盲式双足楼梯穿越：通过仿真到现实的强化学习

November 27, 2024

Abstract

对于现实世界环境中的机器人运动来说，准确和精确的地形估计是一个难题。因此，拥有不依赖于精确估计到脆弱程度的系统是有用的。在本文中，我们通过研究双足机器人在没有任何外部感知或地形模型的情况下穿越类似楼梯地形的问题，探索了这种方法的极限。对于这种盲视双足平台来说，由于意外的高程变化，问题显得困难（甚至对人类来说也是如此）。我们的主要贡献是展示，仿真到现实的强化学习（RL）能够在双足机器人 Cassie 上仅使用本体感觉反馈实现对类似楼梯地形的稳健运动。重要的是，这只要求修改现有的平地训练 RL 框架，以包括类似楼梯地形的随机化，而无需更改奖励函数。据我们所知，这是第一个能够仅使用本体感觉可靠地穿越各种现实世界楼梯和其他类似楼梯干扰的双足人形机器人控制器。

1 介绍

为了在现实世界中发挥作用，双足和仿人机器人需要能够上下楼梯和类似楼梯的地形，如凸起的平台或突然的垂直下降，这些是人类中心环境中的常见特征。能够稳健地导航这些环境对于确保机器人安全地与人类一起工作至关重要。在双足平台上实现这种稳健性并非易事；虽然其他平台如四足机器人由于与地面有多个接触点而具有固有的稳定性，并且能够像桌子一样停下来站立，但像 Cassie 这样的双足机器人完全依赖于动态稳定性（基本上总是处于跌倒状态）。在类似楼梯的环境中，这一点尤为明显，因为用仅有的两条腿从失误中恢复的难度很大。相比之下，具有四足形态的机器人已经能够仅使用本体感觉来处理楼梯 [1, 2]，而六足机器人甚至能够使用开环控制来上下楼梯 [3]。虽然平面双足机器人已被证明能够排除像大型意外的下落台阶这样的干扰 [4]，但在现实世界中使这类机器人能够处理楼梯的绝大多数方法都需要要么准确的视觉系统 [5, 6, 7]，要么在精心控制的实验室环境中操作 [8, 9, 10]，这意味着机器人通过已知的起始位置定位，或者楼梯与机器人形态同时设计。然而，机器人必须能够在受控实验室条件之外操作，并处理现实世界中巨大的条件变化。这个目标与完全依赖外部感知传感器（如 RGB 和深度摄像头）进行准确地形估计并不兼容，因为这会在现实世界条件下引入脆弱性 [11]。例如，如果摄像头受到遮挡、雾霾或光照变化的影响，可能会变得不可靠。此外，将最先进的计算机视觉系统集成到高速控制器中在技术上是困难的，尤其是在像移动机器人这样的计算能力有限的平台上。出于实际考虑，基础控制器应尽可能稳健，同时尽量少依赖于对世界的信息。理想情况下，双足机器人应能够尽可能多地使用本体感觉穿越人类环境的广泛区域，同时依赖外部感知进行进一步的效率提升和高层次规划（并对错误的感知具有鲁棒性）。这就引出了一个问题：盲视双足机器人能有多稳健？基于强化学习（RL）的方法已经开始在稳健的现实世界腿部运动中显示出显著的潜力 [1, 12, 13]。与依赖预定地面接触计划或基于力的事件检测的优化或启发式控制方法不同，RL 可以产生学习本体感觉反射和策略的控制策略，以应对训练期间暴露的各种干扰中的意外早接触或晚接触和粗糙地形。然而，这种方法的局限性尚不清楚，先前的工作尚未在涉及类似楼梯地形的干扰规模和多样性上得到证明。在这项工作中，我们展示了通过现有 RL 框架，几乎不需要修改就可以学习到复杂类似楼梯地形的稳健本体感觉双足控制。特

别是，唯一需要调整的是训练期间使用的地形随机化，我们定义了一个分布，包括上升和下降楼梯的变化，包括接触平面的高度、宽度和坡度。在这个分布上学习，可以允许在未知楼梯上进行盲视运动，以及处理更一般的类似楼梯地形特征，例如原木、路缘、悬崖等。学习到的控制器在模拟和各种现实世界环境中得到演示。据我们所知，这是第一次进行此类演示，并建议继续探索稳健本体感觉双足控制的极限。

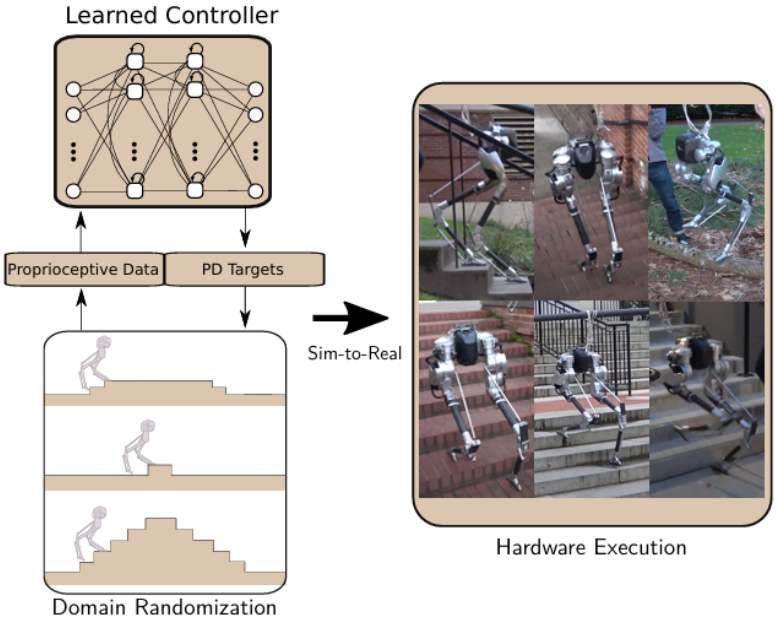


Figure 1: 在这项研究中，我们探索了盲式双足运动的极限。我们提出了一个训练流程，能够产生能够在现实世界中盲目上下楼梯的策略。这些策略学习了基于本体感受的反射动作，以抵御地面高度的重大干扰，从而在许多现实世界环境中展现出极高的鲁棒性。

