计算机科学
COMPUTER SCIENCE

基于拥塞感知和缓存通信的多智能体路径规划

张永良，李子文，许家豪，江雨宸，崔滢

引用本文

张永良，李子文，许家豪，江雨宸，崔滢. 基于拥塞感知和缓存通信的多智能体路径规划[J]. 计算机科学, 2025, 52(8): 317-325.

ZHANG Yongliang, LI Ziwen, XU Jiahao, JIANG Yuchen, CUI Ying. Congestion-aware and Cached Communication for Multi-agent Pathfinding [J]. Computer Science, 2025, 52(8): 317-325.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900012

基于拥塞感知和缓存通信的多智能体路径规划

张永良[1]　李子文[1]　许家豪[1]　江雨宸[2]　崔　滢[1]
1 浙江工业大学计算机科学与技术学院　杭州 310014
2 湖州师范学院信息工程学院　浙江 湖州 313000

摘　要　多智能体路径规划任务(MAPF)是大规模机器人系统的重要组成部分。基于冲突搜索的传统规划器受限于计算时间,导致可扩展性低,而基于通信机制的多智能体强化学习策略显著改善了这一问题。随着任务规模的扩大,如何有效通信和避免拥塞成为基于学习方法的主要障碍。针对这些问题,提出了一种基于缓存通信并具备拥塞感知能力的分布式规划器(C3MAP),在合理降低通信频率的同时保持优异的求解成功率。具体而言,当且仅当智能体的可观测信息与上一次通信内容存在显著差异或接收到其他智能体传来的广播请求信号时,才对局部视野内的智能体进行广播通信;同时,引入拥塞信息作为局部可观测信息,以指导智能体避开拥塞区域。基准测试的实验结果表明,C3MAP 在结构化场景中的求解成功率均高于 90%,显著优于现有基于学习的方法,且在大规模场景实验中进一步验证了缓存通信机制优越的稳定性以及拥塞感知的有效性。

关键词:多智能体系统;路径规划;深度强化学习;拥塞感知;缓存通信
中图分类号　TP391

Congestion-aware and Cached Communication for Multi-agent Pathfinding

ZHANG Yongliang[1], LI Ziwen[1], XU Jiahao[1], JIANG Yuchen[2] and CUI Ying[1]
1 College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310014, China
2 School of Information Engineering, Huzhou University, Huzhou, Zhejiang 313000, China

Abstract　Multi-agent Path Finding(MAPF) is an essential component of large-scale robotic systems. Traditional planners based on conflict search are limited in scalability due to computation time, whereas multi-agent reinforcement learning strategies based on communication mechanisms significantly alleviate this issue. As task complexity and scale increase, how to effectively communicate and avoid congestion becomes significant obstacles for learning-based methods. To overcome these challenges, this paper proposes a decentralized planning method called Congestion-Aware and Cached Communication for Multi-agent Pathfinding (C3MAP), which features cache-based communication and congestion-awareness capabilities. Specifically, agents broadcast communications to neighbors only when the current environmental information significantly differs from the previous communication or when receiving request signals from other agents. Additionally, congestion information is incorporated as locally observable information to guide agents in avoiding congested areas. Experimental results on benchmarks indicate that C3MAP achieves a solution success rate of over 90% in structured scenarios, significantly outperforming existing learning-based methods. Additionally, experiments in large-scale environments confirm the greater stability of the caching communication mechanism and the effectiveness of the congestion awareness strategy.

Keywords　Multi-agent system, Path planning, Deep reinforcement learning, Congestion aware, Cached communication

1　引言

多智能体路径规划任务(Multi-agent Path Finding, MAPF)是大规模机器人系统的重要组成部分[1]。目前,多智能体路径规划已经广泛应用于智能仓储和办公机器人[2],其目的在于为一组智能体从起始点规划前往终点的无冲突路径,同时最大程度地减少每个智能体完成任务的时间成本[3]。求解 MAPF 问题的最优解已经被证明是一个 NP 难问题[4],传统基于搜索的集中式求解器(如 CBS[5],ECBS[6])随着智能体规模的扩大,其计算量会呈指数级增长,难以扩展并应用于大规模复杂场景。

随着神经网络的发展,可分布式执行的深度强化学习方法已被广泛用于求解 MAPF 问题[7]。研究员将 MAPF 视为离散的、部分可观测的马尔可夫决策过程,使各智能体依据自

到稿日期:2024-09-02　返修日期:2024-11-22
基金项目:国家自然科学基金青年基金(62102364);浙江省自然科学基金(LY22F020016)
This work was supported by the National Natural Science Foundation of China(62102364) and Natural Science Foundation of Zhejiang Province (LY22F020016).
通信作者:张永良(titanzhang@zjut.edu.cn)

身决策模型获取传感器的观测信息作为状态，并与环境持续交互，选择到达目标的最优行为[8]。基于深度 Q 网络的强化学习能够利用深度神经网络拟合策略函数和价值函数，根据局部观测信息选择 Q 值最高的动作，最大化多智能体的总收益。这种基于学习的求解方法无需任何先验知识，可以有效扩展到任意数量的智能体中，展现出良好的可扩展性。

近期，多智能体间的可训练通信机制受到广泛关注，旨在 MAPF 训练和执行过程中基于通信获取额外信息并增强多智能体的协作[9]。这些方法通常采用广播通信，将局部观测信息传播至一定范围内的其他智能体。但广播通信的缺点在于需要大量带宽，为减少额外通信开销和噪声影响，基于请求应答的选择性通信机制可以主动选择有效通信对象，减少通信频次[10]。而基于局部视野的选择性通信随着智能体规模的扩大，求解效果显著下降，如何确定有效的通信对象以及避免拥塞仍然存在着巨大的挑战。

针对上述问题，本文提出了基于拥塞感知和缓存通信的多智能体路径规划算法(Congestion-Aware and Cached Communication for Multi-agent Pathfinding，C3MAP)，这是一个基于学习的分布式求解器。不同于选择性通信，C3MAP 将可观测视野(Field of View，FOV)内的所有智能体均作为有效通信对象。与全局信息不同，可观测视野内的智能体通信噪声较小，且相互之间的行为决策存在较高的相关性。同时，受文献[11]的启发，各智能体在连续时间步中的观测信息具有相似性。基于此，本文提出缓存通信机制以降低通信频率。C3MAP 进一步引入拥塞信息作为局部可观测信息，促使智能体感知环境中的拥塞情况，避开拥塞区域探索非拥塞路径。实验表明，拥塞信息感知使得智能体能有效避开拥塞区域，提高求解成功率。本文的主要贡献包括：

1)提出了一种基于缓存通信的分布式多智能体路径规划算法(C3MAP)，旨在减少不必要通信的同时规避大规模环境中选择通信对象的问题。

2)提出了一种基于栅格地图的拥塞信息构建方法，并将其作为可观测信息指导路径规划。

3)在主流 MAPF 基准实验上，将所提出算法与多个最先进的基于学习的方法进行对比，验证了 C3MAP 具有更高的求解成功率和良好的求解质量，且随着任务规模扩大，其能保持稳定的求解效果。

