IMAGINE：基于Godot的智能多智能体室内网络化探索

Tiago Leite[a, *], Maria Inês Conceição[a, b, c], António Grilo[a]

[a]*INESC INOV，Instituto Superior Técnico*，里斯本大学，里斯本，*1000-029*，葡萄牙 *b INESC ID–Instituto de Engenharia de Sistemas e Computadores: Investigação e Desenvolvimento*，*Instituto Superior Técnico*，里斯本，*1000-029*，葡萄牙[c]机器人与系统研究所，*Instituto Superior Técnico*，里斯本大学，里斯本，*1049-001*，葡萄牙

arXiv:2602.02858v1 [cs.RO] 2 Feb 2026

摘要

由一组具备自主能力、通信感知能力且能够协作的无人机（UAV）探索未知且全球导航卫星系统（GNSS）受限的环境，在协调、感知和分散式决策方面面临重大挑战。本文采用多智能体强化学习（MARL），在二维室内环境中应对这些挑战，并使用基于Godot的高保真游戏引擎仿真和连续动作空间。策略训练旨在通过网络分布式部分可观测马尔可夫决策过程（ND-POMDP），实现涌现式协作行为，并在不确定性下进行决策。每架无人机均配备激光雷达传感器，并可与相邻智能体共享数据（传感器测量值和局部占据栅格地图）。智能体间通信受到通信范围、带宽和延迟的限制。我们通过广泛的消融研究评估了多智能体强化学习训练范式、奖励函数、通信系统、神经网络（NN）架构、记忆机制和部分可观测马尔可夫决策过程形式化。本文共同应对了先前研究中的若干关键局限，包括依赖离散动作、单智能体或集中式形式化、假设具备先验知识且连接始终畅通、无法处理动态障碍物、规划时域较短，以及循环神经网络/Transformer架构复杂等问题。结果表明，可扩展的训练范式结合简化的架构，能够实现室内区域的快速自主探索。课程学习的实施（设置五个复杂度逐渐提高的关卡）也使训练更快、更稳健。在三个关卡中，区域覆盖率超过95%；采用卷积神经网络（CNN）则使系统性能提高了20%。这种高保真仿真、多智能体强化学习形式化与计算效率的结合，为在实体机器人系统中部署学习得到的协作策略奠定了坚实基础。

关键词：
多智能体强化学习、Godot、自主探索、未知环境、课程学习

1. 引言

无人机最初用于军事用途，如今已在众多领域日益普及，包括商业用途、安保 [1],监视 [2], 电影制作 [3], 农业 [4], 灾害管理 [5], 以及个人用途 [6]。

无人机团队的部署近年来备受关注，无人机之间的协作是一种极具价值、但颇具挑战性的能力。这一理念受到自然系统的启发，在自然系统中，协作与去中心化是解决复杂任务的必要条件。

本研究立足于利用机器人集群开展任务的背景，尤其是无人机以及可能包括微型飞行器（MAV），用于室内探索。

具体而言，问题的形式化设定是：多个分布式智能体探索一个先验未知的室内二维环境，同时在现实网络约束下彼此通信。

由于环境不确定性和通信约束，协作式多智能体探索颇具挑战。表 1 对最相关的已有研究进行了比较总结。不同方法范式各有取舍。**基于前沿的方法** 将智能体引导至未知区域边界 [7]，并通过地图共享进行协同 [8, 9, 10]，，但可能陷入局部视野。**基于规划的方法** 采用单元分解 [11, 12, 13] 实现系统化覆盖，但需要全局模型。**基于采样的方法**（如快速探索随机树（RRT） [14, 15]）能够探索复杂空间，但通信负担较重。

基于强化学习的方法学习探索策略，要么以两

*通讯作者。电子邮箱： tiagomleite@tecnico.ulisboa.pt (Tiago Leite), ines.conceicao@tecnico.ulisboa.pt (Maria Inês Conceição), antonio.grilo@inov.pt (António Grilo)

阶段方式结合经典规划器 [16] ，或以端到端方式 [17,18]。然而，多智能体强化学习（MARL）会带来协调挑战，这些挑战通常借助计算代价高昂的技术 [19, 20, 21]来解决。在完全不确定性下，通信高效的多智能体策略仍存在显著缺口。

本研究在一个统一框架中整合了当前最先进方法中已有的若干方面，这可视为一项贡献：

(1) 如 [8] 所示，将关注点分离应用于建图和决策。(2) 借鉴 [10], ，在地图层面利用协调机制，并采用如 [20, 21], 所示的传感器信息，从而避免依赖计算开销较大的技术。(3) 基于端到端深度强化学习（DRL）范式，并将其扩展到多智能体领域，采用如 [18], 所示的连续动作空间，而非 [17] 中的离散空间。(4) 将 [17, 18] 中的策略融入奖励设计。

尽管如此，本研究也带来了新的贡献：

(1) 针对探索任务中的无人机群协调采用多智能体强化学习范式，弥补了当前文献中的这一空白。(2) 通过实现更现实的通信模型（包括数据速率和距离限制），评估通信对任务协调的影响。(3) 本研究没有在部分可观测情形下默认使用循环神经网络（RNN），而是成功采用信念马尔可夫决策过程（MDP）形式化，为状态估计提供了更有条理且有效的方法。(4) 本研究力求在保持计算效率的同时兼顾物理真实性，通过开源游戏引擎 Godot [22], 模拟环境；据作者所知，这是该引擎首次应用于这一领域。Godot 提供了一个轻量级开源平台，可进行高保真、游戏引擎级别的模拟，并能准确模拟静态和动态物体的碰撞。(5) 使用工业级工具（Ray/RLlib、WandB），并在高性能计算（HPC）集群上进行验证，以确保可复现性和可扩展性。

本文其余部分的结构如下。第 2 节介绍网络分布式部分可观测马尔可夫决策过程（非分散式部分可观测马尔可夫决策过程）的理论背景。第 3 节描述所提出的方法及其实际实现。第 4 节介绍实验结果。最后，第 5 节总结主要结论并指出未来的研究方向。

2. 网络分布式部分可观测马尔可夫决策过程

标准的多智能体强化学习数学框架是分散式部分可观测马尔可夫决策过程（Dec-POMDP），其定义为元组 $(\mathcal{N}, \mathbb{S}, \mathbb{A}, \mathbb{O}, O, T, R, \gamma)$。通过使用图 $\mathbb{G}$ 对智能体之间的通信约束建模，可将 Dec-POMDP 进一步扩展为非分散式部分可观测马尔可夫决策过程。因此，非分散式部分可观测马尔可夫决策过程定义为元组 $(\mathcal{N}, \mathbb{S}, \mathbb{A}, T, R, \mathbb{O}, O, \mathbb{G}, \gamma)$：

