引用格式：*陈哲瑄，刘付成，孙俊，等.空间机器人强泛化黏附爬行策略生成方法[J].飞控与探测，2025，8(5)：11-24.*
Citation：*CHEN Z X，LIU F C，SUN J，et al. Strong generalization adhesive climbing strategy generation method for space robot[J]. Flight Control & Detection，2025，8(5)：11-24.*

空间机器人强泛化黏附爬行策略生成方法

陈哲瑄[1,2]，刘付成[1,2*]，孙　俊[1,2]，邸昕鹏[1,2]，严余超[1,2]，姚森纯[1,2]

（1. 上海航天控制技术研究所·上海·201109；
2. 上海市空间智能控制技术重点实验室·上海·201109）

摘　要：空间黏附爬行机器人能够附着于航天器外表面，自主完成舱外巡检、作业任务，是实现航天器长期无人在轨服务的重要方式。针对航天器表面特性发生意外变化后黏附爬行机器人控制策略泛化能力不足的问题，在强化学习框架下，构建黏附力作用机制，结合足端接触力“沿用-更新”机制构造密集型奖励函数，并采用近端策略优化-裁剪（Proximal Policy Optimization-clip，PPO-clip）算法训练生成微重力环境下机器人的黏附爬行策略。结果表明，在足端接触力“沿用-更新”机制作用下，策略收敛速率增加约 14.81%；在平坦表面上，获得的爬行策略能够保持机器人的黏附稳定性，并具备抵达误差小于 0.1 m 的目标位置的能力；基于平坦表面生成的爬行策略，在高度意外变化±40 mm、坡度意外变化±18°的表面，均能够实现机器人的稳定黏附爬行。

关键词：空间爬行机器人；微重力；足端黏附；奖励设计；强化学习
中图分类号：TP242.6　　**文献标志码**：A　　**文章编号**：2096-5974(2025)05-0011-14
DOI：10.20249/j.cnki.2096-5974.2025.05.002

Strong Generalization Adhesive Climbing Strategy Generation Method for Space Robot

CHEN Zhexuan[1,2]，LIU Fucheng[1,2]，SUN Jun[1,2]，DI Xinpeng[1,2]，YAN Yuchao[1,2]，YAO Senchun[1,2]

（1. Shanghai Aerospace Control Technology Institute，Shanghai 201109；
2. Shanghai Key Laboratory of Aerospace Intelligent Control Technology，Shanghai 201109）

Abstract：The space adhesive climbing robot can be attached to the outer surface of the spacecraft and complete the external inspection and operation tasks independently，which is an important way to realize the long-term unmanned in-orbit service of the spacecraft. In order to solve the problem of insufficient generalization ability of the control strategy of the adhesive climbing robot after unexpected changes in spacecraft surface characteristics，the mechanism of adhesion force is constructed under the framework of reinforcement learning，and the intensive reward function is constructed by combining the “follow-update” mechanism of the foot contact force，and the proximal policy

基金项目：国家自然科学基金（U20B2056，62204151，12102248）
作者简介：陈哲瑄，男，硕士生。
*通信作者简介：刘付成，男，博士，研究员。

optimization-clip (PPO-clip) algorithm is used to train and generate the adhesion crawling strategy of the robot in microgravity environment. The results show that the strategy convergence rate increases by about 14.81% under the "follow-update" mechanism of foot contact force. The climbing strategy obtained can maintain the adhesion stability of the robot on a flat surface, and has the ability to reach the target position with an arrival error of less than 0.1m. On surfaces with an unpredictable height change of ±40mm and an unpredictable slope change of ±18°, the climbing strategy obtained on the flat surface can achieve stable adhesion climbing of the robot.

Keywords: space climbing robot; microgravity; foot adhesion; reward design; reinforcement learning

0 引言

随着人类对太空探索、研究需求的不断增加，以及航天器设计、制造、驱动、控制技术的飞速发展，航天器的长期在轨维护、维修等需求愈加迫切[1]。航天器在轨维护等任务目前主要依靠航天员出舱作业和空间机器人辅助航天员作业完成，大幅增加航天员自身的安全风险，且任务覆盖的范围有限[2]。因此，使用空间机器人执行无人化的在轨维护任务成为了新的发展趋势。

目前空间机器人在舱外操作的应用主要有空间机械臂[3-5]、自由飞行机器人[6-7]、空间黏附爬行机器人[8-9]等。其中，空间黏附爬行机器人具有体积小、关节自由度丰富的特点，更有利于穿越大型航天器表面的复杂环境，具备较强的大范围抵达能力，适宜执行长期的在轨维护任务。

国内外已相继开展针对黏附爬行机器人运动控制方法的研究，并取得了一定成果。韩国科学技术院[10]依据已知场景设计足端磁吸附机器人爬行步态，成功控制机器人在 90°金属壁面和天花板金属平面的稳定爬行。澳门大学[11]依据爬行表面倾斜角度和障碍物高度信息，设计了黏附爬行机器人的越障步态，在 15°和 45°倾斜表面上实现越障爬行。南京航空航天大学[12]求解机器人在相邻壁面内直角处进行过渡运动时足端黏附力的包络边界，并规划过渡步态，实现了黏附机器人在壁面内直角处的平稳过渡。

然而，一方面，大型航天器表面宏、微观特性复杂，太阳能帆板形成的结构间隙、舱段过渡区域形成的坡面、辅助扶手形成的台阶，以及各类表面材料（太阳能电池片、隔热防护盾等）形成了各种粗糙度变化；另一方面，航天器长期在轨服务过程中，其表面特性往往会由于材料老化或意外碰撞产生不可穷举的改变。现有的黏附爬行机器人运动控制方法，大多依据已知的表面特性参数规划爬行步态并进行精准跟踪，当表面特性发生意外变化后，基于旧表面特性规划的步态无法适应新表面，难以保证机器人的持续稳定爬行。

随着强化学习和神经网络的逐渐兴起，强化学习算法训练获得的机器人运动控制策略因其优秀的多场景适应能力，应用愈加广泛[13-14]。使用强化学习方法提升四足机器人爬行策略泛化性、提高四足机器人在复杂环境下的运动稳定性成为国内外四足机器人运动控制的热门研究方向之一。苏黎世联邦理工大学[15-16]基于强化学习方法，构造了师生学习框架，形成了强泛化性的四足机器人运动控制策略，成功控制 Anymal 四足机器人通过复杂自然地形。雅典国家技术大学[17]结合强化学习框架和足端轨迹规划算法，构造了四足机器人的坡面运动策略，使 Laelaps II 四足机器人在多类坡面（±10°、±15°）上实现稳定爬行。南加州大学[18]采用强化学习策略生成机器人腿足末端运动轨迹，实现了在崎岖表面的快速爬行。在国内，山东大学[19]将强化学习方法与节律运动控制器相结合，使四足机器人成功通过台阶、楼梯等地形。北京化工大学[20]提出了一种基于分布式近端策略优化（Distributed Proximal Policy Optimization，DPPO）算法的分层强化学习框架，采用分层式决策模型实现了四足机器人在多类型表面（台阶、坡面、障碍）的稳定爬行。上海交通大学[21]结合强化学习框架以及难度递增课程学习机制，训练生成了一种四足机器人强泛化爬行策略，成功控制四足机器人在多种表面（斜坡、台阶、崎岖平面等）稳定运动。

然而，现有的基于强化学习算法训练获取的四足机器人爬行策略，均基于地球重力环境下，机器人各部件均受到重力加速度作用，因此无法直接应用于空间微重力环境中，而目前国内外尚缺少针对微重力环境下，基于足端黏附力作用机制和强化学

