使用可伸缩足部提升四足机器人在复杂地形上的运动效率

Lokesh Kumar*TCS Research加尔各答，印度
ku
mar.lokesh7@tcs.com

Sarvesh Sortee*TCS Research加尔各答，印度
sarves
h.sortee@tcs.com

Titas BeraTCS Research加尔各答，印度
titas.bera@tcs.com

Ranjan DasguptaTCS Research加尔各答，印度
ra
njan.dasgupta@tcs.com

arXiv:2305.01998v1 [cs.RO] 3 May 2023

摘要—近年来，腿式运动研究的进展使腿式机器人在穿越复杂地形时，相较于轮式机器人成为更受青睐的选择。本文提出了一种新颖的运动策略，该策略采用深度强化学习进行训练，应用于一种四足机器人；该机器人在每条腿的膝部与足部之间额外配备了一个棱柱关节。训练在 NVIDIA Isaac Gym 仿真环境中进行。我们的研究探讨了这些关节对四足机器人在穿越复杂地形时保持期望高度并跟随指令速度的影响。我们基于运输成本（CoT）指标，对配备和未配备棱柱关节的四足机器人进行了比较。我们在仿真中使用运输成本指标，在一组复杂地形上对学习策略进行评估。结果表明，新增的驱动自由度为运动策略提供了更大的灵活性，使其能够利用额外关节穿越对于传统四足机器人设计而言不可行或成本高得难以接受的地形，从而显著提高效率。

索引词—**用于机器人控制的深度强化学习、腿式机器人、崎岖地形、腿式运动**

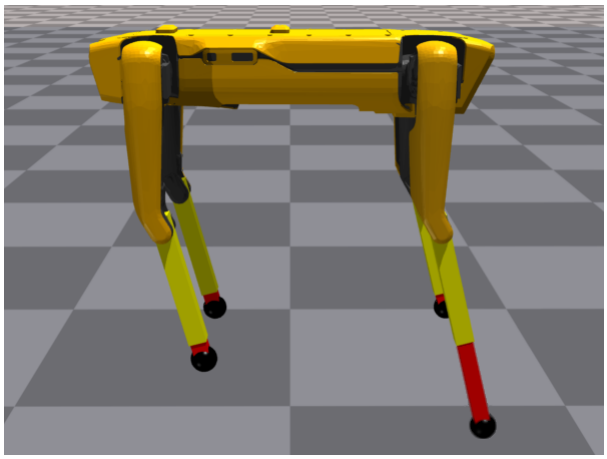


图 1：膝部与足部之间装有驱动棱柱关节的四足机器人，此处通过红色 连杆的伸展示意

I. 引言

在非平坦、无结构地形环境中进行腿式运动是一项复杂的任务。用于此类问题的传统控制方法通常由多个相互连接的不同模块组成，例如感知处理模块、落足点规划模块、状态估计模块、步态/轨迹生成器模块，以及在足部打滑时用于反射控制的反应模块 [1]–[4]。随着所考虑的运行条件日益严苛，这些架构也变得愈加复杂。深度强化学习（DRL）等其他方法，尤其是无模型强化学习，近年来发展迅速，已成为传统方法的一种替代方案。在深度强化学习中，机器人通过与仿真环境持续交互，根据获得的奖励，利用控制策略学习做出决策。本文提出一种在复杂地形上运动的新方法，即使机器人足部能够伸展。具体做法是在膝部之后增加棱柱关节。添加这些关节的目的是赋予机器人穿越高度和深度超过传统四足机器人腿部可达范围的地形，同时利用这些额外的自由度保持期望高度。这对于在煤矿或隧道内创建地形图等任务可能有所帮助；在泥地和雪地等地面不稳固的场景中，四足机器人也必须保持特定的机身高度。据我们所知，现有文献中尚无类似研究评估此类改动的影响，因此本研究具有创新性。后续章节将讨论本项工作的具体内容。

A. 相关工作

在不规则地形上行走对腿式运动构成了复杂的挑战，研究人员已开发出多种方法来应对这一问题。早期研究（如 [5],[6] ）将地形穿越视为一个分层规划问题的求解过程：在每次足地接触事件之间，先选择一个落足点，再规划通往该落足点的腿部轨迹。中央模式发生器（CPG）已被

