中国科学: 技术科学 2026 年 第 56 卷 第 3 期: 809 ~ 822
SCIENTIA SINICA Technologica techcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
论 文 厦门大学建校105周年航空宇航科学与技术学科专刊

基于HCA-MAPPO深度强化学习的多无人机协同路径规划方法

林振希[1], 郭正玉[2,3], 张建[1,4], 韩宇[2,3], 陈才轶[1], 陈惠敏[4], 罗德林[1*]

1. 厦门大学航空航天学院, 厦门 361102
2. 中国空空导弹研究院, 洛阳 471000
3. 空基信息感知与融合全国重点实验室, 洛阳 471000
4. 昌吉学院航空学院, 昌吉 831100
* E-mail: luodelin1204@xmu.edu.cn

收稿日期: 2025-11-30; 接受日期: 2026-01-29; 网络版发表日期: 2026-03-18
空基信息感知与融合全国重点实验室开放课题项目(编号: 202462)和新疆低空空域智能管控及安全保障技术重点实验室项目(编号: XJDX2512)资助

摘要 针对多无人机协同路径规划中存在的感知信息处理扁平化、稀疏奖励下探索效率低下等挑战, 本文提出了一种新的基于多智能体深度强化学习的协同路径规划方法. 本方法将多无人机协同路径规划问题建模为部分可观测的马尔可夫决策过程, 并采用深度强化学习算法进行求解. 为了增强智能体对多源异构观测信息的特征提取与态势感知能力, 本文设计了一种基于分层交叉注意力的执行者网络, 通过序贯处理流程实现了对多源异构信息的优先级感知与处理. 此外, 构建了一种基于协同一致性门控的内在好奇心模块, 将个体探索与集群协同目标有效耦合, 提升了算法在稀疏奖励环境下的有效学习效率. 最后, 通过仿真实验验证了所提出算法的有效性和优越性.

关键词 无人机协同, 路径规划, 多智能体深度强化学习, 分层交叉注意力, 好奇心机制

1 引言

随着无人机(unmanned aerial vehicle, UAV)技术的快速发展, 无人机集群协同路径规划已成为实现复杂任务执行的关键技术[1~7], 该任务要求集群在未知且充满障碍的环境中, 仅依靠局部观测实现安全避障与时空协同, 其本质是一个高维、非线性的序贯决策问题. 尽管基于采样(如RRT算法[8])或启发式搜索(如GA[9])的传统算法在静态场景下表现良好, 但在面对高动态强对抗环境时, 常因计算复杂度过高或对环境适应性不足而难以满足实时协同需求.

深度强化学习(deep reinforcement learning, DRL)的理论与应用[10~15]为解决动态环境下的路径规划问题提供了新的思路. 其中, 多智能体强化学习(multi-agent reinforcement learning, MARL)方法[16~19], 特别是以多智能体近端策略优化(multi-agent proximal

引用格式: 林振希, 郭正玉, 张建, 等. 基于HCA-MAPPO深度强化学习的多无人机协同路径规划方法. 中国科学: 技术科学, 2026, 56: 809–822
Lin Z X, Guo Z Y, Zhang J, et al. A multi-UAV cooperative path planning method based on HCA-MAPPO deep reinforcement learning (in Chinese). Sci Sin Tech, 2026, 56: 809–822, doi: 10.1360/SST-2025-0385

© 2026 《中国科学》杂志社

policy optimization, MAPPO)为代表的算法, 已被证明是解决无人机集群协同问题的有效途径[20]. 然而, 将MARL应用于复杂的无人机协同路径规划时, 现有方法依旧面临一系列深层次的挑战.

首先, 在感知层面, 无人机需要从包含大量邻居与障碍物的局部观测中提取关键特征. 注意力机制[21]因其能够动态分配权重、聚焦关键信息, 本应是处理此类多源异构数据的理想工具. 例如, Wang等人[22]利用注意力机制融合了交互信息. 然而, 现有应用大多采用扁平化处理方式, 即将障碍物威胁与队友协同信息置于同一优先级进行无差别的注意力计算. 这种方式忽略了避撞生存与编队协同在决策逻辑上的层级差异, 导致在复杂态势下关键的避险信息易被淹没, 决策鲁棒性不足. 因此, 设计一种能够体现决策内在逻辑优先级的感知架构显得尤为必要.

其次, 在探索层面, 路径规划属于长时序、稀疏奖励任务. 虽然随机网络蒸馏(RND)[23]等内在好奇心模块能够促进探索, 但其在协同任务中直接应用存在弊端: 通用好奇心机制对新奇性的盲目追求可能诱导个体脱离编队, 严重破坏集群协同性[24]. 因此, 设计一种既能鼓励探索, 又能约束个体行为服务于集体目标的机制, 是提升算法训练效率的关键.

针对上述挑战, 本文提出一种基于分层交叉注意力与多智能体近端策略优化(HCA-MAPPO)相融合的深度强化学习方法. 本文的主要贡献包括两个方面: (1) 提出了一种基于分层交叉注意力(hierarchical cross-attention, HCA)的网络架构, 实现了对异构环境信息的优先级感知与处理; (2) 设计基于协同一致性门控的内在好奇心模块, 将个体内在探索激励与集群协同目标动态耦合. 最终的仿真实验结果验证了所提算法的有效性和优越性.

2 问题描述

本节旨在对本文研究的多无人机协同路径规划问题进行数学描述. 首先, 介绍无人机运动学模型; 随后, 对具体的任务场景进行定义; 最终将其形式化地建模为一个多智能体决策过程.

2.1 无人机运动学模型

本文采用三自由度(3-DOF)运动学模型来描述单个无人机的运动状态. 在该模型中, 无人机的运动由其速度大小v、爬升角γ和航向角χ组成. 无人机的实际状态并非瞬时跟随由智能体输出的指令值$(v_{\mathrm{cmd}}, \gamma_{\mathrm{cmd}}, \chi_{\mathrm{cmd}})$, 而是遵循以下一阶惯性系统在离散时间步长$\Delta t$内进行平滑更新:

$$\begin{cases} v(t+\Delta t) = v(t) + \big(v_{\mathrm{cmd}} - v(t)\big)\big(1 - \mathrm{e}^{-\Delta t/\tau_v}\big), \\ \gamma(t+\Delta t) = \gamma(t) + \big(\gamma_{\mathrm{cmd}} - \gamma(t)\big)\big(1 - \mathrm{e}^{-\Delta t/\tau_\gamma}\big), \\ \chi(t+\Delta t) = \chi(t) + \Delta\chi_{\mathrm{norm}} \cdot \big(1 - \mathrm{e}^{-\Delta t/\tau_\chi}\big), \end{cases} \tag{1}$$

其中, $\tau_v, \tau_\gamma, \tau_\chi$分别是速度、爬升角和航向角响应的时间常数. 对于具有周期性的航向角更新, 用于更新的角度误差$\Delta\chi_{\mathrm{norm}}$是将指令值与当前值的原始差值通过标准化运算, 映射到$[-\pi, \pi]$的区间内, 从而得到两者之间的最短角度差. 根据更新后的状态, 可通过标准三角投影将其分解为三维空间中的速度矢量$\mathbf{V}_{\mathrm{vec}} = [v_x, v_y, v_z]$:

$$\begin{cases} v_x(t+\Delta t) = v(t+\Delta t)\cos(\gamma(t+\Delta t))\sin(\chi(t+\Delta t)), \\ v_y(t+\Delta t) = v(t+\Delta t)\cos(\gamma(t+\Delta t))\cos(\chi(t+\Delta t)), \\ v_z(t+\Delta t) = v(t+\Delta t)\sin(\gamma(t+\Delta t)). \end{cases} \tag{2}$$

最终, 通过对该速度矢量进行欧拉前向积分, 计算出无人机在下一时刻的位置$\mathbf{P} = [x, y, z]$, 从而完成一个仿真步的迭代:

$$\mathbf{P}(t+\Delta t) = \mathbf{P}(t) + \mathbf{V}_{\mathrm{vec}}(t+\Delta t) \cdot \Delta t. \tag{3}$$

2.2 协同约束条件

本文研究的多无人机协同路径规划任务定义为: 在一个包含未知静态障碍物的复杂三维空域中, 由N架无人机(包含领导者与跟随者)组成的集群, 需从分散的初始位置出发, 在满足运动学约束的前提下, 协同规划出安全、高效的路径. 为了评估集群的协同性能, 首先建立空间与时间维度的协同约束.

