控制与决策
Control and Decision

基于非对称强化学习的移动机器人自主导航算法研究

何乃峰, 杨忠, 许诺, 卓浩泽, 徐宏雨, 陈凯

引用本文:
何乃峰， 杨忠， 许诺， 等. 基于非对称强化学习的移动机器人自主导航算法研究[J]. 控制与决策， 2026， 41(6): 1614-1624.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0787

您可能感兴趣的其他文章

Articles you may be interested in

基于改进RRT*FN算法的机器人路径规划

Robot path planning based on improved RRT*FN algorithm

控制与决策. 2021, 36(8): 1834–1840 https://doi.org/10.13195/j.kzyjc.2019.1713

移动机器人运动规划中的深度强化学习方法

Deep reinforcement learning for motion planning of mobile robots

控制与决策. 2021, 36(6): 1281–1292 https://doi.org/10.13195/j.kzyjc.2020.0470

基于 ± 3 σ 正态概率区间分族遗传蚁群算法的移动机器人路径规划

Path planning of mobile robot based on ± 3 σ normal probability interval population division using genetic ant–colony algorithm

控制与决策. 2021, 36(12): 2861–2870 https://doi.org/10.13195/j.kzyjc.2020.0745

机器人信息增益RRT环境探索算法

Robot RRT based on information gain for environment exploration

控制与决策. 2021, 36(11): 2683–2689 https://doi.org/10.13195/j.kzyjc.2020.1007

一种基于免疫机理的确定性移动机器人路径规划算法

A path planning algorithm of deterministic mobile robot based on immune mechanism

控制与决策. 2021, 36(10): 2418–2426 https://doi.org/10.13195/j.kzyjc.2020.0059

基于非对称强化学习的移动机器人自主导航算法研究

何乃峰[1]，杨 忠[1†]，许 诺[1,2]，卓浩泽[1,3]，徐宏雨[1]，陈 凯[1]

(1. 南京航空航天大学 自动化学院，南京 210000；
2. 广东电网有限责任公司东莞供电局 东南区供电局，广东 东莞 523000；
3. 广西电网有限责任公司电力科学研究院 广西电力装备智能控制与运维重点实验室，南宁 530000)

摘 要: 针对动态非结构化环境中移动机器人感知不确定性与策略泛化能力不足的挑战, 提出一种基于非对称强化学习的鲁棒自主导航策略优化框架 (RANav). 该方法融合隐式环境估计、域随机化与非对称强化学习机制, 提升机器人对动态环境的建模与决策能力. 首先, 构建多模态融合的隐式环境估计网络, 以精确提取动态障碍物特征并提升场景表征能力; 其次, 引入基于行为域随机化机制, 提升策略的 Sim-to-Real 迁移能力; 最后, 采用非对称近端策略优化 (PPO) 算法, 利用特权信息优化 Critic 网络提升策略学习效率. 在多组仿真与真实场景实验中, RANav 在导航成功率、避障鲁棒性与路径效率方面均显著优于现有方法, 充分验证了其在复杂非结构环境中的鲁棒泛化能力与实际部署潜力.

关键词: 移动机器人；自主导航；避障；强化学习；环境理解；域随机化

中图分类号: TP242 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0787

引用格式: 何乃峰, 杨忠, 许诺, 等. 基于非对称强化学习的移动机器人自主导航算法研究 [J]. 控制与决策, 2026, 41(6): 1614-1624.

Autonomous navigation algorithm for mobile robots based on asymmetric reinforcement learning

HE Nai-feng[1]，*YANG Zhong*[1†]，*XU Nuo*[1,2]，*ZHUO Hao-ze*[1,3]，*XU Hong-yu*[1]，*CHEN Kai*[1]

(1. College of Automation， Nanjing University of Aeronautics and Astronautics， Nanjing 210000， China；
2. Southeast Power Supply Bureau， Dongguan Power Supply Bureau， Guangdong Power Grid Co.， Ltd.， Dongguan 523000， China； 3. Guangxi Key Laboratory of Intelligent Control and Maintenance of Power Equipment，Electric Power Research Institute of Guangxi Power Grid Co.，Ltd.，Nanning 530000，China)

Abstract: To address the challenges of perceptual uncertainty and limited policy generalization in dynamic unstructured environments, this paper proposes a robust autonomous navigation policy optimization framework based on asymmetric reinforcement learning, termed Robust Asymmetric Navigation (RANav). The framework integrates implicit environment estimation, domain randomization, and asymmetric reinforcement learning to enhance the robot's modeling and decision-making capabilities in dynamic settings. Specifically, a multimodal implicit environment estimation network is designed to accurately extract dynamic obstacle features and improve scene representation. A behavior-driven domain randomization mechanism is introduced to facilitate Sim-to-Real policy transfer. Finally, an asymmetric proximal policy optimization (PPO) algorithm is employed, where privileged information is provided to the Critic network during training to improve policy learning efficiency. Extensive simulations and real-world experiments demonstrate that the RANav significantly outperforms existing methods in terms of navigation success rate, obstacle avoidance robustness, and path efficiency, verifying its strong generalization and deployment potential in complex unstructured environments.

Keywords: mobile robots； autonomous navigation； obstacle avoidance； reinforcement learning； environmental understanding；domain randomization

收稿日期: 2025-07-25；**录用日期:** 2025-11-07.
基金项目: 广西电网公司 2024 年科技创新专业科技项目 (GXKJXM20240152).
责任编委: 张文安.
†通信作者. E-mail: yangzhong@nuaa.edu.cn.

0 引 言

随着人工智能与机器人技术的不断演进, 移动机器人在灾后救援[1]、城市物流[2] 与日常服务[3] 等领域展现出广阔的应用前景. 在实际部署中, 机器人往往需要在动态、非结构化的环境中完成自主导航任务, 例如穿越密集的交通流、复杂室内空间或多智能体系统. 这类场景具有环境状态不可预测、交互对象行为不确定、传感器观测受限等特点, 严重影响了机器人感知与决策系统的稳定性与泛化能力[4-5]. 因此, 如何实现移动机器人在动态环境中的鲁棒感知、环境建模与策略学习, 已成为机器人自主导航研究的核心科学问题之一.

近年来, 深度强化学习 (deep reinforcement learning, DRL) 在机器人导航中得到广泛关注, 凭借其端到端的策略学习能力和非线性环境建模能力, 在静态或规则环境中取得显著成果. 然而, DRL 策略通常依赖仿真环境训练, 难以覆盖现实环境中的行为多样性与感知噪声, 导致显著的“仿真-现实迁移差距” (sim-to-real gap)[6]. 此外, 导航策略多依赖局部、有限的感知信息生成动作, 缺乏对动态环境中潜在结构与交互模式的理解能力, 进一步制约了策略的稳健性与适应性.

为解决上述问题, 研究者从多个方向展开探索. 一方面, 传统导航方法尝试通过规则建模与路径优化提升动态避障能力. 例如, Hoang 等[7] 提出基于时间弹性路径优化方法; Dai 等[8] 结合 Informed-RRT* 与 DWA 实现高效路径生成; Qi 等[9] 引入递归速度障碍模型并设计优先级机制; 此外, MPC [10] 与社会力模型 (social force model, SFM)[11] 也被广泛应用于动态避障与多智能体交互建模. 这些方法在特定环境中具备良好性能, 但其规则设计往往依赖场景先验, 泛化能力有限. 另一方面, 基于学习的环境感知与建模方法在提升策略适应性方面展现出显著潜力. 典型方法如 TerraPN[12]、BADGR[13] 通过监督或自监督方式对环境结构建模, 并以端到端框架提高感知决策效率; 生物结构建模[14] 与激光雷达驱动的自监督分割方法[15] 则增强动态导航中的几何理解能力; PONI[16] 和 TSNav[17] 引入多模态融合机制, 进一步提升在动态场景下的感知鲁棒性. 尽管如此, 这类方法通常依赖大量静态标注数据或高质量传感输入, 难以适应现实世界中目标行为变化快、观测信息不完备等复杂条件, 导致泛化能力与实际部署性能受限[18].

