中国科学:信息科学 2026 年 第 56 卷 第 8 期: 2081–2099
SCIENTIA SINICA Informationis
论文

《中国科学》杂志社
SCIENCE CHINA PRESS
CrossMark click for updates

地下通信小尺度衰落中基于时空图强化学习的多机器人轨迹规划方法

陈岩[1], 张婷婷[1*], 董仁智[2], 陈星宇[1], 杨昊[3], 宋爱国[4]

1. 陆军工程大学指挥控制工程学院, 南京 210007
2. 电科云 (北京) 科技有限公司, 南京 210007
3. 南京信息工程大学软件学院, 南京 210044
4. 东南大学仪器科学与工程学院, 南京 210017
* 通信作者. E-mail: 101101964@seu.edu.cn

收稿日期: 2025–08–17; 修回日期: 2025–11–08; 接受日期: 2025–12–07; 网络出版日期: 2026–05–12

摘要 在地下勘探、矿井测量等探测场景中, 分布式多机器人协作面临无线信号衰减和多径干扰等通信挑战, 导致感知能力受限、信息共享不足, 进而影响轨迹规划效率. 本文提出了一种基于时空图强化学习的多机器人轨迹规划方法, 首先, 采用小尺度衰落信道模型计算通信拓扑的边权重, 构建基于时空图卷积神经网络 (spatio-temporal graph convolutional network, STGCN) 的轨迹规划模型, 利用图卷积的多跳消息传递, 实现动态通信链路下的信息传递, 并引入门控循环单元对观测信息进行记忆重构, 增强机器人在信号衰落环境中的信息感知能力. 接着, 设计了基于时空图神经网络的多智能体双延迟深度确定性策略梯度算法 (spatio-temporal graph-based multi-agent twin delayed deep deterministic policy gradient, STG-MATD3), 采用 Actor- 双 Critic 架构进行分布式策略训练与优化, 实现多机器人轨迹规划. 结果表明, STG-MATD3 算法的任务完成率达 98.8%, 平均轨迹距离较基线算法缩短了 10.78%~18.59%, 显著优于 MADDPG, MASAC 等强化学习算法. 最后, 通过消融实验和泛化性测试验证了 STGCN 模型在多机器人轨迹规划中的有效性和适应性, 增强了多机器人地下作业的应用潜力.

关键词 多机器人, 小尺度衰落, 时空图卷积网络, 图强化学习, 轨迹规划

1 引言

在地下隧道、管廊等环境中, 传统人工巡检面临空间狭窄、危险作业 (如瓦斯泄漏、塌方) 等严峻挑战. 无人系统凭借灵活性、安全性及自主性优势, 通过多机器人协同作业可以有效降低人员伤亡风险, 适合执行繁琐、高危任务 (如排爆、侦察等) [1]. 然而, 在管道、矿井等地下工事中, 空间结构封闭, 几何形态复杂, 存在大量非视距遮挡与不规则反射界面, 导致无线信号传输损耗大、长波导航信号难以穿透、多径效应干扰严重等. 因此, 多机器人系统在执行协同探测、联合搜索等任务时, 往往面临感知信息碎片化与状态共享不充分等问题, 制约了系统整体协同效率与决策可靠性.

在此环境下, 多机器人协同面临的核心挑战在于通信不可靠性对协同决策过程的系统性影响. 机器人个体获得的邻居状态信息呈现出不完整、异步的特征, 导致系统整体难以形成一致、稳定的协同态势认知. 另一方

引用格式: 陈岩, 张婷婷, 董仁智, 等. 地下通信小尺度衰落中基于时空图强化学习的多机器人轨迹规划方法. 中国科学: 信息科学, 2026, 56: 2081–2099, doi: 10.1360/SSI-2025-0193
Chen Y, Zhang T T, Dong R Z, et al. A spatio-temporal graph reinforcement learning-based multi-robot trajectory planning method for small-scale fading in underground communications. Sci Sin Inform, 2026, 56: 2081–2099, doi: 10.1360/SSI-2025-0193

© 2026《中国科学》杂志社

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

面, 机器人间的信息交互关系不断重构, 使依赖全局信息假设的集中式规划方法在实际应用中难以部署, 个体既需要在局部感知与有限通信的基础上独立决策, 又需要通过不稳定的信息交互实现整体协同.

针对上述问题, 本文提出了一种基于时空图强化学习的分布式机器人轨迹规划方法, 通过时空图卷积神经网络 (spatio-temporal graph convolutional network, STGCN) 在有限、不稳定的信息交互条件下实现多机器人状态信息的有效聚合与协同决策, 并利用多智能体强化学习算法对模型进行训练, 学习生成有效的运动轨迹, 同时在算法训练过程中引入图注意力机制, 引导多机器人系统在通信受限环境下优先聚合与路径冲突和协同决策高度相关的信息, 从而有效减少信息冗余, 提升轨迹规划效率. 本文的主要贡献如下.

(1) 针对地下环境中多径传播导致接收信号幅度呈随机分布的特点, 构建了融合路径损耗与 Rician/Rayleigh 小尺度衰落的通信信道模型, 将信噪比 (signal-to-noise ratio, SNR) 映射为通信拓扑边权, 实现对多机器人通信拓扑随环境变化的动态表征.

(2) 构建了基于时空图卷积网络的多机器人轨迹规划策略模型, 针对无线通信链路呈现出强随机性、时变性与频繁间歇性中断的信道特征, 利用图卷积网络 (graph convolutional network, GCN) 有效聚合机器人状态信息. 同时, 引入门控循环单元 (gated recurrent unit, GRU) 构建时序记忆单元, 提升系统对历史关键信息的利用能力, 降低个体对实时通信的依赖.

(3) 设计了一种基于时空图神经网络的多智能体双延迟深度确定性策略梯度算法 (spatio-temporal graph-based multi-agent twin delayed deep deterministic policy gradient, STG-MATD3), 将时空图结构与 Actor- 双 Critic 学习架构相结合, 实现策略的联合训练与优化, 并引入图注意力机制, 增强 Critic 网络对智能体间关联性的评估能力, 实现稳定、高效的多机器人轨迹规划.

2　相关工作

2.1　地下通信模型

随着人工智能应用和 5G 通信技术的蓬勃发展, 网络自组织和自优化能力得到了显著提升 [2]. 然而, 在地下空间中, 探测机器人端到端通信受多径效应的影响, 难以保持信息交互. 钱七虎 [3] 指出地下工程中岩溶隧道突水突泥和岩爆等地质灾害现象日益频发, 突水冲出的巨石和围岩体爆裂物等介质构成了多重信号衰弱源, 为了量化分析地下无线自组网的信道容量、信噪比及路径损耗等关键参数, 现有研究通常基于多径衰落特性进行信道建模.

Ghaddar 等 [4] 考虑了地下隧道的几何非均匀性和表面粗糙不平特性, 构建了多径衰落呈对数正态分布的确定性多输入多输出 (MIMO) 信道模型. Becerra 等 [5] 针对地下矿山路段中光源、反射、光散射以及灰尘引起的衰减, 提出一个波长依赖的 UM-VLC 信道模型, 计算了接收功率、信噪比 (SNR)、信号干扰比 (SIR) 和误码率 (BER) 等参数. Javaid 等 [6] 针对地下煤矿环境的可见光通信系统, 采用 Mie 散射理论分析了煤尘颗粒对信号传输的影响, 通过 Mie 散射模型量化了煤尘导致的路径损耗和多径衰减, 为小尺度衰落的建模提供了依据.

2.2　多机器人轨迹规划

传统的多机器人轨迹规划方法, 例如人工势场法 [7]、动态窗口法 [8] 和模型预测控制 [9] 等, 虽然在路径生成和避障方面展现出一定优势, 但在动态环境中仍面临诸多挑战, 存在路径优化、冲突解决及环境适应性之间的权衡难题. 毛剑琳等 [10] 提出一种基于讨价还价博弈机制的改进层级协作 A* (B-IHCA*) 算法, 引入博弈论和长路径熔断机制到多智能体路径规划框架中, 通过动态协商解决机器人在密集场景中的优先级冲突问题. 王乐乐等 [11] 提出了一种改进的快速扩展随机树 (RRT) 算法, 将编队几何约束与 RRT 路径扩展的随机采样结合, 实现了多机器人编队路径规划中的队形保持与协同避障. 于连波等 [12] 针对智能体和动态障碍物冲突问题, 改进了基于冲突搜索 (CBS) 的算法, 利用障碍速度法提升了在线冲突消解性能, 降低了多智能体全局路径规划的计算量. 传统经典规划算法如 A*, Dijkstra 等在静态或相对简单的动态环境中表现良好, 但在高度动态、复杂、多智能体

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

交互密集的地下环境中仍存在一定的局限性, 强化学习算法 [13, 14] 能够持续优化长期策略, 动态适应复杂动态障碍和环境变化, 具备更强的鲁棒性和泛化能力.

近年来, 图神经网络被广泛应用于多机器人路径规划, 强化学习将感知、决策与动作执行集成到模型中, 实现从原始输入 (如地图或传感器数据) 到路径决策的端到端优化. 牛津大学的李庆标团队 [15] 首次应用图神经网络 (GNN) 到多机器人分布式路径规划中, 在有限观测环境下, 利用 GNN 实现了机器人的多跳通信, 使机器人自主学习信息传递, 替代了人工设计通信规则. 该团队后来又引入类 Transformer 的 Key-Query 注意力机制到图注意力网络 (MAGAT), 使机器人能够优先处理关键邻居的信息, 减少冗余通信, 提升了决策质量 [16]. 杜威等 [17] 利用异构图卷积网络实现了异构信息的融合和传递, 使智能体可以从不同类型信息中学到最有价值和有表达性的信息. Li 等 [18] 首次融合深度强化学习 (DRL) 和图神经网络, 利用图卷积网络聚合来自邻居的状态信息, 使 DRL 可以利用 GNN 提供的信息更全面地了解动态环境, 优化多机器人系统的路径策略. Feng 等 [19] 进一步融合了图注意力网络和多智能体强化学习, 使用图注意力机制改善多智能体强化学习中的评论网络, 提高了 Q 值估计的准确性和多机器人规划策略学习效率. 肖剑等 [20] 针对通信受限环境下的多智能体集群协同控制问题, 在多智能体强化学习的策略网络中引入了距离图注意机制, 降低了通信质量较差的远程邻居对智能体行为决策的影响.

图神经网络与强化学习相结合的技术在多机器人路径规划中已展现初步成效, 但现有研究未考虑带宽受限、信道衰落等复杂场景, 导致模型在实际应用场景下的适应性不足. 此外, 在动态环境中, 环境的时序变化如障碍物移动、机器人位置更新等, 需要模型具备更强的时序建模能力, 以更好地理解环境变化. 为了充分考虑环境的影响, 对动态环境的时序特征进行捕捉, 本文在传统图神经网络中引入了门控循环单元的时序模型, 同时构建小尺度衰落信道模型量化通信链路变化, 帮助解决多机器人在地下环境中同时面临的通信不稳定、协作复杂和环境动态变化的多重挑战.