(1) 空间位置约束

空间位置协同包含两个要素: 第一, 集群必须避免内部碰撞并有效规避环境障碍物; 第二, 跟随者必须保持在领导者或一致性目标的有效范围内, 防止因掉队而导致编队解体.

定义$\mathbf{p}_i(t)$为无人机i在时刻t的位置, $\mathcal{O}_k$为第k个障碍物的区域集合, $\mathbf{p}_i^{\mathrm{target}}(t)$为智能体$i$经一致性算法计

算出的动态目标点. 空间位置协同约束可建模为以下不等式组:

$$\begin{cases} \| \mathbf{p}_i(t) - \mathbf{p}_j(t) \| \geqslant d_{\text{safe}}, & \forall i \neq j, \\ \min_{\mathbf{p}_{\text{obs}} \in \mathcal{O}_k} \| \mathbf{p}_i(t) - \mathbf{p}_{\text{obs}} \| \geqslant d_{\text{obs}}, & \forall i,k, \\ \| \mathbf{p}_i(t) - \mathbf{p}_i^{\text{target}}(t) \| \leqslant R_{\text{follow}}, & \forall i \in N_{\text{fol}}, \end{cases} \tag{4}$$

其中, d_{safe}为机间最小安全间隔, d_{obs}为避障安全阈值, R_{follow}为最大允许跟随半径. 从第一项到第三项分别为机间避撞约束、障碍物避障约束、集群保持约束.

(2) 时间约束

为了满足“多起点、终点同步”的战术需求, 集群需要在时间维度上实现协同. 定义T_i为第i架无人机首次到达目标区域$\mathcal{R}_{\text{goal}}$的时刻. 时间协同要求集群中最早到达时刻$T_{\text{first}}$与最晚到达时刻$T_{\text{last}}$之间的时间差必须控制在允许的阈值$\delta$内:

$$\begin{cases} T_{\text{last}} - T_{\text{first}} \leqslant \delta, \\ T_i = \min\{t \mid \mathbf{p}_i(t) \in \mathcal{R}_{\text{goal}}\}. \end{cases} \tag{5}$$

2.3　协同路径规划的马尔可夫决策过程模型

为了寻求能够同时满足式(4)空间约束与式(5)时间同步约束的最优协同策略, 本文将上述满足协同约束的路径规划问题, 形式化地建模为一个去中心化部分可观马尔可夫决策过程(dec-entralized partially observable Markov decision process, Dec-POMDP), 由元组$\langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{R_i\}_{i=1}^N, \{\mathcal{O}_i\}_{i=1}^N, N, \gamma \rangle$定义, 其中$\mathcal{S}$为全局状态空间, $\mathcal{A}_i$为智能体i的动作空间, 即前述的控制指令, P为状态转移概率, R_i为奖励函数, N为智能体总数, γ为折扣因子. 局部观测空间$\mathcal{O}_i$则被显式地结构化为自身状态s_{self}、领导者实体E_{lead}、障碍物实体E_{obs}及跟随者实体E_{fol}四部分, 以为后续分层交叉注意力网络提供异构特征基础. 所有智能体的共同目标是学习一个联合策略$\pi = \{\pi_1, \ldots, \pi_N\}$, 其中每个智能体的局部策略$\pi_i(a_i \mid o_i)$仅依赖于其局部观测, 以最大化所有智能体的期望累计折扣奖励之和R:

$$R = \max_{\pi} \mathbb{E}_{\mathbf{s}_0, \mathbf{a}_0, \ldots} \left[\sum_{t=0}^{\mathrm{T}} \gamma^t \sum_{i=1}^{N} R_i(\mathbf{s}_t, \mathbf{a}_t) \right]. \tag{6}$$

3　基于HCA-MAPPO深度强化学习的多无人机协同路径规划

为解决在第2节中定义的多无人机协同路径规划问题, 本文提出了一种分层交叉注意力与多智能体近端策略优化相融合的深度强化学习方法(HCA-MAPPO).

3.1　算法基础框架

在算法层面, 本文选择当前在多智能体协同任务中表现先进的多智能体近端策略优化(MAPPO)算法作为基础框架. MAPPO算法遵循“中心化训练、去中心化执行”(CTDE)的范式, 其结构由两类网络构成: 每个智能体独立的执行者(actor)网络$\pi_{\theta_i}(a_i \mid o_i)$和所有智能体共享的评论家(critic)网络$V_\phi(s)$. 评论家网络在训练时利用全局信息$s$对状态价值进行准确评估, 其参数$\phi$通过最小化时序差分误差进行更新, 损失函数$L(\phi)$通常采用均方误差形式:

$$L(\phi) = E[(V_\phi(s_t) - Y_t)^2], \tag{7}$$

其中, Y_t为目标价值, 通常使用广义优势估计(generalized advantage estimation, GAE)方法计算得到. 执行者网络通过最大化一个经过裁剪的代理目标函数来更新参数θ_i, 以在保证策略更新幅度的同时提高样本利用效率. 其目标函数$L^{\text{CLIP}}(\theta_i)$定义如下:

$$\begin{aligned} &L^{\text{CLIP}}(\theta_i) \\ &= \mathbb{E}_t[\min(r_t(\theta_i)\hat{A}_t, \text{clip}(r_t(\theta_i), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)], \end{aligned} \tag{8}$$

其中, $r_t(\theta_i) = \dfrac{\pi_{\theta_i}(a_t \mid o_t)}{\pi_{\theta_{i,old}}(a_t \mid o_t)}$是新旧策略的概率比, $\hat{A}_t$是由评论家网络计算出的优势函数, ε是用于裁剪的超参数. 为鼓励探索, 通常在最终的损失函数中加入策略熵S的正则项.

3.2　基于分层交叉注意力的网络与结构化观测空间

为了提高执行者网络在处理多源异构信息时的效率, 本文设计了一种基于分层交叉注意力(HCA)机制的新型执行者网络结构. 该结构核心思想是, 不再将所有观测信息扁平化地输入网络, 而是根据信息源的类型和决策的内在逻辑优先级, 进行有序的、有重点

的注意力计算, 从而为智能体提供更强的规划决策.

HCA执行者网络接收结构化的观测信息(在第2.2节中定义), 即自身状态$\mathbf{s}_{\text{self}}$, 以及领导者实体$\mathbf{E}_{\text{lead}}$、障碍物实体$\mathbf{E}_{\text{obs}}$和其他跟随者实体$\mathbf{E}_{\text{fol}}$的状态矩阵. 网络通过一个序贯的三层多头交叉注意力流程, 对上述信息进行处理. 首先, 通过独立的、带有非线性激活函数的线性嵌入层将各类原始状态信息映射到注意力机制的高维特征空间:

$$\begin{cases} \mathbf{Q}_1 = \max(0, \mathbf{W}_{\text{self}}\mathbf{s}_{\text{self}} + \mathbf{b}_{\text{self}}), \\ \mathbf{K}_1, \mathbf{V}_1 = \max(0, \mathbf{E}_{\text{lead}}\mathbf{W}_{\text{lead}}{}^{\mathrm{T}} + \mathbf{b}_{\text{lead}}), \\ \mathbf{K}_2, \mathbf{V}_2 = \max(0, \mathbf{E}_{\text{obs}}\mathbf{W}_{\text{obs}}{}^{\mathrm{T}} + \mathbf{b}_{\text{obs}}), \\ \mathbf{K}_3, \mathbf{V}_3 = \max(0, \mathbf{E}_{\text{fol}}\mathbf{W}_{\text{fol}}{}^{\mathrm{T}} + \mathbf{b}_{\text{fol}}), \end{cases} \tag{9}$$

其中, $\mathbf{W}$和$\mathbf{b}$分别是对应嵌入层的可学习权重矩阵和偏置向量, $\mathbf{Q}_1$为初始查询向量. 随后, 网络按以下分层逻辑进行计算, 其详细结构如图1所示.

