

虚拟人运动控制策略学习方法的研究进展与展望

叶永竞^{1,2)}, 许逸文^{1,2)}, 张子豪¹⁾, 胡磊^{1,2)}, 夏时洪^{1,2)*}

¹⁾(中国科学院计算技术研究所前瞻研究实验室 北京 100190)

²⁾(中国科学院大学 北京 100049)

(xsh@ict.ac.cn)

摘要: 虚拟人运动合成是虚拟现实和角色动画领域中的关键问题之一,旨在合成真实自然且能够响应用户输入信息的运动序列。虚拟人运动控制策略根据用户输入约束解算关节力矩,并依托现有的物理引擎更新虚拟人状态,合成的运动序列在满足用户输入约束的同时可以满足物理真实性。近年来,深度强化学习技术因在序列决策和交互任务中的出色表现而备受研究者的关注,为基于物理引擎的虚拟人控制策略学习提供了新途径。文中对虚拟人运动控制策略学习方法进行综述,从理论基础和应用设计等方面介绍相关研究。在应用设计方面,首先基于深度强化学习的基础元素,从状态表示、奖励函数设计、控制策略设计以及物理仿真 4 个角度对现有工作进行梳理总结;其次,分析现有通用技术框架并指出其在控制策略上的拓展方向,并以实际问题为例探讨虚拟人运动控制策略的具体应用。最后,总结当前该领域的研究现状,指出利用丰富的运动捕获数据提升运动控制策略的深度与广度是未来的主要研究方向,展望虚拟人运动控制策略学习方法在多模态的感知与控制、世界模型学习和具身智能等应用方向的发展前景。

关键词: 虚拟人; 运动控制策略; 强化学习; 深度学习; 角色动画

中图分类号: TP391.41

DOI: 10.3724/SP.J.1089.2024-00168

Recent Advances on Motion Control Policy Learning for Humanoid Characters

Ye Yongjing^{1,2)}, Xu Yiwen^{1,2)}, Zhang Zihao¹⁾, Hu Lei^{1,2)}, and Xia Shihong^{1,2)*}

¹⁾(Prospective Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190)

²⁾(University of Chinese Academy of Sciences, Beijing 100049)

Abstract: Motion synthesis for humanoid characters, aimed at generating realistic and natural motion sequences that can respond to user input, has long been a formidable challenge in the fields of virtual reality and character animation. Motion control policies address this challenge by calculating joint torques based on user input constraints, updating the character state using existing physics engines, and synthesizing motion sequences that not only meet user input constraints but also ensure physical realism. In recent years, deep reinforcement learning has garnered significant attention from researchers due to its exceptional performance in sequential decision-making and interactive tasks, providing a novel approach for learning control policies for humanoid characters grounded in physics engines. This paper reviews the advancements in motion control policy learning for humanoid characters and introduces relevant research from both theoretical foundations and practical design perspectives. In terms of practical designs, existing works are examined from four key aspects: state representation, reward function design, control policy design, and the physical simulation engine employed, all based on the fundamental elements of deep reinforcement learning. Furthermore, a comprehensive

收稿日期: 2024-04-07; 修回日期: 2024-05-21. 基金项目: 工业软件重点专项(2022YFB3303202); 国家自然科学基金(62302481).
叶永竞(1995—), 男, 博士研究生, 主要研究方向为计算机图形学、虚拟现实; 许逸文(1997—), 男, 博士研究生, 主要研究方向为计算机图形学、虚拟现实; 张子豪(1993—), 男, 博士, 助理研究员, 主要研究方向为计算机图形学、计算机视觉; 胡磊(1997—), 男, 博士研究生, 主要研究方向为计算机图形学、三维人体运动建模与生成; 夏时洪(1974—), 男, 博士, 研究员, 博士生导师, CCF 高级会员, 论文通信作者, 主要研究方向为计算机图形学、虚拟现实、人工智能。

analysis of a general technical framework is conducted, highlighting potential directions for extending control policies. The specific application of motion control policies for humanoid characters is discussed using practical problems as case studies. Finally, a summary of the current research status is provided, indicating that leveraging extensive motion capture data to enhance the depth and breadth of motion control policies represents a promising future research direction. The paper also outlines the prospects for the development of motion control policy learning for humanoid characters, particularly in the areas of multimodal perception and control, world model learning, and embodied intelligence.

Key words: humanoid character; motion control policy; reinforcement learning; deep learning; character animation

作为人体在虚拟环境中的数字化表达,虚拟人是虚拟现实和角色动画的重要组成部分,合成真实自然且能响应用户输入信息的虚拟人运动,一直是虚拟人研究中的热门问题之一。根据是否建模虚拟人的物理属性,当前的主流方法可分为运动学方法和动力学方法。运动学方法通常采用生成式模型,从运动捕获数据集中学习运动流形,再根据用户的输入信息输出虚拟人的关节位置、朝向等运动学信息,生成的运动往往存在脚步滑动、漂移等物理不真实的现象^[1-2]。动力学方法通过对虚拟人和环境进行物理真实的建模,结合物理约束求解关节驱动信息并模拟真实人体的肌肉发力模式,生成物理真实的虚拟人运动序列,可以有效地克服运动学方法中存在的问题。主流的动力学方法采用物理引擎完成虚拟人的建模和复杂的物理约束计算,并设计运动控制策略,从用户输入的任务目标解算虚拟人关节驱动信息,比其他方法更加简便。传统的运动控制策略设计通常使用回归方法^[3],甚至需要结合复杂的手工设计^[4],这些方法往往存在鲁棒性不足的问题,且只能完成有限的任务目标。因此,迫切需要一种智能化的虚拟人运动控制策略学习框架,在准确地响应多样化的用户输入同时,能够简化控制策略的学习过程并提升其泛化能力和鲁棒性。

近年来,深度强化学习的发展为上述基于物理引擎的虚拟人运动控制策略学习提供了可行的途径。由于强化学习与虚拟人运动控制的基础元素具有高度对应的关系,因此可以很自然地将深度强化学习应用到基于物理引擎的虚拟人运动控制策略学习中。在由特定任务目标定义的奖励的指导下,虚拟人在由物理引擎构建的仿真环境中进行探索,进而学习到最大化累积奖励的运动控制策略。在深度强化学习框架下,基于深度神经网络

建模的控制策略可以充分地利用物理引擎进行大规模样本采集,并拟合高维状态空间和动作空间之间复杂的关系,能够胜任传统控制策略无法实现的复杂任务目标^[5-6]。

当前,已有一些综述文献讨论了基于深度强化学习的虚拟人运动控制策略。Mourot 等^[7]从虚拟人运动合成角度切入,综合性地讨论了基于深度学习和基于深度强化学习的虚拟人运动合成的运动学和动力学方法;Kwiatkowski 等^[8]则是从深度强化学习角度出发,广泛地讨论了单智能体和多智能体强化学习算法的原理,及其在虚拟人运动合成和人群仿真方面的应用。这些综述更多的是从全局角度对虚拟人运动合成或者深度强化学习算法的应用进行回顾与描述,单一虚拟人运动控制问题只作为其中的一部分,并没有深入讨论。

本文将聚焦基于深度强化学习算法的单一虚拟人运动控制问题。(1) 为了准确地定义研究问题,先介绍虚拟人运动控制策略和强化学习的预备知识;在此基础之上,回顾运动控制策略学习算法,即目前主流的深度强化学习算法,并分析其具体应用场景。(2) 将以强化学习基础元素为依据,对主流的方法进行分类,讨论在解决实际虚拟人运动控制问题时相关元素的应用设计,并分析实际应用案例。(3) 总结虚拟人运动控制策略学习方法的研究现状和挑战,并展望发展趋势,为未来的研究提供启示,推动本领域的进一步发展。

1 预备知识与问题定义

1.1 虚拟人运动控制策略

本文讨论的虚拟人运动控制策略特指利用物理引擎学习的虚拟人运动控制策略。物理引擎基于基本物理定律,通过复杂的计算过程模拟真实

世界的各种场景和物理现象。基于物理引擎,可以对真实人体及其所处的环境的几何和物理属性进行建模,构建物理真实的仿真环境和虚拟人。为了响应用户输入的任务目标,需要为虚拟人设计运动控制策略。运动控制策略以虚拟人的状态参数作为输入,根据由任务目标定义的控制参数解算虚拟人的关节驱动信息,再将其交由物理引擎进行仿真,最终完成相应的任务目标。综上所述,要准确地定义基于物理引擎的虚拟人运动控制策略学习问题,需要考虑3个基本问题:(1)虚拟人建模仿真;(2)控制策略设计;(3)物理引擎仿真流程。

1.1.1 虚拟人建模仿真

在物理仿真环境中,虚拟人通常被建模成树形铰接多刚体系统,每个刚体模拟真实人体的特定肢体段,这些刚体的质量、质心和惯性张量等信息参考真实人体的解剖数据进行设置。根据真实人体中对应的骨骼组织和韧带所施加的约束,选择连接这些刚体的仿真关节类型。例如,只有1个自由度的铰接关节用于模拟膝关节,而3自由度的球窝关节则被用来模拟肩关节和髋关节。通过设置每个仿真关节中每个自由度的旋转上下限,可以最大程度地模拟真实人体的关节。依据上述方式建模得到的虚拟人表示形式称为骨架树。图1所示为在PhysX物理引擎^①中建模并绘制的典型虚拟人骨架树,其中有20个可控身体关节,总计46个自由度。通常,用6自由度的自由关节表示位于虚拟人骨盆处的骨架树根节点,用以描述该虚拟人在世界坐标系中的位置与朝向。虚拟人通过关节驱动机制实现主动运动,主流的关节驱动机制包括肌肉发力模型和伺服器方法。肌肉发力模型通过模拟人体生理系统中肌肉收缩发力的过程计算关节力矩,力矩大小取决于关节姿态,以及肌肉收缩的距离和速度;伺服器方法假设每个关节都有一个虚拟的伺服电机,这些电机将直接对每个激活的自由度施加力矩。由于主动运动不依赖根关节处的力或力矩,因此虚拟人的关节驱动机制是欠驱动的。伺服器方法驱动关节运动的方式更为直接,是虚拟人运动控制中采用的主流方法。在伺服器方法中,关节力矩 $\tau = K_p(\bar{q} - q) - K_d\dot{q}$ 的计算可以通过比例微分(proportional derivative, PD)控制器实现。其中, K_p 和 K_d 表示PD控制器的固定

参数; $\bar{q}$ 表示对应自由度的目标旋转, q 和 $\dot{q}$ 则表示自由度当前的旋转和角速度。目前,绝大部分的工作都是通过这种简化的人体表示和基于PD控制的伺服器方法来探究虚拟人运动控制策略,但也有一些工作探究更精细建模下的运动控制策略。

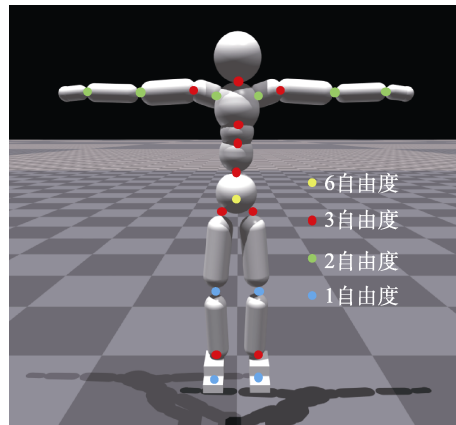


图1 虚拟人骨架树示意图

1.1.2 控制策略设计

由于虚拟人的关节较多,总体自由度较高,手工控制每个关节来驱动人体实现相关任务目标并不现实,需要利用一个高层控制策略 π ,根据当前虚拟人体的状态 s 和任务目标 g 计算输出驱动信息 $a = \pi(s, g)$ 。如果此策略是随机性的,则可以表示为 $a \sim \pi(a|s, g)$ 。虚拟人体状态 s 包括每个关节的位置、旋转和速度等信息,驱动信息可以是关节力矩 τ ,也可以是PD控制器的目标旋转 $\bar{q}$ 。计算得到的关节力矩 τ 交由物理引擎进行仿真,以更新虚拟人体状态。任务目标 g 可以是传统任务,如速度和位置;也可以是更加复杂的视频输入追踪目标。为了简洁起见,下文若不强调任务目标,将在策略表示中省略 g 。

1.1.3 物理引擎仿真流程

在仿真环境中,虚拟人依赖物理引擎进行状态更新,更新人体状态分为3步:

Step1. 碰撞检测。检查仿真环境中不同刚体之间是否存在接触关系,对于判定接触的刚体添加接触关节约束,接触关节一般是自由关节。

Step2. 正向动力学计算。基于施加的外力、关节力矩和关节约束,计算每个刚体的加速度和角加速度。

Step3. 数值积分。根据计算得到的加速度和角加速度信息,更新每个刚体的速度、位置和旋转。

^① <https://developer.nvidia.com/physx-sdk>

目前常用的开源的物理仿真引擎包括 Bullet^①, MuJoCo^[9], DART^[10], PhysX 等.

1.2 强化学习基础

强化学习的主体是智能体, 与智能体交互的对象是环境. 本质上, 强化学习问题就是智能体通过观察环境的状态选择动作, 并根据环境给出的奖励来学习如何在不同的状态下做出最优的决策. 强化学习的目标是找到一个最优策略, 使得智能体在与环境交互的过程中, 始终选择能最大化长期累积奖励的期望值的动作.

