智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于双阶段策略熵自适应Q-learning的智能配送车路径规划方法研究

熊刚, 蔡凯琪, 陈世超, 朱凤华, 郭超, 高超, 陈德旺

引用本文:
熊刚, 蔡凯琪, 陈世超, 等. 基于双阶段策略熵自适应Q-learning的智能配送车路径规划方法研究[J]. 智能系统学报, 2026, 21(5): 1180-1193.
XIONG Gang, CAI Kaiqi, CHEN Shichao, et al. Intelligent delivery vehicle path planning based on dual-phase policy entropy adaptive Q-learning algorithm[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1180-1193.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202511006

DOI: 10.11992/tis.202511006
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260706.1750.006

基于双阶段策略熵自适应 Q-learning 的智能配送车路径规划方法研究

熊刚[1,2], 蔡凯琪[1], 陈世超[2], 朱凤华[2], 郭超[2], 高超[3], 陈德旺[4]
(1. 福建理工大学 人工智能与交通工程学院, 福建 福州 350000; 2. 中国科学院自动化研究所 多模态人工智能系统全国重点实验室, 北京 100190; 3. 北京工商大学 计算机与人工智能学院, 北京 100048; 4. 福建理工大学 计算机与数据科学学院, 福建 福州 350000)

摘 要: 针对传统 Q-learning 算法在智能配送车路径规划中存在收敛速度慢、路径质量不稳定以及动态环境适应性弱的问题, 提出一种基于双阶段策略熵自适应 Q-learning 的路径规划算法 (dual-phase policy entropy adaptive Q-learning, DPEA-QL)。该算法通过引入策略熵驱动的自适应学习率优化, 采用 softmax 策略, 实现在训练初期充分探索和后期高效利用, 提升算法的学习效率和收敛速度; 构建双阶段熵调控机制, 利用 sigmoid 函数平滑过渡策略熵权重, 有效提升策略平稳性和路径最优性; 重塑基于曼哈顿距离的奖励函数, 使奖励信号随配送车与目标点的距离动态变化, 增强目标导向避免陷入局部收敛。在单一静态障碍、混合静态障碍和动态障碍 3 类仿真地图上的实验结果表明: DPEA-QL 在 3 类地图中均能规划出一条最优路径, 最短路径步数为 58 步, 最短路径达成率较传统 Q-learning 算法提高 24.28 百分点, 在收敛速度、路径质量与鲁棒性上明显优于传统 Q-learning、Sarsa 及 SA-QL 算法。

关键词: 路径规划; Q-learning; 双阶段策略熵; 智能配送车; 强化学习; 动态避障; 收敛速度; 路径质量; 鲁棒性
中图分类号: TP181; U495 **文献标志码**: A **文章编号**: 1673−4785(2026)05−1180−14

中文引用格式: 熊刚, 蔡凯琪, 陈世超, 等. 基于双阶段策略熵自适应 Q-learning 的智能配送车路径规划方法研究 [J]. 智能系统学报, 2026, 21(5): 1180–1193.
英文引用格式: XIONG Gang, CAI Kaiqi, CHEN Shichao, et al. Intelligent delivery vehicle path planning based on dual-phase policy entropy adaptive Q-learning algorithm[J]. CAAI transactions on intelligent systems, 2026, 21(5): 1180–1193.

Intelligent delivery vehicle path planning based on dual-phase policy entropy adaptive Q-learning algorithm

XIONG Gang[1,2], CAI Kaiqi[1], CHEN Shichao[2], ZHU Fenghua[2],
GUO Chao[2], GAO Chao[3], CHEN Dewang[4]
(1. School of Artificial Intelligence and Transportation Engineering, FuJian University of Technology, Fuzhou 350000, China; 2. The State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China; 3. School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China; 4. School of Computing and Data Science, FuJian University of Technology, Fuzhou 350000, China)

Abstract: Traditional Q-learning algorithms for intelligent delivery vehicle path planning often suffer from slow convergence, unstable path quality, and poor adaptability in dynamic environments. To address these issues, this paper proposes a dual-phase policy entropy adaptive Q-learning (DPEA-QL) algorithm. First, a policy entropy-driven adaptive learning rate mechanism combined with a softmax policy is introduced to balance sufficient early-stage exploration and efficient late-stage exploitation, thereby accelerating learning efficiency. Second, a dual-phase entropy regulation strategy is constructed, utilizing a sigmoid function to smoothly adjust the entropy weight, which effectively enhances policy stability and global optimality. Furthermore, a Manhattan distance-guided reward function is designed to dynamically reshape the reward signals based on the vehicle-to-goal distance, strengthening target orientation and preventing the algorithm from falling into local optima. Simulation experiments were conducted across three types of environments: single static, hybrid static, and dynamic obstacle maps. The experimental results demonstrate that DPEA-QL consistently plans the optimal path across three types of maps, with a shortest-path length of 58 steps. Compared with traditional Q-learning, the optimal path achievement rate is improved by 24.28 percentage points. Overall, the proposed algorithm significantly outperforms traditional Q-learning, Sarsa, and SA-QL algorithms in terms of convergence speed, path quality, and robustness.

Keywords: path planning; Q-learning; dual-phase policy entropy; intelligent delivery vehicle; reinforcement learning; dynamic obstacle avoidance; convergence speed; path quality; robustness

随着自动驾驶技术的快速发展, 无人车在城市清扫、公交出行、侦察巡逻等城市管理、公共服务和国防安全领域展现出显著的应用价值, 成为现代智能交通体系的重要组成部分[1-3]。尤其是

收稿日期: 2025−11−04. **网络出版日期**: 2026−07−07.
基金项目: 福建省闽江学者讲座教授人才计划项目 (GY-Z24014); 国家自然科学基金项目 (62461160259, U24A20277, 62306319); 福建省第三批创新之星人才计划项目 (003002); 北京市自然科学基金项目 (L241016).
通信作者: 陈德旺. E-mail: dwchen@fjut.edu.cn.

在物流末端配送环节, 利用智能配送车替代传统的末端物流配送方式成为重要的发展方向, 而路径规划正是智能配送车高效应对末端配送难题的关键环节。

然而, 现有的一些传统路径规划算法, 例如 A*算法[4]、动态窗口法 (dynamic window approach, DWA)[5] 与快速扩展随机树 (rapidly-exploring random tree, RRT) 算法[6], 尽管在静态环境下具备较好性能, 但在动态复杂环境中适应性和实时避障性能不足, 其算力难以应对复杂道路条件与多变交通状况。而蚁群算法[7]、遗传算法[8] 等仿生智能算法虽然具备一定的全局优化能力, 但通常依赖环境先验模型且易陷入局部最优, 难以规划实时高效的最优路径。因此, 强化学习算法逐渐成为路径规划研究的主流方向, 代表算法主要包括 Q-learning[9]、深度 Q 网络 (deep Q-network, DQN)[10]、近端策略优化 (proximal policy optimization, PPO)[11]、深度确定性策略梯度 (deep deterministic policy gradient, DDPG)[12] 与柔性行动者−评论家 (soft actor-critic, SAC)[13] 等。

近年来, 数据驱动的路径规划方法取得了重大进展。针对复杂序列决策问题, Chen 等[14] 提出了 Decision Transformer, 将强化学习中的路径规划建模为序列预测问题, 在大规模离线数据集上展现了优于传统时序差分方法的泛化能力; 针对自动驾驶中的多模态感知与规划整合难题, Li 等[15] 提出了 Hydra-MDP(Markov decision process) 框架, 通过多目标知识蒸馏策略, 在复杂动态环境中展现了接近人类驾驶专家的决策水平; 此外, 针对端到端自动驾驶的语义推理瓶颈, Zhou 等[16] 提出了基于 Transformer 架构的视觉−语言−动作 (vision-language-action, VLA) 统一生成模型, 显著提升了智能体在非结构化环境下的场景理解与避障效率。然而, 上述基于深度学习和 Transformer 架构的前沿方法虽然在处理高维感知输入和连续控制任务中表现良好, 但其本质是基于神经网络的近似函数逼近, 存在“黑盒”不可解释性及训练不稳定的问题, 模型参数量巨大, 对计算资源和存储空间要求极高, 往往需要海量离线数据进行预训练, 且模型推理延迟难以满足低成本末端配送的实时响应需求。相比之下, Q-learning 算法作为经典的基于值迭代的强化学习算法之一, 其理论完备的收敛证明、可解释的 Q 表决策机制以及极低的计算开销, 在资源受限的工程场景下具有强大优势。

然而, 传统 Q-learning 算法 (Q-learning, QL) 在大规模或复杂动态环境中仍存在收敛速度慢、探索效率低和容易陷入局部最优等不足[17]。基于此, 田晓航等[18] 根据蚁群信息素浓度的变化动态调整探索因子, 引导智能体从广泛探索转向精细利用, 从而有效减少迭代后期重复探索, 该方法加速了算法收敛进程, 但在训练早期可能导致丢失最优路径, 无法保证寻得最优解。王小康等[19] 通过借鉴模拟退火原理动态调整探索因子, 引入欧氏距离奖励, 在静态环境中有效提高了算法规划效率, 但缺乏动态环境适应性。赵也践等[20] 通过引入以 softmax 函数为主体的动作选择策略, 避免在后期仍出现选择次优动作的情况。文献 [21] 在标准 softmax 策略的基础上提出了自适应的 softmax 策略, 即通过时序差分误差动态调整探索率, 并在探索阶段采用 softmax 分布进行动作选择, 有效提升了奖励收敛速度与动作选择的整体最优性。文献 [22] 使用历史经验中的成功率作为辅助奖励, 提高了稀疏奖励环境中的探索与利用的平衡性, 但其状态定义依赖于任务结构, 导致算法需要大量先验知识定义成功率标签, 泛化能力受限。任伟等[23] 通过动态衰减搜索因子平衡探索与利用之间的关系, 并采用当前位置与目标位置欧氏距离的倒数作为初始 Q 值, 引导智能体在早期阶段更倾向于朝向目标移动, 显著加快了收敛速度, 同时引入自校正估计器, 结合不同时间步的 Q 值估计以修正最大化偏差, 提升了算法的稳定性和学习效率, 然而该改进方法是基于静态环境假设, 难以有效应对动态变化。