第一层(意图感知层): 智能体首先关注领导者的状态以明确任务意图. 利用初始查询向量$\mathbf{Q}_1$对领导者实体$(\mathbf{K}_1, \mathbf{V}_1)$进行多头交叉注意力计算, 生成领导者上下文向量$\mathbf{C}_{\text{lead}}$, 该计算由$h$个并行的注意力头组成, 每个头$\text{head}_j$独立计算一个加权上下文. 对于第$j$个头, 其注意力权重$\alpha_j$的计算如下:

$$\alpha_{jk} = \frac{\exp(e_{jk})}{\sum_{l=1}^{N_{\text{lead}}} \exp(e_{jl})}, \tag{10}$$

其中, e_{jk}是查询向量与第k个领导者实体之间的缩放点积相似度分数:

$$e_{jk} = \frac{(\mathbf{Q}_1\mathbf{W}_j^{\mathbf{Q}})(\mathbf{K}_1[k]\mathbf{W}_j^{\mathbf{K}})^{\mathrm{T}}}{\sqrt{d_k}}. \tag{11}$$

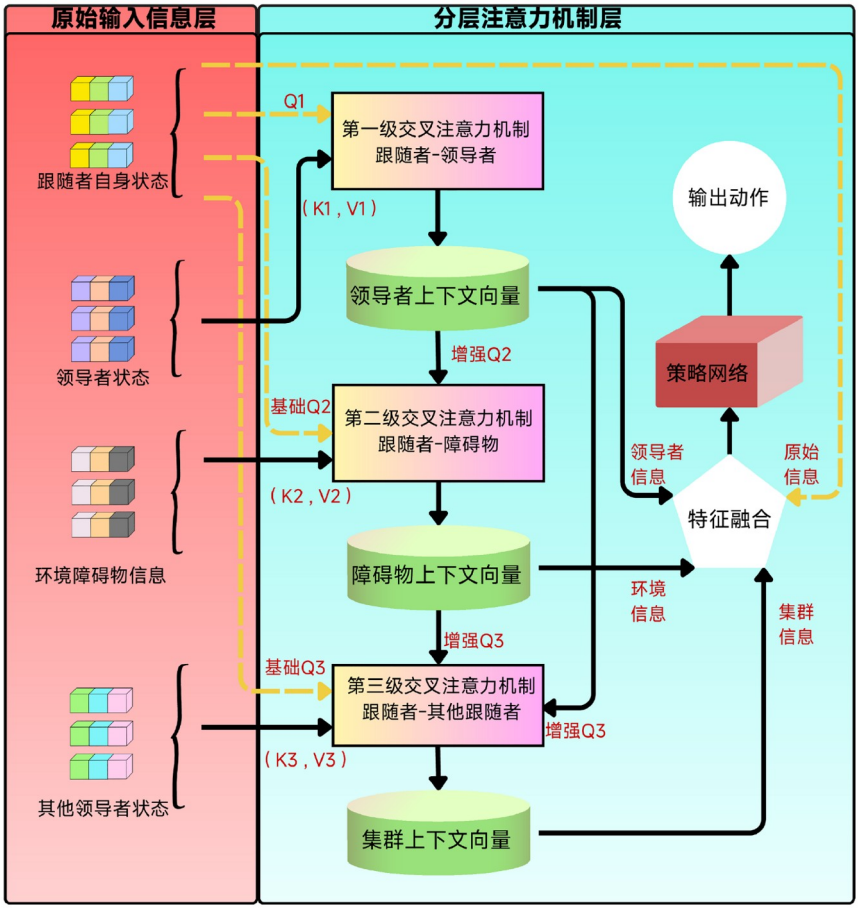


图 1 (网络版彩图)分层交叉注意力机制(HCA)

Figure 1 (Color online) Hierarchical cross-attention mechanism (HCA).

由此权重, 第j个头输出为

$$\text{head}_j = \sum_{k=1}^{N_{\text{lead}}} \alpha_{jk}(\mathbf{V}_1[k]\mathbf{W}_j^{\mathrm{v}}).$$

所有头的输出被拼接并再次通过一个线性投影得到最终的上下文向量:

$$\mathbf{C}_{\text{lead}} = [\text{head}_1, \ldots, \text{head}_h]W^O. \tag{12}$$

第二层(威胁感知层): 在明确任务意图后, 智能体转而关注环境威胁. 将第一层生成的领导者上下文向量$\mathbf{C}_{\text{lead}}$与初始查询向量$\mathbf{Q}_1$进行拼接, 并通过一个更新模块(一个带有非线性激活的线性层)生成一个增强的查询向量:

$$\mathbf{Q}_2 = \max(0, \mathbf{W}_{\text{update2}}[\mathbf{Q}_1^{\mathrm{T}}, \mathbf{C}_{\text{lead}}^{\mathrm{T}}]^{\mathrm{T}} + \mathbf{b}_{\text{update2}}). \tag{13}$$

该增强查询向量对障碍物实体$(\mathbf{K}_2, \mathbf{V}_2)$进行第二次多头交叉注意力计算, 生成障碍物上下文向量$\mathbf{C}_{\text{obs}}$. 计算过程与第一层相同, 仅将输入替换为$(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2)$.

第三层(协同感知层): 在充分感知任务意图和环境威胁后, 智能体最终关注集群内部的协同需求. 再次, 将第二层生成的障碍物上下文向量$\mathbf{C}_{\text{obs}}$与前序的增强查询向量$\mathbf{Q}_2$进行拼接, 通过另一个独立的更新模块生成最终的查询向量:

$$\mathbf{Q}_3 = \max(0, \mathbf{W}_{\text{update3}}[\mathbf{Q}_2^{\mathrm{T}}, \mathbf{C}_{\text{obs}}^{\mathrm{T}}]^{\mathrm{T}} + \mathbf{b}_{\text{update3}}). \tag{14}$$

该最终查询向量对其他跟随者$(\mathbf{K}_3, \mathbf{V}_3)$进行第三次多头交叉注意力计算, 生成集群上下文向量$\mathbf{C}_{\text{fol}}$.

最终, 网络将初始查询向量的嵌入$\mathbf{Q}_1$、领导者上下文向量$\mathbf{C}_{\text{lead}}$、障碍物上下文向量$\mathbf{C}_{\text{obs}}$以及集群上下文向量$\mathbf{C}_{\text{fol}}$进行最终的特征融合, 并将融合后的高维特征向量$\mathbf{F}_{\text{final}} = [\mathbf{Q}_1^{\mathrm{T}}, \mathbf{C}_{\text{lead}}^{\mathrm{T}}, \mathbf{C}_{\text{obs}}^{\mathrm{T}}, \mathbf{C}_{\text{fol}}^{\mathrm{T}}]^{\mathrm{T}}$输入到一个策略网络$\pi_\theta$中, 最终输出无人机的控制动作$a_i \sim \pi_\theta(F_{\text{final}})$.

3.3　基于加权一致性算法的协同目标生成

本文为每个跟随者智能体设计了一种动态协同目标生成机制. 该机制的核心是基于加权一致性算法, 为每个智能体实时地计算出“一致性移动目标点”. 该过程如图2所示, 主要分为权重矩阵求解和目标状态迭代两个步骤.

