MASH：用于单一类人机器人运动的合作-异构多智能体强化学习

Qi Liu [1†], Xiaopeng Zhang[2†] , Mingshan Tan [2] , Shuaikang Ma [2] , Jinliang Ding [1] , Yanjie Li [2] *

***Abstract*—本文提出了一种新的方法，通过协作异构多智能体深度强化学习（MARL）来增强单个类人机器人运动能力。虽然大多数现有方法通常采用单智能体强化学习算法来处理单个类人机器人或使用 MARL 算法来进行多机器人系统任务，我们提出了一种不同的范式：将协作异构 MARL 应用于优化单个类人机器人的运动。该方法——单类人机器人运动的多智能体强化学习（MASH），将每个肢体（腿和手臂）视为独立的智能体，在共享全局评论器进行协同学习的同时探索机器人的动作空间。实验表明，MASH 加速训练收敛并提高了全身合作能力，性能优于传统的单智能体强化学习方法。这项工作推动了 MARL 在单类人机器人控制中的整合，为高效运动策略提供了新的见解。**

***Index Terms*—Robot control, humanoid robot locomotion, reinforcement learning (RL), multi-agent RL.**

I. 引言

深层 强化学习 (RL) 在多个机器人控制领域取得了显著成功 [1], [2] ，例如四足动物运动 [3], [4] 、双足行走 [5], [6] 和自主空中飞行器导航 [7] 。这项工作专注于推进深度强化学习，以改善类人机器人运动，该领域在机器人技术中是一个复杂但至关重要的挑战。

人形机器人运动方法可以分为两大主要范式：基于模型的运动方法和无模型的基于学习的方法。前者以包含模型预测控制和轨迹优化的整体动力学运动方法为代表，在结构化环境中表现出色，但依赖于精确的动态建模。后者包括深度强化学习（RL）和模仿学习。最近的进展表明，基于学习的方法在策略泛化能力方面优于基于模型的运动方法。基于学习的运动方法可总结为两类：（1）针对上半身和下半身的分阶段策略训练与系统集成，以及（2）通过模仿学习或单代理深度 RL 实现的全身运动。

分阶段政策训练和系统集成方法采用模块化分解和递进集成框架，以减轻高自由度（DoF）系统中的运动复杂性。尽管取得了显著进展，这类方法仍面临持续挑战：分阶段训练可能导致上半身和下半身策略之间的协调不足，限制了全身协同作用。而且在复杂任务和场景中的适应性仍然受到限制 [15], [17] 。

全身模仿学习和单智能体深度增强学习方法 [18] 首先通过采集或生成用于人形机器人全身运动的数据，然后利用获取的数据通过模仿学习或单智能体深度增强学习来训练运动策略。然而，获取高质量的人形机器人运动数据面临显著挑战，包括高昂的采集成本、低效率和复杂的后处理要求。虽然基于模仿学习的方法相对直接，但所得策略表现出有限的泛化能力。与模仿学习相比，基于深度增强学习的方法显现出较好的泛化能力，但未能充分利用深度增强学习的试错学习机制。由于这些方法依赖于人类运动数据作为参考轨迹，它们难以探索展示数据之外的动作和技能空间，最终限制了其在复杂环境中的性能上限。

在单人类机器人运动中应用深度强化学习时，主要方法使用的是单智能体深度强化学习算法 [19], [20] 。然而，这些方法在解决复杂机器人系统中固有的协调性挑战时可能存在局限性。现有的解决方案通常是为单个机器人使用单智能体强化学习或为合作的多机器人任务使用多智能体深度强化学习（MARL）[21] 。然而，在利用 MARL 原理增强单一人形机器人实体的协调能力方面仍有未被探索的潜力。因此，开发更高效的整体合作运动方法以增强人形机器人在非结构化环境中的运动与操作协调仍然具有挑战性。

合作多智能体强化学习（MARL）算法在游戏人工智能领域的多智能体合作中表现出显著的成功。Liu 等人提出了一种新方法，使用 MARL 来改进单个四足机器人运动学习。与传统的机器人运动学习方法不同，本文提出了一种新的方法，将运动学习建模为一个合作-异质的 MARL 问题，并使用合作-异质的 MARL 算法来增强单个类人机器人的运动能力。我们的方法通过将每个肢体（双臂和双腿）视为合作 MARL 框架中的独立智能体，实现了在复杂任务中的卓越协调能力。所提出的用于单个类人运动的合作-异质多智能体强化学习（MASH）利用了共享学习结构，其中智能体（肢体）通过经验共享和合作策略学习共同优化运动。

本文的主要贡献如下：

- 我们提出了 MASH，这是一种将类人运动重新表述为合作异构多智能体强化学习（MARL）问题的新框架。MASH 通过将每个肢体（手臂和腿）视为具有不同动作空间的独立智能体，从而实现比传统单智能体强化学习更高效的协调学习。
- 实验结果表明，所提出的方法在步态执行和最终性能方面表现优越，提高了训练效率和样本复杂度，并增强了动态环境中的鲁棒性。这些结果验证了将 MARL 原则应用于单个仿人机器人控制的有效性。

本文的其余部分结构如下。第 II 节回顾了相关工作。第 III 节介绍了马尔科夫决策过程（MDP）和强化学习的基础知识。第 IV 节详细介绍了所提出的 MASH 方法。第 V 节报告了实验结果，这些结果证明了所提出方法的有效性。最后，第 **??** 节总结了本文，并概述了未来工作的方向。

II. 相关工作

A. 深度强化学习在人形机器人运动中的应用

人形机器人运动的方法可以分为两种主要范式：基于模型的运动方法和基于无模型学习的方法。以结合模型预测

This work was supported by Shenzhen Basic Research Program (Grant No. JCYJ20220818102415033, KJZD20230923114220045).
(Corresponding author: Yanjie Li, autolyj@hit.edu.cn)

1 Faculty of Robot Science and Engineering, Northeastern University, Shenyang, 110819, China.

2 Guangdong Key Laboratory of Intelligent Morphing Mechanisms and Adaptive Robotics and School of Intelligence Science and Engineering, the Harbin Institute of Technology Shenzhen, Shenzhen, 518055, China.

† These authors contributed equally to this work.

