arXiv:2303.03724v2 [cs.RO] 7 Aug 2023

利用电流反馈学习人形机器人的双足行走

Rohan P. Singh[1,2], Zhaoming Xie[3],PierreGergondet[1],Fumio Kanehiro[1,2]电子邮件: rohan-singh@aist.go.jp

摘要——近年来，基于深度强化学习（RL）的技术与仿真训练相结合，为开发鲁棒的足式机器人控制器提供了新途径。然而，这类方法在真实硬件上的应用，大多局限于配备直驱执行器的四足机器人，以及配备低齿轮传动比传动系统的轻型双足机器人。可能由于仿真到现实的差距较大，这类方法在接近真人尺寸的人形机器人上的应用仍不常见。

本文提出了一种方法，以有效解决人形机器人因执行器层级的力矩跟踪不准确而产生的仿真到现实差距问题。我们的核心思路是：在仿真环境中对力矩跟踪性能进行人为劣化并训练策略后，利用真实机器人执行器的电流反馈。通过该方法，我们成功在仿真中训练出统一的端到端策略，并将其部署于真实 HRP-5P 人形机器人上，实现双足运动。通过消融实验，我们还表明，将前馈策略架构与有针对性的动力学随机化相结合，足以实现零样本仿真到现实的成功迁移，从而无需使用计算开销巨大的基于记忆的网络架构。最后，我们通过比较所提出的强化学习策略与传统基于模型的控制器在真实机器人不平坦地形行走任务中的表现，验证了该策略的鲁棒性。

I. 引言

尽管用于人形机器人运动的传统基于模型的方法仍在不断改进，例如基于预瞄控制 [1]或模型预测控制（MPC） [2],的方法，但其针对意外扰动和不准确建模的鲁棒性仍然是一个难以实现的研究目标。另一方面，用于腿式运动的基于强化学习的控制方法迅速发展，在非结构化和非受控环境中的四足机器人 [3], [4], [5], [6] ，乃至双足机器人 [7], [8], [9]上，都展现出了卓越的性能。因此，将类似方法应用于开发更大型、更重型人形机器人的行走控制器，也颇具吸引力。

使用深度强化学习训练具备能力的策略需要大量数据，而且可能损坏硬件。物理仿真环境提供了一种安全的方式来收集大量数据，因此通常会先在仿真中训练策略，再将其迁移到真实系统。然而，仿真环境可能无法充分反映真实世界动力学的丰富性。这便产生了“现实差距”，通常也称为仿真到现实 差距。仿真到现实 差距可能导致在仿真中训练的策略部署到真实硬件上时，性能大幅下降。在

[1] CNRS-AIST JRL（联合机器人实验室）IRL，日本产业技术综合研究所（AIST）。[2] 筑波大学，日本茨城。[3] 斯坦福大学计算机科学系，美国。



图1：**HRP-5P 人形机器人** 在 MuJoCo 中通过无模型强化学习训练实现双足运动（上）；强化学习策略迁移至真实机器人（下）。我们利用测得的执行器电流反馈，以应对真实系统中较差的力矩跟踪性能。

对于 HRP系列人形机器人等全尺寸人形机器人，与近期大多数研究使用的四足机器人或轻型双足机器人相比，这种差距对机器人行走时的稳定性影响可能更为关键。此前有人提出了基于记忆的策略架构，利用时间信息来实现在线系统辨识，从而应对这一问题 [10], [5]。但这类网络通常比前馈（FF）网络的计算成本更高，这可能会给强化学习策略的原型开发与部署带来阻碍，尤其是在样本复杂度较高的学习任务中。

此前已有研究指出，与连杆质量及其质心（CoM）位置等其他因素相比，仿真执行器模型对仿真到现实 迁移有重大影响 [4], [11], [12]。我们认为，在执行器模型中，对力矩跟踪行为的建模是造成真实机器人与

仿真机器人之间差异的一个关键来源。在仿真环境中，强化学习策略所期望的关节扭矩会被直接作为控制输入注入动力学计算。然而，对于 HRP-5P 这类采用电流控制的真实机器人，由于电流控制并不完美（且有刷直流电机的扭矩与电流之间近似呈线性关系），电机实际施加到连杆上的扭矩可能与期望扭矩有显著差异。由于机器人只能通过对连杆施加扭矩与环境交互，策略期望的扭矩与实际施加到连杆上的扭矩之间的差异可能对控制产生重大影响。

本文开发了一套系统，用于在仿真中训练双足行走策略，并将其部署到 HRP-5P 人形机器人上（图1）。HRP-5P 是一款大功率电驱动人形机器人，拥有 53 个自由度（DoF），重量为 105 kg，身高为 $182cm$ [13]。我们的关键发现是，对于此类机器人，仿真到现实差距很大程度上源于真实机器人上的电流跟踪不完善，这会导致期望扭矩与施加扭矩不匹配。为此，我们提出在训练期间模拟退化的扭矩跟踪效应，并将电机的电流反馈纳入观测空间。由此得到的策略能够学会主动利用电流反馈信号，并补偿电机驱动器中不准确的力矩跟踪。最终，该策略成功部署在真实的 HRP-5P 机器人上。

具体而言，我们在这项工作中的贡献如下：

- 通过采用所提的仿真到现实方法，我们展示了首批演示之一：端到端策略可使一台真实的电流控制人形机器人实现动态稳定。该策略能够根据用户通过操纵杆发出的指令，实现向前行走、迈步、原地转向以及相当稳定的站立。
- 我们针对前馈多层感知机（MLPs）和长短期记忆网络（LSTMs）开展消融研究，以证明无需依赖基于记忆的策略架构，也无需对动力学参数进行不合理的大范围随机化，就能弥合“现实差距”。这对于在合理的计算资源范围内开发强化学习策略至关重要。
- 我们在真实机器人上验证了所提策略的鲁棒性，并将强化学习策略的定性表现与一个开源基于模型的行走控制器进行比较，考察它们在盲行过程中越过较小的不平整且柔顺的障碍物的能力。

II. 相关工作

A. 足式机器人的强化学习

强化学习已成为为足式机器人合成控制器的一种强大方法。控制策略通常在仿真中训练，然后迁移到硬件上，即仿真到现实。大量此类研究聚焦于四足机器人，例如 ANYmal [4], Laikago[14],A1 [5], 绝影 [15] 和 Mini-Cheetah [3]。还有一些研究也成功地将相同方法应用于双足机器人，例如 Cassie [7], [10], [16], Digit [17], [18],[19] 和 NimbRo-OP2X [20]。DeepWalk [20] 展示了一种单一的学习策略，可用于真实的人形机器人，使其借助 Beta策略 [21]，在平坦地面上实现全向行走，不过研究指出，若要改善迁移效果，还需要更先进的方法。

此前已有研究展示了 HRP-5P 和 JVRC-1 人形机器人的端到端深度强化学习策略可用于按照规划步点行走，但仅限于仿真环境 [22]。而本研究的重点是解决实现真实机器人演示所面临的仿真到现实 问题。

近期，一项在真实 TOCABI 人形机器人上实现的、令人瞩目的基于力矩的深度强化学习运动策略演示，展现了采用力矩动作空间对改善仿真到现实 迁移的优势 [23]。TOCABI 是一款全尺寸人形机器人 [24] ，无需使用关节级力矩传感器即可实现力矩控制顺应性。将这种方法应用于 HRP-5P 人形机器人的低反驱关节，是未来研究的课题。