强化学习的数学基础和建模工具是马尔可夫决策过程(Markov decision process, MDP), 该过程由状态空间 S 、动作空间 A 、状态转移函数 W 、奖励函数 R 和折扣因子 γ 组成. 在一个回合开始, 即时间步 $t=0$ 时, 首先指定或采样获得初始环境状态 $s_t = s_0$, 智能体从动作空间 A 中选取动作 a_t 并执行; 然后观测到新的环境状态 $s_{t+1} \sim W(s_t, a_t)$ 和相应的奖励 $r_t \sim R(s_t, a_t, s_{t+1})$; 从新的环境状态出发不断地重复这一过程, 直至到达终止状态或达到预定义的时长极限, 即可获得此回合的轨迹 $\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T)$. 其中, T 表示轨迹长度. 智能体根据不同状态选择动作的方式称为策略. 在 ρ

上可计算折扣回报(简称回报), 即 $U(\rho) = \sum_{t=0}^{T-1} \gamma^t r_t$.

其中, $\gamma \in [0, 1]$ 是一个折扣因子, 目的是让智能体更关注短期奖励. 智能体的目标就是最大化期望回报, 有 2 种方法可以实现这个目标: (1) 基于价值的算法. 先评估各状态的优劣, 再根据状态转移路径获取最优策略. (2) 基于策略的算法. 直接优化策略, 让智能体根据当前状态选择最优的动作.

1.2.1 价值函数与贝尔曼方程

基于价值的算法使用价值对状态和动作进行评估. 给定状态 s , 智能体基于当前策略 π (初始策略一般为随机策略) 采样得到长度为 T 的轨迹 $\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T)$, 状态价值函数定义为在所有可能轨迹上的平均回报

$$V(s) = E_{\rho \sim \pi | s_0 = s} \sum_{t=0}^{T-1} \gamma^t r_t.$$

其中, $\rho \sim \pi$ 表示轨迹 ρ 由 π 生成. 若智能体在状态 s 下采取的动作 a 也给定, 可类似地定义动作价

值函数为 $Q(s, a) = E_{\rho \sim \pi | s_0 = s, a_0 = a} \sum_{t=0}^{T-1} \gamma^t r_t$.

在基于价值的算法中, 若已估计出价值函数, 则直接通过 $a = \arg\max_{a'} Q(s, a')$ 可获取当前最优策略. 再从新策略出发不断重复上述过程, 在良好的条件下价值函数将收敛至准确的价值, 对应的策略也收敛为最优策略.

贝尔曼方程^[11]可以将不同时刻的价值联系起来, 提供一种通过时间差分(time difference, TD)算法迭代求解价值函数的方法. 状态价值函数和动作价值函数都有对应的贝尔曼方程分别为

$$V(s) = E_{a \sim \pi(a|s; \theta), s' \sim W(s, a)} [R(s, a, s') + \gamma V(s')],$$

$$Q(s, a) = E_{s' \sim W(s, a)} [R(s, a, s') + \gamma E_{a' \sim \pi(a'|s'; \theta)} Q(s', a')].$$

贝尔曼方程表明, 某一时刻的价值函数等于该时刻的奖励值与下一时刻折扣价值函数之和. 下面以动作价值函数为例证明贝尔曼方程: 为简洁起见, 记 $r_t = R(s_t, a_t, s_{t+1})$, 另记 t 时刻至回合结束的回报为 U_t , 根据回报定义可以得到 $U_t = r_t + \gamma U_{t+1}$. 用符号 S_{t+1}, A_{t+1} 表示从 $t+1$ 时刻起所有的状态和动作随机变量. 根据动作价值函数的定义 $Q(s_t, a_t) = E_{S_{t+1}, A_{t+1}} [U_t | s_t, a_t]$, 将 U_t 替换成 $r_t + \gamma U_{t+1}$, 得到

$$Q(s_t, a_t) = E_{S_{t+1}, A_{t+1}} [r_t + \gamma U_{t+1} | s_t, a_t] =$$

$$E_{S_{t+1}, A_{t+1}} [r_t | s_t, a_t] + \gamma E_{S_{t+1}, A_{t+1}} [U_{t+1} | s_t, a_t] \quad (1)$$

假定 r_t 是 s_t, a_t, s_{t+1} 的函数, 如果给定 s_t 和 a_t , 则 r_t 随机性唯一的来源就是 s_{t+1} , 所以

$$E_{S_{t+1}, A_{t+1}} [r_t | s_t, a_t] = E_{s_{t+1}} [r_t | s_t, a_t] \quad (2)$$

式(1)右边 U_{t+1} 的期望可以写成

$$E_{S_{t+1}, A_{t+1}} [U_{t+1} | s_t, a_t] =$$

$$E_{s_{t+1}, a_{t+1}} [E_{S_{t+2}, A_{t+2}} [U_{t+1} | s_{t+1}, a_{t+1}] | s_t, a_t] =$$

$$E_{s_{t+1}, a_{t+1}} [Q(s_{t+1}, a_{t+1}) | s_t, a_t] \quad (3)$$

由式(1)~式(3), 可得动作价值函数的贝尔曼方程.

此外, 在最优策略上定义最优状态价值函数 $V^*(s) = \max_{\pi} V_{\pi}(s), \forall s \in \mathcal{S}$ 和最优动作价值函数

$Q^*(s, a) = \max_{\pi} Q_{\pi}(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}$. 类似地, 可

以得到贝尔曼最优方程为

$$V^*(s) = \max_a E_{s' \sim T(s, a)} [R(s, a, s') + \gamma V^*(s')],$$

$$Q^*(s, a) = E_{s' \sim W(s, a)} [R(s, a, s') + \gamma \max_{a'} Q^*(s', a', s')].$$

① <http://bulletphysics.org>

在基于价值的算法中, 贝尔曼方程等号右侧称为 TD 目标, 而等号两侧的误差则称为 TD 误差, 表示不同状态下的价值函数估计的差异, 在估计价值函数时用作可微的优化目标。

1.2.2 策略函数与策略梯度

与基于价值的算法不同, 基于策略的算法将策略用函数 $\pi(a|s; \theta)$ 近似, 其中, θ 表示函数参数, 称为策略函数。考虑由策略 π 生成的轨迹 $\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T)$, 由轨迹的马尔可夫性可知, 此轨迹生成的概率为

$$p_{\theta}(\rho) = p(s_0) \prod_{t=0}^{T-1} p(s_{t+1} | s_t, a_t) \pi(a_t | s_t; \theta).$$

其中, $p(s_0)$ 表示初始状态分布。下面将从任意初始状态开始, 不断应用策略 π 生成轨迹 ρ 的概率分布, 并将其记为 $\rho \sim \pi(a|s; \theta)$, 在不引起混淆的情况下, 简化为 $\rho \sim \pi$ 。选择期望回报作为策略评估的准则, 将期望回报视为策略函数权重 θ 的函数, 可得优化目标 $J(\theta) = E_{\rho \sim \pi(a|s; \theta)} U(\rho)$ 。此目标函数的梯度表达式称为策略梯度。下面推导其计算方法, 公式为

$$\begin{aligned} \nabla_{\theta} J(\theta) &= \nabla_{\theta} E_{\rho \sim \pi} U(\rho) = \\ &= \nabla_{\theta} \int_{\rho} p_{\theta}(\rho) U(\rho) d\rho = \int_{\rho} \nabla_{\theta} p_{\theta}(\rho) U(\rho) d\rho = \\ &= \int_{\rho} p_{\theta}(\rho) \nabla_{\theta} \ln p_{\theta}(\rho) U(\rho) d\rho = \\ &= E_{\rho \sim \pi} [\nabla_{\theta} \ln p_{\theta}(\rho) R(\rho)] = \\ &= E_{\rho \sim \pi} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \ln \pi(a_t | s_t; \theta) R(\rho) \right]. \end{aligned}$$

策略梯度是一个数学期望, 可以使用蒙特卡罗方法进行估算。假定由策略 π 采样得到 N 个轨迹长度为 T 的回合数据, 则在这些轨迹上有

$$\nabla_{\theta} J(\theta) \approx \frac{1}{N} \sum_{i=0}^N \sum_{t=0}^{T-1} \nabla_{\theta} \ln \pi(a_t^i | s_t^i; \theta) U_t^i.$$

其中, a_t^i 表示第 i 条轨迹 τ^i 在时间步 t 的动作; s_t^i 表示对应的状态; U_t^i 表示从时间步 t 开始的累积回报。这种算法称为 REINFORCE^[12]。

蒙特卡罗方法易受样本方差影响, 导致策略训练过程不稳定。在实际计算策略梯度时, 为了提高稳定性和效率, 一般使用价值函数来近似期望回报 $R(\rho)$ 。例如, 使用动作价值函数

$$\nabla_{\theta} J(\theta) = E_{\rho \sim \pi} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \ln \pi(a_t | s_t; \theta) Q(s_t, a_t) \right].$$

动作价值函数采用与第 1.2.1 节类似的方法学习。

实际应用中, 使用优势函数 $A(s, a) = Q(s, a) - V(s)$ 替代动作价值函数以进一步降低方差。它直观地表明了状态 s 下因采取动作 a 而获得的优势。此时, 策略梯度为

$$\nabla_{\theta} J(\theta) = E_{\rho \sim \pi} \left[\sum_{t=0}^{T-1} \nabla_{\theta} \ln \pi(a_t | s_t; \theta) A_t(s_t, a_t) \right].$$

1.3 问题定义

如图 2 所示, 基于物理引擎的虚拟人运动控制中, 基本组件与强化学习框架的基本元素间有天然的对对应关系, 如虚拟人和智能体、物理引擎和环境等的对应关系。因此, 基于物理引擎的虚拟人运动控制可以很自然地嵌入到强化学习框架中, 建模如下:

- (1) 状态空间 S 。在物理仿真环境中, 状态度、关节角度等信息, 以及虚拟人观测到的环境状态, 如物体的位置和形状、用户输入的控制信号等。
- (2) 动作空间 A 。动作可以表示为虚拟人在物理仿真环境中的运动行为 $a \in A$, 即驱动人体行动的关节力矩信息。
- (3) 奖励函数 R 。奖励函数用于评估智能体在特定状态下采取特定动作的好坏程度。在物理仿真环境中, 可以根据虚拟人的运动效果和目标完成情况设计相应的奖励函数。
- (4) 状态转移函数 W 。在物理仿真方法中, 状态转移通过物理引擎仿真实现。当虚拟人基于当前状态 s_t 采取动作 a_t 之后, 经过物理引擎仿真后可以得到下一时刻的状态 s_{t+1} 。
- (5) 控制策略 π 。使用深度神经网络来近似控制策略 π 相关的策略函数或价值函数。通过与环境的交互和反馈优化网络参数, 虚拟人能够学习到符合任务目标的最优运动控制策略。

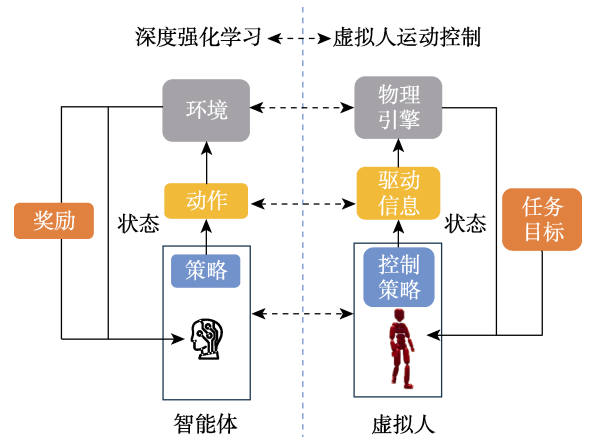


图2 强化学习与虚拟人运动控制之间的关系

在每个控制时间步 t , 虚拟人基于当前状态 s_t 和控制策略选取一个动作 a_t . 经过仿真之后, 虚拟人观测到一个新状态 $s' = s_{t+1}$ 和一个基于奖励函数 R 计算的奖励值 r_t , 这样的交互过程不断重复直至达到终止状态, 最终可以得到一个回合轨迹 $\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_T)$. 无论是求解价值函数 $V(s; \theta)$ 或 $Q(s, a; \theta)$ 还是优化策略函数 $\pi(a|s; \theta)$, 虚拟人的目标都是学习最优参数 $\theta^* = \operatorname{argmax}_{\theta} E_{\tau} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$, 使所有可能轨迹上的期望回报最大化.

虚拟人的运动控制问题涉及高维度的状态空间, 虚拟人的关节角度、速度、位置等都可能是状态的一部分, 导致状态空间的巨大复杂性, 增加了学习的难度. 此外, 虚拟人的运动控制需要在高度连续的动作空间中进行决策, 增加了动作选择的复杂性和计算的难度. 除去选择合适的深度强化学习算法来应对以上挑战外, 解决基于深度强化学习的虚拟人运动控制问题的关键之一是, 如何设计合适的任务目标奖励函数, 它直接影响强化学习算法的学习效果和虚拟人的运动行为. 设计奖励函数需要考虑任务目标、终止条件和奖励函数值范围等; 其次, 随着任务目标的复杂化, 简单的网络结构已经不能满足控制策略学习的需求, 设计更加复杂的网络结构并使用合理的策略学习方案也是需要考虑的问题之一.

