基于可行性约束航点评估的移动机器人分层目标驱动探索方法

李艺博[1]，关 哲[1,2†]，郝 曼[1]，丁伟利[1]，华长春[1]

(1. 燕山大学 电气工程学院，河北 秦皇岛 066000；
2. 河北省工业计算机控制工程重点实验室，河北 秦皇岛 066000)

摘 要: 针对未知环境下移动机器人目标驱动探索中在线地图不完整、候选航点可执行性不足以及局部强化学习策略缺少长程引导的问题, 基于目标驱动自主探索 (GDAE), 设计可行性约束多目标航点评估机制, 并提出改进型目标驱动自主探索方法 (iGDAE). 所提出方法采用中间航点选择-局部航点跟踪的分层结构: 上层根据二维激光雷达 (LiDAR) 观测提取障碍物边界、狭窄通道和开阔区域候选航点, 并结合自由空间、地图路径安全、局部可跟踪范围、通道宽度和航点滞回等约束完成中间航点选择; 下层采用融合复合奖励函数与优先经验回放的双延迟深度确定性策略梯度 (RPER-TD3) 局部控制器输出连续速度指令, 实现航点跟踪和局部避障. 在 ROS/Gazebo 静态仿真场景中, iGDAE 在 4 个未知地图的 40 次测试中均到达目标. 若将失败实验按照预设惩罚值计入统计, iGDAE 相较于分层基线 GDAE-TD3 的平均时间代价与平均路径代价分别降低 19.47% 和 34.96%. 实验结果表明, 所设计的可行性约束多目标航点评估与连续局部控制的协同能够改善所设仿真任务中的目标引导效率和运行稳定性.

关键词: 移动机器人；目标驱动探索；深度强化学习；激光雷达；航点选择；可行性约束

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2026.0705

引用格式: 李艺博, 关哲, 郝曼, 等. 基于可行性约束航点评估的移动机器人分层目标驱动探索方法 [J]. 控制与决策.

A hierarchical goal-driven exploration method for mobile robots based on feasibility-constrained waypoint evaluation

LI Yi-bo[1]，*GUAN Zhe*[1,2†]，*HAO Man*[1]，*DING Wei-li*[1]，*HUA Chang-chun*[1]

(1. School of Electrical Engineering，Yanshan University，Qinhuangdao 066000，China；2. Hebei Province Key Laboratory of Industrial Computer Control Engineering，Qinhuangdao 066000，China)

Abstract: To address incomplete online maps, limited executability of candidate waypoints, and insufficient long-range guidance of local reinforcement learning policies in goal-driven exploration of mobile robots in unknown environments, this paper develops a feasibility-constrained multi-objective waypoint evaluation mechanism based on goal-driven autonomous exploration (GDAE) and proposes an improved GDAE method, termed iGDAE. The method adopts a hierarchical architecture comprising intermediate waypoint selection and local waypoint tracking. At the upper layer, candidate waypoints corresponding to obstacle boundaries, narrow passages, and open areas are extracted from two-dimensional LiDAR observations. Intermediate waypoints are then selected by jointly considering free space, map-based path safety, local trackability, passage width, and waypoint hysteresis. At the lower layer, a twin delayed deep deterministic policy gradient controller integrating a composite reward function and prioritized experience replay (RPER-TD3) generates continuous velocity commands for waypoint tracking and local obstacle avoidance. In static ROS/Gazebo simulations, iGDAE reaches the target in all 40 trials across four unknown maps. When failed trials are included using predefined penalties, iGDAE reduces the mean time cost and mean path cost by 19.47% and 34.96%,

收稿日期: 2026-07-15；**录用日期:** 2026-10-22.

基金项目: 国家自然科学基金项目 (62303398)；骨干人才项目 (留学回国平台)(C2025006)；国家自然科学基金项目 (62303399)；河北省高等学校科学研究项目 (PTX2026003)；河北省自然科学基金项目 (A2026203007)；河北省级重点实验室绩效补助经费项目 (22567612H).

责任编委: 张文安.

†通信作者. E-mail: guan@ysu.edu.cn.

respectively, compared with the hierarchical baseline GDAE-TD3. The results demonstrate that coordinating feasibility-constrained multi-objective waypoint evaluation and continuous local control improves target-guidance efficiency and operational stability in these simulation tasks.

Keywords: mobile robot； goal-driven exploration； deep reinforcement learning； LiDAR； waypoint selection； feasibility constraints

0 引 言

未知环境中的目标驱动探索要求移动机器人在缺少先验地图的条件下, 依靠实时感知和在线建图逐步到达给定目标区域. 与已知地图路径规划相比, 该任务同时涉及未知区域感知、增量地图构建、目标方向推进和局部避障等问题, 是自主导航、主动SLAM 和未知环境探索中的耦合决策问题. 在长走廊、狭窄通道、局部遮挡和障碍物密集场景中, 在线地图的不完整性和传感器感知范围有限会显著增加航点选择与局部控制难度, 机器人容易出现绕行低效、贴障运动、航点频繁切换或陷入局部最优等现象. 因此, 在未知环境中选择安全、可达且有利于目标推进的中间航点, 并由局部控制器稳定执行, 是目标驱动探索中的关键问题[1-2].

现有方法大致可分为基于地图信息的规划方法和基于深度强化学习的局部控制方法. 前沿点探索和 next-best-view 规划通常利用占据栅格地图、前沿边界、信息增益或路径代价选择导航目标[3]. FUEL[4] 通过增量前沿结构和分层规划提升未知空间覆盖效率, TARE[5] 采用局部细粒度与全局粗粒度相结合的分层框架提高复杂三维环境中的探索效率. 这类方法具有较强的可解释性和全局搜索能力, 但在探索初期容易受到 SLAM 误差、未知区域边界和栅格分辨率影响, 导致路径或候选航点不稳定[6-7].

基于深度强化学习的方法能够从 LiDAR 观测和目标相对状态中直接学习连续速度控制策略, 适用于局部避障和短距离目标跟踪[8-9]. 学习型导航方法在处理复杂非线性控制和局部避障方面具有优势, 但泛化性、安全性和可解释性仍需进一步提高[10-11]. 已有研究验证了移动机器人仅依靠二维激光观测进行无地图连续控制和局部避障的可行性[12-13]. 奖励函数设计与知识迁移也被证明有助于提升训练效率和导航性能[14-15]. 此外, DWA-RL[16] 将动态窗口约束与强化学习策略结合, 以增强速度指令的动力学可行性. 改进 TD3 方法[17] 则从策略学习角度改善局部避障性能. 然而, 纯局部策略通常缺乏长程目标引导. 当全局目标被障碍物遮挡或局部通道结构复杂时, 机器人可能出现原地旋转、往复运动或目标引导不足等行为.