* Lokesh Kumar 和 Sarvesh Sortee 对本项工作贡献相同。

与感觉数据结合使用，使四足动物能够穿越未预见的环境 [7]。Righetti 等人 [8] 利用安装在足部的触觉传感器反馈来改变腿部运动并穿越给定地形，而 Hyun 等人 [9]则使用结合本体感受阻抗控制的分层控制器。S. Gay 等人 [10]利用摄像头和陀螺仪的反馈，并结合中央模式发生器（CPG），学习在不同地形上行走。模型预测控制（MPC）方法使用简化的动力学模型，并在每个时间步求解在线优化问题，以实现期望的运动。 [11]提出了一种用于在非结构化地形上进行动态运动的感知耦合模型预测控制方法。这些方法的一个缺点是，必须从外部指定步态，而且可能需要预先定义反应行为，以应对滑移和不正确的接触估计。

近年来，人们对基于学习的控制方法的兴趣激增，这些方法旨在解决四足机器人在非平坦非结构化地形上运动泛化的问题。深度强化学习（DRL）已被用于训练能够适应雪地、泥地和植被等未见环境的鲁棒控制策略 [12]–[14]。S. Chan 等人 [15]尝试直接学习关节力矩，而 Shi 等人 [16] 采用结合式学习方法，同时学习轨迹生成器和关节目标残差，从而成功通过各种地形。一些研究采用机器学习技术进行地形感知 [17]–[19], ，另一些研究则使用深度强化学习实现地形感知运动 [20]–[22]。Wijmans 等人 [23]利用大量仿真样本，为未见环境训练基于视觉的强化学习算法。Octavio V 等人 [18]使用卷积神经网络（CNNs）学习非结构化地形上的落足点适应。

尽管这一领域已取得进展，但关于评估为四足机器人增加棱柱关节，使其无需采用弹跳步态和跳跃等动态步态，便能跨越起伏幅度超过足部可达范围的地形，其效果的研究仍有不足。

B. 贡献

与前文讨论的工作不同，先前的研究主要致力于开发运动策略，而不改变四足机器人的机械设计；我们的工作则采用了一种新颖的方法，通过改变其形态来实现这一目标。具体而言，我们为每条腿增加了一个额外的自由度，使四足机器人不仅能够穿越复杂地形，还能通过标称形态此前无法通行的地形，而无需采用弹跳步态或跳跃。这是因为可行落足点超出了足部的可达区域。尽管 Spro witz 等人 [24] 等工作基于仿生步态改变了机械设计，但并未增加任何额外的自由度，因此与之相比，我们的方法独具特色。

本文其余部分的组织结构如下：第二节简要介绍我们所做的修改，以便于做爰片传统四足机器人模型，并介绍我们用于仿真训练和评估的地形类型。第三节概述强化学习问题，说明状态空间和动作空间的选择、奖励以及训练方法。在第四节中，我们展示训练结果，以及通过运输成本（CoT）衡量的评估结果；评估对象是在各种地形上训练得到的策略，分别适用于带棱柱关节和不带棱柱关节的机器人。

II. 系统设计

我们采用 Boston Dynamics 的 Spot 作为四足机器人基座，并对其小腿和足部连杆进行了改造。在每条腿的膝部与足部之间增加了一个最大行程为 0.15 m 的棱柱关节。机器人共有 16 个驱动关节，每条腿 4 个：3 个旋转关节（髋部内收/外展、髋部屈曲/伸展以及膝部屈曲/伸展）和 1 个线性关节（足部伸展/回缩）。为简化仿真中的接触计算，机器人的足部被简化为球形连杆。机器人如图 1 所示。为了训练四足机器人适应不同地形，我们使用了受 [25] 启发的环境。整个训练环境包含五种不同地形，每种地形的难度均按下文所述从第 1 级提升至第 12 级：

- 平滑斜坡 [图2a]: 随着斜坡角度从 [0.05- 0.4] 弧度逐渐增大。
- 崎岖斜坡 [图 2b] ： 斜坡角度从 [0.05- 0.4]弧度逐渐增加，并加入了一些不均匀性。
- 上楼梯 [图2c]: 楼梯高度变化范围为 [0.05- 0.35]m，楼梯宽度变化范围为 [0.05 - 0.3]m。
- 下楼梯 [图2d] ： 楼梯高度在[0.05 - 0.35]米之间变化，楼梯宽度在 [0.05 - 0.3]米之间变化。
- 随机障碍物 [图2e] ： 障碍物高度在 [0.05 - 0.35]m 之间随机变化。