针对收敛速度、路径质量、平衡探索与利用的问题, 本文在 QL 算法的基础上, 提出一种基于双阶段策略熵自适应 Q-learning(dual-phase policy entropy adaptive Q-learning, DPEA-QL) 的路径规划算法, 并在静态、混合及动态等多类障碍环境下的仿真实验中验证了该算法的可行性和有效性。

1　Q-learning 算法

Q-learning 算法是一种无模型的基于值迭代的离线策略强化学习算法, 主要优点在于不依赖于环境的精确先验模型, 其核心思想是通过智能体与环境持续交互, 迭代更新状态−动作价值函数 (Q 函数), 从而逐步学习并逼近最优策略[24]。Q-learning 的基本框架包含 3 个核心部分。其一是状态与动作空间。智能体与环境基于马尔可夫决策过程 (Markov decision process, MDP) 进行交互[25]。在 MDP 框架下的智能配送场景中, 将智能配送车视为智能体, 配送区域划分为离散的栅格地图。定义状态空间为 S, 智能配送车在 t 时刻的当前状态记为 s_t; 定义动作空间为 A, 包含配送车辆可执行的上、下、左、右 4 个移动方向, 当前时刻执行的动作记为 a_t。其二是策略选择机制, 用于

指导配送车在不同状态下选择动作，实现探索与利用的平衡。其三是价值迭代更新规则，用于学习 Q 函数，当智能体在状态 s_t 执行动作 a_t 后，环境转移至下一状态 s_{t+1} 并反馈即时奖励 r_t，其更新公式为

$$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left[r_t+\gamma\max_{a'}Q(s_{t+1},a')-Q(s_t,a_t)\right] \tag{1}$$

式中：$Q(s_t,a_t)$ 表示在状态 s_t 下执行动作 a_t 的期望价值；$\alpha\in(0,1]$ 为学习率，用于控制新旧 Q 值的更新权重；$\gamma\in[0,1)$ 为折扣率，用于衡量未来奖励对当前决策的影响比重；r_t 为环境反馈的即时奖励；$\max\limits_{a'}Q(s_{t+1},a')$ 表示在下一状态 s_{t+1} 中所有可能动作中的最大 Q 值。

为实现智能配送车路径规划任务，Q-learning 通过与栅格环境的持续交互，自主学习最优路径与避障决策。本文构建了基于传统 Q-learning 算法的路径规划系统框架，如图 1 所示。框架主要包括环境建模与感知模块、奖励函数模块、智能体决策模块与路径执行模块等 4 个核心组成部分，共同实现智能配送车与环境之间的闭环交互。环境建模与感知模块将城市道路及障碍物抽象为栅格状态空间，获取智能配送车的当前状态 s_t；智能体决策模块依据 Q-learning 迭代更新 Q 表，选择最优动作 a_t；路径执行模块将最优动作映射为实际移动；奖励函数模块通过正负奖励 r_t 引导配送车躲避障碍物并接近目标。在训练过程中，配送车通过“感知状态−决策动作−执行移动−获取奖励与新状态”的闭环流程循环迭代，直至智能配送车找到一条能够避开障碍且成功到达目标的路径。

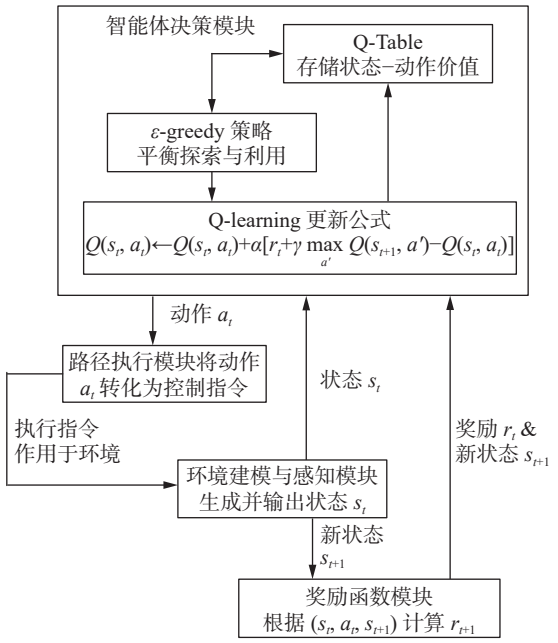


图 1　智能配送车路径规划框架

Fig. 1　Path planning framework for intelligent delivery vehicle

2　双阶段策略熵自适应 Q-learning 算法

2.1　策略熵驱动的自适应优化

在传统 Q-learning 算法中，学习率 α 通常设置为固定值，使得在训练初期智能配送车在未知环境中的探索不足，难以得到最优路径；而在训练后期，固定较大的学习率可能引起 Q 值更新震荡，使得车辆在训练后期错过最优解，从而影响算法收敛[26]。此外，传统 Q 学习进行动作选择时常采用 ε-greedy 策略，以 ε 的概率随机选择动作，$1-\varepsilon$ 的概率选择当前 Q 值最高的动作。该策略在复杂环境下容易出现两类问题：一是探索不足，智能体可能过早收敛到局部最优策略；二是过度随机，使得训练后期策略仍存在较高的不确定性，导致算法学习效率和路径规划性能降低[27]。

在强化学习中，策略的随机性直接影响探索的效率与收敛的稳定性，因此，针对上述问题，本文引入策略熵 (policy entropy)，用于量化智能配送车在当前状态下动作选择的不确定性[28]：

$$H(\pi_t)=-\sum_{a\in A}\pi(a|s_t)\ln\pi(a|s_t) \tag{2}$$

式中：$\pi(a|s_t)$ 表示智能配送车在状态 s_t 下选择动作 a 的概率。本研究动作空间包含 4 个方向 (即 $|A|=4$)，因此策略熵的理论取值范围严格界定在 $[0,\ln 4]$ 之间。策略熵 $H(\pi_t)$ 越高，表示动作越随机，有利于探索未知路径；策略熵 $H(\pi_t)$ 越低，表示动作选择越集中，有利于策略平稳。基于策略熵，本研究设计了自适应学习率衰减机制，具体公式为

$$\lambda_k=\lambda_{\min}+\omega_k(\lambda_{\max}-\lambda_{\min})\frac{H(\pi_t)}{H_{\max}} \tag{3}$$

$$\alpha_t=\max(\alpha_{\min},\alpha_0\exp(-\lambda_k\cdot k)) \tag{4}$$

式中：λ_k 为第 k 回合依据当前策略熵实时计算的动态衰减因子，$H_{\max}=\ln 4$ 为最大策略熵，ω_k 为随训练回合数 k 变化的动态熵权重，α_t 为当前时间步 t 的实时学习率，$\alpha_{\min}$ 为最小学习率界限，α_0 为初始最大学习率。该机制综合考虑训练进度与当前策略熵，使智能配送车在初期保持较高学习率以增强探索，后期逐步降低学习率以提升策略稳定性。

在动作选择方面，为克服贪婪策略的局限性，本文采用 softmax(Boltzmann) 策略[29]：

$$\pi(a_t|s_t)=\frac{\exp(Q(s_t,a_t)/\tau_k)}{\sum\limits_{a'\in\mathcal{A}}\exp(Q(s_t,a')/\tau_k)} \tag{5}$$

式中：τ_k 为第 k 回合的温度参数，用于全局调控动作选择的随机性。结合策略熵的自适应优化机制

执行步骤如下。

步骤 1　策略计算与熵评估。智能配送车在状态s_t计算 softmax 概率分布$\pi(a_t|s_t)$及策略熵$H(\pi_t)$。

步骤 2　学习率调节。根据式 (2) 计算当前回合内的动态衰减因子λ_k, 并根据式 (3) 动态更新当前时间步学习率α_t。

步骤 3　动作选择与执行。根据 softmax 分布采样并执行动作a_t。

步骤 4　更新 Q 值。获取总奖励R_t和新状态s_{t+1}, 使用自适应学习率α_t根据改进后的公式更新 Q 表:

$$\begin{aligned} &Q(s_t,a_t) \leftarrow Q(s_t,a_t)+ \\ &\alpha_t\left[R_t+\gamma\max_{a'}Q(s_{t+1},a')-Q(s_t,a_t)\right] \end{aligned} \tag{6}$$

步骤 5　重复迭代, 直至训练结束。

通过上述优化步骤, DPEA-QL 算法在保证策略稳定收敛的同时增强探索能力, 有效提升智能配送车在复杂环境下的路径规划性能。

2.2　双阶段熵调控机制

在 Q-learning 算法中, 传统的探索策略多通过线性衰减其参数来控制探索程度, 这种静态设计难以兼顾训练早期充分探索与后期稳定利用的动态需求。最大熵强化学习框架的理论证明, 最优策略服从由温度参数决定的 Boltzmann 分布形式[30]。在本文的 softmax 策略中, 直接线性调节温度参数τ_k可能导致策略波动剧烈或收敛不稳定。轨迹平滑方法通过在运动规划中实现连续过渡与稳定性, 验证了平滑调节在动态过程控制中的有效性[31]。

借鉴轨迹平滑控制的思想, 本文提出双阶段熵调控机制, 引入 sigmoid 平滑函数构建全局统一调度因子$\rho(k)$, 具体公式为

$$\rho(k)=\frac{1}{1+\exp\left(c\cdot\left(\dfrac{k}{K}-0.5\right)\right)} \tag{7}$$

式中: k为当前训练回合数; K为最大总训练回合数; c为斜率参数, 用于控制过渡曲线的陡峭程度; 常数 0.5 为设定的过渡中点回合比例。该函数的输出值域严格限制在$(0,1)$区间。

