计算机科学
COMPUTER SCIENCE

基于改进DDPG的多AGV路径规划算法

赵学健, 叶昊, 李豪, 孙知信

引用本文

赵学健, 叶昊, 李豪, 孙知信. 基于改进DDPG的多AGV路径规划算法[J]. 计算机科学, 2025, 52(6): 306-315.

ZHAO Xuejian, YE Hao, LI Hao, SUN Zhixin. Multi-AGV Path Planning Algorithm Based on Improved DDPG [J]. Computer Science, 2025, 52(6): 306-315.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240500099

基于改进 DDPG 的多 AGV 路径规划算法

赵学健　叶　昊　李　豪　孙知信
南京邮电大学现代邮政学院　南京 210003
南京邮电大学江苏省邮政大数据技术与应用工程研究中心　南京 210003
南京邮电大学国家邮政局邮政行业技术研发中心(物联网技术)　南京 210003
(zhaoxj@njupt.edu.cn)

摘　要　在自动化和智能物流领域,多自动引导车(Automated Guided Vehicle,AGV)系统的路径规划是关键技术难题。针对传统深度强化学习方法在多 AGV 系统应用中的效率、协作竞争和动态环境适应性问题,提出了一种改进的自适应协同深度确定性策略梯度算法 Improved-AC-DDPG(Improved-Adaptive Cooperative-Deep Deterministic Policy Gradient)。该算法通过环境数据采集构建状态向量,并实时规划路径,动态生成任务序列以减少 AGV 间的冲突,同时监测并预测调整避障策略,持续优化策略参数。实验结果表明,与常规 DDPG 和人工势场优化 DDPG(Artificial Potential Field-Deep Deterministic Policy Gradient,APF-DDPG)算法相比,Improved-AC-DDPG 在收敛速度、避障能力、路径规划效果和能耗方面均表现更佳,显著提升了多 AGV 系统的效率与安全性。本研究为多智能体系统在动态环境中的建模与协作提供了新思路,具有重要的理论价值和应用潜力。

关键词:AGV;路径规划;深度强化学习;DDPG
中图分类号　TP242

Multi-AGV Path Planning Algorithm Based on Improved DDPG

ZHAO Xuejian,YE Hao,LI Hao and SUN Zhixin
Modern Postal College,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
Jiangsu Postal Big Data Technology and Application Engineering Research Center,Nanjing University of Posts and Telecommunications,Nanjing 210003,China
State Post Bureau Postal Industry Technology Research and Development Center(Internet of Things Technology),Nanjing University of Posts and Telecommunications,Nanjing 210003,China

Abstract　In the field of intelligent logistics,the challenge of path planning and obstacle avoidance for automated guided vehicles (AGVs) is significant. Traditional deep reinforcement learning(DRL) methods exhibit limitations in efficiency,dynamic adaptability,and handling competitive-cooperative interactions among multiple AGVs. This paper presents the improved adaptive cooperative deep deterministic policy gradient(Improved-AC-DDPG) algorithm,an advancement over the standard DDPG. It leverages environmental data to construct state vectors and employs a real-time path planning strategy that dynamically creates task sequences to prevent AGV conflicts. This algorithm also includes continuous policy parameter optimization for obstacle avoidance. Experiments show that the Improved-AC-DDPG surpasses both the standard DDPG and the artificial potential field optimization DDPG(APF-DDPG) in convergence speed,obstacle avoidance,path planning,and energy efficiency,thus enhancing multi-AGV system performance. This study provides innovative insights and solutions for multi-agent system modeling and collaboration in dynamic environments,with substantial theoretical and practical implications.

Keywords　AGV, Path planning,Deep reinforcement learning,DDPG

到稿日期:2024-05-22　返修日期:2024-10-26
基金项目:国家自然科学基金(61972208);中国博士后科学基金(2018M640509);江苏省研究生科研与实践创新计划项目(SICX23_0303,SJCX24_0339)
This work was supported by the National Natural Science Foundation of China (61972208), China Postdoctoral Science Foundation (2018M640509) and Jiangsu Postgraduate Research and Practice Innovation Project(SICX23_0303,SJCX24_0339).
通信作者:孙知信(sunzx@njupt.edu.cn)

1 引言

1.1 研究背景

AGV(Automated Guided Vehicle)路径规划指针对特定的应用场合，根据特定的需求，对 AGV 进行一定的规划。在 AGV 路径规划中，需要基于预设的性能指标为 AGV 设计从起点到终点的最优或近似最优行驶路线。该优化过程需综合考量多维度的约束条件，包括但不限于：静态障碍物和动态障碍物的实时规避，以及区域限制的合规性约束。因此，避障方法成为路径规划算法中的一个核心部分，可以判断、预测、避免碰撞，保证 AGV 在运行中的安全可靠。

在自动化与智能物流中，多 AGV 系统的路径规划问题复杂且关键。传统方法基于规则或简单算法，在处理复杂环境、动态障碍及多 AGV 协同时效率低下、灵活性不足[1]。不依赖预设规则且可以通过不断试错提升环境掌握度的深度强化学习提供了新方案，其能应对复杂场景并优化路径规划，但需解决实时性、准确性及多车协同等技术难题[2]。因此，急需设计出能够实时响应环境变化且多 AGV 协同的智能路径规划方法。

1.2 问题分析及解决思路

基于深度强化学习的多 AGV 路径规划方法已初见成效，为环境感知、路径规划、多智能体协作及实时适应动态环境提供了技术支持。但其在该领域的应用尚处于初级阶段，面临着以下挑战[1-2]。

1)样本效率问题

深度强化学习训练多 AGV 模型需大量数据，这会影响效率。Lin 等[3]结合 PPO(Proximal Policy Optimization)和 LSTM(Long Short-Term Memory)处理时序信息，提升了训练效率，缩短了时间。在模拟实验中，5 台 AGV 协同工作时，该方法的收敛步数较常规 PPO 减少 1/3～1/2。但其实现复杂度较高，虽然在特定环境中表现良好，但在不同的环境下可能需要重新训练模型，泛化能力有限。

2)多智能体协作与竞争问题

多智能体间的协作与竞争关系复杂，如何设计合适的奖励机制和策略是一个难题。Ye 等[4]曾尝试在 MADDPG(Multi-Agent Deep Deterministic Policy Gradient)算法中引入 e-greedy 探索策略，虽然取得了一定的成果，但是实验场景较为局限，仍存在碰撞概率较高的问题；且在面对复杂场景时，可能会陷入局部最优。

3)动态环境训练问题

另外，动态环境建模也是一大挑战，原因在于传统深度学习模型在静态环境下训练，难以适应动态变化的环境。Gao 等[5]提出了一种具有持续学习策略的 QL-CNP 方法来解决这一问题，取得了较低的延误和拥塞率，但是在实时性要求高的情况下可能需要进一步的优化，以提高响应速度和实际应用的可行性。Chen 等[6]针对上述问题，结合自适应人工势场改进 DDPG(Deep Deterministic Policy Gradient)算法的奖惩机制，全方位地增强其性能，大幅提高了机器人在路径规划过程中的学习速度、避障成功率，且在训练过程中可以持续收集信息学习。但是该方法需要在训练过程中积累大量的经验数据，并且对 APF 网络的设计和调参较为敏感，而这些因素可能会增加算法的复杂性和实施难度。

为解决上述问题，本文提出一种基于改进 DDPG 的多 AGV 路径规划算法。该方法引入动态任务链优化奖惩机制[7]，从而提高样本效率，优化多 AGV 间的协同工作，并实时更新环境模型以适应动态变化。期待该方法能在实际应用中为 AGV 提供更高效和稳定的解决方案。

