利用迁移学习促进过程控制中的强化学习：概述与展望

1st Runze Lin
College of Control Science and Engineering
Zhejiang University
Hangzhou, China
rzlin@zju.edu.cn

2nd Junghui Chen
Department of Chemical Engineering
Chung-Yuan Christian University
Taoyuan, Taiwan
jason@wavenet.cycu.edu.tw

3rd Lei Xie
College of Control Science and Engineering
Zhejiang University
Hangzhou, China
leix@iipc.zju.edu.cn

4th Hongye Su
College of Control Science and Engineering
Zhejiang University
Hangzhou, China
hysu@iipc.zju.edu.cn

arxiv:2404.00247v3 中译本

摘要—在工业 4.0 和智能制造的背景下，过程工业优化与控制领域也正在经历数字化转型。随着深度强化学习（DRL）的兴起，其在过程控制中的应用引起了广泛关注。然而，由于探索过程中导致的极低样本效率和安全问题，阻碍了它在工业环境中的实际实施。迁移学习为 DRL 提供了一个有效的解决方案，提升了其在多模式控制场景中的泛化能力和适应性。本文从迁移学习的角度提供了关于将 DRL 用于过程控制的一些见解。我们分析了在过程工业中应用 DRL 的挑战以及引入迁移学习的必要性。此外，还为未来的研究方向提出了建议和展望，探讨如何将迁移学习与 DRL 相结合以提升过程控制。本文旨在为过程工业领域的学者和工程师提供一套有前景的、用户友好的、易于实施且可扩展的人工智能辅助工业控制方法。

***Index Terms*—深度强化学习，工业控制，迁移学习，流程工业**

I. 介绍

IN 在工业 4.0 和智能制造的领域，智能流程在过程控制领域变得越来越重要。科学与工程中人工智能的兴起为现代工艺行业提供了一种新的深度强化学习（DRL）范式，增强了并推动了这些行业的发展。近年来，工艺行业中越来越多地关注 DRL 与过程控制 [1]–[3] 的交叉点。对于连续和批次流程 [4]–[7]，已经开发出了相应的算法改进和应用研究。

利用 DRL 进行过程控制代表着实现智能制造和提高生产自动化水平的一个有希望的方向。然而，与金融、推荐和机器人等领域相比，流程工业对安全性和稳定性的要求更高。这种更高的要求是 DRL 在工业环境中广泛适用面临重大挑战的重要原因。值得注意的是，一些学者已经将迁移学习的概念引入到过程控制的 DRL 中，以增强其安全性和实用性 [8]–[10]。迁移学习的整合有利于过程控制，为工业生产提供了更合理和实用的解决方案 [11]。

不幸的是，在过程控制领域，关于如何将迁移学习整合到深度强化学习中缺乏系统性的分析和综合，特别是在过程系统工程领域的值得探索的研究方向方面。

为了填补这一空白，本文提出了关于利用迁移学习促进过程控制中的强化学习的见解和前景，旨在为本领域的研究人员提供有价值的参考点。

II. 强化学习在过程控制中的应用

强化学习（RL）在学术界和工业界都获得了显著的关注，最近的研究突出了其潜力。特别是深度强化学习（DRL），在 AlphaGo 战胜围棋人类选手 [12] 之后引起了广泛关注。DRL 专注于解决不确定环境中的顺序决策问题 [13]，使其特别适用于过程控制，在过程中不确定性往往起着关键作用 [3]。从数学角度看，DRL 可以被定义为一种能够管理系统随机性的最优决策算法。此外，DRL 不仅支持离线训练策略的存储，还支持适应性和迁移学习 [9], [10]。

在过程工业中，比例-积分-微分（PID）和模型预测控制（MPC）是两种广泛采用的控制算法 [6], [7]。PID 控制通过调整基于误差信号的比例、积分和导数反馈来

This work was supported in part by the National Key R&D Program of China under Grant 2022YFB3305900, and the National Science and Technology Council of Taiwan under Grant NSTC 112-2221-E-033-053.

