arXiv:2304.01159v1 [cs.RO] 3 Apr 2023

DribbleBot：野外动态足式操作

Yandong Ji*、Gabriel B. Margolis*和 Pulkit Agrawal

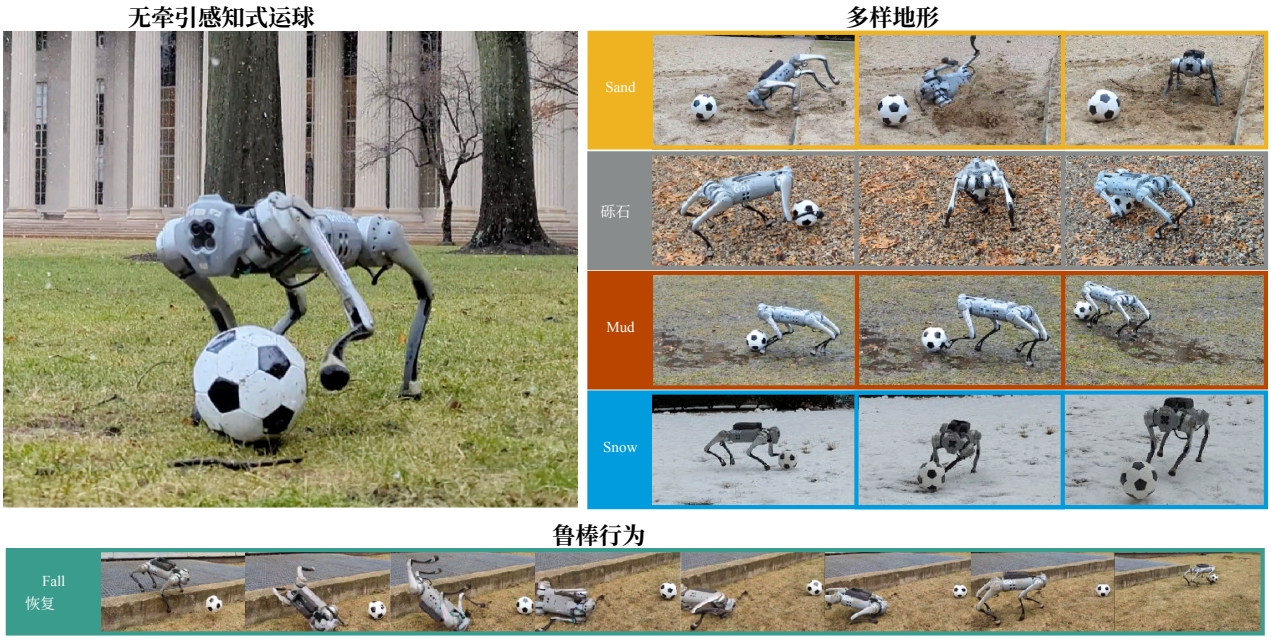


图 1：利用机载感知与计算，在沙地、砾石、泥地和雪地等多种自然地形上进行野外 **带球** 。我们的系统通过强化学习训练，能够成功适应不同地形下变化的球体动力学；在受到极端扰动而跌倒后，还能站起并找回球。

摘要——**DribbleBot（足式机器人灵巧球体操作）是一种足式机器人系统，能够在与人类相同的现实世界条件下（即野外）带足球。我们采用在仿真中通过强化学习训练策略，再将其迁移到现实世界的范式。我们克服了若干关键挑战：如何应对不同地形上多变的球体运动动力学，以及如何在机载计算受限的情况下，利用机身安装摄像机感知球。我们的结果表明，当前的四足平台非常适合研究仅依据感知观测、同时涉及运动与操作的动态全身控制问题。视频和代码见 https://gmargo11.github.io/dribblebot。**

I. 引言

动态移动操作是一类复合任务，需要视觉感知、动态运动和物体操作紧密协同。在包裹配送或搜索与救援等应用中，机器人通常需要携带物体快速移动。由于时间限制，机器人可能需要在移动过程中操作物体，而不是将物体固定在机器人坐标系中的某个位置。

例如，近期的一项演示展示了人形 Atlas 机器人在建筑工地上捡起重物、边跑边携带重物，并将其抛出 [1]。过去，动态移动操作研究一直受到限制，因为这类研究需要昂贵的硬件，而且在估计机器人、物体和环境状态的同时，还要优化运动与操作目标，复杂度很高；在动态场景中，这一挑战尤为突出。因此，先前研究主要关注运动与操作各自的子问题，或关注二者结合后动态性较低的问题。然而，过去几年里，配备视觉和本体感受传感器的足式机器人硬件变得更容易获得。强化学习已成为解决复杂接触密集型问题的强大范式，例如让足式机器人在楼梯、徒步小径、沙地和泥地等复杂地形上稳健运动 [2]–[11]，以及使用灵巧手进行动态物体操控 [12]–[15]。本研究探讨能否将同样的方法拓展到动态移动操作任务中。

作为一个案例研究，我们考虑在野外环境中带足球的任务。这需要机器人各关节面向任务进行协调（全身控制 [16]），在踢球并追球的同时保持平衡。人类运动员则——

* 表示贡献相同。所有作者均供职于美国麻省理工学院 Improbable AI Lab。通讯作者：{yandong, gmargo}@mit.edu

依靠机载感知，能够在草地、泥地、雪地和铺装路面等各种自然地形上动态控球。相比之下，先前关于足式机器人足球的研究受到以下一项或多项限制：(i) 使用平坦光滑的比赛场地表面，无需应对地形变化 [17]–[23]；(ii) 在室内使用多个静态外部摄像机，从而无需处理嘈杂的机载视觉感知和光照变化 [21]–[23]；(iii) 不同时进行运动和踢球，而是用身体推球 [17]–[19]，或进行静态带球 [20]–[22]，即每次踢球前让球停下来。我们提出了一种克服这些限制的方法。

将足球带球从实验室带入野外环境，并非只是将运动与操作技术加以融合，还面临若干独特挑战。其中之一是适应球与地形动力学：由于球的重量较轻，且滚动接触与脚和地面的接触性质不同，球与地形动力学会独立于机器人与地形动力学而变化。在铺装路面上，球可能滚得比机器人跑得更快；在草地上，球会迅速减速，因此需要更频繁、更有力地踢球。我们通过为模拟器添加自定义球体阻力模型来应对这一挑战。另一个需要考虑的问题是机载摄像机的精度和探测范围有限。当小型机器人在身体附近带球时，由于机身安装摄像机的视野狭窄，很难对球进行定位。因此，我们使用来自多个广角鱼眼摄像机的观测，并为其开发了独立的球体检测模块，以促进从模拟到现实的迁移。最后，机器人可能会在运球控制器训练分布之外的复杂地形上发生运动失败，例如遇到陡峭路缘。为应对此类场景，我们集成了一个针对更严苛条件训练的恢复策略，使机器人在跌倒后能够自主站起。我们发现，该控制器有助于重新控球并继续带球。

由此构建的系统 DribbleBot（足式机器人灵巧控球）能够在多种地形上展示动态的真实世界带球动作。通过

通过提供证据表明现有硬件和传感器能够实现成功的行为，我们希望推动机器人足球领域的更多研究，并更广泛地促进对动态移动操作问题的研究。

II. 材料

硬件：所有实验均使用 Unitree Go1 机器人 [24] 和一颗 3 号足球。这款小型四足机器人高 40 厘米。我们使用两台机载 210° 视场鱼眼相机采集图像，一台朝前，一台朝下。所有计算均由两台机载 NVIDIA J etson Xavier NX 单元完成。受机器人计算能力、通信带宽和电力限制，我们在每台 Jetson 单元上本地处理满分辨率图像，仅将足球位置估计发送到运行策略的计算机。