同样，在 Valkyrie 和 HRP系列机器人等笨重的人形机器人的硬件上实现学习得到的策略，比在较轻型机器人上面临的挑战大得多，尤其是安全风险更高。

B. 仿真到现实方法

成功的 仿真到现实 方法所需的重要要素包括：(a) 仔细进行系统辨识，例如，将学习得到的执行器模型纳入仿真器，以考虑难以建模的执行器动力学 [4], [11], (b) 领域自适应，即策略根据观测与动作历史学习进行适应 [5], [6], [10], (c) 动力学随机化，即对质量、摩擦、惯性、质心位置等仿真参数进行随机化，以提升策略的鲁棒性[25]。尽管这些方法已在多个腿式机器人平台上取得成功，但我们尚未看到此类方法成功应用于具有重型肢体的真人尺寸（例如身高超过 170 厘米）人形机器人。这并非因为缺乏尝试，已有研究探索了如何将强化学习应用于美国国家航空航天局的 Valkyrie 机器人，例如：[26],但截至目前，尚未展示任何 仿真到现实 成果。

对于 HRP-5P，由于缺少关节扭矩传感器，无法提供此类执行器模型所需的真实信号，因此端到端监督学习执行器动力学 [4] 并不容易应用。此外，对于大型机器人而言，基于真实数据 [11], [12] 的基于搜索的系统辨识方法可能难以实施，因为这些方法需要在真实机器人上进行多次试运行（5-10 次）。通过使用基于记忆的智能体，领域自适应和在线系统辨识方法已成功应用于双足机器人 [10], [5]。然而，训练基于记忆的网络（如长短期记忆网络）可能会显著增加计算开销，并可能无法在真实机器人上进行实时推理。我们的实验表明，结合针对性动力学随机化，我们可以弥合“现实

差距”，并使用轻量级前馈网络（实验结果见 V-F 小节）。

C. 人形机器人运动的传统控制

人形机器人双足运动的传统基于模型的方法，由局部反馈控制器组成，用于跟踪通过离线过程预先计算的零力矩点（ZMP）或质心（CoM）轨迹。利用发散分量运动（DCM）进行稳定化控制的方案已在先前的研究中得到广泛研究 [27]，[28]，并已应用于真实机器人。该方法依赖于双足行走的线性倒立摆模型 [29], [30]。另一方面，其他近期研究（如 [1], ）不依赖于双足特定的动力学，而是基于预瞄控制在线生成质心轨迹，从而实现出色的多接触运动。该方法使用预瞄控制方案生成质心轨迹，并通过稳定化方案修正轨迹跟踪误差。由于轨迹是在线生成的，机器人能够对环境扰动作出鲁棒响应。

通常，这类控制器的输出形式为期望的ZMP、CoM和/或接触力旋量，随后将其输入二次规划（QP）求解器，以计算期望关节位置。然后，由局部比例-微分（PD）控制器跟踪这些期望关节位置，以实现刚性位置控制。这与上文提到的强化学习方法形成鲜明对比：后者采用较低的PD增益来实现柔顺跟踪，因此在不平坦地形上具有更强的鲁棒性。

为此，我们进行了真实机器人实验，以比较所提出的强化学习策略与一种常用的开源基于模型的控制器在不平坦地形上运动时的鲁棒性（详细方法和结果见第 V-E 小节）。

III. 背景

HRP-5P 机器人的电机控制系统（一般而言，大多数机器人传动系统也是如此）如图 2 所示。该模块结构由一个 PD控制器组成，该控制器根据来自高层控制器的参考位置和来自关节编码器的测量位置，计算所需的扭矩指令。随后，所需的扭矩指令——或者等效地，对于有刷直流电机，假设扭矩与电流成正比关系时的电流指令——被发送至比例-积分（PI）控制器。PI 控制器根据电机中电流传感器提供的测量电流，尝试跟踪该电流指令。PI 控制器的输出被馈送至电机功率放大器，后者继而驱动电机。

这里的关键观察是，PI 控制器无法按照高层控制器或策略的要求精确跟踪扭矩指令，这往往会导致明显的力矩跟踪误差。我们认为，这种跟踪误差很可能主要由反电动势（EMF）的影响造成。电机旋转时，反电动势会产生一个与施加电压方向相反的电压，从而减少流经电枢的电流，进而导致跟踪误差。这使得真实系统与仿真环境大相径庭：在仿真环境中，期望的扭矩指令可以完全无误差地施加到执行器上。电池电压、传动电缆的电阻、负载变化，或 PI 控制器的增益调节不当等其他因素，也可能导致电流跟踪效果不佳。

IV. 方法

本节详细介绍训练强化学习策略所涉及的各个组成部分（见图2）。训练在 MuJoCo 仿真器中进行 [31]。具体而言，我们描述了如何通过模拟反电动势效应并为策略使用电流反馈，来克服真实机器人上较差的力矩跟踪性能。由于假设机器人执行器中的电流与扭矩成正比，因此全文中这两个术语可互换使用。

A. 观测与动作

我们的策略输入由机器人状态、外部状态和时钟信号组成。机器人状态包括每个执行关节（每条腿 6 个）的关节位置和关节速度、根部（骨盆）的横滚和俯仰姿态及角速度，与多项相关工作类似 [22]，[10]。在本研究中，我们还提出将每个执行器的电机级扭矩信号纳入机器人状态 tau_{obs}。在仿真中，该信号等于上一个时间步施加到执行器上的实际扭矩 $\text{tau}_{applied}$。在真实机器人上，tau_{obs} 需要根据原始电流测量值推导得出，具体说明见 V-B。外部状态向量包含一个 $3D$ 独热编码，用于表示行走模式—— $[0,0,1]$ 表示站立， $[0,1,0]$ 表示原地迈步， $[1,0,0]$ 表示向前行走。它还包括一个$1D$ 标量，根据模式充当参考值：如果活动模式为 迈步，则参考值表示转向速度；对于 行走 ，则表示向前行走速度；而在站立模式下则忽略该值。

策略还会观测一个取决于周期相位变量 ϕ的时钟信号。该变量还用于在我们的奖励函数中定义周期性奖励项，以生成行走行为。我们将 ϕ 双射投影到一个$2D$ 单位圆：

$$\text{Clock} = \left\{ \sin\left(\frac{2\pi\phi}{L}\right), \cos\left(\frac{2\pi\phi}{L}\right) \right\}, \tag{1}$$