针对动态环境中的感知误差与策略泛化问题, 研究者引入多种结构性改进手段: 例如, 域随机化机制 (domain randomization, DR)[19-20] 通过扰动仿真环境参数构造多样化训练场景, 以缓解部署阶段策略性能退化; 非对称强化学习框架[21] 则在训练阶段引入特权信息提升 Critic 网络的评估能力, 从而间接增强策略表现; 此外, 激光优化建模[22]、时空注意力机制[23]、图神经网络[24] 与 Transformer 架构[25] 等新兴结构也被用于增强导航策略对环境动态变化的建模能力. 同时, 面向异构多智能体碰撞规避的问题, 也有方法引入定向胶囊网络与风险度量机制[26]. 尽管上述方法在各自领域取得进展, 但在信息不完备、交互不可预测的复杂动态环境中, 如何实现感知-决策一体化的高效建模以及策略的跨场景鲁棒泛化, 仍是当前领域亟待解决的关键难题.

鉴于上述挑战与现有方法的局限性, 本文致力于解决动态非结构环境中的鲁棒自主导航问题. 为此, 提出一种面向动态非结构环境的鲁棒自主导航框架 RANav, 融合隐式环境估计、行为域随机化与非对称 PPO 机制, 旨在系统提升导航策略在动态环境中的感知建模能力、策略鲁棒性与部署泛化能力. 具体贡献如下:

1) 提出隐式环境估计网络, 从异构传感器数据中提取障碍物语义特征, 提升环境建模精度;

2) 引入行为域随机化机制, 构建多样化仿真环境以缓解 Sim-to-Real 迁移问题;

3) 基于非对称 PPO 结构, 训练阶段利用特权信息提升 Critic 评估能力, 优化 Actor 策略.

本文结构如下: 第 1 节介绍问题建模与系统架构; 第 2 节详述所提各子模块设计; 第 3 节展示仿真与真实世界实验及消融研究; 第 4 节总结概括全文.

1 问题建模与方法框架

为实现移动机器人在复杂动态环境中的鲁棒自主导航, 本文将导航任务建模为基于马尔可夫决策过程 (Markov decision process, MDP) 的强化学习问题. 首先对任务目标及状态变量进行形式化描述, 随后基于 MDP 框架定义学习问题, 最后阐述系统整体的感知-决策-控制架构设计.

1.1 问题建模

如图 1 所示, 本文关注复杂动态非结构化环境下的移动机器人自主导航问题. 这类场景普遍具有高动态交互特性, 例如复杂交通流、多智能体系统或密集人群, 为机器人导航带来严峻挑战. 该场景下的核心挑战在于环境结构的理解能力与导航策略的泛化能力[27]. 传统感知模块通常仅抽取障碍物的几何

信息 (如距离或边缘特征), 难以构建反映动态对象分布模式、潜在意图及复杂交互的表示, 限制策略对场景多变性的适应性, 难以实现真正智能的避障行为[28]. 此外, 强化学习策略常因仿真中动态对象行为模型简化且缺乏多样性, 导致策略在仿真环境表现优异, 但在面对真实环境中动态对象行为的高随机性与复杂交互时, 泛化能力不足, 出现“仿真过拟合”现象, 严重影响 Sim-to-Real 迁移性能[29].

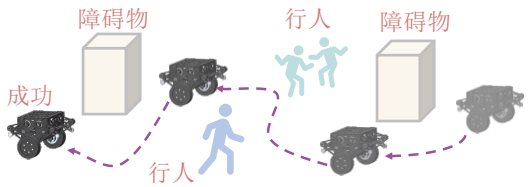


图1 机器人在复杂动态环境中导航示意图

综上, 本文关注的核心问题是: 如何有效增强其对动态环境中特征的感知与理解能力, 并通过系统化训练, 提升导航策略在动态非结构化环境中的泛化性与鲁棒性, 实现安全高效的自主导航.

1.2 马尔可夫决策过程建模

为形式化表达自主导航任务, 本文将其建模为离散时间的马尔可夫决策过程, 定义为五元组$\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma)$. 其中: $\mathcal{S}$表示状态空间, 涵盖机器人自身状态 (如位置、速度、朝向)、导航目标位置及激光雷达、深度相机等传感器获取的环境信息; $\mathcal{A}$为动作空间, 定义为机器人连续控制输入, 包括线速度$v\in\mathbb{R}$与角速度$\omega\in\mathbb{R}$; 状态转移函数$\mathcal{P}(s'|s,a)$描述在状态s下执行动作a后转移至状态s'的概率分布, 受机器人运动学及环境动态共同影响; 奖励函数$\mathcal{R}(s,a)$衡量状态-动作对的即时反馈, 综合反映导航效率、避障鲁棒性及运动平稳性; 折扣因子$\gamma\in(0,1)$调节未来奖励对当前决策的影响程度.

在该 MDP 框架下, 策略函数$\pi_\theta(a|s)$由参数θ控制, 优化目标为最大化期望累积折扣奖励, 即

$$\pi^*=\arg\max_{\pi_\theta}\mathrm{E}\Big[\sum_{t=0}^{T}\gamma^t\mathcal{R}(s_t,a_t)\Big]. \tag{1}$$

该建模为后续基于强化学习的策略训练奠定了理论基础, 尽管实际部署中传感器观测存在噪声和信息不完全, 但系统假设机器人可获得足够的状态表征以指导最优动作选择.

1.3 导航系统框架

基于上述任务建模, 本文设计端到端鲁棒导航系统, 其总体架构如图 2 所示, 包含 4 个核心模块.

机器人的自主导航过程遵循多阶段信息处理与决策流程. 该流程首先起始于环境与机器人自身状态的原始观测数据采集, 由深度相机、激光雷达、惯性测量单元 (inertial measurement unit, IMU)、里程计、目标点及障碍物信息等各类传感器完成. 这些原始数据随后经由预处理模块, 执行特征提取、数据融合及状态估计, 并产生历史信息以及训练阶段所需的特权信息, 为后续决策提供可靠输入. 在此基础上, 环境估计模块中的隐式建模部分, 从预处理后的时序观测中实现障碍物检测与场景理解及编码, 旨在提取动态障碍物的关键特征; 通过多模态信息融合, 这些特征被压缩为紧凑的特征表示, 增强系统对环境的理解能力. 进而, 策略学习通过神经网络模块采用非对称 PPO 架构得以实现: 在训练阶段, Critic 网络被授权访问预处理模块提供的特权信息, 以精确评估价值函数; 而 Actor 网络则仅基于可观测信息输出动作, 从而确保训练与部署之间的一致性. 最终, 策略输出被转换为机器人实际可执行的线速度与角速度指令, 驱动机器人完成实时运动控制.