3　地下通信小尺度衰落模型

在地下隧道等环境中, 由于缺乏网络基础设施, 集中式通信往往无法满足机器人的通信需求 [21]. 为降低通信开销, 通常为机器人搭载无线通信设备, 形成动态自治的移动自组织网络 (ad hoc network). 在无线自组网中, 机器人仅与相邻机器人共享观测值、障碍物位置等信息, 以降低通信带宽的需求.

在时刻 t, 多机器人系统的通信自组网可以表示为拓扑图 $G=(V,E,S)$, 来描述受限条件下通信链路的动态连通性 [22]. 其中节点集合 $V=\{v_1,v_2,\ldots,v_n\}$ 表示 n 个机器人个体, 边集合 $E\subseteq V\times V$ 描述了机器人间的通信连接关系, 边权重使用强度矩阵 $S\in\mathbb{R}^{n\times n}$ 表示, 用于描述通信链路的质量, 边权重越高, 反映机器人能够获取更多来自其他机器人的观测信息.

通信强度的权重值依赖于信噪比 (SNR) 这一参数, 当信噪比 SNR 低于一定阈值时, 通信断开. 为了构建边权重矩阵 $S\in\mathbb{R}^{n\times n}$, 在分析井下巷道电波传播特性的基础上, 综合考虑地下环境中的路径损耗与多径衰落效应, 建立相应的信道模型. 信道衰落存在大尺度衰落与小尺度衰落两种特性, 如图 1 所示, 大尺度衰落主要由阴影效应引起, 小尺度衰落由多径传播引起, 机器人之间的通信链路直射路径较少, 电磁波传播经历反射、折射、绕射和散射后, 不同信号分量的相位、幅度和到达时间存在差异, 导致信号在较短空间尺度内剧烈波动.

对于矿井巷道的无线信道, 接收信号的平均功率随着传播距离呈现对数衰减, 在实际应用中, 可结合测量数据或仿真结果来确定路径损耗指数和参考损耗, 以提高模型精度. 基于加拿大 CANMET 实验室的井下巷道测量数据等, 信道模型考虑了路径损耗指数、阴影衰落标准差以及煤层透射损耗等参数 [23, 24]. 构建的改进型对数距离损耗模型如下:

$$PL_d=PL_0+10n_{\text{env}}\log_{10}\left(\frac{d}{d_0}\right)+X_\sigma+L_{\text{coal}},\tag{1}$$

其中, $PL_0=20\log_{10}\left(\frac{4\pi d_0}{\lambda}\right)$ 为参考距离 d_0 处的自由空间路径损耗, n_{env} 为路径损耗指数, 用于度量路径损耗随传输距离增大的速率, d 为收发两端距离, λ 为电磁波波长, $X_\sigma\sim\mathcal{N}(0,\sigma^2)$ 为阴影衰落, 服从正态分布, L_{coal} 为

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

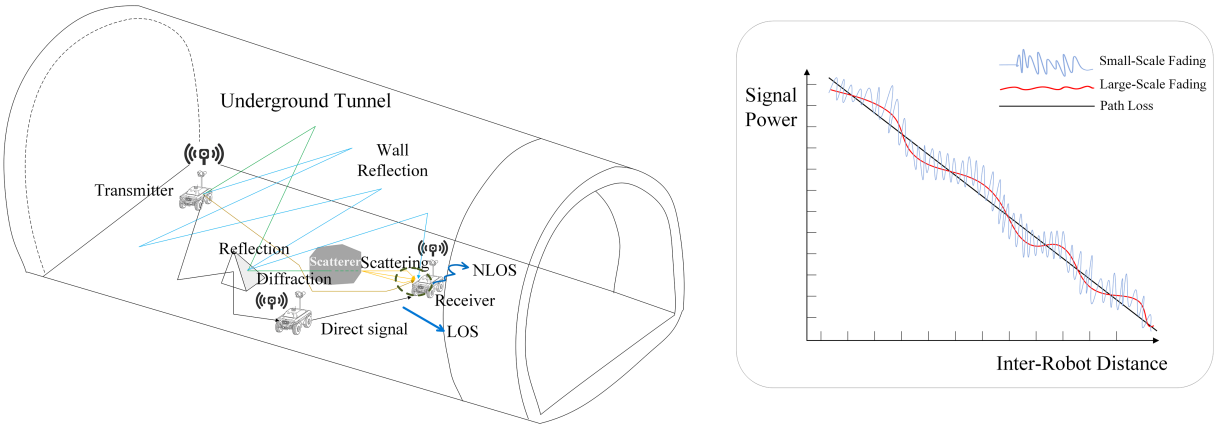


图 1　(网络版彩图) 地下机器人无线通信的多径衰落影响.

Figure 1　(Color online) The multipath fading influence of wireless communication for underground robots.

煤层透射损耗, 与煤层厚度 d_{coal} 和介电常数 ϵ_r 相关, 计算公式为

$$L_{\text{coal}} = (8 + k_\epsilon(\epsilon_r - 2.8))\, d_{\text{coal}}, \tag{2}$$

其中, k_ϵ 为介电常数对透射损耗的影响系数. 根据发射增益 G_r 和接收增益 G_t, 可计算路径增益为

$$G_{\text{path}} = 10^{-\frac{PL_{\text{d}}}{10}} \cdot 10^{\frac{G_r+G_t}{10}}. \tag{3}$$

在多径传播环境中, 机器人接收信号是多个路径信号的矢量叠加. 由于路径时延和相位差异, 合成信号的幅度 (包络) 会呈现随机特性, 其统计规律可通过莱斯 (Rician) 和瑞利 (Rayleig) 分布的概率模型进行描述 [25]. 瑞利分布描述了存在大量散射体且没有直射路径 (non line-of-sight, NLOS) 的信号幅度, 而莱斯分布适用于存在主导视距路径 (line-of-sight, LOS) 的混合传播场景.

考虑天线增益、路径损耗及多径分量等因素, 信道系数 h 可建模为式 (4), 其幅度服从莱斯分布. 当直射分量 h^{LoS} 被严重阻挡, 信号环境变得高度散射和反射, 则衰落符合瑞利分布.

$$h = \sqrt{G_{\text{path}}} \cdot \left(\sqrt{\frac{\beta}{\beta+1}} h^{\text{LoS}} + \sqrt{\frac{1}{\beta+1}} h^{\text{NLoS}} \right), \tag{4}$$

其中, h^{LoS} 为直射链路分量, h^{NLoS} 为多径散射分量, 近似为复高斯分布, $h^{\text{NLoS}} \sim \mathcal{CN}(0, \sigma^2)$, β 为莱斯因子, 表示 LOS 与 NLOS 的分量比, 直射分量 h^{LoS} 具有确定的相位, 一般由传播距离决定, 建模为确定性的复数, 如下所示:

$$h^{\text{LoS}} = \mathrm{e}^{\mathrm{j}\phi}, \tag{5}$$

其中, ϕ 表示信号传播过程中产生的相位变化, 受障碍物密度与厚度影响, 当直射路径被完全阻挡时, 视距分量 h^{LoS}=0.

此时, 根据信道系数 h 可进一步计算当前通信链路的信噪比 SNR, 并将其归一化为通信拓扑的权重. 已知 P_t 为发射功率, P_n 为噪声功率, 信噪比的计算公式如下:

$$\text{SNR} = 10 \log_{10} \left(h \frac{P_t}{P_n} \right). \tag{6}$$

4　基于时空图强化学习的多机器人轨迹规划方法

4.1　基于时空图卷积网络的机器人轨迹规划模型

机器人在地下复杂环境中进行轨迹规划时, 传感器接收到的观测信息是决策的重要输入, 时空图卷积网络

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

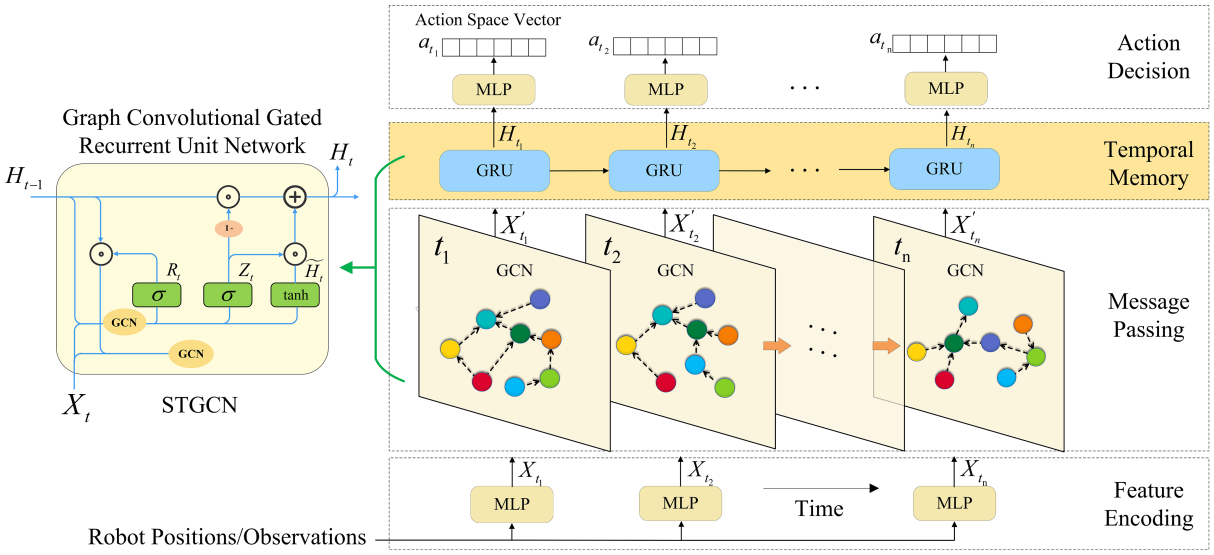


图 2　(网络版彩图) 基于时空图卷积网络的轨迹规划模型.

Figure 2　(Color online) Trajectory planning model based on spatio-temporal graph convolutional networks.

策略模型将观测值表示为图节点特征, 并通过图结构中的消息传递机制实现多机器人间的信息共享, 帮助机器人理解障碍物分布及周围环境结构. 每个机器人节点的特征向量定义为 $x_i = \left[p_i^s, p_i^e, o_i^{(k)}\right]$, 其中 $p_i^s = (x_i^s, y_i^s)$ 为机器人 i 的当前时刻坐标, $p_i^e = (x_i^e, y_i^e)$ 为机器人的目标点坐标, $o_i^{(k)}$ 为机器人 i 观测到的局部环境特征, 包括周围静态障碍物及动态障碍物的信息. 该局部环境特征以一组采样坐标的形式进行编码, 机器人以观测半径 r_{obs} 选取其邻域内障碍物点集, 其中每个障碍点的位置都相对于机器人中心位置进行坐标偏移, 如下所示:

$$o_i^{(k)} = \{(x_{o_j} - x_i^s, y_{o_j} - y_i^s) \mid j = 1, \ldots, k\}. \tag{7}$$

传统时空图神经网络模型主要基于切比雪夫多项式的图卷积实现空间聚合, 而时间建模则多依赖时序卷积或较为简单的递归结构. 针对地下通信受限场景下多机器人协同轨迹规划的需求, 本文在图卷积神经网络中引入动态通信链路的边权变化, 将通信链路质量映射为图结构中的边权, 增强模型在通信受限环境中的适应能力. 模型整体由编码层、信息传递层、时序记忆层和决策层 4 个部分组成, 其结构如图 2 所示.