控制和轨迹优化的全身动态运动方法为例的基于模型的运动方法在结构化环境中表现出稳健的性能，但依赖于精确的动态建模。基于无模型学习的方法包括深度强化学习（RL）[11], [12] 和模仿学习 [13] 。最近的进展 [14]–[16] 显示了学习方法的策略泛化能力优于基于模型的方法。基于学习的运动方法可总结为两类：(1) 针对上半身和下半身的分阶段策略训练和系统集成，以及 (2) 基于模仿学习或单智能体深度 RL 的全身运动学习。

分阶段策略训练和系统集成方法采用模块化分解和渐进集成框架，以减轻高自由度（DoFs）系统中的运动复杂性。这类方法包括三个阶段：（1）解耦策略训练：将全身控制问题划分为下肢（专注于运动稳定性）和上肢（强调操作灵活性）的独立策略训练 [25] 。（2）系统集成：在下肢策略收敛后，逐步整合上肢操作任务 [26] 。（3）任务导向优化：在基本机动性上叠加任务约束（例如末端执行器轨迹跟踪），通过耦合动态优化实现全身控制 [1], [27] 。尽管取得了显著进展，但这类方法面临持续挑战：解耦策略训练可能导致上肢和下肢策略之间协调不足，限制全身协同作用，其适应性在复杂任务和场景中仍受到限制 [15], [17] 。因此，开发更高效的全身协作运动方法以增强人形机器人在运动-操作协调性方面仍然具有挑战性。

基于模仿学习或单代理深度强化学习方法的全身运动学习首先是为人形机器人 [18] 收集或生成全身运动数据，然后通过模仿学习或单代理深度强化学习使用获取的数据来训练运动策略。根据数据来源，这类方法可以分为四种类型：(1) 机器人远程操作数据收集 [18], [28] ：涉及对物理机器人的直接操作以获取运动学准确的运动数据。尽管这可以产生高质量的、特定于平台的数据，但它面临着高获取成本和对硬件的依赖。(2) 人体动作捕捉数据：像 HumanPlus [29] 这样的方法使用动作捕捉系统记录人类动作，并将其映射到机器人关节空间。虽然直观，但这种方法需要专业设备，存在扩展性挑战，并且必须解决人机运动学差异。(3) 来自互联网的人类视频数据：OKAMI [30] 从在线视频中提取人类动作，进行 3D 重建，并将其转移到机器人上。尽管数据丰富，但此方法与噪声输入与运动学不匹配问题作斗争。(4) 合成动画数据：像 OmniH2O [31] 和 PMCP [32] 可以通过动画工具生成运动并重新映射到机器人上。虽然灵活，但这种数据可能缺乏物理现实感。

此外，获取高质量的人形动作数据面临显著的挑战，包括高昂的收集成本、低效率和复杂的后处理要求。尽管基于模仿学习的控制方法相对简单，所产生的策略展示出有限的泛化能力。与模仿学习相比，基于深度强化学习的方法表现出改进的泛化能力，但未能充分利用深度强化学习的试错学习机制。由于这些方法依赖于人类运动数据作为参考轨迹，它们难以探索超出示范数据的动作和技能空间，最终限制了其在复杂环境中的性能上限。

其他方法：[33] 采用传统的轨迹优化来为机器人生成参考轨迹，并结合单代理深度强化学习进行轨迹跟踪控制。尽管这种方法在四足机器人领域已经取得了一定成功，但由于类人机器人的自由度更高、动态约束更复杂，其应用仍然受到限制。一些方法利用广泛的奖励塑造来指导策略学习，包括手动调优的运动跟踪 [34] 、周期性奖励组成 [35] 、对抗性运动先验 [36] 和周期性正弦轨迹 [37] 。然而，手动设计奖励函数往往是费力和耗时的，要求广泛的领域专业知识和反复的调整以实现期望的行为。

B. 多机器人控制的多智能体强化学习

多智能体强化学习在多种多机器人系统中表现出显著的成功，包括机器人群体协作 [38] 、自动驾驶 [39] 、无人飞行器协调 [40] 和智能仓库 [41] 。这些应用展示了多智能体强化学习在需要物理系统中协调复杂行为的能力，这些系统中分散的决策必须调和环境约束与智能体间的协调。与这些多机器人应用形成对比的是，本文通过将单一类人机器人运动形式化为一个协作的多智能体强化学习问题，提出了一种新的范式。我们的方法将每条腿和手臂视作独立的智能体，有别于传统方法将机器人建模为单一统一智能体或集中于多机器人协作。

III. 背景

本节总结了 MDP [42] 和 RL。本文中考虑的 MDP 被建模为一个元组 $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, T)$ ，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\mathcal{P}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ 代表状态转移概率。$\mathcal{R}: \mathcal{S} \times \mathcal{A} \to R$ 是奖励函数，$\gamma \in [0,1)$ 是折扣因子，T 表示时间跨度。在每个时间步 t ，根据策略选择一个动作 $a_t \in \mathcal{A}$ 。然后，代理通过从 $p(s_{t+1} \mid s_t, a_t)$ 采样转移到下一个状态，其中 $p \in \mathcal{P}$ ，并收到标量奖励 $r(s_t, a_t) \in \mathcal{R}$ 。代理继续与环境互动，直到到达终止状态。RL 的目标是学习一个能最大化期望折扣累计奖励的策略 $\pi: \mathcal{S} \times \mathcal{A} \to [0,1]$ 。对于任何策略 π ，状态-动作值函数（Q 函数）定义如下：

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right] \tag{1}$$

近端策略优化（PPO）[43] 通过一个约束优化方法解决了策略梯度方法中的关键稳定性挑战。该算法的核心创新在于其剪辑代理目标函数：

$$L^{CLIP}(\theta) = \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{2}$$

，其中

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \tag{3}$$

以及

$$\hat{A}_t = Q^{\pi}(s_t, a_t) - V_\psi(s_t) \tag{4}$$

，其中 ψ 表示价值函数（V_ψ ）网络的参数，ϵ 表示一个系数。策略参数 θ 更新如下：

$$\theta \leftarrow \theta + \alpha \nabla_\theta L^{CLIP}(\theta) \tag{5}$$

。PPO 的约束更新稳定了训练并提高了性能，使其在复杂的单代理强化学习任务中变得实用。

IV. 单个仿人机器人运动的多智能体强化学习

本文提出了一种新颖的多智能体强化学习框架，通过利用肢体间的协调来增强单个机器人的运动能力。我们的方法将人形机器人的每个肢体（包括手臂和腿）视为合作异构多智能体系统中的独立智能体，每个智能体共享一个全局评判器，同时保持各自的观察和策略。该框架架起了单个机器人运动学习与多智能体协调学习之间的桥梁，展示了合作多智能体学习策略能够提高单个人形机器人的运动能力。

