基于自适应探索与课程学习的多 AGV 路径规划算法

王鉴铸[1]，陶翼飞[1†]，田华亭[1,2]，许　容[3]，钱　傲[1]，佟雨擎[1]

(1. 昆明理工大学 机电工程学院，昆明 650500；2. 云南昆船智能装备有限公司， 昆明 650500；
3. 云南中烟红塔烟草 (集团) 有限责任公司昭通卷烟厂， 云南 昭通 657000)

摘　要: 针对多自动导引车路径规划研究中传统深度强化学习方法收敛速度缓慢、探索效率低、样本利用不充分等问题导致的路径规划成功率低且效果不佳, 提出一种基于自适应探索与课程学习的改进型多智能体深度确定性策略梯度算法 (AECL-MADDPG). 首先, 设计基于动态拥塞感知的自适应探索策略, 将双 Critic 的 Q 值差异作为决策不确定性度量, 与环境拥塞度联合驱动探索强度动态调整, 提高探索策略的动态适应性; 其次, 构建基于课程学习的优先经验回放机制 (PER), 将课程学习的“任务难度递进”与优先经验回放的“样本价值排序”深度融合, 通过课程权重实现跨难度等级样本的平滑过渡, 避免策略震荡; 再次, 设计多维度课程晋级与回调机制, 突破单一成功率阈值的粗糙晋级标准, 引入碰撞率和路径效率等多指标综合评估及性能回退保护, 提升训练稳定性; 最后, 在仓储环境中开展仿真实验, 并将所提算法与主流算法进行对比验证, 通过对比在收敛速度、成功率、平均路径长度等关键指标上的差异, 验证了所提出路径规划算法的可行性和有效性.

关键词: 自适应；课程学习；经验回放；路径规划；仓储环境；深度确定性策略梯度算法

中图分类号: TP18　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.1281

引用格式: 王鉴铸, 陶翼飞, 田华亭, 等. 基于自适应探索与课程学习的多 AGV 路径规划算法 [J]. 控制与决策.

Multi-AGV path planning algorithm based on adaptive exploration and curriculum learning

WANG Yin-zhu[1]，*TAO Yi-fei*[1†]，*TIAN Hua-ting*[1,2]，*XU Rong*[3]，*QIAN Ao*[1]，*TONG Yu-qing*[1]

(1. Faculty of Mechanical and Electrical Engineering，Kunming University of Science and Technology，Kunming 650500，China；2. Yunnan KSEC Intelligent EquipmentCo., Ltd.，Kunming 650500，China；3. Yunnan Hongta Group Zhaotong Cigarette Factory，China Tobacco Yunnan Industrial Co.，Ltd.，Yunnan Zhaotong 657000，China)

Abstract: To address the problems of low path planning success rate and suboptimal performance caused by slow convergence speed, low exploration efficiency, and insufficient sample utilization of traditional deep reinforcement learning methods in multi-automated guided vehicle path planning research, this paper proposes an improved multi-agent deep deterministic policy gradient algorithm based on adaptive exploration and curriculum learning (AECL-MADDPG). Firstly, an adaptive exploration strategy based on dynamic congestion awareness is designed. In this strategy, the Q-value discrepancy between dual-Critic networks is adopted as the metric for decision-making uncertainty, which jointly drives the dynamic adjustment of exploration intensity with the environmental congestion level, thus improving the dynamic adaptability of the exploration strategy. Secondly, a prioritized experience replay (PER) mechanism based on curriculum learning is constructed. This mechanism deeply integrates the“progressive task difficulty”of curriculum learning with the“sample value ranking”of prioritized experience replay, and realizes the smooth transition of samples across difficulty levels through curriculum weighting, so as to avoid policy oscillation. Thirdly, a multi-dimensional curriculum advancement and rollback mechanism is developed. It breaks through the limitation of the crude advancement criterion based on a single success-rate threshold, by introducing a comprehensive multi-metric evaluation system including collision rate and path efficiency, as well as performance rollback protection, thus significantly improving training stability. Finally, simulation experiments are conducted in a warehousing environment, and comparative validation between the proposed algorithm and mainstream baseline algorithms is performed. The comparison results of key indicators including convergence speed, task success rate, and average path

收稿日期: 2025-12-11；**录用日期:** 2026-04-05.
基金项目: 国家自然科学基金项目 (51165014).
责任编委: 蒲志强.
†通信作者. E-mail: 676379098@qq.com.

length fully verify the feasibility and effectiveness of the proposed path planning algorithm.

Keywords: adaptive； curriculum learning； experience replay； path planning； storage environment； deep deterministic policykey gradient

0 引 言

自动导引车 (AGV) 是一种智能运输设备, 广泛应用于物流、制造等行业, 是实现货物高效搬运的关键工具[1]. 智能制造过程中的物料派发、运输、仓储、装卸等环节的自动化和智能化也是连续生产的基础和保证[2]. AGV 作为其中的关键设备之一, 在货物运输和分拣过程中承担着重要作用[3], 凭借自动化程度高、路径调整灵活等优势, 已成为物料搬运的关键设备[4]. 如亚马逊 FBA 仓库采用的“货到人”多 AGV 调度模式, 订单分拣效率较传统模式提升 3 倍[5].

多 AGV 路径规划问题旨在为共享环境中的多台 AGV 规划从起点到目标点的无冲突路径集合[6], 本质是在动态约束环境中, 为每台 AGV 规划满足“安全-效率-协同”三维目标的路径[7]. 从安全性角度, 仓储环境中存在固定货架、临时障碍物及其他 AGV, 目标在于规避碰撞[8]; 从效率角度, 需最小化单台 AGV 的路径长度与任务完成时间, 同时最大化系统吞吐量[9]; 从协同性角度, 多 AGV 需在共享资源竞争中实现全局最优, 避免局部最优导致的系统整体效率下降.

当前多 AGV 路径规划传统方法中, A*算法、Dijkstra 算法等静态路径规划算法, 虽能快速求解单 AGV 最优路径, 但在解决大规模 AGV 的路径规划问题时效率和质量都不高, 难以应对多 AGV 动态交互场景[10], 张新艳等提出一种引入时间因子的改进 A*算法[11], 结合时间窗以及优先级策略实现多 AGV 的动态无碰撞路径规划; 时间窗方法通过为 AGV 分配资源占用时间片避免冲突, 但需提前规划且灵活性差, 难以适应货架动态移动的仓储环境[12], 王彬等[13] 在 A*算法的启发函数中加入父节点信息, 改进了传统动态窗口法的障碍物距离评价子函数, 并将两者成功结合. Petri 网、整数规划等建模方法, 可通过数学模型描述多 AGV 协同约束, 但随着 AGV 数量增加, 模型求解复杂度呈指数级增长, 实时性难以满足[14], 于绍琪等[15] 提出一种用 Petri 网对仓库环境中 AGV 系统进行建模的方法, 以有效解决 AGV 运输货物时产生冲突的问题. 随着物流系统的复杂化与动态化发展, 使得传统静态路径规划方法已无法适配实际应用需求[16].

随着深度强化学习的发展, 通过 AGV 和环境交互试错学习最优策略, 无需预先建立环境模型, 在动态环境中表现出较强的适应性[17]. H.Bae 等[18] 使用基于 DQN(Deep Q-Network) 的单 AGV 路径规划算法, 通过深度神经网络拟合 Q 值函数, 实现了动态避障; 针对传统 Q 学习的局限性, 王岩红等[19] 设计启发式奖励机制引导智能体决策, 同时引入余弦退火学习率与指数衰减探索率动态调节训练过程, 提升算法最终性能. 李佩哲[20] 等设计了一种可在不同的训练阶段分配不同经验池中的经验, 通过对经验做分类存储, 动态调整经验池的采样权重占比, 结合启发式专家经验引导, 让智能体在训练全周期都能高效利用最有价值的经验, 以改善网络的训练效率. 而基于多智能体强化学习的 (MADDPG) 算法通过“集中式训练、分布式执行”范式, 在训练阶段利用全局信息优化策略, 执行阶段仅依赖局部观测决策, 平衡了协同性与灵活性[21]. 但在高密度 AGV 场景中, MADDPG 仍存在三大问题, 一是探索效率低, 固定高斯噪声探索易导致 AGV 在拥堵区域反复碰撞, 学习收敛缓慢[22]; 二是样本利用效率差, 随机经验回放无法优先学习避撞、最优路径等关键经验, 训练数据量需求大; 三是任务难度适应弱, 直接在大规模 AGV 场景中训练, 易出现初期策略混乱, 难以生成有效动作等“冷启动”问题.

