arXiv:2303.01150v1 [cs.RO] 2 Mar 2023

基于深度强化学习的多无人机自适应路径规划

Jonas Westheider、Julius R¨uckin、Marija Popovi´c

摘要—高效的空中数据采集对许多遥感应用至关重要。在大规模监测场景中，部署无人机团队能够提升空间覆盖率，并增强应对个别无人机故障的鲁棒性。然而，一项关键挑战是如何为无人机进行协作路径规划，以高效实现联合任务目标。我们提出了一种基于深度强化学习的全新多智能体信息路径规划方法，适用于无人机团队执行自适应地形监测任务的场景。我们引入了新的网络特征表示，以有效学习三维工作空间中的路径规划。通过利用反事实基线，我们的方法能够明确处理信用分配问题，从而学习协同行为。实验评估表明，与非反事实变体相比，我们的方法规划性能更佳，即能更快地绘制感兴趣区域的地图。在合成数据和真实世界数据上的结果表明，与当前最先进的非学习方法相比，我们的方法性能更优，同时能够适应不同的团队规模和通信约束。

一、引言

高效的空中数据采集对于绘制地图以及监测地球表面的现象至关重要。无人机（UAV）为遥感应用提供了一种灵活、节省人力且成本效益高的解决方案，例如精确农业 [1–3], 野生动物保护 [4], 以及搜寻与救援 [5, 6]。在监测大面积地形时，以多无人机系统取代单架无人机，能够以较低的总体成本提升空间覆盖、灵活性以及应对个体故障的鲁棒性 [7]。然而，要充分发挥其潜力，一项关键挑战是在机载运行效率和通信能力受限的情况下，在复杂环境中协同规划无人机路径。

本文研究在地形监测场景中使用无人机团队进行主动数据采集。我们的目标是利用无人机采集的图像测量值，对二维地形上一个初始未知且非均质的二元目标变量进行制图，例如农业场景中的作物虫害，或灾害场景中等待救援的受害者。我们解决多智能体信息素路径规划（*IPP*）

：我们为无人机规划信息丰富的路径，使其能够在能量、时间或距离限制下协同收集传感器数据。我们的目标是让无人机能够自适应地 监测感兴趣区域内的地形，这些区域的信息价值较高。

传统的数据采集方法是非自适应的，并依赖静态预定义路径。在多无人机覆盖中

本研究由德国研究联合会（DFG）根据德国卓越战略资助，项目编号为 EXC 2070 – 390732324。作者来自波恩大学大地测量与地理信息研究所卓越集群 PhenoRob。通讯邮箱：jwesthei@uni-bonn.de。

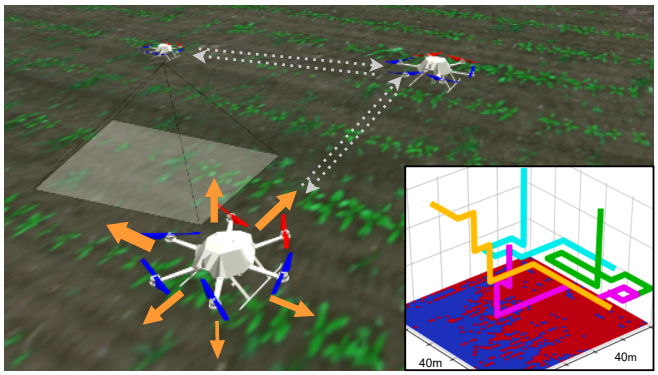


图 1：我们基于强化学习的方法应用于多无人机表面温度制图场景。无人机拍摄地形图像（白色半透明），并相互传输图像（灰色虚线箭头）。每架无人机根据本地可用信息，从一组动作（橙色箭头）中选择下一次测量的位置。插图：4 架已部署无人机的最终轨迹（不同颜色）。通过协同规划路径，我们的方法能够自适应地绘制田野中温暖的感兴趣区域（红色）。

路径规划 [8],会将地形均匀划分，并为每架无人机分配一种扫描模式。这类方法的主要缺点是，它们假设感兴趣的目标变量（如异常、热点、受害者）均匀分布，且无法对特定区域进行定向检查。为解决这一问题，研究者提出了信息素路径规划方法 [1, 9–13] ，使系统能够基于收到的信息进行在线决策。然而，由于这些策略依赖于在线评估大量候选路径，其运行时间通常会随着规划时域的增长而呈指数级增长。在团队场景中，对其他无人机行为的推理会使评估次数进一步随团队规模呈指数级增长，导致运行时间复杂度难以应对。

强化学习（RL）已成为机器人学中高效学习在线决策的一种流行方法 [11–16]。近期研究将强化学习应用于单智能体信息驱动路径规划，以提升自适应数据采集的路径质量并缩短计算时间 [11, 14, 17, 18]。然而，如何为多个协作智能体学习信息丰富的路径，仍是一项开放挑战。初步研究显示出良好的结果 [12, 13],，但仅限于二维动作空间，且未解决信用分配问题 [19],，即每个智能体对整体团队表现的贡献有多大；这会对协作能力产生不利影响。

本文的主要贡献是提出一种新颖的、基于多智能体深度强化学习的信息素路径规划方法，用于无人机团队执行自适应地形监测任务。为弥合强化学习最新进展与机器人应用之间的差距，我们基于反事实多智能体策略梯度

（COMA） [20]，明确解决协作式信息路径规划中的信用分配问题。我们的方法支持去中心化的机载决策，并能在团队规模可变的情况下实现协作式三维路径规划。我们表明：(i) 所设计的网络输入表示适用于三维动作空间中的多智能体信息路径规划；(ii) 基于 COMA 的算法能够处理信用分配问题，从而提升规划性能；(iii) 与非学习方法相比，我们的基于强化学习的方法在团队规模和通信约束各异的情况下，无需重新训练即可提升规划性能。为佐证这些论点，我们在热热点制图场景中使用合成数据和真实世界数据展示了该方法的性能。我们将在以下地址开源代码：
https://github.com/dmar-bonn/ipp-marl。

II. 相关工作

我们的工作结合了用于自主数据采集的多机器人信息路径规划与多智能体强化学习的最新进展。本节概述这两个领域的相关文献，并区分单机器人与多机器人设置。

信息路径规划已被广泛研究，以利用自主机器人实现高效的数据采集 [1, 2, 9, 11–13, 17, 21]。非自适应覆盖率规划器 [8] 根据预定义路径对地形进行全面监测。相比之下，我们关注自适应策略，根据传入的传感器数据进行在线重规划。Hollinger 等人 [9] 研究了在单艇水下巡检任务中，任务预算有限时，通过自适应在线重规划来最大化传感器测量的信息价值所带来的益处。近期研究 [21, 22] 提出了用于单架无人机监测的基于优化的信息路径规划方法。然而，这些方法需要对候选路径的预期信息价值进行计算成本高昂的评估。将其直接扩展到多架无人机，会导致复杂度呈指数级增长，从而限制其在在线应用中的适用性。

