面向复杂地形场景的四轮独立转向重载机器人端到端控制

徐德刚 [1] 王逸之 [1] 谢永芳 [1]

摘 要 针对四轮独立转向与驱动 (4WISD) 移动机器人在复杂地形中依赖理想约束与参数整定、易受轮地非理想效应影响而泛化受限的问题, 构建一种无需显式模型计算的端到端深度强化学习底层控制策略. 该方法基于近端策略优化算法, 结合课程学习与域随机化构建多地形、多扰动训练过程, 并通过结构化状态–动作定义与物理启发奖励塑形, 使策略由本体观测直接生成四轮转向角与轮速命令, 实现多轮协同与侧滑抑制. 基于 Isaac Sim 高保真仿真平台, 在起伏粗糙地形和低摩擦平面两类工况下与 PID-LS、MPC-LS 级联基线进行对比. 结果表明, 该策略在工况变化下性能波动较小, 在粗糙地形中线速度跟踪更接近任务目标, 且在线计算开销较低, 单步耗时 0.304 ms. 室外平整水泥地面实机验证表明, 该策略可基于 ROS-PLC 控制链路稳定在线运行, 三轴速度 RMSE 分别为 0.169 m/s、0.213 m/s 和 0.116 rad/s; 单步策略推理耗时 0.62 ms, 传感器同步与状态构建平均耗时 56.15 ms. 结果验证了端到端数据驱动底层控制在 4WISD 平台无精确建模条件下的可行性, 为高自由度轮式移动平台复杂环境运动控制提供了创新路径.

关键词 四轮独立转向与驱动; 深度强化学习; 端到端控制; 复杂地形

引用格式 徐德刚, 王逸之, 谢永芳. 面向复杂地形场景的四轮独立转向重载机器人端到端控制. 自动化学报, 2026, **52**(9): 1979−1991

DOI 10.16383/j.aas.c250595 **CSTR** 32138.14.j.aas.c250595

End-to-end Control for 4WISD Heavy-duty Robots in Complex Terrains

XU De-Gang[1] WANG Yi-Zhi[1] XIE Yong-Fang[1]

Abstract To address the limited generalization of traditional control methods for four-wheel independent steering and driving (4WISD) mobile robots in complex terrains, caused by their reliance on idealized constraints and parameter tuning as well as their susceptibility to non-ideal wheel-ground interactions, an end-to-end deep reinforcement learning low-level control policy without explicit model-based computation is developed. Based on proximal policy optimization, the method integrates curriculum learning and physics-based domain randomization to construct a multi-terrain and multi-disturbance training process. With a structured state-action formulation and physically motivated reward shaping, the policy directly maps proprioceptive observations to four-wheel steering-angle and wheel-speed commands, enabling coordinated multi-wheel control and slip suppression. Experiments in the high-fidelity Isaac Sim simulator compare the proposed policy with PID-LS and MPC-LS cascaded baselines under rough undulating terrains and low-friction flat terrains. Results show that the proposed policy exhibits smaller performance fluctuations across conditions, achieves more accurate linear-velocity tracking on rough terrains, and maintains low online computational cost, with a single-step latency of 0.304 ms. Real-world validation on an outdoor flat concrete surface further demonstrates stable online execution through a ROS-PLC control pipeline, with velocity root mean square error (RMSE) values of 0.169 m/s, 0.213 m/s, and 0.116 rad/s for the three velocity components. The policy inference latency is 0.62 ms, while sensor synchronization and state construction take 56.15 ms on average. These results verify the feasibility of end-to-end data-driven low-level control for 4WISD platforms without precise modeling, offering an innovative approach to motion control of high-DOF wheeled mobile robots in complex environments.

Keywords four-wheel independent steering and driving; deep reinforcement learning; end-to-end control; complex terrain

Citation Xu De-Gang, Wang Yi-Zhi, Xie Yong-Fang. End-to-end control for 4WISD heavy-duty robots in complex terrains. *Acta Automatica Sinica*, 2026, **52**(9): 1979−1991

收稿日期 2025-11-03 录用日期 2026-04-01
Manuscript received November 3, 2025; accepted April 1, 2026
国家自然科学基金 (62473386), 湖南省重点研发计划项目 (2025RC1009, 2023GK2096), 山东省重点研发计划项目 (2025GNKJHZ0201), 新疆维吾尔自治区重大科技专项项目 (2022A02010) 资助
Supported by National Natural Science Foundation of China (62473386), the Key Research and Development Project of Hunan Province (2025RC1009, 2023GK2096), the Key Research and Development Project of Shandong Province (2025GNKJHZ0201), and the Major Science and Technology Special Project of Xinjiang Uyghur Autonomous Region (2022A02010)
本文责任编委 穆朝絮
Recommended by Associate Editor MU Chao-Xu
1. 中南大学自动化学院 长沙 410083
1. School of Automation, Central South University, Changsha 410083

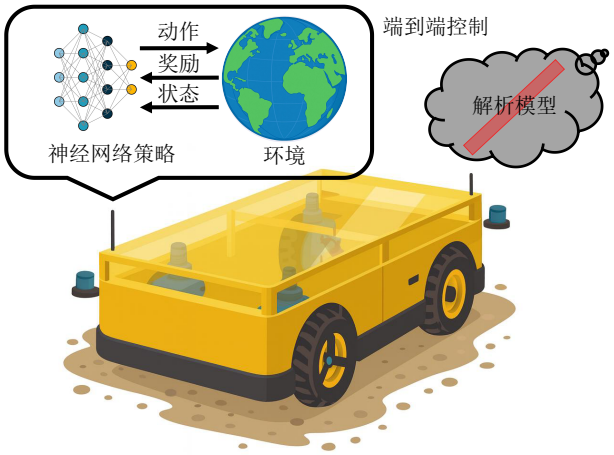
随着物流运输、仓储分拣、野外勘探与灾害救援等典型任务场景对移动机器人在载重能力、机动性与地形适应性方面提出更高要求, 传统差动驱动、阿克曼转向及麦克纳姆轮系统在崎岖、不平整或松软地形中往往难以兼顾“通过效率”与“稳定控制”. 多足或人形机器人确实具备更强的地形适应能力, 但从工程落地角度看, 其在负载能力、能耗效率与控制系统复杂度上, 尚难形成对轮式平台的整体替代. 相比之下, 四轮独立转向与驱动 (four-wheel independent steering and driving, 4WISD) 移动平台凭借重载结构与多模态运动能力, 正在成为重载移动作业的一条现实路线: 其车轮转向轴安装于底盘主梁而非悬臂结构, 更适用于承载; 每个轮组具备独立转向自由度, 使平台可以实现零转弯半径旋转、侧向平移与曲率连续变换等运动模式, 从而提升狭窄空间内的机动性与规划自由度. 此外, 当负载分布不均或地形起伏剧烈时, 4WISD 通过协调各轮驱动与转向角, 有机会维持更均衡的牵引力分配与更可控的姿态响应.

在 4WISD 移动机器人运动控制研究中, 传统方法多从平台的几何结构出发, 在无侧滑等假设下推导运动学闭式解, 并结合模型预测控制 (model predictive control, MPC) 或其他优化/鲁棒控制框架, 实现对位置、航向或轨迹的跟踪[1−3]. 围绕重载场景, Liu 等[4] 基于非线性模型预测控制处理载荷转移带来的动态影响, 以提升重载 AGV 的安全性与控制精度; Ding 等[5] 提出带转向模式约束的速度 MPC (model predictive control under steering strategy constraint, USSC-MPC), 在复杂路径与狭窄环境中提高了实时轨迹跟踪性能; 针对高速运动与转向延迟等因素, Liu 等[6] 构建融合 A* 规划与动力学约束的 MPC 路径跟踪框架; Jeong 等[7] 进一步将 MPC 推广至低附着地面; Nguyen 等[8] 则利用混合整数二次规划 (mixed-integer quadratic programming, MIQP) 实现农业环境下的路径跟踪与转向模式选择. 上述方法在平坦或低复杂度环境中表现稳定, 但多数基于平面运动学假设, 未充分考虑碎石、沙土、不规则凹凸面及斜坡等复杂地形中的轮地非线性交互与随机扰动, 从而限制了其在真实场景中的泛化能力.

为提升控制系统在多样环境下的自适应性, 深度强化学习 (deep reinforcement learning, DRL)[9] 近年来越来越多地被引入机器人控制任务. DRL 的优势在于可以通过交互数据学习高维非线性状态–动作映射, 从而在不依赖精细建模的情况下获得可用策略. 相关进展在足式机器人领域尤为突出: Lee 等[10] 联合训练状态估计与控制网络, 在无地形先验条件下实现了复杂地形稳定运动; Miki 等[11] 结合大规模动作先验预训练与多样扰动机制, 提升了策略在多地形、多任务下的泛化能力; Belmonte-Baeza 等[12] 引入元学习实现对新环境的快速适应; Margolis 等[13] 通过结构化地形表示与轻量感知编码器[14] 增强了仿真到实物的跨地形迁移. 同时, 结构–控制协同优化[15] 与接触状态感知模块[16] 也被用于提升复杂地形下的鲁棒性. 相关思路进一步扩展到人形机器人中, 例如参数化步态策略调节[17]、引入感知噪声与扰动的 sim-to-real 训练[18]、奖励结构与势函数塑形[19] 以及整身对抗模仿学习步态生成[20]. 在操作控制与导航领域, 亦有研究通过强化学习实现空间机器人的柔顺装配控制[21] 以及面向工业自主移动平台的启发式奖励塑形与潜在风险状态增强的无地图导航策略[22−23]. 相比之下, 针对轮式平台, 尤其是 4WISD 此类高自由度轮式平台在复杂地形下的底层运动控制, 仍缺少系统性的端到端研究与可复现的基准对比. 该平台一方面需要面对复杂的轮地接触动力学, 另一方面还必须同时处理四轮转向与驱动的协调耦合, 这对策略收敛、动作物理可行性与部署稳定性提出更高的要求.