综上, 传统算法本质是基于预定义环境模型的静态优化, 面对动态变化时存在突发障碍应对被动, 算法依赖环境先验信息, 当出现未预定义的障碍时, 需重新遍历全局或局部空间重新规划路径, 导致规划延迟, 且动态目标适配性差; 强化学习大多采用固定高斯噪声探索易导致 Q 值过估计, 在经验回放方面, 经典的课程学习基于样本 TD 误差定义难度权重, 仅在样本层面进行难度排序, 未结合多 AGV 任务本身的复杂度递进, 且通过单维度指标实现任务难度递进, 无法解决 AGV 数量增加带来的环境状态空间爆炸问题; 而自适应探索和课程学习方案以强化学习的经验学习为核心, 结合自适应探索的“全局搜索能力”, 填补了传统算法在动态环境中的适配缺口, 既解决了传统算法的反应滞后问题, 又突破了局部最优陷阱, 同时满足多 AGV 协同的实时性与鲁棒性要求.

因此本文提出一种基于自适应探索与课程学习的改进型多智能体深度确定性策略梯度算法 (AECL-MADDPG), 旨在通过优化探索策略与经验

回放机制, 提升多 AGV 在仓储环境中的路径规划性能. 本文的主要内容如下:

1) 提出基于动态拥塞感知的自适应探索策略: 区别于 TD3 算法利用双 Critic 网络解决 Q 值过估计问题, 本文将双 Critic 网络的 Q 值差异作为决策不确定性的量化指标, 并与环境拥塞度融合构建自适应探索因子. 该策略实现了环境拥塞度高或决策不确定时增强探索, 环境畅通且决策确定时减少探索的智能调节, 使 AGV 能够根据实时环境状态动态调整行为模式, 在高密度区域主动寻找替代路径, 有效避免传统固定探索策略的盲目性.

2) 构建基于课程学习的优先经验回放机制: 区别于传统课程学习经验回放 (Curricular PER) 基于样本难度进行优先级排序, 本文将课程学习的任务复杂度递进与优先经验回放的样本价值排序深度融合, 通过设计课程权重函数 Wc 实现跨难度等级样本的高斯平滑采样, 结合事件类型权重 We 强化关键经验学习, 避免了传统方法因难度跳跃导致的策略震荡问题, 显著提升样本利用率与算法收敛稳定性.

3) 设计多维度课程晋级与回调机制: 突破现有课程学习方法以单一成功率阈值作为晋级标准的局限, 引入成功率、碰撞率、路径效率比等多维度综合评估指标, 并设计性能回退保护机制, 晋级后性能显著下降时自动回调至前一课程等级, 有效防止因过早晋级导致的过拟合问题.

4) 开展系统性实验验证与工业适配性分析: 搭建高逼真度仓储仿真环境, 模拟动态货架、多 AGV 高密度场景, 将 AECL-MADDPG 与 DDPG、MADDPG、MAPPO、QMIX 及考虑拥塞感知的强化学习 (CA-MARL) 等基线算法在收敛速度、任务成功率、路径规划效率等维度进行全面对比分析, 同时通过消融实验验证本文所提算法的独立有效性与协同效应.

1　问题描述

多 AGV 路径规划问题可以被定义为一个四元组$\langle G, K, P, T\rangle$其中, $G=(V,E)$是一个无向图, 无向图中的节点$v \in V$是 AGV 可以占据的位置, 边$e=(v_i, v_j) \in E$是v_i和v_j之间的连线, 表示 AGV 可以在v_i和v_j之间移动. K代表问题中的 AGV 数量, 每个 AGV 都有独一无二的起始位置$P_i \in P \in V$和目标位置$g_i \in T \in V$. P是所有 AGV 初始位置的集合, T是所有 AGV 目标位置的集合. 在此问题中, 时间被离散化为时间步的形式$t=\{0,1,2,\ldots\}$, 在栅格地图的路径规划过程中, AGV 可选择向上、向下、向左、向右移动或停留在当前位置的 5 种动作, 且无法进行斜向移动. 栅格图中, AGV 每移动一个栅格花费一个时间步, 最大移动速度 1 栅格/步, 转弯无延迟, 碰撞判定为栅格重叠.

AGV 从p_i移动到g_i的行动序列构成一条路径d_i. 可行解是k个 AGV 的k条路径的集合$D=\{d_1, d_2, \ldots, d_k\}$, 其中, AGV 对应于路径$d_i$, 路径集合中的任 2 条路径$d_i$和$d_j$之间不存在冲突.

2　算法描述

多智能体深度确定性策略梯度算法 (MADDPG) 是一种采用中心化训练去中心化执行 (Centralized Training with Decentralized Execution,CTDE) 范式的多智能体强化学习算法. 在训练阶段, 所有智能体共享信息, 通过中心化的训练方式指导智能体的学习过程以最大化全局性能, 在执行阶段, 各个智能体根据自身的感知和历史经验独立决策. 本文针对多 AGV 路径规划挑战提出核心设计改进策略—基于动态拥塞感知的自适应探索策略与基于课程学习的优先经验回放机制.

2.1　算法框架

在 AECL-MADDPG 算法流程中, 如图 1 所示, 训练阶段每个 AGV 可利用所有 AGV 的全局信息指导学习; 执行阶段每个 AGV 仅依据自身局部观测进行决策, 该机制显著提升了算法稳定性与多 AGV 协同能力.

AECL-MADDPG 为每个 AGV 维护一个确定性策略网络 (Actor) 与一个动作价值网络 (Critic), 网络参数更新原理继承自 DDPG, 具体如下:

1) Critic 网络更新

Critic 网络通过最小化时序差分 (TD) 误差更新, 损失函数定义为均方贝尔曼误差 (MSBE):

$$L(\omega_i)=\mathrm{E}_{\mathrm{x},a,r,\mathrm{x}'\sim\mathcal{D}}[(y_i-Q_{\omega_i}(\mathrm{x},a_1,\ldots,a_N))^2]. \quad (1)$$

其中, ω_i为第i台 AGV 的 Critic 网络的权重参数, r为 AGV 在当前状态执行动作后获得的即时奖励, E为对经验回放池 D 中采样的样本求数学期望, $\mathrm{x}=(s_1,\ldots,s_N)$表示所有 AGV 的联合状态 目标值$y_i$的计算式为:

$$y_i=r_i+\gamma Q_{\omega_{i'}}(\mathrm{x}',a_{1'},\ldots,a_{N'})\big|_{a_{j'}=\mu_{\theta_{j'}}(s_{j'})}. \quad (2)$$

式中, $Q_{\omega_{i'}}$与$\mu_{\theta_{j'}}$分别为 Critic 与 Actor 的目标网络, γ为折扣因子. $a_{j'}$为下一时刻第j台 AGV 的动作, 由目标 Actor 网络输出, $s_{j'}$是第j个 AGV 在下一时刻的局部观测状态. Critic 网络输入为全局信息, 可实

现对环境状态的全面理解.

2) Actor 网络更新

Actor 网络通过策略梯度更新, 优化目标为最大化 Critic 网络对其输出动作的评价值, 策略梯度$\nabla_{\theta_i}J$的计算公式为:

$$\nabla_{\theta_i}J(\mu_i)=\mathrm{E}_{\mathbf{x}\sim\mathcal{D}}[\nabla_{\theta_i}\mu_{\theta_i}(s_i)\nabla_{a_i}Q_{\omega_i}(\mathbf{x},a_1,\ldots,a_N)|_{a_i=\mu_{\theta_i}(s_i)}]. \quad (3)$$

其中, $\mu_{\theta_i}(s_i)$为第i台 AGV 的主 Actor 网络, 输入局部观测状态s_i, 输出确定性基础动作, $a_i=\mu_{\theta_i}(s_i)$表示梯度计算限定在 Actor 网络当前输出的动作处, 确保梯度的有效性, 通过沿 Q 值增大方向更新策略网络参数θ_i, Actor 网络可学习选择能带来更高长期回报的动作.

