DOI: 10.19650/j.cnki.cjsi.J2614858

基于深度强化学习的仿蟹机器人运动控制研究*

王诗雯,张 军,温炜涛,李宏洋,宋爱国
(东南大学仪器科学与工程学院 南京 210096)

摘 要:当前双足和四足机器人的强化学习运动控制研究已取得显著进展,但八足机器人由于自由度冗余高、足端易干涉和动力学建模复杂等问题,其深度强化学习控制仍处于探索阶段。针对这一问题,提出了一种基于深度强化学习的仿蟹八足机器人运动控制框架。首先,通过分析螃蟹的身体结构和运动步态,设计了仿蟹机器人的机构模型和参考步态,为强化学习提供物理先验知识。随后,设计了模仿螃蟹步态参考动作的端到端深度强化学习训练框架,并引入行为克隆初始化策略以加速策略收敛,同时采用分阶段强化学习方法逐步增加地形难度,保证机器人在平地步态基础上稳定迁移至复杂地形。进一步对比分析了不同学习率调整策略、近端策略优化算法(PPO)与优势演员-评论家(A2C)、软演员-评论家(SAC)、深度确定性策略梯度(DDPG)算法的适配性,并通过域随机化训练及行为克隆+分阶段强化学习方法研究了模型在动力学扰动和多样地形下的泛化能力。仿真实验结果显示,所提出框架在训练效率上优于其他算法,步态稳定性指标(即五步质心高度标准差)约为 0.004 m;机器人在斜坡、台阶等复杂地形中均能实现稳定行进,运动性能优于训练的模型。最后,进行了初步样机实验,验证了所提方法在真实物理环境中的可行性和稳定性。所提出的深度强化学习框架结合了生物步态模仿、行为克隆初始化及分阶段强化学习,对八足机器人的运动控制提供了新的技术思路。

关键词:仿生机器人;八足机器人;深度强化学习;神经网络控制器;域随机化
中图分类号:TH113 TP242 TP181 **文献标识码**:A **国家标准学科分类代码**:510.80

Research on the deep reinforcement learning based locomotion control of a crab-like robot

Wang Shiwen, Zhang Jun, Wen Weitao, Li Hongyang, Song Aiguo
(*School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China*)

Abstract: The research of reinforcement learning-based motion control has progressed significantly for the bipedal and quadrupedal robots. However, the deep reinforcement learning control of octopod robots is still in the exploratory stage due to the issues such as high degrees of freedom redundancy, easy interference at the foot end, and complex dynamic modeling. To address this problem, this paper proposes a motion control framework of crab-like octopod robots based on the deep reinforcement learning. Firstly,, the mechanism model and reference gait of the crab-like robot are designed by analyzing the body structure and gait of crabs, which provides the physical prior knowledge for reinforcement learning. Subsequently, an end-to-end deep reinforcement learning training framework is designed to imitate the reference actions of crab gaits, where a behavior cloning initialization strategy is introduced to accelerate the convergence of the policy. At the same time, a phased reinforcement learning method is adopted to gradually increase the terrain difficulty, ensuring that the robot can stably transfer from the flat ground gait to complex terrains. This paper further compares and analyzes different learning rate adjustment strategies as well as the adaptability of the proximal policy optimization (PPO) algorithm with actor-critic (A2C), soft actor critic (SAC), and deep deterministic policy gradient (DDPG) algorithms. Additionally the generalization ability of model under the dynamic disturbances and diverse terrains through domain randomization training and behavior cloning + staged reinforcement learning methods is studied. The simulation experiment results show that the proposed framework outperforms other algorithms in training efficiency, with the gait stability index(i.e., the standard deviation of the five-step center of mass height) of approximately 0.004 m.

收稿日期:2026-01-15 Received Date: 2025-01-15
*基金项目:国家自然科学基金项目(62173090)资助

The robot can achieve the stable movement on complex terrains such as slopes and steps, whose motion performance is superior to that of the trained model. Finally, a preliminary prototype experiment was conducted to verify the feasibility and stability of proposed method in a real physical environment. In conclusion, the proposed deep reinforcement learning framework combing the biological gait imitation, behavior cloning initialization and phased reinforcement learning provides a new technical approach for the motion control of octopod robots.

Keywords: bionic robot; eight-legged robot; deep reinforcement learning; neural network controller; domain randomization

0 引 言

轮式、履带式和足腿式是机器人主要的运动形式[1]。其中足腿式机器人由于其离散的立足点和在任意方向上生成轨迹的能力[2],在复杂地形中展现出巨大的应用潜力。常见的多足机器人主要包括四足、六足和八足机器人。四足机器人典型的代表有波士顿动力的 SpotMini[3]、麻省理工学院的 Cheetah 3[4]、苏黎世联邦理工学院的 ANYMal[5]、宇树科技的 A1 (Unitree)[6]等。六足机器人如中南大学的 PH-ROBOT[7]等。

蟹类独特的肢体结构和运动机制[8]在地球多样化的生态系统中展现出了卓越的适应性,成为八足机器人研究模仿的对象。现有仿蟹八足机器人研究主要集中在机构设计、步态生成、控制器设计和应用场景适配等方面。Liu 等[9]设计了具有 8 条仿生腿和两个前爪的仿蟹机器人,仅用两个电机实现步态控制。Grzelczyk 等[10]在正弦函数步态生成器基础上提出改进型步态生成器,显著提升了八足机器人的运动性能。王刚等[11-12]借鉴螃蟹的生理特征设计了八足机器人机构,通过视频分析研究了螃蟹的运动步态,并构建了不同地形下的机器人足端轨迹库。周敬淞等[13]通过观察中华绒螯蟹的外骨骼与肌肉组织,分析其运动步态,设计了一款刚柔耦合的八足仿蟹机器人,实现了欠驱动和多步态运动控制。

目前的仿蟹八足机器人研究主要基于预编程步态和简单闭环控制方法展开,依赖精确的动力学建模,控制策略在复杂环境中的泛化性不足,对环境和机械参数敏感。近年来,快速发展的深度强化学习方法(deep reinforcement learning, DRL)为上述问题提供了新思路,已成功应用于机械臂和低自由度的足式机器人的控制。杨傲雷等[14]基于人手臂的运动特性设计奖励函数,采用强化学习方法进行机械臂运动模型的训练。苏黎世联邦理工学院进行了无传感器的强化学习[15-16]和基于多传感器的强化学习[17]的四足机器人控制研究。王宇等[18]把强化学习应用在六足机器人上,并进行了陆地和水下两种运动控制模式的探索。此外,在机器人柔顺控制领域,李逃昌等[19]提出了一种基于强化学习的自适应导纳控制方法,有效解决了固定参数控制受机器人建模误差影响的问题。在路径规划领域,曾宁坤等[20]将深度强化学习与哈里斯鹰算法相结合,动态生成关键参数,提升了复杂环境下的路径搜索能力。

鉴于完全依赖自由探索的强化学习训练对控制器精度要求高,奖励函数设计和超参数调节难度大,研究者引入生物稳定的周期性步态作为先验信息,构建基于模仿学习的强化学习框架。Liu 等[21-23]提出将参考运动信息引入奖励函数的 DeepMimic 方法,以减少对复杂技能特异性奖励函数的依赖。基于该方法,Xie 等[24]实现了 Cassie 双足机器人稳定行走策略的仿真训练,并通过系统辨识完成了从仿真到现实的迁移。Peng 等[25]通过模仿拉布拉多犬的运动行为,实现了 Laikago 四足机器人行走、奔跑和转向等的稳定控制。Feng 等[26]结合形态学随机化,提出了一种基于模仿学习的控制器,可适配多种四足机器人。

综上,现有深度强化学习方法主要集中于双足和四足机器人控制,缺少在八足机器人运动控制方面的研究。八足机器人通过更多腿足提升运动稳定性和环境适应性,其在主流研究中都基于预编程步态和关节角度进行控制。然而相较于四足机器人,八足机器人存在自由度冗余、足端干涉、复杂步态时序等特性,使得其深度强化学习运动控制在策略收敛性、步态稳定性和泛化能力方面面临更严峻挑战,直接迁移四足机器人的方法难以获得有效控制策略。本文提出并设计了一种基于深度强化学习的仿蟹八足机器人运动控制方法。根据螃蟹特点设计了仿蟹机器人的机构,并处理螃蟹的运动步态,创新性地提出了根据简单运动步态完成较复杂地形下步态训练的强化学习框架,且不依赖精确动力学模型。框架包含模仿螃蟹步态的强化学习环境设计和通过域随机化与行为克隆-分阶段强化学习进行步态泛化两个部分,在模型训练过程中探究了学习率和强化学习算法对整个框架的影响,并使用域随机化和行为克隆-分阶段强化学习进行模型泛化训练,在仿真实验中证明了所提方法的有效性,并在样机试验中进一步证明所提方法的现实可行性。

1 仿蟹机器人机构及步态设计

1.1 机器人机构设计

已有仿蟹机器人腿部一般简化为三自由度。为了研究螃蟹腿部多自由度的运动性能,本文以中华绒螯蟹为仿生对象,进行生物数据采集,并统计和分析其躯干及腿

足尺寸,在设计仿蟹步行足机构时进行简化处理,将底-基-座节和躯干、底-基-座节和长节、长节和腕-前节、腕-前节和指节分别设计为转动副,由此得到一个具有四自由度的步行足机构,共设计有 4 个关节,称为基节、长节、腕节和指节,如图 1 所示,其中 a_1、a_2、a_3、a_4 分别为基节、长节、腕节和指节的长度。