首先, 将包含机器人位置、传感器观测值的节点特征输入多层感知机 (MLP) 中进行特征编码, 这一过程主要将机器人的位置和环境感知信息映射为适合神经网络处理的特征表示 X_t. 然后, 使用图卷积网络构建模型的信息传递层, 利用通信小尺度衰落模型计算出通信拓扑图 G 的信噪比, 构建具有动态特性的通信边权重矩阵, 图卷积层在每个时间步依据当前边权聚合邻居特征信息, 信息可沿拓扑路径逐层传递, 间接感知其他机器人的状态信息, 并更新自身节点表示为 X_t'. 接着, 利用门控循环单元构建模型的时序记忆层, 记忆机器人接收的信息, 将 X_t' 与历史特征 H_{t-1} 融合为时空特征并传入下一时刻. 最后, 将时空特征 H_t 输入到多层 MLP 中解码为机器人的动作向量 H_t, 输出机器人在时刻 t 的运动决策.

图 2 中, 时空图卷积网络 (STGCN) 由图卷积网络与门控循环单元级联构成, 通过 GCN 对多机器人通信拓扑中的空间信息进行聚合, 并借助 GRU 对连续时间步中的状态信息进行建模, 从而实现对多机器人系统空间 – 时间依赖关系的联合学习.

(1) 图卷积网络. 图卷积网络 [26] 是一种基于卷积计算的图神经网络模型, 其核心思想是通过对节点邻域信息的聚合实现特征更新, 借助 GCN 的邻域聚合机制, 机器人能够在通信网络中共享并融合相邻机器人的观测信息, 实现局部环境感知的扩展与补充.

如图 3 所示, GCN 通过切比雪夫滤波器的卷积操作获取邻域节点的特征向量, 完成 K 跳范围内相邻节点的特征融合. 在动态自组网中, 这种分层递进的信息传播机制使机器人能够通过多跳中继节点逐层整合全局观

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

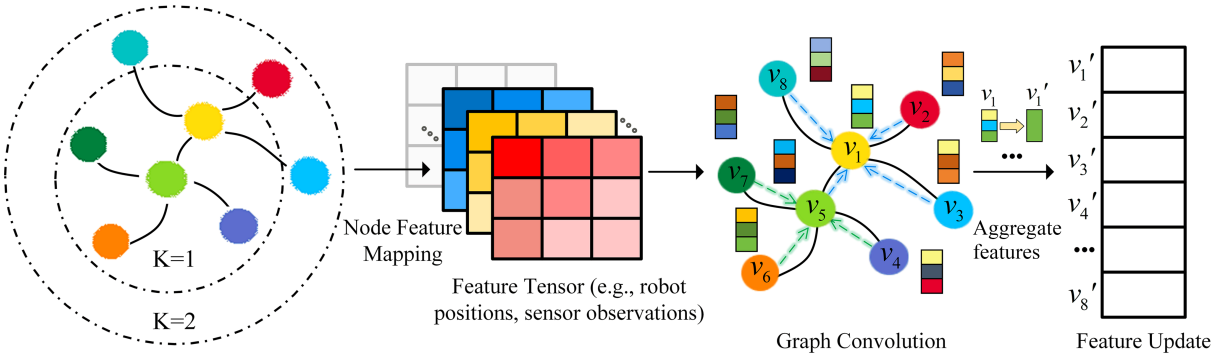


图 3　(网络版彩图) 图卷积聚合节点特征.

Figure 3　(Color online) Graph convolution aggregates node features.

测数据, 即使在通信链路不稳定的复杂场景下仍能构建全局态势认知. 切比雪夫滤波器是一种基于谱图理论的图卷积实现方法, 允许网络聚合多跳邻居信息, 并通过调整多项式阶数 K 灵活控制节点特征的传播范围 [27]. 公式如下:

$$H^{(l+1)}=\sigma\left(\sum_{k=0}^{K}\theta_k^{(l)}T_k(\widetilde{L})H^{(l)}\right), \tag{8}$$

其中, $H^{(l+1)}$ 为 $l+1$ 层网络输出, $H^{(l)}$ 为 l 层的输入, $\theta_k^{(l)}$ 为第 l 层中第 k 个切比雪夫多项式的系数, $T_k(\widetilde{L})$ 是 $\widetilde{L}$ 的 k 阶切比雪夫多项式, $\widetilde{L}$ 是对称归一化后的拉普拉斯矩阵. K 是切比雪夫多项式的最高阶数, σ 为激活函数. 阶数 K 可以控制卷积核的感受野大小, 决定了信息在图结构中的传播范围, 节点可通过多阶邻域逐步感知更远处机器人的状态信息, 维持机器人的协同感知与协同决策.

(2) 门控循环单元. 图卷积网络主要刻画的是单一时间步内的空间依赖关系, 机器人运动状态、通信链路质量以及障碍物分布均随时间动态变化, 轨迹规划过程还需要综合考虑历史运动轨迹以及环境演化趋势. 本文在时空图卷积网络模型中引入门控循环单元 [28] 作为时序记忆模块. GRU 是长短期记忆神经网络 (long short-term memory, LSTM) [29] 的一种简化变体, 通过 “重置门” 和 “更新门” 实现对历史信息保留与更新过程的自适应调控. 通过将 GCN 提取的空间特征与 GRU 建立的时序记忆相结合, 模型能够同时捕获多机器人系统中的空间依赖关系与时间演化特性, 进而增强模型对机器人轨迹的连续优化能力与动态避障性能. 更新门和重置门的状态更新公式如下:

$$\begin{aligned}
Z_t&=\sigma(W_zX_t+U_zH_{t-1}+b_z),\\
R_t&=\sigma(W_rX_t+U_rH_{t-1}+b_r),\\
\widetilde{H_t}&=\tanh(W_hX_t+U_h(R_t\odot H_{t-1})+b_h),\\
H_t&=(1-Z_t)\odot H_{t-1}+Z_t\odot\widetilde{H_t},
\end{aligned} \tag{9}$$

其中, X_t 是输入特征, H_{t-1} 为前一时刻的隐藏状态, W_z 和 U_z 是权重矩阵, b_z 为偏置项, R_t 决定了哪些信息需要被遗忘, Z_t 决定了当前时刻应保留多少前一时刻的记忆信息, 输出 Z_t 值越高, 保存到下一层的信息量越多. $\widetilde{H_t}$ 和 H_t 是候选状态与隐藏状态.

4.2　基于时空图强化学习的轨迹规划算法

在构建的时空图卷积网络 (STGCN) 模型基础上, 提出一种融合时空图神经网络的多智能体双延迟深度确定性策略梯度算法, 通过 STGCN 提取的时空特征作为强化学习中智能体的状态表征, 利用 MATD3 算法 [30] 的分布式 Actor-Critic 架构实现多机器人系统的联合策略优化.

分布式机器人轨迹规划问题可以建模为分布式马尔可夫决策过程 (decentralized partially observable Markov decision process, Dec-POMDP), 使用元组 $G=\langle S,A,P,R,O,N,\gamma\rangle$ 表示, 其中状态空间 $s\in S$ 为地下空间的环境

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

状态, 机器人 i 的部分可观测信息表示为 $o_i \in O$, 联合动作表示为 $a = (a_1, \ldots, a_n)$, 每个机器人 i 的动作空间为 $a_i = (v_i, \theta_i)$, 其中 v_i 和 θ_i 表示机器人的速度和转向角增量.

$P(s'|s, a_1, \ldots, a_N)$ 为状态转移概率, 表示执行联合动作 a 后环境从状态 s 转移到 s' 的概率, 其中机器人的状态变化遵循差分驱动的动力学模型, 每个机器人 i 的位置 (x_i, y_i) 根据速度和转向角进行更新, 新的位置为 (x_i', y_i'), 如下所示:

$$\begin{cases} x_i' = x_i + v_i \cos\theta_i, \\ y_i' = y_i + v_i \sin\theta_i, \\ \theta_i' = \theta_i + \Delta\theta_i, \end{cases} \tag{10}$$

其中, $\Delta\theta_i$ 为转向角增量.

全局奖励函数 R 的设计采用多目标加权奖励机制进行设计, 综合考虑轨迹最短、冲突避免、终点可达等多个规划目标 [31]. 在轨迹规划过程中, 机器人根据环境反馈获得奖励, 通过试错学习, 不断优化自身行为, 每个机器人的目标是学习一组最优策略, 以最大化长期累积奖励. 全局奖励函数 R 的设计如下:

$$\begin{cases} R = R_1 + R_2 + R_3, \\ R_1 = w_1 \times \|\mathbf{p} - \mathbf{p}_{\text{goal}}\|, \\ R_2 = w_2 \times I_2, \\ R_3 = w_3 \times I_3, \end{cases} \tag{11}$$

其中, R_1 是轨迹距离奖励, 机器人与目标位置越近则奖励值越高, R_2 为避障惩罚, R_3 为任务完成奖励, 当机器人成功到达终点时给予较高的奖励值, I_2 和 I_3 为判断碰撞发生和任务完成的因子, w_i 为奖励值的权重参数, 决定了不同目标 (路径长度最短、避障、任务完成) 在总体优化中的优先级和影响力, w_i 设定初步基于任务特性和经验调优, 路径距离奖励权重较大以确保规划效率, 避障和任务完成奖励权重设置保证安全与成功率.

算法框架如图 4 所示. 从图 4(a) 中可以看出, STG-MATD3 使用了双 Critic 网络架构在模型训练时进行策略价值估计, 通过选择较小的 Q 值作为目标值进行梯度更新, 以提升训练稳定性. 同时, STG-MATD3 在 Critic 网络中引入了图注意力机制 (graph attention network, GAT) [32], 使模型能够自适应为每个节点的邻居信息分配不同的权重, 学习节点间的关系强度, 从而更准确地评估多机器人协同行为的整体价值. STG-MATD3 包含 3 个主要网络模块: Actor 网络、Critic 网络和目标网络. 每个模块的结构和功能如下.

(1) 策略 (Actor) 网络. Actor 网络设计为时空图卷积网络的策略网络, 用于生成多个机器人的连续动作向量, 网络结构如图 4(b) 所示, 首先以机器人的局部观测和通信拓扑权重数据作为输入, 通过全连接层将输入特征映射到隐藏维度, 然后经过图卷积门控循环单元层捕捉机器人间的时序依赖性和空间交互, 最后通过全连接层输出机器人的动作策略.

(2) 评价 (Critic) 网络. Critic 网络结构如图 4(c) 所示, 在架构上使用了多层图注意力网络来建模邻域机器人观测信息的重要程度, Critic 网络接收状态数据以及动作输入, 通过注意力机制计算每个机器人与邻居的注意力权重 α_{ij}, 学习机器人 j 的信息对智能体 i 的重要性, 最后经过全连接层输出每个智能体的 Q 值 $Q_\phi(s, a)$, 为 Actor 网络提供价值反馈.