其中 L是周期。 ϕ 在每个控制时间步从 0 增加到 1，并在每经过 L 个时间步后重置为 0。随后，时钟信号将作为策略的输入。

策略输出机器人腿部驱动关节的期望位置。网络的这些预测值会与对应于机器人半蹲姿态的固定电机偏移量相加。随后使用低增益比例微分控制器跟踪期望位置，该控制器按如下方式计算期望关节扭矩：

$$\text{tau}_{pd} = K_p(q_{des} - q) + K_d(0 - \dot{q}), \tag{2}$$

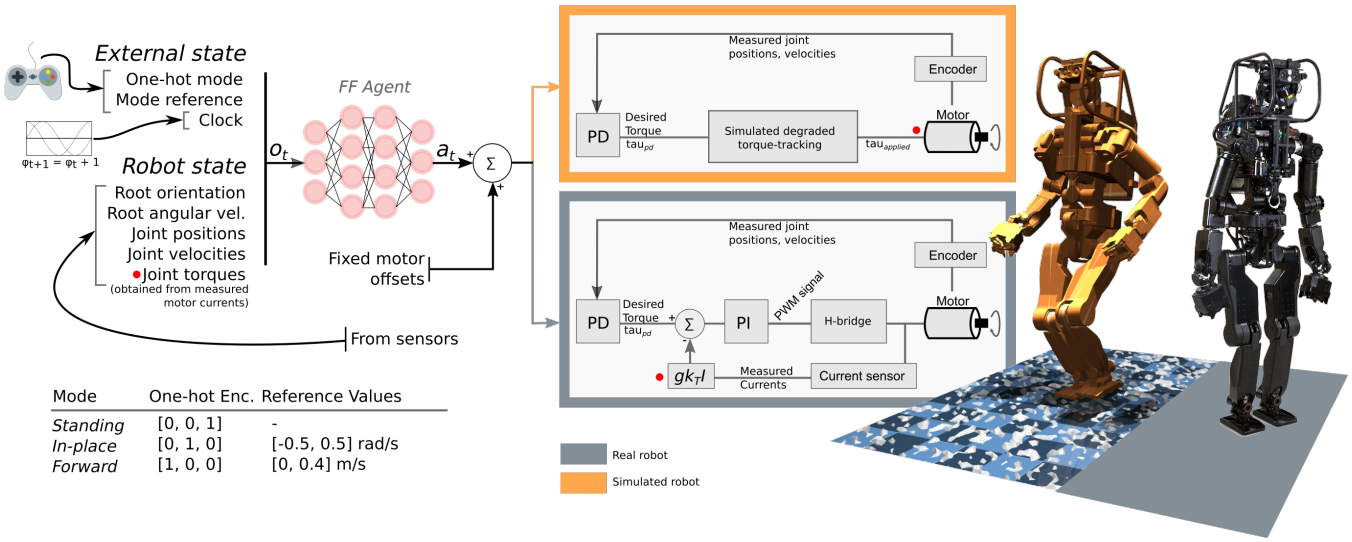

图 2：**HRP-5P 中所提出的强化学习策略及电机控制系统框图概览**。强化学习策略的输出为“期望关节位置”，该输出与固定的电机偏置（对应机器人的标称姿态）相加后，输入关节比例微分控制器。在真实机器人上，比例微分控制器计算出的扭矩由比例积分电流控制器进行跟踪，但存在显著的跟踪误差。这些跟踪误差构成了 仿真到现实 差距的关键部分。为解决这一问题，我们提出：(a) 在训练期间模拟退化的力矩跟踪环境；(b) 观测实际施加的电机扭矩，或等效地，观测真实机器人上测得的电机电流（记为 •）。

其中，K_p 和 K_d 分别表示比例增益系数和微分增益系数。q_{des} 是策略预测值与固定电机偏移量之和。q 和 $\dot{q}$ 分别表示当前关节位置和速度。

B. 奖励函数

我们的奖励设计确保无需参考动作。相反，我们依赖多个手工设计的奖励项，以促进机器人在 3 种模式下实现期望的机器人行为：原地站立、原地踏步（包括转向），以及在给定参考速度的情况下向前行走。这要求机器人形成周期性双足步态，遵循模式和参考速度指令，并保持固定高度。此外，我们引入了一些奖励项，以形成更逼真的动作，从而以真实且安全的方式实现仿真到现实 迁移。

双足行走。我们引入奖励项，以促进对称的双足行走步态，其特征是腿部周期性运动，并在双支撑（DS）相和单支撑（SS）相之间交替。根据相位变量 ϕ 和期望模式（站立 或行走），计算足部地面反作用力（GRF）和身体速度的奖励。

例如，当 ϕ 位于步态周期的第一个单支撑区域（左足处于摆动相，右足处于支撑相）时，左足地面反作用力（GRF）越大，奖励越低；右足 GRF 越大，奖励越高。同时，鼓励左足具有更高的速度，而对右足的高速度予以惩罚。

双足行走术语的定义——足部的地面反作用力 r_{grf} 和足部的身体速度 r_{spd}——完全采用自 [22]:

$$r_{\text{grf}} = I^{grf}_{left}(\phi) \cdot F_{left} + I^{grf}_{right}(\phi) \cdot F_{right} \tag{3}$$

$$r_{\text{spd}} = I^{spd}_{left}(\phi) \cdot S_{left} + I^{spd}_{right}(\phi) \cdot S_{right} \tag{4}$$

其中，F_{left} 和 F_{right} 表示左、右足部的地面反作用力，S_{left} 和 S_{right}分别表示左、右足部的身体速度。关于用于调节地面反作用力和足部速度奖励系数的“相位指示器”函数 I^{grf}_{*}和I^{spd}_{*}的详细解释，请参阅 [22]。

对于站立，双支撑相扩展至覆盖整个步态周期；策略通过最大化双足的地面反作用力，同时尽量降低足部速度来获得奖励。

根部速度、方向和高度。根部线速度奖励项是对机器人根连杆沿 x 方向的全局速度 ${}^{x}v_{root}$ 施加的简单代价。

$$r_{\text{rv}} = \exp(-10 \cdot \| {}^{x}v_{root} - {}^{x}\hat{v}_{root}\|^2) \tag{5}$$

根部偏航速度项鼓励根部 ω_{root} 的角速度接近期望速度$\hat{}\omega_{root}$。

$$r_{\text{av}} = \exp(-10 \cdot \|\omega_{root} - \hat{\omega}_{root}\|^2) \tag{6}$$

在训练期间，每回合开始时，会在站立、迈步和行走 之间随机选择活动模式。根据活动模式，参考值的标量输入从均匀分布中采样，即：当模式为行走时，$\hat{}v_{root}$ 从 $[0, 0.4]\text{m s}^{-1}$ 的范围内采样；当模式为站立时，$\hat{}\omega_{root}$ 从 $[-0.5, 0.5]\text{rad s}^{-1}$ 的范围内采样。

我们还会奖励策略，使其将根部高度 $h_{root}\hat{}$ 维持在期望值 $h_{root} = 0.79\text{m}$:

$$r_{\text{height}} = \exp(-40 \cdot (h_{root} - \hat{h}_{root})^2). \tag{7}$$

安全且逼真的运动。除了上述各项之外，我们还尝试生成一种尽可能接近机器人标称姿态的运动，以避免不必要的摇摆。这对于在真实机器人上安全部署至关重要，因为真实机器人的运动范围很广，且执行器的驱动力显著强劲。

为鼓励机器人保持直立姿态，我们使用一个奖励项来最小化头部位置的地面投影 ${}^{x,\,y}\mathbf{p}_{head}$ 与根部位置 ${}^{x,\,y}\mathbf{p}_{root}$之间的距离。这可以防止机器人出现后仰行为，并避免上身过度摇摆：

$$r_{\text{upper}} = \exp(-10 \cdot \| {}^{x,y}\mathbf{p}_{head} - {}^{x,y}\mathbf{p}_{root} \|^2). \tag{8}$$