2 强化学习公式

我们遵循仿真到现实的强化学习（RL）方法来学习双足运动，并假设对 RL 有基本的了解。在一般的 RL 设置中，每个离散时间步 t，机器人控制策略 π 接收当前状态 s_t 并返回动作 a_t，该动作被应用并导致过渡到下一个状态 s_{t+1}。状态转移动态对机器人来说是未知的，并且由环境条件（如地形类型）和机器人动态的组合所决定。此外，在训练期间，每个状态转移都与一个实值奖励 r_t 相关联。奖励由应用目标所决定，以鼓励在学习期间产生期望的行为。本文考虑的 RL 优化目标是通过与环境的交互来学习一个策略，该策略在有限视界 T 上最大化预期累积折扣奖励。也就是说，找到一个策略 π，使其最大化：$J(\pi) = E\left[\sum_{t=0}^{T} \gamma^t R_t\right]$，其中 $\gamma \in [0, 1]$ 是折扣因子，R_t 是从初始状态分布中抽取的状态，遵循 π 在时间 t 的奖励的随机变量。

对于复杂环境，强化学习通常需要大量的训练经验来识别一个好的策略。此外，对于双足运动，训练将涉及许多跌倒和碰撞，尤其是在训练初期。因此，在现实世界中从头开始训练是不切实际的，我们转而采用仿真到现实的强化学习范式。训练完全在模拟环境中进行，包括动态随机化（见下文），然后将得到的政策应用于现实世界。

在本节的其余部分，我们详细说明了本工作中使用的特定仿真到现实强化学习公式，该公式遵循了最近关于在平坦地形上学习不同双足步态的工作 [13]。令人惊讶的是，为了使策略学习适应本文中更复杂的类似楼梯地形，只需要进行最小的更改。特别是，唯一需要的主要修改是在第三部分后面讨论的，将类似楼梯的地形而不是大多数平坦地形的随机领域生成；不需要新的特定楼梯的奖

励项。

2.1 状态空间

每个时间步输入到控制策略的状态 s_t 包括三个主要组成部分。首先，状态包含关于机器人瞬时物理状态的信息，包括骨盆的方向（以四元数格式表示）、骨盆的角速度、关节位置和关节速度。s_t 的第二个组成部分由人类操作员发出的命令输入组成。这些命令在训练期间会进行随机化，以使策略在尝试以各种速度和接近角度穿越楼梯时获得广泛的经验。这种随机化的详细信息可以在表 I 中看到。

命令	**变化概率**	**范围**
前进速度	1/300	[-0.3m/s, 1.5m/s]
横向速度	1/300	[-0.3m/s, 0.3m/s]
转向速率	1/300	[-90deg/s, 90deg/s]

Table 1: 在每个时间步，输入到策略的每个命令都有 1/300 的概率被改变。当这种情况发生时，新的命令会从最右列参数化的均匀分布中采样。考虑到最大情节长度为 300 个离散时间步，这意味着每个命令平均每个情节至少会改变一次。

第三个组成部分包括两个循环时钟输入，每个输入对应机器人的一条腿，p：

$$p = \begin{cases} \sin(2\pi(\phi_t + 0.0)) \\ \sin(2\pi(\phi_t + 0.5)) \end{cases} \tag{1}$$

这里 ϕ_t 是一个相位变量，它从 0 增加到 1，然后回滚到 0，以跟踪步态的当前相位。常数偏移量 0.0 和 0.5 是相位偏移，用于确保在运动过程中左右腿在相位上始终是直径相对的。

2.2 动作空间

控制策略在每个时间步（以 40Hz 运行）的输出动作 a_t 是一个 11 维向量，其中前 10 个条目对应于关节的 PD 目标，每个目标都输入到一个以 2KHz 运行的 PD 控制器中。先前的研究已经发现，在 PD 目标空间中学习动作比直接学习更高频率的激活命令更为有利 [15]。

a_t 的最后一个维度是一个时钟增量 δ_t（有关时钟的信息，请参见 II-A），它允许策略调节步态的步进频率。直观上，这允许控制器为特定的步态、命令和地形选择适当的步进频率。具体来说，状态表示中的相位变量 ϕ（见 II-A 节）在每个时间步 t 通过以下方式更新：

$$\phi_{t+1} = fmod(\phi_t + \delta_t, 1.0). \tag{2}$$

这个增量被限定在一个范围内，使得策略可以选择将步态周期调节在名义步进频率的 0.5 倍到 1.5 倍之间（大约每 0.7 秒一个步态周期）。虽然这个组件包含在控制策略动作中，但它似乎对性能没有太大影响，并且学习到的策略在响应扰动时也不会太多地改变 δ_t。我们怀疑未来的消融分析将显示它对于真实机器人的性能并不重要。

2.3 奖励函数

我们使用文献 [13] 中介绍的方法来指定我们的奖励函数。简要回顾这种方法，我们希望有一个奖励框架，允许在某些情况下对策略进行惩罚，因为环境中的一些量在某些时候会很大，而在其他

时候允许这些量很大。我们将脚力和脚速度指定为两个这样的量；惩罚脚力激励策略抬起脚，而惩罚脚速度激励策略将脚放置。我们在这些基础奖励项的基础上增加了额外的成本项，包括激励策略匹配平移速度和方向的成本。我们还采用成本来鼓励平滑动作、提高能效和减少骨盆晃动。有关所用奖励函数的详细解释，请参见附录。正如文献 [13] 中所述，我们不依赖专家参考轨迹来学习行为。

参数	**单位**	**范围**
关节阻尼	Nms/rad	$[0.5, 3.5] \times$
关节质量	kg	$[0.5, 1.7] \times$
地面摩擦	–	$[0.5, 1.1]$
关节编码器偏移	rad	$[-0.05, 0.05]$
执行速率	Hz	$[37, 42]$