在 AECL-MADDPG 算法中, 单个 AGV 在训练阶段利用的全局信息包括以下三类:

1) 全局状态信息$\mathrm{x}=(s_1,s_2,\ldots,s_N)$: 包含所有$N$个 AGV 的局部观测向量拼接, 其中每个$s_j$包含 AGV 的归一化坐标$(x_j,y_j)$、目标点相对向量、模拟激光雷达观测及任务进度, 该信息使 Critic 网络能够理解全局环境态势.

2) 全局动作信息$(a_1,a_2,\ldots,a_N)$: 所有 AGV 在当前时刻的动作选择, 用于 Critic 网络评估联合动作的价值$Q_{\omega_i}(\mathrm{x},a_1,\ldots,a_N)$.

3) 其他 AGV 的位置与运动趋势: 通过全局状态中其他 AGV 的坐标信息及前后帧坐标差分, 隐式获取其他 AGV 的运动方向与速度, 用于预测潜在冲突区域.

2.2 自适应探索策略逻辑

针对传统固定高斯噪声探索策略无法适应环境动态变化, 在 AGV 高密度区域易导致拥塞或局部最优的问题, 本文提出自适应探索策略. 该策略中, AGV 最终执行动作a_i由策略网络输出的基础动作与自适应噪声项构成, 而非固定噪声决定, 具体表达式为:

$$a_i=\mu_{\theta_i}(s_i)+\epsilon_{\text{adaptive}}\cdot\mathrm{N}(0,\sigma^2). \quad (4)$$

其中, $\mu_{\theta_i}(s_i)$为 Actor 网络输出的确定性动作, $\mathrm{N}(0,\sigma^2)$为标准高斯噪声, $\epsilon_{\text{adaptive}}$为动态探索因子, 其值由三部分相乘得到, 全面反映探索需求度:

$$\epsilon_{\text{adaptive}}=\epsilon_{\text{base}}\cdot(1+\lambda_C\cdot C(L_i))\cdot(1+\lambda_U\cdot U(s_i)). \quad (5)$$

针对自适应探索策略的设计如下:

1) 基础探索率ϵ_{base}: 提供随训练进程逐步衰减的基准探索强度, 确保训练初期高探索度以探索环境, 后期 逐步倾向于利用已有策略. 实验中设置初始值为 0.7, 每步衰减系数为 0.9999, 最低降至 0.05, 该参数设置参考了多智能体强化学习中探索率衰减的常见策略.

2) 拥塞感知项: 根据环境密度动态调节探索强度. 其中, 拥塞度$C(L_i)$量化 AGVi当前或目标位置L_i周围的繁忙程度, 定义为半径R范围内其他 AGV 的数量, 计算式为:

$$C(L_i)=\sum_{j\neq i}^{N}\mathbb{I}(\parallel L_i-L_j\parallel\leqslant R). \quad (6)$$

$\mathbb{I}(\cdot)$为指示函数, 当$\parallel L_i-L_j\parallel\leqslant R$表示两台 AGV 之间的欧氏距离小于等于拥塞检测$R$, λ_C为拥塞感知权重系数 (实验中取值 0.6~0.8), 拥塞度越高, 该项值越大, 探索因子随之提升, 鼓励 AGV 寻找替代路径.

3) 不确定性感知项: 衡量 AGV 在状态s_i下决策的“把握程度”. 借鉴 TD3 算法处理值函数过高估计的思路, 引入两个独立 Critic 网络Q_{ω_1}与Q_{ω_2}, 当两网络对同一状态-动作对的价值评估差异较大时, 表明决策不确定性高, 需增强探索. 不确定性$U(s_i)$定义为两 Critic 网络输出 Q 值的标准差:

$$U(s_i)=\mathrm{Std}(Q_{\omega_1}(\mathrm{x},a),Q_{\omega_2}(\mathrm{x},a)). \quad (7)$$

其中Std为标准差函数, 用来衡量两个数值的离散程度, $Q_{\omega_1}(\mathrm{x},a)$和$Q_{\omega_2}(\mathrm{x},a)$为两个结构完全独立、参数不同的 Critic 网络 (双 Critic 网络), 对同一全局状态 x、联合动作a输出的 Q 值评估结果, λ_U为不确定性感知权重系数 (实验中取值 0.4~0.6), 该设置平衡了不确定性感知与拥塞感知的影响, 避免单一因素主导探索策略.

2.3 基于课程学习的优先经验回放机制

为解决标准经验回放 (ER) 随机采样导致的样本利用率低, 及复杂仓储任务中从零开始训练因初始阶段难以获得正向反馈而引发的学习停滞问题, 构建融合课程学习与优先经验回放的 CL-PER 机制, 设置课程等级划分和晋级规则, 再进行优先级计算, 通过从易到难的训练范式与聚焦关键的采样策略, 系统性提升训练效率.

2.3.1 课程等级划分与晋级规则

将整个训练过程分解为难度递增的系列课程, 设课程难度等级$\mathrm{c}_{\text{level}}\in\{1,2,\ldots,\mathrm{C}_{\max}\}$, 每个等级对应不同 AGV 数量. 区别于传统课程学习仅以单一成功率作为晋级标准 (易导致过拟合), 本文设计多维度综合评估指标, 晋级条件为:

$$c_{\text{level}}\leftarrow c_{\text{level}}+1 \text{ if } \phi(c_{\text{level}})>\phi_{\text{threshold}}. \quad (8)$$

其中, $\phi_{\text{threshold}}$为晋级阈值, $\leftarrow$表示赋值操作, 即满足

条件时, 课程等级+1, 综合评估分数 ϕ 由三项指标加权构成:

$$\phi(c_{\text{level}}) = \alpha_s \cdot \overline{S} + \alpha_c \cdot (1 - \overline{C}) + \alpha_p \cdot \overline{P}. \quad (9)$$

其中, 任务成功率$\overline{S}$: 最近 100 个回合中所有 AGV 完成任务的回合占比, 权重$\alpha_s = 0.5$;

低碰撞率$(1-\overline{C})$: $\overline{C}$为平均碰撞次数归一化值(除以最大允许碰撞次数 5), 权重$\alpha_c = 0.3$;

路径效率比$\overline{P}$: 实际路径长度与理论最短路径(曼哈顿距离) 的比值取倒数, 即$\overline{P} = L_{\text{optimal}}/L_{\text{actual}}$, 权重$\alpha_p = 0.2$.

当 AGV 在当前等级c_{level}下的平均任务成功率$\overline{S}$超过预设阈值$S_{\text{threshold}}$(实验中设为 0.75) 时, 即综合评估分数需达到75%以上才能晋级.

考虑到是否会出现无兜底的无限无效训练, 首先 3 项指标加权从规则根源上避免了单一指标卡死全局的死循环风险, 其次后续实验明确设置了训练总回合数的全局终止条件, 训练过程中基础探索率会持续衰减至最低, 后期模型会从“盲目探索”转向“策略固化”, 性能会快速收敛稳定, 不会一直无意义波动训练, 最后, 过程中会持续给碰撞、成功、避堵等关键经验赋予高采样权重, 强制模型优先学习“能提升性能的核心经验”, 每一轮训练都会让避障能力、路径效率稳步提升, 会持续向 75% 的阈值逼近.

传统课程学习仅以单一成功率作为晋级标准, 极易出现在低等级仅学到过拟合策略, 晋级到更高难度场景后性能断崖式下跌, 或者过早晋级导致出现策略混乱、碰撞率飙升、任务成功率大幅下滑, 最终训练完全失效, 因此引入性能回退保护机制:

$$c_{\text{level}} \leftarrow c_{\text{level}} - 1 \text{ if } \phi_{\text{new}}(c_{\text{level}}) < \phi_{\text{threshold}} - \Delta_{\text{protect}} \text{且连续} K_{\text{fail}} \text{回合}. \quad (10)$$

其中, $\phi_{\text{new}}(c_{\text{level}})$为晋级后当前课程等级下的实时综合评估分数, Δ_{protect}=0.15 为保护阈值, K_{fail}=50 为连续失败回合数阈值. 当晋级后综合评估分数连续 50 回合低于 60% 时, 自动回调至前一课程等级, 重新夯实基础策略, 避免训练崩溃, 保障全程训练稳定性.

