基于因果影响检测的多无人机海上协同导航策略优化方法

柳文章 [1, 2, 3] 陆建华 [1, 2, 3] 任 璐 [1, 2, 3] 孙长银 [1, 2, 3]

摘 要 多无人机协同导航是实现高效海上协同作业的重要技术. 然而, 在广阔且动态未知的海域中, 受限的感知能力与自主决策机制使无人机之间的协作关系复杂, 难以获取全局信息. 近年来, 基于集中训练与分散执行范式的多智能体强化学习在协作行为学习方面取得显著进展, 并被广泛应用于海上协同导航任务. 但由于智能体交互往往仅在特定情境下发生, 如何有效提升协作效率与探索能力仍是关键挑战. 为解决上述问题, 提出一种基于因果影响检测的多智能体近端策略优化方法. 该方法以智能体之间的因果影响为衡量准则, 引入基于协作规则设计的内在奖励机制, 利用因果推断与条件互信息来检测智能体之间在行为上的因果影响, 从而引导其优先探索对全局状态具有正向影响的动作, 强化多智能体间的合作. 实验结果表明, 所提方法表现出显著的性能提升, 尤其在海上搜救任务中展现出更高的协同效率, 验证了方法的有效性.

关键词 多无人机协同; 近端策略优化; 多智能体强化学习; 因果影响检测

引用格式 柳文章, 陆建华, 任璐, 孙长银. 基于因果影响检测的多无人机海上协同导航策略优化方法. 自动化学报, 2026, **52**(5): 1069−1082

DOI 10.16383/j.aas.c250691 **CSTR** 32138.14.j.aas.c250691

Policy Optimization Method for Multi-UAV Cooperative Maritime Navigation Based on Causal Influence Detection

LIU Wen-Zhang[1, 2, 3] LU Jian-Hua[1, 2, 3] REN Lu[1, 2, 3] SUN Chang-Yin[1, 2, 3]

Abstract Multi-UAV cooperative navigation is a crucial technology for achieving efficient cooperative maritime operations. However, in vast and dynamically unknown maritime environments, limited sensing capabilities and autonomous decision-making mechanism lead to complex cooperation relationships among UAVs, making it difficult to obtain global information. In recent years, multi-agent reinforcement learning under the centralized training and decentralized execution paradigm has achieved remarkable progress in learning cooperative behaviors and has been widely applied to cooperative maritime navigation tasks. Nevertheless, because agent interactions often occur only in specific situations, improving cooperation efficiency and exploration capability remains a major challenge. To address this issue, this paper proposes a causal influence detection for multi-agent proximal policy optimization method. The proposed method uses causal influence among agents as an evaluation metric and introduces an intrinsic reward mechanism designed based on cooperation rules. By leveraging causal inference and conditional mutual information, the method detects behavioral causal influence among agents, guiding them to preferentially explore actions that positively affect the global state and thus enhancing inter-agent cooperation. Experimental results demonstrate that the proposed method achieves significant performance improvements, especially in maritime search and rescue tasks, where it exhibits higher cooperation efficiency, validating the effectiveness of the method.

Keywords multi-UAV cooperation; proximal policy optimization; multi-agent reinforcement learning; causal influence detection

Citation Liu Wen-Zhang, Lu Jian-Hua, Ren Lu, Sun Chang-Yin. Policy optimization method for multi-UAV cooperative maritime navigation based on causal influence detection. *Acta Automatica Sinica*, 2026, **52**(5): 1069−1082

无人机 (unmanned aerial vehicle, UAV) 通常指通过无线电遥控或自主控制系统操纵的无人飞行器, 因其视野广、机动性强、部署灵活等优势, 已在海上巡逻、海洋环境与水质监测、海上搜救以及海上资源探测等典型海洋任务中发挥着重要作用[1−4]. 然而, 单架无人机在续航能力、载荷规模与任务多

收稿日期 2025-11-30 录用日期 2026-03-16
Manuscript received November 30, 2025; accepted March 16, 2026
国家自然科学基金 (62303009, 62495083, 62236002), 安徽省教育厅高校科研项目青年项目 (2025AHGXZK40374) 资助
Supported by National Natural Science Foundation of China (62303009, 62495083, 62236002) and Youth Research Project of Anhui Provincial Department of Education (2025AHGXZK 40374)
本文责任编委 刘志杰
Recommended by Associate Editor LIU Zhi-Jie
1. 安徽大学人工智能学院 合肥 230601 2. 自主无人系统技术教育部工程研究中心 合肥 230601 3. 安徽省安全人工智能重点实验室 合肥 230601
1. School of Artificial Intelligence, Anhui University, Hefei 230601 2. Ministry of Education Engineering Research Center for Autonomous Unmanned Systems Technology, Hefei 230601 3. Anhui Provincial Key Laboratory of Security Artificial Intelligence, Hefei 230601

样性方面存在先天限制, 难以在复杂海域中独立完成长时间、高强度、多目标的作业需求. 因此, 作为群体智能的重要形式, 多无人机 (multi-UAV) 协同被视为提升作业效率与系统鲁棒性的有效途径, 而协同导航则是多无人机协同执行海上任务的关键技术之一. 近年来, 多无人机协同导航的控制方法与路径规划策略已成为研究热点, 并在理论与应用层面得到广泛探讨[5–10]. 在海上搜救这一时效性强、环境复杂且风险较高的典型应用场景中, 如何通过无人机间的协同配合以最快速度、最低成本覆盖目标区域并实现高概率目标发现, 成为亟待解决的关键科学与工程问题.

受自然界群体行为启发的仿生群体智能方法[11–13], 由于能够通过简单的局部规则涌现出复杂而有序的整体协同行为, 在多无人机协同任务中展现出独特优势. 然而, 在真实海域中, 海况变化、目标分布不确定性以及传感噪声等因素使得系统策略必须具备面对未知扰动和任务变化时的快速适应与在线调整能力, 这一需求超出传统基于规则或优化方法在参数设定、泛化能力与鲁棒性方面的能力边界. 为应对上述挑战, 多智能体强化学习 (multi-agent reinforcement learning, MARL)[14–16] 通过交互式试错学习端到端的协同策略, 能够处理部分可观测下多智能体协作系统中的关键问题, 将仿生群体智能的启发式优势与数据驱动的学习能力有效结合起来. 因此, MARL 为实现海上多无人机任务中的高效、鲁棒与可泛化的协同导航提供一条可行的智能化路径.

在算法设计与工程实现方面, 许多经典的 MARL 算法被提出, 如值分解网络 (value decomposition networks, VDN)[17]、Q 混合网络 (Q mixing network, QMIX)[18]、多智能体深度确定性策略梯度 (multi-agent deep deterministic policy gradient, MADDPG)[19]、多智能体近端策略优化 (multi-agent proximal policy optimization, MAPPO)[20]、反事实多智能体策略梯度 (counterfactual multi-agent policy gradient, COMA)[21] 等. 随着无人机平台算力与传感能力提升, MARL 被逐步应用于海上搜救等复杂任务场景. 例如, Rashid 等[22] 将强化学习算法用于无人机与水面无人艇组成的联合网络中以实现协同搜救和路径规划. Lei 等[23] 基于 IPPO (independent proximal policy optimization) 算法, 针对救援无人机、路由无人机以及救援舰艇的轨迹规划与资源调度问题, 提出一种可扩展的 MARL 算法, 实现异构系统的自适应调节. Lei 等[24] 还针对无人机与水面设备之间的通信问题, 提出一种基于 MAPPO 的异构车辆多智能体近端策略优化算法. 同样基于 MAPPO, Wu 等[25] 结合基于种群的学习机制与高斯混合模型 (Gaussian mixture model, GMM) 使具有不同探索偏好的智能体能够发掘无人机间的最优协作模式. 文献 [26] 基于深度 Q 网络 (deep Q-network) 提出一种新的网络结构, 实现多无人机之间的最优飞行轨迹规划. 针对未知区域的多无人机协作, Hou 等[27] 提出基于 MADDPG 的分布式协同搜索并利用卷积神经网络 (convolutional neural networks, CNN) 处理高维地图数据, 实现对未知区域的协同搜索, 有效避免了碰撞与重复搜索.

这些基于集中训练与分散执行 (centralized training and decentralized execution, CTDE) 框架下的 MARL 算法通过智能体与环境的持续交互, 收集经验数据从而学习出接近最优的协同策略. 然而, 在完全协作的多智能体系统中, 由于奖励分配机制往往依赖全局奖励信号, 个体难以准确评估自身行为对整体任务的贡献, 导致一些智能体没有做出对团队奖励提升有用的行为却仍能获得和其他智能体相同的团队奖励, 这种现象称为懒惰智能体 (lazy agent)[17, 28]. 此外, 许多 MARL 算法假设智能体策略相互独立, 从而在优化过程中忽视了其他智能体的决策影响, 这种独立的策略设计进一步限制了智能体间协同行为的学习. 因此, 考虑智能体之间行为的相互影响成为当前 MARL 领域研究的重要方向. 文献 [29–30] 为缓解环境的非平稳性问题并实现智能体之间的协作, 在智能体决策时考虑自身行为对其他智能体行为的影响进而提升了算法的性能. 文献 [31–33] 提出通过互信息 (mutual information, MI) 来量化智能体行为之间的相关性, 以最大化行为相关性从而增强协作. 然而, 在这些方法中如何有效协调多智能体的同步动作仍然是一个具有挑战性的问题.

针对以上问题, 本文从因果推断的角度出发考虑智能体之间的因果影响, 提出一种基于因果影响检测的多智能体近端策略优化 (causal influence detection for multi-agent proximal policy optimization, CID-MAPPO) 方法. 该方法通过使用条件互信息 (conditional mutual information, CMI) 方法计算智能体与团队之间的因果影响因子, 并将其纳入内在奖励机制. 这种内在奖励机制能够有效促进关键动作的识别, 从而增强智能体之间的协作能力. 需要指出的是, 现有因果驱动的强化学习方法可大致分为两类: 一是以文献 [34] 为代表的单智能体因果影响检测方法, 通过衡量动作对状态特征的因果效应加速探索, 但其未涉及多智能体间的联合动作反事实分析, 难以直接应用于多智能体场景;

二是以文献 [30] 为代表的多智能体因果影响方法, 侧重于检测智能体动作对其他智能体行为的因果影响, 并引入情境门控机制实现选择性协作. 与上述方法不同, 本文的因果影响度量聚焦于个体动作对全局状态转移分布的边际干预效应, 采用高斯与混合高斯分布间的 KL (Kullback-Leibler) 散度进行分布层面的量化, 从而为信用分配提供更直接的依据. 本文从状态转移分布层面刻画个体动作对系统演化的结构性影响. 该视角强调因果路径而非统计相关性, 从而为协作任务 MARL 中的信用分配问题提供一种分布级建模方式. 通过在模拟多无人机协同导航任务中的实验验证, 本文所提方法与基准算法相比, 有效提升了多无人机协同导航任务的性能.

