软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(2):711−738 [doi: 10.13328/j.cnki.jos.007006]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向机器人系统的虚实迁移强化学习综述*

林 谦[1], 余 超[1], 伍夏威[1], 董银昭[2], 徐 昕[3], 张 强[4], 郭 宪[5]

[1](中山大学 计算机学院, 广东 广州 510006)

[2](香港大学 机械工程系, 香港 999077)

[3](国防科技大学 智能科学学院, 湖南 长沙 410073)

[4](大连理工大学 计算机科学与技术学院, 辽宁 大连 116081)

[5](南开大学 人工智能学院, 天津 300350)

通信作者: 余超, E-mail: yuchao3@mail.sysu.edu.cn

摘 要: 近年来, 基于环境交互的强化学习方法在机器人相关应用领域取得巨大成功, 为机器人行为控制策略优化提供一个现实可行的解决方案. 但在真实世界中收集交互样本存在高成本以及低效率等问题, 因此仿真环境被广泛应用于机器人强化学习训练过程中. 通过在虚拟仿真环境中以较低成本获取大量训练样本进行策略训练, 并将学习策略迁移至真实环境, 能有效缓解真实机器人训练中存在的安全性、可靠性以及实时性等问题. 然而, 由于仿真环境与真实环境存在差异, 仿真环境中训练得到的策略直接迁移到真实机器人往往难以获得理想的性能表现. 针对这一问题, 虚实迁移强化学习方法被提出用以缩小环境差异, 进而实现有效的策略迁移. 按照迁移强化学习过程中信息的流动方向和智能化方法作用的不同对象, 提出一个虚实迁移强化学习系统的流程框架, 并基于此框架将现有相关工作分为 3 大类: 基于真实环境的模型优化方法、基于仿真环境的知识迁移方法、基于虚实环境的策略迭代提升方法, 并对每一分类中的代表技术与关联工作进行阐述. 最后, 讨论虚实迁移强化学习研究领域面临的机遇和挑战.

关键词: 强化学习; 迁移学习; 虚实迁移; 现实差距; 机器人控制

中图法分类号: TP18

中文引用格式: 林谦, 余超, 伍夏威, 董银昭, 徐昕, 张强, 郭宪. 面向机器人系统的虚实迁移强化学习综述. 软件学报, 2024, 35(2): 711–738. http://www.jos.org.cn/1000-9825/7006.htm

英文引用格式: Lin Q, Yu C, Wu XW, Dong YZ, Xu X, Zhang Q, Guo X. Survey on Sim-to-real Transfer Reinforcement Learning in Robot Systems. Ruan Jian Xue Bao/Journal of Software, 2024, 35(2): 711–738 (in Chinese). http://www.jos.org.cn/1000-9825/7006.htm

Survey on Sim-to-real Transfer Reinforcement Learning in Robot Systems

LIN Qian[1], YU Chao[1], WU Xia-Wei[1], DONG Yin-Zhao[2], XU Xin[3], ZHANG Qiang[4], GUO Xian[5]

[1](School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China)

[2](Department of Mechanical Engineering, University of Hong Kong, Hong Kong 999077, China)

[3](College of Intelligence Science and Technology, University of National Defense Science and Technology, Changsha 410073, China)

[4](School of Computer Science and Technology, Dalian University of Technology, Dalian 116081, China)

[5](College of Artificial Intelligence, Nankai University, Tianjin 300350, China)

Abstract: In recent years, reinforcement learning methods based on environmental interactions have achieved great success in robotic applications, providing a practical and feasible solution for optimizing the behavior control strategies of robots. However, collecting interactive samples in the real world can lead to problems such as high cost and low efficiency. Therefore, the simulation environment is

* 基金项目: 国家自然科学基金面上项目 (62076259, 62073176); 国家自然科学联合基金重点项目 (U1908214); 科技创新 2030—新一代人工智能重大项目 (2021ZD0112400)
收稿时间: 2023-01-13; 修改时间: 2023-06-22; 采用时间: 2023-07-24; jos 在线出版时间: 2023-11-08
CNKI 网络首发时间: 2023-11-10

widely used in the training process of robot reinforcement learning. By obtaining a large number of training samples at a low cost in the virtual simulation environment for strategy training and transferring learning strategies to the real world, the security, reliability, and real-time problems in the real robot training process can be alleviated. However, due to the difference between the simulation environment and the real environment, it is often difficult to obtain ideal performance when directly transferring the strategy trained in the simulation environment to the real robot. To solve this problem, sim-to-real transfer reinforcement learning methods are proposed to reduce the environmental gap, so as to achieve effective strategy transfer. According to the direction of information flow in the process of transfer reinforcement learning and the different objects targeted by intelligent methods, this survey first proposes a sim-to-real transfer reinforcement learning framework, based on which the existing related work is then divided into three categories: the model optimization methods focusing on the real environment, the knowledge transfer methods focusing on the simulation environment, and the iterative policy promotion methods focusing on both simulation and real environments. Then, the representative technologies and related work in each category are described. Finally, the opportunities and challenges in this field are briefly discussed.

Key words: reinforcement learning (RL); transfer learning; sim-to-real transfer; reality gap; robotic control

1 研究现状

当前, 强化学习 (reinforcement learning, RL)[1]方法在一系列复杂决策问题上取得了巨大成功, 如棋牌[2−4]和实时战略类游戏[5,6]、推荐系统[7−9]、自动驾驶[10,11]等. 在诸如机器人运动控制[12,13]、机器人操控[14−17]、运动导航[18−21]和机器人足球[22,23]等任务上, 强化学习也取得了令人瞩目的进展. 为了得到有效机器人控制策略, 强化学习依赖大量交互样本进行训练, 而在真实环境中获取样本具有较高的成本代价与安全风险; 此外, 由于机器人结构复杂且真实世界动态变化, 在机器人控制中运用强化学习依然面临有效性、安全性以及实时性等问题. 为减轻真实样本的需求, 仿真环境被广泛用于机器人策略学习当中. 基于仿真的机器人策略学习有如下优点: (1) 廉价性: 仿真环境的物理引擎能够以比实时更快的速度对真实环境进行计算模拟, 以较低成本生成训练样本用于机器人策略学习, 从而提高机器人策略的训练效率. (2) 真实性: 仿真环境不但能模拟机器人的完整运动特性, 如关节及关节之间的运动关联等, 还能模拟机器人和环境作用之间的物理属性, 如重力、压力、摩擦力等, 从而为真实世界建立逼真的物理模型. (3) 多维性: 在特定的机器人任务中可以利用多个仿真环境对真实世界进行不同粒度建模, 从不同层次反映真实场景的环境属性, 提供与真实世界相关的数据与信息以满足不同的应用需求. (4) 安全性: 在仿真环境中的试错行为没有实际风险, 可以重复不断地执行现实世界中耗时且危险的任务.

综上可知, 基于仿真的强化学习在机器人控制中具有一定优势. 为使机器人成功地完成现实世界中的一系列操作, 需要将仿真环境中学习的策略迁移至真实世界中. 然而, 由于仿真环境和真实世界之间存在现实差距 (reality gap)[24], 包括在不同平台中机器人动力模型的差异以及环境物理属性的差异 (如动作感知延迟、地面状况与大气状况等[25−27]), 即使最高逼真度的仿真环境也难以对真实世界进行完全一致的建模. 因此, 将仿真环境中学到的策略直接迁移到真实机器人上, 效果通常难以达到预期. 为了弥合仿真环境和现实之间的差异, 基于虚实迁移 (sim-to-real transfer)[24,28−30]的机器人强化学习方法通过解决仿真环境和真实环境之间的差异性问题, 从而实现学习策略的有效迁移. 近年来, 一系列虚实混合迁移强化学习方法被提出, 包括系统识别[31−33]、域随机化[34−37]、域自适应[38−40]、多保真度学习[41−43]等, 广泛地应用于运动控制、运动操控以及运动导航等机器人任务上, 取得了巨大的成功, 为机器人行为控制策略优化提供了一个现实可行的解决方案.

已有一些工作对现有的机器人虚实迁移学习方法进行了总结. Zhao 等人[30]对虚实迁移中基本概念与具体技术进行了简要的介绍. Dimitropoulos 等人[29]按照是否需要真实数据将现有虚实迁移方法分类为模拟器方法以及自适应方法, 前者不依赖真实数据, 后者则需要真实数据用于策略迁移. Salvato 等人[24]将虚实迁移方法分为 3 类: 域随机化、对抗强化学习以及迁移学习方法. Zhu 等人[28]按照不同的应用目标, 将仿生机器人研究中使用的虚实迁移方法分为 4 类: 基于精准的模拟器、基于运动学和动力学模型、基于分层与分布式控制器、基于演示的方法. 尽管这些工作对现有的虚实迁移学习研究工作进行了总结, 但缺乏一个通用的框架对现有工作进行全面梳理和分类. 因此, 本文对当前研究进行全面梳理, 从方法执行过程中信息流动和智能化方法作用对象的角度建立一个通用的虚实迁移学习框架, 并基于此框架将当前主要的虚实迁移强化学习方法划分为 3 类: 基于真实环境的模型

优化方法、基于仿真环境的知识迁移方法和基于虚实环境的策略迭代提升方法, 并对相关具体理论和应用进行讨论.

本文第 2 节介绍强化学习与迁移学习中重要的概念. 第 3 节深入探讨为了缩小仿真与现实之间的差异所采取的不同方法, 并提出一个通用的虚实迁移学习框架, 对迁移步骤中的数据信息流动和智能化方法作用的对象进行阐述与分析, 并在此基础上对现有方法进行分类, 分析它们的基本差异与优缺点. 第 4 节对该领域的目前的挑战进行分析, 并对于未来研究进行展望. 最后, 第 5 节对本文工作进行总结.

2 基础知识

2.1 深度强化学习

强化学习[1]是一种基于环境交互的机器学习范式, 智能体不断与环境进行交互并收集样本, 通过样本学习最优策略以实现累积回报的最大化. 智能体与环境的交互被建模为马尔可夫过程 (MDP), 由五元组 $(S,A,\mathcal{P},R,\gamma)$ 所定义, 其中 S 是状态的集合, A 是动作的集合, $\mathcal{P}$ 是状态转移函数/动力学模型 $\mathcal{P}: S\times A\times S\to[0,1]$, R 是奖励函数 $R: S\times A\times S\to\mathbb{R}$, $\gamma(0\leqslant\gamma\leqslant 1)$ 为折扣因子, 决定未来奖励对总回报 $G=\sum_{t=0}^{T}\gamma^t R_t$ 的影响. 强化学习算法使用智能体与环境的交互样本, 即由转移元组 (s_t,a_t,r_t,s_{t+1}) 组成的轨迹样本 τ, 对策略 π 进行优化, 使得总回报期望最大化, 即 $\pi=\underset{\pi}{\operatorname{argmax}}\,\mathbb{E}_{\tau\sim p(\tau|\pi)}[G(\tau)]$.

考虑到在机器人控制等实际问题中, 状态/动作空间通常是连续的且维度较高, 基于表格的方法不能很好地处理这类问题. 此外, 为获得最优控制策略, 通常需要建立一个复杂的非线性映射将状态空间映射到动作空间. 为解决上述问题, 需要将神经网络引入策略与值函数表征中. 深度强化学习 (deep reinforcement learning, DRL)[44,45]结合了强化学习以及深度学习, 不仅提供了用于环境交互和策略优化的强化学习范式, 同时具备深度学习赋予的强大的模型学习与信息提取能力. 从策略的表示形式出发, 深度强化学习可以被分为基于值函数的深度强化学习以及基于策略函数的深度强化学习.

基于值函数的深度强化学习利用深度神经网络以及梯度下降方法建立从状态、动作到期望回报的映射函数, 实现值函数的拟合与学习. 其代表工作包括 DeepMind 提出的深度 Q 网络算法 (deep Q network, DQN)[46], 该工作在利用深度神经网络表示值函数的基础上, 引入经验池回放机制并提出在值函数更新时将原网络与目标网络分离, 提高了样本的利用效率以及训练的稳定性. 在此基础上, 后续提出的一系列深度强化学习方法如 double Q-learning[47] 和 dueling network[48]针对值函数估计中出现的过估计、值函数分解等问题提出了新的解决方案. 虽然 DQN 算法在离散动作任务上具有不错的表现, 但较难应用于如机器人控制等具有连续动作空间的任务上.

在基于策略函数的深度强化学习中, 策略被直接表示为状态到动作的映射, 而深度神经网络则被用于学习该映射. 相较基于值函数的深度强化学习, 基于策略函数的深度强化学习可以解决具有连续动作空间的任务且具有更高的训练效率, 但在训练时也更容易陷入局部最优. 其代表工作包括 Lillicrap 等人提出的深度确定性策略梯度算法 (deep deterministic policy gradient, DDPG)[49], 该方法学习确定性策略并且通过加入噪音提高探索能力, 具有较高的训练效率, 被广泛应用于连续动作控制任务中. Schulman 等人提出的近端策略优化算法 (proximal policy optimization, PPO)[50]提出了 Clip 技巧, 通过限制策略的波动提高策略稳定性. 此外, Haarnoja 等人提出的柔性执行者-评论者算法 (soft actor-critic, SAC)[51]在策略的优化目标中加入了熵正则项, 通过最大化策略的熵提高算法的探索能力. 深度强化学习在诸多领域中都取得了瞩目的成果, 尤其在机器人控制领域, 通过从仿真环境中获取的低成本且数量无限制的样本实现机器人控制策略的高效学习.

2.2 迁移强化学习

在传统机器学习领域中, 迁移学习从不同域之间的数据分布差异出发实现域间的知识迁移. 定义源域为 D_S, 源域中的学习任务为 T_S, 目标域为 D_T, 目标域中的学习任务为 T_T, 迁移学习是指当 $D_S\neq D_T$, $T_S\neq T_T$ 时, 将 D_S 的知识迁移至 D_T, 从而提高目标域的学习效率. 域可表示为 $D=\{\chi,P(X)\}$, 其中 χ 为特征空间, $P(X)$ 为边缘概率分布; 学习任务可表示为 $T=\{y,P(Y|X)\}$, 其中 y 为标记空间, $P(Y|X)$ 为预测函数. 一种经典的迁移场景是在一个分

类任务 (D_T) 中缺少标注数据, 而在另一个相似的分类任务 (D_S) 中有足够的标注数据, 但后者的数据可能遵循不同的数据分布 ($P(X_S) \neq P(X_T)$) 或样本与标注信息间存在不同的关联模式. 在这种情况下, 使用迁移学习完成知识的转移将避免昂贵的数据标记工作, 从而极大地提高学习性能. 传统迁移学习方法可分为 4 类[52]: 基于实例的迁移学习[53,54]、基于特征的迁移学习[55,56]、基于参数的迁移学习[57,58]和基于关系的迁移学习[59,60]. 基于实例的迁移学习方法适用于源域和目标域相似度较高的情况, 其核心是挑选源域数据的某些部分进行复用, 通过改变样本的存在形式来减少源域和目标域的差异; 基于特征的迁移学习算法可应用在域间相似度不太高的情况, 核心是通过特征变换使源域和目标域在某个特征空间下表现出相似的性质; 基于参数的迁移学习方法从模型的角度出发, 共享源域与目标域模型之间的某些参数以达到迁移学习的效果; 基于关系的迁移学习将数据之间的联系从源域迁移到目标域, 通过将两个域之间的相关性知识建立一个映射来实现迁移学习.

在强化学习领域, 迁移学习具体关注不同域之间状态、动作、奖励等数据分布的关联与差异性问题. 源域与目标域是否具有相同的状态空间 S 或动作空间 A 领域将直接影响迁移学习的难度、适用的方法等问题[61,62]. 强化学习问题下迁移学习的目标是从轨迹样本、值函数等构件中提取知识并迁移到目标域, 以提高目标域上策略的性能表现: 令 M_S 表示源域, M_T 表示目标域, $I_S \sim M_S$ 是从源域中获取的外部信息, $I_T \sim M_T$ 是从目标域中的内部信息, 迁移学习的最终目标是获得目标域上的最优策略 π^*, 即 $\pi^* = \underset{\pi}{\operatorname{argmax}} \mathbb{E}_{a\sim\pi}[G]$, 其中 π 是基于 I_S 与 I_T 学到的从状态到动作的映射. 迁移强化学习中常用方法[62,63]包括奖励函数设计[64,65]、示例学习[66,67]、策略迁移[68,69]、任务间映射[70,71]以及表征迁移[71,72]等. 奖励函数设计类方法指通过源域收集的信息设计合理的奖励函数引导在目标域上的策略学习, 以提高学习效率与策略表现; 示例学习类方法通过人类指导或是专家策略与环境交互产生的示例样本实现知识迁移. 在源域与目标域 MDP 相同或是非常相似的情况下, 可以通过模仿学习等方法将示例样本用于学习目标域上的最优策略; 策略迁移方法关注如何将在源域上预先训练好的策略迁移到目标域, 通过策略蒸馏、策略复用等方法提取并迁移源域策略中的知识; 任务间映射方法假设源域与目标域的状态、动作、奖励函数等构成之间存在某种映射关系, 通过手工或学习方法求解该映射关系实现迁移学习; 表征学习方法迁移的对象是从源域中学到的特征表示, 从价值函数、策略网络等部件中提取特征, 然后直接或是进一步解构后迁移到目标域.