课程学习的晋级逻辑如图 2 所示:

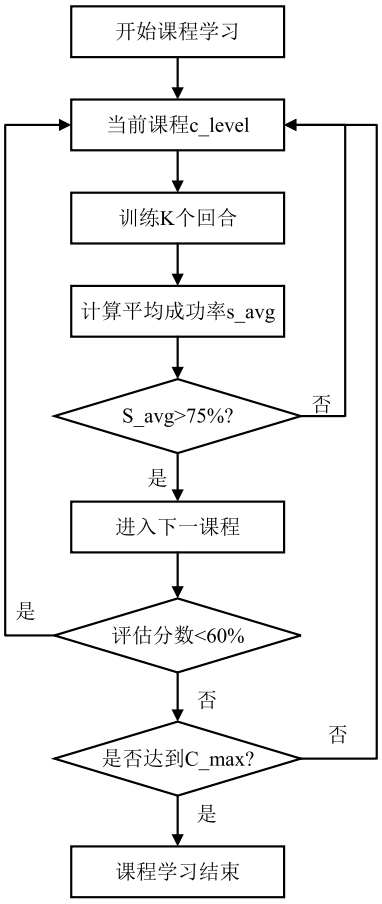

图2　晋级逻辑

2.3.2　综合优先级计算

在课程学习框架下, 对传统优先经验回放进行增强, 为经验池中的每个样本计算综合优先级, 由三部分加权构成: 为经验池中的每个样本j计算综合优先级P_j, 由三部分加权构成:

$$88P_j = (|\delta_j| + \epsilon)^{\alpha} \cdot w_c(c_{\text{level},j}) \cdot w_e(e_{\text{type},j}) \quad (11)$$

各部分含义与计算逻辑如下:

1) 基础优先级: 沿用传统优先经验回放核心思想, 基于 TD 误差δ_j衡量样本“意外程度”, δ_j越大, 样本学习价值越高. δ_j计算式为:

$$\delta_j = r_j + \gamma \min_{k=1,2} Q_{\omega_{k'}}(\mathbf{x}_{j'}, a_{j'}) - Q_{\omega_k}(\mathbf{x}_j, a_j). \quad (12)$$

其中, $\min_{k=1,2} Q_{\omega_{k'}}(\mathbf{x}_{j'}, a_{j'})$为两个目标 Critic 网络对下一状态-动作对价值评估的最小值, $\epsilon = 10^{-6}$为避免优先级为 0 的微小常数, $\alpha = 0.5$为优先级调整因子, 平衡优先级的贫富差距, 避免部分高优先级样本被过度采样.

2) 课程权重: 引导算法优先采样与当前训练课程难度相关的经验, 确保学习过程的“渐进性”. 定义为样本难度等级$c_{\text{level},j}$与当前训练难度等级c_{current}差异的高斯函数:

$$w_c(c_{\text{level},j}) = exp(-\frac{(c_{\text{level},j} - c_{\text{current}})^2}{2\sigma_c^2}). \quad (13)$$

$\exp(\cdot)$为自然指数函数, σ_c为高斯函数标准差(实验中取值 0.5), 当样本难度等级与当前训练等级一致时, w_c取最大值 1; 当难度差异为 1 级时, w_c约为 0.606; 当差异≥2 级时, $w_c \leqslant 0.135$, 大幅降低低相关样本的采样概率.

3) 事件权重: 为对策略优化至关重要的特殊事件赋予更高采样优先级, 强制模型聚焦关键经验. 定义为分段函数:

$$w_e(e_{\text{type},j})=\begin{cases}1.8 & \text{if } e_{\text{type},j}=\text{success}\\ 2.0 & \text{if } e_{\text{type},j}=\text{collision}\\ 1.0 & \text{otherwise}\end{cases}\tag{14}$$

其中, 碰撞事件collision权重最高 (2.0), 因碰撞样本包含 AGV 避障失败的关键信息, 需重点学习以减少后续冲突; 任务成功事件success权重次之 (1.8), 用于强化最优路径决策; 常规行驶事件otherwise权重为1.0, 确保基础经验的正常学习.

2.4 自适应探索与课程学习结合机制

所提出 AECL-MADDPG 算法中, 自适应探索策略 (AE) 与课程学习 (CL) 并非独立功能模块的简单叠加, 而是形成了课程难度正向引导探索强度、探索效果反向驱动课程迭代、关键经验跨课程协同优化的全闭环耦合机制. 二者结合的底层逻辑是课程学习为自适应探索提供难度适配的训练环境与学习节奏, 避免高复杂度场景下的盲目探索与策略崩溃; 自适应探索为课程学习提供高质量的交互样本与性能评估依据, 加速课程内的策略收敛, 同时通过执行效果反向驱动课程等级的平滑晋级与回调, 最终实现探索效率、学习稳定性、策略性能的协同提升. 步骤如下:

1) 初始化阶段: 课程体系与探索策略基线绑定;

2) 执行阶段: 课程约束下的自适应探索动态执行. AGV 基于当前课程的检测半径计算周边拥塞度, 结合双 Critic 网络输出的 Q 值差异计算决策不确定性, 融合课程基线参数得到动态探索因子, 生成最终执行动作. 动作与环境交互后, 获取复合奖励, 同时对本次交互的样本进行双维度标记存入经验池, 完成从探索执行到课程化样本存储的全链路绑定;

3) 优化阶段: 探索经验的课程化采样与网络协同更新, 本阶段实现两大模块的双向协同优化, 将探索经验转化为策略能力提升;

4) 迭代阶段: 探索执行效果反向驱动课程动态调整;

5) 收敛阶段: 双模块协同收敛与策略输出. 当训练总回合数达标, 或课程达到最高级且综合性能连续稳定时, 训练终止: 此时自适应探索的基础探索率衰减至最小值, 课程学习完成从易到难的全流程训练, 二者协同收敛, 最终输出 AGV 的分布式路径规划策略网络.

2.5 状态、动作空间与奖励函数设计

1) 状态空间: 每个 AGV 的局部观测状态$s_i\in \mathrm{R}^{26}$为 26 维向量, 具体构成如下:

1. 自身归一化坐标 (2 维)

$$(x_{\text{norm}},y_{\text{norm}})=(\frac{x_i}{W-1},\frac{y_i}{H-1}).\tag{15}$$

其中$W=H=30$, 归一化后取值 [0,1];

2. 目标点相对向量 (2 维):

距离归一化: $d_{\text{norm}}=d_{\text{goal}}/d_{\max}$, $d_{\max}=\sqrt{W^2+H^2}\approx 42.4$为对角线长度;

角度归一化: $\theta_{\text{norm}}=\theta_{\text{goal}}/\pi$, $\theta_{\text{goal}}\in[-\pi,\pi]$为目标相对方位角;

3. 任务进度信息 (2 维): 已行驶距离比$p_{\text{dist}}=L_{\text{travelled}}/L_{\text{total}}$, 剩余时间比$p_{\text{time}}=(T-t)/T$;

4. 扩展调度信息 (4 维): 任务优先级$\rho_i\in[0,1]$、载重状态$w_i\in\{0,1\}$、电量剩余比$e_i\in[0,1]$等待时间累计$T_{\text{wait},i}\in[0,1]$.

5. 模拟激光雷达对环境感知, 观测分为障碍物检测 (8 维) 和其他 AGV 检测 (8 维), 障碍物检测向 8 个均匀分布方向检测, 检测距离 0 ~ 3 栅格. 每方向返回连续距离值 $d_{\text{obs},k}=\min(d_{\text{hit},k},3)/3$, 仅查询静态障碍物地图 M_{obs}; 其他 AGV 检测独立于障碍物通道, 同样 8 方向检测, 仅查询动态 AGV 位置地图 M_{agv}, 返回 $d_{\text{agv},k}\in[0,1]$; AGV 与障碍物区分采取环境维护两张独立占用地图, 障碍物检测查询 M_{obs}, AGV 检测查询 M_{agv}, AGV 可据此采取差异化避让策略;

2) 动作空间与运动约束: 采用离散动作空间 A = {0,1, 2,3, 4}对应{上、下、左、右、停留}. 考虑 AGV 运动学约束:

1. 速度约束: 空载 $v_{\max}=1$栅格/步, 满载 $v_{\max}=0.8$栅格/步;

2. 加速度约束: $\Delta v_{\max}=0.5$栅格/步^2;

3. 转向约束: 方向变化$>90°$时插入 1 步停留.