1　背景介绍

1.1　马尔科夫决策过程

本文考虑将完全合作的多无人机协同导航任务建模为去中心化的部分观测马尔科夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP). 该过程定义为元组 $\langle \mathcal{I},\ \mathcal{S},\ \{\mathcal{O}_i\}_{i=1}^{N},\ \{\mathcal{A}_i\}_{i=1}^{N},\ R,\ P,\ \gamma\rangle$. 其中, $\mathcal{I}=\{1,\ 2,\ \cdots,\ N\}$ 表示无人机集合, $\mathcal{S}$ 是状态空间, $\mathcal{O}_i$ 是无人机 $i\in\mathcal{I}$ 的局部观测空间, $\mathcal{A}_i$ 是无人机 $i\in\mathcal{I}$ 的动作空间, $R:\mathcal{S}\times\mathcal{A}_1\times\cdots\times\mathcal{A}_N\times\mathcal{S}\to\mathbf{R}$ 表示奖励函数, $P:\mathcal{S}\times\mathcal{A}_1\times\cdots\times\mathcal{A}_N\times\mathcal{S}\to[0,\ 1]$ 表示系统状态转移概率, $\gamma\in(0,\ 1)$ 是折扣因子. 本文分别用 $s_t\in\mathcal{S}$、$a_t^i\in\mathcal{A}_i$ 表示在时刻 t 下系统的全局状态和无人机 i 在时刻 t 所执行的动作, 并且用 $\boldsymbol{a}_t=[a_t^1,\ \cdots,\ a_t^N]$ 表示 N 架无人机联合执行的动作组合, 那么 $R(s_t,\ \boldsymbol{a}_t,\ s_{t+1})$ 则表示多无人机在系统状态 s_t 下执行联合动作 $\boldsymbol{a}_t$ 后状态转移至 s_{t+1} 所获得的奖励, 奖励值用 r_{t+1} 表示. $P(s_{t+1}|s_t,\ \boldsymbol{a}_t)$ 表示多无人机在系统状态 s_t 下执行联合动作 $\boldsymbol{a}_t$ 后系统状态转移至 s_{t+1} 的概率. 用 $o_t^i\in\mathcal{O}_i$ 表示无人机 i 在时刻 t 下的局部观测值, 则 $\pi^i(o_t^i)$ 表示无人机 i 的策略. 在本文, 考虑离散动作空间的情况, 用 $a_t^i\sim\pi^i(o_t^i)$ 表示无人机 i 基于观测 o_t^i 和策略 π^i 采样一个动作 a_t^i, $\pi^i(a_t^i|o_t^i)$ 则表示其采取动作 a_t^i 的概率. 在完全合作的 MARL 设置下, 多无人机系统的目标是学习一个联合策略 $\boldsymbol{\pi}^*=\{\pi^{1*},\ \cdots,\ \pi^{N*}\}$ 来最大化期望折扣回报 $\mathrm{E}_{\boldsymbol{\pi}}[\sum_{t=0}^{\infty}\gamma^t r_{t+1}]$.

1.2　因果图模型

为刻画系统状态转移过程的因果结构, 本文采用因果图模型 (causal graphical model, CGM) 进行建模. CGM 通常用有向无环图 (directed acyclic graph, DAG) 来表示, 其中节点对应随机变量, 有向边表示变量之间的直接因果关系. 在结构因果模型 (structural causal model, SCM) 框架下, DAG 与一组结构方程共同定义变量的生成机制. 具体地, 考虑随机变量集合 $V=\{v_1,\ v_2,\ \cdots,\ v_n\}$ 及条件概率分布 $P(v_i\mid \mathrm{Pa}(v_i))$, $\forall v_i\in V$, 其中 $\mathrm{Pa}(v_i)$ 是节点 $v_i\in V$ 的父节点集合. 在 MDP 假设下, CGM 可用于对状态转移分布进行因果分解, 即联合概率分布 $P(V)$ 可表示为所有节点条件概率乘积[34]:

$$P(v_1,\ v_2,\ \cdots,\ v_{|V|})=\prod_{i=1}^{|V|}P\left(v_i\mid \mathrm{Pa}(v_i)\right)\tag{1}$$

其中, $|V|$ 表示集合 V 中元素的个数. 从直观角度来看, 引入因果图模型的目的在于刻画系统状态转移过程中变量之间的影响, 而不仅仅是统计相关性. 在多智能体系统中, 某一变量 (例如某个智能体的动作) 与其他状态变量之间可能同时受到环境或其他智能体行为的共同影响, 仅依赖联合概率分布难以区分其真实作用路径. 因果图通过有向边显式刻画“谁对谁产生影响”, 使得在固定其他条件不变的情况下, 能够分析单个变量变化对系统演化所产生的因果效应.

在本文中, 基于 CGM 对状态转移过程的结构化分解, 使得后续可以在该因果结构上进一步分析不同智能体动作对整体系统状态变化的影响程度, 为多智能体因果影响检测提供结构基础.

1.3　多无人机协同导航

1.3.1　问题描述

如图 1 所示, 本研究的任务场景如下: N 架无人机随机分布在包含障碍物且具有 M 个目标地点的区域内, 目标地点的初始位置随机. 本文的目的是设计出一种高效的协同导航策略, 使多架无人机在保证不发生碰撞的前提下, 通过协作以最短路径分别抵达各目标地点. 为简化问题从而聚焦算法研

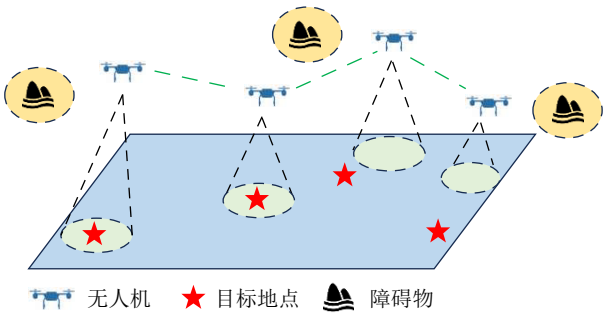


图 1　多无人机协同导航示意图

Fig. 1　Diagram of multi-UAV cooperative navigation

究, 本文仅讨论无人机数量等于目标地点数量的情形, 即 $N = M$. 需要指出, 本文并未假设单架无人机定位能力不足. 协同机制的引入主要源于多目标搜索任务的空间分散性与信息不完全性. 在大范围未知环境中, 多无人机协同能够提高搜索覆盖效率、减少重复探索, 并通过协调机制避免冲突与资源浪费, 从而显著提升整体任务完成效率.

1.3.2 动力学模型

无人机的动力学模型 (dynamics model) 通常由机体固定坐标系和惯性坐标系组成, 如图 2 所示. 机体坐标系 (用下标 b 表示), 其中 O_b 表示原点, 为无人机的质心; 坐标轴 x_b 轴指向无人机的飞行方向, y_b 轴表示与飞行方向垂直的无人机的右方向, z_b 轴对应无人机的下方. 惯性坐标系 (用下标 e 表示) 是相对地球而言, 默认为北东地坐标系, O_e 表示原点, 坐标轴 x_e、y_e、z_e 分别指向北向、东向及地心方向. 在此基础上, 无人机的运动学模型表示为

$$\begin{bmatrix}\dot{x}_e\\ \dot{y}_e\\ \dot{z}_e\end{bmatrix} = \begin{bmatrix} c_{\theta_y}c_{\theta_z} & s_{\theta_x}s_{\theta_y}c_{\theta_z}-c_{\theta_x}s_{\theta_z} & c_{\theta_x}s_{\theta_y}c_{\theta_z}+s_{\theta_x}s_{\theta_z}\\ c_{\theta_y}s_{\theta_z} & s_{\theta_x}s_{\theta_y}s_{\theta_z}+c_{\theta_x}c_{\theta_z} & c_{\theta_x}s_{\theta_y}s_{\theta_z}-s_{\theta_x}c_{\theta_z}\\ -s_{\theta_y} & s_{\theta_x}c_{\theta_y} & c_{\theta_x}c_{\theta_y}\end{bmatrix}\begin{bmatrix}v_x\\ v_y\\ v_z\end{bmatrix} \tag{2}$$

$$\begin{bmatrix}\dot{\theta}_x\\ \dot{\theta}_y\\ \dot{\theta}_z\end{bmatrix} = \begin{bmatrix}1 & s_{\theta_x}\tan\theta_y & c_{\theta_x}\tan\theta_y\\ 0 & c_{\theta_x} & -s_{\theta_x}\\ 0 & s_{\theta_x}\sec\theta_y & c_{\theta_x}\sec\theta_y\end{bmatrix}\begin{bmatrix}\omega_x\\ \omega_y\\ \omega_z\end{bmatrix} \tag{3}$$

其中, c_{θ_x}、s_{θ_x} 分别表示 $\cos\theta_x$、$\sin\theta_x$ 的缩写, 其他姿态角也是如此; v_x、v_y、v_z 为无人机在机体坐标系 x_b、y_b、z_b 三个方向的速度分量; θ_x、θ_y、θ_z 分别代表无人机的三轴姿态角, 即滚转角 (roll)、俯仰角 (pitch) 和偏航角 (yaw); ω_x、ω_y、ω_z 为三轴的姿态角速度, 其中姿态角及姿态角速度的方向如图 2 所示.

1.4 符号说明

本文对主要变量与记号作统一说明, 涉及的主要符号约定如表 1 所示.