此外, 训练过程中引入域随机化策略, 通过扰动仿真中动态障碍物行为模型 (速度范围、反应延迟、协作行为等), 提升策略在现实环境中的泛化能力. 该设计确保训练与部署之间的平滑迁移, 增强系统的实用价值. 具体网络结构与训练算法将在第 3 节

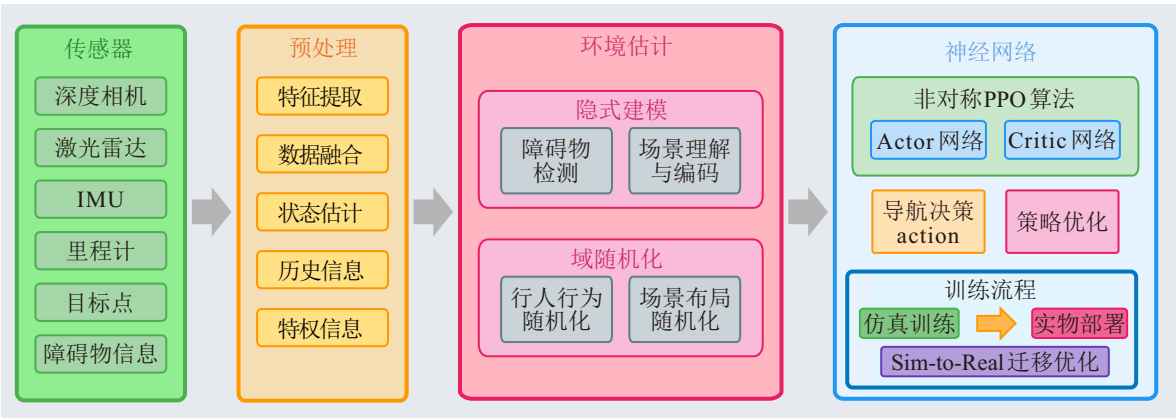


图2 所提出自主导航系统框架

详细展开.

2 方 法

本节详细介绍所提出的机器人鲁棒自主导航决策框架, 旨在使移动机器人能够在复杂、动态且高度不确定的非结构化环境中实现自主导航. 通过融合隐式环境估计、域随机化及非对称 PPO 策略优化, 系统性解决动态环境理解不足、仿真到现实迁移瓶颈以及不确定性下的决策挑战等关键问题.

2.1 系统概述

本文提出一种面向动态非结构化环境的鲁棒导航决策框架. 为促进策略的高效学习与泛化能力提升, 系统构建多模态状态表征结构, 结合训练阶段的特权信息辅助机制, 并采用归一化动作输出方式.

2.1.1 状态空间

机器人状态空间$\mathcal{S}$旨在全面捕捉自身动态状态及局部环境信息以支持策略决策. 具体包括机器人当前线速度$v_r \in [v_{\min}, v_{\max}]$、角速度$\omega_r \in [\omega_{\min}, \omega_{\max}]$, 以及相对目标点的欧氏距离$d_{\text{goal}} \in \mathbb{R}^+$和目标方向角$\theta_{\text{goal}} \in [-\pi, \pi]$. 观测信息融合激光雷达与深度相机数据. 其中: 激光雷达原始数据经栅格化处理并归一化后, 堆叠为尺寸$1 \times 1\,440$的张量; 深度相机数据经处理后提取出深度图特征, 编码为$2\times 80 \times 80$的张量. 两种模态的张量沿通道维度拼接, 输入卷积神经网络 (CNN) 提取融合空间特征. 该结构在保留空间拓扑的同时, 增强异构传感器数据间的交互, 形成表达力强的局部环境表征. 预处理后的特征与机器人运动状态及目标信息拼接, 构成低维且信息丰富的观测向量$s_{\text{obs}} \in \mathbb{R}^D$, 作为策略网络输入.

2.1.2 特权信息

训练阶段, 为提升 Critic 网络的状态价值评估能力, 引入特权信息编码器$O_t^p = [O_t^c \quad O_t^o]$. 其中: O_t^c表示机器人中心半径$r = 0.3\,\text{m}$内的本体感知碰撞信息, 若发生碰撞则对应为单位向量, 否则为零向量, 编码为$1 \times 80 \times 80$的空间图; O_t^o则由仿真环境提供动态障碍物的真实位置与速度, 将其映射并编码为$2 \times 80 \times 80$的张量 (两个通道分为对应位置信息与速度信息). 最后将本体碰撞张量与动态障碍物位置和速度张量组合, 从而形成三通道输入. 特权信息仅在训练阶段提供给 Critic, 辅助其准确评估长期回报. Actor 始终仅依赖基于原始传感器的局部观测, 实现策略的实际部署与 Sim-to-Real 泛化.

2.1.3 动作空间

动作空间定义为机器人连续二维控制指令: 线速度$v \in [v_{\min}, v_{\max}]$和角速度$\omega \in [\omega_{\min}, \omega_{\max}]$. Actor 网络输入当前观测状态, 输出归一化控制量$(\hat{v}, \hat{\omega}) \in [-1, 1]$, 通过线性映射恢复为实际控制指令$v$和$\omega$, 用于驱动机器人执行对应动作. 该归一化过程有助于提升训练稳定性及数值收敛性, 避免动作幅值差异对策略学习产生负面影响.

2.2 隐式环境估计网络

动态非结构化环境中的有效导航高度依赖于对环境变化的准确理解. 然而, 受限于动态障碍物的不可预测性, 机器人往往在不完整的观测下运行, 这将影响策略的评估与决策. 为解决这一关键挑战, 本文提出隐式环境估计网络 (implicit environment estimation network, IEEN).

IEEN 的核心功能是从机器人的状态输入O_t^d中估计潜在状态Z_t. 该潜在状态旨在编码环境的关键特征信息, 例如动态障碍物的精确分布、目标区域的相对方向以及环境的时间动态变化. 所获得的潜在表征作为机器人对环境的内部认知模型, 能够为策略模块提供稳定且信息丰富的环境理解.

IEEN 的设计集成动态编码器[30] 和基于L_2正则化的特征对齐机制[31], 旨在高效处理动态交互与传感器限制导致的感知不确定性问题. 首先, 将机器人状态输入O_t^d映射至潜在状态Z_t, 并通过β-变分自编码器 (β-VAE) 的机制, 同时支持对环境状态的有效估计与输入数据的准确重构; 其次, 引入潜在特征一致性损失, 旨在缓解或消除 Actor 与 Critic 网络从同一观测中提取的潜在特征分布不匹配的问题; 最后, 基于互信息神经网络估计器 (mutual information neural estimator, MINE)[32], 最大化所估计的潜在特征Z_t与动态观测O_t^d之间的互信息, 以确保潜在表征的有效性和信息量. IEEN 的总优化目标为

$$\mathcal{L}_{\text{IEENet}} = \mathcal{L}_{\text{est}} + \beta\mathcal{L}_{\text{KL}} + \lambda\mathcal{L}_{L_2} + \mu\mathcal{L}_{\text{env}}. \tag{2}$$

其中: $\mathcal{L}_{\text{est}}$为重建损失, $\mathcal{L}_{\text{KL}}$为潜在分布正则化, $\mathcal{L}_{L_2}$强制 Actor 和 Critic 特征对齐, $\mathcal{L}_{\text{env}}$用于最大化潜在特征与动态观测的互信息.

信息损失定义为

$$I(Z_t; O_t^d) \geqslant \mathbb{E}_{p(z_t, o_t)}[\log q_\theta(z_t, o_t)] - \mathbb{E}_{p(z_t)p(o_t)}[\log q_\theta(z_t, o_t)]. \tag{3}$$

其中: $I(Z_t; O_t^d)$为潜在状态Z_t与机器人状态输入O_t^d之间的互信息, $p(z_t, o_t)$为Z_t与O_t^d的联合分布, $p(z_t)p(o_t)$为边缘分布的乘积, $q_\theta(z_t, o_t)$为由 MINE 模型训练的判别器.