本文将单个类人机器人运动建模为一个多智能体协作问题，该问题被描述为一个部分可观测的去中心化马尔可夫

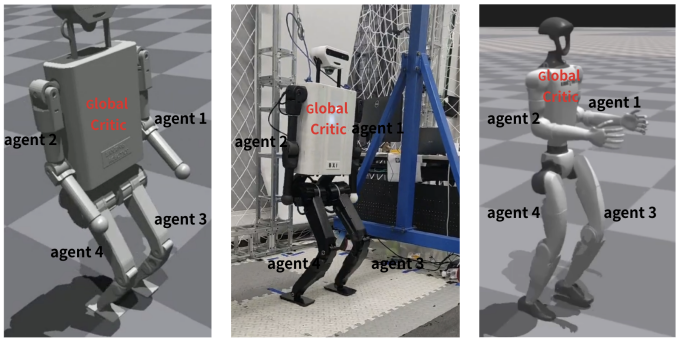

Fig. 1. 用于单个类人机器人的运动的 MARL 模型

决策过程（decPOMDP）。decPOMDP 由一个元组定义。$\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\mathcal{P}$ 是状态转移分布，r 是奖励函数，$\mathcal{Z}$ 是观测空间，$\mathcal{O}$ 是观测函数，N 是智能体的数量，γ 是折扣因子，T 是时间范围。在每一个时间步骤 t，每个智能体 $n \in \{1,\ldots,N\}$ 选择一个动作 $a_t^n \in \mathcal{A}$，从而形成一个联合动作 $\boldsymbol{a_t} = \{a_t^1, a_t^2, \ldots, a_t^N\}$。环境根据 $\mathcal{P}(s_{t+1}|s_t, \boldsymbol{a_t})$ 转变为一个新状态 s_{t+1} 并提供一个共享的奖励 $r(s_t, \boldsymbol{a_t})$。每个智能体从 $\mathcal{O}(s_t, n)$ 接收到一个观测 z_t^n 并保持观测-动作历史 τ_t^n。MARL 旨在学习能够最大化期望累计奖励的策略。

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, \boldsymbol{a_t})\right] \tag{6}$$

本文提出了 MASH，该方法应用了 MARL 算法，将单个人形机器人的不同部分视为独立的智能体，使用共享的奖励进行协作训练。具体而言，本文采用多智能体 PPO [45] (MAPPO) 来解决建模的多智能体问题。MAPPO 在多智能体系统中优化以下目标函数：

$$\begin{aligned} L^{\text{CLIP}}_{\text{MAPPO}}(\theta) = \sum_{i=1}^{n} \mathbb{E}_t \Big[\min \Big(& r_t^i(\theta_i)\hat{A}_t, \\ & \text{clip}(r_t^i(\theta_i), 1-\epsilon, 1+\epsilon)\hat{A}_t \Big) \Big] \end{aligned} \tag{7}$$

，其中

$$r_t^i(\theta_i) = \frac{\pi_{\theta_i}(a_t^i|o_t^i)}{\pi_{\theta_{i,\text{old}}}(a_t^i|o_t^i)} \tag{8}$$

，其中 i 表示 MARL 中的 i -个智能体。每个智能体通过如下方式更新其策略参数：

$$\theta_i \leftarrow \theta_i + \alpha \nabla_{\theta_i} L_i^{CLIP}(\theta_i) \tag{9}$$

本研究采用集中训练与分散执行（CTDE）范式 [46]，以解决 MARL 中的关键挑战，同时确保在动态环境中的训练稳定性。该实现采用双重神经网络，分别学习策略（π_θ）和增强的价值函数（$V_\psi(s)$），后者结合了全局状态信息以提高训练稳定性。通过利用这种架构，所提出的方法可以实现优越的最终性能、加速的收敛速度以及增强的部署鲁棒性，相较于传统方法，特别对于需要协作学习和操作灵活性的复杂机器人控制任务表现出显著的效能。

状态空间与观测：为双足和双臂系统分别构建了一个共享参数的演员网络，接收来自两个代理的串联观测。对于双足系统，每个代理的观测包括：电机位置 $q_t \in \mathbb{R}^6$（表示 6 个腿部电机的角度）、电机速度 $\dot{q}_t \in \mathbb{R}^6$（表示电机的旋转速度）、上一时间步的动作 $a_t \in \mathbb{R}^6$、用于步态序列的三角函数定时指导 $\phi_t \in \mathbb{R}^2$（调节步伐的时间）、躯干的欧拉角 $\theta_t \in \mathbb{R}^3$ 和角速度 $\omega_t \in \mathbb{R}^3$、遥控器提供的控制命令 $c_t \in \mathbb{R}^4$（包括一个二进制站立标签、在 x 和 y 方向的速度以及偏航角速度）、为每个代理的独热编码 $e_t \in \mathbb{R}^2$。因此，每个代理有一个 32 维的观测，导致演员网络的总输入维度为 $o_t^{\text{actor}} \in \mathbb{R}^{64}$（$32 \times 2$）。对于双臂系统，电机仅有 4 个自由度。因此，每个手臂代理接收一个 26 维的观测，导致总输入维度为 $o_t^{\text{actor}} \in \mathbb{R}^{52}$（$26 \times 2$）。

评估网络以整个机器人机身的全局观察作为输入，包括：电机位置 $q_t \in \mathbb{R}^{20}$、电机速度 $\dot{q}_t \in \mathbb{R}^{20}$、上一个时间步的动作 $a_t \in \mathbb{R}^{20}$、电机位置偏差 $d_t \in \mathbb{R}^{20}$、用于四肢的时间指导 $\phi_t \in \mathbb{R}^2$、遥控器提供的控制命令 $c_t \in \mathbb{R}^4$、以及躯干的线速度 $v_t \in \mathbb{R}^3$、欧拉角 $\theta_t \in \mathbb{R}^3$ 和角速度 $\omega_t \in \mathbb{R}^3$。此外，观察还包括外部扰动力 $f_t \in \mathbb{R}^2$ 和扭矩 $\tau_t \in \mathbb{R}^3$、环境摩擦系数 $\mu_t \in \mathbb{R}^1$、机身质量 $m_t \in \mathbb{R}^1$ 和两个二进制掩码，分别表示四肢的姿态状态和接触状态，各为 $\in \mathbb{R}^2$。因此，评估网络的总输入维度为 $o_t^{\text{critic}} \in \mathbb{R}^{106}$。

动作空间：双足演员网络的输出是一个连续动作 $a_t \in \mathbb{R}^{12}$，表示 12 个电机的扭矩。对于双臂系统，演员网络输出 $a_t \in \mathbb{R}^8$。

评论家网络输出值函数 $V_t \in \mathbb{R}^4$，用于计算优势函数。

奖励函数：表格 I 中的奖励函数旨在通过鼓励期望行为和惩罚不期望行为来优化机器人的性能。主要奖励包括：关节位置以跟随参考姿态；速度跟踪项，如跟踪线速度和跟踪角速度以跟随指令运动；与能量相关的惩罚，包括 DOF 扭矩、DOF 速度和 DOF 加速，以鼓励平稳和高效的运动；步态特定奖励，如足部悬空时间、足部清除和足部接触数量促进协调的步伐，同时通过姿态、碰撞、足部滑动和底座高度来解决稳定性和安全性。额外的项如动作平滑性和扭矩率有助于确保运动的一致性并减少驱动峰值。这些项共同塑造了机器人的行为，以在复杂任务中实现稳健和物理上合理的动作。