为实现对动作选择随机性与参数更新步长的同步调控, 本文利用全局调度因子$\rho(k)$同时控制温度参数τ_k与熵权重ω_k。其中, τ_k用于调节动作选择分布的随机性, ω_k用于调节策略熵对学习率衰减因子的影响。具体公式为

$$\tau_k=\tau_{\min}+\rho(k)(\tau_{\max}-\tau_{\min}) \tag{8}$$

$$\omega_k=\rho(k) \tag{9}$$

式中: $\tau_{\max}$为初始高温度参数, 用于强探索; $\tau_{\min}$为最小温度参数, 用于强利用。随着训练回合k增加, $\rho(k)$平滑减小, τ_k与ω_k随之同步衰减, 使算法由前期较充分探索逐步过渡到后期稳定利用, 从而避免传统硬切换造成的策略震荡。

2.3　奖励函数重塑

奖励函数是影响 Q-learning 算法收敛速度和鲁棒性的重要因素。在现有研究中, 传统 Q-learning 算法多采用稀疏奖励机制, 即仅在智能体到达目标点时赋予较大正奖励, 碰撞障碍物或越界时赋予负奖励, 而对正常移动过程则不设置奖励, 导致智能体训练初期出现盲目探索、收敛速度慢等问题。因此, 本文设计基于曼哈顿距离引导的奖励函数。曼哈顿距离D量化了当前位置坐标(x,y)与目标点坐标$(x_{\text{goal}},y_{\text{goal}})$在栅格环境下的最短网格步数:

$$D(s_t,P_{\text{goal}})=|x-x_{\text{goal}}|+|y-y_{\text{goal}}| \tag{10}$$

该距离在不增加环境复杂度的前提下为智能配送车提供了连续的空间引导信号。重塑后的综合奖励函数R_t定义为

$$R_t=r_t+r_{\text{M}} \tag{11}$$

式中: r_t为环境基础奖励, 成功到达目标$r_t=+100$, 越界或碰撞障碍物$r_t=-10$, 基础步进惩罚$r_t=-1$。r_{M}为曼哈顿距离引导奖励, 具体表示为

$$r_{\text{M}}=(D_{\text{old}}-D_{\text{new}})+\eta_k\frac{1}{D_{\text{new}}+\sigma} \tag{12}$$

式中: σ为防止分母为零的微小常数; D_{old}与D_{new}分别为执行动作前后的曼哈顿距离; η_k为动态引导系数。为防止策略后期过度依赖引导信号而干扰全局最优, η_k随训练进度动态衰减:

$$\eta_k=\eta_{\text{base}}\left(1-\frac{k}{K}\right) \tag{13}$$

式中: η_{base}为初始引导系数。在训练初期, 曼哈顿距离项提供明显引导; 训练后期随着η_k的衰减, 奖励函数逐渐退化为基础奖励, 保证算法自主稳定收敛。该设计能够有效提高收敛速度、降低训练成本, 提升配送车在复杂环境下的路径规划性能。

2.4　算法整体框架与流程

基于上述改进点, 本文提出的 DPEA-QL 算法以动态平衡探索与利用为核心, 形成了从环境感知、策略退火、自适应学习率到势能奖励引导的封闭系统。图 2 给出了基于 DPEA-QL 算法的智能配送车路径规划整体训练框架。

在该框架中, 智能体与路径规划环境模拟器构成标准马尔可夫决策过程, 详细伪代码见算法 1。

算法 1　双阶段策略熵自适应 Q-learning 算法

输入　总训练回合数K, 单步限制N, 初始学习率α_0, 熵权边界$\lambda_{\min}$、$\lambda_{\max}$等

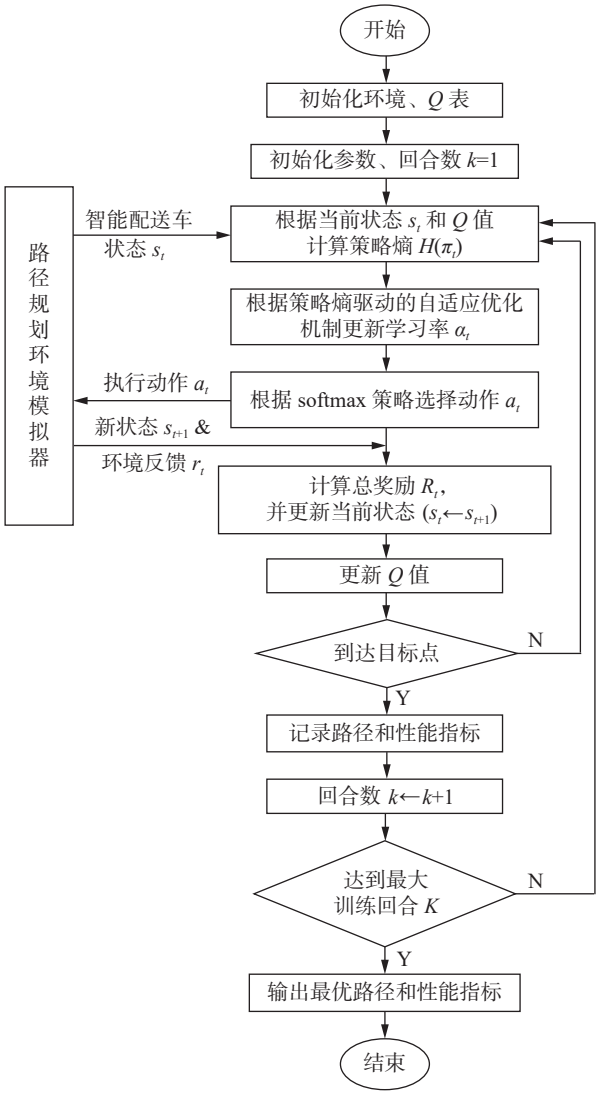


图 2　基于 DPEA-QL 算法的智能配送车路径规划训练框架

Fig. 2　Training framework for intelligent delivery vehicle path planning based on the DPEA-QL algorithm

初始化　动作价值函数 $Q(s,a)=0$

输出　最优 Q 表和规划策略

1)For 回合 $k=1$ to K do

2) 初始化环境，获取当前状态 s_t

3) 根据式 (6) 计算全局调度因子 $\rho(k)$

4) 根据式 (7) 计算当前回合温度参数 τ_k

5) 根据式 (8) 计算当前回合熵权重 ω_k

6) For 单步 $t=1$ to N do

7) 根据式 (4) 计算策略分布 $\pi(a_t|s_t)$，并采样执行动作 a_t

8) 获取新状态 s_{t+1} 及基础反馈 r_t

9) 计算动作执行前后的曼哈顿距离 D_{old} 与 D_{new}

10) 根据式 (12) 计算动态引导系数 η_k

11) 根据式 (11) 计算曼哈顿距离引导奖励 r_{M}

12) 根据式 (10) 计算重塑总奖励 R_t

13) 根据式 (1) 评估当前状态的策略熵 $H(\pi_t)$

14) 根据式 (2) 计算动态衰减因子 λ_k

15) 根据式 (3) 计算当前时间步学习率 α_t

16) 根据式 (5) 更新 Q 值

17) 状态转移：$s_t \leftarrow s_{t+1}$

18) If 到达目标点 then Break

19) End For

20)End For

具体的单步执行流程包括以下 5 个核心阶段。

第一，状态感知与全局调度。在每个训练回合开始时，智能体获取环境反馈的当前状态 s_t。同时，算法根据当前训练回合由式 (6) 计算全局调度因子 $\rho(k)$，并进一步根据式 (7)、(8) 分别计算当前回合的温度参数 τ_k 与熵权重 ω_k，从而实现动作选择随机性与学习率调节权重的同步控制。

第二，策略退火与动作输出。基于当前状态 s_t 下的动作价值 Q 分布与当前温度 τ_k，算法利用式 (4) 计算出 softmax 动作选择概率分布 $\pi(a_t|s_t)$，并采样得到实际动作 a_t，随后将其输出至环境模拟器。

第三，环境基础反馈交互。环境模拟器接收并执行动作 a_t 后，向智能体返回下一时刻的新状态 s_{t+1} 以及基础的环境惩罚或目标奖励 r_t。

第四，奖励动态重塑。智能体获取环境反馈后，根据式 (9) 计算动作执行前后的曼哈顿距离 D_{old} 与 D_{new}，并结合式 (11) 得到曼哈顿距离引导奖励 r_{M}，最终根据式 (10) 获得重塑总奖励 R_t，为智能体提供全局方向引导。

第五，策略熵驱动的 Q 值更新。在获得重塑奖励 R_t 后，算法利用式 (1) 实时评估当前状态的策略熵 $H(\pi_t)$，并结合当前回合熵权重 ω_k，根据式 (2)、(3) 计算当前时间步 t 的自适应学习率 α_t。最终，利用 α_t 按照式 (5) 完成 Q 值更新，并将状态转移至 s_{t+1}，进入下一次循环。

随着训练回合的持续推进，上述闭环流程在时间轴上呈现出由高熵探索向低熵利用的自适应退火特征。在训练前期，高温与高熵权重促使智能体广泛试错并保持较大的 Q 值更新幅度，以避免陷入局部极小值；在训练中后期，随着 $\rho(k)$ 触发双阶段切换，策略迅速冷却收敛，学习率趋于稳健的下界。各改进模块在时序上的协同配合，有效解决传统 Q-learning 在复杂约束环境下收敛缓慢且剧烈震荡的缺陷，实现了高效、安全的智能配送车路径寻优。