2.3 仿真环境

机器人迁移强化学习对仿真环境的逼真度与采样效率提出了较高要求, 仿真环境的选择将对最终的迁移效果产生重要影响. 表 1 列举了当前机器人控制与策略学习中常用的仿真环境[73–92], 并从适用场景、优点与缺点等方面进行对比与分析. 物理模拟器是最重要的一类仿真环境, 其中 Gazebo[73]、Webots[74]以及 V-REP[75]是传统机器人控制领域中常见的物理模拟器, 具有较丰富的生态系统, 被广泛应用于工业机器人的开发以及机器人控制学习中. Gazebo 能够模拟室内外环境中机器人之间的交互, 提供逼真的传感器反馈, 完全开源且可免费访问, 拥有庞大的开发贡献者; Webots 是一个开源机器人模拟器, 为建模、编程和模拟机器人提供了一个完整的开发环境, 拥有友好的使用界面并支持控制算法的快速测试; V-REP 作为一种通用且可扩展的仿真框架, 支持多种不同的编程语言, 允许将控制器和功能嵌入到仿真模型中.

与 Gazebo[73]物理模拟器不同, 基于接触动力学模型的物理引擎如 MuJoCo[78]、PyBullet[83]被广泛地运用机器人强化学习研究领域. 依托于这些物理引擎, 如 OpenAI Gym[79]和 DeepMind Control[80]等强化学习集成环境被不断提出, 其中 MuJoCo[78]专注于机器人和生物力学仿真以及动画和机器学习应用, 支持常见的机器人任务场景仿真, 具有相当高效的样本收集效率, 契合了强化学习对样本的巨大需求; PyBullet[83]基于 Bullet 物理引擎, 同样具有不错的样本收集效率, 将机器人控制与机器学习相结合, 支持加载多种模型格式, 并且拥有一个大型社区为开发者提供支持.

上述机器人仿真平台注重仿真通用性, 在真实性上与特定的实际应用场景还存在一定差距, 导致其在某些特定场景下的仿真存在较大误差. 因此, 一些针对特定机器人场景的高逼真度仿真模拟平台受到关注, 例如专用于机器人足球世界杯的 SimSpark[93]、用于混合四足机器人的 ANYmal[94]、用于两足机器人的 ATRIAS[95], 以及高分辨率光学触觉传感模拟器 Tactile Gym[96]和进一步扩展的 Tactile Gym 2.0[97]、视触觉传感器 GelSight[98]等. 除此之

外, 最近如 Unreal Engine、 Unity3D 等游戏引擎被逐渐应用于机器人强化学习之中, 具有逼真的仿真画质以及丰富的任务场景, 尤其适用于无人驾驶等对画质与训练场景具有要求较高的任务, 其中 AirSim[86]专注无人车以及无人飞行器控制的仿真插件.

表 1 主流仿真环境介绍

类别	仿真环境	适用场景	基于该模拟器的强化学习环境	优点	缺点
物理模拟器	Gazebo[73] + ROS	对真实世界的机器人、传感器、执行器、物体和材料进行仿真, 适用于机械臂、双足机器人、无人机等多类型机器人的操纵、运动、控制、导航等任务	NA	完全开源免费, 算法与模型库丰富, 便于ROS框架下的真机迁移	参考文档不足, 入门难点稍高
传统控制领域	Webots[74]		NA	Webots以及V-REP教育版开源免费, 提供多种编程接口, 支持多种操作系统, 具有良好的ROS接口, 支持对碰撞和接触点进行精确建模	缺少强化学习框架及接口支持, 与ROS系统的兼容略低, 社区规模较小
	V-REP[75]/PyRep[76]		RLBench[77]		
面向强化学习	MuJoCo[78]		1. OpenAI Gym[79] 2. DeepMind control suite[80] 3. Surreal Robotics Suite[81] 4. ROBEL[82]	开源免费, 支持常见的机器人任务场景仿真, 提供便于强化学习的框架与接口, 具有较高的样本收集效率, 具有庞大的社区支持	与真实物理世界存在一定差异, 模拟计算需要专用的硬件, 成本较高, 计算成本与仿真精准度上的不足限制机器人的仿真效果
	PyBullet[83]/Bullet		1. Roboschool 2. PyBullet Gym 3. GibsonEnv[84] 4. PyRoboLearn[85]		
游戏引擎	Unreal Engine/ Unity3D	主要用于针对无人车或飞行无人机导航任务的强化学习	1. AirSim[86] 2. Carla[87]	仿真图像画质逼真, 仿真效果高, 可设置多种任务场景	对计算资源要求较高, 仿真时间较长
可微分模拟器	Brax[88]	支持刚体、柔体、弹性对象的接触、碰撞、反弹等物理过程的仿真, 实验提供了简易的机器人仿真场景	NA	深度神经网络控制器及物理模拟模块相结合, 支持仿真环境的动态学习与提升	技术提出的时间不长, 缺少完善的集成平台或仿真软件, 与上述仿真环境相比技术尚未成熟与推广, 难以运用于机器人强化学习研究之中
	太极[89]		NA		
	JBDL[90]		NA		
GPU模拟器	RaiSim[91]		NA	支持模拟器直接在GPU上进行并行仿真, 极大地提高了仿真的速度	
	Isaac Gym[92]		NA		

在最新的关于仿真环境的研究工作中, 可微分模拟器受到了广泛关注. 可微分模拟器将深度神经网络控制器及物理模拟模块进行结合, 允许梯度通过环境进行传播, 并在训练机器人的同时动态优化仿真环境. 可微分模拟器在仿真速度与效果上都有很大的提升, 其代表工作有可实现自动微分的框架——太极[89]、Google 开发的 Brax[88]以及腾讯的以 Jax[99]为基础的机器人身体动力学算法库 JBDL[90]等. 另外, 不同于以往在 CPU 上用多线程并行采集仿真数据并传送到 GPU 上训练的做法, 以 NVIDIA 开发的 Isaac Gym[92]为代表的 GPU 模拟器, 支持模拟器直接在 GPU 上进行并行仿真与训练, 节省了数据传输时间, 从而极大地提高了样本的采集效率. 与 Isaac Gym 类似, RaiSim[91]是一个开源的物理仿真引擎, 通过 GPU 加速实现高性能的实时物理仿真, 可用于实现高性能、实时的仿真环境. RaiSim 专注于模拟刚体和柔性体的物理行为, 并提供了多种功能和特性, 使其适用于机器人、虚拟现实、仿真训练和物理交互等领域.

3 虚实迁移强化学习

虚实迁移[24,30,100]指机器人在仿真环境中进行训练, 获得的策略被迁移部署至真实机器人控制系统中, 其中仿真环境作为源域, 真实环境作为目标域, 机器人的控制策略作为迁移对象, 核心内容是实现环境信息与控制信息在

真实环境与仿真环境之间的迁移. 虚实迁移对迁移效果提出了 4 个重要的目标: 有效性、效率性、安全性、泛化性. 有效性[101,102]是指在仿真环境中学到的策略能在真实世界中发挥预期作用, 即在仿真环境中获得高回报的策略在真实环境中也能获得较高回报, 针对有效性的研究大多从缩小环境差异角度着手. 效率性[103,104]从采样的时间与经济成本出发, 强调在真实世界中进行尽可能少的探索采样, 针对效率性的研究关注对真实样本中环境信息的充分挖掘与利用. 安全性[105,106]强调现实中的探索与直接的策略部署具有潜在的安全风险及较高的成本代价, 需要指导真实机器人进行安全的探索与样本收集. 泛化性[107,108]针对多任务迁移场景, 强调获得的策略能在尽可能少的微调下快速适应不同环境的任务. 本文主要关注围绕虚实迁移的有效性以及效率性开展的相关研究工作, 在此基础上对现有方法进行梳理总结.

本文从迁移学习过程中的信息流动和智能化方法作用对象的角度提出了一个通用的虚实迁移强化学习系统流程框架, 如图 1 所示. 首先将策略学习与迁移过程分为以下 5 个主要步骤: 基于真实环境的仿真模型优化、仿真策略优化、基于仿真环境的知识迁移、真实环境探索与评估, 以及基于虚实环境的策略迭代提升.

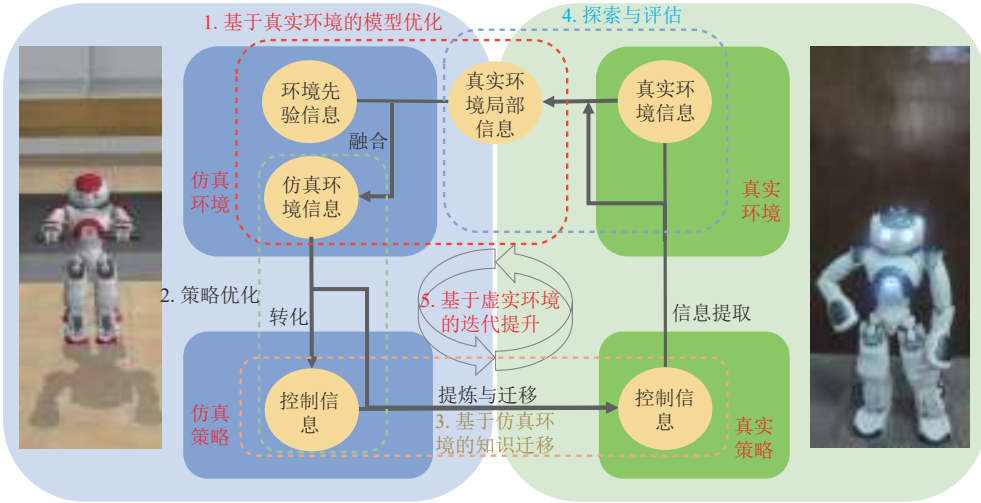


图 1 虚实迁移强化学习流程框架图

(1) 基于真实环境的仿真模型优化: 通过修正或调整仿真模型来缩小仿真环境与真实环境的差异.

(2) 仿真环境策略优化: 通过仿真环境的交互样本对策略进行学习训练.

(3) 基于仿真环境的知识迁移: 将仿真环境中获取的学习知识向真实环境迁移.

(4) 真实环境探索与评估: 对真实环境进行充分探索以及部署策略性能评估.

(5) 基于虚实环境的策略迭代提升: 上述 4 个步骤的循环迭代实现策略的不断提升.

真实环境信息包含了真实任务场景中的动力学模型以及奖励模型, 前者受复杂的环境参数, 如温度、湿度、摩擦力系数等影响, 后者则由最终的任务目标及奖励结构决定. 复杂的真实环境难以被精准且完整地表征, 但蕴含于样本数据、评估指标以及人类对任务的先验认知中的真实环境信息能够被有效利用, 这些信息蕴藏于人类在建立仿真环境时所用的物理方程、环境参数的估计以及对参数范围的约束中. 基于真实环境的仿真模型优化通过真实世界的样本数据修正或调整仿真环境, 其本质是将仿真模型中的先验但不精确信息与真实样本中的真实但局部信息相融合以还原真实环境. 在仿真环境策略优化中, 机器人与仿真环境进行高效交互生成大量廉价样本并利用当前已有的深度强化学习算法如 DDPG[49]、PPO[50]等实现策略优化, 在此过程中融合真实环境信息后的仿真环境信息被转化成仿真控制策略中的控制信息. 在基于仿真环境的知识迁移中, 仿真环境中优化得到的策略控制信息通过分解、组合或者直接迁移等手段从仿真策略中被提炼并迁移到真实策略. 最后, 迁移得到的真实策略又用于真实环境探索与评估, 通过样本采集以及策略评估从真实环境中提取局部信息, 通过真实策略的控制信息引导机

器人探索真实环境的未知部分以建立更加精准的环境模型.

虚实迁移中的仿真环境策略优化、真实环境探索与评估是传统强化学习领域中独立的研究方向. 现有的虚实迁移强化学习方法通常针对基于真实环境的仿真模型优化、基于仿真环境的知识迁移以及基于虚实环境的策略迭代提升这 3 个部分开展, 所以本文着重从上述 3 个角度出发对虚实迁移方法及代表工作进行分类与总结, 并阐述每类方法内在动机与关联.

3.1 基于真实环境的模型优化方法

该类方法旨在根据真实环境实现对仿真环境的优化, 进而缩小仿真环境与真实环境的差异或是通过合理的环境参数设置来提高训练效率. 如图 2 所示, 该类方法根据真实环境确定仿真环境集合, 从中选择特定的仿真环境用于策略训练. 决定环境集合以及控制环境选择的技术包括系统识别、域随机化、课程式学习与多保真度仿真. 此外, 从迁移目标上看, 系统识别、域随机化通过修正或随机化仿真环境以缩小仿真和真实环境的差异, 而课程式学习与多保真度仿真则根据具体任务合理设置仿真环境参数以加速策略学习和训练.

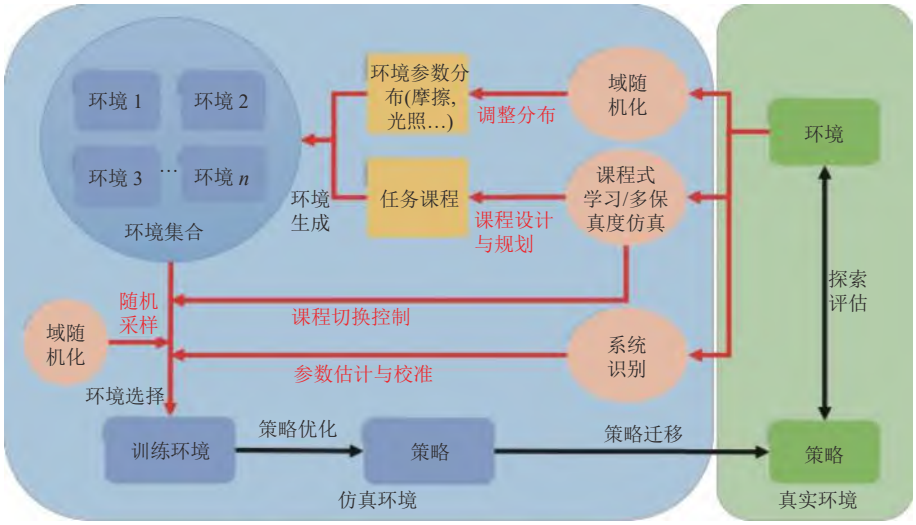


图 2 基于真实环境的模型优化方法的技术路线

3.1.1 系统识别

系统识别[109]类方法通过为物理系统建立一个精确的数学模型并对模型进行精确校准, 从而使模拟器提供的仿真环境更加逼真, 已广泛用于机器人领域[110−112], 涉及运动学校准、自适应控制、环境预测及摩擦估计等一系列问题[31,33]. 如图 2 所示, 系统识别根据真实环境动态调整仿真环境的模型参数, 减小对应的仿真环境与真实环境之间的差异. 通过系统识别构建的模型需要经历不断的修改与调整, 在确定合适的数学模型后可以用静态的参数来表示这些真实系统的组件模型, 并通过最小二乘法、协方差方法、频谱分析和时间序列分析等方法进行参数优化[112]. 此外, 一些端到端的方法, 如人工神经网络, 也被广泛地用于表示系统组件的模型. Allevato 等人[113]构建了网络模型 TuneNet, 使用神经网络直接调整一个模型的参数以匹配另一个模型, 将来自两个不同模型 (即仿真环境和真实世界) 的观察结果作为输入, 估计模型之间参数的差异, 并利用迭代剩余调谐技术进行快速准确的一次性系统识别, 直接调整仿真环境的系统模块参数以匹配真实环境模型, 提高了调优速度并减少了所需的模拟样本数量. 如图 3 所示, 仿真环境和真实世界分别利用当前环境的动力学模型 f_s 、 f_r 以及观测函数 z_s 和 z_r 得到观测值 O_{sim} 和 O_{real}, 根据观测值及神经网络 h_θ 直接估计参数误差, 并通过 $\xi_k=\xi_{k-1}+\Delta\xi$ 更新参数, 如此迭代直至达到以下优化目标, 其中 α 是权重正则化常数.

$$\arg\min_{\pi}\sum_{n=1}^{N}\|(\xi_{k_n}+h_\theta(O_{sim_n},O_{real_n})-\xi_{r_n}\|+\alpha\|\theta\|^2 \tag{1}$$

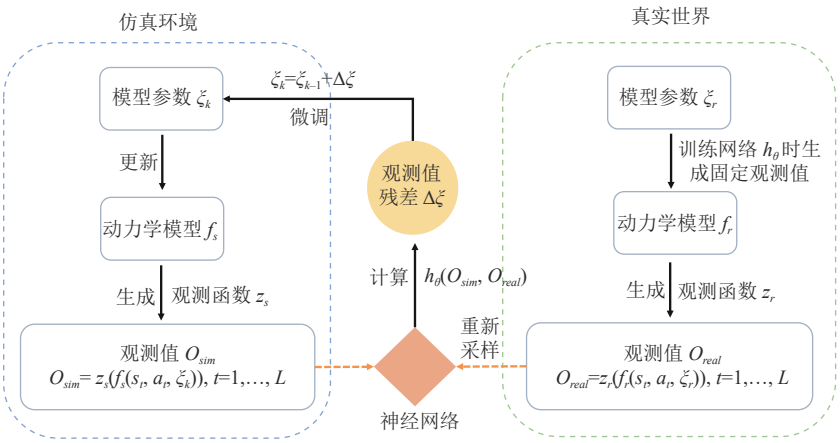


图 3 TuneNet 网络模型结构

Du 等人[114]将系统识别重新建模为参数搜索问题, 建立模型预测当前仿真环境参数与真实环境参数的大小关系, 在此基础上展开搜索求解合适的目标参数. 为了更精确地对真实世界的物体建模, Gao 等人建立了一个由 100 个隐式视觉、听觉和触觉表示的 3D 对象组成的数据集 ObjectFolder[115,116], 促进了机器人抓取、运动协同作用等研究, 在涉及随机动力学的任务上弥合仿真与现实之间的差距. Kumar 等人[117] 针对四足机器人实时在线自适应的问题, 提出一种快速电机自适应算法 RMA. 其中心思想是: 假设系统当前状态是机器人在特定环境下产生的, 则根据过去的状态和动作信息可以推断出当前的环境信息的隐变量 z_t. 在真实部署阶段利用仿真阶段训练的适应模块 Φ 进行在线的系统识别, 输出机器人动作的关节角度. 基于此工作, 研究人员后续分别在双足机器人[118]、机械臂[119]和四轴飞行器上[120]验证了该算法思想的有效性.