Table 2: 为了防止对模拟动力学的过拟合并促进平滑的仿真到现实转移，我们采用动力学随机化。上述范围为每个列出的参数参数化了一个均匀分布。阻尼、质量、摩擦和编码器偏移在每个 rollout 开始时随机化，而执行速率在每个时间步随机化，以模拟真实机器人上可变系统延迟的效果。

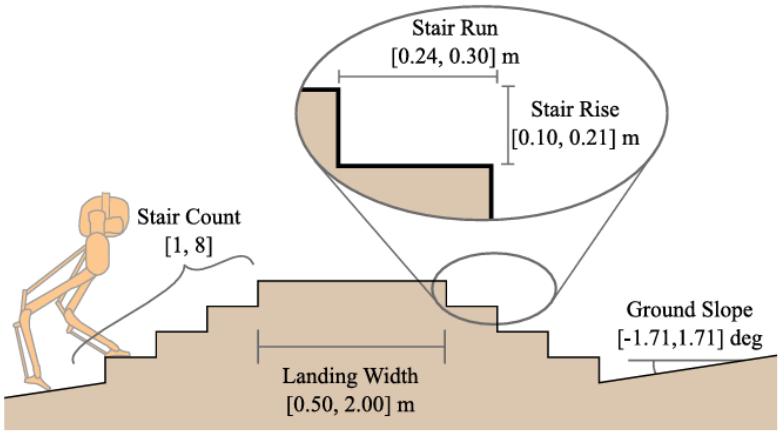


Figure 2: 为了确保在各种可能的类似楼梯地形上的鲁棒性，我们在模拟中每个场景开始时随机化生成楼梯的多个参数。这些参数包括楼梯的数量、每个楼梯的高度、每个楼梯的长度、楼梯顶部平台的长度，以及楼梯前后地面的坡度。

2.4 动力学随机化

为了克服在我们模拟的 Cassie 环境中可能存在的任何建模误差，我们在训练的每个情节开始时随机化几个重要的动力学量，如先前工作 [16][13] 中所述。这些随机化参数列在表 II 中。

2.5 策略表示与学习

我们将控制策略表示为一个 LSTM 循环神经网络 [17]，具有两个维度为 128 的循环隐藏层。我们选择使用一个具有记忆功能的网络，因为先前的工作表明它在处理部分可观测环境方面有更高的熟练度 [18][16][19]。对于消融实验，涉及非记忆基础的控制策略，我们使用一个具有两个维度为 300 的标准前馈神经网络，采用 tanh 激活函数，使得参数数量大约等于 LSTM 网络的参数数量。

对于策略的模拟到真实训练，我们使用近端策略优化（PPO）[20]，这是一种无模型的深度强化学习算法。具体来说，我们使用一个 KL 阈值终止变体，其中每次策略更新时，都会计算更新后的

策略与前一个策略之间的 KL 散度，如果散度过大，则中止更新。在训练过程中，我们使用镜像损失项 [21] 以确保控制策略不会学习到不对称的步态。对于循环策略，我们从重放缓冲区中采样一批情节，如 [19] 所述，而对于前馈策略，我们采样一批时间步。每个情节限制为 300 个时间步，这对应于大约 7.5 秒的模拟时间。

3 地形随机化

先前将强化学习（RL）应用于 Cassie 的工作要么在平坦地面上进行训练 [12][19]，要么在随机轻微倾斜的地面上进行训练 [13]。其他应用深度强化学习的研究探讨了采用逐渐增加难度的粗糙地形课程 [1]。为了简化目的，我们发现在没有课程的情况下，通过与随机楼梯的交互进行训练就足以学习到稳健的行为。

为此，我们在每个 rollout 开始时随机化倾角和横滚轴上的倾斜角度的平面上进行训练。这个倾斜角度在-0.03 弧度和 0.03 弧度之间。作为动态随机化的一部分，地面摩擦力也被随机化，增加了环境的潜在难度。楼梯的起始位置在每个 rollout 开始时随机化，这样可以使策略从楼梯顶部开始，或者楼梯在策略前方最多 10 米的位置开始。这样做是为了确保策略能够在平坦或倾斜的地面上，以及楼梯上获得丰富的经验。

楼梯的尺寸在典型的城市规范尺寸内随机化，每步的升高在 10 厘米到 21 厘米之间，水平距离在 24 厘米到 30 厘米之间。楼梯的数量也被随机化，使得每组楼梯有 1 到 8 个单独的台阶。在每个台阶的升高和水平距离上添加少量噪声（±1 厘米），使得楼梯永远不会完全均匀，以防止策略通过本体感知推断出楼梯的精确尺寸，进而过度拟合到完全均匀的楼梯上。

4 结果

我们训练了四组策略，每组包含五个用不同随机种子初始化的策略。首先，我们训练了一组简单的 LSTM 策略，这些策略具有楼梯地形随机化；在本节中，这些策略被称为 Stair LSTM。为了研究记忆的重要性，我们还训练了一组前馈策略，这些策略也具有楼梯地形随机化；我们将这些策略称为 Stair FF。我们还训练了一组没有楼梯地形随机化的策略，并将这些策略称为 Flat Ground LSTM，以研究本工作中引入的地形随机化的重要性。最后一组策略是用一个简单的额外二进制输入训练的，该输入告知策略一米范围内是否有楼梯，这些策略在这里被称为 Proximity LSTM，目的是为了研究向策略泄露关于世界的信息的好处。

每个策略都训练到从虚拟环境中采样了 3 亿个时间步，这些环境是用 MuJoCo[22] 模拟的。我们为 PPO 算法选择的超参数包括一个大小为 50,000 个时间步的回放缓冲区，对于循环策略是一个包含 64 个轨迹的批次大小，对于前馈策略是一个包含 1024 个时间步的批次大小。每个回放缓冲区最多采样五个 epoch，如果 KL 散度达到了最大允许阈值 0.02，则提前终止优化。我们在每次迭代开始时清除我们的回放缓冲区。我们使用 Adam[23] 优化器，对于演员和评论家都使用 0.0005 的学习率，它们是分别学习的，不共享参数。

