计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):61−73 [doi: 10.15888/j.cnki.csa.009763] [CSTR: 32024.14.csa.009763]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 SPER-TD3 的无人机编队三维航迹规划①

彭　博[1], 王晓波[2], 魏祥麟[2], 成　洁[2], 秦华旺[1], 范建华[2]

[1](南京信息工程大学 电子与信息工程学院, 南京 210044)
[2](国防科技大学 第六十三研究所, 南京 210007)
通信作者: 范建华, E-mail: fjh7659@163.com

摘　要: 复杂地形条件下, 基于深度强化学习的无人机编队航迹规划可以完成无人机编队的轨迹寻优, 路径长度和环境适应性均优于传统启发式算法, 但仍存在训练稳定性不足、规划实时性差等问题. 面向领航者-跟随者模式的无人机集群, 本文提出了一种基于 SPER-TD3 算法的无人机编队实时三维航迹规划方法. 首先, 将基于 SumTree 的优先经验回放机制融入 TD3 算法, 设计了 SPER-TD3 算法, 确定无人机编队的轨迹; 然后, 使用基于角度队形控制方法优化跟随者的飞行轨迹, 并应用动态轨迹平滑算法优化转向角. 为了加快 SPER-TD3 算法的训练收敛速度和稳定性, 解决长时间依赖性问题, 设计了结合 LSTM、自注意力机制以及多重感知机的网络模型结构. 在多种障碍物环境下进行了仿真实验, 结果表明, 所提方法在轨迹安全覆盖率、飞行路径平滑度、成功率、奖励大小等方面综合表现优于 8 种主流的深度强化学习算法, 其重要性综合评估值比当前方法提升 8.5%–72.9% 不等, 且训练稳定性最佳.

关键词: 无人机编队路径规划; 深度强化学习; 三维环境; 实时性; 轨迹平滑

引用格式: 彭博,王晓波,魏祥麟,成洁,秦华旺,范建华.基于 SPER-TD3 的无人机编队三维航迹规划.计算机系统应用,2025,34(2):61–73. http://www.c-s-a.org.cn/1003-3254/9763.html

3D Trajectory Planning for Unmanned Aerial Vehicle Formation Based on SPER-TD3

PENG Bo[1], WANG Xiao-Bo[2], WEI Xiang-Lin[2], CHENG Jie[2], QIN Hua-Wang[1], FAN Jian-Hua[2]

[1](School of Electronic and Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China)

Abstract: In complex terrain conditions, UAV formation path planning based on deep reinforcement learning can optimize the path of UAV formation, with better path length and environmental adaptability than traditional heuristic algorithms. However, it still has problems such as insufficient training stability and poor real-time planning. For UAV clusters with a leader-follower mode, this study proposes a real-time 3D path planning method for UAV formation based on the SPER-TD3 algorithm. Firstly, the prioritized experience replay mechanism based on SumTree is integrated into the TD3 algorithm, and the SPER-TD3 algorithm is designed to determine the path of the UAV formation. Then, an angle formation control method is used to optimize the path of the followers, and a dynamic path smoothing algorithm is applied to optimize the steering angle. To accelerate the training convergence speed and stability of the SPER-TD3 algorithm, and solve the long-term dependence problem, a network model structure combining LSTM, self-attention mechanism, and multiple perceptrons is designed. Simulation experiments are conducted in environments with various obstacles. Results show that the method mentioned above is superior to eight mainstream deep reinforcement learning algorithms in terms of path safety coverage rate, flight path smoothness, success rate, and reward size. Its comprehensive evaluation value of importance is 8.5% to 72.9% higher than existing methods, and it has the best training stability.

① 收稿时间: 2024-07-10; 修改时间: 2024-08-01; 采用时间: 2024-08-27; csa 在线出版时间: 2024-12-19
CNKI 网络首发时间: 2024-12-20

Key words: multi UAV path planning; deep reinforcement learning; 3D environment; real time; trajectory smoothing

多无人机空中编队协同完成大范围、高动态、强鲁棒的复杂任务逐渐成为无人系统应用的新范式. 为了完成给定任务, 无人机编队通常需要在给定障碍区域内, 在多个指定点之间移动. 因此, 如何为无人机编队规划碰撞风险小、训练稳定性高、轨迹平滑的飞行轨迹是业界关注的焦点问题之一.

基于领航者-跟随者模式的无人机编队中, 领航者负责整个编队的路径规划, 跟随者跟随领航者的路径和动作并保持相对距离, 显著降低了编队形成和管理的复杂度. 在此场景下, 业界设计了多种基于人工势场法、群体智能和深度强化学习的编队轨迹规划方法[1−3]. 人工势场法[4,5]通过模拟物理势场的概念确定飞行路径, 基于目标位置的引力和障碍物的斥力计算合力, 并基于作用力来规划无人机运动路径, 具有简单直观、路径平滑、实时性强等特性, 但也易出现局部最小值、狭窄通道中摆动或在障碍物附近徘徊等问题[6]. 群体智能类算法通过模拟生物群落中的集群行为寻找最优或近优路径, 例如蚁群、粒子群算法等[7,8]. 这类算法能够适应非线性和动态变化环境, 但其要求环境已知且参数配置和调整困难. 深度强化学习方法将轨迹规划问题建模为马尔可夫决策过程, 将无人机的移动或转向看作动作选择, 将减少移动距离、避免与障碍物碰撞等作为动作奖励, 通过探索+利用的模型训练模式, 训练得到无人机集群的飞行轨迹. 但是, 基于现有深度强化学习方法进行编队轨迹规划时, 面临训练时间长、训练过程不稳定甚至无法收敛的问题[9].

针对这些问题, 本文提出了一种基于 SPER-TD3 算法的无人机编队实时三维航迹规划方法. 首先, 将基于 SumTree 的优先回放机制融入 TD3 算法, 设计了 SPER-TD3 算法, 确定无人机编队的轨迹; 然后, 使用基于角度队形控制方法优化跟随者的飞行轨迹, 并应用动态轨迹平滑算法优化其转向角. 为了加快 SPER-TD3 算法的训练收敛速度和稳定性, 解决长时间依赖性问题, 设计了结合 LSTM、自注意力机制以及多重感知机的网络模型结构. 本文贡献包括 3 个方面.

(1) 将 SumTree 优先回放机制与 TD3 算法相融合, 提出了 SPER-TD3 算法, 整体框架使用完全去中心化模式, 增加 LSTM 和注意力模块改进了 Actor-Critic 网络模型, 增强了模型训练稳定性, 提升了收敛速度.

(2) 设计了一种基于角度的虚拟领航者法对编队进行队形控制, 灵活调整编队阵型, 极大提高了编队整体的鲁棒性, 同时提出了一种动态轨迹平滑算法, 减小跟随者急剧转弯时的转向角度, 大幅优化其轨迹平滑度.

(3) 为了验证所提算法的有效性, 建立了两个无人机路径规划三维障碍物环境, 对编队转向角进行轨迹平滑度对比, 并设立重要性综合评估值对算法性能进行评估. 结果显示, 所提方法的轨迹平滑度优于传统算法, 且在各方面综合表现优于 8 种主流的深度强化学习算法.

本文第 1 节介绍现有深度强化学习无人机路径规划问题的相关进展. 第 2 节介绍无人机路径规划问题建模与优化, 以及编队控制算法与跟随者路径平滑算法. 第 3 节详细介绍所提 SPER-TD3 算法的具体框架和模型结构. 第 4 节介绍实验参数并分析实验结果. 第 5 节总结全文并对下一步进行展望.

1 相关工作

基于深度强化学习的路径规划方法将任务区域以及无人机在该区域的位置等信息作为状态, 将无人机的轨迹选择作为动作, 将规划目标作为奖励函数, 然后基于探索-利用迭代寻找最优路径.

表 1 总结对比了当前业界提出的轨迹规划方法. 文献[10]提出一种具有事后经验重放 (HER) 的软行动者-批评家 (SAC) 算法, 使用 SACHER 来提高 SAC 的学习性能, 并应用于无人机的路径规划和防撞控制. 文献[11]提出一种混合动作空间的动作解耦 SAC (AD-SAC) 算法, 使离散动作和连续动作彼此独立, 可以同时执行. 文献[12]提出一种基于 DDQN 的改进路径规划算法, 提出随机障碍训练方法, 提高算法的鲁棒性和适应性, 并采用线性软更新策略实现神经网络参数的平滑更新, 提高了训练的稳定性和收敛性. 文献[13]利用 DDQN 训练无人机的覆盖策略, 得到基于 DDQN 的无人机未知区域覆盖路径规划框架. 文献[14]提出一种基于扰动流体与 TD3 算法的无人机路径规划框架, TD3 算法中依正态分布进行动作选取, 提高动作选取随机性.

