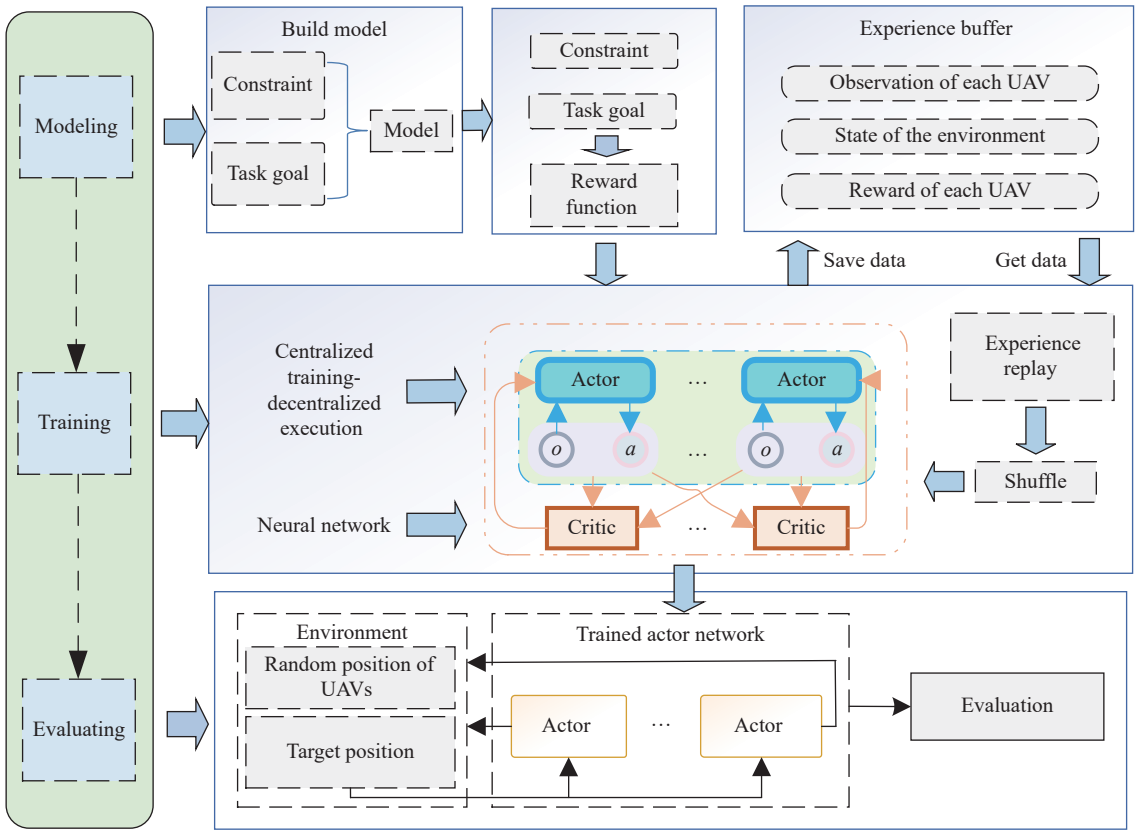

图 1　多无人机协同探索方法实施框架

Fig.1　Multi-UAV cooperative exploration method

步骤三: 在线决策. 在仿真平台中随机给出己方无人机和待救援目标的初始位置, 随机初始化地图障碍物分布, 采用训练好的决策网络实时产生无人机决策结果, 完成面向搜索任务的无人机集群运动规划.

2.2　基于 MASAC−Discrete 的无人机集群协同探索算法

SAC(Soft actor critic) 算法是应用较广泛的强化学习算法, 采用 Actor−Critic 网络架构. Actor−Critic 架构由 Actor 和 Critic 两个部分组成: Actor 部分是策略网络, Critic 部分是价值网络. Actor 网络的输入是观测, 输出是动作, 它的优化目标是能够根据观测输入选择最优动作; Critic 网络的输入是状态和动作, 输出是价值, 它的优化目标是能够最准确地评判状态−动作对的价值.