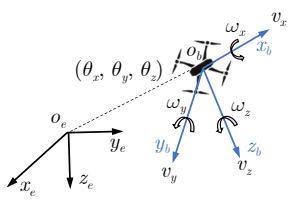


图 2 无人机的机体固定坐标系和惯性坐标系

Fig. 2 The body-fixed coordinate system and inertial coordinate system of UAV

表 1 数学符号说明

Table 1 Explanation of mathematical symbols

数学符号	符号说明
$\mathcal{I} = \{1, 2, \cdots, N\}$	无人机集合
s_t	时刻 t 的系统状态
o_t^i	无人机 i 在时刻 t 的局部观测
a_t^i	无人机 i 在时刻 t 的动作
π^i	无人机 i 的策略
$\boldsymbol{a}_t = [a_t^1, \cdots, a_t^N]$	联合动作
$\boldsymbol{a}_t \backslash a_t^i$	去除无人机 i 动作后的联合动作
$P(s_{t+1}\|s_t, \boldsymbol{a}_t)$	状态转移概率
γ	折扣因子
$C^i(s_t, s_{t+1})$	无人机 i 的因果影响度量
$D_{\mathrm{KL}}(\cdot\|\cdot)$	KL 散度
β	β-VAE 正则系数
r_t^i	无人机 i 在时刻 t 的奖励
$r_{\mathrm{cid},\,t}^i$	无人机 i 在时刻 t 的因果内在奖励
λ	内在奖励权重

2 方法

为解决局部最优以及信用分配等问题, 本文介绍了 CID-MAPPO 这一新的 MARL 算法, 以提升传统 MARL 算法在完成多无人机协同导航任务时的性能. 在该方法中, 无人机通过最大化自身动作的因果影响, 同时学习策略和内在奖励函数. 通过预训练一个动力学模型, 基于因果干预和 CMI 检测无人机之间的因果影响, 进而更好地指导无人机探索能够影响全局状态的策略, 促进无人机之间的协作.

2.1 多智能体因果图模型

将 CGM 推广到 CTDE 框架下多智能体协作任务的情形, 得到多智能体因果图模型 (multi-agent causal graph model, MACGM), 用 $\mathcal{G}$ 来表示. $\mathcal{G}$ 是由一组随机变量 $V = \{s_t, a_t^1, a_t^2, \cdots, a_t^N, s_{t+1}, r_{t+1}\}$ 以及条件概率分布 $P(v_i|\mathrm{Pa}(v_i)), \forall v_i \in V$ 构成的一个有向无环图. 以包含 3 个智能体的多无人机系统举例, 假设不同智能体的动作相互独立, 其单步状态转移过程可通过图 3 进行描述. 当状态以概率 $P(s_{t+1}|s_t, \boldsymbol{a}_t)$ 从 s_t 转移至 s_{t+1} 时, 各智能体将获得一个共享奖励 r_{t+1}. 在该模型中, 当前状态 s_t 以及各智能体的动作 a_t^1、a_t^2、a_t^3 是下一状态 s_{t+1} 和奖励 r_{t+1} 的成因.

在结合 CGM 的 MARL 中, 本文研究更关注

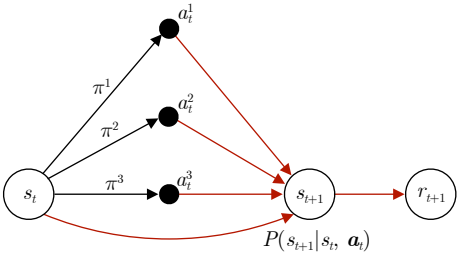


图 3　具有 3 个智能体的单步状态转移因果图模型

Fig. 3　The causal graphic model of one-step state transition with three agents

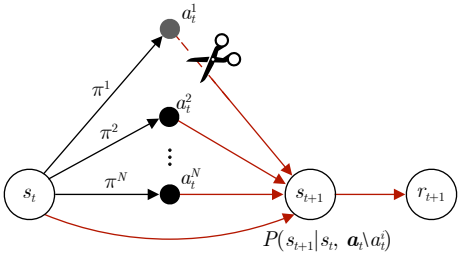


图 4　不考虑智能体 i 动作的因果图模型

Fig. 4　The causal graphic model without considering the action of agent i

的是对应于动作或环境的干预, 这些干预可能会改变状态 s_{t+1} 的分布. 在形式上, 定义 $do(I)$ 为对应事件 I 的一次干预. 例如, $do(a_t^i = \pi^j(o_t^i))$ 表示对变量 a_t^i 的干预, 将智能体 i 的策略替换为 π^j. 因此, 通过比较分布 $P(s_{t+1}|\mathrm{Pa}(s_{t+1}))$ 与 $P(s_{t+1}|\mathrm{Pa}(s_{t+1}), do(I))$ 可以衡量由干预 $do(a_t^i = \pi^j(o_t^i))$ 引起的影响.

在一些将 CGM 用于 MARL 的情形中, 下一状态的节点并不一定与所有智能体的动作完全相连. 换言之, 某些智能体的动作对下一状态分布没有影响. 因此, 本文给出因果极小性的定义[34].

定义 1 (因果极小性). 设 $\mathcal{G}$ 为描述 MARL 单步状态转移的 CGM, 节点集合为 V. 若对任意 v_i、$v_j \in V$ 且 $v_i \neq v_j$, 当 v_j 在给定 $\mathrm{Pa}(v_j)$ 的条件下仍然条件依赖于 v_i 时, 称图模型 $\mathcal{G}$ 满足因果极小性.

因果极小性的定义有助于识别 CGM 中的噪声变量. 例如, 若 MARL 系统中存在“懒惰智能体”, 则这些智能体的动作对下一步状态转移或团队回报没有任何影响. 因此, 包含这些无效动作的 CGM 不满足因果极小性.

基于上述因果极小性定义, 视角自然转向干预的另一面——反事实推断: 即在给定已观测到的历史 (状态、动作、噪声) 条件下, 若某个智能体未采取动作, 下一步状态或团队回报会如何变化. 这种问法正是反事实推断在因果推断中最直接的应用, 也是判断智能体在完成协作任务中是否存在信用分配问题的关键. 同样以图 3 为例, 通过假设在智能体 i 未施加干预的情况下, 仅由其余智能体执行动作, 环境的状态演化以及最终的奖励可能发生的变化如图 4 所示. 通过比较两种情况下系统状态分布和奖励分布的变化, 计算出智能体 i 的行为对团队的贡献程度.

2.2　预训练环境模型

针对多无人机系统, 为计算无人机 i 的动作对全局状态的因果影响, 需要刻画系统在当前状态下如何在外界扰动或联合动作条件下转移至下一状态. 基于反事实分析思想, 本文首先研究状态转移分布 $P(s_{t+1}|s_t,\ \boldsymbol{a}_t)$. 本文采用 β- 变分自编码器 (β-variational autoencoder, β- VAE)[35] 框架来建模状态转移过程. 通过在 β- VAE 中引入时间依赖的潜在变量, 近似刻画潜在变量随时间演化的后验分布, 同时, 将重参数化噪声与联合动作相结合, 使模型能够学习动作对状态变化的影响规律, 如图 5 所示.

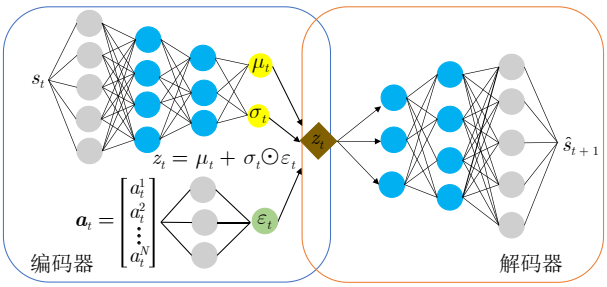


图 5　β- VAE 框架示意图

Fig. 5　Diagram of β- VAE framework

编码器部分通过两个多层感知机 (multi-layer perceptron, MLP) 将输入的状态 s_t 和动作 $\boldsymbol{a}_t$ 分别编码为一个状态分布 $(\mu_t,\ \sigma_t)$ 和对应的编码 ε_t. 编码器的这一约束使得通过添加噪声生成新数据成为可能. 在本研究中, β- VAE 的目标是预测下一状态, 因此本文将其中的噪声项替换为联合动作. 随后, 潜在向量 z_t 按以下公式计算:

$$z_t = \mu_t + \sigma_t \odot \varepsilon_t \tag{4}$$

解码器则同样是由一个 MLP 用于预测下一时刻状态 $\hat{s}_{t+1}$. β- VAE 在无人机学习策略的同时更新自身参数, 为最小化下一时刻真实状态 s_{t+1} 与预测状态 $\hat{s}_{t+1}$ 之间的差异, 其损失函数设计如下:

$$\mathcal{L}(\theta) = \Bigg[\frac{(s_{t+1} - \mu_\theta(s_t,\ \boldsymbol{a}_t))^2}{2\sigma_\theta^2(s_t,\ \boldsymbol{a}_t)} + \frac{1}{2}\log\left(2\pi\sigma_\theta^2(s_t,\ \boldsymbol{a}_t)\right) + \beta\, D_{\mathrm{KL}}\big(\mathrm{N}(\mu_\theta(s_t,\ \boldsymbol{a}_t),\ \sigma_\theta^2(s_t,\ \boldsymbol{a}_t))\|\mathrm{N}(0,\ 1)\big)\Bigg] \tag{5}$$

其中, β 为超参数且 $\beta \geq 0$; $\mu_\theta(s_t,\ \boldsymbol{a}_t)$、$\sigma_\theta(s_t,\ \boldsymbol{a}_t)$ 表示编码器输出的均值和方差. 前两项采用似然损失来表示编码–解码过程中的重建损失, 最后一项则通过计算 KL 散度来衡量潜在变量的后验分布与先验分布之间的差异, 确保潜在空间的结构是均匀和连续的, 这种约束可以帮助模型在学习过程中避免不规则或高度不规范的潜在空间结构. 在实际使用中, 本文将编码器部分作为独立模块用于预测环境下一时刻状态分布, 其在整体框架中的功能如图 6 中间部分所示.