为兼顾长程引导和局部连续控制, 分层目标驱动探索方法将远距离导航任务分解为中间航点选择和局部航点跟踪. GDAE[18] 通过兴趣点选择和深度强化学习局部控制实现未知环境下的目标驱动探索. 图结构强化学习探索方法从不确定性和信息增益角度建模探索决策[19-20], Active Neural SLAM[21] 采用模块化结构缓解端到端策略在长程规划中的困难. 但是, 现有分层方法仍存在上层航点选择与下层控制执行协同不足的问题[22-23]. 一是上层若主要依据目标距离、候选方向或局部开阔度选择航点, 容易忽略航点可执行性、路径安全余量、局部控制器可跟踪范围和时间连续性. 二是下层若只依赖到达奖励和碰撞惩罚, 则难以同时约束目标趋近、避障安全、速度效率和动作平滑[24-25].

针对上述问题, 本文基于 GDAE 分层目标驱动探索框架, 构建可行性航点引导的改进方法 iGDAE. 上层基于二维 LiDAR 局部几何信息生成候选航点, 利用自由空间、路径安全、局部可跟踪范围和通道宽度等约束筛除不可行航点, 再通过多目标效用函数选择中间航点. 下层采用 RPER-TD3 控制器, 通过复合奖励函数引导目标趋近、安全避障和动作平滑, 并利用优先经验回放提高关键样本利用效率.

本文主要工作如下:

1. 改进了面向未知环境目标驱动探索的分层导航框架, 通过增强上层航点选择与下层局部控制之间的协同, 提高目标引导的稳定性.

2. 设计可行性约束的多目标航点评估方法, 将 LiDAR 局部几何结构、地图路径安全、局部可跟踪性和航点滞回机制纳入同一评分框架, 以降低不可执行航点和频繁切换的风险.

3. 面向中间航点跟踪任务, 设计融合复合奖励函数与优先经验回放的 RPER-TD3 局部控制器, 并在 ROS/Gazebo 平台上从局部航点跟踪和完整目标驱动探索两个层面进行验证.

1 方 法

本文方法采用中间航点选择-局部航点跟踪的分层结构. 上层根据二维 LiDAR 观测、在线 SLAM 地图和全局目标生成可行中间航点. 下层以中间航点为局部目标, 利用 RPER-TD3 控制器输出连续速

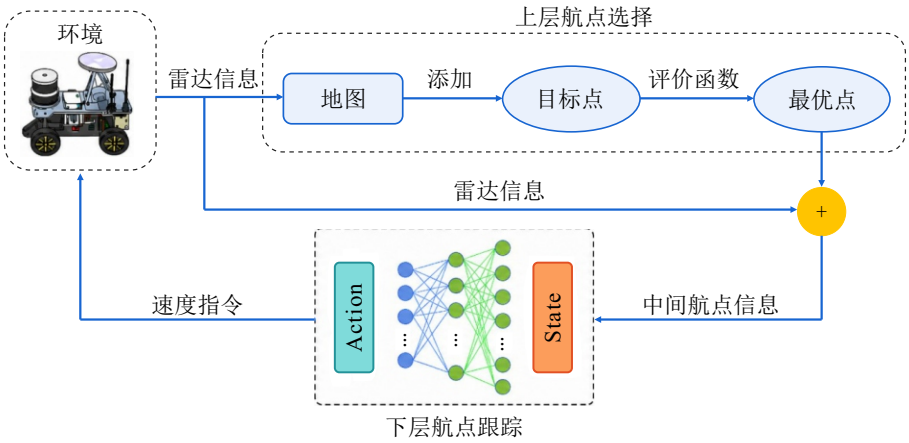


图 1　分层目标驱动探索框架

度指令, 实现航点跟踪与局部避障. 方法整体框架如图 1 所示.

1.1　问题描述与总体框架

设移动机器人在未知二维环境中运动, 时刻t的机器人位姿为$p_t = (x_t, y_t, \theta_t)$, 其中$(x_t, y_t)$为机器人位置, θ_t为航向角. 给定全局目标点$g = (x_g, y_g)$, 机器人需要在无先验地图条件下, 依靠实时感知、在线建图和局部避障逐步到达目标区域.

机器人搭载二维 LiDAR, 当前观测记为$D_t = \{d_1, d_2, \ldots, d_N\}$, 其中$d_k$为第$k$束激光的测距值, N为激光束数量. 在线 SLAM 模块同步更新增量地图M_t. 在每个决策周期, 上层根据D_t、M_t、p_t和g生成候选航点集合C_t, 并选择中间航点w_t^*. 下层以w_t^*为短期目标, 输出控制动作$a_t = (v_t, \omega_t)$, 其中v_t和ω_t分别为线速度和角速度.

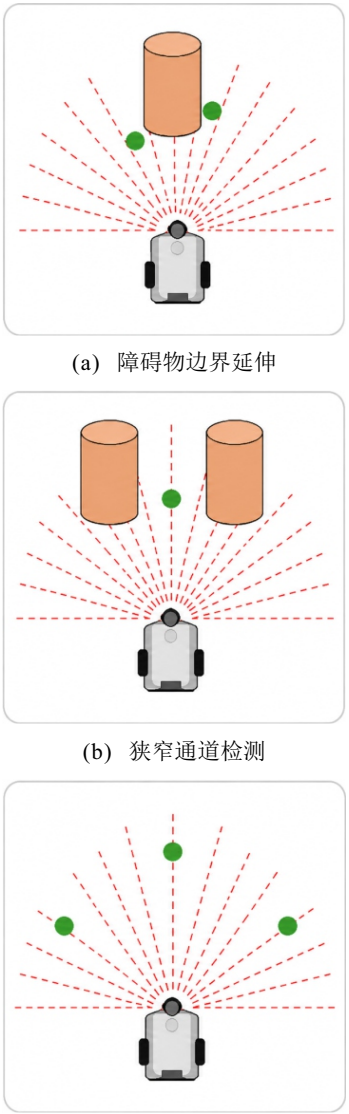
1.2　LiDAR 候选航点生成

未知环境探索初期在线地图通常不完整, 若完全依赖栅格地图生成航点, 容易受到建图误差和未知区域边界影响. 为此, 本文基于 GDAE 的局部观测候选航点生成机制, 直接从二维 LiDAR 测距序列中提取局部几何结构, 生成候选航点集合$C_t = \{c_i\}_{i=1}^{n_c}$. 候选航点包括障碍物边界点、狭窄通道点和开阔区域点三类, 其类型标签记为$\tau_i \in \{\text{edge}, \text{gap}, \text{open}\}$, 如图 2 所示.

