工程科学学报, 第 47 卷, 第 4 期: 768−779, 2025 年 4 月
Chinese Journal of Engineering, Vol. 47, No. 4: 768−779, April 2025
https://doi.org/10.13374/j.issn2095-9389.2024.10.29.006; http://cje.ustb.edu.cn

深度强化学习及其在工业场景的应用与展望

谭　靖[1,2,3)], 杨利刚[4)], 李潇睿[2,3,5)], 袁兆麟[2,3,5)], 崔允端[6)], 姚　超[1,2,3)], 王宗杰[1)]✉, 班晓娟[2,3,5,7)]✉

1) 北京科技大学计算机与通信工程学院, 北京 100083　2) 北京科技大学北京材料基因工程高精尖创新中心, 北京 100083　3) 北京科技大学材料领域知识工程北京市重点实验室, 北京 100083　4) 北京科技大学土木与资源工程学院, 北京 100083　5) 北京科技大学智能科学与技术学院, 北京 100083　6) 中国科学院深圳先进技术研究院, 深圳 518055　7) 辽宁材料实验室材料智能技术研究所, 沈阳 110004
✉通信作者, 王宗杰, E-mail: wangzj@ustb.edu.cn; 班晓娟, E-mail: banxj@ustb.edu.cn

摘　要　工业控制系统(Industrial control systems, ICS)在现代工业生产中发挥关键作用, 负责监控和控制工业过程, 确保高效、安全和稳定的生产. 随着工业 4.0 和智能制造的发展, 传统工业控制方法难以应对日益复杂且动态变化的生产环境. 深度强化学习(Deep reinforcement learning, DRL)结合了深度学习与强化学习的优势, 在工业智能控制领域展现出巨大潜力. 本文综述了 DRL 在工业智能控制中的应用现状和研究进展. 首先介绍了 DRL 的基本原理及相关算法, 并简述工业控制的背景, 分析智能控制的应用需求与现存挑战. 随后, 详细综述了 DRL 在工业领域的应用, 并对当前研究进行了总结, 最后对未来研究方向提出了展望.

关键词　深度强化学习; 在线强化学习; 离线强化学习; 工业控制系统; 智能控制
分类号　TP391.9

Deep reinforcement learning applications and prospects in industrial scenarios

TAN JING[1,2,3)], *YANG Ligang*[4)], *LI Xiaorui*[2,3,5)], *YUAN Zhaolin*[2,3,5)], *CUI Yunduan*[6)], *YAO Chao*[1,2,3)], *WANG Zongjie*[1)]✉, *BAN Xiaojuan*[2,3,5,7)]✉

1) School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
2) Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China
3) Beijing Key Laboratory of Knowledge Engineering for Materials Science, University of Science and Technology Beijing, Beijing 100083, China
4) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
5) School of Civil and Resource Engineering, University of Science and Technology Beijing, Beijing 100083, China
6) Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
7) Institute of Materials Intelligent Technology, Liaoning Academy of Materials, Shenyang 110004, China
✉Corresponding author, WANG Zongjie, E-mail: wangzj@ustb.edu.cn; BAN Xiaojuan, E-mail: banxj@ustb.edu.cn

ABSTRACT　Industrial production is fundamental to human society. Industrial control systems (ICS) serve as the cornerstone of modern industrial processes and are responsible for monitoring and controlling operations to ensure efficiency, safety, and stability. Central to these systems are control algorithms, which enable the automation of operations, optimization of process parameters, and reduction of operational costs. However, with the rapid advancements in Industry 4.0 and smart manufacturing, traditional control methods are increasingly inadequate to address the growing complexity, high dynamics, and real-time demands of modern industrial environments. Deep reinforcement learning (DRL), which integrates the high-dimensional feature extraction of deep learning with the adaptive decision-making capabilities of reinforcement learning, has emerged as a transformative technology in intelligent industrial