Haarnoja 等提出了基于最大熵的 SAC 强化学习算法[24], 之后又提出了该算法的具有自适应温度参数α的改进版本[25]. 由于 SAC 算法只适用于连续动作空间, 为了将其应用于离散动作空间, Christodoulou [26] 在 SAC 的基础上提出了 SAC−Discrete 算法, Critic 网络基于输入的状态一次性输出所有动作的价值Q. Wang 等[27] 将 SAC 算法扩展到多智能体环境, 提出了 MASAC(多智能体柔性策略−评判)算法. 本文研究的是离散动作空间下无人机集群探索路径规划问题, 因此基于已有算法提出 MASAC−Discrete 算法, 适用于离散动作空间的多智能体决策. 此外, 对于多智能体环境而言, 由于每个智能体的动作都会对环境产生影响, 如果简单地把单智能体强化学习算法应用于每个智能体的决策, 会使得环境始终处于不稳定的状态, 难以进行学习[28]. 而如果把所有智能体当作一个整体来训练, 会导致状态空间和动作空间的维度指数上升, 出现“维度爆炸”. 为了解决这种两难问题, 通常采用集中式训练和分布式执行框架(CTDE), 即训练时考虑其余智能体的状态, 执行决策时仅以自身的观测作为输入.

因此, 基于 MASAC−Discrete 算法与 CTDE 框架, 本文强化学习算法总体架构如图 2 所示.

为了更好地完成决策模型训练, 优化无人机集群决策性能, 做出如下改进:

(1)环境方面.

在强化学习迭代训练过程中, 如果无人机与环境交互时障碍物信息总是固定的, 那么虽然输入了栅格地图信息, 无人机很可能没有学习到理解

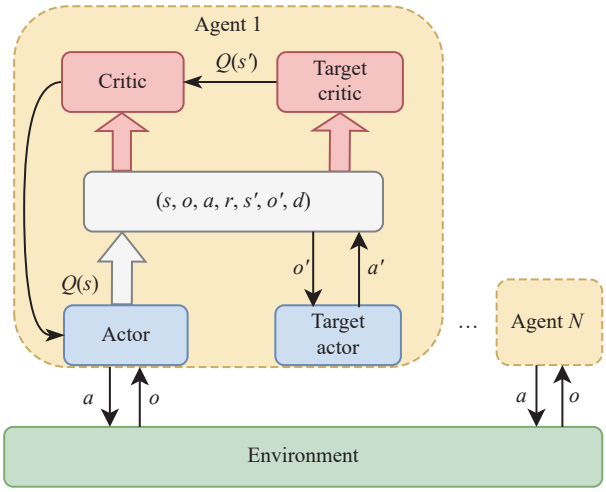

图 2 MASAC−Discrete 算法架构

Fig.2 Framework of the MASAC−Discrete algorithm

地图的能力, 只是记忆了某些固定的最优路径[29]. 以上问题很可能导致算法的泛化能力不足, 遇到新的未知环境时性能出现大幅下降. 因此, 在强化学习迭代训练过程中, 在每个回合开始时随机生成地图, 以保证无人机在每次迭代中都能学习到新的环境信息.

(2)网络结构方面.

由于地图信息是以栅格化的二维数组存储的, 且存在包括障碍物、探索区域等在内的多个二维数组, 这种数据类型与多通道图像十分相似. 因此将多个二维数组堆叠起来, 使用卷积神经网络(CNN)提取地图信息的高维特征再输入全连接网络(MLP), 可以更好地利用地图特征.

详细的网络训练过程如算法 2 所述.