TABLE I
奖励函数

REWARD SETTINGS, CORRESPONDING EQUATIONS, AND THEIR SCALES

Reward Term	Equation	Scale
Joint Position	$\exp\left(-\lVert \boldsymbol{q} - \boldsymbol{q}_{\text{default}} - \boldsymbol{q}_{\text{ref}} \rVert\right)$	3.5
Tracking Linear Velocity	$\exp\left(-\frac{\lVert \boldsymbol{v}_{\text{cmd}} - \boldsymbol{v}_b \rVert^2}{\sigma_t}\right)$	1.5
Tracking Angular Velocity	$\exp\left(-\frac{(\boldsymbol{\omega}_{\text{cmd},z} - \boldsymbol{\omega}_{b,z})^2}{\sigma_{\text{yaw}}}\right)$	1.4
DOF Torques	$\sum\left(\frac{\tau}{\tau_{\max}}\right)^2$	-2.0×10^{-3}
DOF Velocity	$\sum_i \lVert \boldsymbol{v}_{d,i} \rVert^2$	-5×10^{-4}
DOF Acceleration	$\sum_i \left(\frac{\boldsymbol{v}_{d,i,t} - \boldsymbol{v}_{d,i,t-1}}{\Delta t}\right)^2$	-1.0×10^{-7}
Feet Air Time	$\sum_{i=1}^{2} t_{\text{air},i} \cdot \exp\left(-\alpha \cdot \lVert \Delta \boldsymbol{x}_i \rVert\right)$	2.0
Feet Clearance	$\sum(\lvert z_{feet} - h_{\text{target\_feet}} \rvert < \delta)$	2.0
Feet Contact Number	$\sum \begin{cases} +1, & \text{if contact}_i = \text{stance}_i \\ -0.3, & \text{otherwise} \end{cases}$	1.2
Orientation	$\exp(\lvert \theta_{\text{pitch, roll}} \rvert) + \exp(\lVert g_{\text{proj}} \rVert)$	1.0
Collision	$\sum(1.0 \cdot (\lVert \boldsymbol{f}_c \rVert > 0.1))$	-1.0
Feet Slip	$\sum(c_f \cdot \lVert \boldsymbol{v}_f \rVert^2)$	-5×10^{-2}
Base Height	$\exp\left(-\lvert h_{base} - h_{target_base} \rvert\right)$	0.2
Action Smoothness 1	$\sum(\boldsymbol{a}_t - \boldsymbol{a}_{t-1})^2$	-0.1
Action Smoothness 2	$\sum(\boldsymbol{a}_t - 2\boldsymbol{a}_{t-1} + \boldsymbol{a}_{t-2})^2$	-0.1
Torque Rate	$\sum\left(\frac{\tau_t - \tau_{t-1}}{\tau_{\max} \cdot \Delta t}\right)^2$	-2×10^{-4}

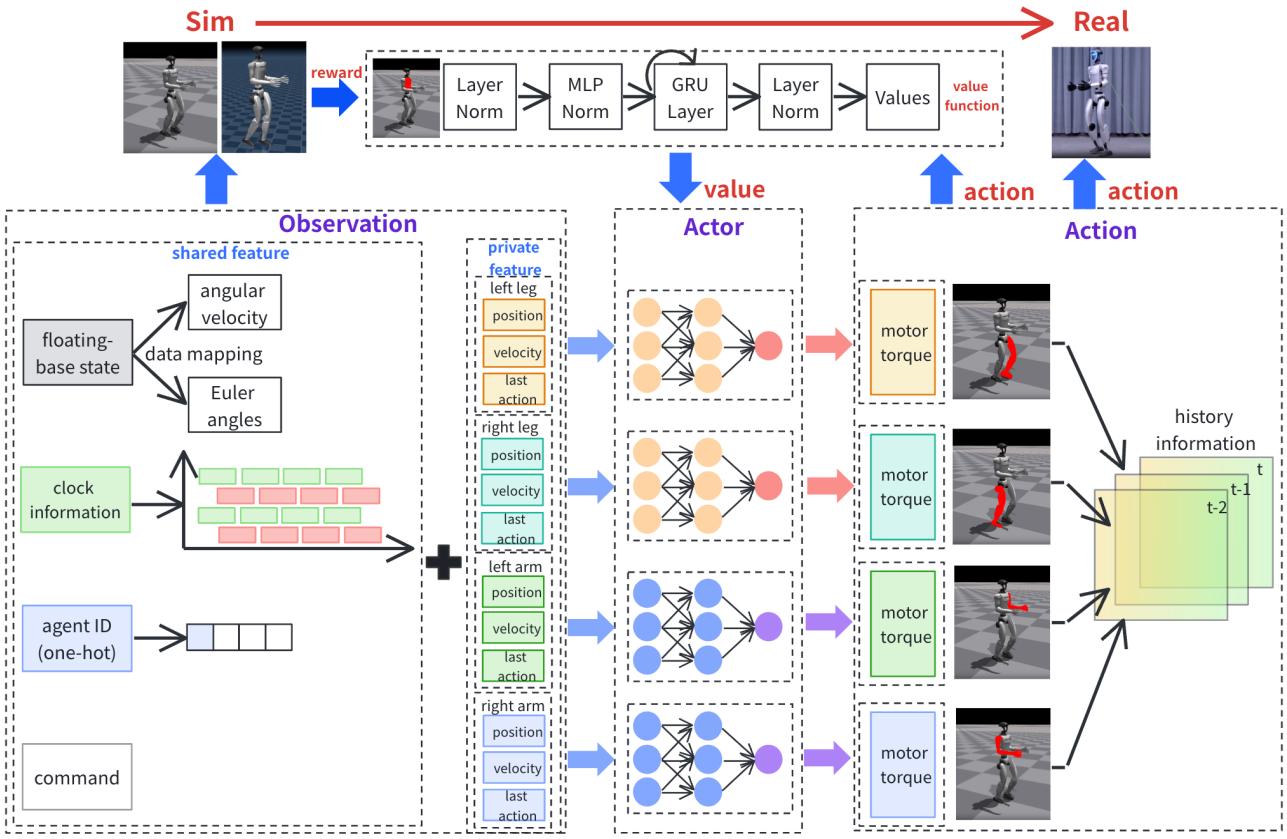

Fig. 2. MASH 的框架

A. 多智能体演员和全局评论员网络

在 Isaac Gym [47] 仿真环境中，机器人接收观测和奖励以促进学习。观测分为共享和私有特征，演员和评论家网络都使用奖励来优化策略。

演员网络采用 MLP/RNN 架构，并具有一个生成行动及其对数概率的激活层，而评论家网络使用类似的 MLP/RNN 结构来估计状态值。在训练过程中，演员网络根据环境观察选择行动，评论家提供价值估计以指导策略更新。行动输出对应于期望的关节位置。一个比例-微分（PD）控制器根据这些目标计算相应的关节扭矩，然后将其应用到物理模拟中的机器人模型上。训练完成后，优化的演员网络被直接应用于机器人平台，以进行实地操作。完整的学习流程如图 2 所示。