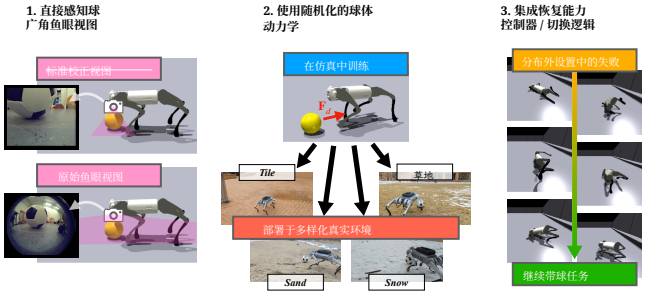


图 2：**DribbleBot 中为仿真到现实迁移所做的改进。**

模拟器：我们使用制造商提供的 URDF 模型，在 Isaac Gym [25]中仿真 Unitree Go1 机器人。仿真和训练在单台 NVIDIA RTX 3090 上运行。

目标检测模块：我们使用在 COCO 数据集 [28]上预训练的 YOLOv7 [26], [27]模型权重，并对其进行微调，以执行第 III-B.1 节所述的球检测。我们使用 TensorRT 加速目标检测，从而能够以 30Hz 的速度轻松处理分辨率为 400×480 像素的图像。

III. 方法

概述：我们在仿真中训练控制策略 $\mathbf{a}_t = \pi_d(\mathbf{o}_t, \mathbf{c}_t)$，并将其迁移到现实世界。观测$\mathbf{o}_t$ 由本体感知数据和球位置$\mathbf{b}_t$组成。球速度指令$\mathbf{c}_t$ 作为输入提供（例如，部署期间由人类用户提供）。我们没有选择在摄像机图像上端到端训练策略，因为这会面临诸多挑战，包括仿真速度慢、基于高维视觉观测进行训练时样本效率低，以及图像观测中的仿真到现实差距。相反，我们使用在仿真中易于获取的球位置来训练策略。在真实世界部署时，单独训练的目标检测模型（$\mathbf{Y}$）会根据机载摄像头拍摄的图像（$\mathbf{o}_t^v$）预测球位置：$\hat{\mathbf{b}}_t = \mathbf{Y}(\mathbf{o}_t^v)$。策略 π_d 输出动作$\mathbf{a}_t$，即 12 个电机（每条腿 3 个电机）的关节位置目标，频率为 50Hz。 π_d 采用近端策略优化（PPO）强化学习算法进行训练 [29]。

A. 带球策略训练

1) 环境设计： 我们在仿真中训练机器人在平坦地面上带球，并按照第八节所述随机改变物理参数。每个回合开始时，机器人的偏航方向随机初始化，其初始腿部位置则在标称姿态附近随机设定。足球在机器人周围 2 m 范围内的随机位置初始化。目标球的速度也均匀随机设定。这些设置确保机器人学会全向运动和带球。每个回合时长为 40 s，控制时间步长为 50 Hz。

2）野外环境中的带球控制接口：成功带球需要在机器人移动、保持平衡并调整自身位置以应对移动中的球时，调整腿部摆动以施加目标力。以往关于仿真到现实的腿式运动研究通常采用身体速度

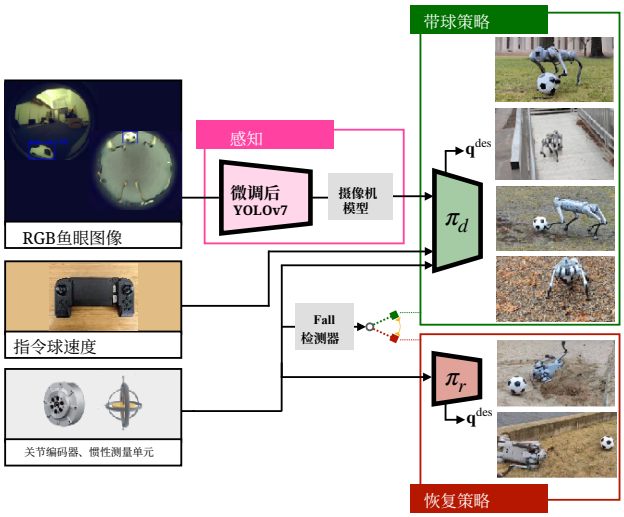


图 3：**DribbleBot 的系统架构**。 π_d 和 π_r 是通过强化学习在模拟环境中训练的多层感知机。YOLOv7 是一种目标检测网络 [27]，我们使用监督学习在来自我们所在领域的图像上对其进行微调。

命令或位置 [30] 作为控制接口，只有少数例外允许用户调整步态 [9], [31] 或足部落点 [32] 参数。仅用步态或身体层级的控制指令难以表达盘带技能，而且对于部署机器人的人类用户而言，发出此类指令并不直观。相反，我们直接命令球在全局参考坐标系中、二维地面平面上的线速度。由于球具有完全的旋转对称性，而且机器人在踢球动作中朝向可能迅速变化，机器人和球的局部参考坐标系都会不断变化，对人类操作者而言不太实用。我们选择在全局坐标系中向机器人发出命令，因为全局坐标系不会随机器人的运动而变化，因此人类更容易控制机器人。机器人在第一个时间步的局部机体坐标系作为全局参考坐标系。

3）观测与动作空间：策略π_d的输入$\mathbf{o}_t$由以下内容组成：命令$\mathbf{c}_t$的 15 步历史记录、球位置$\mathbf{b}_t$、关节位置和速度$\mathbf{q}_t$, $\dot{\mathbf{q}}_t$、机体坐标系中的重力单位向量$\mathbf{g}_t$、全局机体偏航角ψ_t，以及 [9]中定义的时序参考变量$\boldsymbol{\theta}_t^{\text{cmd}}$。命令$\mathbf{c}_t$由全局坐标系中的目标球速度$\mathbf{v}_x^{\text{cmd}}$,$\mathbf{v}_y^{\text{cmd}}$组成（第 III-A.2 节）。为了让机器人处理全局坐标系中的命令，它必须知道自身在全局坐标系中的朝向，因此我们将从惯性测量单元获取的全局机体偏航角ψ_t作为输入提供给策略。动作空间$\mathbf{a}_t$是由高频比例微分控制器维持的十二个关节的目标位置，控制器参数为 $k_p = 20.0$、 $k_d = 0.5$ [9]。

4）奖励模型：表 III（附录）列出了用于控球的任务奖励项。任务奖励包括一项奖励，用于跟踪全局参考坐标系中的球的速度，以及一项奖励塑形项，鼓励机器人靠近球。由于摄像机的观测范围有限，为了让球更容易出现在摄像机视野中，机器人面向球时也会获得奖励。一组步态奖励项 [9], [31] 鼓励机器人采用一致的地面接触时序，并生成良好的步态。此外，还加入了额外的标准安全奖励项 [8],[9], [33], [34]，以惩罚危险指令并促进仿真到现实迁移。

5）策略架构与优化：我们使用近端策略优化（PPO） [29] 训练足球带球策略，该策略由一个包含三个全连接隐藏层的神经网络表示，隐藏层大小为 $[512, 256, 128]$。经过 70 亿个时间步（约 48 小时训练）后，策略收敛。为了便于可视化和调试，最好能够获取状态变量：机器人的身体速度、球的速度以及球地面阻力系数。我们在仿真中训练了一个由两层神经网络表示的回归模型，其包含 $[256, 128]$ 个单元，用于根据$\mathbf{o}_t$预测这些参数 [34] （第 III-A.3 节）。此外，还将这些预测参数加入策略网络的输入中。尽管先前研究发现这项技术有助于跑步策略的仿真到现实迁移 [34],，但我们并未评估它对带球任务仿真到现实性能的影响。