部署协作机器人团队有助于在大面积地形上执行监测任务，例如灾后评估 [5, 6] 和精准农业 [1–3]。我们重点研究分散式规划策略：机器人在机载端根据局部信息做出决策，从而增强系统对个体故障的鲁棒性，并使其能够随团队规模扩大。Best 等人 [23]提出了一种分散式蒙特卡洛树搜索变体，基于动作序列的联合概率分布，为主动感知规划各自的路径。与我们的多无人机配置类似，Albani 等人 [1]和 Carbone 等人 [2] 提出了用于监测农田中感兴趣区域的方法。前者采用受生物学启发的群体行为进行自适应信息素路径规划（IPP）。尽管结果颇具前景，但其方法依赖人工进行特定问题的参数调优。

另一类研究采用基于学习的方法进行主动数据采集。遵循集中式训练和分散式策略执行的范式，Li 等人 [24] 学习机器人间通信策略，以交换机器人本地信息。Tzes 等人 [10]提出了一种更通用的多机器人框架，该框架依赖于图神经网络。两种方法均通过监督式模仿学习进行训练，并且需要专家作为学习对象。我们使用深度强化学习来学习所需的团队行为，从而能够在各种应用中灵活地进行前瞻性规划。

强化学习（RL）日益广泛地应用于机器人学和无人机领域 [11, 14, 17, 18, 25]。Pirinen 等人 [25] 提出了一种策略，利用无人机根据有限的视觉线索寻找未知的目标区域。近年来，强化学习也被应用于信息素路径规划（IPP），以便在线高效地重新规划机器人路径。Chen 等人 [14] 提出了一种基于图的深度强化学习方法，用于探索并选择能够降低地图不确定性和行进时间的地图前沿。然而，他们的方法仅适用于二维工作空间，而我们考虑的是三维规划。其他研究则对能够带来高信息增益 [18]以及降低目标区域地图不确定性 [17]的智能体动作给予奖励。在与我们类似的问题设置中，R¨uckin 等人 [11] 提出了一种结合深度强化学习与基于采样的规划的信息素路径规划（IPP）方法，用于在三维工作空间中进行无人机自适应地形监测。天真地将单智能体算法扩展到多智能体设置，会导致复杂度随智能体数量呈指数级增长。

基于多智能体强化学习的信息素路径规划是一个相对缺乏探索的研究领域。现有的协作团队应用方法未考虑信用分配问题 [19], ，即无法区分单个智能体对整体团队表现的贡献。大多数研究关注相关应用，包括导航 [15], 、目标分配 [26], 和覆盖规划 [27]。与我们类似，Bayerlein 等人 [13] 提出了一种基于多智能体强化学习的信息素路径规划方法，在不进行智能体间通信的情况下最大化采集的数据量。Viseras 和 Garcia [12] 允许智能体通过通信模块交换信息。这两项研究均局限于无人机飞行高度恒定的情况，忽略了传感器噪声可能随高度变化这一因素。近期，Luis 等人 [16] 提出了一种深度 Q 学习算法，通过根据无人机间距离对冗余测量进行惩罚，支持学习协同行为。他们设计的奖励针对纯探索，而非对感兴趣区域进行自适应监测。

这些研究分别为单个智能体设定了人工设计的奖励。因此，它们未能充分考虑信息素路径规划的协作特性，并且需要手动调整奖励。相比之下，我们提出了一种新方法，仅依赖普遍适用的全局奖励来奖励整个无人机团队。我们将反事实多智能体（COMA）强化学习算法 [20]应用于三维工作空间中的主动机器人数据采集。通过这种方式，我们在训练过程中明确地为各个智能体分配信用。实验结果强调了明确进行信用分配对于实现协同行为的必要性，并验证了这有助于提升信息素路径规划的性能。

III. 问题陈述

我们研究一组同质无人机对平坦地形进行监测的情形。目标是在新的测量值到达时即时规划信息丰富的无人机路径，从而在给定有限任务预算的情况下，准确且高效地绘制地形上的感兴趣区域。我们简要介绍一般的多智能体信息路径规划问题、我们的制图策略，以及如何为强化学习方法量化信息价值。

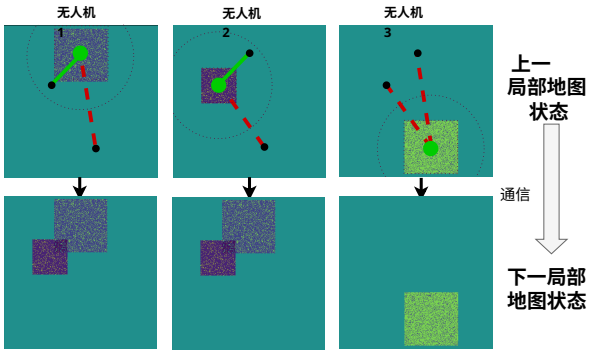


图 2：无人机间通信。当无人机彼此距离小于有限通信范围（黑色虚线圆圈）时，它们（实心圆）会相互交换当前测量值（方形覆盖区域）。绿色实线和红色虚线分别表示范围内和范围外的通信。无人机利用收到的测量值更新各自的局部地图状态（从上到下排列）。

A. 多智能体信息路径规划

我们研究多智能体信息路径规划（信息素路径规划，IPP）问题，即在所有 N无人机路径 $\psi = \{\psi_1, \ldots, \psi_N\}$中优化信息论准则 $\mathrm{I}: \Psi^N \to \mathbb{R}^+$，其中 Ψ 是所有可能的单架无人机路径的集合：

$$\psi^* = \underset{\psi \in \Psi^N}{\operatorname{argmax}}\, \mathrm{I}(\psi), \text{ s.t. } \mathrm{C}(\psi_i) \leq B \;\forall\; i \in \{1, \ldots, N\}\,. \quad (1)$$

每组无人机路径 ψ 由各条独立路径$\psi_i = (\boldsymbol{p}_i^0$、.....、$\boldsymbol{p}_i^B) \in \Psi$ 组成，每条路径长度为 B，其三维测量位置 $\boldsymbol{p}_i^t \in \mathbb{R}^3$ 位于分辨率为r_P 的等距网格 P 中，分布于地形上方的多个高度。函数 $\mathrm{C}: \Psi \to \mathbb{R}^+$ 将无人机路径 ψ_i 映射为相应的执行成本；在我们的工作中，沿一条路径 ψ_i 进行的测量次数最多为 $B \in \mathbb{N}^+$。