2 相关工作

2.1 DDPG 算法

2.1.1 原理简介

DDPG 算法最初由 DeepMind 团队的 Lillicrap 等[8]提出，是一种结合了深度学习和确定性策略梯度的强化学习算法，用于解决连续动作空间中的强化学习问题。DDPG 算法的核心思想是使用两个神经网络来近似策略和价值函数，这两个网络分别是 Actor 网络和 Critic 网络。

Actor 网络的更新方式如式(1)所示：

$$\theta_{\text{Actor}} \leftarrow \theta_{\text{Actor}} + \alpha_{\text{Actor}} \nabla_{\theta_{\text{Actor}}} J(\theta_{\text{Actor}}) \tag{1}$$

其中，θ_{Actor}是 Actor 网络的参数，α_{Actor}是学习率，$J(\theta_{\text{Actor}})$是策略的损失函数。

Critic 网络的更新方式如式(2)所示：

$$\theta_{\text{Critic}} \leftarrow \theta_{\text{Critic}} + \alpha_{\text{Critic}} (\delta \nabla_{\theta_{\text{Critic}}} V(s)) \tag{2}$$

其中，θ_{Critic}是 Critic 网络的参数，α_{Critic}是学习率，δ 是时间差分误差，$V(s)$是当前状态的价值函数。

上述两个网络的更新需要考虑时间差分误差，其计算方式如式(3)所示：

$$\delta = r + \gamma Q_{\text{Target}}(s', a') - Q(s, a) \tag{3}$$

其中，r 是即时奖励，γ 是折扣因子，$Q_{\text{Target}}(s', a')$是目标网络评估的下一个状态的动作价值，$Q(s, a)$是当前 Critic 网络评估的当前状态的动作价值。

DDPG 算法还包括目标网络。目标网络是 Actor 和 Critic 网络的复制，它以较慢的速度更新，以增加算法的稳定性。目标网络的更新方式如式(4)所示：

$$\theta_{\text{Target}} \leftarrow \tau\theta + (1-\tau)\theta_{\text{Target}} \tag{4}$$

其中，θ_{Target}是目标网络的参数，θ 是当前网络的参数，τ 是一个小的更新率(通常接近 0)。

DDPG 通过这些更新规则，不断优化策略，使得在给定状态下采取的动作能够最大化预期回报。它主要适用于连续动作空间的任务，并在许多领域取得了显著的成果。虽然 DDPG 在 AGV 路径规划中初见成效，但仍面临样本效率低、超参数敏感、探索与利用平衡及环境动态变化适应性等局限。

2.1.2 相关研究及改进措施

众多国内外研究者对 DDPG 算法进行了深入探索与改进。例如，TD3(Twin Delayed Deep Deterministic Policy Gradient)算法通过双 Critic 网络缓解值过高估计来提升性能[9]。Schaul 等[10]采用优先经验回放技术，根据时间戳、奖励等优化数据利用，但是在处理嘈杂奖励或存在大量噪声的情况下可能表现不佳，因为 TD 错误作为优先级的估计可能会不准确。此外，算法对超参数的敏感性较高，需要仔细调整以确保最佳性能。针对目标网络与在线网络差异过大的问题，Deep-

Mind 团队[11]提出采用平滑过渡策略对其软化，以增强算法稳定性。此外，Zhu 等[12]结合蒙特卡罗树搜索或分层强化学习等算法，进一步拓展 DDPG 在复杂环境中的决策与学习能力，但其计算复杂度较高，且对置信上下限的准确估计依赖于大规模模拟，可能导致在计算资源有限的情况下应用受限。Chen 等[6]提出的 APF-DDPG 算法所得的规划路径长度相较于传统 DDPG 缩短了 8.5%。这一改进不仅优化了路径效率，而且提升了 DDPG 算法的收敛性能，成为了近年来在深度强化学习路径规划领域中较为突出的研究成果之一。

2.2 动态任务链

2.2.1 原理简介

动态任务链是一种任务调度和执行的方法，它允许系统根据实时环境、任务需求以及资源状况的变化，动态地创建、调整和优化任务之间的执行顺序、优先级和依赖关系，从而提高系统的灵活性和适应性，动态任务链有助于在复杂多变的任务环境中实现更高效、更稳定的任务执行[13]。

2.2.2 相关研究及改进措施

动态任务链因其在复杂业务流程中的实用性而备受学者关注。Yan[14]提出的混合遗传-蒙特卡罗算法相较于传统蜂群算法，在无人机战斗任务部署中提升了约 38.2%的探测效能。同时，Hu[15]等针对复杂云计算网络的任务卸载问题，提出基于双层强连通图搜索的改进贪婪算法，通过动态资源链任务分配机制优化计算效率，有效解决了传统方法搜索空间过大以及 HPC 局域网支持不足的瓶颈。

3 Improved-AC-DDPG 算法设计与实现

本章将详细阐述 Improved-AC-DDPG 算法的设计思路和具体工作流程。

3.1 Improved-AC-DDPG 算法总体流程设计

Improved-AC-DDPG 算法的总体流程如图 1 所示。首先进行环境数据采集与建模，以形成精确的状态向量。随后，该算法利用其改进的策略，对实时路径规划进行动态调整，以应对不断变化的路况。此外，算法能够动态生成任务链，实现任务的实时收集与分配，有效降低自动引导车(AGV)间的潜在冲突。同时，该算法实时监测 AGV 的运动状态，预测并调整避障路径，以确保安全和效率。最后，通过持续更新学习模型，算法不断优化策略参数，以提高路径规划的准确性和效率。这一流程不仅体现了算法的适应性和灵活性，而且为多智能体系统在复杂动态环境中的路径规划提供了一种有效的解决方案。