B. 缓解仿真到现实差距的措施

1) 鱼眼图像中的视觉感知：当足球距离机器人最近仅 10 cm 时，DribbleBot 需要通过机身安装摄像机定位一个 3 号足球（直径 18 cm）。由于视野狭窄，RealSense 等标准深度摄像头无法满足这一需求（105°）。球靠近时，这类摄像头只能拍到足球的一小部分，不足以对其进行定位（图 2）。因此，我们的机器人配备了超广角前向和下向鱼眼摄像头，各自具有210°的视野。我们没有在策略学习期间模拟这些图像，因为这会拖慢学习速度，并导致较大的仿真到现实差距；取而代之的是，我们基于 YOLOv7 [27]，构建了一个独立的足球检测模块，该模块采用在大型图像数据集上预训练的目标检测模型 [28]）。由于视野极广，校正会使球形足球产生明显变形，因此，在网上获取的窄视野校正图像数据集上训练的模型，其现成模型性能并不理想。不过，我们发现，只需使用来自机器人摄像机的 254 张人工标注的足球鱼眼图像（包括足球位于画面边缘的图像）对 YOLOv7 进行微调，就能轻松恢复整个视野范围内的良好检测性能。我们对图像采用了标准的数据增强方法，包括水平翻转、最多 15° 的旋转以及模糊处理。

目标检测会输出一个边界框，但我们的控制策略接收的输入是球位置。我们通过近似应用等距鱼眼镜头模型，从单目RGB图像中获取球位置。给定足球像素坐标，我们首先使用等距鱼眼模型计算摄像头主轴到足球中心的角度 θ，其中 $r = f\theta$ 中的 r 是图像中心的像素距离，f 是摄像头焦距。忽略变形效应后，我们根据其边界框计算图像中球的像素尺寸 Δr，并且已知其物理

球半径 R，我们再次应用等距鱼眼模型，计算球与摄像机之间的近似距离 z：$z = f\frac{R}{\Delta r}$。已知摄像机到球的距离以及从摄像机指向球的方向，即可确定球在相机坐标系中的位置。我们利用已知的相机外参，将观测到的球位置转换到机体坐标系中。最后，如果前置摄像机和底部摄像机都检测到球，则选择置信度更高的检测结果，得到一个位置估计。

2) 视觉噪声模型：在仿真中，可以获得准确的球位置估计；而在现实世界中，球位置估计存在噪声。为确保策略能够应对视觉感知噪声，我们在训练期间对球位置添加从均匀分布中采样的噪声。此外，为模拟因人踢球或球移出机器人视野而可能发生的球位置大幅变化，我们还会在地面平面上随机传送球。最后，由于摄像机的数据速率有限，我们会模拟相机通信延迟。噪声模型的详细信息见第八节（附录）。

3) 机器人系统辨识：为缩小机器人动力学方面的仿真到现实差距，我们采用两种标准且有效的系统辨识措施：(a) 使用现实世界的扭矩数据训练执行器网络，以补偿非理想的电机动力学 [9], [15], [35]；以及 (b) 辨识并建模观测测得时刻与动作施加时刻之间的延迟 [9], [15], [36]。

4) 球与地形交互模型：我们使用的刚体仿真器未对足球与不同地形滚动接触时产生的可变阻力进行建模。人类足球运动员能够迅速适应球体动力学的变化，这些变化可能由气压、尺寸等物理因素，以及不平坦地形或对方球员对球造成的外部扰动引起。我们通过实施自定义域随机化方案，让 DribbleBot 具备鲁棒性。我们的域随机化包括一种球体阻力模型，该模型施加的阻力与速度的平方成正比：$F_D = C_D v^2$，遵循空气阻力的标准方程。C_D 取不同值，以模拟阻力各异的地形，例如高草场地（C_D较高）或铺装路面（C_D较低）。此外，我们还会随机化球体质量，并在训练期间对球的速度施加随机变化。球速随机化用于模拟球所受的外部扰动，例如人为干预或与不平坦地形接触。随机化范围的详细信息见表 IV（附录）。

5）跌倒恢复控制器：如果机器人遭遇强烈扰动，例如被推搡或遇到陡峭路缘，并发生运动失败，也会导致控球失效。在这种场景下，我们希望机器人能够从跌倒中站起来并恢复带球。与先前研究类似 [35], [37], ，我们训练了一种专用恢复策略，使机器人能够从各种跌倒场景中恢复到站立姿势。该恢复策略在

Tile		草地	
Full	4/4	Full	4/4
-R	4/4	-R	4/4
-Y	0/4	-Y	0/4
-D	4/4	-D	4/4
Sand		**Snow**	
Full	4/4	Full	3/4
-R	4/4	-R	3/4
-Y	0/4	-Y	0/4
-D	2/4	-D	-
路缘石降坡		**Ramp**	
Full	2/4	Full	0/4
-R	1/4	-R	0/4
-Y	-	-Y	-
-D	-	-D	-

表 I：**真实世界运球性能评估**。机器人在每次试验中执行固定的运球轨迹。我们在每种场景下测试完整系统设计（满）及以下消融设置：无恢复控制器（-R）；无 YOLO 微调（-Y）；训练期间使用无阻力模型（-D）

更严苛的环境中训练，包括不平坦地形和随机重力；这些场景对运球控制器而言过于困难。我们首先通过从不同姿态随机放倒机器人，生成一组包含 1000 种初始跌倒构型的数据，然后训练策略，使其获得身体方向、基座高度和动作平滑度方面的奖励。恢复策略的奖励函数和训练流程详情见表 III（附录）。为了在带球策略和恢复策略之间切换，我们定义了一个基于身体方向进行状态转换的有限状态机。当滚转角或俯仰角大于 1.0 弧度时，表明发生运动失败，恢复策略便会执行动作，使机器人恢复到站立姿态。当滚转角和俯仰角均小于 0.5 弧度时，则重新激活带球策略。

IV. 结果

A. 仿真性能

1）运球控制：首先，我们在与训练期间相同的条件下，于仿真中评估带球策略。如果机器人带球表现良好，我们可以通过其对球速度指令的跟踪精度、带球速度以及转弯的灵敏程度来评估其性能。图 4 展示了一段持续 60 秒的长时间运行，并呈现了运球路径、指令跟踪性能和关节运动。相应的行为见视频 S1。

在仿真实验中，机器人能够跟踪最高达 1.5m/s 的运球速度，并覆盖指令方向瞬时变化的整个范围，最高可达 180°。与常规运动不同，改变球的速度这一动作可能持续数秒，在此期间，机器人会建立对球的控制并完成多次踢球（图 4b、4c）。