B. 地形映射

无人机使用朝下摄像机拍摄图像，以映射地形。图像信息经过处理（例如语义分割），随后传输并融合到地形图中。无人机i 在时间步 t 获取的测量值 z_i^t 是投影到平坦地形上的似然估计。我们考虑逐像素二元分类，并使用概率占据栅格映射 [28]对 z_i^t 进行顺序融合。每架无人机 i 都会存储一个局部后验地图信念$\mathcal{M}_i$，该信念离散为 M 个分辨率为 r_M的网格单元 $\mathcal{M}_i^j$。当前测量位置 $\boldsymbol{p}_i^t$ 的高度决定了映射的视场和映射分辨率。我们以最低高度下的映射分辨率进行对齐 r_M，并对更高高度下的测量值进行上采样至 r_M，以便在 $\mathcal{M}_i$中融合不同的映射分辨率。与 Popovi´c 等人 [21], 的方法类似，我们利用一种高度相关的传感器模型，指定 $p(z_i^t \mid \mathcal{M}_i^j, \boldsymbol{p}_i^t)$，以在每个时间步 t根据 $\mathcal{M}_i^j$更新无人机的局部后验地图信念。

图 2 展示了无人机间通信协议。所有无人机 i 发送其测量值 z_i^t 给另一架无人机 k并接收测量值 z_k^t 以更新各自的局部地图置信度 $\mathcal{M}_i$。无人机 i 和 k 共享测量值，若$\|\boldsymbol{p}_i^t - \boldsymbol{p}_k^t\|_2 \leq D$，其中 $D \in \mathbb{R}^+$ 是近似表示范围受限的通信信道。通信消息包含无人机标识符 i、其位置 $\boldsymbol{p}_i^t$以及其收集的测量值 z_i^t。接收无人机 k 利用 z_i^t 和 p_i^t 来更新其对 $\mathcal{M}^k$的局部地图置信度。

C. 自适应地形监测的效用定义

我们的目标是为每架无人机 i 规划其在时间步 t 的未来测量位置 $\psi_i^{t+1} = (\boldsymbol{p}_i^{t+1}, \ldots, \boldsymbol{p}_i^B)$，使得从这些位置获得的下一组测量值 $\{z_i^{t+1}, \ldots, z_i^B\}$ 能最大限度地降低当前全局地图 $\mathcal{M}$ 后验信念的不确定性。全局地图 $\mathcal{M}$ 包含截至当前时间步 t 由 N 无人机采集的测量值 $z_i^{0:t}$。我们通过计算融合 z^{t+1} 后在 $\mathcal{M}$ 上的熵降低，来量化在下一时间步 $t+1$ 采集的一组测量值 $z^{t+1} = \{z_1^{t+1}, \ldots, z_N^{t+1}\}$ 所带来的不确定性降低。然后，将信息准则 $\mathrm{I}(\psi^{t+1})$ 计算为沿路径 ψ^{t+1} 的熵降低总和：

$$\mathrm{I}(\psi^{t+1}) = \sum_{m=0}^{B-t-1} H(\mathcal{M} \mid z^{0:t+m}, \boldsymbol{p}^{0:t+m}) - H(\mathcal{M} \mid z^{0:t+m+1}, \boldsymbol{p}^{0:t+m+1})\,. \quad (2)$$

我们考虑具有一个感兴趣目标类别的自适应建图，即一个具有信息价值的类别，以及一个互补类别。因此，时间步 t 的地图熵 $H(\mathcal{M} \mid z^{0:t}, \boldsymbol{p}^{0:t})$ 按感兴趣类别和不感兴趣类别的重要性加权：

$$\begin{aligned} H(\mathcal{M} \mid z^{0:t}, \boldsymbol{p}^{0:t}) &= \sum_{j=1}^{M} H(\mathcal{M}^j \mid z_{0:t}, \boldsymbol{p}_{0:t}) \\ &= -\sum_{j=1}^{M} \mathrm{W}(\mathcal{M}^j)\, p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t}) \log\left(p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t})\right) \\ &\quad + \mathrm{W}(\overline{\mathcal{M}^j})\, p(\overline{\mathcal{M}^j} \mid z^{0:t}, \boldsymbol{p}^{0:t}) \log\left(p(\overline{\mathcal{M}^j} \mid z^{0:t}, \boldsymbol{p}^{0:t})\right), \end{aligned} \quad (3)$$

其中 $p(\overline{\mathcal{M}^j} \mid z^{0:t}, \boldsymbol{p}^{0:t}) = 1 - p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t})$。重要性加权 $W(\mathcal{M}^j)$ 定义为：

$$W(\mathcal{M}^j) = \begin{cases} w_1 & \text{if } p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t}) > 0.5\,, \\ w_2 & \text{if } p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t}) < 0.5\,, \\ 0.5 & \text{else}\,, \end{cases} \quad (4)$$

其中 $w_1, w_2 \geq 0$ 和 $w_1 + w_2 = 1$。重要性权重 w_1，w_2 鼓励无人机规划路径 ψ^{t+1}，以根据当前后验地图信念 $p(\mathcal{M}^j \mid z^{0:t}, \boldsymbol{p}^{0:t})$，瞄准潜在的感兴趣区域。

IV. 我们的方法

我们提出了一种面向无人机团队的新型基于多智能体强化学习的信息路径规划方法。我们的目标是在三维工作空间中规划无人机路径，以实现协作式自适应地形测绘。如图 3所示，我们利用强化学习，根据全局地形信息对智能体进行离线训练，以集中式方式学习无人机路径。其中一个关键环节是引入反事实基线，从而能够估计每个智能体对整体团队表现的测绘贡献，并改善协作。在任务执行期间，我们利用训练得到的智能体行为，部署完全分布式系统，在线重新规划具有信息价值的测量位置。

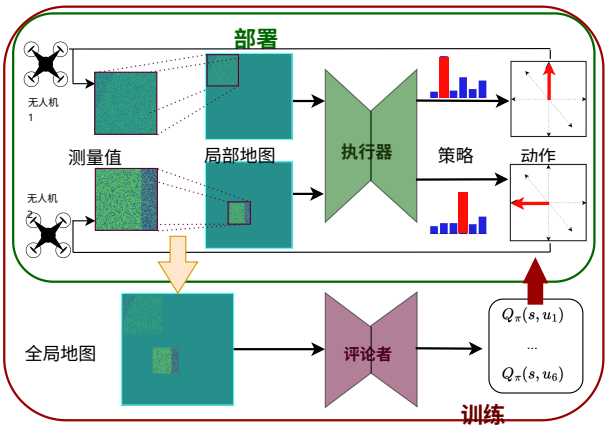


图 3：我们方法的概览。在任务期间的每个时间步，每架无人机都会进行测量并更新其局部地图状态。局部地图作为行动者网络的输入，网络会输出一个策略，并据此采样一个动作。在训练期间，还会利用全局地图信息训练集中式评论者网络，并根据当前状态为每个动作输出 Q 值，即预期未来回报。