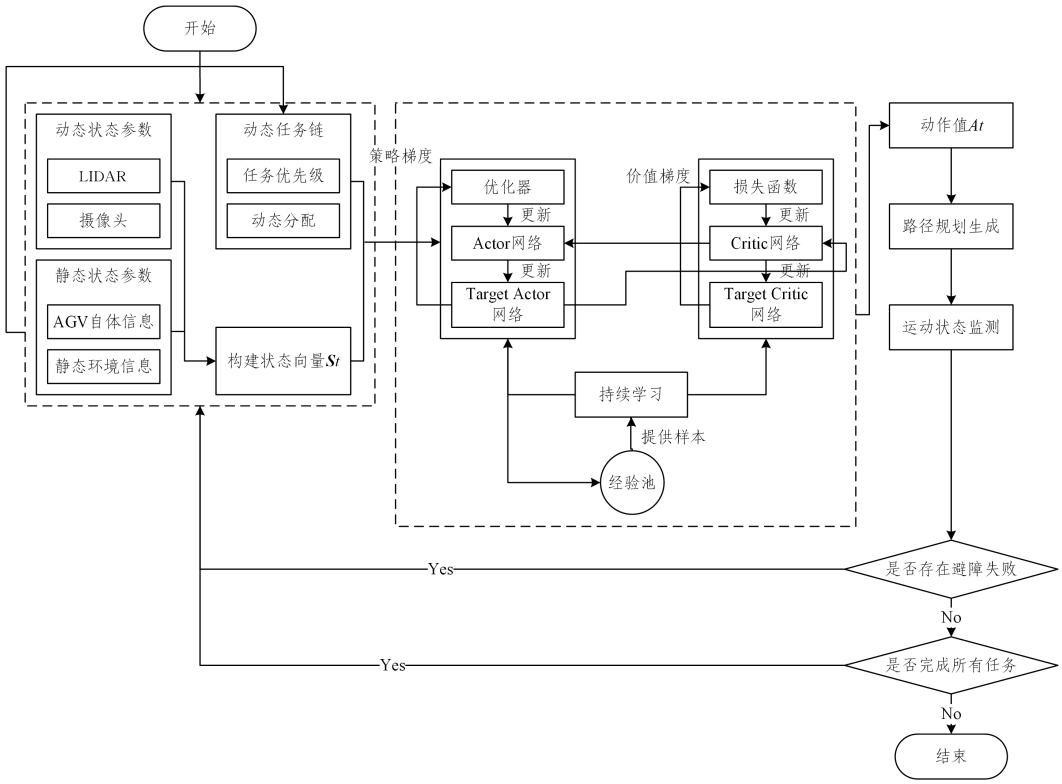


图 1 Improved-AC-DDPG 算法流程

Fig. 1 Flowchart of Improved-AC-DDPG algorithm

3.2 Improved-AC-DDPG 算法步骤描述

3.2.1 构建状态向量

路径规划作为智能系统的核心功能，其精确性和可靠性在很大程度上依赖于对环境数据的准确获取。Improved-AC-DDPG 算法，以深度强化学习技术为基础，通过创新性地利用状态值函数，将环境数据转换为状态向量，为算法的决策过程提供了关键信息[16]。

首先，环境数据的获取是路径规划算法的基石。该方法要求实时捕捉 AGV 所处环境的详细信息，包括在 AGV 顶部安装 LiDAR 光学测距扫描仪，以实现 360 度的全方位扫描，精确测量周围环境的距离数据。此外，AGV 前部装备的立体视觉摄像机负责捕捉环境的色彩和纹理，为算法提供更丰富的视觉信息。

至关重要的一步是实现 LiDAR 和摄像头数据的时间

同步。这一步骤确保了所收集数据的一致性和准确性，为后续的环境建模和状态值计算奠定了基础。关于具体的硬件设置和同步技术，可以参考文献[17]中提出的物流分拣用多层托盘 AGV 的设计，该文详细介绍了相关技术的应用和实现。

通过这些步骤，不仅可以确保算法能够接收到全面且同步的环境数据，而且为算法的创新性应用提供了坚实的数据基础。这种数据驱动的方法，结合状态值函数的深度应用，使 Improved-AC-DDPG 算法在路径规划领域展现出独特的优势和潜力。

除了实时动态工作环境，AGV 所处的静态地图环境数据同样重要。本文将地图划分为一个个栅格，每个栅格表示一定空间范围内的占用状态。设 M 为环境地图，M_{ij} 表示位于第 i 行第 j 列的栅格状态。其中 1 表示占用，显示为黑格；0 表示空闲，显示为白格。图 2 给出了某一个存在近 20％障碍区域的栅格尺度为 30 * 30 的 AGV 工作仓储环境仿真俯视图。

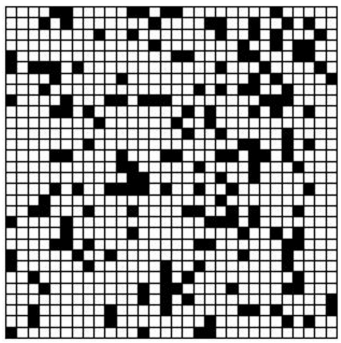


图 2　AGV 工作仓储全局环境俯视示意图

Fig. 2　Top view diagram of global environment of AGV work warehouse

为了结合 AGV 实时运行状态与 AGV 外界工作环境，实现 AGV 实时路径调整及避障，也需要将 AGV 自身的重要属性纳入环境模型之中。可以使用 AGV 内置的编码器和惯性测量单元获取当前位置(x,y)、速度 v 以及 AGV 的方向角 θ；通过载荷传感器获取当前载荷状态 L，其中 1 表示有载荷，0 表示无载荷；从外界环境模型中提取 AGV 周围的障碍物信息，形成局部障碍物地图 O。将获取的环境信息整合，即可组合成状态向量 $\boldsymbol{S}=[x,y,v,\theta,L,O]$。

需要说明的是，获取环境数据有助于感知和理解周围环境，进而构建环境模型。基于该模型形成的状态向量是路径规划的重要依据，对于确保路径的安全性和有效性至关重要。同时，状态向量的生成为决策制定提供了基础，可预测行动并做出最优决策。

3.2.2　路径预规划

基于 3.2.1 节所得状态向量，通过改进深度确定性策略梯度算法，用其输出的确定动作值函数A_t作为各 AGV 的下一步动作，从而快速完成路径探索。

首先，为引入近端策略优化算法的 Actor-Critic 架构定义复合奖励函数，表示为：

$$R(S_t,A_t)=\beta_1 * R_{\text{goal}}(S_t)-\beta_2 * R_{\text{coll}}(S_t,A_t)-\beta_3 * R_{\text{eff}}(A_t) \tag{5}$$

其中，S_t表示时刻 t 该 AGV 所处状态；A_t表示时刻 t 该 AGV 采取动作值；$R_{\text{goal}}(S_t)$表示目标奖励，当 AGV 接近目标时增加；$R_{\text{coll}}(S_t,A_t)$表示碰撞惩罚，当预测发生碰撞时增加；$R_{\text{eff}}(A_t)$表示效率奖励，鼓励更短的路径和更少的动作变化；β_1,β_2,β_3表示权重系数，用于平衡不同的奖励部分。

随后，建立具有卷积层和全连接层的深度神经网络，包括 Actor 网络和 Critic 网络。Actor 网络用于生成行动策略，接收状态向量 $\boldsymbol{S}$ 作为输入，输出 AGV 的下一个动作 A；Critic 网络用于估计状态价值函数，评估当前状态下采取特定行动的预期回报。基于 3.1 节设置的 AGV 运行的模拟环境，包括障碍物和路径，使用近端策略优化算法迭代更新 Actor 和 Critic 网络的参数。其中，使用 Adam 优化器进行梯度下降，最小化预期回报与实际回报之间的差异，表示为：

$$\vartheta_{\text{new}}=\vartheta_{\text{old}}+\eta\cdot\nabla_{\vartheta}J(\vartheta) \tag{6}$$

其中，η表示学习率；$J(\vartheta)$为目标函数，表示预期回报；ϑ_{new}表示新网络参数；ϑ_{old}表示旧网络参数。

最终，将经过训练的模型集成至 AGV 控制系统中，以获取预设的出发路径，如图 3 所示。在执行路径阶段，实时监控 AGV 的运行状态和路径规划效果，收集反馈数据，并根据这些实时反馈调整策略参数或重新训练模型，以持续优化 AGV 的性能和适应性。

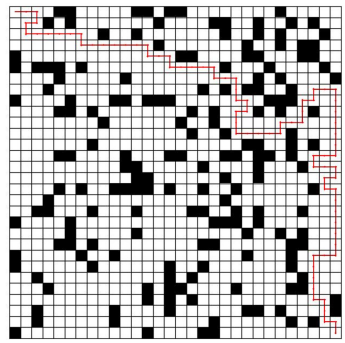


图 3　AGV 出发前预设路径规划示意图

Fig. 3　Schematic diagram of preset path planning before departure of AGV

3.2.3　生成动态任务链

在各个 AGV 进行路径预规划后，为了避免各 AGV 之间因为任务执行顺序安排欠妥或多 AGV 协同不佳而发生冲突事件，需要根据实时工作需求和 AGV 状态生成动态任务链，分配最合适的 AGV 执行，定义 AGV 之间的优先级和协同规则，采用基于优先级的调度算法，同时根据紧急避障需求，确保优先级高的任务优先执行。