基于上述背景, 本文关注一个更贴近工程落地的问题: 在不依赖显式运动学或动力学模型的前提下, 能否学习到可执行、可部署的 4WISD 底层控制策略, 并在复杂地形扰动下保持稳定通过? 为此, 本文在 NVIDIA Isaac Sim 中构建高保真并行仿真训练环境, 通过可控地形参数生成不规则/起伏地形, 并结合课程机制与域随机化完成训练与评估. 因此, 本文采用近端策略优化 (proximal policy optimization, PPO) 构建端到端底层控制框架, 使策略直接由本体观测 (关节编码器与 IMU 等) 生成四轮转向角与轮速命令, 从而避免传统“机体速度控制器 + 解析分配器”级联结构在模型失配下的误差累积. 相较 SAC、TD3 等 off-policy 方法, PPO 作为 on-policy 算法通过截断代理目标限制更新幅度, 通常在大规模并行采样下更易稳定收敛, 且不依赖回放缓冲区的数据分布管理. 图 1 给出本文研究框架的直观示意. 需要强调的是, 本文所称“无模型”并非完全不使用模型表达, 而是指策略学习与在线控制阶段不依赖显式模型计算控制量. 文中相关运动学描述更多作为领域知识, 用于启发奖励函数的塑形与约束项设计. 本文的主要贡献如下:

1) 提出一种面向重载 4WISD 平台的端到端底层控制方法, 直接输出轮组转向与轮速命令, 不显式构造机体速度控制器与解析分配器, 以降低模型



注: 策略在复杂地形中由本体观测直接生成轮组命令, 无需解析模型在线求解.

图 1　端到端底层控制框架

Fig. 1　End-to-end low-level control framework

失配与级联结构带来的性能折损;

2) 提出一套物理启发的奖励塑形设计, 将速度跟踪、姿态动态稳定与侧滑趋势统一纳入学习目标, 作为学习信号引导策略收敛, 并强化控制行为的动态合理性与物理可行性;

3) 提出一种课程机制与域随机化结合的训练框架, 在高保真仿真中构建多地形、多扰动的可复现评测方法, 以系统地检验策略在非结构化地形与扰动条件下的稳定性与泛化能力.

1　四轮独立转向与驱动移动机器人系统建模

1.1　机器人系统结构与控制任务描述

在固定世界坐标系 $\{W\}$ 下, 4WISD 移动机器人的平面位姿记为 $(x,\ y,\ \theta)$, 其中 $(x,\ y)$ 为质心位置, θ 表示绕 z 轴的航向角. 定义机体坐标系 $\{B\}$, 其原点位于质心位置. 记机体在 $\{B\}$ 中的广义速度为

$$\boldsymbol{\xi} = \begin{bmatrix} v_x & v_y & \omega_z \end{bmatrix}^{\mathrm{T}} \tag{1}$$

其中 v_x, v_y 分别表示机体坐标系下的线速度分量; ω_z 为偏航角速度.

机器人共有四个车轮, 编号 $i = 1, \cdots, 4$. 第 i 个车轮在机体坐标系中的几何位置为

$$\boldsymbol{r}_i = \begin{bmatrix} x_i & y_i \end{bmatrix}^{\mathrm{T}} \tag{2}$$

每个车轮具备独立驱动与转向自由度, 其控制输入分别为转向角位置 δ_i (相对于机体 x 轴) 与驱动角速度 ψ_i. 图 2 为系统结构的示意.

控制任务为在给定期望机体速度 $\boldsymbol{\xi}^{\star} = [v_x^{\star},\ v_y^{\star},\ \omega_z^{\star}]^{\mathrm{T}}$ 的条件下, 底层控制策略输出 $\{\delta_i,\ \psi_i\}_{i=1}^{4}$, 使机器人实现期望的三自由度平面运动, 并在不规则复杂地面中尽可能抑制打滑与姿态扰动.

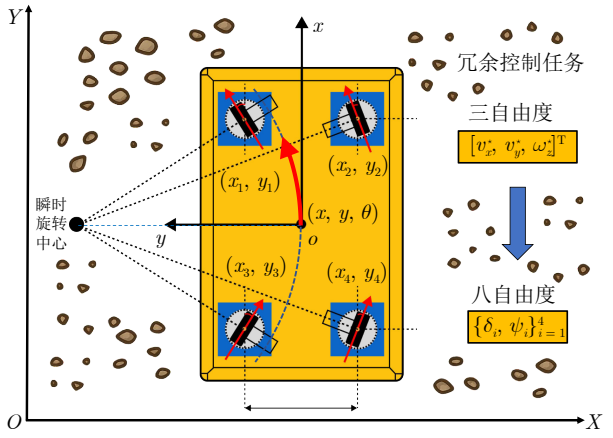


图 2　四轮独立转向与驱动移动机器人系统结构

Fig. 2　Schematic of the four-wheel independent steering and driving mobile robot system architecture

1.2　复杂地形下的模型失配与学习问题形式化

在理想平坦、刚性地面且满足无侧滑条件下, 车轮–地面接触可近似为纯滚动约束. 设机体速度为 $\boldsymbol{\xi}$, 则第 i 个车轮的轮心速度可写为

$$\boldsymbol{v}_{w,\,i} = \begin{bmatrix} v_x - \omega_z y_i \\ v_y + \omega_z x_i \end{bmatrix} \tag{3}$$

定义车轮滚动方向的单位向量

$$\boldsymbol{t}_i = \begin{bmatrix} \cos\delta_i \\ \sin\delta_i \end{bmatrix},\ \boldsymbol{s}_i = \begin{bmatrix} -\sin\delta_i \\ \cos\delta_i \end{bmatrix} \tag{4}$$

其中 $\boldsymbol{t}_i$ 为滚动切向方向, $\boldsymbol{s}_i$ 为其在平面内的法向 (侧向) 方向. 在无侧滑条件下, 轮心速度在滚动方向的分量与轮缘线速度一致, 即

$$u_i = \boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{v}_{w,\,i},\ u_i = r_w\,\psi_i \tag{5}$$

其中 r_w 为轮半径. 将四个车轮的滚动约束写为矩阵形式, 有

$$\boldsymbol{u} = \begin{bmatrix} u_1 & u_2 & u_3 & u_4 \end{bmatrix}^{\mathrm{T}} = K(\boldsymbol{\delta})\,\boldsymbol{\xi} \tag{6}$$

其中运动学映射矩阵 $K(\boldsymbol{\delta}) \in \mathbf{R}^{4\times 3}$ 的第 i 行可写为

$$K_i(\delta_i) = \begin{bmatrix} \cos\delta_i & \sin\delta_i & -y_i\cos\delta_i + x_i\sin\delta_i \end{bmatrix} \tag{7}$$

上述映射在本文中主要有两类目的: 一是刻画 4WISD 的几何约束与可达运动形式; 二是为后续传统基线 (机体速度控制 + 轮组分配) 提供一致的映射关系[24].

在理想条件下, 若进一步考虑加速度响应与外界扰动, 系统通常还需要引入动力学描述, 涉及轮

地作用力、法向载荷分配与摩擦极限等因素. 此类动力学建模往往隐含若干前提: 接触几何相对稳定, 地面材料参数 (如摩擦系数、刚度) 可近似为常值或可辨识, 载荷与质心变化可被准确建模或在线估计. 然而在不规则/起伏地形中, 上述前提往往难以同时满足, 典型失配来源包括:

- 轮地相互作用的强非线性. 沉陷、弹跳与侧滑会改变有效接触条件, 使得“纯滚动”约束与力学近似同时偏离.
- 地形与材料参数的不确定性. 摩擦系数、地面刚度等呈空间非均匀分布, 且随工况变化, 使模型对参数误差高度敏感.
- 载荷与质心状态的时变. 任务装载变化引起质量与质心漂移, 进一步放大运动响应的不确定性.

因此, 系统表现为强非线性、强耦合且时变的动态过程, 依赖精确模型与参数整定的控制器在复杂场景中容易出现性能退化甚至失稳. 为此, 本文将 4WISD 的底层运动控制问题形式化为马尔科夫决策过程 (Markov decision process, MDP), 通过策略学习构建状态–动作映射, 从而降低对显式模型与精确参数的依赖.