4.1 仿真

1. *成功上下楼梯的概率*：为了理解记忆和地形随机化的重要性，我们评估了三组策略在成功爬升和下降一组楼梯的任务上的表现。在模拟中，我们比较了 Stair FF、Stair LSTM 和 Flat Ground LSTM 策略在这项任务上的表现。具体来说，我们进行了 150 次试验，测试策略能够成功爬上一组有五级台阶的楼梯的频率，每级台阶的踏步为 17 厘米，深度为 30 厘米（典型的现实世界

中相对平缓的楼梯几何形状）。这应该能给我们一个估计，即每组控制策略在盲目接近时可靠地爬上一段楼梯的能力。成功的定义是到达楼梯顶部而没有摔倒。我们同样对下楼梯的程序进行了 150 次试验，记录了每组策略能够到达底部而没有摔倒的比率。

图 4 显示了三种不同训练条件下的这些测试结果。我们注意到，Stair LSTM 策略具有最高的整体成功率。然而，成功率在很大程度上取决于接近速度。策略在低速时失败率较高，因为它们可能缺乏推动自己越过不良脚步放置的动量。在高速时，它们也经历了较高的失败率，可能是由于高速步态的动态性更强。Flat Ground LSTM 策略在训练期间从未见过类似楼梯的地形，无法补偿，并且在上升和下降时都经历了较高的失败率。尽管 Stair FF 策略在训练期间遇到了楼梯，但它们无法学习到有效处理楼梯的策略，这表明记忆可能是对楼梯状地形具有鲁棒性的重要机制。

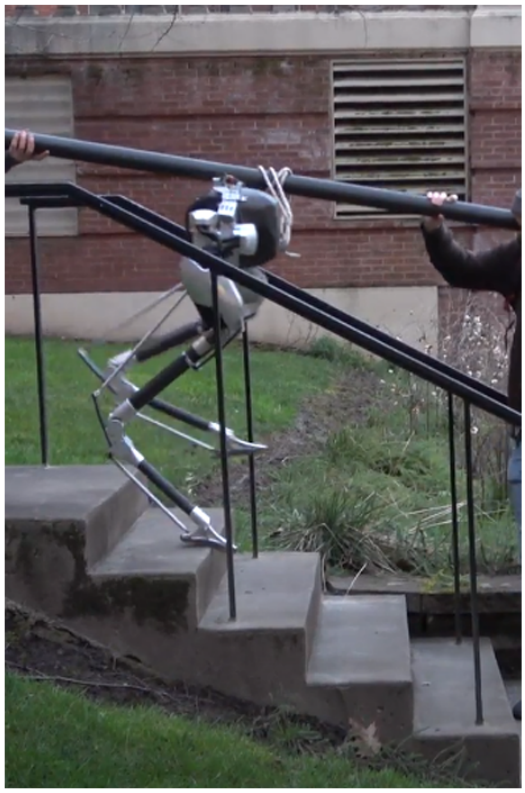


Figure 3: 学习到的策略展现出对各种类似楼梯地形的高度盲目鲁棒性，并且能够可靠地上下人类环境中常见的典型尺寸楼梯。

2. 能量效率比较：为了理解地形随机化训练的后果，我们还比较了 Flat Ground LSTM 策略、Stair LSTM 策略和 Proximity LSTM 策略之间的运输成本。运输成本（CoT）是腿式机器人、人类和动物效率的常见衡量标准。它是每单位距离使用的能量，按重量归一化以使单位无关。它定义为

$$CoT = \frac{E_m}{Mgd}, \tag{3}$$

其中 E_m 是电机使用的能量，M 是机器人的总质量，g 是重力加速度，d 是行进的距离。Cassie 使用的能量是通过正向执行器工作和通过

$$E_m = \int_0^T \left(\sum_i \max(\tau_i \cdot \omega_i, 0) + \frac{\omega_i^2}{p_i^{max}} \right) dt. \tag{4}$$

这里 τ_i 是施加在电机 i 上的扭矩，ω_i 是其旋转速度。我们使用两个参数来定义扭矩方面的电阻损耗，p_i^{max} 是电机 i 的最大输入功率，ω_i^{max} 是电机 i 的最大速度。在 1 m/s 的平坦地面上测试稳态 CoT 的结果可以在表 III 中看到。这些 CoT 的计算不包括来自计算和控制电子设备的额外功耗，因此它们不应该用来在机器人之间进行比较，而只应该在控制策略之间进行比较。

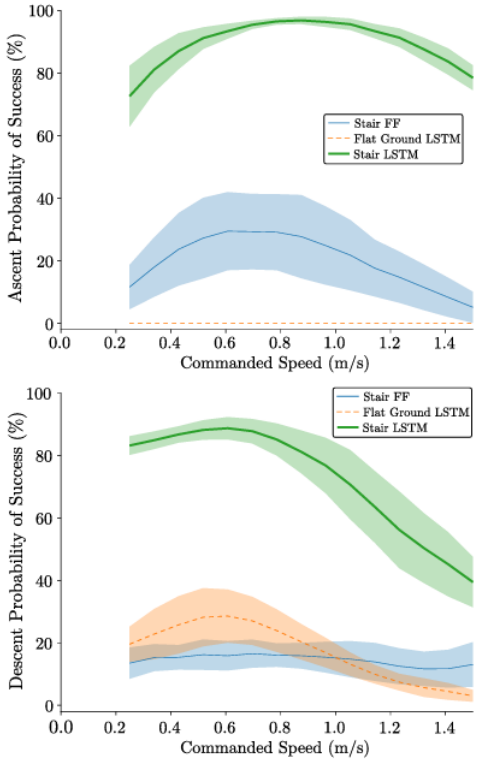


Figure 4: 我们评估了在 0.25 米/秒到 1.5 米/秒的命令速度下，成功上下楼梯而不跌倒的概率，共进行了 150 次试验。对于楼梯 LSTM 策略来说，似乎存在一个最佳的上楼梯速度和一个单独的最佳下楼梯速度。楼梯 FF（前馈）策略没有达到高性能，这意味着记忆可能是学习行为的一个重要组成部分。平地 LSTM 策略在训练中从未遇到过楼梯，几乎无法成功上楼梯，但在下楼梯时能够避免跌倒，取得了一定的成功。