2 虚拟人运动控制策略学习主流算法

本节讨论的虚拟人运动控制策略学习主流算法特指控制策略学习使用的主流深度强化学习算法. 基于 3 种不同的分类标准, 深度强化学习算法可以分成基于价值的和基于策略的、同策略和异策略, 以及基于模型的和无模型的等不同类型, 这些分类标准不是互斥的, 许多算法可以同时满足多个标准.

如图 3 所示, 本文依据基于价值的算法和基于策略的算法对深度强化学习算法进行分类, 分类的依据是智能体的决策是由策略函数或价值函数主导. 在基于策略的算法中, 策略函数起主导作用, 通过优化策略函数学习最优策略, 以最大化期望回报; 典型的算法有 REINFORCE^[12]、置信域策略优化(trust region policy optimization, TRPO)^[13]和近端策略优化(proximal policy optimization, PPO)^[14]

等. 在基于价值的算法中, 价值函数起主导作用, 通过学习和优化价值函数指导决策, 典型的算法有深度 Q 网络(deep Q-network, DQN)^[15-16]及其后续改进^[17-22]. 需要注意的是, 当前的深度强化学习算法通常会结合以上 2 种算法. 在 actor-critic 架构中, 算法使用策略函数生成动作, 并使用价值函数评估动作的价值, 且同时学习和优化这 2 部分; 典型的算法有深度确定性策略梯度(deep deterministic policy gradient, DDPG)算法^[23]和软性策略评估(soft actor-critic, SAC)^[24]等.

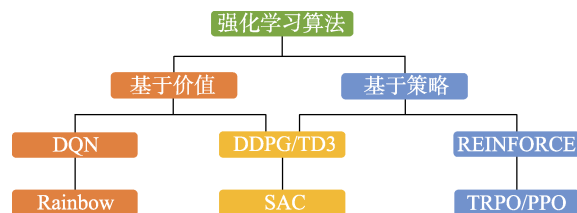


图 3 深度强化学习算法分类

2.1 DQN

DQN^[15-16]首次将深度神经网络与经典的基于价值的强化学习算法 Q 学习^[25]结合, 解决传统的强化学习算法难以应对高维状态空间的问题. DQN 使用深度神经网络作为最优价值函数的逼近器, 能够处理更复杂的环境和更大的状态空间.

DQN 中智能体的决策基于价值网络 $Q(s, a; \omega)$ 和 ε -greedy 策略. 价值网络由一个参数为 ω 的深度神经网络表示. 将状态 s 作为输入, 神经网络输出每个可能动作对应的动作价值, 称为 Q 值, 智能体以概率 ε 选择随机动作, 以概率 $1-\varepsilon$ 选择具有最高 Q 值的动作. 训练价值网络的目标函数为贝尔曼最优方程的 TD 误差的平方; 同时, DQN 还通过经验回放和目标网络^[16]这 2 个关键技术解决学习稳定性的问题. 把智能体在一个回合中的轨迹记作 $\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots, s_n)$, 将其分为 n 个 (s_t, a_t, r_t, s_{t+1}) 元组并存入数组. 其中, $t = 0, 1, \dots, n-1$, 这个数组称为经验回放数组, 通过随机抽样存储在经验回放数组中的数据减少样本间的时间相关性, 有助于提高训练的稳定性. 目标网络 $Q(s, a; \omega^-)$ 是一个与 Q 网络结构相同的神经网络, 其参数 ω^- 与 Q 网络的参数 ω 在训练过程中是分离的. 目标网络的参数通过软更新(soft update)的方式从 Q 网络中定期更新. 目标网络通过旧参数计算更加稳定的目标值, 提高了训练的稳定性.

使用经验回放和目标网络的 DQN 算法步骤

如下:

- Step1. 初始化容量为 N 的经验回放数组 D .
- Step2. 随机初始化价值网络 Q , 权重为 ω .
- Step3. 初始化 Q 的目标网络, 权重 $\omega^- \leftarrow \omega$.
- Step4. 判断迭代次数是否达到上限, 若达到, 转 Step15.
- Step5. 初始化时间步 $t = 0$, 获取初始状态 $s_t = s_0$.
- Step6. 判断轨迹长度达到上限, 若达到, 转 Step4.
- Step7. 根据价值网络 Q 和 ε -greedy 策略选取动作 a_t .
- Step8. 执行动作 a_t , 获得状态 s_{t+1} 和奖励 r_t .
- Step9. 将 (s_t, a_t, r_t, s_{t+1}) 保存到 D 中.
- Step10. 从 D 中随机批量采样.
- Step11. 若回合中止, 计算 TD 目标 $y_j \leftarrow r_j$; 否则, 计算 TD 目标 $y_j \leftarrow r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \omega^-)$.
- Step12. 以 $(y_j - Q(s_j, a_j; \omega))^2$ 为目标执行梯度下降, 更新价值网络 Q 的权重 ω .
- Step13. 每隔若干步数软更新目标网络参数.
- Step14. 更新 $t \leftarrow t + 1$, 转 Step6.
- Step15. 输出价值网络 Q 的权重 ω .

虽然 DQN 在解决各种强化学习问题方面取得了显著的成功, 但是它的局限性在于只能解决离散动作空间的问题, 对于连续动作空间, 需要将其离散化或者采用其他的算法.

已有一些算法在 DQN 上进行改进, 如 Double DQN^[17], Dueling DQN^[18], 优先经验回放^[19], 分布式强化学习^[20], 噪声网络^[21]等, 这些算法被集成在 Rainbow^[22]框架中, 使得 DQN 在处理复杂问题时更加高效和鲁棒.

2.2 TRPO 和 PPO

第 1.2.2 节所述, TRPO^[13]和 PPO^[14]使用优势函数替换回报, 进而估计策略梯度的算法, 并通过引入置信域限制策略更新的幅度, 以提升算法的训练稳定性和收敛性.

TRPO 在每次策略更新时, 除了兼顾最大化策略性能, 还要保持策略更新在一个可接受的置信域范围内. 使用更新前后策略的 KL 散度衡量策略更新大小, 确保更新后的策略 $\pi(a|s; \theta)$ 不会相距之前的策略 $\pi(a|s; \theta')$ 太远, 公式为

$$\max_{\theta} L(\theta, \theta') \quad \text{s.t.} \quad D_{\text{KL}}(\theta \| \theta') \leq \delta.$$

其中, δ 是表示置信域大小的超参数; $L(\theta, \theta') = E_{\rho \sim \pi(a|s; \theta)} [\pi(a|s; \theta) / \pi(a|s; \theta') A(s, a)]$ 表示新策略相对于旧策略的优势.

TRPO 使用共轭梯度法计算在置信域约束下

的策略更新最优步长, 导致计算复杂度较高. 同时, 置信域大小 δ 是一个重要的超参数, 它决定了策略更新的幅度. 选择合适的置信域大小对算法的性能和收敛性至关重要, 然而确定合适的置信域大小并不是一项容易的任务, 需要进行实验和调优. 由于这种复杂性, 目前 TRPO 在实践中已经很少使用.

为了简化优化问题求解, PPO 在 TRPO 的基础上舍弃了显式的约束条件, 而是通过引入近端优化机制避免过大的策略更新, 进而保持训练的稳定性. PPO 的目标函数使用 PPO-Clip 形式^[14], 即

$$L(\theta, \theta') = E_{\rho \sim \pi(a|s; \theta)} \left[\min(r(\theta) A(s, a), \text{clip}(l(\theta), 1 - \varepsilon, 1 + \varepsilon) A(s, a)) \right].$$

其中, $r(\theta) = \pi(a|s; \theta) / \pi(a|s; \theta')$ 表示 2 个策略的比值; $\varepsilon \ll 1$ 是一个超参数; $\text{clip}(r(\theta), 1 - \varepsilon, 1 + \varepsilon)$ 为截断函数, 将 $r(\theta)$ 截断在 $[1 - \varepsilon, 1 + \varepsilon]$, 保证新旧策略的相似性. PPO 在保证计算效率和更易实现的同时, 能够得到比肩 TRPO 的结果, 在许多强化学习任务中表现出色, 包括连续动作和离散动作的环境. 此外, 近端优化机制使 PPO 对超参数的选择相对不敏感, 使其成为一个强大且通用的强化学习算法.

广义优势估计 (generalized advantage estimation, GAE)^[26] 也被广泛应用于 PPO 等需要估计优势函数的算法中, 其通过组合优势函数的多步估计值来更好地估计当前优势函数. 在从当前时刻 t 开始的长度为 k 的轨迹上, 优势函数的 k 步折扣估计值公式为

$$A_t^k = \sum_{i=0}^{k-1} \gamma^i (Q(s_{t+i}, a_{t+i}) - V(s_{t+i})) = \sum_{i=0}^{k-1} \gamma^i (r_{t+i} + \gamma V(s_{t+i+1}) - V(s_{t+i})) = \sum_{i=0}^{k-1} \gamma^i \delta_{t+i}.$$

其中, 折扣因子取回报的折扣因子 γ , $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$ 表示 t 时刻状态价值函数的 TD 误差. GAE 通过引入 λ 参数, 以平衡优势函数在长期和短期的估计值. 在无穷轨迹上, 以 λ 为折扣因子, 将所有步长的优势函数进行加权求和, 即

$$A_t^{\text{GAE}} = (1 - \lambda) \sum_{k=1}^{+\infty} \lambda^{k-1} A_t^k = (1 - \lambda) \sum_{k=1}^{+\infty} \lambda^{k-1} \sum_{i=0}^{k-1} \gamma^i \delta_{t+i} = (1 - \lambda) \sum_{i=0}^{+\infty} \gamma^i \delta_{t+i} \sum_{k=i+1}^{+\infty} \lambda^{k-1} = (1 - \lambda) \sum_{i=0}^{+\infty} \gamma^i \delta_{t+i} \frac{\lambda^i}{1 - \lambda} = \sum_{i=0}^{+\infty} (\gamma \lambda)^i \delta_{t+i}.$$

在有限轨迹上,可以由 $A_t^{\text{GAE}} \approx \sum_{i=0}^{T-t-1} (\gamma\lambda)^i \delta_{t+i}$ 近似

计算优势函数,其中, T 为轨迹长度. 使用 GAE 估计优势函数,可以提高算法的性能和训练的稳定性,使得策略的更新更加准确和有效.

使用 GAE 的 PPO 算法步骤如下:

Step1. 初始化策略网络参数 θ .

Step2. 判断迭代次数是否达到上限,若达到,转 Step6.

Step3. 记当前策略参数 $\theta' = \theta$, 使用策略 $\pi(a|s; \theta')$ 收集一组回合轨迹 D .

Step4. 使用 GAE 估计所有轨迹上的优势函数 A_t^n , 其中, n 和 t 分别表示轨迹索引和时间步索引.

Step5. 在所有时间步上使用梯度下降算法更新策略参数, $\theta \leftarrow \arg\max_{\theta} L(\theta, \theta')$; 转 Step2.

Step6. 输出策略参数 θ .

2.3 DDPG

DDPG 算法^[23]是 DQN 在连续动作空间的扩展,其基于确定性策略梯度(deterministic policy gradient, DPG)^[27]和 actor-critic 框架,使用神经网络学习确定性策略函数 $\pi(s; \theta)$, 该函数将状态 s 映射到动作 a , 称为 actor 网络; 同时, 使用神经网络估计动作价值函数 $Q(s, a; \omega)$, 称为 critic 网络. DDPG 的训练方式与 DQN 类似, 不同之处在于 DDPG 使用 actor 网络与环境交互收集样本, 且在更新 critic 网络后需要更新 actor 网络. actor 网络的优化目标函数为所有状态的平均价值

$$J(\theta) = E_s [Q(s, \pi(s; \theta); \omega)].$$

通过链式求导法则可得策略梯度

$$\nabla_{\theta} J(\theta) = E_s [\nabla_{\theta} \pi(s; \theta) \cdot \nabla_{\omega} Q(s, a; \omega)],$$

以更新 actor 网络参数. 为了平衡确定性策略的探索和利用, DDPG 在训练中会对 actor 的输出添加基于 Ornstein-Uhlenbeck 过程的噪声. DDPG 虽然可以应用于连续动作空间, 但由于其渐进性能较差, 在虚拟人运动控制领域的应用较少.

双重延迟深度确定性策略梯度(twin delayed deep deterministic policy gradient, TD3)^[28]是 DDPG 的改进, 它引入双 critic 网络和延迟更新的方法提高算法的稳定性和收敛性. 需要注意的是, 尽管 TD3 在某些情况下可以提供更高的性能和稳定性, 但它仍然受到超参数选择和调整的影响, 不同的环境和任务可能需要不同的超参数设置.

2.4 SAC

SAC^[24]可以看作 DDPG 的一种改进算法, 同

样也适用于解决连续动作空间的强化学习问题, 通过引入熵正则化来提高训练效果和鲁棒性.

SAC 类似于 TD3 使用了截断双 Q 学习, 但是它额外采用了熵作为对策略探索的奖励机制. 其动作价值函数 Q 的 TD 目标和策略函数 π 的优化目标中均引入了熵正则项, 分别记为

$$\text{TD}_{\text{entropy}}(Q) = E[-\alpha \ln \pi(a'|s')],$$

$$L_{\text{entropy}}(\pi) = E[-\alpha \ln \pi(a'|s)].$$

其中, s 和 s' 分别表示当前状态和下一时刻状态; a' 表示状态 s' 时由 π 采样的动作, 并不真正执行, 是熵正则化系数.