设第k束激光在机器人坐标系下的角度为$\vartheta_k = \vartheta_{\min} + (k-1)\Delta\vartheta$, 其端点为$q_k^r = d_k[\cos\vartheta_k, \sin\vartheta_k]^{\mathrm{T}}$. 若候选航点在机器人坐标系下的方向为$\psi_i$, 投影距离为$r_i^c$, 则其世界坐标为

$$c_i = \begin{bmatrix} x_t \\ y_t \end{bmatrix} + r_i^c \begin{bmatrix} \cos(\theta_t + \psi_i) \\ \sin(\theta_t + \psi_i) \end{bmatrix}. \tag{1}$$

障碍物边界候选点由相邻 LiDAR 测距突变生成. 若



(a)　障碍物边界延伸

(b)　狭窄通道检测

(c)　开阔区域引导

图 2　基于 LiDAR 的候选航点生成示意图

$$|d_k - d_{k+1}| > \delta_{\text{dis}}, \tag{2}$$

则认为该方向附近存在障碍物边界或遮挡边界, 并取测距较大的一侧作为候选方向, 即$m = \arg\max_{j\in\{k,k+1\}} d_j$, 令$\psi_i = \vartheta_m$, $r_i^c = \min(d_m - r_s, d_{\text{proj}})$, $\tau_i = \text{edge}$. 其中r_s为机器人半径与安全裕度之

和, d_{proj}为候选航点最大投影距离. 若$r_i^c \leqslant 0$, 则不生成该候选点.

狭窄通道候选点由 LiDAR 测距序列中的近—远—近结构生成. 设满足该结构的连续激光区间为$\mathcal{G}=[k_l,k_r]$, 通道宽度估计为

$$b_i = \|q_{k_l}^r - q_{k_r}^r\|_2. \tag{3}$$

候选点沿通道中心方向生成, 即$\psi_i=(\vartheta_{k_l}+\vartheta_{k_r})/2$, $r_i^c=\min(\text{median}_{k\in\mathcal{G}}\, d_k - r_s, d_{\text{proj}})$, 并令$\tau_i=\text{gap}$.

开阔区域候选点由连续大测距区间生成. 设满足自由空间阈值的连续激光区间为$\mathcal{O}=[k_l,k_r]$, 且该区间角宽不小于$\Delta\theta_{\text{open}}$, 则沿区间角平分线生成候选点, 即$\psi_i=(\vartheta_{k_l}+\vartheta_{k_r})/2$, $r_i^c=\min(|\mathcal{O}|^{-1}\sum_{k\in\mathcal{O}} d_k - r_s, d_{\text{proj}})$, 并令$\tau_i=\text{open}$. 三类候选航点同时出现时, 统一筛选并保留综合效用最高的可行航点.

1.3 可行性约束多目标航点评估

候选航点生成后, 需要从C_t中选择当前中间航点. 若仅依据目标距离或局部开阔度进行选择, 容易产生不可达航点、贴障航点或频繁切换航点. 为此, 本文将航点评估设计为可行性约束筛选—多目标效用评分的两阶段过程:

$$w_t^* = \arg\max_{c_i\in C_t^{\text{f}}} J_t(c_i), \tag{4}$$

其中C_t^{f}为满足可行性约束的候选集合, $J_t(c_i)$为候选航点综合效用函数.

记机器人当前位置为$p_t^{xy}=(x_t,y_t)$. 候选航点$c_i=(x_i,y_i)$到机器人当前位置的距离和相对方位角分别为

$$\begin{aligned}&\rho_i = \|c_i - p_t^{xy}\|_2,\\ &\phi_i = \text{wrap}(\text{atan2}(y_i-y_t, x_i-x_t) - \theta_t).\end{aligned} \tag{5}$$

其中$\text{wrap}(\cdot)$表示将角度归一化到$(-\pi,\pi]$.

本文从距离范围、LiDAR 自由空间、地图路径安全、局部可跟踪范围和通道宽度五个方面筛选候选航点. 首先, 候选航点距离需满足$\rho_{\min}\leqslant\rho_i\leqslant\rho_{\max}$. 其次, 以候选航点方向$\phi_i$为中心构建 LiDAR 扇形窗口

$$\Omega_i = \{k \mid |\text{wrap}(\vartheta_k - \phi_i)| \leqslant \Delta\theta_i\}, \tag{6}$$

并记窗口内最小测距为$d_i^{\min}=\min_{k\in\Omega_i} d_k$, 要求$\rho_i < d_i^{\min} - r_s$, 以保证候选方向前方具有足够自由空间.

进一步地, 利用在线 SLAM 地图进行局部路径安全校验. 记$D_M(x)$为地图中位置x到最近已知障碍物的距离, $\mathcal{P}_i$为线段$\overline{p_t^{xy}c_i}$上按地图分辨率离散得到的采样点集合, 则路径最小安全距离为

$$D_{\text{path}}(c_i) = \min_{x\in\mathcal{P}_i} D_M(x). \tag{7}$$

若线段采样点落入已知占据或障碍膨胀区域, 则候选航点判为不可行. 未知栅格不直接作为障碍处理, 但候选航点仍需满足 LiDAR 自由空间约束. 综合上述条件, 可行候选集合定义为

$$\begin{aligned}C_t^{\text{f}} = \{&c_i\in C_t \mid \rho_{\min}\leqslant\rho_i\leqslant\rho_{\max}, \rho_i < d_i^{\min} - r_s,\\ &D_{\text{path}}(c_i) > r_s, |\phi_i|\leqslant\phi_{\max}, \Gamma_i = 1\},\end{aligned} \tag{8}$$

其中

$$\Gamma_i = \begin{cases}1, \tau_i \neq \text{gap},\\ 1, \tau_i = \text{gap 且 } b_i > 2r_s+\delta_b,\\ 0, \tau_i = \text{gap 且 } b_i \leqslant 2r_s+\delta_b.\end{cases} \tag{9}$$

当C_t^{f}为空时, 机器人依次执行安全回退策略: 若上一航点仍满足当前可行性约束, 则继续保持. 否则, 选择最大自由空间方向的短距离临时航点. 若仍无可行方向, 则低速停止并等待下一轮规划.

对于每个可行候选航点$c_i\in C_t^{\text{f}}$, 本文构建综合效用函数

$$\begin{aligned}J_t(c_i) = {}&\omega_g S_g(c_i) + \omega_o S_o(c_i) + \omega_s S_s(c_i) +\\ &\omega_r S_r(c_i) + \omega_h S_h(c_i),\end{aligned} \tag{10}$$

其中各评分项均归一化到$[0,1]$, 且$\omega_g+\omega_o+\omega_s+\omega_r+\omega_h=1$. 本文用$\text{clip}(x,a,b)=\min(\max(x,a),b)$表示截断函数.