为了评估学习策略，我们使用了第四节中讨论的难度等级最高的地形。

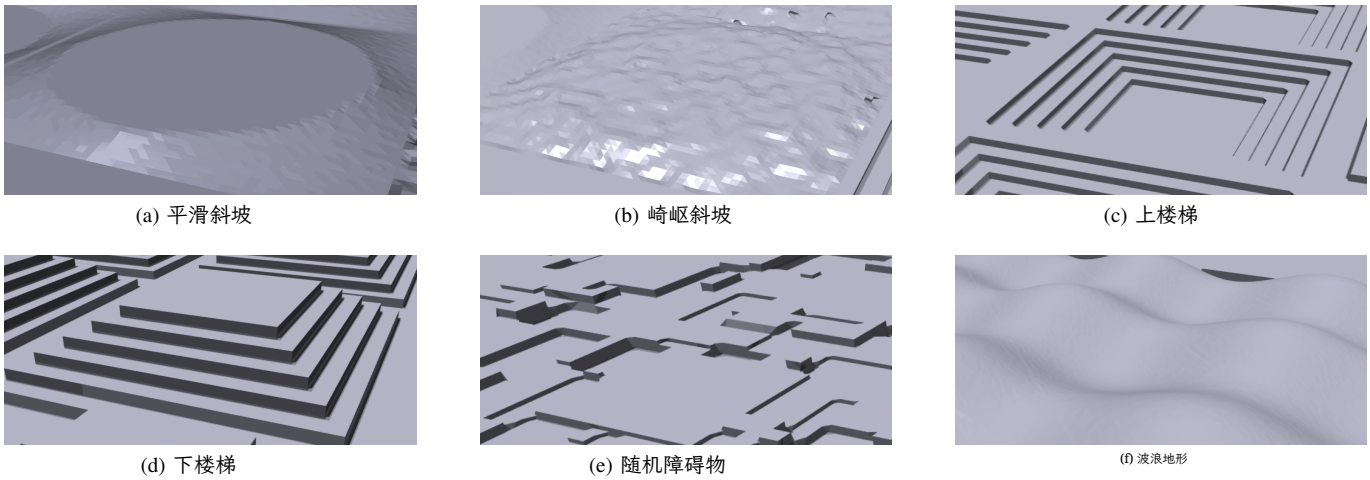
III. 强化学习框架

本节介绍了采用深度强化学习在仿真环境中训练四足机器人的方法。首先给出问题的基本定义，随后详细说明观测空间和动作空间、学习算法以及其他具体参数。

A. 定义

在强化学习设定中，智能体的决策过程被建模为马尔可夫决策过程（MDP）。MDP 是由状态集合 $\mathcal{S}$、$\mathcal{A}$、$\mathcal{P}$、$\mathcal{R}$、动作集合 $\mathcal{S}$（即决策空间）、转移函数 $\mathcal{A}$ 以及奖励函数 $\mathcal{P}$ (a_t、 $\mathcal{P}$) 组成的元组。转移函数描述了在时间步 t 选择动作 s_{t+1} 如何影响下一时间步 $s_{t+1}|s_t$ 的状态，通常由分布 (、s_t、a_t) 给出；奖励函数 $\mathcal{R}$(s_t、a_t、s_{t+1}) 则表示动作的价值。

强化学习智能体的目标是在一个时间范围内最大化总期望奖励。在深度强化学习中，通常通过寻找一个策略 π_θ : $[\mathcal{S}, \mathcal{A}] \rightarrow [0, 1]$ 参数化



(a) 平滑斜坡 (b) 崎岖斜坡 (c) 上楼梯

(d) 下楼梯 (e) 随机障碍物 (f) 波浪地形

Fig. 2：本文所考虑的典型地形。机器人仅在地形 (a)–(e) 上进行训练 .

由参数向量 θ 表示。最优参数 θ^* 由强化学习算法确定，该算法会最大化按照策略 π_θ 行动所产生轨迹的期望奖励，具体如下：

$$\theta^* = \arg\max_{\theta} \sum_{t=1}^{T-1} \mathbb{E}_{\{(s,a)\}_t \sim p_{\pi_\theta}(s_t,a_t)} \mathcal{R}(s_t, a_t, s_{t+1}) \quad (1)$$

B. 动作与观测空间

1)动作空间：如前所述，我们对波士顿动力公司的 Spot 机器人进行了改造，在每条腿的膝部和足部之间增加了线性执行器。现在每条腿有 4 个驱动关节。我们选择直接学习这 16 个关节的目标值，再由执行器级控制器对其进行跟踪。