算法2 无人机决策网络和评价网络训练更新算法

输入:

初始化Actor网络参数$\boldsymbol{\theta}=\{\theta^1,\cdots,\theta^N\}$, Critic网络参数$\boldsymbol{\phi}=\{\phi^1,\cdots,\phi^N\}$, 空的经验池$\boldsymbol{D}$, 温度系数$\alpha^n=1$, 最小期望熵$\bar{H}$

输出:

最优Actor网络参数$\boldsymbol{\theta}^*=\left\{\theta_1^*,\cdots,\theta_N^*\right\}$

1: 将在线网络参数复制到目标网络: $\phi_{\text{targ}}^n\leftarrow\phi^n$, $\theta_{\text{targ}}^n\leftarrow\theta^n$, $n\in[1\sim N]$为对应智能体的编号

2: repeat

3: 获取观测$\boldsymbol{o}=\{o_1,\cdots,o_N\}$、状态$s$, 并在环境中执行动作$\boldsymbol{a}=\{a_1\sim\pi_{\theta^1}(o_1),\cdots,a_N\sim\pi_{\theta^N}(o_N)\}$

4: 获取下一时刻观测$\boldsymbol{o}'=\{o_1',\cdots,o_N'\}$, 状态$s'$, 奖励$r=\{r_1,\cdots,r_N\}$和回合结束终止信号$d$

5: 在经验池$\boldsymbol{D}$中存储$(s,\boldsymbol{o},\boldsymbol{a},\boldsymbol{r},s',\boldsymbol{o}',d)$

6: 如果s'为回合结束, 重置环境

7: if 更新网络参数 then

8: for 更新次数 do

9: for N个智能体 do

10: 从经验池$\boldsymbol{D}$中随机抽出一批经验$\boldsymbol{B}=\{(s,o,a,r,s',o',d)\}$

11: 计算目标Critic网络的Q值:

$$y=r_n+\gamma(1-d)(\pi_{\theta_{\text{targ}}^n}(o_{n'})^{\mathrm{T}}(\min_{i=1,2}Q_{\phi_{\text{targ}}^n}^i(s',\tilde{\boldsymbol{a}}'_{\text{other}})-\alpha\ln\pi_{\theta_{\text{targ}}^n}(o_n')))$$

其中, $\tilde{\boldsymbol{a}}'_{\text{other}}=\{\cdots,\tilde{a}'_{n-1}\sim\pi_{\theta_{\text{targ}}^{n-1}}(o'_{n-1}),\tilde{a}'_{n+1}\sim\pi_{\theta_{\text{targ}}^{n+1}}(o'_{n+1}),\cdots\}$为其他智能体下一时刻的动作

12: 使用梯度下降更新Critic网络参数:

$$\nabla_{\phi^n}\frac{1}{|\boldsymbol{B}|}\sum_{(s,a,r,s',o',d)\in B}\left(Q_{\phi^n}^i(s,\boldsymbol{a}_{\text{other}})(a_n)-y\right)^2\qquad\text{for } i=1,2$$

其中, $\boldsymbol{a}_{\text{other}}=\{\cdots,a_{n-1},a_{n+1},\cdots\}$

13: 使用梯度上升更新Actor网络参数:

$$\nabla_{\theta^n}\frac{1}{|\boldsymbol{B}|}\sum_{(s,o,a)\in B}(\pi_{\theta^n}(o_n)^{\mathrm{T}}(\min_{i=1,2}Q_{\phi^n}^i(s,\boldsymbol{a}_{\text{other}})-\alpha\ln\pi_{\theta^n}(o_n)))$$

14: 使用梯度下降更新温度系数α:

$$\nabla_{\alpha}\frac{1}{|\boldsymbol{B}|}\sum_{o\in B}(\pi_{\theta^n}(o_n)^{\mathrm{T}}(-\alpha(\ln(\pi_{\theta^n}(o_n)+\bar{H}))))$$

15: 软更新目标网络参数:

$$\phi_{\text{targ}}^n\leftarrow\rho\phi_{\text{targ}}^n+(1-\rho)\phi^n\qquad\theta_{\text{targ}}^n\leftarrow\rho\theta_{\text{targ}}^n+(1-\rho)\theta^n$$

16: end for

17: end for

18:end if

19:until 网络收敛

3 算法仿真验证

3.1 基于 Unity 搭建虚拟仿真环境

虚拟仿真作为一种最经济、最高效的实验手段, 是多无人机协同探索研究过程中不可缺少的验证方式. 一方面, 多无人机协同探索概念尚处于初期阶段, 利用仿真方法可以对多无人机协同探索相关问题进行预先研究; 另一方面, 实体无人机集群规模大、成本高, 用仿真模型替代实体无人机, 在实物验证前开展大量仿真实验可节省成本, 加快研究进度.