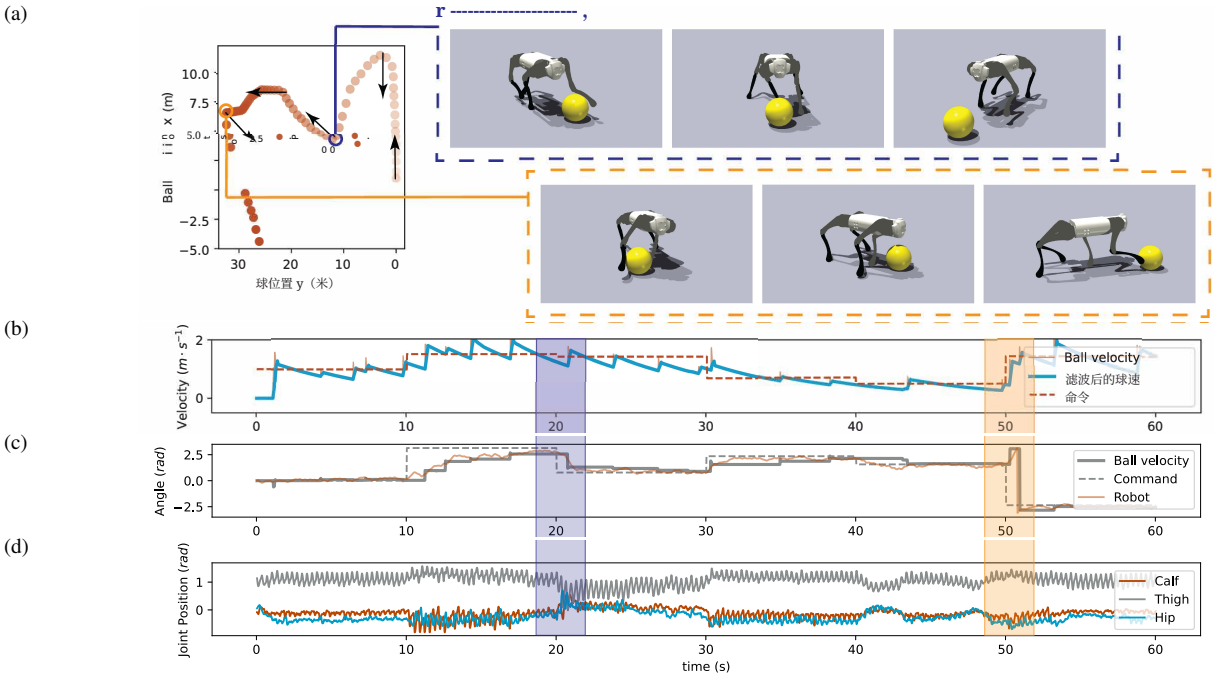


图 4：**仿真运球表现评估**。(a) 世界坐标系中的球位置及转弯时刻快照。红点表示球的位置，并随着时间推移颜色逐渐加深。深色箭头表示指令速度的大致方向。(b)、(c) 极坐标中的球速跟踪表现。这里，首先命令机器人以 1m/s 的速度向前带球，然后以不同速度依次转弯。(d) 我们展示了右前腿的关节位置，该腿通常用于带球和左转。约 20 s 处的蓝色高亮对应于上方图像第一行所示的左转。约 50 s 处的橙色高亮对应于第二行所示的右转。

我们观察到，以这种方式动态带球需要机器人所有关节进行面向任务的协调（全身控制）。图 4d 显示，在带球时，右前腿在左转和右转中的参与方式各不相同。左转时，这条腿用于踢球（蓝色高亮，图 4a 上方图像），其动作在这一过程中发生了显著变化。随后，机器人右转时（橙色高亮，图 4a 下方图像），左前腿用于踢球，而右腿的动作也会相应调整，以稳定这一动作。

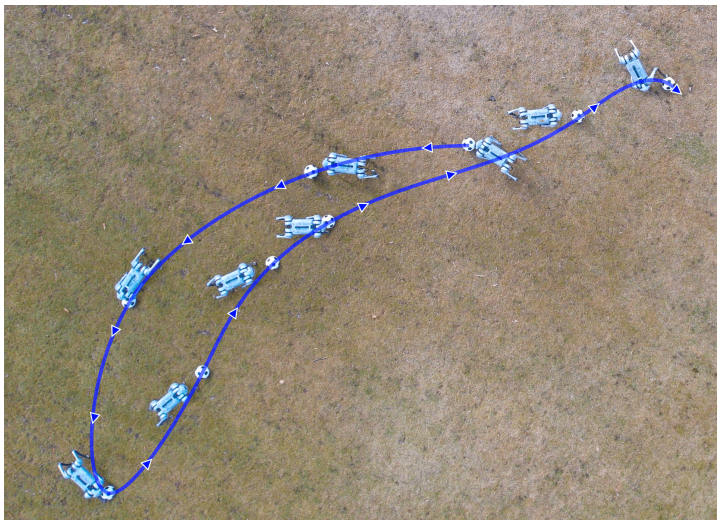
B. 真实世界部署

1) 多样地形上的定性结果：我们通过遥操作，在具有不同球与地形动力学特性的多样地形上，对运球控制器进行了定性评估。当地形导致足部打滑或绊 stumbling 时，运动控制器必须进行适应。带球还要求根据球与地形的相互作用来调整奔跑和踢球动作；由于球的质量较轻且会与表面发生滚动接触，这种相互作用可能与机器人足部和地形之间的作用大不相同。例如，在草地上，较大的阻力往往会使滚动的球减速；在铺装路面上，较小的阻力可能会使球加速远离机器人；在砾石上，球往往会滚动，但机器人会打滑；在雪地或沙地上，球和机器人都会打滑；在崎岖不平的地形上，球撞击地形表面时会不可预测地改变方向。DribbleBot 能够在每种地形上完成带球和转向动作，并跟踪球的速度指令，这些指令由人类发送。这些测试地形如图 1 所示，完整行为请见补充视频。

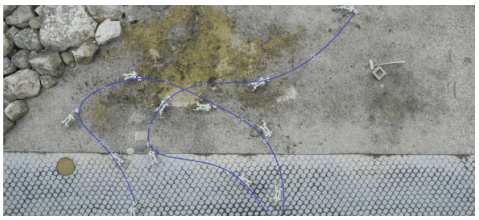
由于我们的系统无需系绳或外部传感，因此能够在广阔的户外空间中操控足球。为展示这一点，我们收集了无人机拍摄的视频，记录机器人经过 (a) 一个 180° 转弯、(b) 包含瓷砖、砾石和凹凸苔藓地面在内的一系列多样地形，以及 (c) 朝足球球门奔跑 10 米并进行躲避性转向动作的完整路线。图5展示了拼接后的俯拍照片，以呈现机器人在真实环境中的带球表现。

2) 定量结果与消融实验：我们对 DribbleBot 在多样地形上沿预设轨迹运动时的全自主行为进行了定量评估。我们向机器人下达预定轨迹命令：以 1.5 m/s 的速度向前带球 10 s，然后停球 5s，接着以 1.5 m/s 的速度返回起始线，直至抵达该线。如果机器人失去对球的控制，我们就将该次试验计为失败；但如果失控是由机器人跌倒造成的，我们允许它自主恢复并继续尝试。

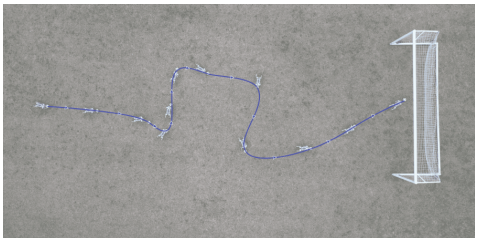
如表 I 所示，机器人使用完整系统设计，在瓷砖地面、草地和沙地上连续四次成功完成动作。雪地、下台阶和坡道的挑战性依次增加，表现也随之下降。由于训练期间从未接触过台阶或坡道，它们对我们的策略而言属于分布外地形。在下台阶任务中，机器人曾一次跌进一堆雪里且未能恢复，还曾在跌倒时把球远远撞开



(a) 180° 转向机动。



(b) 在瓷砖、砾石和凹凸苔藓地面上控球。



(c) 带有规避性转向地带球冲向球门。

图 5：**DribbleBot 在真实世界部署期间的俯视图像。**

，恢复后也无法感知到球。在遥操作下，机器人成功通过了坡道；但在标准实验中，预先指定带球指令时，机器人并未取得明显的前进进展。每次踢球后，球都会滚回坡道下方，机器人不得不转身将球找回。