3　实验与结果

3.1　实验条件设定

为验证本文所提出的双阶段策略熵自适应 Q-learning 算法的有效性和可行性，选择在 Pycharm

2024.1.4 集成开发环境下进行仿真实验, 其操作系统环境为 Windows11 x64, 使用软件工具包版本为 Python 3.12。针对在复杂城市交通中智能配送车对前方局部视野进行高频动态避障的真实工程需求, 实验采用 30×30 网格地图作为城市道路环境的抽象模型, 通过不同障碍物配置场景模拟智能配送车在城市复杂环境中的路径规划任务。

3.2　实验参数设置

为确保实验的可比性, 传统 QL 算法、Sarsa 算法、基于模拟退火的 QL 算法以及本文提出的 DPEA-QL 算法均在相同实验条件下运行, 最终参数设置如表 1 所示。

表 1　实验参数设置
Table 1　Experimental parameter settings

参数	值
最大训练回合数K	5 000
最大单步限制N	500
折扣因子γ	0.95
探索率ε	0.1
QL固定学习率α	0.1
SA-QL初始温度T_0	10
SA-QL最低温度$T_{\min}$	0.1
初始引导系数η_{base}	5
斜率参数c	1.0
熵权上界$\lambda_{\max}$	0.001
熵权下界$\lambda_{\min}$	0.000 1
DPEA-QL初始学习率α_0	0.9
DPEA-QL温度上界$\tau_{\max}$	10
DPEA-QL温度下界$\tau_{\min}$	0.1

3.3　仿真实验结果与分析

为全面评估 DPEA-QL 算法在智能配送车路径规划中的性能, 本文在单一静态障碍物地图、混合静态障碍物地图和动态障碍物地图上进行仿真实验。实验采用平均奖励、平均步数和最短路径累计次数作为主要评价指标, 并选取 QL、Sarsa 和 SA-QL 作为对比算法。QL 与 Sarsa 均属于基于时序差分的值迭代算法, 但二者在策略更新上存在显著差异: QL 采用离策略 (off-policy), Sarsa 采用在策略 (on-policy), 而 SA-QL 引入了温度退火机制以改善后期收敛。通过多维度的基准对比, 针对强化学习算法早期随机探索阶段存在震荡问题, 本文对所有奖励与步数收敛曲线均采用了滑动平均滤波进行平滑处理 (滑动窗口 w=50), 旨在充分验证 DPEA-QL 算法中双阶段熵控机制的优越性。

3.3.1　单一静态障碍物地图 Ⅰ 搭建与结果分析

为检验 DPEA-QL 改进算法的可行性, 在单一静态障碍物环境中与 QL、Sarsa 及 SA-QL 算法对比分析, 评估其路径规划性能。本文首先搭建仅包含单一静态障碍物的基础地图环境 Ⅰ, 地图整体采用 30×30 的规则栅格构建, 如图 3 所示, 该场景中黑色区域为静态障碍物, 代表固定建筑物或无法通行的道路, 白色区域为可通行的道路。起始点以蓝色圆圈表示, 目标点以红色圆圈表示, 智能配送车从起点出发, 沿上下左右 4 个方向单步移动, 最终到达目标点。

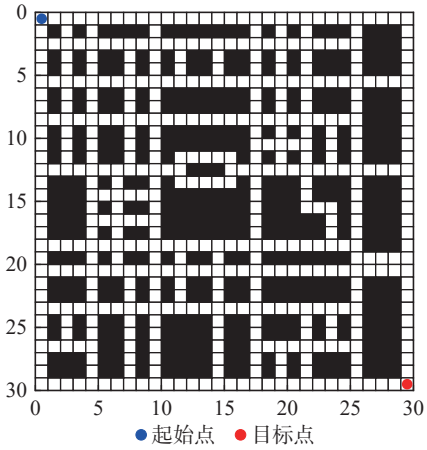
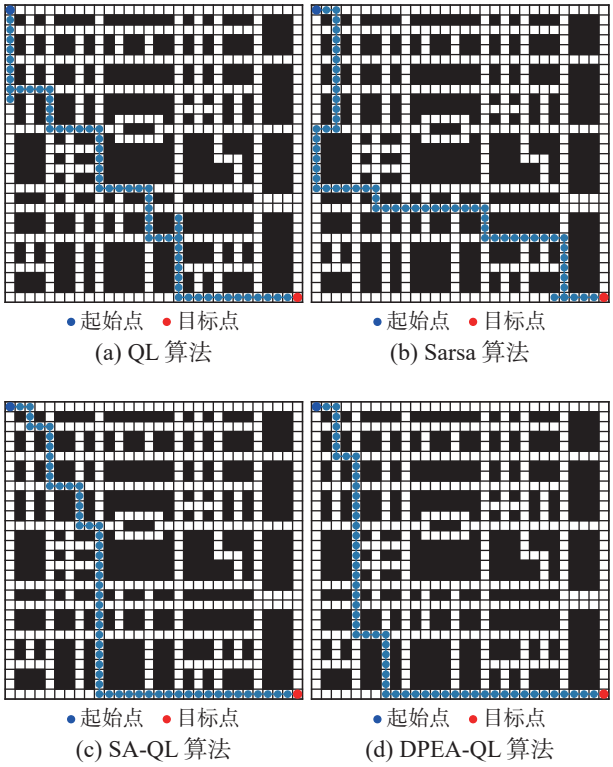


图 3　单一静态障碍物地图 Ⅰ
Fig. 3　Single static obstacle map Ⅰ

图 4 是 4 种算法在地图 Ⅰ 的路径规划结果。如图 4(a) 所示, QL 算法在初期探索阶段存在较强的随机性和明显冗余路径, 整体路径较长, 拐点个数共计 11 个, 路径步数为 64 步; 如图 4(b) 所示, Sarsa 算法由于在更新过程中考虑了动作依赖性, 路径较为平滑, 但在部分区域仍出现非最优绕行, 拐点个数为 11 个, 路径步数为 64 步; 如图 4(c) 所示, SA-QL 算法虽然在训练后期引入模拟退火凭借低温贪婪策略找到了较为平滑的最短路径, 但前期缺乏空间引导, 拐点个数为 8 个, 路径步数为 58 步; 如图 4(d) 所示, DPEA-QL 算法通过策略熵驱动的自适应学习率优化机制, 能够在较少的回合内收敛至更短路径, 拐点个数为 6 个, 路径步数为 58 步。图 4 表明, QL 和 Sarsa 算法在部分障碍区域存在冗余绕行现象, 而 DPEA-QL 通过策略熵机制实现探索与利用的平衡, 使智能配送车能够在有限的训练回合内找到更短且更平滑的最优路径, 相比其他 3 种算法表现出更高的路径质量和更优的路径规划性能。



(a) QL 算法　(b) Sarsa 算法

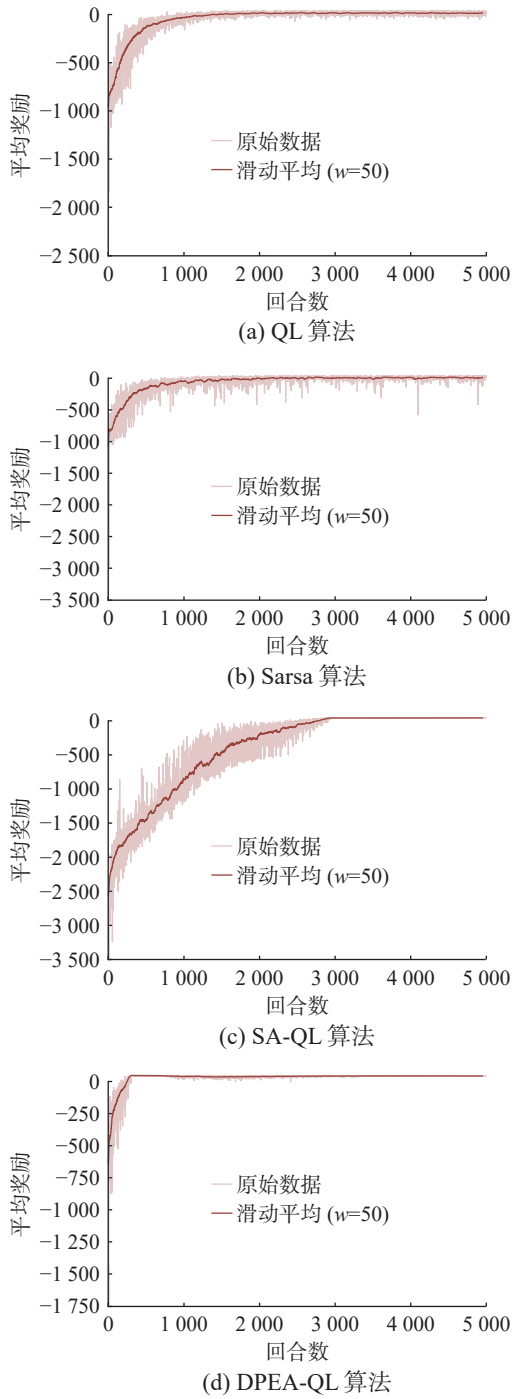
(c) SA-QL 算法　(d) DPEA-QL 算法

图 4　4 种算法的路径规划结果

Fig. 4　Path planning results of the four algorithms

图 5 是 4 种算法在训练过程中的奖励变化曲线。如图 5(a) 所示，QL 算法从初期较大负奖励 −2 450 开始缓慢上升且波动较大，训练后期逐渐趋于平缓，奖励值在 10 左右；如图 5(b) 所示，Sarsa 算法的初期奖励极低，约为−4 150，且训练后期奖励水平有限，奖励值在−40~20 且仍存在小幅震荡。如图 5(c) 所示，SA-QL 算法由于在高温阶段进行纯随机探索，初期试错成本极高，起始奖励低至−3 540，收敛速度极其缓慢，直至近 3 000 回合后才随温度下降逐渐攀升并稳定；如图 5(d) 所示，DPEA-QL 算法在早期阶段由策略熵驱动的自适应优化机制实现了高效探索，奖励值从初期奖励−2 000 快速上升，后期通过 sigmoid 函数平滑熵权重实现稳定收敛，并设计曼哈顿距离引导奖励贯穿完整训练阶段，最终奖励值平稳趋于 45 左右，显著优于对比算法。图 5 表明，DPEA-QL 在初期阶段奖励快速提升，且后期保持稳定收敛，说明其自适应学习率机制有效加速了 Q 值更新过程。