我们使用一个惩罚项来惩罚当前关节位置 $\mathbf{q}$ 与标称“半蹲姿态”$\mathbf{q_{nominal}}$之间的距离：

$$r_{\text{posture}} = \exp(-\|\mathbf{q} - \mathbf{q_{nominal}}\|^2). \tag{9}$$

我们还会对超过最大关节速度 $\dot{\mathbf{q}}_{\mathbf{lim}}$ 50% 的关节速度 $\dot{\mathbf{q}}$ 施加惩罚。

$$r_{\text{jv}} = \exp\left(-5 \times 10^{-6} \sum_{\dot{\mathbf{q}} > k \cdot \dot{\mathbf{q}}_{\mathbf{lim}}} \|\dot{\mathbf{q}}\|^2 \right). \tag{10}$$

完整的奖励函数如下：

$$\begin{aligned} r = w_1 r_{\text{grf}} + w_2 r_{\text{spd}} + w_3 r_{\text{rv}} + \dots \\ w_4 r_{\text{av}} + w_5 r_{\text{height}} + w_6 r_{\text{upper}} + \dots \\ w_7 r_{\text{posture}} + w_8 r_{\text{jv}}, \end{aligned} \tag{11}$$

其中，权重 w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8 分别设为 0.225、0.225、0.100、0.100、050、0.100、0.100、0.100。

C. 动力学随机化

由于在仿真中训练的策略与不完美的世界表征交互，因此容易过拟合，并且在硬件上表现得不够稳健。克服这一问题的一种常见方法是随机化各种机器人模型和环境参数，例如质量、惯性、电机强度、延迟、地面摩擦力等 [10], [5]。

在我们的工作中，我们仔细选择了为实现更好的迁移而需要随机化的变量。首先，由于连杆内部电子元件和机械部件的分布，真实系统中每个连杆的质量和质心位置（CoM）可能与仿真中的不同。训练期间，我们在每个回合开始时将每个连杆的质量随机化 5%，并将质心位置随机化为 $5cm$。

其次，先前的研究表明，在齿轮传动系统中，电机与负载之间存在相当大的摩擦 [32]。这种摩擦力矩难以辨识，甚至难以在仿真中建模。MuJoCo 支持模拟静摩擦和黏性摩擦。因此，我们将静摩擦幅值随机设定在均匀范围 (2，8) N·m 内，并将黏性摩擦系数随机设定在均匀范围 (0。5，5) Nm/rad s^{-1} 内，依据 [32] 中进行的粗略辨识结果。

除了质量、质心位置和关节摩擦外，训练期间我们不会随机化其他机器人动力学参数。

D. 地形随机化

为了使真实机器人能够稳健地在不平坦地形上行走，我们在训练期间让策略接触随机化地形。在 MuJoCo 中，地形使用高度场表示——一个包含高程数据的 $2D$ 矩阵。由于在训练过程中在线生成新的高度场可能会拖慢训练速度，我们改为在训练开始时生成一个随机高度场，然后在每个回合中随机改变它相对于平坦地面的平移位置。平坦地面平面和高度图会同时进行仿真，使地面呈现为一个散布着不同高度障碍物的地形。这样，机器人接触到的不平坦地形最大高度为 3.5 厘米。本文在整个训练过程中均匀引入地形随机化；也可以根据课程学习逐步引入，以促进更平稳的学习。

E. 模拟反电动势

为了模拟扭矩跟踪不佳这一现象（在真实系统中观察到），我们在训练阶段对每个仿真时间步中各关节的施加扭矩进行修正。具体而言，通过注入一个随关节速度缩放的反扭矩来实现这一修正，所用方程如下：

$$\text{tau}_{applied} = \text{tau}_{pd} - \text{k}_{bemf} \cdot \dot{q}, \tag{12}$$

其中，tau^{pd} 表示PD控制器输出端的扭矩，$\dot{q}$ 表示关节速度。对于真实系统，阻尼系数k^{bemf}是未知的。在训练期间，我们对k^{bemf} 进行随机化，以模拟不同的跟踪行为。每隔100 ms，都会为每个关节从 $[5, 40]$ 中均匀采样一个系数。该范围是通过比较仿真环境与真实机器人实验日志，以经验方式确定的。

在训练期间，策略直接观测到$\text{tau}^{applied}$ ，即$\text{tau}_{obs} = \text{tau}^{applied}$。

V. 实验

A. 强化学习策略

训练细节。与 [22],一样，actor 和 critic 策略均由多层感知机架构表示，用于参数化策略和价值函数，并采用近端策略优化（PPO） [33] 进行训练。两个多层感知机网络都包含 2 个隐藏层，每层有 256 个单元，并使用 *ReLU*激活函数。每次回合展开最多包含 400 个控制时间步（相当于 10 秒的模拟时间）；若满足终止条件，则可能重置。每个训练批次包含 64 次这样的回合展开。学习率设为 0.0001。

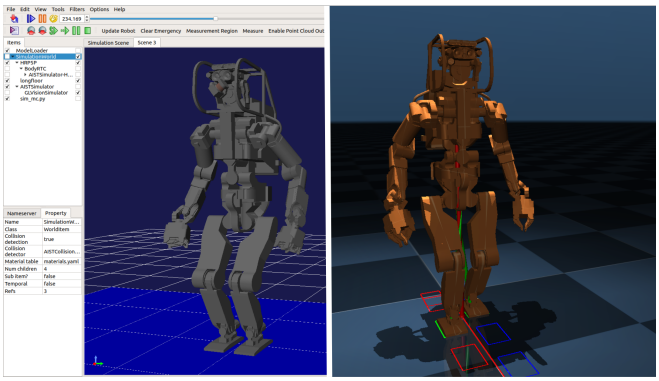

图 3：**仿真到仿真验证**。分别使用 mc-openrtm 和 mc-mujoco 接口，在 Choreonoid（左）和 MuJoCo（右）中对 HRP-5P 进行仿真。

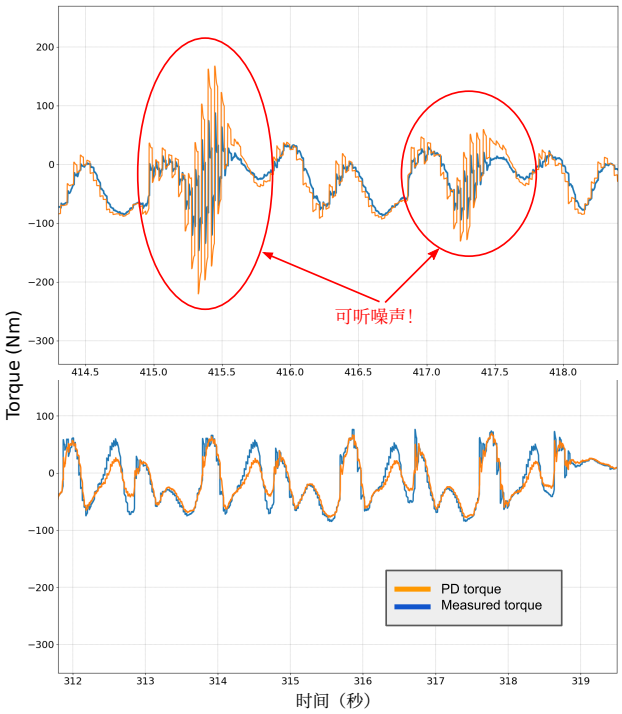

图 4：**基线策略（上）与所提策略（下）：右髋关节侧摆关节力矩跟踪的真实实验日志**。我们观察到，当机器人转向时，基线策略（圆圈标出）的力矩跟踪变得不稳定，并且可以听到关节发出的可听噪声。使用所提方法时未观察到这些现象，且从视觉上看，机器人的摆动腿行为似乎更加稳定。