工作，这可能导致更慢且更被动的控制，并可能产生诸如振荡或超调等问题。相比之下，MPC 是一种基于模型的方法，它预测系统的未来行为并实时优化控制输入 [14]。然而，MPC 需要一个准确的系统模型，对于复杂或动态系统这可能是不可行的。

在此背景下，DRL 通过提供一种数据驱动的无模型方法，提供了有希望的替代方案。它通过与环境互动，在尝试和错误中学习，使其非常适合处理具有未知或时变动态的复杂系统。与依赖于预测模型的 MPC 不同，DRL 特别适用于难以建模或随时间变化的行为控制问题。这使得 DRL 成为解决过程控制挑战的强大候选者。最近的进步已导致了许多基于 DRL 的过程工业控制应用和改进 [8]–[11], [15]–[17]。

形式上，DRL 是一种学习范式，与传统的监督学习和无监督学习不同。强化学习（RL）旨在通过与环境 E 的交互来训练一个智能体以最大化长期的折扣奖励。通常，RL 任务被建模为一个马尔可夫决策过程（MDP），具有状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$。在每个时间步骤中，智能体根据其策略选择一个动作 $\boldsymbol{a}_t$，给定当前状态 $\boldsymbol{s}_t$，转换到一个新的状态 $\boldsymbol{s}_{t+1}$，同时从环境接收到一个奖励信号 r_t。

RL 策略可以是随机的或确定性的。随机策略表示为动作的条件概率密度，$\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) : \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$，其中 θ 参数化该策略。相比之下，确定性策略由 $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) : \mathcal{S} \mapsto \mathcal{A}$ 给出，其中动作 $\boldsymbol{a}_t$ 被选为其概率分布的最大值。即时奖励，$r_t = r(\boldsymbol{s}_t, \boldsymbol{a}_t)$，是状态-动作对的标量函数。一个 episode 的总奖励，R_t，是即时奖励的折扣和：

$$R_t = \sum_{t=1}^{T} \gamma^t r_t = \sum_{t=1}^{T} \gamma^t r(\boldsymbol{s}_t, \boldsymbol{a}_t). \tag{1}$$

RL 代理的性能目标是最大化期望折扣回报：

$$J(\pi_\theta) = \max_{\pi_\theta} \mathbb{E}_{\boldsymbol{s}\sim\rho^\pi, \boldsymbol{a}\sim\pi_\theta} \left[R_t \,|\boldsymbol{s}_t, \boldsymbol{a}_t \right]. \tag{2}$$

动作值函数 $Q^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$ 是在状态 $\boldsymbol{s}_t$ 中采取行动 $\boldsymbol{a}_t$ 并遵循策略 π 时的期望回报：

$$Q^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}_\pi \left[R_t \,|\boldsymbol{s}_t, \boldsymbol{a}_t \right]. \tag{3}$$

贝尔曼方程提供了一种递归计算动作值函数和求解公式 (2) 中的优化问题的方法。

$$Q^\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = \mathbb{E}_{r_t, \boldsymbol{s}_{t+1}\sim E} \left[r_t + \gamma \mathbb{E}_{\boldsymbol{a}_{t+1}\sim\pi} \left[Q^\pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}) \right] \right]. \tag{4}$$

III. DRL 在过程控制中的实际障碍

尽管深度强化学习在视频游戏和推荐系统等领域展示了令人印象深刻的成功，这些领域数据丰富且安全问题较少，但在流程制造中的应用，尤其是在化工过程控制中，面临着巨大的挑战。现实世界的工业系统的复杂性为广泛采用深度强化学习带来了显著障碍。两大主要问题阻碍了其实际使用：采样效率低和在不同运行条件下泛化能力差。特别是化工过程是大规模系统，动态特性缓慢，这使得深度强化学习代理的训练、验证和部署既耗时又耗费资源。