具体地, 定义 MDP 为五元组 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma \rangle$:

- 状态空间 $\mathcal{S}$: 包含与姿态稳定、速度响应及执行器状态相关的关键本体观测量;
- 动作空间 $\mathcal{A}$: 由四个车轮的转向角位置与驱动角速度组成的连续 8 维向量 $\boldsymbol{a} = [\delta_1, \cdots, \delta_4, \psi_1, \cdots, \psi_4]^{\mathrm{T}}$, 并满足物理约束 $\delta_i \in [\delta_{\min}, \delta_{\max}]$, $\psi_i \in [\psi_{\min}, \psi_{\max}]$;
- 状态转移 $\mathcal{P}$: 由高保真仿真引擎或实际系统诱导, 隐式包含轮地相互作用、执行器动力学、地形扰动等非线性因素;
- 奖励函数 $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$: 刻画速度跟踪、姿态稳定与打滑抑制等目标, 并以标量形式反馈给策略优化;
- 折扣因子 $\gamma \in (0, 1)$: 用于平衡短期与长期性能. 强化学习的目标是寻找策略 $\pi : \mathcal{S} \to \mathcal{A}$, 最大化期望的累积折扣回报:

$$J(\pi) = \mathrm{E}\left[\sum_{t=0}^{T-1} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t)\right] \tag{8}$$

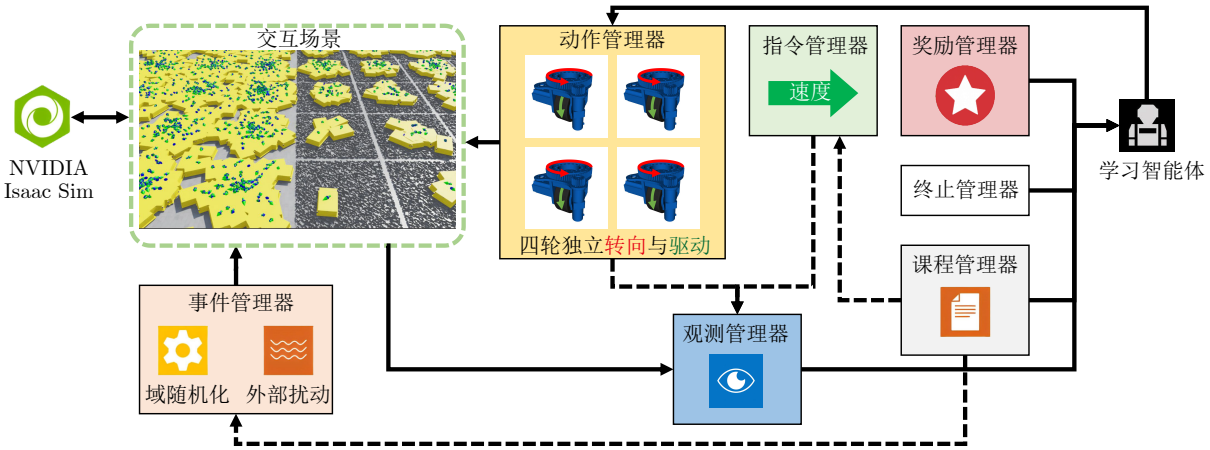
其中期望是关于策略 π 与环境动态 $\mathcal{P}$ 的联合分布.

2 面向 4WISD 移动机器人的端到端控制策略学习方法

为求解上述 MDP 问题, 本节提出一种基于近端策略优化算法的端到端控制策略学习方法, 策略直接从机器人可观测的本体状态映射至四个车轮的转向角与轮速命令, 以在不规则/起伏地形引起的扰动下提升运动稳定性与控制鲁棒性. 整个训练过程基于 Isaac Lab 框架[25] 中的管理器式强化学习环境 (manager-based RL environment) 构建而成 (见图 3), 该框架以模块化方式组织观测、奖励、动作执行、重置与课程调度等组件, 并支持多环境并行交互采样, 从而提高采样效率与训练稳定性.

2.1 状态与动作空间设计

端到端控制策略直接作用于物理系统, 因此需要合理设计状态观测与动作变量. 在状态空间建模中, 本文定义状态向量 $\boldsymbol{s}_t$, 由多个物理量构成, 如下



注: 训练环境通过域随机化和外部扰动构造多地形交互场景, 并结合观测、动作、指令、奖励、终止与课程管理完成策略训练.

图 3 基于 Isaac Lab 的端到端训练流程

Fig. 3 End-to-end training pipeline based on Isaac Lab

所示:

$$\boldsymbol{s}_t = \left[\boldsymbol{\xi}_t^{\star},\ \boldsymbol{v}_t^{b},\ \boldsymbol{\omega}_t^{b},\ \boldsymbol{g}_t^{b},\ \boldsymbol{\delta}_t,\ \boldsymbol{\psi}_t,\ \boldsymbol{a}_{t-1}\right]^{\mathrm{T}} \in \mathbf{R}^{28} \tag{9}$$

其中各子向量定义如下:

- $\boldsymbol{\xi}_t^{\star} \in \mathbf{R}^3$: 当前时间步的目标速度命令 $(v_x^{\star},\ v_y^{\star},\ \omega_z^{\star})$;
- $\boldsymbol{v}_t^{b} \in \mathbf{R}^3$: 机器人在机体坐标系下的线速度;
- $\boldsymbol{\omega}_t^{b} \in \mathbf{R}^3$: 机体三轴角速度;
- $\boldsymbol{g}_t^{b} \in \mathbf{R}^3$: 重力方向在机体坐标系的投影;
- $\boldsymbol{\delta}_t \in \mathbf{R}^4$: 四个转向关节转向角位置;
- $\boldsymbol{\psi}_t \in \mathbf{R}^4$: 四个驱动关节驱动角速度;
- $\boldsymbol{a}_{t-1} \in \mathbf{R}^8$: 上一时刻策略输出的动作向量.

在动作空间设计上, 考虑到 4WISD 底盘结构的控制特性, 本文将动作向量 $\boldsymbol{a}_t$ 构建为如下形式:

$$\boldsymbol{a}_t = [\delta_1,\ \delta_2,\ \delta_3,\ \delta_4,\ \psi_1,\ \psi_2,\ \psi_3,\ \psi_4]^{\mathrm{T}} \in \mathbf{R}^8 \tag{10}$$

其中 δ_i 为第 i 个转向关节的转向角位置 (rad), ψ_i 为第 i 个驱动关节的驱动角速度 (rad/s). 策略输出动作在执行前直接按关节物理范围进行裁剪:

$$\delta_i \leftarrow \mathrm{clip}(\delta_i,\ \delta_{\min},\ \delta_{\max}),\ \psi_i \leftarrow \mathrm{clip}(\psi_i,\ \psi_{\min},\ \psi_{\max}) \tag{11}$$

在控制接口方面, 本文采用“转向位置控制 + 驱动速度控制”的底层执行方式: δ_i 作为转向关节的转向角位置设定值, ψ_i 作为驱动关节的驱动角速度设定值, 从而保证指令形式与实际执行器接口一致.

2.2　奖励函数设计

为引导策略在端到端输出轮组命令时仍能形成符合物理直觉的协调行为, 本文所提奖励函数的核心思路是将“可控目标”与“可行约束”统一编码为复合型学习信号: 任务项驱动速度跟踪以完成指令运动, 约束项则从垂向冲击、姿态动态、控制平滑性与侧滑趋势四个方面抑制不稳定与非物理行为. 完整的时刻奖励 r_t 定义为

$$r_t = r_{\mathrm{task}}(t) - p_{\mathrm{cons}}(t) \tag{12}$$

其中 $r_{\mathrm{task}}(t)$ 表示任务相关的正向奖励项, $p_{\mathrm{cons}}(t)$ 为稳定性与物理合理性相关的约束惩罚项.

正向奖励项旨在驱动策略跟踪目标速度命令 $\boldsymbol{\xi}_t^{\star} = (v_x^{\star},\ v_y^{\star},\ \omega_z^{\star})$, 鼓励机器人在地形中按指令运动. 本文采用线速度与角速度两项跟踪奖励:

$$r_{\mathrm{lin}} = \exp\left(-\frac{1}{2\sigma^2}\left\|\boldsymbol{v}_{xy} - \boldsymbol{v}_{xy}^{\star}\right\|_2^2\right) \tag{13}$$

$$r_{\mathrm{ang}} = \exp\left(-\frac{1}{2\sigma^2}(\omega_z - \omega_z^{\star})^2\right) \tag{14}$$

其中 $\boldsymbol{v}_{xy} = (v_x,\ v_y)$ 为机体坐标系下的线速度分量, ω_z 为偏航角速度, σ^2 为速度误差的标准化系数. 该类指数型奖励可视作对速度误差施加平滑软约束: 在误差较小时提供连续且稳定的梯度信号, 利于策略在训练后期收敛到可跟踪、可执行的稳态控制行为, 同时避免线性误差项在大偏差区域对优化过程造成过强拉扯.

总任务正向奖励项为

$$r_{\mathrm{task}} = w_1 \cdot r_{\mathrm{lin}} + w_2 \cdot r_{\mathrm{ang}} \tag{15}$$

其中 w_1, w_2 用于平衡平移与转向跟踪的相对重要性, 使策略在不同指令组合下保持一致的控制偏好.