(3) 目标网络. Actor 和 Critic 的目标网络结构与原始网络相同, 算法通过定期将策略网络和评价网络的参数复制到目标网络来稳定训练过程.

STG-MATD3 算法训练过程如图 4(a) 所示, 在探索阶段, 机器人观测值 o_i 输入到 Actor 网络 $\pi_i(a_i|o_i;\theta_i)$ 执行当前动作, 其中 θ_i 是策略网络的参数, Actor 网络使用策略梯度方法更新参数 θ_i 时, 损失函数为

$$J_\pi(\theta) = -\mathbb{E}_{s\sim\mathcal{D}} \min\left[Q_{\phi_1}(s, a), Q_{\phi_2}(s, a)\right], \tag{12}$$

其中, a 是 Actor 网络输出的动作策略, Q_{ϕ_1} 和 Q_{ϕ_2} 是两个 Critic 网络估计的 Q 值, 反映了策略网络在状态 s 下采取特定动作 a 的长期收益.

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

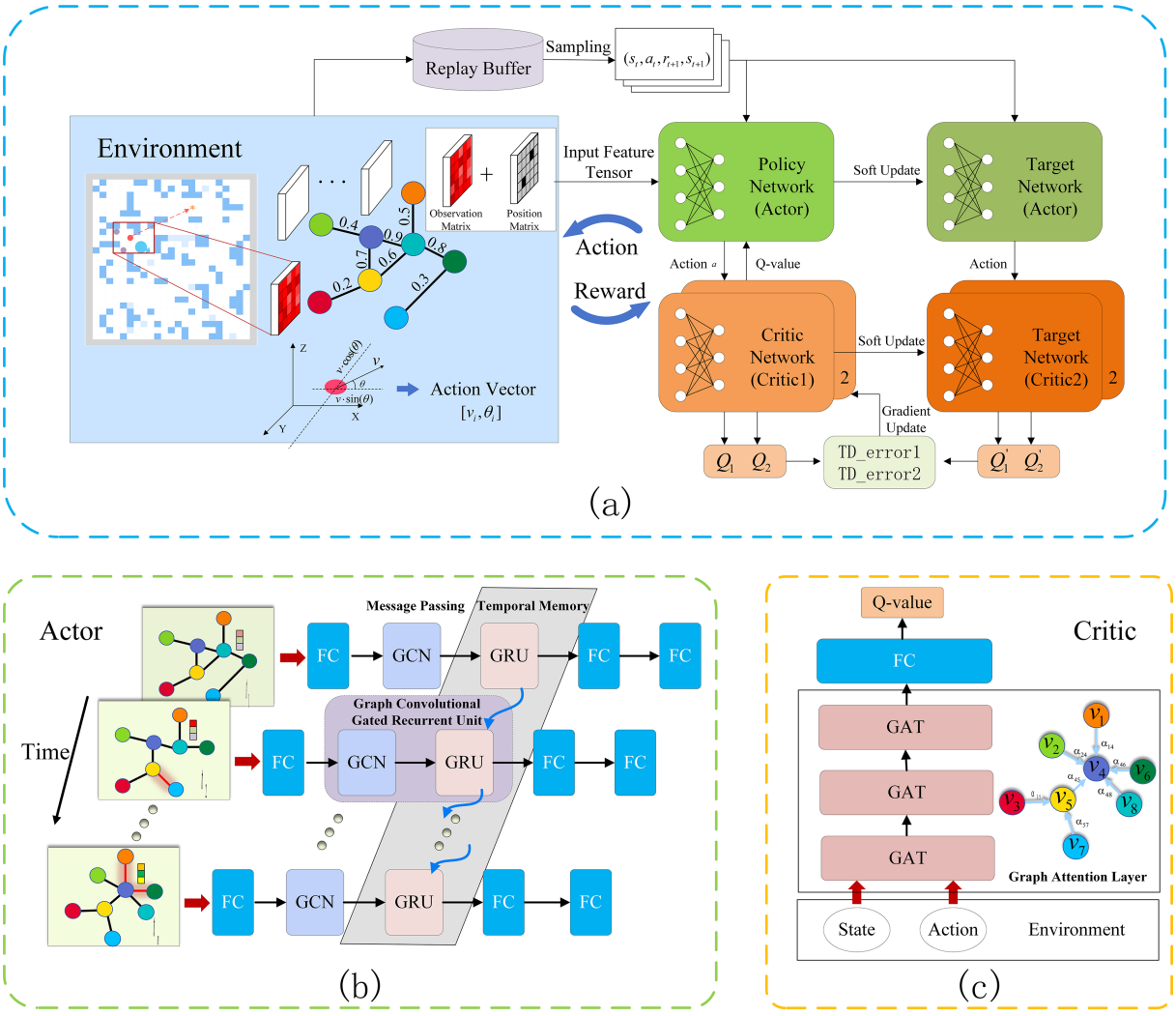


图 4　(网络版彩图) 基于 STG-MATD3 的多机器人轨迹规划算法训练框架. (a) Actor-Critic 架构下算法训练过程; (b) 策略网络架构; (c) 评价网络架构.

Figure 4　(Color online) Training framework of the STG-MATD3-based multi-robot trajectory planning algorithm. (a) Algorithm training process under the Actor-Critic architecture; (b) policy network architecture; (c) critic network architecture.

STG-MATD3 算法使用双 Critic 网络评估每个机器人状态 – 动作对的预期累积回报值, Critic 的目标值主要基于时序差分 (temporal difference, TD) 学习计算, 表示当前机器人状态 – 动作对的期望回报, STG-MATD3 通过目标 Q 值和当前 Q 值之间的均方误差损失函数来优化 Critic 网络, 目标值 y 计算公式与损失函数分别如下所示:

$$y = r + \gamma(1-d)\min\left(Q_{\phi_1'}(s', a'+\epsilon), Q_{\phi_2'}(s', a'+\epsilon)\right), \tag{13}$$

$$J_Q(\phi) = \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}}\left[\frac{1}{2}\sum_{i=1}^{2}\left(Q_{\phi_i}(s,a)-y\right)^2\right], \tag{14}$$

其中, r 为当前状态和动作下的即时奖励; γ 为折扣因子, 用于平衡即时奖励和未来奖励的权重; d 为回合结束的终止标志; $Q_{\phi_i}(s', a')$ 为目标 Critic 网络对一状态 s' 和下一动作的 Q 值估计; ϵ 为动作噪声值; ϕ_i' 和 ϕ 为目标 Critic 网络与当前 Critic 网络的参数.

STG-MATD3 的算法流程如算法 1 所示.

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

Algorithm 1 Graph-based reinforcement learning.

Input: Environmental parameters, hyperparameters, policy network π_θ, target actor $\pi_{\theta'}$, critic networks Q_{ϕ_1} and Q_{ϕ_2}, target critics $Q_{\phi_1'}$ and $Q_{\phi_2'}$, replay buffer $\mathcal{D}$.

Main Loop:

for episode = 1 **to** max_episodes **do**
  Reset environment to get initial observations $\{o_i\}$;
  Construct communication graph G_t with node features $\mathbf{x}_i = [p_i^s, p_i^e, o_i^{(k)}]$;
  Initialize hidden state $\mathbf{h}_0$;
  for time step $t = 1$ **to** max_steps **do**
    Generate action a_t via Actor network;
    Execute a_t, receive reward r_t, next observation o_{t+1}, done flag d_t;
    Update graph to G_{t+1}, store transition $(G_t, h_{t-1}, a_t, r_t, G_{t+1}, h_t, d_t)$ in $\mathcal{D}$;
    Update hidden state: $\mathbf{h}_t = \mathrm{GConvGRU}(\mathbf{x}_t, G_t, \mathbf{h}_{t-1})$;
  end for
  for gradient update **do**
    Sample batch $B = \{(s_i, h_i, a_i, r_i, s_i', h_i', d_i)\}$ from $\mathcal{D}$;
    Compute target $y = r + \gamma(1-d)\min\left(Q_{\phi_1'}(s', a' + \epsilon), Q_{\phi_2'}(s', a' + \epsilon)\right)$;
    Update critic parameters ϕ_1, ϕ_2;
    Update actor parameter θ via policy gradient;
    Soft update target networks:
    $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$;
    $\phi_1' \leftarrow \tau\phi_1 + (1-\tau)\phi_1'$;
    $\phi_2' \leftarrow \tau\phi_2 + (1-\tau)\phi_2'$;
  end for
end for

Output: Trained policy network π_θ and critic networks Q_{ϕ_1}, Q_{ϕ_2}.

表 1　环境参数.

Table 1　Environmental parameters.

Parameter	Value	Physical description
$L \times W$	15 m × 50 m	Size of the underground space
Robot diameter	0.5 m	Robot diameter
Robot observation range	5	Robot perception radius
$[v_{\min}, v_{\max}]$	[0.2, 0.5] m/s	Robot velocity range
$[\theta_{\min}, \theta_{\max}]$	[0, 2π] rad	Steering angle range
O	10	Number of obstacle

5　实验分析

5.1　实验设置

仿真环境的参数设置如表 1 所示, 环境设置为 15 m × 50 m 的地下隧道, 包含 4 个静态障碍物区, 实验在基于网格结构的虚拟环境中进行, 其中机器人、障碍物均在二维网格地图上模拟. 设置 5 个机器人的任务目标为: 从隧道一端的起始点出发, 在避免碰撞的条件下穿越 50 m 长的空间并到达目标区域. 为了增加环境的随机性, 测试算法在复杂动态环境中的适应性, 环境增加了 6 个动态障碍物 (其中 2 个匀速直线移动、1 个随机速度直线移动、2 个随机方向移动), 算法运行的硬件环境为: NVIDIA RTX 4090 显卡、Xeon Gold 6430 处理器、64 GB 内存, 软件环境为 Ubuntu.

STG-MATD3 算法的核心参数设置如表 2 所示, 包括学习率、经验缓存区大小等. 机器人动作控制通过策略网络生成, 网络结构为图卷积门控循环结构, 输入维度和隐藏层维度分别为 16 和 64, 图卷积核大小 K 为 3,

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

表 2　算法参数.

Table 2　Algorithm parameters.