收稿日期: 2024−10−29
基金项目: 国家重点研发计划资助项目(2022YFE0129200)

control. This paper provides a comprehensive review of DRL's principles, methodologies, and applications in industrial scenarios. The review begins with an introduction to the fundamental concepts of DRL, including the Markov decision process (MDP) framework and the Bellman equation for optimizing decision-making strategies, followed by an exploration of the latest advancements in both online and offline reinforcement learning algorithms. The paper systematically examines the background and challenges of industrial control systems, highlighting the limitations of traditional methods such as proportional–integral–derivative (PID) control and rule-based systems when faced with multi-variable, nonlinear, and dynamic processes. By analyzing the evolving demands of intelligent control, the review underscores the necessity for advanced, self-learning approaches, such as DRL, that are capable of operating effectively in environments with incomplete information, real-time constraints, and multiple conflicting objectives. As a key contribution, a novel classification framework for DRL applications in industrial scenarios is proposed. Current research is categorized into three domains: (1) adaptive optimization in dynamic environments, enabling systems to respond to changes such as market fluctuations, equipment degradation, and operational disturbances; (2) decision-making under multi-objective and constrained conditions, with DRL balancing competing goals such as efficiency, cost, and sustainability while adhering to technical constraints; and (3) performance enhancement in complex systems, whereby DRL tackles high-dimensional, nonlinear, and coupled processes to improve stability, scalability, and operational excellence. This framework provides new perspectives for designing control algorithms tailored to specific industrial contexts. The review also synthesizes key findings from recent DRL studies, presenting a detailed evaluation of their achievements, limitations, and opportunities for improvement. Case studies across sectors such as energy management, manufacturing, and process optimization illustrate the versatility and effectiveness of DRL in solving diverse industrial problems. However, challenges remain, including the need for high-quality training data, computational efficiency in high-dimensional spaces, and robust algorithms capable of handling uncertainties and safety-critical conditions. To address these challenges, future research directions that are essential for advancing DRL in industrial applications are outlined. These include the development of high-fidelity industrial process simulators, techniques to improve sample efficiency and generalization across varying conditions, and methods to enhance interpretability and transparency in DRL decision-making processes. In conclusion, this paper emphasizes DRL's transformative potential in redefining industrial control paradigms. By overcoming current limitations and fostering interdisciplinary collaboration, DRL is well-positioned to drive innovation in industrial intelligence and automation. The insights and frameworks presented in this review offer a valuable foundation for future research, accelerating the adoption of DRL technologies in real-world industrial settings and paving the way for the next generation of smart manufacturing systems.

KEY WORDS　deep reinforcement learning; online reinforcement learning; offline reinforcement learning; industrial control systems; intelligent control

随着工业 4.0 和智能制造的快速发展, 工业系统的复杂性与动态性日益提高, 对控制优化方法提出了更高的要求. 传统控制方法尽管在特定场景中表现良好, 但在面对复杂多变、非线性和实时性需求高的工业环境时, 其适应性和效率受到了显著限制. 深度强化学习(Deep reinforcement learning, DRL)结合了深度学习的高维感知能力与强化学习的自适应决策能力, 为复杂工业系统的优化提供了一种全新的技术途径. 现有文献中虽有关于 DRL 的综述, 但大多集中于特定算法[1−2]或特定行业[3−4], 缺乏对整体应用场景的全面梳理与分类, 难以系统揭示 DRL 在工业中的实际优势及面临的挑战. 此外, 许多研究仅从理论层面探讨方法[5−6], 未能充分结合实际工业需求, 导致在指导实际应用时存在局限性. 基于此, 本文从在线强化学习与离线强化学习两大方向出发, 聚焦 DRL 在工业场景的不同应用, 包括动态环境下的适应性优化、多目标和约束条件下的决策以及复杂系统的性能增强. 进一步按照行业领域对研究进行分类, 以系统呈现 DRL 在工业中的广泛应用与差异化特性.

其余章节的安排如下: 第一章介绍 DRL 的基本原理与主要算法; 第二章分析工业控制系统中的需求与挑战; 第三章详细综述 DRL 在工业场景的具体应用; 第四章提出未来研究方向并总结全文.