约束惩罚项用于防止策略在追踪速度目标的同时引发不稳定行为 (如上下颠簸、姿态波动、指令频繁跳变或轮胎侧滑等). 本文设计了四项物理惩罚项.

为抑制机器人在粗糙地形中因地形落差或结构回弹所导致的垂直方向不稳定运动, 引入垂直振动惩罚项:

$$p_z = v_z^2 \tag{16}$$

其中 v_z 为机器人在机体坐标系下的垂直线速度, 其平方用于反映地形冲击所引发的上下抖动程度.

为避免横滚与俯仰姿态剧烈波动造成运动不稳, 设置横滚/俯仰惩罚项:

$$p_{\omega_{xy}} = \omega_x^2 + \omega_y^2 \tag{17}$$

其中 ω_x 和 ω_y 分别为绕横轴与纵轴的角速度, 平方和可以抑制急剧姿态变化, 提升底盘动态稳定性.

考虑控制输出平滑性对于物理执行的重要性, 本文引入动作变化率惩罚项, 用于限制连续时间步之间的动作剧烈波动:

$$p_{\mathrm{rate}} = \left\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\right\|_2^2 \tag{18}$$

该项一方面约束控制输出的时间连续性, 降低因指令突变导致的系统振荡; 另一方面在工程上可视为对执行器高频驱动的抑制, 从而在一定程度上兼顾能耗与机械磨损.

最后, 为抑制地形扰动或驱动–转向不协调导致的轮胎横向滑移, 本文设计了轮胎侧滑惩罚项. 侧滑现象不仅会增加姿态扰动, 而且会降低控制精度与能源利用效率. 为量化该行为, 定义轮胎侧滑速度为:

$$v_i^{\perp} = \boldsymbol{s}_i^{\mathrm{T}} \boldsymbol{v}_{w,\,i} \tag{19}$$

其中 $\boldsymbol{v}_{w,\,i}$ 表示第 i 个车轮的轮心速度 (见式 (3)), $\boldsymbol{s}_i = [-\sin\delta_i,\ \cos\delta_i]^{\mathrm{T}}$ 为与滚动方向正交的侧向单位向量. 该定义直接来源于无侧滑 (纯滚动) 假设下的运动学一致性: 当发生侧滑时, $v_i^{\perp}$ 将显著偏离零,

因此可作为侧滑趋势的可计算代理量. 轮胎侧滑惩罚项定义为:

$$p_{\text{slide}} = \sum_{i=1}^{4} F_{z,\,i} \cdot |v_i^{\perp}| \tag{20}$$

其中 $F_{z,\,i}$ 为第 i 个车轮与地面的法向接触力. 该项可视作对“侧向滑移强度”的载荷加权度量: 法向载荷越大且侧滑越明显, 通常对应越高的牵引损失、能耗代价与姿态扰动风险, 因此采用 $F_{z,\,i}$ 加权更贴近轮地相互作用的工程直觉. $F_{z,\,i}$ 在本文中仅用于仿真训练阶段的奖励计算 (学习信号), 不作为策略观测输入, 策略部署时不依赖该信息.

最终约束惩罚项为

$$p_{\text{cons}} = w_3 \cdot p_z + w_4 \cdot p_{\omega_{xy}} + w_5 \cdot p_{\text{rate}} + w_6 \cdot p_{\text{slide}} \tag{21}$$

权重设置遵循“稳定性优先、兼顾跟踪精度”的整定原则; 具体权重超参数在本文中明确列出, 以确保训练配置可复现、结果可对照. 表 1 汇总了奖励函数相关的关键超参数取值.

2.3 策略网络与训练算法

为实现高效、可部署的端到端底层控制策略, 本文采用轻量化多层感知机 (multi-layer perceptron, MLP) 分别表示策略网络与价值网络, 并基于 PPO 算法进行训练. 所选网络结构需在表达能力与推理效率之间取得平衡, 以满足轮组命令级控制对实时性的要求.

策略网络与价值网络使用相同的前向编码结构, 分别独立输出动作分布参数与状态价值估计. 具体而言, 策略网络由三层全连接层构成, 维度依次为 [512, 256, 128]; 激活函数采用 ELU (exponential linear unit); 输出层给出动作均值 $\mu(\boldsymbol{a}_t) \in \mathbf{R}^8$ 与对数标准差 $\ln\sigma(\boldsymbol{a}_t) \in \mathbf{R}^8$, 用于定义高斯策略分布并采样连续动作; 价值网络结构与之相同, 仅将输出层替换为标量 $V(\boldsymbol{s}_t)$, 用于估计状态的长期回报.

表 1 奖励函数与训练相关关键超参数汇总

Table 1 Summary of key hyperparameters related to the reward function and training

符号	含义	取值
σ^2	速度误差标准化系数	0.25
w_1	线速度跟踪权重	2.0
w_2	角速度跟踪权重	1.0
w_3	垂直振动惩罚权重	1.0
w_4	横滚/俯仰角速度惩罚权重	1.0
w_5	动作变化率惩罚权重	0.2
w_6	侧滑惩罚权重	0.001

采用全连接 MLP 的主要考虑在于: 本文观测维度较低 ($\mathbf{R}^{28}$) 且为本体量, 控制任务不涉及图像/点云等高维感知编码, 因而无需引入 CNN 等重型特征提取器. 同时, 端到端轮组命令对推理耗时较敏感, MLP 在保持足够非线性表达能力的同时具有更低的计算开销, 更符合工程部署需求. 类似的配置已在多种机器人本体控制任务中得到验证, 并具有良好的工程复用性[10−11, 14, 25].

训练过程中, 采用 PPO 算法在并行环境中进行样本采集与网络更新, 其优化目标为最大化截断的优势函数代理目标:

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathrm{E}_t\Big[\min\big(\rho_t(\theta)\hat{A}_t,\ \text{clip}\big(\rho_t(\theta),\,1-\epsilon,\,1+\epsilon\big)\hat{A}_t\big)\Big] \tag{22}$$

其中 $\rho_t(\theta) = \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)\,/\,\pi_{\theta_{\text{old}}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ 为策略概率比率, 用于度量新旧策略在同一采样动作上的相对概率; $\hat{A}_t$ 为优势函数的估计值, 用于衡量当前动作相对基准策略的增益; ϵ 为截断系数 (本文采用 $\epsilon = 0.2$), 用于限制策略更新幅度以提升训练稳定性.

优势函数采用广义优势估计 (generalized advantage estimation, GAE) 方法计算:

$$\hat{A}_t = \sum_{l=0}^{T-t-1} (\gamma\lambda)^l \delta_{t+l},\ \delta_t = r_t + \gamma V(\boldsymbol{s}_{t+1}) - V(\boldsymbol{s}_t) \tag{23}$$

式中, δ_t 为单步时序差分残差, $\gamma = 0.99$ 为折扣因子, $\lambda = 0.95$ 为优势估计的衰减系数. GAE 在采样轨迹上进行滑动窗口式累加, 可有效衡量策略在未来的增量收益, 平衡偏差与方差.

策略训练的总损失由策略优化项、状态值函数回归损失项和策略熵项组成, 其表达式为

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PPO}} + c_v\mathcal{L}_{\text{value}} - c_e\mathcal{H}[\pi_\theta] \tag{24}$$

其中, $\mathcal{L}_{\text{PPO}}$ 表示 PPO 策略优化项, $\mathcal{L}_{\text{value}}$ 表示状态值函数回归损失项, $\mathcal{H}[\pi_\theta]$ 为策略熵项, c_v 和 c_e 分别为状态值函数回归损失项与策略熵项的权重系数. 状态值函数回归损失定义为

$$\mathcal{L}_{\text{value}} = \frac{1}{2}\Big(V_\phi(\boldsymbol{s}_t) - \hat{R}_t\Big)^2 \tag{25}$$

式中, $V_\phi(\boldsymbol{s}_t)$ 为价值网络在状态 $\boldsymbol{s}_t$ 下的状态值估计, ϕ 为价值网络参数; $\hat{R}_t$ 为通过采样轨迹计算得到的目标回报估计.

策略熵项用于鼓励策略输出的随机性, 防止策略在早期收敛到局部最优, 其定义为:

$$\mathcal{H}[\pi_\theta] = -\int \pi_\theta(\boldsymbol{a}_t\,|\,\boldsymbol{s}_t)\ln\pi_\theta(\boldsymbol{a}_t\,|\,\boldsymbol{s}_t)\mathrm{d}\boldsymbol{a}_t \tag{26}$$

对于高斯策略分布, 熵可显式计算为:

$$\mathcal{H}[\pi_\theta] = \sum_{i=1}^{n} \ln\left(\sigma_i \sqrt{2\pi \mathrm{e}}\right) \tag{27}$$

其中 σ_i 为第 i 维动作的标准差参数, $n = 8$ 表示动作空间维度. 本文设置各项损失权重为: $c_v = 1.0$, $c_e = 0.01$. 在训练初期, 策略熵项可提高策略探索能力; 训练后期则逐渐收敛至确定性控制行为.

2.4　课程学习与域随机化机制

为了提升策略的泛化能力和训练效率, 本文结合了课程学习 (curriculum learning) 和域随机化 (domain randomization), 从地形结构与机器人物理属性两个层面构建可控扰动, 辅助策略从基本控制向复杂适应逐步过渡.