2.3 因果影响检测

本文的目标是通过计算无人机 i 的动作对系统全局状态的因果影响来衡量该无人机对团队任务的贡献度, 而 CMI 是衡量这种因果效应的常用方法. 因此, 本研究采用 CMI 作为度量无人机因果影响的工具. 当该变量大于 0 时, 表明 a_t^i 对预测 s_{t+1} 是必要的, 即存在因果路径 $(a_t^i \to s_{t+1})$, 计算公式如下:

$$C^i(s_t,\ s_{t+1}) := \mathrm{I}(a_t^i; s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i) = \mathrm{E}_{a_t^i\sim\pi^i}\left[D_{\mathrm{KL}}(P(s_{t+1}|s_t,\ \boldsymbol{a}_t)\|P(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i))\right] \tag{6}$$

其中, $\boldsymbol{a}_t\backslash a_t^i$ 表示去除无人机 i 动作后的联合动作.

在进行反事实分析时, 为了计算不考虑无人机 i 动作的状态转移分布, 边际概率可以写为:

$$P(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a^i) = \sum_{a\sim\mathcal{A}_i}\pi^i(a|s_t)P\big(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i,\ a\big) \tag{7}$$

对于连续动作空间和高斯策略, 式 (7) 中的加权求和可以通过蒙特卡洛 (Monte Carlo, MC) 采样来估计:

$$P\big(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i\big) \approx \frac{1}{K}\sum_{k=1}^{K}P\big(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i,\ a^{(k)}\big) \tag{8}$$

其中, K 是从分布 $\pi^i(\cdot|s_t)$ 采样的动作数量; $a^{(k)}$ 为第 k 个采样动作. 因此, 给定分布 $P(s_{t+1}|s_t,\ \boldsymbol{a}_t)$ 与分布 $P(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i)$, 无人机 i 在时间步 t 的因果影响可通过以下公式度量:

$$C^i(s_t,\ s_{t+1}) = D_{\mathrm{KL}}\big(P(s_{t+1}|s_t,\ \boldsymbol{a}_t)||P(s_{t+1}|s_t,\ \boldsymbol{a}_t\backslash a_t^i)\big) \tag{9}$$

其中, 由于 $P(s_{t+1}|s_t,\ \boldsymbol{a}_t)$ 是高斯分布, 而式 (7) 和 (8) 中的加权求和构成混合高斯模型, 本文给出以下定理来近似计算高斯分布与高斯混合分布之间的 KL 散度.

定理 1. 给定一个高斯分布 $f \sim \mathrm{N}(\mu,\ \Sigma)$ 和一个高斯混合分布 $g \sim \sum_{k=1}^{K} w_k\mathrm{N}(\mu_k,\ \Sigma_k)$, 两者之间的 KL 散度可以近似计算为:

$$D_{\mathrm{KL}}(f\|g) \approx -\frac{1}{2}\log\left(\sum_{k=1}^{K}w_k\exp\big(\mathrm{E}_{s\sim f}[\log g_k(s)]\big)\right) - \frac{1}{2}H(f) - \frac{1}{2}\log\left(\sum_{k=1}^{K}w_k\mathrm{e}^{-D_{\mathrm{KL}}(f\|g_k)}\right) \tag{10}$$

其中, $g_k \sim \mathrm{N}(\mu_k,\ \Sigma_k)$; w_k 为超参数权重且 $\sum_{k=1}^{K} w_k = 1$; $H(f) = -\mathrm{E}_{s\sim f}[\log(f(s))]$.

证明. 根据 KL 散度的定义, 有:

$$D_{\mathrm{KL}}(f\|g) = \int_s f(s)\log\frac{f(s)}{g(s)}\mathrm{d}s = H(f,\ g) - H(f) \tag{11}$$

其中,

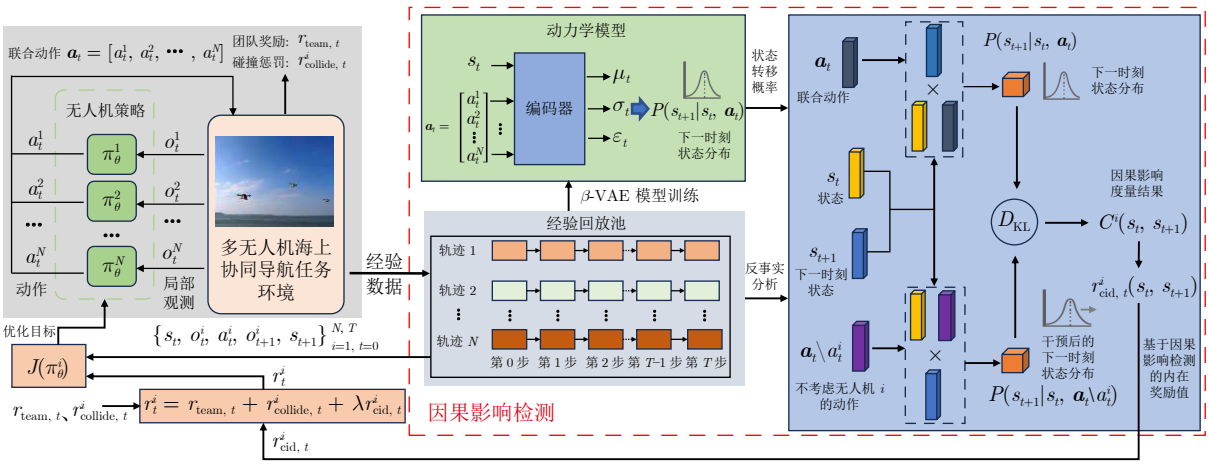


图 6 CID-MAPPO 示意图

Fig. 6 Diagram of CID-MAPPO

$$H(f,\ g) = -\mathrm{E}_{s\sim f}\left[\log g(s)\right] \tag{12}$$

$$H(f) = -\mathrm{E}_{s\sim f}\left[\log f(s)\right] = \frac{1}{2}\log\det\left(2\pi\mathrm{e}\Sigma\right) \tag{13}$$

其中, $\det(2\pi\mathrm{e}\,\Sigma) = (2\pi\mathrm{e})^d \det\Sigma$, d 表示随机变量 s 的维度.

虽然目前还没有 $H(f,\ g)$ 的解析表达式, 但在工程上可以通过对其上下界取平均来近似估计.

首先, 将式 (12) 展开如下:

$$H(f,\ g) = -\mathrm{E}_{s\sim f}\left[\log\sum_{k=1}^{K} w_k g_k(s)\right] \tag{14}$$

下界. 根据琴生不等式 (Jensen′s inequality), 有

$$H(f,\ g) \geq -\log\mathrm{E}_{s\sim f}\left[\sum_{k=1}^{K} w_k g_k(s)\right] = -\log\sum_{k=1}^{K} w_k \mathrm{E}_{s\sim f}\left[g_k(s)\right] \tag{15}$$

这样便得到 $H(f,\ g)$ 的下界

$$H(f,\ g)_{-} = -\log\sum_{k=1}^{K} w_k \mathrm{E}_{s\sim f}[g_k(s)] \tag{16}$$

在式 (16) 中, $\mathrm{E}_{s\sim f}[g_k(s)]$ 可由下式计算得到:

$$\mathrm{E}_{s\sim f}[g_k(s)] = \left((2\pi)^d|\Sigma_k'|\mathrm{e}^{\Delta\mu_k^{\mathrm{T}}\Sigma_k'^{-1}\Delta\mu_k}\right)^{-\frac{1}{2}} \tag{17}$$

其中, $\Delta\mu_k = \mu_k - \mu$; $\Sigma_k' = \Sigma + \Sigma_k$ 并且 $|\Sigma_k'| = \det\Sigma_k'$.

上界. 为了计算 $H(f,\ g)$ 的上界, 可以采用权重集合 $\{\phi_k|0<\phi_k\leq 1\}_{k=1}^{K}$ 重写式 (14) 如下:

$$H(f,\ g) = -\mathrm{E}_{s\sim f}\left[\log\sum_{k=1}^{K}\phi_k\frac{w_k}{\phi_k}g_k(s)\right] \tag{18}$$

其中, $\sum_{k=1}^{K}\phi_k = 1$.

然后, 反向使用琴生不等式, 得到

$$\begin{aligned} H(f,\ g) &\leq -\mathrm{E}_{s\sim f}\left[\sum_{k=1}^{K}\phi_k\log\left(\frac{w_k}{\phi_k}g_k(s)\right)\right] = \\ &\sum_{k=1}^{K}\phi_k\mathrm{E}_{s\sim f}\left[\log g_k(s) + \log\frac{w_k}{\phi_k}\right] = \\ &-\sum_{k=1}^{K}\phi_k\mathrm{E}_{s\sim f}\left[\log f(s) - \log\frac{f(s)}{g_k(s)} + \log\frac{w_k}{\phi_k}\right] = \\ &H(f) + \sum_{k=1}^{K}\phi_k\left(D_{\mathrm{KL}}(f\|g_k) - \log\frac{w_k}{\phi_k}\right) \end{aligned} \tag{19}$$

式 (19) 右侧是关于 ϕ_k 的变量, 为了找到 $H(f,\ g)$ 的上界, 需计算出式 (19) 右侧部分的下界. 这样 $H(f,\ g)$ 的上界计算转化为以下同时包含等式和不等式约束的优化问题:

$$\begin{aligned} \min_{\phi_1,\cdots,\phi_K}\quad & H(f) + \sum_{k=1}^{K}\phi_k\left(D_{\mathrm{KL}}(f\|g_k) - \log\frac{w_k}{\phi_k}\right) \\ \text{s.t.}\quad & \sum_{k=1}^{K}\phi_k = 1 \\ & 0<\phi_k\leq 1,\ \ \forall k = 1,\cdots,K \end{aligned} \tag{20}$$

引入拉格朗日乘子 ν, $\eta_1,\cdots,\eta_K$, $\xi_1,\cdots,\xi_K$, 分别对应等式约束和两个不等式约束, 则优化问题 (20) 的拉格朗日函数为:

$$\begin{aligned} \mathcal{L}(\phi_1,\cdots,\phi_K,\nu,\eta_1,\cdots,\eta_K,\xi_1,\ldots,\xi_K) = \\ H(f) + \sum_{k=1}^{K}\phi_k\left(D_{\mathrm{KL}}(f\|g_k) - \log\frac{w_k}{\phi_k}\right) + \\ \nu\left(\sum_{k=1}^{K}\phi_k - 1\right) - \sum_{k=1}^{K}\eta_k\phi_k + \sum_{k=1}^{K}\xi_k(\phi_k - 1) \end{aligned} \tag{21}$$