1　深度强化学习

DRL 近年来得到了广泛的关注, 在各个领域涌现出大量的研究成果. 特别在围棋[7]、电子游戏[8]等具有封闭、确定性的环境下, DRL 能取得超过人类的表现. 然而在真实开放世界的环境中, DRL 的表现还有很大提升空间. 针对现实世界中的诸多问题, 研究人员对 DRL 进行了广泛探索, 包括推动在线强化学习算法的发展和提出离线强化学

习[9]的概念. 在线强化学习与离线强化学习的对比如图 1 所示.

1.1　DRL 基础

强化学习[10]是一种通过与环境交互并基于奖励和惩罚机制学习最优策略的方法. 强化学习一般被建模成马尔科夫决策过程(Markov decision process, MDP). 马尔可夫决策过程由元组(S,A,P,r,γ)构成, 其中S是状态的集合, A是动作的集合, γ是折扣因子, r是奖励函数, P状态转移函数. MDP 具有马尔可夫性, 即下一个状态s_{t+1}的每个可能的值出现的概率只取决于当前状态s_t和动作a_t, 并与历史的状态和动作完全无关. 策略π表示状态到动作的映射函数$\pi: S \to A$, 可分为确定性策略$a_t=\pi(s_t)$和随机性策略$\pi(a_t|s_t)$. 带折扣回报G_t定义为从t时刻开始到终止状态的累计奖励:

$$G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k} \tag{1}$$

其中, R_{t+k}为第$t+k$时刻的奖励. 强化学习的目标是最大化期望回报, 以此获得最优策略. 在状态s_t下, 智能体在策略π下得到的期望回报被定义为状态值函数(State value function), 即:

$$V^{\pi}(s_t)=E_{\tau\sim\rho_{\pi}(\tau|s_t)}[R_t|s_t=s] \tag{2}$$

其中, E表示数学期望, τ为状态–动作序列$\tau=s_t,a_t,s_{t+1},a_{t+1}\cdots$, 轨迹分布$\rho_{\pi}(\tau|s_t)$表示在策略$\pi$下, 给定初始状态$s_t$后, 状态动作序列$\tau$发生的概率. 在状态$s_t$下执行动作$a_t$, 智能体在策略$\pi$下得到的期望回报被定义为状态–动作值函数: $Q^{\pi}(s_t,a_t)=E_{\tau\sim\rho_{\pi}(\tau|s_t,a_t)}[R_t\mid s_t=s,a_t=a]$. 在有限状态和动作集合的 MDP 中, 至少存在一个策略比其他所有策略都好, 这个策略就是最优策略 (Optimal policy), 表示为$\pi^*(s)$. 最优状态–动作值函数:

$$Q^*(s_t,a_t)=\max_{\pi}E_{\tau\sim\rho_{\pi}(\tau|s_t,a_t)}[R_t|s_t=s,a_t=a] \tag{3}$$

最优状态–动作值函数遵循贝尔曼最优方程(Bellman optimality equation):

$$Q^*(s_t,a_t)=r(s_t,a_t)+\gamma E_{s_{t+1}\sim p(s_{t+1}|s_t,a_t)}\times\left[\max_{a_{t+1}}Q^*(s_{t+1},a_{t+1})\right] \tag{4}$$

求解贝尔曼最优方程即可得到最优 Q 值, 进而得到最优策略.

深度学习拥有强大的感知能力[11]. 它通过深度神经网络处理和提取高维数据中的特征, 在众多领域[12]取得了成功. 深度强化学习将两者结合, 常见的方式是使用神经网络来拟合学习状态–动作值函数或者策略函数, 使其能够在复杂和高维的环境中有效地学习控制策略.