相对于单个无人机而言, 基于多智能体强化学习

(MADRL) 的无人机编队系统具备冗余性和鲁棒性, 同时能够覆盖更大的区域, 提高效率, 还能够协同收集和共享数据. 这些特点使得无人机编队在复杂任务和挑战性环境下表现出明显优势. 文献[15]利用改进的 HEED 动态聚类算法对网络模型进行优化, 提出一种基于智能算法和深度强化学习 IA-DRL 的智能路径优化算法. 文献[16]提出经验共享互惠奖励 (multi agent Actor-Critic (MAAC-R)) 算法来学习所有同质无人机的合作共享策略, 该算法能更有效地提高无人机的协同性能, 并能激发出丰富形式的无人机群协同跟踪行为. 文献[17]介绍了一种参数共享的非策略多智能体路径规划方法, 通过智能体之间的经验共享, 加速了策略融合. 文献[18]提出了一种新的任务分解 MATD3 (TD-MATD3) 算法, 将路径规划任务分解为飞向目标的导航任务模块和避开障碍物和其他无人机的避障任务模块, 使无人机能够在复杂的多障碍环境中执行路径规划. 文献[19]提出一种基于深度强化学习 (DRL) 的特征融合近端策略优化 (FF-PPO) 算法, 将图像识别信息与原始图像融合, 实现多无人机路径规划算法. 文献[20]为提高编队学习效率, 设计了相应的解耦型奖励函数和神经网络结构, 并采用解耦型多智能体深度确定性策略梯度 (MADDPG) 算法对模型进行自适应训练, 以生成无人机编队自主跟踪与避障最优机动策略. 文献[9]提出基于网络剪枝的多智能体近端策略优化 (network pruning-based multi-agent proximal policy optimization, NP-MAPPO) 算法, 提高了训练效率.

表 1 基于深度强化学习的路径规划方法对比

类型	DRL算法	文献	路径长度	成功率	奖励大小	是否平滑	收敛速度	训练稳定性
单无人机	HER-SAC	[10]	√	√	√	√	—	—
	AD-SAC	[11]	—	—	√	√	√	√
	DDQN	[12]	√	—	√	—	√	√
	DDPG	[13]	—	√	—	√	√	√
	TD3	[14]	√	—	√	√	√	—
多无人机	DQN	[15]	√	—	√	—	√	√
	MAAC	[16]	√	√	√	√	—	—
	SAC	[17]	—	√	√	√	√	—
	MATD3	[18]	—	√	√	—	√	√
	PPO	[19]	√	√	√	—	√	—
	MADDPG	[20]	√	—	√	—	√	√
	MAPPO	[9]	√	—	√	—	√	√
	SPER-TD3	本文	√	√	√	√	√	√

可以看出, 当前基于 DRL 进行无人机轨迹规划已经取得了显著进展, 但有两个问题仍关注较少: 一是 DRL 算法研究多数集中于算法效率提升, 对编队中无人机的碰撞安全性考虑不足; 二是面对复杂地形环境, 无人机航迹规划稳定性和轨迹平滑性较差.

2 问题建模

2.1 环境模型

本文考虑无人机编队在地形未知的三维任务区域中, 在保持编队队形和避免碰撞的条件下, 以尽可能短的路径解决从出发点到达目标点的轨迹规划问题. 轨迹规划场景的一个典型示例如图 1 所示.

从图 1 可以看出, 任务区域中存在多种类型的障碍物. 简单起见, 建立任务区域的三维坐标系, 无人机的位置建模为坐标点, 障碍物简化成规则的立体几何形状, 如球形、圆柱形和圆锥形. 当无人机位置为(x,y,z)时, 其与球形、圆柱形和圆锥形障碍物的位置关系表示为:

$$\Phi(Q)=\begin{cases}\left(\dfrac{x-x_{\rm sp}}{r}\right)^2+\left(\dfrac{y-y_{\rm sp}}{r}\right)^2+\left(\dfrac{z-z_{\rm sp}}{r}\right)^2\\ \left(\dfrac{x-x_{\rm cy}}{r}\right)^2+\left(\dfrac{y-y_{\rm cy}}{r}\right)^2\\ \dfrac{1}{1-\dfrac{z}{H}}\left(\left(\dfrac{x-x_{\rm co}}{r}\right)^2+\left(\dfrac{y-y_{\rm co}}{r}\right)^2\right)\end{cases} \tag{1}$$

其中, 从上至下分别为球形、圆柱形和圆锥形障碍物的障碍判断方程. 球形障碍物由其中心点的坐标$(x_{\rm sp},y_{\rm sp},z_{\rm sp})$和半径$r$来定义. 圆柱形和圆锥形障碍物由底面中心点坐标$(x_{\rm cy},y_{\rm cy},z_{\rm cy})$、$(x_{\rm co},y_{\rm co},z_{\rm co})$以及半径$r$和高度$H$来定义, 且$0\leqslant z\leqslant H$. $\Phi(Q)<1$、$\Phi(Q)=1$以及$\Phi(Q)>1$分别表示无人机位于障碍物的内部, 表面以及外部, 从而可以根据$\Phi(Q)$的取值情况判断无人机与障碍物是否发生碰撞.

2.2 运动模型

无人机的运动可以表示为 3 个部分, 即三维坐标、航迹角和爬升角.

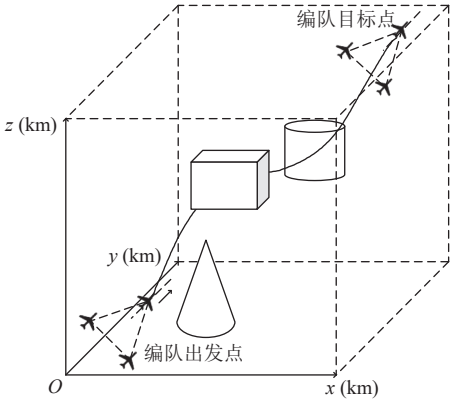


图 1　所考虑场景的典型示例

$$\begin{cases} q_t = q_{t-1} + \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} \\ \psi = \arcsin\left(\dfrac{|\Delta y|}{\sqrt{\Delta x^2 + \Delta y^2}}\right) \\ \gamma = \arcsin\left(\dfrac{\Delta z}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}\right) \end{cases} \tag{2}$$

其中, q_t 表示无人机在 t 时刻得到的三维坐标, Δx、Δy 和 Δz 是 $t-1$ 时刻与 t 时刻的坐标点间在 x、y 和 z 轴方向的变化量. ψ 表示无人机在 t 时刻的航迹角, γ 表示无人机在 t 时刻的爬升角. 航迹角是飞行路径在 $x-y$ 平面的投影与 x 轴正方向之间的角度, 爬升角是飞行路径与其在 $x-y$ 平面的投影之间的角度.

由式 (2) 可以计算出在给定的速度约束和角度约束下从 t 时刻到 $t+1$ 时刻无人机位置的变化量, 即:

$$\begin{cases} \Delta x_{\text{res}} = R\cos(\psi_{\text{res}})\cos(\gamma_{\text{res}}) \\ \Delta y_{\text{res}} = R\cos(\psi_{\text{res}})\sin(\gamma_{\text{res}}) \\ \Delta z_{\text{res}} = R\sin(\gamma_{\text{res}}) \\ R = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \end{cases} \tag{3}$$

其中, ψ_{res} 表示无人机经过运动学约束过的航迹角, γ_{res} 表示无人机经过运动学约束过的爬升角. Δx_{res}, Δy_{res}, Δz_{res} 是在给定的速度约束和角度约束下计算出的 x 轴、y 轴和 z 轴坐标位置的变化量. R 表示 q 从 t 时刻到 $t+1$ 时刻位置之间的欧几里得距离, 即最短直线距离. 计算得到经过运动学约束后无人机在 $t+1$ 时的位置:

$$q_{t+1} = q_t + \begin{bmatrix} \Delta x_{\text{res}} \\ \Delta y_{\text{res}} \\ \Delta z_{\text{res}} \end{bmatrix} \tag{4}$$

其中, q_t 表示无人机在 t 时刻的位置, q_{t+1} 表示无人机经过运动学约束后在 $t+1$ 时刻的位置.