SAC 利用熵正则项引入随机性, 促使策略更倾向于探索, 提高了算法的鲁棒性和泛化能力; 另外, 使用随机策略, 不再需要类似于 TD3 中在策略网络输出上添加额外的随机噪声, 以提高策略的探索能力. 这些改进使得 SAC 成为处理连续动作空间的强化学习问题的一种有效算法.

2.5 应用场景

由以上讨论可知, 深度强化学习算法依赖奖励信号优化策略, 当智能体在与环境交互过程中能持续地获取奖励信号时, 这些算法的学习效率将大大提高. 在虚拟人运动控制实际应用中, 如果虚拟人运动模拟易于实现且有充足的运动参考数据可供模仿或追踪, 则可以很容易地设计可持续获取的奖励函数. 在这种情况下, 研究者倾向于选择 PPO 算法, 利用其并行化能力和高学习效率快速、稳定地学习运动控制策略. 然而, 由于有些任务场景没有明确的运动参考或模拟过程成本较高, 或者应用运动控制策略去完成下游任务时只能获取稀疏的奖励信号, 因此需要使用更适合的策略, 以收集更多样本且充分探索动作空间; 如使用带有熵正则项的 SAC 算法促进策略的探索倾向, 并利用其可反复利用旧策略采样样本的特性提高样本利用率. 此外, 虽然虚拟人运动控制大多是连续的动作空间, 但如果研究者将其建模为离散空间、DQN 及其变体, 也是可选的方案.

3 虚拟人运动控制策略学习实际应用设计

利用深度强化学习解决虚拟人运动控制策略学习问题除了选择合适的算法外, 根据具体的任务目标设计强化学习的各个基础元素也至关重要. 如图 4 所示, 本文将基于深度强化学习框架下解决虚拟人运动控制问题分为 4 个组成部分: 状态表

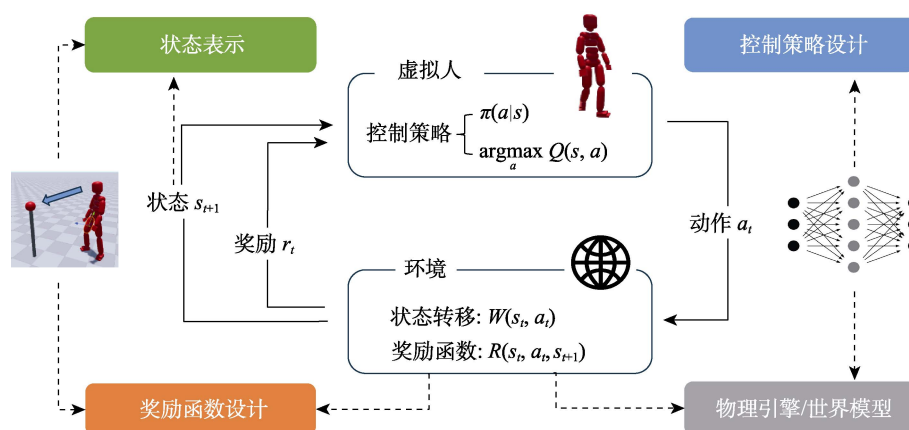


图4 基于深度强化学习框架的虚拟人运动控制策略学习应用设计

示、奖励函数设计、物理引擎/世界模型、控制策略设计;讨论相关工作在实际应用时的具体设计和解决思路。

3.1 状态表示

在基于物理引擎的虚拟人运动控制问题中,强化学习框架中的状态定义为虚拟人当前的状态以及虚拟人可以感知到的任务目标,如需要交互的物体、地形变化、用户输入的约束等。确定虚拟人的状态表示之前,需要在仿真环境中建模虚拟人并确定其关节驱动机制;此外,还需要考虑对任务目标进行合理的表示。合理的状态表示可以帮助控制策略更好地理解环境和任务要求,从而更好地进行运动控制。

3.1.1 虚拟人建模与状态表示

如第1.1.1节所述,由于常规的虚拟人控制问题更多的是考虑复杂的运动控制目标,因此通常将虚拟人建模成简化的树形铰接多刚体系统,并设置合理的物理参数使虚拟人状态近似真实人体;同时,使用基于PD控制的伺服器方法进行驱动。虚拟人状态通常使用冗余表示,即将虚拟人各关节在仿真环境中所能观测到的数据,如位置、速度、旋转角速度等都包含在内,这种冗余的状态表示提供更加丰富的环境感知信息,从而让控制策略更好地理解虚拟人状态的变化。需要注意的是,虚拟人状态信息通常表示在虚拟人根关节的局部坐标系下,可以减少全局信息的影响,使得控制策略更加专注于局部动作的控制,提高控制策略的泛化能力。

虚拟人运动技能是依赖双腿发力来实现的,如行走、跑或者跳等。为此,一些方法使用简化的上半身建模方法^[29-30],如质量相同的立方体近似替代上半身的整体运动。这种简化的方法一方面

可以减少虚拟人的状态和动作空间维度,降低策略的学习难度;另一方面,可以使得控制策略更加专注于腿部运动的控制,增加控制策略的鲁棒性。

随着简化的虚拟人体运动控制策略学习方法日渐成熟,也有一些方法开始考虑更加精细化的人体建模或者驱动方法。手掌作为虚拟人与外界交互的重要窗口,是首先考虑的目标之一。通过对手掌关节的精细化建模,可以模拟真实人体手掌的发力模式,完成更加精细的控制任务,如运球技巧^[31]。另一方面,与PD控制直接计算关节力矩相比,通过建模肌肉和肌腱的生理学模型来模拟人体的发力模式,可以更好地考虑肌肉的生物力学特性和力量输出的非线性关系,生成更加自然、合理的虚拟人运动,使其更接近真实人体的运动方式。这种肌肉驱动方法通过使用神经网络建模关节力矩和肌肉激活信号之间的关系^[32-33]来实现。在肌肉驱动的基础上, Park 等^[34]进一步研究生理构造参数与运动技能之间的关系,拓展了控制策略对于生理结构变化的适应性。

真实人体虽然有相同的骨架结构,但骨长比例却随着个体而有所差异。在仿真环境中,可以通过调整虚拟人的几何参数(如身高、腿长等)和物理参数(如质量、惯性等),模拟不同个体的身体比例差异,实现对虚拟人的多样化建模,以满足用户的个性化需求。具有可变人体参数的虚拟人控制策略学习中,将人体参数加入状态表示中,以便控制策略能够感知和适应不同参数的虚拟人^[35-36];但会导致状态空间维度进一步扩大,需要复杂的策略学习方案才能完成对参数空间的探索。

3.1.2 任务目标

控制策略的学习目标包括2个方面:一是让虚拟人掌握尽可能多的、鲁棒的运动技能;二是让虚

拟人响应各种外部环境变化,如用户输入的控制信号、地形变化、需要交互的物体或其他虚拟人的状态变化等.从传统的用户键盘输入控制虚拟人的运动方向与速度^[5],到更加复杂的多虚拟人交互任务^[37]等,虚拟人运动控制问题中的外部环境变化多种多样,为此需要对其进行参数化,并与虚拟人本身的状态一起作为强化学习的状态输入来学习控制策略.为了方便理解,本节将用任务目标指代环境变化,并将任务目标分类,如图 5 所示,具体讨论每类任务目标的参数化方法.

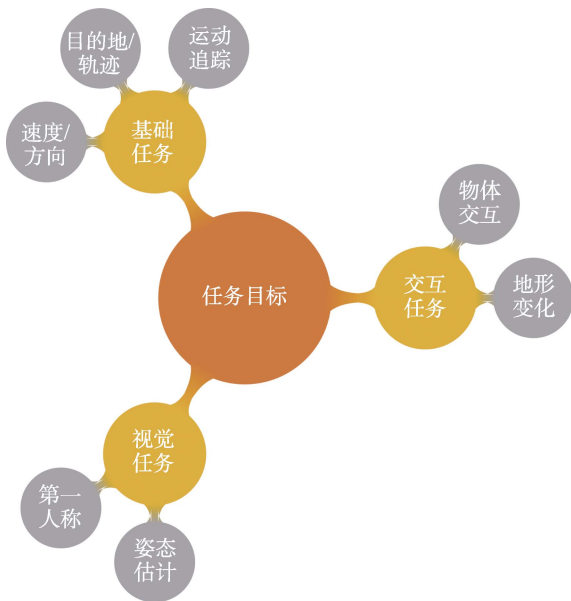


图 5 任务目标分类

(1) 基础任务. 最基础的任务目标来源于游戏控制,即用键盘控制虚拟人前进的方向和速度等.这种目标的参数空间较小,相对容易学习,在深度强化学习算法之前已经有了较稳定的解决方案^[3,38].在深度强化学习算法的帮助下,通过高参数量的控制策略网络,可以更准确和鲁棒地完成此任务.进阶的任务包括轨迹控制和目的地控制^[39-41],这 2 类任务比较类似,轨迹控制的目标通常形式化为连续的目的地控制点^[41].这些基础的任务目标很容易用数值表示,需要注意的是,全局坐标系表示下的任务目标需要转化为相对于虚拟人根关节位置的局部任务目标差距,以便控制策略更好地理解人体运动,相应的奖励函数设计也是鼓励虚拟人消除这种相对的差距.

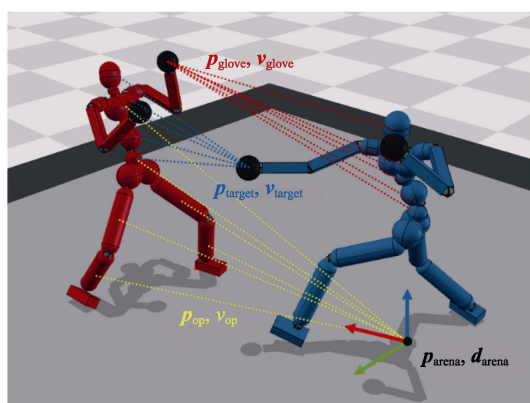
随着运动捕获数据的逐渐丰富,运动追踪任务也成为研究者关心的问题之一.运动追踪的目标是在仿真环境中控制虚拟人准确、真实地重现给定的真实人体的运动捕获序列,包括时序密集的

全身关节的位置和朝向等信息.运动追踪是从运动数据中学习运动技能的基础,通过模仿运动捕获数据,控制策略建立虚拟人状态与动作之间的映射关系.在运动追踪任务中,目标通常被形式化为下一个时间步的、从运动捕获数据中计算的虚拟人状态.由于虚拟人的状态空间维度较大,策略学习相对困难,传统方法通过在目标状态附近采样来寻找最优的控制输入,可以实现对一些相对简单的运动片段的追踪^[3].在深度强化学习框架下,DeepMimic^[5]通过手工设计的模仿运动捕获数据的奖励函数,并将运动的相位作为策略输入来约束动作空间,基于 PPO 算法实现对单一运动短片段的准确鲁棒的追踪.在此基础上,后续的工作或是将整个目标状态加入策略网络输入^[42],或是结合生成式方法^[43],实现了更长、更复杂序列的运动追踪;还有的工作尝试解决稀疏的追踪目标,如 VR 输入的头盔和手柄相关信息^[44-45].

(2) 交互任务. 基于物理仿真的虚拟人运动控制的优势之一是具有较强的交互性.通过利用物理引擎模拟真实世界,可以更准确、便捷地建模虚拟人与周围环境中的人或者物体的交互.对于静止的物体,交互的任务目标包括搬运^[46]、击打^[40,47]等;与运动物体的交互任务则包括各种球类运动,如篮球^[31]、足球^[40,48]和网球^[49]等.在这类交互任务中,目标输入被形式化为交互的目标物体相对于需要交互的虚拟人肢体部位的相对状态,包括相对位置、朝向、速度和角速度等,这种形式化的目标输入与真实人体在面对交互物体时的行为相符.

虽然基于深度强化学习的控制策略较为鲁棒,能够应对一定的地形变化,但是当地形较为复杂时,需要将地形因素也加入目标输入中.地形因素通常形式化为虚拟人周围一定范围内采样的地形高度信息^[29-30,41,50],这些信息可以提供地形起伏、坡度和障碍物位置等,通过将地形的高度信息加入目标输入,控制策略可以更好地感知和理解周围地形的特征,从而更好地做出决策.