我们使用 *LOSS* 方法 [8] , , 该方法通过添加一个辅助损失项（在原有 PPO 损失项之外）来强制实现对称性。在 AMD Ryzen Threadripper PRO 5975WX CPU（32 核）上，训练 FF 策略约需 12 小时，以收集总计 1.2 亿个样本来学习所有模式。

B. 实机器人上的实现

我们提议将实际施加扭矩纳入强化学习策略的观测向量中。如前所述，真实机器人上的施加扭矩是从电机驱动器中的电流传感器测量值中提取的。与关节对应的测量电流 I_{meas} 乘以扭矩常数k_T 和齿轮比 g，然后输入策略，如下所示：

$$\text{tau}_{mot} = I_{meas} \cdot k_T \tag{13}$$

$$\text{tau}_{obs} = g \cdot \text{tau}_{mot} \tag{14}$$

需要注意的是，此处的施加扭矩是指在执行器层面施加、并缩放到关节空间的扭矩。由于 HRP-5P 机器人没有关节扭矩传感器，因此无法测量施加在负载（即机器人连杆）上的扭矩。由于存在静摩擦和黏性摩擦，缩放后的执行器扭矩与关节扭矩之间的差异实际上可能相当显著。

该策略在机器人的控制PC上执行（规格：配备 2 个核心的 Intel NUC5i7RYH i7-5557U CPU、Ubuntu 18.04 LTS PREEMPT-RT内核），并在 C++ 中实现为 *mc-rtc*[1]，控制器。推理频率为 40 Hz，而 PD控制器以 1000 Hz 运行。策略推理速度非常快，仅耗时约 0.2 ms。控制器的整体运行时间约为 1 ms。

C. 仿真到仿真验证

虽然我们使用 MuJoCo 作为训练环境，但在部署到真实机器人之前，也会在 Choreonoid 模拟器中进行全面评估。Choreonoid 一直是用于模拟人形机器人控制器 {v1>。我们使用 *mc-rtc* 控制框架在机器人的控制PC上执行策略。

这使我们能够在 3 种不同环境中透明地评估同一个控制器代码：(1) 在 MuJoCo 中使用 *mc-mujoco* 接口 [35]，；(2) 在 Choreonoid 中；以及 (3) 在真实机器人上，后两者均使用 *mc-openrtm* 接口与 OpenRTM 中间件进行通信。[36] 该中间件用于 HRP 机器人和 Choreonoid 仿真器。

我们发现，Choreonoid 中的接触建模比 MuJoCo 中针对高度场的接触建模更加稳定。因此，在进行实机器人部署之前，Choreonoid 构成了我们在不平坦地形上评估策略的重要流程组成部分。尽管如此，除了一些细微差异外，我们在评估过程中并未观察到 MuJoCo 与 Choreonoid 之间的策略行为存在任何重大差异。

D. 消融研究

我们针对本文提出的 sim2real 成功的两个主要要素进行消融研究：(1) 使用模拟的较差力矩跟踪进行训练；以及 (2) 向策略提供力矩反馈。

尽管真实机器人实验成本高昂，而且测试容易失败的策略可能很危险，但在仿真中构建测试环境并不是一种可靠的替代方案。我们发现，即使在极具挑战性的情况下（例如性能退化的

[1]https://jrl-umi3218.github.io/mc_rtc/index.html

力矩跟踪、不平坦地形、外部扰动），在仿真中表现成功的策略部署到真实机器人上时，仍可能出现不理想的行为，这表明存在巨大且关键的现实差距。因此，研究这些策略在真实系统上的行为十分重要。我们训练了 3 种不同的策略，以分析所提方法的影响：

1)**策略 A** 构成基线策略。训练时未模拟较差的力矩跟踪，也未观测来自执行器的电流反馈。在真实机器人上部署时，该策略的表现最差。摆动腿落地时，机器人容易出现足部之间的自碰撞。这表明策略在控制真实机器人的腿部摆动运动时面临困难。这是因为策略是在仿真环境中以完美跟踪为条件进行训练的，但在真实机器人上却面临跟踪性能下降的问题。此外，我们尝试训练另一个策略，并增加足部间距的终止条件（$d < 0.2\text{m}$），以促进形成更宽的站姿。在这种情况下，真实机器人上不会发生自碰撞，但我们观察到，在运动的某些区域，力矩跟踪变得不稳定，关节还会发出可听噪声（见图 4）。

2)**策略 B** 是在力矩跟踪较差的情况下训练的，但未使用力矩反馈。训练和评估时，我们将力矩反馈对应的输入替换为 **0** 向量，同时保持其他所有参数不变。该策略在仿真环境中表现出较强的鲁棒性。然而，在真实机器人上部署时，我们再次观察到足部之间发生自碰撞。摆动腿的速度也明显更高，这意味着策略更难补偿指令与施加扭矩之间不断变化的偏差。根据这一观察，我们得出结论：提供力矩反馈（由测量电流得到）对于策略适应力矩跟踪性能下降的环境至关重要。

3) **策略 C** 采用所提出的方法进行训练：模拟较差的力矩跟踪性能，并向策略提供来自电流测量值的反馈。该策略在真实机器人上表现得明显更加稳定。在整个运动过程中未观察到自碰撞，也没有可听噪声。机器人能够成功行走数米，包括转向、原地踏步和站立。机器人还能够在由刚性障碍物和柔性障碍物组成的最高 2 厘米不平坦地形上行走。

我们进一步分析了图 5 中对应 3 种策略在真实机器人上进行“RKP”（右膝俯仰）关节力矩跟踪的实验日志。选择“RKP”关节是因为该关节的跟踪误差更明显。此外，膝关节预计会对行走行为产生更大的影响（例如，相比之下，髋关节偏航的影响较小）。我们观察到，策略 A 和策略 B 都出现了足部之间的自碰撞，而策略 C 则能够稳定行走并应对