图 6 是 4 种算法的步数随训练回合变化曲线。在训练初期，所有算法都达到最大步数限制。如图 6(a) 所示，QL 算法步数收敛较慢，在 2000 回合数左右逐渐平稳，后期步数稳定在 64 步左右；如图 6(b) 所示，Sarsa 算法步数变化较大，且在训练后期仍频繁出现较大震荡，训练过程中无明显收敛；如图 6(c) 所示，SA-QL 算法则在近 3 000 个回合内基本处于 500 步的最大限制中，直至训练末期才平稳收敛；如图 6(d) 所示，DPEA-QL 算法在高熵阶段探索充分，步数波动略大，但随着双阶段机制的切换和自适应学习率调整，在回合数为 300 左右时步数迅速下降后稳定在最优路径附近。图 6 表明，DPEA-QL 在约 300 回合后步数趋于稳定，而对比算法仍处于较大波动或极其缓慢的收敛阶段，说明双阶段机制在提升训练稳定性与收敛速度方面的重要作用。



(a) QL 算法

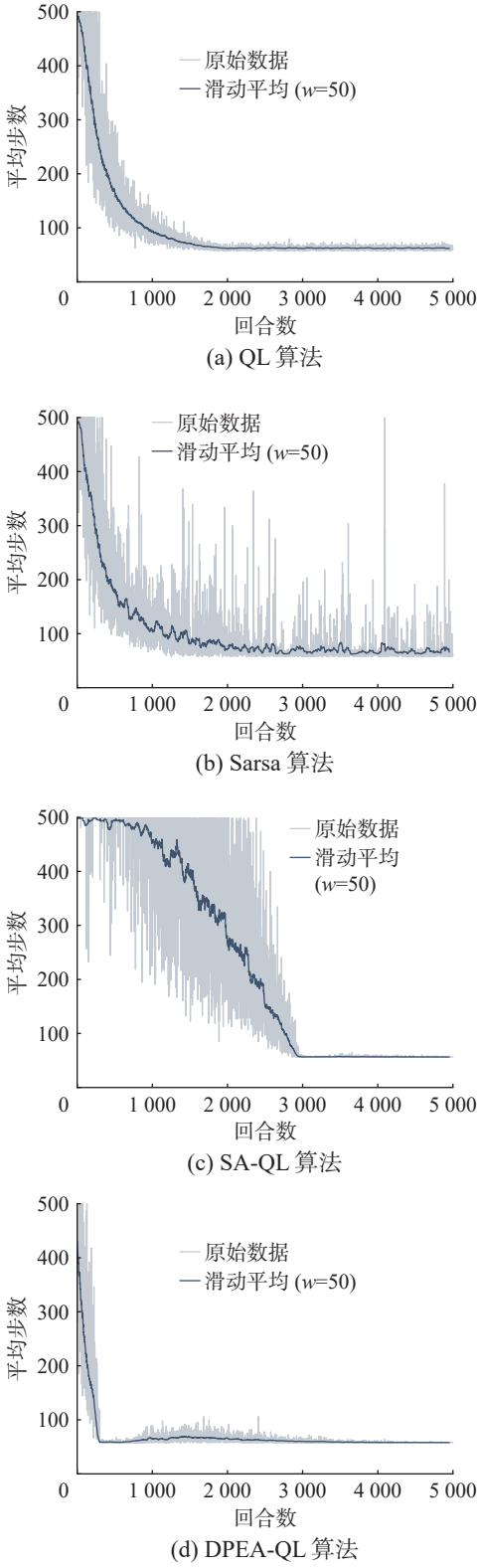
(b) Sarsa 算法

(c) SA-QL 算法

(d) DPEA-QL 算法

图 5　4 种算法的奖励收敛曲线

Fig. 5　Reward convergence curves of the four algorithms



(a) QL 算法

(b) Sarsa 算法

(c) SA-QL 算法

(d) DPEA-QL 算法

图 6　4 种算法的步数收敛曲线

Fig. 6　Step convergence curves of the four algorithms

实验表明，在单一静态障碍物地图环境中，DPEA-QL 算法较 QL、Sarsa 及 SA-QL 算法具备更高的路径最优性与学习收敛效率。

3.3.2　混合静态障碍物地图 Ⅱ 搭建与结果分析

为能够更真实地模拟城市交通中的复杂情况，在地图 Ⅰ 的基础上随机增加紫色区域，搭建混合静态障碍物地图 Ⅱ 如图 7 所示，用于模拟城市突发的交通封锁或施工场景等临时交通管制区，为不可通行区域，并评估算法在复杂静态环境中的适应性和路径重规划能力。

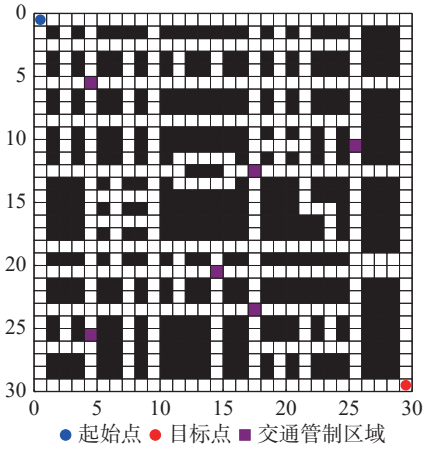
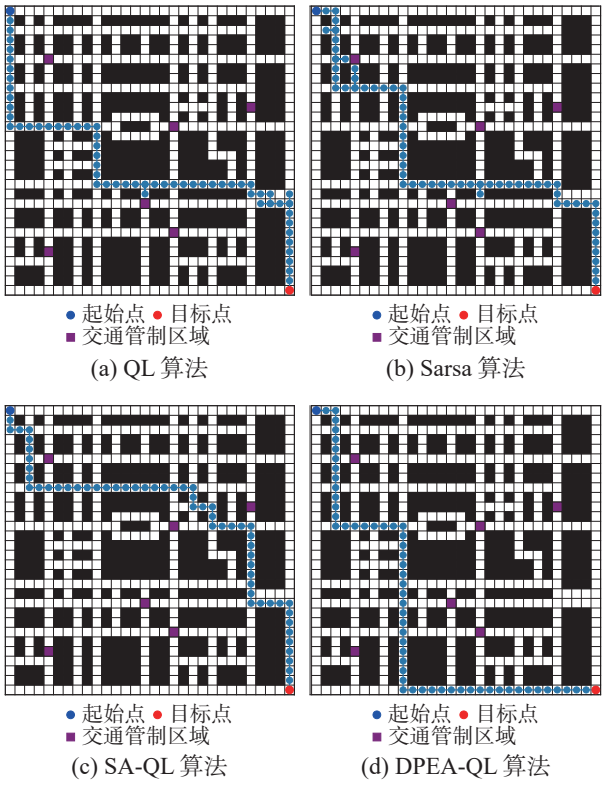


图 7　混合静态障碍物地图 Ⅱ

Fig. 7　Hybrid static obstacle map Ⅱ

图 8 给出了 4 种算法在混合静态障碍物环境下的路径规划结果。QL 算法在交通管制区域边界附近仍出现不必要的回环，导致整体路径存在冗余，路径步数为 64 步。Sarsa 算法在管制区附近容易陷入重复避障，寻优能力进一步下降，路径步数为 68 步。SA-QL 算法成功避开障碍并找到最短路径，但路径较为曲折。DPEA-QL 算法使智能配送车成功避开管制区域，且路径更平滑，达到理论最短路径 58 步。



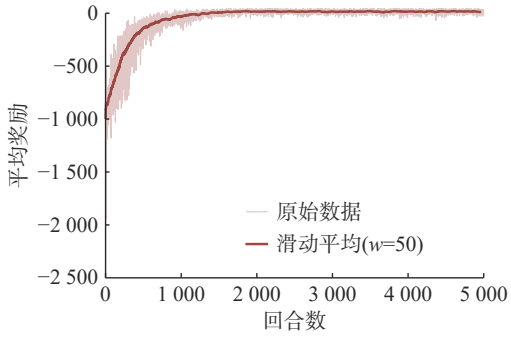
(a) QL 算法　(b) Sarsa 算法

(c) SA-QL 算法　(d) DPEA-QL 算法

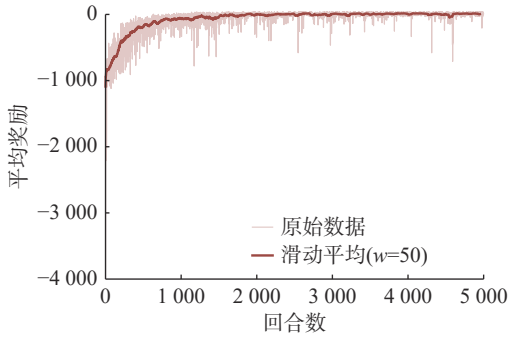
图 8　4 种算法的路径规划结果

Fig. 8　Path planning results of the four algorithms

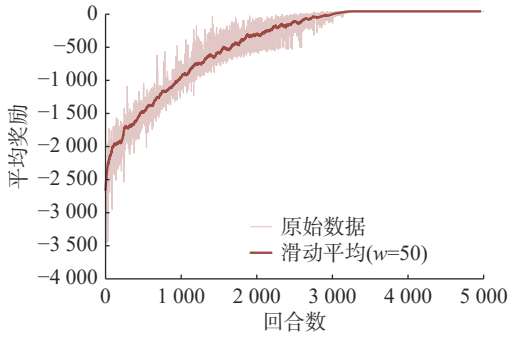
图 9 所示为奖励变化曲线。如图 9(a) 所示，QL 算法由于贪婪策略的探索随机性，容易碰撞障碍物，在初期奖励值为较大负值，为−2 750，在中后期奖励迅速上升并稳定于 20 左右，虽然存在轻微震荡，但算法整体在区间内趋于平稳。Sarsa 算法作为 on-policy 算法，容易陷入局部最优策略，如图 9(b) 所示，其初期奖励极低，约为−4 180，随着训练的进行，智能配送车的奖励没有出现稳定收敛的趋势。如图 9(c) 所示，SA-QL 算法在前期近 1 400 个回合内因高温随机探索而处于极低的负奖励，直到第 3 200 回合后才随温度下降缓慢收敛至 40 左右。如图 9(d) 所示，DPEA-QL 算法的奖励曲线，其初期奖励为−2000 并迅速上升后稳定收敛于 50 左右，在前期上升速度显著优于其他对比算法，且中后期波动较小，说明其在复杂混合环境下具有更好的稳定性和鲁棒性。