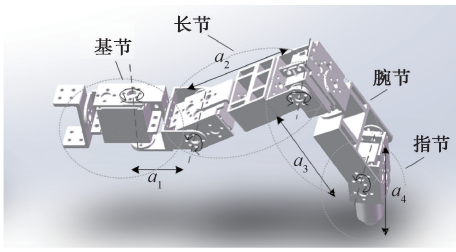


图 1　仿蟹机器人步行足结构及旋转关节示意图

Fig. 1　Schematic diagram of the walking leg structure and rotational joints of the crab-like robot

为简化设计流程,仿蟹机器人基于以上步行足结构,采用左右侧对称布局,4 对相同步行足等间距分布在躯体两侧,机器人整体设计图如图 2 所示。

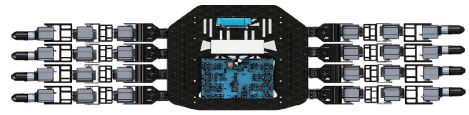


图 2　仿蟹机器人整体设计图

Fig. 2　Overall design of the crab-like robot

1.2　机器人步态设计

基于对螃蟹行走数据的处理,设计了机器人的仿生步态,分为仿蟹腿关节运动轨迹处理和多足运动步态规划两部分。本研究基于螃蟹运动数据,提取并重构其足端相对于甲壳的运动轨迹。在此基础上,依据螃蟹步行足与躯体的尺寸比例关系,对单步轨迹进行归一化处理。再结合机器人步行足与躯体结构参数,对无量纲轨迹进行运动学放缩,得到机器人离散足端轨迹。最后,根据如图 3 的 D-H 法建立的机器人腿部运动学模型[27]。

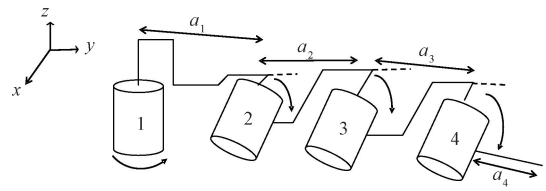


图 3　仿蟹机器人腿部运动学模型

Fig. 3　Kinematic model of the leg of the crab-like robot

通过逆运动学由足端轨迹得到关节转动角度[27],即:

$$\begin{cases}\theta_1 = \tan^{-1}\left(\dfrac{y}{x}\right)\\ \theta_2 = \tan^{-1}\dfrac{(c_3a_3+a_2)(z-s_{234}a_4)-(s_3a_3)(xc_1+ys_1-a_4c_{234})}{(c_3a_3+a_2)(xc_1+ys_1-a_4c_{234})+(s_3a_3)(z-s_{234}a_4)}\\ \theta_3 = \cos^{-1}\left(\dfrac{(xc_1ys_1-a_4c_{234}-a_1)^2+(z-s_{234}a_4)^2-a_2^2-a_3^2}{2a_2a_3}\right)\\ \theta_4 = \tan^{-1}\left(\dfrac{z}{c_1x-s_1y}\right)-\theta_3-\theta_2\end{cases} \tag{1}$$

式中:(x, y, z) 表示足端坐标;$s_i=\sin\theta_i$,$c_i=\cos\theta_i$,$i=1, 2, 3, 4$;$s_{234}=\sin(\theta_2+\theta_3+\theta_4)$;$c_{234}=\cos(\theta_2+\theta_3+\theta_4)$。由于本文主要研究仿蟹机器人的横向运动控制,因此基节角度固定不变,即 θ_1 设置为常量。此外,根据生物步行足特性对其他关节角度进行限幅,$\theta_2\in[-90°, 0°]$,θ_3、$\theta_4\in[0°, 90°]$,又根据采集得到的生物数据中足端与地面夹角 $\alpha=(\theta_2+\theta_3+\theta_4)$,计算出 s_{234} 与 c_{234} 简化求解过程,从而得到唯一的逆解。得到单腿单步的各关节角度数据之后,需要设计 8 条腿的运动步态。根据文献[11],波形步态是八足机器人综合最优步态,生物数据也同样表明波形步态为中华绒螯蟹的主要步态[27],因此本文中仿蟹机器人采用波形步态。波形步态是一种对称规则步态,是足式机器人的基本运动模式,指机器人处于躯体对称轴两侧的步行足相位相差半个步态周期,同侧步行足按其在躯体上的排列顺序由后至前依次抬起和落下[28]。为减少能量损耗和机体稳定,令同侧步行足构成等相位步态,根据定义 $2n$ 足步行机器人的相对相位[29]为:

$$\varphi_{2m+1} = 1 - m/2n,\ m = 1, 2, \cdots,\ n-1 \tag{2}$$

式中:相对相位 φ_i 是指在一个步态周期中第 i 条步行足着地时刻滞后于第 1 条步行足着地时刻的时间占整个步态周期的比例,$2m+1$ 表示左侧步行足的序数,对八足仿蟹机器人来说 $n=4$,同组同侧相邻步足的相位差 $\Delta\phi=\pi/2$,且根据先后摆动的两步行足相位相差 $\Delta\times\pi$,$\Delta\phi=2\Delta\times\pi$,计算出相位因数 $\Delta=0.25$,据此设计了如图 4 所示的等相位的波形步态用于后续强化学习训练。横轴表示时间,纵轴表示不同的步行足,将步行足分为引导侧与跟随侧两类,引导侧步行足定义为与螃蟹横向运动方向同侧的步行足,L1、L2、L3、L4 表示引导侧的 4 条步行足,T1、T2、T3、

T4 表示跟随侧的 4 条步行足,黑色块表示支撑相,斜线块表示摆动相。

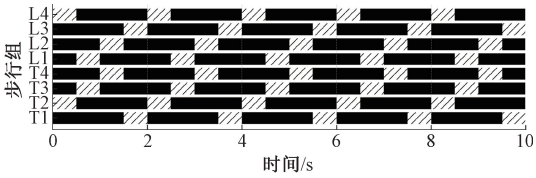


图 4 等相位波形步态图

Fig. 4 Equal-phase waveform of gait

2 强化学习环境设计

2.1 基于 PPO 的强化学习

近端策略优化算法(proximal policy optimization, PPO)[30]是一种基于策略梯度(policy gradient, PG)的演员-评论家(actor-critic,AC)强化学习方法。通过引入裁剪机制约束策略更新幅度,在保证训练稳定性的同时高效优化策略。策略梯度算法的核心思想是直接优化策略 $\pi_\theta(a|s)$,通过计算期望回报对策略参数 θ 的梯度 ∇_θ,沿着梯度方向更新策略,使得未来能获得更高回报的动作的概率增加,使得累积奖励 $J(\theta)$ 最大化。为解决传统的 PG 方法面临较高的方差问题,演员-评论家算法引入一个“评论家”来评估动作的优劣,加入状态价值函数 $V(s_t)$ 来计算得出优势函数 A_t,并用 A_t 代替蒙特卡洛回报 G_t,将策略梯度公式更新为:

$$\nabla_\theta J(\theta) = \mathrm{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\nabla_\theta \log\pi_\theta(a_t|s_t)A_t\right] \tag{3}$$

在演员-评论家算法基础上,PPO 通过裁剪系数 ε 限制策略更新幅度,将目标函数优化为:

$$L^{CLIP}(\theta) = \mathrm{E}_t[\min(r_t(\theta)A_t, \mathrm{clip}(r_t(\theta),1-\varepsilon, 1+\varepsilon)A_t)] \tag{4}$$

式中:$r_t(\theta)$ 为新旧策略的概率比,$\mathrm{clip}(x, k_1, k_2)$ 将 x 的值限制在 k_1 与 k_2 之间,这样的修改有效地限制了每次更新的幅度,使得通过梯度下降法对策略参数 θ 进行优化时避免策略更新过度,从而提高训练的稳定性。考虑到仿蟹机器人的运动控制涉及高维连续的状态-动作空间,采用直接输出连续动作的演员-评论家算法显得尤为重要。为了进一步提高训练过程的稳定性,本研究选择 PPO 算法作为强化学习的基础方法。

2.2 环境设计

强化学习作为一种以试错交互为核心的学习范式,其性能和效果在很大程度上依赖于合理的环境设计,针对具体问题,需精确设置马尔可夫决策过程(Markov decision process, MDP)五元组中的状态空间 S,动作空间 A 和奖励函数 $R(s, a)$。状态空间是控制策略的输入,即机器人通过传感器观测到的数据。在本研究中强化学习环境仅使用机器人的本体状态,参考四足机器人模仿生物步态的强化学习框架[22-23]并进行简化,将状态定义为 $\boldsymbol{S}_t = (\boldsymbol{q}_{t-2:t}, \boldsymbol{a}_{t-2:t}, \varphi)$,其中 $\boldsymbol{q}_{t-2:t}$ 是过去 N 步中机器人基座姿态四元数,$\boldsymbol{a}_{t-2:t}$ 是过去 N 步中机器人 32 个关节的旋转角度,φ 是当前步态在参考步态中的时间戳(假设参考步态有 n 帧,φ 代表当前步态对应 0 到 $n-1$ 中的某一帧)。通过预实验发现增加历史帧数会显著提高状态维度,导致训练不稳定,且对性能提升有限。同时步态相位变量 φ 已对周期信息进行了编码,使策略能够隐式推断当前运动阶段,从而降低对更长历史序列的依赖,确保生成的步态符合周期性规律。因此,在处理机器人本体状态的历史数据时 N 选取 3。