A. 用于序贯决策的强化学习

我们将多无人机信息路径规划任务表述为由智能体团队执行的序贯决策问题，并使用强化学习对其进行训练。无人机执行任务，在每个时间步 t 每架无人机 i 根据当前的局部机载状态 ω_i^t同时规划其下一个测量位置$\boldsymbol{p}_i^{t+1}$ 。局部状态 ω_i^t 包括局部地图置信度 $\mathcal{M}_i$ ，该置信度依次融合过去的（已通信）测量值 $z^{0:t}$，以编码测量历史。这样，我们考虑了由通信约束和含噪传感器测量引起的部分可观测性。在每个时间步，所有无人机动作，即下一个测量位置，共同定义联合动作 $\boldsymbol{u}^t = (u_1^t, \ldots\ldots, u_N^t) \in U^N$。无人机获得一个全局团队奖励 R^t：$S \times U^N \times S \to \mathbb{R}$ 用于量化已测绘测量值的联合信息价值 $\{z_1^t, \ldots, z_N^t\}$ $G^t = \sum_{k=0}^{B-t} \gamma^k R^{t+k}$ 。折扣回报，其中 $\gamma \in [0,1)$ 为折扣因子，用于衡量从 t 开始直至任务预算耗尽时团队路径的信息价值。尽管每架无人机以分布式方式选择动作，我们仍将团队作为整体评估其表现，以强化协作式信息路径规划行为。

状态。全局环境状态 $s^t \in S$ 定义为 $s^t = \{\mathcal{M}^t, \boldsymbol{p}_{1:N}^t, b\}$。 $\mathcal{M}^t$ 描述当前的全局地图置信度$p(\mathcal{M} \mid z_{1:N}^{0:t}$， $\boldsymbol{p}_{1:N}^{0:t})$， $\boldsymbol{p}_{1:N}^t$表示当前智能体位置，$b \le B$ 表示剩余任务预算。 s^{t+1} 是智能体执行动作 $\boldsymbol{u}^t$后移动至新位置 $\boldsymbol{p}_{1\,N}^{t+1}$：所到达的下一状态。 $\mathcal{M}^{t+1}$ 根据在 $\boldsymbol{p}_{1\,N}^{t+1}$：处获取的测量值 z^{t+1} 进行更新。

动作。智能体在离散位置网格 P 内移动，该网格受环境边界以及离散的最低和最高高度限制。每架无人机 i 从离散三维动作空间 U 中选择一个动作 u_i，其中包含 $\{up, north, east, south, west, down\}$ 固定步长的移动。我们禁止会导致移动到环境之外或无人机具有相同二维地形坐标的动作。**奖励**。我们明确设计奖励函数 R ，以反映式（2）中定义的信息准则，从而自适应且快速地降低目标区域内的地图不确定性：

$$R^t(s^t, \boldsymbol{u}^t, s^{t+1}) = \alpha \frac{H(\mathcal{M}^{t+1}) - H(\mathcal{M}^t)}{H(\mathcal{M}^t)} + \beta\,. \qquad (5)$$

对于新的测量值 z^{t+1}，我们奖励从当前地图状态到下一地图状态 $\mathcal{M}^{t+1}$ 的加权地图熵降低。我们将奖励除以当前地图熵进行归一化，使其在任务期间的幅值近似保持恒定，并采用仿射缩放因子 α 和β 以提高训练稳定性。请注意，对于 $\gamma = 1$，回报 G^t 与式（2）中的信息路径规划准则相似，但进行了归一化并应用了缩放因子。

B. 算法

我们的目标是学习一种策略，使无人机团队能够在自适应监测任务中协同行动。为此，我们在 Foerster 等人提出的 COMA 强化学习算法基础上进行扩展[20]，以适用于无人机团队的信息素路径规划应用。COMA 是一种行动者-评论家算法，它使用集中式评论者网络评估每个智能体由策略$\pi(\cdot \mid \omega_i^t)$描述、由行动者网络参数化的行为，并据此优化策略。评论家通过估计智能体在给定其他智能体动作 $\boldsymbol{u}_{-i}^t$时的 i 状态-动作价值 $Q_\pi\big(s^t, (u_1^t, \ldots, u_i^t, \ldots, u_N^t)\big)$，来评估当前策略 π 。评论家网络通过 TD(λ) [29]进行同策略训练，以估计从当前状态 s^t 采取联合动作 $\boldsymbol{u}^t$ 、并在此后遵循当前策略 π 所获得的折扣回报 G_t；该回报在第 IV-A 节中介绍。训练期间，评论家使用全局信息 s^t ，而执行器仅利用机载信息 ω_i^t ，在训练和部署期间以分散式方式预测下一个最佳测量位置。我们利用反事实基线，根据单个智能体对团队表现的贡献为其分配功劳，从而促进团队协作行为。团队动作 $\boldsymbol{u}^t$ 中，智能体 i采取动作 u_i^t 时的优势 A_i^t 为：

$$A_i^t(s^t, \boldsymbol{u}^t) = Q_\pi(s^t, \boldsymbol{u}^t) - \sum_{u'^t_i \in U} \pi(u'^t_i \mid \omega_i^t)\, Q_\pi\big(s^t, (\boldsymbol{u}_{-i}^t, u'^t_i)\big)\,. \qquad (6)$$

智能体 i 对团队动作 $\boldsymbol{u}^t = (u_1^t, \ldots, u_i^t, \ldots, u_N^t)$的联合状态-动作价值 $Q_\pi(s^t, \boldsymbol{u}^t)$ 的贡献，通过执行动作 u_i^t 进行估计：对所有可能的个体动作 $u'^t_i \in U$ 求边际化，同时保持其他智能体的动作 $\boldsymbol{u}_{-i}^t$ 不变。为优化 $\pi(\cdot \mid \omega_i^t)$，我们使用 公式 (6)应用策略梯度定理，并最小化：

$$\mathcal{L} = -\log \pi(u_i^t \mid \omega_i^t)\, A_i^t(s^t, \boldsymbol{u}^t)\,, \qquad (7)$$

使用小批量随机梯度下降。更多详情请参阅 Foerster 等人 [20]。

C. 网络架构与特征设计

我们提出了新的行动者和评论者表示方式，以便在三维机器人应用中利用 COMA。如图 4所示，行动者和评论者由神经网络 f_{θ^π}和 f_{θ^c}表示，其参数分别为 θ^π 和 θ^c。行动者网络以智能体 i 及其局部状态 ω_i^t为条件。特征

输入包括智能体标识符 i、剩余任务预算 b以及以下空间输入：(a) 以智能体位置为中心的位置图，其中编码了地图边界和已通信的其他智能体的位置，图中数值表示智能体高度；(b) 局部地图状态 $\mathcal{M}_i$；(c) 局部地图状态的加权熵$H(\mathcal{M}_i \mid z^{0:t}, \boldsymbol{p}^{0:t})$，见公式 (3)；(d) 测量值 $H(z_i^t \mid \boldsymbol{p}_i^t)$的加权熵；以及 (e) 通信范围内所有智能体视野当前覆盖的地图单元格（“覆盖范围图”）。