尽管精确的仿真模型能有效地缩小虚实环境之间的差异, 但直接对真实环境下的机器人系统建模仍极具挑战, 存在诸如缺乏观测样本和存在估计误差等问题, 且机器人运行机理通常十分复杂, 需要考虑如响应延迟、关节阻尼等[121,122]因素影响. 此外, 系统识别建模的准确性通常取决于研究人员的经验, 对于一些动态环境建模的准确度仍有所欠缺. 因此, 提出误差更小且具有自主性的建模方法将是系统识别下一阶段的研究重点.

3.1.2 域随机化

由于真实机器人系统难以被精确建模, 而且随着温度、湿度、位置或时间的改变, 同一机器人系统的物理参数可能发生显著变化, 从而加剧系统识别的难度. 此外, 不断地调整仿真环境参数以适应环境变化需要付出昂贵的代价, 且难以保证其准确性[123]. 域随机化方法旨在通过对仿真环境进行随机化, 生成鲁棒的控制策略以解决系统识别难以适应动态环境以及识别准确度不足的问题, 被广泛应用于不同的机器人控制任务[34−37]. 域随机化原理如下: 尽管仿真环境和真实世界之间存在差异, 但不必对真实世界所有的参数进行精确的识别与建模, 而是在训练过程中根据环境参数的分布对仿真环境进行随机化, 以覆盖真实的动态环境, 获得更鲁棒的策略[30,124]. 如图 2 所示, 在迁移框架中域随机化根据真实环境调整环境参数的随机分布, 每次训练根据该分布随机采样一组环境参数并生成对应仿真环境用于机器人训练与学习. 具体而言, 一组环境参数 ξ 决定了仿真环境的动力学模型 $\mathcal{P}_{\xi\sim p(\xi)}=\mathcal{P}(s_{t+1}|s_t,a_t,\xi)$, 仿真策略的训练目标重新定义为在分布 $\xi\sim p(\xi)$ 对应的环境中实现回报最大化, 即: $\max_{\pi}\mathbb{E}_{\xi\sim p(\xi)}[\mathbb{E}_{\tau\sim p(\tau)}[G(\tau)]]$. 域随机化的核心在于参数 ξ 的选择以及分布 $p(\xi)$ 的确定. 常见的方式包括通过专家知识与手工设计[123,125]或者通过学习的方式获得[101] 获得环境参数及其分布. 由于随机参数的选择对最终的迁移效果有较大的影响, 因此通常根据任务的不同进行不同的随机参数选择. 例如, 在基于图像输入的虚实迁移任务中[35,126], 颜色、光照属性、纹理以及干扰噪声等基础的视觉属性被广泛地挑选作为随机化属性. Alghonaim 等人[35]针对基于图像的姿态估计任务中不同环境随机参数的重要性开展研究, 发现图像质量、干扰噪音以及纹理对域随机化效果有较大影响. James 等人[104]结合域随机化和自适应方法, 通过对抗网络将随机渲染图像和现实世界图像转化为统一的标准模拟图像,

大幅提高策略训练效率.

对于随机参数的分布, 目前域随机化方法多从人工设计的固定先验分布中采样域参数, 例如低保真度仿真的动态随机化方法[125]、应用域随机化进行姿态估计[127]、基于域随机化并具有可转移性的策略优化算法 (SPOTA)[128]等. 而 Muratore 等人[124]提出了贝叶斯域随机化 (BayRn) 策略搜索算法, 其核心是在随机仿真环境中进行策略训练, 同时在学习过程中使用贝叶斯优化来适应源域分布. 与以往的工作相比, 该算法不局限于固定的分布, 利用域分布参数在仿真环境中构建了域分布参数和策略收益之间的概率模型, 因此在 BayRn 优化中只需要与真实环境进行少量的交互. 此外, Muratore 等人[129]同样使用了贝叶斯优化提出了神经后域随机化 (NPDR) 来调整仿真环境参数, 摆脱了域随机化中对域参数分布的一些限制性假设. 作为高效便捷的虚实迁移方法, 域随机化已被广泛用于弥合仿真与现实的差距, 但当下仍面临一些亟待解决的问题. 例如, 许多方法仍从固定分布中采样域参数; 部分方法中域随机化参数分布的调整缺乏主动性, 过分依赖人工经验和启发式方法, 在复杂高维场景中的效率也有待提升.

3.1.3 课程学习与多保真度仿真

课程式学习的概念最早由 Bengio 等人在 2009 年提出[130], 借鉴人类由易到难的学习顺序, 主张让模型先从容易的样本开始学习, 并逐渐进阶到复杂的样本与知识. 课程式学习为不同难易的训练样本赋予不同的权重, 随着训练过程的持续, 简单样本的权重下降而困难样本的权重上升, 对样本进行权重动态分配的过程被称为课程. 在机器人迁移强化学习中, 课程式学习成为有效解决高样本复杂度以及稀疏奖励值等问题的有效方法之一[131,132]. 如图 2 所示, 针对虚实迁移场景, 课程式学习根据真实的目标任务手工或自动设计由易到难的一组学习任务, 每个任务被建模为马尔可夫过程, 并对应一个独立的任务环境, 共同构成仿真环境集合. 作为课程式学习中的关键部分, 课程的设计以及切换机制直接作用于仿真环境的模型优化中, 通过设置合理的任务环境减少高成本样本的需求, 进而解决稀疏奖励值的问题和提高训练效率.

针对课程设计与规划问题, Florensa 等人[133]针对稀疏奖励值问题提出了一种课程设计方法, 根据不同的课程任务中机器人的初始状态与目标状态的距离来实现任务由简单过渡到复杂. Nguyen 等人[134]将课程式学习与机械臂、移动机器人的分层结构相结合提出分层强化学习框架, 利用学习任务的依赖关系以及层次结构生成课程, 通过内在动机探索以及主动模仿学习实现任务的切换. Sukhbaatar 等人[135]和 He 等人[136]采用了对抗学习的思想, 利用上层的生成策略产生课程任务供底层的行动策略进行课程式学习, 两级策略同时且独立地训练并保证生成的课程任务难度适中以满足训练效率要求. 针对课程的切换控制问题, Shukla 等人[137]从特定任务的奖励回报以及任务特征出发设计了多种在线方法来规划课程学习的顺序, 在多目标导航任务上完成了从仿真到实物的迁移. Matiisen 等人[138]将任务的切换建模为多臂赌博机问题, 根据历史的训练记录拟合策略在某一任务中的下一次训练回报, 并采用探索方法挑选出在该任务上表现增长或下降最快的任务作为下一个训练任务.

按照课程学习的思想, 多保真度学习在不同保真度的仿真环境中进行策略训练, 其中, 低保真度与高保真度的仿真环境分别对应简单与困难的训练任务. 通过合理地安排策略在不同仿真环境的切换, 可以极大减少物理机器人所需的样本数量, 节省训练成本与时间. Cutler 等人[41]在 2015 年提出的多保真度学习框架允许智能体停留在仍为其提供有利用价值信息的最低保真度模拟器上进行策略训练, 从而减少高保真度环境上所需要的昂贵的样本. 作者在之后的工作中[42]将不同保真度的仿真环境与真实世界上的策略学习形成闭环, 如图 4 所示, 在简单仿真环境 Σ_s 中通过如经典、自适应和最优控制等传统控制方法得到最优策略 π_s^*, 结合 K-means 方法和正则化径向基函数 (RBF) 网络将 π_s^* 转化为复杂仿真环境 Σ_c 的初始策略 π_c^{init}. 之后在复杂仿真环境中使用基于学习的方法如强化学习 PILCO 算法[139]等来优化初始策略, 得到的策略 π_c^{new} 再用于指导真实世界 Σ_{rw} 中的学习. 来自真实世界的观测数据用于评估 π_c^{new} 是否是最优, 若不是最优则使 $\pi_c^{\text{init}} = \pi_c^{\text{new}}$, 结合观测数据继续迭代更新策略, 从而实现虚实闭环. Di Castro 等人[43]提出一种有效方法来平衡高吞吐量、低成本、低保真度的仿真样本以及低吞吐量、高成本、高保真度的真实样本, 为每个环境维护单独的经验池, 并设计训练权重以平衡不同样本对策略训练的影响. 与传统的不断增加保真度的方法相比, Truong 等人[140]提出了通过降低保真度以提升虚实迁移效果的思想, 并将策略分解为仅在仿真中训练的“高级策略”和完全在硬件上设计的“低级控制器”以缓解仿真训练缓慢和过拟合的问题.

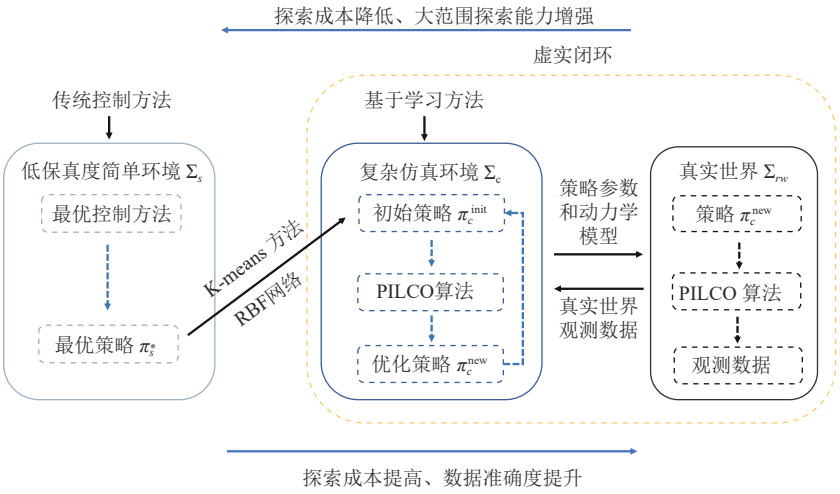


图 4 多保真度学习闭环框架

随着课程学习的应用场景日益丰富, 对该领域的总结和梳理也至关重要. Wang 等人[141]从动机、定义、理论和应用等各个方面全面回顾了课程学习, 详细说明如何设计一个手动预定义课程或自动课程, 提出了统一的课程学习框架将课程学习算法分为两个大类和多个小类. Zhou 等人进一步发布了全球首个课程学习开源库 CurML[142], 将现有课程学习算法集成至一个高度封装的统一框架中, 具有较高的易用性. 当前, 课程学习和多保真度仿真方面的研究仍面临一些问题. 例如, 在预设课程时, 由于人类思维与机器思维的差异, 许多人类认为简单的样本对模型来说较为困难, 这种人机决策边界的不一致容易造成课程梯度设置不合理等问题. 此外, 课程学习还存在专家知识昂贵、灵活性欠缺等问题. 针对课程学习的未来研究方向集中在建立评价数据集和指标、提供更完善的理论分析和算法等方面.

3.2 基于仿真环境的知识迁移方法

基于仿真环境的知识迁移方法关注如何将仿真环境获得的知识向真实策略进行迁移, 如图 5 所示, 可分为基于样本的知识迁移以及基于策略的知识迁移, 前者的迁移对象是包含了环境信息的仿真样本, 后者的迁移对象是包含了控制信息的仿真策略. 基于样本的知识迁移关注仿真轨迹的迁移与利用, 具体是通过重要性权重以及模仿学习将仿真样本向真实样本经验池进行迁移. 基于策略的知识迁移关注仿真策略分解与部件迁移, 按功能或层次对策略进行划分与迁移.

3.2.1 仿真轨迹的迁移与利用

由于仿真环境与真实环境存在内在关联性, 仿真环境获取的大量丰富的样本数据一定程度上能包含真实环境的动力学模型与奖励模型信息, 如何利用仿真样本将是真实策略优化的重点, 也是提高训练效率的关键. 如图 5 所示, 通过重要性权重与模仿学习实现仿真样本向真实样本经验池的迁移是利用仿真轨迹的重要方法. 通过这些方法, 仿真样本极大地拓展真实策略的训练样本池, 为真实策略提供丰富的演示信息, 进而提高训练效率并减少对真实样本的需求.

重要性权重是迁移学习领域中一种通用方法[143,144], 该方法通过估计目标域样本与源域样本的似然比对源域样本进行加权, 将加权后的源域样本用于目标域的训练中. 针对强化学习的应用场景, 设训练样本为 $\tau_i=(s_i,a_i,s_i',r_i)$, 则源域 M_S 中的样本 τ_i 向目标域 M_T 迁移后的重要性权重应设置为 $w_i=\dfrac{p(\tau_i|M_T)}{p(\tau_i|M_S)}$, 其中 $p(\tau_i|M)=\mathcal{P}(s_i'|s_i,a_i)R(r_i|s_i,a_i)$, 重要性权重由动力学模型 $\mathcal{P}(s_i'|s_i,a_i)$ 和奖励模型 $R(r_i|s_i,a_i)$ 共同决定. Lazaric 等人[145]计算重要性权重时首先基于奖励模型与动力学模型计算不同域之间整体样本的相关性度量, 从整体上度量样本迁移的契合度, 然后计算样本间个体的相似性度量, 用以反映相同域中不同样本的独特性, 将两者结合得到最终的重要性权重. Tirinzoni 等

人[146]采用高斯过程从样本中估计奖励模型 $R(r_i|s_i,a_i)$ 和动力学模型 $\mathcal{P}(s'_i|s_i,a_i)$, 重要性权重被分为奖励权重 $w_{r,i}=\dfrac{R_T(r_i|s_i,a_i)}{R_S(r_i|s_i,a_i)}$ 与动力学权重 $w_{p,i}=\dfrac{\mathcal{P}_T(s'_i|s_i,a_i)}{\mathcal{P}_S(s'_i|s_i,a_i)}$, 在 Q 值迭代算法中前者用于估计迁移样本在目标域中奖励值, 后者用于调整迁移样本在目标域更新 Q 函数时的影响大小.

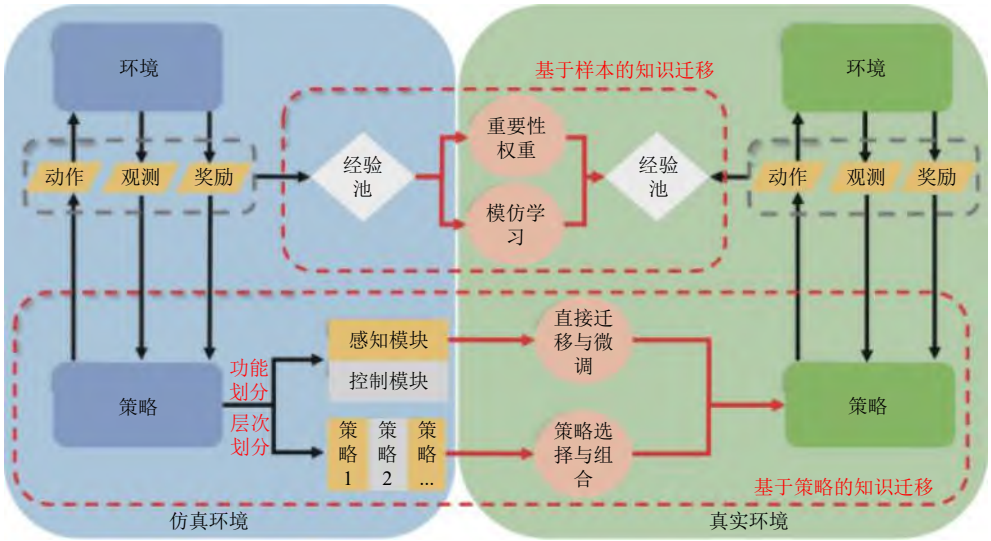


图 5 基于仿真环境的知识迁移方法的技术路线

模仿学习是另一种基于轨迹样本的迁移方法, 通过采用专家演示或轨迹代替人工设计的固定奖励函数来训练智能体, 将专家演示数据作为真实环境智能体的模仿范例来训练智能体的策略网络逼近演示效果, 进而大幅缩短训练时间. 常用的专家演示数据包括人类演示和仿真环境生成的轨迹样本. 其方法大致可分为两类: 行为克隆, 即智能体学习从观测到行为演示的映射[147,148]; 逆强化学习, 即智能体尝试估计专家策略的奖励函数[149]. 模仿学习为智能体提供关于奖励函数的环境信息, 可被用于鲁棒的强化学习以及虚实迁移[150]. Christiano 等人[151]假设仿真环境产生的状态轨迹在真实环境中出现的概率总不为 0, 但是相同的状态轨迹在仿真与真实中对应的动作序列可能不同, 基于上述假设训练真实环境中的逆动力学模型 ϕ, 该模型基于时刻 H 到 $H+k$ 的历史观测 $(o_H,o_{H+1},\ldots,o_{H+k})$ 以及下一时刻的观测 o_{H+k+1} 预测当前时刻真实环境做出的动作 a_{H+k}, 即 $\phi:(o_H,o_{H+1},o_{H+k},o_{H+k+1})\rightarrow a_{H+k}$, 如果学习到的逆动力学模型足够准确, 那么仿真环境中采取行动后的下一个观测 o_{H+K+1} 将与 o_{H+k} 相似, 如图 6 所示.