人与人之间的交互则相对复杂,对于特定问题需要单独设计目标输入.Won 等^[37]提出一种双人拳击运动的控制策略,可以控制虚拟人用拳头去击打对抗虚拟人的关键身体部位并躲避对抗虚拟人的进攻.在此任务背景下,单个虚拟人控制策略的目标输入需要包括:对抗虚拟人关键关节相对于控制虚拟人平面坐标和手部的相对位置和速度,以及控制虚拟人关键关节相对于对抗虚拟人手部的相对位置和速度,如图 6 所示.对于更复杂

图6 拳击运动的控制策略目标输入^[37]

的多人交互运动,则需要考虑使用多智能体深度强化学习算法^[46]。

(3) 视觉任务. 从视频中重建三维人体运动是近年来的研究热点,然而由二维的图像信息推断三维的人体姿态具有歧义性,是一个病态问题. 基于运动学的估计方法在人体与其他物体(如地面)发生接触时,估计结果中往往会产生不合理的脚步滑动,同时难以应对人体被遮挡的情况. 通过学习虚拟人运动控制策略,将重建问题转化为运动追踪问题,是解决上述挑战的途径之一. 运动追踪的效果受到输入的运动参考影响,主流的方法通常是先基于成熟的姿态估计算法^[51]从视频中提取出二维或者三维的运动参考序列,再对其进行优化修正,优化的手段包括时序平滑^[52]、利用环境信息修正^[53]和全局轨迹优化^[49]. 获取运动参考序列之后,该问题就简化为运动追踪问题,任务目标的输入可以是运动的相位或者是具体的运动参考序列. Yu 等^[54]则在姿态估计的基础上,先基于轨迹优化学习基础的开环控制策略,再用强化学习提升策略的鲁棒性和泛化能力. 然而在控制虚拟人去追踪视频中的高敏捷度的运动时,虚拟角色与真实人体在全局位置和局部动作上可能无法完美匹配. 为此, Yuan 等^[55]通过向人体施加虚拟的残差力(residual force contrl, RFC)来解决这种动力学不匹配的难题,并使用追踪奖励以及 RFC 正则化项来构成奖励函数优化修正后的策略. 在 RFC 的基础上, Yuan 等^[56]提出的 SimPoE 进一步学习 PD 控制器的参数,实现了根据角色的状态动态地调整 PD 控制器参数,以提高追踪精度. 如图 7 所示,在之前工作的基础上,目前较新的工作已经能较准确地从视频中重建出三维人体姿态^[57-58].

第一人称视角下的人体姿态估计问题是 VR/AR 领域中的重要研究方向之一. 与第三人称视角的

图7 基于运动追踪的姿态估计^[57-58]

问题不同,第一人称视角下无法直接利用姿态估计方法来估计全身运动学姿态,故无法直接学习虚拟人控制策略进行运动追踪. Yuan 等^[59]通过从穿戴式相机提供的视觉输入中,基于光流提取视觉上下文,并将其作为控制策略的输入,该方法的奖励函数仍然是基于运动参考数据的模仿性质函数. 而 Luo 等^[60]则利用监督学习预训练方法,基于第一人称视角的视觉输入生成全身运动参考,再利用运动追踪方法学习虚拟人运动控制策略.

3.2 奖励函数设计

应用深度强化学习算法学习虚拟人运动控制策略时,奖励函数的设计至关重要,该函数定义了控制策略在特定状态下选择的动作的好坏程度. 通过奖励函数,控制策略可以根据环境的反馈调整和优化行为决策. 奖励函数的设计包括与任务目标相关的稀疏或稠密的奖励项、鼓励运动自然的能量惩罚项等. 在运动捕获数据充足的情况下,模仿性质的奖励函数也能有效地提升运动的自然程度;同时,中止条件的设计也很重要,它通过提供非正值的奖励来约束动作的搜索空间,可以有效地提升学习效率. 奖励项通过负指数函数映射到固定范围内,并通过加权求和或者相乘的方式进行组合. 本节将详细讨论运动捕获数据充足和缺失时相关奖励设计方法.

3.2.1 运动数据充足时的奖励函数

运动数据充足时,模仿性质的奖励函数通常占据主导地位,该函数是时序稠密的,需要在每个时间步都进行计算,包括局部姿态约束、全局位置约束和物理属性约束^[5]. 局部姿态约束指虚拟人各关节的局部旋转和角速度与运动捕获数据对应关节之间的误差;全局位置约束表示为虚拟人末端关节位置,通常是手和脚与运动数据中对应关节位置的误差;物理属性约束表示为虚拟人的质心与运动捕获数据中对应的质心的误差,其中,质心近似表达人体整体的运动状态,可以从整体上为模仿学习提供参考. 当控制策略与生成式方法结合时,模仿的对象变为生成的运动数据^[43,61]. 同时,当虚拟人状态与模仿目标差距超过一定阈值

或者虚拟人在仿真环境中摔倒无法继续进行控制时,会提前中止当前仿真回合^[5],并添加惩罚性质的负值奖励^[61]。如果没有提前中止,训练的早期阶段会采集到大量的无效样本,影响控制策略的学习效率。

手工设计模仿性质的奖励函数需要对运动数据有着较深的理解,且不同部分之间的权重也是需要考虑的因素,比较费时费力。与之相反,生成式对抗模仿学习(generative adversarial imitation learning, GAIL)^[62]则省去这一步骤,通过学习一个判别器来判断策略给出的状态-动作样本对与专家策略给出的样本对之间的真伪,并将其形式化为奖励函数来引导控制策略模仿专家样本,实现行为克隆。在 GAIL 框架的基础上,AMP^[40]将样本观测建模为虚拟人相邻 2 个时间步的状态,可以直接将运动捕获数据作为专家样本,通过对抗性的奖励函数引导控制策略学习运动捕获数据隐含的动作空间;同时,AMP 还使用梯度惩罚^[63]来保证判别器学习的稳定性。AMP 可以在给定少量运动捕获数据的情况下学习到其中包含的运动技能,并可以结合任务相关的奖励函数完成类型丰富的目标任务。凭借其易用性和稳定性,AMP 成为后续复杂运动策略研究中学习底层运动控制策略的基础^[6,47,57,64-65]。

3.2.2 运动数据缺失时的奖励函数

与有运动捕获数据作为指导的模仿学习相对应的是无参考数据的虚拟人运动控制策略学习。此类方法的奖励函数设计中,更依赖对于人类运动的理解和与任务目标相关的奖励函数设计。Duan 等^[66]设计奖励函数来鼓励虚拟人实现位移而非原地静止,同时,惩罚脚步与地面的接触,并防止虚拟人偏离前进的方向,虚拟人能够实现在两点之间移动而不摔倒,但生成的动作生涩且不像真实人体。为了解决无运动参考下生成的人体运动不自然的情况,Yu 等^[67]根据观测发现,对称性是健康步态的一个重要的特征,并设计相应的对称性损失项与 PPO 常规损失函数一同优化;同时,在 Duan 等^[66]工作的基础上,在奖励函数中额外地引入平衡奖励以及过大扭矩的惩罚项,以尽可能小的能量驱动虚拟人的位移。Abdolhosseini 等^[68]的工作进一步证明,在策略学习过程中,引入对称性约束能够产生更高质量的运动。

除了在常规平整的地形上学习虚拟人的行走策略,AllSteps^[30]实现了一种在无运动捕获数据参考的情况下虚拟人的跨阶梯位移控制策略,在保证生成的运动与真实人体相似的奖励函数^[66-67]的

基础上,同时将虚拟人脚部关节与目标阶梯的接触作为奖励函数的一部分,并鼓励人体重心与阶梯目标靠近,以完成跨越阶梯的任务。除了位移运动,Tao 等^[69]提出一种由任意跌倒状态起身的运动控制策略,将起立动作的学习分为 3 个阶段,且为每个阶段设计不同的奖励函数以达到阶段目标:第 1 阶段的奖励函数鼓励智能体提高头部的位置、保持躯干垂直直立、惩罚位移以及限制双脚之间的距离;第 2 阶段通过课程学习的方式逐步地减少关节的力矩,同时设立奖励函数模仿第 1 阶段的运动;第 3 阶段学习更慢的起立速度来模仿人类跌倒后缓慢恢复站立的动作,同时通过奖励函数鼓励人体最终保持预设的站立姿态。

3.3 物理引擎与世界模型

在深度强化学习中,环境是智能体所处的外部世界,包含智能体所能感知和交互的一切;状态转移则描述智能体在环境中从一个状态转移到另一个状态的过程。在虚拟人运动控制问题中,环境和状态转移都基于物理引擎实现。物理引擎以物理规律为基础,模拟真实世界中的物理行为,它可以模拟重力、碰撞、摩擦等各种物理效应,使虚拟环境或虚拟角色的行为更加逼真和可信;其提供一组函数和工具,用于创建和管理物理对象,并基于约束计算物体的运动轨迹等。物理引擎是基于深度强化学习的虚拟人运动控制的基础,无论是无模型的方法还是基于模型的方法,都依赖物理引擎对虚拟人和环境进行建模,并收集交互样本来学习相关策略。

目前,常用的开源物理仿真引擎有 Bullet, MuJoCo^[9], DART^[10], PhysX 等,它们的目标都是模拟真实世界的各种物理规律,但可能在底层的近似计算中有所不同。这些物理引擎没有明显的优劣,熟练掌握其中一种即可。MuJoCo 是深度强化学习中使用较多的物理引擎之一,被广泛地用于学习各种虚拟人控制策略^[70-73];DART 提供高性能的动力学仿真和控制功能,适用于机器人仿真等场景,同时也能提供肌肉驱动的仿真控制功能;Bullet 提供丰富的物理效果模拟功能,集成了丰富的 API。限制强化学习效率的一个主要因素就是采样效率低下,因为物理引擎仿真的计算通常只能在 CPU 上进行。近年来,基于 PhysX 引擎的 Isaac Gym^[74]利用 GPU 的计算能力同时进行大量的仿真计算,大大提升了采样效率,可以在较短的时间内进行大规模的强化学习实验。对于数据需求量较大的任务,研究者更倾向于使用 Isaac Gym 来节省

训练时间^[57,64-65]。

物理引擎虽然能够提供精细的运动仿真结果,但是其计算过程是不可微分的。近年来,基于监督学习的生成式模型,如变分自编码器(variational autoencoder, VAE)^[75]在运动学方法中取得了出色的效果^[76]。物理仿真不可微分的特性,已成为利用生成式模型进一步提升基于强化学习的控制策略性能的阻碍。若要回避这种阻碍方式,常用的方法将生成式模型的目标函数形式化为强化学习中的一个特别奖励项^[47];然而这种方法需要以非监督的方式同时训练2个耦合的系统,训练的过程和目标函数需要仔细地设计^[77]。另一种方法是利用基于模型的强化学习的思想,通过学习一个针对当前任务的世界模型神经网络,以拟合当前虚拟人的状态转移方程。世界模型提供了可微的物理仿真计算,使得端到端的结合生成式模型的控制策略学习成为可能。通常,训练世界模型的样本是基于控制策略进行虚拟人的物理运动仿真得到的。世界模型的训练可以与控制策略的学习交替进行^[58,78-80],也可以通过专家控制策略进行采样和预训练,并在后续控制策略学习的过程中固定其参数^[41]。基于世界模型学习到的控制策略有效地结合了生成式模型和物理仿真方法的优势,是未来虚拟人控制策略学习的重要发展方向之一。

3.4 控制策略设计

目前,主流的深度强化学习算法为 actor-critic 结构,需要同时学习策略网络(actor)和价值网络(critic)。策略网络根据当前状态选择动作,而价值网络则评估当前状态的价值,这2个网络相互协作,共同优化智能体的决策策略。在虚拟人运动控制问题中,鉴于运动捕获数据集的富足,主流的方法是利用 PPO 算法学习数据集中隐含的动作分布。当任务目标比较简单且数据集运动类型单一时,策略网络可以用简单的多层感知机(multi-layer perceptron, MLP)表示^[5],但随着运动种类的丰富以及任务目标的复杂化,需要更加复杂的网络架构和训练方法。

3.4.1 与生成式模型结合

为了提升控制策略对运动技能类型的包容性,早期的工作直接利用运动学的生成方法。在运动匹配^[61]或者基于循环神经网络的生成式方法^[43]中,先根据任务目标生成运动预测,再训练控制策略追踪预测,进而实现任务导向的长序列运动控制。为了保证追踪的稳定性,一段时间内的运动预测

也需要作为状态输入。在 DeepMimic^[5]的基础上,AMP^[40]结合生成式对抗学习^[81]的思想,通过额外训练一个判别器,并使用对抗损失函数直接学习给定数据片段中的动作分布,省去了复杂的手工奖励函数设计步骤。

另一种结合生成式模型的方法是将生成式模型直接嵌入控制策略网络中。Xu 等^[82]使用循环神经网络作为控制策略,通过隐藏变量传递输入状态的时序信息,使得生成的动作更加平滑。潜在空间模型在成功建模多种运动类型的同时,还能结合高层策略在潜在空间中进行采样,完成多种具有挑战性的运动学下游任务^[76]。潜在空间模型需要通过重建误差进行监督学习,然而由于物理仿真的不可微性,因此当结合基于强化学习的控制策略时,潜在空间的目标损失函数通常被转化为策略学习相关的奖励函数项,以间接地建模运动技能连续潜在空间^[6,47]。Zhu 等^[77]利用矢量量化编码^[83]构建一个运动片段相关的离散隐空间,为了解决使用该空间信息存在的数据多样性不足问题,通过先验转移并鼓励转移后的分布均匀覆盖所有可能的运动片段。PULSE(physics-based universal motion latent space)^[65]则利用 PHC(perpetual humanoid control)^[57]学习得到的专家策略,先大量获取状态-动作对的真值序列,再直接基于监督学习建模控制策略中的潜在空间。学习世界模型神经网络^[79]是另一条实现控制策略中潜在空间监督学习的途径。基于在仿真环境下采样的虚拟人状态-动作样本,世界模型可以模拟物理引擎的仿真计算,实现可微的状态转移过程。PhysicsVAE^[41]首先利用模仿学习的方法^[5]学习专家策略,生成专家轨迹并预训练世界模型,再基于世界模型学习控制策略中的潜在空间表示。ControlVAE^[78]则是将潜在空间与世界模型的学习结合起来,在每个训练步交替进行;同时采用了一个基于虚拟人当前状态的运动先验^[84]替代传统 VAE 中的零均值高斯分布,这种先验可以根据当前状态学习不同的分布,更好地建模多运动类型的隐空间。在 ControlVAE 学习框架基础上,MoConVQ^[58]利用矢量量化编码的方法将运动变化映射到离散的潜在空间,进一步提升了潜在空间对运动类型的容量,并且这种离散的表示可以结合生成式预训练转换器(generative pre-trained transformer, GPT)模型完成更高阶的跨模态生成任务^[85]。

3.4.2 策略组合

与在控制策略中通过潜在空间隐式建模多种

的运动类型时,学习的运动控制策略只能控制虚拟人以走路和跑步的方式完成对应的任务目标;还提供了一种从少量运动捕获数据中,稳定学习底层运动控制策略的技术路线.在此基础上,可以对策略网络模块进行拓展,通过从运动捕获数据集中学习更多的运动技能完成更复杂的任务目标.在利用此框架学习底层控制策略的基础上,PHC基于混合专家模型学习高层策略,以调度各底层策略,最终实现对大范围运动类型的鲁棒追踪.另一类方法通过拓展控制策略提升对多样化运动类型的学习能力,这些方法在学习控制策略的同时建模策略的潜在空间^[6,47].由于物理仿真的不可微性,潜在空间的重建误差被转换成相关奖励函数来实现.因此建模潜在空间后,可以通过学习高层策略调制潜在空间,完成特定的任务目标;同时,也可以利用增量学习^[95]提升控制策略的性能.Xu等^[82]将此框架学习到的控制策略作为元策略,通过学习元策略输出的偏置策略完成新的任务目标.