训练地形采用生成式结构, 并划分为多个子区域, 在每个训练环境中, 由系统按比例布置平坦地形 (60%) 与粗糙地形 (40%). 平坦地形区域具备良好的几何连续性与动力学稳定性, 主要用于训练机器人的基础运动能力. 粗糙地形区域模拟具有起伏不平表面的复杂地貌, 用以训练机器人在非结构化表面上的姿态稳定性与动作鲁棒性. 粗糙地形的高度扰动噪声范围为 [0.05, 0.08] m, 步进间距为 0.05 m, 法向斜率不超过 0.75. 因此, 在训练初期即可同时提供“可控性强”和“扰动性高”的混合样本.

随着训练的进行, 课程机制根据每个子环境中机器人在当前地形等级下的累计移动距离动态调整地形难度. 当机器人在当前地形上的移动距离超过一定阈值 (如地形长度的五分之一) 时, 系统自动提升其对应地形等级, 使其面临起伏更加复杂的地形; 反之, 若其移动距离远低于预期速度下的参考值, 且未触发上调条件, 则适当降低其地形等级. 该机制通过在线难度调节, 确保每个子环境中策略始终面对“略具挑战”的训练情形, 有效避免了策略陷入过拟合平坦场景或在早期遭遇过度扰动所导致的学习崩溃.

同时, 为增强策略的鲁棒性, 在每一回合初始化时, 系统对机器人模型的若干物理属性进行均匀采样扰动, 主要包括:

- 轮地摩擦扰动: 静摩擦系数 μ_s 和动摩擦系数 μ_d 分别在区间 [0.5, 0.9] 内均匀采样;
- 载荷扰动: 在机器人主结构上施加额外质量扰动 $\Delta m \sim \mathrm{U}(-10,\ 30)$ kg;
- 质心扰动: 中心坐标在 x, y 轴方向最大扰动为 ±8 cm, 在 z 轴方向为 ±3 cm.

该机制模拟了现实环境中常见的物理建模偏差、地面材质变化、负载转移等现象.

通过将课程学习与域随机化有机结合, 训练策略得以在具备控制性、稳定性与挑战性的仿真环境中高效进化, 最终形成在多地形与多动力学扰动下均能稳定运行的鲁棒控制策略.

3　实验验证与方法可行性分析

3.1　训练配置与机器人平台

为验证端到端控制策略在复杂地形下的可行性与鲁棒性, 本文基于 Isaac Sim 5.1.0 构建高保真仿真训练与评测平台. 部署环境为 Ubuntu 22.04, 硬件平台为 Intel i9-14900KF、NVIDIA RTX 4080 Super 与 64 GB DDR5 内存. Isaac Sim 支持基于 PhysX 5 的刚体动力学与接触仿真, 并可在并行环境中高效运行训练任务, 从而提高实验验证的效率与可重复性.

仿真平台虽然允许访问机器人系统的完整状态量 (如速度、姿态角、接触相关量等), 但本文端到端控制策略的观测输入严格采用与实机一致的本体量 (关节编码器与 IMU 等, 见状态定义式 (9)), 不依赖仿真特权观测. 其中法向接触力等信息仅用于训练阶段的奖励计算 (学习信号), 不作为策略输入, 以避免策略在部署时对不可获得信息产生依赖.

本文采用并行多环境交互采样进行训练: 并行环境数 $N_{\mathrm{env}} = 4\,096$, 控制频率 $f_{\mathrm{ctrl}} = 20$ Hz, 每次更新采样步数 $N_{\mathrm{step}} = 24$, 总迭代次数 $N_{\mathrm{iter}} = 2\,000$, 则总交互步数约为 $N_{\mathrm{env}} \times N_{\mathrm{step}} \times N_{\mathrm{iter}} \approx 1.97 \times 10^8$. 在上述硬件平台上, 单次完整训练的时间约为 2 h. 该配置明确给出了训练数据规模与计算开销, 使实验设置与结论具备可复现性.

机器人建模方面, 本文使用自主设计的 4WISD 移动机器人作为控制对象, 如图 4 所示. 该机器人具备四个独立转向与驱动关节, 底盘建模基于 URDF/Xacro 格式, 并导出为 USD 文件以集成至仿真环境中. 机器人物理属性设置为: 尺寸 3 020 mm × 1 920 mm × 715 mm, 轮距 2 030 mm, 轴距 1 020 mm, 轮半径 155 mm. 每个轮组具有独立的舵机与驱动器控制接口, 支持精确设定动作指令.

本文控制策略输出的轮组命令为转向关节的转向角位置 δ_i 与驱动关节的驱动角速度 ψ_i. 在仿真与实机部署中, 该命令均通过底层执行器闭环 (位置与速度控制链路) 实现跟踪. 本文假设执行层在其工作带宽内具备足够的响应速度与跟踪精度, 使轮组能够快速且准确地跟随上层给定的目标. 因此研究重点集中于底盘层面的多轮协同控制与稳定通过能力, 而电机级动态建模与控制器整定不作为本文

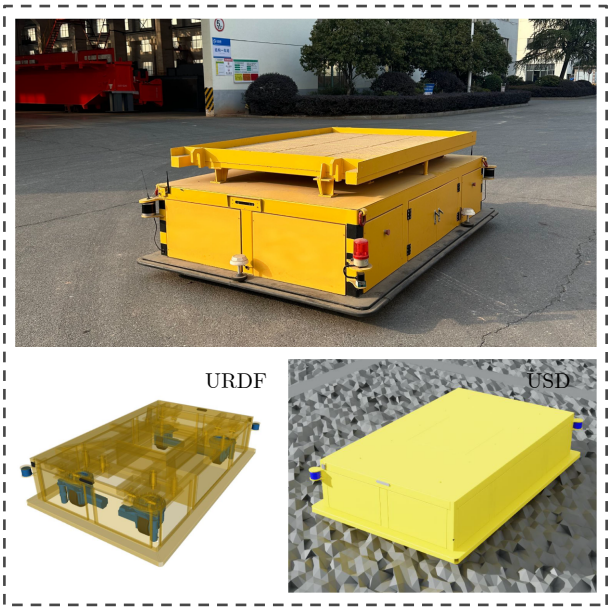


图 4 4WISD 移动机器人实物图 (上) 以及基于 URDF (左下) 与 USD (右下) 格式的仿真模型

Fig. 4 Physical photograph of the 4WISD mobile robot (top), and its simulation models in URDF (bottom left) and USD (bottom right) formats

的实验重点.

3.2 策略训练与收敛性分析

为评估所提端到端控制策略的训练动态与收敛表现, 本文在训练过程中持续记录并可视化关键统计量, 包括平均回报 (mean reward)、课程等级 (curriculum level) 以及由奖励函数分解得到的若干子项指标: 线速度跟踪奖励、角速度跟踪奖励、垂直振动惩罚、姿态扰动惩罚、动作变化率惩罚与侧滑惩罚. 相应曲线如图 5 与图 6 所示.

图 5 左图给出了课程等级随训练推进的变化. 训练初期策略尚未形成稳定控制, 易在粗糙区域出现失稳或低效推进, 课程机制因此将地形难度下调至较低水平. 随着策略逐步学会基本速度跟踪与姿态抑制能力, 课程等级快速回升, 并且后期稳定在约 4.5 的位置上, 这表明策略能够在更具扰动性的

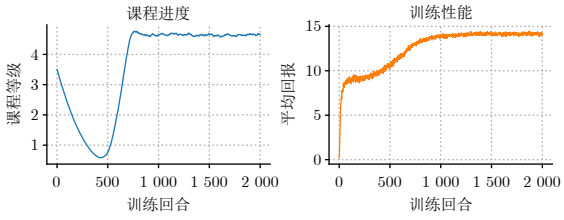


图 5 训练过程中课程等级与平均回报的变化趋势

Fig. 5 The changing trends of curriculum level and mean reward during training

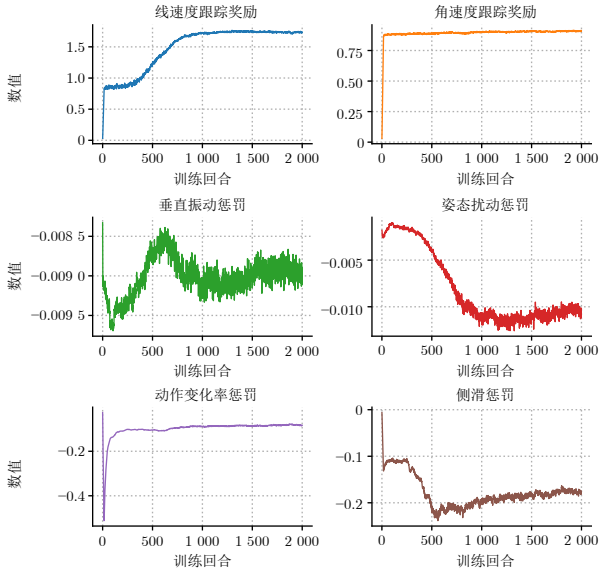


图 6 训练过程中六项关键奖励与惩罚指标的变化趋势

Fig. 6 The changing trends of six key reward and penalty components during training

地形条件下保持可持续的推进能力. 右图为平均回报随训练回合数的变化趋势. 平均回报在前期快速增长, 约在第 750 回合后进入相对平稳阶段, 并最终稳定在 14 附近. 需要指出的是, 该数值与本文奖励尺度及权重设置相对应, 其主要意义在于反映训练过程的相对收敛趋势与性能稳定性, 而非跨任务的绝对可对比指标.