- $\mathcal{N}$ 是智能体集合，索引为 $i \in \mathcal{N}$，其中智能体总数为 $n = |\mathcal{N}|$。
- $\mathbb{S}$ 是状态空间，即所有可能状态的集合。
- $\mathbb{A}$，其中 $\mathbb{A}_i$ 是智能体 i 可用的动作集合，各智能体的动作集合可能不同，而 $\mathbb{A} = \mathbb{A}_1 \times \mathbb{A}_2 \times \cdots \times \mathbb{A}_{|\mathcal{N}|}$ 是联合动作空间。在每个时间步 t，每个智能体 i 执行一个动作[1] a_{ti}，从而在每个时间步 [23]产生一个联合动作 $a = (a_1, \ldots, a_{|\mathcal{N}|})$。
- T： $\mathbb{S} \times \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ 是状态转移概率函数 $T(s', s, a) = P(s'|s, a)$，描述在当前状态 s 和联合动作 $a = (a_1, a_2, \ldots, a_n)$ 下转移到状态 s' 的概率。
- R： $\mathbb{S} \times \mathbb{A} \to \mathbb{R}$ 是共享奖励函数 $R(s, a, s')$，根据当前状态 s、联合动作 a 以及环境转移到状态 s' 的情况，为所有智能体提供相同的标量奖励。
- $\mathbb{O}$，其中 $\mathbb{O}_i$ 是智能体 i 可用的观测集合，每个智能体的观测集合可能不同，而 $\mathbb{O} = \mathbb{O}_1 \times \mathbb{O}_2 \times \cdots \times \mathbb{O}_{|\mathcal{N}|}$ 是联合观测空间。
- O： $\mathbb{O} \times \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ 是联合观测概率函数 $O(o, s', a) = P(o|s', a)$，描述在下一个状态 s' 和联合动作 a 条件下，观测到 $o = (o_1, o_2, \ldots, o_{|\mathcal{N}|})$ 的概率。
- $\mathbb{G}$：所有可能通信图的集合，其中每个图 $g = (V, E)$ 表示给定时刻的通信拓扑。顶点集 V 对应于智能体 $\mathcal{N}$，边集 E 表示通信链路。在每个时间步 t，当前通信

[1] 本文采用一种约定：当时间步下标的含义明确时，可将其省略。例如，智能体 $i \in \mathcal{N}$ 在时间 t 采取的动作，既可显式表示为 a_i^t，也可隐式表示为 a_i。类似地，本文使用相对时间记法，其中 s 表示当前状态，a 表示所选的联合动作，s' 表示后继状态。这样，奖励函数便可简洁地表示为 $R(s, a, s')$，而不必写作 $R(s^t, a^t, s^{t+1})$。

表 1：多智能体探索领域相关关键工作的比较。

Ref	Name	问题	解决方案	MA
[17]	自学习 DRL	通过寻找最佳 感知动作来探索未知地图。	使用和不使用循环神经网络的深度Q学习。一个 CNN 根据局部 地图以及局部地图完全已知时的 Frontier Rescue，输出最佳感知动作 是完全已知的。	no
[9]	i-VFM	减少多智能体的重复探索	状态表示（地图），称为 i-VFM, 它整合了智能体此前 访问的次数。	yes
[8]	多智能体八叉树地图	未知地下环境的自主探索 地下环境。	使用共享地图进行导航 基于图和基于前沿的规划。	yes
[10]	Multi S-Graphs	协同高效地构建地图 使用多个智能体。	语义图。	yes

缩写：*MA* - 多智能体。

图[2] $g_t \in \mathbb{G}$ 包含一条边（i，j），当且仅当智能体 i与 j 在时刻 t 能够通信。

- $\gamma \in [0, 1]$ 是折扣因子，用于决定未来奖励的重要性。

解决这些复杂多智能体问题的两种主要方法是：（i）将其表述为分布式约束优化问题（DCOPs），或（ii）借助图神经网络（GNNs）利用通信图的结构，使智能体能够高效地聚合并处理来自其邻居的信息 [20, 24, 21]。

本文提出的方法不会采用上述任一方案。它将通过允许相邻智能体共享观测来简化该问题。这样既能避免计算复杂性，又能仍然受益于智能体之间的通信与协调。

3. 提出的方法

所实现的系统架构由两个核心组件组成：基于 Godot 的环境仿真，以及使用 Ray/RLlib 的基于 Python 的学习算法，二者通过 Godot RL Agents 库连接，如图 1 所示。完整实现可在 GitHub 仓库 [25] 中获取。

3.1. 架构

Python 端实现为大规模实验奠定了基础：

- 单次试验：可通过包含所有超参数（例如训练批次大小、智能体数量、范式等）的 YAML 文件进行配置。
- 多次试验：利用 Ray 的分布式框架，并行执行网格搜索、特定配置和超参数调优。
- 研究：预配置的多次试验。
- 硬件抽象：通过环境检测和 UV 包管理，确保在高性能计算集群（Deucalion 和 Cirrus）之间的可移植性。
- 实验跟踪：集成 Weights & Biases [26]，以实时监控每次试验的 ~200 指标。

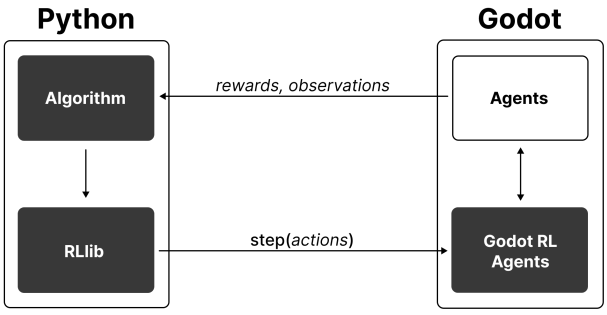


图 1：Godot-Python 强化学习流程。

这一完整的架构栈展示了本地和高性能计算执行路径，如图 2 所示。

环境仿真 基于 Godot 游戏引擎 [22], 构建，可提供具有逼真约束的基于物理的仿真。在不失一般性的前提下，由前文所述 Python 框架控制的仿真无人机以 DJI Tello 为原型。现实世界无人机及其仿真对应体的对比如图 3a 和图 3b 所示。

仿真比例为 $^{16}\text{cm}_{\,80\text{px}} = 0.2$ cm/px，无人机的运动学模型也参照 DJI Tello 的规格，并对速度进行了调整（线速度为 0.8 m/s，角速度为 3 rad/s）。

[2]可能的通信配置会随智能体数量呈指数增长 $|\mathbb{G}| = 2^{\binom{n}{2}} = 2^{n(n-1)/2}$ 对于$n = |\mathcal{N}|$ 个智能体。

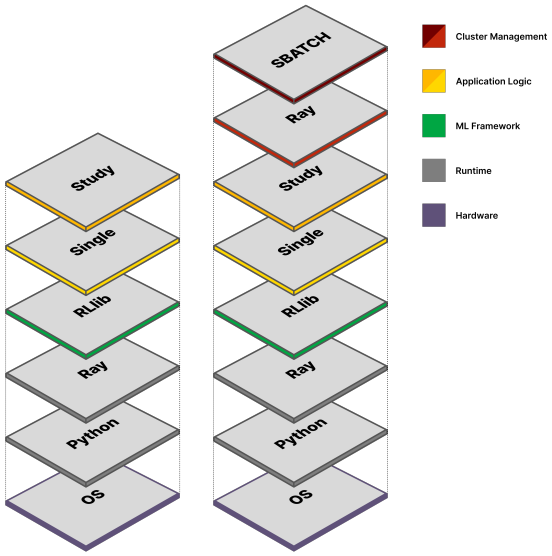
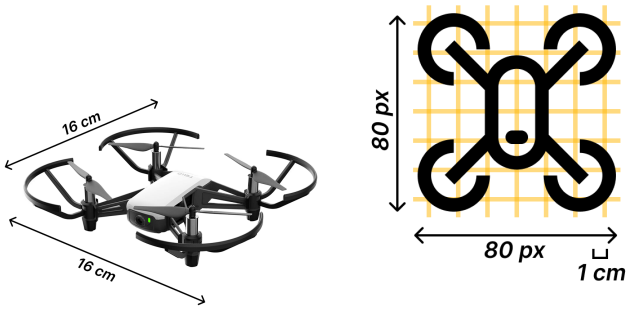


图 2：架构栈。



(a) 真实世界的 DGI Tello 无人机。

(b) 模拟无人机。

图 3：真实世界无人机及其仿真对应体。

假设无人机模型配备了单一传感器：激光雷达。激光雷达通过 RayCast2D 节点进行仿真，其量程和视场均有限。利用这些信息更新其局部地图，即采用贝叶斯对数几率更新的占据栅格地图。策略网络接收固定尺寸的自我中心地图作为输入，该地图是以智能体为中心截取的局部地图区域。图 4 展示了集成的感知与建图流程，包括激光雷达射线、颜色调制的局部地图，以及提供给策略网络的自我中心地图。