2) 观测空间：我们的观测空间包括外部指令，指定所需的基座在*xy*方向上的速度以及绕*z*轴的角速度；所有驱动关节的位置和速度；所有驱动关节的先前动作；在机器人基座坐标系中表示的重力向量；以及同样在基座坐标系中表示的基座线速度和角速度。由于本文不考虑盲态运动，我们会对机器人当前位置周围测得的地形高度进行网格采样。本文使用一个 21×11 网格，尺寸为 $10cm$ 。观测空间总计为 291 维。

C. 奖励

我们定义奖励函数，以鼓励机器人保持特定的目标基座高度和朝向（仅允许偏航 运动），并跟随指令给定的 *xy* 线速度和角速度。机器人基座的不良运动，即垂直速度以及沿机器人*x* 轴和*y* 轴的角速度，都会受到惩罚。通过惩罚执行器扭矩和执行器目标的变化来鼓励运动平滑。机器人上、下腿与地形的碰撞】【。足部与地形发生碰撞时会受到惩罚，而机器人基座与地形发生碰撞则被视为终止条件，并重置环境。与 [25], 类似，我们还添加了一项奖励项，以鼓励每只足部与地形相邻两次接触之间保持更长的时间间隔。

D. 学习算法与策略网络

我们采用近端策略优化算法 [26]，该算法被认为是连续状态和动作空间中首选的在策略算法。该算法会训练两个网络：由 θ 参数化的策略网络 π_θ ，以及由 ϕ 参数化的评论家网络V_ϕ 。在我们的工作中，我们假设 π_θ 和 V_ϕ 神经网络都是全连接的，且规模相同。

在每次学习迭代中，当前策略都会针对 N 个回合进行展开，每个回合最多运行 T 个时间步。展开过程中收集的经验用于更新评论家网络 V_ϕ 和策略/行动者网络 π_θ。与原始策略梯度算法不同，PPO 通过求解无约束优化问题，确保对更新后的策略与旧策略之间较大的 KL 散度进行惩罚。表 I 给出了训练参数、PPO 超参数以及神经网络的维度。

E. 训练方法论

我们使用 NVIDIA 的 Isaac Gym [27] 与 PhysX 作为仿真引擎，以利用其并行仿真能力。地形设置为由 8 m×8m 瓦片组成的网格，其中每个瓦片属于五种特征类型之一 [25]。每种地形的难度沿行递增。在每次训练迭代中，*xy* 线速度和 z 角速度分别从 $[v_x^{min}, v_x^{max}]$、$[v_y^{min}, v_y^{max}]$ 和 $[\omega_z^{min}, \omega_z^{max}]$ 中均匀采样。为确保训练稳定性并防止灾难性遗忘，我们采用地形课程[14]：预先为每个机器人分配一种地形类型及其难度级别，并根据其使用当前策略时的表现，将其晋级到所分配地形类型的下一级，或降级到较低级别。对于

训练，我们并行模拟 $n = 2048$ 个机器人，这意味着在每次学习迭代中，当前策略会同时运行 $N = n = 2048$ 个回合。

表 I：PPO 超参数与神经网络结构

机器人数量（$n = N$）	2048
批次大小（$N \times T$）	2048×24
小批次大小	2048×6
学习轮数	5
熵系数	0.01
折扣因子	0.99
优势估计器折扣因子	0.95
目标 KL 散度	0.01
每个隐藏层的单元数	[512,256,128]
激活	指数线性单元
学习率	自适应

为增强学习策略抵御外部扰动的能力，每隔 15 秒，机器人基座都会被随机方向的外力推动，推力最高达 $1m/s$ 。此外，还会在关节和基座位置及速度的测量值中加入噪声。为实现仿真到现实的迁移，每个回合都会在 $0.3 - 1.25$ 范围内随机调整地形的摩擦系数，并将机器人基座质量扰动 ± 5 千克。

IV. 结果与讨论

A. 训练

训练是在一台配备 Intel Core-i9 处理器、32GB RAM 和 6GB Nvidia A3000 GPU 的机器上完成的。

为便于比较，我们也采用前文所述的相同方法，为传统四足形态训练了一个策略（由于执行器较少，观测空间更小）。两种形态结构在训练过程中的平均总奖励如图 3 所示。请注意，此处的平均值是对每次学习迭代中用于收集经验数据的 2048 个机器人仿真实例取平均所得。