我们将每条腿和手臂建模为仿人机器人系统的一个独立代理，同时在两条腿和两条手臂之间分别采用共享参数的 actor 网络（即，两条手臂共享一个共享参数 actor 网络，而两条腿共享另一个）。这个架构选择提供了两个主要优势：(1) 与为每个肢体维护单独网络相比，它显著减少了计算开销，(2) 它正确地捕捉到仿人机器人在运动过程中左右肢体之间固有的对称性和协调性。与传统的多智能体系统（如 StarCraft [22] ）中智能体独立运作不同，我们的仿人机器人的腿和手臂形成一个本质上耦合的系统，对称地排列在身体中心周围，并机械地限制为协调移动。共享参数网络自然地编码了这些物理对称性和协调要求，使其特别适合于机器人控制，同时保持 MARL 框架的优势。

我们将每条腿和手臂的策略表示如下：

$$\pi^{leg}_{\theta_i}(a_{i,t} \mid s_{i,t}) = \pi^{leg}_{\theta}(a_{i,t} \mid s_{i,t}) \tag{10}$$

$$\pi^{arm}_{\theta_i}(a_{i,t} \mid s_{i,t}) = \pi^{arm}_{\theta}(a_{i,t} \mid s_{i,t}) \tag{11}$$

，其中 $i = 1, 2$ 为左右肢体的索引。腿的策略 $\pi^{leg}_{\theta_i}$ 在两条腿之间共享，手臂的策略 $\pi^{arm}_{\theta_i}$ 在两只手臂之间共享。

为了增强代理之间的协调性，我们在每个独立观察中加入了共享观察，包括从惯性测量单元数据计算的欧拉角和角速度、时间引导器，以及代理标识符（ID）。时间引导器 $T_i(t)$ 指导每条腿的步态顺序，并在不同的运动姿势下协调每只手臂的运动模式，而代理 ID 对于共享参数网络是必需的。此设置在确保每个代理独立性的同时，提高了它们的合作能力。时间引导器有助于同步不同腿的运动，并协调手臂的动作，确保步态模式的流畅和平衡。其定义如下：

$$T_i(t) = \sin\left(2\pi(kt + \Delta_i)\right) \tag{12}$$

其中

- k 是步态周期的缩放因子。
- Δ_i 是第 i 个肢体（腿或手臂）的相位偏移，这决定了它在步态周期中的相对时序，以确保协调运动。

全局评论家在 CTDE 框架中采用集中式价值函数，利用结合个体肢体观察和共享系统范围信息的全局观察。这种架构在保持像 PPO 这样的单代理优化方法的优点的同时，实现了双腿和双臂代理之间的有效协调。评论家网络处理这一综合的状态表示，以学习用于指导策略更新的价值函数。这确保了在考虑机器人的全局状态和肢体之间依赖关系的情况下，所有代理的同步学习。

B. 从模拟到现实

为了增强从模拟到现实的迁移，我们采用领域随机化。为了增强从模拟到现实的迁移，我们采用领域随机化。根

据应用的时机，所有领域随机化参数可以分为两种类型。第一种类型包括在环境初始化时随机化的物理参数，比如链接质量和惯性、关节阻尼和摩擦力、地面摩擦力以及重力。第二种类型涉及在每一次模拟步骤中随机化的参数，包括动作延迟、扭矩噪声和外部扰动。这些随机化增强了策略在不同物理条件下的鲁棒性和适应性。

V. 实验

我们使用 BanXing 类人机器人进行了 MASH 实验，涉及多种实验。第 V-A 节提出了实验设置。第 V-B 节展示了模拟实验结果。第 **??** 节展示了真实世界的实验和比较。

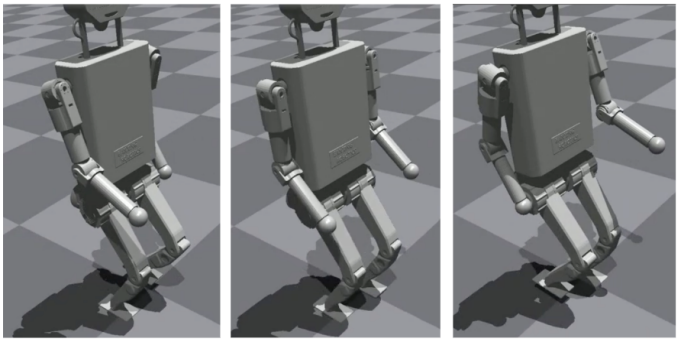

Fig. 3. MASH 的模拟实验

A. 实验设置

1) 任务描述： 在模拟环境中，我们训练一个单一的人形机器人以快速且稳定的方式在平坦地形上向前行走。该任务作为人形机器人控制的标准基准，用于评估机器人在速度和平衡方面的能力。我们的训练流程从 Isaac Gym 中的直立移动训练开始，包括：(i) 双足行走（双腿组成一个多代理组）和 (ii) 摆臂行走（双腿和双臂组成两个多代理组）。与双足行走相比，摆臂行走对四肢协调性的要求更高。然后，我们在 MuJoCo 模拟器上进行 Sim-to-Sim 转移，以严格评估所学策略在不同模拟环境下的鲁棒性。最终，该策略被部署在物理机器人上，以展示其在现实场景中的有效性。

2) 评估指标： 为了全面展示 MASH 的优越性，我们建立了以下评估指标：

- 收敛时间 T_{Conv}：我们将收敛时间定义为在训练期间，为了平均每集奖励稳定达到渐近奖励值的 95 % 所需的训练迭代次数。
- 动作平滑性 S_{action}：我们衡量控制动作（例如，关节扭矩或目标关节位置）的连续性和稳定性，旨在避免动作序列的突然变化。通过动作序列二阶差分的平方 L2 范数来量化：

$$S_{action} = \frac{1}{T}\sum_{t=1}^{T-1}\sum_{i=1}^{N}(a_{i,t+1} - a_{i,t})^2 \tag{13}$$

- 躯干稳定性 S_{torso}：通过测量躯干在运动过程中高度和方向（包括俯仰、横滚和偏航角）的波动来评估躯干的稳定性：

$$S_{torso} = w_h \cdot \mathrm{Var}(h_t) + w_\theta \cdot \mathrm{Var}(\theta_t) \tag{14}$$

- 肢体协调 C_{limb}：通过分析腿部关键关节（例如臀部或膝盖）的运动轨迹，并计算它们在步态周期内的相对相位差，可以定量评估肢体协调：

$$C_{limb} = \frac{1}{T}\sum_{t=1}^{T}|\phi_{\text{left}}(t) - \phi_{\text{right}}(t) - \phi_{\text{target}}| \tag{15}$$

3) 基线： 为了更好地展示 MASH 的强大潜力，我们将其与传统的单智能体 PPO 算法进行比较，在模拟环境中训练这两者 3000 集，每集长度为 48 步。

B. 模拟实验

实验评估在相同的模拟条件下将 MASH 与其他基线进行比较。图 4 展示了在训练过程中 (a) 腿部和 (b) 全身的平滑奖励增长趋势，比较了提出的 MASH 框架与单智能体 PPO 基线。在这两种情况下，MASH 在早期训练阶段的奖励增加得明显更快，表明学习效率的加速。此外，MASH 比基线收敛到更高的渐近奖励，特别是在腿部训练中，改进的幅度更加明显。这些结果显示了多智能体分层结构在捕捉肢体间协调模式和优化控制性能方面的优势。训练完成后，我们评估了所有方法的部署性能。所有评估指标总结在表 II 中。