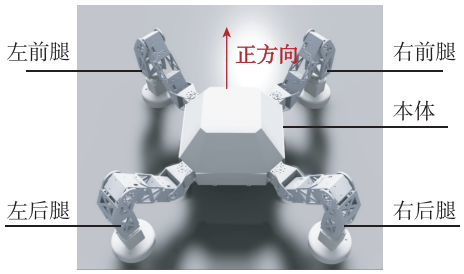
习框架的四足机器人黏附爬行策略生成方法研究。

针对以上问题，本文提出了一种基于机器人足端黏附力作用机制和强化学习框架的空间黏附爬行机器人爬行策略生成方法，主要内容包括：1）构建爬行机器人结构模型及机器人足端黏附力约束模型，并构造基于足端接触力反馈的黏附力作用机制；2）基于强化学习框架和黏附力作用机制设计密集型黏附爬行奖励函数，并构建机器人黏附爬行策略训练生成框架；3）基于 Isaac Sim 仿真平台，验证爬行策略控制机器人在微重力环境下的黏附爬行稳定性，并验证爬行策略在不同类型表面的泛化能力。

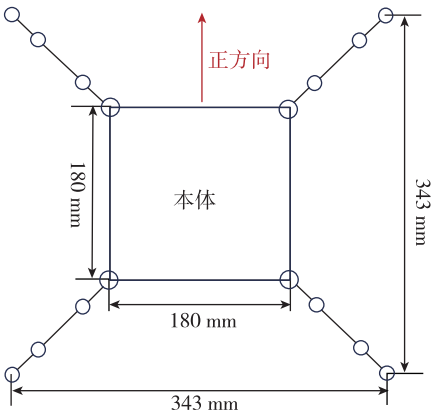
1　空间爬行机器人结构及足端黏附机制建模

1.1　机器人结构建模

空间爬行机器人由机器人本体及分布于本体几何顶点的腿足结构组成，腿足结构和本体呈昆虫形态分布，其整体结构如图 1 所示。



（a）空间爬行机器人整体结构模型

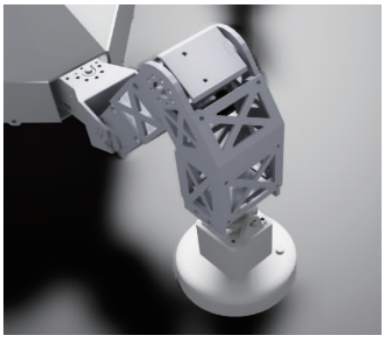


（b）空间爬行机器人整体结构简图

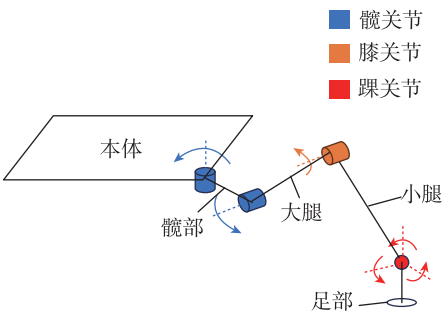
图 1　空间爬行机器人整体结构模型及其简图

Fig. 1　Overall structure model of space climbing robot and its schematic diagram

空间爬行机器人单腿由 4 个刚性构件及 6 个关节自由度组成，刚性构件包括髋部、大腿、小腿和足部，6 个关节自由度则由 2 个髋关节自由度、1 个膝关节自由度以及 3 个踝关节自由度组成，单腿刚性构件和关节分布如图 2 所示。机器人运动过程中，通过调节各关节电机输出转矩，即可驱动各腿部刚性构件运动，最终支撑机器人本体进行平稳移动。表 1 详细列举了机器人各部件的尺寸信息，包括各主要腿部部件的长度以及足部半径。



（a）空间爬行机器人单腿结构模型



（b）空间爬行机器人单腿结构简图

图 2　空间爬行机器人单腿结构模型及其简图

Fig. 2　Single-leg structure model of space climbing robot and its schematic diagram

表 1　空间爬行机器人单腿各部位尺寸

Tab. 1　Dimensions of each part of the space climbing robot's single leg

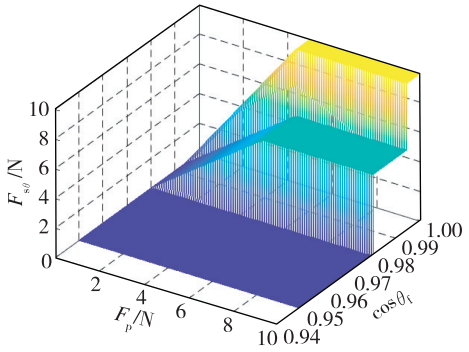
部位	尺寸/mm
髋部	40.42
大腿	111.00
小腿	104.57
足部	20.00

爬行机器人 4 条腿足部位结构相同，材料相

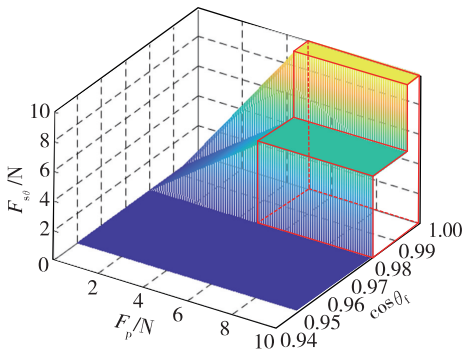
同，具备相同的运动能力和关节自由度；机器人整体具备 24 个关节自由度，因此整体运动灵活性较强。

1.2 机器人足端黏附机制建模

空间爬行机器人足端黏附材料基于壁虎足底的范德华力黏附作用机理研制而成，因此足端黏附作用力受足端施加于爬行表面的预压力以及接触角度的影响[22]。足端黏附作用力与预压力、接触角度的数学关系具有极强的非线性，为便于仿真且满足精度要求，本文仅采用黏附力与预压力、接触角度的准静态模型，其对应关系如图 3 所示。



(a) 足端黏附作用力准静态模型



(b) 足端黏附作用力准静态简化模型

图 3 黏附作用力与预压力和接触角度对应关系的准静态模型

Fig. 3 Quasi-static models of adhesion forces corresponding to pre-pressure and contact angle

图中，$F_{s\theta}$表示考虑接触角度时黏附材料产生的宏观黏附作用力大小，F_p 表示足末端对运动所在表面施加的预压力大小，$\cos\theta_f$ 为足端与表面接触角度的余弦值。设 F_s 表示不考虑接触角度时黏附材料产生的宏观黏附作用力大小，F_{st}为黏附作用力的最大值，F_{pt}为黏附力达到最大值时所施加的预压力，α，β，k_1，k_2 为标量。则足端黏附作用力的准静态模型可量化为

$$F_{st}=\alpha F_{pt} \tag{1}$$

$$F_s=\begin{cases}\alpha F_p & F_p\leqslant F_{pt}\\ F_{st} & F_p>F_{pt}\end{cases} \tag{2}$$

$$F_{s\theta}=\begin{cases}F_s & \cos\theta_f\geqslant k_1\\ \beta F_s & k_1>\cos\theta_f\geqslant k_2\\ 0 & \cos\theta_f<k_2\end{cases} \tag{3}$$

当机器人足端施加的预压力处于 0 至 F_{pt}之间，足端与环境表面产生的黏附作用力大小与施加的预压力成正比关系。当机器人足端施加的预压力大于 F_{pt}时，黏附作用力大小停留在最大值处，并保持不变，黏附作用力进入饱和状态。综合考虑计算复杂度及仿真精度，本文进一步简化黏附力作用模型，如图 3（b）红色曲线包络区域所示。简化后黏附力作用模型如下

$$F_{s\theta}=\begin{cases}F_{st} & \cos\theta_f\geqslant k_1,F_p>F_{pt}\\ \beta F_{st} & k_1>\cos\theta_f\geqslant k_2,F_p>F_{pt}\\ 0 & \text{其他}\end{cases} \tag{4}$$

黏附材料特性参数如下

$$F_{st}=10,F_{pt}=5,\beta=0.55,k_1=0.995,k_2=0.98$$

基于上述足端黏附力作用简化模型，本文提出了一种基于足端接触力反馈(Contact Force Feedback)的黏附作用机制。如图 4 所示，本文基于作用力与反作用原理，将足端对爬行表面的预压力转换为表面对足端的接触力 F_{cf}，将足端接触角度转换为足底接触面法向量与接触力矢量的夹角，并据此计算黏附作用力，黏附作用力方向反向于接触力矢量，并最终作用于足部刚体。