Parameter	Value	Description
α_a	1e−4	Actor network learning rate
α_c	1e−4	Critic network learning rate
Batch size	512	Number of batch samples
Num-episode	3500	Total number of training episodes
σ	0.3	Exploration noise intensity
λ	0.98	Noise decay coefficient per episode
γ	0.9	Reward discount factor
Node dimension	16	Dimension of the state space
Action dimension	2	Dimension of the action space

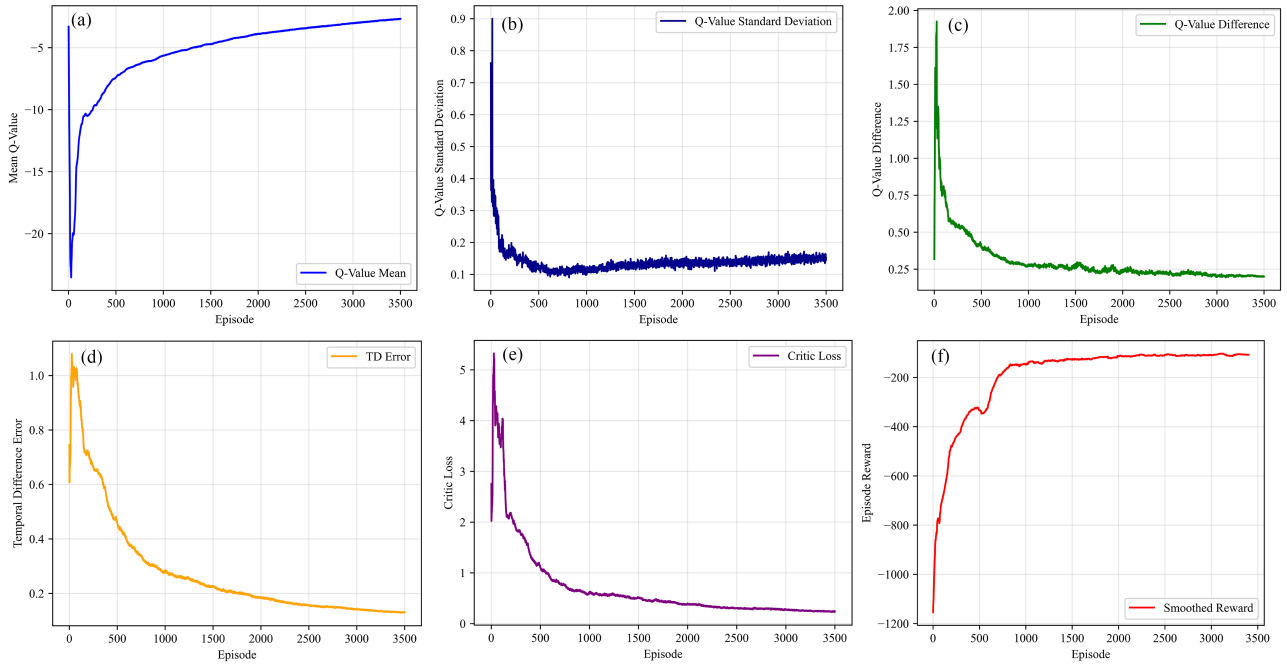


图 5　(网络版彩图) STG-MATD3 训练稳定性指标趋势图. (a) Q 值均值; (b) Q 值标准差; (c) 双 Critic 的 Q 值差异; (d) TD 误差; (e) 目标 Q 值与预测 Q 值的差值; (f) 回合奖励值.

Figure 5　(Color online) Trends of training stability metrics for STG-MATD3. (a) Q-value mean; (b) Q-value standard deviation; (c) Q-value difference between critics; (d) TD error; (e) difference between target and predicted Q-values; (f) episode reward.

输出为二维动作向量. 评价网络采用 3 层图注意力网络, 输入为状态与动作拼接的 18 维特征, 输出为当前动作的值函数估计, 整个训练为深度强化学习过程. 在多机器人协同轨迹规划的场景中, 机器人节点之间的通信负载主要受到邻居节点连接数、消息体积以及通信频率的影响, 实验限制每个机器人最大邻居连接数为 3, 以简化和控制单节点所需处理的通信负载.

5.2　算法性能与稳定性分析

为了评估所提出的 STG-MATD3 算法的性能与稳定性, 图 5 统计了 Q 值的均值和标准差、双 Critic 的 Q 值差异、目标 Q 值与预测值差异、TD 误差和奖励曲线等变化趋势, 以观察算法是否出现训练发散、欠拟合等问题, 为进一步优化算法参数提供依据.

从图 5 中可以看出, 图 5(a) 的 Q 值均值从初期 -23 稳步上升至 -2.7 左右, Q 值估计逐步优化, 使策略

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

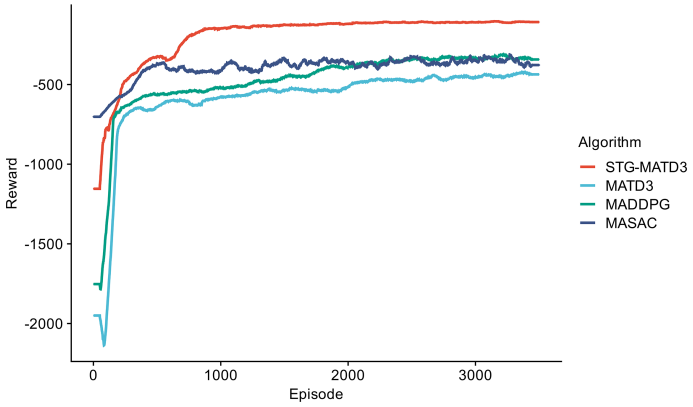


图 6　(网络版彩图) 4 种算法的训练累积奖励曲线对比.

Figure 6　(Color online) Comparison of the cumulative reward curves for the four algorithms.

表 3　STG-MATD3 与基准算法规划效率对比.

Table 3　Comparative analysis of planning efficiency between STG-MATD3 and benchmark algorithms.

Algorithm	η_{task} (%)	L_{avg} (m)	T_{task} (s)	R_{avoid} (%)	T_{step} (s)
STG-MATD3	98.8	186.70	25.90	98.2	0.006
MATD3	0	267.91	48.67	62.8	0.00003
MADDPG	61.8	285.19	44.59	59.6	0.00002
MASAC	72.6	289.80	43.11	45.4	0.001

性能趋于收敛. 图 5(b) 的 Q 值标准差随着训练稳定在接近 0 的水平, 说明 Q 值预测较为稳定, 模型已经收敛, 未出现过拟合问题. 图 5(c) 中双 Critic 差异的变化显著减小, 表明两个 Critic 网络逐渐收敛到一致的 Q 值估计. 在训练中, 双 Critic 的 Q 值差异、TD 误差和 Critic Loss 均收敛在低值 (双 Critic 的 Q 值差异约 0.2, TD 误差与 Critic Loss 约 0.13), 表明 Critic 未出现过估计现象, 预测 Q 值高度接近目标 Q 值, 反映了估计值接近实际动作价值. 图 5(f) 的奖励值在 1000 回合后进入平稳阶段, 无震荡或发散现象, 所有关键指标趋势说明了 STG-MATD3 算法在轨迹规划任务中的训练稳定性和持续优化能力, 为后续部署提供了可靠依据.

5.3　对比实验

为了验证所提算法的性能优势, 选取 3 种主流的多智能体强化学习算法作为基线算法, 将 STG-MATD3 算法与 MATD3 算法 [30]、MADDPG 算法 [33] 和 MASAC 算法 [34] 进行对比, 强化学习算法的目的是通过最大化长期累积奖励, 并根据接收到的奖励信号进行策略更新 [35]. 4 种算法的平均累积奖励曲线如图 6.

图 6 中的红色曲线为 STG-MATD3 算法的奖励曲线, 与其他算法奖励曲线相比, 其奖励值在训练初期迅速超越其他算法的峰值水平, 并在 1000 回合时就趋于稳定, 平均累积奖励值最高. 其次是 MASAC 算法与 MADDPG 算法, 相较于 MATD3 算法, MASAC 算法与 MADDPG 算法收敛速度更快, 但 3 个算法的奖励曲线上升过程都较为平缓, 无法跳出局部最优, 最后收敛时奖励值均低于 STG-MATD3 算法. STG-MATD3, MATD3, MADDPG 和 MASAC 算法训练时间分别为 59155, 70120, 69345 和 60122 s. STG-MATD3 训练时间最短, 说明了 STG-MATD3 算法的策略优化能力更强, 显著提高了机器人群体的决策能力.

在强化学习中, 奖励收敛是策略稳定的必要条件, 但是机器人的轨迹规划策略最优性需通过多维指标体系交叉验证, 为了评估算法在任务效率、安全性与实时性等方面的综合表现, 选取任务完成率 η_{task}、平均任务时间 T_{task}、避障率 R_{avoid}、平均轨迹距离 L_{avg} 和平均决策时间 T_{step} 作为性能指标, 采用蒙特卡罗 (Monte Carlo) 模拟方法进行 500 次独立仿真测试. 表 3 给出了 4 种算法在测试中的性能表现. 其中, 任务完成率 η_{task} 表示机器人成功到达目标点的回合比例, 避障率 R_{avoid} 为无碰撞回合比例, T_{task} 为平均任务完成时间, L_{avg} 为平均轨迹

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

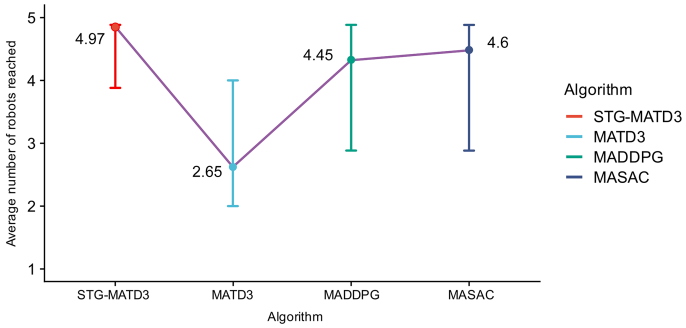


图 7　(网络版彩图) 不同算法在 100 次测试中平均到达目标点的机器人数量, 上界和下界为测试中到达数量的最大值与最小值.

Figure 7　(Color online) Average number of robots reaching the target point over 100 tests for different algorithms, where upper and lower bounds represent the maximum and minimum number of robots reaching the target during the tests.

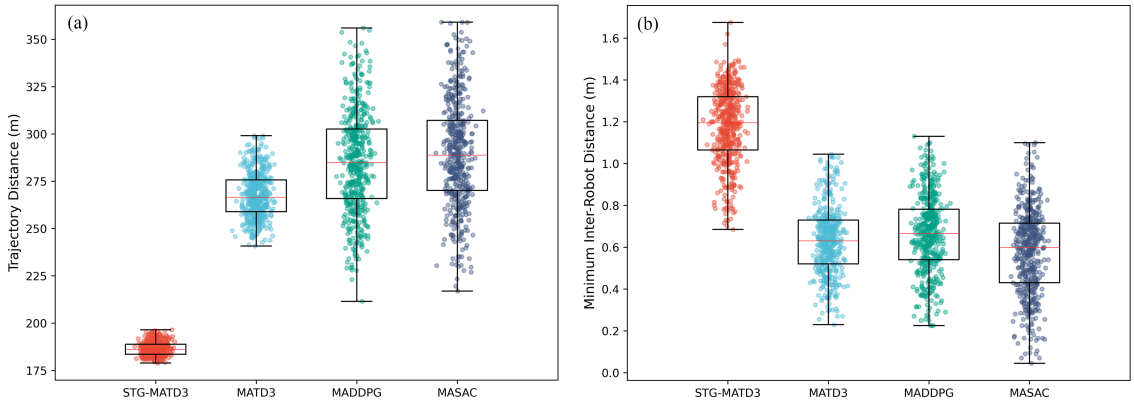


图 8　(网络版彩图) 不同算法的轨迹长度与安全距离分布对比. (a) 轨迹距离分布箱线图; (b) 机器人间最小距离分布箱线图.

Figure 8　(Color online) Comparison of trajectory length and safe distance distributions for different algorithms. (a) Trajectory distance distribution; (b) minimum inter-robot distance distribution.