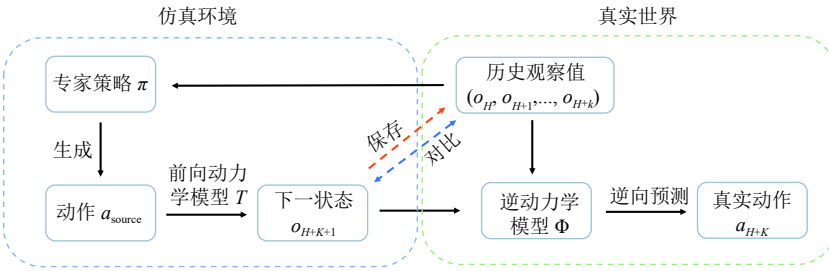


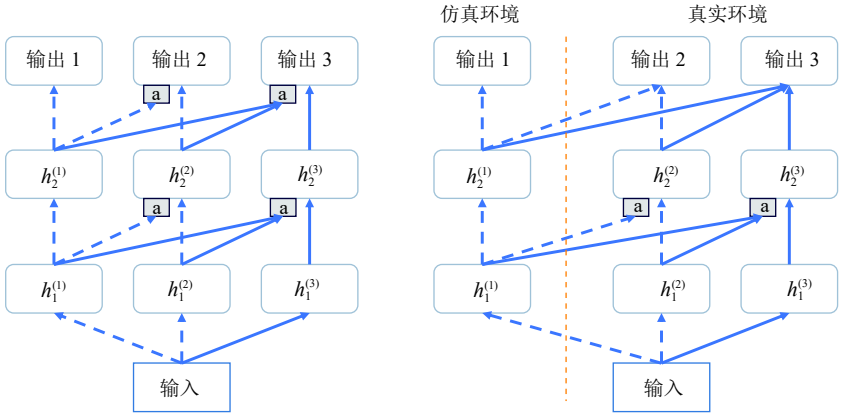
图 6 逆动力学模型结构

仿真环境训练得到的策略被视作专家策略, 用以输出预期的高奖励值的状态轨迹序列, 真实策略则通过逆动力学模型 ϕ 求解对应的动作序列并执行. 另外, 仿真轨迹样本还被用于数据收集, 例如, Rahmatizadeh 等人[152]从机器人训练的安全和成本出发, 在真实环境的示例数据难以被直接采集的场景下选择在虚拟环境中进行人工演示, 得到的仿真轨迹作为专家轨迹用于训练真实机器人的控制策略. Kaushik 等人[153]提出了 SafeAPT 算法, 充分利用

仿真环境数据作为真实世界的先验知识, 从真实世界的观察中迭代学习奖励概率和安全模型, 再利用奖励模型进行贝叶斯优化, 在保持指定安全约束的同时, 利用安全模型从集合中选择最优策略. Lai 等人[154]提出了一种两阶段训练框架 TERT, 该框架由离线预训练阶段和在线校正阶段组成. 在第 1 阶段的离线预训练中, 教师策略与仿真环境交互并收集训练轨迹, 然后训练 Transformer 根据教师的观察动作序列给出相对准确的预测; 在第 2 阶段的在线校正中, Transformer 策略的与仿真环境进行交互, 与此同时, 教师给出相应的参考动作. 然后将教师的观察动作序列作为条件, 训练 Transformer 策略以适应教师的动作. 该模型将 Transformer 与特权训练相结合, 使四足机器人能在不同地形上稳定运动.

3.2.2 仿真策略分解与部件迁移

针对仿真策略的迁移, 最常见的方式是直接地将仿真训练得到的整个策略网络迁移到真实环境中, 根据真实样本对策略进行微调. 但这种方式可能导致模型无法收敛, 或者遗忘仿真环境学到的知识. 基于这一思想, 许多工作通过功能划分或者层次划分将仿真策略进行分解并对其中的部件进行迁移. 如图 5 所示, 当按功能划分时, 策略网络被分解为感知模块与控制模块, 前者对观测输入进行高效的编码与特征提取, 被直接的迁移与微调, 后者通过提取的特征实现策略控制并输出动作, 通常需要重新设置与训练. Rusu 等人[72]提出渐进式网络, 在保留源域学习到的网络的基础上在目标域上训练新网络, 通过侧边连接将旧网络输出的各层特征作为新网络对应层输入的一部分, 新任务 k 的第 i 层特征 $h_i^{(k)}$ 由新任务策略网络的上一层特征 $h_{i-1}^{(k)}$ 以及旧任务 j 对应的策略网络的上一层特征 $h_{i-1}^{(j)}$ 共同决定, 即 $h_i^{(k)}=f\left(W_i^{(k)}h_{i-1}^{(k)}+\sum_{j<k}U_i^{(k:j)}h_{i-1}^{(j)}\right)$, 其中 W 和 U 分别为新网络权重以及侧边连接权重. Rusu 等人[155]将上述的渐进式网络运用到机器人的迁移任务上, 冻结仿真策略的参数并训练真实策略, 在真实策略对应的下一层网络中输入仿真策略输出的每一层状态表示, 从而避免迁移后出现知识遗忘问题 (如图 7 所示), 并在机械臂抓取任务上的实验说明该方法能很好地提高样本效率与策略表现. Kang 等人[156]针对飞行器的视觉任务, 提出了一种将大量模拟数据与少量真实世界经验相结合的方法. 飞行器在真实世界学习机器的物理特性及其动力学, 同时从仿真环境中学习视觉不变量和模式. 迁移时将仿真环境中近似 Q 函数的神经网络按层划分为视觉感知模型以及基于动作的奖励预测模型, 其中视觉感知模型用以处理逼真和多样的视觉场景, 被直接迁移至现实世界, 并在现实世界策略训练期间保持这些感知层参数不变, 防止策略过拟合; 奖励预测模型在仿真环境的训练任务与真实世界机器人任务一致, 被迁移至真实世界后再利用真实世界数据进行训练优化.



(a) 原始渐进网格结构[72] (b) 机器人迁移学习的渐进网格结构[155]

图 7 渐进式网络及其在机器人迁移任务中的应用

当进行层次划分时, 策略的结构按语义层次进行划分, 得到层次不同的若干独立策略, 在进行选择与组合后得到真实环境中的控制策略. Wulfmeier 等人[157]提出的分层结构化策略包含了一个高层策略与若干低层的策略, 在一组任务上训练时, 高层策略在任务已知的条件下选择低层策略, 低层策略则在任务未知的条件下做出决策. 前者

获取任务相关的高层语义信息用以控制低层策略的切换, 后者捕获任务通用的底层语义信息用以特征提取与策略控制, 迁移时保留低层策略重新训练高层策略. 此外, 仿真策略还可以按行为模式进行划分与迁移, Yu 等人[158]结合域随机化技术, 在仿真环境中学习一组高期望回报且行为模式不同的策略, 根据策略在目标任务上的表现在策略组中搜索最佳策略组合进行迁移.

3.3 基于虚实环境的策略迭代提升方法

由于真实环境的复杂性以及静态探索策略的局限性, 离线探索往往无法获得真实环境的全部信息, 需要动态地改进策略以便更好地探索环境, 与此同时, 策略的进一步提升又对环境信息提出了更高的要求. 基于虚实环境的策略迭代提升方法强调了模型优化与知识迁移的循环迭代, 其中基于真实环境的仿真模型优化、仿真环境中策略提升、基于仿真环境的知识迁移以及真实环境下探索与策略评估 4 个步骤不断地迭代执行, 在此过程中仿真环境与真实环境的差异不断减小, 真实策略的表现也不断提升. 根据不同的优化模板, 该类方法又可分为两类: 环境的在线探索与对齐方法以及基于轨迹分布的域自适应方法. 图 8 展示了这两类优化方法的技术路线, 前者以缩小环境差异为目标进行迭代优化, 包括环境对齐以及探索策略的迁移; 后者则从拉近仿真与真实的轨迹分布出发进行迭代优化, 包括轨迹分布的近似、分布对齐以及奖励补偿等机制.

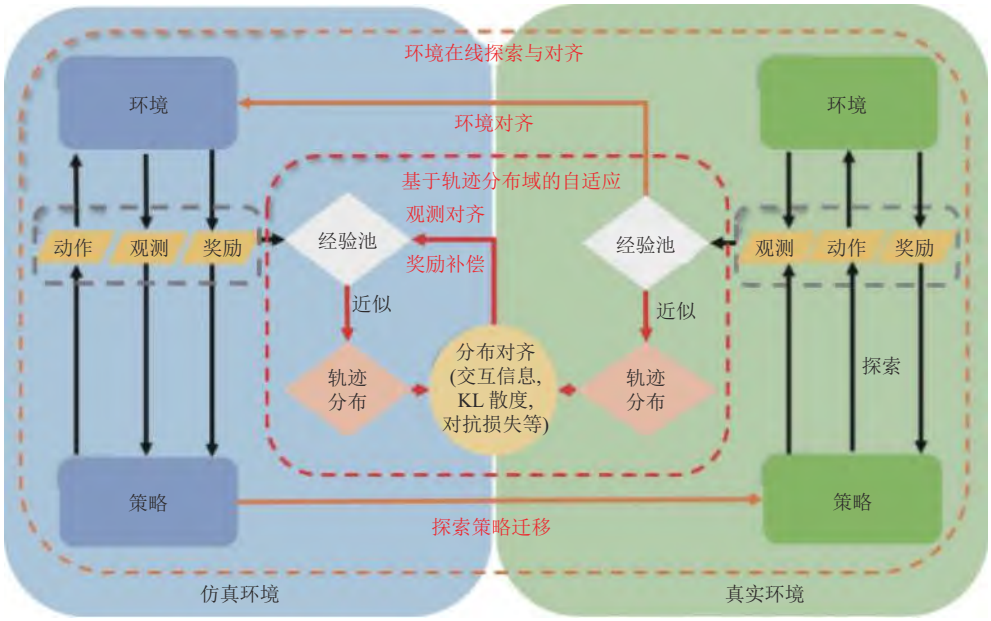


图 8 基于虚实环境的策略迭代提升方法技术路线

3.3.1 环境在线探索与对齐方法

缩小仿真环境和真实环境的差异是确保迁移策略有效性的重要方法, 但通过固定策略的离线探索难以实现这个目标, 所以在策略学习过程中需要对真实环境进行持续探索, 在此基础上对训练环境进行在线对齐, 即, 仿真环境训练得到的策略用于真实环境的探索, 收集的样本用于仿真环境中动力学模型与奖励模型的修正, 反复迭代直至策略的表现不再提升.

在 Farchy 等人[159]于 2013 年提出的 GSL (grounded simulation learning) 框架中, 策略在仿真环境中训练并在真实世界中采样与评估, 从中获取真实动力学模型的信息. GSL 针对机器人行走任务, 通过真实样本建立输入为关节状态和关节命令, 输出为下一关节状态的映射, 并利用该映射调整仿真环境中的力矩参数使仿真机器人到达与真实机器人相同的目标关节状态, 修正后的动力学模型被用于重新学习控制策略, 新策略被继续用于样本收集.

与 GSL[159]中调整环境参数的对齐机制不同, Hanna 等人提出的 GAT (grounded action transformation)[160]不直接对仿真环境的动力学模型进行修正, 而是在控制策略后接入了一个动作转化层, 通过监督方法学习真实环境的动力学模型建立从 (s_t, a_t) 到 s_{t+1} 的映射 f, 并学习仿真环境的逆动力学过程以建立从 (s_t, s_{t+1}) 到 a_t 的映射 f_{sim}^{-1}. 动作转化层输出的动作为 $\hat{a}_t = f_{sim}^{-1}(s_t, f(s_t, a_t))$, 使得在仿真环境中执行动作 $\hat{a}_t$ 后的状态与在真实世界中执行动作 a_t 后的状态一致, 进而缩小动力学模型的差异, 如图 9 所示. 一系列建立在 GAT 基础上的工作[161−163]将适用场景拓展到随机环境, 通过深度学习建立动力模型并使用强化学习进行策略学习, 提高了方法的泛化性与迁移表现.

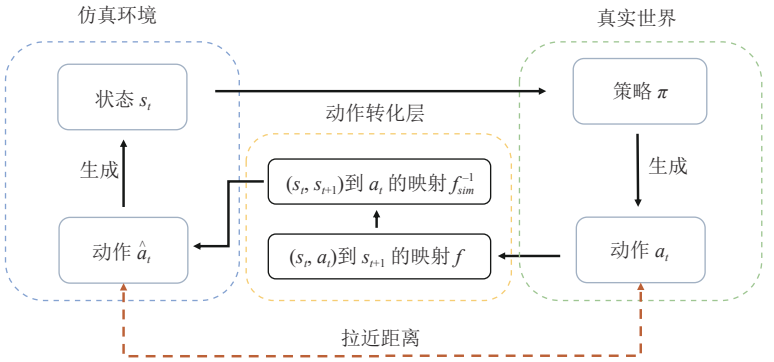


图 9 可修改模拟器 GAT 结构框图

除了直接学习动力学模型, 另一种常见的环境对齐方式是对真实环境与仿真环境的动力学差异进行建模[103,164−166]. 为降低迁移学习中对真实样本的巨大需求以及对硬件的严格要求, Ha 等人[164]以及 Rastogi 等人[103]都对仿真环境和真实环境的动力学差异进行建模与学习, 建立预测转移状态差值 $d(s_t, a_t) = s_{t+1}^{\text{sim}} - s_{t+1}^{\text{real}}$ 的模型. 在策略优化与环境探索的迭代中, 新策略在该模型修正的状态转移样本上进行训练并用于真实环境的探索. Liu 等人[167]提出了一种数字孪生系统, 将真实世界采取到的数据作为孪生数据传输到仿真环境中, 同时将仿真环境中的数据提取和计算传递给真实环境中的物理实体, 使虚拟机器人的运动也能影响物理实体, 从而保证物理实体和虚拟实体的运动过程同步. Abeyruwan 等人[168]设计了虚实迭代模型 i-Sim2Real, 将仿真环境中训练的初始机器人策略应用于现实世界进行人机交互, 收集交互数据的同时在现实世界微调此策略, 再将交互数据用于更新仿真环境中使用的人类行为模型, 最后将微调后的策略重新部署在仿真环境中, 与更新后的人类行为模型进行交互, 虚实迭代训练出与人类球员合作打乒乓球的机械臂.

与关注动力学模型不同, 另一类工作侧重于奖励模型的探索与对齐[169,170]. Laroche 等人[169]假设源域和目标域的动力学模型相同, 通过源域样本与目标域奖励函数直接学习目标域策略, 在目标域奖励函数未知的情况下, 根据不确定性乐观原则探索目标域内未知的状态. Barreto 等人[170]针对奖励函数发生变化但环境动态保持不变的场景提出了一个探索与学习框架, 在一系列导航任务和仿真机械臂的控制中实验并取得了不错的效果. 为了解决在仿真环境中难以准确地渲染真实世界场景的问题, Chen 等人[171]提出了 Real2Sim 方法, 通过仿真环境产生的虚拟图像和真实世界的图像的相互转换, 缩小现实和仿真环境图像的差异.

3.3.2 基于轨迹分布的域自适应方法

基于轨迹分布的域自适应方法利用与环境交互产生的样本估计轨迹分布, 并通过互信息、KL 散度、对抗损失等指标[101,172−174]将缩小源域轨迹分布和目标域轨迹分布的差异作为优化目标. 进而实现对策略的优化与迁移. 在此过程中, 策略的更新优化会导致轨迹分布的改变, 需要重新估计轨迹分布并计算分布差异, 因此, 轨迹估计与策略优化的过程将在源域与目标域之间不断迭代, 进而提高策略的性能以及迁移的有效性.