第一步, 构建归一化权重矩阵$\overline{\mathbf{W}}$. 系统根据输入的无人机集群拓扑图$\mathcal{G}$(表明无人机两两之间是否可以信息传递)、领导者信息和信息传递因子$\eta \in (0,1)$进行计算: 首先, 根据集群拓扑$\mathcal{G}$构建邻接矩阵$\mathbf{S}$(其中元素$S_{ij}=1$表示无人机i可以感知到无人机j的状态信息); 其次, 计算任意两个节点间的图距离d_{ij}(d_{ij}越短, 无人机信息来源越接近领导者, 其传递的信息越重要, 记领导者自身图距离为0); 接着, 结合邻接关系、图距离和信息传递因子来计算初始权重$\mathbf{W}_{ij} = \mathbf{S}_{ij} \cdot \eta^{d_{ij}}$, 该公式意味着来自更远邻居的信息权重会相应衰减; 最后, 对W_{ij}

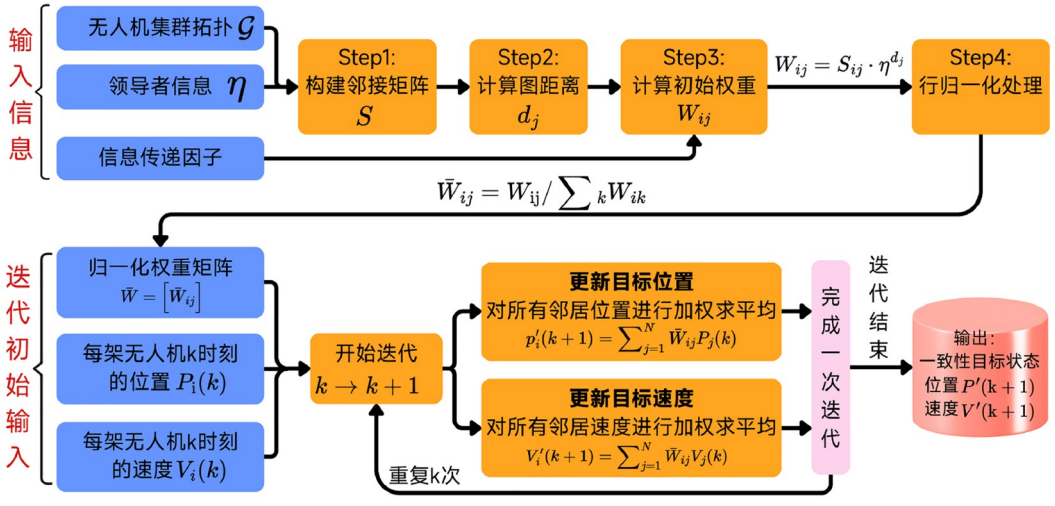


图 2　(网络版彩图)一致性算法求目标点流程图

Figure 2　(Color online) Flowchart of the consistency algorithm for finding the target point.

进行归一化处理得到$\overline{W}_{ij}=W_{ij}/\sum_k W_{ik}$, 确保权重和为1.

第二步, 进行一致性目标状态的迭代计算. 在每个决策时刻k, 以归一化权重矩阵$\overline{\mathbf{W}}$、所有邻居无人机当前时刻的位置$\mathbf{P}_j(k)$和速度$\mathbf{V}_j(k)$作为迭代初始输入. 对所有邻居的位置和速度进行加权平均, 更新得到当前智能体i下一时刻的一致性目标位置$\mathbf{P}'_i(k+1)$和目标速度$\mathbf{V}'_i(k+1)$:

$$\begin{cases}\mathbf{P}'_i(k+1)=\sum_{j=1}^{N}\overline{\mathbf{W}}_{ij}\mathbf{P}_j(k),\\ \mathbf{V}'_i(k+1)=\sum_{j=1}^{N}\overline{\mathbf{W}}_{ij}\mathbf{V}_j(k).\end{cases}\tag{15}$$

该动态目标状态$(\mathbf{P}',\mathbf{V}')$既作为结构化观测信息输入HCA执行者网络以增强态势感知, 又作为协同门控模块与奖励函数的计算基准以引导编队保持.

3.4 协同一致性门控的内在好奇心模块

为解决集群在长航程、稀疏奖励任务中探索效率低下问题, 并克服传统好奇心机制在协同任务中的无效探索, 本文在HCA-MAPPO框架中设计了基于协同一致性门控的内在好奇心模块, 如图3所示.

该模块基于随机网络蒸馏(RND)原理, 通过一个固定的、随机的目标网络f_{target}和一个可学习的预测网络$f_{\text{predictor}}$之间的预测误差来量化状态的新奇性. 对于智能体i遇到的新状态$o_{i,t+1}$, 其原始新奇性信号r_{novelty}被定义为两个网络输出向量之间的欧氏距离的平方:

$$r_{\text{novelty}}(o_{i,t+1})=\| f_{\text{predictor}}(o_{i,t+1};\theta_p)-f_{\text{target}}(o_{i,t+1})\|_2^2,\tag{16}$$

其中, θ_p是预测网络的可学习参数. 然而, 在多智能体协同任务中, 并非所有新奇的状态都是有益的. 一个为了满足自身好奇心而严重偏离编队的智能体, 其探索行为对整个集群是不利的. 为解决这一问题, 本文提出了一种协同一致性门控机制(cooperative-consistency gating mechanism), 用于调节和筛选原始的新奇性信号.

该机制为每个跟随者智能体i引入协同门控函数$G_i(t)$, 该函数的值取决于智能体当前行为与集群协同目标的一致性程度. 在本发明中, 该一致性程度通过智能体当前位置$\mathbf{p}_i(t)$与其动态的"一致性移动目标点"$\mathbf{p}_i^{\text{target}}(t)$之间的欧氏距离$d_i(t)$来量化. 该动态目标点是通过加权一致性算法, 根据预设的通信拓扑, 对智能体自身及其邻居节点的状态进行加权平均而实时计算得出的. 门控函数$G_i(t)$被设计为一个随距离$d_i(t)$增大而单调递减的函数, 具体形式为

$$G_i(t)=\exp\left(-\beta\cdot\left(\frac{d_i(t)}{R_{\text{follow}}}\right)^2\right),\tag{17}$$

其中, $d_i(t)=\|\mathbf{p}_i(t)-\mathbf{p}_i^{\text{target}}(t)\|_2$. R_{follow}是一个用于控制衰减速度的特征跟随半径, β是一个正的超参数. 当智能体紧密跟随时($d_i(t)$趋近于0), $G_i(t)$值趋近于1, 此时

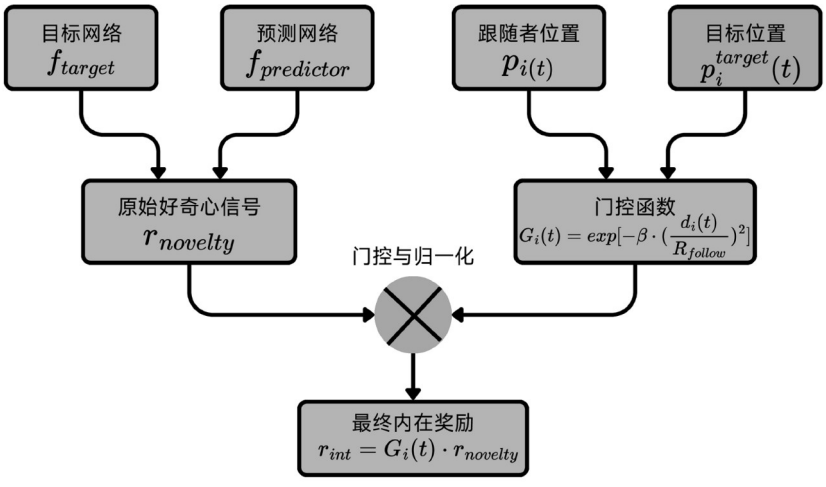


图 3 改进的好奇心模块图

Figure 3 Diagram of the improved curiosity module.

允许智能体获得完整的新奇性奖励; 当智能体严重偏离其协同目标时($d_i(t)$远大于0), $G_i(t)$值趋近0, 从而抑制其因偏离航线产生的无效新奇性奖励.

最终, 智能体i获得的、经过门控调节的有效内在奖励r_{int}被定义为原始新奇性信号与协同门控函数的乘积, 并通过运行时统计进行归一化后, 再与任务的外在奖励进行加权组合, 共同用于策略更新:

$$r_{\text{int}}(o_{i,t+1}) = G_i(t) \cdot r_{\text{novelty}}(o_{i,t+1}). \tag{18}$$

通过引入此协同一致性门控机制, 本文将智能体的探索行为与集群的宏观协同目标进行有效耦合, 激励智能体在保持良好协同状态的前提下进行探索, 从而提升了探索的有效性, 加速了协同策略的学习.