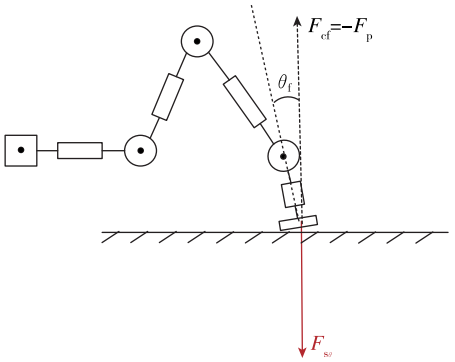


图 4 基于足端接触力反馈的黏附作用机制示意图

Fig. 4 Schematic diagram of adhesion mechanism based on foot contact force feedback

2　机器人黏附爬行策略生成方法设计

空间机器人的黏附爬行策略生成方法，基于强化学习框架设计而成。强化学习框架通常由两部分组成：智能体与仿真环境。仿真环境又由物理仿真器与奖励函数组成。智能体通过当前状态进行动作决策，并于物理仿真器执行动作，生成新状态。奖励函数依据该动作和新状态计算奖励值，并反馈智能体，引导智能体训练生成更优的策略。

2.1　马尔可夫决策过程

采用强化学习框架求解空间黏附爬行机器人的运动控制问题，首先需将空间爬行机器人的黏附爬行过程抽象为马尔可夫决策过程，其分别包括机器人状态空间 $\boldsymbol{S}$、机器人腿足动作空间 $\boldsymbol{A}$、机器人状态的转移概率 $\boldsymbol{P}$ 以及动作空间获得的奖励 $\boldsymbol{R}$[23]。决策模型依据机器人当前状态决策出相应的腿足动作，并依据状态和动作获取下一时刻各状态出现的概率，以及当前状态下执行当前动作的奖励值。

空间机器人黏附爬行的马尔可夫过程关注机器人运动过程的未来累积奖励，即机器人当前时刻 t 至最终时刻 T（环境重置）过程中获得的奖励之和。未来累积奖励可以表示为

$$R_t=\sum_{k=0}^{T}\gamma^k r_{t+k} \tag{5}$$

式中：γ 为折扣因子，为 0 至 1 的小数；r_t 表示 t 时刻机器人获取的奖励值，奖励值计算函数由人为设计给出。折扣因子越大，则决策模型对未来奖励的重视程度越高。机器人运动的马尔可夫决策模型迭代目标是使决策的机器人关节动作能够获得尽量大的未来累积奖励，即要使下述状态-动作价值函数最大化

$$Q(s_t,a_t)=\sum_{s_{t+1}\in S}P(s_{t+1}\mid s_t,a_t)[r(s_{t+1}\mid s_t,a_t)+\gamma V(s_{t+1})] \tag{6}$$

式中：$Q(s_t,\ a_t)$ 表示机器人在当前时刻状态下，执行策略生成的腿足动作，而能够获得的未来累积奖励的期望；$r(s_{t+1}\mid s_t,\ a_t)$ 表示当前状态 s_t 下执行生成动作 a_t，达到下一时刻状态 s_{t+1} 的奖励值；$P(s_{t+1}\mid s_t,\ a_t)$ 表示在 s_t 下执行动作 a_t，使机器人达到 s_{t+1} 的概率。

$V(s_t)$ 为状态值函数，表示机器人在当前状态下能够获得的未来累积奖励概率期望，其可以表达为

$$V(s_t)=\sum_{a_t\in A}\pi(a_t\mid s_t)\sum_{s_{t+1}\in S}P(s_{t+1}\mid s_t,a_t)\cdot[r(s_{t+1}\mid s_t,a_t)+\gamma V(s_{t+1})] \tag{7}$$

式中，$\pi(a_t\mid s_t)$ 为决策模型依据状态 s_t 生成动作的概率分布。

2.2　密集型奖励函数设计

本文针对机器人的黏附爬行全过程设计密集型奖励函数，即每当机器人进入新的状态，奖励均随之变化。相较于目标导向的稀疏型奖励，本文所设计的奖励更关注机器人达到目标的过程状态，更有利于策略学习迭代，降低策略寻优、收敛难度。

机器人黏附爬行奖励函数按照目标抵达情况、本体稳定性、运动能耗等方面划分为以下几部分。

（1）目标位置抵达奖励

$$r_{\mathrm{g}}=\|\boldsymbol{p}_{\mathrm{goal}}-\boldsymbol{p}_{\mathrm{base\text{-}pre}}\|_2-\|\boldsymbol{p}_{\mathrm{goal}}-\boldsymbol{p}_{\mathrm{base}}\|_2 \tag{8}$$

式中，$\boldsymbol{p}_{\mathrm{goal}}$ 为给定目标位置，$\boldsymbol{p}_{\mathrm{base\text{-}pre}}$ 为上一时刻机器人本体位置，$\boldsymbol{p}_{\mathrm{base}}$ 为当前机器人本体位置，$\|\boldsymbol{a}\|_2$ 表示矢量 $\boldsymbol{a}$ 的 2-范数。当机器人运动时靠近目标位置，则给予相应奖励；若远离目标位置，则奖励变为惩罚。目标位置抵达奖励为最主要的任务奖励，其鼓励机器人向给定目标位置移动。

（2）时间步奖励

$$r_t=0.5 \tag{9}$$

该奖励为常数，若机器人爬行过程中未触发重置条件，则在每个时间步中均可获得此奖励，此奖励旨在鼓励机器人在爬行过程中尽量不触发环境重置条件。本文中，当机器人在爬行过程中与表面脱附，则仿真环境重置，因此时间步奖励鼓励机器人不脱附。

（3）平稳性奖励

$$r_{\mathrm{s}}=\begin{cases}\cos\theta_{\mathrm{b\text{-}f}} & \cos\theta_{\mathrm{b\text{-}f}}<0.95\\ 4 & \cos\theta_{\mathrm{b\text{-}f}}\geqslant 0.95\end{cases} \tag{10}$$

式中：$\theta_{\mathrm{b\text{-}f}}$ 为机器人本体坐标系的 Z 轴正方向与四足所受接触力合力矢量 $\boldsymbol{f}$ 的夹角。本文针对机器人单足提出了足端接触力沿用-更新机制，若足部与表面接触，则实时获取足端所受接触力，更新足端接触力矢量；若该足部未与表面接触，则沿用上一次接触时获取的接触力矢量。足端接触合力矢量 $\boldsymbol{f}$ 可表示为

$$\boldsymbol{f}=\boldsymbol{f}_{\mathrm{bl}}+\boldsymbol{f}_{\mathrm{br}}+\boldsymbol{f}_{\mathrm{fl}}+\boldsymbol{f}_{\mathrm{fr}} \tag{11}$$

式中，$\boldsymbol{f}_{\mathrm{bl}}$，$\boldsymbol{f}_{\mathrm{br}}$，$\boldsymbol{f}_{\mathrm{fl}}$，$\boldsymbol{f}_{\mathrm{fr}}$ 分别表示机器人左后、

右后、左前、右前足端所受接触力矢量，$\boldsymbol{f}_{\mathrm{bl}}$ 可表示为

$$\boldsymbol{f}_{\mathrm{bl}}=\begin{cases}\boldsymbol{f}_{\mathrm{bl}} & \|\boldsymbol{f}_{\mathrm{bl}}\|_2^2>0\\ \boldsymbol{f}_{\mathrm{bl\text{-}pre}} & \|\boldsymbol{f}_{\mathrm{bl}}\|_2^2=0\end{cases} \tag{12}$$

式中，$\boldsymbol{f}_{\mathrm{bl\text{-}pre}}$ 为上一次接触时左后足端所受接触力矢量。其余三足接触力矢量计算形式与上式相同。本文以此接触力更新机制构造奖励，引导机器人在足端未接触表面时依旧按照过往的运动方式爬行。