1.2　在线强化学习算法

近年来, DRL 是在线强化学习的一个重要研究方向. DQN 由 Mnih 等[13]在 2015 年提出, 开创了将深度学习应用于强化学习的新纪元. DQN 通过使用卷积神经网络来估计 Q 值函数, 使其在 Atari 游戏中取得了突破性进展. Sutton 等[14]提出直接优化策略的方法, 即策略梯度算法. 这种方法通过直接对策略进行优化, 克服了 Q 学习中动作选择的不连续性问题. 然而策略梯度方法的高方差问题限制了其效率, 为了解决策略梯度方法的高方差问题, Konda 和 Tsitsiklis[15]提出了 Actor–Critic 算法. 这个方法结合了策略梯度和价值函数逼近, 通过引入一个 Critic 网络来估计价值函数, 从而降低了策略更新的方差. Lillicrap 等[16]提出了深度确定性策略梯度(DDPG), 专为处理连续动作空间而设计. Schulman 等[17]引入了信赖域策略优化(TRPO), 通过约束每次策略更新的步长, 确保策略更新的稳定性. Schulman 等[18]在 TRPO 的基础上进一步提出了近端策略优化算法(PPO), 简化了优化过程. Hiraoka 等[19]提出的 DroQ 算法通过使用小规模的

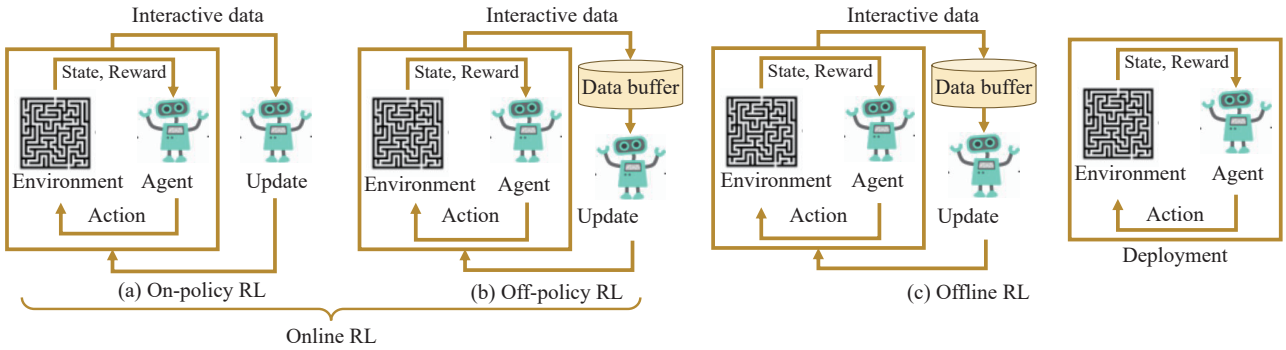

图 1　离线强化学习与在线强化学习架构对比

Fig.1　Comparison of offline RL and online RL architecture

dropout Q 函数集, 在保证样本效率的同时, 显著降低了计算成本. 此外, Bhatt[20] 等通过引入批归一化和去除目标网络的方式提出 CrossQ 算法, 进一步提升了样本效率并降低了复杂度. 同时, TD7 算法[21] 通过在 TD3 框架中整合状态–动作表示学习, 大幅提升了在连续控制任务中的表现. 这些算法在样本效率, 计算复杂度, 算法鲁棒性等方便不断优化, 推动了 DRL 在更广泛场景中的应用.

1.3　离线强化学习算法

在线强化学习需要智能体与环境在线交互, 但是在众多场景中这种在线交互是代价昂贵甚至安全风险很大. 与在线强化学习不同, 离线强化学习不需要在训练过程中与环境实时交互, 而是依赖于已有的经验数据进行策略优化. 这种方法特别适用于那些可以提前获取大量数据的场景, 避免了在实际操作中进行高风险的探索. 由此离线强化学习成为一个重要的研究方向.