其中一个最紧迫的挑战是安全问题和试错学习的高昂成本。DRL 通常需要与系统进行大量交互以学习有效的控制策略，这通常涉及数百万次与环境的互动。在工业环境中，这可能转化为几天甚至几周的训练时间，在这段时间里，安全风险和财务成本变得难以承受。在化学工艺环境中进行实际培训往往不切实际，因为存在潜在危险并且需要受控条件，使得直接将 DRL 应用于工业生产中的物理系统变得困难。

一个解决这一挑战的潜在方案涉及使用仿真环境和精确的系统模型进行 DRL 训练。虽然这些模型可以帮助缓解安全风险并降低培训成本，但开发复杂化学系统的详细机制模型是一项劳动密集型且高度专业化的任务。此外，这些模型通常需要大量的计算资源来进行训练，这进一步使其在实际应用中变得复杂。即使有了准确的模型，训练过程仍然具有挑战性并且计算成本高昂，特别是在试图复制现实世界过程中缓慢、非线性的动态特性时。

另一个关键问题是 DRL 代理的泛化能力。DRL 代理对其训练时的具体条件非常敏感，环境中的微小变化——例如系统参数的变化或干扰——可能导致性能显著下降。在过程控制中，操作条件可能会随时间及不同模式发生变化，这种缺乏鲁棒性构成了一个主要限制。在一个操作条件下训练的 DRL 代理可能无法有效适应新条件，在时变或高度非线性的环境中这将影响其性能。

此外，许多现有的关于过程控制中 DRL 的研究未能充分解决其在工业环境中应用的可行性。它们常常忽视安全性和泛化等关键因素，这些对于化学过程行业中大规模部署 DRL 至关重要。如果不解决这些挑战，DRL 在过程控制中的实际应用仍将受到限制。直到克

服这些障碍，DRL 在化学过程控制中的潜力才会得到充分发挥，它在这种领域的使用才会逐渐扩大。

因此，本文的下一部分内容将从迁移学习的新视角探讨 DRL 在过程控制中的潜在应用。

IV. 迁移学习如何促进工业过程控制中的强化学习的视角

在以下章节中，我们将深入讨论并展望如何将迁移学习与强化学习相结合，以更好地促进过程工业下一代智能制造的发展。

A. 世界模型赋能的 RL 训练加速

将 RL 应用于化工过程控制的主要挑战之一是训练所需的时间和数据量庞大。为了解决这个问题，可以使用世界模型（如数字孪生）来创建环境的内部模拟。世界模型利用代理模型来表示化工过程的基本动态，这些代理模型基于历史数据进行训练以生成合成过渡，从而使 RL 代理能够在不与物理系统交互的情况下执行“梦想”和“想象”的滚动，并完善控制策略。

例如，作为工业 4.0 智能制造的关键组成部分，数字孪生可以为深度强化学习迁移学习提供多维、高精度和细粒度的虚拟现实集成仿真模型。构建工业过程的数字孪生将极大有助于在源域和目标域之间转移学习、适应和泛化强化学习代理，因为切换一个或多个不同领域的操作需要由数字孪生提供的可扩展多模式流程描述。

我们的方法 [9], [10], 受到模型辅助的 RL 和 Dream to Control 的启发，通过最小化对实时数据收集的依赖显著加速了学习。通过利用世界模型，RL 智能体可以更快地适应系统动力学和操作模式的变化，在训练过程中确保更大的灵活性和鲁棒性。

B. 使用预训练+微调的 RL 代理从仿真到现实的迁移学习

从仿真到现实的迁移学习对于使强化学习代理适应工业设置中的多模式和跨过程控制至关重要。这使得预训练的强化学习模型能够在最少重新训练的情况下调整到新的操作条件或流程，增强其在不同任务之间的泛化能力。它涵盖了当深度强化学习策略从模拟世界过渡到现实世界 [18] 时，尽量减少性能下降的方法。通常情况下，当环境发生变化时，深度强化学习代理的性能会显著下降。