对应损失为

$$\mathcal{L}_{\text{env}} = -(\mathbb{E}_{p(z_t, o_t)}[\log q_\theta(z_t, o_t)] - \mathbb{E}_{p(z_t)p(o_t)}[\log q_\theta(z_t, o_t)]), \tag{4}$$

其中$\mathcal{L}_{\rm env}$为用于最大化潜在特征与动态观测互信息的损失函数.

$\mathcal{L}_{\rm KL}$正则化项为

$$\mathcal{L}_{\rm KL}=\beta D_{\rm KL}(q(z_t|o_t)\|p(z_t)). \tag{5}$$

其中: $D_{\rm KL}(\cdot)$为 KL 散度, 衡量后验分布$q(z_t|o_t)$与先验分布$p(z_t)$之间的差异; β为总优化目标中用于调节 KL 散度项权重的超参数.

特征一致性损失定义为

$$\mathcal{L}_{L_2}=\|\tau_t^a-\tau_t^c\|_2^2, \tag{6}$$

其中τ_t^a和τ_t^c分别为 Actor 与 Critic 网络的潜在特征.

2.3 域随机化

针对策略泛化能力不足的问题, 本文提出行为域随机化 (dynamic pedestrian behavior domain randomization, DPB-DR) 机制, 旨在缓解因动态障碍物行为建模简化导致的 Sim-to-Real 迁移瓶颈. 该机制通过不同的行为扰动, 提升策略在动态障碍物分布、行为差异及交互不确定性方面的应对能力.

与传统静态或规则动态障碍物模型不同, DPB-DR 机制在训练时引入多个关键扰动维度: 动态障碍物速度$v\in[0.3,1.8]\,{\rm m/s}$在每轮训练中进行随机采样; 行为目标 (包括目标位置、路径规划及换向频率) 动态变化, 促使动态障碍物运动模式呈现更高的多样性; 响应延迟基于高斯分布模型进行模拟, 以体现感知冲突和注意力分散导致的行为滞后; 动态障碍物数量$N\in[5,25]$进行动态采样, 其初始位置服从空间泊松过程生成. 所有这些扰动通过统一的随机种子进行控制, 从而在确保仿真环境动态稳定性的同时保持行为的高度多样性.

此外, 为模拟动态障碍物群体内个体差异及交互复杂性, 动态障碍物行为的生成还采用社会力模型和规则驱动模型.

2.4 非对称 ppo 网络

本文采用非对称结构的 PPO 框架, 用于在动态非结构化环境中进行策略学习. 与标准 PPO 方法中 Actor 和 Critic 共享输入的结构不同, 所提出方法在训练阶段为 Critic 网络提供仿真环境中的特权信息, 如全局坐标系下的障碍物位置与速度等, 而 Actor 仅基于局部传感器数据进行策略学习. 该结构通过分离训练阶段的信息源, 使 Critic 具备更强的状态评估能力, 同时保持 Actor 在部署时的约束, 确保其依赖于实际可获取的观测信息进行动作生成.

2.5 奖励函数设计

在自主导航任务中, 奖励函数的设计直接定义智能体在复杂动态环境中学习的目标行为. 本文设计一个多目标复合奖励函数, 其各项子奖励旨在驱动机器人实现向目标高效前进、安全避障、运动平稳、航向与目标一致以及接近目标时的稳定性. 该复合奖励函数定义为

$$\mathcal{R}(s_t,a_t)=r'_g+r'_c+r'_w+r'_\theta+r'_s. \tag{7}$$

1) 目标接近奖励r'_g: 激励机器人持续向目标前进, 并在超时或未成功抵达目标时进行惩罚, 有

$$r'_g=\begin{cases}r_{\rm goal},\ \|p'_g\|<g_m;\\ -r_{\rm path}(\|p'_g\|-\|p'_g\|),\ t\geqslant t_{\max};\\ 0,\ \text{otherwise}.\end{cases} \tag{8}$$

其中: p'_g为目标距离, g_m为成功距离阈值, $t_{\max}$为最大时间, $r_{\rm goal}=20$, $g_m=0.3\,{\rm m}$, $t_{\max}=25\,{\rm s}$, $r_{\rm path}=3.2$.

2) 碰撞规避奖励r'_c: 强烈惩罚接近或碰撞障碍物, 有

$$r'_c=\begin{cases}r_{\rm collision},\ \|p'_o\|\leqslant d_r;\\ r_{\rm obstacle}(d_m-\|p'_o\|),\ d_r<\|p'_o\|\leqslant d_m;\\ 0,\ \text{otherwise}.\end{cases} \tag{9}$$

其中: p'_o为最近障碍物距离, $d_r=0.3\,{\rm m}$, $d_m=1.2\,{\rm m}$, $r_{\rm collision}=-20$, $r_{\rm obstacle}=-0.2$.

3) 路径平滑奖励r'_w: 惩罚角速度剧烈变化, 提升运动平稳性, 有

$$r'_w=-\omega_t^2, \tag{10}$$

其中ω_t为当前时刻角速度.

4) 航向一致奖励r'_θ: 惩罚航向偏离, 提升目标导向性, 有

$$r'_\theta=-\|\theta_t-\theta_{\rm goal}\|, \tag{11}$$

其中θ_t和$\theta_{\rm goal}$分别为当前航向与目标方向.

5) 稳定性奖励r'_s: 惩罚高加速度及过度摇晃, 保障机器人动态平稳, 有

$$r'_s=-\|a_t\|^2, \tag{12}$$

其中a_t为机器人加速度向量.

2.6 网络架构与训练范式

如图 3 所示, 本文构建面向动态环境的自适应策略学习流程, 通过引入非结构化特征提取模块与特权辅助训练机制, 用于对动态场景中的潜在风险进行建模与响应. 多模态输入由激光雷达、深度相机以及历史观测序列组成. 视觉信息通过多层残差卷积编码器提取高维语义特征, 采用渐近式下采样 ($S=2$) 与通道扩展 ($C=\{64,128,256\}$) 以保留空间层级结构与纹理信息; 激光点云输入则经由紧