我们还进行了消融研究，以量化各项设计选择对真实世界性能的影响。我们按照上述相同方法评估消融配置：无恢复控制器（-R）；无 YOLO 微调（-Y）；训练期间无自定义球体阻力模型（-D）。消融结果表明，YOLO 微调（-Y）对性能至关重要，而恢复策略（-R）在具有挑战性的路缘环境中帮助机器人完成了一次运行。未使用球体阻力模型（-D）的系统在草地和瓷砖上都能保持控球，但补充视频显示，尽管速度指令相同，它在草地上的带球速度明显更慢。在沙地上，未使用阻力模型的策略在转向机动时因未能踢到球而两次失败。这表明，阻力模型带来的额外鲁棒性也可能有助于应对由湿滑地形等其他变化导致的意外球体运动。

3）与人类协作及涌现行为：一场真正的足球比赛并非独自进行，而是有另一位试图控制球的参与者。为了了解机器人在这一场景中的行为，我们探索了机器人与人类伙伴进行交互式比赛的设置。与典型的运动任务不同，控制球的速度这一任务赋予机器人很大的行为自由度，即使用户没有改变命令也是如此。成功的带球并非单一技能：机器人往往需要进行长时间、非周期性的动作，以靠近球、调整身体朝向以便踢球；如果球因意外扰动、暂时性的视觉感知故障或控制故障而丢失，还需要折返追球。我们发现，DribbleBot 会表现出这些行为，以便在人机互动过程中保持控球。

V. 相关工作

A. 足式机器人的足球技能

足球长期以来一直是机器人研究人员感兴趣的领域。今年迎来第 26 届赛事的机器人世界杯竞赛，每年都吸引数千名参与者。在过去的 [17]–[19] 年里，机器人世界杯参赛队伍已采用有效的基于规则的方法进行踢球、传球和射门。

近期，一些研究已将学习方法应用于仿真环境 [38]–[40] 和外部仪器化室内环境 [22], [23], [41]中的足式球操控任务。 [41] 展示了一种仰卧的四足机器人如何用腿控制并改变球的方向。研究者还利用强化学习，为物理仿真角色展示了多种足球技能，例如带球 [38], [42] 和颠球 [40], 。 [21] 利用模仿学习，在动作捕捉辅助下于现实世界的室内环境中实现了静态带球。 [22] 将分层框架应用于现实世界中的足球射门任务，选择前右脚的贝塞尔曲线参数作为低层命令输入，并通过真实世界微调阶段提高射门精度。 [23] 在外部仪器化的实验室环境中，利用仿真到真实强化学习训练了一个用于跳起拦截来球的控制策略。 [43] 在仿真训练期间，通过渲染目标实验室环境神经辐射场中的图像，直接根据机载RGB图像学习人形机器人带球策略。

除了低层技能学习，一些研究还致力于端到端地学习高层足球比赛。值得注意的是，[44]通过先利用人类足球运动员的动作捕捉视频片段预训练低层技能，再使用强化学习在低层控制空间中寻找多智能体协调目标的解决方案，从而应对肌肉级控制和长时域决策问题。为了学习带球等低层技能， [44] 依赖人类足球运动员的动作捕捉数据，而这类数据并不适用于四足机器人形态。

B. 动态物体操控

先前研究探索了使用固定底座或全驱动底座动态操控物体。 [45] 通过开环策略控制机械臂盲目地进行颠球。另一项关于机器人乒乓球的研究 [46]使用扩展卡尔曼滤波器估计球状态，该滤波器内部利用了飞行和弹跳行为的模型。 [47]学习了残差物理模型，随机抓起刚体并将其扔进盒子里。 [48] 基于人类行为模型进行自举训练，并使用仿真和真实数据训练乒乓球机器人学习控制策略。

另一项相关研究方向探讨了如何利用搭载机械臂的四足机器人操控物体。 [49]利用四足机器人搭载的机械臂操控物体，并通过学习得到的估计模块协调机身与机械臂的运动。 [50]实现了一种基于模型的控制器，用于在站立姿态下通过搭载机械臂的四足机器人操控物体。 [51] 通过强化学习训练了一个端到端控制器，使搭载机械臂的四足机器人能够在遥操作下协调执行操作，并展示了利用 AprilTag 标记进行视觉引导到达的能力。这些研究探讨了动态移动操作领域中互为补充的问题。

VI. 讨论

DribbleBot 存在若干局限性，我们希望在未来的工作中进一步探索并加以改进。下面列举其中几项；失败案例的视频可在项目网站上查看。转向响应缓慢：我们的系统能够使球急转弯，但命令发出后到球实际转向之间存在延迟（图 4）。视觉感知对光照敏感：我们发现，在明亮的直射阳光下，球和摄像机上的反射会产生眩光，导致视觉感知模块表现不佳。使用更多样化的户外图像对视觉感知网络进行微调，或许可以解决这一问题。高速时精度不足：如果球在低阻力地形上移动过快，或在高速行进时突然收到急转命令，机器人可能无法成功拦停球，从而失去对球的视线。缺乏几何感知：虽然机器人能够在湿滑且不平坦的地形上运球，但无法稳定地通过陡坡和楼梯等较大的障碍物。此外，它无法感知环境中的杆和墙壁等物体。未来的工作可以将更多环境几何结构信息纳入控制器，以提升机器人在杂乱且恶劣环境中的控球能力。

我们相信，凭借可靠的真实环境中的运球基线，还有许多令人兴奋的前沿领域值得探索。带球只是足球的一部分。尤其是，要参与一场有竞争力的比赛，除了射门 [22] 和守门 [23] 技能，还需要高层次比赛策略以及对其他智能体的感知。同样，要将动态移动操作应用于配送和应急响应等实际任务，也需要多种技能、高层次规划和丰富的世界理解。真实环境中的足球或许也能成为研究人机交互的有趣场景。尽管人与足式机器人的直接身体交互通常较为有限，但通过足球这一共享媒介进行交互，既丰富又有趣。未来的工作可以探索机器人在比赛过程中在人类眼中的形象。

致谢

我们感谢 Improbable AI Lab 的成员提供有益的讨论和反馈。我们感谢 MIT Supercloud 和林肯实验室超级计算中心提供高性能计算资源。本研究得到 DARPA 机器常识计划、MIT-IBM Watson 人工智能实验室以及美国国家科学基金会的支持，资助协议为合作协议 PHY-2019786（美国国家科学基金会人工智能与基础相互作用人工智能研究所，http://iaifi.org/）。本研究还由美国空军研究实验室和美国空军人工智能加速器资助，并根据合作协议编号 FA8750-19-2-1000 开展。本文所载观点和结论仅代表作者，不应被解读为美国空军或美国政府明示或暗示的官方政策。尽管本文包含任何版权声明，美国政府仍有权为政府用途复制和分发再版。

作者贡献

- **Yandong Ji**参与了整个系统的构思与实现、实验评估和论文撰写。
- **Gabriel B. Margolis** 参与了整个系统的构思与实现、实验评估和论文撰写。
- **Pulkit Agrawal** 指导了该项目，并参与了项目开发、实验设计和论文撰写。