目标推进、空间开阔、安全余量、局部可跟踪和航点稳定评分分别定义为

$$S_g(c_i) = \text{clip}(\frac{\|p_t^{xy}-g\|_2 - \|c_i-g\|_2}{\rho_i+\epsilon}, 0, 1), \tag{11}$$

$$S_o(c_i) = \frac{1}{|\Omega_i|}\sum_{k\in\Omega_i}\text{clip}(\frac{d_k - d_{\text{safe}}}{R_{\max}-d_{\text{safe}}}, 0, 1), \tag{12}$$

$$S_s(c_i) = \text{clip}(\frac{D_{\text{path}}(c_i) - r_s}{d_{\text{safe}} - r_s}, 0, 1), \tag{13}$$

$$S_r(c_i) = \exp(-\frac{(\rho_i-\rho_0)^2}{2\sigma_\rho^2})\exp(-\frac{\phi_i^2}{2\sigma_\phi^2}), \tag{14}$$

$$S_h(c_i) = \begin{cases}\exp(-\dfrac{\|c_i - w_{t-1}\|_2^2}{2\sigma_h^2}), & w_{t-1}\text{ 存在},\\ 1, & w_{t-1}\text{ 不存在}.\end{cases} \tag{15}$$

其中, S_g衡量目标推进程度, S_o和S_s分别反映候选方向开阔度和路径安全余量, S_r用于约束局部控制器可跟踪性, S_h用于减少航点跳变.

设初步最优航点为$\hat{w}_t=\arg\max_{c_i\in C_t^{\text{f}}} J_t(c_i)$. 上一时刻航点$w_{t-1}$不受当前候选集合$C_t$限制, 若其在当前观测下重新满足可行性约束, 且满足

$$J_t(w_{t-1}) \geqslant J_t(\hat{w}_t) - \Delta_h, \tag{16}$$

则保持上一航点, 即$w_t^* = w_{t-1}$. 否则更新为$w_t^* = \hat{w}_t$.

1.4 RPER-TD3 局部控制器

上层输出中间航点w_t^*后, 下层局部控制器根据当前 LiDAR 观测和航点相对状态输出连续速度指令. 本文在 TD3[26] 框架中引入复合奖励函数和优先经验回放机制[27], 构建 RPER-TD3 局部控制器. 局部控制器网络结构如图 3 所示.

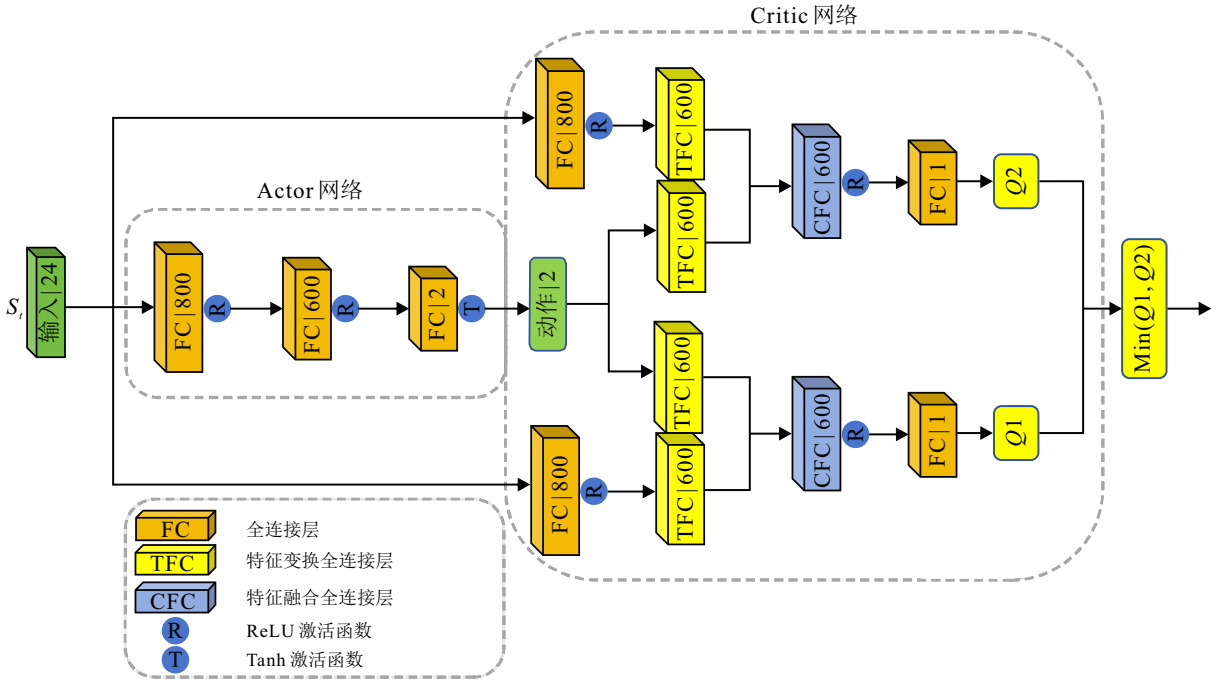


图 3　局部规划网络结构

局部控制器状态定义为

$$s_t = [\hat{l}_t, d_t^w, \alpha_t^w, v_{t-1}, \omega_{t-1}], \tag{17}$$

其中$\hat{l}_t$为 LiDAR 扫描下采样后的距离向量, d_t^w为机器人到当前航点的距离, α_t^w为航点相对方位角, v_{t-1}、ω_{t-1}为上一时刻动作. 动作定义为$a_t = (v_t, \omega_t)$.

奖励函数由终止奖励和密集奖励组成:

$$r_t = \begin{cases} R_g, d_t^w < d_g \text{ 且未发生碰撞}, \\ R_c, d_{\min,t} < d_c, \\ r_t^{\text{dense}}, \text{其他情况}, \end{cases} \tag{18}$$

其中R_g为到达航点奖励, R_c为碰撞惩罚, $d_{\min,t}$为当前 LiDAR 最小测距. 非终止状态下的密集奖励定义为

$$\begin{aligned} r_t^{\text{dense}} = {} & k_p \Delta d_t + k_v v_t A_t I_t - k_\omega |\omega_t| - c_{\text{step}} - \\ & k_s |\omega_t - \omega_{t-1}| - k_o \max(0, d_{\text{safe}} - d_{\min,t}), \end{aligned} \tag{19}$$

其中$\Delta d_t = d_{t-1}^w - d_t^w$表示距离进展, $A_t = \max(0, \cos\alpha_t^w)$为方向一致性因子. 当$\Delta d_t > 0$时$I_t = 1$, 否则$I_t = 0$. 该奖励函数从目标趋近、方向一致性、速度效率、动作平滑和障碍物安全距离等方面约束局部控制行为.