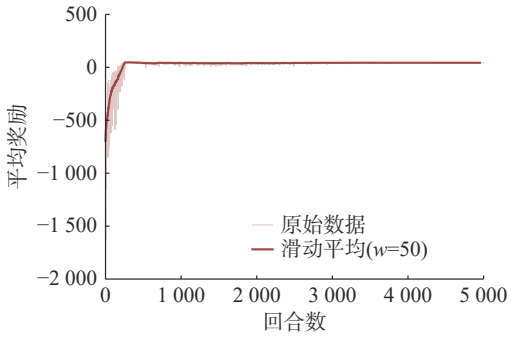
(a) QL 算法



(b) Sarsa 算法



(c) SA-QL 算法



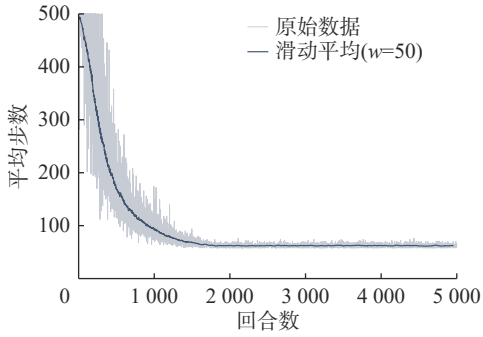
(d) DPEA-QL 算法

图 9　4 种算法的奖励收敛曲线

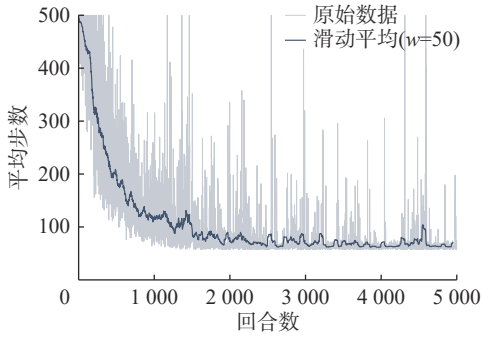
Fig. 9　Reward convergence curves of the four algorithms

图 10 为步数变化曲线，QL 算法中智能配送车在早期阶段广泛探索导致前期波动频繁，随着训练回合的增加，在回合数为 1 950 左右开始趋于平缓，但仍存在少量回合的步数突变。Sarsa 算法在完整训练过程中步数变化曲线不存在稳定趋势，说明 Sarsa 算法在复杂约束环境中稳定性不足。SA-QL 算法在训练初期 1 400 个回合内步数持续达到最大限制，到后期 3 200 回合后才随退火机制稳定在 58 步。而 DPEA-QL 算法中策略熵驱动自适应优化能够在训练中后期有效抑制震荡，步数曲线整体趋于平稳，稳定步数为 58 步，其路径规划更趋于最优。

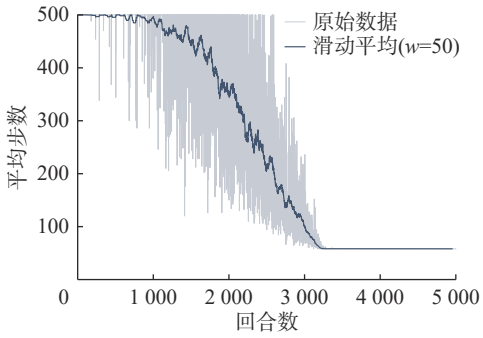
实验表明，DPEA-QL 算法在混合静态障碍物地图环境下，较对比算法具备更强的收敛稳定性与鲁棒性。



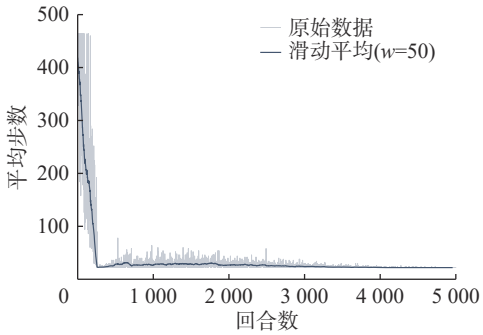
(a) QL 算法



(b) Sarsa 算法



(c) SA-QL 算法



(d) DPEA-QL 算法

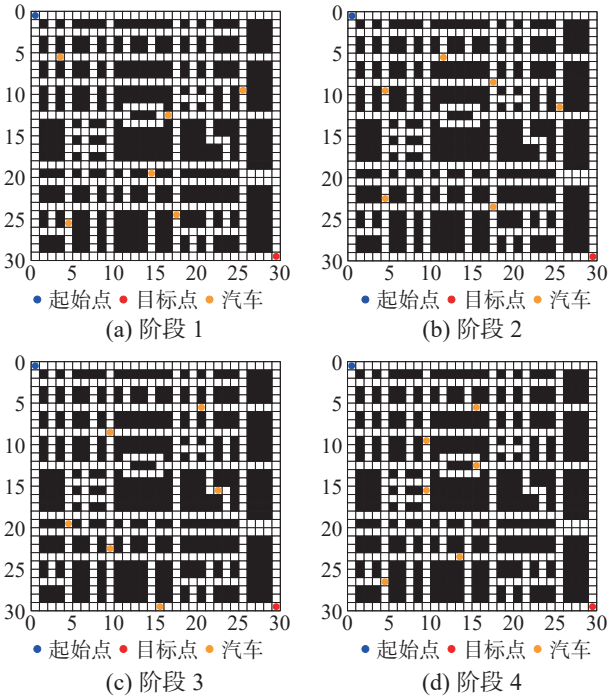
图 10　4 种算法的步数收敛曲线

Fig. 10　Step convergence curves of the four algorithms

3.3.3　动态障碍物地图 Ⅲ 搭建与结果分析

为进一步验证 DPEA-QL 的鲁棒性, 搭建动态障碍物地图 Ⅲ, 模拟真实城市交通场景中行驶车辆对路径规划性能的干扰, 其分阶段动态障碍物地图 Ⅲ 如图 11 所示。地图中以橙色圆圈表示动态车辆, 为了更好地模拟真实交通流, 本文设计了“基准路径+随机扰动+状态重生”的混合运动逻辑。具体而言: 1) 基准路径, 每辆动态车在生成时随机分配起始点与目标点, 并由 A*算法预先计算出一条全局最短路径作为其行驶意图; 2) 随机扰动, 在每个时间步中, 动态车辆大概率沿 A*路径行驶, 但设定了一定的随机扰动概率以模拟现实交通中前车急刹、避让行人或临时偏航等, 使其可能在原地停留或向周围可行栅格随机移动一步; 3) 状态重生, 当动态车辆到达其目标点后, 触发重生机制, 立即在地图可行区域内随机分配新的起点与终点。该机制确保了环境中动态障碍物的密度恒定, 从而高度模拟停车、避让或临时绕行等非平稳、高复杂度的现实交通情况。

为更加客观评估算法的稳定性与泛化性能, 确保实验结果的可靠性, 本研究设计多种子实验机制。由于动态环境中汽车的运动轨迹与初始状态存在随机性, 单一随机种子可能导致路径规划结果受偶然因素影响。通过在相同环境参数下设置 5 组不同的随机种子, 分别独立运行训练, 动态地图 Ⅲ 能高度模拟真实环境中的复杂性, 全面评估不同算法的动态避障与自适应规划能力。



(a) 阶段 1　(b) 阶段 2

(c) 阶段 3　(d) 阶段 4

图 11　动态障碍物分阶段地图 Ⅲ

Fig. 11　Dynamic obstacle phased map Ⅲ

图 12 给出了各算法在多种子实验条件下的最短路径累计次数对比结果。其中, 由于 Sarsa 算法策略更新依赖当前动作序列, 对动态环境变化的适应性较弱, 因此在相同条件下的最短路径次数最少; 而 QL 算法受动态环境扰动影响, 容易产生碰撞惩罚, 其最短路径的频率仍比较低; 引入模拟退火的 SA-QL 算法凭借训练后期的低温贪婪策略, 获得相对较高的最短路径累计次数, 但在动态环境中由于缺乏先验引导, 导致 SA-QL 在中前期存在巨大试错, 步数持续达到上限; DPEA-QL 算法规划得到最短路径的次数显著高于对比算法。如图 13 和图 14 所示, DPEA-QL 在面对高频动态扰动时, 能够在初期快速探索并收敛, 形成稳定的优化策略, 表明其在动态环境中具备更强的收敛稳定性与路径优化能力。

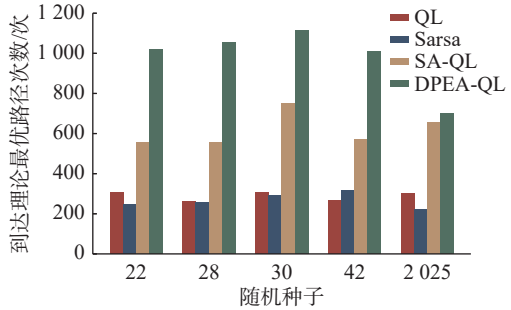


图 12　4 种算法在多种子实验下最短路径累计次数对比

Fig. 12　Comparison of cumulative optimal path counts of fouralgorithms under multi-seed experiments