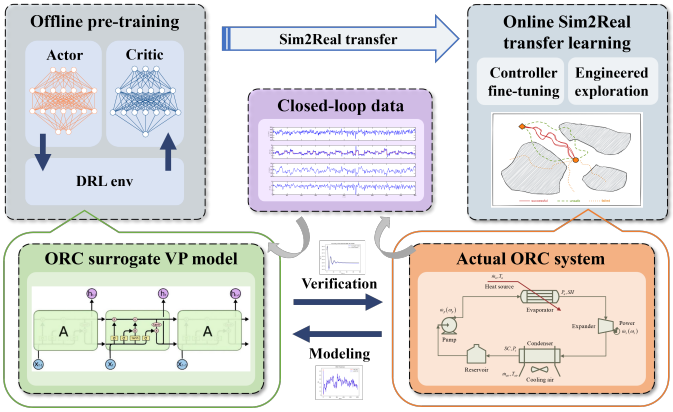

图 1. 基于 DRL 的 Sim2Real 控制设计框架 [8]。

在源领域中预训练 RL 控制器，并结合在目标域中微调基于 RL 的控制策略，是转移 RL 代理最常见的方法。这种概念与工业过程控制器的设计原则非常吻合。事实上，预训练+微调早已应用于流程工业优化控制中，其扩展到现代 RL 显示出明显潜力。最初在机器人领域强调的 RL 代理的 Sim2Real 迁移，也为我们的过程控制系统提供了一种高度可行的解决方案。

我们之前的工作 [8] 提出了一种用于能源领域热力学循环过程的基于 DRL 控制器的简单 Sim2Real 迁移学习方法，解决了与实际过程控制系统交互训练中遇到的样本效率低、耗时和安全风险等问题。框架如图 1所示。我们首次使用 DRL Sim2Real 迁移学习的概念来解决能源系统控制问题，创新性地提出了利用虚拟原型进行离线预训练并随后转移到真实系统的 Sim2Real 迁移方法。

C. 模仿学习启发的 RL 控制器先验学习

从演示中学习涵盖了许多利用专家演示对 RL 控制器进行训练的方法，其中模仿学习是一种常见做法。在工业生产中，存在丰富的闭环操作历史数据，从中可以推导出控制器的先验知识。随后，初始化的 RL 控制器可以在实际过程中进行迁移学习，从而提高 DRL 训练的安全性。例如，行为克隆是最简单的模仿学习技术，其本质上是拟合专家轨迹中的“状态 s_t- 动作 a_t”数据对。模仿学习的结果可以作为 Sim2Real 预训练的基础，随后进行精细调整 [19]。

D. 从（人类）专家中提炼控制器的学徒学习

在过程控制系统的历史操作中，存在大量由人工操作员生成或由经典控制器如 PID 或 MPC 存储的数据。

尽管这些数据包含了有关隐式控制模式的有价值信息，但通常被利用不足。通过从这类历史闭环操作数据中提取潜在的专家控制器行为，我们可以帮助 DRL 代理在无需在线探索的情况下做出明智决策。

学徒学习 [20], [21]，如逆向 RL [22]–[24]，可用于帮助 RL 代理从人类专业知识或现有控制器中学习，而不依赖危险的试错方法。逆向 RL 代表了一种特殊的模仿学习形式，旨在将控制器与专家匹配，同时恢复用于合理化最优策略的奖励函数。与一般的模仿学习相比，逆向 RL 更适合 DRL 迁移学习，因为它可以推断输入输出数据 [15] 中隐藏的闭环控制逻辑。估计奖励函数类似于理解专家为什么执行此类行为背后的意图，而不仅仅是拟合控制器的行为。这有助于在迁移学习过程中将 RL 代理从源域扩展和泛化到目标域。

这样，可以通过离线方式训练 RL 代理，学习来自专家人类操作员或经过验证的控制方案的最佳策略。该方法将使 DRL 代理能够为工业控制系统执行有效的“热/暖启动”操作，减少对昂贵实时实验的需求，并提高生产自动化的安全性和效率。