2.3　编队模型

本文设计了一种基于角度的虚拟领航者法控制无人机编队, 如图 2 所示. 具体来说即由领航者的运动方向产生的方向向量按照指定的角度和轴进行旋转, 计算出虚拟领航者的位置. 跟随者将各自的虚拟领航者作为目标点持续跟随, 同时保持和领航者的相对位置. 此方法能灵活调整编队阵型, 易于控制, 且跟随者只需和各自的虚拟领航者进行通信, 更具鲁棒性, 最大程度降低了跟随者出现故障干扰整个编队的问题.

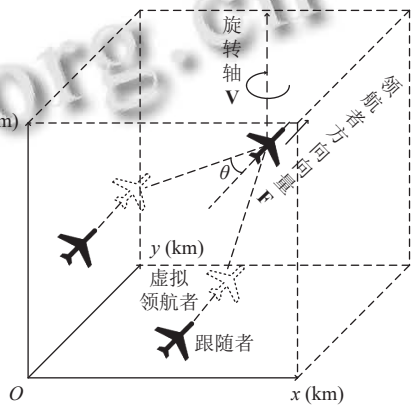


图 2　无人机编队控制算法示意图

若要领航者与跟随者形成编队队形, 则需确定领航者与虚拟领航者的位置关系. 计算领航者前进方向的方向向量 $\mathbf{F}$.

$$\mathbf{F} = \frac{q_{t+1} - q_t}{\|q_{t+1} - q_t\|} \tag{5}$$

其中, q_{t+1} 和 q_t 分别为领航者在 $t+1$ 和 t 时刻的位置.

计算垂直于领航者前进方向的轴来确定旋转轴, 旋转指定的角度来确定虚拟领航者与领航者之间的相对位置:

$$\mathbf{V} = \mathbf{F} \times \mathbf{K} \tag{6}$$

$$\mathbf{V}' = \frac{\mathbf{V}}{\|\mathbf{V}\|} \tag{7}$$

其中, $\mathbf{V}$ 为经过计算得到的旋转轴, 将 $\mathbf{V}$ 经过单位化得到 $\mathbf{V}'$. $\mathbf{K}$ 为非平行于方向向量 $\mathbf{F}$ 的单位向量, $\times$ 表示向量叉乘.

使用罗德里格斯旋转公式计算旋转后的向量:

$$\begin{aligned} \mathbf{V}_{\text{rot}} =& \mathbf{V}\cos\theta + (\mathbf{V}' \times \mathbf{V})\sin\theta \\ &+ \mathbf{V}'(\mathbf{V}' \cdot \mathbf{V})(\mathbf{1} - \cos\theta) \end{aligned} \tag{8}$$

其中, $\mathbf{V}$ 是原始向量, $\mathbf{V}'$ 是单位旋转轴向量 (长度为 1). θ 是旋转角度 (弧度制). $\mathbf{V}_{\text{rot}}$ 是旋转后的方向向量. $\cdot$ 表示向量点乘.

最后计算出虚拟领航者的最终位置:

$$Q = q_t + D \cdot \mathbf{V}_{\text{rot}} \tag{9}$$

其中, D为虚拟领航者与领航者之间的指定距离, Q为虚拟领航者的最终位置. 沿着旋转后的方向向量从领航者当前位置出发, 移动指定距离, 得到虚拟领航者的位置.

由于领航者使用基于角度的虚拟领航者法控制无人机编队, 因此跟随者跟随虚拟领航者转向时难免会产生跟随者剧烈转向的情况, 增加了能量损耗, 严重时可能导致无法及时转向从而造成无人机与障碍物或其他跟随者碰撞的情况. 因此, 需要对跟随者轨迹进行优化以减少转向角度, 从而平滑轨迹使跟随者正常转向.

为了解决上述问题, 本文提出了一种基于曲率的轨迹动态平滑算法来平滑轨迹, 在跟随者进行位置更新时加入上述轨迹平滑算法.

计算曲率 (转向角度):

$$\theta_{\text{curvature}} = \arccos\left(\frac{\mathbf{F}_{\text{ave}} \cdot \mathbf{F}'}{\|\mathbf{F}_{\text{ave}}\| \ \|\mathbf{F}'\|}\right) \tag{10}$$

其中, $\mathbf{F}_{\text{ave}}$为过去方向向量, $\mathbf{F}'$为经过单位化后的总合力.

基于曲率动态调整平滑因子:

$$\alpha = \begin{cases} 0.5, & \text{if } \theta > \dfrac{\pi}{4} \\ 1, & \text{otherwise} \end{cases} \tag{11}$$

其中, α表示平滑因子, 当$\theta > \dfrac{\pi}{4}$时, 调整α为 0.5, 其他时候为 1.

计算平滑后的方向并更新跟随者位置:

$$\mathbf{F}_{\text{smooth}} = \alpha\mathbf{F}' + (1-\alpha)\mathbf{F}_{\text{ave}} \tag{12}$$

$$q_{t+1} = q_t + v \cdot \mathbf{F}_{\text{smooth}} \tag{13}$$

其中, $\mathbf{F}_{\text{smooth}}$为平滑后的方向向量, v为每一时刻移动的步长.

当无人机超过最大爬升角和最大俯冲角时, 需要经过动力学约束来限制无人机的飞行范围. 即通过式 (4) 来计算跟随者最终的位置.

设q_{leader}表示领航者的位置, 每个跟随者的目标位置$q_{i,\text{follower}}$根据领航者的位置、虚拟领航者相对于领航者的位置偏移d_i、避障偏移向量a_i计算得出:

$$q_{i,\text{follower}} = q_{\text{leader}} + d_i + a_i \tag{14}$$

其中, d_i由虚拟领航者与领航者之间的相对位置D及虚拟领航者与跟随者之间的相对位置相加而得出. a_i由$\mathbf{F}_{\text{smooth}}$与$v$相乘得出.

2.4 问题描述

无人机编队航迹规划的目标是找到一条从起点到目标点的最短路径, 且在飞行过程中维持队形并避免与障碍物碰撞.

为了规避风险, 优化问题需要受到一系列约束条件的限制:

$$\begin{cases} -\psi_{\max} \leqslant \psi \leqslant \psi_{\max}, & \forall i \\ -\gamma_{\max} \leqslant \gamma \leqslant \gamma_{\max}, & \forall i \\ q_i \in A, & \forall i \\ \|q_i - o\| \geqslant d_{\min}, & \forall i, \forall o \end{cases} \tag{15}$$

路径长度通过计算路径上连续两点间距离的累计和来表示:

$$L = \sum_{I=1}^{N-1} \|q_{i+1} - q_i\| \tag{16}$$

式 (16) 在每个时间步重新计算一次路径长度, 保证路径的实时性和有效性.

碰撞概率通过与障碍物的最小距离的函数来估计:

$$P_{\text{collision}}(q) = \exp(-\lambda \min_{o \in obstacles} \|q - o\|) \tag{17}$$

其中, λ是衰减系数, 用以调整距离对碰撞概率的影响. λ越大碰撞概率随距离的变化更加敏感, 即无人机在接近障碍物时, 碰撞概率迅速增加.

综合上述目标, 形成无人机编队航迹规划加权优化问题:

$$\omega_1 \min_{q_1, q_2, \cdots, q_n} L + \omega_2 \sum_{I=1}^{N} P_{\text{collision}}(q) + \omega_3 \sum_{I=1}^{N} \left\| q_{i,\text{follower}} - q_i \right\|^2 \tag{18}$$

其中, ω_1, ω_2, ω_3分别是路径长度、碰撞概率和编队保持的权重系数.

3 SPER-TD3 算法

基于人工势场法的轨迹规划方法当前广泛使用, 但其存在局部最小值、目标位置不可达等问题. 式 (18) 所示的优化问题属于多目标优化问题, 难以用多项式优化算法进行解决. 而深度强化学习的目标是通过最大化累积奖励来找到最优策略. 因此, 将多目标优化问题转化为累积奖励的形式能有效解决式 (18) 所示的优化问题. 基于深度强化学习的方法通过训练无人机学习最优路径规划和决策策略, 提高无人机在复杂环境中的适应能力, 求得的轨迹质量较高, 但其搜索时间很长, 过程不稳定, 甚至无法收敛. 为了降低搜索时间并

保证求解质量, 本文提出 SPER-TD3 算法, 首先使用具有实时性强、避障效果好等特点的人工势场法快速找到初始轨迹, 并避免无人机编队与障碍物相撞; 然后利用深度强化学习方法搜索更优的轨迹. 此外, 对搜索确定的轨迹进行了平滑处理.