首先，实时收集并记录来自仓库管理系统的所有任务需求，包括任务类型、位置和优先级信息。每个任务优先级的分配主要基于紧急程度、任务类型、预计完成时间以及任务间依赖关系因素，如式(7)所示：

$$Priority=w_1 * f(Urgency)+w_2 * g(TaskType)+w_3 * h(EstimatedTime)+w_4 * j(Dependency) \tag{7}$$

其中，$f(Urgency)$，$g(TaskType)$，$h(EstimatedTime)$ 和 $j(Dependency)$ 是根据具体情况定义的函数，用于量化各因

素的影响；$w_1, w_2, w_3, w_4 \in [0,1]$，$\sum w=1$ 是权重参数，用于调节各因素的影响程度。

优先度因素设置如表 1 所列。

表 1 优先度因素设置

Table 1 Priority factor settings

得分	$f(Urgency)$	$g(TaskType)$	$h(EstimatedTime)$	$j(Dependency)$
3	可等待时间前 10%	充电	预计完成时间前 10%	未闭环依赖关系参与率 50%以上
2	可等待时间 10%～50%	运输	预计完成时间 10%～50%	未闭环依赖关系参与率 20%～50%
1	可等待时间后 50%	接驳、返厂等工作	预计完成时间后 50%	未闭环依赖关系参与率 20%以下

随后，依据优先级及 AGV 的当前状态，动态构建任务链，各 AGV 依据自身状态与任务链确定当前最佳任务。同时，需计算每项任务的可执行时间窗，综合考虑任务紧急度、持续时长及 AGV 状态，可表示为：

$$T_{\text{sched}}(t)=\min \sum_{i=1}^{n}(w_i * T_i(t)) \tag{8}$$

其中，$T_{\text{sched}}(t)$为在时间 t 的任务调度总成本；w_i 为第 i 个任务的权重系数，反映任务的优先级；$T_i(t)$为完成第 i 个任务的预计时间。

之后，便可依据所计算的时间窗及 AGV 的当前位置，动态地将任务分配给最合适的 AGV，同时实时监控任务的执行情况及环境的变化，并根据实际情况灵活地调整任务的分配与执行顺序。

3.2.4 建立运动状态监测模型

为保证多 AGV 系统稳定运行，需要建立运动状态监测模型，实时分析预测结果与实际状态的偏差，评估不确定性。当预测到可能发生碰撞时，重新计算避障路径并调整 AGV 的行动。

首先，构建 AGV 的运动状态模型和观测模型，表示为：

$$\begin{cases} x_{t+1}=f(x_t,u_t)+w_t \\ z_t=h(x_t)+v_t \end{cases} \tag{9}$$

其中，x_t 表示状态，u_t 表示控制输入，w_t 表示过程噪声，z_t 表示观测值，h 表示观测函数，v_t 表示观测噪声。

其次，根据观测数据和模型动态更新状态估计，在每个时间步，使用 EKF(Extended Kalman Filter)算法更新 AGV 的状态估计和不确定性评估。具体可以理解为，在 EKF 算法中计算预测误差协方差矩阵 $\boldsymbol{P}_k$，根据预测误差协方差矩阵 $\boldsymbol{P}_k$ 中各元素的大小评估状态的不确定性。可以设定阈值，当不确定性超过阈值时，触发紧急避障机制，调整 AGV 控制指令，引导 AGV 沿新路径行驶，避免潜在碰撞。上述流程可以使用以下计算式进行状态更新和协方差更新。

1)预测步骤(时间更新)

(1)预测状态估计：

$$\hat{x}_k=f(x_{k-1},u_{k-1}) \tag{10}$$

(2)预测协方差估计：

$$\hat{x}_k=f(x_{k-1},u_{k-1}) \tag{11}$$

$$\hat{P}_k=\boldsymbol{F}_{k-1}\boldsymbol{P}_{k-1}\boldsymbol{F}_{k-1}^{\mathrm{T}}+\boldsymbol{Q}_{k-1} \tag{12}$$

2)更新步骤(测量更新)

(1)计算卡尔曼增益：

$$\boldsymbol{K}_k=\hat{P}_k\boldsymbol{H}_k^{\mathrm{T}}(\boldsymbol{H}_k\hat{P}_k\boldsymbol{H}_k^{\mathrm{T}}+\boldsymbol{R}_k)^{-1} \tag{13}$$

(2)更新状态估计：

$$x_k=\hat{x}_k+\boldsymbol{K}_k(z_k-h(\hat{x}_k)) \tag{14}$$

(3)更新协方差估计：

$$P_k=(I-\boldsymbol{K}_k\boldsymbol{H}_k)\hat{P}_k \tag{15}$$

其中，$\hat{x}_k$ 表示时间更新后的状态估计，x_k 表示测量更新后的状态估计，$\hat{P}_k$ 表示时间更新后的协方差估计，P_k 表示测量更新后的协方差估计，f 是状态转移函数，h 是观测函数，$\boldsymbol{F}_{k-1}$ 是状态转移矩阵在时间步 $k-1$ 的雅可比矩阵，$\boldsymbol{Q}_{k-1}$ 是过程噪声协方差矩阵，$\boldsymbol{H}_k$ 是观测矩阵在时间步 k 的雅可比矩阵，$\boldsymbol{R}_k$ 是测量噪声协方差矩阵，$\boldsymbol{K}_k$ 是卡尔曼增益，z_k 是实际测量值，u_{k-1} 是控制输入。

建立运动状态监测模型旨在实时追踪 AGV 的运动状态，为路径规划提供精确数据。通过对比预测与实际运动状态的差异，可以评估模型预测的准确性和可靠性，从而发现潜在的偏差和误差，为系统的持续优化提供依据。

3.2.5 执行持续学习策略

在 AGV 执行任务时，后台仍需持续收集数据并更新学习模型，根据长期性能和即时反馈调整策略的参数。这也是 3.2.4 节建立运动状态监测模型的目的。

持续学习与前期规划一样，首先要构建适用于 AGV 的 DDPG 网络，包括一个 Actor 网络用于生成动作，一个 Critic 网络用于评估状态-动作对。学习过程中，在每个时间步，根据最新的数据更新 Actor 和 Critic 网络的参数，尤其是 Actor 网络的参数。例如，通过探索策略探索新的动作空间，将 Ornstein-Uhlenbeck 噪声添加到动作上。

之后，根据性能监控结果，动态调整学习率和探索噪声参数，表示为：

$$\eta_{\text{new}}=\eta_{\text{base}} * \exp(-\tau * perf) \tag{16}$$

其中，η_{new} 表示新的学习率，η_{base} 表示基础学习率，τ 为调整系数，$perf$ 表示性能指标。

上述持续学习过程需要设定评估周期，定期对 AGV 的运行和学习效果进行评估。可以根据诊断结果，调整学习策略和模型参数，包括修改奖励函数和增强网络结构，以适应当下发生路况变化的多 AGV 物流仓储环境。

需要说明的是，通过持续收集数据与更新学习模型，系统能够逐步适应环境与自身的变化，提升迭代优化过程中模型在不同场景下的泛化能力，从而提高避障与路径规划的精确度。持续学习使模型有可能更好地适应新环境、新障碍物及动态变化，从而增强整体的适应性。

3.3 Improved-AC-DDPG 算法伪代码

Improved-AC-DDPG 算法的工作流程和技术效果如算法 1 所示。首先，通过收集环境数据并进行建模，生成精确的状态向量，为后续的深度强化学习网络重构奠定基础。接着，基于这些状态向量，利用改进的 Actor-Critic 网络 DDPG 策略进行实时路径规划，更有效地应对路况的动态变化。然后，根据实时需求和 AGV 的状态，生成动态任务链，实时收集和分