（4）前进方向奖励

$$r_{\mathrm{h}}=\begin{cases}\dfrac{\cos\theta_{\mathrm{head}}}{0.9} & \cos\theta_{\mathrm{head}}\leqslant 0.93\\ 1 & \cos\theta_{\mathrm{head}}>0.93\end{cases} \tag{13}$$

机器人足端接触力方向总是与爬行表面法向量方向相同，因此本文利用足端接触合力矢量，计算目标位置在机器人爬行所在平面的投影，而式（13）中的 θ_{head} 即为目标投影相对于机器人本体的位置矢量与机器人前进正方向的夹角。前进方向奖励旨在引导机器人始终面向目标位置运动。

（5）驱动力矩惩罚

$$p_{\mathrm{a}}=\sum_{j=0}^{23}a_j^2 \tag{14}$$

式中，a_j 表示第 j 个关节的驱动力矩。该惩罚项旨在防止机器人运动过程中因关节控制力矩过大而发生倾覆或脱附。

（6）能耗惩罚

$$p_{\mathrm{e}}=\sum_{j=0}^{23}\left|a_j\omega_j\right| \tag{15}$$

式中，ω_j 表示第 j 个关节的转动角速度。该惩罚项旨在引导机器人在运动过程中以尽量小的功率驱动关节转动。

（7）关节角度限位惩罚

$$p_{\mathrm{l}}=\sum_{j=0}^{23}l_j \tag{16}$$

$$l_j=\begin{cases}0 & \left|\dfrac{2\theta_j}{\theta_{j-\mathrm{uplimit}}-\theta_{j-\mathrm{lowlimit}}}\right|\leqslant 0.92\\ 1 & \left|\dfrac{2\theta_j}{\theta_{j-\mathrm{uplimit}}-\theta_{j-\mathrm{lowlimit}}}\right|>0.92\end{cases} \tag{17}$$

式中，θ_j 为第 j 个关节的转动角度，$\theta_{j\text{-uplimit}}$ 为第 j 个关节的转动角度上限位，$\theta_{j\text{-lowlimit}}$ 为转动角度下限位。该惩罚项旨在限制机器人各关节转动角度在限位区间内。

将上述各奖励、惩罚项依据其重要程度加权求和，构成机器人运动过程的整体奖励

$$\boldsymbol{R}=[k_{\mathrm{g}},k_{\mathrm{t}},k_{\mathrm{s}},k_{\mathrm{h}},-k_{\mathrm{a}},-k_{\mathrm{e}},-k_{\mathrm{l}}]\begin{bmatrix}r_{\mathrm{g}}\\ r_{\mathrm{t}}\\ r_{\mathrm{s}}\\ r_{\mathrm{h}}\\ p_{\mathrm{a}}\\ p_{\mathrm{e}}\\ p_{\mathrm{l}}\end{bmatrix} \tag{18}$$

式中，k 表示各奖励、惩罚项的权重。

（8）脱附惩罚

本文旨在实现微重力环境下机器人的稳定黏附爬行，爬行过程中保持黏附是最主要、最基础的要求，因此奖励设计中，脱附惩罚必须具备最高优先级，最终机器人运动奖励设计为

$$\boldsymbol{R}_{\mathrm{total}}=\begin{cases}R & n_{\mathrm{d}}<n_{\mathrm{reset}}\\ -2 & n_{\mathrm{d}}\geqslant n_{\mathrm{reset}}\end{cases} \tag{19}$$

式中，n_{d} 为机器人脱附所持续的时间步。若机器人四足所受接触合力为 0，则判定其进入脱附状态，此时 n_{d} 开始计数，当机器人累积脱附时间步数达到 n_{reset} 步后视为机器人完全脱附，进入不可控状态，此时最大惩罚覆盖所有奖励，爬行策略输出任何动作均无法获得正向反馈。该奖励设计旨在防止机器人在爬行过程中完全脱附。

2.3 基于黏附作用机制和密集型奖励的爬行策略训练框架设计

空间机器人黏附爬行过程中的腿足控制，是一种复杂的连续控制问题。现有研究针对连续控制问题，基于马尔可夫决策过程提出了多类强化学习算法，其中深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法与置信域策略优化（Trust Region Policy Optimization，TRPO）算法均具有较强的代表性。近端策略“优化-裁剪”（Proximal Policy Optimization-clip，PPO-clip）算法[24]基于 Actor-Critic 网络框架，采用旧策略的交互样本训练新策略，相较于 DDPG 具备更高的样本利用效率；同时采用裁剪措施限制迭代过程中新、旧策略输出的动作概率分布的距离，相较于 TRPO 减小了计算复杂度，更适合解决机器人黏附爬行的复杂连续控制问题。

训练过程中，策略网络（Actor 网络）更新前通过与环境交互，收集基于旧策略参数的交互数据样本（“状态-动作-奖励”），并依据旧策略在特定状态下输出对应动作的概率来引导网络参数更新，形

成新策略。策略参数更新完毕后，新策略网络再次与环境交互，进入下一轮的参数迭代更新过程，直至获得的累积奖励收敛。其目标函数可表示为

$$J^{\theta'}_{\text{PPO-CLIP}}(\theta)=\sum_{(s_t,a_t)\sim\pi_{\theta'}}\min\left\{\frac{p_{\theta'}(a_t\mid s_t)}{p_\theta(a_t\mid s_t)}A_\theta(s_t,a_t),\right.$$
$$\left.\text{clip}\left[\frac{p_{\theta'}(a_t\mid s_t)}{p_\theta(a_t\mid s_t)},1-\varepsilon,1+\varepsilon\right]A_\theta(s_t,a_t)\right\}\quad(20)$$

$$A_\theta(s_t,a_t)=Q_\theta(s_t,a_t)-V_\theta(s_t)\quad(21)$$

$$\text{clip}(x,x_1,x_2)=\begin{cases}x_1 & x<x_1\\ x & x_1\leqslant x\leqslant x_2\\ x_2 & x>x_2\end{cases}\quad(22)$$

其中，$p_\theta(a_t\mid s_t)$ 表示旧策略依据 s_t 生成 a_t 的概率，$p_\theta(a_t\mid s_t)$ 表示新策略依据 s_t 生成 a_t 的概率，$A_\theta(s_t,\ a_t)$ 表示旧策略生成动作 a_t 的优势函数，即 s_t 下生成 a_t 的动作值函数与 s_t 的状态值函数的差值，$\varepsilon\in(0,\ 1)$ 由人为给定，限制新、旧策略输出动作概率分布的比值处于$[1-\varepsilon,\ 1+\varepsilon]$区间。

价值网络（Critic 网络）的输入为当前状态 s_t 和对应动作 a_t，以及当前奖励值 $r(s_t,\ a_t)$，输出状态值函数 $V(s_t)$。策略网络迭代更新一轮，价值网络依据策略网络输出动作和对应状态的奖励，依据基于状态价值函数的损失函数

$$L_{\text{Critic}}=\frac{1}{2}[r(s_t,a_t)+\gamma V(s_{t+1})-V(s_t)]^2\quad(23)$$

采用梯度下降方式自我迭代，使价值网络输出的状态值函数尽可能趋于准确，以引导策略网络在迭代过程中逐步收敛至最优。

策略网络与价值网络的结构与机器人动作空间、状态空间维度相关。机器人整体包含 24 个关节自由度，每个关节自由度通过输出驱动力矩来控制机器人腿足运动，最终构成 24 维的动作空间。

机器人黏附爬行过程中，可观测包括机器人运动过程中本体相对于爬行表面的高度、机器人本体移动线速度和角速度、机器人本体的滚转角度、偏航角度、机器人前进方向与目标方向的夹角、机器人本体到目标位置的距离、机器人腿足关节转动角度和角速度、机器人足端所受接触力方向以及上一时刻动作指令等状态变量，构成 98 维状态空间。