该方法将创建一个数据驱动的控制器，利用工业大数据和专家知识来提升复杂过程中 RL 代理的智能性和安全性。此外，离线 IRL 训练的结果可以用作后续在线迁移学习的基础，并可能实现少量样本适应。

E. 离线强化学习面向安全关键的实际应用

将 DRL 在流程工业中实际应用的主要障碍是在线学习和随机探索相关的显著安全风险。假设我们已经从工厂的历史操作中积累了丰富的数据集，解决 DRL 中的安全问题的一个直接方法是利用这些数据集直接训练过程控制器。

正如名称所暗示的，离线 RL [25]–[27] 涉及完全以离线方式训练控制策略，无需与环境进行交互，而是依赖于一个离线数据集。这种方法的优势在于其在迁移学习场景中的实用性，作为 RL 代理源域训练阶段，允许通过跨越状态-动作空间的大型数据集来训练控制器。

然而，缺点是显而易见的：假设离线数据集涵盖了所有可能的情况是不现实的，因为 DRL 在探索过程中难以预测它将要遍历的状态-动作空间。此外，当前关于离线 RL 的研究主要集中在解决训练期间的价值函数高估问题上。然而，针对这一问题的传统解决方案往往会引入保守性，影响最终的控制性能 [28], [29]。尽管存在这些挑战，离线 RL 仍然是一个有前景的研究领域，能够将 DRL 应用于实际的过程控制应用中。

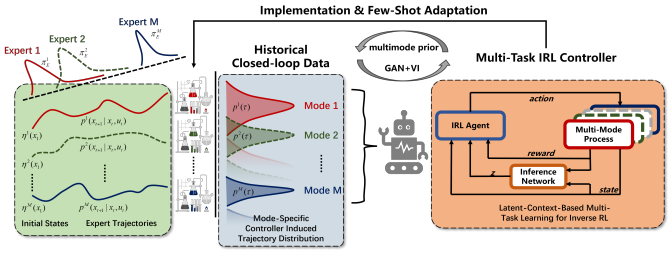

图 2. 多任务逆强化学习框架用于多模式过程控制设计。

F. 元强化学习或多项任务强化学习用于多模式控制

过程系统经常表现出多模式特征，例如设定点或系统参数的变化，以及操作条件的大幅度波动。这里，模式的定义是广泛的，涵盖任何改变过程系统分布 [8]–[10] 的情况。传统的深度强化学习（DRL）难以很好地适应模式的突然变化，因为它对模型-植物不匹配敏感，并且其泛化性能在不同场景中往往较差。一种非常有前景的战略是元学习 RL [30], [31] 或多任务 RL [32], [33]，能够训练一个通用控制器来适应多个任务/模式 [34], [35]。鉴于过程系统的多模式特性，可以整合数据解析中的高级技术，如潜在空间建模和变分推理，通过多任务学习解决多模式控制问题。因此，当源域的培训涵盖一些代表性的模式时，DRL 智能体能够快速适应那些未见过的目标域。

G. 多任务逆强化学习在过程控制中的应用及多模式历史大数据分析

这一新颖的想法结合了多模式控制器学习与逆强化学习 [15]，主要用于从涵盖各种操作模式的大型闭环数据中恢复奖励函数和控制策略。该框架如图 2所示。

对于多模式控制系统，存在对应不同模式的多个专家控制器，导致专家轨迹分布不同。我们可以利用历史工业闭环大数据和多任务 IRL 从特定模式的轨迹分布中学习控制器先验。这些多模式先验将在逆 RL 框架内提供一个通用的控制器架构，在实现过程中促进跨模式少量样本适应。

H. 基于模型的 RL 或基于 MPC 的 RL 以实现更快适应

基于模型的强化学习（MBRL）[36] 也可作为基于强化学习的过程控制的迁移学习解决方案。MBRL 通过结合系统建模的优势和强化学习的适应性，为过程工

业优化提供了显著优势。通过纳入已学习或预定义的过程模型，MBRL 提高了样本效率，使在较少的真实世界交互中更快地收敛到最优控制策略成为可能。MBRL 还可以促进迁移学习，因为世界模型捕捉了可以在不同领域应用的系统动态特性，减少了对相似过程重新训练模型的需求。这增强了可扩展性，使得能够部署鲁棒且适应性强的控制策略跨越不同的生产线或工厂。