在经验回放阶段, 本文采用基于 TD 误差的优先经验采样. 设经验池容量为N_D, 第i个样本的优先级和采样概率分别为

$$p_i = |\delta_i| + \epsilon_p, \ P(i) = \frac{p_i^{\alpha_{\text{PER}}}}{\sum_{k=1}^{N_D} p_k^{\alpha_{\text{PER}}}}. \tag{20}$$

为修正非均匀采样引入的估计偏差, 采用重要性采样权重

$$\mu_i = \frac{(N_D P(i))^{-\beta_{\text{PER}}}}{\max_\ell (N_D P(\ell))^{-\beta_{\text{PER}}}}. \tag{21}$$

对于采样得到的转移样本$(s_i, a_i, r_i, s_i', m_i)$, 其中$m_i = 1$表示终止状态, TD3 目标动作为

$$\tilde{a}_i = \text{clip}(\pi_{\bar{\eta}}(s_i') + \epsilon_i, a_{\min}, a_{\max}), \tag{22}$$

其中$a_{\min}$和$a_{\max}$分别为动作空间下界和上界, ϵ_i为截断高斯平滑噪声. 目标 Q 值为

$$y_i = r_i + \gamma(1 - m_i) \min_{j=1,2} Q_{\bar{\theta}_j}(s_i', \tilde{a}_i). \tag{23}$$

两个 Critic 网络的加权损失为

$$L_j(\theta_j) = \frac{1}{B} \sum_{i=1}^{B} \mu_i (Q_{\theta_j}(s_i, a_i) - y_i)^2, \ j = 1, 2. \tag{24}$$

Actor 网络采用延迟更新, 其损失函数为

$$L_\pi(\eta) = -\frac{1}{B} \sum_{i=1}^{B} Q_{\theta_1}(s_i, \pi_\eta(s_i)). \tag{25}$$

目标网络采用软更新, 样本优先级根据双 Critic 中较大的 TD 误差更新:

$$\delta_i = \max_{j=1,2} |y_i - Q_{\theta_j}(s_i, a_i)|. \tag{26}$$

通过复合奖励函数与优先经验回放机制的结合, RPER-TD3 能够提高关键经验利用效率, 并增强中间航点跟踪过程中的安全性和稳定性.

1.5 上下层协同与算法流程

RPER-TD3 控制器首先在随机局部航点跟踪环境中进行离线训练. 训练完成后, 在完整目标驱动探索测试阶段固定 Actor 网络参数, 控制器仅根据当前状态输出控制动作, 不再更新经验回放池和网络参数. iGDAE 的在线执行流程如算法 1 所示.

算法1 iGDAE在线执行流程

输入: 全局目标g, 初始位姿p_0, 初始地图M_0, 训练好的actor网络π_η.

输出: 控制动作序列$\{a_t\}$和任务执行结果.

初始化: $t=0$, 上一中间航点$w_{\text{pre}}=\varnothing$.

while 未到达目标区域且未触发终止条件**do**

获取当前位姿p_t和LiDAR观测D_t, 并由在线SLAM更新地图M_t.

根据D_t生成候选航点集合C_t.

根据自由空间、路径安全、可跟踪范围和通道宽度等约束, 构造可行候选集合C_t^f.

if $C_t^f=\varnothing$ **then**

按上述顺序执行安全回退策略, 得到临时航点w_t^*.

else

计算各可行候选航点的综合效用$J_t(c_i)$.

结合上一中间航点w_{pre}和滞回机制确定w_t^*.

end if

根据当前观测和航点相对状态构造局部控制状态s_t.

由actor网络输出控制动作$a_t=\pi_\eta(s_t)$, 并执行该动作.

更新$w_{\text{pre}}\leftarrow w_t^*$, $t\leftarrow t+1$.

end while

return 控制动作序列$\{a_t\}$和任务执行结果.

2 仿真实验与结果分析

为验证所提 iGDAE 方法在未知环境目标驱动探索任务中的有效性, 本文基于 ROS/Gazebo 构建移动机器人仿真实验平台, 并从局部航点跟踪和完整目标驱动探索两个层面进行实验. 首先, 在随机局部航点跟踪环境中比较不同 TD3 控制器的训练性能, 以验证 RPER-TD3 局部控制器的有效性. 随后, 在完整未知环境目标驱动探索任务中, 将本文方法与纯局部控制、传统地图规划和分层基线方法进行对比, 评价其任务可靠性、导航效率和运行稳定性.

2.1 仿真平台与实验设置

仿真实验采用两轮差速驱动移动机器人模型, 机器人搭载二维 LiDAR 作为主要环境感知传感器. Gazebo 用于模拟机器人运动、传感器观测和障碍物交互过程, ROS 用于实现 LiDAR 数据订阅、在线 SLAM、航点生成与评价、局部控制以及速度指令发布. 机器人控制输入为线速度v和角速度ω, 局部控制器根据当前 LiDAR 观测、航点相对距离、航点相对方位角以及上一时刻动作输出连续控制指令.

局部控制器训练环境如图 4 所示. 训练过程中随机生成机器人初始位姿、局部目标点和障碍物布局, 使控制器在不同局部几何结构下学习目标趋近与避障行为. 当机器人到达局部目标点、发生碰撞或达到最大步数时, 当前训练回合结束.

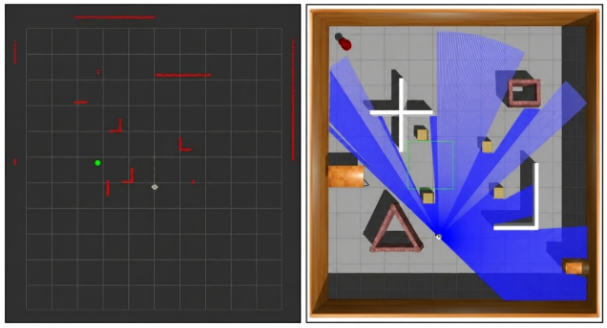


图 4 局部规划器随机训练环境

完整目标驱动探索实验采用四个未知地图, 如图 5 所示. 四个地图由简单到复杂, 包含开阔区域、长走廊、岔路口、狭窄通道、局部遮挡和中心障碍物等典型结构, 用于考察不同方法在目标推进、局部绕行、通道穿越和航点切换等场景下的表现. 所有方法均在无先验地图条件下从指定起点出发, 依靠在线 SLAM 构建增量地图. 四个地图采用相同的机器人模型、LiDAR 参数及方法参数, 速度约束和最大运行时间保持一致.

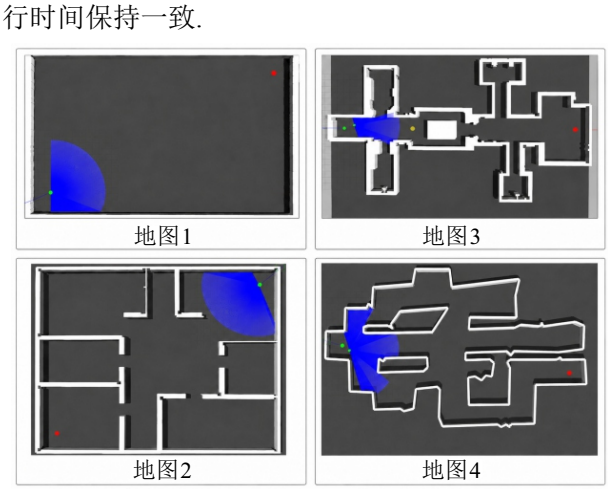


图 5 完整目标驱动探索仿真环境

2.2　对比方法与评价指标

为分析复合奖励函数和优先经验回放机制对局部航点跟踪性能的影响, 本文设置四组局部控制器进行对比: TD3 采用原始 TD3 算法和基础奖励函数, 并使用均匀经验回放. PER-TD3 在 TD3 基础上引入优先经验回放机制. NR-TD3 采用本文设计的复合奖励函数, 但仍使用均匀经验回放. RPER-TD3 同时采用复合奖励函数和优先经验回放机制. 四组控制器采用相同的网络结构、状态输入、动作输出、训练环境和基础超参数, 仅改变奖励函数和经验采样机制.