本文基于 Unity 引擎开发了多无人机协同探索虚拟仿真环境. 使用 City Generator 插件构建城市场景, 每个回合开始时, 可在城市区域内随机生成街区、道路和建筑物, 产生不同的训练场景, 帮助提高强化学习算法面对未知环境的泛化能力.

仿真平台使用 ML-Agents[30] 插件实现 Unity 环境与 Python 算法的高速、双向的数据通信, 交互过程如图 3 所示. 首先, Unity 仿真环境端向 Python 强化学习算法端传输每架无人机的观测信息和环境的状态信息; 然后, 强化学习算法基于仿真环境的观测和状态信息使用神经网络推理出每架无人机的最优动作决策; 最后, 在仿真环境中执行返回的动作决策, 并将执行动作的奖励及新的观测和状态信息发送到算法端. 不断执行上述循环, 直到

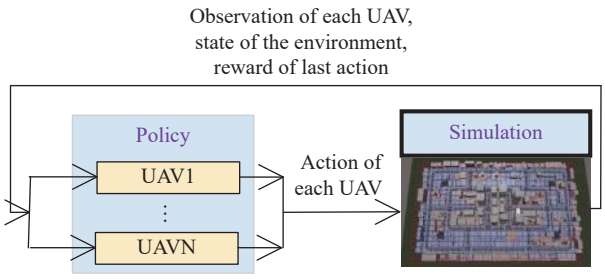

图 3　仿真平台数据交互过程

Fig.3　Data interaction process of simulation platform

满足任务成功条件或时间耗尽后任务失败.

3.2　虚拟仿真验证与分析

3.2.1　仿真参数设置

场景设置如 1.1.1 节所述, 5 架无人机被置于城市环境内, 初始位置在 8 km×5 km 的城市范围内随机生成, 无人机需要在与城市建筑物和其他无人机保持安全距离(大于 0.5 m)的情况下飞行到目标地点. 当 3 架及以上无人机到达目标地点并维持 3 个时间步时任务成功, 若 300 个时间步后仍未完成任务则判定超时, 任务失败.

本文算法的超参数如表 4 所示.

表 4　算法参数设置

Table 4　Algorithm parameter settings

Discount factor	Batch size	Capacity of replay buffer	Random action sampling episodes	Max steps per episode	Learning rate of actor network	Learning rate of critic network
0.99	128	1×10^{6}	100	300	1×10^{-4}	3×10^{-4}

3.2.2　仿真结果分析

在多无人机搜索仿真平台上训练, 下面从任务成功率、回合步数以及回合奖励三个评判指标对本文提出的 MASAC−Discrete 算法与单智能体的 SAC 算法进行比较.

(1)任务成功率.

无人机集群完成任务的成功率如图 4 所示. 为了衡量无人机当前策略的改进, 此处的成功率指之前 10 个回合完成任务的成功率. 由曲线可以看出, 在随机动作采样阶段与训练初期, 成功率一直为 0, 随着策略的优化, 无人机集群完成任务的成功率不断上升, 最后稳定在 90% 左右.

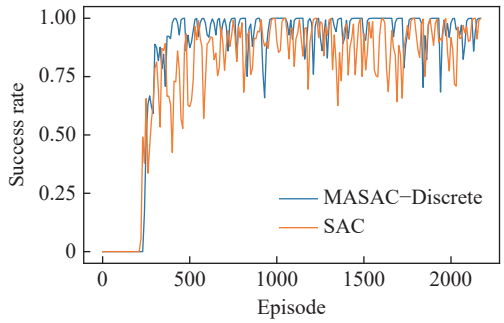

图 4　任务成功率曲线

Fig.4　Task success rate curve

(2)回合步数.