鉴于 MPC 在工业应用中的主导地位，近期的研究开始探索将 MPC 与 DRL 潜在整合的可能性。这两种方法各有优缺点，但能有效互补。一方面，MPC 能够考虑不确定性、系统复杂性和长期预测范围，其性能很大程度上依赖于所用过程模型的准确性 [37]。另一方面，DRL 可以自然处理复杂系统的不确定性并管理无限预测范围。然而，DRL 在约束满足方面存在困难且缺乏可解释性，而 MPC 则提供安全保证和透明度。近年来，一些研究人员提出了基于 RL 的 MPC，将 MPC 的稳定性与 RL 的自主探索相结合。通过详细的数学推导，已经证明了非线性 MPC 与 RL 之间的等价条件 [38]，并引入了使用鲁棒 MPC 的安全 RL 概念 [39]。这样，MPC 的目标函数和 DRL 的价值函数可以被整合 [40], [41]，确保约束满足并提供类似 MPC 的性能，同时实现持续学习和不确定性处理。

I. 基于物理信息的 RL 用于将先验/物理知识融入控制器

广义上讲，任何整合了物理信息（如已知的物理定律、系统动力学或经验过程模型）的强化学习算法都可以被归类为基于物理学的强化学习。通过将领域特定的知识嵌入到学习过程中，它显著提高了智能体的学习效率和准确性，特别是在仅靠数据驱动的方法可能不够复杂的系统中。

从更具体的意义上说，物理信息强化学习指的是将强化学习与物理信息神经网络（PINNs）[42] 结合起来。PINNs 是一类直接在其训练过程中纳入诸如质量、能量或动量守恒等物理定律的神经网络架构。这种整合确保了所学策略遵循系统的支配性物理动力学，从而提高了控制策略的鲁棒性和现实性。

物理学知识强化的 RL 的一个显著优势是其增强迁移学习的能力。在将学习到的策略在不同领域之间转移时，嵌入的物理知识充当一种通用特征，为 RL 代理提供了一个稳定的基石，即使是在不同的操作环境中也是如此。这使得更容易将策略适应到新的但物理特性相似的系统中，减少了大量重新训练的需求。

V. 结论

本文介绍了一组开创性的化工过程控制方法，解决了传统方法和当前 RL 技术中的关键挑战。通过将 RL 与转移学习和学徒学习等先进技术相结合，这一框架有可能彻底改变工业控制系统的工作方式。这为 AI 在各行各业的更广泛应用铺平了道路，开启了效率、可持续性和安全性的新水平。旨在推进化工行业的人工智能驱动技术，本文提出的理念将在现代化工过程控制中推动可能性的边界，帮助塑造工业自动化的未来。

参考文献

[1] J. Shin, T. A. Badgwell, K.-H. Liu, and J. H. Lee, "Reinforcement learning – overview of recent progress and implications for process control," *Computers & Chemical Engineering*, vol. 127, pp. 282–294, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135419300754

[2] R. Nian, J. Liu, and B. Huang, "A review on reinforcement learning: Introduction and applications in industrial process control," *Computers & Chemical Engineering*, vol. 139, p. 106886, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135420300557

[3] H. Yoo, H. E. Byun, D. Han, and J. H. Lee, "Reinforcement learning for batch process control: Review and perspectives," *Annual Reviews in Control*, vol. 52, pp. 108–119, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S136757882100081X

[4] S. Spielberg, A. Tulsyan, N. P. Lawrence, P. D. Loewen, and R. Bhushan Gopaluni, "Toward self-driving processes: A deep reinforcement learning approach to control," *AIChE Journal*, vol. 65, no. 10, p. e16689, 2019. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.16689