进一步, 图 6 展示了奖励函数各子项在训练过程中的变化. 线速度与角速度跟踪奖励在训练初期快速上升, 并在后期保持高位, 说明策略逐步学会在扰动环境中稳定响应速度指令. 与之对应, 垂直振动惩罚、姿态扰动惩罚、动作变化率惩罚与侧滑惩罚整体呈下降趋势, 表明策略在优化过程中不仅提高了跟踪能力, 而且逐渐抑制了由地形冲击、姿态波动、控制指令突变与轮地不协调引起的非理想行为. 尤其在动作变化率和轮胎侧滑相关项上, 策略表现出显著抑制作用, 反映出本文所设计的奖励结构能够有效引导策略学习到更平滑、更符合轮地物理直觉的多轮协同控制模式.

3.3 控制行为对比与分析

为评估本文端到端控制策略在复杂地形条件下的性能表现, 本文将其与两类经典级联控制基线 (PID-LS 与 MPC-LS) 进行量化对比. 两种基线均采用“机体速度层控制 $\rightarrow$ 轮组分配”的结构: 首先在车体速度层对期望指令 $\boldsymbol{\xi}^{\star}=(v_x^{\star},\ v_y^{\star},\ \omega_z^{\star})$ 进行跟踪控制, 得到机体速度控制量 $\hat{\boldsymbol{\xi}}$ (或其增量); 其次基于式 (6) 所描述的理想运动学约束, 对轮组转向与

轮速命令 $\{\delta_i,\ \psi_i\}_{i=1}^{4}$ 进行运动学逆解, 以生成可执行的轮组控制输入. 相比之下, 本文端到端控制策略直接从本体观测生成轮组命令, 不显式求解上述运动学过程, 从而避免级联结构在复杂轮地交互与地形扰动条件下可能出现的误差积累与性能折损.

为覆盖典型的轮地条件变化与扰动强度差异, 本文选取两类场景进行对比评测: $\mathcal{T}_1$ 为起伏粗糙地形, $\mathcal{T}_2$ 为低摩擦平面地形. 每类场景进一步设置两种强度配置, 分别记为 $\mathcal{T}_{1a}/\mathcal{T}_{1b}$ 与 $\mathcal{T}_{2a}/\mathcal{T}_{2b}$. 粗糙地形以高度扰动范围 Δh 与步进间距 Δs 表征, 其中 $\mathcal{T}_{1a}$ 为 $\Delta h\in[0.05,\ 0.08]$ m、$\Delta s=0.03$ m, $\mathcal{T}_{1b}$ 为 $\Delta h\in[0.02,\ 0.10]$ m、$\Delta s=0.05$ m; 低摩擦平面以静/动摩擦系数 $(\mu_s,\ \mu_d)$ 表征, 其中 $\mathcal{T}_{2a}$ 为 $(0.9,\ 0.9)$, $\mathcal{T}_{2b}$ 为 $(0.2,\ 0.2)$. 上述设置通过几何起伏与接触摩擦条件的系统变化引入轮地非理想效应, 用于检验不同方法在复杂工况下的稳健性差异. 本文选取以下指标对比不同方法的控制性能:

- 速度跟踪指标: 线速度误差 E_{lin} (m/s) 与角速度误差 E_{ang} (rad/s), 用于评估对期望速度指令的跟踪能力;
- 运动稳定性指标: 垂直振动均方根值 Z_{rms} (m/s) 与横滚/俯仰角速度均方根值 RP_{rms} (rad/s), 用于评估在扰动地形下的姿态稳定性;
- 计算效率指标: 控制命令输出时延 T_{ctrl} (ms), 用于评估控制器的实时性.

针对每种方法与每种地形配置 $\mathcal{T}$, 本文将连续 100 个控制步记为一次任务片段, 并重复运行 100 次. 各指标先在单次片段内按时间步计算并求平均, 再对 100 次片段结果统计均值与标准差, 以量化同一配置下的平均性能与波动范围.

表 2 汇总了端到端控制策略与两类传统级联基线 (PID-LS、MPC-LS) 在四种地形配置下的统计结果 (均值 ± 标准差). 整体来看, 当地形几何起伏强度变化 ($\mathcal{T}_{1a}\rightarrow\mathcal{T}_{1b}$) 或接触条件显著变化 ($\mathcal{T}_{2a}\rightarrow\mathcal{T}_{2b}$) 时, 端到端控制策略各项指标的均值与方差变化相对温和, 体现出较好的鲁棒性与一致性. 图 7 展示了端到端控制策略在起伏粗糙地形与平坦室内地面的代表性运动过程.

在起伏粗糙地形 $\mathcal{T}_{1a}/\mathcal{T}_{1b}$ 中, 端到端控制策略在线速度跟踪误差 E_{lin} 上取得最优 (分别为 0.124 m/s 与 0.149 m/s), 且波动幅度较小, 表明其在几何扰动下仍能维持稳定推进并较好地完成速度跟踪任务. 进一步可见, 地形由 $\mathcal{T}_{1a}$ 增强至 $\mathcal{T}_{1b}$ 时, 端到端策略的 E_{lin} 与 E_{ang} 仅出现有限幅度变化 (0.124 m/s → 0.149 m/s, 0.116 rad/s → 0.119 rad/s), 从侧面说明策略对起伏强度的提升具有一定适应能力. 作为对照, MPC-LS 在角速度跟踪 E_{ang} 上更优, 这与 MPC 在机体速度层对角速度误差进行显式优化的机制一致. PID-LS 在 Z_{rms} 与 RP_{rms} 上取得最小值, 但其 E_{lin} 明显更大, 表明其更偏向以保守推进换取姿态平稳. 需要强调的是, 姿态指标更优并不必然意味着整体控制性能更优: 在粗糙地形中, 推进更少通常也更不容易激发较大的姿态动态扰动. 因此, 端到端控制策略在粗糙地形下的稳定性指标虽非最优, 但仍处于可接受范围, 并在保持稳定性的同时取得了更接近任务目标的速度跟踪表现.

在平面地形 $\mathcal{T}_{2a}/\mathcal{T}_{2b}$ 中, MPC-LS 在跟踪精度上总体占优, 尤其是 E_{ang} (0.050 rad/s 与 0.063 rad/s), 反映出当轮地交互更接近理想假设时, 传统级联控制结构能够更充分发挥模型与优化求解的优势. 端

表 2　两类地形及其子配置下端到端控制策略与基线方法的控制性能对比 (均值 ± 标准差)

Table 2　Control performance comparison between the end-to-end control policy and baseline methods under two terrain types and their sub-configurations (mean ± standard deviation)

地形	方法	E_{lin} (m/s)	E_{ang} (rad/s)	Z_{rms} (m/s)	RP_{rms} (rad/s)	T_{ctrl} (ms)
$\mathcal{T}_{1a}$	PID-LS	0.246 ± 0.102	0.120 ± 0.062	**0.086 ± 0.008**	**0.130 ± 0.026**	0.850 ± 0.021
	MPC-LS	0.145 ± 0.125	**0.070 ± 0.064**	0.101 ± 0.012	0.160 ± 0.044	2.913 ± 0.088
	end-to-end (PPO)	**0.124 ± 0.057**	0.116 ± 0.066	0.106 ± 0.009	0.173 ± 0.055	**0.308 ± 0.018**
$\mathcal{T}_{1b}$	PID-LS	0.291 ± 0.137	0.132 ± 0.067	**0.106 ± 0.012**	**0.159 ± 0.036**	0.851 ± 0.015
	MPC-LS	0.217 ± 0.151	**0.098 ± 0.062**	0.121 ± 0.015	0.209 ± 0.045	2.880 ± 0.089
	end-to-end (PPO)	**0.149 ± 0.070**	0.119 ± 0.069	0.125 ± 0.018	0.205 ± 0.071	**0.303 ± 0.020**
$\mathcal{T}_{2a}$	PID-LS	0.222 ± 0.088	0.101 ± 0.053	**0.075 ± 0.007**	**0.061 ± 0.010**	0.852 ± 0.019
	MPC-LS	**0.096 ± 0.094**	**0.050 ± 0.049**	0.083 ± 0.008	0.064 ± 0.017	2.866 ± 0.094
	end-to-end (PPO)	0.105 ± 0.070	0.105 ± 0.064	0.085 ± 0.008	0.081 ± 0.029	**0.305 ± 0.019**
$\mathcal{T}_{2b}$	PID-LS	0.210 ± 0.090	0.136 ± 0.062	**0.078 ± 0.009**	**0.060 ± 0.013**	0.848 ± 0.021
	MPC-LS	0.098 ± 0.079	**0.063 ± 0.040**	0.085 ± 0.009	0.062 ± 0.019	2.870 ± 0.100
	end-to-end (PPO)	**0.098 ± 0.057**	0.102 ± 0.067	0.084 ± 0.006	0.085 ± 0.029	**0.301 ± 0.030**

到端控制策略在低摩擦配置 $\mathcal{T}_{2b}$ 下的 E_{lin} 与 MPC-LS 持平 (同为 0.098 m/s), 同时其 Z_{rms} 与 RP_{rms} 虽不及 PID-LS 的最小值, 但仍处于同一量级, 未体现出明显的稳定性退化, 说明策略在接触条件变化下仍能维持基本一致的控制表现.