缩小轨迹分布差异的一种简单且直观的方法是在训练时对与目标域轨迹有较大差异的源域轨迹样本在奖励函数上给与额外的惩罚, 进而抑制仿真策略产生在真实世界中出现概率较低的轨迹. Eysenbach 等人[173]推导得到,

最小化源域轨迹和目标域轨迹的 KL 散度等价于最大化熵正则化的回报加上奖励补偿, 即 $\min\limits_{\pi} D_{\mathrm{KL}}(p_{\mathrm{source}}(\tau)\|p_{\mathrm{target}}(\tau))$ 等价于:

$$\max_{\pi}\mathbb{E}_{p_{\mathrm{source}}}\left[\sum_{t} r(s_t,a_t)+H_{\pi}[a_t|s_t]+\Delta r(s_{t+1},s_t,a_t)\right]+c \tag{2}$$

其中, $\Delta r(s_{t+1},s_t,a_t)\triangleq \log p_{\mathrm{target}}(s_{t+1}|s_t,a_t)-\log p_{\mathrm{source}}(s_{t+1}|s_t,a_t)$, 而 $\sum\limits_{t} r(s_t,a_t)+H_{\pi}[a_t|s_t]$ 合称熵正则化的回报, 是传统的强化学习最大熵算法的优化目标. 另外, 算法还学习两个区分源域轨迹和目标域轨迹的辅助分类器, 以此对奖励补偿项 Δr 进行估计. Wulfmeier 等人[174]提出相互对齐转移学习模型, 该模型通过学习区分仿真轨迹状态序列和真实轨迹状态序列的辅助分类器来估计轨迹分布的差异, 这一差异被作为额外的奖励函数, 同时作用于仿真机器人和真实机器人训练中, 并指导对现实世界的探索. 如图 10 所示, 分别在仿真环境和真实环境训练策略 π_{sim} 和 π_{real}, 产生相应轨迹, 利用辅助分类器 D_w 判别轨迹以得到辅助奖励值 r', 其中 $r'_{sim}=-\log(D_w(\xi_k))$ 而 $r'_{real}=\log(D_w(\xi_k))$. 辅助奖励与环境反馈一起组成用以调整策略的完整奖励 $r=r_t+\lambda r'$. 再利用 TRPO 算法根据 $\mathbb{E}_{\pi_\theta}[\nabla_\theta \log\pi_\theta(a_t|s_t)r_t]$ 分别调整仿真环境和真实世界的策略. Wu 等人[175]同样将基于奖励的自适应方法应用于仿真推荐系统中, 用以缩小动态差异, 同时将对抗学习与奇异值分解 (SVD) 结合用于观察的表示学习. Chung 等人[176]针对自动驾驶任务提出了分段编码虚实迁移算法 SESR, 通过多类分割网络将仿真环境和真实环境中分割后的 RGB 图像转移到同一个域, 最小化编码器和并行网络 VAE 所产生向量的 KL 散度, 同时保留初级对象如行人、道路和汽车的必要信息.

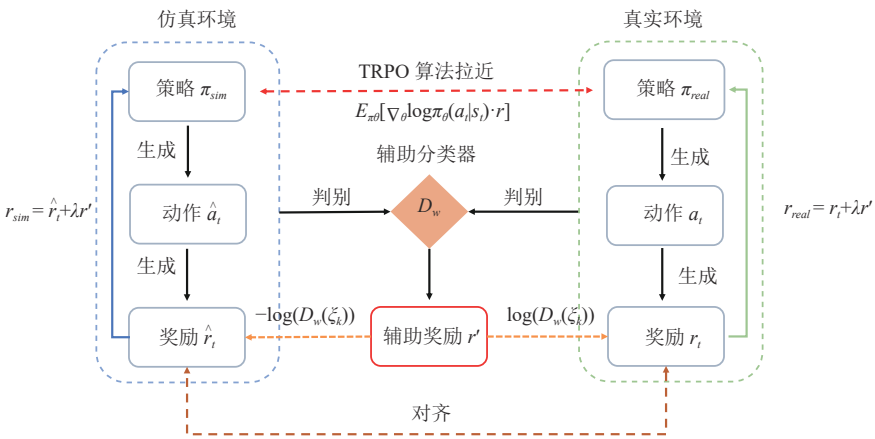


图 10 相互对齐转移学习结构图

另一种常见的迁移方法是基于观测分布的域自适应方法[38−40]. 可以证明在源域和目标域状态转移概率相同的假设下, 基于观测的域自适应是基于轨迹分布的域自适应的一个特例[173]. Gamrian 等人[177]提出基于对抗思想建立源域到目标域的视觉映射, 目标是使经过视觉映射后的图片难以被分类器区分来源. Bousmalis 等人[178]针对基于像素图片的机器人抓取任务, 结合合成图片与域随机化技术, 训练了用于视觉观测自适应对齐的对抗生成网络, 实验证明其方法在实物机器人上有效并降低了真实样本的需求量. 从轨迹分布差异出发, Chebotar 等人[101]提出基于仿真轨迹和真实轨迹中观测状态序列的度量距离, 并利用域随机化方法调整仿真环境参数的随机分布. Truong 等人[179]提出一种双向的域自适应方法, 对仿真中的状态以及真实环境的观察输入采用域自适应方法弥合差异.

3.4 总 结

以上从基于真实环境的仿真模型优化、基于仿真环境的知识迁移、基于虚实环境的策略迭代提升 3 个方向

总结了现有的虚实迁移方法, 并对虚实迁移强化学习研究领域建立了一个统一的通用系统框架. 表 2 从实验的角度总结了上述 3 个方向代表性工作的内容与成果, 罗列了实验的仿真环境、真实环境及实验任务和结果等信息, 并对各个方向现存挑战与未来可能的研究方向进行概述.

表 2 代表性工作汇总

类别	文献	实验任务	策略优化	仿真环境	真实机器人	实验结果概况	困难挑战	未来方向
真实到仿真的环境优化	Allevato 等人[113]	1. 建立真实机械臂与弹跳球对应仿真环境 2. 机械臂弹跳击球(将球从斜面上弹回篮筐)虚实迁移	—	PyBullet	Kinova Jaco 7DOF robot arm	系统识别得到仿真环境与真实环境参数误差小, 弹跳击球任务的成功率达到87%	1. 环境建模依赖较强的先验知识或专家经验 2. 模型拟合需要大量真实样本并存在较大的估计误差	1. 设计自主的环境建模方法, 从多个环境因素中自动识别关键因素 2. 研究计算高效、建模精确以及高样本效率的模拟器和仿真环境
	Peng 等人[125]	机械臂抓取	RDPG	MuJoCo	7DOF Fetch Robotics arm	仿真与真实环境中机械臂抓取成功分别达到91%与89%		
	Cutler 等人[42]	机器人小车的受控漂移	RBF network	—	RC car	仿真策略控制真实小车获得漂移轨迹达到预期效果		
仿真到真实的知识迁移	Rusu 等人[155]	基于图像输入的机械臂抓取	A3C	MuJoCo	Jaco arm	方法有效, 与直接迁移微调相比平均奖励更高	1. 仿真样本与真实样本间存在的分布偏移导致策略性能低下等问题 2. 仿真策略的分解与迁移需要较强的先验假设以及启发式设计	1. 引入重要性采样、数据增强、域自适应等方法减少不同环境样本间分布偏移 2. 进一步研究对抗学习、域自适应等方法, 提出更灵活通用的策略分解和迁移框架
	Kang 等人[156]	飞行器避障	deep Q-learning	Gibson simulator	Crazyflie 2.0 nano quadcopter	以避障时间作为指标, 提出的模型优于直接迁移微调、监督迁移与无监督, 避障时间长达85.8 s, 在复杂场景中也能有效避障		
	Christiano 等人[151]	机械臂抓取	TRPO	MuJoCo	Physical Fetch robot	其策略在环境参数的干扰下具有鲁棒性, 真实轨迹和仿真轨迹误差在3.7%以内		
仿真与真实的迭代提升	Farchy 等人[159]	双足机器人行走	CMA-ES	SimSpark	Bipedal robot NAO	与基线策略相比机器人行走速度增加26.7%	1. 在多轮迭代中, 策略性能提升依赖高效探索与安全性之间的权衡 2. 仿真环境需要动态地适应每轮迭代中新收集的真实样本并做出相应调整	1. 将安全强化学习引入策略学习中, 实现对真实环境高效探索的同时保证策略的安全性 2. 设计可动态调整的仿真环境模型或模拟器, 以实现对真实环境的精确拟合
	Hanna 等人[160]	双足机器人行走	CMA-ES	SimSpark、Gazebo	Bipedal robot NAO	SimSpark到Gazebo的迁移中行走行走速度增加 34.6% (一次迭代)和 43.3% (二次迭代), Gazebo到真实机器人的迁移中速度增加了37.8%		
	Chebotar 等人[101]	基于深度图的机械臂控制摆锤进洞以及拉开抽屉	基于GPU并行仿真器的PPO[180]	NVIDIA Flex	7-DoF Franka Panda and ABB Yumi robots	在少量真实样本的支持下, 最终进洞率达90%, 并且策略学习到拉开抽屉的规范动作		

表 2 中第 1 类基于真实环境的模型优化方法主要在 PyBullet、MuJoCo 等仿真环境中通过系统识别通过系统识别、域随机化等代表性方法使仿真环境不断逼近现实世界, 通过课程式学习和多保真度仿真合理地设置环境, 提高样本效率与训练速度. 该类方法在机械臂、机器人汽车等真实环境上得到验证, 但仍存在过度依赖先验知识、真实样本需求大、环境建模精度低等问题. 针对这些问题, 可能的解决方案包括研究更通用的环境建模方法以及更精准的仿真模拟器等; 表 2 中第 2 类基于仿真环境的知识迁移方法, 则从另一角度出发, 利用重要性采样、模仿学习、策略分解与部件迁移等方法, 将仿真样本中的环境信息以及仿真策略中控制信息向真实策略迁移, 在器械臂、飞行器等机器上取得了良好效果, 但该类方法面临仿真样本与真实样本间分布偏移、策略分解依赖人类知识等问题, 需要更有效的方法进行样本甄别及策略或部件划分; 表 2 中第 3 类基于虚实环境的策略迭代提升方法将前二者相结合, 通过仿真环境与真实环境的在线对齐、基于轨迹分布的域自适应等方法, 在迭代优化中不断减小仿真环境与真实环境差异, 从而实现真实策略表现的不断提升, 已能够较好应用于真实机械臂、双足机器人等的控制行为的学习训练. 该类方法中存在包括安全探索、环境模型动态适应等问题. 针对这些问题的一些解决思路包括将约束强化学习引入策略学习以及设计可动态调整的仿真环境模型或模拟器等.

4 挑战与展望

尽管当前虚实迁移强化学习领域研究取得了一些进展, 但仍有许多问题未能得到有效解决, 本节梳理当下所面临的挑战及未来具有潜力的研究方向.

4.1 挑　战

4.1.1 缺乏理论分析

当前已有方法的迁移效果通常仅通过有限、理想化的实验进行验证, 存在严格的前提假设或是缺少相应的理论分析, 导致算法在实际应用中存在局限性以及盲目性, 算法精确性、稳定性和灵活性得不到理论保证. 针对这一问题, 一些工作尝试对迁移学习过程的样本复杂性等算法性能进行了理论分析, 例如 Cutler 等人[41]在建立多保真度学习模型时, 利用 KWIK 探索框架对基于模型学习的样本复杂性进行了深入分析; Modi 等人[181]提出了一种基于样本有效模型且具有多项式样本复杂性保证的算法; Du 等人[182]从样本复杂性角度提出了设计高效强化学习算法的必要条件等, 以确保在虚实迁移场景中现实可行. 但这些工作大多基于某些特定的任务场景或者算法类型, 而且仅考虑样本复杂度这一单一维度, 如何从样本的重要性、多样性等角度进一步开展相关的理论研究和论证, 是当前虚实迁移学习领域的一大挑战性难题.

4.1.2 依赖人工经验

当前, 迁移强化学习方法大多依赖某种程度的人工经验知识, 如域随机化方法中的随机参数与随机分布的选择大多依赖研究人员手工设计, 而选择一个或多个辅助任务作为源任务进行迁移时也通常依赖专家知识, 因此, 最终迁移性能易受研究人员专业知识及偏好等影响, 造成较高的时间与安全成本, 无法完全保证所设计的策略能够有效地迁移至现实世界. 此外, 域随机化类方法在人工设计初期常假设模拟任务和实际任务来自相同的分布. 若真实环境与随机模拟有很大不同, 或在真实任务中遇到意外情况, 脱离人工经验辅助的迁移策略性能可能会在真实环境出现显著下降[183]. 在此情况下, 一个有效的解决途径是依靠真实世界样本来提高迁移策略的初始性能, 如将域随机化与第 3.3.1 节环境在线探索与对齐方法相结合, 通过域随机化在仿真环境建立初始模型, 再将仿真环境训练得到的策略用于真实环境的探索, 收集的真实环境样本用于仿真环境中模型的修正. 如何让模型自主选择合适的源数据来实现知识的自动有效迁移仍是亟待解决的问题.

4.1.3 缺少评估指标

目前虚实迁移方法大多针对特定任务量身研制, 不同的迁移方法之间缺少统一的指标用于评估不同方法的性能、适用程度以及策略鲁棒性等, 导致缺少能够保证迁移有效性的, 且与任务无关的通用方法. 如何设计一些通用指标如一致性、适用性、鲁棒性和安全性指标等, 在保证最终迁移效果的同时实现对不同算法性能的更全面评估, 仍然是一个挑战性难题. 例如, 一致性指标可通过表征或量化性能差距用以评价不同测量方法或观察者对同一研究对象进行实验的结果一致性; 适用性指标则通过衡量迁移任务的难度以及迁移效果, 较客观地评估该迁移方法能否适应当前场

景, 同时还可被直接用于衡量迁移方法的优化目标以缩小现实差距, 有助于提出更加有效的解决方案; 而鲁棒性指标可用于评估迁移策略是否足够健壮, 能否在异常或扰动的条件下保证理想的迁移效果. 此外, 因为虚实迁移涉及复杂过程, 在迁移之前设计相应安全性指标以评估迁移可行性, 可以保障迁移的安全性, 避免迁移意外情况的发生.

4.2 展 望

4.2.1 自动迁移学习和终生迁移学习

自动迁移学习和终身迁移学习[184,185]能较好地解决依赖人工知识这一问题. 自动迁移学习旨在从过去已实施过的一系列迁移学习任务中学习总结经验, 并把这些经验应用到未来可能出现的迁移学习任务中; 终身迁移学习的核心思想是通过不断学习不同任务自适应地使用各种迁移学习技术来提高终身学习的有效性. 以上方法均通过建立学习目标域与源域的联系, 让机器人在新环境下也能自主决定何时迁移、如何迁移和迁移什么, 自动化地设计随机过程, 尽可能少地使用人类知识和干预, 避免研究人员花费过多时间在参数设定、源域选择等工作上. 如 Bou Ammar 等人[184]率先提出了自主高效的跨域终身迁移强化学习框架, 使智能体能够自主在交织的任务域中学习, 迅速在目标域中获得较高性能的策略; Wei 等人[186]提出了自动迁移学习框架 L2T, 利用先前的迁移学习经验, 自动优化源域和目标域之间的迁移目标及迁移方法. 如何设计高效的自动迁移学习模型并与终身迁移学习相结合将是虚实迁移的强化学习领域拥有巨大挖掘潜力的研究方向之一. 此外, 可进一步将其他方法与自动迁移学习和终生迁移学习相结合, 例如将潜在特征因素纳入迁移内容进行分层迁移或引入对抗性鲁棒迁移学习[187]等.

4.2.2 基于元强化学习的迁移学习

在多任务或环境动态变化的机器人应用场景中引入元强化学习[107,188,189], 可通过在多任务上进行训练和优化以快速适应新任务[107], 进而提升迁移学习的泛化能力. 元强化学习可以从有限的真实样本中学习到具有较高泛化性的策略, 对于具有相似的动力学或奖励模型的新任务, 该策略在从新任务中收集的少量样本上进一步训练就能取到不错的表现, 进而提高真实样本利用率. 基于元强化学习的迁移学习[190]可以看作自动迁移学习的一种具象实现, 该方法学习将源域的何种知识迁移至目标域的何处, 进而使智能体具备自动迁移学习的能力. 目前关于元强化学习的研究大多都集中在固定的任务分布上, 仍需要再更广泛的任务分布上对策略进行评估以提高迁移策略的泛化性和鲁棒性, 例如全面设计适用于更大任务范围的元强化学习基准和评估指标[189]等. 在算法层面, 当前算法通常难以适用于元训练任务高度多样化的环境中, 如何在元学习过程中充分利用一些智能化手段和方法在适应多样化环境方面的显著优势, 例如多智能体强化学习算法[189], 是未来值得探索的方向之一. 此外, 元学习严重依赖奖励信号, 这使其在稀疏奖励环境下表现不佳, 如何引入辅助密集奖励等方法克服稀疏奖励困难也将是该领域未来研究的重点之一[191].

4.2.3 对抗强化学习

对抗强化学习的基本思想是通过在一组环境模型而不是单个环境中训练智能体从而增强策略鲁棒性, 从而提升策略的鲁棒性. 与第 3.3 节所述对抗方法类似, 对抗强化学习方法设置一个待训练的智能体和一个对抗者, 通过二者博弈改变目标项目特征的分布, 使其接近于源项目特征的分布, 不断演化出更复杂的模型以提升虚实迁移效果. 在训练过程中, 一些复杂环境的模拟可以转换成对智能体施加额外干扰, 对抗者会对智能体施加干扰以逼近真实环境, 迫使智能体学习高级策略以应对对抗者的干扰. 通过二者联合训练、协同优化, 智能体最终将学会一种最优的、泛化性强的策略. 一些研究已证明可将对抗性学习嵌入到深度网络中, 以学习更多的可转移特征, 从而减少域之间的分布差异[192]. 在迁移过程中, Pinto 等人[193]将建模误差、虚实环境差异等视为系统中的额外干扰, 利用对抗强化学习思想设计对手施加干扰, 演化出鲁棒对抗性强化学习框架 RARL. Fu 等人[194]则提出了一种基于对抗性奖励学习公式且可扩展的逆强化学习算法 AIRL, 其迁移效果在迷宫、蚂蚁奔跑等任务上得到了验证. 尽管对抗性思想可以有效提高迁移策略的鲁棒性, 但多数工作仅停留在仿真环境的不同任务域中进行迁移实验, 在仿真至真实环境下的迁移验证是一个有意义的研究方向.