训练过程中每个回合时间步的长度如图 5 所示. 在训练初期, 无人机集群无法完成任务, 因此回合长度到达 300 步后, 环境被重置, 开始下一回合; 随着无人机策略不断改进, 回合长度逐渐缩短, 表明无人机完成任务所需的时间减少.

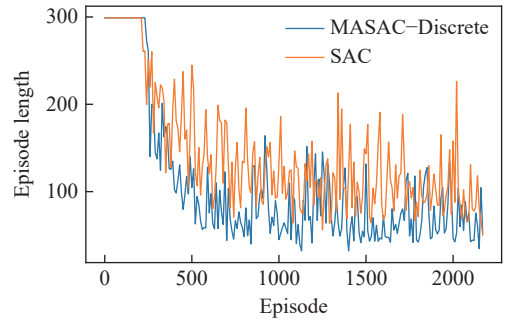

图 5　回合长度曲线

Fig.5　Episode length curve

(3)回合奖励.

无人机在回合内的平均总奖励如图 6 所示. 在训练初期, 无人机不能很好地靠近目标地点, 总是受到惩罚, 因此回合奖励为负值; 随着策略改进, 回合开始时迅速靠近目标地点避免较大惩罚, 同时完成任务得到大的奖励, 回合奖励逐渐上升.

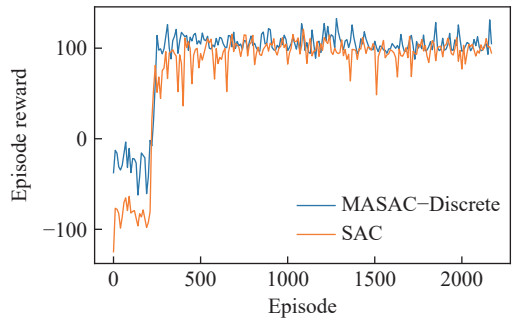

图 6　回合奖励曲线

Fig.6　Episode reward curve

从 3 条训练曲线与表 5 的性能指标对比可以看出, 本文所采用的基于 MASAC−Discrete 算法的

无人机集群协同探索路径规划方法在任务成功率、任务执行时间和任务奖励方面均优于单智能体的 SAC 算法.

表 5 本文采用 MASAC−Discrete 算法与 SAC 算法性能指标

Table 5 Performance metrics of the MASAC−Discrete algorithm and SAC algorithm

Algorithm	Task success rate/%	Episode length	Episode reward
MASAC−Discrete	92.55	121.46	109.51
SAC	83.23	85.79	87.05

无人机集群协同探索三维视景演示结果如图 7 所示, 无人机在探索飞行过程中逐步收集环境信息并在集群内部共享, 当收集到足够的环境信息后, 无人机飞向目标地点附近, 判定为任务成功.

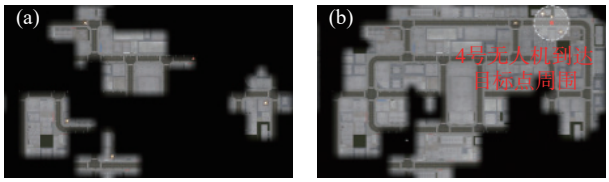

图 7 三维视景场景演示. (a) 无人机探索未知环境; (b) 无人机到达目标地点

Fig.7 3D visual simulation platform snapshots: (a) UAVs exploring unknown environment; (b) UAVs arrived at the target location

4 结论

针对传统的路径规划方法需要先验地图信息的问题, 本文考虑未知环境下无人机集群协同探索任务需求, 提出了一种基于强化学习的动态路径规划方法. 考虑无人机集群协同探索任务特点与无人机自身约束条件, 建立了协同探索博弈模型, 提出了基于 MASAC−Discrete 的无人机集群协同探索算法, 并结合未知环境下的无人机集群协同探索任务特点对算法提出了相应改进, 通过采用随机训练地图增强算法泛化能力, 利用卷积神经网络提取地图特征. 最后, 使用 Unity 搭建了无人机集群协同探索虚拟仿真环境, 并将 MASAC−Discrete 算法与 SAC 算法进行了对比实验, MASAC−Discrete 算法在任务成功率、任务完成效率、回合奖励都具有优势, 验证了本文提出方法的可行性和有效性.