3.5　改进算法框架

综上, 本文提出的HCA-MAPPO深度强化学习方法的改进框架如图4所示. 整个框架分为决策流与学习流两个部分. 在决策流中, 跟随者智能体接收来自环境的局部观测信息(自身信息、障碍物、领导者、其他跟随者), 并通过核心的分层交叉注意力模块进行优先级感知与上下文特征提炼, 最终由策略网络输出动作. 在学习流中, 联合的全局状态S被用于价值网络以评估全局态势, 同时也被好奇心模块用于生成内在奖励. 最终, 外在奖励、内在奖励、状态价值和动作概率共同汇集到策略优化核心(MAPPO), 通过中心化的训练, 计算出对策略网络和价值网络的更新信号, 从而驱动整个框架的学习与进化. 具体训练与更新流程总结于算法1.

3.6　奖励函数设计

为了在多目标约束下引导智能体学习最优策略, 本文采用线性加权的复合奖励函数. 领导者与跟随者总奖励分别为$R_{\text{leader}} = \sum_{k\in\mathcal{K}_L} w_k R_k$和$R_{\text{follower}} = \sum_{k\in\mathcal{K}_F} w_k R_k$, 其中$\mathcal{K}_L = \{1,2,3,4,5,6,7,8\}$为领导者奖励项集合, $\mathcal{K}_F = \{1,2,3,4,5,6,9\}$为跟随者奖励项集合, w_k为各分量的权重系数. 将各奖励项归纳为任务完成、安全引导、编队协同三类进行阐述.

(1) 任务完成类奖励: 此类奖励旨在引导智能体快速接近并到达目标.

到达奖励(R_1): 包含阶段性奖励与终点奖励. 阶段性奖励在无人机i首次穿越距离阈值d_j时触发, 提供稀疏引导; 终点奖励在进入目标半径R_{goal}时给予.

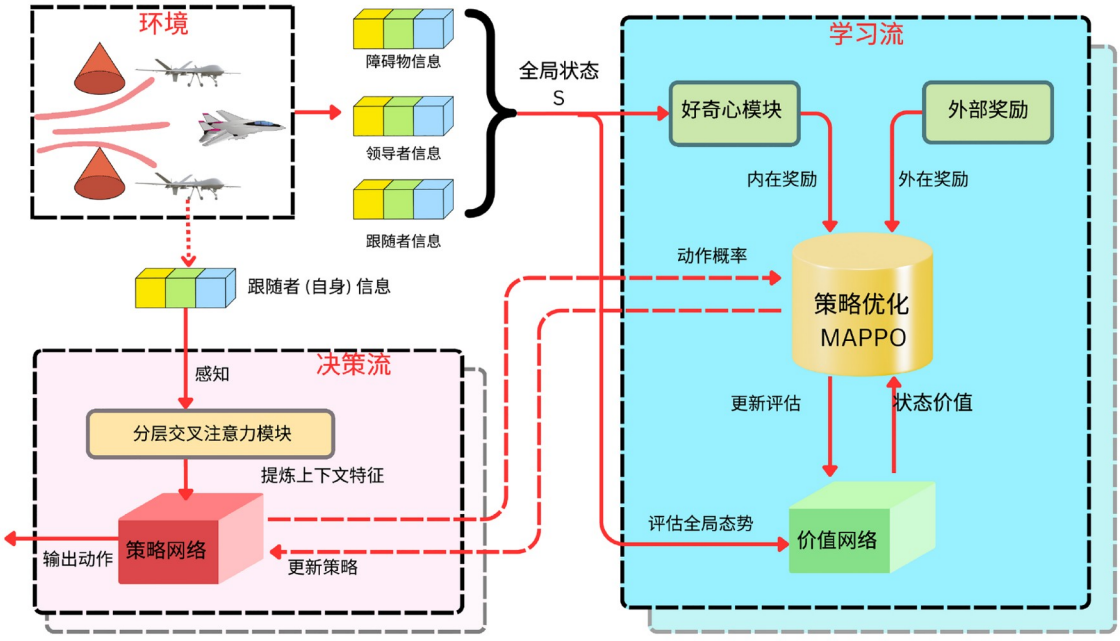


图 4　(网络版彩图)基于HCA-MAPPO深度强化学习的多无人机协同路径规划框架图

Figure 4　(Color online) Framework diagram of multi-UAV cooperative path planning based on HCA-MAPPO deep reinforcement learning.

算法1: HCA-MAPPO算法

对于每个智能体$i \in \{1, \cdots, N\}$, 初始HCA执行者网络π_{θ_i};
初始化共享评论家网络V_ϕ的参数θ_i和ϕ;
for episode = 1, max_episode **do**
初始化一个空的轨迹存储器$\mathcal{D}$;
并行重置个环境, 获得初始观测$\left\{o_{i,m}^{0}\right\}_{i=1,m=1}^{N,M}$;
for t = 1, max_t **do**
for i = 1, N **do**
从策略$\pi_{\theta_i}\left(a_{i,m}^{t}\middle|o_{i,m}^{t}\right)$中采样动作$a_{i,m}^{t}$;
end for
在所有并行环境中执行联合动作$\left\{a_{i,m}^{t}\right\}_{i=1,m=1}^{N,M}$;
获得新的观测$\left\{o_{i,m}^{t+1}\right\}$, 奖励$\left\{r_{i,m}^{t}\right\}$和完成标志$\left\{d_{i,m}^{t}\right\}$;
将轨迹数据$\left(\left\{o_{i,m}^{t}\right\}, \left\{a_{i,m}^{t}\right\}, \left\{r_{i,m}^{t}\right\}, \left\{o_{i,m}^{t+1}\right\}, \left\{d_{i,m}^{t}\right\}\right)$存入$\mathcal{D}$;
end for
使用存储器$\mathcal{D}$中的轨迹数据, 计算优势函数估计$\hat{A}_i^t$;
for epoch = 1, max_epoch **do**
从$\mathcal{D}$中随机抽取小批量数据;
最小化损失函数$L(\phi)$更新共享评论家网络V_ϕ;
最大化目标函数$L^{\mathrm{CLIP}}(\theta_i)$更新执行者网络$\pi_{\theta_i}$;
end for
end for

$$R_1 = \sum_j r_j \cdot \mathbb{I}(\| \mathbf{P}_i(t) - \mathbf{P}_{\mathrm{goal}} \| < d_j) + k_{\mathrm{task}} \cdot \mathbb{I}(\mathrm{arrived}). \quad (19)$$

距离逼近奖励(R_2): 根据每一步与目标点距离的缩减量给予密集奖励, 形式为$R_2 = k_2(d_{t-1} - d_t)$, 其中d_t为当前时刻与目标的距离.

(2) 安全引导类奖励: 此类奖励旨在确保生存并解决路径规划的死锁问题.

避撞与避障惩罚(R_3, R_4): 采用非线性斥力场模型. 当智能体与障碍物或地面距离$d < d_{\mathrm{safe}}$时, 施加指数增长的惩罚R_3; 若发生碰撞($d \leqslant 0$), 则施加截断惩罚R_4. 具体公式如下:

$$\begin{cases} R_3 = -k_3\left(\dfrac{d_{\mathrm{safe}}}{d} - 1\right), \\ R_4 = -k_4. \end{cases} \quad (20)$$

基于冲突感知的智能引导奖励(R_5): 针对传统势场法易陷入局部最优的缺陷[12], 该奖励项的作用在于主动识别并打破这种僵局. 当检测到吸引力与排斥力矢量的夹角θ超过阈值时, 该机制强制将引导方向$\mathbf{v}_{\mathrm{guide}}$从指向目标切换为切线方向, 从而引导无人机在障碍物前做出果断的侧向绕行动作, 而非停滞不前. 奖励为当前速度与引导方向的余弦相似度:

$$R_5 = k_5 \cdot \frac{\mathbf{v}_{\mathrm{uav}} \cdot \mathbf{v}_{\mathrm{guide}}}{\| \mathbf{v}_{\mathrm{uav}} \| \| \mathbf{v}_{\mathrm{guide}} \| + \epsilon}. \quad (21)$$

(3) 编队协同类奖励: 此类奖励用于维持集群时空特征与动作稳定性.