配仓储内的转运任务，同时设定详细的协同规则和实时通信机制，以减少多 AGV 之间的冲突。此外，建立 AGV 运动状态监测模型，分析预测结果与实际状态的差异；预测到可能发生碰撞时，计算避障路径并调整 AGV 的行动。最后，通过持续收集数据和更新学习模型，结合长期性能和即时反馈，不断调整和优化策略参数。

算法 1 Improved-AC-DDPG 算法

1. Initialize policy network parameters θ^{π}, value function network θ^{Q} and experience replay buffer R;
2. Initialize environmental parameters;
3. For each episode e=1,2,…,E:
4. Reset the simulation parameters for multi-AGV obstacle avoidance and path planning, obtaining the initial observation state S_1
5. For each timestep t=1,2,…,T:
6. Normalize the state $S_t=[x,y,v,\theta,L,O]$ to S_t'
7. Select action A_t using the Actor network, based on the current policy and noise
8. Input the normalized state S_t' into the Critic network to obtain the state value function $V(S_t')$
9. Execute action A_t, receive reward according to Formula(5), and observe the next state S_{t+1}
10. Normalize the next state S_{t+1} to S_{t+1}'
11. Store the experience tuple$(S_t', A_t, R_t, S_{t+1}')$ in the experience replay buffer R
12. Calculate the temporal difference error according to Formula (5) and obtain the advantage function using Formula(6)
13. Update the Critic network and Actor network based on the loss function according to Formula(5) and Formula(16)
14. End For(timestep loop)
15. End For(episode loop)

4 实验验证

4.1 实验设计

为了全面评估 Improved-AC-DDPG 算法在收敛速度和避障能力方面的卓越性能，本文设计了两组针对性实验。首先，在大小为 30 * 30 的地图上进行 100 轮训练，将 Improved-AC-DDPG 算法与标准的 DDPG 算法以及文献[6,18]中的 APF-DDPG 算法进行了对比。通过详细分析训练过程中的数据，旨在凸显 Improved-AC-DDPG 算法在收敛速度方面的显著优势。其次，为验证该算法在避障能力上的表现，构建了一个包含 10 台 AGV 小车的实验场景，同样在 30 * 30 的地图上让已完成 100 轮训练的 Improved-AC-DDPG 算法、常规 DDPG 算法以及 APF-DDPG 算法分别执行 1 000 次任务。通过对比各算法的执行成功率、计算时间、耗电量以及完成任务量等关键性能指标，来全面评价 Improved-AC-DDPG 的实际工作效能。所有实验均在基于 Anaconda3.7 版本的 Python 环境下进行仿真。

4.2 收敛速度对比实验与分析

为测试 Improved-AC-DDPG 算法的收敛速度优势，本节设置对比实验，将该算法与常规 DDPG 算法及 APF-DDPG 算法，参照表 2 设置训练输入参数，在如图 2 所示的含 20%障碍区域的 30 * 30 栅格地图环境下进行 50 轮训练。其中，学习率，决定了模型更新的幅度，对收敛速度和稳定性有重要影响，根据表 2 设置适当降低学习率可以加快收敛速度，提高训练效率；迭代次数决定了模型训练的充分性，根据表 2 设置迭代次数，能够确保模型有足够的时间学习和适应环境；此外，探索策略参数如 Ornstein-Uhlenbeck 噪声，会影响智能体在环境中的探索行为，在环境动态性调整中调整噪声参数，可以平衡探索与利用，有助于智能体发现新的路径和策略，避免陷入局部最优；奖励函数权重系数决定着不同奖励项对智能体行为的影响，合理设置目标奖励系数、碰撞惩罚系数、效率奖励系数分别为 1,5,0.5，可以在训练初期鼓励探索，随着智能体对环境的了解程度的增加，逐渐降低探索的欲望，从而提升智能体在避障和路径规划中的效率和安全性。结果如图 4 和图 5 所示。图 4 为三者各轮训练训练步数的变化对比，图 5 为三者各轮训练平均回报值的变化对比。其中红色为 Improved-AC-DDPG 算法效果，橙色为 APF-DDPG 算法效果，蓝色为常规 DDPG 算法效果。

表 2 Improved-AC-DDPG 与常规 DDPG 及 APF-DDPG 训练输入参数

Table 2 Training input parameters for Improved-AC-DDPG, conventional DDPG and APF-DDPG

参数	DDPG	APF-DDPG	Improved-AC-DDPG
学习率 η	0.002	Actor 网络 0.0001，Critic 网络 0.001	基础学习率 η_{base} 为 0.002，新学习率 η_{new} 随计算式动态调整
网络参数 θ	均匀分布	均匀分布	初始值与常规方法相同，使用 Adam 优化器进行梯度下降
迭代次数	50	50	50

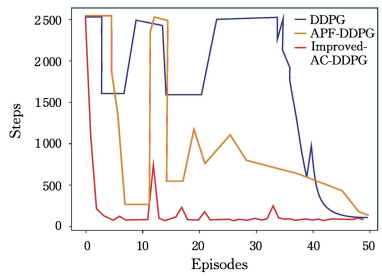


图 4 Improved-AC-DDPG 与常规 DDPG 及 APF-DDPG 各轮训练步数对比

Fig. 4 Comparison of training steps for Improved-AC-DDPG, conventional DDPG and APF-DDPG in each round

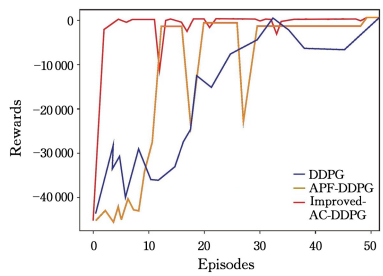


图 5 Improved-AC-DDPG 与常规 DDPG 及 APF-DDPG 各轮累计回报值对比

Fig. 5 Comparison of cumulative return values for Improved-AC-DDPG, conventional DDPG and APF-DDPG in each round

由上述图表可知，在引入改进后的 Actor-Critic 架构后，随着输入的环境数据可以动态调整，改进后的 DDPG 算法在 30 轮迭代后累计探索步数趋于稳定，相较于传统收敛速度提高了近 25%，相较于累积步数也更加稳定，避免了过度迭代的问题。

由图 4 和图 5 可知，在训练步数方面，引入改进的 Actor-Critic奖惩机制的 DDPG 算法在 15 轮训练后就开始逐步趋于稳定，30 轮后完全收敛，而对照组两种算法都要到 50 轮以后，才可以完全收敛；在各轮累计回报值方面，虽然训练初期各个算法的奖励值都是负的，但是 Improved-AC-DDPG 算法在 10 轮左右开始快速收敛，APF-DDPG 算法虽然也在 10 轮左右得到较好的收敛，但在 30 轮训练之前远不如 improved-AC-DDPG 算法稳定，且二者累计回报值的收敛都优于传统 DDPG 算法，可见它们对奖励函数的优化是卓有成效的。