本文所述策略网络和价值网络均由多层感知机构成，且均具备 4 层隐含层，每层隐藏层包括 256 个网络节点，隐藏层共包含 256×256×256×256 个网络权重。策略网络结构为 98×256×256×256×256×24，包括 98 个输入节点和 24 个输出节点，对应 98 维状态数据以及 24 维腿足关节驱动数据；价值网络结构为 98×256×256×256×256×1，包括 98 个输入节点和 1 个输出节点，对应 98 维状态数据以及 1 维状态价值数据。策略网络与价值网络均采用 elu 函数作为激活函数。

微重力环境下基于足端黏附作用机制和密集型奖励的机器人黏附爬行策略训练框架如图 5 所示：决策网络输出机器人腿足关节控制力矩至物理仿真器，物理仿真器依据关节控制力矩进行机器人运动仿真，并更新机器人状态数据。同时，仿真器依据足端黏附作用机制，采用机器人足端接触力数据在每一个仿真时间步内计算黏附作用力，并施加该力于机器人足部刚体。密集型奖励函数依据机器人状态和动作数据，辅以足端接触力“沿用-更新”机制，计算当前时间步内机器人获得的奖励数据，并与机器人状态、动作数据共同构成四元组 $[s_t,\ a_t,\ s_{t+1},\ r_t]$ 存入采样池。决策网络与价值网络从采样池中抽取序列数据，采用 PPO-clip 算法完成迭代更新。

通过大量试验测试，密集型奖励函数中各项奖励与惩罚的权重系数、脱附时间步数确定为

$$k_{\text{g}}=1.8,k_{\text{t}}=1,k_{\text{s}}=0.05,k_{\text{h}}=0.577\ 5$$
$$k_{\text{a}}=3.5\times10^{-4},k_{\text{e}}=8\times10^{-4},k_{\text{l}}=0.012$$
$$n_{\text{reset}}=35$$

在该权重系数与参数构成的奖励下，训练获取的黏附爬行策略较优。

3　仿真结果分析

本文拟进行 3 类仿真实验，验证基于黏附作用机制和密集型奖励函数训练生成的黏附爬行策略的性能，以及奖励函数设计的有效性。实验主要包括：1）在训练轮数相同条件下，验证足端接触力数据“沿用-更新”机制的作用效果；2）将训练获取的爬行策略应用于空间爬行机器人黏附任务，验证爬行策略的性能；3）改变爬行表面特性，验证固化爬行策略在不同爬行表面的泛化能力。

本文仿真实验均在 64 位 Ubuntu20.04 操作系统下进行，中央处理器型号为 Intel（R）Core（TM）i7-10750 H CPU @ 2.60 GHz，运行内存为 64 GB，显卡为 NVIDIA GeForce RTX 2080 Super，程序运行环境基于 python3.7，仿真器采用 NVIDIA Isaac Sim，强化学习仿真环境基于 NVIDIA Isaac Orbit 开发。

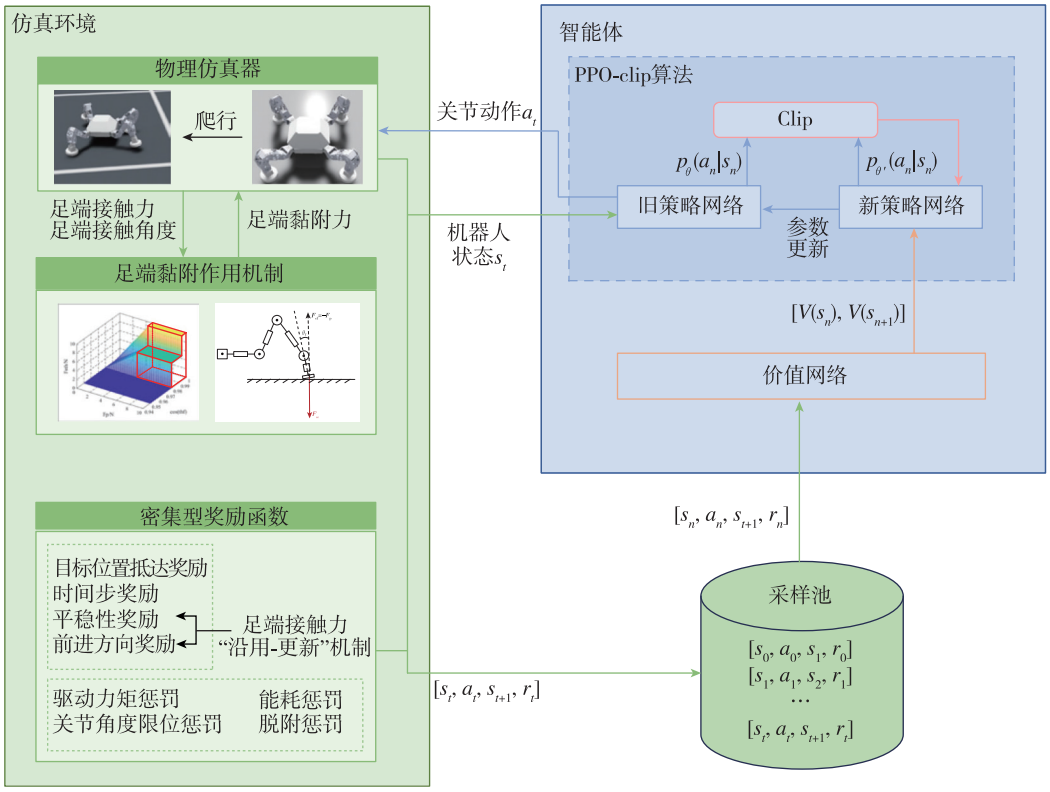


图 5　基于足端黏附作用机制和密集型奖励的爬行策略训练框架示意图

Fig. 5　Schematic diagram of crawling strategy training framework based on foot adhesion mechanism and intensive reward

3.1　环境配置及超参数设置

强化学习仿真环境需设定重置条件以重置机器人状态，为新策略提供新的交互数据，本文开发的空间机器人黏附爬行仿真环境满足以下条件之一即触发重置：1）机器人该轮仿真达到最大时间步；2）机器人完全脱附。

NVIDIA Isaac Orbit 是由英伟达公司开发的并行强化学习环境，能够同时对千级数量级的机器人进行仿真，极大提高样本收集的效率。在初始时刻，机器人以固定间距按照方阵排列。为使尽量多的机器人能够在环境重置前达到目标点，同时避免大幅降低训练速率，本文将最大仿真时间步设定为 250 0 步，每个时间步持续时间 0.008 3 s，其他环境参数和算法超参数如表 2、表 3 所示。

表 2　仿真环境配置参数

Tab. 2　Simulation environment configuration parameters

参数名称	参数含义/单位	参数值
num _ envs	并行机器人个数/个	800
env _ spacing	机器人间距/m	5
power _ scale	最大驱动力矩系数×10^{-1}/（N·m）	0.3
gravity	重力加速度/（m·s^{-2}）	0.0

表 3　算法超参数

Tab. 3　Algorithm hyperparameter

参数名称	参数含义	参数值
gamma	累积奖励折扣因子	0.99
Learning rate	学习率	6×10^{-3}
e _ clip	裁剪因子	0.2
max _ epochs	最大迭代轮数	3 000
batch _ size	样本批量大小	51 200
minibatch _ size	小样本批量大小	6 400

3.2　奖励函数对比仿真验证

空间机器人黏附爬行策略训练过程的累积平均奖励变化曲线如图 6 所示。original _ reward 表示未加入“沿用-更新”机制时爬行策略训练过程的累积奖励变化曲线，improved _ reward表示加