无人机能够使用有限范围网络进行通信（如图 5 所示），并根据消息大小和带宽模拟传输延迟。它们采用简单的单跳协议，持续与范围内的智能体共享激光雷达数据，并在（重新）建立连接时共享整个局部地图 。

图 6 展示了模块化的 Godot 架构，它通过专门的管理器分离各项职责，分别负责集群控制、通信、奖励计算和日志记录。

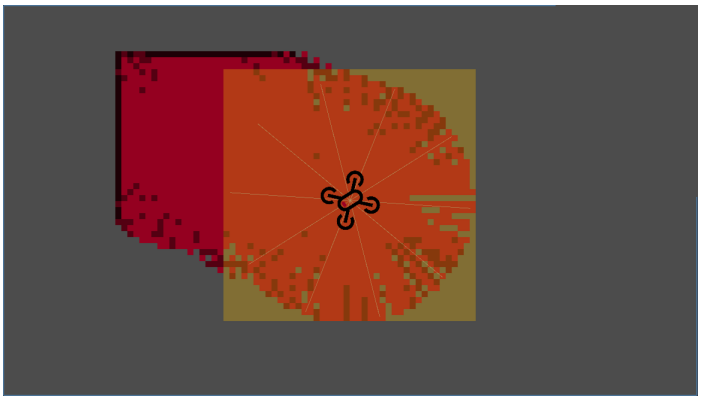


图 4：无人机感知与建图：(1) 模拟激光雷达射线（灰线），(2) 智能体的局部地图（红色区域），以及 (3) 提供给策略网络的自我中心地图（黄色区域）。

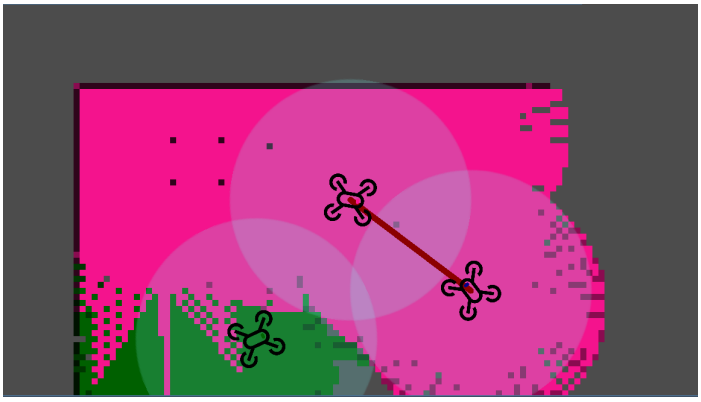


图 5：无人机网络：通信范围（蓝色圆圈）、活动链路（红色线条）以及智能体局部地图（粉色/绿色）。

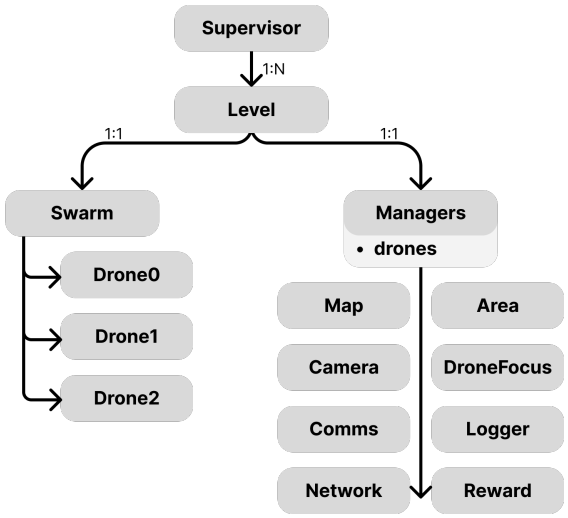


图 6：Godot 架构。

多智能体强化学习范式的实现。多智能体强化学习范式根据训练和执行两个阶段采用的方法进行区分。在训练阶段，目标是学习有效的策略，通常可以访问全局状态。在执行阶段，目标是在没有全局状态信息的情况下，将训练好的策略部署到去中心化的现实环境中。三种多智能体强化学习范式 [27]均在 Godot 端实现，以避免修改 RLlib：

- 集中式训练、集中式执行（CTCE）是一种用于多智能体问题的单智能体解决方案。所有智能体的观测和动作都会被拼接，因此这种解决方案无法扩展，且执行依赖于集中式系统。
- 分散式训练、分散式执行（DTDE）：一种完全分散式的方法，每个智能体都基于自身的局部观测进行训练和执行。这种范式具有很强的可扩展性和鲁棒性 [28]，但智能体之间往往缺乏协调行为。
- 集中式训练、分散式执行（CTDE）：一种混合范式，在训练期间利用集中式信息来学习更加协调的策略，并以分散式方式执行。这种方法旨在兼顾集中式方法的训练稳定性与分散式执行的可扩展性。

3.2. 多智能体强化学习形式化

多智能体探索问题被形式化为非分散式部分可观测马尔可夫决策过程（ND-POMDP），以考虑部分可观测性和智能体间通信。该框架由元组 $(\mathcal{N}, \mathbb{S}, \mathbb{A}, T, R, \mathbb{O}, O, \gamma, \mathbb{G})$ 定义，其中同质智能体共享相同的动作空间和观测空间（$\mathbb{A}_i = \mathbb{A}_j$, $\mathbb{O}_i = \mathbb{O}_j$，对所有$i$、$j \in \mathcal{N}$）。

奖励函数。主要目标是在避免重复探索的同时最大化区域覆盖率。奖励函数旨在激励持续探索：

$$R(s, a, s') = W_{\text{area}} \cdot \Delta Area(s, s'), \tag{1}$$

其中，系数 W_{area} 用于缩放区域奖励，而$\Delta Area(s,s')$表示整个团队在状态转移过程中探索到的新区域，并按每个时间步 Δt的理论最大可探索区域 $A_{\max}$ 进行归一化。这种归一化可确保稳定学习并防止梯度爆炸。

具体而言，区域(s) $= A_{\text{old}}$ 和 区域(s') $= A_{\text{new}}$ 分别表示状态转移前后团队覆盖的已绘制区域总面积。逐步奖励 $R^{\text{step}} = R(s,a,s')$ 随后根据新探索区域的归一化面积计算，定义如下：

$$\Delta Area(s, s') = \frac{Area(s') - Area(s)}{A_{\max}} = \frac{A_{\text{new}} - A_{\text{old}}}{A_{\max}}, \tag{2}$$

奖励计算如图7所示，将传感器覆盖范围建模为半径为r的圆形视场，相当于具有无限条射线的理想激光雷达。根据此覆盖模型和最大行进距离，计算每个智能体每个时间步可探索的理论最大区域$A_{\max}$，其计算依据为$A_{\max} = 2r \cdot v_{\max}\Delta t$。加权系数$W_{\text{area}}$ 可用于明确校准其与其他潜在目标之间的权衡（若使用）。例如，若引入权重为$W_{\text{collision}}$的碰撞惩罚$R_{\text{collision}} = -1$，则这种定量权衡由比值$W_{\text{area}}/W_{\text{collision}}$定义。若配置$W_{\text{area}} = 1$ 和$W_{\text{collision}} = 3$，则意味着智能体必须探索3单位的新区域，才能抵消一次碰撞。