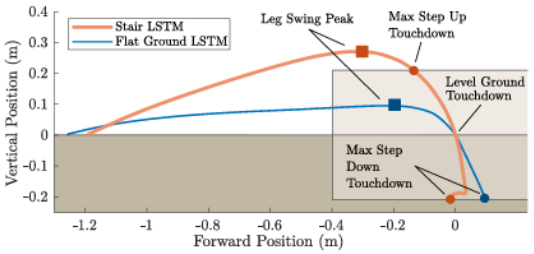
策略组	**平均 CoT**	**标准差 CoT**
Proximity LSTM (楼梯)	0.47	0.0086
Stair LSTM	0.46	0.0323
Proximity LSTM (平坦)	0.39	0.0257
Flat Ground LSTM	0.38	0.0205

Table 3: 在模拟中以 1 m/s 的速度在平坦地面上行走的运输成本（CoT），在所有五个随机种子上测试三组策略。我们注意到，未在楼梯地形随机化上训练的策略倾向于学习更节能的步态，尽管可以通过向楼梯训练的策略提供二进制楼梯存在/不存在输入来恢复一些能效。

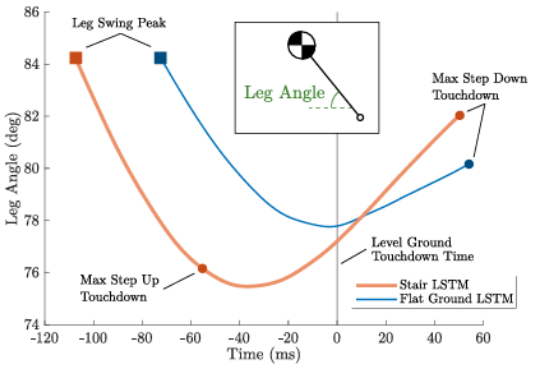
我们发现，Flat Ground LSTM 策略学习了在平坦地面上行走的最节能步态。Stair LSTM 策略学习了为了对楼梯具有鲁棒性而效率较低的平坦地面步态；然而，楼梯接近输入可以帮助 Proximity LSTM 恢复一些失去的能效，通过允许学习控制器在楼梯准备步态和更节能的平坦地面步态之间切换。

4.2 行为分析

为了理解策略所采用的策略，我们可以从实验生物学的角度受益。我们特别关注机器人在平坦地面上行走后首次接触台阶上或下的步态行为。首先，我们将分析摆动腿的运动，以了解机器人如何在台阶上或下放置其脚。一旦摆动脚接触到台阶上或下，站立阶段施加在地面上的力可以调节，以更好地为未来的步骤做准备。我们分析了在台阶上或下的情况下，地面反作用力和总冲量的变化。



(a) Swing foot paths for the stair trained policy and the flat ground policy overlaid on example step ups and step downs.



(b) The leg angle between the robot body and the swing foot as the foot descends toward touchdown.

Figure 5: 在以 1.0 米/秒的速度移动时，楼梯 LSTM 策略和平地 LSTM 策略的摆动脚运动进行了比较。由于在随机化楼梯上的训练，腿部摆动策略发生了显著变化。最显著的变化包括更高的脚部离地间隙、更陡的脚部下降角度以及更快的腿部角度回缩速度。

1. *摆动脚运动：*为了理解楼梯地形对脚摆动路径的影响，我们比较了平坦地面 LSTM 策略和楼梯 LSTM 策略在遇到台阶下降时的结果。下降步骤期间的脚摆动路径让我们看到，如果遇到台阶上升或下降，策略会将脚放在哪里。图 5a 显示了这两种策略相对于地面的脚摆动路径。我们可以看到，楼梯 LSTM 策略采取了更高的步幅，与平坦地面 LSTM 策略相比，它提供了额外的间隙，使其能够踏上大台阶。第二个有趣的观察是摆动脚的更陡峭路径。对于楼梯 LSTM 策略，摆动脚仅向前移动 14 厘米，而它处于可能遇到台阶上升的前缘的高度范围内。我们假设这是一种策略，防止脚在台阶的前缘上过于用力，导致机器人向前绊倒。

   第二种理解腿摆动运动的观点是观察腿摆动回缩。在人类和双足鸟类中，观察到摆动腿在站立结束时相对于身体向后摆动，朝向地面。这有助于减少脚相对于地面的速度，从而减少冲击，并通过自动改变腿部触地角度来改善地面高度干扰的排斥。

我们的训练程序没有明确激励机器人表现出这些腿摆动回缩行为，但我们确实看到它们在图 5b 中出现。这个图显示了腿摆动相对于身体的角度，从腿摆动的峰值到与最大步幅下降的接触。楼梯 LSTM 策略的腿摆动回缩速度比平坦地面 LSTM 策略更快。仅凭这些数据，我们无法说这种回缩曲线是否是最佳的，甚至是否是楼梯上改进性能的原因。然而，由于在楼梯训练中腿回缩曲线发生了显著变化，这是一个有趣的观察。