4.2.4 虚实迁移学习通用实验平台

由于真实环境复杂多变且仿真环境难以对其进行完全模拟, 在验证算法或迁移方法效果时, 为了保证迁移的安全性, 通常先将策略从一个仿真平台迁移至另一仿真平台, 在多个仿真平台验证后再迁移至真实世界机器人, 以免对机器人造成损伤. 图 11(a) 展示了特定条件下 (以固定仿真平台、算法、迁移方法举例) NAO 机器人的虚实

迁移过程, 大致可分为如下步骤: (1) 配置仿真环境 PyBullet 及仿真环境针对机器人的接口 qibullet; (2) 重写 Gym 环境; (3) 实现深度强化学习算法如 DDPG; (4) 通过域随机化方法将训练后的模型先迁移至 Webot 仿真平台观察效果, 避免直接迁移的安全性问题; (5) 通过域随机化方法将训练后的模型迁移至实体机器人.

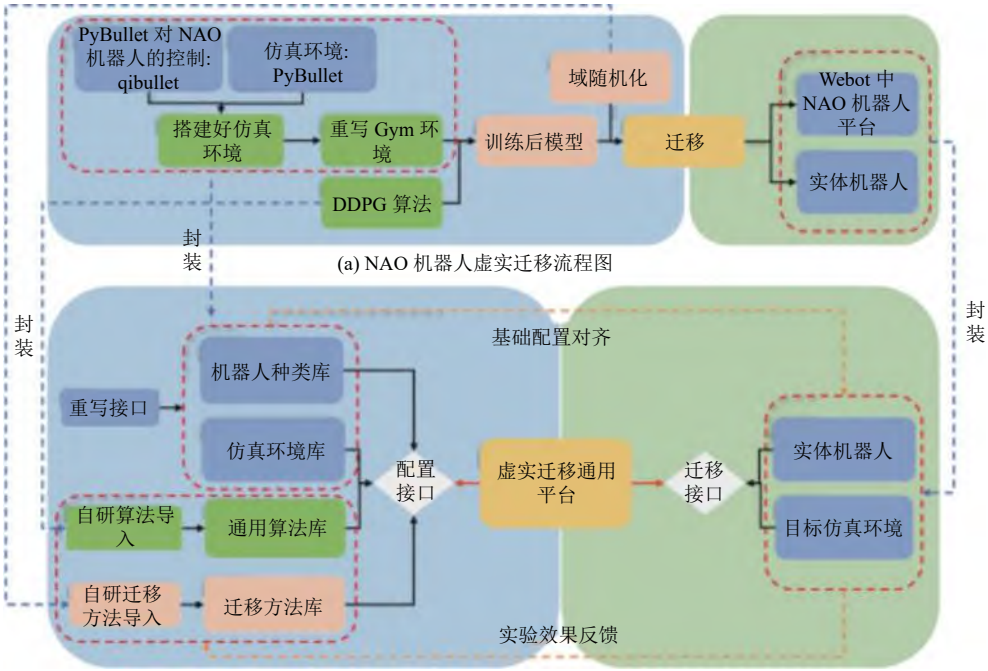


图 11 特定条件下 NAO 机器人的虚实迁移过程

由于机器人种类、仿真环境、强化学习算法及虚实迁移方法具有多样性, 在不同条件下迁移需要手动配置相关环境, 流程较为繁琐复杂, 耗费大量时间精力, 急需一个自主可控、开发便捷、简单易用的虚实迁移通用平台. 尽管迁移学习任务不同, 但整体流程大致可如图 11(b) 所示, 将虚实迁移中涉及的一些功能操作进行封装, 过该平台相应功能接口调用指定库, 包括机器人种类库 (如 NAO 机器人、机械臂等)、仿真环境库 (如 PyBullet、MuJoCo 等)、通用算法库 (如 DDPG、PPO 等) 和迁移方法库 (如域自适应、域随机化、课程学习等). 研究人员只需在该平台调用配置接口完成的环境信息将同步至实体机器人或目标仿真平台, 待模型训练完成后调用迁移接口即可实施迁移. 构建一套虚实迁移的通用实验平台将有助于开发人员专注于迁移学习算法本身的研究, 节约在环境配置方面的时间与精力, 这将大幅缩短科研周期, 提高研究效率.

5 总 结

本文总结了有关于机器人虚实迁移学习的主要工作, 从迁移学习过程中数据信息流动和智能化方法作用对象的角度提出一个虚实迁移的流程框架, 并在此基础上提出了当前虚实迁移学习技术的 3 个主要方向: 基于真实环境的仿真模型优化、基于仿真环境的知识迁移、基于虚实环境的策略迭代提升. 接着, 对每个方向中的主要方法以及相关工作进行阐述, 并对比了一些代表性工作的实验内容. 最后, 介绍虚实迁移学习领域未来面临的挑战, 给出了相应的解决思路与发展方向. 值得说明的是, 除了机器人领域, 虚实迁移强化学习相关思想与技术也可广泛应用于其他领域, 以避免高风险的真实环境交互并提高训练效率. 如在金融交易中, 虚实迁移强化学习可应用于虚拟金融市场环境中的交易策略开发和优化, 以提高交易策略的性能和鲁棒性, 并将其应用于实际市场中进行交易. 另外, 在医疗决策场景中, 虚实迁移强化学习可用于在虚拟环境中训练智能代理来模拟医疗治疗决策, 并将其迁移到

实际临床实践中, 以提高医疗决策的准确性和效果. 我们希望通过对当前相关工作的分类与总结, 为相关研究人员提供一种新的视角解读虚实迁移强化学习领域的研究现状与方向.

References:

[1] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed., Cambridge: MIT Press, 2018.

[2] Li MY, Huang WZ. Research and implementation of Chinese chess game algorithm based on reinforcement learning. In: Proc. of the 5th Int'l Conf. on Control, Robotics and Cybernetics (CRC). Wuhan: IEEE, 2020. 81–86. [doi: 10.1109/CRC51253.2020.9253458]

[3] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen YT, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D. Mastering the game of Go without human knowledge. Nature, 2017, 550(7676): 354–359. [doi: 10.1038/nature24270]

[4] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 2018, 362(6419): 1140–1144. [doi: 10.1126/science.aar6404]

[5] Vinyals O, Babuschkin I, Czarnecki WM, *et al*. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 2019, 575(7782): 350–354. [doi: 10.1038/s41586-019-1724-z]

[6] Berner C, Brockman G, Chan B, *et al*. Dota 2 with large scale deep reinforcement learning. arXiv:1912.06680, 2019.

[7] Chen XC, Yao LN, McAuley J, Zhou GJ, Wang XZ. A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. arXiv:2109.03540, 2021.

[8] Wang L, Zhang W, He XF, Zha HY. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 2447–2456. [doi: 10.1145/3219819.3219961]

[9] Zhao XY, Xia L, Zhang L, Ding ZY, Yin DW, Tang JL. Deep reinforcement learning for page-wise recommendations. In: Proc. of the 12th ACM Conf. on Recommender Systems. Vancouver: ACM, 2018. 95–103. [doi: 10.1145/3240323.3240374]

[10] Osiński B, Jakubowski A, Zięcina P, Miłoś P, Galias C, Homoceanu S, Michalewski H. Simulation-based reinforcement learning for real-world autonomous driving. In: Proc. of the 2020 IEEE Int'l Conf. on Robotics and Automation (ICRA). Paris: IEEE, 2020. 6411–6418. [doi: 10.1109/ICRA40945.2020.9196730]

[11] Pan XL, You YR, Wang ZY, Lu CW. Virtual to real reinforcement learning for autonomous driving. In: Proc. of the 2017 British Machine Vision Conf. London: BMVA Press, 2017.

[12] Hwang KS, Lin JL, Yeh KH. Learning to adjust and refine gait patterns for a biped robot. IEEE Trans. on Systems, Man, and Cybernetics: Systems, 2015, 45(12): 1481–1490. [doi: 10.1109/TSMC.2015.2418321]

[13] García J, Shafie D. Teaching a humanoid robot to walk faster through safe reinforcement learning. Engineering Applications of Artificial Intelligence, 2020, 88: 103360. [doi: 10.1016/j.engappai.2019.103360]

[14] Ding ZH, Tsai YY, Lee WW, Huang BD. Sim-to-real transfer for robotic manipulation with tactile sensory. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems. Prague: IEEE, 2021. 6778–6785. [doi: 10.1109/IROS51168.2021.9636259]

[15] Kalashnikov D, Irpan A, Pastor P, Ibarz J, Herzog A, Jang E, Quillen D, Holly E, Kalakrishnan M, Vanhoucke V, Levine S. Scalable deep reinforcement learning for vision-based robotic manipulation. In: Proc. of the 2nd Conf. on Robot Learning. Zurich: PMLR, 2018. 651–673.

[16] Quillen D, Jang E, Nachum O, Finn C, Ibarz J, Levine S. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 6284–6291. [doi: 10.1109/ICRA.2018.8461039]

[17] Zeng A, Song SR, Welker S, Lee J, Rodriguez A, Funkhouser T. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In: Proc. of the 2018 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 4238–4245. [doi: 10.1109/IROS.2018.8593986]

[18] Lobos-Tsunekawa K, Leiva F, Ruiz-del-Solar J. Visual navigation for biped humanoid robots using deep reinforcement learning. IEEE Robotics and Automation Letters, 2018, 3(4): 3247–3254. [doi: 10.1109/LRA.2018.2851148]

[19] Kahn G, Villaflor A, Ding BS, Abbeel P, Levine S. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 5129–5136. [doi: 10.1109/ICRA.2018.8460655]

[20] Chen CG, Liu YJ, Kreiss S, Alahi A. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In: Proc. of the 2019 IEEE Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 6015–6022. [doi: 10.1109/

ICRA.2019.8794134]

[21] Zhu MX, Wang XS, Wang YH. Human-like autonomous car-following model with deep reinforcement learning. Transportation Research Part C: Emerging Technologies, 2018, 97: 348–368. [doi: 10.1016/j.trc.2018.10.024]

[22] Abreu M, Lau N, Sousa A, Reis LP. Learning low level skills from scratch for humanoid robot soccer using deep reinforcement learning. In: Proc. of the 2019 IEEE Int'l Conf. on Autonomous Robot Systems and Competitions (ICARSC). Porto: IEEE, 2019. 1–8. [doi: 10.1109/ICARSC.2019.8733632]

[23] Shi HB, Lin ZQ, Zhang SG, Li XS, Hwang KS. An adaptive decision-making method with fuzzy Bayesian reinforcement learning for robot soccer. Information Sciences, 2018, 436–437: 268–281. [doi: 10.1016/j.ins.2018.01.032]

[24] Salvato E, Fenu G, Medvet E, Pellegrino FA. Crossing the reality gap: A survey on sim-to-real transferability of robot controllers in reinforcement learning. IEEE Access, 2021, 9: 153171–153187. [doi: 10.1109/Access.2021.3126658]

[25] Ghadirzadeh A, Maki A, Kragic D, Björkman M. Deep predictive policy training using reinforcement learning. In: Proc. of the 2017 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 2351–2358. [doi: 10.1109/IROS.2017.8206046]

[26] Hu H, Zhang KC, Tan AH, Ruan M, Agia CG, Nejat G. A sim-to-real pipeline for deep reinforcement learning for autonomous robot navigation in cluttered rough terrain. IEEE Robotics and Automation Letters, 2021, 6(4): 6569–6576. [doi: 10.1109/LRA.2021.3093551]

[27] Reddy G, Wong-Ng J, Celani A, Sejnowski TJ, Vergassola M. Glider soaring via reinforcement learning in the field. Nature, 2018, 562(7726): 236–239. [doi: 10.1038/s41586-018-0533-0]

[28] Zhu W, Guo X, Owaki D, Kutsuzawa K, Hayashibe M. A survey of sim-to-real transfer techniques applied to reinforcement learning for bioinspired robots. IEEE Trans. on Neural Networks and Learning Systems, 2023, 34(7): 3444–3459. [doi: 10.1109/Tnnls.2021.3112718]

[29] Dimitropoulos K, Hatzilygeroudis I, Chatzilygeroudis K. A brief survey of Sim2Real methods for robot learning. In: Proc. of the 2022 Int'l Conf. on Robotics in Alpe-Adria Danube Region. Klagenfurt: Springer, 2022. 133–140. [doi: 10.1007/978-3-031-04870-8_16]

[30] Zhao WS, Queralta JP, Westerlund T. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In: Proc. of the 2020 IEEE Symp. Series on Computational Intelligence (SSCI). Canberra: IEEE, 2020. 737–744. [doi: 10.1109/SSCI47803.2020.9308468]

[31] Wu TF, Movellan J. Semi-parametric Gaussian process for robot system identification. In: Proc. of the 2012 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vilamoura-Algarve: IEEE, 2012. 725–731. [doi: 10.1109/IROS.2012.6385977]

[32] van Baar J, Sullivan A, Cordorel R, Jha D, Romeres D, Nikovski D. Simulation to real transfer learning with robustified policies for robot tasks. In: Proc. of the 2019 Int'l Conf. on Robotics and Automation (ICRA). IEEE, 2019. 6001–6007

[33] Ye F, Zhang S, Wang P, Chan CY. A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles. In: Proc. of the 2021 IEEE Intelligent Vehicles Symp. (IV). Nagoya: IEEE, 2021. 1073–1080. [doi: 10.1109/IV48863.2021.9575880]

[34] Li ZY, Cheng XX, Peng XB, Abbeel P, Levine S, Berseth G, Sreenath K. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 2811–2817. [doi: 10.1109/ICRA48506.2021.9560769]

[35] Alghonaim R, Johns E. Benchmarking domain randomisation for visual sim-to-real transfer. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 12802–12808. [doi: 10.1109/ICRA48506.2021.9561134]

[36] Siekmann J, Green K, Warila J, Fern A, Hurst JW. Blind bipedal stair traversal via sim-to-real reinforcement learning. In: Robotics: Science and System XVII. 2021. [doi: 10.15607/RSS.2021.XVII.061]

[37] Anderson P, Shrivastava A, Truong J, Majumdar A, Parikh D, Batra D, Lee S. Sim-to-real transfer for vision-and-language navigation. In: Proc. of the 4th Conf. on Robot Learning. Cambridge: PMLR, 2021. 671–681.

[38] Fernando B, Habrard A, Sebban M, Tuytelaars T. Unsupervised visual domain adaptation using subspace alignment. In: Proc. of the 2013 IEEE Int'l Conf. on Computer Vision. Sydney: IEEE, 2013. 2960–2967. [doi: 10.1109/ICCV.2013.368]

[39] Hoffman J, Wang DQ, Yu F, Darrell T. FCNs in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv:1612.02649, 2016.

[40] Zhang DD, Barbot A, Seichepine F, Lo FPW, Bai WJ, Yang GZ, Lo B. Micro-object pose estimation with sim-to-real transfer learning using small dataset. Communications Physics, 2022, 5(1): 80. [doi: 10.1038/S42005-022-00844-Z]

[41] Cutler M, Walsh TJ, How JP. Real-world reinforcement learning via multifidelity simulators. IEEE Trans. on Robotics, 2015, 31(3): 655–671. [doi: 10.1109/tro.2015.2419431]

[42] Cutler M, How JP. Autonomous drifting using simulation-aided reinforcement learning. In: Proc. of the 2016 IEEE Int'l Conf. on Robotics and Automation (ICRA). Stockholm: IEEE, 2016. 5442–5448. [doi: 10.1109/ICRA.2016.7487756]

[43] Di Castro S, Di Castro D, Mannor S. Sim and real: Better together. In: Proc. of the 35th Conf. on Neural Information Processing Systems (NIPS). 2021. 6868–6880.

[44] Wang X, Wang S, Liang XX, Zhao DW, Huang JC, Xu X, Dai B, Miao QG. Deep reinforcement learning: A survey. IEEE Trans. on Neural Networks and Learning Systems, 2022, 1–15. [doi: 10.1109/TNNLS.2022.3207346]

[45] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M. Playing Atari with deep reinforcement learning. arXiv:1312.5602, 2013.

[46] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529–533. [doi: 10.1038/nature14236]

[47] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proc. of the 30th AAAI Conf. on Artificial Intelligence. Phoenix: AAAI, 2016. 2094–2100. [doi: 10.1609/aaai.v30i1.10295]

[48] Wang ZY, Schaul T, Hessel M, Van Hasselt H, Lanctot M, De Freitas N. Dueling network architectures for deep reinforcement learning. In: Proc. of the 33rd Int'l Conf. on Machine Learning. New York: JMLR.org, 2016. 1995–2003.

[49] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D. Continuous control with deep reinforcement learning. arXiv:150902971, 2015.

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

[51] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 1861–1870.