3.1 马尔可夫决策过程建模

在和环境交互的过程中, 无人机下一时刻的位置和状态主要由当前的位置、状态和控制输入决定, 而马尔可夫决策过程 (Markov decision process, MDP) 未来的状态 (无人机的位置和状态) 仅依赖于当前状态和采取的行动这一性质与无人机环境交互过程相符合. 因此, 无人机轨迹规划问题可以建模为马尔可夫决策过程, 可表示为:

$$M = (S, A, P, R) \tag{19}$$

其中, 状态空间S表示无人机获取所有信息的集合, 动作空间A表示在给定状态下执行所有动作的集合, 转移概率P表示系统转移到各个可能新状态的概率, 奖励函数R表示从一个状态转移到另一个状态的过程中获得的奖励.

3.1.1 状态空间

状态空间S代表了无人机所处环境的信息, 包括无人机相对于每个障碍物的位置信息以及无人机相对于目标点的位置信息. 状态空间使无人机能够评估当前环境, 并据此做出是否需要调整飞行路径以避免障碍物的判断.

状态$s \in S$对于每个障碍物i的状态s_i可以表示为:

$$s_i = [\Delta x_i, \Delta y_i, \Delta z_i, \Delta x_{\text{goal}}, \Delta y_{\text{goal}}, \Delta z_{\text{goal}}] \tag{20}$$

其中, $\Delta x_i, \Delta y_i, \Delta z_i$表示无人机相对于障碍物$i$的位置差, 而$\Delta x_{\text{goal}}, \Delta y_{\text{goal}}, \Delta z_{\text{goal}}$表示无人机相对于目标位置的位置差.

3.1.2 动作空间

动作空间是通过每个障碍物类型 (球形、圆柱形、圆锥形) 对应的 TD3 网络来决定的. 对于每个障碍物, 动作是由对应的 TD3 网络生成的转向角度, 其范围在给定的动作边界内. 这表示无人机对每个障碍物的反应程度, 进而影响无人机的移动方向和速度. 动作a_i可以表示为:

$$a_i = \eta_i \tag{21}$$

其中, η_i是由 TD3 网络针对第i个障碍物生成的转向角度.

3.1.3 奖励函数

奖励函数是强化学习中的关键组成部分, 它为每个决策反馈所采取行动的效果. 神经网络基于当前状态做出决策, 环境根据这些决策进行更新, 并反馈奖励值. 这个奖励值用于调整神经网络的参数, 以期在未来能做出更好的决策. 因此, 奖励函数的设计直接决定了学习过程的效果和算法的收敛速度.

碰撞检测奖励R_{col}设计如下:

$$R_{\text{col}} = \begin{cases} \dfrac{d(q_t, obs) - R_{obs}}{R_{obs}} - 1, & \text{if } d(q_t, obs) \leqslant R_{obs} \\ k \cdot [d(q_t, obs) - d_{\text{safe}}], & \text{if } d(q_t, obs) > d_{\text{safe}} \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

其中, $d(q_t, obs)$表示无人机与障碍物中心之间的距离, R_{obs}是障碍物的半径或相关尺寸参数. 当无人机与障碍物发生碰撞时, 给予一个负值奖励. 当无人机与障碍物的距离大于设定的安全距离d_{safe}时, 给予一个安全奖励, 目的是让无人机与障碍物保持一个安全距离.

距离目标奖励R_{len}, 定义为:

$$R_{\text{len}} = \begin{cases} -\dfrac{d(q_t, q_{\text{goal}})}{d(q_{\text{start}}, q_{\text{goal}})} + 3, & \text{if } d(q_t, q_{\text{goal}}) < threshold \\ -\dfrac{d(q_t, q_{\text{goal}})}{d(q_{\text{start}}, q_{\text{goal}})}, & \text{otherwise} \end{cases} \tag{23}$$

其中, $d(q_{\text{start}}, q_{\text{goal}})$表示无人机与目标点之间的距离, $d(q_{\text{start}}, q_{\text{goal}})$是起点到目标点的距离. 当无人机与目标点之间的距离小于 *threshold*, 则视为到达目标点, 给予一个正值奖励. 其他情况视为未到达目标点, 给予负值奖励.

角度平滑奖励R_{ang}, 定义为:

$$R_{\text{ang}} = -\left(\frac{|\Delta\psi|}{\psi_{\max}} + \frac{|\Delta\gamma|}{\gamma_{\max}}\right) \cdot 0.1 \tag{24}$$

其中, $|\Delta\psi|$和$|\Delta\gamma|$分别是航迹角变化和爬升角变化的绝对值, $\psi_{\max}$和$\gamma_{\max}$是这些角度变化的最大允许值. 航迹角、爬升角变化过大时, 给予负值奖励, 目的是减少角度变化, 平滑无人机轨迹.

总的奖励函数R_{total}定义为:

$$R_{\text{total}} = R_{\text{col}} + R_{\text{en}} + R_{\text{ang}} \tag{25}$$

3.2 基于深度强化学习的轨迹规划

3.2.1 基于 SumTree 的优先经验回放机制

在强化学习中, 使用 TD 误差来表示样本的重要程度, TD 误差的值越大则说明该样本的重要性高. 在传统的随机采样中, 所有经验被等概率地重复采样, 这可能导致对于一些不太重要 (低 TD 误差) 的经验过度学习. 优先经验回放 (prioritized experience replay, PER) 通过调整采样概率减少了这种冗余, 确保算法能够集中注意力在最需要学习的经验上. 因此本文将这种优先经验回放思想与 TD3 算法相结合, 让无人机能够区分经验样本的重要性, 优先学习那些具有高 TD 误差的经验, 从而加快学习进度, 提高样本利用效率, 帮助无人机更快地学习如何避开障碍物、优化路径等.

为了实现基于优先级的快速采样, 提升学习效率, 引入了 SumTree 优先级采样结构. SumTree 是一种特殊的二叉树, 它将经验的优先级作为树的叶子节点, 非叶子节点存储其所有子节点值的和, 如图 3 所示. 根节点代表整棵树的总优先级值, 等于其所有子节点优先级值之和; 每个内部节点的值是其所有子节点的优先级值之和; 叶子节点代表存储的经验及其对应的优先级值.