[5] P. Petsagkourakis, I. O. Sandoval, E. Bradford, D. Zhang, and E. A. del Rio-Chanona, "Reinforcement learning for batch bioprocess optimization," *Computers & Chemical Engineering*, vol. 133, p. 106649, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0098135419304168

[6] N. P. Lawrence, M. G. Forbes, P. D. Loewen, D. G. McClement, J. U. Backström, and R. B. Gopaluni, "Deep reinforcement learning with shallow controllers: An experimental application to PID tuning," *Control Engineering Practice*, vol. 121, p. 105046, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066121002963

[7] O. Dogru, K. Velswamy, F. Ibrahim, Y. Wu, A. S. Sundaramoorthy, B. Huang, S. Xu, M. Nixon, and N. Bell, "Reinforcement learning approach to autonomous PID tuning," *Computers & Chemical Engineering*, vol. 161, p. 107760, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135422001016

[8] R. Lin, Y. Luo, X. Wu, J. Chen, B. Huang, H. Su, and L. Xie, "Surrogate empowered Sim2Real transfer of deep reinforcement learning for ORC superheat control," *Applied Energy*, vol. 356, p. 122310, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261923016744

[9] R. Lin, J. Chen, L. Xie, and H. Su, "Accelerating reinforcement learning with case-based model-assisted experience augmentation for process control," *Neural Networks*, vol. 158, pp. 197–215, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608022004129

[10] R. Lin, J. Chen, L. Xie, and H. Su, "Accelerating reinforcement learning with local data enhancement for process control," in *2021 China Automation Congress (CAC)*, 2021, Conference Proceedings, pp. 5690–5695.

[11] R. Lin, J. Chen, L. Xie, H. Su, and B. Huang, "Facilitating Reinforcement Learning for Process Control Using Transfer Learning: Perspectives," *arXiv e-prints*, p. arXiv:2404.00247, Mar. 2024.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015. [Online]. Available: https://doi.org/10.1038/nature14236

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaa8415

[14] H. Hassanpour, X. Wang, B. Corbett, and P. Mhaskar, "A practically implementable reinforcement learning-based process controller design," *AIChE Journal*, vol. 70, no. 1, p. e18245, 2024. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.18245

[15] R. Lin, J. Chen, B. Huang, L. Xie, and H. Su, *Developing Purely Data-Driven Multi-Mode Process Controllers Using Inverse Reinforcement Learning*. Elsevier, 2024, vol. 53, pp. 2731–2736. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780443288241504567

[16] L. Zhang, R. Lin, L. Xie, W. Dai, and H. Su, "Event-triggered constrained optimal control for organic rankine cycle systems via safe reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 7126–7137, 2024.

[17] H. Chang, Q. Chen, R. Lin, Y. Shi, L. Xie, and H. Su, "Controlling pressure of gas pipeline network based on mixed proximal policy optimization," in *2022 China Automation Congress (CAC)*, Conference Proceedings, pp. 4642–4647.

[18] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-Real transfer in deep reinforcement learning for robotics: A survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, Conference Proceedings, pp. 737–744.

[19] Y. Nakagawa, H. Ono, Y. Hazui, and S. Arai, "Imitation of piping warm-up operation and estimation of operational intention by inverse reinforcement learning," *Journal of Process Control*, vol. 122, pp. 41–48, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959152422002335

[20] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ser. ICML '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 1. [Online]. Available: https://doi.org/10.1145/1015330.1015430

[21] M. Mowbray, R. Smith, E. A. Del Rio-Chanona, and D. Zhang, "Using process data to generate an optimal control policy via apprenticeship and reinforcement learning," *AIChE Journal*, vol. 67, no. 9, p. e17306, 2021. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.17306

[22] S. Levine, "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review," *arXiv e-prints*, p. arXiv:1805.00909, May 2018.

[23] J. Fu, K. Luo, and S. Levine, "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning," *arXiv e-prints*, p. arXiv:1710.11248, Oct. 2017.