参考文献

[1] R. Deits 和 T. Koolen，《拾取动量》，2023年1月。[线上]。网址：https://www.bostondynamics.com/resources/blog/ picking-momentum [2] J. Tan、T. Zhang、E. Coumans、A. Iscen、Y. Bai、D. Hafner、S. Bo-hez 和 V. Vanhoucke，《仿真到现实：为四足机器人学习敏捷运动》，载于机器人：科学与系统会议（*RSS*）论文集，美国宾夕法尼亚州匹兹堡，2018年6月，第1–9页。[3] 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而动态的运动技能》，《科学机器人》，第4卷，第26期，文章编号 aau5872，2019年1月。[4] J. Lee、J. Hwangbo、L. Wellhausen、V. Koltun 和 M. Hutter，《学习在复杂地形上进行四足运动》，《科学机器人》，第5卷，第47期，文章编号 eabc5986，2020年10月。[5] A. Kumar、Z. Fu、D. Pathak 和 J. Malik，《RMA：足式机器人的快速运动适应》，载于机器人：科学与系统会议（*RSS*）论文集，线上，2021年7月。[6] J. Siekmann、K. Green、J. Warila、A. Fern 和 J. Hurst，《通过仿真到真实强化学习实现盲态双足楼梯行走》，载于机器人：科学与系统会议（*RSS*）论文集，线上，2021年7月。[7] T. Miki、J. Lee、J. Hwanbo、L. Wellhausen、V. Koltun 和 M. Hutter，《学习适用于野外环境的四足机器人鲁棒感知运动》，《科学机器人》，第7卷，第62期，文章编号 abk2822，2022年1月。[8] G. B. Margolis、G. Yang、K. Paigwar、T. Chen 和 P. Agrawal，《通过强化学习实现快速运动》，载于机器人：科学与系统会议（*RSS*）论文集，2022年6月。[9] G. B. Margolis 和 P. Agrawal，《照这样走：利用多样化行为调整机器人控制以实现泛化》，载于机器人学习会议（*CoRL*）论文集，新西兰奥克兰，2022年12月。[10] A. Agarwal、A. Kumar、J. Malik 和 D. Pathak，《利用自我中心视觉在复杂地形上进行足式运动》，载于机器人学习会议（*CoRL*）论文集，新西兰奥克兰，2022年12月。[11] S. Choi、G. Ji、J. Park、H. Kim、J. Mun、J. H. Lee 和 J. Hwangbo，《学习在可变形地形上运动》，《科学机器人》，第8卷，第74期，文章编号 eade2256，2023年。[12] O. M. Andrychowicz、B. Baker、M. Chociej、R. Jozefowicz、B. Mc-Grew、J. Pachocki、A. Petron、M. Plappert、G. Powell、A. Ray、等，《学习灵巧手内操作》，《国际机器人研究杂志》（*IJRR*），第39卷，第1期，第3–20页，2020年。

[13] OpenAI、I. Akkaya、M. Andrychowicz、M. Chociej、M. Litwin、B. McGrew、A. Petron、A. Paino、M. Plappert、G. Powell、R. Ribas、J. Schneider、N. Tezak、J. Tworek、P. Welinder、L. Weng、Q. Yuan、W. Zaremba 和 L. Zhang，《用机器人手玩魔方》，arXiv预印本，2019年。[14] T. Chen、J. Xu 和 P. Agrawal，《一种用于手内物体重新定向的系统》，载于机器人学习会议（CoRL）论文集，英国伦敦，2021年11月，第297–307页。[15] T. Chen、M. Tippur、S. Wu、V. Kumar、E. Adelson 和 P. Agrawal，《视觉灵巧性：基于深度信息的手内灵巧操作》，arXiv预印本 arXiv:2211.11744，2022年。[16] O. Khatib、L. Sentis、J. Park 和 J. Warren，《类人机器人的全身动态行为与控制》，《国际人形机器人学杂志》，第1卷，第01期，第29–43页，2004年。[17] M. Veloso、W. Uther、M. Fijita、M. Asada 和 H. Kitano，《与足式机器人一起踢足球》，载于 IEEE/RSJ智能机器人与系统国际会议（IROS）论文集，第1卷，IEEE，1998年，第437–442页。[18] P. Stone，《智能自主机器人技术：机器人足球案例研究》，《人工智能与机器学习综合讲义》，第1卷，第1期，第1–155页，2007年。[19] M. Friedmann、J. Kiener、S. Petters、D. Thomas、O. Von Stryk 和 H. Sakamoto，《人形足球机器人的多功能高质量运动与行为控制》，《国际人形机器人学杂志》，第5卷，第03期，第417–436页，2008年。[20] D. L. Leottau、J. Ruiz-del Solar、P. MacAlpine 和 P. Stone，《分层学习策略在机器人足球个体行为中的应用研究》，载于《机器人世界杯2015：第十九届机器人世界杯》，Springer，2015年，第290–302页。[21] S. Bohez、S. Tunyasuvunakool、P. Brakel、F. Sadeghi、L. Hasen-clever、Y. Tassa、E. Parisotto、J. Humplik、T. Haarnoja、R. Hafner、M. Wulfmeier、M. Neunert、B. Moran、N. Siegel、A. Huber、F. Romano、N. Batchelor、F. Casarini、J. Merel、R. Hadsell 和 N. Heess，《模仿并重新利用：从人类和动物行为中学习可复用的机器人运动技能》，arXiv预印本 arXiv:2203.17138，2022年。[22] Y. Ji、Z. Li、Y. Sun、X. B. Peng、S. Levine、G. Berseth 和 K. Sreenath，《利用四足机器人，通过分层强化学习实现精准足球射门技能》，IEEE/RSJ智能机器人与系统国际会议（IROS）论文集，2022年10月。[23] X. Huang、Z. Li、Y. Xiang、Y. Ni、Y. Chi、Y. Li、L. Yang、X. B. Peng 和 K. Sreenath，《利用强化学习打造动态四足机器人守门员》，arXiv预印本 arXiv:2210.04435，2022年。[24] 宇树科技，Go1，2023年，https://www.unitree.com/go1，[线上；访问日期：1月 2023]。[25] V. Makoviychuk、L. Wawrzyniak、Y. Guo、M. Lu、K. Storey、M. Mack-lin、D. Hoeller、N. Rudin、A. Allshire、A. Handa 和 G. State，《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，2021年。[26] J. Redmon、S. Divvala、R. Girshick 和 A. Farhadi，《只看一次：统一的实时目标检测》，载于 IEEE计算机视觉与模式识别会议论文集，2016年，第779–788页。[27] C.-Y. Wang、A. Bochkovskiy 和 H.-Y. M. Liao，《YOLOv7：可训练的免费午餐技巧集实现实时目标检测器的新最优性能》，arXiv预印本 arXiv:2207.02696，2022年。[28] T.-Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan、P. Dollár 和 C. L. Zitnick，《Microsoft COCO：上下文中的常见物体》，载于《计算机视觉——ECCV 2014：第13届欧洲计算机视觉会议，瑞士苏黎世，2014年9月6日至12日，论文集，第V部分，第13卷》，Springer，2014年，第740–755页。[29] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，《近端策略优化算法》，arXiv预印本，2017年。[30] N. Rudin、D. Hoeller、M. Bjelonic 和 M. Hutter，《通过端到端学习运动与局部导航实现高级技能》，载于2022年 IEEE/RSJ智能机器人与系统国际会议（IROS），IEEE，2022年，第2497–2503页。[31] J. Siekmann、Y. Godse、A. Fern 和 J. Hurst，《通过周期性奖励组合进行仿真到现实学习，实现所有常见双足步态》，载于 IEEE机器人与自动化国际会议（ICRA）论文集，中国西安，2021年6月，第7309–7315页。[32] H. Duan、A. Malik、J. Dao、A. Saxena、K. Green、J. Siekmann、A. Fern 和 J. Hurst，《通过仿真到现实学习实现足步约束双足动态行走》，载于 IEEE机器人与自动化国际会议（ICRA）论文集，2022年，第10 428–10 434页。