参 考 文 献

[1] Peng Y L, Duan H B, Wei C. UAV swarm task allocation algorithm based on the alternating direction method of multipliers network potential game theory. *Chin J Eng*, 2022, 44(4): 792
(彭雅兰, 段海滨, 魏晨. 基于交替方向网络进化博弈的无人机集群任务分配. 工程科学学报, 2022, 44(4): 792)

[2] Tao L, Hong T, Chao X. Drone identification and location tracking based on YOLOv3. *Chin J Eng*, 2020, 42(4): 463
(陶磊, 洪韬, 钞旭. 基于 YOLOv3 的无人机识别与定位追踪. 工程科学学报, 2020, 42(4): 463)

[3] Duan H B, Qiu H X. *Unmanned Aerial Vehicle Swarm Autonomous Control Based on Swarm Intelligence*. Beijing: Science Press, 2018
(段海滨, 邱华鑫. 基于群体智能的无人机集群自主控制. 北京: 科学出版社, 2018)

[4] Wang R H, Gao X Y, Xiang Z R. Review on the manned/unmanned aerial vehicle cooperative system and key technologies. *J Ordnance Equip Eng*, 2023, 44(8): 72
(王荣浩, 高星宇, 向峥嵘. 有人/无人机协同系统及关键技术综述. 兵器装备工程学报, 2023, 44(8): 72)

[5] Fransen K J C, van Eekelen J A W M, Pogromsky A, et al. A dynamic path planning approach for dense, large, grid-based automated guided vehicle systems. *Comput Oper Res*, 2020, 123(1): 105046

[6] Li J A, Zhang W J, Hu Y T, et al. RJA-star algorithm for UAV path planning based on improved R5DOS model. *Appl Sci*, 2023, 13(2): 1105

[7] Qian X, Peng C, Nong C, et al. Dynamic obstacle avoidance path planning of UAVs // 34*th Chinese Control Conference*. Hangzhou, 2015: 8860

[8] Abeywickrama H V, Jayawickrama B A, He Y, et al. Potential field based inter-UAV collision avoidance using virtual target relocation // *IEEE* 87*th Vehicular Technology Conference*. Porto, 2018: 1

[9] Peng Z H, Sun L, Chen J. Online path planning for UAV low-altitude penetration based on an improved differential evolution algorithm. *J Univ Sci Technol Beijing*, 2012, 34(1): 96
(彭志红, 孙琳, 陈杰. 基于改进差分进化算法的无人机在线低空突防航迹规划. 北京科技大学学报, 2012, 34(1): 96)

[10] Phung M D, Ha Q P. Safety-enhanced UAV path planning with spherical vector-based particle swarm optimization. *Appl Soft Comput*, 2021, 107: 107376

[11] Sonny A, Yeduri S R, Cenkeramaddi L R. Autonomous UAV path planning using modified PSO for UAV-assisted wireless networks. *IEEE Access*, 2023, 11: 70353

[12] Huang C, Zhou X B, Ran X J, et al. Adaptive cylinder vector particle swarm optimization with differential evolution for UAV path planning. *Eng Appl Artif Intell*, 2023, 121: 105942

[13] Shen Y K, Duan H B, Deng Y M, et al. Verification of a UAV swarm flight simulating the passive inertial emergency obstacle avoidance behavior of a pigeon flock. *Sci Sin Inf*, 2019, 49(10): 1343
(申燕凯, 段海滨, 邓亦敏, 等. 仿鸽群被动式惯性应急避障的无人机集群飞行验证. 中国科学:信息科学, 2019, 49(10): 1343)