动作空间是控制策略的输出,代表机器人驱动关节的控制命令。为实现仿蟹机器人的端到端 DRL 方法,动作空间定义 a_t,表示 32 个关节的控制变量。在仿真中,强化学习策略并不直接输出关节力矩,而是输出关节目标位置指令。该指令通过 PyBullet 提供的内置位置控制器映射为关节力矩。

奖励函数设计是实现仿生步态的重要环节。本研究奖励函数通过对与螃蟹运动数据接近的行为赋予高回报来促进步态模仿螃蟹的效果。奖励函数设计参考足式机器人模仿生物步态的强化学习研究[22,24-26],并根据螃蟹机器人的特点和任务需求进行相应调整:由 5 个部分组成,分别是关节角度奖励 r_t^p、关节速度奖励 r_t^v、基座位姿奖励 r_t^{rp}、单一方向前进奖励 r_t^f 和基座高度奖励 r_t^h。最终的奖励值 r_t 为:

$$r_t = w^p r_t^p + w^v r_t^v + w^e r_t^e + w^f r_t^f + w^h r_t^h \tag{5}$$

其中,加权系数 $w^p = 1$、$w^v = 0.1$、$w^e = 0.1$、$w^f = 0.05$、$w^h = 0.07$。前 3 项奖励为鼓励模仿参考运动的奖励函数,采用负指数函数形式,进行了量纲归一化处理;后两项奖励分别针对侧行任务和八足高自由度在训练中突破局部最优策略而设计,参数通过预实验测试进行了调节,可在上述设计的控制器中表现出稳定鲁棒性。r_t^p 鼓励机器人最小化参考运动指定的关节旋转量与机器人关节旋转量之间的差值,即:

$$r_t^p = \exp\left[-2.1\sum \| \hat{q}_t - q_t \|^2\right] \tag{6}$$

式中:$\hat{q}_t$ 表示 t 时刻关节在参考运动的一维局部旋转,q_t 表示机器人的关节实际的一维局部旋转。

类似地,根据关节速度计算 r_t^v:

$$r_t^v = \exp\left[-0.1\sum \| \hat{q}_{tv} - q_{tv} \|^2\right] \tag{7}$$

式中:$\hat{q}_{tv}$ 和 q_{tv} 分别为参考运动和机器人关节的角速度。

除关节模仿奖励外,设置基座位姿奖励 r_t^{rp} 激励机器人跟踪参考运动中的基座位姿,基座位置选取机器人躯干的底板面中心,r_t^{rp} 为:

$$r_t^{rp} = 0.5\exp[-10\sum \| \hat{x}_t^{root} - x_t^{root} \|^2)] + 0.5\exp[-10\sum \| \hat{q}_t^{root} - q_t^{root} \|^2] \quad (8)$$

式中:$\hat{x}_t^{root}$ 和 x_t^{root} 表示基座在参考运动和实际运动中的全局位置;$\hat{q}_t^{root}$ 和 q_t^{root} 则表示基座在参考运动和实际运动中的姿态四元数。

单一方向前进奖励 r_t^f 由激励机器人沿着轴向前行的部分和惩罚机器人偏离路线的部分组成:

$$r_t^f = \log[0.5(v_t^{root} + x_t^{root} - x_{t-1}^{root}) + 1] - |v_t^{root'}| \quad (9)$$

式中:v_t^{root} 为沿轴速度,x_t^{root} 为轴向位置,$v_t^{root'}$ 为垂直于目标轴的速度。基座高度奖励 r_t^h 是为了缩短策略前期机器人通过基座贴地面移动来获取稳定性回报而设置的鼓励基座抬高的奖励函数,即:

$$r_t^h = h_t^{root} - 0.07 \quad (10)$$

式中:h_t^{root} 为机器人运动过程中基座的高度值,0.07 m 是根据仿真实验得到的高度阈值。

3　行为克隆与分阶段强化学习

由于螃蟹的体型和关节较小,且难以引导和训练,四足机器人研究[23,26]中直接采集猫或狗在复杂地形上的运动步态数据的方法难以实现。因此本研究提出先在平地上训练稳定模型,采集步态数据后通过行为克隆技术初始化策略网络,再针对不同地形设计新的奖励函数,实现爬坡、爬楼梯等复杂地形的模型训练。

首先,从平地训练的稳定模型中收集高质量状态-动作对(s_i, a_i),存储为字典格式,作为专家演示数据供后续行为克隆训练使用,其中 s_i 表示状态,a_i 是专家在该状态下选择的动作。

其次,在行为克隆(behavior cloning, BC)阶段,使用(s_i, a_i)来训练策略网络。训练目标是让策略网络输出的动作尽可能接近专家动作,从而模仿专家行为。由于动作空间是连续的,训练时采用均方误差衡量预测动作与专家动作间的差距。损失函数为:

$$L_{BC} = \frac{1}{N}\sum_{i=1}^{N} \| a_i^{expert} - a_i^{\pi_\theta}(s_i) \|^2 \quad (11)$$

式中:N 是样本数;a_i^{expert} 是专家在状态 s_i 下的动作,$a_i^{\pi_\theta}(s_i)$ 是策略 π_θ 在状态 s_i 下生成的动作。该过程目标是求解策略网络参数 θ^*,即:

$$\theta^* = \arg_\theta \min E_{(s,a)B}[L_{BC}(\pi_\theta(s), a)] \quad (12)$$

在训练中,批量化处理状态和动作数据,并通过多层感知机(multi-layer perceptron, MLP)生成动作预测。最终,经过行为克隆训练的策略网络可生成与专家相似的步态,确保机器人实现稳定的运动。训练完成后,行为克隆模型的参数 θ^* 被保存为网络文件。

在后续的强化学习阶段,PPO 算法的策略网络会通过策略迁移来加速学习过程,并提高初始性能。策略迁移的核心是加载已有神经网络的权重。在 Actor-Critic 框架中,模型由多个部分组成:特征提取器用于处理输入的状态空间数据,策略网络负责选择动作并优化动作的概率分布,体现了行为模式,价值网络评估状态-动作对的价值,与奖励函数密切相关。

模仿生物步态的奖励函数记作 R_A,即第 2 章中的奖励函数,为新地形设计的奖励函数记作 R_B。实验目标是利用 R_A 训练的策略在 R_B 的引导下继续学习。由于价值网络与奖励函数密切相关,因此在进行策略迁移时,需要避免对与奖励函数密切相关的价值网络及其输出层进行迁移。具体的网络结构如图 5 所示。

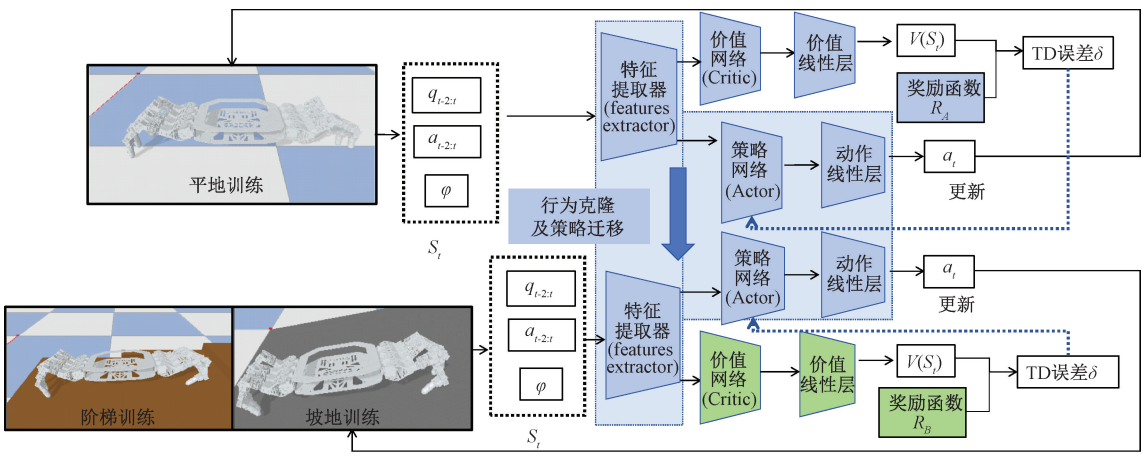


图 5　行为克隆及策略迁移逻辑

Fig. 5　Logic diagram of behavior cloning and strategy transfer

为避免机器人在复杂地形直接训练过程中破坏已形成的稳定步态结构,导致步态模式发生非预期的畸形变化,本文采用基于课程学习思想的分阶段强化学习训练策略,使策略在继承平地稳定步态的基础上,逐步适应新