距离, T_{step} 为平均决策耗时. 由表中数据可知, STG-MATD3 表现最佳, 其任务完成率高达 98.8%, 在平均距离、任务时间及安全性方面均优于基准算法, 虽然平均决策时间 (0.006 s) 略高于其他算法, 但仍满足机器人在动态场景下的实时性需求.

图 7 展示了各算法在 100 测试回合内的机器人成功到达目标点的平均个数, 其中 STG-MATD3 的最小值为 4, 最大值为 5, 平均值达到 4.97, 说明该算法在几乎所有测试中, 均能使所有机器人成功到达目标点, 而且最差情况下仍有较多机器人完成任务. 而 MATD3 算法在测试仅最多到达 4 个机器人, 因此表 3 中该算法的任务完成率为 0%. 从安全性能上来看, STG-MATD3 的安全性最高, 其避障率为 98.2%, 明显高于 MATD3, MADDPG 和 MASAC 这 3 种传统多智能体强化学习算法. 从时间效率上看, STG-MATD3 的平均任务时间为 25.90 s, 耗时最短, 因此在多机器人协同救援、环境探测等对规划效率和稳定性要求较高的地下任务场景中, STG-MATD3 算法可以有效减少机器人的能耗和运行时间, 具备了较强的实用性.

表 3 数据显示, STG-MATD3 算法的 L_{avg} 为 186.70 m. 相较于传统 MADDPG 和 MASAC 算法, STG-MATD3 的平均轨迹距离分别缩短了 18.59% 和 10.78%. 为了更直观地观察 Monte Carlo 测试的轨迹数据分布情况, 图 8(a) 箱线图绘制了 500 次测试的轨迹距离分布, 与基准算法对比, STG-MATD3 生成的轨迹距离更短、数据分布更加集中, 说明该算法输出策略更加稳定. 而 MADDPG 和 MASAC 的轨迹距离上限较高, 说明在部分测试回合中轨迹规划效率较差. 图 9 展示了不同算法的轨迹图, STG-MATD3 规划的轨迹更光滑, 其他算法有抖动和绕远的轨迹存在冗余移动或次优轨迹. MATD3 算法的平均轨迹距离较短是由于该算法在动态障碍物随机移动的环境下, 部分机器人无法成功到达目标点, 环境适应性差. 因此, STG-MATD3 算法能够为机器人更高

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

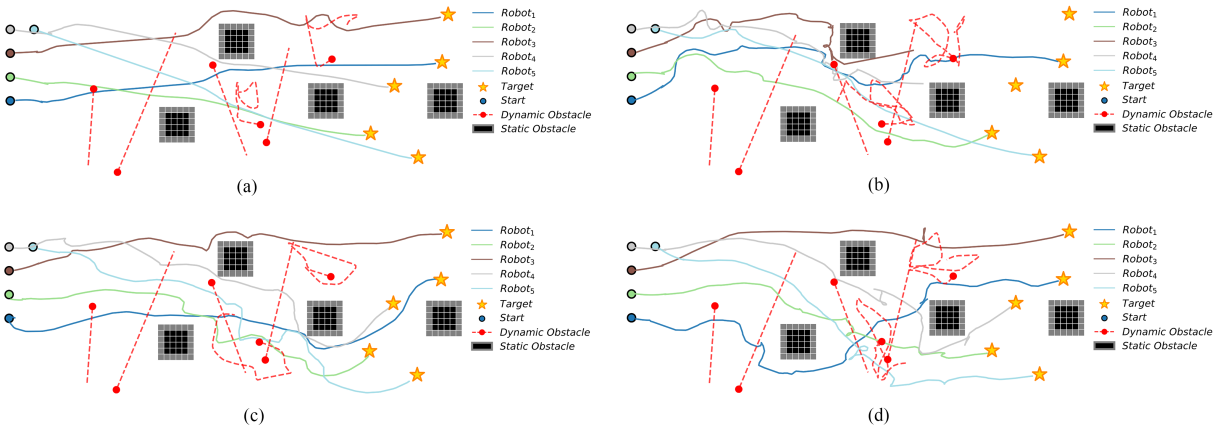


图 9　(网络版彩图) 不同算法的轨迹规划结果. (a) STG-MATD3; (b) MATD3; (c) MADDPG; (d) MASAC.

Figure 9　(Color online) Trajectory planning results of different algorithms. (a) STG-MATD3; (b) MATD3; (c) MADDPG; (d) MASAC.

表 4　不同算法的协同避障效率对比.

Table 4　Comparative of collaborative obstacle avoidance efficiency of different algorithms.

Algorithm	$D_{\min}$ (m)	N_{inter}	N_{col}	CI_{col}
STG-MATD3	1.17	1	11	(0.006, 0.037)
MATD3	0.63	113	557	(0.95, 1.27)
MADDPG	0.66	97	571	(0.96, 1.31)
MASAC	0.57	177	765	(1.35, 1.70)

效地进行轨迹规划, 适合在动态、复杂环境中部署.

表 4 的 D_{safe} 记录了不同算法在 500 次测试中平均最小机间距离 $D_{\min}$、机间冲突回合数 N_{inter}、总碰撞次数 N_{col} 和 95% 碰撞置信区间 CI_{col} 指标. 表中数据显示, STG-MATD3 算法具备更好的协同避障能力, 图 8(b) 箱线图显示了该算法在 500 次测试中机器人间最小距离分布的四分位距集中于 [1.0, 1.4] m 的安全范围, STG-MATD3 的 $D_{\min}$ 为 1.17 m, 高于 MATD3 (0.63 m), MADDPG (0.66 m) 和 MASAC (0.57 m) 算法, 表明 STG-MATD3 算法能够在机器人轨迹规划过程中, 使机器人能够维持更高的安全间距. 从 N_{col} 的数据统计来看, STG-MATD3 仅发生 11 次碰撞事件, 其 95% 碰撞置信区间 CI_{col} 为 (0.006, 0.037), 而对比算法碰撞次数高达 557~765 次, 置信区间均超过 0.95 次. 此外, STG-MATD3 的机间冲突回合数 Ninter 低至 1 次, 进一步证明了该算法在多机器人轨迹规划, 既能使机器人维持 1.17 m 的平均安全间距, 又能将碰撞风险控制在 3.7% 以下, 在轨迹规划的安全性方面展现出显著优势.

5.4　消融实验

所提 STG-MATD3 算法通过结合 GCN 与 GRU 提取多机器人系统中的时空特征, 并作为策略网络用于决策输出. 在局部观测受限的多机器人环境中, STGCN 利用通信拓扑结构和多跳邻居信息传播, 有效支持机器人间的信息扩散和融合, 即使在通信链路受限或断开的情况下, 也能够通过中间节点传递环境和状态信息, 保证全局的态势感知. 为了更准确评估联合动作的价值, 算法在训练阶段引入注意力机制到 Critic 网络中进行决策评估, 改善神经网络训练的收敛速度和稳定性. 为了定量评估图注意力 (GAT)、时序门控循环单元 (GRU)、时空图卷积网络 (STGCN) 和双 Critic 结构在整体框架性能中的独立贡献, 在保持表 2 中算法参数不变的条件下, 针对 STG-MATD3 的核心模块 STGCN-GAT 开展了消融实验. 实验分别设置仅使用图注意力网络 (STGCN)、仅使用时空图卷积网络 (GAT)、仅使用图卷积网络 (GCN-GAT), 以及使用单 Critic 网络 (single Critic) 的对比方案, 系统分析各模块对算法性能的影响, 以验证各模块的有效性. 算法训练过程中的奖励曲线如图 10 所示.

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

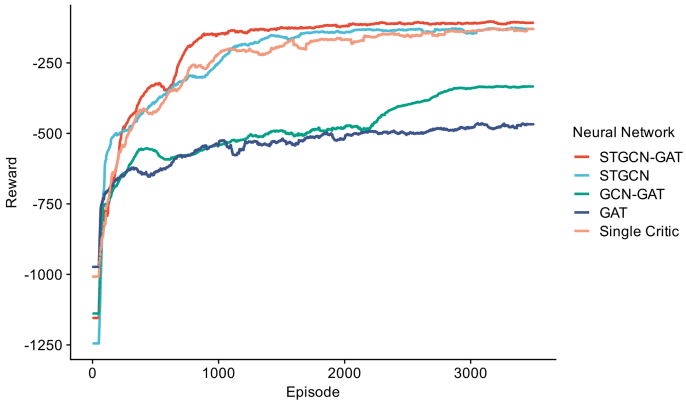


图 10　(网络版彩图) 不同消融策略的累计奖励曲线.

Figure 10　(Color online) Cumulative reward curves for different ablation strategies.

表 5　消融模型指标统计.

Table 5　Ablated models' metric statistics.

Neural network	η_{task} (%)	L_{avg} (m)	T_{task} (s)	R_{avoid} (%)	$D_{\min}$ (m)	T_{step} (s)
STGCN-GAT	98.8	186.70	25.90	98.2	1.17	0.006
STGCN	97	203.14	27.58	98.4	1.12	0.011
GCN-GAT	0	311.98	48.63	87.2	1.30	0.002
GAT	0	270.37	48.71	92.6	1.26	0.00002
Single Critic	95.8	195.32	28.22	96	1.12	0.006

奖励曲线表明, STGCN-GAT 的表现最优, 其整合了时空图卷积网络与图注意力机制, 训练过程中的奖励值增长明显优于其他模型, 在训练约 1000 步后达到一个高且稳定的奖励水平, 证明该结构的策略优化能力最强. 其次是仅使用 STGCN 的浅蓝色曲线, 相比 STGCN-GAT, STGCN 的奖励提升速度略慢, 但仍比仅使用 GAT 和 GCN 的模型好, 说明 STGCN 对时空特征提取的能力较强, 但缺少注意力机制可能导致决策网络无法有效筛选关键信息, 导致收敛速度变慢. 同样将双 Critic 结构变为单 Critic 结构后, 其性能表现与 STGCN 接近, 说明双 Critci 结构与 GAT 作用相同, 可以有效提升算法训练速度和策略优化性能. 将 GRU 模块消融后, GCN-GAT 的奖励提升明显低于 STGCN-GAT, 仅介于 STGCN 和 GAT 之间, 说明 GCN 能够捕捉一定的结构信息, 但缺少时序信息的特征记忆会影响算法的长期决策能力. GAT 的奖励曲线表明仅有 Critic 网络的注意力机制无法完成高效的特征提取, 缺乏时空图卷积网络的模型难以优化策略. 因此, STGCN-GAT+ 双 Critic-MATD3 算法框架的配置在多机器人协同强化学习任务中表现出显著的性能优势, 提升了算法的鲁棒性和适应性.