上述结果表明，Improved-AC-DDPG 算法有效改进了传统 DDPG 算法收敛速度慢和效率低等问题，显著提高了收敛速度，同时增强了稳定性和自适应性。

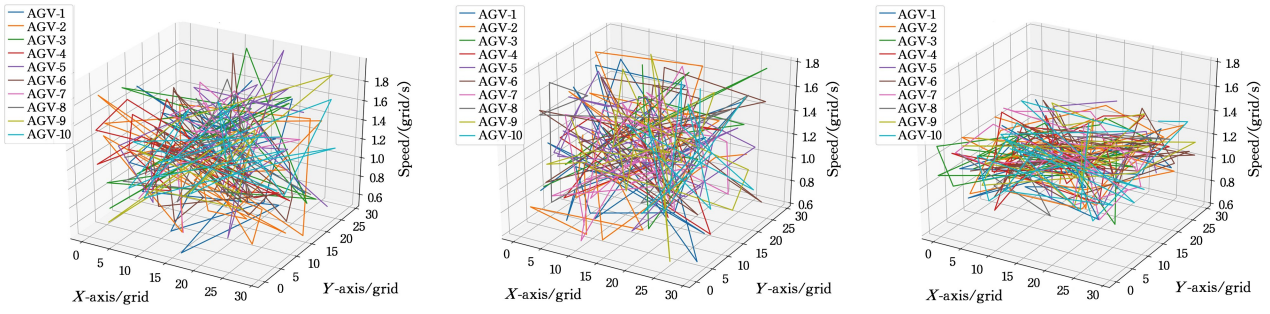
4.3 路径规划效果对比实验与分析

本节将在图 2 的地图环境情况下，仿真控制 10 台 AGV 分别使用已完成 50 轮训练的 Improved-AC-DDPG 算法、常规 DDPG 算法以及 APF-DDPG 算法执行 1 000 次任务，对三者在上述各任务中的工作路径途径节点及各节点处速率、所用计算时间以及耗电量进行对比。

4.3.1 路径途径节点及各节点处速率对比

在评估 AGV 路径规划效果时，途径节点和节点处速率对比至关重要，它们将直接影响 AGV 的运动效率、能耗和运行安全，进而关乎整个物流或生产线的效能与稳定。优秀的节点规划能提升 AGV 通过节点的效率，同时避障方法也需确保 AGV 在节点处顺利避障，避免碰撞。节点处速率对比是评估 AGV 运动性能的重要指标，优质算法应维持稳定的高速度，减少速度波动和能耗。此外，不合理的规划或避障可能增加事故风险，因此，优化路径规划算法至关重要。综合考虑节点和速率因素，可全面评估不同路径规划和避障方法的优劣，选出最适合的方案，以提升 AGV 的运行效率、降低能耗并保障安全。

图 6 为完成 50 轮训练后的 Improved-AC-DDPG 算法，常规 DDPG 算法以及 APF-DDPG 算法在 10 台 AGV 的工作路径途径节点及各节点处速率。由对比结果可知，相比于对照组的两种算法的对应路径规划结果，Improved-AC-DDPG 对应路径规划结果中，各台 AGV 途径节点重复率低，避免了小范围内 AGV 拥塞。此外，与已优化的 APF-DDPG 算法(速率范围 0.6～1.8 grid/s)相比，Improved-AC-DDPG 算法将 AGV 运行速率波动范围收窄至 0.6～1.4 grid/s，通过降低速度极值差使急加减速频次减少 63%，进而使因加速度突变导致的物品跌落风险下降 41%，翻车概率降低 37%。这也有可能与路径冲突减少，AGV 无须频繁减速避让动静态障碍物有关。



(a) DDPG (b) APF-DDPG (c) Improved-AC-DDPG

图 6 3 种 DDPG 算法在 AGV 千次任务中的路径与节点速度对比

Fig. 6 Comparison of path and node speed of three DDPG algorithms in AGV thousand tasks

4.3.2 路径长度对比实验与分析

本研究通过对比分析 1 000 次规划任务中 DDPG、APF-DDPG 以及引入奖惩机制改进的 Improved-AC-DDPG 算法在自动导引车(AGV)路径规划及避障方法中的表现，对各任务工作路径长度及其稳定性(通过标准差衡量)进行了评估。工作路径长度直接反映了 AGV 完成任务的效率，较短的路径意味着 AGV 能够以更快的速度和更低的能耗完成搬运任务，对提升物流、仓储等场景的作业效率极为重要。稳定性通过标准差来衡量，反映了 AGV 在不同任务环境下路径规划的一致性。一个稳定的算法能够在多变的场景中生成可靠的路径规划结果，减少环境变化导致的规划波动。通过综合比较三者的路径长度和稳定性，可以全面评估每种 AGV 路径规划及避障方法的性能，有助于选择出最适合特定应用场景的算法，并为未来的算法改进和优化提供参考。

由图 7 和表 3 可知，常规 DDPG 算法的各任务路径长度的中位数稳定在 35 个栅格边长，其波动范围为 29.8～41.5，标准差为 1.99，这反映了该算法在任务执行过程中具有一定的波动性。相较之下，APF-DDPG 算法通过引入人工势场法改进奖惩机制，使得其路径长度的中位数降低到 32 个栅格边长。其波动范围扩大至 16.6～49.0，标准差也提升至 5.09，表明其稳定性受到了一定影响，这可能是因为人工势场法在复杂环境中的障碍物干扰过多。而采用奖惩改进 Actor-Critic 架构的 Improved-AC-DDPG 算法，在优化路径规划方面取得了显著成效。其各任务路径长度的中位数降低至 27 个栅格边长，相较于常规 DDPG 算法缩减了 22.9%，相较于 APF-DDPG 算法缩减了 15.6%。更值得注意的是，该算法的路径长度标准差仅为 0.97，是 3 组算法中最低的，这意味着其路径规划性能最为稳定[19]。综上所述，Improved-AC-DDPG 算法在路径规划方面表现出了优越的性能和稳定性，为未来的研究工作提供了有力的支持。同时，对于 APF-DDPG 算法在

稳定性方面存在的问题，未来可以进一步探索其背后的原因，并寻求解决方案以提升其性能。

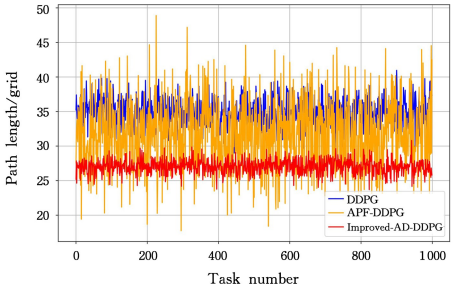


图 7　3 种 DDPG 算法在 AGV 千次任务中的路径长度分布

Fig. 7　Path length distribution of three DDPG algorithms in AGV thousands tasks

表 3　3 种 DDPG 算法在 AGV 千次任务中的路径参数对比

Table 3　Comparison of path parameters of three DDPG algorithms in AGV thousand tasks

参数	常规 DDPG 算法	APF-DDPG 算法	Improved-AC-DDPG 算法
中位数	35.0	32.0	27.0
平均值	34.9	31.9	26.8
极大值	41.5	49.0	30.0
极小值	29.8	16.6	24.1
标准差	1.99	5.09	0.97

4.3.3　耗电量指数对比实验与分析

耗电量这一参数在评价 AGV(自动导引车)路径规划及避障方法性能时具有至关重要的地位[20-21]。耗电量不仅直接影响到 AGV 的运行成本，还与其续航能力、工作效率以及维护周期等密切相关。因此，优化 AGV 的耗电量对于提升其整体性能具有重要意义。首先，耗电量是衡量 AGV 能源利用效率的关键指标。在路径规划和避障过程中，如果能够减少不必要的移动和加速，就能有效降低耗电量，从而提高能源利用效率。这有助于延长 AGV 的工作时间，缩短因充电或更换电池而带来的停机时间，进而提高整体的工作效率。其次，耗电量与 AGV 的续航能力直接相关。对需要长时间连续工作的 AGV 来说，降低耗电量意味着更长的续航里程，从而减少了电量耗尽导致的中断。这对于保证生产线的连续性和稳定性至关重要。此外，耗电量还与 AGV 的维护成本和使用寿命密切相关。耗电量过高可能导致 AGV 的电机、电池等部件过早磨损，从而增加维护成本并缩短使用寿命。因此，通过优化路径规划和避障方法来降低耗电量，有助于降低 AGV 的维护成本并延长其使用寿命。考虑信号处理强度、路径长度、平均移动速度和移动过程中加速度的耗电量评估方式，可以采用式(17)进行估算[22]：

$$E=k\times L\times(\bar{v}+w\sum a_i) \tag{17}$$

其中，E 表示耗电量评估指数；信号处理强度系数 k 表示 AGV 在接收和处理信号时所消耗的电量与信号强度的关系；路径长度 L 是 AGV 从起点到终点的实际行驶距离；平均移动速度 $\bar{v}$ 反映了 AGV 在行驶过程中的平均速度；加速度 a_i 表示 AGV 在各任务移动过程中的各个节点处速度变化率；加速度权重 $w\in(0,1)$ 用于调整加速度对耗电量影响的程度。