完整目标驱动探索实验设置五组方法进行对比: Pure-TD3 直接以全局目标作为局部控制目标, 不设置中间航点. Pure-Navfn 采用传统 Navfn 方法基于在线地图进行目标导航. Navfn-TD3 由 Navfn 提供导航引导, 并由 TD3 控制器执行局部连续控制. GDAE-TD3 采用分层目标驱动探索基线方法生成中间航点, 并由 TD3 控制器完成局部航点跟踪. iGDAE 为本文完整方法, 上层采用可行性约束多目标航点评估选择中间航点, 下层采用 RPER-TD3 控制器完成局部航点跟踪. 完整探索测试阶段, 所有学习型控制器均固定网络参数, 不再更新经验回放池和网络权重.

每种方法在四个地图上分别重复运行 10 次, 共 40 次实验. 若机器人在限定时间内到达目标区域, 则判定为成功. 若发生碰撞、长时间停滞或运行时间超过 300 s, 则判定为失败. 为兼顾导航效率与任务可靠性, 本文采用失败惩罚下的全样本任务代价统计: 成功样本计入实际完成时间和实际路径长度, 失败样本按统一任务预算计入 $T_{\max}=300\,\mathrm{s}$ 和 $L_{\max}=v_{\max}T_{\max}=300\,\mathrm{m}$, 其中 $L_{\max}$ 仅表示失败样本的路径代价上限, 不作为其真实行驶距离解释. 评价指标包括任务成功次数、平均时间代价、平均路径代价及其标准差. 设全部实验集合为 S, 实验次数为 N_s, 第 i 次实验用于统计的完成时间和路径长度分别为 T_i 和 L_i, 则

$$T_i=\begin{cases}T_i^{\mathrm{real}}, & \text{成功},\\ T_{\max}, & \text{失败},\end{cases}\quad L_i=\begin{cases}L_i^{\mathrm{real}}, & \text{成功},\\ L_{\max}, & \text{失败},\end{cases}$$

其中 T_i^{real} 和 L_i^{real} 分别为成功实验的实际完成时间和实际行驶路径长度. 则

$$\bar{T}=\frac{1}{N_s}\sum_{i\in S}T_i,\ \bar{L}=\frac{1}{N_s}\sum_{i\in S}L_i, \tag{27}$$

$$\sigma_T=\sqrt{\frac{1}{N_s-1}\sum_{i\in S}(T_i-\bar{T})^2},$$

$$\sigma_L=\sqrt{\frac{1}{N_s-1}\sum_{i\in S}(L_i-\bar{L})^2}. \tag{28}$$

2.3　局部控制器训练过程与稳定性分析

图 6 给出了四种局部控制器在随机航点跟踪环境中的训练曲线, 其中图 6(a) 为评估阶段平均回合奖励, 图 6(b) 为训练过程平均 Q 值图, 图中曲线均未进行额外平滑处理. TD3 与 PER-TD3 采用基础奖励函数, NR-TD3 与 RPER-TD3 采用本文设计的复合奖励函数. 由于不同奖励函数的组成和数值尺度存在差异, 本文不对两类奖励结构下的奖励值和 Q 值进行直接比较, 而主要分析其收敛趋势, 并在相同奖励结构下比较优先经验回放机制的作用.

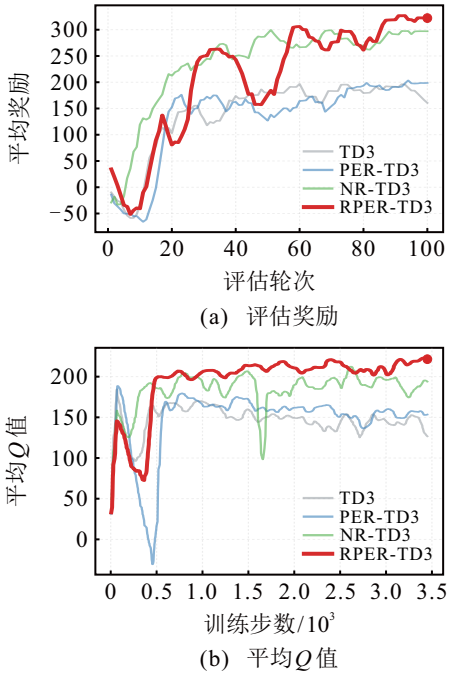
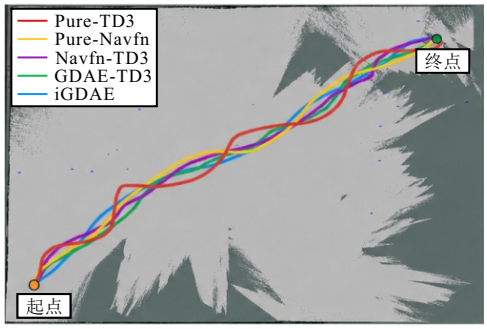


图 6　局部控制器训练过程中奖励值与平均 Q 值变化曲线

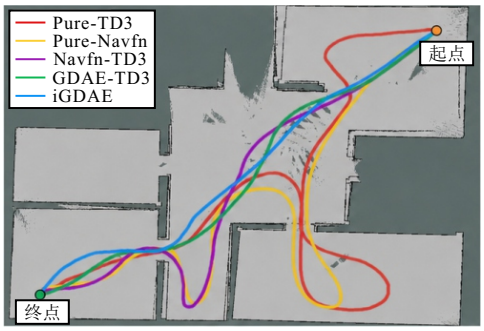
在基础奖励函数条件下, PER-TD3 相较于 TD3 在训练中后期表现出较好的回报变化趋势. 在复合奖励函数条件下, RPER-TD3 相较于 NR-TD3 获得了较高且较稳定的评估回报, 表明优先经验回放有助于提高关键样本的利用效率并改善策略训练过程. 图 6(b) 中, 各控制器的平均 Q 值在训练后期波动幅度逐渐减小, 表明价值估计过程趋于稳定. RPER-TD3 在完整目标驱动探索任务中的实际作用, 将结合第 2.5 节的消融实验进一步分析.