[14] Ji J, Khajepour A, Melek W W, et al. Path planning and tracking

for vehicle collision avoidance based on model predictive control with multiconstraints. *IEEE Trans Veh Technol*, 2016, 66(2): 952

[15] Li Y, Lei J Q. Formation model predictive control of multi-mobile robots under disturbance. *Inf Control*, 2023, 52(2): 166
(李艳, 雷佳琦. 扰动作用下的多移动机器人编队模型预测控制. 信息与控制, 2023, 52(2): 166)

[16] Zhang S W, Wang H, Chen P, et al. Overview of the application of neural networks in the motion control of unmanned vehicles. *Chin J Eng*, 2022, 44(2): 235
(张守武, 王恒, 陈鹏, 等. 神经网络在无人驾驶车辆运动控制中的应用综述. 工程科学学报, 2022, 44(2): 235)

[17] Wang Q, Liu M W, Ren W J, et al. Overview of common algorithms for UAV path planning. *J Jilin Univ Inf Sci*, 2019, 37(1): 58
(王琼, 刘美万, 任伟建, 等. 无人机航迹规划常用算法综述. 吉林大学学报(信息科学版), 2019, 37(1): 58)

[18] Ait Saadi A, Soukane A, Meraihi Y, et al. UAV path planning using optimization approaches: A survey. *Arch Comput Methods Eng*, 2022, 29(6): 4233

[19] Wu J, Shin S, Kim C G, et al. Effective lazy training method for deep Q-network in obstacle avoidance and path planning // 2017 *IEEE International Conference on Systems, Man, and Cybernetics*. Banff, 2017: 1799

[20] Li H R, Zhang Q C, Zhao D B. Deep reinforcement learning-based automatic exploration for navigation in unknown environment. *IEEE Trans Neural Netw Learn Syst*, 2019, 31(6): 2064

[21] Lu Y, Chen Y R, Zhao D B, et al. MGRL: Graph neural network based inference in a Markov network with reinforcement learning for visual navigation. *Neurocomputing*, 2021, 421: 140

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

[23] Zhao Z Y, Wu N, Wang X X. Autonomous optimization algorithm for UAV path planing based on breadth-first search // 6*th China Conference on Command and Control*. Beijing, 2018: 492
(赵真一, 吴娜, 王晓璇. 基于广度优先搜索的无人飞行器航路自主寻优算法//第六届中国指挥控制大会. 北京, 2018: 492)

[24] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor // *Proceedings of the* 35*th International Conference on Machine Learning*. Stockholm, 2018: 1861

[25] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications [J/OL]. *arXiv preprint* (2019-01-29) [2023-10-03]. https://arxiv.org/abs/1812.05905

[26] Christodoulou P. Soft actor-critic for discrete action settings [J/OL]. *arXiv preprint* (2019-10-18) [2023-10-15]. https://arxiv.org/abs/1910.07207

[27] Wang Z H, Zhang Y X, Yin C K, et al. Multi-agent deep reinforcement learning based on maximum entropy // *IEEE* 4*th Advanced Information Management, Communicates, Electronic and Automation Control Conference*. Chongqing, 2021: 1402

[28] Abdallah S, Kaisers M. Addressing environment non-stationarity by repeating Q-learning updates. *J Mach Learn Res*, 2016, 17(1): 1582

[29] Wu Y H. *Research on Reactive Obstacle Avoidance Using Deep Reinforcement Learning and Transfer Learning* [Dissertation]. Wuhan: Huazhong University of Science and Technology, 2019
(吴宇豪. 基于深度强化学习和迁移学习的反应式避障方法研究[学位论文]. 武汉: 华中科技大学, 2019)

[30] Juliani A, Berges V P, Teng E, et al. Unity: A general platform for intelligent agents [J/OL]. *arXiv preprint* (2020-05-06) [2023-10-15]. https://arxiv.org/abs/1809.02627