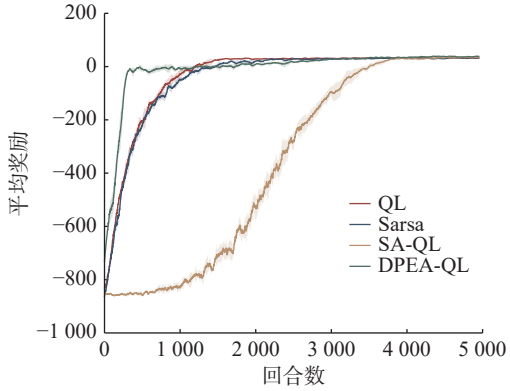


图 13　4 种算法在多种子实验下的平均奖励变化曲线

Fig. 13　Average reward variation curves of foural-gorithms under multi-seed experiments

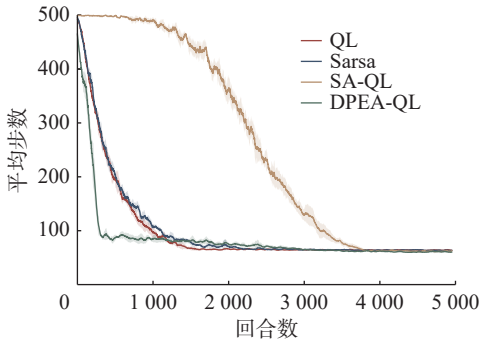


图 14　4 种算法在多种子实验下的平均步数变化曲线

Fig. 14　Average step variation curves of four algorithms

实验表明，DPEA-QL 在多随机种子条件下表现出更高的最短路径达成能力和收敛稳定性，说明其双阶段熵调节机制与自适应优化设计有效提升了算法的收敛性与鲁棒性。

3.3.4　消融实验设计及结果分析

为评估 DPEA-QL 算法中各改进模块对路径规划性能的影响，本研究在地图 Ⅲ 进行消融实验。实验以 DPEA-QL 作为基准算法，逐一移除改进模块，进而分析各模块的贡献度。本文采用成功路径平均步数指标，即在成功到达目标点的回合中，智能配送车从起点到目标点实际走过的步数均值，用于衡量算法在完成路径规划任务后的路径执行效率。为进一步量化算法的最优收敛性能，本文引入性能指标最短路径达成率 (optimal path achievement rate, OPAR)，定义为算法在训练过程中达到理论最短路径的累计次数占总成功次数的比例，用以衡量在去除失败回合干扰的基础上，算法在不同扰动条件下的全局最优收敛性能。

从表 2 消融实验结果得，各改进模块对算法性能提升的贡献程度不同。

表 2　消融实验结果

Table 2　Results of ablation experiment

分组	QL	adaptive-α	softmax-policy	sigmoid-smooth	manhattan-reward	平均奖励	成功路径平均步数	成功率/%	OPAR/%
基准算法	√	√	√	√	√	**1.34±8.34**	**80±4**	**98.45±0.46**	30.95±7.72
逐一移除改进模块	√		√	√	√	−118.16±18.17**	144±10**	91.92±1.16**	**35.61±9.04**
	√	√		√	√	−128.67±231.10*	150±83**	90.58±21.36	1.12±0.30**
	√	√	√		√	−30.06±16.92**	97±9**	98.43±0.37	1.57±2.53**
	√	√	√	√		−11.61±8.01**	85±4**	97.83±0.42**	26.61±8.15
基线算法	√					−40.39±8.57**	103±5**	97.07±0.74**	6.67±1.76**

注：粗字体为最佳结果。显著性分析基于独立样本t检验(Welch t检验)，表中“*”和“**”分别表示在p<0.05和p<0.01水平上与基准算法(DPEA-QL)存在显著差异；无标注表示与基准算法差异不显著。所有指标均为多种子实验平均值±标准差。

具体地，相对于 DPEA-QL 算法：

1) 移除 adaptive-α 机制后算法平均奖励从 1.34 下降至−118.16，减少了 119.5；成功路径平均步数从 80 上升至 144，增加了 64 步；成功率从 98.45% 下降至 91.92%，降低了 6.53 百分点；OPAR 从 30.95% 上升至 35.61%，提升了 4.66 百分点，且对成功率在 p<0.01 水平上与 DPEA-QL 存在显著差异，说明策略熵驱动学习率自适应优化模块能够有效提升学习效率和收敛速度，而对路径最优性影响有限。

2) 移除 softmax-policy 机制后算法平均奖励从 1.34 下降至−128.67，减少了 130.01；成功路径平均步数从 80 上升至 150，增加了 70 步；成功率从 98.45% 下降至 90.58%，降低了 7.87 百分点；OPAR 从 30.95% 下降至 1.12%，降低了 29.83 百分点，且对 OPAR 在 p<0.01 水平上与 DPEA-QL 存在显著差异，证明 softmax 策略对路径优化具有重要作用。

3) 移除 sigmoid-smooth 机制后算法平均奖励从 1.34 下降至−30.06，减少了 31.4；成功路径平均

步数从 80 上升至 97, 增加了 17 步; 成功率从 98.45% 下降至 98.43%, 降低了 0.02 百分点; OPAR 从 30.95% 下降至 1.57%, 降低了 29.38 百分点, 且对 OPAR 在 p<0.01 水平上与 DPEA-QL 存在显著差异, 说明设计平滑机制模块在平稳策略和提升路径最优性方面发挥关键作用。

4) 移除 manhattan-reward 机制后算法平均奖励从 1.34 下降至−11.61, 减少了 12.95; 成功路径平均步数从 80 上升至 85, 增加了 5 步; 成功率从 98.45% 下降至 97.83%, 降低了 0.62 百分点; OPAR 从 30.95% 下降至 26.61%, 降低了 4.34 百分点, 且对成功率在 p<0.01 水平上与 DPEA-QL 存在显著差异, 表明曼哈顿距离引导奖励模块对全局路径引导和避免局部收敛具有重要意义。

此外, 与基线算法 QL 算法相比, DPEA-QL 算法在各指标上均有显著提升, 平均奖励从−40.39 上升至 1.34, 增加了 41.73; 成功路径平均步数从 103 降低至 80, 减少了 23 步; 成功率从 97.07% 上升至 98.45%, 提升了 1.38 百分点; OPAR 从 6.67% 上升至 30.95%, 提升了 24.28 百分点, 整体算法的路径最优性与全局收敛性能显著提升, 验证了各模块协同优化的合理有效性。

4　结束语

针对传统 Q-learning 算法在智能配送车路径规划中存在收敛速度慢、路径质量不佳以及动态环境适应性差的问题, 本文在传统 Q-learning 算法的基础上提出了基于双阶段策略熵自适应 Q-learning(DPEA-QL) 的路径规划算法。该算法通过设计策略熵驱动的自适应优化机制, 提高了算法的探索效率和收敛速度; 构建双阶段熵调控机制, 改善了智能配送车的策略平稳过渡和路径优化能力; 同时引入基于曼哈顿距离的奖励重塑, 增强了配送车的目标导向信号, 有效避免算法陷入局部收敛。通过在 3 类仿真地图上进行实验, 在单一静态障碍物环境下对比分析 DPEA-QL、传统 Q-learning、Sarsa 和引入模拟退火的 SA-QL 算法, 验证 DPEA-QL 算法的可行性。在混合静态障碍物环境下分析对比 4 种算法, 表明 DPEA-QL 算法环境适应性、路径规划质量和收敛速度都优于对比算法。在动态障碍物环境下对比分析 4 种算法, 说明 DPEA-QL 算法较对比算法具备更强的环境适用性、收敛稳定性与路径最优性。虽然改进后的算法在提升路径规划性能的同时实现了更智能的策略调整, 但在复杂场景下计算开销有所增加, 尤其是在动态环境下的训练时间较长。未来的研究主要沿着两个维度展开, 一方面, 在理论架构上推动自适应优化机制向深度强化学习框架拓展, 结合轻量化神经网络与多模态感知技术, 解决大规模连续拓扑交通场景引发的“维度灾难”问题; 另一方面, 在工程应用上加速算法的实物落地与交叉编译, 将优化后的模型移植至真实的智能配送车硬件平台, 通过在封闭园区或复杂非结构化城市道路中开展真车实地测试, 进一步校准并提升算法在真实物理世界中的实时规划与安全避障能力, 从而更好地满足现代智能物流配送体系规模化应用的战略需求。

参考文献:

[1] 孙智威, 裴晓飞, 刘一平, 等. 无人驾驶清扫车的路径跟踪及远程控制[J]. 汽车安全与节能学报, 2022, 13(4): 729−737.
SUN Zhiwei, PEI Xiaofei, LIU Yiping, et al. Path tracking and remote control of driverless sweeper[J]. Journal of automotive safety and energy, 2022, 13(4): 729−737.

[2] 李玉龙, 谢辉, 宋康. 无人驾驶公交车基于循迹误差观测和目标测量误差观测的避障路径规划算法[J]. 汽车安全与节能学报, 2024, 15(4): 579−590.
LI Yulong, XIE Hui, SONG Kang. An obstacle avoidance path planning algorithm for autonomous buses based on tracking error observation and target measurement error observation[J]. Journal of automotive safety and engergy, 2024, 15(4): 579−590.

[3] 夏雨奇, 黄炎焱, 陈恰. 基于深度 Q 网络的无人车侦察路径规划[J]. 系统工程与电子技术, 2024, 46(9): 3070−3081.
XIA Yuqi, HUANG Yanyan, CHEN Qia. Path planning for unmanned vehicle reconnaissance based on deep Q-network[J]. Systems engineering and electronics, 2024, 46(9): 3070−3081.