地形条件。课程学习从简单的样本开始,逐渐学习复杂样本,实现更快收敛并提高泛化能力。本研究通过调整坡度、台阶高度等环境参数逐步增加难度。在每个阶段,训练模型一定步数后保存模型,并将其作为下一阶段训练的起点,实现从简单环境到复杂环境的平滑过渡。斜坡地形的奖励函数 R_B 设置如表 1 所示。其中,r_{Fp} 通过设定目标速度来控制机器人沿坡方向的速度 $\hat{V}_t$,在引导其沿坡爬行的同时将速度控制在合理范围内,防止其通过跳跃或振动实现前进。r_{Sp} 中 q_{pitch}、q_{yaw}、q_{roll} 表示当前机器人基座姿态角,r_{Dp} 中 V_{lateral} 指垂直于行走方向的速度,这两项鼓励机器人保持基座稳定和行走路线不偏移目标轴。r_{Ap} 可减小不必要的动作能耗,配合前进奖励抑制关节高频振动,其中 V_{joint} 是各关节速度。r_{Jp} 则延续模仿步态的思路,保证模型步态的周期性。

表 1 斜坡地形分阶段强化学习奖励函数表

Table 1 Table of phased reinforcement learning reward functions for the slope terrain

奖励函数类型	奖励函数公式	权重
沿坡前进奖励 r_{Fp}	$\exp(-5\mid\hat{V}_t-0.3\mid^2)$	0.70
姿态稳定奖励 r_{Sp}	$-(\mid q_{\text{pitch}}\mid+\mid q_{\text{yaw}}\mid+\mid q_{\text{roll}}\mid)$	0.50
动作平滑奖励 r_{Ap}	$-(\sum V_{\text{joint}}^2)^{0.5}$	0.06
偏移惩罚 r_{Dp}	$-V_{\text{lateral}}$	1.00
关节模仿奖励 r_{Jp}	$\exp(-1.5\sum\|\hat{q}_t-q_t\|^2)$	0.70

类似于斜坡地形奖励函数的设计,阶梯地形奖励函数的设计如表 2 所示。

表 2 阶梯地形分阶段强化学习奖励函数表

Table 2 Table of the phased reinforcement learning reward functions for the steps terrain

奖励函数类型	奖励函数公式	权重
阶梯前进奖励 r_{Ft}	$\begin{cases}\exp(-5\mid\hat{V}_t-0.3\mid^2), & \hat{V}_t\geqslant 0\\ 4\hat{V}_t, & \hat{V}_t<0\end{cases}$	0.70
姿态稳定奖励 r_{St}	$-(\mid q_{\text{pitch}}\mid+2\mid q_{\text{yaw}}\mid+\mid q_{\text{roll}}\mid)$	0.50
动作平滑奖励 r_{At}	$-(\sum V_{\text{joint}}^2)^{0.5}$	0.06
偏移惩罚 r_{Dt}	$-V_{\text{lateral}}$	1.00
基座高度奖励 r_{Bt}	h_t	1.00
关节模仿奖励 r_{Jt}	$\exp(-1.0\sum\|\hat{q}_t-q_t\|^2)$	0.70

阶梯地形下的奖励函数 R_B 主要做了 4 处优化:1)由于台阶易发生刚性碰撞导致倒退,为防止通过高速的前进后退获得奖励,r_{Ft} 中增加倒退惩罚项;2)加重 r_{St} 中偏航角的比重,防止机器人侧翻下楼梯;3)增加 r_{Bt},鼓励运动策略抬高基座和跨越更高台阶;4)通过改变指数系数适当降低 r_{Jt} 的精度比重,鼓励策略在步幅上更新出激进的动作。

4 仿真实验研究

为验证本文机器人及深度强化学习控制方法设计的正确性,本文采用基于 PyBullet 仿真引擎,并建立 Gym 强化学习环境,进行模型训练和仿真实验研究。

4.1 实验设计与数据采集

在搭建的强化学习环境中,Gym 接口连接 PyBullet,使强化学习算法与仿真环境交互,实现状态更新、动作选择和奖励计算等功能。在本文所有实验中,控制器参数均采用 PyBullet 的位置控制器,该控制器在物理引擎内部通过约束求解实现关节电机控制,而非显式解析 PD 公式,只提供了位置增益与速度增益两个参数,其中位置增益 K_p 设为 0.1,速度增益 K_v 设为 1.0。此外,在预实验中也对 PPO 算法的核心参数裁剪系数 ε 进行了多组训练,结果表明在高维动作空间下,较大的 ε 虽可加快初期收敛,但易导致训练不稳定,较小的 ε 降低收敛速率;而 $\varepsilon=0.2$ 能在收敛速度与训练稳定性之间取得平衡。上述参数在所有训练和测试实验中保持不变,未针对不同地形或任务进行单独调节,以确保实验结果的可重复性。

为加快训练过程,本研究采用并行环境训练策略,使用 6 个独立的仿真环境同时进行独立训练。并行化通过 Gym 的 SubprocVecEnv 实现,每个进程在独立 CPU 上运行,充分利用硬件资源缩短训练时间。在训练过程中鉴于监控模型性能需求,额外建立了一个独立的回调环境用于可视化和数据记录。训练指标通过 TensorBoard 实时展示,其中奖励曲线是主要关注的指标,用于反映策略在训练过程中的收敛性与稳定性。

4.2 学习率调整实验

学习率是强化学习训练中的关键超参数,其大小直接决定策略梯度更新幅度。在 PPO 算法中,学习率与裁剪机制共同作用,对训练效率和收敛稳定性产生重要影响。特别是在仿蟹机器人这种高维连续动作空间任务中:较高的学习率虽然可以缩短训练时间,但过高会导致策略更新波动较大,奖励曲线震荡甚至崩溃;较低的学习率较为稳定,有助于逐步提升奖励,但过低则会导致训练停滞,周期延长。表 3 列出了在不同学习率下的实验模型参数。为优化训练过程,实验采用了 4 种学习率调整策略:固定学习率、分段学习率、线性下降学习率和余弦退火学习率。

1)固定学习率及分段学习率

在使用固定学习率时,结果表明高学习率会导致策

表 3　学习率调整实验模型参数

Table 3　The model parameters utilized during the learning rate adjustment experiment

参数	数值
学习率范围	$1\times10^{-5}\sim1\times10^{-3}$
批次大小	128
训练总步数	1×10^{7}
折扣因子	0.97
衰减因子	0.95
裁剪系数	0.2
神经网络结构	[248,64,64,64]
激活函数	tanh

略不稳定,低学习率使训练缓慢,都不利于模型训练。适当的学习率可以在训练初期稳定且快速提升奖励,但长期训练后可能陷入局部最优,导致奖励振荡,奖励曲线如图 6 所示。因此学习率需随训练步数进行相应调整。

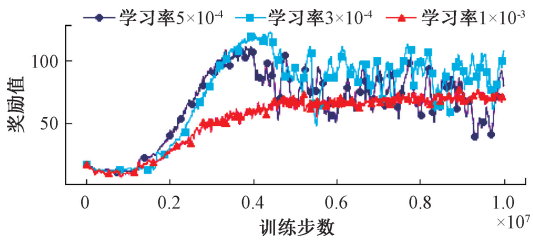


图 6　单一学习率训练的奖励值

Fig. 6　The reward value of single learning rate training

分段学习率方法设置了初始学习率为固定学习率下表现最佳的值(3×10^{-4}),终止学习率为较低的值(5×10^{-5})。实验分为 3 组,每组在不同训练节点调整学习率,实验结果如图 7 所示,结果显示当奖励曲线出现波动时,逐步降低学习率有助于稳定训练并引导策略探索更高奖励。

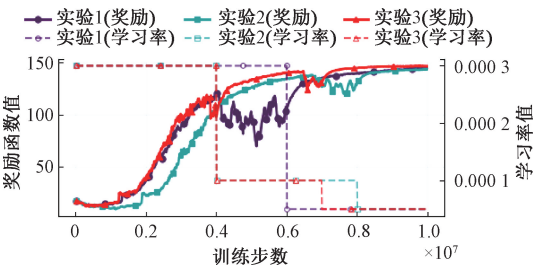


图 7　分段学习率训练的奖励值及对应学习率

Fig. 7　The reward value of segmented learning rate training process and the corresponding learning rate

2)线性下降学习率及余弦退火学习率

线性下降学习率方法指学习率从初始值线性递减至最终值,适合基础任务,能够在前期快速收敛,后期稳定。余弦退火学习率是通过余弦函数调整学习率,适合复杂任务,能够帮助模型跳出局部最优,即:

$$\eta_t=\eta_{\min}+\frac{1}{2}(\eta_{\max}-\eta_{\min})\left[1+\cos\left(\frac{t\pi}{T}\right)\right] \tag{13}$$

式中:η_t 为当前学习率;T 为训练周期。以上两种方法实现了对学习率的连续调节,且仅需起始和终止学习率数据。实验给两种方法分别设定两组参数,如图 8 所示,奖励函数曲线平滑且少有振荡阶段,在 1×10^{7} 步的训练中稳步得到接近 150 的奖励值,达到较理想效果。

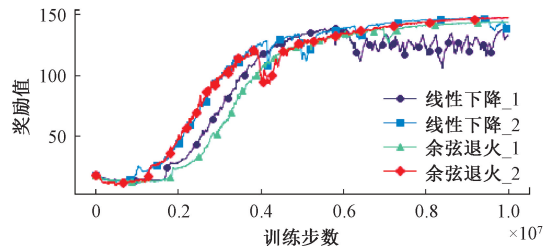


图 8　线性下降及余弦退火学习率训练的奖励值

Fig. 8　The reward values of linear descent and cosine annealing learning rate training process