3) 奖励函数: 设计复合奖励函数引导 AGV 高效学习:

$$r=r_{\text{goal}}+r_{\text{collision}}+r_{\text{step}}+r_{\text{shape}}.\tag{16}$$

r_{goal}: AGV 到达终点时奖励+20, 所有 AGV 完成任务额外奖励+50;

$r_{\text{collision}}$: 与障碍物或其他 AGV 碰撞惩罚−10, 碰撞后停留 1 步;

r_{step}: 每步基础惩罚−0.01, 避免 AGV 原地徘徊;

r_{shape}: 每步移动后与目标点距离缩短时, 奖励+0.1×缩短距离; 距离增加时, 惩罚−0.1×增加距离 (距离单位为栅格).

2.6 与现有方法的对比分析

本节将 AECL-MADDPG 与 TD3、Curricular PER、CA-MARL 等方法进行对比分析, 如表 1 所示.

表1　各算法对比

对比算法	探索策略	样本优先级	课程设计	不确定性估计	适用场景
TD3	固定高斯噪声+目标策略平滑	无	无	双Critic用于价值下界估计	连续控制单AGV
Curricular PER	固定噪声	TD误差+课程难度	样本级难度递进	无	加速单AGV学习
CA-MARL	拥塞感知调节 (仅环境因素)	TD误差	无	无	多AGV避免拥塞
AECL-MADDPG (本文)	双因子自适应(环境 拥塞+决策不确定性)	TD误差+事件类型+ 课程等级(三维融合)	任务级难度递进+ 多维晋级评估+回退保护	双Critic用于 探索强度调节	多AGV协作 规划

1) 与 TD3 的区别: TD3 算法使用双 Critic 网络估计价值下界以解决过估计问题, 本文将双 Critic 的 Q 值方差作为决策不确定性度量, 驱动自适应探索. 这是双 Critic 架构的全新应用方式;

2) 与 Curricular PER 的区别: Curricular PER 在样本级进行难度调节, 本文在此基础上引入任务级课程学习, 并设计多维度晋级评估与回退保护机制, 形成"任务-样本"双层课程体系;

3) 与 CA-MARL 的区别: CA-MARL 基于环境拥塞度调节行为, 本文进一步融合 AGV 决策不确定性, 实现"环境感知+自我认知"的双因子自适应探索, 在策略未成熟时增强探索以避免局部最优.

2.7　算法伪代码

本文算法在 MADDPG“集中式训练, 分布式执行”框架基础上, 集成自适应探索模块与 CL-PER 模块, 形成完整优化流程.

算法1　ACEL-MADDPG算法求解路径规划

输入 AGV运行环境地图, 训练总回合数M=5 000, 每回合最大步数T=200

初始化段

1. 初始化课程等级c_{level}= 1, 最大课程等级C_{max}= 3

2. 初始化AGV数量N = 2~3(对应c_{level}=1)

3. 为每台AGV初始化:

3.1　Actor网络μ_{θ_i}(输入: 观测s_i, 输出: 动作a_i)

3.2　双Critic网络Q_{ω_1}、Q_{ω_2}(输入: (s_i, a_i), 输出: 价Q)

3.3　目标网络μ_{θ_i}'、Q_{ω_1}'、Q_{ω_2}'(参数初始同步主网络)

4. 初始化CL-PER经验池D(容量=5×10^4, 存储元组: (s_i, a_i, r_i, $s_{i'}$, c_{level},e_{type})

5. 设置超参数: γ=0.98(折扣因子), T=0.005(软更新系数), batch_size=256(批次大小), ϵ_{base}=0.7(初始探索率)

主训练循环(每回合)

6. for episode = 1 to M do

7.　重置环境Env: 生成30×30栅格地图(20%障碍物), 随机分配AGV起点/终点

8　获取初始观测集S = {s_1, s_2, ..., s_N}

9.　total_reward = 0,success_flag = False

单回合内步数循环(执行阶段)

10.　for step = 1 to T do

11.　　AGV分布式决策: 生成动作

12.　　动作集A = {}

13.　　for each AGV_i in 1...N do

14.　　　计算拥塞度C与决策不确定性U

15. 输出最终策略

16. return {u_{θ_1}, u_{θ_2}, ..., u_{θ_N}}

3　实验与结果分析

3.1　实验环境与设置

3.1.1　仿真环境搭建

基于 Python 3.8 搭建实验平台, 深度学习框架采用 PyTorch1.12, 仿真环境基于 OpenAI Gym 扩展开发, 模拟智能仓储场景, 图 3 为基础、进阶、高级三个等级的课程场景的仿真环境示意图, 障碍物由静态货架和动态货架共同组成.

地图规格: 30×30 二维栅格, 栅格尺寸对应实际仓储中 0.5 m×0.5 m 区域, 实际覆盖面积 15 m×15 m;

障碍物设置: 随机生成占总面积 15%~25% 的静态障碍物 (模拟货架), 障碍物分布采用聚类生成策略以模拟真实货架排列, 边界不可穿越;

AGV 参数: 最大移动速度 1 栅格/步 (对应实际速度 0.5m/s), 满载时降速至 0.8 栅格/步, 转向延迟 1 步 (方向变化>90°时), 碰撞判定为栅格重叠;

动态元素: 高级课程 (3 级) 中加入动态货架, 每 50 步随机移动 1 个栅格 (非障碍物区域) 模拟"货到人"仓储模式.

测试场景复杂度量化: 为系统评估算法在不同复杂度场景下的性能, 定义场景复杂度指数κ:

$$\kappa = \frac{N_{\mathrm{AGV}}}{N_{\max}} \times (1+\rho_{\mathrm{obs}}) \times (1+\lambda_{\mathrm{dyn}} \cdot f_{\mathrm{dyn}}). \quad (17)$$

其中N_{AGV}为AGV数量, $N_{\max}=20$为归一化基准, ρ_{obs}为障碍物密度, $\lambda_{\mathrm{dyn}}=0.5$为动态因子权重, $f_{\mathrm{dyn}} \in \{0,0.5,1\}$为动态元素级别 (无/少量/完整).各课程等级的场景配置与复杂度指数见表 2

3.1.2　超参数设置与调优依据

为保障 AECL-MADDPG 算法在路径规划任务中的训练稳定性、收敛效率与路径规划性能, 本文针

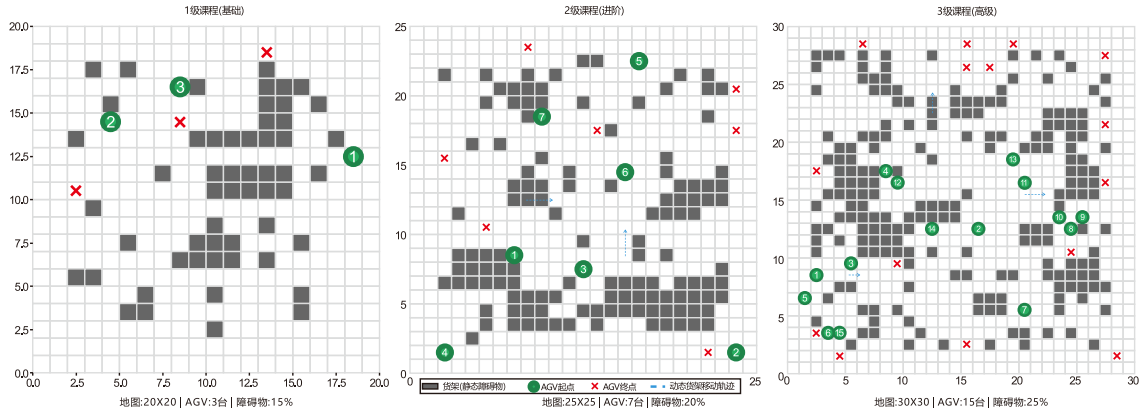

图3 智能仓储仿真环境示意图

表2 场景配置

场景等级	AGV数量	障碍密度	动态元素	通道宽度	复杂度κ
1级	2~3	15%	无	3栅格	0.17~0.26
2级	5~8	20%	少量	2栅格	0.45~0.72
3级	10~15	25%	完整	2栅格	0.94~1.41

对算法核心模块的超参数进行系统性设置与针对性调优, 所有超参数取值均结合仓储场景实际应用需求, 并通过网格搜索、敏感性分析等方法验证最优取值, 同时兼顾参数的普适性与场景适配性, 确保算法在不同复杂度的仓储路径规划场景中均能保持良好表现, 具体设置和依据见表 3.