在离线强化学习中, 智能体利用预先收集的数据 $\mathcal{D}\{(s_t^i, a_t^i, s_{t+1}^i, r_t^i)\}$ 进行策略学习, 其中 i 表示第 i 条轨迹, t 表示该轨迹中的时间步数. 生成数据集的策略称为行为策略 π_β, 优化时的策略称为目标策略 π. 离线强化学习的主要问题是离线数据集会产生分布漂移[22]. 具体而言是在训练时的策略与行为策略不一致, 导致训练策略访问的状态–动作对数据集中可能不存在, 由此对 Q 值的估计产生不可估量的误差. 现有研究主要利用策略约束方法解决该问题. Kumar 等[23] 提出了 BEAR, 这是一种支持集匹配方法, 通过最大均值差异(MMD)将动作限制在数据集的支持集上. Wu 等[24] 提出了 BRAC, 在策略和价值函数上使用 KL 散度约束, 并进行详细的实验验证不同差异度量(如 KL 散度、MMD 和 Wasserstein 距离)的效果. Fujimoto 等[25] 提出了一种简约的离线强化学习方法 TD3+BC, 直接将行为克隆损失添加到策略优化目标中, 尽管其方法简单, 但仍能达到最先进的方法效果. 上述算法直接对行为策略进行约束, 除了直接约束方法, 许多离线强化学习算法通过隐式策略约束来匹配数据集分布. 如 Kumar[26] 等提出了保守 Q 学习(CQL)算法, 该算法学习了一个保守的价值函数, 目标策略被间接约束为行为策略. Kostrikov[27] 等提出了隐式 Q 学习(IQL)算法, 通过将状态值函数视为随机变量来隐式地估计策略改进步骤, 并通过优势加权行为克隆来提取策略, 这也避免了对样本外行为的查询. 然而离线强化学习的算法性能受制于离线数据集的质量, 如何在中等或者低质量的数据集下依旧取得令人满意的性能是之后研究的重点.

2　工业控制

2.1　工业控制系统

工业控制系统(ICS)是用于过程控制的多种控制系统及相关仪器设备的总称, 涵盖了监控和数据采集系统(SCADA)、分散式控制系统(DCS)和可编程逻辑控制器(PLC)等多种配置形式[28]. 这些系统通过远程传感器测量过程变量(PV), 并将测量数据与设定值(SV)进行比较, 生成控制指令以调节终端控制元件(如控制阀)来实现工业目标. 结合相关文献[28], ICS 可大致划分为感知层、执行层、控制层、通信层和监控层. 感知层通过嵌入式设备和传感器采集工业生产中的实时数据. 执行层根据控制层发出的指令, 执行具体操作, 直接干预生产过程, 确保设备按计划运行并实现反馈调节. 控制层是系统的核心, 通过处理感知层提供的数据执行控制逻辑, 生成指令, 并利用可编程逻辑控制器(PLC)或分布式控制系统(DCS)实现智能控制和优化分析. 通信层负责各层之间的信息传递, 确保高效、安全的网络连接. 监控层通过人机接口(HMI)界面向操作员提供系统状态的可视化信息, 实现全面的系统管理与远程支持. 这些系统广泛应用于化工、造纸、发电、石油天然气提炼、电信等行业. 随着计算机技术、通信技术和控制技术的发展, 传统的控制领域向智能控制方向的变革, 拓展了工业控制领域的发展空间, 带来了新的发展机遇.

2.2　工业控制的需求和挑战

随着工业 4.0 和智能制造的推进, 传统的工业控制方法面临新的挑战和需求. 现代工业生产中, 生产过程变得更加复杂和动态, 传统的控制方法, 如 PID 控制[29]、模糊控制[30] 和专家系统[31], 尽管在特定场景中表现优异, 但在面对高度复杂和动态变化的工业环境时, 其性能和适应性往往受限. 智能控制系统需要应对多种挑战: 首先是复杂性, 现代工业过程通常具有多变量、强耦合和非线性的特点, 这使得建立精确的数学模型变得困难; 其次是动态性, 生产环境和条件可能会随着时间变化, 要求控制系统具有高度的自适应性; 此外, 实时性也是一个重要的挑战, 特定场景下要求控制系统能够在毫秒级的时间内响应环境的变化, 确保系统的稳定和高效运行. 基于 DRL 的控制算法的出

现为解决这些挑战提供了新的途径. 其自适应性和自主学习能力使其能够在动态的环境中有效地优化控制策略, 同时能够处理高维度的数据, 并从中提取有用的特征, 这使得其在处理多变量和非线性控制问题上具有独特的优势. 基于深度强化学习的控制与传统控制方式的特性对比见表 1.