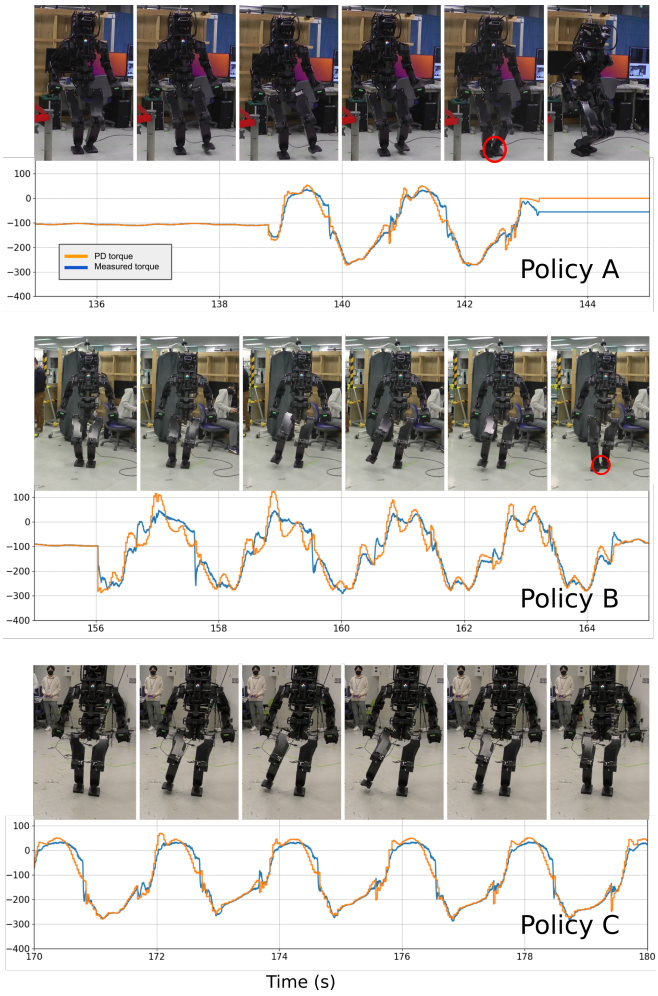

图 5：**RKP关节的力矩跟踪性能**（针对 3 种策略）。策略 A 在仿真中训练时未考虑反电动势效应，且未使用反馈。策略 B 在训练时考虑了反电动势，但未使用反馈。策略 C 在训练时考虑了反电动势，并使用了扭矩（电流）反馈。策略 A 和 B 都会导致足部之间发生自碰撞（红圈所示），而策略 C 能够实现稳定行走。

不平坦地形。策略 A 和策略 C 的“RKP”关节跟踪情况看起来有些相似；不过，对于策略 C，反馈使策略能够根据上一个时间步的跟踪误差做出反应，从而更好地控制摆动运动。策略 B 的跟踪效果则明显更差。策略 A（针对更大的足间距重新训练）和策略 C 的“RCR”（右髋滚转）关节跟踪情况也显示在图 4 中。

值得注意的是，提供扭矩反馈并不能消除跟踪误差，因为策略难以预判未来的误差。我们认为，将历史数据纳入观测空间也会有所帮助。

E. 与基于模型的控制器的比较

我们将策略在不平坦地形上进行运动的鲁棒性，与现有的人形机器人运动基于模型的控制器进行比较。测试中，我们使用开源的

BaselineWalkingController [2]，该控制器实现了基于线性倒立摆模式（LIPM）的行走控制。该方法结合了基于预瞄控制的在线质心轨迹生成、基于 LIPM 发散运动分量（DCM）的零力矩点反馈，以及基于阻尼控制律的足部接触力控制[28], [29], [30]。

我们的测试环境由一叠带衬垫地毯块构成，每块厚度为 0.6 厘米，铺在平坦地面上。机器人从地毯块前方一段距离处出发，需要迈步越过障碍并踏上地毯块。由于地毯块由柔软材料制成，这个障碍形成了柔顺支撑面，从平衡角度来看更具挑战性。*BaselineWalkingController* 能够成功越过由 3 块地毯组成的地毯堆，但未能越过由 4 块地毯组成的地毯堆（高度接近 2.4 厘米）——当支撑腿踩在叠放的地毯上时，机器人失去平衡并摔倒。另一方面，采用我们的方法训练的强化学习策略，在 2 次试验中有 2 次成功越过由 5 块地毯组成的地毯堆（高度为 $= 3$ 厘米）。该策略也能成功在障碍物上实现多次部分接触，即足部仅部分落在地毯堆上。（测试结果见补充视频。）

尽管已有更新的、旨在提高鲁棒性的人形机器人运动基于模型的方法 [1]，[2]，我们的测试仍然为了解无模型强化学习策略相对于基于线性倒立摆模式的双足运动控制器的鲁棒性提供了有价值的见解。对于基于深度强化学习的控制器，其鲁棒性较高的关键原因仍有争议，但我们认为，较低的 PD增益、策略的反馈特性，以及对足部轨迹缺乏严格约束，可能发挥了重要作用。

F. 前馈策略与基于记忆的策略

在本小节中，我们通过仿真和真实机器人实验，研究不包含任何历史状态信息的前馈（FF）多层感知机策略，与基于记忆的策略架构（如长短期记忆网络（LSTM））以及使用观测历史的前馈策略的行为。具体而言，我们分析了以下 3 种网络架构：

(a) FF策略（基础型），由一个仅观测当前观测状态的多层感知机构成 o_t(b) 带历史信息的FF策略，观测当前状态以及之前的 3 个状态 $\{o_t, o_{t-1}, o_{t-2}, o_{t-3}\}$(c) 具有两个隐藏层的 LSTM策略，每个隐藏层包含 128 个单元。

两种前馈策略均具有 2 个隐藏层，每层包含 256 个节点。每个策略的评论家网络与智能体采用相同的架构，但输出层除外。我们分别在采用和不采用动力学随机化（DR）的情况下训练了每种架构。所有策略均未获得任何有关动力学扰动的信息。所有实验均在平坦且坚硬的地面上进行。奖励曲线绘制于图 6 中。

在训练过程中，我们发现，就奖励收集而言，前馈策略的表现略优于 LSTM策略。

[2]https://github.com/isri-aist/BaselineWalkingController

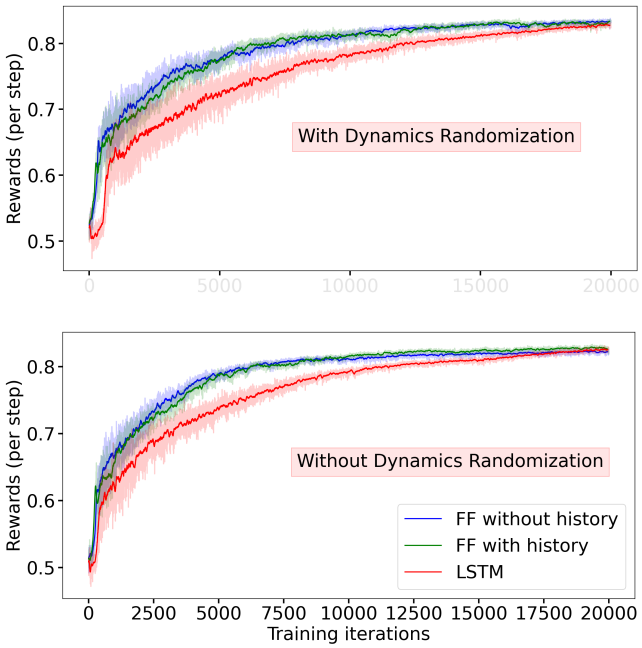

图 6：**不同策略架构的训练奖励对比上：** 所有策略均采用动力学随机化进行训练。不带历史信息的 FF 策略（标记为 FF（原版））和带有观测历史的 FF 策略，其奖励曲线相似。长短期记忆网络策略起初获得的奖励最低，但最终似乎收敛到大致相同的水平。两种 FF 策略都在真实机器人上进行了测试，其中 FF（原版）的表现最为理想。**下：** 不采用动力学随机化进行训练可以加快收敛速度，而且长短期记忆网络策略的奖励仍然最低。未采用动力学随机化的策略没有部署到真实机器人上。