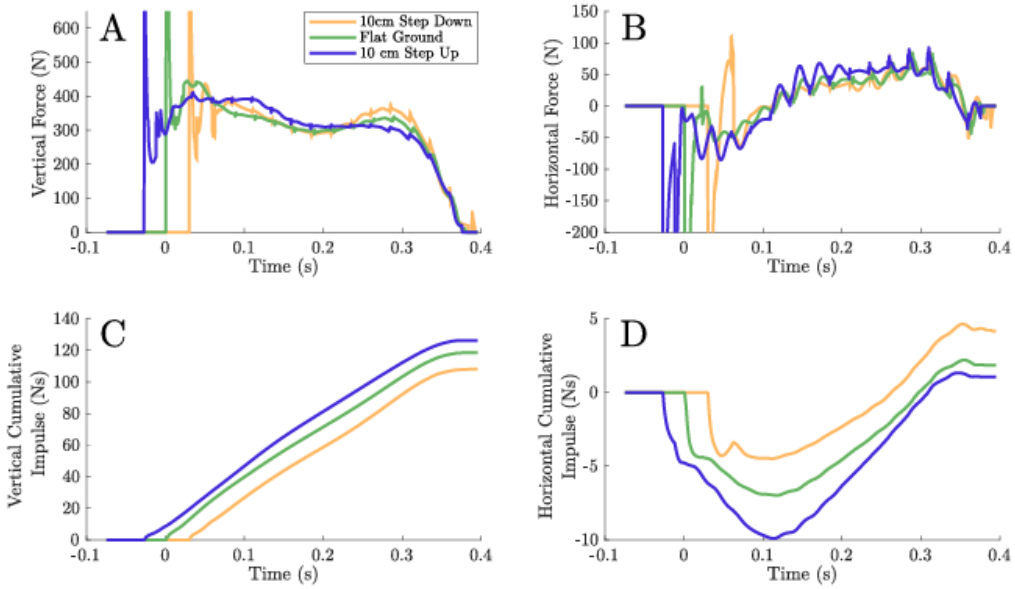


Figure 6: 楼梯 LSTM 策略在遇到不同地面高度时的地面反作用力和累积冲量如下所述：峰值垂直力（A）：在冲击后，峰值垂直力大致保持相等，而支撑阶段后半部分的力则有所调节。这表明在初次接触地面后，策略能够维持一个相对稳定的垂直力峰值，但在后续的支撑阶段会根据地面情况调整力的大小。水平力（B）：显示出与学习策略执行频率相匹配的振荡。这可能是策略在控制身体姿态，通过调整水平力来维持平衡。总垂直冲量（C）：呈现出预期的结果，即上楼梯时有更大的垂直冲量，下楼梯时有更小的垂直冲量。这与我们对上下楼动作的直观理解相符，即上楼需要更大的力来克服重力，而下楼则需要较小的力。水平冲量（D）：显示出由腿部摆动回缩预测的结果。在下楼梯时，脚相对于身体向后移动，这导致了一个净的向前加速度，在这里通过正的水平冲量表现出来。这意味着策略通过调整脚步的位置和摆动来控制身体的前进速度和稳定性，确保在不同地面高度变化时能够有效地控制身体的运动。

2. *地面反作用力：*一旦机器人的脚触地，由于双足运动的欠驱动特性，其控制能力受到限制。然而，机器人仍然通过地面反作用力有相当大的控制能力。为了了解楼梯 LSTM 策略如何对 10 厘米的台阶上升或下降做出反应，我们在图 6 中绘制了水平和垂直地面反作用力在矢状平面上的变化。在站立开始时，有一个大的力峰值，使正常力在站立期间变矮。这个峰值主要由模拟接触模型的调整定义，因此它不是理解策略行为的主要兴趣点。第一个有趣的观察是，在子图 A 中，我们看到最大名义腿力保持相对恒定，这是经过良好调整的腿摆动策略的预测结果。其次，我们看到在台阶下降时，双峰地面反作用力的第二个峰值增加，而在台阶上升时减少。在水平力（子图 B）中，我们看到振荡信号，其振荡频率与策略评估的频率相匹配。我们假设这是策略在控制骨盆姿态方面的工作。优先考虑身体姿态而不是前进速度，将类似于虚拟模型控制中明确优先考虑单次站立的行为。图 6 中的最后两个子图（C 和 D）显示了在站立阶段垂直和前进方向上的累积冲量。我们可以看到，台阶上升施加了更大的垂直冲量，而台阶下降施加了较小的垂直冲量。这与机器人应该施加较小的垂直冲量以降低自身下降台阶的直觉相符，而

不是提升自身上升台阶。水平冲量告诉我们机器人在站立阶段前进方向上加速或减速。我们看到台阶下降导致前进冲量显著增加，而台阶上升略微减少垂直冲量。这与从良好调整的摆动腿回缩策略中预测的行为一致。

3. *硬件设计：* 当前策略在硬件上转移时没有遇到任何显著困难。我们能够使用随机选择的楼梯 LSTM 策略，带机器人在一所大型大学校园内行走，并尝试攀登我们遇到的楼梯。我们观察到稳健且纠错的行为，以及成功且可重复的楼梯攀登和下降。此外，我们注意到对不平坦地形、原木和路缘的稳健性，这些在训练中都没有被建模。策略同样对倾斜和可变形地形具有稳健性，这通过在湿草地上行走和上小山的演示得以证明。这些实验可以在我们提交的视频中看到，其中一个实验的静态图像可以在图 3 中看到。

   除了在大学校园内测试一次性地形外，我们还在户外真实世界的楼梯上进行了十次上楼梯和十次下楼梯的试验。使用选定的楼梯 LSTM 策略，我们在上楼梯时记录了 80% 的成功率，在下楼梯时记录了 100% 的成功率。这次试验的完整视频可以在我们提交的附件中看到。我们注意到，学习到的行为对错误具有稳健性，并且可以从错误中快速恢复，尽管策略并非完全不可破坏，如果犯下特别严重的错误，它仍然会跌倒。这个实验可以在我们的补充视频 4 中看到。盲视、前馈学习策略似乎依赖于坚固的楼梯面；在模拟中评估倾斜楼梯上的政策导致更高的失败率，这指出了这种方法的局限性。即使在训练中明确包含，倾斜楼梯在上升时往往会使政策绊倒。相比之下，具有随机倾斜步骤的楼梯（例如，每步都有独特的俯仰和横滚方向）似乎对上升或下降没有困难。同样，以一定角度接近和攀登楼梯似乎对策略来说不是问题。