我们的评论家网络接收相同的输入 (a)-(e)，此外还会接收训练期间可获取的全局环境信息。具体而言，它还会接收：(f) 对所有智能体位置进行编码的全局位置图 $\boldsymbol{p}_{1:N}^t$；(g) 全局地图状态 $\mathcal{M}$；(h) 其加权熵 $H(\mathcal{M} \mid z^{0:t}, \boldsymbol{p}^{0:t})$；(i) 当前处于所有智能体视野范围内的地图单元格；以及，为了在公式 (6)中学习反事实基线，(j) 其他智能体的动作。输入采用位置网格分辨率 r_P。我们将地图分辨率 r_M下采样 $\frac{r_P}{r_M}$ 倍，以使两种分辨率一致。标量输入 i 和 b会扩展为数值恒定的二维特征图。

为处理地形监测任务所需的空间信息，这些网络由卷积编码器和多层感知机头组成，分别用于预测策略π和Q值 Q_π。图 4 上部展示了两个网络的架构。执行器的逻辑值 $f_{\theta^\pi}(x_i^t)$以特征图 (a)-(e) 的集合 x_i^t 为输入，经过有界 softmax 函数 $\pi(u_i \mid x_i^t) = (1-\epsilon)\,\mathrm{softmax}\big(f_{\theta^\pi}(x_i^t)\big) + \frac{\epsilon}{|U|}$，以预测随机策略。超参数 $\epsilon \in [0,1]$ 有助于训练期间的探索，在部署时设为 0。经过最后一个线性层后，评论家会针对智能体 i的每个动作 $u_i^t \in U$输出 Q 值 $Q_\pi\big(s^t, (u_1^t, \ldots, u_i^t, \ldots, u_N^t)\big)$。

D. 网络训练

我们离线训练一个执行器和一个评论家网络，并在部署时在线使用训练好的执行器。我们模拟无人机任务，并从将测量结果 z 融合到全局地图 $\mathcal{M}$后获得的奖励中学习。每次任务前，我们生成一个分辨率为 r_M 的新地形。这些地形被划分为相连的感兴趣区域和互补的非感兴趣区域。划分方向是随机的，感兴趣区域覆盖地形的30%至60%，以促进泛化。我们固定初始无人机位置并执行

一次任务，直至预算 B 耗尽。训练期间，按照第 IV-C 节所述，从执行器的策略中采样动作，其中 ϵ 在前0.5 至0.02 的10,000 次任务中线性递减。

我们交替进行3，000 次同策略环境交互与策略优化，采用公式（7）。两个网络均通过随机小批量梯度下降进行优化，共训练5 个周期，使用 Adam优化器，学习率分别为 $1e^{-5}$ （执行器）和 $1e^{-4}$ （评论家），批量大小为600。评论家网络通过应用 TD（λ），并结合 $\lambda = 0.8$ 和 $\gamma = 0.99$ 进行训练，以估计期望回报 G_t；目标评论家网络每经过 30，000次环境交互，就复制一次评论家网络。

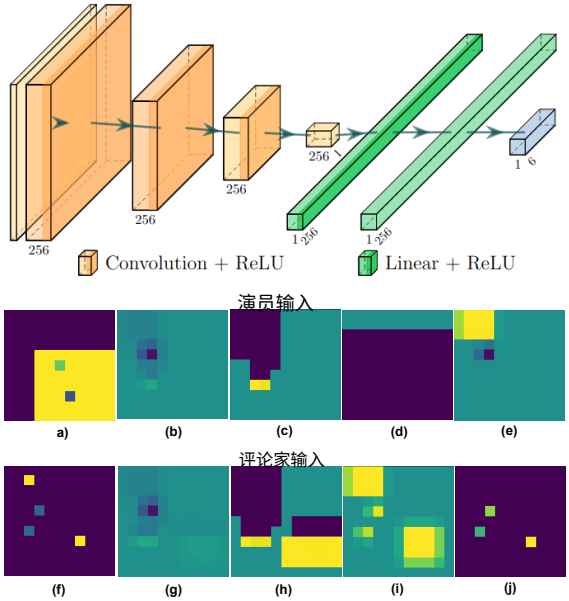


图4：我们的执行器和评论家的架构及输入。两个网络均由卷积编码器和预测头中的线性层组成。对于执行器，输出由softmax层（灰色）生成。执行器和评论家均接收局部输入（“演员输入”）。评论家还接收用于集中式训练的全局输入（“评论家输入”）。所有输入(a)-(j)详见第 IV-C 节。

V. 实验结果

我们通过实验展示了所提出的基于多智能体强化学习的信息路径规划方法在利用无人机团队进行自适应地形监测方面的能力。实验结果支持以下结论：（i）我们设计的网络表示对于无人机在三维工作空间中的多智能体信息路径规划是有效的；（ii）通过反事实基线解决信用分配问题能够提升规划性能；（iii）在不同团队规模和通信约束下，就规划性能而言，我们的方法优于不基于学习的最先进方法。此外，我们还展示了该方法在真实世界地表温度监测场景中的应用。

A. 实验设置

对于每项实验，我们按照第 IV-D 节所述，在不断变化的感兴趣区域中执行 50 次地形监测任务。地形尺寸为 50 m $\times$ 50m，地图分辨率为 $r_M = 10$cm。我们将规划分辨率设置为 $r_P = 5$m，将高度限制在 5 m 至 15 m 之间，并使用视场为 60°的相机，从而确保在最低高度进行测量时，相邻测量值不会重叠。为考虑较高高度下增加的传感器噪声，我们模拟的 $p(z_i^t \mid \mathcal{M}_i^j, \boldsymbol{p}_i^t)$ 在 $\{5, 10, 15\}$ m 高度时为 $\{0.99, 0.735, 0.625\}$。无人机团队由 4 个智能体组成，除非另有说明，否则通信半径限制为 25 m。作为指标，我们使用地图状态的熵 $H(\mathcal{M} \mid z^{0:t}, \boldsymbol{p}^{0:t})$ 来评估地图不确定性，并使用地图状态 $\mathcal{M}$ 与真实地图之间的 F1分数来评估 $\mathcal{M}$ 的正确性。我们报告

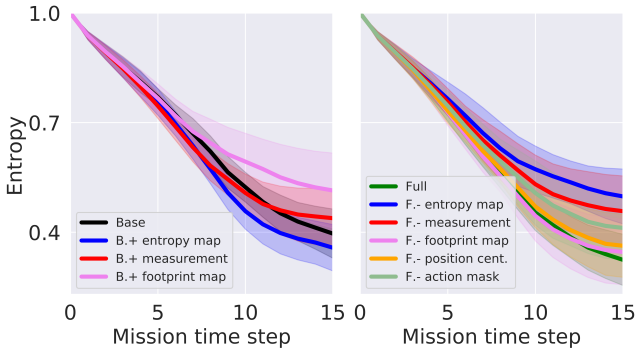


图 5：特征输入消融。我们系统地添加（左）和移除（右）方法中的组成部分，并展示任务时间内地图熵降低的变化。结果表明，熵地图输入的影响最大，而覆盖范围图仅在任务末期有助于提升性能。我们提出的完整网络输入（绿色）表现最佳。