2 相关工作

2.1 基于强化学习的多智能体路径规划

随着深度学习的进一步发展，使用深度强化学习来解决 MAPF 问题受到业内研究人员的广泛关注。相较于传统的集中式规划方法[5-6]，基于学习的方法旨在学习一个通用的去中心化规划策略，使得智能体依赖各自的局部可观测信息做出行为决策，这种集中训练分散执行的分布式策略可以更好地扩展到更大规模的任务中，且不会增加计算的复杂度[9]。PRIMAL[8]开创性地将强化学习和模仿学习相结合，仅依赖局部观测信息，就能提供完全去中心化的解决方案。PICO[12]基于 ODrM* 获取智能体的优先级和通信拓扑，以建立有效的避碰机制。TIRL[13]提出基于 Transformer 的策略网络进行特征提取，增强环境中的信息融合。这种结合模仿学习的训练方法通常需要使用 ODrM* 等集中式规划器指导模型的训练，而集中式算法受限于较高的计算耗时，导致模型训练过程缓慢。DHC[14]和 DCC[10]的实验表明，采用单智能体最短路径距离作为启发式指导相比遵循集中式规划器的专家路径更加高效，此外 DCC 提出了请求-应答两阶段选择性通信策略，显著减少了智能体间的通信次数。这种基于通信并使用独立 Q 学习的训练方式，随着任务规模扩大易出现行为振荡、拥堵和选择通信对象的问题。SCRIMP[15]提出基于状态值的优先级策略，允许在冲突发生前，使高优先级智能体先行，减少振荡行为。HELSA[16]基于层次结构，使上层控制器生成子目标，下层控制器采用两阶段注意力机制指导路径规划。但是上述方法通常仅考虑如何解除冲突，忽略了拥塞区域的影响，缺乏对环境信息的有效利用。CRAMP[17]提出基于人群的密度感知策略，改进了智能体在拥挤环境中的行为决策。Alexey 等[18]根据栅格地图中各单元格到可达终点的最短距离获取潜在拥塞值，提出 Follower 模型减少拥塞情况发生。受 CRAMP 和 Follower 启发，本文提出一种基于栅格地图的拥塞信息构建方法，使模型具备拥塞感知能力，可指导智能体避开拥塞区域，减少拥塞情况的发生。

2.2 多智能体强化学习的通信技术

部分可观测性是多智能体强化学习中的一个基本假设，每个智能体都面临一个局部动态的环境，该环境会因其他智能体的不断变化而受到影响，导致多智能体强化学习训练的不平稳问题[19]。智能体间的有效通信是解决该问题的重要策略。通信过程通常会受到成本和噪声影响，SchedNet[20]基于信息重要性，选择部分智能体进行信息传递，减少通信总量，缺点在于需要中央调度器收集权重和通信决策。基于方差控制(Variance-based Control，VBC)[21]将信息向量的方差作为重要性指标，通过预定义阈值过滤方差较小的信息。单独推断通信(Independent Inference Communication，I2C)[22]通过评估智能体通信信息对行为决策的影响，允许单方面决定通信对象。Zhang 等[11]认为智能体周边的环境信息在连续时间上通常具有相似性，提出了时间消息控制(TMC)限制非必要的信息交互，仅在环境信息发生显著变化时进行通信。MAPF 可以被视为合作型多智能体强化学习任务[9]，例如 DCC 基于 I2C 设计了满足 MAPF 的选择性通信机制，显著降低了智能体之间的通信次数。但随着任务规模的扩大，选择性通信会受到信息相似对象的影响，难以确定有效的通信对象，导致求解成功率显著下降。受 TMC 启发，MAPF 中每个智能体在连续的时间步上同样具有环境相似性，本文提出基于缓存的通信机制替换选择性通信，规避通信对象选择的问题，以提高算法的求解成功率和大规模任务中的求解稳定性。

3 问题形式化

3.1 多智能体路径规划问题定义

MAPF 有多个变体，本文主要考虑 Stern 等[23]定义的经典 MAPF 情况，即每个智能体会在每个时间步中同时运动。具体而言，给定一个无向连通图 $G=(V,E)$，其中部分顶点被定义为无法访问的障碍 $V_0 \in V$，每个任务会产生一组 N 个

智能体，索引为 $i\in\{1,2,\cdots,N\}$。每个智能体都有其唯一对应的起点 $s_i\in V$ 和终点 $g_i\in V$，$s_i\neq g_i$。运行时间会被离散为时间步，$t=0,1,\cdots,t_{\max}$。在每个时间步内，智能体可以根据当前环境选择移动到相邻顶点（四邻域），或者驻留在当前顶点。智能体 i 从起点 s_i 到终点 g_i 的路径由一串连续的顶点组成，依次表示每个时间步中智能体在地图上的位置。智能体在抵达各自终点后允许继续移动，仅当所有智能体均到达各自目标位置时，才表示任务完成，任务完成时间为 T。不同任务根据其场景的复杂程度存在最大时间步长 $t_{\max}$ 的时间限制。整个任务中不允许出现智能体之间的行为冲突：顶点冲突和边冲突。顶点冲突指两个智能体在同一时刻 t 移动到同一个顶点 v。边冲突指两个智能体从 t 到 $t+1$ 时间步移动的过程当中，使用相同的边 (u,v) 逆向移动。MAPF 的解是一组无冲突的路径，每个智能体有一个解路径。任务的目标是在给定最大时间步长 $t_{\max}$ 内，找到一组任务完成时间 T 最小的解。

3.2 环境设置

根据上述经典的 MAPF 任务，本文仅考虑离散的 2D 网格环境，其中每个智能体、目标点和障碍物分别占据一个网格单元，且互不重叠。形式上，地图是一个 $m\times m$ 的矩阵，其中 0 表示空单元，1 表示障碍物。对于每一个任务，本文为 n 个智能体随机在空单元中选取 n 个互不重叠的起点和相对应的 n 个目标终点，并且确保每对起点和终点是从同一个连通区域中采样得到的。在每个时间步中，所有的智能体同时进行行为决策，每个智能体可以向上、下、左、右 4 个方向移动，或者保持在原地。此外，在任务求解过程中，智能体不仅可能与障碍物和地图边界碰撞，还可能和其他智能体行为产生冲突，因此本文考虑两种类型的冲突：顶点冲突和边冲突。如果发生冲突，智能体当前行为会被阻止，直至下次移动时冲突解除。