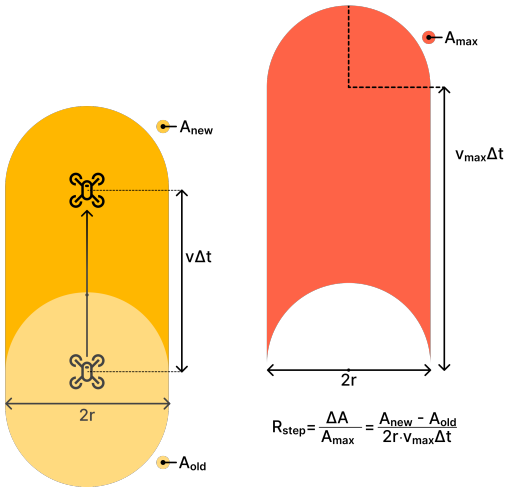


图7：每步奖励。

奖励设计对于避免不良涌现行为至关重要。早期采用碰撞惩罚的实现导致智能体为了避免碰撞而停留不动。仅对新探索到的区域进行奖励，而非奖励累计总区域，有效地克服了智能体被动性并促进了持续探索。较差的探索策略会通过机会成本（探索区域减少）受到隐式惩罚。

状态。系统采用同质智能体和连续动作空间，以实现逼真的无人机运动学。状态$s \in \mathbb{S}$ 包含完整的二维地图环境，以及由以下特征描述的各智能体内部状态：(i) 位置 $p = (p_x, p_y)$，使用地图 m的 x 和 y 坐标表示；(ii) 旋转 θ；(iii) 线速度 $v = (v_x, v_y)$；(iv) 角速度 ω。如图8所示。

动作。公式（3）中定义的动作空间输出归一化速度指令：

$$a_i = (v_x, v_y, \omega) \in [-1, 1]^3 \tag{3}$$

这些数值经过缩放，以匹配现实中的无人机速度（$v_{\max} = 400\text{px/s}, \omega_{\max} = 3\text{rad/s}$）。选择连续动作空间而非

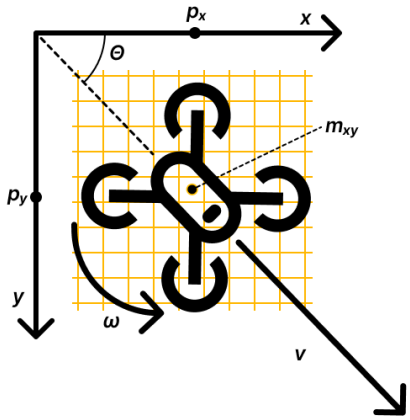


图8：无人机状态。

离散替代方案，是为了实现平滑轨迹并避免维度呈指数级增长，但这也将算法选择限制在策略梯度方法，例如近端策略优化（PPO）。

观测。智能体通过激光雷达传感器感知其环境。智能体的位置和速度也被视为观测值。在真实世界场景中，这些信息可能来自全球导航卫星系统（GNSS）（室外场景）和/或组合导航系统（INS） [29]。本研究使用来自仿真的完美、无噪声观测。这一选择是有意为之的，因为研究重点在于多智能体强化学习面临的挑战，而非感知噪声鲁棒性——在现实世界的部署中，这将由现成的同步定位与建图（SLAM）方法处理。

此外，通信协议可以为智能体提供访问其他智能体的激光雷达数据和局部地图的权限。这些信息用于更新智能体自身的局部地图。

信念。更新后的局部地图构成了智能体对环境的信念。智能体的信念以占据栅格地图（OGM）的形式实现，并通过以下两种方式之一进行更新。首先，使用布雷森汉姆直线算法融合自身和共享的激光雷达扫描，以确定被占据的单元格和空闲单元格。其次，通过对其他智能体共享的局部地图的对数优势概率求和来进行融合[3]。由于生成的局部地图在理论上是无界的，策略网络接收以智能体为中心的固定大小自我中心地图作为输入。

神经网络架构。由于 PPO 是一种演员-评论家算法，每个智能体都会使用两个神经网络：

- 动作策略（π_a）：将历史映射为连续动作。
- 价值函数（π_v）：估计当前状态下的预期累积奖励。

动作策略[4] 网络架构如图9所示，通过卷积神经网络（CNN）处理具有空间结构的自我中心地图，同时通过全连接层处理激光雷达观测。这种混合架构有效地将地图中的空间推理与激光雷达的即时感官数据相结合。该策略的输入和输出配置可以修改。例如，观测空间可以扩展以包含智能体间距离，动作空间则可以通过排除旋转而简化为纯二维平移。

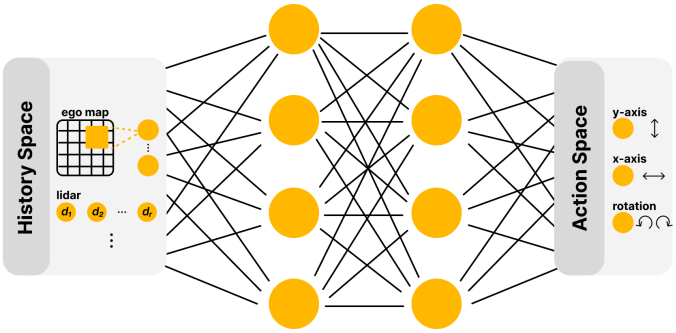


图9：动作策略神经网络。

4. 结果

考虑到多智能体强化学习固有的随机性，本文将所有环境级指标报告为多个回合的平均值。这种做法可以提供更稳定、可靠的性能估计，其中较高的数值表示该策略性能更优且更稳定。

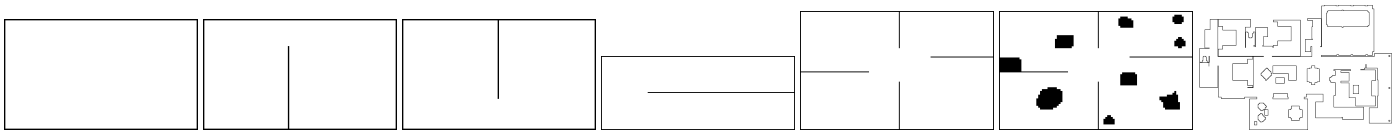
4.1. 关卡

智能体在七个复杂程度逐步增加的环境中进行了测试（面积及障碍物/房间数量逐渐增加，如图10所示），测试涵盖单智能体/多智能体（1、2、3个智能体）、范式（CTCE、CTDE、DTDE）和特定关卡（0–6）的所有组合。每次试验包含1000个回合，每个回合包含1000步。

所有试验均显示奖励持续提升，最终发现的区域汇总于图11a。性能从 L0 提升至 L4，但在 L5 急剧下降至 8.4%，这是障碍物带来的导航挑战所致；随后在 L6 再度上升。图11b 进一步展示了这一难度变化过程。

[3]这等价于在条件独立假设下的贝叶斯更新。给定来自两幅地图的单元格概率P_1和P_2，融合概率为$P_c = P_1P_2P_1P_2+(1-P_1)(1-P_2)$，这对应于对数几率空间中的$L_1 + L_2$。由于每架无人机都使用自己的传感器，该假设是成立的。

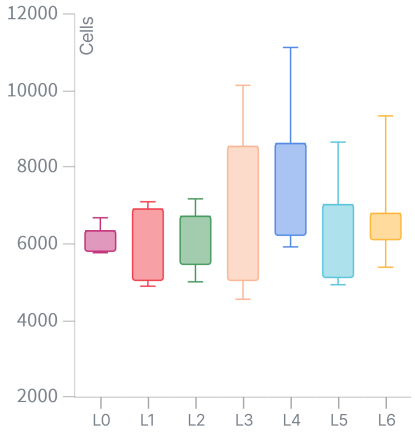
[4]价值策略网络 π_v 未在图中展示，因为其结构完全相同，只有输出层不同；输出层由一个神经元构成，用于表示当前状态的价值估计。