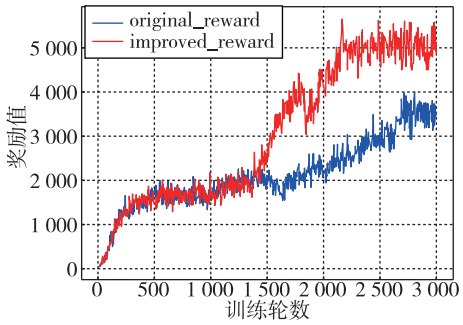


图 6　累积奖励变化曲线

Fig. 6　Cumulative reward change curve

入“沿用-更新”机制后的累积奖励变化曲线。

在训练初期，由于策略并不具备先验知识，神经网络初始参数随机，策略对关节动作进行随机探索，因此策略所获得的累积奖励较低。在此阶段，两类累积奖励变化曲线趋势相似。

随着训练轮数增加，策略探索获取的高奖励动作及状态样本数量增多，累积奖励增长速率增加。在此阶段，加入“沿用-更新”机制前，累积奖励变化曲线增长速度较为缓慢，在约 1 600 个训练轮数后开始加速上升。在约 2 700 个训练轮数后，累积奖励值增长速率放缓，并最终小幅震荡，策略进入收敛阶段。

而加入“沿用-更新”机制后，在约 1 400 个训练轮数后，累积奖励开始快速上升。在约 2 100 个训练轮数后，累积奖励上升放缓，在约 2 300 个训练轮数后，累积奖励值仅小幅度震荡，爬行策略收敛。可见，加入上述机制后，策略所获得的奖励提早约 6.67%进入快速增长阶段，约 13.33%进入收敛阶段，策略收敛速率增加了约 14.81%。

3.3　爬行策略性能仿真验证

图 7 为四足机器人在爬行策略控制下，在空间微重力环境中进行黏附爬行的示意图。可见机器人在前进过程中并未出现翻倒、脱附的情况。

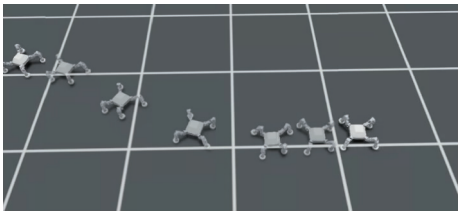
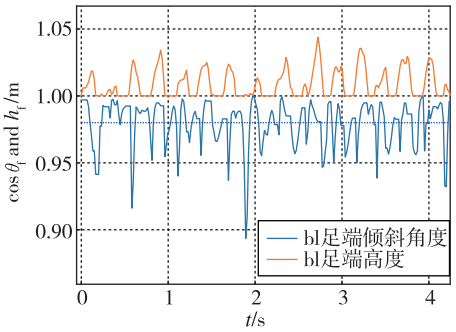


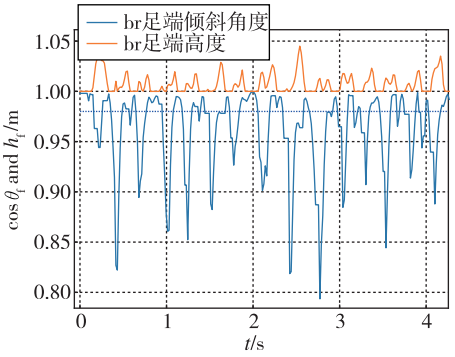
图 7　微重力环境下四足机器人黏附爬行示意图

Fig. 7　Schematic diagram of adhesive climbing of quadruped robot in microgravity environment

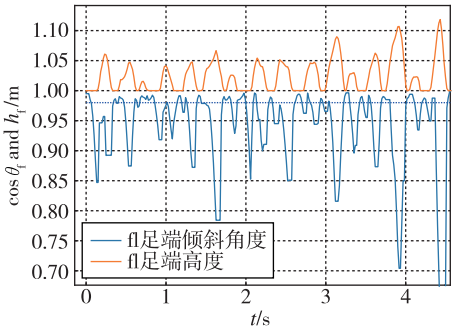
图 8 为机器人爬行过程中足端高度 h_f 和足底倾斜角度 θ_f 随时间 t 的对应变化关系，其中 bl，br，fl，fr 分别指代机器人的左后、右后、左前、右前足。为便于比较，本文绘制足端高度变化曲线时，将其统一上移 1 m，因此该图中，足端高度值 1 m 对应足底触地状态，足底倾斜角度采用足端刚体模型坐标系的 z 轴正方向与爬行表面法向量夹角的余弦值（即足端倾斜角的余弦值）表示。由图 8 可以看出，当机器人落足时，足端倾斜角度余弦值迅速增加，并且余弦值在足端触地时能够达到 0.98（蓝色虚线）以上，即倾斜角度小于 12°，符合黏附力作用的足端倾斜角约束。



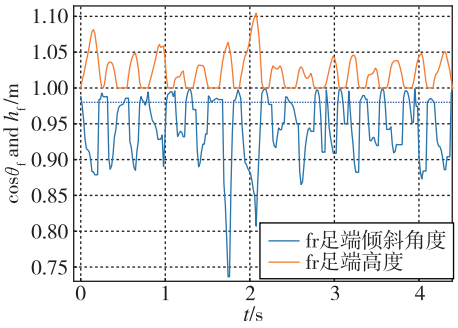
(a) 左后足高度、倾斜角变化曲线



(b) 右后足高度、倾斜角变化曲线



(c) 左前足高度、倾斜角变化曲线



(d) 右前足高度、倾斜角变化曲线

图 8 机器人四足高度、倾斜角变化曲线

Fig. 8 Robot quadruped height and tilt angle change curve

图 9、图 10、图 11 分别为机器人爬行过程中各足端高度、接触力、黏附作用力的变化情况。通过图 9 高度变化曲线可知，爬行策略控制机器人以“对足起-对足落”的形式向前运动。首先，策略同时控制右前、左后足抬起，保持左前、右后足触地不动，当右前、左后足回落触地；随后，策略同时控制左前、右后足抬起迈步，如此循环往复。通过图 10 足端接触力变化曲线可知，当足端触地时，产生的接触力均大于 5 N，符合黏附作用力触发的预压力约束。通过图 11 足端黏附作用力变化曲线可知，当机器人足端触地时，黏附作用力成功作用于对应足端；当足端抬起，黏附作用力消失。

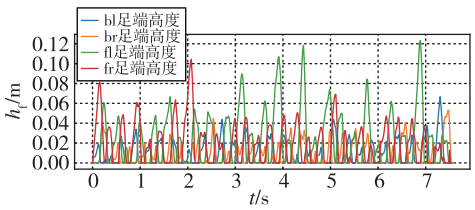


图 9 足端高度变化曲线

Fig. 9 Foot height change curve

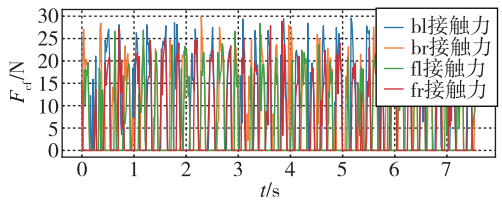


图 10 足端接触力变化曲线

Fig. 10 Foot contact force change curve

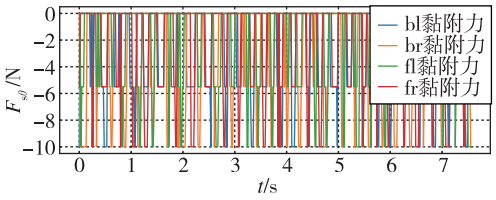


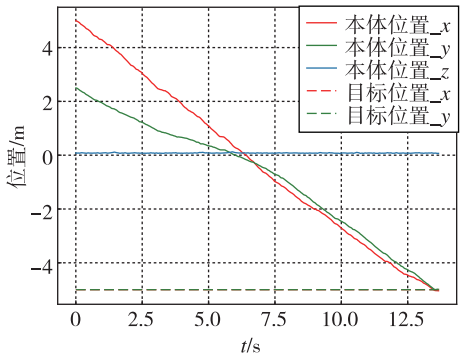
图 11 足端黏附作用力变化曲线

Fig. 11 Foot adhesion force change curve

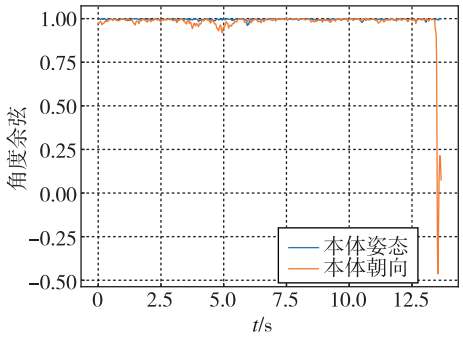
爬行策略控制机器人以“对足起-对足落”的类对角步态运动，以此确保在爬行过程中，当一侧对角的双足向前迈步以驱动机器人本体前进时，另一侧对角的双足始终接触表面，为机器人提供黏附作用力，以实现机器人在微重力环境下的稳定黏附爬行。