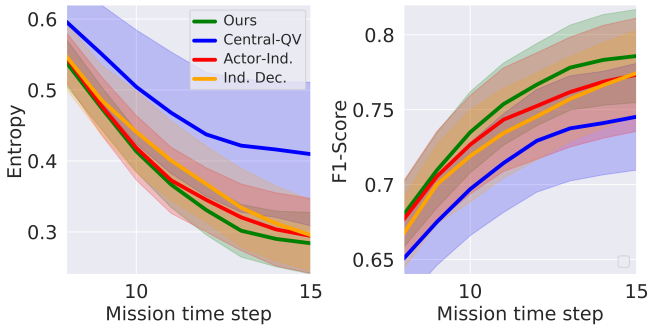


图 6：信用分配研究。我们比较了 COMA方法（绿色）与未针对地图熵（左）和 F1分数（右）中的信用分配问题进行显式处理的变体。在任务后期，当环境中的大部分区域都已探索过时，我们的方法能够更好地协同进行自适应建图。

在给定 $B = 15$ 测量值的真实值感兴趣区域中，这些性能指标的均值和标准差。

B. 网络特征设计消融研究

本节通过消融实验表明，我们提出的网络输入表示对三维工作空间中的多无人机信息素路径规划有效。我们系统地逐一添加和移除输入特征图，以评估它们对整体规划性能的影响。

图 5 展示了任务期间采用不同网络输入特征时，平均地图熵降低量及其标准差。曲线下降越陡，规划性能越好。在 图 5-左图中，我们在基础输入特征（a）—（d）上添加单个输入特征图，具体如第 IV-C 节所述；这些特征是对信息路径规划问题进行建模所必需的。在图 5-右图中，我们移除单个输入特征图。

如预期，所提出的完整输入（绿色）表现最佳，即最终地图熵最低。熵地图特征（蓝色）对规划性能的提升最为显著，这表明它包含了学习信息丰富路径所需的大部分相关信息。有趣的是，加入覆盖范围图（粉色）会略微损害规划性能，但从输入中移除它又会导致最终性能变差。这说明，与其他输入结合时，覆盖范围图在任务后期的贡献更大。直观地说，这是因为随着测量数据不断映射，识别当前已观测的地图单元格会变得更加困难，而这对于规划下一步的测量位置至关重要。加入当前测量的熵（红色）会在最初的 ~ 12个任务时间步内更快降低不确定性，但会损害最终性能，推测是因为这一局部特征造成了短视规划偏差。然而，作为包含足够全局特征的完整输入的一部分，该特征对规划性能有显著作用，能够提供邻近区域的额外地图信息。此外，不屏蔽无效动作（浅绿色）以及不对智能体位置图进行居中处理（橙色）都会损害性能，因为这会移除规划所需的宝贵空间信息。总之，结果证实，我们的网络输入特征（“完整”）最适用于三维工作空间中无人机的多智能体信息路径规划，使我们提出的基于强化学习的方法相比可能的变体展现出更优异的规划性能。

C. 信用分配机制研究

本节实验表明，考虑信用分配问题能够提升规划性能，这验证了在基于强化学习的协作式多智能体信息素路径规划中采用显式信用分配机制的必要性。我们系统地研究了无人机团队在式（6）中采用不同优势函数时的信息素路径规划性能，并由此改变式（7）中的策略梯度更新。为此，我们将所提出的方法（第 IV-B 节）与其自身的三个变体进行比较，具体如下。Central-QV。与 Foerster 等人 [20]，类似，我们在自适应信息素路径规划场景中，通过将式（6）中的基线替换为状态价值 $V_\pi(s^t)$，验证反事实基线的有效性。这样，我们不再估计单个智能体的贡献，而只考虑团队的联合性能。我们训练两个评论家网络，分别估计Q 和 V，二者均如第 IV-D 节所述，并将优势函数改为 $A_i^t(s^t, \boldsymbol{u}^t) = Q_\pi(s^t, \boldsymbol{u}^t) - V_\pi(s^t)$。

独立演员。 如第 IV-B 节所述，我们在训练期间利用全局状态信息，采用集中式评论家。然而，为了研究推理其他团队成员动作所产生的影响，我们现在不考虑这些动作，并将其从评论家网络输入中排除。我们调整优势函数，使其仅取决于智能体自身的动作： $A_i^t(s^t, u_i^t) = Q_\pi(s^t, u_i^t) - \sum_{u_i'^t \in U} \pi(u_i'^t \mid \omega_i^t) Q_\pi(s^t, u_i'^t)$。

分散式。在此变体中，我们移除所有全局信息，仅考虑基于局部智能体信息 ω_i^t 的分散式评论家。与独立演员变体类似，评论家使用该智能体自身的状态价值作为基线，从而忽略其他智能体的动作。优势函数简化为： $A_i^t(\omega_i^t, u_i^t) = Q_\pi(\omega_i^t, u_i^t) - \sum_{u_i'^t \in U} \pi(u_i'^t \mid \omega_i^t) Q_\pi(\omega_i^t, u_i'^t)$。

图 6 展示了使用不同优势函数时的规划性能。我们重点关注任务后期的性能，此时环境的大部分区域已

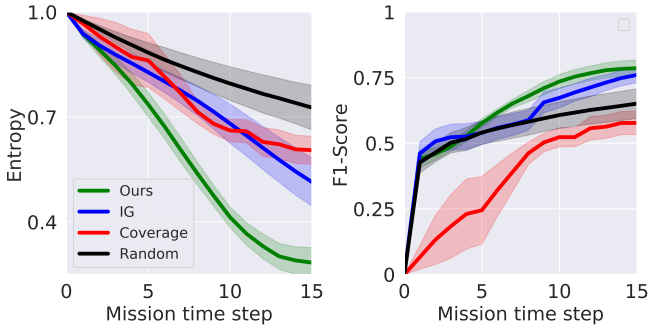


图 7：4 个智能体、通信范围为 $25m$时的规划方法比较。通过协同规划路径，我们的方法能最快降低地图不确定性（左图）并提高地图精度（右图），表明其性能更优。

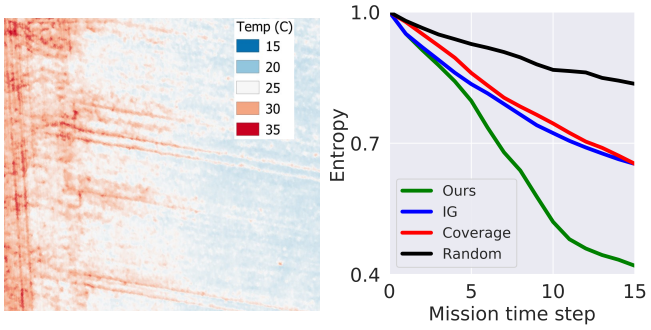


图 8：真实世界评估。我们将方法应用于表面温度数据（左），并比较所考察规划方法的地图熵降低情况。较暖的区域被视为更值得关注。尽管我们的方法是在合成数据上训练的，但其表现优于所有其他方法。图 1 展示了本实验中规划得到的无人机路径。