3.5.2 实际应用举例与分析

在解决基于视频的姿势估计和基于VR传感器的全身姿势重建等实际应用问题时,基于虚拟人运动控制策略学习的解决思路是将其转化为运动追踪问题,获取对应的物理真实且准确的虚拟人运动序列^[44-45,57].对于视频输入,利用先进的姿势估计方法^[98]获取运动学姿势序列,并将其作为运动追踪的目标;而对于VR传感器输入,则是将传感器的位置、朝向和速度等信息作为追踪目标进行控制策略的求解.运动追踪问题是虚拟人运动控制的经典问题,旨在学习控制策略追踪特定的运动捕获数据片段,生成物理真实的虚拟人运动序列.在深度强化学习框架下,当学习控制策略追踪仅包含单一运动技能的较短运动片段时,可以将运动相位作为状态输入约束动作空间^[5].当要实现对包含多种运动技能的长运动片段的追踪时,需要将未来一小段时间的目标姿态作为任务目标输入,保证追踪的稳定性^[43,61].设计奖励函数时,如果追踪的运动序列复杂多变,可以使用手工设计的模仿性质的奖励函数^[43];如果追踪目标只包含少量的运动片段,设计对抗损失是更省时高效的选择^[40].在环境方面,可以使用PPO算法进行无模型的控制策略学习,也可以利用采样数据监督式地训练世界模型^[79],进行基于模型的端到端训练.对于数据集较大的情况,单一控制策略难以

处理数据集中包含的所有运动类型,可以在学习多个底层控制策略的基础上,利用高层策略进行调度,实现对多种运动类型的追踪^[42].

对于复杂的任务目标,运动捕获数据往往难以获取,在这种缺乏运动捕获数据的情况下,控制策略需要从无到有地学习运动技能完成任务目标,因此需要更加仔细地设计奖励函数,引导虚拟人对运动空间的探索;同时,也需要分解任务目标,从易到难渐进式地学习控制策略.为了解决在无运动捕获数据参考的跨阶梯位移运动问题,Xie等^[30]提出一种基于课程式学习的控制策略学习方法AllSteps,将控制策略的学习由易到难分成4个课程阶段,并在每个阶段通过动态地调整需要跨越的台阶地形难度,对任务目标的参数空间进行渐进式的探索.因为任务本身的状态空间不是特别复杂,只使用5层的MLP作为控制策略网络,并使用PPO算法学习控制策略,所以在奖励函数设计上,除了为每个脚步的目标位置设定稀疏奖励函数外,还设计了与目标距离相关的稠密奖励函数鼓励虚拟人去接近目标;同时,设计了惩罚性质的奖励函数,如能量惩罚项、关节约束惩罚项和速度惩罚项,以提升运动的自然程度.

3.6 数据集

本节介绍针对运动控制策略学习中常用的运动捕获数据集,以及针对特定姿势估计任务常用的数据集.

3.6.1 常用运动捕获数据集

当前,已有一系列虚拟人运动控制策略的研究工作关注对特定运动技能的学习^[5,31,45,49,54],它们使用独立采集或从视频中重建获得的运动数据作为数据集.如何学习通用的运动控制策略一直是计算机图形学和机器人领域面临的挑战.随着动作捕捉技术的发展,多个大规模人体运动数据集的开放为通用策略的学习提供了基础性的支持.下面详细介绍虚拟人运动控制策略学习相关的数据集.

CMU-Mocap数据集^①.由卡内基梅隆大学图形实验室提出,是利用光学标记运动捕获的大规模三维人体运动数据集,内容涵盖移动运动、体育活动、人与人交互、人与物体交互、情景交流、测试数据等6大类别.该数据集包含111个不同尺寸的人体骨架,共计2605段运动(120帧/s,543.49min);因其三维运动具有高精度、高质量的特征,被广泛应用于三维人体运动建模、控制、预

① <http://mocap.cs.cmu.edu>

测、生成等任务。

AMASS 数据集^[99]。马普所智能系统研究团队将包括 CMU-Mocap 数据集在内的多个基于光学标记捕捉的数据集进行整合,并以 SMPL 参数模型^[100]对人体骨架结构、形态表示进行统一,形成了 AMASS 数据集。截至目前,AMASS 整合了 24 个数据集,涵盖 500 个主题、17 916 段运动,时间超过 62 h,并且它仍处于更新维护状态。AMASS 的统一表示特点,使其很容易用于动画、可视化和生成深度学习的训练数据。

LAFANI 数据集^[101]。由育碧公司于 2020 年提出的三维人体运动数据集。该数据集以 30 帧/s 的帧率捕获,共有 5 个角色骨架、77 段运动,共计 496 672 帧(约 276 min);其主题主要针对移动运动,包括平地和非平地的移动,同时还包含少量搏击、瞄准以及跳舞等艺术动作。

3.6.2 姿态估计数据集

在姿态估计任务中,运动控制策略根据视觉特征操控虚拟人重现运动序列。姿态估计任务是当前计算机视觉领域备受关注的研究方向,不断有包含运动视频和运动序列数据,且场景、视角与内容各异的数据集发布。下面介绍目前应用最普遍的 2 大代表性数据集。

Human3.6M 数据集^[102]。同时包含三维人体运动数据以及 4 个视角下高分辨率的视频数据。该数据集包含 360 万帧人体姿态及其对应的图像,由 11 名专业的动作捕捉演员(6 男 5 女)完成,包含 7 个场景(讨论、吸烟、拍照、打电话等),捕获帧率为 50 帧/s。数据捕获中,动作捕捉演员穿着各自的日常服装,并在身体上直接粘贴反光材质以尽可能保持真实感。该数据集在基于二维视频的动力学建模中广泛使用。

3DPW 数据集^[103]。不同于 Human3.6M 数据集主要集中在室内环境,3DPW 使用单目摄像机与惯性测量单元(inertia measurement unit, IMU)的复合传感系统进行采集,旨在同时涵盖室内和室外环境,且重点聚焦于室外复杂环境中人的自然动作。该数据集以 30 帧/s 的帧率采集,共包含 51 000 个运动帧和 60 段视频序列,每帧都包含对应的单目摄像机参数。

3.7 评估标准

从训练与推断时间、定性标准和定量标准 3 个方面,讨论虚拟人运动控制策略的评估标准,并介绍相关工作在标准上的进展情况。

3.7.1 训练与推断时间

在基于深度强化学习的虚拟人运动控制策略学习中,虚拟人每次更新策略参数时都需要与物理仿真环境进行交互,以收集足够的训练样本,而物理仿真涉及复杂费时的约束求解过程,大大限制了控制策略的学习效率。早期的工作受限于 CPU 的计算性能,即使是简单的运动片段,都需要花费数个小时进行学习^[5]。随着训练数据增多,耗费的时间会增至数十小时^[43]甚至上百小时^[42]。最近,Isaac Gym^[74]的出现极大地缓解了这一问题,其利用 GPU 的并行计算能力在大量仿真环境中同时进行采样,大大提升了效率,在相同的任务上,通常可将训练速度提升 2 个数量级^[44],在短时间内即可完成基于大规模数据的运动控制策略学习^[6,57,65]。

由于运动控制策略网络本身的结构并不复杂,且大多数运动控制策略都是以用户交互控制的任务目标进行学习的,因此绝大多数的运动控制策略都可以做到实时交互控制,生成用户需求的运动序列。

3.7.2 定性标准

由于虚拟人运动控制策略的相关研究工作没有固定的任务目标,因此研究者更倾向于从视觉上的运动生成效果来评价工作。参考 Geijtenbeek 等^[104]的工作,本文将定性标准总结为 3 点。

(1) 运动技能。评估控制策略的一个重要指标是判断其是否具备足够丰富的运动技能。控制策略掌握的运动技能越丰富,其能完成的任务目标也会随之复杂。早期的方法受限于稀少的运动捕获数据,只能从运动片段中学习单一的运动技能^[5]。随着运动捕获数据集的丰富,越来越多的工作从海量数据中学习尽可能多的运动技能。从多个运动片段^[40,43],到单一运动数据集^[6,42,78],再到多个数据集的集合^[57-58,65],已有方法已经能成功地从数十小时的运动捕获数据中学习种类多样、风格各异的运动技能。

(2) 鲁棒性。在运动控制过程中维持平衡也是评估控制策略的一个重要指标。在结果展示中,研究者通过给虚拟人添加额外扰动,如突然的推力,考察控制策略的鲁棒性。目前,基于深度强化学习的控制策略都有较强的鲁棒性,能应对一定程度的外力扰动。如果虚拟人意外摔倒,有的工作也会额外学习起身策略^[43],保证控制策略能够继续进行之前的任务。

(3) 自然程度。虚拟人体运动控制的要求是使得生成的动作尽可能真实。现阶段,主流的从运动捕获数据中学习运动控制策略的方法使用模仿性

质的奖励函数^[5]或者对抗性质的奖励函数^[40],都能够较好地重现运动数据集中的运动序列;对于不使用运动捕获数据的方法,会额外设计约束动作自然程度的能量^[30]或者对称性奖励函数^[67],提升运动质量。

3.7.3 定量标准

由于虚拟人运动控制策略没有固定的任务目标,研究者更倾向于定义新的研究问题并生成视觉自然的虚拟人运动序列,因此相关的定量标准较少。近年来,随着相关研究方法的进步和数据集的丰富,一些工作如 PHC^[57], UHC^[60], PULSE^[65]参考计算机视觉领域的定量标准,在大量数据集上验证其方法在追踪运动参考时的准确性和鲁棒性:

(1) 平均关节位置误差(mean per joint position

error, MPJPE). 计算生成的运动与参考运动之间的关节位置的平均误差来衡量相似程度,分为在根关节坐标系下的局部 MPJPE 和世界坐标系下的全局 MPJPE(G-MPJPE)。

(2) 速度误差(velocity error, Vel). 计算生成运动与参考运动关节速度的平均误差。

(3) 加速度误差(acceleration error, Acc). 计算生成运动与参考运动关节加速度的平均误差。

(4) 追踪成功率(success rate, Succ). 在任何时间点,若生成运动与参考运动的关节位置平均距离小于某一阈值,则视为追踪成功。

在同等条件下,PHC^[57],UHC^[60]和PULSE^[65]在 AMASS 数据集上的运动追踪量化性能对比如表 1 所示。可以看出它们在定性指标上的进步情况。

表 1 在 AMASS 数据集上的运动追踪量化性能对比

方法	Succ/% ↑	G-MPJPE/mm ↓	MPJPE/mm ↓	Acc/(mm·帧 ⁻²) ↓	Vel/(mm·帧 ⁻¹) ↓
UHC ^[60]	62.6	58.2	98.1	22.8	21.9
PHC ^[57]	97.1	53.1	39.5	7.5	10.4
PULSE ^[65]	99.2	36.1	24.1	2.7	3.9

注: 粗体表示最优值, ↑ 表示该项数值越大越好, ↓ 表示该项数值越小越好。

4 结 语

近年来,深度强化学习算法已经逐步发展成熟,并在解决虚拟人运动控制问题上展现了巨大优势和潜力。基于深度神经网络表示的运动控制策略能够有效地处理虚拟人复杂的高维状态空间和动作空间,并通过学习复杂的非线性关系优化决策过程,控制虚拟人完成多样化的任务目标。

4.1 研究现状与挑战

在虚拟人运动控制策略学习方法的相关工作中,研究者通常根据具体的任务目标,设计合理的虚拟人状态表示和奖励函数,利用物理引擎,并应用合适的深度强化学习算法学习控制策略。目前,从基础的速度控制到与视觉输入结合的任务目标,该领域的工作呈现复杂化、多样化的趋势。本文将重点工作按照出版年标注在图 9 中;并将现有相