图 12、图 13、图 14 分别依据 4 种目标位置，绘制了机器人的本体位置变化、姿态稳定性以及运动朝向曲线。图 12 (a)、图 13 (a)、图 14 (a) 中，红色、绿色、蓝色实线分别表示机器人本体在世界坐标系下的 x 轴、y 轴、z 轴坐标值，红色、绿色虚线分别表示目标位置在世界坐标系下的 x 轴、y 轴坐标值；本体姿态和本体朝向曲线部分，本体朝向表示机器人本体正方向与目标方向之间的夹角余弦值，本体姿态表示机器人本体坐标系的 z 轴正方向与表面法向之间的夹角余弦值。

机器人初始位置坐标设定为 [5，2.5，0]，由图 12 (a)、图 13 (a)、图 14 (a) 中位置曲线变化趋势可知，机器人爬行过程中持续趋近于目标位置，最终抵达目标位置附近（误差小于 0.1 m，小于机器人单腿长度的 1/2，满足机器人的后续操作要求），由机器人的爬行距离及消耗时间的变化关



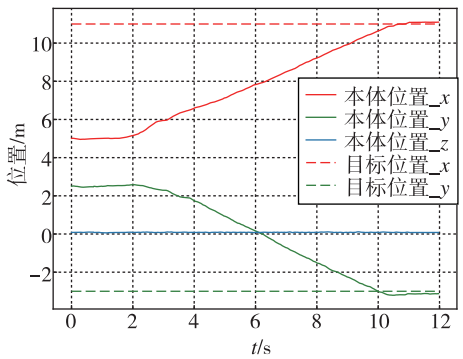
(a) 本体位置变化曲线



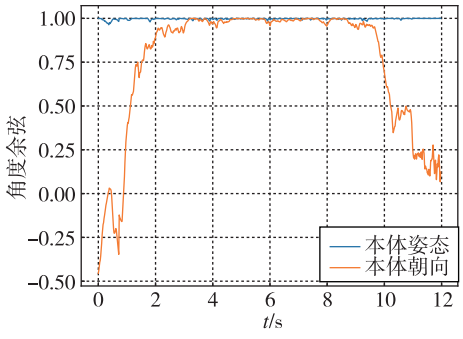
(b) 本体姿态、朝向角度变化曲线

图 12　目标位置［−5，−5，0］时机器人本体各变化曲线

Fig. 12　Each change curve of the robot body at the target position［−5，−5，0］



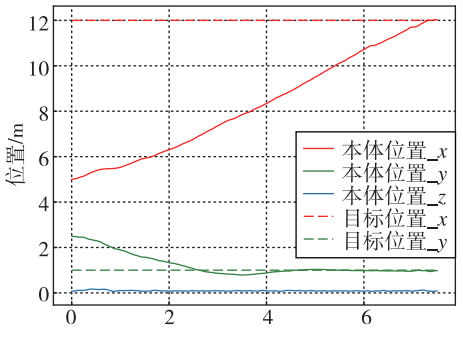
(a) 本体位置变化曲线



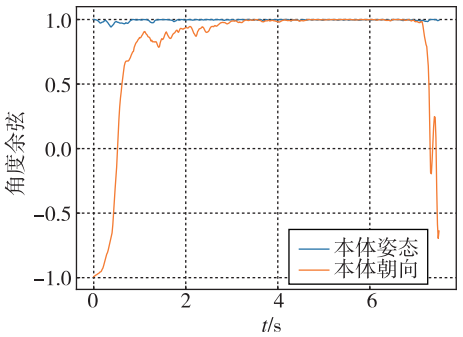
(b) 本体姿态、朝向角度变化曲线

图 13　目标位置［11，−3，0］时机器人本体各变化曲线

Fig. 13　Each change curve of the robot body at the target position［11，−3，0］



(a) 本体位置变化曲线



(b) 本体姿态、朝向角度变化曲线

图 14　目标位置［12，1，0］时机器人本体各变化曲线

Fig. 14　Each change curve of the robot body at the target position［12，1，0］

系，可知机器人在调整完毕前进方向后，黏附爬行的平均速度可达 0.8～0.9 m/s，且机器人本体高度曲线（z 轴坐标值变化曲线）平稳，本体高度始终维持于 60～80 mm；同时，由本体朝向曲线和本体姿态曲线变化趋势可知，机器人抵达目标位置过程中，能够自主调整本体的前进方向，保持机器人本体面向目标位置爬行，朝向角度余弦值误差不超过 0.06 且爬行过程中姿态角度余弦值误差不超过 0.01，能够保持本体运动的姿态稳定性。

3.4　爬行策略泛化能力仿真验证

本文的黏附爬行策略完全基于微重力环境下平坦爬行表面训练生成，且训练过程中完全没有为其提供其他类型表面的黏附爬行数据。本文通过改变机器人爬行表面特征，分别构造台阶和坡面，以验证基于平坦表面训练固化的黏附爬行策

略在爬行表面特性发生意外变化后的适应能力。

图 15 分别展示了四足爬行机器人在爬行策略控制下穿越台阶和坡面地形的情况。其中，台阶表面可划分为两部分，由高度 40 mm 的矩形刚体和水平表面构成；坡面则可划分为 3 部分，分别由倾斜角度为±18°的斜面以及水平面构成。

机器人在台阶表面黏附爬行，其面临的挑战主要为台阶上下的高度落差导致的四足落足高度不均匀，从而增加脱附风险。由图 15（a）可知，当机器人前足踏于台阶上方、后足仍处于台阶下方时，在爬行策略的控制下，本体向台阶上方倾斜；当机器人四足完全处于台阶之上，机器人本体姿态恢复水平。当机器人到达台阶下降沿，进入回落阶段，其前足位于台阶下方，后足位于台阶上方，本体向台阶下方倾斜。当回落阶段结束，机器人调整本体姿态，使其恢复水平位置。

机器人在斜坡表面黏附爬行，影响黏附稳定性的因素除落足高度变化外，还包括落足角度的变化。由图 15（b）可知，当机器人开始进入倾斜角度变化的表面时，其前足落于倾斜角度变化的新表面上，而后足仍处于原先表面，在爬行策略的控制下，机器人处于黏附状态的前、后足均能够贴合其所接触表面，同时机器人本体姿态沿爬行表面坡度的变化的方向进行调整。当机器人完全处于新表面后，机器人本体姿态完成调整过渡，与新表面坡度基本保持一致。

(a) 台阶表面爬行

(b) 斜坡表面爬行

图 15 微重力环境下机器人在多类表面黏附爬行仿真

Fig. 15 Simulation of robot adhesive climbing on multiple surfaces in microgravity environment

综上，基于密集奖励和黏附力作用机制训练生成的机器人爬行策略，具备在微重力平坦表面实现机器人稳定黏附爬行的能力。当爬行表面高度及坡度在平坦表面基础上发生意外的高度及坡度变化，即使爬行策略在生成过程中并未采集过坡面爬行和台阶表面爬行数据，其仍具备控制机器人在高度变化±40 mm、坡度变化±18°范围内的爬行表面实现稳定黏附爬行的能力，验证了该爬行策略的泛化能力。