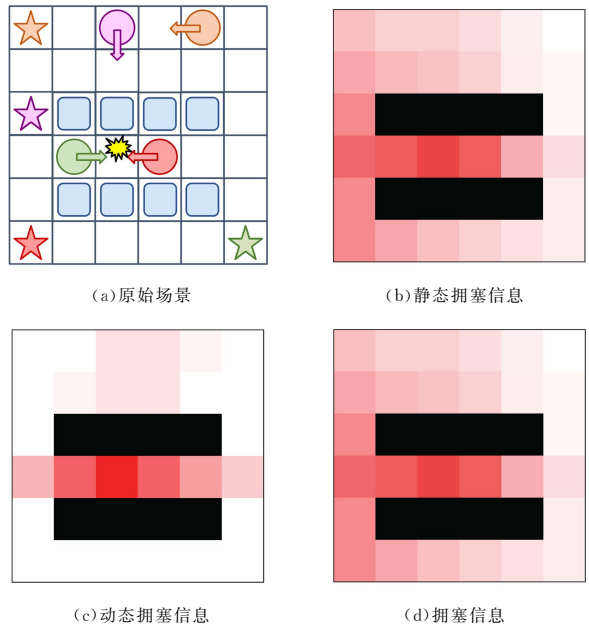
在本文中，考虑 MAPF 任务是一个部分可观测的环境。每个智能体只能获取到 $\xi\times\xi$ 有限区域视野（FOV）内的信息，实验中需要保证 ξ 是奇数，确保智能体位于 FOV 的中心处。可观测信息包含 8 个通道：4 个与 DHC 一致的启发式通道，以及 4 个分别用于表示 FOV 内的其他智能体、障碍物、各单元格拥塞信息和已访问信息的环境信息通道。

4 基于拥塞感知和缓存通信的多智能体路径规划算法

本文提出了一种基于拥塞信息感知和缓存通信机制的多智能体路径规划算法（C3MAP），下面将详细介绍该方法：4.1 节阐述了如何构建拥塞信息，并将其作为智能体局部观测信息的一部分；4.2 节详细介绍了模型的架构设计和训练方式；4.3 节进一步介绍了缓存通信机制的实现细节。

4.1 拥塞信息感知

此前，基于强化学习的 MAPF 求解器通常将智能体局部视野内的其他智能体、障碍物、子目标[15]或者启发式路径信息[14]作为可观测信息输入求解器进行行为决策。但是，随着智能体规模的扩大和环境复杂度的加剧，仅基于上述观测信息的智能体易发生阻塞或死锁现象，导致任务无法在有限时间内完成。受交通路网中车辆拥塞信息的启发，栅格地图中各区域（单元格）的拥塞情况也同样适用于 MAPF 任务，是影响智能体行为决策的重要因素。与此同时，现有研究表明[18]，栅格地图内每个单元到所有可达终点的最短路径的均值越小，则该单元潜在的拥塞可能性越大。基于此，本文提出了拥塞感知的概念，即从环境中抽象出拥塞信息作为可观测信息，用于指导智能体避开拥塞区域，从而减少拥塞情况的发生并辅助拥塞区域内的智能体解除阻塞状态。如图 1 所示，拥塞信息由静态信息和动态信息两个部分组成。



(a)原始场景　(b)静态拥塞信息　(c)动态拥塞信息　(d)拥塞信息

图 1　拥塞信息的组成

Fig. 1　Composition of congestion information

静态信息用于表示地图中每个单元潜在的拥塞可能性。首先在任务初始化，采用广度优先搜索策略计算地图中各单元格 (x,y) 到其可达终点 p 的最短步长 $path\_step^{p}_{x,y}$。而后对各单元格到所有其可达智能体终点的最短步长求平均值 $agv\_step(x,y)$：

$$agv\_step(x,y)=\sum_{p\in V_{\text{arr}}(x,y)}\frac{path\_step^{p}_{x,y}}{|V_{\text{arr}}(x,y)|} \tag{1}$$

其中，$V_{\text{arr}}(x,y)$ 表示 (x,y) 可达的智能体终点集合；$x\in H$，$y\in W$ 表示地图中单元格的坐标位置；W，H 分别表示地图横、纵坐标的长度。

对于地图中的一个网格单元 (x,y)，其 $agv\_step(x,y)$ 越小，则可能选择该单元作为解路径的智能体数量就越多。因此，在不考虑智能体初始分布位置的前提下，距离所有可达目标终点的平均最短步长越小的单元，其潜在拥堵的可能性也就越大。同时，为了确保模型训练的稳定性，对 $agv\_step(x,y)$ 进行标准化处理得到静态拥塞值：

$$cong_{\text{st}}(x,y)=\frac{agv\_step(x,y)-\mu_{\text{agv}}}{\sigma_{\text{agv}}} \tag{2}$$

其中，μ_{agv} 和 σ_{agv} 分别是地图中网格单元的平均最短步长的期望和标准差。

静态信息反映了先验的潜在拥塞可能性。本文进一步引入动态信息，用于表示每个时间步中各网格单元的实时拥塞情况。已知智能体的下一步位置只可能在 4 个相邻单元或者停留在原地，因此如果某网格单元被越多智能体作为下一可

能位置，则该单元在下一时间步的拥塞概率也就越大。同时，一旦多个智能体在地图中发生碰撞，则碰撞点周边易造成小范围的拥塞。未发生碰撞的智能体应尽量避开此类碰撞或拥塞区域，给予拥塞区域内的智能体足够空间进行合理避让。因此，在每个时间步后，针对存在碰撞行为的智能体，从碰撞点开始向外扩散碰撞损失，扩散规则如下。

1）如果被扩散单元是障碍物，则停止扩散。

2）如果被扩散单元是智能体，则根据该智能体向 4 个邻域继续扩散 K 个单元。

3）如果被扩散单元是窄道（一次只可通行一个方向的智能体），则沿着窄道扩散至窄道尽头。

4）如果被扩散单元是普通空单元，且当前相对碰撞点已扩散 E 个单元，则继续向 4 个邻域扩散 $K-E$ 个单元。

若最大扩散步数为 K_m，则该碰撞损失将持续$\frac{K_m}{2}$个时间步。综上所述，可以得到动态拥塞值的计算式：

$$cost_{\text{col}}(x,y)=\begin{cases}\dfrac{K_m}{S}, & \text{if } (x,y)\in V_{\text{diffused}}\\ 0, & \text{if } (x,y)\notin V_{\text{diffused}}\end{cases} \tag{3}$$

$$cost(x,y)=Count(x,y)+cost_{\text{col}}(x,y) \tag{4}$$

$$cong_{\text{dyn}}(x,y)=\frac{cost(x,y)-\mu_{\text{cost}}}{\sigma_{\text{cost}}} \tag{5}$$

其中，$cost_{\text{col}}(x,y)$表示单元格(x,y)的碰撞损失，V_{diffused}表示因碰撞而被罚值扩散到的网格单元集合，S 表示当前单元相对碰撞点扩散了多少步，$Count(x,y)$表示该点作为下一可能位置的智能体数量。对每个单元格的拥塞损失 $cost$ 进行标准化处理，可得到动态拥塞值，其中 μ_{cost} 和 σ_{cost} 表示各单元拥塞损失的期望和标准差。将静态拥塞值和动态拥塞值进行加权求和，得到最终的拥塞信息：

$$congestion(x,y)=\alpha\cdot cong_{\text{st}}(x,y)+(1-\alpha)\cdot cong_{\text{dyn}}(x,y) \tag{6}$$