被探索，协作对于自适应建图至关重要。我们的方法中采用的反事实基线（绿色）表现最佳，表明它有助于实现协同行为。所有变体都缺少明确的信用分配机制，无论采用集中式还是分散式评论家，都会对规划性能产生不利影响。这证实了我们的基于 COMA 的算法通过解决信用分配问题提升了规划性能。

D. 与非学习方法的比较

以下实验支持我们的主张：基于强化学习的方法优于最先进的非学习型多智能体信息素路径规划方法。我们将本方法与三种方法进行比较：（i）Carbone 等人提出的一种用于无人机群的自适应信息增益方法。[2]，该方法通过最大化估计的地图熵降低量，贪心地选择无人机动作，且不进行显式的信用分配（‘IG’）；（ii）非自适应覆盖式割草机路径模式，在表现最佳的高度上设置等距（5米）测量位置（‘覆盖率’）；（iii）非自适应随机探索，以均匀随机方式采样无人机动作（‘Random’）。所有方法使用相同的状态空间、动作空间和动作屏蔽策略。

图 7 展示了所考察方法在任务时间内的评估指标。基于信息增益的策略（蓝色）和我们基于强化学习的方法（绿色）优于非自适应方法，因为它们能够主动聚焦于感兴趣区域。我们方法的优异表现证实了解决信用分配问题的益处，也证明了基于学习的规划方法适用于不同地形。

接下来，我们展示基于学习的方法对不同团队规模和通信设置的泛化能力。我们将针对通信半径为 25m 的 4智能体设置训练的执行器部署到包含 2、4 和 8 个智能体的场景中，并设置通信半径为 0m、25 m 或无限制通信，且不进行重新训练。表 I展示了我们的方法与基于信息增益的方法和覆盖率方法相比的规划性能。正如预期，智能体数量越多，建图速度越快，即熵降低、F1分数提高。尽管改变通信半径会导致性能略有下降，我们的方法始终优于这两种方法。这证实了我们的强化学习方法能够适应不同数量的智能体和通信要求，无需重新训练，展现出广泛的适用性，且不增加额外的训练成本。

E. 温度制图场景

我们使用真实世界数据，在地表温度监测场景中展示了我们方法的性能。数据来自德国 J40 m ×40m 农田附近¨ulich。该农田数据由搭载 Vue Pro R 640 热传感器的 DJI Matrice 600 无人机采集，见 图 8-左侧。我们将地形离散化为分辨率为 500×500的网格地图，分辨率为 $r_M = 8$厘米。规划网格的分辨率为 $r_P = 4$米，以确保网络输入维度与训练时相同。地表温度为 $\geq 25°$ 摄氏度的区域被视为值得关注，可用于自适应热点制图。图 8-右侧 展示了采用我们的方法时，$B = 15$ 次测量下，4 个智能体在通信半径为 25m 时的地图熵降低情况，并与 第 V-D 节介绍的方法进行了比较。请注意，如 第 IV-D 节所述，我们的方法仅在模拟环境中使用合成数据进行训练。我们的方法明显优于所有其他方法，展示了其应用于真实世界地形监测任务的能力。

VI. 结论与未来工作

本文提出了一种新颖的多智能体深度强化学习信息素路径规划方法，利用无人机团队开展自适应地形监测。该方法采用了新的网络表示，并利用反事实基线解决三维工作空间中协同无人机路径规划的信用分配问题。因此，我们的方法在监测效率方面成功超越了最先进的非学习式方法，同时无需重新训练即可适应不同的任务设置。使用无人机采集的热数据进行的实验验证了该方法在现实世界中的适用性。未来的工作将研究异构机器人团队，以及边界未知环境中的动态扩展地图。

表 I：在不同团队规模和通信约束下的鲁棒性。我们报告任务时间经过 33%、66% 和 100% 后，50 次试验中熵和 F1分数的均值和标准差。最佳结果以粗体标出。

设置	方法	33% 熵 ↓	67% 熵 ↓	100% 熵 ↓	33% F1 ↑	67% F1 ↑	100% F1 ↑
2 个智能体	Ours	**0.8826** ± 0.0267	**0.7343** ± 0.0548	**0.5802** ± 0.0451	**0.3608** ± 0.0661	**0.5035** ± 0.0768	**0.6268** ± 0.0436
	IG	0.9155 ± 0.0154	0.8449 ± 0.0410	0.7432 ± 0.0801	0.3275 ± 0.0386	0.4353 ± 0.0539	0.5340 ± 0.0857
	覆盖率	0.9163 ± 0.0472	0.7845 ± 0.0628	0.6970 ± 0.0332	0.1603 ± 0.0871	0.3687 ± 0.0912	0.4863 ± 0.0418
4 个智能体	Ours	**0.7360** ± 0.0349	**0.4137** ± 0.0294	**0.2842** ± 0.0408	**0.5765** ± 0.0168	**0.7350** ± 0.0268	**0.7858** ± 0.0308
	IG	0.8302 ± 0.0298	0.6805 ± 0.0554	0.5176 ± 0.0700	0.5396 ± 0.0435	0.6728 ± 0.0375	0.7599 ± 0.0289
	覆盖率	0.8615 ± 0.0762	0.6614 ± 0.0317	0.6052 ± 0.0396	0.1603 ± 0.0870	0.3687 ± 0.0913	0.4864 ± 0.0418
8 个智能体	Ours	**0.5621** ± 0.0235	**0.2952** ± 0.0403	**0.2209** ± 0.0459	**0.7115** ± 0.0336	0.7859 ± 0.0322	0.8171 ± 0.0283
	IG	0.6887 ± 0.0234	0.4886 ± 0.0460	0.3077 ± 0.0446	0.6957 ± 0.0260	**0.7881** ± 0.0194	**0.8576** ± 0.0151
	覆盖率	0.8185 ± 0.0832	0.6072 ± 0.0274	0.5149 ± 0.0309	0.2455 ± 0.1281	0.5246 ± 0.0358	0.5800 ± 0.0433
零通信	Ours	**0.7638** ± 0.0408	**0.5445** ± 0.0674	**0.3699** ± 0.0486	**0.5476** ± 0.0256	**0.6846** ± 0.0327	**0.7620** ± 0.0282
	IG	0.8284 ± 0.0283	0.6814 ± 0.0529	0.5286 ± 0.0775	0.5396 ± 0.0453	0.6510 ± 0.0442	0.7151 ± 0.0432
有限通信 我们的方法		**0.7360** ± 0.0349	**0.4137** ± 0.0294	**0.2842** ± 0.0408	**0.5765** ± 0.0168	**0.7350** ± 0.0268	**0.7858** ± 0.0308
	IG	0.8302 ± 0.0298	0.6805 ± 0.0554	0.5176 ± 0.0700	0.5396 ± 0.0435	0.6728 ± 0.0375	0.7599 ± 0.0289
完整通信	Ours	**0.7428** ± 0.0276	**0.4623** ± 0.0602	**0.3674** ± 0.0843	**0.5818** ± 0.0294	**0.7077** ± 0.0481	0.7524 ± 0.0465
	IG	0.8294 ± 0.0290	0.6809 ± 0.0501	0.5266 ± 0.0593	0.5389 ± 0.0449	0.6721 ± 0.0432	**0.7606** ± 0.0306