为从物理性能角度分析不同学习率策略的影响,引入 5 个步态周期内机器人质心高度的标准差 σ_h 作为补充指标。实验结果显示,无论采用何种学习率调整策略,奖励函数收敛至近似极限值的各模型在步态稳定性指标上表现一致:$\sigma_h\approx0.004$ m,说明在收敛至最优奖励水平后,不同调整策略训练模型获得的步态在稳定性上具有相近性能。相比之下,采用单一学习率且奖励远低于 150 的模型,其稳定性指标 σ_h 范围是 0.01~0.024 m,表现出较大的步态震荡。由此可见,步态稳定性与最终收敛奖励值呈正相关,学习率策略的主要作用在于加速训练收敛及提升训练过程稳定性,而非改变最终收敛步态的物理性能上限。

4.3　算法对比实验

为评估 PPO 算法的适配性,将其与优势演员-评论家算法(advantage actor-critic, A2C)、软演员-评论家算法(soft actor-critic, SAC)和深度确定性策略梯度(deep deterministic policy gradient, DDPG)3 种支持并行训练、适用于连续动作和状态空间的算法进行对比实验。如表 4 所示,对比算法均采用 Stable-Baselines 3 算法库中的官方实现及其默认超参数设置,未进行额外调参,只同时增加了基于默认参数的线性下降学习率算法以加速训练,保证对比的公平性与可复现性。

表 4 算法对比实验超参数

Table 4 Hyperparameters of the algorithm comparison experiment

参数	A2C	SAC	DDPG
策略类型	MlpPolicy	MlpPolicy	MlpPolicy
学习率范围	$1\times10^{-4}\sim7\times10^{-4}$	$5\times10^{-5}\sim3\times10^{-4}$	$3\times10^{-4}\sim1\times10^{-3}$
折扣因子	0.99	0.99	0.99
批次大小	–	256	256
熵系数	0.01	自动调整	–
采样步长	5	–	–
经验回放池大小	–	1×10^{6}	1×10^{6}
目标网络更新系数	–	0.005	0.005

对比实验结果如图 9 所示,PPO 在奖励函数收敛性方面表现优异,最终奖励值稳定在 150 左右。尽管 DDPG 在 1 000 万步的训练中也能获得较高的奖励,但训练过程仍未收敛,最终奖励值约为 80。A2C 和 SAC 的奖励值增长较慢,几乎没有明显的提升趋势。

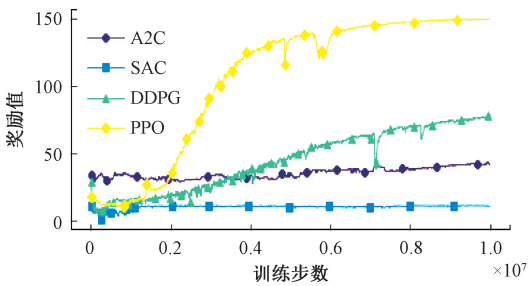


图 9 多种强化学习算法的对比实验

Fig. 9 Comparative experiments of multiple reinforcement learning algorithms

4.4 域随机化训练及实验

域随机化(domain randomization, DR)是实现仿真-现实迁移的关键技术,通过在仿真中引入多样化的环境变化,迫使策略学习适应广泛的参数范围,从而提升机器人在真实环境中的适应能力。对于仿蟹机器人的复杂运动控制,域随机化通过动态扰动仿真环境关键参数,有效模拟真实世界不确定性。训练过程中,动力学参数和地形参数在每一轮训练开始前随机初始化,如表 5 所示。地形随机化通过设置高度场的高度取值范围来实现。通过这种方式,模型能够在训练中适应多变的地形,进而学习到更为鲁棒的控制策略。

表 5 域随机化参数

Table 5 Domain randomization parameter

	参数	域随机化范围	单位
动力学参数	机器人质量	[0.7, 1.3]×真实值	kg
	横向接触摩擦系数	[0.5, 1.5]	
	滑动摩擦系数	[0.001, 0.01]	
	恢复系数	[0.0, 0.4]	
	重力系数	[−11.0, −8.0]	N/kg
	地面摩擦系数	[0.5, 1.5]	
	地面恢复系数	[0.0, 0.4]	
地形参数	地形高度	[0, 50]	mm

为了评估域随机化的效果,仿真实验对比了在应用域随机化前后机器人的运动能力。图 10 展示了 3 个模型在物理条件变化下完成沿 Y 轴负方向运动时的位移变化。结果显示,加入动力学和地形随机化训练的模型表现最优,不仅前进速度稳定且行进距离较长,达到 2 m,在 Z 轴(基座高度)的稳定性上也最优,保持在 0.125 m 左右。然而,在 X 轴偏移量上,未进行域随机化的模型表现更优,这可能是因为该模型仅在稳定环境下训练,策略单一,轨迹偏离较小,精确性更高。可以从物理与控制原理角度理解这一现象:域随机化引入扰动,使策略必须适配更多物理环境,从而增强鲁棒性,但在单一稳定环境下可能产生轨迹偏差,即策略的轨迹精确度或最优性略有下降。因此,在训练设计中,可根据任务需求调节参数随机化幅度或在训练后期减弱扰动以保证轨迹精确度,从而在两者之间取得平衡。

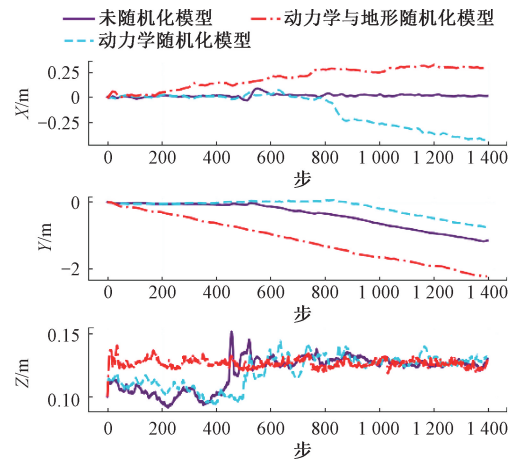


图 10 域随机化前后模型的基座三维轨迹

Fig. 10 Three-dimensional trajectories of the base for the model before and after domain randomization

图 11 展示了域随机化前后 3 种模型在地形高度随

机化场景下运动仿真视频序列图,可看出动力学和地形均随机化的模型的运动位移最大,效果最优。

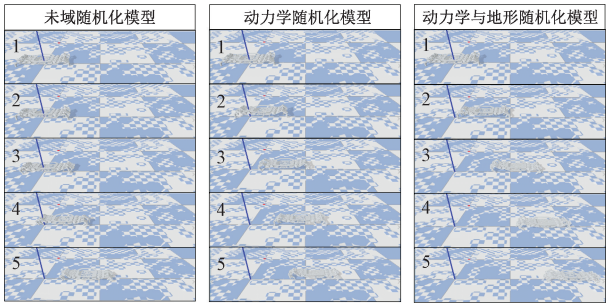


图 11　仿蟹机器人平地行走域随机化前后运动序列图

Fig. 11　Locomotion sequence before and after the domain randomization

为进一步验证域随机化对模型稳定性的影响,本研究固定了 5 组随机化参数种子,使用未进行域随机化、仅进行动力学随机化和同时进行动力学与地形随机化的 3 种模型进行对比测试。每个模型在每轮随机化测试中行走 10 步,统计它们的三维轨迹和姿态角。三维轨迹分布如图 12(a)所示,域随机化后的模型在 X 轴(运动偏移量)上有所增加,表明模型在适应环境变化时偏移较大;在 Y 轴(运动目标)上,域随机化后的模型行走更远,均值更高;在 Z 轴(高度稳定性)上,域随机化模型的异常值更少,数据分布更加集中,表现出更强的高度稳定性。欧拉角分布如图 12(b)所示,域随机化后模型的俯仰角和翻滚角取值范围更广,偏航角分布集中,这表明域随机化模型牺牲了部分姿态稳定性来保证侧行任务顺利完成。其中,A 为未域随机化,B 为动力学随机化,C 为动力学与地形随机化。

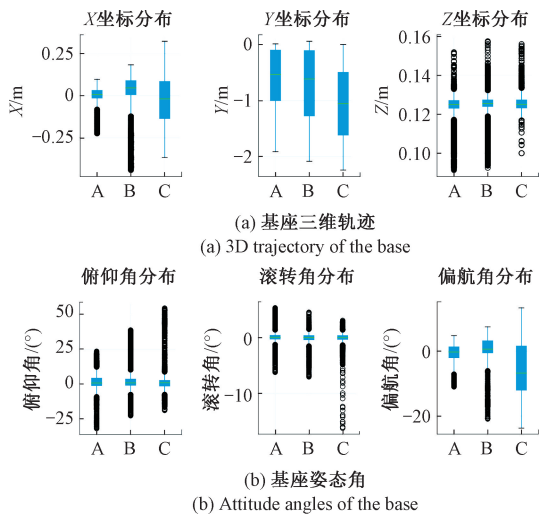


图 12　域随机化前后结果对比

Fig. 12　Comparison of the results before and after domain randomization

4.5　行为克隆与分阶段强化学习爬坡和台阶实验

为实现爬坡和楼梯的复杂地形运动,采用行为克隆和分阶段强化学习,基于已有模型进行策略迁移和持续优化,训练过程主要分为两个阶段:数据采集与行为克隆、分阶段强化学习。在第 1 阶段,针对平地环境下已训练完成的稳定模型进行专家数据采集,并据此训练行为克隆策略。为适配策略网络中的特征提取器结构,通过多层展平(Flatten)操作,将多维状态特征统一映射为一维向量作为网络输入,保证网络结构上的一致性。该过程训练参数配置如表 6 所示。