3.1.3 对比算法与评估指标

选取 6 种主流多智能体强化学习算法作为基线, 涵盖不同技术路线:

1) DDPG: 单智能体深度确定性策略梯度, 多 AGV 场景采用“独立学习者”模式 (各 AGV 独立训练);

2) MADDPG: 标准多智能体深度确定性策略梯度, 采用固定高斯噪声 (ε=0.1) 探索;

3) MADDPG+PER: 标准 MADDPG 集成传统优先经验回放 (仅基于 TD 误差排序).

4) MAPPO: 多智能体近端策略优化算法,采用集中式价值函数与分布式策略,代表当前主流 on-policy 方法;

5) QMIX: 基于值分解的多智能体 Q 学习, 通过单调性约束实现集中式训练分布式执行, 适用于协作场景;

6) CA-MARL: 考虑拥塞感知的多智能体强化学习方法, 在奖励函数中显式建模拥塞惩罚, 但探索策略固定;

所有对比算法均采用相同网络结构 (3 层全连接, 隐藏层 256 单元) 与训练超参数 (学习率、批次大小等), 确保对比公平性.

表3 CL-MADDPG 算法超参数设置

超参数	符号	取值	设置依据
折扣因子	γ	0.98	参考MADDPG原文, 平衡即时与长期回报
软更新系数	τ	0.005	经验值, 确保目标网络平滑更新
批次大小	batch_size	256	网格搜索[64,128,256,512], 256时收敛最稳定
经验池容量	CD	5×10^4	约100回合数据量, 平衡内存与样本多样性
初始探索率	εbase	0.7	确保初期充分探索, 高于DDPG默认值0.1
探索率衰减	decay_rate	0.9999	约5 000回合衰减至0.05, 与训练总回合数匹配
最小探索率	εmin	0.05	保留少量探索以应对环境变化
拥塞感知权重	λC	0.7	敏感性分析[0.4,0.6,0.8,1.0], 0.7时碰撞率最低
不确定性权重	λU	0.5	敏感性分析[0.3,0.5,0.7], 0.5时收敛最快
拥塞检测半径	R	3栅格	对应实际1.5m, 覆盖AGV两倍安全距离
优先级因子	α	0.5	PER原文推荐值, 平衡优先级差异
重要性采样因子	β	0.4→1.0	线性退火, 初期偏差校正弱, 后期完全校正
课程权重标准差	σc	0.5	确保相邻课程样本权重差异适中
晋级阈值	Φthreshold	0.75	综合评估分数需达75%, 避免过早晋级
回退保护阈值	Δprotect	0.15	晋级后分数低于60%触发回退
连续失败回合数	Kfail	50	约10%训练回合, 避免噪声触发误回退

从收敛性、成功率、路径效率、安全性维度设置 5 项核心指标:

1) 平均奖励: 每回合所有 AGV 奖励的平均值, 反映策略整体优劣;

2) 任务成功率:规定步数 (200 步) 内所有 AGV 完成任务的回合占比,反映鲁棒性;

3) 平均路径长度: 每回合成功完成任务的 AGV 路径长度平均值 (单位: 栅格), 反映路径效率;

4) 平均碰撞次数: 每回合 AGV 间或 AGV 与障碍物的碰撞总次数, 反映安全性.

5) 路径效率比: $\eta = L_{\text{optimal}}/L_{\text{actual}}$, 理论最短路

径与实际路径的比值, 值越接近 1 表示路径越优.

3.2　收敛性对比分析

如图 4 示, 选取 4 种算法, 进行 5000 个回合训练, 观察平均奖励变化曲线 (采用 50 回合滑动窗口平滑).

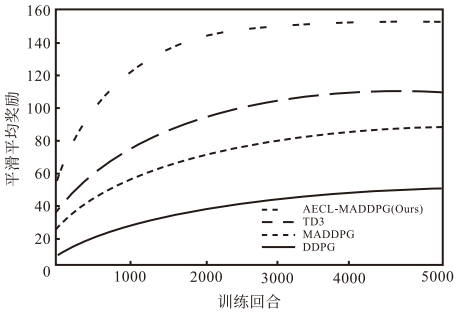

图4　平均奖励变化曲线

从图 4 见: AECL-MADDPG 收敛速度最快, 训练至 1500 回合时平均奖励达到 125, 2000 回合后稳定在 132 左右; MADDPG+PER 收敛速度次之, 3000 回合后稳定在 108; 标准 MADDPG 收敛较慢, 4000 回合后稳定在 86; DDPG 表现最差, 5000 回合仍未完全收敛, 稳定阶段平均奖励仅为 54.可以看出传统经验回放存在样本浪费问题, MADDPG 的随机采样使碰撞、成功等关键样本 (占比仅 15%) 与常规样本 (占比 85%) 被同等对待, 模型需遍历大量无效样本才能学到有效策略, 导致收敛延迟; 而课程学习的渐进式学习具有独特优势, 高斯平滑机制避免了难度跳跃导致的策略震荡, 1 级课程先掌握少 AGV+静态环境的基础避碰, 2 级课程加入动态货架, 3 级课程提升 AGV 密度, 每阶段的策略都能基于前一阶段的经验迭代, 减少从零开始的学习成本.

3.3　规划性能对比分析

在最高难度课程 (3 级,15 台 AGV) 下, 对 7 种算法的训练成熟模型进行 1000 次独立测试, 统计关键性能指标, 结果如表 4 所示. 所有测试保持环境参数一致, 动态货架每 50 步随机移动, AGV 起点和终点随机生成, 每回合最大步数 200.

表4　各算法性能对比

算法	任务成功率	平均路径长度(栅格)	平均碰撞次数(次/回合)	平均完成时间(步)	路径效率比
DDPG	61.32.5%	45.83.2	3.120.45	148.512.3	0.68
MADDPG	85.71.8%	41.22.7	1.050.21	121.89.7	0.76
MADDPG+PER	91.21.5%	39.52.1	0.880.17	105.38.2	0.79
MAPPO	88.41.6%	40.32.4	0.950.19	112.68.9	0.77
QMIX	86.91.7%	42.12.8	1.020.20	118.210.1	0.74
TD3	88.51.3%	39.72.0	0.920.15	104.47.6	0.79
AECL-MADDPG(本文)	99.10.8%	36.41.9	0.120.09	89.66.5	0.86

在最高难度课程下从表 4 据可得出以下结论:

1) 任务成功率: AECL-MADDPG 达到 99.1%, 较 MADDPG 提升 13.4 个百分点, 较 MADDPG+PER 提升 7.9%, 较 MAPPO 提升 10.7%, 较 QMIX 提升 12.2%, 表明其在高密度动态环境中具有极强的鲁棒性, 几乎可完全避免因碰撞、路径拥堵导致的任务失败.

2) 平均路径长度: 每回合中, AECL-MADDPG 最短为 36.4 栅格, 较 MADDPG 缩短 11.6%, 较 MADDPG+PER 缩短 7.8%, 这是因为 MADDPG 和 TD3 的固定高斯噪声无法区分“环境复杂度”, 在拥堵区域仍维持低探索强度, 导致 AGV“硬闯”冲突区域, 碰撞后被迫后退重试, 而自适应探索的主动预判机制使 AGV 在进入冲突区域前即触发强探索, 在拥堵区域通过提前绕行, 动态货架移动后通过探索新路径, 从根源上避免碰撞到重试的低效循环;

3) 平均碰撞次数: 每回合中, AECL-MADDPG 仅为 0.12 次/回合,较 MADDPG 降低 88.6%, 较 MADDPG+PER 降低 86.4%;这是因为动态拥塞感知能实时避开高密集区域,减少 AGV 间冲突.

上述性能优势的原因在于 AECL-MADDPG 的自适应探索策略使 AGV 能动态调整行为, 拥堵区域主动绕行, 而 CL-PER 机制确保模型优先学习避撞、最优路径等关键经验, 二者协同优化了“决策效率”与“策略安全性”. 确保在各项核心性能指标中处于领先.