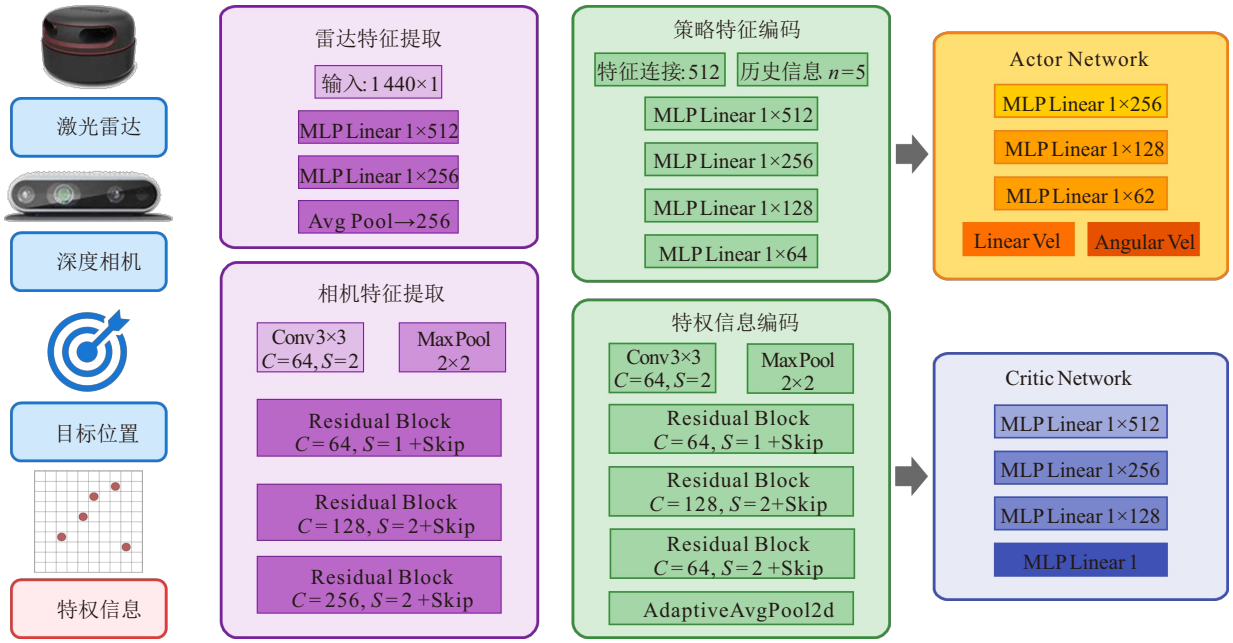


图3　机器人在动态环境中的自主导航网络流程

凑型 MLP 结构 ($1 \times 512 \to 1 \times 256$) 实现压缩表达. 所有模态特征被统一映射对齐至策略特征空间. 同时, 在 IEEN 中, β-VAE 的潜在空间维度固定为 64, 以保证潜在表征具有足够的表达能力来刻画动态障碍物的分布与环境变化. 在策略学习部分, 设计解耦式 Actor-Critic 网络结构, 其中 Actor 仅基于局部可观测信息进行控制决策, 输出连续动作向量, 以满足部署阶段的感知约束; Critic 则引入训练阶段的特权信息, 用于价值估计. 特征融合模块采用多层感知机对多源输入与历史状态进行联合编码, 形成统一、压缩的策略表示, 作为上下文输入驱动 Actor 与 Critic 优化.

3　实验分析

本节详细阐述实验配置. 依次介绍策略训练与评估的仿真环境、真实环境中机器人及传感器参数、非对称 PPO 的训练超参数, 以及用于对比验证的基准方法.

3.1　实验设置

3.1.1　仿真环境

所有训练与测试均在基于 Gazebo 构建的高保真仿真平台上完成, 该平台支持精确的动力学建模. 如图 4 所示, 构建多个场景, 包括封闭式走廊、开放空间与半结构化环境, 具有多样的空间拓扑结构与障碍物分布. 动态障碍物采用 SFM 模型建模, 其数量在5 ～ 25名之间随机分布, 包含稀疏到高密度的动态实体交互情景. 每个动态障碍物的初始位置、目标点、速度、反应时间及交互半径等参数均通过大规模随机采样设定, 以实现 DPB-DR. 导航任务目标在可达区域内随机采样生成, 涵盖单目标导航与多目标切换等不同配置.

3.1.2　机器人与传感器模型

为验证策略的实际部署能力, 在真实环境的机器人平台上进行实验, 如图 5 所示. 机器人搭载有 360° 激光雷达, 最大探测距离为10 m, 波束数1 440;

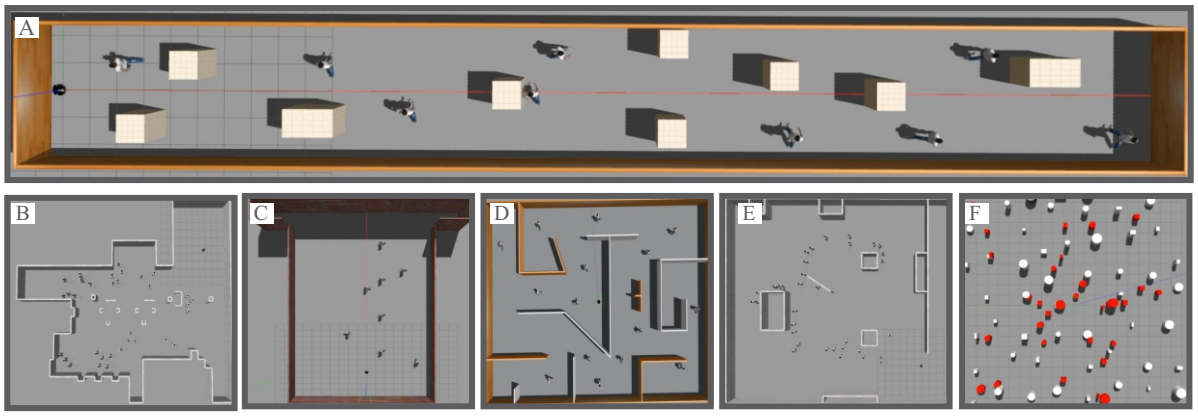


图4　机器人导航仿真环境示意图

前向深度相机, 分辨率为640 × 480, 水平视场角85°; IMU 与轮式里程计. 机器人本体参数为半径 $r = 0.2\,\mathrm{m}$, 质量$m = 6\,\mathrm{kg}$, 最大线速度$v_{\max} = 2\,\mathrm{m/s}$, 最大角速度$\omega_{\max} = 6\,\mathrm{rad/s}$. 在仿真阶段, 模拟现实中感知不确定性, 向传感器中注入高斯噪声, 以增强策略对感知噪声的鲁棒性.

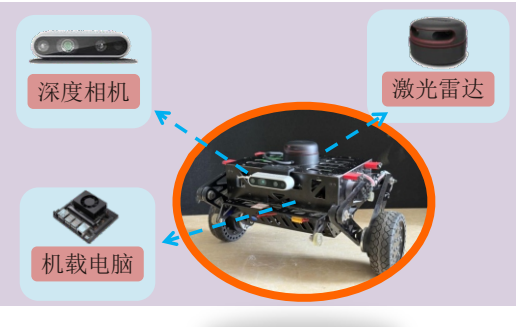


图5 真实环境机器人验证平台

3.1.3 训练参数

基于 Stable-Baselines3 框架实现定制化的 PPO 算法, 用于训练不对称 Actor-Critic 架构. 训练环境涵盖密集动态人群、多目标切换等任务类型, 训练超参数经过调优, 以在鲁棒性、安全性与训练效率之间取得平衡. 最终采用的主要训练参数如表 1 所示.

表1 超参数设计

变量名称	值	变量名称	值
GAE因子	0.95	批量大小	256
裁剪范围	0.2	学习率	adaptive*
折扣因子	0.99	熵系数	0.001
最大批量大小	256	时间步长$\mathrm{d}t$ /s	0.01
每次更新的轮数	1 000	激活函数	ReLU
梯度裁剪范数	1.0	β	0.2
λ	0.5	μ	0.01
反应延迟	0.1 s	舒适距离	0.5 m
排斥力	(2.0, 0.3)	行人数量	5至25人

3.2 对比方法与消融实验

3.2.1 对比方法

为系统评估所提出导航策略在动态非结构化环境中的表现, 本文选取 3 类代表性的方法进行对比. 涵盖传统规划范式、DRL 基线以及融合式智能导航策略, 这些方法旨在从多个维度对本文方法进行评估. 1) 首先, 引入 DWA [33], 作为一种局部路径规划方法; 2) 其次, 使用 3D-DRL[34] 的强化学习导航策略作为基线, 该策略采用与本文方法相同的网络结构与超参数设置, 但不包含 IEEN 模块、DPB-DR 以及非对称结构; 3) 最后, 引入 SAGE 方法[24], 该方法通过迁移优化规划器的离线经验并结合图神经网络建模环境状态, 从而实现样本高效的导航.