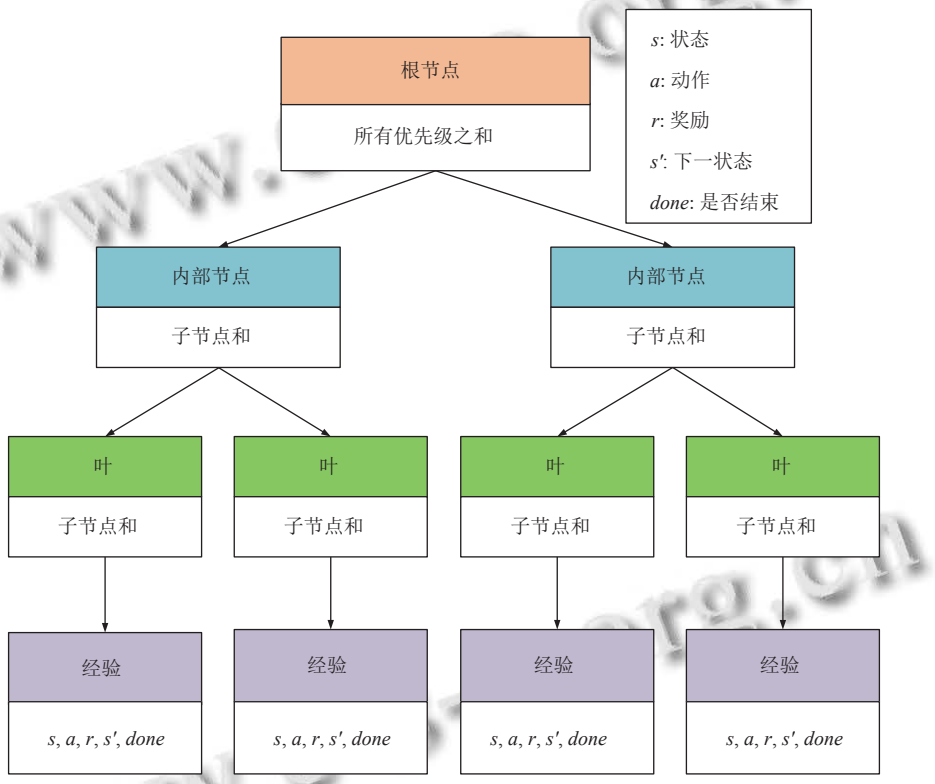


图 3 基于 SumTree 的优先级计算结构

TD3 算法整体框架如图 4 所示. 首先从经验缓冲区中采样一批经验, 并将s'输入到目标策略网络. 然后, 在下一次获得a', 并将状态-动作对(s',a')输入到目标 Q 网络. 在得到两个目标Q值$(Q_1(s'\cdot a'))$和$(Q_2(s'\cdot a'))$后, 选择较小的一个来计算值函数目标y. 另一方面, 将(s,a)输入到 Q 网络中, 并获得两个Q值 $(Q_1(s,a))$和$(Q_2(s,a))$. 然后, 使用它们来计算y的 MSE, 并反向传播 MSE 的总和来更新两个 Q 网络的参数. 接下来, 将从第 1 个 Q 网络获得的Q值输入到策略网络中, 并在Q值增加的方向上更新策略网络参数 (每 2 次迭代更新 1 次). 最后, 采用“软更新”方法对所有目标网络进行更新. TD3 为了实现优先经验回放机制, 经验回放缓冲区首先收集每一步的经验元组, 这些经验随后存储到 SumTree 中, 每个经验都有一个与其相关的优先级值. 当算法需要进行训练时, 它会从 SumTree 中采样一批经验, 采样概率基于经验的优先级. 采样得到的经验用于训练更新网络参数, 即 Q 网络和策略网络 (Actor) 的参数. 在更新网络参数的过程中, TD3 算法会计算这些经验的新 TD 误差. 最后, 算法将这些计算出的 TD 误差反馈给 SumTree, 以更新相应经验的优先级, 确保 SumTree

中的优先级信息始终是最新的.

SPER-TD3 整体算法流程如算法 1 所示. 第 1–7 步初始化网络; 第 8–11 步观测得到奖励和状态; 第 12–16 步储存经验; 第 17–29 步更新网络.

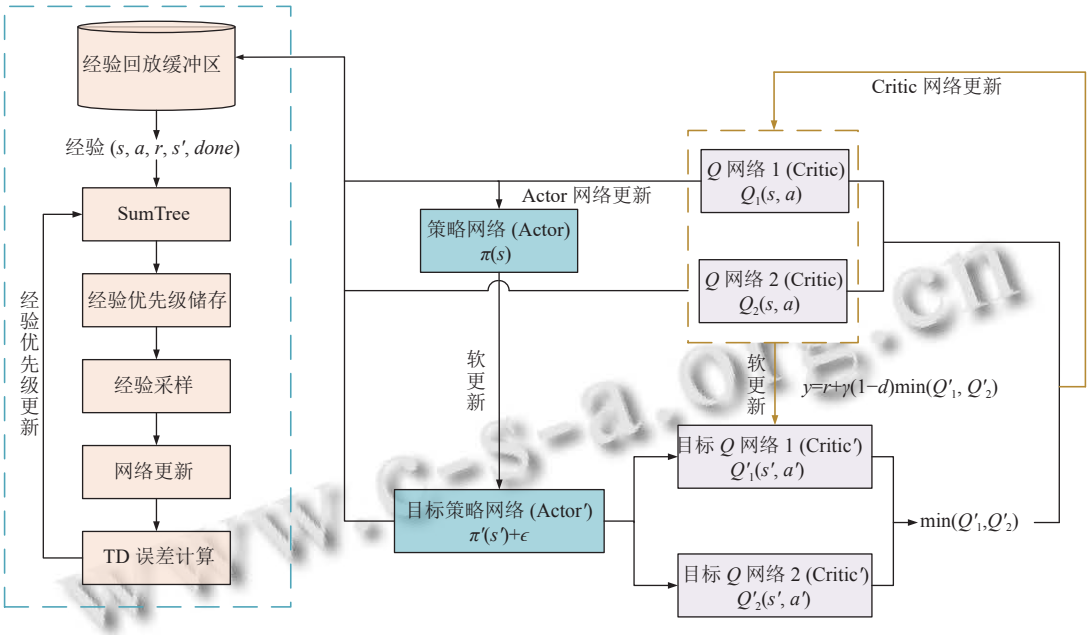


图 4 SPER-TD3 算法整体框架图

算法 1. SPER-TD3 整体算法

// 网络初始化:

1. Actor 网络ac, 输出动作$a=\pi(\theta,s)$
2. Critic 网络$q_1=\pi(s,a)$, $q_2=\pi(s,a)$, 输出价值 *value*
3. 初始化目标网络参数$\theta'\leftarrow\theta$, $\varphi'\leftarrow\varphi$
4. 初始化优先级经验回放缓冲区优先级指数α, 重要性采样指数β
5. 设置训练参数: γ (gamma), τ (polyak), 学习率I_a, I_c
6. 设置噪声参数: 动作噪声σ, 目标策略噪声σ', 噪声剪切ε.
7. 设置策略延迟更新参数 policy_delay

// 训练开始:

8. 对每个训练步骤
9. 观测状态s
10. 根据当前策略$\pi(\theta,s)$ 选择动作a, 并加入噪声ε, $\varepsilon\sim N(0,\sigma^2)$
11. 执行动作a, 观测得到奖励r和新状态s'
12. 计算 TD 误差:
13. $\delta=|r+\gamma\min(Q_1(\phi',s',\pi(\theta',s')),Q_2((\phi',s',\pi(\theta',s')))$
14. $\quad -\min(Q_1(\phi,s,a),Q_2(\phi,s,a))|$
15. 计算存储优先级$p=(|\delta|+\varepsilon)^{\alpha}$
16. 存储转换(s,a,r,s',p)到 SumTree
17. 如果 SumTree 中经验足够, 根据存储优先级采样经验
18. 计算样本权重$\omega=\left(\frac{1}{N}\cdot\frac{1}{p}\right)^{\beta}$, 更新$\beta$
19. 使用 Actor 目标网络计算s'对应的动作
20. $a'=\pi'(s')+\varepsilon'$, $\varepsilon'\sim\text{clip}(N(0,\sigma'^2),-\varepsilon,\varepsilon)$
21. 使用 Critic 目标网络计算
22. $y=r+\gamma\min(Q_1(\phi',s',a'),Q_2(\phi',s',a'))$
23. 更新主 Critic 网络, 最小化损失:
24. $Loss\_Q=\omega\cdot(MSE(Q_1(\varphi,s,a)-y)+MSE(Q_2(\varphi,s,a)-y))$
25. 每 policy_delay 步更新 Actor 网络, 最大化 $Q_1(\phi,s,a(\theta,s))$
26. 软更新目标网络参数:
27. $\theta'\leftarrow\tau\theta+(1-\tau)\theta'$
28. $\phi'\leftarrow\tau\phi+(1-\tau)\phi'$
29. 重复更新 Critic 网络并计算新的 TD 误差, 更新 SumTree 中的优先级, 如达到终止条件或完成目标, 则结束训练, 否则继续

3.2.2 网络模型设计

现有 TD3 算法虽然在 DDPG (deep deterministic policy gradient) 算法的基础上引入了较大改进, 但算法的稳定性不足且收敛速度较慢. LSTM 处理时间依赖性, 记忆过去的关键信息. 其通过其内部记忆单元, 能够捕捉和记忆这些长期依赖, 从而帮助无人机做出更好的决策. 此外, LSTM 可以生成更平滑的动作序列, 避免了不必要的动作抖动, 从而提高稳定性和效率. 自注意力机制在当前决策时选择性关注重要信息. 其在计算注意力权重时考虑整个输入序列, 从而捕捉到长距离的依赖关系, 而不像传统的 RNN 那样局限于邻近的时间步. 这使得模型能够更全面地理解和利用输入信息, 提高了性能. 在高维状态和动作空间中, MLP 能够处理多种输入特征, 并将其映射到低维空间, 简化策略和价值函数的学习. 因此本文引入 LSTM、自注意力机制及多层感知机等结构, 设计了新的 Actor 网络和 Critic 网络结构, 如图 5 所示.

Actor 网络将观测的状态s作为输入. 第 1 层为 LSTM 层, 神经元个数为 128, 通过学习何时“记住”或

“忘记”某些信息, 能够更好地处理长时间依赖性的问题. 第 2 层为 MLP (多层感知机) 层, 由多个全连接层组成, 每层都采用激活函数进行输出处理. MLP 第 1 层连接到 LSTM 层, 神经元个数为 128, 接收 LSTM 层输出的时间依赖性特征. 第 2 层和第 3 层为隐藏层, 神经元个数为 256, 使用 ReLU 激活函数, 用于动作决策. 最后一层为输出层, 神经元个数为 1, 使用激活函数 tanh 根据动作维度输出动作.