[4] 成怡, 肖宏图. 融合改进 A*算法和 Morphin 算法的移动机器人动态路径规划[J]. 智能系统学报, 2020, 15(3): 546−552.
CHENG Yi, XIAO Hongtu. Mobile-robot dynamic path planning based on improved A* and Morphin algorithms [J]. CAAI transactions on intelligent systems, 2020, 15(3): 546−552.

[5] QU Tianci, XIONG Gang, ALI H, et al. USV path planning under marine environment simulation using DWA and safe reinforcement learning[C]//2023 IEEE 19th International Conference on Automation Science and Engineering. Auckland: IEEE, 2023: 1−6.

[6] 李娟, 张子浩, 张宏瀚. 复杂环境下 DWA 与 RRT 算法融合的 AUV 局部路径规划[J]. 智能系统学报, 2024, 19(4): 961−973.
LI Juan, ZHANG Zihao, ZHANG Honghan. Local path planning for AUV with fusion of DWA and RRT algorithms in a complex environment[J]. CAAI transac-

tions on intelligent systems, 2024, 19(4): 961−973.

[7] 蔡军, 钟志远. 改进蚁群算法的送餐机器人路径规划[J]. 智能系统学报, 2024, 19(2): 370−380.
CAI Jun, ZHONG Zhiyuan. Path planning of a meal delivery robot based on an improved ant colony algorithm[J]. CAAI transactions on intelligent systems, 2024, 19(2): 370−380.

[8] CHEN Zenghua, XIONG Gang, LIU Sheng, et al. Path planning of mobile robot based on an improved genetic algorithm[C]//2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence. Boston: IEEE, 2022: 1−6.

[9] WANG Chunlei, YANG Xiao, LI He. Improved Q-learning applied to dynamic obstacle avoidance and path planning[J]. IEEE access, 2022, 10: 92879−92888.

[10] ZHANG Kehan. Path planning of intelligent new energy vehicles based on DQN[C]//2025 5th International Conference on Electronics, Circuits and Information Engineering. Guangzhou: IEEE, 2025: 845−848.

[11] 伍锡如, 沈可扬. 基于人工势场的防疫机器人改进近端策略优化算法[J]. 智能系统学报, 2025, 20(3): 689−698.
WU Xiru, SHEN Keyang. Improved proximal policy optimization algorithm for epidemic prevention robots based on artificial potential fields[J]. CAAI transactions on intelligent systems, 2025, 20(3): 689−698.

[12] 张森, 代强强. 改进型深度确定性策略梯度的无人机路径规划[J]. 系统仿真学报, 2025, 37(4): 875−881.
ZHANG Sen, DAI Qiangqiang. UAV path planning based on improved deep deterministic policy gradients[J]. Journal of system simulation, 2025, 37(4): 875−881.

[13] 李永迪, 李彩虹, 张耀玉, 等. 基于改进 SAC 算法的移动机器人路径规划[J]. 计算机应用, 2023, 43(2): 654−660.
LI Yongdi, LI Caihong, ZHANG Yaoyu, et al. Mobile robot path planning based on improved SAC algorithm[J]. Journal of computer applications, 2023, 43(2): 654−660.

[14] CHEN Lili, LU K, RAJESWARAN A, et al. Decision transformer: reinforcement learning via sequence modeling[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 2021: 15084−15097.

[15] LI Zhenxin, LI Kailin, WANG Shihao, et al. Hydra-MDP: end-to-end multimodal planning with multi-target hydra-distillation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 15823−15833.

[16] ZHOU Zewei, CAI Tianhui, ZHAO S Z, et al. AutoVLA: a vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning[C]//Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego: Curran Associates Inc. , 2025: 1−30.

[17] 宋丽君, 周紫瑜, 李云龙, 等. 改进 Q-Learning 的路径规划算法研究[J]. 小型微型计算机系统, 2024, 45(4): 823−829.
SONG Lijun, ZHOU Ziyu, LI Yunlong, et al. Research on path planning algorithm based on improved Q-learning algorithm[J]. Journal of Chinese computer systems, 2024, 45(4): 823−829.

[18] 田晓航, 霍鑫, 周典乐, 等. 基于蚁群信息素辅助的 Q 学习路径规划算法[J]. 控制与决策, 2023, 38(12): 3345−3353.
TIAN Xiaohang, HUO Xin, ZHOU Dianle, et al. Ant colony pheromone aided Q-learning path planning algorithm[J]. Control and decision, 2023, 38(12): 3345−3353.

[19] 王小康, 冀杰, 刘洋, 等. 基于改进 Q 学习算法的无人物流配送车路径规划[J]. 系统仿真学报, 2024, 36(5): 1211−1221.
WANG Xiaokang, JI Jie, LIU Yang, et al. Path planning of unmanned delivery vehicle based on improved Q-learning algorithm[J]. Journal of system simulation, 2024, 36(5): 1211−1221.

[20] 赵也践, 王艳红, 张俊, 等. 改进 Q 学习算法在作业车间调度问题中的应用[J]. 系统仿真学报, 2022, 34(6): 1247−1258.
ZHAO Yejian, WANG Yanhong, ZHANG Jun, et al. Application of improved Q learning algorithm in job shop scheduling problem[J]. Journal of system simulation, 2022, 34(6): 1247−1258.

[21] ZANGIROLAMI V, BORROTTI M. Dealing with uncertainty: Balancing exploration and exploitation in deep recurrent reinforcement learning[J]. Knowledge-based systems, 2024, 293: 111663.

[22] MA Haozhe, LUO Zhengding, VO T V, et al. Highly efficient self-adaptive reward shaping for reinforcement learning[EB/OL]. (2025−02−28)[2025−11−04]. https://doi.org/10.48550/arXiv.2408.03029.

[23] 任伟, 朱建鸿. 改进的自校正 Q-learning 应用于智能机器人路径规划[J]. 机械科学与技术, 2025, 44(1): 126−132.
REN Wei, ZHU Jianhong. Improved self-tuning Q-learning algorithm applied to path planning of intelligent robot[J]. Mechanical science and technology for aerospace engineering, 2025, 44(1): 126−132.

[24] 温广辉, 杨涛, 周佳玲, 等. 强化学习与自适应动态规划: 从基础理论到多智能体系统中的应用进展综述[J]. 控制与决策, 2023, 38(5): 1200−1230.
WEN Guanghui, YANG Tao, ZHOU Jialing, et al. Reinforcement learning and adaptive/approximate dynamic programming: a survey from theory to applications in multi-agent systems[J]. Control and decision, 2023, 38(5): 1200−1230.

[25] 段建民, 陈强龙. 利用先验知识的 Q-Learning 路径规划算法研究[J]. 电光与控制, 2019, 26(9): 29−33.
DUAN Jianmin, CHEN Qianglong. Prior knowledge based Q-learning path planning algorithm[J]. Electronics optics & control, 2019, 26(9): 29−33.

[26] 蔡静雯, 马玉敏, 黎声益, 等. 基于 Q 学习的智能车间自适应调度方法[J]. 计算机集成制造系统, 2023, 29(11): 3727−3737.
CAI Jingwen, MA Yumin, LI Shengyi, et al. Self-adaptive scheduling method for smart shop floor based on Q-learning[J]. Computer integrated manufacturing systems, 2023, 29(11): 3727−3737.

[27] 杨秀霞, 高恒杰, 刘伟, 等. 基于阶段 Q 学习算法的机器人路径规划[J]. 兵器装备工程学报, 2022, 43(5): 197−203.
YANG Xiuxia, GAO Hengjie, LIU Wei, et al. Robot path planning based on stage Q learning algorithm[J]. Journal of ordnance equipment engineering, 2022, 43(5): 197−203.

[28] MAHRAN Y, GAMAL Z, EL-BADAWY A. Dynamic entropy tuning in reinforcement learning low-level quadcopter control: stochasticity vs determinism[C]//2024 34th International Conference on Computer Theory and Applications. Alexandria: IEEE, 2024: 185−192.

[29] 陈永, 康婕. 玻尔兹曼优化 Q-learning 的高速铁路越区切换控制算法[J]. 控制理论与应用, 2025, 42(4): 688−694.
CHEN Yong, KANG Jie. Boltzmann optimized Q-learning algorithm for high-speed railway handover control[J]. Control theory & applications, 2025, 42(4): 688−694.

[30] HAARNOJA T, TANG Haoran, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]// International Conference on Machine Learning. Sydney: PMLR, 2017: 1352−1361.

[31] ALI H, XIONG Gang, WU Huaiyu, et al. Multi-robot path planning and trajectory smoothing[C]//2020 IEEE 16th International Conference on Automation Science and Engineering. Hong Kong: IEEE, 2020: 685−690.

作者简介:

熊刚, 研究员, 博士生导师, 兼任北京市智能化技术与系统工程技术研究中心常务副主任、中国人工智能学会高级会员、中国自动化学会会士、中国计算机学会杰出会员等。主要研究方向为人工智能、智能控制与管理。先后获得吴文俊人工智能科技进步奖二等奖、中国自动化学会科技进步奖特等奖和一等奖等 10 余项。通过专利合作条约 (patent cooperation treaty, PCT) 途径申请并获授权 6 项, 获专利授权 100 余项, 登记软著 90 余项, 发表学术论文 500 余篇, 出版专著 3 部。E-mail: gang.xiong@ia.ac.cn。

陈德旺, 博士, 电气电子工程师学会高级会员, 中国自动化学会高级会员, 福建省“闽江学者”特聘教授。主要研究方向为人工智能算法、模糊系统和智能交通系统。获得省部级和国家一级学会科技奖励 10 余项。发表学术论文 200 余篇, 出版学术专著 4 部。E-mail: dwchen@fjut.edu.cn。

蔡凯琪, 硕士研究生, 主要研究方向为平行交通与物流系统。E-mail: 397159947@qq.com。