其中，α 是权重因子，用于调整静态和动态拥塞信息的影响比例，在实验中 $\alpha=0.5$。C3MAP 将生成的拥塞信息作为智能体可观测信息的一部分，与其他环境信息共同送入模型进行特征提取。

4.2 模型设计和训练

不同于为多个智能体分别训练多个策略，C3MAP 从单个智能体的角度出发，将其他智能体作为环境的一部分训练单个策略。训练好的模型可以使用相同的模型参数部署到多个智能体上，每个智能体根据各自的观测信息做出独立的行为决策。整个模型主要由 3 个模块组成：观测编码器、信息汇集模块和 Q 网络，如图 2 所示。下面依次介绍上述 3 个模块的设计细节以及模型的训练策略。

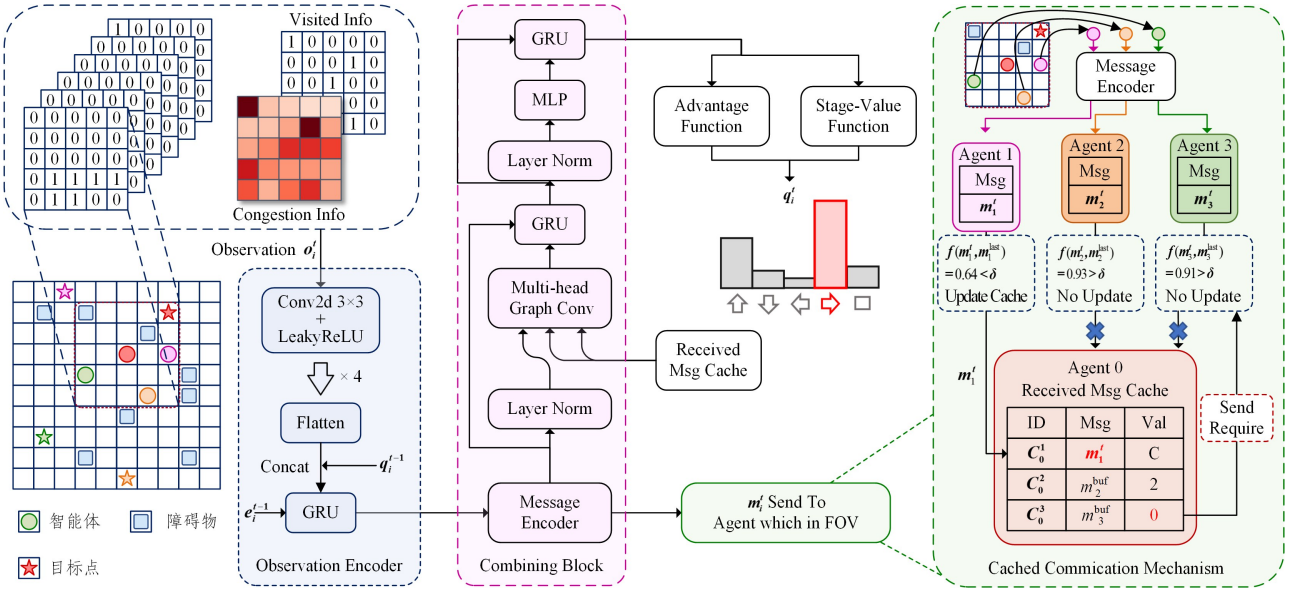


图 2 C3MAP 结构图

Fig. 2 Structure diagram of C3MAP

4.2.1 观测编码器

与 DCC 保持一致，本文同样采用 FOV 大小为 $\xi\times\xi$ 的障碍物通道、智能体通道和 4 个启发式通道作为智能体的局部观测信息。4 个启发式通道分别对应上、下、左和右 4 个动作，每个通道具有和 FOV 相同的大小，仅当智能体采取通道关联的行为能够更加接近目标时，当前位置被标记为 1，此时与最佳行为相关联的通道除去障碍物以外，其余位置应尽可能多的为 1[14]。此外，本文进一步引入 4.1 节中所描述的拥塞信息以及已访问信息作为新增的观测通道。其中已访问信息的计算式如下：

$$Visit_k(x,y)=\frac{Visit\_times_k(x,y)}{Max(Visit\_times_k)} \tag{7}$$

其中，$Visit_k(x,y)$和 $Visit_times_k(x,y)$分别表示智能体 k 在(x,y)单元上的已访问信息和已访问次数。

观测编码器将智能体 i 上述形为 $8\times\xi\times\xi$ 的局部观测信息 o_i^t 作为输入，通过 4 个连续的 3×3 卷积和 LeakyReLU 激活函数得到 $o_i'^t$。将 $o_i'^t$ 展平后与上一时间步的各行为 Q 值 q_i^{t-1} 进行拼接，并与行为编码 e_i^{t-1} 一起送入 GRU 单元，实现连续时间步的特征融合，得到最终的观测编码$\hat{o}_i^t$。

4.2.2 消息汇集模块

在该模块中，首先需要用信息编码器将智能体观测编码

$\hat{o}_i^t$ 转换为消息编码 m_i^t。信息编码器仅由一个全连接层和 GRU 单元组成。m_i^t 会根据缓存通信机制进行消息广播，同时智能体 i 会接收在 FOV 内的其他智能体广播的信息并存放在自身消息缓存中，缓存通信的具体实现细节见 4.3 节。

消息汇集模块的核心功能是基于多头注意力机制[24]的信息汇集。其目的是将智能体的本地观测信息和缓存中其他邻近智能体的消息编码进行信息融合。具体而言，假设在 t 时间步时，智能体 a_i 的 FOV 内有智能体集合 A_i^t，a_i 的本地消息编码 m_i^t 经过层归一化后在每个注意力头上通过矩阵 $\boldsymbol{W}_Q^h$ 映射成 Query，同时将消息缓存中的有效信息 $c_j^t(j\in A_i^t)$ 分别通过 $\boldsymbol{W}_K^h$ 和 $\boldsymbol{W}_V^h$ 映射成 Key 和 Value。智能体 a_i，$a_j\in A_i^t$ 在第 h 个头上的注意力计算式如下：

$$\alpha_{i,j}^h=\mathrm{softmax}\left(\frac{\boldsymbol{W}_Q^h m_i^t\cdot(\boldsymbol{W}_K^h c_j)^{\mathrm{T}}}{\sqrt{d_k}}\right) \tag{8}$$

$$\hat{m}_i^t=Concat\left(\sum_{j\in A_i^t}\alpha_{i,j}^h\cdot\boldsymbol{W}_V^h c_j,\forall h\in\mathcal{H}\right) \tag{9}$$