[52] Pan SJ, Yang Q. A survey on transfer learning. IEEE Trans. on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

[53] Huang JY, Smola AJ, Gretton A, Borgwardt KM, Schölkopf B. Correcting sample selection bias by unlabeled data. In: Proc. of the 20th Annual Conf. on Neural Information Processing Systems. Vancouver: MIT Press, 2006. 601–608.

[54] Liao XJ, Xue Y, Carin L. Logistic regression with an auxiliary data source. In: Proc. of the 22nd Int'l Conf. on Machine Learning. Bonn: ACM, 2005. 505–512. [doi: 10.1145/1102351.1102415]

[55] Daumé III H. Frustratingly easy domain adaptation. In: Proc. of the 45th Annual Meeting of the Association of Computational Linguistics. Prague: Association for Computational Linguistics, 2007. 256–263.

[56] Dai WY, Xue GR, Yang Q, Yu Y. Co-clustering based classification for out-of-domain documents. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. San Jose: ACM, 2007. 210–219. [doi: 10.1145/1281192.1281218]

[57] Chattopadhyay R, Sun Q, Fan W, Davidson I, Panchanathan S, Ye JP. Multisource domain adaptation and its application to early detection of fatigue. ACM Trans. on Knowledge Discovery from Data, 2012, 6(4): 18. [doi: 10.1145/2382577.2382582]

[58] Zhuang FZ, Luo P, Xiong H, He Q, Xiong YH, Shi ZZ. Exploiting associations between word clusters and document classes for cross-domain text categorization. Statistical Analysis and Data Mining, 2011, 4(1): 100–114. [doi: 10.1002/sam.10099]

[59] Li FT, Pan SJ, Jin O, Yang Q, Zhu XY. Cross-domain co-extraction of sentiment and topic lexicons. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island: Association for Computational Linguistics, 2012. 410–419.

[60] Mihalkova L, Huynh T, Mooney RJ. Mapping and revising Markov logic networks for transfer learning. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI, 2007. 608–614.

[61] Lazaric A. Transfer in reinforcement learning: A framework and a survey. In: Wiering M, Otterlo M, eds. Reinforcement Learning. Berlin: Springer, 2012. 143–173. [doi: 10.1007/978-3-642-27645-3_5]

[62] Taylor ME, Stone P. Transfer learning for reinforcement learning domains: A survey. The Journal of Machine Learning Research, 2009, 10: 1633–1685.

[63] Zhu ZD, Lin KX, Jain AK, Zhou JY. Transfer learning in deep reinforcement learning: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13344–13362. [doi: 10.1109/TPAMI.2023.3292075]

[64] Erez T, Smart WD. What does shaping mean for computational reinforcement learning? In: Proc. of the 7th IEEE Int'l Conf. on Development and Learning (ICDL). Monterey: IEEE, 2008. 215–219. [doi: 10.1109/DEVLRN.2008.4640832]

[65] Marom O, Rosman B. Belief reward shaping in reinforcement learning. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 3762–3769. [doi: 10.1609/aaai.v32i1.11741]

[66] Zhu ZD, Lin KX, Dai B, Zhou JY. Learning sparse rewarded tasks from sub-optimal demonstrations. arXiv:2004.00530, 2020.

[67] Ho J, Ermon S. Generative adversarial imitation learning. In: Proc. of the 30th Int'l Conf. on Neural Information Processing Systems (NIPS). Barcelona: Curran Associates Inc., 2016. 4572–4580.

[68] Fernández F, Veloso M. Probabilistic policy reuse in a reinforcement learning agent. In: Proc. of the 5th Int'l Joint Conf. on

Autonomous Agents and Multiagent Systems. Hokkaido: ACM, 2006. 720–727. [doi: 10.1145/1160633.1160762]

[69] Czarnecki WM, Pascanu R, Osindero S, Jayakumar S, Swirszcz G, Jaderberg M. Distilling policy distillation. In: Proc. of the 22nd Int'l Conf. on Artificial Intelligence and Statistics. Okinawa: PMLR, 2019. 1331–1340.

[70] Gupta A, Devin C, Liu YX, Abbeel P, Levine S. Learning invariant feature spaces to transfer skills with reinforcement learning. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[71] Devin C, Gupta A, Darrell T, Abbeel P, Levine S. Learning modular neural network policies for multi-task and multi-robot transfer. In: Proc. of the 2017 IEEE Int'l Conf. on Robotics and Automation (ICRA). Singapore: IEEE, 2017. 2169–2176. [doi: 10.1109/ICRA.2017.7989250]

[72] Rusu AA, Rabinowitz NC, Desjardins G, Soyer H, Kirkpatrick J, Kavukcuoglu K, Pascanu R, Hadsell R. Progressive neural networks. arXiv:1606.04671, 2016.

[73] Koenig N, Howard A. Design and use paradigms for Gazebo, an open-source multi-robot simulator. In: Proc. of the 2004 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Sendai: IEEE, 2004. 2149–2154. [doi: 10.1109/IROS.2004.1389727]

[74] Michel O. Cyberbotics Ltd. Webots™: Professional mobile robot simulation. Int'l Journal of Advanced Robotic Systems, 2004, 1(1): 39–42. [doi: 10.5772/5618]

[75] Rohmer E, Singh SPN, Freese M. V-REP: A versatile and scalable robot simulation framework. In: Proc. of the 2013 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Tokyo: IEEE, 2013. 1321–1326. [doi: 10.1109/IROS.2013.6696520]

[76] James S, Freese M, Davison AJ. PyRep: Bringing V-REP to deep robot learning. arXiv:1906.11176, 2019.

[77] James S, Ma ZC, Arrojo DR, Davison AJ. RLBench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters, 2020, 5(2): 3019–3026. [doi: 10.1109/Lra.2020.2974707]

[78] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control. In: Proc. of the 2012 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vilamoura-Algarve: IEEE, 2012. 5026–5033. [doi: 10.1109/IROS.2012.6386109]

[79] Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W. OpenAI Gym. arXiv:1606.01540, 2016.

[80] Tassa Y, Doron Y, Muldal A, Erez T, Li YZ, de Las Casas D, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T, Riedmiller M. DeepMind control suite. arXiv:1801.00690, 2018.

[81] Fan LX, Zhu YK, Zhu JR, Liu ZH, Zeng O, Gupta A, Creus-Costa J, Savarese S, Fei-Fei L. SURREAL: Open-source reinforcement learning framework and robot manipulation benchmark. In: Proc. of the 2nd Conf. on Robot Learning. Zurich: PMLR, 2018. 767–782.

[82] Ahn M, Zhu H, Hartikainen K, Ponte H, Gupta A, Levine S, Kumar V. ROBEL: Robotics benchmarks for learning with low-cost robots. In: Proc. of the 2020 Conf. on Robot Learning. Osaka: PMLR, 2020. 1300–1313.

[83] Coumans E, Bai YF. PyBullet, a Python module for physics simulation for games, robotics and machine learning. 2016. http://pybullet.org

[84] Xia F, Zamir AR, He ZY, Sax A, Malik J, Savarese S. Gibson env: Real-world perception for embodied agents. In: Proc. of the 2018 Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 9068–9079. [doi: 10.1109/CVPR.2018.00945]

[85] Delhaisse B, Rozo L, Caldwell DG. PyRoboLearn: A python framework for robot learning practitioners. In: Proc. of the 2020 Conf. on Robot Learning. Osaka: PMLR, 2020. 1348–1358.

[86] Shah S, Dey D, Lovett C, Kapoor A. AirSim: High-fidelity visual and physical simulation for autonomous vehicles. In: Hutter M, Siegwart R, eds. Field and Service Robotics. Cham: Springer, 2018. 621–635. [doi: 10.1007/978-3-319-67361-5_40]

[87] Dosovitskiy A, Ros G, Codevilla F, Lopez A, Koltun V. CARLA: An open urban driving simulator. In: Proc. of the 1st Conf. on Robot Learning. Mountain View: PMLR, 2017. 1–16.

[88] Freeman CD, Frey E, Raichuk A, Girgin S, Mordatch I, Bachem O. Brax—A differentiable physics engine for large scale rigid body simulation. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021.

[89] Hu YM, Anderson L, Li TM, Sun Q, Carr N, Ragan-Kelley J, Durand F. DiffTaichi: Differentiable programming for physical simulation. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[90] Zhou C, Han L, Mao YZ, Ye G, Lin QJ, Ding WB, Liu H, Wang ZR, Zhang ZY. JBDL: A JAX-based body dynamics algorithm library forrobotics. 2021. https://github.com/Tencent-RoboticsX/jbdl

[91] Hwangbo J, Lee J, Hutter M. Per-contact iteration method for solving contact dynamics. IEEE Robotics and Automation Letters, 2018, 3(2): 895–902. [doi: 10.1109/Lra.2018.2792536]

[92] Makoviychuk V, Wawrzyniak L, Guo YR, Lu M, Storey K, Macklin M, Hoeller D, Rudin N, Allshire A, Handa A, State G. Isaac Gym: High performance GPU-based physics simulation for robot learning. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021.

[93] Boedecker J, Asada M. SimSpark—Concepts and application in the RoboCup 3D soccer simulation league. In: Proc. of the 2008 Int'l

Conf. on Simulation, Modeling and Programming for Autonomous Robot. 2008. 174–181.

[94] Hutter M, Gehring C, Jud D, Lauber A, Bellicoso CD, Tsounis V, Hwangbo J, Bodie K, Fankhauser P, Bloesch M, Diethelm R, Bachmann S, Melzer A, Hoepflinger M. ANYmal—A highly mobile and dynamic quadrupedal robot. In: Proc. of the 2016 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Daejeon: IEEE, 2016. 38–44. [doi: 10.1109/IROS.2016.7758092]

[95] Martin WC, Wu A, Geyer H. Robust spring mass model running for a physical bipedal robot. In: Proc. of the IEEE Int'l Conf. on Robotics and Automation (ICRA). Seattle: IEEE, 2015. 6307–6312. [doi: 10.1109/ICRA.2015.7140085]

[96] Church A, Lloyd J, Lepora NF. Tactile sim-to-real policy transfer via real-to-sim image translation. In: Proc. of the 5th Conf. on Robot Learning. London: PMLR, 2022. 1645–1654.

[97] Lin YJ, Lloyd J, Church A, Lepora NF. Tactile Gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch. IEEE Robotics and Automation Letters, 2022, 7(4): 10754–10761. [doi: 10.1109/LRA.2022.3195195]

[98] Yuan WZ, Dong SY, Adelson EH. GelSight: High-resolution robot tactile sensors for estimating geometry and force. Sensors, 2017, 17(12): 2762. [doi: 10.3390/s17122762]

[99] Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, Necula G, Paszke A, VanderPlas J, Wanderman-Milne S, Zhang Q. JAX: Composable transformations of Python+numpy programs. 2018. https://github.com/google/jax

[100] Höfer S, Bekris K, Handa A, Gamboa JC, Mozifian M, Golemo F, Atkeson C, Fox D, Goldberg K, Leonard J, Karen Liu C, Peters J, Song SR, Welinder P, White M. Sim2Real in robotics and automation: Applications and challenges. IEEE Trans. on Automation Science and Engineering, 2021, 18(2): 398–400. [doi: 10.1109/TASE.2021.3064065]

[101] Chebotar Y, Handa A, Makoviychuk V, Macklin M, Issac J, Ratliff N, Fox D. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In: Proc. of the 2019 IEEE Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 8973–8979. [doi: 10.1109/ICRA.2019.8793789]

[102] Valassakis E, Ding ZH, Johns E. Crossing the gap: A deep dive into zero-shot sim-to-real transfer for dynamics. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 5372–5379. [doi: 10.1109/IROS45743.2020.9341617]

[103] Rastogi D, Koryakovskiy I, Kober J. Sample-efficient reinforcement learning via difference models. In: Proc. of the 2018 Machine Learning in Planning and Control of Robot Motion Workshop at ICRA. 2018.

[104] James S, Wohlhart P, Kalakrishnan M, Kalashnikov D, Irpan A, Ibarz J, Levine S, Hadsell R, Bousmalis K. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12619–12629. [doi: 10.1109/CVPR.2019.01291]

[105] Moldovan TM, Abbeel P. Safe exploration in Markov decision processes. In: Proc. of the 29th Int'l Conf. on Machine Learning. Edinburgh: icml.cc/Omnipress, 2012.

[106] Mannucci T, van Kampen EJ, de Visser C, Chu QP. Safe exploration algorithms for reinforcement learning controllers. IEEE Trans. on Neural Networks and Learning Systems, 2018, 29(4): 1069–1081. [doi: 10.1109/TNNLS.2017.2654539]

[107] Gupta A, Mendonca R, Liu YX, Abbeel P, Levine S. Meta-reinforcement learning of structured exploration strategies. In: Proc. of the 32nd Int'l Conf. on Neural Information Processing Systems. Montreal: Curran Associates Inc., 2018. 5307–5316.

[108] Nair A, McGrew B, Andrychowicz M, Zaremba W, Abbeel P. Overcoming exploration in reinforcement learning with demonstrations. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 6292–6299. [doi: 10.1109/ICRA.2018.8463162]

[109] Kristinsson K, Dumont GA. System identification and control using genetic algorithms. IEEE Trans. on Systems, Man, and Cybernetics, 1992, 22(5): 1033–1046. [doi: 10.1109/21.179842]

[110] Johansson R, Robertsson A, Nilsson K, Verhaegen M. State-space system identification of robot manipulator dynamics. Mechatronics, 2000, 10(3): 403–418. [doi: 10.1016/S0957-4158(99)00049-5]

[111] Akanyeti O, Nehmzow U, Billings SA. Robot training using system identification. Robotics and Autonomous Systems, 2008, 56(12): 1027–1041. [doi: 10.1016/j.robot.2008.09.007]

[112] Al-Dabbagh RD, Kinsheel A, Mekhilef S, Baba MS, Shamshirband S. System identification and control of robot manipulator based on fuzzy adaptive differential evolution algorithm. Advances in Engineering Software, 2014, 78: 60–66. [doi: 10.1016/j.advengsoft.2014.08.009]

[113] Allevato A, Short ES, Pryor M, Thomaz A. TuneNet: One-shot residual tuning for system identification and sim-to-real robot task transfer. In: Proc. of the 2020 Conf. on Robot Learning. Osaka: PMLR, 2020. 445–455.

[114] Du YQ, Watkins O, Darrell T, Abbeel P, Pathak D. Auto-tuned sim-to-real transfer. In: Proc. of the 2021 IEEE Int'l Conf. on Robotics and Automation (ICRA). Xi'an: IEEE, 2021. 1290–1296. [doi: 10.1109/ICRA48506.2021.9562091]

[115] Gao RH, Chang YY, Mall S, Fei-Fei L, Wu JJ. ObjectFolder: A dataset of objects with implicit visual, auditory, and tactile representations. In: Proc. of the 2021 Conf. on Robot Learning. London: PMLR, 2021. 466–476.

[116] Gao RH, Si ZL, Chang YY, Clarke S, Bohg J, Fei-Fei L, Yuan WZ, Wu JJ. ObjectFolder 2.0: A multisensory object dataset for Sim2Real transfer. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10588–10598. [doi: 10.1109/CVPR52688.2022.01034]

[117] Kumar A, Fu ZP, Pathak D, Malik J. RMA: Rapid motor adaptation for legged robots. In: Robotics: Science and Systems XVII. 2021. [doi: 10.15607/RSS.2021.XVII.011]

[118] Kumar A, Li ZY, Zeng J, Pathak D, Sreenath K, Malik J. Adapting rapid motor adaptation for bipedal robots. In: Proc. of the 2022 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Kyoto: IEEE, 2022. 1161–1168. [doi: 10.1109/Iros47612.2022.9981091]

[119] Qi HZ, Kumar A, Calandra R, Ma Y, Malik J. In-hand object rotation via rapid motor adaptation. In: Proc. of the 2022 Conf. on Robot Learning. Auckland: PMLR, 2022. 1722–1732.

[120] Zhang DQ, Loquercio A, Wu XY, Kumar A, Malik J, Mueller MW. Learning a single near-hover position controller for vastly different quadcopters. In: Proc. of the 2022 IEEE Int'l Conf. on Robotics and Automation (ICRA). London: IEEE, 2022. 1263–1269. [doi: 10.1109/ICRA48891.2023.10160836]

[121] Ruderman M, Hoffmann F, Bertram T. Modeling and identification of elastic robot joints with hysteresis and backlash. IEEE Trans. on Industrial Electronics, 2009, 56(10): 3840–3847. [doi: 10.1109/Tie.2009.2015752]

[122] Park HW, Sreenath K, Hurst JW, Grizzle JW. Identification of a bipedal robot with a compliant drivetrain. IEEE Control Systems Magazine, 2011, 31(2): 63–88. [doi: 10.1109/Mcs.2010.939963]

[123] Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In: Proc. of the 2017 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Vancouver: IEEE, 2017. 23–30. [doi: 10.1109/IROS.2017.8202133]

[124] Muratore F, Eilers C, Gienger M, Peters J. Data-efficient domain randomization with Bayesian optimization. IEEE Robotics and Automation Letters, 2021, 6(2): 911–918. [doi: 10.1109/LRA.2021.3052391]

[125] Peng XB, Andrychowicz M, Zaremba W, Abbeel P. Sim-to-real transfer of robotic control with dynamics randomization. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 3803–3810. [doi: 10.1109/ICRA.2018.8460528]

[126] Valassakis E, Di Palo N, Johns E. Coarse-to-fine for sim-to-real: Sub-millimetre precision across wide task spaces. In: Proc. of the 2021 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Prague: IEEE, 2021. 5989–5996. [doi: 10.1109/IROS51168.2021.9636388]

[127] Ren XY, Luo JL, Solowjow E, Ojea JA, Gupta A, Tamar A, Abbeel P. Domain randomization for active pose estimation. In: Proc. of the 2019 IEEE Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 7228–7234. [doi: 10.1109/ICRA.2019.8794126]

[128] Muratore F, Treede F, Gienger M, Peters J. Domain randomization for simulation-based policy optimization with transferability assessment. In: Proc. of the 2nd Conf. on Robot Learning. Zurich: PMLR, 2018. 700–713.