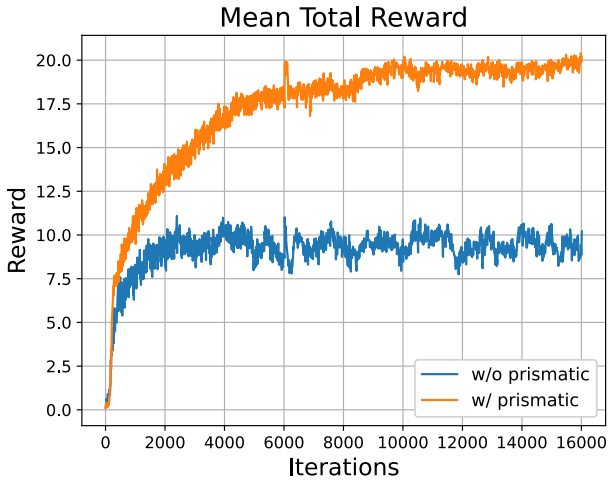


图 3：训练期间的平均总奖励

奖励图显示，采用此处使用的训练方法后，配备额外棱柱关节的智能体能够获得显著更高的奖励。

如图 4 所示，用于跟踪速度指令的奖励组成部分表明，配备额外棱柱关节的智能体所学习的策略平均而言更擅长跟踪所指令的基座速度。我们使用指数函数 $f(v - v^*) = \exp(\frac{-\|v-v^*\|^2}{\sigma^2})$ 作为速度跟踪奖励，因此完美跟踪对应的奖励为 1。