如表 II 所示，我们对比我们的提出的方法 MASH 与单智能体 PPO 基线，在二足和全身（摆臂）运动任务中进行定量评估。评估指标——收敛时间、动作平滑性、躯干稳定性和肢体协调性，几乎在所有方面都展示了我们方法的明显优越性。

图 5 比较了我们的 MASH 控制器与基线单代理 PPO 在髋部俯仰轨迹跟踪任务中的性能。正如所示，PPO 控制器（橙色线）在跟踪参考轨迹（绿色虚线）时表现不佳，出现显著的超调和相位滞后。这导致了不稳定和不精确的步态。相反，我们的 MASH 控制器（蓝色线）表现出卓越的跟踪精度，在相位和幅度上都紧密贴合参考轨迹。这强调了我们的方法在生成稳定、平滑和精确的机器人运动方面的优越性。

我们在一个人形机器人上部署了 MASH。为了促进从仿真到现实的转移，我们在训练过程中加入了域随机化，参数如表 **??** 所示。

图 6 展示了使用 MASH 训练的人形机器人行走步态的现实世界硬件验证。从运动学角度来看，频闪图像清晰地捕捉到了一个连贯且流畅的步态循环，展示了熟练的动态平衡。膝关节角度（以弧度为单位）随时间（以秒为单位）变化的相应图表量化了这种稳定的性能。曲线表现出高度的周期性和平滑性，膝盖角度在约 1.2 到 1.7 弧度的范围内规律地振荡，每个步态循环持续约 0.6 秒。曲线的光滑、非突发性特征反映了电机控制的精确性以及生成策略的健全性，有效避免了冲击和不稳定性。运动和数据之间的这种强有力相关性证实了 MASH 可以成功生成一个复杂的电机控制策略，不仅在仿真中有效而且可以直接移植到现实世界的物理系统中。结果表明，训练出的策略具有优异的鲁棒性，能够应对现实世界的物理限制。

在这项工作中，我们提出了 MASH，这是一种新颖的合作异构多智能体强化学习框架，旨在增强单个仿人机器人行走能力。通过重新定义控制问题并将机器人的每个肢体视为合作系统中的一个独立智能体，我们的方法有效利用了多智能体强化学习的原则来促进优越的肢体间协调。在 Isaac Gym 中实现的该方法展示了三个主要优势：(1) 通过协调的肢体学习加速训练收敛；(2) 通过多智能体设计

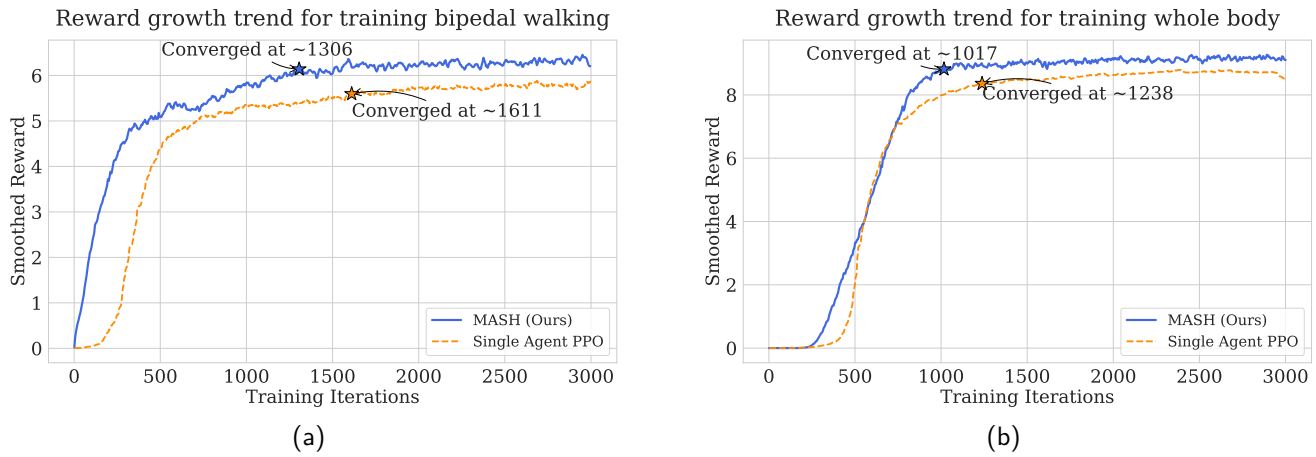

Fig. 4. 奖励增长趋势：(a) 腿部训练和 (b) 全身训练，比较 MASH 和单智能体 PPO 基线。平滑后的奖励与训练迭代次数的关系被绘制出来。

TABLE II
评估指标结果

Method	T_{Conv}		S_{action}		S_{torso}		C_{limb}	
	Bipedal	Arm-swing	Bipedal	Arm-swing	Bipedal	Arm-swing	Bipedal	Arm-swing
MASH (Ours)	$\sim$ **1306** (↓)	$\sim$ **1017** (↓)	**0.107** (↓)	**0.124** (↓)	$\mathbf{8.240 \times 10^{-4}}$ (↓)	$\mathbf{2.720 \times 10^{-3}}$ (↓)	**0.612** (↓)	**0.421** (↓)
Single-agent PPO	$\sim$ 1661	$\sim$ 1238	0.547	0.546	3.398×10^{-3}	2.802×10^{-2}	0.974	0.951