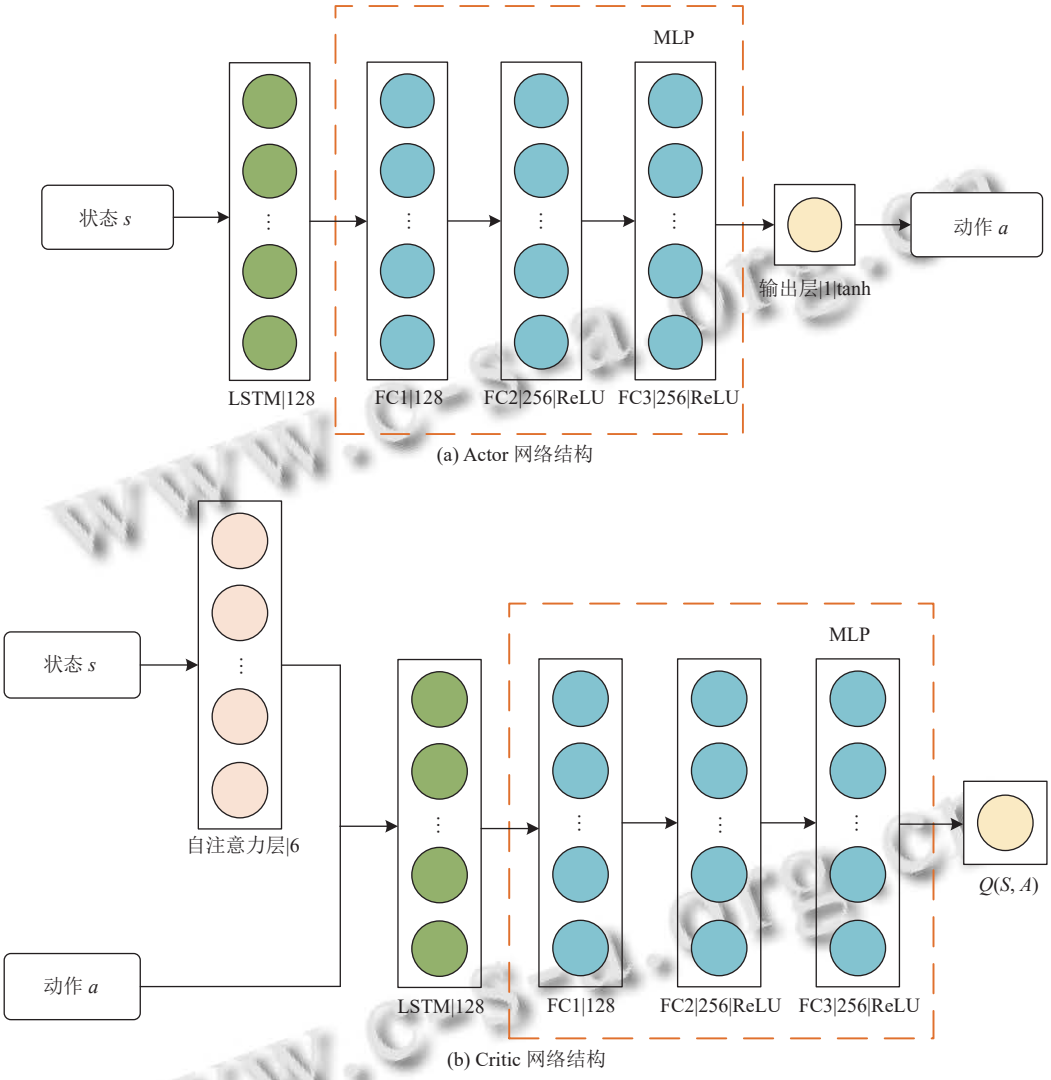


图 5 Actor 和 Critic 网络结构图

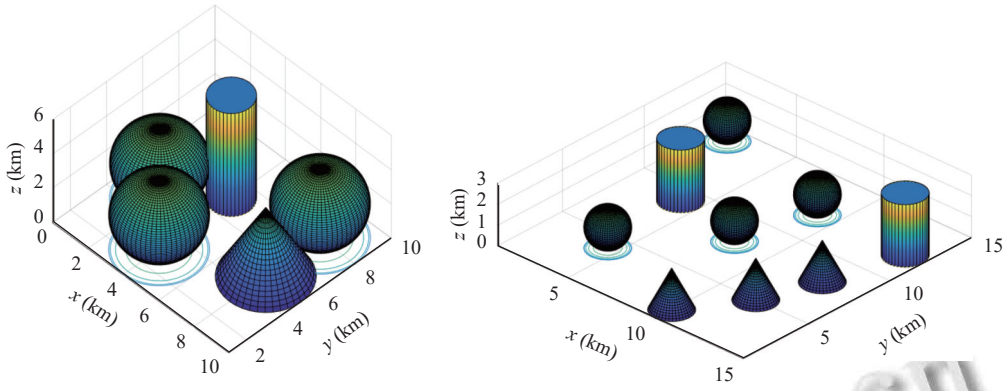
Critic 网络 (或称 Q 网络) 的输入是状态s和动作α. 状态s首先通过自注意力层, 神经元个数为 6, 增强网络对环境感知能力, 提升路径规划的效率. 然后将状态s和动作α进行拼接, 通过 LSTM 层和 MLP 层, 神经元个数和激活函数与 Actor 网络中相同. 最后输出一个Q值, 用于评估给定状态和对动作的价值.

4 实验仿真

为了对所提算法进行性能评估, 基于 Python 设计了仿真环境. 仿真实验环境硬件配置为: CPU 是 12th Gen Intel(R) Core(TM) i7-12700H, GPU 是 GeForce GTX3060, 软件环境为: 操作系统是 Windows 11, 软件配置为 CUDA 11.8、PyTorch 2.0.1、Python 3.10、Matlab 2022b. 第 4.1 节介绍了实验环境、参数配置、对比基准和对比测度; 第 4.2 节展示了编队轨迹平滑性与算法性能测试的仿真结果及分析.

4.1 实验环境和参数配置

本节将在大型障碍物、小型障碍物场景中进行路径规划实验, 对实验结果进行分析总结. 大、小障碍物环境示意图如图 6 所示.



(a) 大型障碍物 (b) 小型障碍物

图 6 大、小障碍物环境示意图

大型障碍物环境位于 10×10×6 km^3 的三维空间内, 包含 3 个半径为 2 km 的球形障碍物、1 个底面半径为 2 km, 高为 2 km 的圆锥形障碍物和 1 个底面半径为 1 km, 高为 6 km 的圆柱形障碍物. 小型障碍物环境位于 15×15×3 km^3 的三维空间内, 包含 4 个半径为 1 km 的球形障碍物、3 个底面半径为 1 km, 高为 1 km 的圆锥形障碍物和 2 个底面半径为 1 km, 高为 3 km 的圆柱形障碍物. SPER-TD3 模型训练的主要参数如表 2 所示.

表 2 实验参数设置

参数名称	意义	参数值
γ	折扣因子	0.99
l_a	Actor学习率	0.001
l_c	Critic学习率	0.001
τ	软更新系数	0.005
α	优先级系数	0.6
β	重要性采样指数	0.4
Max capacity	经验回放池容量	10^6
Batch size	采样大小	512
Adam optimizer	优化器	—
Max step	最大步数	500
Max Episode	最大轮数	500

4.2 对比基准和测度

4.2.1 对比基准

TD3 主要针对单一无人机的决策问题, 而对于多无人机, 通常有完全中心化和完全去中心化两种算法框架. 完全中心化框架使用一个中心化的控制器接收所有无人机的状态信息. 完全去中心化框架中, 为了提高学习的灵活性和适应性, 每个无人机拥有单独的决策模型, 根据自己的对环境的局部观察进行动作决策.

在两种不同的框架下, 本文实现了 TD3、DDPG、SAC、PPO 等主流 DRL 方法. 此外, 还实现了 MADDPG 算法, 该算法中, 每个无人机拥有自己的决策模型, 相互之间通过共享动作信息实现彼此间的协同.

4.2.2 对比测度

为综合考虑无人机算法的安全性与最佳性能, 本文设定路径长度、奖励大小、成功率、轨迹安全覆盖率和平均回合奖励作为评价指标, 定义分别如下.

路径长度: 无人机从出发点到目标点所移动的距离.

奖励大小: 无人机在环境中到达目标时收到反馈信号的总和.

成功率: 无人机到达目标点的次数与总测试次数的百分比.