从计算效率看, 端到端控制策略在所有地形配置下均取得最小的 T_{ctrl} (约 0.304 ms), 明显低于 PID-LS (约 0.850 ms) 与 MPC-LS (约 2.882 ms). 这表明端到端控制策略在在线推理侧具有较大的计算余量, 为后续在更高频控制或与更复杂上层模块耦合时留出了更充分的实时性空间. 综上所述, 端到端控制策略在地形与摩擦条件变化时性能波动较小, 体现出较好的稳健性, 而 PID-LS、MPC-LS 级联基线在稳定性或角速度精度上各有优势, 但往往伴随更保守的任务执行或更高的在线计算代价.

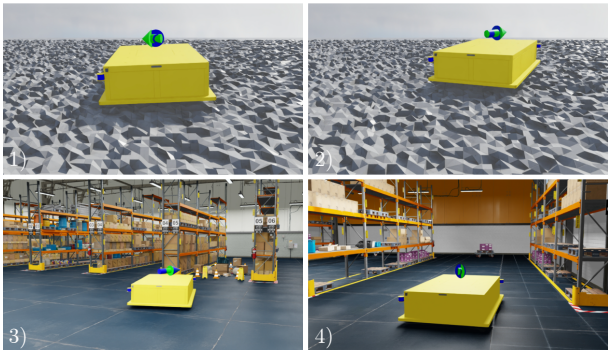
3.4 实机部署验证

为验证所学端到端控制策略在真实机器人上的可执行性与 sim-to-real 的可迁移性, 进一步在室外平整水泥地面开展实机部署实验. 本节重点检验策略在真实硬件闭环中的在线运行能力与速度跟踪效果, 包括传感器数据同步、状态构建、策略推理以及命令下发在 ROS-PLC 控制链路中的一致性与稳定性. 考虑端到端控制策略直接输出轮组命令在实机上存在潜在的安全风险, 复杂地形与高动态工况的系统性验证不作为本节的目标. 图 8 展示了实机平台与室外实验环境.

实机部署采用车载工控机运行策略推理节点, 并通过 ROS-PLC 控制链路完成轮组控制指令的下发与驱动器写入. 策略输入由实机可获得的本体传感与里程计估计量构建, 并与训练阶段的输入形式保持一致; 策略输出为四轮转向角与轮速命令, 经物理范围裁剪后交由底层控制器执行. 图 9 给出了实机端到端控制链路与节点通信关系的概览.

考虑到工控机与 PLC 之间的通信与执行延迟, 系统实际闭环更新受链路约束, 控制周期固定为 10 Hz. 每个周期内依次完成传感器数据同步与状态构建、策略前向推理、命令发布以及执行器闭环跟踪. 为量化在线计算与链路开销, 本文对关键环节耗时进行统计: 单步策略前向推理的平均耗时为 0.62 ms, 而传感器同步与状态构建环节的平均耗时为 56.15 ms, 表明实机闭环周期主要受数据同步与链路时延限制, 策略推理本身仍保有充足计算余量.

实验过程中通过遥控手柄在线发布速度指令, 并连续记录约 120 s 的指令与反馈数据, 用于离线分析. 图 10 给出了实机速度跟踪曲线示例: 策略能够在真实硬件闭环中持续稳定运行, 三轴速度测量



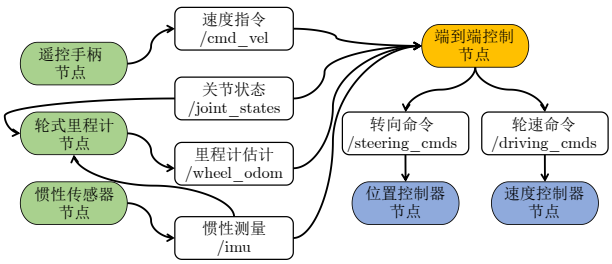
注: 绿色箭头表示期望机体速度方向, 蓝色箭头表示实际执行速度方向.

图 7 不同地形下的端到端控制执行过程

Fig. 7 End-to-end control execution process under different terrains

图 8 在室外平整水泥地面上的实机速度指令跟踪实验

Fig. 8 Real-world velocity-command tracking experiment on an outdoor flat concrete surface



注: 遥控手柄输入、轮式里程计、惯性传感器、关节状态和里程计估计构成策略输入, 端到端控制节点输出轮速命令, 并由位置控制器和速度控制器执行.

图 9 实机端到端控制链路与 ROS 节点通信关系

Fig. 9 Real-world end-to-end control pipeline and ROS node communication relationship

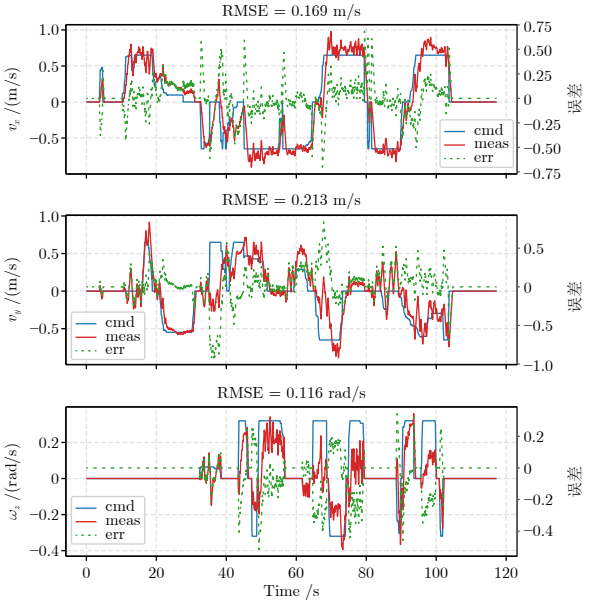


图 10　实机速度跟踪结果

Fig. 10　Real-world velocity tracking results

曲线整体随指令变化呈现一致的跟随趋势. 其中, cmd 为指令值 (蓝实线), meas 为实测值 (红实线), err 为误差 (绿虚线). 对应的速度跟踪 RMSE (root mean square error) 分别为 v_x : 0.169 m/s、v_y : 0.213 m/s、ω_z : 0.116 rad/s, 其主要表现为指令切换与快速换向阶段的短时滞后与波动. 鉴于实机速度反馈依赖轮式里程计估计且不可避免地受噪声与通信时延影响, 同时本文未针对执行器与状态估计进行额外工程优化, 上述误差量级在当前控制链路与频率约束下具有合理解释. 总体而言, 该实机结果直接验证了端到端控制策略的 sim-to-real 可部署性: 在仅使用本体传感与标准 ROS-PLC 控制链路的条件下, 策略能够实现可复现的在线运行与速度跟踪, 并为后续提升闭环频率、延长运行时程及扩展更复杂场景验证提供了工程基础.

3.5　方法讨论与局限性

本文提出的端到端控制策略在无需显式求解运动学模型与精确参数标定的前提下, 能够直接从本体观测学习轮组命令, 并在起伏粗糙与低附着两类典型工况中保持较为一致的控制表现. 与 PID-LS、MPC-LS 等传统基线相比, 端到端控制策略在粗糙地形下更接近任务目标的线速度跟踪, 并在地形强度与摩擦条件变化时呈现较小的性能波动, 体现出一定的稳健性与多轮隐式协调能力. 同时, 实机部署实验验证了策略可在 ROS-PLC 控制链路中稳定在线运行, 并在真实平台上实现可复现的速度跟踪, 为 sim-to-real 工程落地提供了有力支撑.

尽管如此, 本文仍存在若干局限性与待扩展方向. 首先, 仿真实验虽覆盖了起伏粗糙与低附着两类场景, 但仍难以充分反映真实环境中的材料非均匀性、局部沉陷、碎石滚动等更复杂的轮地交互. 本文亦未对大坡度斜坡等 3 维地形进行系统验证, 后续需结合更完备的状态估计与地形表示能力扩展场景覆盖. 其次, 实机实验目前主要在平整地面上进行, 核心量化指标聚焦于速度跟踪误差与在线推理耗时; 在更高速度、更强扰动或更长时程运行条件下, 仍需引入更完善的安全监控与在线约束机制, 以降低潜在的失控风险并提升工程可靠性. 最后, 本文策略训练依赖大规模仿真交互数据, 并通过课程机制与域随机化提升稳健性. 当平台结构、载荷分布或执行器特性发生显著变化时, 策略仍可能需要再训练或采用迁移学习与微调, 以维持性能与稳定性.