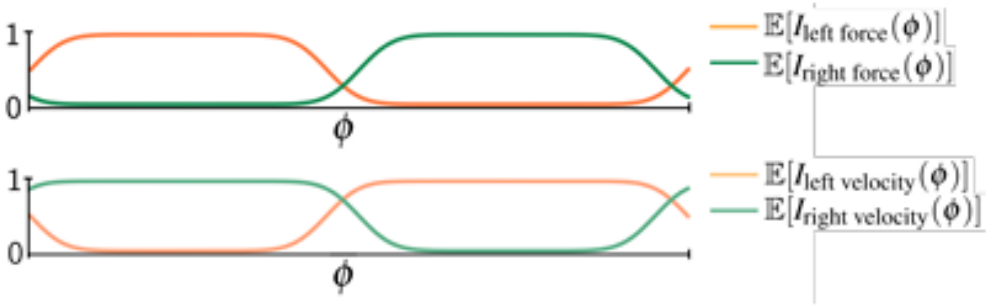


Figure 7: 通过在“支撑”阶段交替惩罚脚部力量来教导策略抬脚，在“摆动”阶段惩罚脚部速度来教导策略将脚放在地上，我们可以构建一个学习简单行走行为的基础。沿着前人工作的路径，我们将这些周期性系数定义为阶段的随机指示函数，并取其期望值。

5 结论

在这项工作中，我们论证了一个高度稳健但盲目的步行控制器的可取性，并展示了这样一个盲目的双足步行控制器能够攀登各种现实世界的楼梯。此外，我们注意到，产生这样的控制器只需要对现有的训练流程进行很少的修改，特别是不需要楼梯特定的奖励项；只需在环境中添加楼梯而无需进一步信息，就足以学习楼梯攀登的控制策略。这种学习能力的一个重要要求似乎是一种记忆机制，可能是由于在盲目行走未知地形的任务中部分可观察的性质。在未来的工作中，研究如何最有效地利用视觉来提高盲目双足机器人的效率和/或性能将是有趣的。此外，这项工作展示了盲目运动的惊人能力，并留下了关于极限在哪里的问题。

6 附录

6.1 奖励函数

为了简要回顾 [13] 中采取的方法，我们希望利用步行过程中脚力和脚速度的互补性质，构建一个奖励函数，该函数将在步态的关键间隔内惩罚一个并允许另一个，反之亦然。我们使用概率框架来表示这些间隔时间的不确定性。更具体地说，我们使用一个二值随机指示函数 $I_i(\phi)$ 来表示每个量 q_i，该函数在步态周期中的某个时间希望对其进行惩罚。这个指示函数在活跃期间可能为 1，在不活跃期间可能为 0。这个二值随机函数的分布通过 Von Mises 分布定义；有关更全面的描述，请参见 [19]。此外，我们不是在奖励中使用实际的随机变量，而是选择使用其期望值，以实现更稳定的学习；有关此期望的图示，请参见图 7。

我们的完整奖励函数如下：

$$R(x,\phi) = 1 - E[\rho(s,\phi)] \tag{5}$$

也就是说，我们的奖励是偏差与概率惩罚项 $\rho(s,\phi)$ 的期望值之间的差异，如 [13] 中所述。有关使用的确切量和权重的详细信息，请参见表 4。

Weight	Cost Component
0.140	$1 - E[\lvert leftforce(\phi)\rvert \cdot \exp(-0.1\lvert I_2\rvert)]$
0.140	$1 - E[\lvert rightforce(\phi)\rvert \cdot \exp(-0.1\lvert I_2\rvert)]$
0.140	$1 - E[\lvert leftvelocity(\phi)\rvert \cdot \exp(-\lvert I_2\rvert)]$
0.140	$1 - E[\lvert rightvelocity(\phi)\rvert \cdot \exp(-\lvert I_2\rvert)]$
0.140	$1 - \exp(-\epsilon_a)$
0.140	$1 - \exp(-\lvert \dot{x}_{desired} - \dot{x}_{actual}\rvert)$
0.078	$1 - \exp(-\lvert \dot{y}_{desired} - \dot{y}_{actual}\rvert)$
0.028	$1 - \exp(-5 \cdot \lvert \alpha_z - \alpha_z\rvert)$
0.028	$1 - \exp(-0.05 \cdot \lvert \tau\rvert)$
0.028	$1 - \exp(-0.1(\lVert \phi_{desired}\rVert + \lVert \phi_{actual}\rVert))$

Table 4: 成本项，它们被加在一起以构成期望惩罚 $E[\rho(s,\phi)]$。涉及变量 $I_i(\phi)$ 期望的项在步态周期中变化，目的是在关键间隔内惩罚脚力和脚速度，以教导策略定期抬起和放置脚以行走。其他项存在是为了命令策略向前、向后或侧向移动，或使机器人面向期望的方向。最后，剩余的项旨在减少摇摇晃晃的行为，从而不太可能在硬件上良好工作。

我们定义 F_l 和 F_r 为施加在左脚和右脚上的平移力的向量，v_l 和 v_r 类似地为左脚和右脚速度的向量。为了保持稳定的方向，使用了一个方向误差 ϵ_o，其等于：

$$\epsilon_o = 3(1 - \hat{q}^T q_{body})^2 + 10((1 - \hat{q}^T q_l)^2 + (1 - \hat{q}^T q_r)^2) \tag{6}$$

其中 q_l 和 q_r 是左脚和右脚的四元数方向，q_{body} 是骨盆的四元数方向，$\hat{q}$ 是一个期望的方向（为了我们的目的，固定为总是面向前方）。量 $\dot{x}_{desired}$ 和 $\dot{y}_{desired}$ 对应于命令的平移速度，而 $\dot{x}_{actual}$ 和 $\dot{y}_{actual}$ 是机器人的实际平移速度。项 $\dot{y}_{desired}$ 表示角速度，而 $\dot{y}_{actual}$ 表示平移加速度；这些项用于成本组件中，以减少步行行为的摇晃。项 a_t 和 a_{t-1} 指的是当前时间步的动作和前一时间步的动作，它们在成本组件中的使用是为了鼓励平滑的行为。项 τ 是施加在关节上的净力矩向量，其在成本组件中的使用是为了鼓励能效高的步态。

7 Reference

[1] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. Science Robotics, 5(47):eabc5986, Oct 2020. ISSN 2470-9476. doi: 10.1126/scirobotics.abc5986.