即使策略是在未进行动力学随机化的情况下训练的，这一结论仍然成立。不过，不出所料，所有策略在未进行动力学随机化时似乎都收敛得更快。原始前馈策略和带历史信息的前馈策略的奖励曲线几乎重合。我们在训练 20000 次迭代后停止了训练，但由于对仿真动力学过拟合，LSTM策略的奖励最终仍有可能超过前馈策略。

未经过动力学随机化训练的前馈策略，在原型策略的初步结果不理想后，未部署到真实机器人上。从视觉上看，摆动腿运动与仿真机器人不同，而且机器人容易出现轻微不稳定。不过，即使没有动力学随机化，我们也没有观察到抖动或严重不稳定的情况。经过动力学随机化后，带观测历史的前馈策略在真实机器人上原地踏步时，表现似乎与原始前馈策略相近。然而，在向前行走和转向时，我们注意到前者的表现会逐渐退化——步态对称性似乎被破坏，身体振荡也似乎不断加剧。我们推测，这是由于仿真到现实 误差随时间累积所致；如果网络只观测当前状态 o_t，则不会出现这种情况。我们没有在硬件上部署长短期记忆网络策略，因为其较慢的推理速度可能会妨碍实时执行。我们预计其表现会与带历史信息的前馈策略相似。经过动力学随机化训练的原始前馈策略在所有真实机器人实验中表现最佳，因此成为我们首选的

策略架构。

我们的结果似乎与此前针对 Cassie 的研究发现存在一定矛盾 [10]，此前研究报告称，基于记忆的网络能够有效实现 仿真到现实 的成功迁移。我们认为，这可能是由于针对我们的机器人在训练期间进行的随机化范围较窄，因此基于记忆的智能体所具备的状态历史压缩能力并未带来额外益处。由于训练长短期记忆网络所需时间几乎是前馈策略的 3 倍（36 小时，而前馈策略为 12 小时），我们的结果显示，结合针对性动力学随机化使用前馈策略来实现仿真到现实方法具有潜在优势。我们还注意到，对于实现同样的目标，准确的机器人初始模型也十分重要。

VI. 讨论与结论

在本研究中，我们开发了一套用于训练全尺寸人形机器人 HRP-5P 控制策略的系统。我们发现，对于此类大型机器人而言，主要的 仿真到现实 差距源于电机控制系统的力矩跟踪性能较差，而这又是由较高的齿轮传动比导致的。我们对反电动势进行了仿真，并向策略中加入力矩反馈，以弥补仿真到现实差距。策略在仿真中完成训练后，直接迁移到了硬件上。

我们的实验表明，提供电流反馈是实现可靠仿真到现实迁移的关键因素。在没有所提出的反馈信号时，策略在控制摆腿运动方面容易失败，常常导致双腿之间发生自碰撞。如果不在训练过程中模拟较差的力矩跟踪效果，我们无法实现仿真到现实成功。对于配备关节级力矩传感器的机器人，我们相信，所提出的方法通过考虑关节中的摩擦力矩，能够实现更好的性能。

我们还对基于记忆的策略架构的必要性进行了消融分析。结果表明，当动力学随机化在较窄范围内进行时，前馈多层感知机足以将在仿真中学习到的策略成功迁移。由于长短期记忆网络等基于记忆的网络训练起来可能计算成本更高，并且容易过拟合，因此在本研究中，我们选择多层感知机策略用于真实机器人部署。

我们无需对奖励函数进行激进的手动调节，也无需将动力学变量随机化到宽泛且不合理的范围（文献中往往会省略这一点），便实现了零样本迁移。这表明，使用准确的机器人模型进行训练，以及仔细识别克服现实差距所需的关键因素，可能具有显著的效果。然而，策略似乎会使摆动腿产生较大的摆动运动；通过降低髋部侧倾关节的PD增益，并放宽关节速度奖励项中惩罚项的系数，可以消除这种现象。

我们在真实的人形机器人平台上，将强化学习策略与一种传统的基于模型的双足运动方法进行了比较，并取得了令人鼓舞的结果。强化学习策略能够应对高度超过 3 厘米的障碍物，而使用基于模型的控制器时，机器人遇到高度超过 2 厘米的障碍物便会失去平衡并摔倒。尽管有更新的基于模型的方法能够应对更高的障碍物，但我们的测试为强化学习方法所提供的柔顺关节跟踪、闭环控制和宽松的轨迹约束提供了有力证据。从视觉效果来看，与强化学习策略下明显过度摇摆的步态相比，基于模型的控制器使机器人的步态更加流畅、高效。未来，可以通过采用参考动作并微调学习参数来解决这一问题。

我们公开了用于在 MuJoCo 中进行强化学习训练和评估的源代码，以便复现 [3]。

未来工作。未来，我们计划扩展框架，以开发用于后退运动的策略，并应对更加复杂的地形。我们也希望找出并克服妨碍仿真到现实迁移效果提升的其他因素。

致谢

作者感谢 JRL 的所有成员在开展本研究期间进行的机器人实验中给予的支持。我们尤其感谢 Hiroshi Kaminaga、Mitsuharu Morisawa 和 Mehdi Benallegue，感谢他们进行的许多富有见地的讨论。本研究部分获得 JST SPRING 奖学金计划资助，项目编号 JPMJSP2124。

参考文献

[1] M. Murooka、M. Morisawa 和 F. Kanehiro，《基于预瞄控制的人形机器人多接触运动质心轨迹生成与稳定》，*IEEE*《机器人与自动化快报》，第 7 卷第 3 期，8225–8232 页，2022 年。[2] G. Romualdi、S. Dafarra、G. L'Erario、I. Sorrentino、S. Traversaro 和 D. Pucci，《用于人形机器人运动并带有步态调整的在线非线性质心模型预测控制》，载于*2022* 年国际机器人与自动化会议（*ICRA*）。IEEE，2022 年，第 10 412–10 419 页。[3] G. B. Margolis、G. Yang、K. Paigwar、T. Chen 和 P. Agrawal，《通过强化学习实现快速运动》，*arXiv* 预印本 *arXiv:2205.02824*，2022 年。[4] J. Hwangbo、J. Lee、A. Dosovitskiy、D. Bellicoso、V. Tsounis、V. Koltun 和 M. Hutter，《为足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，2019 年。[5] A. Kumar、Z. Fu、D. Pathak 和 J. Malik，《Rma：足式机器人的快速运动适应》，*arXiv* 预印本 *arXiv:2107.04034*，2021 年。[6] J. Lee、J. Hwangbo、L. Wellhausen、V. Koltun 和 M. Hutter，《学习足式机器人在复杂地形上的四足运动》，《科学·机器人》，第 5 卷，第 47 期，2020 年。 [在线]。网址：https://robotics.sciencemag.org/content/ 5 47 5986 / /eabc Z. [7] Xie、P. Clary、J. Dao、P. Morais、J. Hurst 和 M. van de Panne，《Cassie 的运动技能学习：迭代设计与仿真到现实》，载于机器人学习会议论文集（*CORL 2019*），2019 年。[8] J. Siekmann、Y. Godse、A. Fern 和 J. Hurst，《通过周期性奖励组合实现所有常见双足步态的仿真到现实学习》，载于*2021* 年 *IEEE* 机器人与自动化国际会议（*ICRA*）。IEEE，2021 年，第 7309–7315 页。[9] J. Siekmann、K. Green、J. Warila、A. Fern 和 J. Hurst，《通过仿真到现实强化学习实现盲行双足楼梯通行》，*arXiv* 预印本 *arXiv:2105.08328*，2021 年。[10] J. Siekmann、S. Valluri、J. Dao、L. Bermillo、H. Duan、A. Fern 和 J. Hurst，《学习用于人尺度双足运动的基于记忆的控制》，*arXiv* 预印本 *arXiv:2006.02402*，2020 年。[11] W. Yu、V. C. Kumar、G. Turk 和 C. K. Liu，《双足运动的仿真到现实迁移》，*arXiv* 预印本 *arXiv:1903.01390*，2019 年。[12] S. Masuda 和 K. Takahashi，《在无扭矩传感器齿轮驱动人形机器人上进行鲁棒柔顺双足运动的仿真到现实学习》，*arXiv* 预印本 *arXiv:2204.03897*，2022 年。