4　结束语

本文面向复杂地形条件下四轮独立转向与驱动移动机器人的底层运动控制问题, 提出一种不依赖显式运动学与动力学在线计算的端到端控制策略学习方法. 该方法基于深度强化学习框架, 通过课程学习与域随机化逐步提升策略对地形扰动与参数不确定性的适应能力. 同时结合物理启发的奖励函数设计, 在速度跟踪、姿态稳定与控制可执行性之间建立一致的学习目标, 从而引导策略学习到更符合工程直觉的多轮协同行为. 策略网络采用轻量化 MLP 结构, 并基于 PPO 算法训练, 兼顾收敛稳定性与在线推理效率.

实验结果表明, 所提策略在课程机制驱动下能够持续提升回报并稳定收敛. 在仿真评估中, 本文进一步引入 PID-LS 与 MPC-LS 级联基线, 并在起伏粗糙地形与低摩擦平面地形 (含不同参数子配置) 下进行对比, 结果显示, 端到端控制策略在地形扰动与参数偏差条件下具备良好的稳定性与执行鲁棒性. 实机部署验证的结果进一步表明, 该策略可在 ROS-PLC 控制链路中稳定运行, 并实现对速度指令的基本跟踪, 同时单步推理耗时处于可接受范围内, 为工程落地提供了初步依据.

未来研究可在本文框架基础上进一步扩展任务闭环能力, 引入更强的感知与表示学习模块 (如融合视觉/点云等高维输入的编码网络), 将端到端学习从底盘速度跟踪拓展到更贴近真实任务的导航避障、目标导向探索与鲁棒路径跟随等场景, 从而提升 4WISD 平台在复杂环境中的自主性与任务完成度.

参考文献

1 Li D Y, Song Y D, Huang D, Chen H N. Model-independent adaptive fault-tolerant output tracking control of 4WS4WD road vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2013, **14**(1): 169−179

2 Guo J H, Luo Y G, Li K Q. An adaptive hierarchical trajectory following control approach of autonomous four-wheel independent drive electric vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2018, **19**(8): 2482−2492

3 Liang Y, Li Y, Zheng L, Yu Z, Ren Y. Yaw rate tracking-based path-following control for four-wheel independent driving and four-wheel independent steering autonomous vehicles considering the coordination with dynamics stability. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, 2021, **235**(1): 260−272

4 Liu X J, Wang G L, Chen K. Nonlinear model predictive tracking control with C/GMRES method for heavy-duty AGVs. *IEEE Transactions on Vehicular Technology*, 2021, **70**(12): 12567−12580

5 Ding T, Zhang Y H, Ma G C, Cao Z H, Zhao X W, Tao B. Trajectory tracking of redundantly actuated mobile robot by MPC velocity control under steering strategy constraint. *Mechatronics*, 2022, **84**: Article No. 102779

6 Liu X X, Wang W, Li X L, Liu F S, He Z H, Yao Y Z, et al. MPC-based high-speed trajectory tracking for 4WIS robot. *ISA Transactions*, 2022, **123**: 413−424

7 Jeong Y, Kim D H, Youn M H, Park S, Li X J, Lee T. Model predictive control-based path tracking with four-wheel independent steering, driving, and braking autonomous vehicles on low friction road. In: Proceedings of the 23rd International Conference on Control, Automation and Systems (ICCAS). Busan, Korea: IEEE, 2023. 1427−1432

8 Nguyen N T, Gangavarapu P T, Mandel N, Bruder R, Ernst F. Motion planning for 4WS vehicle with autonomous selection of steering modes via an MIQP-MPC controller. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 2024. 9765−9771

9 Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T, Harley T, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the International Conference on Machine Learning (ICML). New York, USA: PMLR, 2016. 1928−1937

10 Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 2020, **5**(47): Article No. eabc5986

11 Miki T, Lee J, Hwangbo J, Wellhausen L, Koltun V, Hutter M. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 2022, **7**(62): Article No. eabk2822

12 Belmonte-Baeza A, Lee J, Valsecchi G, Hutter M. Meta reinforcement learning for optimal design of legged robots. *IEEE Robotics and Automation Letters*, 2022, **7**(4): 12134−12141

13 Margolis G B, Agrawal P. Walk these ways: Tuning robot control for generalization with multiplicity of behavior. In: Proceedings of the 6th Conference on Robot Learning (CoRL). Atlanta, USA: PMLR, 2023. 22−31

14 Margolis G B, Yang G, Paigwar K, Chen B, Agrawal P. Rapid locomotion via reinforcement learning. *The International Journal of Robotics Research*, 2024, **43**(4): 572−587

15 Chen Ci, Yu Ji-Yu, Li Chao, Lu Hao-Jian, Gao Hong-Bo, Xiong Rong, et al. Structure-control co-design of quadruped robots based on pre-training-fine-tuning framework. *Robot*, 2025, **47**(5): 625−635
(陈词, 余纪宇, 李超, 陆豪健, 高洪波, 熊蓉, 等. 基于预训练−微调框架的四足机器人结构−控制协同设计. 机器人, 2025, **47**(5): 625−635)

16 Zhang Nan-Jie, Chen Yu-Quan, Ji Mao-Qin, Sun Yun-Kang, Wang Bing. Adaptive control method for quadruped robot facing floors of different roughness. *Acta Automatica Sinica*, 2025, **51**(7): 1585−1598
(张楠杰, 陈玉全, 季茂沁, 孙运康, 王冰. 面向不同粗糙程度地面的四足机器人自适应控制方法. 自动化学报, 2025, **51**(7): 1585−1598)

17 Li Z Y, Cheng X X, Peng X B, Abbeel P, Levine S, Berseth G, et al. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi′an, China: IEEE, 2021. 2811−2817

18 Siekmann J, Godse Y, Fern A, Hurst J. Sim-to-real learning of all common bipedal gaits via periodic reward composition. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi′ an, China: IEEE, 2021. 7309−7315

19 Jeon S H, Heim S, Khazoom C, Kim S. Benchmarking potential based rewards for learning humanoid locomotion. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). London, UK: IEEE, 2023. 9204−9210

20 Zhang Q, Cui P, Yan D, Sun J K, Duan Y Q, Han G, et al. Whole-body humanoid robot locomotion with human reference. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE, 2024. 11225−11231

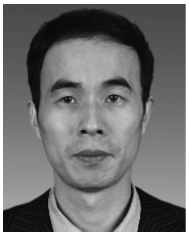
21 Wu Li-Ming, Wang Ming-Ming, Luo Jian-Jun, Zhang Da-Yu, Liang Cheng-Xi. Compliant control method for on-orbit assembly by space robots based on reinforcement learning. *Robot*, 2025, **47**(2): 239−248
(武黎明, 王明明, 罗建军, 张大羽, 梁澄汐. 基于强化学习的空间机器人在轨装配柔顺控制方法. 机器人, 2025, **47**(2): 239−248)

22 Xu D G, Chen P, Zhou X H, Wang Y Z, Tan G Z. Deep reinforcement learning based mapless navigation for industrial AMRs: Advancements in generalization via potential risk state augmentation. *Applied Intelligence*, 2024, **54**(19): 9295−9312

23 Wang Y Z, Xie Y F, Xu D G, Shi J H, Fang S Y, Gui W H. Heuristic dense reward shaping for learning-based map-free navigation of industrial automatic mobile robots. *ISA Transactions*, 2025, **156**: 579−596

24 Wang Y Z, Xu D G, Xie Y F. Universal multi-mode kinematic controller for four-wheel independent steering and driving mobile robots. In: Proceedings of the China Automation Congress (CAC). Harbin, China: IEEE, 2025.

25 Mittal M, Yu C, Yu Q, Liu J, Rudin N, Hoeller D, et al. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 2023, **8**(6): 3740−3747



徐德刚 中南大学自动化学院教授. 2007 年获得浙江大学信息学院工业控制技术国家重点实验室控制科学与工程专业博士学位. 主要研究方向为复杂工业过程优化控制和智能机器人控制. E-mail: dgxu@csu.edu.cn

(**XU De-Gang** Professor at the School of Automation, Central South University. He

received his Ph.D. degree in control science and engineering from the College of Information Science and Engineering and the State Key Laboratory of Industrial Control Technology, Zhejiang University in 2007. His research interests include optimal control of complex industrial processes and intelligent robot control.)

王逸之 2026 年获得中南大学自动化学院控制科学与工程专业博士学位. 主要研究方向为机器人与人工智能, 学习决策与自主控制. 本文通信作者.
E-mail: wangyizhi@csu.edu.cn
(**WANG Yi-Zhi** Received his Ph.D. degree in control science and engineering from the School of Automation, Central South University in 2026. His research interests include robotics and artificial intelligence, learning-based decision making, and autonomous control. Corresponding author of this paper.)

谢永芳 中南大学自动化学院教授. 1993、1996、1999 年分别获得中南大学信息科学与工程学院控制科学与工程专业学士、硕士、博士学位. 主要研究方向为工业人工智能, 计算机视觉和鲁棒控制.
E-mail: yfxie@csu.edu.cn
(**XIE Yong-Fang** Professor at the School of Automation, Central South University. He received his bachelor, master, and Ph.D. degrees in control science and engineering from the College of Information Science and Engineering, Central South University in 1993, 1996, and 1999, respectively. His research interests include industrial artificial intelligence, computer vision, and robust control.)