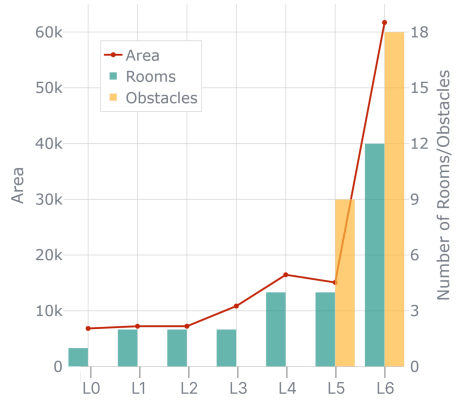
(a) 关卡0 (b) 关卡1 (c) 关卡2 (d) 关卡3 (e) 关卡4 (f) 关卡5 (g) 关卡6

图10：所有训练关卡。

如图12a 所示，策略损失持续下降，表明训练过程中不断取得进步。此外，价值函数损失呈现出预期的模式[5]，图12b 展示了关卡5的情况，这也代表了所有关卡。为使图表更加清晰，图12a 和图12b 使用了指数加权移动平均（EWMA） [30]。



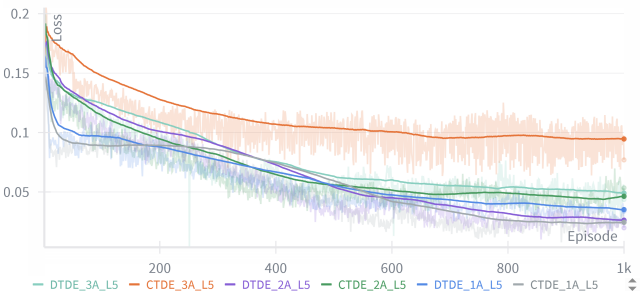
(a) 已探索关卡单元格的进展。



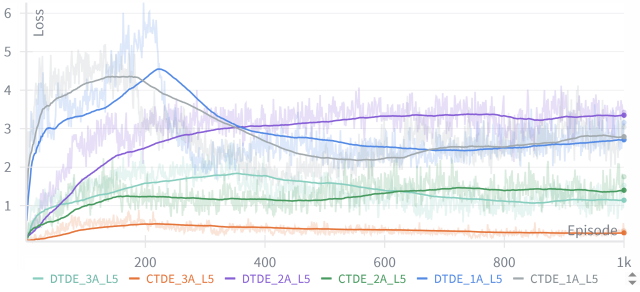
(b) 通过增加区域和障碍物/房间来逐步提高难度。格。

图11：环境探索与难度缩放的比较。

[5]先上升后下降。这是因为随着智能体开始学习并获得更高的奖励，价值函数必须适应这些新的、更高的价值估计，因此损失最初会增加。一旦学习趋于稳定，奖励变得更可预测，价值函数便会收敛到准确的状态价值估计，损失随之下降。



(a) 策略损失



(b) 价值函数损失

图 12：策略损失和价值函数损失。

主要发现如下：

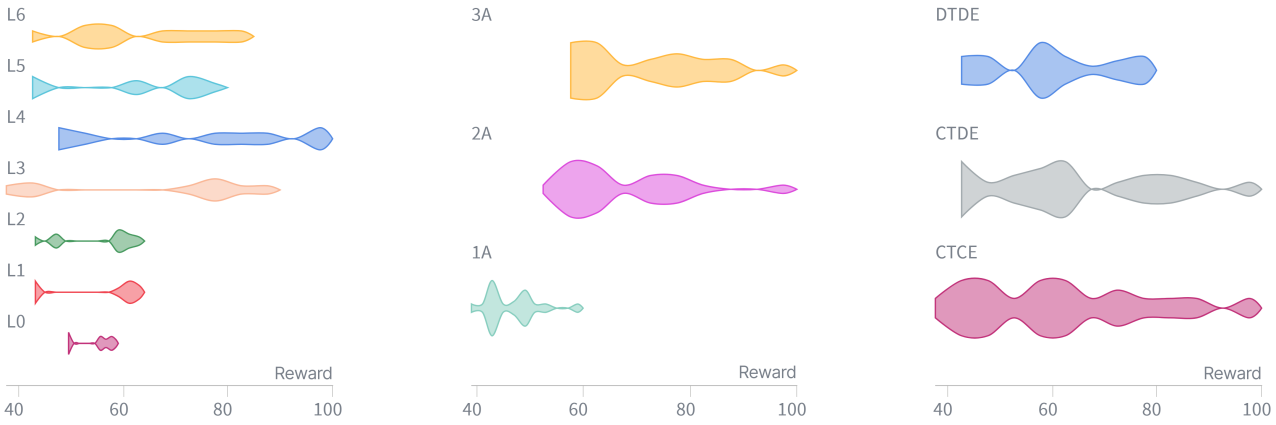
- 可扩展性：多智能体配置的表现始终优于单智能体设置，在较简单的关卡中覆盖率最高可达 98%（图 10a 至 10c）。

- 多智能体复杂性：双智能体配置偶尔优于三智能体系统，这表明协调复杂性会随着团队规模的扩大而非线性增长。

- 范式：作为单智能体范式，CTCE 大幅简化了问题，这解释了其出色的表现。不过，CTDE 成为最有效的训练范式，尤其是在复杂环境中，集中式价值估计相较于完全分散式方法 DTDE 更具优势。

- 环境复杂性：随着障碍物密度和探索区域的增加，性能自然会下降；不过，随着智能体在更大的环境中发现更多单元格，绝对探索能力有所提升。

跨关卡分析（图13a至13c）证实，较高的



(a) 关卡奖励。 (b) 范式奖励。 (c) 智能体数量奖励。

图13：每个关卡、范式及智能体数量的奖励。

智能体数量能带来更好的结果，而且较高关卡更具挑战性，因为奖励范围逐渐扩大。请注意，较高关卡的规模也更大，这意味着善于导航的智能体在更难的关卡中能获得更多奖励。关卡6是最大且要求最高的关卡，这一点尤为明显：其奖励范围较大，且分布均匀。

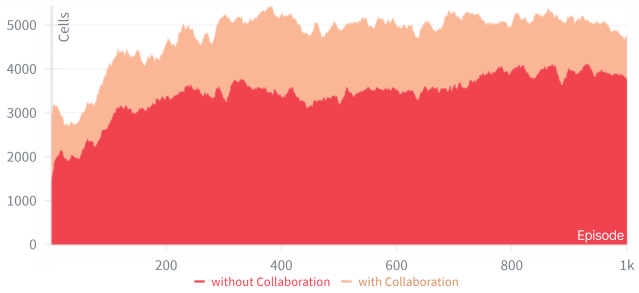
4.2. 通信与协调

对智能体间通信的分析揭示了显著的协作优势：

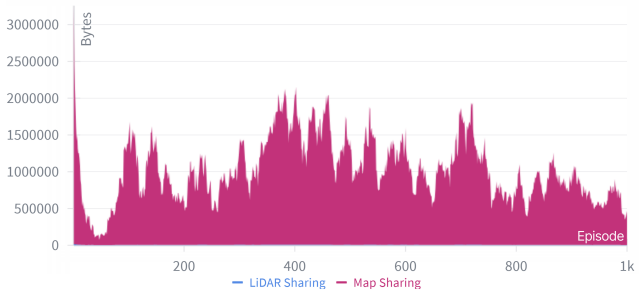
- 智能体有效地共享了环境信息，通信贡献了多达 40% 的已发现单元格（如图 14a 所示）。
- 激光雷达共享被证明比完整地图共享更有效（其贡献的共享单元格比例从 85% 逐渐降至 60%），也更高效[6]（占带宽的比例为 1% 对 99%，如图14b所示），不过两种策略都促进了探索。
- 假设的涌现行为包括连续运动协调（通过保持彼此接近来实现激光雷达共享）或周期性重新集结以同步地图。
- 跨关卡分析表明，探索和共享动态受到限制：智能体之间的比率始终保持在[0.5, 2]以内。这意味着在任意时刻，一个智能体发现的单元格数量最多是另一个智能体的两倍，同样，其分享的新单元格数量也最多是另一个智能体的两倍。