[3]https://github.com/rohanpsingh/LearningHumanoidWalking

[13] K. Kaneko、H. Kaminaga、T. Sakaguchi、S. Kajita、M. Morisawa、I. Kumagai 和 F. Kanehiro，《人形机器人 HRP-5P：采用高功率、大运动范围关节的电驱动人形机器人》，*IEEE*《机器人与自动化快报》，第 4 卷，1438，2019 2，1431 第 – 页。[14] X. B. Peng、E. Coumans、T. Zhang、T.-W. Lee、J. Tan 和 S. Levine，《通过模仿动物学习敏捷的机器人运动技能》，*arXiv* 预印本 *arXiv:2004.00784*，2020 年。[15] C. Yang、K. Yuan、Q. Zhu、W. Yu 和 Z. Li，《自适应足式运动的多专家学习》，《科学·机器人》，第 5 卷，第 49 期，文章编号 eabb2174，2020 年。[16] A. Kumar、Z. Li、J. Zeng、D. Pathak、K. Sreenath 和 J. Malik，《将快速运动适应拓展至双足机器人》，载于*2022* 年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）。IEEE，2022 年，第 1161–1168 页。[17] G. A. Castillo、B. Weng、W. Zhang 和 A. Hereid，《基于强化学习的级联运动策略设计，实现鲁棒三维双足运动》，*IEEE*开放获取，第 10 卷，第 20 135–20 148 页，2022 年。[18] L. Krishna、U. A. Mishra、G. A. Castillo、A. Hereid 和 S. Kolathaya，《学习线性策略，实现不同坡度地形上的鲁棒双足运动》，载于*2021* 年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）。IEEE，2021 年，第 5159–5164 页。[19] I. Radosavovic、T. Xiao、B. Zhang、T. Darrell、J. Malik 和 K. Sreenath，《利用 Transformer 学习人形机器人运动》，*arXiv* 预印本 *arXiv:2303.03381*，2023 年。[20] D. Rodriguez 和 S. Behnke，《DeepWalk：通过深度强化学习实现全向双足步态》，载于*2021* 年 *IEEE* 机器人与自动化国际会议（*ICRA*）。IEEE，2021 年，第 3033–3039 页。[21] P.-W. Chou、D. Maturana 和 S. Scherer，《利用贝塔分布改进连续控制中的随机策略梯度与深度强化学习》，载于国际机器学习会议。PMLR，2017 年，第 834–843 页。[22] R. P. Singh、M. Benallegue、M. Morisawa、R. Cisneros 和 F. Kanehiro，《为人形机器人学习基于规划步点的双足行走》，载于*2022* 年 *IEEE-RAS* 第 *21* 届人形机器人国际会议（*Humanoids*）。IEEE，2022 年，第 686–693 页。[23] D. Kim、G. Berseth、M. Schwartz 和 J. Park，《通过仿真到现实迁移，在双足机器人上实现基于扭矩的深度强化学习，以进行任务与机器人无关的学习》，*arXiv* 预印本 *arXiv:2304 09434*，2023。。[24] M. Schwartz、J. Sim、J. Ahn、S. Hwang、Y. Lee 和 J. Park，《人形机器人 TOCABI 的设计》，载于*2022* 年 *IEEE-RAS* 第 *21* 届人形机器人国际会议（*Humanoids*）。IEEE，2022 年，第 322–329 页。[25] J. Tan、T. Zhang、E. Coumans、A. Iscen、Y. Bai、D. Hafner、S. Bohez, and V. Vanhoucke，“仿真到现实：为四足机器人学习敏捷运动”，*arXiv*预印本 *arXiv:1804.10332*，2018年。[26] C. Yang、K. Yuan、W. Merkt、T. Komura、S. Vijayakumar和Z. Li，“为人形机器人学习全身运动技能”，载于*2018*年*IEEE-RAS*第*18*届人形机器人国际会议（*Humanoids*）。IEEE，2018年，第270–276页。[27] J. Englsberger、C. Ott和A. Albu-Sch¨affer，“基于发散运动分量的三维双足行走控制”，*IEEE*机器人学汇刊，第31卷，第2期，第355–368页，2015年。[28] S. Caron、A. Kheddar和O. Tempier，“采用全身导纳控制实现HRP-4人形机器人的爬楼梯稳定控制”，载于*2019*年国际机器人与自动化会议（*ICRA*）。IEEE，2019年，第277–283页。[29] S. Kajita、F. Kanehiro、K. Kaneko、K. Fujiwara、K. Harada、K. Yokoi和H. Hirukawa，“使用零力矩点预瞄控制生成双足行走模式”，载于*2003*年*IEEE*机器人与自动化国际会议（第*03CH37422*号），第2卷。IEEE，2003年，第1620–1626页。[30] S. Kajita、M. Morisawa、K. Miura、S. Nakaoka、K. Harada、K. Kaneko、F. Kanehiro和K. Yokoi，“基于线性倒立摆跟踪的双足行走稳定控制”，载于*2010*年*IEEE/RSJ*智能机器人与系统国际会议。IEEE，2010年，第4489–4496页。[31] E. Todorov、T. Erez和Y. Tassa，“MuJoCo：用于基于模型控制的物理引擎”，载于*2012*年*IEEE/RSJ*智能机器人与系统国际会议。IEEE，2012年，第5026–5033页。[32] R. Cisneros、M. Benallegue、R. Kikuuwe、M. Morisawa和F. Kane- hiro，“对存在高静摩擦的关节摩擦力矩进行可靠、无颤振仿真”，载于*2020*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2020年，第6318–6325页。[33] 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。[34] S. Nakaoka，“Choreonoid：基于集成式GUI框架构建的可扩展虚拟机器人环境”，载于*2012*年*IEEE/SICE*系统集成国际研讨会（*SII*）。IEEE，2012年，第79–85页。[35] R. P. Singh、P. Gergondet和F. Kanehiro，“mc-mujoco：在MuJoCo中使用有限状态机控制器仿真关节型机器人”，载于*2023*年*IEEE/SICE*系统集成国际研讨会（*SII*）。IEEE，2023年，第1–5页。[36] N. Ando、T. Suehiro、K. Kitagaki、T. Kotoku和W.-K. Yoon，“RT组件对象模型：用于RT（机器人技术）的RT-Middleware分布式组件中间件”，载于*2005*年机器人与自动化计算智能国际研讨会。IEEE，2005年，第457–462页。