因此, 优化问题 (20) 的 Karush-Kuhn-Tucker (KKT) 条件可列出如下:

$$\begin{cases} \nabla_{\phi_k^*}\mathcal{L} = D_{\mathrm{KL}}(f\|g_k) - \log\dfrac{w_k}{\phi_k^*} + 1 + \\ \qquad \nu^* - \eta_k^* + \xi_k^* = 0 \\ \displaystyle\sum_{k=1}^{K}\phi_k^* = 1 \\ 0<\phi_k^*\leq 1,\ k=1,\cdots,K \\ \eta_k^*\geq 0,\ \xi_k^*\geq 0,\ k=1,\cdots,K \\ \eta_k^*\phi_k^* = 0,\ \xi_k^*(\phi_k^* - 1) = 0,\ k=1,\cdots,K \end{cases}$$

通过求解以上 KKT 条件, 得出优化问题 (20) 的最优解:

$$\phi_k^* = \frac{w_k\mathrm{e}^{-D_{\mathrm{KL}}(f\|g_k)}}{\displaystyle\sum_{j=1}^{K} w_j\mathrm{e}^{-D_{\mathrm{KL}}(f\|g_j)}},\quad k=1,\cdots,K \tag{22}$$

联立式 (19) 与式 (22) 进而得到

$$\begin{aligned} H(f,\ g) &\leq H(f) + \sum_{k=1}^{K}\phi_k^*\left(D_{\mathrm{KL}}(f\|g_k) - \log\frac{w_k}{\phi_k^*}\right) = \\ &H(f) - \sum_{k=1}^{K}\phi_k^*\left(\log\sum_{j=1}^{K} w_j\mathrm{e}^{-D_{\mathrm{KL}}(f\|g_j)}\right) = \\ &H(f) - \log\sum_{j=1}^{K} w_j\mathrm{e}^{-D_{\mathrm{KL}}(f\|g_j)} \end{aligned} \tag{23}$$

因此得到 $H(f,\ g)$ 的上界:

$$H(f,\ g)_+ = H(f) - \log \sum_{k=1}^{K} w_k \mathrm{e}^{-D_{\mathrm{KL}}(f\|g_k)} \tag{24}$$

利用式 (16)、(24), 通过平均 $H(f,\ g)$ 的下界和上界, 便得到 $H(f,\ g)$ 的近似值为

$$\begin{aligned} D_{\mathrm{KL}}(f\|g) = H(f,\ g) - H(f) \approx \\ \frac{1}{2}\left(H(f,\ g)_- + H(f,\ g)_+\right) - H(f) \approx \\ -\frac{1}{2}\log \sum_{k=1}^{K} w_k \mathrm{E}_{s\sim f}[g_k(s)] - \frac{1}{2}H(f) - \\ \frac{1}{2}\log \sum_{k=1}^{K} w_k \mathrm{e}^{-D_{\mathrm{KL}}(f\|g_k)} \end{aligned} \tag{25}$$

□

根据定理 1, 式 (9) 可通过计算一个高斯分布和一个混合高斯分布的 KL 散度得到, 而这些高斯分布的均值和方差信息则通过预训练环境动力学模型计算获得. 因此可以定量地衡量出无人机 i 的动作对系统全局状态的因果影响, 进而为奖励函数的分配提供依据. 该部分计算流程如图 6 右侧部分所示.

需要指出的是, 定理 1 中的 KL 近似建立在状态转移分布可由高斯或混合高斯模型的假设之上. 在连续控制任务中, 该假设在局部线性或平滑动力学条件下具有较好适用性. 若真实分布存在强非高斯特性, 则该近似可能引入偏差, 但由于因果影响最终用于奖励重塑而非精确概率估计, 该误差对策略优化的影响相对可控.

2.4 奖励函数

奖励函数的设计是深度强化学习算法性能的关键, 对无人机的策略评估与改进有着直接影响. 本文将奖励函数分为环境奖励和内在奖励两部分进行设计. 环境奖励由两部分组成: 一是采用基于距离的团队奖励; 二是碰撞惩罚项, 用以抑制无人机之间以及无人机与环境障碍物间的碰撞. 具体而言, 基于距离的团队奖励针对的是环境中所有目标点 j ($j = 1,\ 2,\ \cdots,\ M$), 计算该目标位置与其最近无人机的距离 d_t^j. 若 $d_t^j < r_s$ (r_s 为超参数, 表示成功靠近的阈值), 则视为已接近并给予固定正奖励 +5; 否则以距离的负值作为惩罚, 即为 $-d_t^j$. 因此, 时刻 t 的团队距离奖励定义为:

$$r_{\mathrm{team},\ t} = \sum_{j=1}^{M} r_{\mathrm{dist},\ t}^{j} \tag{26}$$

其中,

$$r_{\mathrm{dist},\ t}^{j} = \begin{cases} +5, & d_t^j < r_s \\ -d_t^j, & d_t^j \geq r_s \end{cases} \tag{27}$$

根据式 (26) 不难发现, 每架无人机的距离奖励均相同, 因此它是团队共享的奖励部分. 可以看出, 当一个目标地点被某架无人机靠近时, 团队奖励 $r_{\mathrm{team},\ t}$ 增大 (趋于正值). 只有当每个目标地点都被一架无人机覆盖时, 表示无人机之间通过协作完成了多个地点的导航任务.

令 $r_{\mathrm{collide1},\ t}^{i}$ 表示无人机 i 与其他无人机发生碰撞后获得的惩罚, $r_{\mathrm{collide2},\ t}^{i}$ 表示无人机 i 与障碍物发生碰撞时获得的惩罚. 因此, 无人机的碰撞惩罚定义如下:

$$r_{\mathrm{collide},\ t}^{i} = r_{\mathrm{collide1},\ t}^{i} + r_{\mathrm{collide2},\ t}^{i} \tag{28}$$

当发生碰撞时, $r_{\mathrm{collide1},\ t}^{i}$ 与 $r_{\mathrm{collide2},\ t}^{i}$ 对应的取值为 −5, 否则其取值为 0.

以上是环境奖励的设计, 而本文的重点在于内在奖励的设计, 该奖励是基于因果效应检测的结果计算得出, 用以衡量单架无人机动作对环境 (下一状态分布) 的贡献, 从而鼓励无人机执行对团队有因果影响的行为. 具体地, 基于因果效应的内在奖励项记为 $r_{\mathrm{cid},\ t}^{i}(s_t,\ s_{t+1})$, 表示第 i 架无人机的因果影响度 $C^i(s_t,\ s_{t+1})$ 在所有无人机影响度之和中的归一化比重:

$$r_{\mathrm{cid},\ t}^{i}(s_t,\ s_{t+1}) = \frac{C^i(s_t,\ s_{t+1})}{\sum\limits_{j=1}^{N} C^j(s_t,\ s_{t+1})} \tag{29}$$

其中, $C^i(s_t,\ s_{t+1})$ 表示第 i 架无人机的因果影响值, 其定义如式 (6) 所示, 并且由定理 1 给出了近似计算方法. 这样, 就得到了第 i 架无人机在时刻 t 的奖励函数 r_t^i, 其计算如下:

$$r_t^i = r_{\mathrm{team},\ t} + r_{\mathrm{collide},\ t}^{i} + \lambda r_{\mathrm{cid},\ t}^{i} \tag{30}$$

其中, λ 为超参数, 用于平衡环境奖励和内在奖励的权重关系. 由于内在奖励采用归一化形式, 并通过 λ 控制其强度, 当 λ 较小时不会改变原始最优策略集合, 而主要起到探索引导作用. 因此, 该机制不会破坏原问题的最优性, 仅在训练阶段影响策略搜索路径.

需要说明的是, 本文所提出的因果影响度量机制本质上基于多智能体因果图结构进行建模, 其核心在于刻画个体动作对整体状态转移分布的因果贡献, 因此并不依赖于 $N = M$ 的特定设置 (即无人机个数和目标个数相同). 当目标数量大于或小于无人

机数量时, 因果影响的估计仍然通过对个体干预与联合分布变化进行计算获得, 其计算形式保持不变. 不同规模关系可能会影响协作策略的复杂度与分工结构, 但不会改变因果贡献估计机制本身.

2.5　基于因果影响检测的多智能体近端策略优化

本文所提 CID-MAPPO 方法基于 MAPPO 算法框架, 主要考虑在第 i 架无人机训练中引入一项具有因果效应的函数项, 使得在训练阶段每架无人机能够学习到更优的策略, 从而缓解完全协作任务中常见的“懒惰智能体”问题.

本文采用 actor-critic 架构: actor 是由参数 θ 表征的策略网络, 一共有 N 个, 分别用于近似随机策略 $\pi_\theta^i(a_t^i \mid o_t^i)$, $i=1, \cdots, N$, 其功能是将第 i 架无人机在时刻 t 的局部观测 o_t^i 映射为相应的动作分布, 从而生成决策动作; critic 则是由参数 ψ 表征的价值网络, 同样有 N 个, 用于估计第 i 架无人机的价值函数 $V_\psi^i(s_t)$, 以评估当前状态下策略 π_θ^i 的长期回报, 并为策略更新提供梯度信息. 为提升训练稳定性与收敛效率, CID-MAPPO 采用 CTDE 框架. 算法具体框架如图 6 所示.