表 1　控制算法对比

Table 1　Comparison of control algorithms

Algorithms	Model dependency	Adaptability	Complexity handling	Real-time response
PID Control	High	Low	Low	High
Fuzzy Control	Medium	Medium	Medium	High
Expert Systems	High	Low	High	Medium
Based on DRL	Low	High	High	High

3　DRL 在工业场景中的应用

在工业自动化和智能化的快速发展中, DRL 已成为解决复杂工业问题的强大工具. 通过对现有文献的广泛综述, 本章将探讨 DRL 在工业场景中的多样化应用, 并将其归纳为三个主要类型: 动态环境下的适应性优化、多目标和约束条件下的决策、以及复杂系统的性能增强. 这些类型反映了 DRL 在工业领域的核心作用, 涵盖了从即时适应到长期性能提升的全方位能力.

3.1　动态环境下的适应性优化

在不断变化的工业环境中, 算法必须能够快速适应新的挑战, 如市场需求的波动、供应中断或操作条件等变化. DRL 能帮助工业系统实现动态环境下的适应性优化, 使系统能够实时响应外部变化, 优化生产调度、资源分配和能源管理. 图 2 直观展示了在当前环境下强化学习框架中的问题.

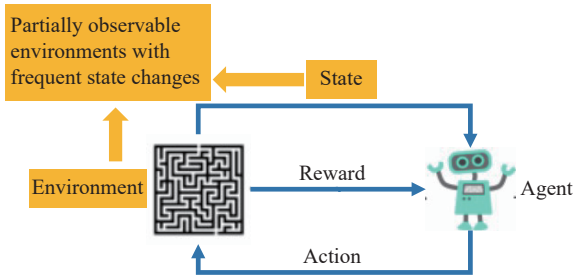

图 2　强化学习框架中动态环境的问题

Fig.2　Problems with dynamic environments in the reinforcement learning framework

Kumar 和 Dimitrakopoulos[32] 开发了一种结合 DRL 与蒙特卡洛树搜索(MCTS)的短期生产调度系统, 能够实时更新设备性能和供应链的不确定性, 优化采矿顺序和矿石目的地决策. 类似的, Levinson 等[33] 进一步将 DRL 与随机优化结合, 提出了通过预浓缩设施优化采矿调度的系统, 提升了整体经济效益. Huo 等[34] 和 Noriega 等[35] 的研究均针对露天矿场的车队调度系统. Huo 等[34] 通过 Q-learning 算法在不确定条件下优化卡车的行驶路线和作业优先级, 实时适应现场情况的变化, 从而减少燃料消耗. 而 Noriega 等[35] 使用双深度 Q 学习算法并引入去耦合的 Q 值更新机制, 降低 Q 值过高估计问题, 显著提高了在不确定性环境下卡车调度效率, 减少了卡车等待时间和排队现象. Cao 等[36] 提出了基于近端策略优化(PPO)算法的矿山通风阻力系数反演方法, 解决了传统的阻力系数反演方法因不确定性和复杂的地下环境面临精度和效率问题. Kumar 和 Dimitrakopoulos[37] 的另一项研究专注于通过 DRL 实时更新矿床地统计模型. 他们提出了一种基于深度确定性策略梯度(DDPG)的算法, 结合高阶空间统计, 用于优化矿床模型的更新. Jiang 等[38] 针对高炉炼铁过程中存在复杂的动态性、滞后性和原料不确定性等难题, 提出了一种递归监督双深度确定性策略梯度(iRSDDPG)的离线强化学习方法, 实现了在确保安全的前提下实时优化能耗控制. Liu 等[39] 则针对因高温和复杂的物理化学反应具有极大的动态性和不确定性的炼钢过程, 利用 DDPG 算法结合能量驱动的限制玻尔兹曼机(EDRBM)和多目标进化算法(MODE), 在钢水成分和温度控制方面取得了精确调节, 提升了操作效率和生产质量. Shi 等[40] 针对氢电耦合系统控制问题提出了一种基于 DQN 的能源优化管理方法, 通过经验回放和参数冻结机制提高算法性能. Meng 等[41] 同样针对可再生能源的不稳定性和间歇性问题, 优化微电网系统的能源管理系统, 采用 SARSA 在线强化学习算法, 迭代学习微电网系统的调度策略, 确保在动态环境下的实时优化. Yin 和 Lei[42] 则聚焦于离岸风能与光伏发电的联合系统, 通过 DDPG 算法实现了多目标优化,