3.2.2 消融实验

为分析所提出方法中各关键模块的贡献, 本文设计一系列消融实验, 采用控制变量策略分析不同模块的作用. 具体包括: 1) 移除 IEEN 模块, 使策略直接基于原始传感器观测进行决策; 2) 移除训练阶段的 DPB-DR 机制, 使策略在动态障碍物行为固定下进行训练; 3) 将非对称 Actor-Critic 架构替换为对称结构, 使 Actor 与 Critic 网络基于相同的可观测状态空间进行训练; 4) 所有消融模型在与完整策略一致的训练参数与测试设置下进行, 其对比结果将在后续章节中详细呈现.

3.3 实验结果与分析

本节展示导航策略的定量实验结果, 分析各对比方法与消融配置在不同指标下的性能表现. 训练环境选用图 4 所示的仿真场景, 测试环境为策略未曾见过的新结构, 以验证其泛化能力与鲁棒性.

3.3.1 消融实验结果对比

为量化各关键模块对策略性能的贡献, 本文设计 3 类控制变量的消融模型, 并在与完整方法一致的训练配置下进行评估. 表 2 展示了各模型在导航成功率 (SR)、碰撞率 (CR)、平均路径长度 (AL)、线速度变化量 (MC) 与角速度变化量 (MS) 共 5 个核心指标上的表现. 从结果可见, 完整策略在所有评估指标上均表现最优, 验证了各模块之间的协同增益作用: IEEN 模块对动态障碍物建模尤为关键, 缺失时成功率下降6.8%, 碰撞率上升90.6%, 角速度波动加剧, 表明缺乏高层语义引导将削弱策略对复杂动态结构的适应能力. DR 模块的移除主要影响泛化性能, 在未见场景中路径长度增长, 速度控制不稳定, 表明多样化训练经验对现实感知噪声具有重要意义. 相比之下, 对称 PPO 模型虽具有一定的鲁棒性, 但在平均路径长度和速度平稳性方面仍不及完整架构, 验证了引入特权信息训练的非对称结构在提高策略收敛性与稳定性方面的作用. 综上, IEEN、DR 以及非对称结构均能够在提升导航效率、避障稳定性与策略泛化方面发挥重要作用, 验证了所提出方法的必要性与有效性.

3.3.2 基线方法对比

为全面评估 RANav 策略的性能, 在典型仿真环境中对所选取的基线算法与 RANav 策略进行对比实验, 并对各项性能指标进行量化分析, 结果如表 3 所示. 从实验数据可以看出, RANav 在 SR、CR 与 AL 等关键指标上均明显优于所有对比方法, 尤其在

表2　消融实验结果对比: 各模块在移除后的性能影响

Model Configuration	SR	CR	AL	MC	MS
完整策略	92.5	3.2	25.8	0.27	0.17
移除IEEN	85.7	12.1	26.5	0.28	0.28
移除DPB-DR	88.3	8.9	26.2	0.30	0.25
对称PPO	89.1	7.5	26.0	0.32	0.23

高动态交互场景中仍保持优异的避障稳定性. DWA 作为传统方法, 在静态环境中表现尚可, 但在动态实体环境中路径易失效、碰撞率较高. 3D-DRL 和 SAGE 在考虑动态因素后性能有所提升, 但在复杂多变场景中仍存在泛化能力不足的问题. 而 RANav 通过引入结构建模与策略约束, 能够显著提升策略鲁棒性与环境适应性, 展现出较强的任务完成效率与通用性.

表3　不同算法仿真实验结果对比

Algorithm	SR	CR	AL	MC	MS
Ours	95.2	1.8	15.2	0.42	0.15
DWA	68.3	9.7	17.8	0.22	0.35
3D-DRL	82.1	5.6	16.7	0.30	0.27
SAGE	86.5	4.2	16.3	0.34	0.21

3.3.3　泛化能力与鲁棒性分析

如表 4 所示, RANav 在所有测试环境中都展现出卓越的性能. 其成功率始终保持在 93.8% 以上, 且碰撞率始终控制在 2.1% 以下, 是所有方法中最低的. 这充分表明其具备良好的空间结构理解能力和动态障碍物适应性. 此外, 在路径效率与运动平滑性方面, RANav 同样表现出色. 其平均路径长度显著低于其他算法, 而线速度变化量和角速度变化率运动平滑性指标也远低于其他方法. 这些数据均表明, RANav 生成的路径不仅高效, 且更加平滑、稳定, 从而提升机器人导航的安全性和稳定性.

3.3.4　可视化导航效果

为进一步展示各算法在交互场景中的表现, 在两个仿真环境中对导航轨迹进行可视化对比, 结果如图 6 所示. 在第 1 个环境中, 动态障碍物分布较为稀疏, RANav 能够快速生成一条平滑且避障合理的路径, 表现出对空间结构与动态目标的高效响应. 而 DWA 方法轨迹明显弯曲, 且在接近动态障碍物时出现多次停顿与转向, 反映出其对动态变化的处理迟滞. 3D-DRL 在全局策略上表现较好, 但局部路径存在明显冗余和抖动. SAGE 在避障策略上具备一定的稳定性, 但路径长度与效率仍不及 RANav. 在第 2 个更具挑战性的高密度场景中, RANav 通过预测动态障碍物运动趋势, 合理选择穿行路线, 展现出卓越的适应性与路径规划能力. 而其他方法均在狭窄空间中出现路径回退或轨迹重复, 任务完成效率与稳定性显著下降. 综合来看, RANav不仅能在宽松场景中实现快速避障, 也能在复杂结构中保持高效与稳定, 体现出其在动态环境下良好的鲁棒性与泛化能力.

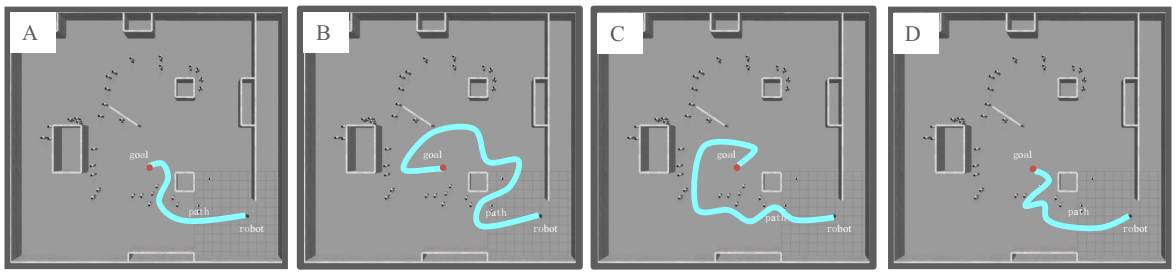
表4　多个测试环境下导航测试结果

环境	Algorithm	SR	CR	AL	MC	MS
环境A	Ours	95.2	1.8	5.2	0.42	0.15
	DWA	68.3	9.7	9.8	0.22	0.35
	3D-DRL	82.1	5.6	8.7	0.30	0.27
	SAGE	88.5	3.6	6.2	0.36	0.20
环境C	Ours	93.8	2.1	12.5	0.40	0.16
	DWA	65.2	10.3	18.1	0.20	0.37
	3D-DRL	80.5	5.9	15.9	0.28	0.28
	SAGE	85.2	4.1	15.2	0.34	0.23
环境D	Ours	94.5	1.9	10.3	0.41	0.15
	DWA	66.7	9.9	14.9	0.21	0.36
	3D-DRL	81.8	5.7	12.8	0.29	0.27
	SAGE	86.9	3.8	11.2	0.34	0.19
环境E	Ours	95.0	1.7	9.1	0.43	0.14
	DWA	67.5	9.6	13.7	0.23	0.34
	3D-DRL	82.3	5.5	11.6	0.31	0.26
	SAGE	87.3	3.9	10.7	0.34	0.22
环境F	Ours	94.2	2.0	11.4	0.41	0.15
	DWA	67.0	9.8	23.8	0.22	0.35
	3D-DRL	81.5	5.7	12.8	0.30	0.27
	SAGE	87.3	4.5	12.3	0.33	0.25