其中，d_k 是 Key 的维度。随后将 $\mathcal{H}$ 个注意力头上的输出进行拼接得到 $\hat{m}_i^t$。最后，将 $\hat{m}_i^t$ 和 m_i^t 共同输入 GRU 单元，其输出 $\dot{m}_i^t$ 继续通过层归一化、MLP 和 GRU 单元得到消息汇集模块的输出 e_i^t。

$$\dot{m}_i^t=GRU(\hat{m}_i^t,m_i^t) \tag{10}$$

$$e_i^t=GRU(\dot{m}_i^t,MLP(LN(\dot{m}_i^t))) \tag{11}$$

4.2.3 Q 网络和训练过程

基于学习的 MAPF 算法通常采用独立的 Q-Learning 策略进行训练[25]，将训练好的模型使用相同的权重分别部署在各个智能体上。C3MAP 基于 D3QN 进行模型训练。D3QN 结合了 Double DQN[26] 和 Dueling DQN[27] 的优势，将隐藏层输出分别输入状态价值函数 Val 和优势函数 Adv，使用训练网络估计行为，并使用目标网络估计目标 Q 值来改善行为价值高估的情况。智能体 a_i 在状态 s 下采取行为 w 的 Q 值，最终可以通过式(12)计算得到：

$$Q_{\eta,\alpha,\beta}^i(s,w)=Val_{\eta,\alpha}(e_i^t)+Adv_{\eta,\beta}(e_i^t)_w-\frac{1}{|\mathcal{A}|}\sum_{w'\in A}Adv_{\eta,\beta}(e_i^t)w' \tag{12}$$

其中，η 是状态价值函数和优势函数共享的网络参数，α 和 β 分别是状态价值函数和优势函数的网络参数，$\mathcal{A}$ 是智能体的动作空间。获得 Q 值后，基于均方误差(MSE)计算 n 步时序差分误差来训练 C3MAP：

$$\mathcal{L}(\theta)=\mathbb{E}[(R_t^i-Q^i(s_t,w_t;\theta))^2] \tag{13}$$

其中，$R_t^i=r_t^i+\gamma r_{t+1}^i+\cdots+\gamma^n Q^i(s_{t+n},w_{t+n};\theta_t)$，$r_t^i$ 是智能体 i 在时间步 t 获得的奖励，θ 表示训练网络的参数，θ_t 表示目标网络的参数，γ 是折扣因子。训练网络的参数将在预先定义的迭代次数下更新目标网络。

如表 1 所列，本文对 MAPF 任务中的 4 个主要行为进行奖励设计，即移动、停留、发生碰撞、所有智能体到达目标点。为增强环境中拥塞信息对行为决策的影响，额外针对采取何种拥塞行为设计奖罚，若避免拥塞行为表示下一步动作前往拥塞值减少的单元则给予奖励。其中奖励低于罚值，即智能体通过多次正向行为才能抵消一次负面操作，可使智能体的行为决策更加谨慎，避免拥塞。同时，为缓解 MAPF 中的行为振荡，本文设计探索奖励，鼓励智能体探索未访问的次优路径，即当下一步动作前往已访问值减少的单元则给予奖励。探索奖励和惩罚略高于拥塞奖励和惩罚，旨在拥塞和振荡同时产生时优先探索新路径，减少振荡行为，避免出现死锁现象。

表 1 奖励设计

Table 1 Reward design

行为	奖励
移动(上/下/左/右)	−0.055
停留(在目标点/不在目标点)	0，−0.055
发生碰撞	−0.5
所有智能体到达目标点	+10
拥塞行为(避免，陷入)	+0.01，−0.05
探索行为(是，不是)	+0.05，−0.5

4.3 缓存通信机制

本文提出了缓存通信机制，用于替代 DCC 基于 I2C 的选择通信机制。选择通信显著减少了智能体之间的通信次数，其实现过程可以简述为：智能体 a_i 暂时屏蔽其 FOV 内智能体 a_j 的通信信息，得到行为 w_i^{-j}，如果 $w_i^{-j}=w_i$，即屏蔽 a_j 通信信息的前后行为保持一致，则认为 a_j 的信息不会对智能体 i 造成直接影响，故忽略与 a_j 的通信。但是这种决策方式存在一定的弊端，若智能体 a_i 的 FOV 内存在两个智能体对象 a_1，a_2 发送的信息具有较高相似性，则在通信决策上会因为屏蔽 a_1 后存在另一个高相似对象 a_2 而导致 a_i 的行为决策没有变化，从而忽略与 a_1 通信；同理，在继续决策是否与 a_2 通信时受 a_1 影响，进而造成 a_1，a_2 两个有效信息对象都被屏蔽，且不同智能体间消息的联合作用对行为决策的影响难以界定，导致选择通信的有效性难以保证。因此，随着任务规模扩大，选择通信往往会由于邻近智能体信息的相似性而错误地屏蔽了重要的通信对象，使求解效果快速下降。

对于 MAPF 任务，智能体在每个时间步上的观测信息具有时间连续性，因此通过消息编码器得到的消息编码在连续的时间步上具有相似性。基于上述理念，仅当某一时刻的消息编码与此前存在显著差异时，才有必要在 FOV 内进行广播通信。同时，为每个智能体设计独立的消息接收缓存，用于存放各自 FOV 视野内其他智能体广播的信息。算法 1 详述了 C3MAP 的通信流程。

算法 1 智能体 i 的缓存通信过程

输入：智能体 i 在 t 时刻前最新的已广播信息 m_i^{last}，最大有效时间步长 C，通信门限值 δ，智能体 i 局部视野内可观测智能体集 A_i^t，智能体 i 在 t 时刻的观测信息 $\hat{o}_i^t$

1. For t=1，T do
2. 基于当前观测信息 $\hat{o}_i^t$ 生成广播信息 m_i^t
3. If $\mathrm{Cos}(m_i^t,m_i^{last})>\delta$ or $b_{required}$ then
4. 向可观测智能体集 A_i^t 广播信息 m_i^t
5. 将智能体 i 的最近广播信息 m_i^{last} 设置为 m_i^t
6. 接收可观测视野内其他智能体 $a_j\in A_i^t$ 的广播信息
7. 令智能体 i 缓存中存放智能体 j 的消息单元 $c_i^j=m_j^t$，并设置该消息有效值 $val_j=C$
8. For $a_j\in A_i^t$ do
9. If 没有接受到消息 m_j^t then

10. If 缓存中对应消息有效值 $val_j <= 0$ then
11. 向智能体 j 发送广播请求信号 $b_{required}$
12. Else
13. 设置缓存中消息有效值 $val_j = val_j - 1$
14. End for
15. End for

具体而言，在时间步为 t 时，智能体 a_i 会将自身的观测编码 $\hat{o}_i^t$ 通过消息编码器得到消息编码 m_i^t，随后将 m_i^t 和上一次广播信息 m_i^{last} 计算余弦相似度，如果相似度小于 δ，则将 m_i^t 广播至 FOV 内的其他智能体。在同一时刻的广播决策后，智能体 a_i 会接收来自 FOV 内其他智能体 a_j 的广播信息。在接收到 a_j 的广播信息 m_j^t 后，a_i 会将其存放在自身的接收缓存中，同时更新信息有效值 $Val_j = C$；如果未收到 a_j 的广播信息，则将缓存中 a_j 信息的有效值减 1。在 a_i 进行行为决策时，会从消息接收缓存中取出所有在 FOV 内且有效值大于 0 的缓存信息，与自身消息编码 m_i^t 通过消息汇集模块进行信息融合，实现多智能体通信合作。如果在 t 时刻 a_i 未收到在 FOV 内 a_j 的广播信息，且缓存中 a_j 信息有效值为 0，则主动向 a_j 发送广播请求，a_j 在收到请求信号后重新将 m_j^t 广播至 FOV 内其他智能体，以保证信息的完整性。