3.4　主动绕行机制分析

绕行通常会增加路径长度,但 AECL-MADDPG 的平均路径长度最短,主要是因为:

1) 传统"被动碰撞-后退-重试"模式的路径损耗: 在 MADDPG 等算法中, AGV 采用固定探索策略进入拥塞区域后发生碰撞, 典型行为模式为 AGV 沿最短路径直行进入拥塞区域, 与其他 AGV 发生碰撞, 触发碰撞惩罚, 被迫停留 1 步; 后退 1~2 栅格寻找替代路径, 产生额外路径长度; 重新尝试通行, 可能再次碰撞. 实验统计显示,MADDPG 每次碰撞导致平

均路径增量约 4.3 栅格 (含后退、等待、绕行).

2)AECL-MADDPG"主动预判-提前绕行"模式: 本文算法通过拥塞感知项 C(L_i) 与不确定性感知项 U(s_i), 在 AGV 进入拥塞区域前即触发强探索, 当 C(L_i)> =2(即 3 栅格内存在> =2 台其他 AGV) 时, 探索因子增大 40%~60%; AGV 在距离拥塞区域 2-3 栅格时即开始绕行, 绕行距离约 2-3 栅格; 由于提前绕行, 避免了碰撞后的后退与重试.

3.5 拓展性分析

为验证算法的可扩展性, 本文在不同 AGV 数量的场景下进行了对比实验, 如表 5 所示.

表5 同 AGV 数量获得奖励

AGV数量	10	15	25	30	40
MADDPG	6.7	8.6	10.1	11.2	10.7
TD3	8.6	10.5	11.4	12.3	11.6
AECL-MADDPG	10.2	13.2	16.3	19.2	20.7

在核心测试的 15 台 AGV 场景下, 算法性能已较基线有明显优势; 当 AGV 规模扩大至 40 台的超大规模集群时, AECL-MADDPG 较 MADDPG 性能提升近 1 倍, 较 TD3 提升超 78%, 传统 MADDPG、TD3 均采用固定高斯噪声探索策略, AGV 数量越多, 环境拥塞度越高, 固定探索策略要么探索不足, 导致 AGV 陷入局部拥堵死锁, 频繁触发碰撞惩罚; 要么探索过度, 导致路径冗余、任务完成效率下降, 最终在 30 台以上出现性能下滑.

3.6 消融实验

为验证自适应探索 (AE) 与课程学习优先经验回放 (CL-PER) 两项改进点的独立有效性及协同效应, 设计 4 组消融实验:

1) 基线组 (MADDPG): 标准 MADDPG, 固定高斯噪声 (ε=0.1)+随机经验回放;

2) 仅 AE 组 (MADDPG+AE):标准 MADDPG+自适应探索策略+随机经验回放;

3) 仅 CL-PER 组 (MADDPG+CL-PER): 标准 MADDPG+固定高斯噪声+CL-PER 机制;

4) 完整组 (AECL-MADDPG).

所有实验组在相同参数下训练 5000 回合, 测试场景为 15 台 AGV 高密度环境, 结果如图 5 示 (任务成功率随训练回合变化曲线).

如图 5 示基线组 (MADDPG) 稳定成功率仅 86%; 仅 AE 组成功率提升至 93%, 证明虽能动态调整探索, 但缺乏足够的关键样本支撑, 策略优化缓慢; 仅 CL-PER 组成功率提升至 97%, 收敛速度较基线组加快约 30%, 表明 CL-PER 能显著提升样本效率,

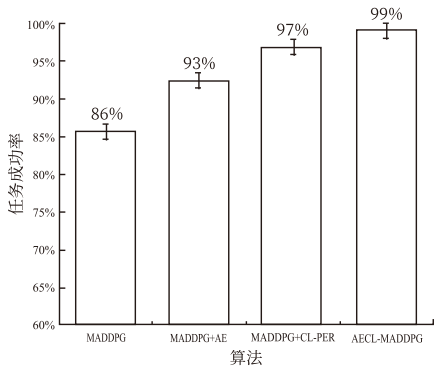

图5 任务成功率

但学到的避堵规则无法灵活适配动态环境, 在高拥堵场景仍易陷入局部最优; 完整组 (AECL-MADDPG) 成功率达到 99%, 且性能提升超过两项改进点单独作用和, 验证了 CL-PER 为自适应探索提供“高质量经验库”, 使探索行为更具针对性, 避免盲目探索; 自适应探索则为 CL-PER 提供动态场景反馈, 使样本库持续更新动态环境下的有效经验, 形成闭环.

3.7 路径规划效果可视化

为直观展示 AECL-MADDPG 的路径规划能力, 选取 15 台 AGV 测试中的典型成功案例, 结果如图 6 示 (彩色线条为 AGV 轨迹, 方形为起点, 叉形为终点, 灰色栅格为障碍物).

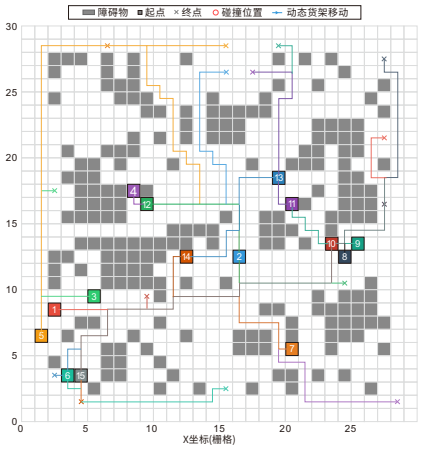

图6 CL-MADDPG 运行轨迹图

从图 6 观察到以下特征:

1) 路径平滑性: 所有 AGV 轨迹无明显迂回, 均沿“起点-终点”直线方向附近规划, 平均路径长度接近理论最短路径 (曼哈顿距离), 验证路径优化效果;

2) 避障协同性: 在地图中部、右下角等交叉路径区域, AGV 通过“时序避让”(如部分 AGV 短暂停留 0.5~1 步) 或“路径绕行”(如偏离主路径 1~2 栅格) 实现无碰撞通行, 无两台 AGV 在同一栅格重叠;

3) 动态适应能力: 动态货架移动后 (如地图左侧货架从 (5,8) 移至 (6,8)), 附近 AGV(轨迹编号 7) 能

实时调整路径, 避免与移动后的货架碰撞, 体现对动态环境的适应性.

MADDPG 的轨迹如图 7 示, 对比可见其存在明显缺陷, 部分 AGV(轨迹编号 3、9) 在交叉区域发生短暂拥堵, 路径迂回导致平均长度增加约 12%, 且存在 2 次轻微碰撞 (轨迹编号 5 与 8 在 (18,12) 栅格重叠).

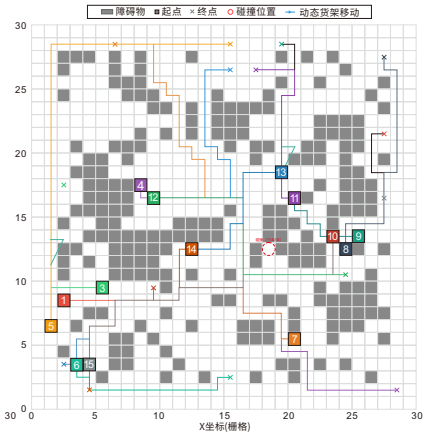

图7　DDPG 运行轨迹图

3.8　工业适配性分析

本文算法具备以下工业适配特性:

1) 实时性: 单步决策延迟<10ms, 满足仓储 AGV 50ms 响应要求;

2) 可扩展性: 状态空间设计包含任务优先级、载重状态、电量等调度信息, 可无缝对接 WMS/WCS 系统;

3) 鲁棒性: 课程学习机制使算法在不同复杂度场景间迁移时无需重新训练.