关工作按照第 3 节提到的 4 个维度进行总结,如表 2 所示,其中,CL 表示课程学习,IL 表示增量学习,TL 表示迁移学习,为后续的研究提供参考。

随着运动捕获数据集的发展,主流的方法更倾向于从运动捕获数据中利用模仿性质的奖励函数来学习运动技能。如何利用丰富的运动捕获数据提升控制策略的深度与广度是未来的主要研究方向。在广度方面,因为不同运动技能的动作空间分布是不同的,如何合理利用大规模数据集让控制策略掌握尽可能多的运动技能,并使用学习到的技能去完成更加复杂的任务目标是当前研究的挑战之一。在深度方面,探索单一运动类型控制策略的个性化参数空间,让控制策略能够建模尽可能多的可控参数,例如虚拟人的身体参数或物理仿真环境参数,以实现个性化的运动控制也是一个亟待解决的问题。

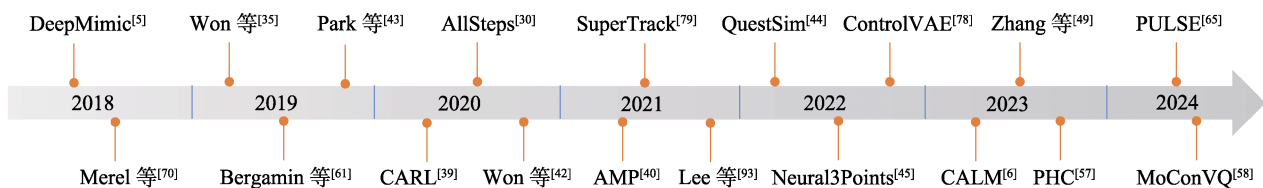


图 9 虚拟人控制策略学习研究时间线

表 2 虚拟人控制策略学习实际应用设计总结

相关工作	状态表示										奖励	环境	控制策略设计			
	虚拟人建模			基础任务		交互任务		视觉任务		深度强化学习算法			复杂策略			
	全身模型	简化人体	精细建模	速度/方向	目的地/轨迹	运动追踪	物体交互	地形变化	姿态估计		第一人称	数据依赖	物理引擎	世界模型	生成式模型	策略组合
Liu 等 ^[87]	√			√		√	√	√			√	ODE	DQN		√	
Heess 等 ^[86]	√	√		√				√				MuJoCo	PPO		√	
Merel 等 ^[70]	√			√		√		√			√	MuJoCo	TRPO			
Liu 等 ^[31]	√		√				√				√	ODE	DDPG			
DeepMimic ^[5]	√			√	√	√	√				√	Bullet	PPO		√	
SFV ^[52]	√								√		√	Bullet	PPO			
Yu 等 ^[54]	√					√					√	DART	PPO			CL
Yu 等 ^[67]	√	√		√								DART	PPO			
Abdolhosseini 等 ^[68]	√	√		√	√	√		√				MuJoCo	PPO			
Bergamin 等 ^[61]	√			√	√		√				√	Bullet	PPO	√		
Yuan 等 ^[59]	√									√	√	MuJoCo	PPO			
MCP ^[89]	√	√			√		√				√	Bullet	PPO		√	
Won 等 ^[35]	√	√				√					√	DART	PPO			
Park 等 ^[43]	√							√			√	DART	PPO	√		
Won 等 ^[42]	√					√					√	Bullet	PPO	√	√	
CARL ^[39]						√					√	Bullet	PPO	√	√	
RFC ^[55]	√									√	√	MuJoCo	PPO			
AllSteps ^[30]	√	√						√			√	MuJoCo	PPO			CL
Lee 等 ^[93]				√	√			√			√	DART	PPO			CL
Won 等 ^[37]	√					√	√				√	Bullet	PPO			TL
AMP ^[40]	√			√	√	√	√	√			√	Isaac Gym	PPO	√	√	
Xu 等 ^[88]	√			√		√		√			√	Bullet	PPO	√	√	
SimPoE ^[56]	√								√		√	MuJoCo	PPO			
UHC ^[60]	√									√	√	MuJoCo	PPO			
Luo 等 ^[53]	√								√		√	MuJoCo	PPO			
ASE ^[47]	√			√	√	√					√	Isaac Gym	PPO	√		
PhysicsVAE ^[41]	√			√	√		√	√			√	Bullet	√	PPO	√	
QuestSim ^[44]	√					√					√	Isaac Gym	PPO			
Neural3Points ^[45]	√					√					√	ODE	PPO	√		
Tao 等 ^[69]	√											MuJoCo	SAC			CL
Park 等 ^[34]			√			√					√	DART	PPO			
ControlVAE ^[78]	√			√	√		√				√	ODE	√		√	√
CALM ^[6]	√					√					√	Isaac Gym	PPO	√		
Zhang 等 ^[49]	√						√				√	Isaac Gym	PPO			CL
QuestEnvSim ^[50]	√					√		√			√	Isaac Gym	PPO			
PHC ^[57]	√					√					√	Isaac Gym	PPO		√	IL
PULSE ^[65]	√					√					√	Isaac Gym	PPO	√		
Trace and Pace ^[64]		√			√		√	√			√	Isaac Gym	PPO	√		
Xu 等 ^[82]	√			√	√	√	√				√	Isaac Gym	PPO	√		IL
AdaptNet ^[97]	√						√	√			√	Isaac Gym	PPO	√		TL
MoConVQ ^[58]	√			√		√		√	√		√	ODE	√		√	
Zhu 等 ^[77]	√						√				√	Isaac Gym	PPO	√		

4.2 未来工作展望

目前,虚拟人运动控制策略主要应用于角色动画合成等领域。随着人工智能的快速发展和应用的推广,针对虚拟人控制策略的深度和广度有以下应用方向的发展前景。

(1) 多模态的感知与控制。已有的运动控制策略已经能够从运动数据集中学习到丰富多样的运动技能,如何拓展控制策略对于多模态输入(如声音、图像和文字等)的理解,是未来的重要研究方向之一。目前的工作在基于自身视觉输入、音乐律动输入和文本输入的虚拟人运动控制问题上取得了一定的进展,这些工作为拓展控制策略对多模态输入的理解提供了一些启示。未来可以探索如何更好地整合多模态输入的信息,实现更全面和多样化的虚拟人运动控制,进一步提升控制策略的广度。

(2) 世界模型学习。已有的基于深度强化学习的虚拟人运动控制策略依赖于与仿真环境进行交互采样来感知环境信息,而物理引擎不可微分的特性导致虚拟人运动控制策略很难与具有强大性能的生成式模型进行结合,限制了控制策略的广度。虽然已有工作通过学习虚拟人强相关的世界模型,并结合深度生成式模型在一些具有挑战性的任务目标取得了进展,但这样的世界模型只能模拟单一虚拟人在一定状态范围内的状态转移过程,具有较大的局限性。进一步拓展基于物理仿真环境下采样学习的世界模型的泛化能力,甚至是利用真实世界提供的样本来学习建模世界模型,是未来的研究方向。

(3) 具身智能。近年来,随着具身智能的发展,人们对于智能化的人形机器人控制的需求也在不断提高。虽然基于深度强化学习的控制策略已经在虚拟人运动控制问题上取得了显著的进展,但由于物理仿真引擎的局限性和关节驱动方式的不同,将控制策略直接应用到真实人形机器人的控制上并非易事,已经有工作在这方面取得了一定的进展^[105-106]。如何进一步打破虚拟与现实的界限,拓展控制策略的深度,将目前成熟的虚拟人控制策略应用到人形机器人运动控制领域,也是未来的研究方向。

本文对虚拟人运动控制问题的相关研究进行梳理与分析,以期后续的研究工作提供参考,也希望虚拟人的运动控制策略可以应用到更多创新领域。

参考文献(References):

- [1] Holden D, Komura T, Saito J. Phase-functioned neural networks for character control[J]. *ACM Transactions on Graphics*, 2017, 36(4): Article No.42
- [2] Starke S, Zhao Y W, Komura T, *et al.* Local motion phases for learning multi-contact character movements[J]. *ACM Transactions on Graphics*, 2020, 39(4): Article No.54
- [3] Liu L B, Yin K K, van de Panne M, *et al.* Sampling-based contact-rich motion control[C] // *Proceedings of the ACM SIGGRAPH Papers*. New York: ACM Press, 2010: Article No.128
- [4] Liu L B, Panne M V D, Yin K K. Guided learning of control graphs for physics-based characters[J]. *ACM Transactions on Graphics*, 2016, 35(3): Article No.29
- [5] Peng X B, Abbeel P, Levine S, *et al.* DeepMimic: example-guided deep reinforcement learning of physics-based character skills[J]. *ACM Transactions on Graphics*, 2018, 37(4): Article No.143
- [6] Tessler C, Kasten Y, Guo Y R, *et al.* CALM: conditional adversarial latent models for directable virtual characters[C] // *Proceedings of the ACM SIGGRAPH Conference Proceedings*. New York: ACM Press, 2023: Article No.37
- [7] Mourot L, Hoyet L, Le Clerc F, *et al.* A survey on deep learning for skeleton-based human animation[J]. *Computer Graphics Forum*, 2022, 41(1): 122-157
- [8] Kwiatkowski A, Alvarado E, Kalogeiton V, *et al.* A survey on reinforcement learning methods in character animation[J]. *Computer Graphics Forum*, 2022, 41(2): 613-639
- [9] Todorov E, Erez T, Tassa Y. MuJoCo: a physics engine for model-based control[C] // *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Los Alamitos: IEEE Computer Society Press, 2012: 5026-5033
- [10] Lee J, X. Grey M, Ha S, *et al.* Dart: dynamic animation and robotics toolkit[J]. *The Journal of Open Source Software*, 2018, 3(22): Article No.500
- [11] Bellman R. Dynamic programming[J]. *Science*, 1966, 153(3731): 34-37
- [12] Sutton R S, McAllester D, Singh S, *et al.* Policy gradient methods for reinforcement learning with function approximation[C] // *Proceedings of the 12th International Conference on Neural Information Processing Systems*. Cambridge: MIT Press, 1999: 1057-1063
- [13] Schulman J, Levine S, Moritz P, *et al.* Trust region policy optimization[C] // *Proceedings of the International Conference on Machine Learning*. Cambridge MA: JMLR, 2015: 1889-1897
- [14] Schulman J, Wolski F, Dhariwal P, *et al.* Proximal policy optimization algorithms[OL]. [2024-04-07]. <https://arxiv.org/abs/1707.06347>
- [15] Mnih V, Kavukcuoglu K, Silver D, *et al.* Playing atari with deep reinforcement learning[OL]. [2024-04-07]. <https://arxiv.org/abs/1312.5602>
- [16] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning[J]. *Nature*, 2015, 518(7540): 529-533
- [17] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning[C] // *Proceedings of the 30th AAAI Con-*