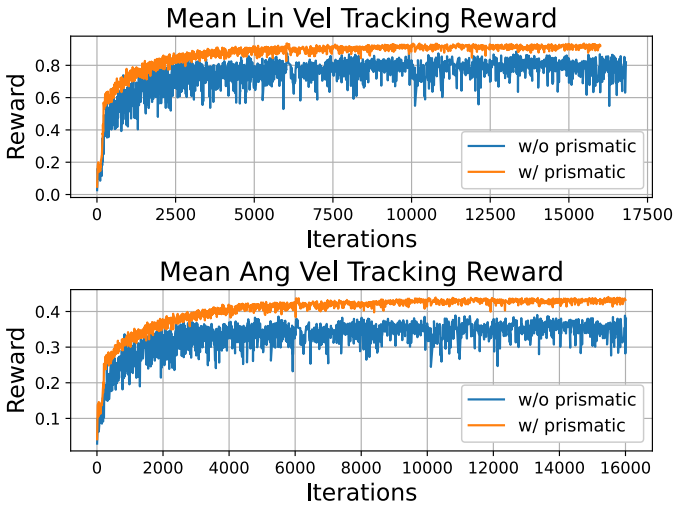


图 4：训练期间的平均速度跟踪奖励

B. 评估

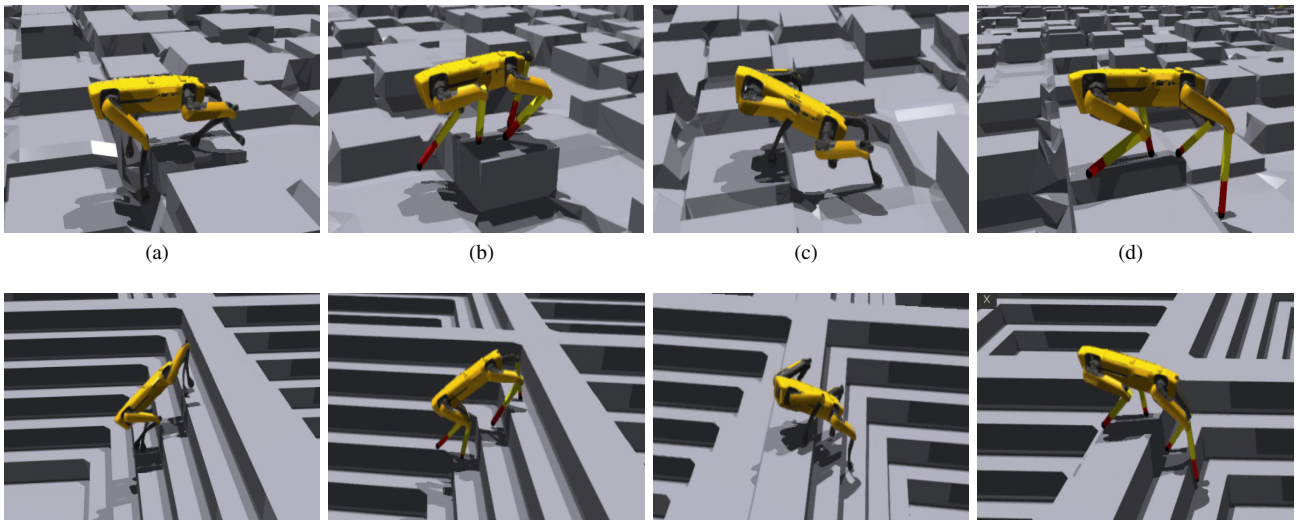
我们根据下述无量纲运输成本指标，对两种形态在不同地形上的表现进行比较：

$$CoT = \sum_{actuators} \frac{[\tau\dot{\theta}]_i^+}{MgV_b} \tag{2}$$

其中，$[\tau\dot{\theta}]_i^+$ 是 i号执行器输出的正机械功率，M 是总质量，g 和 V_b 分别是重力加速度和基座运动速度。

运输成本研究所考虑的地形包括第 II 节中提到的 5 种类型、如图 2f 所示的波浪地形以及平坦地形。实验通过在某种地形的 5×5 平铺区域中生成机器人来进行。每次实验持续 20 秒。前 10 秒，机器人接收一组随机的线速度和角速度指令，并在这段时间内保持不变。10 秒时，机器人接收另一组速度指令，并保持不变直至实验结束。实验中使用的地形参数如下：

1) 上楼梯： 楼梯宽度：0.3米，楼梯高度：0.35米 2) 下楼梯： 楼梯宽度：0.3米，楼梯高度：0.35米 3) 平滑斜坡： 坡度：0.4弧度 4) 粗糙斜坡： 坡度：0.4弧度，高度在 [0.05- 0.15]米之间均匀扰动 5)随机障碍物： 最大障碍物高度：0.35米，障碍物尺寸在 [0.1- 0.8]米之间



(a) (b) (c) (d)

(e) (f) (g) (h)

图 5：四足机器人利用棱柱连杆，在不同地形上迈步上下。

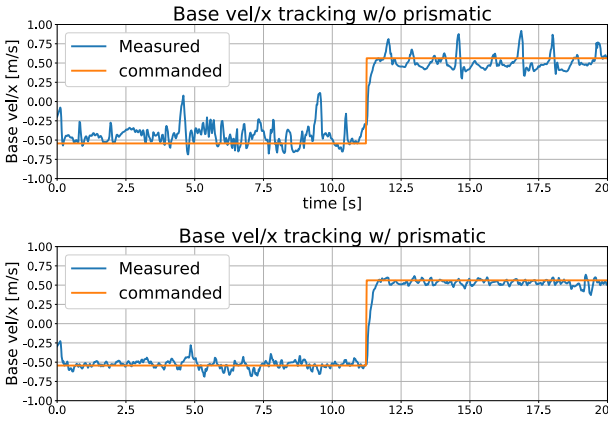


图 6：测得的和指令的 x 方向基座线速度

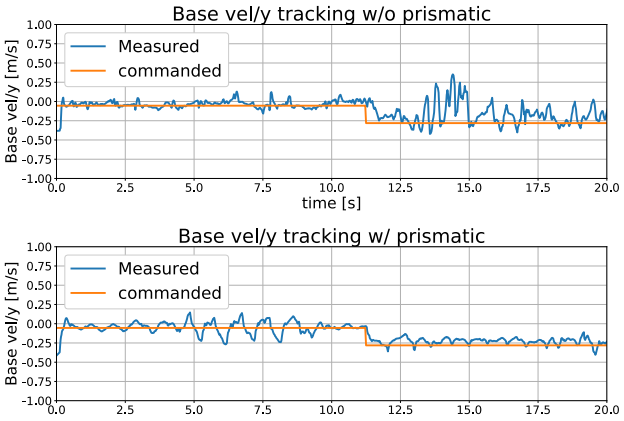


图 7：y 方向测得的和指令给定的基座线速度

6) 波浪地形： 波数：3，振幅：0.35 米

这些数值对应于第 II 节中提到的最高地形等级，但波浪地形地形除外，其参数是任意选取的。

详细结果见表 II。显然，使用棱柱关节显著降低了 *Stair*-中的运输成本

表 II：五次 20 秒运行的平均运输成本

地形	运输成本	
	不带棱柱形关节	带棱柱形的
上楼梯	5.06	1.84
下楼梯	3.33	1.60
平滑斜坡	0.86	1.04
粗糙斜坡	2.74	1.90
随机障碍物	2.88	1.59
Wavy	1.17	1.40
Flat	0.52	0.72