5 实验与分析

在本章中，本文实现了 C3MAP，并与其他基于学习的方法在多个基准环境上进行实验评估，验证了 C3MAP 出色的求解效果。

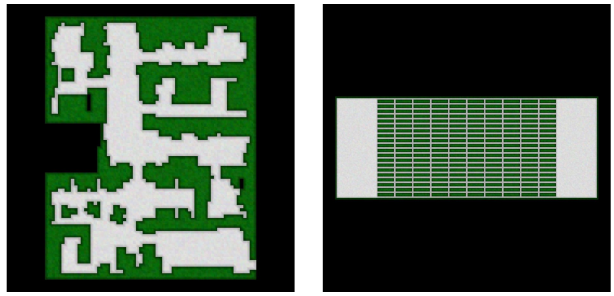
5.1 实验设置

5.1.1 训练环境

C3MAP 采用带有优先经验回放[28]的 D3QN 进行模型训练，并引入课程学习[29]稳定学习过程。课程学习从在 10×10 大小的地图上训练单个智能体开始，到在 40×40 大小的地图上训练 16 个智能体为止，且只有在当前环境中的任务成功率达到 0.9 时，才增加环境的难度(扩大地图大小和增加智能体数量)。该策略可以使模型逐步地从经验回放池中获取并学习越来越复杂的任务。本文在批量大小为 128、序列长度为 20 的情况下对 C3MAP 进行训练，训练过程中 FOV 的大小选用 9×9，最大经验步长为 256。在单张 NVIDIA RTX 4090 和 AMD EPYC 7K62@2.8GHz 服务器上训练 300 000 步大约需要 22 小时。

5.1.2 测试环境

如图 3 所示，本实验从基准测试[3]中选择了两张广泛使用的结构化地图 den312d(65×81)和 warehouse(161×63)作为主要评估场景。对于每一个地图，分别设置{4,8,16,32,64}个智能体以构建不同难度的任务场景，并在每个场景中随机创建 1000 个任务实例对模型进行评估。同时，重新训练了 DHC 和 DCC 以确保代码的正确性，并基于已有的实验数据，将最新基于学习的方法与 C3MAP 在成功率(SR)和任务平均完成步长(EL)指标上进行对比。SR 表示算法在有限时间内解决 MAPF 任务的成功率，EL 表示所有智能体到达其目标位置所花费的平均时间步长。此外，针对不同通信门限值 δ，在结构化地图中对通信频率进行了对比验证。



(a) den312d (b) warehouse

图 3 基准测试地图

Fig. 3 Benchmark maps

为了进一步探索 C3MAP 在大规模 MAPF 场景中的求解稳定性，本文采用和 HELSA[16]相同的实验设置。如表 2 所列，固定环境中障碍物占比为 0.2，针对地图大小、智能体数量、最大时间步长创建不同规模的任务场景，并针对每个场景随机生成 200 个实例用于对比评估 C3MAP 和其他模型的求解稳定性。

表 2 大规模场景参数设置

Table 2 Parameter settings for large-scale scenarios

map_size	n_robots	n_obstacles	max_epi_len
40×40	8	320	256
80×80	32	1280	512
160×160	128	5120	1024
240×240	288	11520	1536
320×320	512	20480	2048
400×400	800	32000	2560

5.2 实验结果

5.2.1 结构化场景求解质量实验

在结构化地图对比实验中，本文将经典的 PRIMAL，DHC，DCC，以及最新的效果优越的 SACHA[30]，SCRIMP 作为基准模型。表 3 列出了 C3MAP 对比基准方法在有限时间内完成任务的成功率和平均完成时间步长。实验结果表明，在环境相对开阔的 den312d 地图上，基于通信机制的学习方法均表现出较为优异的求解效果，且 SCRIMP 和 C3MAP 保持着 100%的求解成功率，这是由于通信机制使得单个智能体可获得超视距的环境信息，因而做出更合理的行为决策，且在较为开阔的环境下振荡、拥塞情况较少。但由于对拥塞区域的感知避让，C3MAP 在提高求解成功率的同时，往往会在面临拥塞时选择次优路径以降低拥塞情况的发生，导致任务平均完成步长变高。

对于 warehouse 地图，受限于环境中存在大量同一时间内一次仅能单向通行的窄道，使得此前基于通信的求解方法随着智能体数量增加，求解效果大幅度下降且成功率低。主要原因是窄道内可移动空间少，当智能体规模较大时，易在窄道处发生边冲突导致阻塞，且仅根据局部视野内的有限信息进行决策，易造成智能体避障后又反复进入窄道，再次出现边冲突的振荡或死锁现象。相较而言，C3MAP 会根据实时环境获取周边拥塞情况作为可观测信息，发生碰撞的窄道和拥塞区域根据所提出拥塞信息计算方式可得到更高的拥塞值，促使智能体在避免拥塞的奖励引导下进行拥塞避让，同时，探索奖励会鼓励智能体选择未行驶过的次优路径，以避免振荡

行为。因此，C3MAP 在 warehouse 地图上有着更好的表现，在各个场景的求解成功率均超过 90%。但是实验显示，即便 C3MAP 在一定程度上可以规避拥塞，但是受限于 FOV 的信息限制，随着智能体数量的增多，智能体仍会因为反复避障而发生振荡或死锁现象，即智能体缺乏长期有效的行为信息记忆能力。我们会在后续的工作进一步探讨这个问题。

表 3 不同多智能体路径规划算法的求解质量

Table 3 Solution quality for different MAPF methods

Map	Agent	PRIMAL		DHC		DCC		SACHA		SCRIMP		C3MAP	
		SR/%	*EL*	*SR*/%	*EL*	*SR*/%	*EL*	*SR*/%	*EL*	*SR*/%	*EL*	*SR*/%	*EL*
den312d	4	40	196.54	**100**	86.56	**100**	82.99	**100**	**81.43**	**100**	82.34	100	83.93
	8	8	245.02	**100**	100.70	99	97.95	**100**	**89.73**	**100**	99.58	**100**	96.06
	16	0	~~256.00~~	**100**	109.24	97	108.29	**100**	**96.74**	**100**	105.78	**100**	107.26
	32	0	~~256.00~~	98	124.38	97	119.15	98	**104.30**	**100**	115.39	**100**	120.12
	64	0	~~256.00~~	93	153.17	93	145.21	94	142.97	**100**	**131.59**	**100**	140.96
warehouse	4	42	355.80	99	146.12	99	135.89	99	134.59	87	197.79	**100**	**133.46**
	8	18	451.82	91	198.82	96	169.50	93	166.72	66	304.22	**100**	**153.58**
	16	8	492.04	74	281.37	90	208.72	76	198.72	48	366.98	**100**	**172.58**
	32	4	505.58	28	432.28	58	335.81	48	354.33	21	451.40	**99**	**193.59**
	64	0	~~512.00~~	1	~~512.00~~	14	473.92	28	437.29	4	504.26	**94**	**245.57**