表 6　专家数据收集及行为克隆策略训练参数

Table 6　The training parameters of expert data collection and behavioral cloning strategy

参数	数值/类型
专家数据收集轮数	200
专家数据每轮步数	350
策略类型	MultiInputActorCriticPolicy
特征提取器	CombinedExtractor
策略训练批次大小	64
学习率	3×10^{-4}
损失函数类型	MSE
策略神经网络结构	[178,64,64,64]
激活函数类型	tanh

分阶段强化学习适用于任务难度逐步提升的训练场景。不同于前述步态模仿任务中单周期即可评估运动性能,斜坡与台阶地形需要更长时间才能体现机器人运动能力的差异,因此将单轮训练步数扩展为 5 个步态周期。为充分利用已有步态策略,在分阶段强化学习中对超参数进行了针对性设置,尤其是动作分布的标准差。通过将 $\log(std)$ 设为 −1,有效限制初期探索幅度,避免策略输出出现剧烈震荡。分阶段强化学习的具体训练参数如表 7 所示。

表 7　分阶段强化学习实验模型参数

Table 7　Model parameters of the phased reinforcement learning experimental

参数	数值
批次大小	128
每轮周期	5
折扣因子	0.97
衰减因子	0.95
裁剪系数	0.2
初始动作分布标准差	0.37
神经网络结构	[178,64,64,64]
激活函数	tanh

为保证复杂地形难度设置的物理合理性，本文从几何与能量角度，对地形引起的质心抬升量进行归一化分析，其量纲基准选取为机器人在平地行走时的质心高度 $H_{CoM}=0.04$ m，行走时机器人体长 $L_{body}\approx1$ m。在斜坡地形实验中，机器人质心平均抬升高度近似为：

$$\Delta h_{slope}\approx\frac{L_{body}\sin\theta}{2} \tag{14}$$

式中：θ 是斜坡角度，当坡度逐渐增大时，该抬升量达到或成倍超过 H_{CoM} 的量级，将显著增加系统的重力势能变化和姿态稳定控制难度。斜坡角度依次设置为 0°、5°、10°、15°，对应 4 个难度阶段，质心平均高度约为 1～3 倍 H_{CoM}。每个阶段训练 2.0×10^6 步，阶段结束后保存模型，并在此基础上继续训练下一阶段，使策略逐步适应坡地行走任务。为验证方法有效性，设计了 4 种对比模型：模型 A 指仅在平地训练的基础模型；模型 B 指在斜坡地形重新开始训练的模型；模型 C 指经行为克隆后直接在 15°斜坡训练的模型；模型 D 指经行为克隆并采用分阶段强化学习训练的模型。其中，模型 B、C 和 D 均在相同超参数条件下训练 8.0×10^6 步，并在 15°斜坡条件下统计平均奖励值及实际前进距离。

实验结果如图 13 所示。模型 A 虽保留周期性的生物步态，但不具备爬坡能力；模型 B 未能形成稳定步态，无法爬坡行走；模型 C 在保持生物步态的基础上具备一定爬坡能力，但运动速度较慢；模型 D 在爬坡距离、爬坡速度及坡地奖励值等指标上均表现最优。

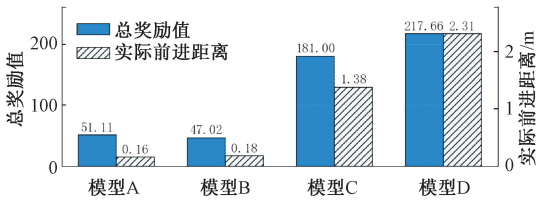


图 13 斜坡地形下不同模型奖励值和前进距离对比

Fig. 13 Comparison chart of reward values and forward distances for different models in sloping terrain

图 14 展示了 4 种模型在斜坡地形下完成 5 个步态周期时的三维轨迹对比。机器人沿 Y 轴负方向行走，其中 X 轴位移的周期性波动表明除模型 B 外，其余模型均成功学习到稳定的周期性步态；Y 轴和 Z 轴的变化则反映了爬坡能力，模型 D 明显优于模型 C，而模型 A 和 B 未能实现有效爬坡。该结果与奖励函数分析结论一致。表明采用行为克隆和分阶段强化学习从平地步态成功迁移实现了爬坡运动功能。

在阶梯地形实验中，台阶高度设置为 0、0.04、0.06 和 0.08 m，对应 4 个难度阶段，约为机器人平地行走时质心高度 H_{CoM} 的 1～2 倍，对步态稳定性和足端协调提出较

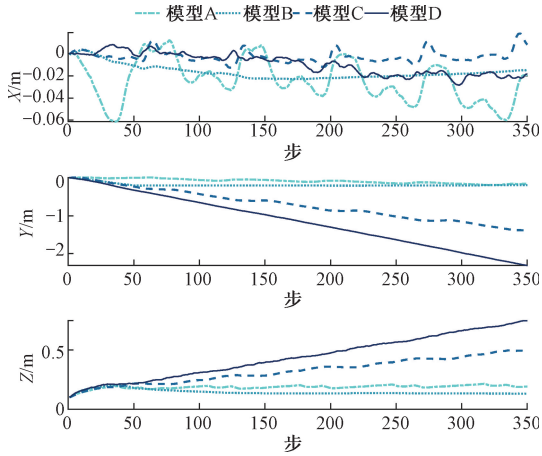


图 14 斜坡地形下不同模型三维轨迹对比

Fig. 14 Three-dimensional trajectory comparison chart of different models in sloping terrain

高要求，在几何尺度与能量扰动意义上属于高难度工况。每阶段训练 2.5×10^6 步，其余训练方法与斜坡实验相同。为验证方法效果，设计了 4 种对比模型：模型 A 指仅在平地训练的基础模型；模型 B 指在阶梯地形从零开始训练的模型；模型 C 指行为克隆后直接在台阶高度 0.08 m 下训练的模型；模型 D 指行为克隆后采用分阶段强化学习训练的模型。模型 B、C、D 在相同超参数条件下训练 1.0×10^7 步，并在台阶高度 0.08 m 下采集奖励值和实际前进距离，结果如图 15 所示。结果显示，总奖励值和前进距离呈逐步上升趋势，模型 D 表现最佳。

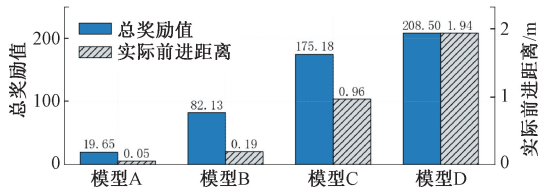


图 15 阶梯地形下不同模型奖励值和前进距离对比

Fig. 15 Comparison chart of reward values and forward distances for different models in stepped terrain

图 16 展示了 4 种模型在阶梯地形下完成 5 个步态周期的三维轨迹。经过地形训练的模型 B、C、D 的轨迹周期性明显弱化，尤其模型 D 的轨迹显示 3 轮周期而非五步周期，反映其策略根据台阶高度调整步态。Y 轴位移变化表明，模型 C 和 D 成功获得爬阶能力，其中模型 D 优于 C，而模型 A 和 B 未能获得爬阶能力。

为全面评估 4 种模型在不同难度地形下的适应性和行走能力，实验统计了斜坡地形下 4 种模型在 0°、5°、10°、15°坡度获得的平均奖励和阶梯地形下 4 种模型在 0、0.04、0.06 和 0.08 m 台阶高度获得的平均奖励，数据以热力图形式呈现，如图 17 所示。结果显示只有模型 D

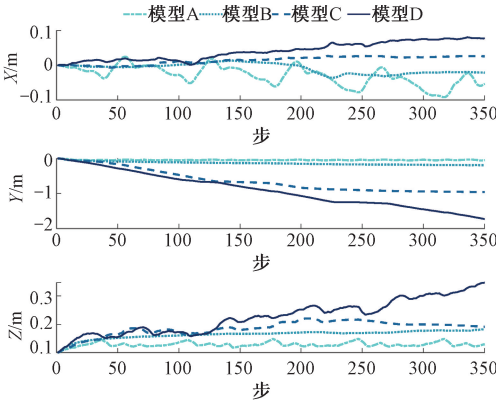
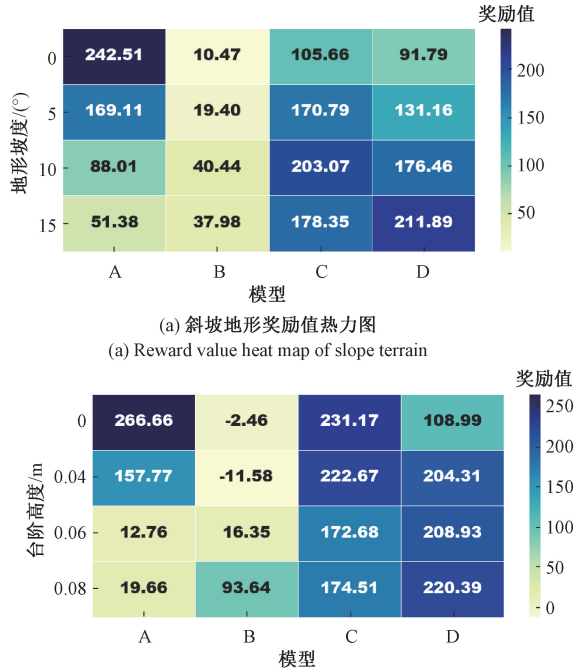