4 结论

针对现阶段微重力环境下空间黏附爬行机器人控制策略泛化能力不足的问题，本文提出了一种结合足端黏附作用机制和强化学习框架的机器人黏附爬行策略：依据足端黏附材料特性，构造机器人足端黏附作用约束模型，通过施加外部作用力的方式为机器人足端添加黏附机制；依据机器人爬行的黏附要求，在奖励设计中构造脱附惩罚，同时参考生物学习运动技能过程中考虑邻近历史状态的特性，提出了足端接触力的“沿用-更新”机制，并以此构造密集型机器人爬行奖励；基于黏附作用机制和密集型奖励，采用 PPO-clip 算法训练爬行策略，并在仿真平台上完成爬行策略的验证。由仿真实验结果可知：所设计的密集型奖励能够增加爬行策略的收敛速率，所获取的爬行策略能够控制机器人在微重力环境下的平坦表面实现稳定黏附爬行，爬行过程中未出现脱附情况，且机器人本体高度和姿态变化幅度较小，并展现出了较为准确的目标位置抵达能力；同时，基于平坦表面训练生成的爬行策略在高度、坡度发生意外变化的爬行表面均能控制机器人实现稳定黏附爬行，验证了爬行策略的泛化能力。

本文仅针对单一材料表面的机器人黏附爬行策略生成方法开展研究，因此在后续工作中，拟在训练过程中多样化爬行表面特性，考虑不同表面材料对足端黏附作用力的影响；考虑引入视觉信号，与机器人状态数据共同训练生成爬行策略，增强爬行策略泛化性；深入研究目标位置抵达奖励的设计方法，提升机器人目标位置抵达的准确性；计划进行策略的实机部署，完成物理样机试验。

参考文献(References)

[1] 岳晓奎，张滕. 在轨服务软体机器人应用展望[J]. 飞控与探测，2020，3(1)：1-7.
YUE X K，ZHANG T. Soft robots for on-orbit service [J]. Flight Control & Detection，2020，3(1)：1-7 (in Chinese).

[2] 赵亮亮，李雪皑，赵京东，等. 面向航天器自主维护的空间机器人发展战略研究[J]. 中国工程科学，2024，26(1)：149-159.
ZHAO L L，LI X A，ZHAO J D，et al. Development strategy of space robots for autonomous repair and maintenance of spacecraft [J]. Strategic Study of CAE，2024，26(1)：149-159 (in Chinese).

[3] BLAISE J，BAZZOCCHI M C F. Space manipulator collision avoidance using a deep reinforcement learning control[J]. Aerospace，2023，10(9)：778.

[4] ZHAO J D，TANG J W，ZHAO Z Y，et al. Inverse kinematics of a reconfigurable redundant manipulator with lockable passive telescopic links [C]// 2022 IEEE International Conference on Mechatronics and Automation (ICMA)，Guilin，Guangxi，China：IEEE，2022：557-562.

[5] 石凯，李军，王树兵，等. 基于 PWM 的气动软体空间机械臂压力控制系统[J]. 飞控与探测，2021，4(6)：61-69.
SHI K，LI J，WANG S B，et al. Pressure control system of pneumatic soft space manipulator based on PWM [J]. Flight Control & Detection，2021，4(6)：61-69 (in Chinese).

[6] 张欧阳. 自由飞行空间机器人目标捕获运动规划与控制研究[D]. 哈尔滨：哈尔滨工业大学，2022.
ZHANG O Y. Research on motion planning and control of free-flying space robot for target capturing [D]. Harbin：Harbin Institute of Technology，2022 (in Chinese).

[7] MOGHADDAM B M，CHHABRA R. On the guidance，navigation and control of in-orbit space robotic missions：a survey and prospective vision [J]. Acta-Astronautica，2021，184(1)：70-100.

[8] TANG T，HOU X，XIAO Y，et al. Research on motion characteristics of space truss-crawling robot [J]. International Journal of Advanced Robotic Systems，2019，16(1)：1729881418821578.

[9] 唐玲，王克鹏，张彬，等. 空间黏附足式爬行机器人的稳定性判据及蠕动步态[J].宇航学报，2022，43(9)：1186-1195.
TANG L，WANG K P，ZHANG B，et al. Stability criterion and creep gait of a space robot with adhesive feet [J]. Journal of Astronautics，2022，43 (9)：1186-1195 (in Chinese).

[10] HONG S，UM Y，PARK J，et al. Agile and versatile climbing on ferromagnetic surfaces with a quadrupedal robot [J]. Science Robotics，2022，7 (73)：eadd1017.

[11] LI Z Y，LI Z J，TAM L M，et al. Design and development of a versatile quadruped climbing robot with obstacle-overcoming and manipulation capabilities [J]. IEEE/ASME Transactions on Mechatronics，2023，28(3)：1649-1661.

[12] 宋益帆，王炳诚，段晋军，等. 基于黏附稳定包络边界的爬壁机器人竖直壁面过渡策略[J]. 机器人，2023，45(5)：532-545.
SONG Y F，WANG B C，DUAN J J，et al. A vertical wall transition strategy with reliable adhesion envelope boundaries for wall-climbing robots [J]. Robot，2023，45(5)：532-545 (in Chinese).

[13] YIN F L，TANG A N，XU L W，et al. Run like a dog：learning based whole-body control framework for quadruped gait style transfer [C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)，Prague，Czech Republic：IEEE，2021：8508-8514.

[14] SHENG J P，CHEN Y Y，FANG X，et al. Bio-inspired rhythmic locomotion for quadruped robots [J]. IEEE Robotics and Automation Letters，2022，7(3)：6782-6789.

[15] LEE J，HWANGBO J，WELLHAUSEN L，et al. Learning quadrupedal locomotion over challenging terrain [J]. Science Robotics，2020，5(47)：eabc5982.

[16] MIKI T，LEE J，HWANGBO J，et al. Learning robust perceptive locomotion for quadrupedal robots in the wild [J]. Science Robotics，2022，7(62)：eabk2822.

[17] MASTROGEORGIOU A S，ELBAHRAWY Y S，KECSKEMÉTHY A，et al. Slope handling for quadruped robots using deep reinforcement learning and toe trajectory planning [C]// 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)，Las Vegas，Nevada，USA：IEEE，2020：3777-3782.

[18] BELLEGARDA G，CHEN Y Y，LIU Z C，et al. Robust high-speed running for quadruped robots via deep reinforcement learning [C]// 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)，Kyoto，Japan：IEEE，2022：10364-10370.

[19] 盛嘉鹏. 基于深度强化学习的四足机器人节律运动控

制方法研究[D]. 济南:山东大学, 2023.
SHENG J P. Research on rhythmic motion control method of quadruped robot based on deep reinforcement learning [D]. Jinan: Shandong University, 2023 (in Chinese).

[20] XIAO H, SHAO S, ZHANG D. Agile control for quadruped robot in complex environment based on deep reinforcement learning method [C]// 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China: IEEE, 2021: 1065-1070.

[21] WU J Z, XIN G Y, QI C K, et al. Learning robust and agile legged locomotion using adversarial motion priors [J]. IEEE Robotics and Automation Letters, 2023, 8(8): 4975-4982.

[22] MURPHY M P, SITTI M. Waalbot: an agile small-scale wall climbing robot utilizing dry elastomer adhesives [J]. IEEE/ASME Transactions on Mechatronics, 2007, 12 (3): 330-338.

[23] 张柄汉, 王琛, 彭兆涛, 等. 一种面向空间非合作目标的强化学习多臂协同俘获策略研究[J]. 宇航学报, 2023, 44(12): 1934-1943.
ZHANG B H, WANG C, PENG Z T, et al. A reinforcement learning capture strategy for non-cooperative targets based on a multi-arm synergy method [J]. Journal of Astronautics, 2023, 44(12): 1934-1943 (in Chinese).

[24] ZHU W S, ROSENDO A. A functional clipping approach for policy optimization algorithms [J]. IEEE Access, 2021, 9(1): 96056-96063.