在集中训练阶段, critic 网络利用系统的全局状态 s_t 及奖励信息, 分别为无人机 i ($i=1, \cdots, N$) 学习对应的价值函数 $V_\psi^i(s_t)$. critic 通过最小化如下损失函数进行更新:

$$Loss(V_\psi^i) = \mathrm{E}_{(s_t,\, s_{t+1},\, r_{t+1}^i)\sim\mathcal{D}}\left[\left(V_\psi^i(s_t) - \hat{y}_t^i\right)^2\right] \tag{31}$$

其中, $\mathcal{D}$ 表示经验回放池; $\hat{y}_t^i = r_{t+1} + \gamma V_\psi^i(s_{t+1})$. actor 网络则通过最大化如下目标函数实现对无人机 i 策略 π_θ^i 的更新:

$$J(\pi_\theta^i) = \mathrm{E}\left[\min(\rho_t^i(\theta)A_\theta^i,\ \mathrm{clip}(\rho_t^i(\theta),\ 1-\varepsilon,\ 1+\varepsilon)A_\theta^i)\right] \tag{32}$$

其中,

$$\rho_t^i(\theta) = \frac{\pi_\theta^i(a_t^i|o_t^i)}{\pi_{\theta_{\mathrm{old}}}^i(a_t^i|o_t^i)} \tag{33}$$

$$\mathrm{clip}(\rho_t^i(\theta),\ 1-\varepsilon,\ 1+\varepsilon) = \begin{cases} 1+\varepsilon, & \rho_t^i(\theta) > 1+\varepsilon \\ \rho_t^i(\theta), & 1-\varepsilon \le \rho_t^i(\theta) \le 1+\varepsilon \\ 1-\varepsilon, & \rho_t^i(\theta) < 1-\varepsilon \end{cases} \tag{34}$$

A_θ^i 表示无人机 i 的优势函数, $A_\theta^i = r_{t+1}^i + \gamma V_\psi^i(s_{t+1}) - V_\psi^i(s_t)$; π_θ^i 表示无人机 i 的当前策略; $\pi_{\theta_{\mathrm{old}}}^i$ 表示无人机 i 的旧策略; ε 为裁剪系数, 本文设置 $\varepsilon = 0.2$.

需要强调的是, 本文引入因果视角并非为了解决隐藏变量问题, 而是为了区分统计相关性与真实协同贡献. 在多智能体协作场景中, 仅基于奖励相关性难以刻画个体对系统状态演化的边际干预效应.

3　仿真实验和分析

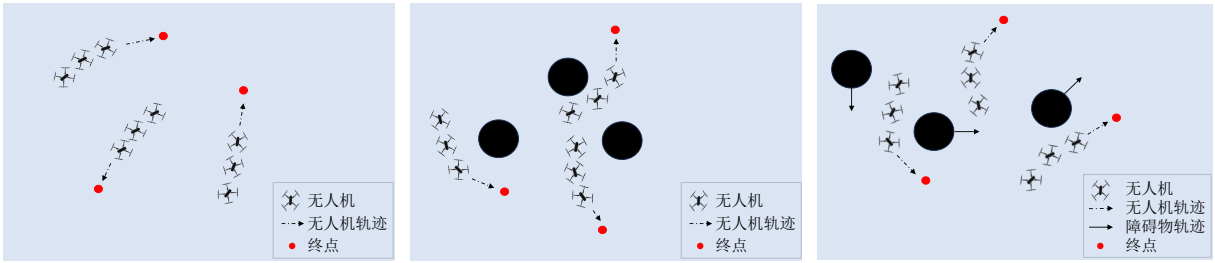
3.1　环境设定

本文基于 Gym 环境构建了多无人机协同导航的仿真平台, 包含多架无人机、多个目标地点以及动态与静态障碍物. 为了训练出具备更好泛化能力、适应真实环境的策略, 在搭建仿真环境时本文考虑了风速、礁石等因素对多无人机系统协同控制的影响. 对于复杂地形和人为设施, 实验中引入了诸如鸟群或其他飞行器的动态障碍, 以评估无人机在遇到各类障碍时的避障与导航能力. 如图 7 所示, 以 3 架无人机的海上救援任务场景为例: 场景 1 包含 3 架无人机与 3 名待救援人员 (红色圆点所示); 场景 2 包含 3 架无人机、多个静态障碍物 (黑色圆点所示) 与 3 名待救援人员; 场景 3 包含 3 架无人机、3 个待救援人员及多个动态障碍物 (带有移动方向的黑色圆点所示). 为聚焦协同导航策略的学习效果, 在仿真实现中采用了简化的平面运动模型, 仅考虑位置与速度的二维演化, 而忽略姿态耦合效应. 该简化不影响因果奖励机制的构造. 在仿真环境中, 多个目标点用于模拟待救援目标的位置, 无人机通过覆盖与接近目标完成“发现”过程, 累积奖励反映搜救效率与协同质量. 因此, 实验任务可视为对多无人机协同搜救场景的抽象建模.

3.2　实验超参数设置

为保证实验结果的可复现性, 客观地评估所提方法的性能, 本文对算法所涉及各模块的超参数设置如下.

在所有对比实验中, 各基准算法均采用相同的网络结构, 学习率 α 设置为 0.000 5, 折扣因子 $\gamma = 0.99$. 各智能体的 actor 网络和 critic 网络分别为包含 3 层和 2 层隐层的 MLP 网络作为全连接层, 最后一层为循环神经网络 (recurrent neural network, RNN) 层, 各隐藏层均包含 64 个节点, 每个隐藏层均使用 ReLU 函数激活. 在训练阶段, 经验回放池大小设置为 3 200, 每轮训练采样的批量数据大小设置为 1 024. 近端策略优化中的裁剪系数设置为 $\varepsilon = 0.2$. 针对环境动力学模型训练, 本研究使用 MLP 拟合 β- VAE 中的编码器和解码器, 其



(a) 场景 1
(a) Scene 1

(b) 场景 2
(b) Scene 2

(c) 场景 3
(c) Scene 3

图 7 实验场景示意图

Fig. 7 Diagram of the experimental scenes

中编码器分为状态编码和动作编码, 其隐层节点及以上超参数设置如表 2 所示. β- VAE 模块每当经验回放池 (experience replay buffer) 存满了更新一次. 针对式 (8) 所描述的蒙特卡洛采样, 本文采用均匀采样策略, 并且在所有实验中动作采样数量 K 均设置为 128.

3.3 实验结果

为了验证本文方法的有效性, 本文对比不同实验场景下的学习曲线和累积回合奖励来验证所提方法的有效性. 如图 8 所示, 其中每幅子图的横轴为训练时间步数, 纵轴为每回合的平均累积奖励, 数值越高表示策略的性能越好. 首先, 分别在包含 3、5 和 7 架无人机的场景 1 任务上评估了本文所提方法与其他基线方法的表现, 图 8(a)、图 8(b) 和图 8(c) 分别展示了本文所提方法与 MAPPO[20]、VDN[17]、QMIX[18]、MADDPG[19] 与 PMIC[31] 在没有障碍物情况下完成救援任务的回报比较结果. 观察实验曲线可以发现, 本文所提 CID-MAPPO 相较于其他算法在学习性能上表现更好, VDN、QMIX 和 MADDPG 在训练过程中的噪声干扰较大导致方差较大. 相比于 MAPPO 和 PMIC, 尽管 CID-MAPPO 在场景 1 中训练前期的收敛速度上表现相当, 但其收敛后的性能明显优于其他算法, 这表明内在奖励确实有助于提升智能体之间的协作. 从另一个角度看, 增强智能体之间的因果影响明显比仅仅增强它们的相关性更有助于协作.

此外, 本研究还在包含 3、5、7 架无人机的静态障碍物场景中对所提方法进行了评估. 该任务要求各无人机在到达救援人员位置的同时规避环境中的固定障碍物. 从图 8(d)、图 8(e) 和图 8(f) 可以观察到, CID-MAPPO 在不同规模的协作导航任务中均取得了高于其他基线方法的回报, 表现出良好的可扩展性. 在此基础上, 本文进一步考察了 CID-MAPPO 在更具挑战性的动态障碍物场景 (同样包含 3、5 和 7 架无人机) 中的性能. 该任务代表了极高难度的环境配置, 需要无人机在协作搜索的同时对动态风险做出及时响应. 图 8(g)、图 8(h) 和图 8(i) 表明, CID-MAPPO 不仅收敛速度更快, 而且在总体性能上依旧显著优于对比算法. 为提供直观的定性结果, 图 9(a)、图 9(b) 和图 9(c) 展示了 CID-MAPPO 在三个实际任务场景中的轨迹示例 (如虚线所示). 可以看到, 在场景 2 中, 即使布设了多种静态障碍物, 无人机仍能够通过协作高效抵达目标位置; 在带有动态障碍的场景 3 中, CID-MAPPO 亦能成功规避多类动态干扰并顺利完成任务. 表 3 给出了各场景下不同算法的平均性能和方差对比 (三个不同随机种子). 综合以上实验结果可以看出, CID-MAPPO 在多种任务类型、不同规模的无人机

表 2 实验超参数设置

Table 2 Experimental hyper-parameters setting

参数名称	取值
学习率 α	0.000 5
VAE 模块学习率 α_{vae}	0.000 5
折扣因子 γ	0.99
裁剪系数 ε	0.2
激活函数	ReLU
批量数据大小	1 024
经验回放池 $\mathcal{D}$ 大小	3 200
奖励权重 λ	1.0
VAE 参数 β	0
蒙特卡洛采样数 K	128
actor 网络全连接层节点数	[64, 64, 64]
actor 网络 RNN 隐层节点数	64
critic 网络全连接层节点数	[64, 64]
critic 网络 RNN 隐层节点数	64
β-VAE 编码器隐层节点数 (状态编码)	[256, 128, 64]
β-VAE 编码器隐层节点数 (动作编码)	[64, 128]
β-VAE 解码器隐层节点数	[64, 128, 256]