参考文献

[1] D. Albani, T. Manoni, D. Nardi, and V. Trianni, “动态无人机集群部署以实现非均匀覆盖”，载于自主智能体与多智能体系统国际会议（*AAMAS*）论文集，2018年。[2] C. Carbone, D. Albani, F. Magistri, D. Ognibene, C. Stachniss, G. Kootstra, D. Nardi, and V. Trianni, “基于信息增益的无人机群农田监测与制图”，载于分布式自主机器人系统。Springer，2022年，第306–319页。[3] W. H. Maes and K. Steppe, “精准农业中使用无人机进行遥感的前景”，《植物科学趋势》，第24卷，第152–164页，2019年。[4] E. Bondi, D. Dey, A. Kapoor, J. Piavis, S. Shah, F. Fang, B. Dilkina, R. Hannaford, A. Iyer, L. Joppa, and M. Tambe, “AirSim-W：面向无人机野生动物保护的仿真环境”，载于*ACM*计算与可持续社会*SIGCAS*会议论文集，2018年。[5] M. A. Goodrich, B. S. Morse, D. Gerhardt, J. L. Cooper, M. Quigley, J. A. Adams, and C. Humphrey, “使用配备摄像头的小型无人机支持荒野搜寻与救援”，《现场机器人学期刊》（*JFR*），第25卷，第1–2期，第89–110页，2008年。[6] S. Hayat, E. Yanmaz, T. X. Brown, and C. Bettstetter, “面向搜寻与救援的多目标无人机路径规划”，载于*IEEE*机器人与自动化国际会议（*ICRA*）论文集，2017年。[7] Z. Yan, N. Jouandeau, and A. A. Cherif, “多机器人协同综述与分析”，《国际先进机器人系统期刊》，第10卷，第12期，第399页，2013年。[8] R. B¨ahnemann, D. Schindler, M. Kamel, R. Siegwart, and J. Nieto, “用于搜索、拾取和转移物体的分布式多智能体无人机系统”，载于国际机器人研究研讨会（*ISRR*）论文集，2017年。[9] G. A. Hollinger, B. Englot, F. S. Hover, U. Mitra, and G. S. Sukhatme, “水下检测的主动规划及自适应性的益处”，《国际机器人研究期刊》（*IJRR*），第32卷，18，2013年3页.–。[10] M. Tzes, N. Bousias, E. Chatzipantazis, and G. J. Pappas, “用于多机器人主动信息获取的图神经网络”，*arXiv*预印本 *arXiv:2209.12091*，2022年。[11] J. R¨uckin, L. Jin, and M. Popovi´c, “基于深度强化学习的自适应信息路径规划：用于基于无人机的主动感知”，载于*IEEE*机器人与自动化国际会议（*ICRA*）论文集，2022年。[12] A. Viseras and R. Garcia, “DeepIG：基于深度强化学习的多机器人信息收集”，《*IEEE*机器人与自动化快报》（*RA-L*），第4卷，第3期，第3059–3066页，2019年。[13] H. Bayerlein, M. Theile, M. Caccamo, and D. Gesbert, “基于深度强化学习的无线数据采集多无人机路径规划”，《*IEEE*通信学会开放期刊》（*OJ-COMS*），第2卷，第1171–1187页，2021年。[14] F. Chen, J. D. Martin, Y. Huang, J. Wang, and B. Englot, “通过深度强化学习实现不确定环境下的自主探索图，”载于*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）论文集，2020年。[15] T. Fan、P. Long、W. Liu 和 J. Pan，“面向复杂场景导航的深度强化学习多机器人分布式避碰”，国际机器人研究期刊（*IJRR*），第39卷，第7期，第856–892页，2020年。[16] S. Y. Luis、M. P. Esteve、D. G. Reina 和 S. T. Marın，“将深度强化学习应用于环境任务中的多智能体信息路径规划。”[17] Y. Cao、Y. Wang、A. Vashisth、H. Fan 和 G. A. Sartoretti，“CAt-NIPP：用于信息路径规划的上下文感知注意力网络”，载于机器人学习会议（*CoRL*）论文集。[18] M. Lodel、B. Brito、A. Serra-G´omez、L. Ferranti、R. Babuˇska 和 J. Alonso-Mora，“接下来该看哪里：学习用于信息轨迹规划的视点建议”，载于*IEEE*机器人与自动化国际会议（*ICRA*）论文集。IEEE，2022年。[19] A. K. Agogino 和 K. Tumer，“统一时间信用分配与结构信用分配问题”，载于自主智能体与多智能体系统国际会议（*AAMAS*）论文集，2004年。[20] J. Foerster、G. Farquhar、T. Afouras、N. Nardelli 和 S. Whiteson，“反事实多智能体策略梯度”，载于人工智能进展会议（*AAAI*）论文集，2018年。[21] M. Popovi´c、T. Vidal-Calleja、G. Hitz、J. J. Chung、I. Sa、R. Siegwart 和 J. Nieto，“基于无人机的地形监测信息路径规划框架”，自主机器人，第44卷，第889–911页，2020年。[22] A. Blanchard 和 T. Sapsis，“用于环境探索与监测中异常检测的信息路径规划”，海洋工程，第243卷，文章编号110242，2022年。[23] G. Best、O. M. Cliff、T. Patten、R. R. Mettu 和 R. Fitch，“Dec-MCTS：用于多机器人主动感知的分布式规划”，国际机器人研究期刊（*IJRR*），第38卷，第316–337页，2019年。[24] Q. Li、F. Gama、A. Ribeiro 和 A. Prorok，“用于分布式多机器人路径规划的图神经网络”，载于*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）论文集，2020年。[25] A. Pirinen、A. Samuelsson、J. Backsund 和 K. ˚Astr¨om，“基于强化学习的俯视图目标定位”，*arXiv*预印本 *arXiv:2209.03694*，2022年。[26] H. Qie、D. Shi、T. Shen、X. Xu、Y. Li 和 L. Wang，“基于多智能体强化学习的多无人机目标分配与路径规划联合优化”，*IEEE Access*，第7卷，第146 264–146272页，2019年。[27] A. Puente-Castro、D. Rivero、A. Pazos 和 E. Fernandez-Blanco，“面向野外勘探的基于强化学习的无人机集群路径规划”，应用智能，第52卷，第14 101–14118页，2022年。[28] A. Elfes，“利用占据栅格实现移动机器人感知与导航”，*Computer*，第22卷，第6期，第46–57页，1989年。[29] R. S. Sutton 和 A. G. Barto，强化学习：导论。麻省理工学院出版社，2018 年。