轨迹安全覆盖率: 设定离障碍物的一定区间内的路径为安全路径, 轨迹安全覆盖率为安全路径与整条路径的百分比.

路径长度、奖励大小为算法模型测试结果, 奖励范围为 $(-\infty, 0)$, 越接近于 0 奖励越大, 为方便对比在图中取绝对值. 为了便于评估和比较不同算法的性能, 本文将多个评估指标合并成一个单一的数值, 设计了重要性综合评估值. 首先对评估指标进行归一化, 将数值缩放到[0, 1]区间.

$$\begin{cases} L' = 1 - \dfrac{|L - L_{\text{safe}}|}{L_{\max} - L_{\min}} \\ R' = \dfrac{R - R_{\min}}{R_{\max} - R_{\min}} \\ S' = S \\ C' = C \end{cases} \tag{26}$$

其中, L是路径长度, R是奖励值, S 是成功率, C是轨迹安全覆盖率. $L_{\min}$是路径长度的最小值, $L_{\max}$是路径长度的最大值, L_{safe}是安全路径长度, $R_{\min}$是奖励值的最小值, $R_{\max}$是奖励值的最大值. L'、R'、S'、C'分别为 L、R、S、C归一化后的值. 轨迹安全覆盖率和成功

率本身在[0, 1]区间内, 因此可以直接作为归一化的值.

最后计算重要性综合评估值.

$$I = \omega_L \cdot L' + \omega_R \cdot R' + \omega_S \cdot S' + \omega_C \cdot C' \tag{27}$$

其中, I为重要性综合评估值, ω_L、ω_R、ω_S、ω_C分别是路径长度、奖励值、成功率、轨迹安全覆盖率的权重. ω_L较大时更短路径的算法I值更大, 从而减少行驶时间和能源消耗; ω_R较大时高奖励路径的算法I值更大, 更适用于高奖励驱动的任务; ω_S较大时成功率更高的路径I值更大, 从而减少任务失败的风险. ω_C较大时轨迹安全覆盖率高的路径I值更大, 有助于提高任务的安全性. 由于本文算法和其他深度强化学习算法进行了路径长度、奖励值、成功率和轨迹安全覆盖率指标的对比, 各算法指标中突出的指标各不相同. 为了保证算法指标对比的公平性和均衡性, 均等权重提供一个平衡的路径选择, 路径长度、奖励值、成功率和轨迹安全覆盖率均可得到重视. 权重之和相加为 1, 因此每个权重的值为 1/4.

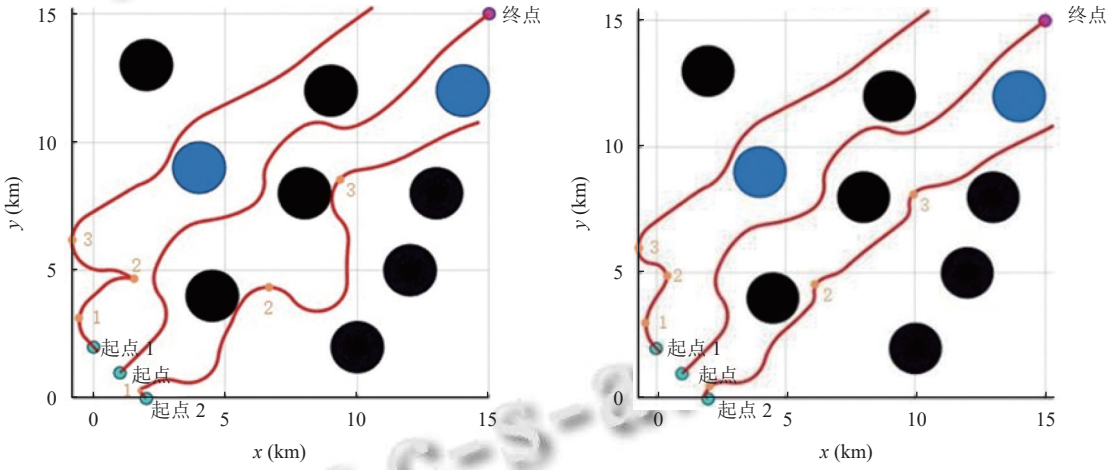
4.3 仿真结果及分析

4.3.1 编队轨迹平滑性

在三维复杂障碍物环境下, 跟随者转向角经过本文提出的动态轨迹平滑算法平滑后生成的无人机编队路径由图 7 所示.

由于篇幅限制, 仅展现对领航者之间算法指标的分析. 编队由一架领航者和两架跟随者构成, 领航者左右两侧分别为跟随者 1、2. 从图 7 中可以看出, 本文提出的基于角度的虚拟领航者法使跟随者能很好地完成跟随任务, 并与领航者保持一定距离.

为了对比动态轨迹平滑算法平滑路径前后的转向角大小区别, 在表 3 加入图 7 中算法平滑前后转向角 1、2、3 处的局部放大图. 动态轨迹平滑算法使跟随者转向剧烈的转向角均得到显著的平滑, 转向角越小平滑效果越好, 范围在 (0°, 180°) 内, 避免了由剧烈转向而引起的损伤情况出现, 达到安全执行任务和减低能量损耗的目的.



(a) 人工势场法-平滑前 (b) 人工势场法-平滑后

图 7 跟随者平滑算法对比俯视图

表 3 跟随者转向角对比 (°)

算法	跟随者1				跟随者2			
	1	2	3	平均	1	2	3	平均
人工势场法-平滑前	59.89	158.95	50.00	89.61	177.75	76.26	88.08	114.03
人工势场法-平滑后	44.63	53.65	56.72	**51.67**	58.08	52.35	53.62	**54.68**

4.3.2 算法性能测试

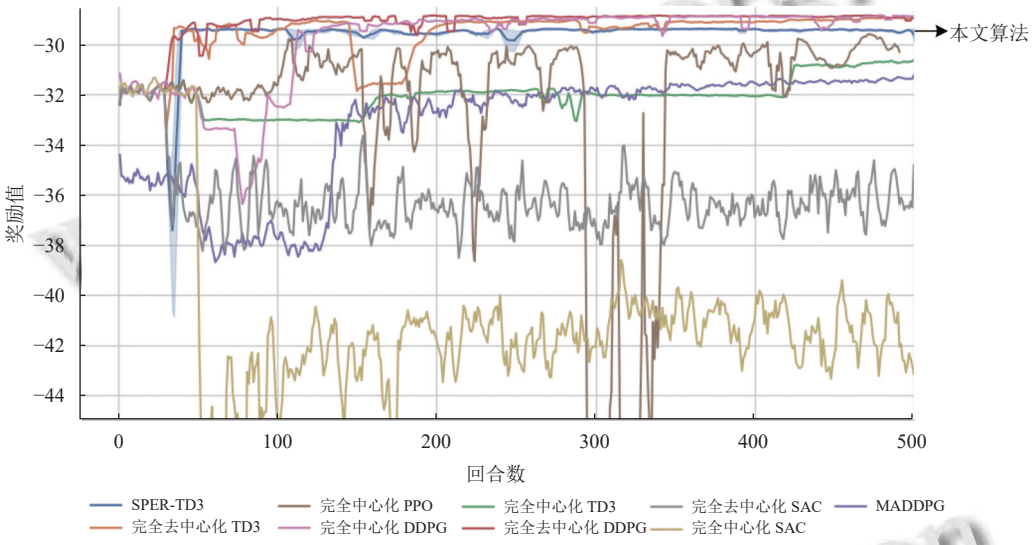
算法性能测试如表 4 所示, 可见本文所提算法在所有算法性能指标中路径长度适中, 保证了安全性; 奖励大小、轨迹安全覆盖率、成功率指标均位于前列;

与现在流行的 MADDPG 算法对比, 所提算法成功率略低, 而其他性能指标均优于 MADDPG, 故相较而言综合性能优于 MADDPG. 对比原始的完全去中心化算法, 所有指标均有所提升; 本文所提算法的综合优势, 对于 4 种算法指标形成的重要性综合评估值, 高于其他深度强化学习算法, 性能表现优秀.