[24] N. Ab Azar, A. Shahmansoorian, and M. Davoudi, "From inverse optimal control to inverse reinforcement learning: A historical review," *Annual Reviews in Control*, vol. 50, pp. 119–138, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1367578820300511

[25] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems," *arXiv e-prints*, p. arXiv:2005.01643, May 2020.

[26] R. Figueiredo Prudencio, M. R. O. A. Maximo, and E. L. Colombini, "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 8, pp. 10 237–10 257, 2024.

[27] H. He, "A survey on offline model-based reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2305.03360

[28] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative Q-learning for offline reinforcement learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[29] Y. Mao, Q. Wang, Y. Qu, Y. Jiang, and X. Ji, "Doubly mild generalization for offline reinforcement learning," in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, Eds., 2024.

[30] J. Beck, R. Vuorio, E. Zheran Liu, Z. Xiong, L. Zintgraf, C. Finn, and S. Whiteson, "A Survey of Meta-Reinforcement Learning," *arXiv e-prints*, p. arXiv:2301.08028, Jan. 2023.

[31] K. Rakelly, A. Zhou, C. Finn, S. Levine, and D. Quillen, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5331–5340. [Online]. Available: https://proceedings.mlr.press/v97/rakelly19a.html

[32] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," in *Proceedings of the 38th International Conference on Machine Learning*, ser.

Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 9767–9779. [Online]. Available: https://proceedings.mlr.press/v139/sodhani21a.html

[33] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of the Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, L. P. Kaelbling, D. Kragic, and K. Sugiura, Eds., vol. 100. PMLR, 30 Oct–01 Nov 2020, pp. 1094–1100. [Online]. Available: https://proceedings.mlr.press/v100/yu20a.html

[34] D. G. McClement, N. P. Lawrence, J. U. Backström, P. D. Loewen, M. G. Forbes, and R. B. Gopaluni, "Meta-reinforcement learning for the tuning of PI controllers: An offline approach," *Journal of Process Control*, vol. 118, pp. 139–152, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959152422001445

[35] D. G. McClement, N. P. Lawrence, P. D. Loewen, M. G. Forbes, J. U. Backström, and R. B. Gopaluni, "A meta-reinforcement learning approach to process control," *IFAC-PapersOnLine*, vol. 54, no. 3, pp. 685–692, 2021, 16th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896321010958

[36] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "Model-based reinforcement learning: A survey," *Found. Trends Mach. Learn.*, vol. 16, no. 1, p. 1 – 118, Jan. 2023. [Online]. Available: https://doi.org/10.1561/2200000086

[37] Y. Shi, R. Lin, X. Wu, Z. Zhang, P. Sun, L. Xie, and H. Su, "Dual-mode fast DMC algorithm for the control of ORC based waste heat recovery system," *Energy*, vol. 244, p. 122664, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544221029133

[38] S. Gros and M. Zanon, "Data-driven economic NMPC using reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 65, no. 2, pp. 636–648, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8701462

[39] M. Zanon and S. Gros, "Safe reinforcement learning using robust MPC," *IEEE Transactions on Automatic Control*, pp. 1–1, 2020.

[40] T. H. Oh, H. M. Park, J. W. Kim, and J. M. Lee, "Integration of reinforcement learning and model predictive control to optimize semi-batch bioreactor," *AIChE Journal*, vol. 68, no. 6, p. e17658, 2022. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.17658

[41] J. Arroyo, C. Manna, F. Spiessens, and L. Helsen, "Reinforced model predictive control (RL-MPC) for building energy management," *Applied Energy*, vol. 309, p. 118346, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261921015932

[42] A. Ramesh and B. Ravindran, "Physics-informed model-based reinforcement learning," in *Proceedings of The 5th Annual Learning for Dynamics and Control Conference*, ser. Proceedings of Machine Learning Research, N. Matni, M. Morari, and G. J. Pappas, Eds., vol. 211. PMLR, 15–16 Jun 2023, pp. 26–37. [Online]. Available: https://proceedings.mlr.press/v211/ramesh23a.html