注：加粗表示同比最佳结果，删除线表示无效求解。

5.2.2 结构化场景通信频率实验

为深入研究 C3MAP 在求解 MAPF 任务过程中，其不同通信阈值 δ 对通信频率和求解成功率的影响，将阈值 δ 分别按 0.95，0.9，0.8，0.7 设置对照组，并与 DHC 和 DCC 在结构化地图中进行对比实验。形式上，DHC 选择最邻近的两个智能体作为通信对象，DCC 则基于因果决策进行选择性通信。

如表 4 所列，当 δ 等于 0.9 时，C3MAP 已趋近于 DCC 的通信开销，且在 warehouse 地图上表现出更少的通信频率，原因是该地图中窄道居多，智能体间通常被障碍物阻拦，而障碍物是影响决策的主要信息，所以传递的抽象环境信息相似度更高，由此可以证明缓存通信可以有效减少通信开销。

表 4 结构化地图中不同算法的通信频率

Table 4 Communication frequency of different algorithms in structured maps

Map	Agent	DHC	DCC	C3MAP			
				δ=0.95	δ=0.9	δ=0.8	δ=0.7
Den312d	4	42.0	2.4	7.2	3.5	1.8	1.2
	8	210.7	11.7	34.8	16.8	8.6	5.5
	16	894.9	51.5	162.5	78.1	38.4	23.5
	32	3335.7	218.8	760.1	357.3	166.3	94.2
	64	11870.6	1306.1	4205.1	2001.9	815.8	453.9
warehouse	4	27.8	7.1	3.9	2.2	1.2	0.7
	8	130.2	40.7	17.7	7.9	5.4	3.3
	16	692.9	148.6	75.2	44.9	24.1	14.4
	32	3619.2	583.6	317.6	184.2	100.3	60.8
	64	19589.3	2341.9	1499.6	878.8	455.7	275.8

结合图 4 发现，在任务运行过程中存在大量与上次信息相似度介于 0.95 到 0.9 之间的冗余信息，这表明降低阈值后通信频率出现了大幅度下降且求解成功率无显著降低。由此可以认为，智能体的可观测信息即便在连续时间步中其具体环境稍有不同，但其所抽象的通信信息具有较高相似性，这种微弱的信息变化并不会显著影响其他智能体的行为决策和任务协作，故可以用缓存的相似信息来替代，以较大程度地降低通信开销，同时保持成功率。对比通信频率和求解成功率可知，当通信阈值低于 0.8 时，虽然通信频率进一步降低，但是该阈值下的信息已经存在明显的协作关系，故求解成功率出现了显著下滑。综上所述，当通信阈值为 0.9 时，C3MAP 在通信频率和求解成功率上达到了良好的平衡，同时我们基于该阈值在后续实验中对模型进行更深入的研究。

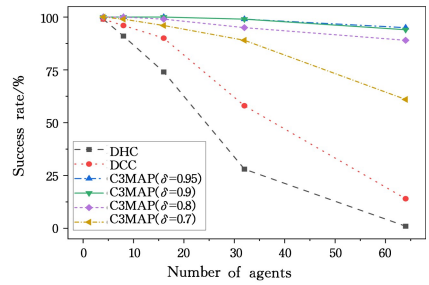


图 4 仓库环境中不同通信阈值下的求解成功率

Fig. 4 Success rate at different thresholds in warehouse

5.2.3 大规模场景实验

对于大规模场景，本文主要对比评估基于部分通信的 DHC 和基于选择通信的 DCC 与 C3MAP 的求解稳定性。如图 5 所示，PRIMAL，DHC 和 DCC 随着 MAPF 任务规模的扩大，其求解成功率出现明显下降，尤其是 DCC，其选择性通信在大规模任务中存在大量的相似智能体信息，易导致重要通信对象被错误地屏蔽，因此其求解效果下降更为明显。HELSA 采用分层结构，基于选择通信融合时间和空间信息获取奖励并鼓励智能体探索新的区域[16]，在大规模任务中保持良好的求解成功率但仍有不足。

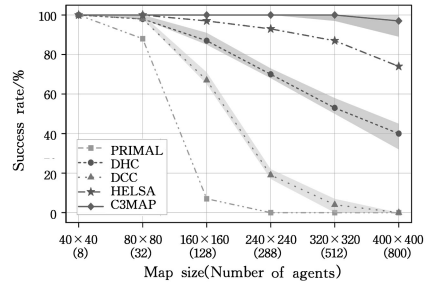


图 5 大规模场景下的求解成功率

Fig. 5 Success rate of solution in large-scale scenarios

相较而言，C3MAP 的求解曲线均高于基线方法且较为平稳，原因主要在于，其采用缓存通信机制，基于连续时间步

中的环境相似性，从缓存或通过广播通信获取局部视野内所有智能体的有效信息进行信息汇聚而做出行为决策，规避了环境中选择通信对象的难题，故在大规模场景中，随着智能体数量的增多，其仍能获取有效通信信息并保持较高的求解成功率。此外，拥塞信息的引入，减少了阻塞、振荡现象的发生，使 C3MAP 相较此前算法有着更高的求解成功率。

5.2.4 消融实验

C3MAP 主要基于两个关键模块：拥塞信息感知和缓存通信机制。为验证这两个模块对模型性能的影响，在(240×240,288)和(320×320,512)两个大规模场景下将 DCC 作为基线，并基于是否使用拥塞信息感知和是否使用缓存通信替换选择性通信引入 4 组不同的模型配置进行消融实验。

表 5 列出了各个模块在大规模环境中对求解 MAPF 任务的影响。由实验数据可见，原始基线方法在大规模环境下求解效果较差，在采用缓存通信机制后，显著提高了任务求解成功率，说明基于环境相似性的缓存通信更适用于具有高可扩展需求的 MAPF 任务。动态拥塞信息的引入进一步提高了求解成功率，验证了环境中实时拥塞状态对求解 MAPF 问题的重要性。然而，仅依赖动态拥塞信息可能会导致智能体从某一拥塞区域逃离至潜在拥塞风险更大的区域，从而增加平均完成步长。对比引入所提拥塞信息构建方法可知，通过将动态和静态拥塞信息相融合，可以有效改善这一问题，静态拥塞信息预先估计了地图中每个单元潜在的拥塞风险，引导智能体选择不易阻塞的次优路径，减少了拥塞情况的发生，保证了解的质量，降低了任务平均完成步长。