为了进一步评估不同模块在多机器人轨迹规划中的实际效果, 基于 5.2 小节中提出的部分性能指标, 对各个消融后的模型进行 500 次 Monte Carlo 测试, 结果如表 5 所示. 整体来看, 每个模型的单步决策时间 T_{step} 都在毫秒级相应延迟内, 使用 STGCN-GAT 的模型在大多数指标上均表现最佳, 其任务完成率 η_{task} 达到 98.8%, 保持 186.70 m 的最优平均轨迹距离 L_{avg}, 同时避障率 R_{avoid}、平均任务时间 T_{task} 的指标性能均优于其他消融后的模型. 从箱线图 11(a) 显示的轨迹距离分布情况来看, STGCN-GAT 输出最为稳定, 轨迹距离集中在 190 m 左右的窄箱体分布. 仅用单 Critic 的网络架构和仅用 STGCN 模型的算法任务完成率稍低, 分别为 95.8% 和 97%, 且轨迹距离分布更为分散, 表明其决策稳定性相对不足, 说明了 MATD3 的双 Critic 网络架构通过最小化双 Q 值估计偏差, 能够有效抑制过估计偏差, 同时结合 GAT 提高了策略优化过程的稳定性. 表 5 中, 仅使用 GCN-GAT 和 GAT 的模型平均轨迹距离明显较高, 且任务完成率均为 0%, 表明 GRU 和 STGCN 在小尺度通信衰落环境下对保持任务可执行性具有关键作用. 从表 5 中的避障率来看, GCN-GAT 的避障率为 87.2%, 虽然该模型的平均最小机间距离最高, 但图 11(b) 的距离分布中最为分散, 进一步验证了仅靠空间特征提取, 缺少 GRU 的时空

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

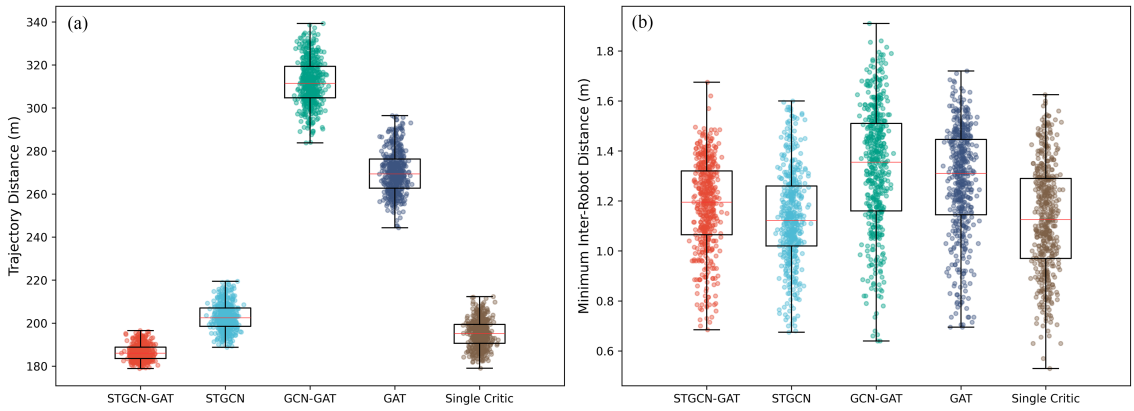


图 11　(网络版彩图) 不同消融模型的轨迹距离与安全距离分布对比. (a) 轨迹距离分布箱线图; (b) 机器人间最小距离分布箱线图.

Figure 11　(Color online) Comparison of trajectory and safe distance distributions across ablated models. (a) Trajectory distance distribution; (b) minimum inter-robot distance distribution.

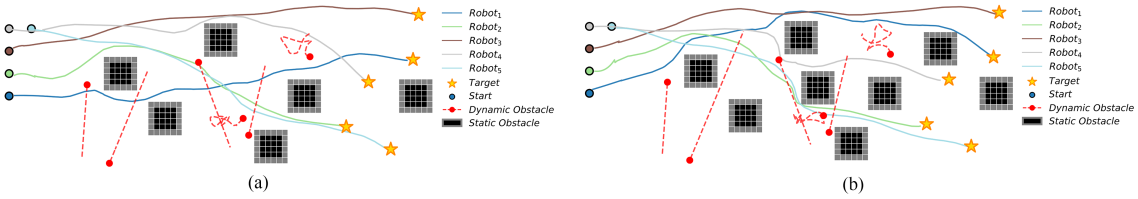


图 12　(网络版彩图) 障碍物密度提升下的多机器人轨迹规划图. (a) 障碍物密度提升 50%; (b) 障碍物密度提升 100%.

Figure 12　(Color online) Multi-robot trajectory planning visualization with increased obstacle densities. (a) 50% increase; (b) 100% increase.

特征记忆在一定程度上降低了机器人的协同避障能力.

通过 GAT 模型其他模型比较分析可知, 在通信受限场景下, 时空图卷积神经网络是保证多机器人系统鲁棒性的核心模块. STGCN-GAT 模块基于当前动态通信拓扑构建图结构边权, 能够有效聚合多机器人系统的环境感知信息, 实现动态多跳信息融合和邻居重要性评估, 为机器人的策略网络提供更丰富的输入特征, 使模型训练时更快收敛. STGCN 中的时序记忆模块通过长短期信息融合, 增强了机器人的协同安全控制能力, 进一步融合时空特征强化了机器人的全局态势感知能力, 有效缓解了小尺度衰落环境下通信不稳定导致的感知受限问题, 从而提高多机器人轨迹规划的协调性和安全性, 减少规划冲突和冗余路径.

5.5　算法泛化性测试

为了评估 STG-MATD3 算法在环境中的鲁棒性与泛化能力, 将静态障碍物密度分别提升 50% 和 100% 后, 验证算法的轨迹规划性能, 测试同样采用 500 次独立蒙特卡罗仿真, 以确保统计显著性. 轨迹规划可视化结果如图 12 所示. 在障碍物密度提升至原来的 1.5 倍和 2 倍的情况下, STG-MATD3 仍能生成平滑的多机器人协同轨迹, 轨迹曲率变化自然, 体现了时空图注意力机制对障碍物的感知与动态规避能力.

表 6 给出了 500 次蒙特卡罗测试的量化指标统计, 增加了 95% 碰撞置信区间 $\mathrm{CI}_{\mathrm{col}}$ 这一指标量化分析系统安全性. 结果显示 STG-MATD3 在障碍物密度提升 50% 和 100% 的场景下, 仍能实现高于 90% 的任务完成率. 与原始场景相比, 障碍物密度的增加仅导致路径长度增幅为 4.68%, 路径长度依然接近全局较优水平. 与表 3 中数据相比, 障碍物密度提升时, 机器人避障率有所降低, 但维仍处在安全可控范围内, 在 +100% 密度下, 95% 的碰撞置信区间为 [0.216, 0.347], 平均每轮碰撞 0.28 次, 相比 +50% 障碍物密度的场景, 碰撞置信区间收窄, 系统稳定性增强, 说明该算法在高密度障碍物环境下采用的是 “预防性避障” 策略, 而非 “反应性避障”, 有效提升了系统的安全性.

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

表 6　障碍物密度对规划性能的影响.

Table 6　Impact of obstacle density on planning performance.

Obstacle density increase	η_{task} (%)	L_{avg} (m)	T_{task} (s)	R_{avoid} (%)	CI_{col}
+50%	99.4	193.13	24.16	80.2	(0.277, 0.43)
+100%	92.0	195.44	26.17	83.8	(0.216, 0.347)

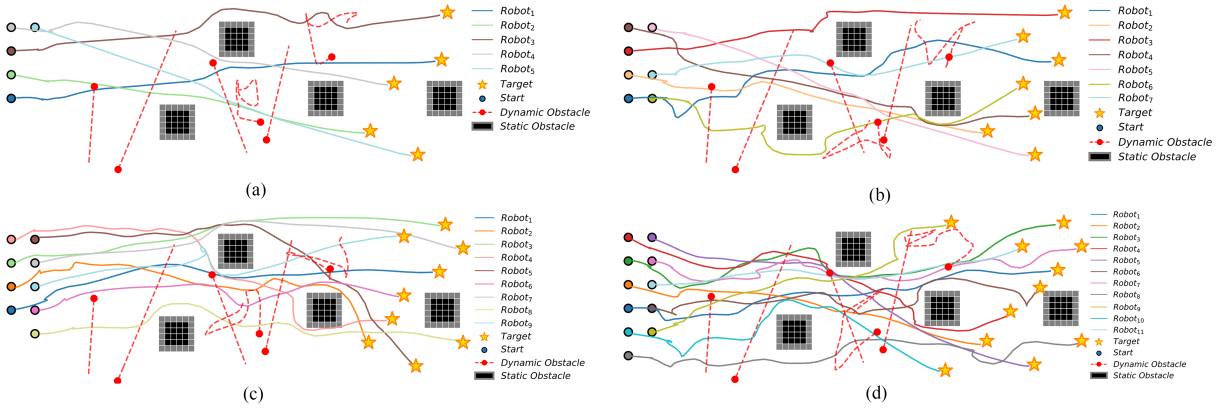


图 13　(网络版彩图) 4 种数量规模下多机器人轨迹规划图. **(a)** $n = 5$; **(b)** $n = 7$; **(c)** $n = 9$; **(d)** $n = 11$.

Figure 13　(Color online) Multi-robot trajectory planning visualization under four robot population scales. (a) $n = 5$; (b) $n = 7$; (c) $n = 9$; (d) $n = 11$.

表 7　不同机器人数量下的规划性能与冲突统计.

Table 7　Planning performance and conflict statistics under different numbers of robots.

Number of robots	η_{task} (%)	R_{avoid} (%)	$D_{\min}$ (m)	N_{inter}	CI_{col}
5	98.8	98.2	1.17	1	(0.006, 0.037)
7	99.4	94.2	0.88	11	(0.058, 0.137)
9	96.0	79.8	0.86	16	(0.267, 0.412)
11	96.8	64.0	0.69	61	(0.801, 1.222)

为了进一步测试 STG-MATD3 算法在机器人数量上的泛化性能, 确定该地下空间中机器人的最优数量, 分别将机器人数量增加至 7, 9, 11 进行多次实验. 由于在探测、救援等任务场景下, 每个机器人独立搜索不同废墟区域、待搜救位置, 因此实验中为每个机器人设置不同的目标点. 如图 13 所示.

图 13 展示了不同机器人数量下的轨迹图. 可以观察到, 随着机器人数量的增加, 部分机器人的轨迹出现了抖动. 为了评估机器人数量对任务执行率和协同避障效率的影响, 表 7 统计了 500 次蒙特卡罗实验中, 机器人数量为 5, 7, 9 和 11 时的关键指标, 新增了机间冲突回合数 N_{inter} 这一指标统计. 根据表 7 的实验数据分析可以看出, 当机器人数量增加到 10 以上时, 任务完成率依然维持在 96.0%~99.4% 的较高水平. 但是机器人的避障率发生下降, 机间冲突回合数急剧增加, 此时机器人间平均最小距离仅为 0.69 m, 95% 置信区间的碰撞次数为 0.801~1.222 次/轮, 说明算法在中等规模任务下具有良好的泛化性, 但当机器人数量增加至临界点 10 以上时, 机间冲突频率的显著上升, 导致轨迹规划的解空间收缩, 局部避障策略的协调难度增加. 未来研究可重点优化协同避障框架, 通过引入分布式冲突消解机制和动态优先级分配等, 进一步提升多机系统的鲁棒性.