[6]这种低效源于传输了整张地图，而非仅传输已探索的部分，同时也缺乏游程编码或 LZMA2 等压缩技术（例如 7zip 中使用的技术）。

为进一步量化协作的影响，我们在集中训练分散执行范式下，测试了由 2–4 个智能体组成的团队在三种通信协议中的表现：无通信（关）、受限单跳协议（单跳）以及多跳网络（多跳），在多跳网络中，所有智能体都会转发收到的消息。



（a）通过探索和协作获取的单元格。



（b）协作中的传输字节数。

图14：探索与协作中的单元格，以及传输的字节数。

图15所示的结果揭示了一种违背直觉的模式：通信增加反而导致整体性能下降。

据推测，这一悖论的产生是因为智能体通过共享地图信息，学会避开标记为已探索的区域。在关卡4中，必须穿过中央房间才能进入其他房间，而这种行为会变得

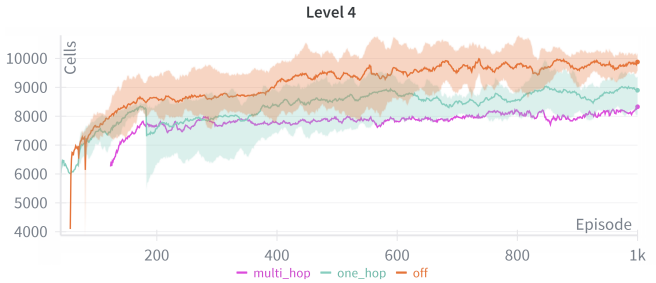


图15：不同通信协议下的性能 ols.

有害。如果多个智能体在最初探索后学会避开这一中心区域，它们就会被困住，无法抵达其外更广阔的未探索区域。

因此，减少通信会限制共享地图数据，反而帮助智能体识别未探索区域。这进一步表明，当前策略架构的复杂度不足以支持智能体穿越已知区域，抵达未探索区域。

4.3. 消融研究

碰撞即终止。本文研究了在智能体发生碰撞时提前终止的可变长度回合，这种机制称为碰撞即终止，并假设该机制能够激励智能体避免碰撞、提升导航性能。与预期相反，固定长度回合的表现比这种方法高出 5 倍。

基于单智能体表述的研究结果 [18],，本文引入了碰撞惩罚$R_{\text{collision}} = -3$。然而，这导致智能体学会停滞不动，以避免可能发生的碰撞。

这些发现表明，多智能体强化学习表述的复杂性远高于单智能体场景，因此需要仔细考量奖励结构和终止条件，以避免出现非预期行为。

并行与顺序。鉴于环境关卡的层级结构[7],，评估了两种课程学习策略：(i) 在所有关卡上同时进行并行训练；(ii) 顺序训练，智能体逐一通过各个关卡（达到 80% 区域覆盖率后进入下一关）。

结果证明，顺序推进比并行策略更有效，性能高出 250%。这一结果归因于并行设置中的灾难性干扰，

[7]各关卡的可探索区域面积和障碍物密度逐步增加。关卡1引入了基本墙体，关卡3增加了障碍物，关卡4通过增加房间数量扩大了规模，关卡5提高了障碍物密度，而关卡6则是最逼真、最复杂的场景，依据房屋蓝图创建。

在这种设置下，所有关卡中多样且相互冲突的动态会同时压垮智能体的策略，阻碍稳定学习。相比之下，顺序课程使智能体能够逐步积累知识，掌握较早关卡中的基础技能，并将其稳健地迁移到更复杂的关卡中。

课程学习。此前顺序训练的一个主要局限是通过条件过于简单。只需达到一次 80% 的区域覆盖率，这一结果可能源于策略探索，而非展示出稳定一致的能力。

通过参数化设置 pass_area（覆盖率阈值）和 pass_x_times（一致性要求），例如，达到 80% 区域覆盖率 20 次，这种课程学习方案得以提升性能并加快训练。

图16展示了通过_区域比例在60%、70%、80%和90%之间变化，以及通过_x_次数在1、10、15和20之间变化的情况。关卡3揭示了阈值难度与所需一致性之间的明显交互作用。



图16：关卡3。

图16展示了因通过条件变化而产生的不同间隔。研究发现了几个关键行为：

(1) 更高难度的进阶标准能够促进学习，80%(20x)展现出相较于80% (15x)的即时性能优势，而后者又优于60% (20x)。同样，90%(15x)>(10x)>(1x)，证实了进阶难度与学习效果之间的正相关关系。

(2) 可靠性优先于早期成功：90% (1x) 条件比80% (20x) 提前约50个回合进入关卡3，但后者能够实现更优的即时性能。单次成功条件仅呈现出渐进式改进，并且始终落后于多次成功变体（15x、10x），验证了相较于偶然运气，可靠的性能更为重要。

(3) 完美主义：过高的覆盖率阈值（90%）及其变体(15x)、(10x)和(1x) 整体表现低于80%(20x)。达到更高的

要达到90-100%的覆盖率非常困难，因为移动过程中会出现建图缺口，而且可能需要回溯已探索区域才能找到剩余的小缺口。

（4）平衡：80%（20 倍）条件的区域平均值达到 80%，表明这些超参数可能实现了最佳平衡，在避免过度完美主义的同时，确保性能一致。

4.4. 优化动作和观测

智能体的策略对动作和观测的形式化表现出显著敏感性。对这些设计可能性进行网格搜索后，呈现出清晰的性能层级。

按动作空间分组后，呈现出一致的排名：在二维平面上不进行旋转的移动优于带独立旋转的一维移动，而这两者又显著优于基于航点的方法；在该方法中，智能体指定目标方向，无人机则自动旋转并移动[8]。

按观测空间分析时，差异不太明显，但趋势线始终更倾向于将自我中心地图与激光雷达数据结合使用。明确纳入智能体间距离被证明不利于学习效率。推测激光雷达测量已经隐式编码了邻近信息，但无法区分智能体接触和障碍物接触。与此同时，仅使用激光雷达的观测被证明不足，这与预期一致，因为只有自我中心地图纳入了历史观测，为有效导航提供了必要的时间上下文。

4.5. 架构优化

多项神经网络增强显著提升了性能。

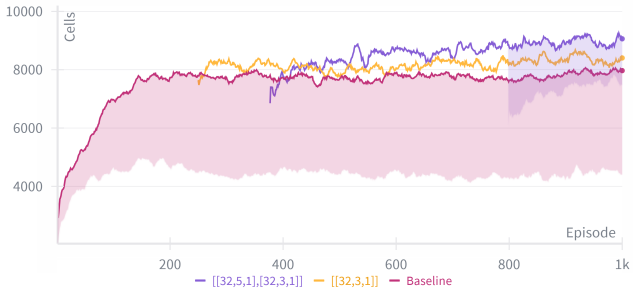
集中式评论家。在价值函数之间共享参数带来了小幅提升（区域覆盖率提高了 1%）。

激光雷达卷积。一维卷积神经网络对激光雷达数据的处理使性能提升了20%，使智能体能够克服以往的性能障碍，并推进到更复杂的关卡5（图17a）。测试了两种不同的卷积滤波器：单层和双层，其中层数更深的一种表现更好。