[33] 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”，载于机器人学习会议（*CoRL*）论文集，英国伦敦，2021年11月，第 91–100 页。[34] G. Ji、J. Mun、H. Kim 和 J. Hwangbo，“动态且稳健的腿式运动中控制策略与状态估计器的并行训练”，*IEEE*机器人与自动化快报（*RA-L*），第 7 卷，第 2 期，第 4630–4637 页，2022年4月。[35] J. Lee、J. Hwangbo 和 M. Hutter，“基于深度强化学习的四足机器人稳健恢复控制器”，*arXiv* 预印本 *arXiv:1901.07517*，2019年。[36] Z. Xie、X. Da、M. van de Panne、B. Babich 和 A. Garg，“重新审视动力学随机化：四足运动案例研究”，载于 *IEEE*机器人与自动化国际会议（*ICRA*）论文集，线上，2021年5月，第 4955–4961 页。[37] L. Smith、J. C. Kew、X. B. Peng、S. Ha、J. Tan 和 S. Levine，“持续学习的足式机器人：在现实世界中微调运动策略”，载于 *IEEE*机器人与自动化国际会议（*ICRA*）论文集。IEEE，2022年，第 1593–1599 页。[38] X. B. Peng、M. Chang、G. Zhang、P. Abbeel 和 S. Levine，“MCP：通过乘法组合策略学习可组合的分层控制”，神经信息处理系统进展，第 32 卷，2019年。[39] A. F. Muzio、M. R. Maximo 和 T. Yoneyama，“人形机器人行为的深度强化学习”，智能与机器人系统期刊，第 105 卷，第 1 期，第 1–16 页，2022年。[40] Z. Xie、S. Starke、H. Y. Ling 和 M. van de Panne，“利用逐层混合专家学习足球颠球技能”，载于 *ACM SIGGRAPH 2022* 会议论文集，2022年，第 1–9 页。[41] F. Shi、T. Homberger、J. Lee、T. Miki、M. Zhao、F. Farshidian、K. Okada、M. Inaba 和 M. Hutter，“Circus ANYmal：利用四肢学习灵巧操作的四足机器人”，载于*IEEE*机器人与自动化国际会议（*ICRA*）论文集。IEEE，2021年，第 2316–2323 页。[42] X. B. Peng、Z. Ma、P. Abbeel、S. Levine 和 A. Kanazawa，“AMP：用于风格化、基于物理的角色控制的对抗运动先验”，*ACM*图形学汇刊（*TOG*），第 40 卷，第 20 期，2021年，第 4:1–4:1 页。[43] A. Byravan、J. Humplik、L. Hasenclever、A. Brussee、F. Nori、T. Haarnoja、B. Moran、S. Bohez、F. Sadeghi、B. Vujatovic，等，“Nerf2real：利用神经辐射场实现视觉引导双足运动技能的仿真到现实迁移”，*arXiv* 预印本 *arXiv:2210.04932*，2022年。[44] S. Liu、G. Lever、Z. Wang、J. Merel、S. M. A. Eslami、D. Hennes、W. M. Czarnecki、Y. Tassa、S. Omidshafiei、A. Abdolmaleki、N. Y. Siegel、L. Hasenclever、L. Marris、S. Tunyasuvunakool、H. F. Song、M. Wulfmeier、P. Muller、T. Haarnoja、B. Tracey、K. Tuyls、T. Graepel 和 N. Heess，“从运动控制到模拟人形机器人足球中的团队配合”，《科学·机器人》，第 7 卷，第 eabo0235 号，2022年。[45] K. Ploeger、M. Lutter 和 J. Peters，“利用二元奖励进行现实世界颠球的高加速强化学习”，*arXiv* 预印本 *arXiv:2010.13483*，2020年。[46] K. Mülling、J. Kober、O. Kroemer 和 J. Peters，“学习在机器人乒乓球中选择和泛化击打动作”，国际机器人研究期刊（*IJRR*），第 32 卷，第 3 期，第 263–279 页，2013年。[47] A. Zeng、S. Song、J. Lee、A. Rodriguez 和 T. Funkhouser，“Tossing-bot：利用残差物理学习投掷任意物体”，*IEEE*机器人学汇刊（*T-RO*），第 36 卷，第 4 期，第 1307–1319 页，2020年。[48] S. Abeyruwan、L. Graesser、D. B. D’Ambrosio、A. Singh、A. Shankar、A. Bewley 和 P. R. Sanketi，“i-sim2real：在人机交互紧密闭环中进行机器人策略的强化学习”，*arXiv* 预印本 *arXiv:2207.06572*，2022年。[49] M. Mittal、D. Hoeller、F. Farshidian、M. Hutter 和 A. Garg，“在未知场景中通过全身移动操作实现关节物体交互”，载于*2022*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2022年，第 1647–1654 页。[50] Y. Ma、F. Farshidian、T. Miki、J. Lee 和 M. Hutter，“将基于学习的运动策略与基于模型的操作相结合，用于足式移动操作机器人”，*IEEE*机器人与自动化快报，第 7 卷，第 2 期，第 2377–2384 页，2022年。[51] Z. Fu、X. Cheng 和 D. Pathak，“深度全身控制：为操作与运动学习统一策略”，*arXiv* 预印本 *arXiv:2210.10044*，2022年。

表 II：符号说明。

参数	定义	单位	维度
	系统组件		
π_d	带球策略	-	-
π_r	恢复策略	-	-
$\mathbf{Y}$	YOLOv7 网络 [27]	-	-
	机器人状态		
$\mathbf{q}$	关节角度	rad	12
$\dot{\mathbf{q}}$	关节速度	弧度/s	12
$\ddot{\mathbf{q}}$	关节加速度	弧度/s	12
$\boldsymbol{\tau}$	关节力矩	弧度/s	12
$\mathbf{g}$	重力单位向量，机体坐标系	$\mathrm{m/s^2}$	3
$\boldsymbol{\psi}_t$	机体偏航角，全局坐标系	rad	1
$\mathbf{q}_{\text{des}}$	关节位置目标	rad	12
$\boldsymbol{\theta}^{\text{cmd}}$	时序参考变量 [9]	-	4
$\mathbf{p}^{\text{FRHip}}$	右前髋部位置	m	3
	球状态		
$\mathbf{o}^v$	鱼眼相机图像	-	400×480
$\mathbf{b}$	球在机身坐标系中的位置	m	3
$\hat{\mathbf{b}}$	估计的球在机身坐标系中的位置	m	3
$\mathbf{v}^b$	全局坐标系中的球速度	m/s	2
$\mathbf{v}^{\text{cmd}}$	全局坐标系中的命令球速度	m/s	2
ψ_b	球速度方向	rad	1
ψ_{cmd}	命令球速度方向	rad	1
	控制策略		
$\mathbf{o}$	策略观测	-	37×15
$\mathbf{a}$	策略动作	-	12
$\mathbf{c}$	命令	-	2