2.4　完整目标驱动探索结果分析

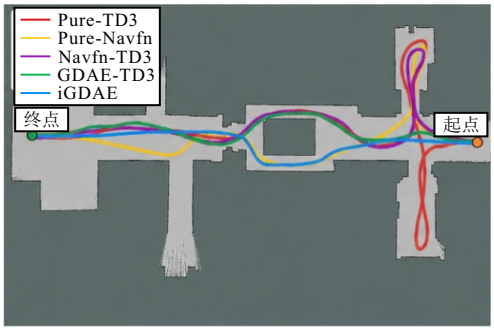
图 7 给出了五种方法在四个测试地图中的典型运行轨迹, 表 1 给出了 40 次重复实验在失败惩罚统计口径下的统计结果. 从轨迹结果可以看出, Pure-TD3 缺乏中间航点引导, 在复杂地图中容易出现往复运动、绕行过长或停滞. Pure-Navfn 依赖在线



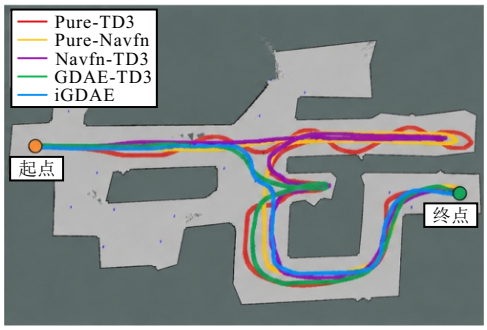
(a) 地图1



(b) 地图2



(c) 地图3



(d) 地图4

图 7 5 种方法在 4 个测试地图中的典型运行轨迹

SLAM 地图, 在地图尚不完整时易受未知区域和局部建图误差影响. Navfn-TD3 虽结合了地图规划和学习型局部控制, 但整体效果仍受上层规划质量限制. GDAE-TD3 通过中间航点提升了目标推进能力, 但在岔路、狭窄通道和障碍物密集区域仍存在航点切换不稳定和绕行偏长问题.

由表 1 可知, Pure-TD3、Pure-Navfn、Navfn-TD3 和 GDAE-TD3 的总成功次数分别为 21/40、25/40、31/40 和 36/40, 而 iGDAE 在四个地图中均取得 10/10 的成功次数, 总成功次数达到 40/40. 由于 iGDAE 未出现失败, 其时间代价和路径代价等同于实际完成时间和实际行驶路径. 其他方法的均值则同时反映成功效率和失败代价. iGDAE 的全样本平均时间代价为 101.26 s, 全样本平均路径代价为 63.12 m, 均为各方法最优. 相较于分层基线方法 GDAE-TD3, iGDAE 的平均时间代价降低 19.47%, 平均路径代价降低 34.96%. 同时, 时间代价标准差和路径代价标准差分别降至 42.72 s 和 24.67 m. 结果表明, 所提方法不仅提高了任务成功率和导航效率, 也增强了不同复杂度地图下的运行稳定性.

2.5 消融实验分析

为进一步分析上层可行性约束航点评估和下层 RPER-TD3 控制器对完整目标驱动探索性能的影响, 本文在复杂度较高的地图 3 和地图 4 上设置消融实验. 对比方法包括: GDAE-TD3, 即原 GDAE 航点选择与 TD3 局部控制器组合. GDAE-RPER-TD3, 即保持 GDAE 航点选择不变, 仅将下层控制器替换为 RPER-TD3. iGDAE-TD3, 即采用本文可行性约束多目标航点评估, 但下层仍使用 TD3. iGDAE 为本文完整方法. 各方法在两个地图上分别运行 10 次, 共 20 次实验, 并同样采用第 2.2 节定义的失败惩罚统计口径.

由表 2 可知, 仅将下层控制器由 TD3 替换为 RPER-TD3 后, GDAE-RPER-TD3 相比 GDAE-TD3 的平均时间代价和平均路径代价均有所降低, 说明复合奖励函数与优先经验回放能够改善局部航点跟踪性能. 但其在地图 4 中仍存在失败情况, 表明仅改进局部控制器难以完全解决复杂环境中的航点选择问题. 相比之下, iGDAE-TD3 在两个复杂地图中均

表 1 不同方法在四个地图上的完整目标驱动探索结果对比 ($T_{\max}=300\,\mathrm{s}$, $L_{\max}=300\,\mathrm{m}$)

方法	平均时间代价/s	平均路径代价/m	时间代价标准差/s	路径代价标准差/m	地图1成功次数	地图2成功次数	地图3成功次数	地图4成功次数	总成功次数
Pure-TD3	186.19	167.20	113.43	128.03	10/10	6/10	3/10	2/10	21/40
Pure-Navfn	175.72	148.49	105.39	120.37	10/10	8/10	3/10	4/10	25/40
Navfn-TD3	143.59	109.17	97.70	105.50	10/10	10/10	5/10	6/10	31/40
GDAE-TD3	125.74	97.05	79.88	80.75	10/10	10/10	8/10	8/10	36/40
iGDAE	**101.26**	**63.12**	**42.72**	**24.67**	10/10	10/10	**10/10**	**10/10**	**40/40**

表 2　复杂地图中的消融实验结果 ($T_{\max}=300\,\mathrm{s}$, $L_{\max}=300\,\mathrm{m}$)

方法	平均时间代价/s	时间代价标准差/s	平均路径代价/m	路径代价标准差/m	地图3成功次数	地图4成功次数	总成功次数
GDAE-TD3	177.055	86.815	146.546	90.540	8/10	8/10	16/20
GDAE-RPER-TD3	136.810	85.514	101.794	90.881	9/10	8/10	17/20
iGDAE-TD3	139.740	79.229	78.478	47.520	10/10	10/10	20/20
iGDAE	**118.160**	**46.471**	**70.768**	**27.577**	**10/10**	**10/10**	**20/20**

取得 10/10 的成功次数, 且路径代价明显降低, 说明可行性约束多目标航点评估能够提高中间航点的安全性、可达性和稳定性. 在此基础上, 完整 iGDAE 进一步将平均时间代价降低至 118.160 s, 平均路径代价降低至 70.768 m, 表明上层可行性航点评估与下层 RPER-TD3 控制器具有协同作用, 能够进一步提升复杂未知环境中的导航效率和任务稳定性.

3　结　论

本文面向给定目标坐标的未知环境目标驱动探索任务, 提出一种可行航点引导的分层方法 iGDAE. 该方法基于 LiDAR 生成候选航点, 综合考虑自由空间、路径安全、局部可跟踪性、通道宽度和航点滞回等因素完成中间航点选择, 并采用 RPER-TD3 控制器实现连续航点跟踪与局部避障. 仿真结果表明, iGDAE 在 4 个未知仿真环境的 40 次测试中均成功到达目标. 相较于 GDAE-TD3, 其平均时间代价和平均路径代价分别降低 19.47% 和 34.96%, 表明可行性约束航点评估与局部强化学习控制的协同能够提高目标驱动探索的任务可靠性、导航效率和运行稳定性. 后续将基于实体移动机器人平台开展实验, 进一步验证所提方法在传感器噪声、建图误差及真实复杂环境下的有效性和鲁棒性.

参考文献 (References)

[1] Brugali D, Muratore L, De Luca A. Mobile robots exploration strategies and requirements: A systematic mapping study[J]. The International Journal of Robotics Research, 2025, 44(9): 1461-1506.