6　总结

本文针对地下空间多机器人轨迹规划面临的感知与通信双重受限难题, 构建了一种基于小尺度衰落模型的时空图卷积神经网络, 使各机器人基于局部观测与动态通信拓扑实现信息聚合与时空特征记忆. 在此基础上, 构

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

建了多智能体强化学习框架, 实现了分布式轨迹优化与协同避障. 仿真实验表明, STG-MATD3 算法能够在信号衰减环境中, 使机器人安全高效地抵达目标点, 且具备一定的决策稳定性与泛化能力. 未来工作可以引入元学习 (meta-learning) 机制, 使算法能够从少量小规模场景经验快速适应大规模集群任务, 进一步提升其在复杂地下环境中的泛化性能与可扩展性.

参考文献

1 Chen J, Xin B. Key scientific problems in the autonomous cooperation of manned-unmanned systems. Sci Sin Inform, 2018, 48: 1270–1274 [陈杰, 辛斌. 有人/无人系统自主协同的关键科学问题. 中国科学: 信息科学, 2018, 48: 1270–1274]

2 Zhang T T, Song A G, Lan Y S. Adaptive structure modeling and prediction for swarm unmanned system. Sci Sin Inform, 2020, 50: 347–362 [张婷婷, 宋爱国, 蓝羽石. 集群无人系统自适应结构建模与预测. 中国科学: 信息科学, 2020, 50: 347–362]

3 Qian Q H. Challenges faced by underground projects construction safety and countermeasures. Chin J Rock Mech Eng, 2012, 31: 1945–1956 [钱七虎. 地下工程建设安全面临的挑战与对策. 岩石力学与工程学报, 2012, 31: 1945–1956]

4 Ghaddar M, Molina-García-Pardo J M, Mabrouk I B, et al. UTD-based ray-tracing MIMO channel modeling for the next-generation communications within underground tunnels. IEEE Trans Antennas Propagat, 2023, 71: 5235–5245

5 Becerra R, Azurdia-Meza C, Palacios Játiva P, et al. A wavelength-dependent visible light communication channel model for underground environments and its performance using a color-shift keying modulation scheme. Electronics, 2023, 12: 577

6 Javaid F, Wang A, Sana M U, et al. Characteristic study of visible light communication and influence of coal dust particles in underground coal mines. Electronics, 2021, 10: 883

7 Lin H I, Shodiq M A F, Hsieh M F. Robot path planning based on three-dimensional artificial potential field. Eng Appl Artif Intell, 2025, 144: 110127

8 Wang H B, Yin P H, Zheng W, et al. Mobile robot path planning based on improved A* algorithm and dynamic window method. Robot, 2020, 42: 346–353 [王洪斌, 尹鹏衡, 郑维, 等. 基于改进的 A* 算法与动态窗口法的移动机器人路径规划. 机器人, 2020, 42: 346–353]

9 Chen Y, Guo M, Li Z. Deadlock resolution and recursive feasibility in MPC-based multirobot trajectory generation. IEEE Trans Automat Contr, 2024, 69: 6058–6073

10 Zhang K X, Mao J L, Xiang F H, et al. B-IHCA*, a bargaining game based multi-agent path finding algorithm. Acta Autom Sin, 2023, 49: 1483–1497 [张凯翔, 毛剑琳, 向凤红, 等. 基于讨价还价博弈机制的 B-IHCA* 多机器人路径规划算法. 自动化学报, 2023, 49: 1483–1497]

11 Wang L L, Sui Z Z, Pu Z Q, et al. An improved RRT algorithm for multi-robot formation path planning. Acta Electron Sin, 2020, 48: 2138–2145 [王乐乐, 眭泽智, 蒲志强, 等. 一种改进 RRT 的多机器人编队路径规划算法. 电子学报, 2020, 48: 2138–2145]

12 Yu L B, Cao P Z, Shi L, et al. An improved conflict-based search algorithm for multi-agent path planning. Chin J Aeronaut, 2023, 44: 101–113 [于连波, 曹品钊, 石亮, 等. 基于改进冲突搜索的多智能体路径规划算法. 航空学报, 2023, 44: 101–113]

13 Yang Y, Kiumarsi B, Modares H, et al. Model-free λ-policy iteration for discrete-time linear quadratic regulation. IEEE Trans Neural Netw Learn Syst, 2023, 34: 635–649

14 Yang Y, Pan Y, Xu C Z, et al. Hamiltonian-driven adaptive dynamic programming with efficient experience replay. IEEE Trans Neural Netw Learn Syst, 2024, 35: 3278–3290

15 Li Q B, Gama F, Ribeiro A, et al. Graph neural networks for decentralized multi-robot path planning. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020. 11785–11792

16 Li Q, Lin W, Liu Z, et al. Message-aware graph attention networks for large-scale multi-robot path planning. IEEE Robot Autom Lett, 2021, 6: 5533–5540

17 Du W, Ding S F, Guo L L, et al. Heterogeneous multi-agent reinforcement learning method based on value function decomposition and communication learning mechanism. Chin J Comput, 2024, 47: 1304–1322 [杜威, 丁世飞, 郭丽丽, 等. 基于价值函数分解和通信学习机制的异构多智能体强化学习方法. 计算机学报, 2024, 47: 1304–1322]

18 Li Z, Shi N, Zhao L, et al. Deep reinforcement learning path planning and task allocation for multi-robot collaboration. Alexandria Eng J, 2024, 109: 408–423

19 Feng Z, Wu D, Huang M, et al. Graph-attention-based reinforcement learning for trajectory design and resource assignment in multi-UAV-assisted communication. IEEE Int Things, 2024, 11: 27421–27434

20 Xiao J, Yuan G, He J, et al. Graph attention mechanism based reinforcement learning for multi-agent flocking control in communication-restricted environment. Inf Sci, 2023, 620: 142–157

21 Hao Z T, Guo B, Zhao K X, et al. From rule-driven to collective intelligence emergence: a review of research on multi-robot air-ground collaboration. Acta Autom Sin, 2024, 50: 1877–1905 [郝肇铁, 郭斌, 赵凯星, 等. 从规则驱动到群智涌现: 多机器人空地协同研究综述. 自动化学报, 2024, 50: 1877–1905]

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

22 Zhang Z Y, Duan H B. Virtual error-based distributed cooperative control for multiple UAVs under time-varying switching topologies. Sci Sin Tech, 2024, 54: 1637–1651 [张兆宇, 段海滨. 时变切换拓扑下基于虚拟误差的无人机集群分布式协同控制. 中国科学: 技术科学, 2024, 54: 1637–1651]

23 Wang Y F. Study on channel models of the ultra-wideband wireless communication in underground mines. Dissertation for Ph.D. Degree. Xuzhou: China University of Mining and Technology, 2009 [王艳芬. 矿井超宽带无线通信信道模型研究. 博士学位论文, 徐州: 中国矿业大学, 2009]

24 Chehri A, Fortier P, Tardif P M. CTHp1-8: measurements and modeling of line-of-sight UWB channel in underground mines. In: Proceedings of the IEEE Global Telecommunications Conference, 2006

25 Iqbal N, Luo J, Muller R, et al. Multipath cluster fading statistics and modeling in millimeter-wave radio channels. IEEE Trans Antennas Propagat, 2019, 67: 2622–2632

26 Bruna J, Zaremba W, Szlam A, et al. Spectral networks and locally connected networks on graphs. 2014. ArXiv:1312.6203

27 Liu Y, Wu L, Zhang C Y, et al. Research on node classification and link prediction of knowledge graph in complex network analysis. In: Proceedings of the 2nd International Conference on Electrical, Automation and Computer Engineering, 2024. 1691–1695

28 Cho K, Merrienboer B, Gülehre A, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014. 1724–1734

29 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

30 Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods. In: Proceedings of the International Conference on Machine Learning, 2018

31 Zhang T T, Yang X J. Autonomous coordination saturation attacks method for loitering munitions in urban scenarios based on reinforcement learning. J Command Control, 2023, 9: 457–468 [张婷婷, 杨学军. 基于强化学习的城市场景下巡飞弹自主协同饱和攻击方法. 指挥与控制学报, 2023, 9: 457–468]

32 Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. 2018. ArXiv:1710.10903

33 Lowe R, Wu Y I, Tamar A. Multi-agent actor-critic for mixed cooperative-competitive environments. 2020. ArXiv:1706.02275

34 Fang C L, Yang F S, Pan Q. Multi-UAV collaborative path planning based on multi-agent soft actor critic. Sci Sin Inform, 2024, 54: 1871–1883 [方城亮, 杨飞生, 潘泉. 基于 MASAC 强化学习算法的多无人机协同路径规划. 中国科学: 信息科学, 2024, 54: 1871–1883]

35 Chai L, Zhang T T, Dong H. et al. Multi-agent deep reinforcement learning algorithm based on partitioned buffer replay and multiple process interaction. Chin J Comput, 2021, 44: 1140–1152 [柴来, 张婷婷, 董会, 等. 基于分区缓存区重放与多线程交互的多智能体深度强化学习算法. 计算机学报, 2021, 44: 1140–1152]

https://www.sciengine.com/doi/10.1360/SSI-2025-0193

A spatio-temporal graph reinforcement learning-based multi-robot trajectory planning method for small-scale fading in underground communications

Yan CHEN[1], Tingting ZHANG[1*], Renzhi DONG[2], Xingyu CHEN[1], Hao YANG[3] & Aiguo SONG[4]

1. *College of Command and Control Engineering, Army Engineering University, Nanjing 210007, China*
2. *Cetccloud (Beijing) Technology Co., Nanjing 210007, China*
3. *College of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China*
4. *College of Instrumental Science and Engineering, Southeast University, Nanjing 210017, China*

* Corresponding author. E-mail: 101101964@seu.edu.cn

Abstract In scenarios such as underground exploration and mine surveying, distributed multi-robot cooperation faces communication challenges like wireless signal attenuation and multipath interference, which limit perception capabilities and reduce information sharing, thereby affecting trajectory planning efficiency. This paper proposes a multi-robot trajectory planning method based on spatio-temporal graph reinforcement learning. First, a small-scale fading channel model is used to calculate the edge weights of the communication topology, and a trajectory planning model based on spatio-temporal graph convolutional networks (STGCN) is constructed. Through multi-hop message passing with graph convolution, information transmission across dynamic communication links is achieved, and a gated recurrent unit is introduced to memorize and reconstruct observation information, enhancing the robot's information perception in fading signal environments. Next, a spatio-temporal graph-based multi-agent twin delayed deep deterministic policy gradient (STG-MATD3) algorithm is designed based on spatio-temporal graph neural networks. It uses an Actor-Critic architecture to perform distributed strategy training and optimization, enabling multi-robot trajectory planning. The results show that the STG-MATD3 algorithm achieves a task completion rate of 98.8%, with an average trajectory distance shortened by 10.78%∼18.59% compared to baseline algorithms, significantly outperforming reinforcement learning algorithms such as MADDPG and MASAC. Finally, ablation experiments and generalization tests validate the effectiveness and adaptability of the STGCN model in multi-robot trajectory planning, enhancing the application potential of multi-robot underground exploration.

Keywords multi-robot, small-scale fading, spatio-temporal graph convolutional network, graph reinforcement learning, trajectory planning

https://www.sciengine.com/doi/10.1360/SSI-2025-0193