动作平滑奖励(R_6): 惩罚当前动作$a_i(t)$与最近W步动作移动平均值$\bar{a}_i(t)$的偏差, 抑制控制震荡:

$$R_6 = -k_6 \left\| a_i(t) - \frac{1}{W} \sum_{j=0}^{W-1} a_i(t-j) \right\|_2^2. \quad (22)$$

领导者协同奖励(R_7, R_8): R_7用于惩罚不同领导者(L0, L1)到达目标点的距离差, 以实现时间同步; R_8用于维持领导者间的安全间距$d_{\mathrm{safe},L}$.

$$\begin{cases} R_7 = -k_7 \left| \| \mathbf{P}_{\mathrm{L0}}(t) - \mathbf{P}_{\mathrm{goal}} \| - \| \mathbf{P}_{\mathrm{L1}}(t) - \mathbf{P}_{\mathrm{goal}} \| \right|, \\ R_8 = -k_8\left(\dfrac{d_{\mathrm{safe},L}}{\| \mathbf{P}_{\mathrm{L0}}(t) - \mathbf{P}_{\mathrm{L1}}(t) \|} - 1\right), \\ \qquad \text{if } \| \mathbf{P}_{\mathrm{L0}} - \mathbf{P}_{\mathrm{L1}} \| < d_{\mathrm{safe},L}. \end{cases} \quad (23)$$

跟随者掉队惩罚(R_9): 当跟随者与一致性目标点的距离超过最大跟随半径R_{follow}时, 施加惩罚:

$$R_9 = -k_9 \cdot \| \mathbf{P}_i(t) - \mathbf{P}_i^{\mathrm{target}}(t) \|, \ \text{if } d > R_{\mathrm{follow}}. \quad (24)$$

4 仿真验证

4.1 实验环境与参数设置

本文仿真实验基于Python3.9与PyTorch框架搭建. 运行于配备Intel Core i9-12900K CPU、64GB RAM及RTX 3090 GPU的服务器. 训练过程采用16个并行环境实例同步采样以提升效率.

为确保对比公平性, 所有基准算法(MAPPO, MAPPO+RND和HCA-MAPPO)与本文提出的算法(HCA-MAPPO+RND)均采用完全一致的训练配置. 具

体而言，所有算法均训练了20000万个环境步(200M steps)，且其Actor和Critic网络均采用了相同的隐藏层结构，学习率、折扣因子等通用超参数也保持一致，详细参数设置如表1所示.

为了确保算法的收敛性与性能，关键参数选取遵循以下原则: (1) MAPPO基础参数(如学习率、折扣因子、裁剪阈值)直接参考文献[25]在多智能体协作任务中的推荐配置，以保证基准性能的稳定性. (2) HCA网络结构(交叉注意力维度128，MLP [128, 256, 128])是基于高维输入状态空间大小，在特征提取能力与计算开销之间权衡确定的. (3) 任务特定权重(如w_k及门控系数β)采用经验性网格搜索方法确定，优先选取能满足避障与到达约束的参数范围，随后微调以优化协同效率. 例如，内在好奇心系数设为0.05，旨在平衡探索激励与外在任务奖励.

4.2　结果与分析

4.2.1　消融对比实验分析

为验证算法的可行性与各模块贡献，本文设置了三种实验场景: 场景1含1架领导者与3架跟随者; 场景2含1架领导者与5架跟随者; 场景3含2架领导者与10架跟随者(分为两组协同任务).

同时，设计了4种对比算法进行消融分析: 算法1为本文完整方法(HCA-MAPPO+改进好奇心); 算法2为HCA-MAPPO+原始好奇心; 算法3为标准MAPPO+改进好奇心; 算法4为标准MAPPO基准. 基于训练好的模型，在随机初始化条件下(起始点附近200 m内)对每种组合进行100次独立测试. 测试指标如下.

成功率J_{SR}(success rate). 在无人机飞行过程中，没有碰撞，跟随者没有掉队，视为任务成功，成功率为$J_{\mathrm{SR}} = N_S / N$，其中$N$为测试次数，$N_S$为成功次数.

平均步长J_{AS}(average step). 表示无人机从起始点到达目标点所需的平均步长数. 平均步长$J_{\mathrm{AS}}=\left(\sum_{i=1}^{N_S} L_S\right)/N_S$，其中$L_S$表示在任务成功的测试中，每次测试的步长数.

平均轨迹长度J_{ATL}(average trajectory length). 表示无人机从起始点到达目标点的路径的平均轨迹长度. $J_{\mathrm{ATL}}=\left(\sum_{i=1}^{N_S} S\right)/N_S$，其中$S$表示在任务成功的测试中，每次测试的路径长度(m).

将4种算法与3种场景分别一一组合. 测试的结果如表2所示. 从表中可以看出，无论哪种场景(场景1、场景2、场景3)，使用本文改进内在好奇心模块的HCA-MAPPO算法(算法1)的实验结果具有一定优越性: 相比于其他三种算法，本文算法成功率J_{SR}更高，平均步长J_{AS}更小，平均轨迹长度J_{ATL}更短.

如图5所示，在最复杂的场景3下，算法1在收敛速度与最终性能上均表现最优. 深入分析其机理可知: 首先，引入HCA架构的算法1相比基础MAPPO(算法3)获得了显著更高的奖励上限. 这归因于HCA的序贯处理结构，通过“意图-威胁-协同”的优先级逻辑，有效实现了对多源异构信息的解耦，避免了关键避障信息被协同信号淹没. 其次，使用改进门控机制的算法1相比原始好奇心(算法2)在训练后期的震荡幅度更小. 其原因在于标准RND存在盲目性，易诱导偏离航线的无效探索; 而协同门控机制相当于引入了任务滤波器，利用一致性目标动态抑制了与集群协同相悖的奖励，从而增强了策略稳定性.

表 1　HCA-MAPPO的参数与训练参数

Table 1　Parameters of HCA-MAPPO and training parameters

参数	值
时间步长dt (s)	0.1
训练总步长	200000000
每回合最大步长	400
GAE参数	0.95
协同门控系数	1.0
内在好奇心模块系数	0.1
折扣因子	0.99
交叉注意力特征维度	128
MLP层神经元个数	[128, 256, 128]
PPO裁剪值	0.2
梯度裁剪值	10
策略熵损失权重	0.01

4.2.2　HCA-MAPPO实验结果分析

为直观展示算法的协同性能，本文利用Tacview对算法1/场景3(2架领导者、10架跟随者)的飞行轨迹进行了三维可视化(图6). 实验设置两组编队分别从

表 2 HCA-MAPPO算法(算法1)及消融对比实验

Table 2 HCA-MAPPO algorithm (Algorithm 1) and ablation comparison experiment

场景	J_{SR} (%)	J_{AS}	J_{ATL} (m)
算法1/场景1	96.0	184.6	3215.4
算法1/场景2	94.0	191.5	3390.2
算法1/场景3	91.0	193.9	3470.0
算法2/场景1	87.0	190.2	3311.7
算法2/场景2	88.0	199.7	3457.7
算法2/场景3	85.0	205.0	3602.1
算法3/场景1	65.0	207.2	3744.6
算法3/场景2	62.0	211.2	3812.8
算法3/场景3	45.0	211.5	3821.5
算法4/场景1	66.0	213.8	3812.0
算法4/场景2	55.0	215.4	3922.4
算法4/场景3	40.0	220.5	4117.1