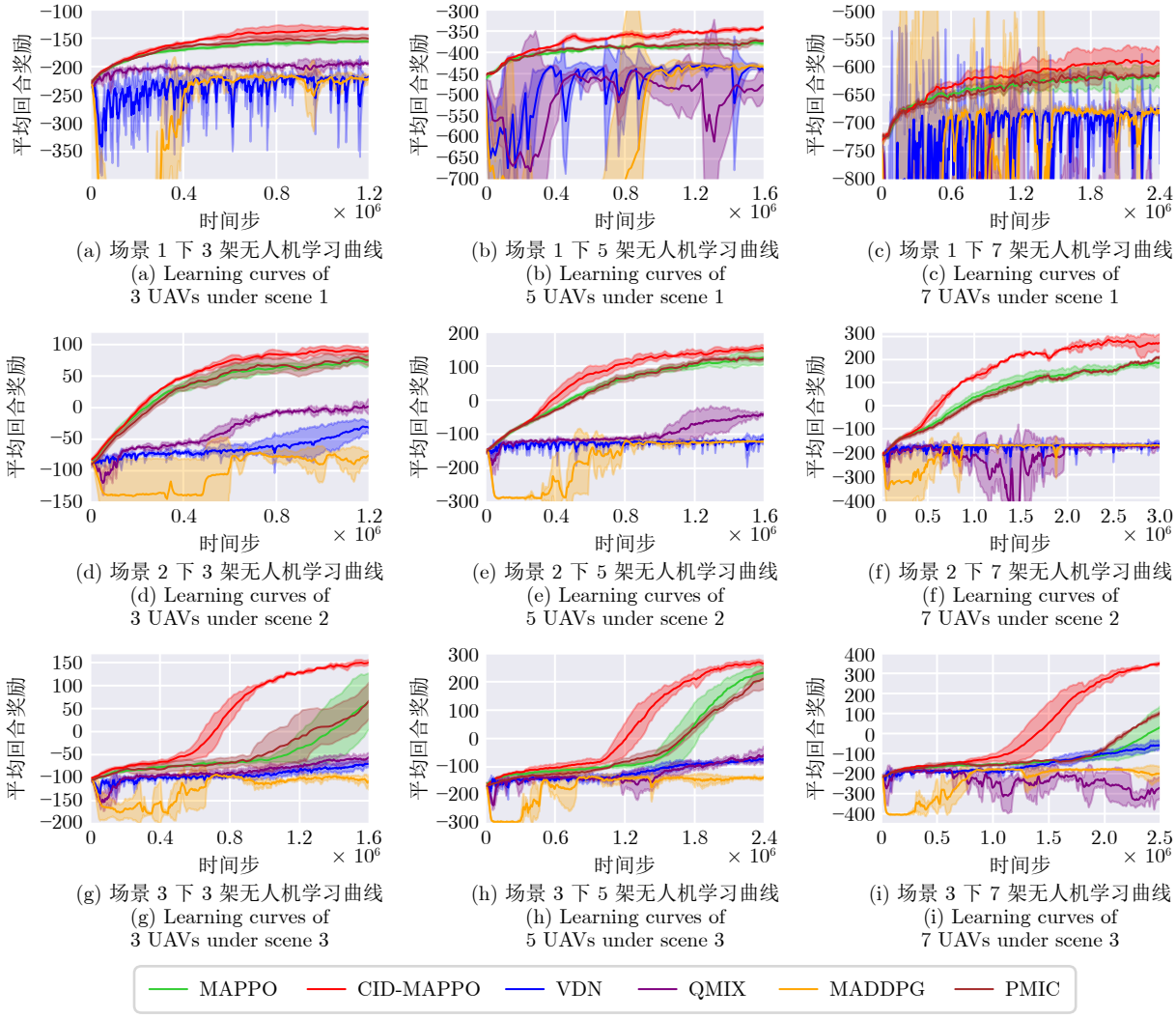


图 8　不同场景下各算法学习曲线

Fig. 8　Learning curves of different algorithms under various scenarios

场景以及复杂干扰条件下均表现出稳定且显著的优势, 验证了其卓越的鲁棒性与协作能力.

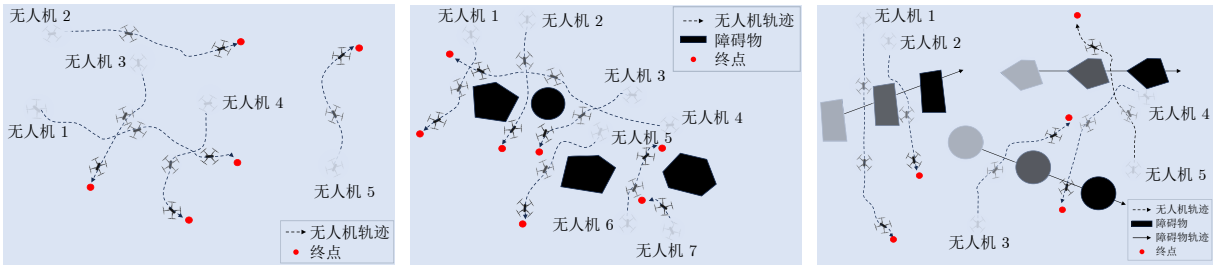
从方法结构上看, 本文所提出的因果影响估计机制基于 CTDE 框架构建, 不依赖于特定任务结构, 因此在理论上可迁移至通用多智能体基准环境. 受篇幅与实验资源限制, 本文未在额外公开基准环境或真实飞行数据上展开系统验证, 未来工作将进一步在标准多智能体测试平台及更接近实际场景的数据环境中评估算法的泛化能力.

3.4　消融实验

为进一步分析本方法中的内在奖励权重 λ 以及 β- VAE 中 β 系数对算法性能的影响, 本研究对这两个超参数进行了消融实验. 选择场景 2 下 3 架无人机作为消融实验环境, 保持其他训练设置不变, 实验结果如图 10 所示, 其中横轴表示训练时间步数, 纵轴为每回合的平均累积奖励, 数值越高表示策略的性能越好.

内在奖励权重. 从图 10(a) 可以看出, 内在奖励权重 λ 对算法性能具有显著影响. 当 $\lambda = 0$ 时, 算法退化为无因果内在奖励版本, 整体收敛速度较慢且最终性能有限. 随着 λ 增大, 算法性能明显提升, 说明因果内在奖励能够有效促进协作探索. 然而当 λ 进一步增大至 1.5 或 2.0 时, 性能出现下降趋势, 这表明过强的内在驱动会削弱对外部任务目标的优化. 实验结果表明 $\lambda = 1$ 时在探索效率与任务收益之间取得较优平衡.

β- VAE 正则系数. 如图 10(b) 所示, 随着 β 增大, 算法收敛速度与最终性能均呈下降趋势, 且在 $\beta = 0$ 时取得最佳结果. 这表明在本文设定下, KL 正则项对潜变量分布的约束会引入额外的信息压缩, 从而削弱潜在表示对控制相关信息的保留能力,



(a) 场景 1 下 5 架无人机的轨迹示意图
(a) Trajectories diagram of 5 UAVs under scene 1

(b) 场景 2 下 7 架无人机的轨迹示意图
(b) Trajectories diagram of 7 UAVs under scene 2

(c) 场景 3 下 5 架无人机的轨迹示意图
(c) Trajectories diagram of 5 UAVs under scene 3

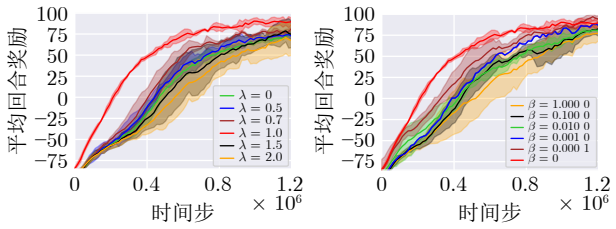
图 9 部分实验场景轨迹示意图

Fig. 9 Trajectories diagram in some experimental scenes

表 3 各算法在不同实验配置下的累积回报统计结果

Table 3 Statistical results of cumulative rewards for different algorithms under various experimental configurations

场景	CID-MAPPO	MAPPO	VDN	QMIX	MADDPG	PMIC
场景 1 下 3 架无人机	**−131.89 ± 22.66**	−155.67 ± 15.71	−225.21 ± 39.60	−194.05 ± 5.96	−225.02 ± 9.64	−147.94 ± 17.65
场景 1 下 5 架无人机	**−342.43 ± 26.52**	−377.48 ± 16.83	−442.86 ± 11.65	−548.82 ± 116.16	−425.96 ± 6.70	−371.49 ± 17.98
场景 1 下 7 架无人机	**−590.88 ± 43.33**	−616.58 ± 35.03	−673.14 ± 211.54	−916.50 ± 239.70	−673.55 ± 28.43	−610.09 ± 26.67
场景 2 下 3 架无人机	**89.79 ± 10.45**	65.35 ± 12.62	−43.16 ± 16.48	7.42 ± 32.91	−75.18 ± 29.76	74.46 ± 18.73
场景 2 下 5 架无人机	**156.93 ± 16.15**	128.10 ± 15.49	−117.25 ± 23.16	−49.41 ± 29.84	−120.12 ± 33.18	125.02 ± 21.03
场景 2 下 7 架无人机	**259.88 ± 41.40**	194.71 ± 23.77	−176.29 ± 26.21	−180.94 ± 31.54	−173.93 ± 24.84	228.79 ± 16.97
场景 3 下 3 架无人机	**149.63 ± 12.43**	71.39 ± 63.32	−73.31 ± 14.59	−55.00 ± 21.38	−106.13 ± 33.02	89.83 ± 48.67
场景 3 下 5 架无人机	**279.15 ± 11.32**	232.04 ± 23.27	−70.00 ± 29.73	55.69 ± 35.57	−141.55 ± 14.94	209.06 ± 28.05
场景 3 下 7 架无人机	**332.24 ± 12.04**	92.45 ± 41.34	−58.83 ± 35.12	−276.00 ± 76.42	−197.56 ± 27.56	115.33 ± 34.52



(a) 奖励权重 λ
(a) Reward weight λ

(b) 参数 β
(b) Parameter β

图 10 超参数消融实验结果

Fig. 10 Results of the ablation experiments on hyper-parameters

进而影响策略优化效果. 相比之下, 移除 KL 约束 ($\beta = 0$) 能够保留更充分的状态表征信息, 反而更有利于稳定学习与最终性能. 需要指出, 当 $\beta = 0$ 时, 该模块退化为等价的自编码形式, 其目标更偏向重构而非分布正则化. 基于该观察, 本文在其余实验中默认采用 $\beta = 0$ 以获得更稳定且更优的控制性能.

4 结束语

为促进多无人机之间的有效协作并增强探索能力, 本文提出了 CID-MAPPO 方法. 该方法从因果关系视角度量无人机行为对团队整体的影响, 并在此基础上设计内在奖励机制, 将无人机之间的因果影响显式引入策略更新过程. 通过鼓励无人机优先探索对全局状态具有正向贡献的动作, CID-MAPPO 能够显著提升协作效率与任务表现. 相较于现有基于相关性或经验设计奖励的方法, CID-MAPPO 能够更直接地刻画个体行为对团队协作的实际作用, 尤其适用于多智能体强耦合、协作关系复杂的任务场景. 在模拟的多无人机海上救援任务场景中, 本文对所提算法进行了性能验证. 实验结果表明, 与现有主流方法相比, CID-MAPPO 在协作性、稳定性以及综合性能方面均取得了显著提升, 有效验证了方法的可行性与优势.