4　结　论

针对大型智能仓储多 AGV 路径规划中“探索低效”“收敛缓慢”“协同不足”三大核心问题,本文提出基于自适应探索与课程学习的改进型 MADDPG 算法 (AECL-MADDPG); 首先通过融合环境拥塞度与双 Critic 决策不确定性,实现探索强度动态调节,与固定噪声策略不同的是能根据局部环境状态实时调整,在拥塞区域增强探索以发现绕行路径,在空旷区域降低探索以提升效率,平均碰撞次数较标准 MADDPG 降低 88.6%,解决传统固定探索的盲目性问题;其次将课程学习与 PER 机制融合, 通过多维度晋级评估与回退保护机制, 实现"从易到难"的稳健学习,与传统 Curricular PER 仅关注样本难度不同,本文方法同时考虑任务难度递进与样本优先级,使算法收敛速度较标准 MADDPG 加快 50%,样本利用率提升约 40%,解决复杂任务中“从零开始”训练的停滞问题;最后在协同效应方面, AE 与 CL-PER 的结合使算法在 15 台 AGV 高密度场景下任务成功率达到 99.1%,平均路径长度缩短 11.6%,综合性能优于 DDPG、MADDPG 等基线算法,为多 AGV 协同路径规划提供高效解决方案.

参考文献 (References)

[1] 宋莹, 杨金波, 胡东东. 基于改进 CBS 算法的多 AGV 路径规划研究[J]. 机床与液压, 2026, 54(5): 93-102.
(Song Y, Yang J B, Hu D D. Multi-AGV path planning based on improved CBS algorithm[J]. Machine Tool & Hydraulics, 2026, 54(5): 93-102.)

[2] 孙孝飞, 郭捷, 魏灿名, 等. 3C 智能制造工厂的 AGV 智慧物料传输与调度综述[J]. 中南大学学报: 自然科学版, 2025, 56(2): 514-535.
(Sun X F, Guo J, Wei C M, et al. Review of AGV smart material transmission and dispatching in 3C smart manufacturing factories[J]. Journal of Central South University: Science and Technology, 2025, 56(2): 514-535.)

[3] 司明, 邬伯藩, 胡灿, 等. 智能仓储交通信号与多 AGV 路径规划协同控制方法[J]. 计算机工程与应用, 2024, 60(11): 290-297.
(Si M, Wu B F, Hu C, et al. Collaborative control method of intelligent warehouse traffic signal and multi-AGV path planning[J]. Computer Engineering and Applications, 2024, 60(11): 290-297.)

[4] 林国义, 黄千禧, 谢帅, 等. 面向智能制造的 AGV 与柔性作业车间协同调度模型与算法[J]. 控制与决策, 2026, 41(4): 1166-1175.
(Lin G Y, Huang Q X, Xie S, et al. A model and algorithm for coordinated scheduling of AGV and flexible job shop in intelligent manufacturing[J]. Control and Decision, 2026, 41(4): 1166-1175.)

[5] dos Reis W P N, Couto G E, Junior O M. Automated guided vehicles position control: A systematic literature review[J]. Journal of Intelligent Manufacturing, 2023, 34(4): 1483-1545.

[6] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in Actor-Critic methods[C]. Proceedings of the 35th International Conference on Machine Learning. Stockholm, 2018: 1587-1596.

[7] 张书凡, 毛剑琳, 张凯翔, 等. 面向不确定性的多机器人路径鲁棒规划研究综述[J]. 控制与决策, 2024, 39(12): 3873-3888.
(Zhang S F, Mao J L, Zhang K X, et al. Survey on robust multi-robot path planning under uncertainty[J]. Control and Decision, 2024, 39(12): 3873-3888.)

[8] 熊骏, 张文博, 熊智, 等. 多智能体协同路径规划综述[J]. 系统仿真学报, 2025, 37(12): 3033-3049.
(Xiong J, Zhang W B, Xiong Z, et al. Survey of cooperative multi-agent path finding[J]. Journal of System Simulation, 2025, 37(12): 3033-3049.)

[9] Ho G T S, Tang Y M, Leung E K H, et al. Integrated reinforcement learning of automated guided vehicles

dynamic path planning for smart logistics and operations[J]. Transportation Research — Part E: Logistics and Transportation Review, 2025, 196: 104008.

[10] 刘志飞, 曹雷, 赖俊, 等. 多智能体路径规划综述[J]. 计算机工程与应用, 2022, 58(20): 43-62.
(Liu Z F, Cao L, Lai J, et al. Overview of multi-agent path finding[J]. Computer Engineering and Applications, 2022, 58(20): 43-62.)

[11] Zhang X Y, Zou Y S. Collision-free path planning for automated guided vehicles based on improved A* algorithm[J]. Systems Engineering — Theory & Practice, 2021, 41(1): 240-246.

[12] Mehrdadi B, Lockwood S. Design of a navigation system for automated guided vehicles operating in a man-machine shared environment[C]. Laser Metrology and Machine Performance III. West Yorkshire, 1997: 453-462.

[13] 王彬, 聂建军, 李海洋, 等. 优化 A*与动态窗口法的移动机器人路径规划[J]. 计算机集成制造系统, 2024, 30(4): 1353-1363.
(Wang B, Nie J J, Li H Y, et al. Mobile robot path planning based on optimized A*and dynamic window approach[J]. Computer Integrated Manufacturing Systems, 2024, 30(4): 1353-1363.)

[14] Wang D K, Wu H X, Zheng W G, et al. A mixed-integer linear programming model for addressing efficient flexible flow shop scheduling problem with automatic guided vehicles consideration[J]. Applied Sciences, 2025, 15(6): 3133.

[15] 于绍琪, 田玉平. 基于 Petri 网与多智能体深度强化学习的 AGV 路径规划[J]. 控制与决策, 2025, 40(5): 1438-1446.
(Yu S Q, Tian Y P. AGV path planning based on Petri net and multi-agent deep reinforcement learning[J]. Control and Decision, 2025, 40(5): 1438-1446.)

[16] 孙哲, 马胜男, 解相朋, 等. 基于仿生算法的多式联运路径规划方法综述[J]. 控制与决策, 2025, 40(2): 375-386.
(Sun Z, Ma S N, Xie X P, et al. Bio-inspired optimization-based path planning algorithms in multimodal transportation: A survey[J]. Control and Decision, 2025, 40(2): 375-386.)

[17] Huang J, Liu H, Junginger S, et al. Mobile robots in automated laboratory workflows[J]. SLAS Technology, 2025, 30: 100240.

[18] Bae H, Kim G, Kim J, et al. Multi-robot path planning method using reinforcement learning[J]. Applied Sciences, 2019, 9(15): 3057.

[19] 王岩红, 钟颖, 张允华. 基于改进 Q 学习的电动冷藏车多目标跨区域路径优化[J]. 控制与决策, 2026, 41(3): 741-753.
(Wang Y H, Zhong Y, Zhang Y H. Multi-objective cross-regional path optimization for electric refrigerated vehicles based on improved Q-learning[J]. Control and Decision, 2026, 41(3): 741-753.)

[20] 李佩哲, 张文彪. 基于改进经验回放策略的路径规划算法[J]. 控制与决策, 2025, 40(8): 2545-2552.
(Li P Z, Zhang W B. A path planning algorithm based on improved experience replay strategy[J]. Control and Decision, 2025, 40(8): 2545-2552.)

[21] Campuzano G, Lalla-Ruiz E, Mes M. The two-tier multi-depot vehicle routing problem with robot stations and time windows[J]. Engineering Applications of Artificial Intelligence, 2025, 147: 110258.

[22] Zafar M, Khan R A, Fedoseev A, et al. HetSwarm: Cooperative navigation of heterogeneous swarm in dynamic and dense environments through impedance-based guidance[C]. International Conference on Unmanned Aircraft Systems. Charlotte, 2025: 309-315.

作者简介

王崟铸 (1999−), 男, 硕士生, 主要研究方向为物流系统建模与路径优化, E-mail: 1292299791@qq.com;

陶翼飞 (1983−), 男, 讲师, 博士, 主要研究方向为复杂系统建模、调度与智能算法, E-mail: 676379098@qq.com;

田华亭 (1984−), 男, 高级工程师, 博士生, 主要研究方向为移动机器人导航、控制等关键技术, E-mail: ksectian@foxmail.com;

许容 (1998−), 男, 助理工程师, 硕士, 主要研究方向为生产车间建模与布局优化, E-mail: 344646060@qq.com;

钱傲 (2002−), 男, 硕士生, 主要研究方向为物流系统建模与路径优化, E-mail: 2423320850@qq.com;

佟雨擎 (2002−), 女, 硕士生, 主要研究方向为生产车间建模与调度, E-mail: 1844772141@qq.com.