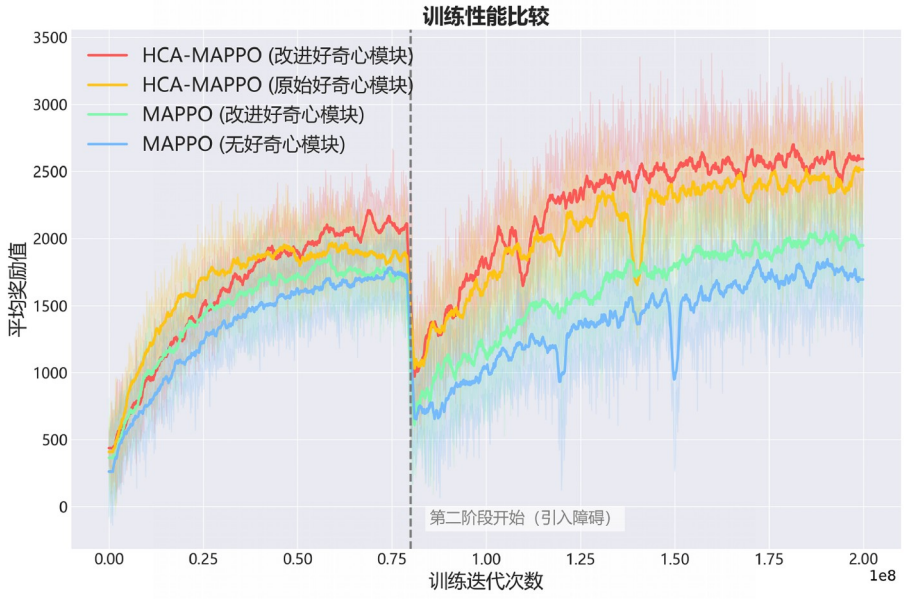


图 5 (网络版彩图)奖励函数曲线图

Figure 5 (Color online) Reward function curve.

(120.008°E, 60°N, 800 m)和(120°E, 60.005°N, 800 m), 协同飞往距起点约3.5 km的同一目标点(120.05°E, 60°N, 1500 m).

首先, 在时间协同方面, 得益于领导者协同奖励机制, 两架领导者到达目标点的时间高度同步. 如图7所示, 两者分别耗时183与190步, 时间差仅为0.7 s(7步), 有效验证了时间协同约束的满足. 其次, 在空间安全与编队保持方面, 如图8所示, 所有无人机在飞行全程中的机间距离均保持在安全阈值(25 m)以上, 且未超过最大跟随半径(400 m). 这表明在HCA-MAPPO算法驱动下, 集群不仅成功规避了碰撞风险, 还始终维持了紧密的编队构型, 未出现掉队现象.

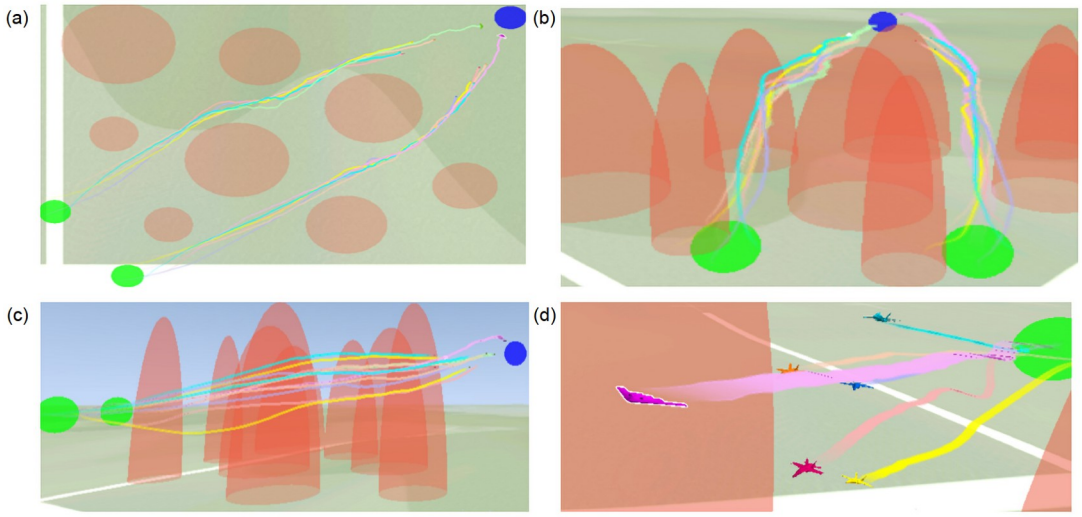


图 6　(网络版彩图)领导者-跟随者路径规划轨迹图. (a) 俯视图; (b) 主视图; (c) 侧视图; (d) 局部集群图

Figure 6　(Color online) Leader-follower path planning trajectory. (a) Top view; (b) main view; (c) side view; (d) local cluster diagram.

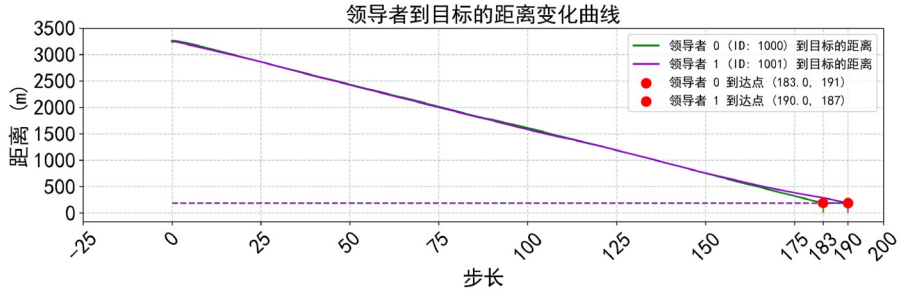


图 7　(网络版彩图)领导者L0与领导者L1距离目标点的距离变化曲线

Figure 7　(Color online) Distance variation curves between Leader L0 and Leader L1 and the target point.

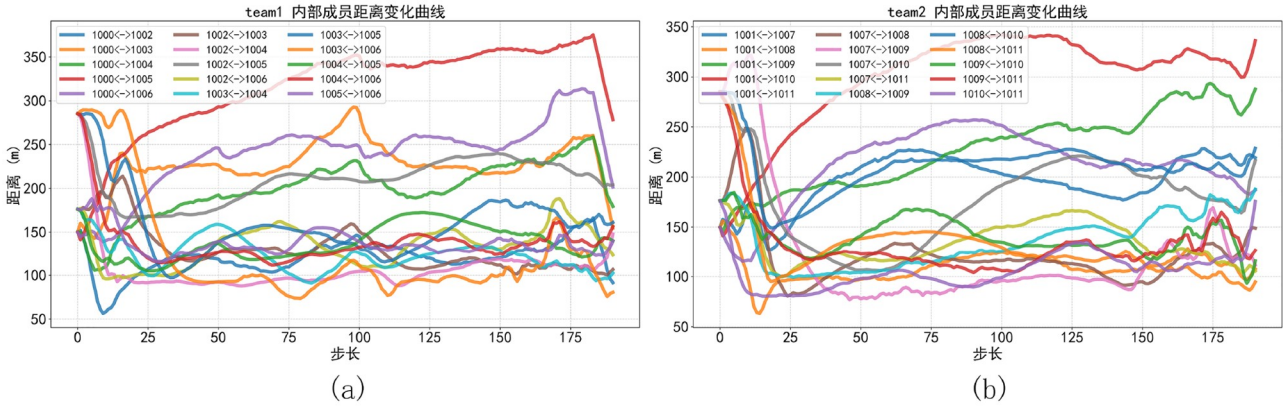


图 8　(网络版彩图)每个队伍的内部成员之间的距离变化曲线. (a) 队伍一; (b) 队伍二

Figure 8　(Color online) Distance variation curves between internal members of each team. (a) Team 1; (b) Team 2.

5 总结

本文针对复杂动态环境下多无人机协同路径规划任务中存在的感知信息处理扁平化、稀疏奖励下探索效率低下以及决策易陷入局部最优等关键问题, 提出了一种基于HCA-MAPPO的深度强化学习方法. 本文的主要工作总结如下.