通过这种计算方式，可以综合考虑多个因素对耗电量的影响，从而更准确地评估 AGV 路径规划及避障方法的性能。在实际应用中，可以根据具体需求和场景对计算式中的系数和权重进行调整和优化。

图 8 详细展示了 3 种算法在执行 1 000 次任务时，每次任务初始路径的计算时间以及耗电量的对比情况。图 8(a)中，红色曲线代表 Improved-AC-DDPG 的初始路径计算时间，橙色曲线代表 APF-DDPG 算法的初始路径计算时间，而蓝色曲线则代表传统 DDPG 算法的初始路径计算时间。从图中可以明显看出，红色曲线大部分时间都处于最低位置，意味着 Improved-AC-DDPG 算法在计算时间上具有显著优势。图 8(b)中，红色曲线表示 Improved-AC-DDPG 算法各迭代轮次的耗电量，橙色曲线表示 APF-DDPG 算法各迭代轮次的耗电量，而蓝色曲线则代表常规 DDPG 算法各迭代轮次的耗电量。从柱状图的对比中可以观察到，Improved-AC-DDPG 算法的耗电量也相对较低。综上所述，Improved-AC-DDPG 算法在计算时间方面，位数相较于传统 DDPG 算法减少了约 71.4%，相较于 APF-DDPG 算法减少了约 51.1%；在耗电量方面，位数相较于传统 DDPG 算法缩减了约 3.3%，相较于 APF-DDPG 算法缩减了约 1.3%。这些数据充分证明，Improved-AC-DDPG 算法不仅在计算时间上表现优越，而且耗电量也相对较低，展现了其高效性和出色的性能。综上所述，Improved-AC-DDPG 算法在任务计算时间和耗电量方面均优于其他两种算法，具有显著的优势和实用性。

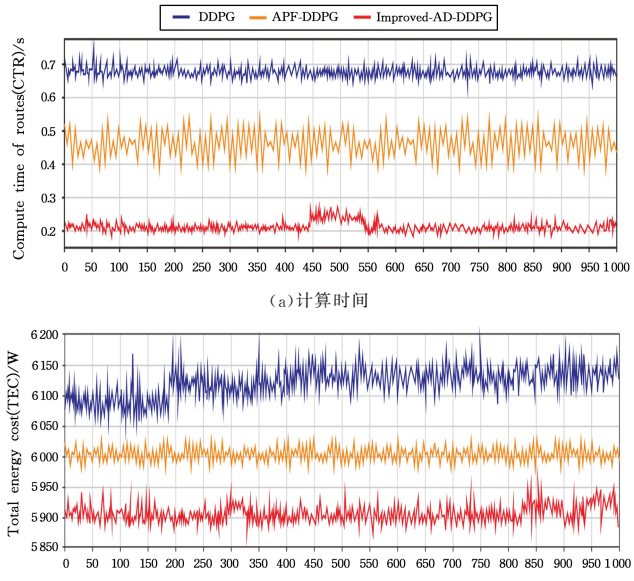


图 8　3 种 DDPG 算法在 AGV 千次任务中的计算与能耗对比

Fig. 8　Calculation and energy consumption comparison of three DDPG algorithms in thousands of tasks of AGV

4.3.4　避障效果对比实验与分析

在评估 AGV 路径规划算法性能时，常规 DDPG 算法、APF-DDPG 算法以及 Improved-AC-DDPG 算法占据着重要地位。为了全面评估各算法的性能，需要深入分析多个关键指标，包括有碰撞静态障碍物记录次数、有碰撞动态障碍物记录次数、有急刹避让记录次数、有平稳避让记录次数，以及累积全程成功执行次数(无碰撞)[23]，如表 4 所列。

表 4 本文方法与常规 DDPG 方法路径规划累积成功次数与碰撞动态障碍物次数对比

Table 4 Comparison of cumulative success times and collision times of dynamic obstacles between the proposed method and conventional DDPG methods

算法	任务次数	有碰撞静态障碍物记录次数	有碰撞动态障碍物记录次数	有急刹避让记录次数	有平稳避让记录次数	累积全程成功执行次数(无碰撞)
DDPG	1000	2	31	310	943	967
APF-DDPG	1000	0	17	256	762	983
Improved-AC-DDPG	1000	0	12	128	1000	505

在对 AGV 路径规划算法的性能评估中，Improved-AC-DDPG 算法展示了显著的优势和独特的优点。首先，它在避免碰撞静态障碍物和动态障碍物方面表现出色，碰撞记录次数分别为 0 和 12，明显优于 DDPG 和 APF-DDPG 算法。此外，Improved-AC-DDPG 算法在急刹避让记录次数上也表现最佳，仅有 128 次，显著少于其他两种算法，表明其路径规划更为平滑和有效。值得注意的是，Improved-AC-DDPG 的平稳避让记录次数达 1000 次，完美地实现了平稳避让，远超 DDPG 和 APF-DDPG 算法。尽管其累积全程成功执行次数(505 次)相对较低(可能反映了某些复杂情况下的稳定性问题)，但其在关键性能指标上的卓越表现，突显了其在路径规划平滑性和动态障碍物处理方面的优越性。

综合来说，Improved-AC-DDPG 算法在避障性能、运行效率、收敛速度和成功执行次数等方面均优于传统 DDPG 算法和 APF-DDPG 算法。该算法在训练步数和累计回报值方面展现出更快的收敛速度和更高的稳定性，在路径规划方面实现了更短的路径长度和更高的稳定性，有效降低了 AGV 的耗电量，并显著提高了避障成功率。

4.3.5 动态参数效果分析

智能体在全局地图上进行路径规划时，依赖于构建的状态向量，该向量融合了 AGV 的实时工作环境数据和静态地图环境数据，即使环境发生动态变化，智能体通过持续学习策略和动态调整任务链，也能够快速适应这些变化，而不会对全局路径规划的稳定性造成负面影响。实际上，这种动态调整参数的能力增强了智能体对环境变化的响应能力，确保了路径规划的实时性和适应性。

当环境中出现新的障碍物时，智能体利用安装的 LiDAR 光学测距扫描设施和立体视觉摄像机实时捕捉环境变化。这些传感器数据被用于即时更新状态向量，从而触发避障机制。智能体的 Actor 网络根据新的状态向量输出动作，而 Critic 网络评估这些动作的预期回报，确保智能体选择出最优避障路径。

智能体能在有限的优化迭代步数中成功避障的关键在于，DDPG 算法的高效性和 Actor-Critic 架构的持续学习策略。通过不断迭代更新 Actor 和 Critic 网络的参数，智能体能够在每次迭代中学习并优化其避障策略。此外，通过设置合适的奖励函数权重系数，在保证安全的前提下，智能体被激励选择更短、更有效的避障路径，从而在有限的迭代次数内实现成功避障。

仿真实验结果表明，引入奖惩改进 Actor-Critic 架构后的 DDPG 算法，在避障性能、运行效率等方面均表现出明显的优势。特别是在碰撞动态障碍物次数上，改进后的算法显著降低了碰撞率，同时在平稳避让记录次数和全程成功执行次数上也优于其他方法，证明了算法在动态环境中的有效性和鲁棒性。