3.4　真实环境下的导航实验

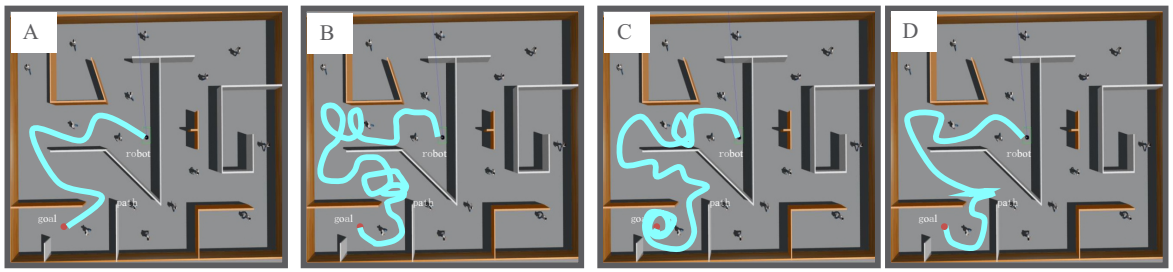
为评估 RANav 在真实室内环境中的部署能力与泛化性能, 设计 3 类未参与训练的现实场景, 基于固定起止点对 RANav 策略开展 10 次重复导航实验, 并采用与仿真实验一致的指标体系. 第 1 个场景为“开阔区人机交互环境”, 模拟办公室或大厅等宽敞空间, 其中存在零散静态障碍物与随机动态行人干扰, 旨在验证机器人在低约束环境中应对人-物混合动态的能力. 第 2 个场景为“狭窄走廊避障环境”, 布设密集静态障碍物构成非凸通道结构, 同时伴随部分动态障碍物随机穿行, 旨在验证策略在高约束空间中的路径效率与避障安全性. 第 3 个场景为“多行人复杂交互场景”, 该场景在走廊区域引入 4 名动态行人进行高频率对向与穿插移动, 旨在验证策略在动态障碍物数量增加和复杂交互冲突下的实时决策鲁棒性与泛化性能.

如图 7 所示, 这些真实导航测试场景对策略的实时感知与决策能力提出挑战. 图片清晰展示了 RANav 策略在 3 类未见环境中的导航序列: 1) 开阔交互场景; 2) 狭窄走廊场景; 3) 多行人交互场景. 表



PDNav方法 DWA方法 3D-DRL方法 SAGE方法

(a) 场景 1



PDNav方法 DWA方法 3D-DRL方法 SAGE方法

(b) 场景 2

图6 不同导航算法可视化结果

5 汇总了 RANav 与多种主流导航方法在上述 3 类场景下的性能评估结果. 实验表明, RANav 在 3 类复杂动态环境中均表现出卓越的导航性能: 首先, 在开阔交互场景中, RANav 保持 93.0% 的高成功率与 3.3% 的低碰撞率, 显著优于各基线方法; 其次, 在更具挑战性的狭窄走廊场景中, 尽管场景难度提升, 其成功率和碰撞率仍分别维持在 90.0% 和 6.6%, 性能依然远超其他对比方法; 最后, 在多行人复杂交互场景中, 面对动态障碍物数量的增加和高频率的交互冲突, RANav 的导航成功率仍能达到 86.7%, 碰撞率控制在 5.1%, 充分验证了策略在非训练场景下的鲁棒泛化能力. 此外, RANav 在路径效率与运动平滑性方面亦展现出显著优势. RANav 在开阔交互场景中的平均路径长度为 8.7 m, 平均平滑度为 0.18; 在狭窄走廊场景中, 路径长度为 15.3 m, 平滑度为 0.17; 在多行人复杂交互场景中, 路径长度为 9.8 m, 平滑度为 0.19. 这些数据表明 RANav 能够持续生成最短且最平稳的避障轨迹, 其运动平稳性指标远低

(a) 开阔交互场景

(b) 狭窄走廊场景

(c) 动态交互场景

图7 真实环境场景下导航测试

于基线方法, 充分验证了 RANav 策略在真实复杂动态环境中的实际部署可行性与鲁棒性.

表5 真实世界不同场景下的导航性能评估结果

真实环境场景	Algorithm	SR	CR	AL	MC	MS
开阔交互场景	Ours	92.7	3.3	15.3	0.35	0.18
	DWA	65.5	11.2	25.1	0.50	0.30
	3D-DRL	80.5	7.4	22.8	0.45	0.25
	SAGE	88.3	4.6	18.5	0.38	0.22
狭窄走廊场景	Ours	89.6	6.6	8.7	0.38	0.17
	DWA	60.1	13.5	15.4	0.55	0.35
	3D-DRL	75.8	10.1	12.1	0.48	0.23
	SAGE	85.2	8.7	10.1	0.42	0.20
动态交互场景	Ours	**90.5**	**5.1**	**16.5**	**0.37**	**0.19**
	DWA	62.0	12.5	23.0	0.52	0.32
	3D-DRL	78.5	8.5	20.5	0.47	0.24
	SAGE	86.5	6.8	19.1	0.40	0.21

4 结 论

本文提出了一种面向动态不确定环境的非对称强化学习导航框架 RANav. 该框架系统整合隐式环境估计、动态障碍行为域随机化与非对称强化学习三大核心机制, 可显著提升移动机器人在部分可观测性场景下的环境理解能力、策略泛化性与行为鲁棒性. 仿真实验结果表明, RANav 在未见测试环境中实现超过 90% 的导航成功率与低于 5% 的碰撞率, 且在路径效率、避障距离及动作平滑性等关键指标上均优于主流基线方法. 实地测试进一步验证了其出色的 Sim-to-Real 迁移能力: 在开阔交互、狭窄走廊以及多行人交互 3 类未参与训练的真实场景中, RANav 的导航成功率最高可达 93.0%, 最低保持在 86.7%; 同时, 其碰撞率在所有场景中均控制在 7% 以下, 充分展现出该策略在动态环境下的运行稳定性与安全性. 上述综合结果表明, RANav 在动态不确定环境中具有良好的实用性、安全性与推广潜力, 可为高可靠性移动机器人自主导航任务提供切实可行的解决方案.

参考文献 (References)

[1] Ren Y F, Zhu F C, Lu G Z, et al. Safety-assured high-speed navigation for MAVs[J]. Science Robotics, 2025, 10(98): eado6187.

[2] Lee J, Bjelonic M, Reske A, et al. Learning robust autonomous navigation and locomotion for wheeled-legged robots[J]. Science Robotics, 2024, 9(89): eadi9641.

[3] Samavi S, Han J R, Shkurti F, et al. SICNav: Safe and interactive crowd navigation using model predictive control and bilevel optimization[J]. IEEE Transactions on Robotics, 2025, 41: 801-818.

[4] 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.)

[5] 何玉庆, 赵忆文, 韩建达, 等. 与人共融 —— 机器人技术发展的新趋势[J]. 机器人产业, 2015(5): 74-80.
(He Y Q, Zhao Y W, Han J D, et al. Harmony with people-a new trend of robot technology development[J]. Robot Industry, 2015(5): 74-80.)