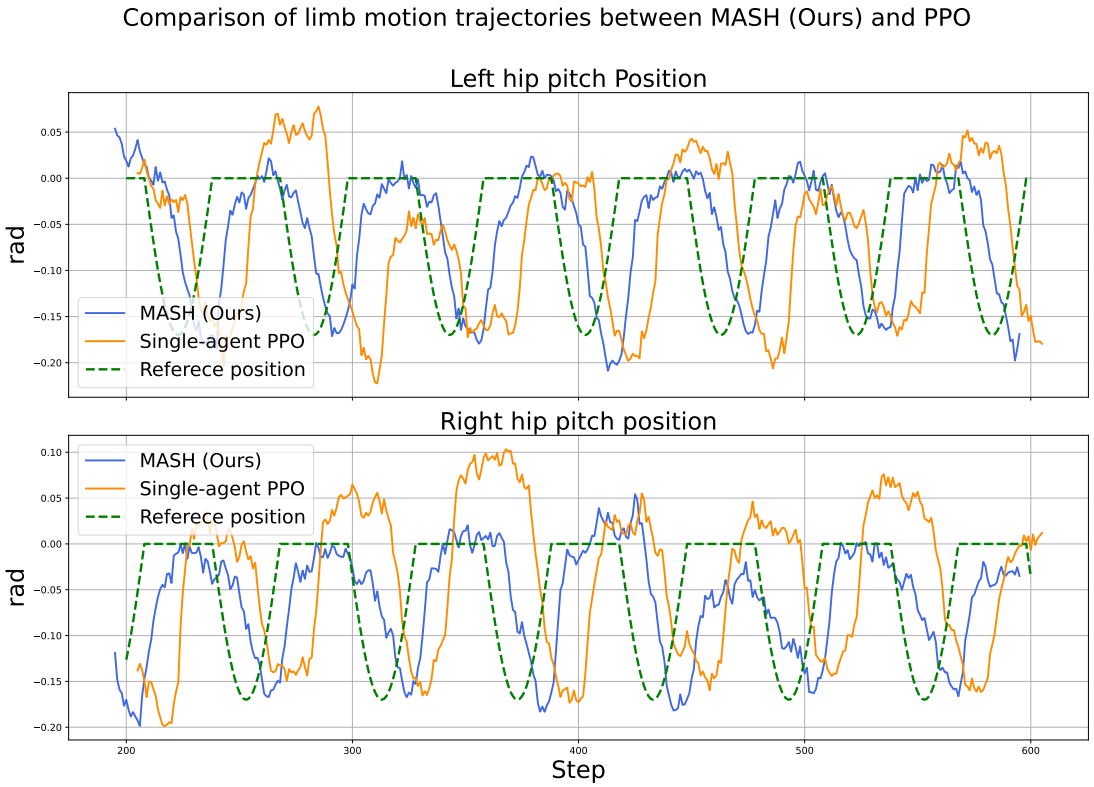

Fig. 5. 髋关节俯仰轨迹的比较。图表展示了我们的 MASH 控制器与用于左髋和右髋关节的单智能体 PPO 基线的跟踪性能。我们的方法（蓝色）紧密跟随参考轨迹（绿色，虚线），而基线（橙色）则表现出显著的超调和相位误差。

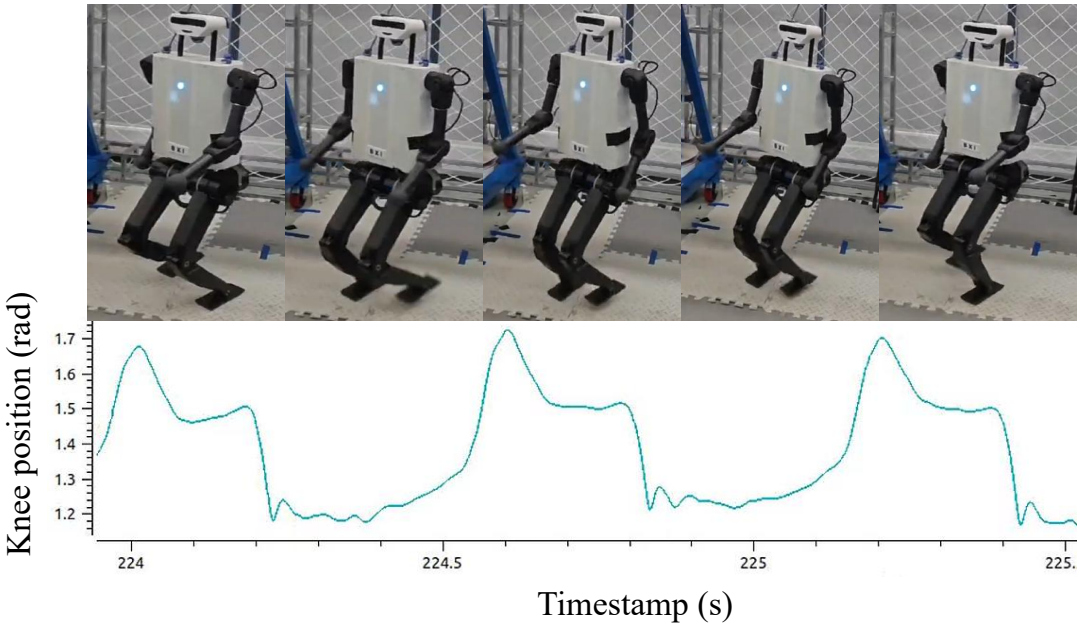

Fig. 6. 使用 MASH 训练的策略在物理仿人机器人上稳定步行步态。频闪图像（顶部）和对应的膝关节轨迹（底部）显示出顺畅、周期性的运动，确认了仿真到现实的成功转移。

改善肢体运动协调；(3) 通过域随机化增强鲁棒性。我们的实验结果表明，MASH 显著优于传统的单智能体 PPO 基线，实现了更快的训练收敛和更高的渐近奖励。学习到的策略在动作平滑性、躯干稳定性和肢体协调性方面表现出定量上的优越表现。重要的是，我们的方法的有效性和鲁棒性通过成功将学习到的策略转移到物理仿人机器人上得到了验证，该机器人执行了稳定且流畅的行走步态。这项研究不仅为复杂的单机器人控制引入了一个强大且高效的策略，还为应用多智能体学习范式推进仿人机器人行走提供了新的见解。未来的工作中，我们将把 MASH 应用于其他配置的机器人。

References

[1] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang et al. , "HOVER: Versatile neural whole-body controller for humanoid robots," in 2025 IEEE International Conference on Robotics and Automation (ICRA) , 2025.

[2] Q. Liu, Y. Li, X. Shi, K. Lin, Y. Liu, and Y. Lou, "Distributional policy gradient with distributional value function," IEEE Transactions on Neural Networks and Learning Systems , vol. 36, no. 4, pp. 6556–6568, 2025.

[3] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in Conference on Robot Learning . PMLR, 2023, pp. 22–31.

[4] I. M. Aswin Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in 2023 IEEE International Conference on Robotics and Automation (ICRA) , 2023, pp. 5078–5084.

[5] H. Benbrahim and J. A. Franklin, "Biped dynamic walking using reinforcement learning," Robotics and Autonomous Systems , vol. 22, no. 3-4, pp. 283–302, 1997.

[6] H. Duan, B. Pandit, M. S. Gadde, B. Van Marum, J. Dao, C. Kim, and A. Fern, "Learning vision-based bipedal locomotion for challenging terrain," in 2024 IEEE International Conference on Robotics and Automation (ICRA) , 2024, pp. 56–62.

[7] H. Lu, Y. Li, S. Mu, D. Wang, H. Kim, and S. Serikawa, "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning," IEEE Internet of Things Journal , vol. 5, no. 4, pp. 2315–2322, 2018.

[8] P. M. Wensing, M. Posa, Y. Hu, A. Escande, N. Mansard, and A. D. Prete, "Optimization-based control for dynamic legged robots," IEEE Transactions on Robotics , vol. 40, pp. 43–63, 2024.