表 5 大规模场景下的消融实验结果

Table 5 Ablation experiments results in large scenarios

Cache Comm	Congestion Dynamic	Static Info	240×240,288		320×320,512	
			SR/%	*EL*	*SR*/%	*EL*
—	—	—	19	1375.04	4	2020.76
√	—	—	81	710.85	64	1244.63
√	√	—	100	464.00	99	694.23
√	√	√	100	409.20	100	556.49

结束语 本文提出了一种基于学习的多智能体路径规划求解器，其引入拥塞信息作为可观测信息，指导智能体避开拥塞区域，减少拥塞的发生。本文对选择性通信的潜在问题进行了思考，提出缓存通信机制进行替代，在减少连续时间内不必要通信的同时保证智能体间的有效信息交互。实验结果表明，C3MAP 在复杂的结构化地图中，在保证求解质量的同时显著提高了成功率，并随着任务规模的扩大仍能保持稳定的求解效果。在未来的工作中，我们将继续深入研究智能体行为的长期记忆能力，探索更有效的防振荡、死锁策略，同时期待将该算法扩展至长期多智能体路径规划任务当中，并移植到真实 AGV 上进行实物实验和验证。

参 考 文 献

[1] WANG Z H, TONG X R. Research Progress of Multi-agent Path Finding Based on Conflict-based Search Algorithms[J]. Computer Science, 2023, 50(6): 358-368.

[2] XIN Y X, HUA D Y, ZHANG L. Multi-agent Reinforcement Learning Algorithm Based on AI Planning [J]. Computer Science, 2024, 51(5): 179-192.

[3] STERN R, STURTEVANT N, FELNER A, et al. Multi-agent pathfinding: Definitions, variants, and benchmarks[C]// Proceedings of the International Symposium on Combinatorial Search. AAAI, 2019: 151-158.

[4] BANFI J, BASILICO N, AMIGONI F. Intractability of time-optimal multirobot path planning on 2D grid graphs with holes [J]. IEEE Robotics and Automation Letters, 2017, 2(4): 1941-1947.

[5] SHARON G, STERN R, FELNER A, et al. Conflict-based search for optimal multi-agent pathfinding[J]. Artificial Intelligence, 2015, 219: 40-66.

[6] BARERM, SHARON G, STERN R, et al. Suboptimal variants of the conflict-based search algorithm for the multi-agent pathfinding problem[C]// Proceedings of the International Symposium on Combinatorial Search. AAAI, 2014: 19-27.

[7] SHI D X, PENG Y X, YANG H H, et al. DQN-based Multi-agent Motion Planning Method with Deep Reinforcement Learning[J]. Computer Science, 2024, 51(2): 268-277.

[8] SARTORETTI G, KERR J, SHI Y, et al. Primal: Pathfinding via reinforcement and imitation multi-agent learning[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2378-2385.

[9] CHUNG J, FAYY AD J, YOUNES Y A, et al. Learning team-based navigation: a review of de-ep reinforcement learning techniques for multi-agent pathfinding[J]. Artificial Intelligence Review, 2024, 57(2): 41.

[10] MA Z, LUO Y, PAN J. Learning selective communication for multi-agent path finding[J]. IEEE Robotics and Automation Letters, 2021, 7(2): 1455-1462.

[11] ZHANG S Q, ZHANG Q, LIN J. Succinct and robust multi-agent communication with temporal message control[J]. Advances in Neural Information Processing Systems, 2020, 33: 17271-17282.

[12] LI W, CHEN H, JIN B, et al. Multi-agent path finding with prioritized communication learning[C]// 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022: 10695-10701.

[13] CHEN L, WANG Y, MIAO Z, et al. Transformer-based imitative reinforcement learning for multirobot path planning[J]. IEEE Transactions on Industrial Informatics, 2023, 19(10): 10233-10243.

[14] MA Z, LUO Y, MA H. Distributed heuristic multi-agent path finding with communication[C]// 2021 IEEE International Conference on Robotics and Automation(ICRA). IEEE, 2021: 8699-8705.

[15] WANG Y, XIANG B, HUANG S, et al. SCRIMP: Scalable communication for reinforcement-and imitation-learning-based multi-agent pathfinding[C]// 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2023: 9301-9308.

[16] SONG Z, ZHANG R, CHENG X. HELSA: Hierarchical Reinforcement Learning with Spatiotemporal Abstraction for Large-Scale Multi-Agent Path Finding[C]// 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS).

IEEE,2023:7318-7325.

[17] PHAN T,DRISCOLL J,ROMBERG J,et al. Confidence-Based Curriculum Learning for Multi-Agent Path Finding[J]. arXiv:2401.05860,2024.

[18] SKRYNNIK A,ANDREYCHUK A,NESTEROVA M,et al. Learn to Follow:Decentralized Lifelong Multi-Agent Pathfinding via Planning and Learning [C]// 2024 AAAI Conference on Artificial Intelligence. AAAI,2024:17541-17549.

[19] SHI D X,HU H M,SONG L N,et al. Multi-agent Reinforcement Learning Method Based on Observation Reconstruction [J]. Computer Science,2024,51(4):280-290.

[20] KIM D,MOON S,HOSTALLERO D,et al. Learning to schedule communication in multi-agent reinforcement learning[J]. arXiv:1902.01554. 2019.

[21] ZHANG S Q,ZHANG Q,LIN J. Efficient communication in multi-agent reinforcement learning via variance based control [C] // Advances in Neural Information Processing Systems (NeurIPS). MIT Press,2019:3230-3239.

[22] DING Z,HUANG T,LU Z. Learning individually inferred communication for multi-agent cooperation[C]// Advances in Neural Information Processing Systems(NeurIPS). MIT Press,2020:22069-22079.

[23] STERN R,STURTEVANT N,FELNER A,et al. Multi-agent pathfinding:Definitions,variants,and benchmarks[C] // Proceedings of the International Symposium on Combinatorial Search(SoCS). AAAI,2019:151-158.

[24] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]// Advances in Neural Information Processing Systems(NeurIPS). MIT Press,2017:5998-6008.

[25] TAN M. Multi-agent reinforcement learning:Independent vs. cooperative agents[C]// Proceedings of the Tenth International Conference on Machine Learning. ACM,1993:330-337.

[26] VAN HASSELT H,GUEZ A,SILVER D. Deep reinforcement learning with double q-learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI,2016:2094-2100.

[27] WANG Z,SCHAUL T,HESSEL M,et al. Dueling network architectures for deep reinforcement learning[C]// International Conference on Machine Learning. ACM,2016:1995-2003.

[28] HORGAN D,QUAN J,BUDDEN D,et al. Distributed prioritized experience replay[J]. arXiv:1803.00933,2018.

[29] BENGIO Y,LOURADOUR J,COLLOBERT R,et al. Curriculum learning[C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM,2009:41-48.

[30] LIN Q,MA H. SACHA:Soft actor-critic with heuristic-based attention for partially observable multi-agent path finding[J]. IEEE Robotics and Automation Letters,2023,8:5100-5107.

ZHANG Yongliang, born in 1977, Ph.D, associate professor. His main research interests include biometric features recognition, artificial intelligence and multirobot system.

(责任编辑:喻藜)