图 16　阶梯地形下不同模型三维轨迹对比

Fig. 16　Three-dimensional trajectory comparison chart of different models in stepped terrain

的奖励值随难度增大有明显提升的趋势,证明行为克隆+分阶段强化学习的训练思路可以引导模型逐步适应高难度地形的策略。

图 18 为仿蟹机器人爬坡和爬阶梯的时序步态图。爬坡任务中,模型 A 与 B 没有爬坡能力;模型 D 爬坡能力强于模型 C。爬楼梯任务中,模型 A 和 B 均无法跨越第 1 级台阶;模型 B 通过伸展一侧腿和抬高机体获得有



(a) 斜坡地形奖励值热力图

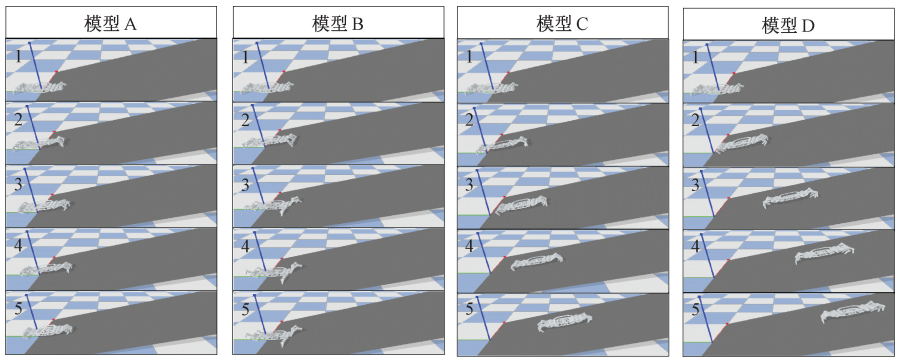
(a) Reward value heat map of slope terrain

(b) 阶梯地形奖励值热力图

(b) The reward value heat map of staircase-shaped terrain

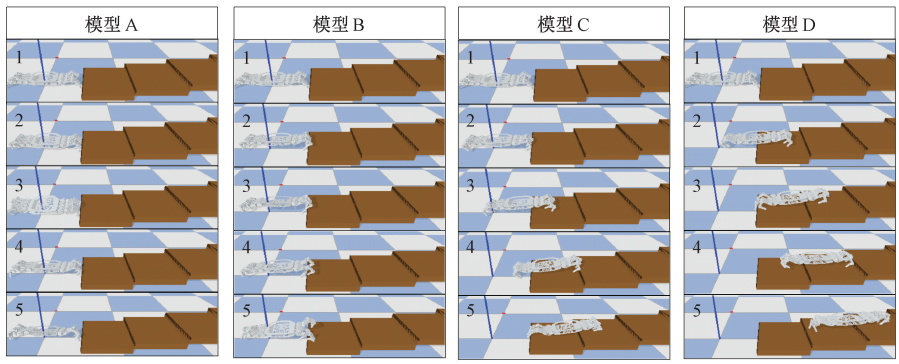
图 17　不同难度复杂地形下模型奖励值热力图

Fig. 17　Heat maps of model reward values under complex terrains of different difficulties



(a) 多模型爬坡实验序列图

(a) The sequence of multi-model climbing experiment



(b) 多模型爬阶梯实验序列图

(b) The sequence of multi-model stepped experiment

图 18　仿蟹机器人爬坡和爬阶梯步态仿真实验序列图

Fig. 18　Sequence diagram of gait simulation experiments for the crab-like robot when climbing slopes and stairs

限奖励;模型 C 在跨过前两级台阶后左侧腿抬升不足导致停滞;模型 D 则在 5 个周期步内连续爬阶,表现出最佳的爬台阶能力。

综上,仿真实验表明行为克隆能够有效迁移平地训练模型的仿生步态策略,为复杂地形训练提供合理初始化;分阶段强化学习能够使模型在平地步态上逐步适应复杂地形,实现坡地和阶梯地形上的稳定运动。

5 样机实验研究

在仿真基础上本文开展了样机试验。受限于样机与仿真模型的材料、驱动等的差异,本文仅开展了平地域随机化实验,爬坡和爬楼梯实验将在未来工作中开展。实验前首先对关节舵机进行了参数辨识,通过阶跃响应拟合得到位置增益 $K_p=0.5$,速度增益 $K_v=1.0$,在此参数下重新训练对应模型,并根据实际舵机的力矩上限设置动作空间。实验采用的路面类型及其摩擦系数如表 8 所示,实验场景如图 19 所示。

表 8 样机实验路面类型和测得的摩擦系数

Table 8 Road types and measured surface friction coefficients for prototype experiment

路面类型	摩擦系数
灰地毯铺设路面	0.86
光滑水泥地路面	0.57
白地毯正面铺设路面	0.65
白地毯反面铺设路面	0.82

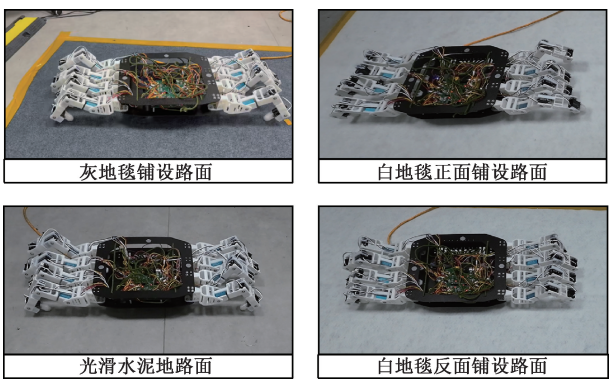


图 19 仿蟹机器人样机实验路面

Fig. 19 Experimental road map of the crab-like robot

实验过程中,模型在 PC 端运行并通过网络接口发送运动指令给机器人,机器人回传模型相关数据作为下一次模型计算的输入参数。采用红外动作捕捉设备完成三维位置的采集,并与仿真数据进行对比。机器人仿真与水泥地实验轨迹对比图如图 20 所示。

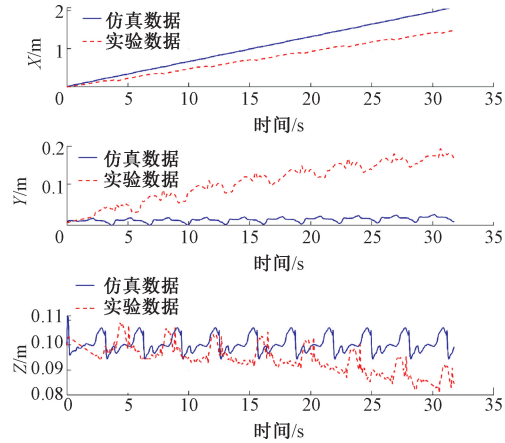


图 20 仿蟹机器人仿真与实验轨迹对比

Fig. 20 Comparison chart of the simulation and experimental trajectory for the crab-like robot

实验结果显示,样机实际运动轨迹与仿真模型基本一致,机器人沿着 X 方向前进,运动位移为 1.48 m,误差为 0.6 m,Y 轴偏移量在 0.2 m 以内,Z 轴偏移量在±0.02 m 以内,基本还原了仿真中稳定行走的效果。此外,为评估机器人姿态稳定性,计算了姿态角误差模,即同一时间下仿真与实验数据欧拉角插值平方和的平方根,该指标全程保持在[3°,15°]范围,表明实验中强化学习模型对姿态的控制效果良好,在姿态层面与仿真结果具有良好的一致性。

4 种路面运动的前进速度、横向偏移和竖直波动最大值如表 9 所示。实验数据表明,运动过程中平均速度与地面摩擦系数基本成正比,垂直震荡与横向摆动指标与地面类型未表现明显相关性,但都保持较低水平,说明强化学习模型在多种路面行走实验中都表现出稳定性与可靠性。

表 9 多路面样机实验数据统计表

Table 9 Statistical table of multiple road sample machine test data

路面类型	平均速度/(m·s^{-1})	横向摆动/m	竖直波动最大值/m
灰地毯铺设路面	0.087	0.035	0.036
光滑水泥地路面	0.081	0,026	0.024
白地毯正面铺设路面	0.078	0.019	0.029
白地毯反面铺设路面	0.084	0.020	0.033

实验结果与仿真结果差异原因为:1)样机存在装配误差;2)关节舵机角度开环控制,导致腿部运动控制误差;3)关节舵机执行指令的频率低于仿真,导致运动柔顺

性低和控制误差;4)机器人运动轨迹开环控制,存在轨迹误差。未来研究中将采用具有角度反馈的大力矩总线舵机,并开展机器人的运动位姿闭环控制,提升机器人的运动精度,实现爬坡和爬楼梯运动。

6 结　论

本文提出了一种基于螃蟹生物特性及步态的高效可泛化深度强化学习训练框架。设计仿蟹机器人模型、提取可用于模仿的运动步态,搭建强化学习训练环境,通过并行训练、学习率调整、奖励函数优化等方法加速模型训练,再使用域随机化、行为克隆及分阶段强化学习等方法增强模型泛化能力,扩大其应用场景,并对其进行了模型训练和仿真实验测试。结果表明,本研究提出的强化学习框架适用于八足仿蟹机器人的运动控制场景,在模型训练效率、控制稳定性及模型泛化能力上相较于未改进的强化学习方法都有明显优势,可以通过简单生物步态训练出多地形下的端到端控制策略。此外,本研究进行了多种路面情况的样机试验,实验结果与仿真数据呈现一致性,且在 4 种路面情况下都保持了良好的运动功能和姿态稳定性,初步证明了所提框架的实际可行性。未来,本研究计划为模型增加更多适应性训练,并增加环境传感器以增强其感知能力,同时改进仿蟹机器人软硬件设备,在复杂地形上进一步验证所提强化学习框架的可行性与稳定性。

参考文献

[1] RUBIO F, VALERO F, LLOPIS-ALBERT C. A review of mobile robots: Concepts, methods, theoretical framework, and applications[J]. International Journal of Advanced Robotic Systems, 2019, 16(2):1-22.