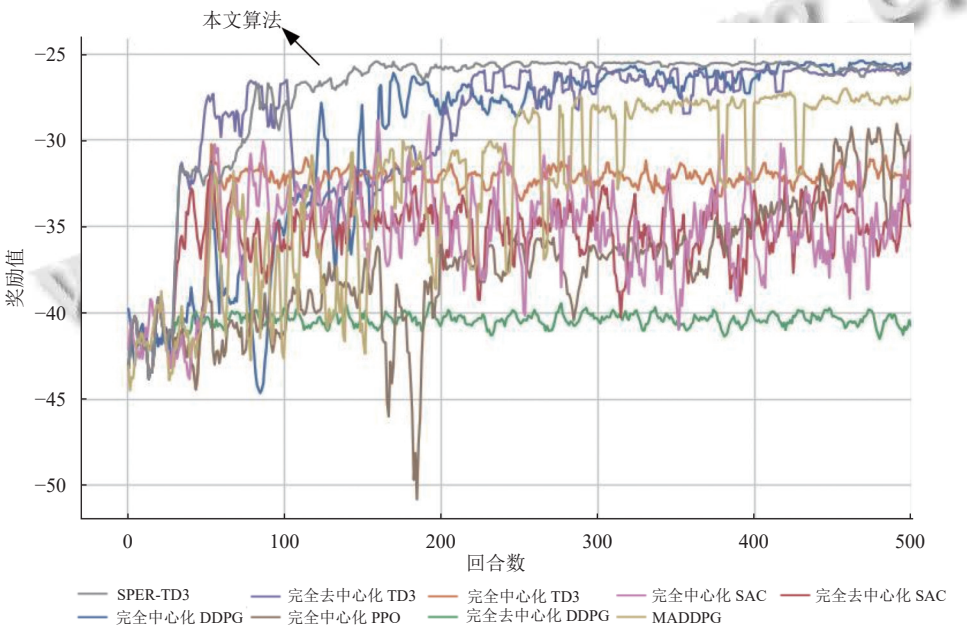
如图 8 所示, 在平均回合奖励指标中, 本文所提出算法的平均回合奖励在所有算法中位于第 1 梯队, 且在小型障碍物环境中最高. 对比其他算法来看, 两种环境内收敛速度均达到最快, 并且稳定性最高.

表 4　深度强化学习算法指标对比

算法	路径长度 (km)	奖励大小	成功率 (%)	轨迹安全覆盖率 (%)	重要性综合评估值
SPER-TD3	13.26	29.96	98.80	**33.33**	**0.650**
MADDPG	13.17	29.92	**99.00**	19.40	0.596
完全去中心化 TD3	13.33	30.92	98.40	24.84	0.585
完全中心化TD3	13.60	30.70	98.00	31.43	0.642
完全去中心化 DDPG	12.84	**28.54**	97.40	22.72	0.599
完全中心化 DDPG	13.78	32.01	94.00	31.22	0.477
完全去中心化 SAC	**12.74**	29.86	92.20	30.43	0.478
完全中心化 SAC	13.77	31.47	90.60	26.57	0.498
完全中心化 PPO	13.92	31.91	72.20	26.57	0.376



(a) 大型障碍物环境



(b) 小型障碍物环境

图 8　平均回合奖励曲线图

5 结论

针对未知障碍物环境, 本文基于 SPER-TD3 算法进行无人机编队航迹规划研究, 首先对无人机编队控制算法进行改进并提出了基于角度的虚拟领航者法控制算法, 对于跟随者提出了动态轨迹平滑算法, 在保证跟随者安全、轨迹平滑的同时保持编队的队形统一, 并提升了算法的鲁棒性. 其次, 在提出的 SPER-TD3 算法中, 采用了基于 SumTree 的优先回放机制对经验进行采样, 加快了学习进度, 提高样本利用效率. 提出了结合 LSTM、自注意力机制以及多重感知机的网络模型结构, 加快收敛速度, 提升算法训练稳定性. 实验结果表明, 在大型障碍物、小型障碍物实验环境中, 编队的轨迹平滑度得到明显提升, 队形保持良好. 算法性能层面, 所提算法在收敛速度、安全性、稳定性等方面均得到有效验证. 在进一步的研究工作中, 可以将本文的算法针对大型障碍物环境进行改进, 以获取更好的算法性能.

参考文献

1 吴杰宏, 李丹阳. 无人机集群编队控制方法研究综述. 无线电通信技术, 2023, 49(4): 589–596. [doi: 10.3969/j.issn.1003-3114.2023.04.001]

2 Tang J, Duan HB, Lao SY. Swarm intelligence algorithms for multiple unmanned aerial vehicles collaboration: A comprehensive review. Artificial Intelligence Review, 2023, 56(5): 4295–4327. [doi: 10.1007/s10462-022-10281-7]

3 Saadi AA, Soukane A, Meraihi Y, *et al*. UAV path planning using optimization approaches: A survey. Archives of Computational Methods in Engineering, 2022, 29(6): 4233–4284. [doi: 10.1007/s11831-022-09742-7]

4 Guo YC, Liu XX, Jiang W, *et al*. Collision-free 4D dynamic path planning for multiple UAVs based on dynamic priority RRT* and artificial potential field. Drones, 2023, 7(3): 180. [doi: 10.3390/drones7030180]

5 Fang YX, Yao YP, Zhu F, *et al*. Piecewise-potential-field-based path planning method for fixed-wing UAV formation. Scientific Reports, 2023, 13(1): 2234. [doi: 10.1038/s41598-023-28087-0]

6 陈博琛, 唐文兵, 黄鸿云, 等. 基于改进人工势场的未知障碍物无人机编队避障. 计算机科学, 2022, 49(S1): 686–693. [doi: 10.11896/jsjkx.210500194]

7 Chen YQ, Yu QZ, Han D, *et al*. UAV path planning: Integration of grey wolf algorithm and artificial potential field. Concurrency and Computation: Practice and Experience, 2024, 36(15): e8120. [doi: 10.1002/cpe.8120]

8 Meng QC, Chen K, Qu QJ. PPSwarm: Multi-UAV path planning based on hybrid PSO in complex scenarios. Drones, 2024, 8(5): 192. [doi: 10.3390/drones8050192]

9 司鹏搏, 吴兵, 杨睿哲, 等. 基于多智能体深度强化学习的无人机路径规划. 北京工业大学学报, 2023, 49(4): 449–458. [doi: 10.11936/bjutxb2022080007]

10 Lee MH, Moon J. Deep reinforcement learning-based model-free path planning and collision avoidance for UAVs: A soft actor-critic with hindsight experience replay approach. ICT Express, 2023, 9(3): 403–408. [doi: 10.1016/j.icte.2022.06.004]

11 Xu YH, Wei YR, Jiang KY, *et al*. Action decoupled SAC reinforcement learning with discrete-continuous hybrid action spaces. Neurocomputing, 2023, 537: 141–151. [doi: 10.1016/j.neucom.2023.03.054]

12 Zhu YF, Tan YJ, Chen YF, *et al*. UAV path planning based on random obstacle training and linear soft update of DRL in dense urban environment. Energies, 2024, 17(11): 2762. [doi: 10.3390/en17112762]

13 沈骁, 赵彤洲. 基于 DDQN 的无人机区域覆盖路径规划策略. 电子测量技术, 2023, 46(14): 30–36.

14 陈康雄, 刘磊. 基于扰动流体与 TD3 的无人机路径规划算法. 电光与控制, 2024, 31(1): 57–62. [doi: 10.3969/j.issn.1671-637X.2024.01.009]

15 Shan TL, Wang Y, Zhao CX, *et al*. Multi-UAV WRSN charging path planning based on improved heed and IA-DRL. Computer Communications, 2023, 203: 77–88. [doi: 10.1016/j.comcom.2023.02.021]

16 Zhou WH, Li J, Liu ZH, *et al*. Improving multi-target cooperative tracking guidance for UAV swarms using multi-agent reinforcement learning. Chinese Journal of Aeronautics, 2022, 35(7): 100–112. [doi: 10.1016/j.cja.2021.09.008]

17 Zhao XR, Yang RN, Zhong LS, *et al*. Multi-UAV path planning and following based on multi-agent reinforcement learning. Drones, 2024, 8(1): 18. [doi: 10.3390/drones8010018]

18 Zhou YT, Kong XR, Lin KP, *et al*. Novel task decomposed multi-agent twin delayed deep deterministic policy gradient algorithm for multi-UAV autonomous path planning. Knowledge-based Systems, 2024, 287: 111462. [doi: 10.1016/j.knosys.2024.111462]

19 Xu YH, Wei YR, Wang D, *et al*. Multi-UAV path planning in GPS and communication denial environment. Sensors, 2023, 23(6): 2997. [doi: 10.3390/s23062997]

20 文超, 董文瀚, 解武杰, 等. 基于解耦型 MADDPG 的无人机集群自主跟踪与避障. 飞行力学, 2022, 40(6): 24–31.

(校对责编: 张重毅)