附录

VII. 奖励结构

带球策略：表 III 给出了足球带球学习的奖励项。$\mathbf{v}^b$ 和 $\mathbf{v}^{\text{cmd}}$ 分别是全局参考坐标系中的期望球速度和指令球速度。ψ_b 和 ψ_{cmd} 分别是全局参考坐标系中期望球速度和指令球速度的方向，可由 $\mathbf{v}^b$ 和 $\mathbf{v}^{\text{cmd}}$ 直接计算。$\mathbf{b}$ 是机体坐标系中的球位置。$\mathbf{p}^{\text{FRHip}}$ 是机体坐标系中的右前髋部位置。e_{rbcmd} 表示机器人至球的向量与 ψ_b 之间的角度差。e_{rbbase} 表示基座偏航角与 ψ_b 之间的角度差。$\boldsymbol{\kappa}$ 是目标接触状态 $\mathbf{C}^{\text{cmd}}_{\text{foot}}(\boldsymbol{\theta}^{\text{cmd}}, t)$；与先前研究 [9] 一致，$\mathbf{f}^{\text{foot}}$ 表示支撑相的足部接触力，$\mathbf{v}^{\text{foot}}$ 表示摆动相的足部速度，$\mathbf{C}^{\text{cmd}}_{\text{foot}}$ 表示期望足部接触序列，$\boldsymbol{\theta}^{\text{cmd}}$ 表示时序参考变量。$\boldsymbol{\tau}$、$\dot{\mathbf{q}}$、$\ddot{\mathbf{q}}$ 分别表示关节力矩、位置、速度和加速度。i 是关节索引。$\mathbf{g}^{\text{xy}}_t$ 表示投影到机器人横向平面上的重力单位向量。$\mathbf{a}_t$ 表示时间步 t 上的动作。$|\cdot|$ 表示 l^2-范数。时间步 t 的总奖励表示为 $r_t = r^{\text{pos}}_t \exp(r^{\text{neg}}_t)$，其中 r^{pos}_t 和 r^{neg}_t 分别表示正向任务奖励和负向惩罚奖励 [34]。

恢复策略：表 III 给出了用于学习恢复策略的奖励项，其灵感来源于 [35], [37]。这里，$\mathbf{g}_z$ 表示机体坐标系中重力单位向量的垂直分量。当机器人完全直立时，$\mathbf{g}_z = -1$。$\mathcal{I}$ 是指示变量 $\mathbb{1}_{\mathbf{g}_z < -0.6}$ 的简写，用于表示只有当机器人的身体接近直立时，身体高度、身体姿态和脚部高度奖励才会被激活。clamp 会将其输入值限制在 0 到 1 之间。

VIII. 噪声模型

机器人物理随机化：我们对机器人的有效载荷质量、电机强度、关节标定、足部摩擦力、足部恢复系数和质心位移进行随机化。表 IV 列出了随机化参数的范围。

表 III：控球和恢复策略的奖励项。

控球策略		
Term	表达式	权重
投影球速	$\exp\{-\delta_v \lvert\mathbf{v}^b - \mathbf{v}^{\text{cmd}}\rvert^2\}$	0.5
机器人与球的距离	指数$\{-\delta_p \lvert\mathbf{b} - \mathbf{p}^{\text{FRHip}}\rvert^2\}$	4.0
偏航对齐	$\exp\{-\delta_\psi(e^2_{\text{rbcmd}} + e^2_{\text{rbbase}})\}$	4.0
球速范数	$\exp\{-\delta_n(\lvert\mathbf{v}^{\text{cmd}}\rvert - \lvert\mathbf{v}^b\rvert)^2\}$	4.0
球速度角	$1-(\psi_b - \psi_{\text{cmd}})^2 / \pi^2$	4.0
摆动相位调度	$[1-\boldsymbol{\kappa}]\exp\{-\delta_{\text{cf}} \lvert\mathbf{f}^{\text{foot}}\rvert^2\}$	4.0
支撑相位调度	$\boldsymbol{\kappa}\exp\{-\delta_{\text{cv}} \lvert\mathbf{v}^{\text{foot}}_{\text{xy}}\rvert^2\}$	4.0
关节限位违规	$\mathbb{1}_{q_i > q_{\max} \,\|\|\, q_i < q_{\min}}$	-10.0
关节力矩	$\lvert\boldsymbol{\tau}\rvert^2$	-0.0001
关节速度	$\lvert\dot{\mathbf{q}}\rvert$	-0.0001
关节加速度	$\lvert\ddot{\mathbf{q}}\rvert$	-2.5e-7
髋部/大腿碰撞	$\mathbb{1}_{\text{collision}}$	-5.0
投影重力	$\lvert\mathbf{g}_{\text{xy}}\rvert^2$	-5.0
动作平滑	$\lvert\mathbf{a}_{t-1} - \mathbf{a}_t\rvert^2$	-0.1
动作平滑 2	$\lvert\mathbf{a}_{t-2} - 2\mathbf{a}_{t-1} + \mathbf{a}_t\rvert^2$	-0.1
恢复策略		
Term	表达式	权重
身体方向	$(0.5 - 0.5\mathbf{g}_z)^2$	1.0
身体高度	$\mathcal{I}(1.0 - \text{钳制}(\frac{h^{\text{身体}}_{\text{目标}} - h^{\text{身体}}}{h^{\text{身体}}_{\text{目标}}}^2))$	1.0
身体姿态	$\mathcal{I}(1.0 - \text{限制}(\lvert q - q_{\text{standing}}\rvert^2 / 20.0))$	1.0
脚部高度	$\mathcal{I}(\exp -10 \sum_i (h^{\text{foot}}_i)^2)$	1.0
动作	$\lvert\mathbf{a}_t\rvert^2$	$-1e{-3}$
关节力矩	$\lvert\boldsymbol{\tau}\rvert^2$	$-1e{-5}$

球体物理随机化：我们对球体质量和球体-地形阻力系数进行随机化。表 IV 给出了随机化参数的范围。

地形物理随机化：在训练带球策略时，我们没有对地形几何形状和重力进行大幅随机化，因为我们发现，许多此类组合会使成功带球变得不可行。不过，我们确实在随机地形上训练了恢复策略，以便在此类场景下实现恢复。为模拟崎岖地形，我们对地形高度上佩林噪声的幅值进行随机化。为模拟倾斜地形，我们每隔 6.0 秒对各轴上的重力向量施加均匀随机扰动。表 IV 给出了随机化参数的范围。

球体传送：我们每隔 7.0 秒将球传送到 1.0 米范围内均匀随机采样的位置。

相机延迟：我们将下一次观察的到达时间建模为泊松分布，每个回合的平均到达时间在 20 毫秒至 60 毫秒之间随机设定。

IX. 策略优化

我们使用同一组 PPO 超参数来训练带球策略和恢复策略。表 V 列出了这些超参数值。它们与先前研究中用于在这台机器人上训练运动策略的设置相同 [9]。

表 IV：训练期间机器人动力学、球动力学和指令的随机化范围。

动力学参数	范围	单位
机器人动力学		
有效载荷质量	$[-1.0, 3.0]$	kg
电机强度	$[90, 110]$	%
关节标定	$[-0.02, 0.02]$	rad
机器人-地形摩擦力	$[0.40, 1.00]$	–
机器人-地形恢复系数	$[0.00, 1.00]$	–
机器人质心位移	$[-0.15, 0.15]$	m
球体动力学		
Mass	$[0.159, 0.254]$	kg
相机帧到达率	$[0.3, 0.7]$	–
瞬移位置	$[0.0, 1.0]$	m
扰动速度	$[0.0, 0.3]$	m/s
球体-地形阻力系数	$[0.0, 1.5]$	–
地形动力学 （仅恢复控制器）		
柏林噪声幅度	$[0, 10]$	cm
重力噪声	$[-1.0, 1.0]$	$\mathrm{m/s^2}$
命令		
$\mathbf{v}_x^{\mathrm{cmd}}$	$[-1.5, 1.5]$	m/s
$\mathbf{v}_y^{\mathrm{cmd}}$	$[-1.5, 1.5]$	m/s

超参数	值
折扣因子	0.99
GAE 参数	0.95
每次 rollout 的时间步数	21
每次 rollout 的轮数	5
每轮的小批次	4
熵奖励 (α_2)	0.01
价值损失系数 (α_1)	1.0
裁剪范围	0.2
奖励归一化	yes
学习率	1e−3
# 环境	4096
# 总时间步数	7B
优化器	Adam

表 V：PPO 超参数。