[2] Gerardo Bledt, Matthew J Powell, Benjamin Katz, Jared Di Carlo, Patrick M Wensing, and Sangbae Kim. MIT Cheetah 3: Design and control of a robust, dynamic quadruped robot. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2245–2252. IEEE, 2018.

[3] EZ Moore and M Buehler. Stable stair climbing in a simple hexapod robot. Technical report, McGill Research Centre for Intelligent Machines, 2001.

[4] Hae-Won Park, Alireza Ramezani, and Jessy W Grizzle. A finite-state machine for accommodating unexpected large ground-height variations in bipedal robot walking. IEEE Transactions on Robotics, 29(2):331–345, 2012.

[5] J-S Gutmann, Masaki Fukuchi, and Masahiro Fujita. Stair climbing for humanoid robots using stereo vision. In 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566), volume 2, pages 1407–1413. IEEE, 2004.

[6] Amos Albert, Michael Suppa, and Wilfried Gerth. Detec tion of stair dimensions for the path planning of a bipedal robot. In 2001 IEEE/ASME International Conference on Advanced Intelligent Mechatronics. Proceedings (Cat. No. 01TH8556), volume 2, pages 1291–1296. IEEE, 2001.

[7] Philipp Michel, Joel Chestnutt, Satoshi Kagami, Koichi Nishiwaki, James Kuffner, and Takeo Kanade. GPU accelerated real-time 3D tracking for humanoid locomo tion and stair climbing. In 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 463–469. IEEE, 2007.

[8] St´ ephane Caron, Abderrahmane Kheddar, and Olivier Tempier. Stair climbing stabilization of the HRP-4 humanoid robot using whole-body admittance control. In 2019 International Conference on Robotics and Automa tion (ICRA), pages 277–283. IEEE, 2019.

[9] Agility Robotics. Cassie: Dynamic Planning on Stairs.

[10] Giorgio Figliolini and Marco Ceccarelli. Climbing stairs with EP-WAR2 biped robot. In Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automa tion (Cat. No. 01CH37164), volume 4, pages 4116–4121. IEEE, 2001.

[11] Michele Focchi, Romeo Orsolino, Marco Camurri, Vic tor Barasuol, Carlos Mastalli, Darwin G. Caldwell, and Claudio Semini. Heuristic Planning for Rough Terrain Locomotion in Presence of External Distur bances and Variable Perception Quality, volume 132, pages 165–209. Springer International Publishing, Cham, 2020. ISBN 978-3-030-22327-4. doi: 10.1007/ 978-3-030-22327-4 9.

[12] Zhaoming Xie, Glen Berseth, Patrick Clary, Jonathan Hurst, and Michiel van de Panne. Feedback control for cassie with deep reinforcement learning. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1241–1246. IEEE, 2018.

[13] Jonah Siekmann, Yesh Godse, Alan Fern, and Jonathan Hurst. Sim-to-Real Learning of All Common Bipedal Gaits via Periodic Reward Composition. In IEEE International Conference on Robotics and Automation (ICRA), 2021.

[14] Richard S Sutton and Andrew G Barto. Reinforcement learning: An introduction. MIT press, 2018.

[15] Xue Bin Peng and Michiel van de Panne. Learning Locomotion Skills Using DeepRL: Does the Choice of Action Space Matter? CoRR, abs/1611.01055, 2016.

[16] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. Sim-to-Real Transfer of Robotic Control with Dynamics Randomization. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528.

[17] Sepp Hochreiter and J¨ urgen Schmidhuber. Long short term memory. Neural computation, 9(8):1735–1780, 1997.

[18] Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. Memory-based control with recurrent neural networks, 2015.

[19] Jonah Siekmann, Srikar Valluri, Jeremy Dao, Lorenzo Bermillo, Helei Duan, Alan Fern, and Jonathan Hurst. Learning memory-based control for human-scale bipedal locomotion. In Proceedings of Robotics: Science and Systems, July 2020.

[20] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017.

[21] Farzad Adbolhosseini, Hung Yu Ling, Zhaoming Xie, Xue Bin Peng, and Michiel van de Panne. On Learning Symmetric Locomotion. In Proc. ACM SIGGRAPH Motion, Interaction, and Games (MIG 2019), 2019. A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026–5033. IEEE, 2012.

[22] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco A physics engine for model-based control. In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 5026–5033. IEEE, 2012.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[24] KL Poggensee, MA Sharbafi, and A Seyfarth. Charac terizing swing-leg retraction in human locomotion. In Mobile Service Robotics, pages 377–384. World Scien tific, 2014.

[25] Monica A. Daley and Andrew A. Biewener. Running over rough terrain reveals limb control for intrinsic sta bility. Proceedings of the National Academy of Sciences, 103(42):15681–15686, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0601473103.

[26] Yvonne Blum, Susanne W Lipfert, Juergen Rummel, and Andr´ e Seyfarth. Swing leg control in human running. Bioinspiration biomimetics, 5(2):026006, 2010. [27] Andr´ e Seyfarth, Hartmut Geyer, and Hugh Herr. Swing leg retraction: a simple control model for stable running. Journal of Experimental Biology, 206(15):2547–2555, 2003.

[28] Aleksandra V. Birn-Jeffery, Christian M. Hubicki, Yvonne Blum, Daniel Renjewski, Jonathan W. Hurst, and Monica A. Daley. Don't break a leg: running birds from quail to ostrich prioritise leg safety and economy on uneven terrain. Journal of Experimental Biology, 217(21):3786–3796, 2014. ISSN 0022-0949. doi: 10.1242/jeb.102640.

[29] Jerry Pratt and Chee-Meng Chew and Ann Torres and Peter Dilworth and Gill Pratt. Virtual model control: An intuitive approach for bipedal locomotion. The International Journal of Robotics Research, 20(2):129 143, 2001. doi: 10.1177/02783640122067309.