[2] 常路, 单梁, 张伟龙, 等. 未知环境下基于激光雷达的启发式导航算法[J]. 控制与决策, 2024, 39(9): 2903-2912.
(Chang L, Shan L, Zhang W L, et al. Heuristic navigation based on LiDAR in unknown environments[J]. Control and Decision, 2024, 39(9): 2903-2912.)

[3] 曹路阳, 周乐来, 戴骁蒙, 等. 基于分层边界与可视图的移动机器人自主探索算法[J]. 控制与决策, 2025, 40(4): 1207-1216.
(Cao L Y, Zhou L L, Dai X M, et al. An autonomous exploration algorithm of mobile robots based on hierarchical frontier and visibility graph[J]. Control and Decision, 2025, 40(4): 1207-1216.)

[4] Zhou B Y, Zhang Y C, Chen X Y, et al. FUEL: Fast UAV exploration using incremental frontier structure and hierarchical planning[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 779-786.

[5] Cao C, Zhu H, Choset H, et al. TARE: A hierarchical framework for efficiently exploring complex 3D environments[C]. Proceedings of Robotics: Science and Systems XVII. Virtual, 2021: p018.

[6] Zhu H B, Cao C, Xia Y K, et al. DSVP: Dual-stage viewpoint planner for rapid exploration by dynamic expansion[C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, 2021: 7623-7630.

[7] Cao C, Zhu H B, Yang F, et al. Autonomous exploration development environment and the planning algorithms[C]. International Conference on Robotics and Automation. Philadelphia, 2022: 8921-8928.

[8] 王童, 李骜, 宋海荦, 等. 基于分层深度强化学习的移动机器人导航方法[J]. 控制与决策, 2022, 37(11): 2799-2807.
(Wang T, Li A, Song H L, et al. Navigation method for mobile robot based on hierarchical deep reinforcement learning[J]. Control and Decision, 2022, 37(11): 2799-2807.)

[9] 户高铭, 蔡克卫, 王芳, 等. 基于深度强化学习的无地图移动机器人导航[J]. 控制与决策, 2024, 39(3): 985-993.
(Hu G M, Cai K W, Wang F, et al. Mapless navigation based on deep reinforcement learning for mobile robots[J]. Control and Decision, 2024, 39(3): 985-993.)

[10] 刘京奇, 韩奇松, 胡春鹤, 等. 基于动态演化的机器人安全强化学习运动规划方法[J]. 控制与决策, 2026, 41(8): 2332-2344.
(Liu J Q, Han Q S, Hu C H, et al. Dynamic evolution-based safe reinforcement learning for robotic motion planning[J]. Control and Decision, 2026, 41(8): 2332-2344.)

[11] 刘永康, 张严心, 柳向斌, 等. 基于自适应控制障碍函数的安全强化学习路径规划框架[J]. 控制与决策, 2026, 41(8): 2345-2352.
(Liu Y K, Zhang Y X, Liu X B, et al. An adaptive control barrier function-based safe reinforcement learning framework for path planning[J]. Control and Decision, 2026, 41(8): 2345-2352.)

[12] Dobrevski M, Skočaj D. Deep reinforcement learning for map-less goal-driven robot navigation[J]. International Journal of Advanced Robotic Systems, 2021, 18(1): 1729881421992621.

[13] Wu K Y, Wang H, Esfahani M A, et al. Learn to navigate autonomously through deep reinforcement

learning[J]. IEEE Transactions on Industrial Electronics, 2022, 69(5): 5342-5352.

[14] Miranda V R F, Neto A A, Freitas G M, et al. Generalization in deep reinforcement learning for robotic navigation by reward shaping[J]. IEEE Transactions on Industrial Electronics, 2024, 71(6): 6013-6020.

[15] Zhu Z D, Lin K X, Jain A K, et al. Transfer learning in deep reinforcement learning: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13344-13362.

[16] Patel U, Kumar N K S, Sathyamoorthy A J, et al. DWA-RL: Dynamically feasible deep reinforcement learning policy for robot navigation among mobile obstacles[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 6057-6063.

[17] Li P, Chen D H, Wang Y C, et al. Path planning of mobile robot based on improved TD3 algorithm in dynamic environment[J]. Heliyon, 2024, 10(11): e32167.

[18] Cimurs R, Suh I H, Lee J H. Goal-driven autonomous exploration through deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 730-737.

[19] Zhang Z W, Shi C H, Zhu P M, et al. Autonomous exploration of mobile robots via deep reinforcement learning based on spatiotemporal information on graph[J]. Applied Sciences, 2021, 11(18): 8299.

[20] Cao Y H, Hou T X, Wang Y Z, et al. ARiADNE: A Reinforcement learning approach using Attention-based Deep Networks for Exploration[C]. IEEE International Conference on Robotics and Automation. London, 2023: 10219-10225.

[21] Chaplot D S, Gandhi D, Gupta S, et al. Learning to explore using active neural SLAM[C]. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, 2020.

[22] Jang Y, Baek J, Han S. Hindsight intermediate targets for mapless navigation with deep reinforcement learning[J]. IEEE Transactions on Industrial Electronics, 2022, 69(11): 11816-11825.

[23] Zhu W, Hayashibe M. A hierarchical deep reinforcement learning framework with high efficiency and generalization for fast and safe navigation[J]. IEEE Transactions on Industrial Electronics, 2023, 70(5): 4962-4971.

[24] Chai R Q, Niu H L, Carrasco J, et al. Design and experimental validation of deep reinforcement learning-based fast trajectory planning and control for mobile robot in unknown environment[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5778-5792.

[25] Cai B L, Wei C Y, Ji Z. Deep reinforcement learning with multiple unrelated rewards for AGV mapless navigation[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 4323-4340.

[26] 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.

[27] Schaul T, Quan J, Antonoglou I, et al. Prioritized experience replay[C]. Proceedings of the 4th International Conference on Learning Representations. San Juan, 2016.

作者简介

李艺博 (2002−), 男, 硕士生, 主要研究方向为移动机器人导航、深度强化学习, E-mail: liyibo@stumail.ysu.edu.cn;

关哲 (1988−), 男, 讲师, 博士, 主要研究方向为数据驱动理论及其应用、工程机械控制、过程控制, E-mail: guan@ysu.edu.cn;

郝曼 (1994−), 女, 讲师, 博士, 主要研究方向为人机交互、意图理解、疲劳检测与情感识别、多模态信息融合, E-mail: haoman@ysu.edu.cn;

丁伟利 (1979−), 女, 教授, 博士, 博士生导师, 主要研究方向为机器人智能感知与控制、计算机视觉、模式识别与空间智能, E-mail: weiye51@ysu.edu.cn;

华长春 (1979−), 男, 教授, 博士, 博士生导师, 主要研究方向为非线性动力系统的控制及应用、网络化控制系统的分析与综合、基于数据驱动的故障诊断和容错控制, E-mail: cch@ysu.edu.cn.