[129] Muratore F, Gruner T, Wiese F, Belousov B, Gienger M, Peters J. Neural posterior domain randomization. In: Proc. of the 5th Conf. on Robot Learning. London: PMLR, 2021. 1532–1542.

[130] Bengio Y, Louradour J, Collobert R, Weston J. Curriculum learning. In: Proc. of the 26th Annual Int'l Conf. on Machine Learning. Montreal: ACM, 2009. 41–48. [doi: 10.1145/1553374.1553380]

[131] Narvekar S, Peng B, Leonetti M, Sinapov J, Taylor ME, Stone P. Curriculum learning for reinforcement learning domains: A framework and survey. The Journal of Machine Learning Research, 2020, 21(1): 181.

[132] Narvekar S, Stone P. Learning curriculum policies for reinforcement learning. In: Proc. of the 18th Int'l Conf. on Autonomous Agents and MultiAgent Systems. Montreal: Int'l Foundation for Autonomous Agents and MultiAgent Systems, 2019. 25–33.

[133] Florensa C, Held D, Wulfmeier M, Zhang M, Abbeel P. Reverse curriculum generation for reinforcement learning. In: Proc. of the 1st Conf. on Robot Learning. Mountain View: PMLR, 2017. 482–495.

[134] Nguyen SM, Duminy N, Manoury A, Duhaut D, Buche C. Robots learn increasingly complex tasks with intrinsic motivation and automatic curriculum learning: Domain knowledge by emergence of affordances, hierarchical reinforcement and active imitation learning. KI—Künstliche Intelligenz, 2021, 35(1): 81–90. [doi: 10.1007/s13218-021-00708-8]

[135] Sukhbaatar S, Lin ZM, Kostrikov I, Synnaeve G, Szlam A, Fergus R. Intrinsic motivation and automatic curricula via asymmetric self-play. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[136] He ZH, Gu CC, Xu R, Wu KJ. Automatic curriculum generation by hierarchical reinforcement learning. In: Proc. of the 27th Int'l Conf. on Neural Information Processing. Bangkok: Springer, 2020. 202–213. [doi: 10.1007/978-3-030-63833-7_17]

[137] Shukla Y, Thierauf C, Hosseini R, Tatiya G, Sinapov J. ACuTE: Automatic curriculum transfer from simple to complex environments. In: Proc. of the 21st Int'l Conf. on Autonomous Agents and MultiAgent Systems. Int'l Foundation for Autonomous Agents and Multiagent Systems, 2022. 1192–1200.

[138] Matiisen T, Oliver A, Cohen T, Schulman J. Teacher-student curriculum learning. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(9): 3732–3740. [doi: 10.1109/TNNLS.2019.2934906]

[139] Deisenroth MP, Rasmussen CE. PILCO: A model-based and data-efficient approach to policy search. In: Proc. of the 28th Int'l Conf. on Machine Learning. Washington: Omnipress, 2011. 465–472.

[140] Truong J, Rudolph M, Yokoyama NH, Chernova S, Batra D, Rai A. Rethinking Sim2Real: Lower fidelity simulation leads to higher Sim2Real transfer in navigation. In: Proc. of the 6th Conf. on Robot Learning. PMLR, 2023. 859–870.

[141] Wang X, Chen YD, Zhu WW. A survey on curriculum learning. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2022, 44(9): 4555–4576. [doi: 10.1109/TPAMI.2021.3069908]

[142] Zhou YW, Chen H, Pan ZR, Yan CH, Lin FQ, Wang X, Zhu WW. CurML: A curriculum machine learning library. In: Proc. of the 30th ACM Int'l Conf. on Multimedia. Lisboa: ACM, 2022. 7359–7363. [doi: 10.1145/3503161.3548549]

[143] Cortes C, Mohri M. Domain adaptation and sample bias correction theory and algorithm for regression. Theoretical Computer Science, 2014, 519: 103–126. [doi: 10.1016/j.tcs.2013.09.027]

[144] Lu N, Zhang TY, Fang TT, Teshima T, Sugiyama M. Rethinking importance weighting for transfer learning. arXiv:2112.10157, 2021.

[145] Lazaric A, Restelli M, Bonarini A. Transfer of samples in batch reinforcement learning. In: Proc. of the 25th Int'l Conf. on Machine Learning. Helsinki: ACM, 2008. 544–551. [doi: 10.1145/1390156.1390225]

[146] Tirinzoni A, Sessa A, Pirotta M, Restelli M. Importance weighted transfer of samples in reinforcement learning. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 4936–4945.

[147] Pomerleau DA. ALVINN: An autonomous land vehicle in a neural network. In: Proc. of the 1st Int'l Conf. on Neural Information Processing Systems (NIPS). Denver: MIT Press, 1988. 305–313.

[148] Ross S, Gordon GJ, Bagnell D. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proc. of the 14th Int'l Conf. on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale: JMLR.org, 2011. 627–635.

[149] Ng AY, Russell S. Algorithms for inverse reinforcement learning. In: Proc. of the 17th Int'l Conf. on Machine Learning. Stanford: Morgan Kaufmann, 2000. 663–670.

[150] Yan MY, Frosio I, Tyree S, Kautz J. Sim-to-real transfer of accurate grasping with eye-in-hand observations and continuous control. arXiv:1712.03303, 2017.

[151] Christiano P, Shah Z, Mordatch I, Schneider J, Blackwell T, Tobin J, Abbeel P, Zaremba W. Transfer from simulation to real world through learning deep inverse dynamics model. arXiv:1610.03518, 2016.

[152] Rahmatizadeh R, Abolghasemi P, Behal A, Bölöni L. From virtual demonstration to real-world manipulation using LSTM and MDN. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. New Orleans: AAAI, 2018. 6524–6531. [doi: 10.1609/aaai.v32i1.12099]

[153] Kaushik R, Arndt K, Kyrki V. SafeAPT: Safe simulation-to-real robot learning using diverse policies learned in simulation. IEEE Robotics and Automation Letters, 2022, 7(3): 6838–6845. [doi: 10.1109/LRA.2022.3177294]

[154] Lai H, Zhang WN, He XL, Yu C, Tian Z, Yu Y, Wang J. Sim-to-real transfer for quadrupedal locomotion via terrain transformer. In: Proc. of the 2023 IEEE Int'l Conf. on Robotics and Automation (ICRA). London: IEEE, 2023. 5141–5147. [doi: 10.1109/ICRA48891.2023.10160497]

[155] Rusu AA, Večerík M, Rothörl T, Heess N, Pascanu R, Hadsell R. Sim-to-real robot learning from pixels with progressive nets. In: Proc. of the 1st Conf. on Robot Learning. Mountain View: PMLR, 2017. 262–270.

[156] Kang K, Belkhale S, Kahn G, Abbeel P, Levine S. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight. In: Proc. of the 2019 IEEE Int'l Conf. on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 6008–6014. [doi: 10.1109/ICRA.2019.8793735]

[157] Wulfmeier M, Abdolmaleki A, Hafner R, Springenberg JT, Neunert M, Hertweck T, Lampe T, Siegel N, Heess N, Riedmiller M. Regularized hierarchical policies for compositional transfer in robotics. arXiv:1906.11228, 2019.

[158] Yu WH, Liu CK, Turk G. Policy transfer with strategy optimization. In: Proc. of the 7th Int'l Conf. on Learning Representations. New Orleans: OpenReview.net, 2019.

[159] Farchy A, Barrett S, MacAlpine P, Stone P. Humanoid robots learning to walk faster: From the real world to simulation and back. In: Proc. of the 2013 Int'l Conf. on Autonomous Agents and Multi-agent Systems (AAMAS). Saint Paul: IFAAMAS, 2013. 39–46.

[160] Hanna J, Stone P. Grounded action transformation for robot learning in simulation. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 3834–3840. [doi: 10.1609/aaai.v31i1.11044]

[161] Desai S, Karnan H, Hanna JP, Warnell G, Stone P. Stochastic grounded action transformation for robot learning in simulation. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 6106–6111. [doi: 10.1109/IROS45743.2020.9340780]

[162] Hanna JP, Desai S, Karnan H, Warnell G, Stone P. Grounded action transformation for sim-to-real reinforcement learning. Machine Learning, 2021, 110(9): 2469–2499. [doi: 10.1007/s10994-021-05982-z]

[163] Karnan H, Desai S, Hanna JP, Warnell G, Stone P. Reinforced grounded action transformation for sim-to-real transfer. In: Proc. of the 2020 IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems (IROS). Las Vegas: IEEE, 2020. 4397–4402. [doi: 10.1109/IROS45743.2020.9341149]

[164] Ha S, Yamane K. Reducing hardware experiments for model learning and policy optimization. In: Proc. of the 2015 IEEE Int'l Conf. on Robotics and Automation (ICRA). Seattle: IEEE, 2015. 2620–2626. [doi: 10.1109/ICRA.2015.7139552]

[165] Golemo F, Taiga AA, Courville A, Oudeyer PY. Sim-to-real transfer with neural-augmented robot simulation. In: Proc. of the 2nd Conf. on Robot Learning. Zurich: PMLR, 2018. 817–828.

[166] Ota K, Jha DK, Romeres D, van Baar J, Smith KA, Semitsu T, Oiki T, Sullivan A, Nikovski D, Tenenbaum JB. Data-efficient learning for complex and real-time physical problem solving using augmented simulation. IEEE Robotics and Automation Letters, 2021, 6(2): 4241–4248. [doi: 10.1109/Lra.2021.3068887]

[167] Liu YK, Xu H, Liu D, Wang LH. A digital twin-based sim-to-real transfer for deep reinforcement learning-enabled industrial robot grasping. Robotics and Computer-integrated Manufacturing, 2022, 78: 102365. [doi: 10.1016/j.rcim.2022.102365]

[168] Abeyruwan SW, Graesser L, D'Ambrosio DB, Singh A, Shankar A, Bewley A, Jain D, Choromanski KM, Sanketi PR. I-Sim2Real: Reinforcement learning of robotic policies in tight human-robot interaction loops. In: Proc. of the 2022 Conf. on Robot Learning. Auckland: PMLR, 2023. 212–224.

[169] Laroche R, Barlier M. Transfer reinforcement learning with shared dynamics. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 2147–2153. [doi: 10.1609/aaai.v31i1.10796]

[170] Barreto A, Dabney W, Munos R, Hunt JJ, Schaul T, van Hasselt H, Silver D. Successor features for transfer in reinforcement learning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4058–4068.

[171] Chen YW, Li X, Guo S, Ng XY, Ang MH. Real2Sim or Sim2Real: Robotics visual insertion using deep reinforcement learning and Real2Sim policy adaptation. In: Petrovic I, Menegatti E, Marković I, eds. Intelligent Autonomous Systems 17. Cham: Springer, 2023. 617–629. [doi: 10.1007/978-3-031-22216-0_41]

[172] Zhu GX, Zhang MH, Lee H, Zhang CJ. Bridging imagination and reality for model-based deep reinforcement learning. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 754.

[173] Eysenbach B, Chaudhari S, Asawa S, Levine S, Salakhutdinov R. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. In: Proc. of the 9th Int'l Conf. on Learning Representations. OpenReview.net, 2021.

[174] Wulfmeier M, Posner I, Abbeel P. Mutual alignment transfer learning. In: Proc. of the 1st Conf. on Robot Learning. Mountain View: PMLR, 2017. 281–290.

[175] Wu JD, Xie ZH, Yu T, Li QZ, Li S. Sim-to-real interactive recommendation via off-dynamics reinforcement learning. In: Proc. of the 35th Conf. on Neural Information Processing Systems. 2021.

[176] Chung SH, Kong SH, Cho S, Nahrendra IMA. Segmented encoding for Sim2Real of RL-based end-to-end autonomous driving. In: Proc. of the 2022 IEEE Intelligent Vehicles Symp. (IV). Aachen: IEEE, 2022. 1290–1296. [doi: 10.1109/IV51971.2022.9827374]

[177] Gamrian S, Goldberg Y. Transfer learning for related reinforcement learning tasks via image-to-image translation. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 2063–2072.

[178] Bousmalis K, Irpan A, Wohlhart P, Bai YF, Kelcey M, Kalakrishnan M, Downs L, Ibarz J, Pastor P, Konolige K, Levine S, Vanhoucke V. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: Proc. of the 2018 IEEE Int'l Conf. on Robotics and Automation (ICRA). Brisbane: IEEE, 2018. 4243–4250. [doi: 10.1109/ICRA.2018.8460875]

[179] Truong J, Chernova S, Batra D. Bi-directional domain adaptation for Sim2Real transfer of embodied navigation agents. IEEE Robotics and Automation Letters, 2021, 6(2): 2634–2641. [doi: 10.1109/Lra.2021.3062303]

[180] Liang J, Makoviychuk V, Handa A, Chentanez N, Macklin M, Fox D. GPU-accelerated robotic simulation for distributed reinforcement learning. In: Proc. of the 2nd Conf. on Robot Learning. Zurich: PMLR, 2018. 270–282.

[181] Modi A, Jiang N, Tewari A, Singh S. Sample complexity of reinforcement learning using linearly combined model ensembles. In: Proc. of the 23rd Int'l Conf. on Artificial Intelligence and Statistics. Palermo: PMLR, 2020. 2010–2020.

[182] Du SS, Kakade SM, Wang RS, Yang LF. Is a good representation sufficient for sample efficient reinforcement learning? In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[183] Ju H, Juan R, Gomez R, Nakamura K, Li GL. Transferring policy of deep reinforcement learning from simulation to reality for robotics. Nature Machine Intelligence, 2022, 4(12): 1077–1087. [doi: 10.1038/s42256-022-00573-6]

[184] Bou Ammar H, Eaton E, Luna JM, Ruvolo P. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. Buenos: AAAI, 2015. 3345–3351.

[185] Sheikhlar A, Thórisson KR, Eberding LM. Autonomous cumulative transfer learning. In: Proc. of the 13th Int'l Conf. on Artificial General Intelligence. St. Petersburg: Springer, 2020. 306–316. [doi: 10.1007/978-3-030-52152-3_32]

[186] Wei Y, Zhang Y, Huang JZ, Yang Q. Transfer learning via learning to transfer. In: Proc. of the 35th Int'l Conf. on Machine Learning. Stockholm: PMLR, 2018. 5085–5094.

[187] Shafahi A, Saadatpanah P, Zhu C, Ghiasi A, Studer C, Jacobs DW, Goldstein T. Adversarially robust transfer learning. In: Proc. of the 8th Int'l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[188] Schweighofer N, Doya K. Meta-learning in reinforcement learning. Neural Networks, 2003, 16(1): 5–9. [doi: 10.1016/s0893-6080(02)00228-9]

[189] Yu TH, Quillen D, He ZP, Julian R, Hausman K, Finn C, Levine S. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Proc. of the 3rd Conf. on Robot Learning. Osaka: PMLR, 2020. 1094–1100.

[190] Jang Y, Lee H, Hwang SJ, Shin J. Learning what and where to transfer. In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 3030–3039.

[191] Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I, Abbeel P. Continuous adaptation via meta-learning in nonstationary and competitive environments. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[192] Yu CH, Wang JD, Chen YQ, Huang MY. Transfer learning with dynamic adversarial adaptation network. In: Proc. of the 2019 IEEE Int'l Conf. on Data Mining (ICDM). Beijing: IEEE, 2019. 778–786. [doi: 10.1109/ICDM.2019.00088]

[193] Pinto L, Davidson J, Sukthankar R, Gupta A. Robust adversarial reinforcement learning. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: PMLR, 2017. 2817–2826.

[194] Fu J, Luo KT, Levine S. Learning robust rewards with adversarial inverse reinforcement learning. arXiv:1710.11248, 2017.

林谦(1999－), 男, 硕士, CCF 学生会员, 主要研究领域为离线强化学习, 安全强化学习, 机器人虚实迁移.

余超(1985－), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为强化学习, 智能机器人, 博弈智能.

伍夏威(2001－), 男, 硕士, 主要研究领域为强化学习, 机器人控制, 机器人虚实迁移.

董银昭(1995－), 男, 博士, 主要研究领域为强化学习, 四足机器人, 触觉感知.

徐昕(1974－), 男, 博士, 教授, 主要研究领域为机器学习, 智能无人系统, 智能控制.

张强(1971－), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为机器行为与人机协同, 生物计算.

郭宪(1985－), 男, 博士, 副教授, 博士生导师, CCF 专业会员, 主要研究领域为强化学习, 多智能体技术、博弈论等在机器人领域中的研究和应用.