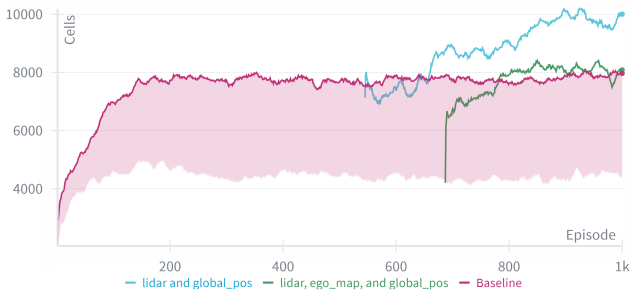
循环神经网络集成。时序处理使性能提升了13%，最优配置结合了激光雷达与全局定位数据（图17b），但代价是训练收敛更慢。基线在70个回合内实现了相当的单元格探索，而循环神经网络方法需要220个回合才能超过这一阈值。

表现最佳的配置使用了全局定位数据。这使智能体能够利用全局位置的时间序列，学习平移运动，从而系统地实现区域覆盖。此外，通过分析位置历史，智能体可以推断未探索区域的情况，这引出了全局规划与局部规划策略孰优的问题。当前的形式化通过激励即时区域发现来优化局部探索，而全局方法或许会优先探索奖励较少的区域，以尽量减少未来的回溯。全局规划方法更符合实现完整区域覆盖（90%–100%）这一目标；这是一个相关但超出本研究范围的问题。

仅使用激光雷达观测的循环神经网络，其表现不如本文提出的不使用循环神经网络的基于信念的方法（即自我中心地图）。这表明，不加区分地使用循环神经网络——正如其他研究中常见的那样，用以应对部分可观测性——可能并非最优。如第 2 节所述，循环神经网络会根据激光雷达数据隐式地构建信念状态，其功能与*ego_map*中的显式地图表示相似。然而，显式地图具有更强的可解释性，并且能够通过地图融合轻松实现智能体间通信。相比之下，循环神经网络的潜在空间缺乏固有的可解释性，而在智能体之间融合这些表示则面临巨大挑战。



(a) 激光雷达卷积。



(b) 循环神经网络。

图17：关卡5中的架构优化：(a) 对激光雷达数据进行卷积处理，或 (b) 集成循环神经网络以进行时序推理。

[8]在这种模式下，手动控制颇具挑战性且不够直观，因此有必要进一步改进。这类控制机制类似于两阶段强化学习形式化，会增加学习复杂度，与本研究采用的端到端方法相冲突。

有的可解释性，而在智能体之间融合这些表示则面临巨大挑战。

5. 结论

本文提出了一种用于未知室内环境中多智能体通信感知探索的框架。我们在受限通信网络中构建了一个去中心化且可扩展的多智能体系统，以近似真实任务条件。

本文通过在高保真游戏引擎仿真中应用多智能体强化学习，训练一组无人机协同探索未知的 GNSS拒止室内环境，展示了该框架的能力。每个无人机智能体都配备了激光雷达传感器，用于构建局部占据栅格地图。每个智能体还可在有限范围内与邻近智能体交换数据。无人机智能体在连续动作空间中接受训练，利用非分散式部分可观测马尔可夫决策过程在不确定条件下进行导航。

通过广泛的消融研究，本文系统评估了训练范式、奖励函数和网络架构等关键因素。研究还表明，课程学习能够显著加快训练并提升其稳定性。此外，所得到的策略克服了先前研究中的若干关键局限，例如依赖离散动作和集中式控制，从而为将学习得到的策略部署到实体机器人系统奠定了有依据的基础。

实验结果得出了若干重要发现：

（1）多智能体视角至关重要。例如，基于惩罚的碰撞规避导致智能体学会了不移动。（2）与标准循环神经网络相比，SLAM-RL集成在处理部分可观测性方面更为有效。（3）专用的激光雷达卷积网络带来了 20% 的性能提升，凸显了领域特定感知架构的价值。（4）利用 Godot引擎构建了稳健且易于使用的仿真框架，在高性能计算集群上通过工业级工具验证，为用户提供了一个轻量级开源平台，并具备高保真物理效果。（5）通过结构化的课程学习方法成功完成训练，使智能体能够逐步掌握在日益复杂的环境中的导航。

本研究承认其固有局限，包括多智能体强化学习（MARL）样本复杂度较高，因而需要较长的训练时间，以及环境泛化方面可能面临的挑战。此外，本研究侧重于比较多智能体学习范式，而非与经典覆盖算法（例如基于前沿的方法）进行直接基准比较。

这项研究作为未来工作的基石，勾勒出后续研究的全面路线。未来的研究方向包括：

- **改进传感器与观测**，包括支持视觉传感器和视觉同步定位与建图，在自我中心地图中实现动态旋转，以实现更平滑的无人机轨迹 [18], ，并将激光雷达视场限制在合理范围内，而非不切实际的 360°。
- **扩展到三维**设置，调整智能体运动学的形式化，考虑观测及其产生的信念所带来的不同信息增益，以及通信限制对任务性能的影响。
- **SLAM-RL**。尝试不同的选项，利用它们的共性，例如图优化同步定位与建图和强化学习侧的图神经网络。
- **增强的智能体模型和学习算法**，包括通过价值分解（例如，Q值混合（QMIX） [31]）改进信用分配；用于多层级决策的分层策略；通过专家示范进行模仿学习预训练；通过博弈论自我博弈实现竞争动态 [32]；以及采用受 AlphaZero 启发的搜索算法（如蒙特卡洛树搜索）进行集成规划。
- **时序推理**。将循环神经网络应用于自我_地图时观察到的性能提升，可能源于对智能体运动模式的时序推理，这类似于强化学习中的帧堆叠。可以通过应用三维卷积神经网络来实现。鉴于这些因素，采用激光雷达卷积的形式化可能更为可取，因为它适用于拒止全球导航卫星系统环境、架构简单且训练速度更快，尽管与最佳循环神经网络配置相比，其最终性能降低了12.5%。未来对改进观测表示的研究可以进一步缩小这一性能差距。

致谢

作者谨此感谢 HPCvLAB，其计算资源使本研究得以形成目前的成果。感谢其提供对 Deucalion [33] 和 Cirrus [34]高性能计算集群的访问权限。

本研究得到葡萄牙科学技术基金会（FCT）2023.04842.BD 项目的支持。

本工作亦由FCT——科学与技术基金会（Fundação para a Ciência e a Tecnologia, I.P.）通过国家资金资助，资助项目为

项目/支持 UID/6486/2025（https://doi.org/10.54499/UID/06486/2025）和UID/PRR/6486/2025（https://doi.org/10.54499/UID/PRR/06486/2025）。