上坡、下楼梯 和随机障碍物 地形。这可以通过图5c–5h来解释，这些图展示了在类似情况下，两种形态结构的学习策略所采取的动作。没有棱柱关节时，机器人必须大幅活动髋关节，才能找到合适的足部支撑姿态。由于髋关节的折算惯量较大，这会导致机械功耗更高。另一方面，配备棱柱关节时，策略会学会让髋关节进行相对较小的运动，而足部的伸展则有助于找到可行的足部支撑姿态。在平滑斜坡 地形上，这种差异不太明显；而在平坦 地形上，结果则转而不利于使用棱柱关节。这可能是因为在较为平缓的环境中，额外的4个关节所提供的额外自由度可能派不上用场，反而增加了运输成本。在评估训练后的策略时，新增关节还有助于四足机器人保持期望高度（见图8）。在

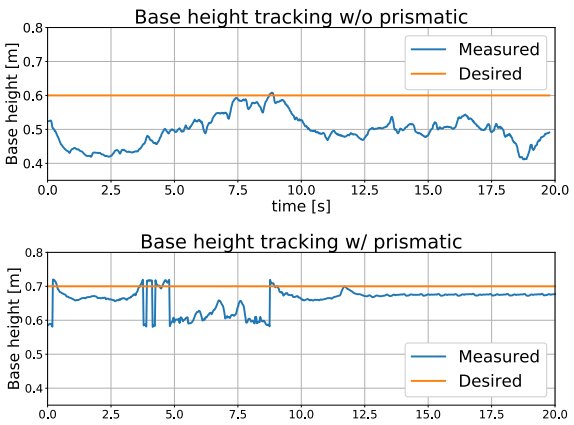


图 8：测得的基座高度与期望基座高度

四足机器人在给定地形上行进的同时，维持给定的线速度控制（见图6和图7）。

五、结论与未来工作

所 प्रस्तुत的结果表明，将棱柱关节纳入四足机器人的设计具有诸多益处。具体而言，配备棱柱关节的机器人获得了更高的奖励，并在保持指令速度和期望机身高度的同时，展现出更好的复杂地形穿越能力；与更为传统的设计相比，其运输成本更低。然而，本研究所采用的深度强化学习方法可能存在一个潜在缺点：即使在不需要的情况下，棱柱关节也可能被激活，从而导致运输成本增加。未来的工作可以重点开发更受控的驱动策略，以确保仅在必要时激活棱柱关节。此外，还可以探索将学习策略迁移至与感知系统耦合的硬件上，作为本研究的延伸。