[2] ZHAO Y, CHAI X, GAO F, et al. Obstacle avoidance and motion planning scheme for a hexapod robot Octopus-III[J]. Robotics and Autonomous Systems, 2018, 103: 199-212.

[3] BOUMAN A, GINTING M F, ALATUR N, et al. Autonomous Spot: Long-range autonomous exploration of extreme environments with legged locomotion[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020: 2518-2525.

[4] BLEDT G, POWELL M J, KATZ B, et al. MIT Cheetah 3: Design and control of a robust, dynamic quadruped robot[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018: 2245-2252.

[5] HUTTER M, GEHRING C, JUD D, et al. ANYmal-a highly mobile and dynamic quadrupedal robot[C]. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2016: 38-44.

[6] QI SH H, LIN W CH, HONG Z J, et al. Perceptive autonomous stair climbing for quadrupedal robots[C]. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021: 1813-1820.

[7] DENG H, XIN G Y, ZHONG G L, et al. Gait and trajectory rolling planning and control of hexapod robots for disaster rescue applications[J]. Robotics and Autonomous Systems, 2017, 95: 13-24.

[8] CHEN X, WANG L Q, YE X F, et al. Prototype development and gait planning of biologically inspired multi-legged crablike robot[J]. Mechatronics: The Science of Intelligent Machines, 2013, 23(4): 429-444.

[9] LIU K, YANG H Y, WANG L, et al. Design and implementation of a dual-drive bionic crab robot[J]. Journal of the Brazilian Society of Mechanical Sciences and Engineering, 2024, 46(7): 387.

[10] GRZELCZYK D, AWREJCEWICZ J. Modeling and control of an eight-legged walking robot driven by different gait generators[J]. International Journal of Structural Stability and Dynamics, 2019, 19(5): 19410098.

[11] 王刚. 仿蟹机器人步态规划及复杂地貌行走方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2010.
WANG G. Research on gait planning and walking method in complex terrains of crab-like robot[D]. Harbin: Harbin Engineering University, 2010.

[12] WANG G, CHEN X, YANG SH X, et al. Subsea crab bounding gait of leg-paddle hybrid driven shoal crablike robot[J]. Mechatronics: The Science of Intelligent Machines, 2017, 48: 1-11.

[13] 周敬淞,张军,肖毅, 等. 基于仿蟹刚柔耦合机构的搜救机器人设计[J]. 仪器仪表学报,2023,44(6): 11-20.
ZHOU J S, ZHANG J, XIAO Y, et al. Design of search and rescue robot based on crab-like rigid-flexible coupling mechanism[J]. Chinese Journal of Scientific Instrument, 2023, 44(6): 11-20.

[14] 杨傲雷,陈燕玲,徐昱琳. 基于强化学习的机器人手臂

仿人运动规划方法[J]. 仪器仪表学报, 2021, 42(12):136-145.
YANG AO L, CHEN Y L, XU Y L. Human-like motion planning method for robot arm based on reinforcement learning[J]. Chinese Journal of Scientific Instrument, 2021, 42(12): 136-145.

[15] HWANGBO J, LEE J, DOSOVITSKIY A, et al. Learning agile and dynamic motor skills for legged robots[J]. Science Robotics, 2019, 4(26): aau5872.

[16] LEE J, HWANGBO J, WELLHAUSEN L, et al. Learning quadrupedal locomotion over challenging terrain[J]. Science Robotics, 2020, 5(47): abc5986.

[17] MIKI T, LEE J, HWANGBO J, et al. Learning robust perceptive locomotion for quadrupedal robots in the wild[J]. Science Robotics, 2022, 7(62): abk2822.

[18] 王宇,杜艾芸,李亚鑫. 两栖六足仿生机器人的水陆运动控制研究[J]. 仪器仪表学报,2022,43(11):274-282.
WANG Y, DU AI Y, LI Y X. Study on the terrestrial and underwater motion control of the amphibious hexapod bionic robot [J]. Chinese Journal of Scientific Instrument, 2022, 43(11): 274-282.

[19] 李逃昌,李健璋,侯利民, 等. 多智能体深度强化学习优化的机器人导纳控制[J]. 电子测量与仪器学报, 2025,39(5):134-143.
LI T CH, LI J ZH, HOU L M, et al. Robot admittance control optimized by multi-agent deep reinforcement learning[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(5): 134-143.

[20] 曾宁坤,胡朋,梁竹关, 等. 基于深度强化学习哈里斯鹰算法的路径规划[J]. 电子测量技术, 2023, 46(12):69-76.
ZENG N K, HU P, LIANG ZH G, et al. Path planning based on deep reinforcement learning Harris Hawks algorithm [J]. Electronic Measurement Technology, 2023, 46(12): 69-76.

[21] LIU L B, VAN DE PANNE M, YIN K K. Guided learning of control graphs for physics-based characters[J]. ACM Transactions on Graphics, 2016, 35(3): 1-14.

[22] PENG X B, ABBEEL P, LEVINE S, et al. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills [J]. ACM Transactions on Graphics, 2018, 37(4): 1-14.

[23] PENG X B, KANAZAWA A, MALIK J, et al. SFV: Reinforcement learning of physical skills from videos[J]. ACM Transactions on Graphics, 2018, 37(6): 1-14.

[24] XIE ZH M, CLARY P, DAO J, et al. Learning locomotion skills for Cassie: Iterative design and sim-to-real[C]. Conference on Robot Learning, 2019: 317-329.

[25] PENG X B, COUMANS E, ZHANG T N, et al. Learning agile robotic locomotion skills by imitating animals[C]. Robotics: Science and Systems, 2020:64.

[26] FENG G, ZHANG H B, LI ZH Y, et al. GenLoco: Generalized locomotion controllers for quadrupedal robots[C]. Proceedings of the 6th Conference on Robot Learning, 2023, 205: 1893-1903.

[27] 程思源. 仿蟹机器人三维步态设计与运动性能实验研究[D]. 南京: 东南大学,2025.
CHENG S Y. 3D gait design and locomotion performance experimental research on a crab-inspired robot [D]. Nanjing: Southeast University, 2025.

[28] WANG P F, SUN L N. The stability analysis for quadruped bionic robot [C]. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006: 5238-5242.

[29] 王刚, 张立勋, 王立权. 仿蟹机器人交错等相位波形步态研究[J]. 机器人, 2011, 33(2): 237-243.
WANG G, ZHANG L X, WANG L Q. On alternating equal-phase wave gait of crab-like robot [J]. Robot, 2011, 33(2): 237-243.

[30] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [J]. ArXiv preprint arXiv:1707.06347, 2017.

作者简介

王诗雯,2023 年于南京邮电大学获得学士学位,现为东南大学硕士研究生,主要研究方向为多足机器人的强化学习。
E-mail:220233649@seu.edu.cn

Wang Shiwen received her B.Sc. degree from Nanjing University of Posts and Telecommunications in 2023. She is currently a master's student at Southeast University. Her main research interest includes reinforcement learning for multi-legged robots.

张军(通信作者),2008 年于南京理工大学获得学士学位,2013 年于东南大学获得博士学位,现为东南大学教授,主要研究方向为仿生机器人、康复机器人、深空探测机器人和医疗机器人等。

E-mail:j. zhang@ seu. edu. cn

Zhang Jun (Corresponding author) received his B. Sc. degree from Nanjing University of Science and Technology in 2008, and his Ph. D. degree from Southeast University in 2013. He is currently a professor at Southeast University. His main research interests include bio-inspired robotics, rehabilitation robotics, space robotics, and medical robotics.

宋爱国,分别在 1990 年和 1993 年于南京航空航天大学获得学士学位和硕士学位,1996 年于东南大学获得博士学位,现为东南大学教授、博士生导师,主要研究方向为机器人传感与控制技术、信号处理、遥操作技术等。

E-mail:a. g. song@ seu. edu. cn

Song Aiguo received his B. Sc. and M. Sc. degrees both from Nanjing University of Aeronautics and Astronautics in 1990 and 1993, respectively, and his Ph. D. degree from Southeast University in 1996. He is currently a professor and Ph. D. advisor at Southeast University. His main research interests include robotic sensor and control, signal processing, teleoperation, etc.