结束语 本研究针对自动引导车(AGV)在复杂动态环境中的路径规划问题，提出了一种创新的 Improved-AC-DDPG 算法。该算法基于深度确定性策略梯度(DDPG)并结合动态任务链，旨在提高 AGV 的运行效率和安全性。通过深入分析现有深度强化学习方法的局限性，设计了一种新的算法框架，以增强样本学习效率、优化多 AGV 间的协同工作[24]，并实时更新环境模型以适应动态变化。

Improved-AC-DDPG 算法的核心在于，通过构建状态向量来有效表示 AGV 的当前状态与环境信息，利用改进的 DDPG 策略进行实时路径规划，动态生成任务链以减少 AGV 间的冲突，监测 AGV 的运动状态并预测调整避障路径，通过持续学习策略优化模型参数。算法流程设计包括状态向量构建、路径预规划、动态任务链生成、运动状态监测模型建立和持续学习策略执行等关键步骤。

仿真实验结果表明，Improved-AC-DDPG 算法在避障性能、运行效率、收敛速度和成功执行次数等方面均优于传统 DDPG 算法和 APF-DDPG 算法。该算法在训练步数和累计回报值方面展现出了更快的收敛速度和更高的稳定性，在路径规划方面实现了更短的路径长度和更高的稳定性，有效降低了 AGV 的耗电量，显著提高了避障成功率。

本文的研究成果为多 AGV 路径规划提供了一种高效、稳定且适应性强的解决方案，具有重要的理论意义和实际应用价值。未来的工作将集中于进一步提升算法性能，探索其在更复杂场景下的应用，并优化算法以适应不断变化的物流仓储需求，推动多 AGV 避障技术向更高水平发展[25]。

参 考 文 献

[1] ZHAO X J, YE H, JIA W, et al. A review of AGV path planning and obstacle avoidance algorithms [J]. Microcomputer Systems, 2024, 45(3): 529-541.

[2] AIZAT M, QISTINA N, RAHIMAN W. A Comprehensive Review of Recent Advances in Automated Guided Vehicle Technologies: Dynamic Obstacle Avoidance in Complex Environment Toward Autonomous Capability [J/OL]. https://www.researchgate.net/publication/376154191_A_Comprehensive_Review_of_Recent_Advances_in_Automated_Guided_Vehicle_Technologies_Dynamic_Obstacle_Avoidance_in_Complex_Environment_Toward_Autonomous_Capability.

[3] LIN Y, HU G, WANG L, et al. A multi-AGV routing planning method based on deep reinforcement learning and recurrent neural network [J]. IEEE/CAA Journal of Automatica Sinica, 2023, 11(7): 1720-1722.

[4] YE X,DENG Z,SHI Y,et al. Toward energy-efficient routing of multiple AGVs with multi-agent reinforcement learning [J]. Sensors,2023,23(12):5615.

[5] GAO Y,CHEN C H,CHANG D. A Machine Learning-Based Approach for Multi-AGV Dispatching at Automated Container Terminals [J]. Journal of Marine Science and Engineering,2023,11(7):1407.

[6] CHEN Y,SCHOMAKER L,CRUZ F. Boosting Reinforcement Learning Algorithms in Continuous Robotic Reaching Tasks using Adaptive Potential Functions [J]. arXiv: 2402. 04581,2024.

[7] BHADAURIA S,PLAKU K,DESHPANDE Y,et al. Evaluation of NR-Sidelink for Cooperative Industrial AGVs [J]. arXiv: 2309. 02949,2023.

[8] LILLICRAP T P,HUNT J J,PRITZEL A,et al. Continuous control with deep reinforcement learning [J]. arXiv: 1509. 02971,2015.

[9] DOBREV D. Formal Definition of Artificial Intelligence and an Algorithm Which Satisfies This Definition [C]// XII-th International Conference. 2006.

[10] SCHAUL T, QUAN J, ANTONOGLOU I,et al. Prioritized experience replay [J]. arXiv:1511. 05952,2015.

[11] KALIDINDI H T,CROSS K P,LILLICRAP T,P et al. Rotational dynamics in motor cortex are consistent with a feedback controller [J]. Elife,2021,10:e67256.

[12] ZHU H,XIE Y,ZHENG S. A double Actor-Critic learning system embedding improved Monte Carlo tree search [J]. Neural Computing and Applications,2024,36:8485-8550.

[13] LI C. Research on Multi-AGV Scheduling System of Intelligent Warehouse Based on Dynamic Task Chain [D]. Hangzhou:Zhejiang University,2023.

[14] YAN J D. Modeling and deployment optimization of "low,slow and small" UAV bee colony counterwarfare mission chain [D]. Nanjing:National University of Defense Technology,2021.

[15] HU B,TIAN X L,YANG C,et al. A Dynamic Resource Chain Task Unloading Method Based on Improved Greedy Algorithm [J]. Journal of Physics: Conference Series, 2021, 1883(1): 012021.

[16] XIONG J T,LI Z X,CHEN S M,et al. Obstacle avoidance planning of virtual robot picking path based on deep reinforcement learning [J]. Journal of Agricultural Machinery,2020,51(S2): 1-10.

[17] YE H,ZHANG X,FAN F. A fast mounting structure of multilayer pallet and AGV trolley:CN220244403[P]. 2023-12-26.

[18] GUO S,ZHANG X,ZHENG Y,et al. An autonomous path planning model for unmanned ships based on deep reinforcement learning [J]. Sensors,2020,20(2):426.

[19] RUPAPARA V,RAJEST S S,RAJAN R,et al. A dynamic perceptual detector module-related telemonitoring for the intertubes of health services [M]//Artificial Intelligence for Smart Healthcare. Cham:Springer International Publishing,2023:245-274.

[20] CHEN X,LIU S,ZHAO J,et al. Autonomous port management based AGV path planning and optimization via an ensemble reinforcement learning framework [J]. Ocean & Coastal Management,2024,251:107087.

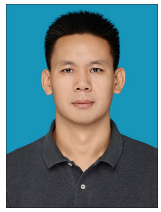
[21] GONG L,HUANG Z,XIANG X,et al. Real-time AGV scheduling optimisation method with deep reinforcement learning for energy-efficiency in the container terminal yard [J]. International Journal of Production Research,2024,62(21):7722-7742.

[22] ISLAM F,BALL J E,GOODIN C T. Enhancing Longitudinal Velocity Control With Attention Mechanism-Based Deep Deterministic Policy Gradient(DDPG) for Safety and Comfort [J]. IEEE Access,2024,12:30765-30780.

[23] HAZARIKA B,SAIKIA P,SINGH K,et al. Enhancing Vehicular Networks With Hierarchical O-RAN Slicing and Federated DRL [J]. IEEE Transactions on Green Communications and Networking,2024,8(3):1099-1117.

[24] LI H. Research on Multi-task Allocation and Path Planning of Multi-AGV [D]. Nanjing:Nanjing University of Posts and Telecommunications,2019.

[25] TIAN S H,SHEN Y F,OU L Y,et al. AGV Task Assignment Optimization of Automatic Picking System Considering Load Balancing [J]. Computer Application Research, 2024, 41(8): 2366-2373.



ZHAO Xuejian, born in 1982, Ph.D, associate professor, is a member of CCF (No. 88401M). His main research interests include data mining and wireless sensor networks.

SUN Zhixin, born in 1964, Ph.D, professor, doctoral supervisor. His main research interests include the theory and technology of network communication, computer network and security.

(责任编辑:何杨)