参考文献

1 Chang Y Y, John A, Hsiung P A. Maritime UAV patrol tasks based on YOLOv4 object detection. In: Proceedings of the 2022 International Conference on Computational Science and Computational Intelligence. Las Vegas, USA: IEEE, 2022. 1484–1490

2 Xu P F, Fang Y, Jiang Q Y, Lu H F, Li G X, Zhou H. Design and research of a water quality monitoring system for aquaculture using UAV integrated with 3D GIS. *Advances in Transdisciplinary Engineering*, DOI: 10.3233/atde241331

3 Sendner F. An energy-autonomous UAV swarm concept to support sea-rescue and maritime patrol missions in the Mediter-

ranean Sea. *Aircraft Engineering and Aerospace Technology*, 2022, **94**(1): 112−123

4 Nomikos N, Gkonis P K, Bithas P S, Trakadas P. A survey on UAV-aided maritime communications: Deployment considerations, applications, and future challenges. *IEEE Open Journal of the Communications Society*, 2022, **4**: 56−78

5 Liu H D, Long X L, Li Y, Yan J J, Li M Y, Chen C, et al. Adaptive multi-UAV cooperative path planning based on novel rotation artificial potential fields. *Knowledge-Based Systems*, 2025, **317**: Article No. 113429

6 Zhao H M, Gu M X, Qiu S P, Zhao A, Deng W. Dynamic path planning for space-time optimization cooperative tasks of multiple unmanned aerial vehicles in uncertain environment. *IEEE Transactions on Consumer Electronics*, 2025, **71**(3): 7673−7682

7 Hu W J, Yu Y, Liu S M, She C Y, Guo L, Vucetic B. Multi-UAV coverage path planning: A distributed online cooperation method. *IEEE Transactions on Vehicular Technology*, 2023, **72**(9): 11727−11740

8 Sil M, Rakshit P, Chatterjee S, Ghosh R A, Chowdhury A. Multi-UAV cooperative path-planning in complex terrain: A multi-objective optimization approach. *IETE Journal of Research*, 2025, **71**(8): Article No. 2606

9 Sun W M, Hao M R. A survey of cooperative path planning for multiple UAVs. In: Proceedings of the 2021 International Conference on Autonomous Unmanned Systems. Singapore: Springer, 2021. 189–196

10 Liu W, Cai W Z, Jiang K, Cheng G R, Wang Y D, Wang J W, et al. Xuance: A comprehensive and unified deep reinforcement learning library. arXiv preprint arXiv: 2312.16248, 2023.

11 Alexandros T, Georgios D. A comprehensive survey on the applications of swarm intelligence and bio-inspired evolutionary strategies. *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications*, DOI: 10.1007/978-3-030-49724-8_15

12 Dario F, Claudio M. *Bio-inspired Artificial Intelligence: Theories, Methods, and Technologies*. Cambridge: MIT press, 2008.

13 Haldorai A, Kandaswamy U. A bio-inspired swarm intelligence technique for social aware cognitive radio handovers. *Computers & Electrical Engineering*, 2018, **71**: 925−937

14 Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301−1312
(孙长银, 穆朝絜. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301−1312)

15 Luo Biao, Hu Tian-Meng, Zhou Yu-Hao, Huang Ting-Wen, Yang Chun-Hua, Gui Wei-Hua. Survey on multi-agent reinforcement learning for control and decision-making. *Acta Automatica Sinica*, 2025, **51**(3): 510−539
(罗彪, 胡天萌, 周育豪, 黄廷文, 阳春华, 桂卫华. 多智能体强化学习控制与决策研究综述. 自动化学报, 2025, **51**(3): 510−539)

16 Chen Kai, Lei Yi-Chen, Li Yan-Ze, Fang Guo-Yu, Hu Zi-Zhuo, Yang Ming-Shi, et al. Multi-object trajectory planning method based on improved MADDPG. *Journal of Beijing University of Aeronautics and Astronautics*, DOI: 10.13700/j.bh.1001-5965.2025.0636
(陈凯, 雷一辰, 李琰泽, 方国宇, 胡子卓, 杨明实, 等. 基于改进MADDPG 的多目标航迹规划方法. 北京航空航天大学学报, DOI: 10.13700/j.bh.1001-5965.2025.0636)

17 Sunehag P, Lever G, Gruslys A, Czarnecki W M, Zambaldi V, Jaderberg M, et al. Value-decomposition networks for cooperative multi-agent learning. arXiv preprint arXiv: 1706.05296, 2017.

18 Rashid T, Samvelyan M, de Witt C S, Farquhar G, Foerster J, Whiteson S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 2020, **21**(178): 1−51

19 Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-agent actor-critic for mixed cooperative competitive environments. *Advances in Neural Information Processing Systems*, DOI: 10.48550/arXiv.1706.02275

20 Yu C, Velu A, Vinitsky E, Gao J X, Wang Y, Bayen A, et al. The surprising effectiveness of PPO in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 2022, **35**: 24611−24624

21 Foerster J, Farquhar G, Afouras T, Nardelli N, Whiteson S. Counterfactual multi-agent policy gradients. In: Proceedings of the 2018 AAAI Conference on Artificial Intelligence. Louisiana, USA: AAAI Press, 2018. 2974–2982

22 Rashid T, Samvelyan M, de Witt C S, Farquhar G, Foerster J, Whiteson S. Maritime search and rescue based on group mobile computing for unmanned aerial vehicles and unmanned surface vehicles. *IEEE Transactions on Industrial Informatics*, 2020, **16**(12): 7700−7708

23 Lei C J, Wu S H, Yang Y, Xue J Y, Zhang Q Y. Maritime search and rescue leveraging heterogeneous units: A multi agent reinforcement learning approach. In: Proceedings of the 12th IEEE/CIC International Conference on Communications. Dalian, China: IEEE, 2018. 1–6

24 Lei C J, Wu S H, Yang Y, Xue J Y, Zhang Q Y. Joint trajectory and communication optimization for heterogeneous vehicles in maritime sar: Multi-agent reinforcement learning. *IEEE Transactions on Vehicular Technology*, 2024, **73**(9): 12328−12344

25 Wu X, Yan Q Z, Wang J C, Zhou Y H, Huang Q L, Jiang C H. Dynamic task allocation for UAV swarms in maritime rescue scenarios based on PG-MAPPO. *IEEE Internet of Things Journal*, 2025, **12**(18): 38073−38087

26 Luo Q Y, Luan T H, Shi W S, Fan P Z. Deep reinforcement learning based computation offloading and trajectory planning for multi-UAV cooperative target search. *IEEE Journal on Selected Areas in Communications*, 2022, **41**(2): 504−520

27 Hou Y K, Zhao J, Zhang R Q, Cheng X, Yang L Q. UAV swarm cooperative target search: A multi-agent reinforcement learning approach. *IEEE Transactions on Intelligent Vehicles*, 2022, **9**(1): 568−578

28 Panait L, Luke S. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 2005, **11**(3): 387−434

29 Liu M H, Zhou M, Zhang W N, Zhuang Y Z, Wang J, Liu W L, et al. Multi-agent interactions modeling with correlated policies. arXiv preprint arXiv: 2001.03415, 2020.

30 Du X, Ye Y T, Zhang P Y, Yang Y N, Chen M S, Wang T. Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI, 2024. 17362–17370

31 Li P Y, Tang H Y, Yang T P, Hao X T, Sang T, Zheng Y, et al. PMIC: Improving multi-agent reinforcement learning with progressive mutual information collaboration. In: Proceedings of the 39th International Conference on Machine Learning. Seoul, South Korea: PMLR, 2022. 12979–12997

32 Kim W J, Jung W Y, Cho M S, Sung Y C. A variational approach to mutual information-based coordination for multi-agent reinforcement learning. arXiv preprint arXiv: 2303.00451, 2023.

33 Kim W J, Jung W Y, Cho M S, Sung Y C. Signal instructed coordination in cooperative multi-agent reinforcement learning. arXiv preprint arXiv: 1909.04224, 2019.

34 Seitzer M, Schölkopf B, Martius G. Causal influence detection for improving efficiency in reinforcement learning. In: Proceedings of the 2021 Advances in Neural Information Processing Systems. Virtual Event: NeurIPS, 2021. 22905–22918

35 Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In: Proceedings of the 2017 International Conference on Learning Representations. Toulon, France: ICLR, 2017. 60–81

柳文章 安徽大学人工智能学院讲师. 2021 年获得东南大学控制科学与工程专业博士学位. 主要研究方向为深度强化学习, 具身智能系统, 多智能体强化学习, 迁移强化学习.
E-mail: wzliu@ahu.edu.cn
(**LIU Wen-Zhang** Lecturer at the School of Artificial Intelligence, Anhui University. He received his Ph.D. degree in control science and engineering from Southeast University in 2021. His research interests include deep reinforcement learning, embodied intelligent systems, multi-agent reinforcement learning, and transfer reinforcement learning.)

陆建华 安徽大学人工智能学院硕士研究生. 主要研究方向为多智能体强化学习, 多无人机协同控制.
E-mail: jhlu@stu.ahu.edu.cn
(**LU Jian-Hua** Master student at the School of Artificial Intelligence, Anhui University. His research interests include multi-agent reinforcement learning and multi-UAV cooperative control.)

任 璐 安徽大学人工智能学院副教授. 2021 年获得东南大学控制科学与工程专业博士学位. 主要研究方向为自主无人系统分布式协同控制, 深度强化学习和多智能体强化学习.
E-mail: penny_lu@ahu.edu.cn
(**REN Lu** Associate professor at the School of Artificial Intelligence, Anhui University. She received her Ph.D. degree in control science and engineering from Southeast University in 2021. Her research interests include distributed cooperative control of autonomous unmanned systems, deep reinforcement learning, and multi-agent reinforcement learning.)

孙长银 安徽大学人工智能学院教授. 2004 年获得东南大学电子工程专业博士学位. 主要研究方向为智能控制, 飞行器控制, 模式识别和优化理论. 本文通信作者.
E-mail: cysun@seu.edu.cn
(**SUN Chang-Yin** Professor at the School of Artificial Intelligence, Anhui University. He received his Ph.D. degree in electrical engineering from Southeast University in 2004. His research interests include intelligent control, flight control, pattern recognition, and optimal theory. Corresponding author of this paper.)