- ference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 2094-2100
- [18] Wang Z Y, Schaul T, Hessel M, *et al.* Dueling network architectures for deep reinforcement learning[C] //Proceedings of the 33rd International Conference on International Conference on Machine Learning. Cambridge MA: JMLR, 2016: 1995-2003
- [19] Schaul T, Quan J, Antonoglou I, *et al.* Prioritized experience replay[OL]. [2024-04-07]. <https://arxiv.org/abs/1511.05952>
- [20] Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning[C] //Proceedings of the 34th International Conference on Machine Learning. Cambridge MA: JMLR, 2017: 449-458
- [21] Fortunato M, Azar M G, Piot B, *et al.* Noisy networks for exploration[OL]. [2024-04-07]. <https://openreview.net/pdf?id=rywHCPkAW>
- [22] Hessel M, Modayil J, van Hasselt H, *et al.* Rainbow: combining improvements in deep reinforcement learning[C] //Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 3215-3222
- [23] Lillicrap T P, Hunt J J, Pritzel A, *et al.* Continuous control with deep reinforcement learning[OL]. [2024-04-07]. <https://arxiv.org/abs/1509.02971>
- [24] Haarnoja T, Zhou A, Abbeel P, *et al.* Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C] //Proceedings of the 35th International Conference on Machine Learning. Cambridge MA: JMLR, 2018: 1861-1870
- [25] Watkins C J C H, Dayan P. Q-learning[J]. *Machine Learning*, 1992, 8(2-3): 279-292
- [26] Schulman J, Moritz P, Levine S, *et al.* High-dimensional continuous control using generalized advantage estimation[OL]. [2024-04-07]. <https://arxiv.org/abs/1506.02438>
- [27] Silver D, Lever G, Heess N, *et al.* Deterministic policy gradient algorithms[C] //Proceedings of the 31st International Conference on International Conference on Machine Learning. Cambridge MA: JMLR, 2014: 387-395
- [28] Fujimoto S, Hoof H, Meger D. Addressing function approximation error in actor-critic methods[C] //Proceedings of the 35th International Conference on Machine Learning. Cambridge MA: JMLR, 2018: 1587-1596
- [29] Peng X B, Berseth G, Yin K K, *et al.* DeepLoco: dynamic locomotion skills using hierarchical deep reinforcement learning[J]. *ACM Transactions on Graphics*, 2017, 36(4): Article No. 41
- [30] Xie Z M, Ling H Y, Kim N H, *et al.* AllSteps: curriculum-driven learning of stepping stone skills[J]. *Computer Graphics Forum*, 2020, 39(8): 213-224
- [31] Liu L B, Hodgins J. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning[J]. *ACM Transactions on Graphics*, 2018, 37(4): Article No. 142
- [32] Lee S, Park M, Lee K, *et al.* Scalable muscle-actuated human simulation and control[J]. *ACM Transactions on Graphics*, 2019, 38(4): Article No. 73
- [33] Jiang Y F, van Wouwe T, de Groote F, *et al.* Synthesis of biologically realistic human motion using joint torque actuation[J]. *ACM Transactions On Graphics*, 2019, 38(4): Article No. 72
- [34] Park J, Min S, Chang P S, *et al.* Generative gaitNet[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2022: Article No. 22
- [35] Won J, Lee J. Learning body shape variation in physics-based characters[J]. *ACM Transactions on Graphics*, 2019, 38(6): Article No. 207
- [36] Lee S, Lee J, Lee J. Learning virtual chimeras by dynamic motion reassembly[J]. *ACM Transactions on Graphics*, 2022, 41(6): Article No. 182
- [37] Won J, Gopinath D, Hodgins J. Control strategies for physically simulated characters performing two-player competitive sports[J]. *ACM Transactions on Graphics*, 2021, 40(4): Article No. 146
- [38] Yin K K, Loken K, van de Panne M. SIMBICON: simple biped locomotion control[J]. *ACM Transactions on Graphics*, 2007, 26(3): Article No. 105
- [39] Luo Y S, Soeseno J H, Chen T P C, *et al.* CARL: controllable agent with reinforcement learning for quadruped locomotion[J]. *ACM Transactions on Graphics*, 2020, 39(4): Article No. 38
- [40] Peng X B, Ma Z, Abbeel P, *et al.* AMP: adversarial motion priors for stylized physics-based character control[J]. *ACM Transactions on Graphics*, 2021, 40(4): Article No. 144
- [41] Won J, Gopinath D, Hodgins J. Physics-based character controllers using conditional vaes[J]. *ACM Transactions on Graphics*, 2022, 41(4): Article No. 96
- [42] Won J, Gopinath D, Hodgins J. A scalable approach to control diverse behaviors for physically simulated characters[J]. *ACM Transactions on Graphics*, 2020, 39(4): Article No. 33
- [43] Park S, Ryu H, Lee S, *et al.* Learning predict-and-simulate policies from unorganized human motion data[J]. *ACM Transactions on Graphics*, 2019, 38(6): Article No. 205
- [44] Winkler A, Won J, Ye Y T. QuestSim: human motion tracking from sparse sensors with simulated avatars[C] //Proceedings of the SIGGRAPH Asia Conference Papers. New York: ACM Press, 2022: Article No. 2
- [45] Ye Y J, Liu L B, Hu L, *et al.* Neural3Points: learning to generate physically realistic full-body motion for virtual reality users[J]. *Computer Graphics Forum*, 2022, 41(8): 183-194
- [46] Zhang Y B, Gopinath D, Ye Y T, *et al.* Simulation and retargeting of complex multi-character interactions[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2023: Article No. 65
- [47] Peng X B, Guo Y R, Halper L, *et al.* ASE: large-scale reusable adversarial skill embeddings for physically simulated characters[J]. *ACM Transactions On Graphics*, 2022, 41(4): Article No. 94
- [48] Xie Z M, Starke S, Ling H Y, *et al.* Learning soccer juggling skills with layer-wise mixture-of-experts[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2022: Article No. 25
- [49] Zhang H T, Yuan Y, Makovychuk V, *et al.* Learning physically simulated tennis skills from broadcast videos[J]. *ACM Transactions on Graphics*, 2023, 42(4): Article No. 95
- [50] Lee S, Starke S, Ye Y T, *et al.* QuestEnvSim: environment-aware simulated motion tracking from sparse sensors[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2023: Article No. 62
- [51] Kanazawa A, Black M J, Jacobs D W, *et al.* End-to-end recovery of human shape and pose[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

- Los Alamitos: IEEE Computer Society Press, 2018: 7122-7131
- [52] Peng X B, Kanazawa A, Malik J, *et al.* SFV: reinforcement learning of physical skills from videos[J]. *ACM Transactions on Graphics*, 2018, 37(6): Article No.178
- [53] Luo Z Y, Iwase S, Yuan Y, *et al.* Embodied scene-aware human pose estimation[C] //Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2022: Article No.494
- [54] Yu R, Park H, Lee J. Figure skating simulation from video[J]. *Computer Graphics Forum*, 2019, 38(7): 225-234
- [55] Yuan Y, Kitani K M. Residual force control for agile human behavior imitation and extended motion synthesis[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.1826
- [56] Yuan Y, Wei S E, Simon T, *et al.* SimPoE: simulated character control for 3D human pose estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 7155-7165
- [57] Luo Z Y, Cao J K, Winkler A, *et al.* Perpetual humanoid control for real-time simulated avatars[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 10861-10870
- [58] Yao H Y, Song Z H, Zhou Y Y, *et al.* MoConVQ: unified physics-based motion control via scalable discrete representations[OL]. [2024-04-07]. <https://arxiv.org/abs/2310.10198>
- [59] Yuan Y, Kitani K. Ego-pose estimation and forecasting as real-time PD control[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 10081-10091
- [60] Luo Z Y, Hachiuma R, Yuan Y, *et al.* Dynamics-regulated kinematic policy for egocentric pose estimation[C] //Proceedings of the 35th Conference on Neural Information Processing Systems. San Diego: NIPS, 2021: 25019-25032
- [61] Bergamin K, Clavet S, Holden D, *et al.* DReCon: data-driven responsive control of physics-based characters[J]. *ACM Transactions on Graphics*, 2019, 38(6): Article No.206
- [62] Ho J, Ermon S. Generative adversarial imitation learning[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2016: 4572-4580
- [63] Gulrajani I, Ahmed F, Arjovsky M, *et al.* Improved training of wasserstein GANs[C] //Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 5769-5779
- [64] Rempe D, Luo Z Y, Peng X B, *et al.* Trace and pace: controllable pedestrian animation via guided trajectory diffusion[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 13756-13766
- [65] Luo Z Y, Cao J K, Merel J, *et al.* Universal humanoid motion representations for physics-based control[OL]. [2024-04-07]. <https://arxiv.org/abs/2310.04582>
- [66] Duan Y, Chen X, Houthoofd R, *et al.* Benchmarking deep reinforcement learning for continuous control[C] //Proceedings of the 33rd International Conference on International Conference on Machine Learning. Cambridge MA: JMLR, 2016: 1329-1338
- [67] Yu W H, Turk G, Liu C K. Learning symmetric and low-energy locomotion[J]. *ACM Transactions on Graphics*, 2018, 37(4): Article No.144
- [68] Abdolhosseini F, Ling H Y, Xie Z M, *et al.* On learning symmetric locomotion[C] //Proceedings of the 12th ACM SIGGRAPH Conference on Motion, Interaction and Games. New York: ACM Press, 2019: Article No.19
- [69] Tao T X, Wilson M, Gou R Y, *et al.* Learning to get up[C] //Proceedings of the ACM SIGGRAPH Conference Proceedings. New York: ACM Press, 2022: Article No.47
- [70] Merel J, Tassa Y, Dhruva T B, *et al.* Learning human behaviors from motion capture by adversarial imitation[OL]. [2024-04-07]. <https://arxiv.org/abs/1707.02201>
- [71] Merel J, Hasenclever L, Galashov A, *et al.* Neural probabilistic motor primitives for humanoid control[OL]. [2024-04-07]. <https://arxiv.org/abs/1811.11711>
- [72] Merel J, Ahuja A, Pham V, *et al.* Hierarchical visuomotor control of humanoids[OL]. [2024-04-07]. <https://arxiv.org/abs/1811.09656>
- [73] Wang Z, Merel J S, Reed S E, *et al.* Robust imitation of diverse behaviors[C] //Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 5320-5329
- [74] Makovychuk V, Wawrzyniak L, Guo Y, *et al.* Isaac Gym: high performance GPU-based physics simulation for robot learning[OL]. [2024-04-07]. <https://arxiv.org/abs/2108.10470>
- [75] Kingma D P, Welling M. Auto-encoding variational bayes[OL]. [2024-04-07]. <https://arxiv.org/abs/1312.6114>
- [76] Ling H Y, Zinno F, Cheng G, *et al.* Character controllers using motion vaes[J]. *ACM Transactions on Graphics*, 2020, 39(4): Article No.40
- [77] Zhu Q X, Zhang H, Lan M T, *et al.* Neural categorical priors for physics-based character control[J]. *ACM Transactions on Graphics*, 2023, 42(6): Article No.178
- [78] Yao H Y, Song Z H, Chen B Q, *et al.* ControlVAE: model-based learning of generative controllers for physics-based characters[J]. *ACM Transactions on Graphics*, 2022, 41(6): Article No.183
- [79] Fussell L, Bergamin K, Holden D. SuperTrack: motion tracking for physically simulated characters using supervised learning[J]. *ACM Transactions on Graphics*, 2021, 40(6): Article No.197
- [80] Yuan Y, Song J M, Iqbal U, *et al.* PhysDiff: physics-guided human motion diffusion model[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 15964-15975
- [81] Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial nets[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680
- [82] Xu P, Shang X M, Zordan V, *et al.* Composite motion learning with task control[J]. *ACM Transactions on Graphics*, 2023, 42(4): Article No.93
- [83] van den Oord A, Vinyals O, Kavukcuoglu K. Neural discrete representation learning[C] //Proceedings of the 31st Interna-

- tional Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 6309-6318
- [84] Rempe D, Birdal T, Hertzmann A, *et al.* HuMoR: 3D human motion model for robust pose estimation[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 11488-11499
- [85] Zhong C Y, Hu L, Zhang Z H, *et al.* AttT2M: text-driven human motion generation with multi-perspective attention mechanism[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 509-519
- [86] Heess N, Dhruva T B, Sriram S, *et al.* Emergence of locomotion behaviours in rich environments[OL]. [2024-04-07]. <https://arxiv.org/abs/1707.02286>
- [87] Liu L B, Hodgins J. Learning to schedule control fragments for physics-based characters using deep Q-learning[J]. ACM Transactions on Graphics, 2017, 36(3): Article No.29
- [88] Xu P, Karamouzas I. A GAN-like approach for physics-based imitation learning and interactive character control[J]. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 2021, 4(3): Article No.44
- [89] Peng X B, Chang M, Zhang G, *et al.* MCP: learning composable hierarchical control with multiplicative compositional policies[C] //Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2019: 3686-3697
- [90] Shazeer N, Mirhoseini A, Maziarz K, *et al.* Outrageously large neural networks: the sparsely-gated mixture-of-experts layer [OL]. [2024-04-07]. <https://arxiv.org/abs/1701.06538>
- [91] Jacobs R A, Jordan M I, Nowlan S J, *et al.* Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79-87
- [92] Bengio Y, Louradour J, Collobert R, *et al.* Curriculum learning [C] //Proceedings of the 26th Annual International Conference on Machine Learning. New York: ACM Press, 2009: 41-48
- [93] Lee S, Lee S, Lee Y, *et al.* Learning a family of motor skills from a single motion clip[J]. ACM Transactions on Graphics, 2021, 40(4): Article No.93
- [94] Yin Z Q, Yang Z S, Van De Panne M, *et al.* Discovering diverse athletic jumping strategies[J]. ACM Transactions on Graphics, 2021, 40(4): Article No.91
- [95] Parisi G I, Kemker R, Part J L, *et al.* Continual lifelong learning with neural networks: a review[J]. Neural Networks, 2019, 113: 54-71
- [96] Zhuang F Z, Qi Z Y, Duan K Y, *et al.* A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76
- [97] Xu P, Xie K X, Andrews S, *et al.* AdaptNet: policy adaptation for physics-based character control[J]. ACM Transactions on Graphics, 2023, 42(6): Article No.177
- [98] Li J F, Xu C, Chen Z C, *et al.* HybriK: a hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 3382-3392
- [99] Mahmood N, Ghorbani N, Troje N F, *et al.* AMASS: archive of motion capture as surface shapes[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 5441-5450
- [100] Loper M, Mahmood N, Romero J, *et al.* SMPL: a skinned multi-person linear model[J]. Seminal Graphics Papers: Pushing the Boundaries, 2023, 2: Article No.88
- [101] Harvey F G, Yurick M, Nowrouzezahrai D, *et al.* Robust motion in-betweening[J]. ACM Transactions on Graphics, 2020, 39(4): Article No.60
- [102] Ionescu C, Papava D, Olaru V, *et al.* Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325-1339
- [103] von Marcard T, Henschel R, Black M J, *et al.* Recovering accurate 3D human pose in the wild using imus and a moving camera[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 601-617
- [104] Geijtenbeek T, Pronost N. Interactive character animation using simulated physics: a state-of-the-art review[J]. Computer Graphics Forum, 2012, 31(8): 2492-2515
- [105] Radosavovic I, Xiao T T, Zhang B K, *et al.* Real-world humanoid locomotion with reinforcement learning[OL]. [2024-04-07]. <https://arxiv.org/abs/2303.03381>
- [106] Haarnoja T, Moran B, Lever G, *et al.* Learning agile soccer skills for a bipedal robot with deep reinforcement learning[J]. Science Robotics, 2024, 9(89): Article No.eadi8022