参考文献

[1] H. A. Foudeh、P. C.-K. Luk、J. F. Whidborne，基于学习控制的先进无人驾驶飞行器（uav）用于架空电力线监测的方法：综合综述，《IEEE Access》9（2021）130410–130433。doi:10.1109/ACCESS.2021.3110159。[2] N. Boonyathanmig、S. Gongmanee、P. Kayunyeam、P. Wutticho、S. Prongnuch，微型无人机的设计与实现：用于室内监控，载于：2021年第9届国际电气工程大会（iEECON），2021，9 440350 305 308 10 1109 51072 2021，第 – 页。doi: 。/iEECON 。 。 。[3] I. Mademlis、V. Mygdalis、N. Nikolaidis、M. Montagnuolo、F. Negro、A. Messina、I. Pitas，用于覆盖户外活动的高级多无人机电影摄影工具，《IEEE广播学报》65（3）（2019）627–635。doi:10.1109/TBC.2019.2892585。[4] D. Rakesh、N. Akshay Kumar、M. Sivaguru、K. V. R. Keerthivaasan、B. Rohini Janaki、R. Raffik，无人机在创新农业及未来应用中的作用：综述，载于：2021年国际电气、电子、通信、计算与自动化进展会议（ICAECA），2021，第1–6页。doi:10.1109/ICAECA52838。2021.9675612。[5] J. Dong、K. Ota、M. Dong，基于无人机的灾后搜救行动实时幸存者探测系统，《IEEE微型化航空航天系统学报》2（4）（2021）209–219。doi：10.1109/JMASS.2021.3083659。[6] 2014年新兴及跨领域航空问题；无人驾驶航空器系统（UAS）的日益广泛应用——icao.int，https://www.icao.int/annual-report-2021/Pages/emerging-and-cross-cutting-aviation-issues-increased-use-of-unmanned-aircraft-syste ms-uas.aspx，[访问日期：12-01-2025]。[7] B. Yamauchi，基于前沿的方法实现自主探索，载于：1997年IEEE机器人与自动化计算智能国际研讨会论文集（CIRA'97）。"面向机器人与自动化的新计算原理"，1997，第146–151页。doi:10.1109/CIRA.1997.613851。[8] M. T. Ohradzansky、E. R. Rush、D. G. Riley、A. B. Mills、S. Ahmad、S. McGuire、H. Biggie、K. Harlow、M. J. Miles、E. W. Frew、C. Heckman、J. S. Humbert，多智能体自主性：地下探索的进展与挑战（2021）。arXiv:2110.04390。URL https://arxiv.org/abs/2110 04390。X. Chen、A. N. Iyer、Z. [9] Wang、A. H. Qureshi，基于访问频率地图的高效Q学习：用于未知环境中的多智能体探索（2023）。arXiv:2307.16318。URL https://arxiv.org/abs/2307 16318。M. Fernandez-Cortizas、[10] H. Bavle、D. Perez-Saura、J. L. Sanchez-Lopez、P. Campoy、H. Voos，Multi S-Graphs：一种高效的分布式语义关系协同SLAM，《IEEE机器人与自动化快报》9（6）（2024）6004–6011。doi:10.1109/lra.2024.3399997。URL http://dx.doi.org/10 1109 2024 3399997。/LRA。。H. Choset、[11] P. Pignon，覆盖路径规划：往复式细胞分解法，载于：A. Zelinsky（编），《野外与服务机器人学》，Springer London，伦敦，1998，第203–209页。[12] J. Wang、R. Wang，基于沃罗诺伊图和注意力机制的多无人机区域覆盖航迹规划，《应用科学》14（17）（2024）。doi:10.3390/app14177844。URL https://www.mdpi.com/2076 3417 14 17 7844 - / / / E. U. Acar、[13] H. Choset、A. A. Rizzi、P. N. Atkar、D. Hull，用于覆盖任务的Morse分解，《国际机器人研究杂志》21（2002）331 – 344。URL https://api.semanticscholar.org/CorpusID:12656982 [14] S. Karaman、E. Frazzoli，最优运动规划的基于采样的算法（2011）。arXiv:1105.1186。URL https://arxiv.org/abs/1105 1186。Z. Shen、J. P. Wilson、S. [15] Gupta，C*：一种利用快速覆盖图实现未知环境覆盖路径规划的算法（2025）。arXiv:2505.13782。URL https://arxiv.org/abs/2505 13782。X. Chen、T. M. Tucker、T. R. [16] Kurfess、R. Vuduc，Adaptive deep path：各种配置下已知环境的高效覆盖，载于：2019年IEEE/RSJ智能机器人与系统国际会议（IROS），IEEE Press，2019，第3549–3556页。doi:10.1109/IROS 40897.2019.8967793。URL https://doi.org/10 1109 40897 2019 8967793。/IROS。。F. [17] Chen、S. Bai、T. Shan、B. Englot，通过深度强化学习实现移动机器人的自学习探索与地图构建，AIAA Scitech 2019论坛（2019）。URL https://api.semanticscholar.org/CorpusID:68149666 [18] A. Jonnarth、J. Zhao、M. Felsberg，在未知环境中通过深度强化学习学习覆盖路径（2024 1 6978 2306）。arXiv：。。URL https://arxiv.org/abs/2306 16978。C. Guestrin、D. Koller、R. E. [19] Parr，基于因子化马尔可夫决策过程的多智能体规划，载于：神经信息处理系统，2001。URL https://api.semanticscholar.org/CorpusID:6487585 [20] Z. Liu、J. Zhang、E. Shi、Z. Liu、D. Niyato、B. Ai、Xuemin，Shen，图神经网络与多智能体强化学习相结合：基础、应用与未来方向（2024）。arXiv:2404.04898。URL https://arxiv.org/abs/2404 04898。C. Sun、[21] M. Shen、J. P. How，面向机器人系统扩展多智能体强化学习：学习自适应稀疏通信图（2020）。arXiv:2003.01040。URL https://arxiv.org/abs/2003 01040。G. Engine，Godot引擎——免费[22] 开源的二维和三维游戏引擎——godotengine.org，https://godotengine.org/，[访问日期：02-12-2024]。[23] F. Oliehoek、C. Amato，《去中心化部分可观测马尔可夫决策过程简介》，Springer智能系统简报，Springer International Publishing，2016。URL https://books.google.pt/books?id=FZRPDAAAQBAJ [24] A. Goeckner、Y. Sui、N. Martinet、X. Li、Q. Zhu，基于图神经网络的多智能体强化学习：面向多机器人系统的韧性分布式协同（2024）。arXiv:2403.13093。URL https://arxiv.org/abs/2403 13093。T. Leite，[25] GitHub - TiagoMLeite/IMAGINE——github.com，https://github.com/TiagoMLeite/IMAGINE（2025）。[26] wandb.ai，https://wandb.ai/site/，[访问日期：16-10-2025]。[27] Z. Ning、L. Xie，多智能体强化学习及其应用综述，《自动化与智能学报》3（2）（2024）73–91。doi:https://doi.org/10.1016/j.jai.2024.02.003。URL https://www.sciencedirect.com/science/article/pii/ S2949855424000042[28] K. Zhang、Z. Yang、H. Liu、T. Zhang、T. Ba.sar，网络化智能体的完全去中心化多智能体强化学习（2018）。arXiv: 1802.08757。URL https://arxiv.org/abs/1802.08757[29] 无人机、无人驾驶飞行器、无人地面车辆、无人水面艇与自主水下航行器的自主导航与定位——unmannedsystemstechnology.com，https://www.unmannedsystemstechnology.com/expo/autonomous-navigation/，[访问日期：11-01-2025]。[30] 指数加权移动平均（EWMA）——corporatefinance- institute.com，https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/exponentially - weighted-moving-average-ewma/，[访问日期：16-10-2025]。[31] T. Rashid、M. Samvelyan、C. S. de Witt、G. Farquhar、J. Foerster、S. Whiteson，QMIX：用于深度多智能体强化学习的单调价值函数分解（2018）。arXiv:1803.11485。URL https://arxiv.org/abs/1803 11485。R. [32] Zhang、Z. Xu、C. Ma、C. Yu、W.-W. Tu、S. Huang、D. Ye、W. Ding、Y. Yang、Y. Wang，强化学习中自对弈方法综述（2024）。arXiv:2408.01072。URL https://arxiv.org/abs/2408 01072。MACC，[33] Deucalion——Deucalion用户指南——docs.macc.fccn.pt，https://docs.macc.fccn.pt/deucalion/，[访问日期：04-10-2025]。[34] INCD-Lisbon（cirrus.a.... | INCD用户文档——wiki.incd.pt，https://wiki.incd.pt/books/compute-node-specs-and-infor mation，[访问日期：04-10-2025]。