[6] Jiang H G, Bhujel N, Lin Z Y, et al. Learning relation in crowd using gated graph convolutional networks for DRL-based robot navigation[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(6): 5085-5095.

[7] Hoang V B, Nguyen V H, Ngo T D, et al. Socially aware robot navigation framework: Where and how to approach people in dynamic social environments[J]. IEEE Transactions on Automation Science and Engineering, 2023, 20(2): 1322-1336.

[8] Dai J, Li D F, Zhao J W, et al. Autonomous navigation of robots based on the improved informed-RRT algorithm and DWA[J]. Journal of Robotics, 2022, 2022(1): 3477265.

[9] Qi Y, He B B, Wang R D, et al. Hierarchical motion planning for autonomous vehicles in unstructured dynamic environments[J]. IEEE Robotics and Automation Letters, 2023, 8(2): 496-503.

[10] Wurts J, Stein J L, Ersal T. Collision imminent steering at high speed using nonlinear model predictive control[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8278-8289.

[11] 胡郑希, 张千一, 翟晓琳, 等. 考虑人类视线区域约束的机器人社交导航[J]. 机器人, 2023, 45(6): 670-682.
(Hu Z X, Zhang Q Y, Zhai X L, et al. Socially-aware robot navigation considering human gaze-related area constraints[J]. Robot, 2023, 45(6): 670-682.)

[12] Sathyamoorthy A J, Weerakoon K, Guan T R, et al. TerraPN: Unstructured terrain navigation using online self-supervised learning[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, 2022: 7197-7204.

[13] Kahn G, Abbeel P, Levine S. BADGR: An autonomous self-supervised learning-based navigation system[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1312-1319.

[14] Wang D S, Hu Y H, Ma T L. Mobile robot navigation with the combination of supervised learning in cerebellum and reward-based learning in basal ganglia[J]. Cognitive Systems Research, 2020, 59: 1-14.

[15] Thomas H, Agro B, Gridseth M, et al. Self-supervised learning of lidar segmentation for autonomous indoor navigation[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 14047-14053.

[16] Ramakrishnan S K, Chaplot D S, Al-Halah Z, et al. PONI: Potential functions for objectgoal navigation with interaction-free learning[C]. IEEE/CVF Conference on

Computer Vision and Pattern Recognition. New Orleans, 2022: 18868-18878.

[17] Xie Z T, Xin P J, Dames P. Towards safe navigation through crowded dynamic environments[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague, 2021: 4934-4940.

[18] 孙虎, 金宇强, 张文安, 等. 基于多引导结构感知网络的深度补全[J]. 控制与决策, 2024, 39(2): 401-410.
(Sun H, Jin Y Q, Zhang W A, et al. Depth completion method based on multi-guided structure-aware networks[J]. Control and Decision, 2024, 39(2): 401-410.)

[19] Sen N A, Kulić D, Carreno-Medrano P. Domain randomization for learning to navigate in human environments[J]. IEEE Robotics and Automation Letters, 2025, 10(2): 1625-1632.

[20] 倪浩, 章胜, 刘福炜, 等. 基于域随机化增强EfficientZero的无人机空战智能决策[J]. 控制与决策, 2025, 40(11): 3273-3286.
(Ni H, Zhang S, Liu F W, et al. UAV air combat intelligent decision-making based on domain randomization enhanced EfficientZero[J]. Control and Decision, 2025, 40(11): 3273-3286.)

[21] Tang C, Abbatematteo B, Hu J H, et al. Deep reinforcement learning for robotics: A survey of real-world successes[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(27): 28694-28698.

[22] Cui Y X, Zhang H D, Wang Y, et al. Learning world transition model for socially aware robot navigation[C]. IEEE International Conference on Robotics and Automation. Xi'an, 2021: 9262-9268.

[23] de Heuvel J, Zeng X Y, Shi W X, et al. Spatiotemporal attention enhances lidar-based robot navigation in dynamic environments[J]. IEEE Robotics and Automation Letters, 2024, 9(5): 4202-4209.

[24] Liu H J, Dong W, Mao S R, et al. Sample-efficient learning-based dynamic environment navigation with transferring experience from optimization-based planner[J]. IEEE Robotics and Automation Letters, 2024, 9(8): 7055-7062.

[25] Wang H T, Tan A H, Nejat G. NavFormer: A transformer architecture for robot target-driven navigation in unknown and dynamic environments[J]. IEEE Robotics and Automation Letters, 2024, 9(8): 6808-6815.

[26] Zhu K, Li B, Zhe W M, et al. Collision avoidance among dense heterogeneous agents using deep reinforcement learning[J]. IEEE Robotics and Automation Letters, 2023, 8(1): 57-64.

[27] Zhu W, Hayashibe M. Autonomous navigation system in pedestrian scenarios using a dreamer-based motion planner[J]. IEEE Robotics and Automation Letters, 2023, 8(6): 3836-3843.

[28] Liu Z H, Na W J, Yao C P, et al. Relaxing the limitations of the optimal reciprocal collision avoidance algorithm for mobile robots in crowds[J]. IEEE Robotics and Automation Letters, 2024, 9(6): 5520-5527.

[29] Song Y C, Wang R H, Bi Q C, et al. STVO: Spatial-temporal constrained velocity obstacle for safe navigation among pedestrians[J]. IEEE Transactions on Vehicular Technology, 2025, 74(9): 13580-13591.

[30] Fu Z, Kumar A, Malik J, et al. Minimizing energy consumption leads to the emergence of gaits in legged robots[J/OL]. 2021, arXiv: 2111.01674.

[31] Margolis G B, Yang G, Paigwar K, et al. Rapid locomotion via reinforcement learning[J]. The International Journal of Robotics Research, 2024, 43(4): 572-587.

[32] Belghazi M I, Baratin A, Rajeswar S, et al. MINE: mutual information neural estimation[J/OL]. 2018, arXiv: 1801.04062.

[33] 王洪斌, 刘德垚, 郑维, 等. 异构多目标差分-动态窗口算法及其在移动机器人中的应用[J]. 控制与决策, 2023, 38(12): 3390-3398.
(Wang H B, Liu D Y, Zheng W, et al. Heterogeneous multi-objective differential evolution-dynamic window algorithm and application for energy-saving motion planning of mobile robot[J]. Control and Decision, 2023, 38(12): 3390-3398.)

[34] Akmandor N Ü, Li H Y, Lvov G, et al. Deep reinforcement learning based robot navigation in dynamic environments using occupancy values of motion primitives[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems. Kyoto, 2022: 11687-11694.

作者简介

何乃峰 (1993−), 男, 博士生, 主要研究方向为机器人控制与决策、强化学习、特种机器人, E-mail: nfhe@nuaa.edu.cn;

杨忠 (1969−), 男, 教授, 博士, 主要研究方向为智能机器人及飞行器的设计与控制, E-mail: yangzhong@nuaa.edu.cn;

许诺 (1986−), 男, 高级工程师, 博士生, 主要研究方向为无人机导航、多传感器融合、配电网数字孪生技术, E-mail: xunuo-nuaa@nuaa.edu.cn;

卓浩泽 (1986−), 男, 正高级工程师, 博士生, 主要研究方向为无人机系统控制与管理、智能检查与控制技术, E-mail: zhuohaoze@nuaa.edu.cn;

徐宏雨 (1999−), 男, 博士生, 主要研究方向为空地跨模态机器人、无人机系统设计与控制, E-mail: hongyuxu@nuaa.edu.cn;

陈凯 (2002−), 男, 硕士生, 主要研究方向为足式机器人智能控制, E-mail: sz2403024@nuaa.edu.cn.