参考文献

[1] R. J. Griffin、G. Wiedebach、S. McCrory、S. Bertrand、I. Lee 和 J. Pratt，“崎岖地形自主行走的落足点规划”，载于*2019*年*IEEE-RAS*第*19*届人形机器人国际会议（*Humanoids*）。IEEE，2019年，第9–16页。[2] G. Bledt、P. M. Wensing、S. Ingersoll 和 S. Kim，“非结构化地形中事件驱动运动的接触模型融合”，载于*2018*年*IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2018年，第4399–4406页。[3] M. Focchi、V. Barasuol、M. Frigerio、D. G. Caldwell 和 C. Semini，“四足机器人的滑移检测与恢复”，《机器人研究：第*2*卷》，第185–199页，2018年。[4] M. Camurri、M. Fallon、S. Bazeille、A. Radulescu、V. Barasuol、D. G. Caldwell 和 C. Semini，“用于四足机器人状态估计的概率接触估计与冲击检测”，*IEEE*机器人与自动化快报，第2卷，第2期，第1023–1030页，2017年。[5] M. Kalakrishnan、J. Buchli、P. Pastor、M. Mistry 和 S. Schaal，“四足机器人在复杂地形上的快速、鲁棒运动”，载于*2010*年*IEEE*国际机器人与自动化会议。IEEE，2010年，第2665–2670页。[6] J. Z. Kolter、M. P. Rodgers 和 A. Y. Ng，“四足机器人在崎岖地形上运动的控制架构”，载于*2008*年*IEEE*国际机器人与自动化会议。IEEE，2008年，第811–818页。[7] Y. Fukuoka、H. Kimura 和 A. H. Cohen，“基于生物学概念的四足机器人不规则地形自适应动态行走”，《国际机器人研究杂志》，第22卷，第3-4期，第187–202页，2003年。[8] L. Righetti 和 A. J. Ijspeert，“带感官反馈的模式发生器在四足机器人运动控制中的应用”，载于2008年IEEE国际机器人与自动化会议。IEEE，2008年，第819–824页。[9] D. J. Hyun、S. Seok、J. Lee 和 S. Kim，“高速小跑：在MIT Cheetah上采用本体感受阻抗控制实现分层控制器”，《国际机器人研究杂志》，第33卷，第11期，第1417–1445页，2014年。[10] S. Gay、J. Santos-Victor 和 A. Ijspeert，“利用神经网络作为中央模式发生器的感官反馈函数，学习机器人步态稳定性”，载于2013年IEEE/RSJ智能机器人与系统国际会议。IEEE，2013年，第194–201页。[11] R. Grandia、F. Jenelten、S. Yang、F. Farshidian 和 M. Hutter，“通过非线性模型预测控制实现感知运动”，arXiv预印本arXiv:2208.08373，2022年。[12] J. Hwangbo、J. Lee、A. Dosovitskiy、D. Bellicoso、V. Tsounis、V. Koltun 和 M. Hutter，“为腿式机器人学习敏捷动态运动技能”，《科学·机器人》，第4卷，第26期，文章编号eaau5872，2019年。[13] A. Kumar、Z. Fu、D. Pathak 和 J. Malik，“RMA：腿式机器人的快速运动适应”，arXiv预印本arXiv:2107.04034，2021年。[14] J. Lee、J. Hwangbo、L. Wellhausen、V. Koltun 和 M. Hutter，“学习四足机器人在复杂地形上的运动”，《科学·机器人》，第5卷，第47期，文章编号eabc5986，2020年。[15] S. Chen、B. Zhang、M. W. Mueller、A. Rai 和 K. Sreenath，“学习四足机器人运动的力矩控制”，arXiv预印本arXiv:2203.05194，2022年。[16] H. Shi、B. Zhou、H. Zeng、F. Wang、Y. Dong、J. Li、K. Wang、H. Tian 和 M. Q.-H. Meng，“结合进化轨迹生成器的强化学习：一种通用的四足机器人运动方法”，IEEE机器人与自动化快报，第7卷，第2期，第3085–3092页，2022年。[17] M. Kalakrishnan、J. Buchli、P. Pastor、M. Mistry 和 S. Schaal，“四足机器人在复杂地形上运动的学习、规划与控制”，《国际机器人研究杂志》，第30卷，第2期，第236–258页，2011年。[18] O. A. V. Magana、V. Barasuol、M. Camurri、L. Franceschi、M. Focchi、M. Pontil、D. G. Caldwell 和 C. Semini，“通过卷积神经网络实现动态运动中的快速连续落足点适应”，IEEE机器人与自动化快报，第4卷，第2期，第2140–2147页，2019年。[19] T. Klamt 和 S. Behnke，“探索学习用于高维状态空间运动规划的抽象表示”，载于2019年国际机器人与自动化会议（ICRA）。IEEE，2019年，第922–928页。[20] N. Heess、D. TB、S. Sriram、J. Lemmon、J. Merel、G. Wayne、Y. Tassa、T. Erez、Z. Wang、S. Eslami 等，“丰富环境中运动行为的涌现”，arXiv预印本arXiv:1707.02286，2017年。[21] X. B. Peng、G. Berseth、K. Yin 和 M. Van De Panne，“DeepLoco：采用分层深度强化学习的动态运动技能”，《ACM图形学汇刊》（TOG），第36卷，第4期，第1–13页，2017年。[22] T. Miki、J. Lee、J. Hwangbo、L. Wellhausen、V. Koltun 和 M. Hutter，“学习四足机器人在野外环境中的鲁棒感知运动”，《科学·机器人》，第7卷，第62期，文章编号eabk2822，2022年。[23] E. Wijmans、A. Kadian、A. Morcos、S. Lee、I. Essa、D. Parikh、M. Savva 和 D. Batra，“DD-PPO：利用25亿帧学习近乎完美的点目标导航器”，arXiv预印本arXiv:1911.00357，2019年。[24] A. Spr¨owitz、A. Tuleu、M. Vespignani、M. Ajallooeian、E. Badri 和 A. J. Ijspeert，“迈向动态小跑步态运动：与Cheetah-Cub柔顺四足机器人的设计、控制和实验”，《国际机器人研究杂志》，第32卷，第8期，第932–950页，2013年。[25] N. Rudin、D. Hoeller、P. Reist 和 M. Hutter，“利用大规模并行深度强化学习，在几分钟内学会行走”，ArXiv，第abs/2109.11978卷，2021年。[26] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，“近端策略优化算法”，arXiv预印本arXiv:1707 06347，2017年。[27] V. Makoviychuk、L. Wawrzyniak、Y. Guo、M. Lu、K. Storey、M. Macklin、D. Hoeller、N. Rudin、A. Allshire、A. Handa 等，“Isaac Gym：用于机器人学习的高性能基于GPU的物理仿真”，arXiv预印本arXiv:2108.10470，2021年。