[9] J.-P. Sleiman, F. Farshidian, and M. Hutter, "Versatile multicontact planning and control for legged loco-manipulation," Science Robotics , vol. 8, no. 81, p. eadg5014, 2023.

[10] X. Da and J. Grizzle, "Combining trajectory optimization, supervised machine learning, and model structure for mitigating the curse of dimensionality in the control of bipedal robots," The International Journal of Robotics Research , vol. 38, no. 9, pp. 1063–1097, 2019.

[11] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," Science Robotics , vol. 9, no. 89, p. eadi9579, 2024.

[12] Y. Kim, H. Oh, J. Lee, J. Choi, G. Ji, M. Jung, D. Youm, and J. Hwangbo, "Not only rewards but also constraints: Applications on legged robot locomotion," IEEE Transactions on Robotics , 2024.

[13] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," The International Journal of Robotics Research , p. 02783649241273668, 2023.

[14] S. Ha, J. Lee, M. van de Panne, Z. Xie, W. Yu, and M. Khadiv, "Learning-based legged locomotion: State of the art and future perspectives," The International Journal of Robotics Research , p. 02783649241312698, 2024.

[15] K. Jiang, Z. Fu, J. Guo, W. Zhang, and H. Chen, "Learning whole-body loco-manipulation for omni-directional task space pose tracking with a wheeled-quadrupedal-manipulator," IEEE Robotics and Automation Letters , 2024.

[16] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner et al. , "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," Science Robotics , vol. 9, no. 89, p. eadi8022, 2024.

[17] M. Liu, Z. Chen, X. Cheng, Y. Ji, R. Qiu, R. Yang, and X. Wang, "Visual whole-body control for legged loco-manipulation," in Conference on Robot Learning , 2024.

[18] K. Darvish, L. Penco, J. Ramos, R. Cisneros, J. Pratt, E. Yoshida, S. Ivaldi, and D. Pucci, "Teleoperation of humanoid robots: A survey," IEEE Transactions on Robotics , vol. 39, no. 3, pp. 1706–1727, 2023.

[19] B. Jia and D. Manocha, "Sim-to-real robotic sketching using behavior cloning and reinforcement learning," in 2024 IEEE International Conference on Robotics and Automation (ICRA) , 2024, pp. 18 272–18 278.

[20] S. Lyu, H. Zhao, and D. Wang, "A composite control strategy for quadruped robot by integrating reinforcement learning and model-based control," in 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) , 2023, pp. 751–758.

[21] Y. Chen, T. Wu, S. Wang, X. Feng, J. Jiang, Z. Lu, S. McAleer, H. Dong, S.-C. Zhu, and Y. Yang, "Towards human-level bimanual dexterous manipulation with reinforcement learning," in Advances in Neural Information Processing Systems , vol. 35, 2022, pp. 5150–5163.

[22] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," Journal of Machine Learning Research , vol. 21, no. 178, pp. 1–51, 2020.

[23] Q. Liu, Y. Li, Y. Liu, K. Lin, J. Gao, and Y. Lou, "Data efficient deep reinforcement learning with action-ranked temporal difference learning," IEEE Transactions on Emerging Topics in Computational Intelligence , vol. 8, no. 4, pp. 2949–2961, 2024.

[24] Q. Liu, J. Guo, S. Lin, S. Ma, J. Zhu, and Y. Li, "Masq: Multi-agent reinforcement learning for single quadruped robot locomotion," arXiv preprint arXiv:2408.13759 , 2024.

[25] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," in Robotics Science and Systems , 2024.

[26] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang, "Mobile-television: Predictive motion priors for humanoid whole-body control," in 2025 IEEE International Conference on Robotics and Automation (ICRA) , 2025.

[27] C. Zhang, W. Xiao, T. He, and G. Shi, "Wococo: Learning whole-body humanoid control with sequential contacts," in Conference on Robot Learning , 2024.

[28] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi, "Learning human-to-humanoid real-time whole-body teleoperation," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) , 2024, pp. 8944–8951.

[29] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn, "Humanplus: Humanoid shadowing and imitation from humans," in Conference on Robot Learning , 2024.

[30] J. Li, Y. Zhu, Y. Xie, Z. Jiang, M. Seo, G. Pavlakos, and Y. Zhu, "Okami: Teaching humanoid robots manipulation skills through single video imitation," in Conference on Robot Learning , 2024.

[31] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi, "OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," in Conference on Robot Learning , 2024.

[32] Z. Luo, J. Cao, K. Kitani, W. Xu et al. , "Perpetual humanoid control for real-time simulated avatars," in Proceedings of the IEEE/CVF International Conference on Computer Vision , 2023, pp. 10 895–10 904.

[33] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Robust and versatile bipedal jumping control through reinforcement learning," in Robotics Science and Systems , 2023.

[34] Z. Xie, G. Berseth, P. Clary, J. Hurst, and M. Van de Panne, "Feedback control for cassie with deep reinforcement learning," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) . IEEE, 2018, pp. 1241–1246.

[35] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in 2021 IEEE International Conference on Robotics and Automation (ICRA) . IEEE, 2021, pp. 7309–7315.

[36] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character control," ACM Transactions on Graphics (ToG) , vol. 40, no. 4, pp. 1–20, 2021.

[37] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," arXiv e-prints , pp. arXiv–2408, 2024.

[38] M.-A. Blais and M. A. Akhloufi, "Reinforcement learning for swarm robotics: An overview of applications, algorithms and simulators," Cognitive Robotics , vol. 3, pp. 226–256, 2023.

[39] M. Zhou, J. Luo, J. Villella, Y. Yang, D. Rusu, J. Miao, W. Zhang, M. Alban, I. Fadakar, Z. Chen et al. , "Smarts: An open-source scalable multi-agent rl training school for autonomous driving," in Conference on robot learning . PMLR, 2021, pp. 264–285.

[40] S. Lim, H. Yu, and H. Lee, "Optimal tethered-uav deployment in a2g communication networks: Multi-agent q-learning approach," IEEE Internet of Things Journal , vol. 9, no. 19, pp. 18 539–18 549, 2022.

[41] Q. Liu, J. Gao, D. Zhu, Z. Qiao, P. Chen, J. Guo, and Y. Li, "Multi-agent target assignment and path finding for intelligent warehouse: A cooperative multi-agent deep reinforcement learning perspective," arXiv preprint arXiv:2408.13750 , 2024.

[42] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction . MIT press, 2018.

[43] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347 , 2017.

[44] S. C. Ong, S. W. Png, D. Hsu, and W. S. Lee, "Pomdps for robotic tasks with mixed observability." in Robotics: Science and Systems , vol. 5, 2009, p. 4.

[45] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," in Advances in Neural Information Processing Systems , vol. 35, 2022, pp. 24 611–24 624.

[46] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in International Conference on Machine Learning . PMLR, 2017, pp. 1146–1155.

[47] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa et al. , "Isaac gym: High performance gpu based physics simulation for robot learning," in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) , 2021.