(1) 在感知与架构层面, 提出了一种基于分层交叉注意力(HCA)的执行者网络. 该架构通过序贯的注意力计算流程, 实现了对障碍物、队友及领导者等多源异构信息的优先级感知与特征解耦, 显著增强了智能体在复杂态势下的环境理解能力.

(2) 在训练与探索层面, 设计了一种基于协同一致性门控的内在好奇心模块. 该模块创新性地将个体内在探索激励与集群协同目标进行动态耦合, 有效抑制了破坏编队协同的无效探索行为, 提升了算法在稀疏奖励环境下的收敛速度与稳定性.

(3) 在实验验证层面, 多场景仿真结果表明, 与MAPPO及其变体相比, 本文所提方法在任务成功率、路径平滑度及训练效率上均表现出显著优势. 消融实验进一步证实了上述各创新模块在提升决策鲁棒性和协同稳定性方面的有效性.

未来的研究将致力于解决深度强化学习模型在环境迁移和多任务场景下的泛化难题. 计划引入元强化学习(meta-reinforcement learning)机制, 赋予智能体“学会学习”的能力, 使其能够利用少量的交互经验, 快速适应从未见过的复杂地形特征或突变的任务约束, 从而进一步提升无人机集群在非平稳环境下的持续学习与自适应协同能力.

参考文献

1 Duan H B, Mei Y, Niu Y F, et al. Review of technological hotspots of unmanned aerial vehicle in 2024 (in Chinese). Sci Technol Rev, 2025, 43: 143–156 [段海滨, 梅宇, 牛轶峰, 等. 2024年无人机热点回眸. 科学导报, 2025 , 43: 143–156]

2 Shakhatreh H, Sawalmeh A H, Al-Fuqaha A, et al. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. IEEE Access, 2019, 7: 48572–48634

3 Hayat S, Yanmaz E, Muzaffar R. Survey on unmanned aerial vehicle networks for civil applications: A communications viewpoint. IEEE Commun Surv Tutor, 2016, 18: 2624–2661

4 Lv J H, Ran M P, Wang C C, et al. Manned/unmanned aerial vehicle intelligent cooperation: Opportunities and challenges (in Chinese). Sci Sin Tech, 2024, 54: 968–978 [吕金虎, 冉茂鹏, 王成才, 等. 有人/无人机智能协同: 机遇与挑战. 中国科学: 技术科学, 2024, 54: 968–978]

5 Kurunathan H, Huang H, Li K, et al. Machine learning-aided operations and communications of unmanned aerial vehicles: A contemporary survey. IEEE Commun Surv Tut, 2024, 26: 496–533

6 Niu Y F, Shen L C, Li J, et al. Key scientific problems in cooperation control of unmanned-manned aircraft systems (in Chinese). Sci Sin Inf, 2019, 49: 538–554 [牛轶峰, 沈林成, 李杰, 等. 无人-有人机协同控制关键问题. 中国科学: 信息科学, 2019, 49: 538–554]

7 Chen Y, Luo G, Mei Y, et al. UAV path planning using artificial potential field method updated by optimal control theory. Int J Syst Sci, 2016, 47: 1407–1420

8 Shi Y, Li Q, Bu S, et al. Research on intelligent vehicle path planning based on rapidly-exploring random tree. Math Probl Eng, 2020, 2020: 1–14

9 Alhijawi B, Awajan A. Genetic algorithms: Theory, genetic operators, solutions, and applications. Evol Intel, 2024, 17: 1245–1256

10 Wang X, Wang S, Liang X, et al. Deep reinforcement learning: A survey. IEEE Trans Neural Netw Learn Syst, 2024, 35: 5064–5078

11 Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep reinforcement learning: A brief survey. IEEE Signal Process Mag, 2017, 34: 26–38

12 Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search. Nature, 2016, 529: 484–489

13 Wang X, Gursoy M C, Erpek T, et al. Learning-based UAV path planning for data collection with integrated collision avoidance. IEEE Internet Things J, 2022, 9: 16663–16676

14 Zhang Y, Zhang Y, Yu Z. Path following control for UAV using deep reinforcement learning approach. Guid Navigat Control, 2021, 01: 2150005

15 Li S, Wu F, Luo S, et al. Dynamic online trajectory planning for a UAV-enabled data collection system. IEEE Trans Veh Technol, 2022, 71: 13332–13343

16 Yuan L, Zhang Z Q, Li L H, et al. Progress on cooperative multi-agent reinforcement learning in open environment (in Chinese). Sci Sin Inf, 2025, 55: 217–268 [袁雷, 张子谦, 李立和, 等. 开放环境下的协作多智能体强化学习进展. 中国科学: 信息科学, 2025, 55: 217–268]

17 Zhang S, Li Y, Dong Q. Autonomous navigation of UAV in multi-obstacle environments based on a deep reinforcement learning approach. Appl Soft Comput, 2022, 115: 108194

18 Xu Y. New explorations in autonomous low-altitude area defense decision-making and control methods of UAV swarms. Guid Navigat Control, 2025, 05: 285–289

19 Wang L, Wang K, Pan C, et al. Multi-agent deep reinforcement learning-based trajectory planning for multi-UAV assisted mobile edge computing. IEEE Trans Cogn Commun Netw, 2020, 7: 73–84

20 Wang Y, Cui Y, Yang Y, et al. Multi-UAV path planning for air-ground relay communication based on mix-greedy MAPPO algorithm. Drones, 2024, 8: 706

21 Shiri H, Seo H, Park J, et al. Attention-based communication and control for multi-UAV path planning. IEEE Wirel Commun Lett, 2022, 11: 1409–1413

22 Wang T, Zhang B, Zhang M, et al. Multi-UAV collaborative path planning method based on attention mechanism. Math Probl Eng, 2021, 2021: 1–8

23 Yin H, Su S, Lin Y, et al. Random network distillation based deep reinforcement learning for AGV path planning. In: Proceedings of the IEEE Intelligent Vehicles Symposium. Jeju Island, 2024. 2667–2673

24 Yang H, Shi D, Zhao C, et al. Ciexplore: Curiosity and influence-based exploration in multi-agent cooperative scenarios with sparse rewards. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York: Association for Computing Machinery, 2021. 2321–2330

25 Kang J J, Wu J H. Multiple UAVs formation control based on MIX-MAPPO (in Chinese). Appl Res Comput 2025, 42: 231–235 [康家境, 吴杰宏. 基于MIX-MAPPO算法的多无人机编队控制. 计算机应用研究, 2025, 42: 231–235]

A multi-UAV cooperative path planning method based on HCA-MAPPO deep reinforcement learning

LIN ZhenXi[1], GUO ZhengYu[2,3], ZHANG Jian[1,4], HAN Yu[2,3], CHEN CaiYi[1], CHEN HuiMin[4] & LUO DeLin[1]

[1] *School of Aerospace Engineering, Xiamen University, Xiamen 361102, China*
[2] *China Airborne Missile Academy, Luoyang 471000, China*
[3] *National Key Laboratory of Air-based Information Perception and Fusion, Luoyang 471000, China*
[4] *School of Aeronautics, Changji University, Changji 831100, China*

To address the challenges of multi-UAV cooperative path planning, such as the flattening of perception information processing and low exploration efficiency under sparse rewards, this paper proposes a novel cooperative path planning method based on multi-agent deep reinforcement learning. This method models the multi-UAV cooperative path-planning problem as a partially observable Markov decision process and employs a deep reinforcement learning algorithm for solution. To enhance the agent's ability to extract features and assess situational awareness from multi-source heterogeneous observation information, a hierarchical cross-attention-based agent network is designed, which achieves priority perception and processing of multi-source heterogeneous information through a sequential processing flow. Furthermore, an intrinsic curiosity module based on cooperative consistency gating is constructed to effectively couple individual exploration with cluster cooperative goals, improving the algorithm's effective learning efficiency under sparse reward environments. Finally, simulation experiments verify the effectiveness and superiority of the proposed algorithm.

UAV cooperation, path planning, multi-agent deep reinforcement learning, hierarchical cross-attention, curiosity mechanism

doi: 10.1360/SST-2025-0385