有效提高了发电效率并抑制了功率振荡. Li 等[43]采用了深度学习与强化学习结合的混合策略, 采用径向基函数网络(RBFN)进行空气动力学预测, 平衡了风能系统的功率输出和稳定性. Wang 等[44]研究了云制造中传统调度算法存在响应慢、无法适应环境变化的问题, 通过离线 DRL 优化了云制造中的物流任务调度, 在提高任务调度效率的同时实现了较好的泛化能力和调度性能. Yun 等[45]提出了一种解释性多智能体 DRL 实时控制框架, 在实时需求响应中优化制造业中的能源管理, 确保生产要求的同时大幅降低能耗, 并提高了控制策略的可解释性. Zhu 等[46]针对装配式建筑场景, 通过 DRL 优化机器人装配计划, 利用 BIM 模拟器提高了装配式建筑现场的实时规划效率. Cui 等[47]利用 DRL 结合模型预测控制(MPC)优化了无人表面车辆(USV)的控制系统. 研究通过引入过滤的概率模型预测控制(FPMPC), 提升了系统的控制鲁棒性和精确性, 使 USV 能在动态、复杂环境中自主运行. 表 2 总结了动态环境下的适应性优化的相关研究.

3.2　多目标和约束条件下的决策

工业决策过程中往往需要在多个目标之间进行权衡, 同时还要满足各种技术、经济和环境约束. DRL 能在这些复杂的决策空间中寻找最优解, 拥有处理多目标优化问题时的能力, 在满足约束的同时实现决策的优化. 图 3 直观展示了在多目标和约束条件下强化学习框架中的问题.

Liang 等[48]和 Tang 等[49]聚焦于锌焙烧和锌氧化物挥发回转窑的温度控制. Liang 等[48]通过卷积 Q 学习网络(CQLN)结合计算流体力学(CFD)模型, 维持焙烧过程中平均温度的同时还优化了整个温度场的分布, 提高了温度控制的稳定性和抗

表 2　动态环境下的适应性优化研究总结

Table 2　Summary of research on adaptive optimization in dynamic environments

Research problem	Field	Algorithm	Algorithm innovation
Mine production scheduling[32]	Mining engineering	Self-play RL	Combines DRL with Monte Carlo tree search (MCTS) and ensemble Kalman filter (EnKF), to update equipment performance and supply chain uncertainties in real time.
Mine production scheduling[33]	Mining engineering	Actor–Critic	Integrates preconcentration facilities and stochastic optimization.
Fleet dispatching in mines[34]	Mining engineering	Q-learning	Optimizes truck dispatch using Q-learning.
Ventilation resistance coefficient optimization[36]	Mining engineering	PPO	Designs DRL to optimize ventilation networks.
Truck dispatch in mines[35]	Mining engineering	DDQN	Uses double deep Q-network (DDQN) to improve equipment utilization.
Real-time geostatistical model optimization[37]	Mining engineering	DDPG	Combines higher-order spatial statistics with deep deterministic policy gradient (DDPG) for dynamic sensor data integration.
Energy consumption control in blast furnaces[38]	Metallurgical engineering	DDPG	Combines long shot-term memory (LSTM) and dual critic networks to handle partially observable problems.
Steel composition and temperature optimization[39]	Metallurgical engineering	DDPG	Integrates DDPG, energy driven restricted boltzmann machine (EDRBM), and multi-objective differential evolution (MODE) for multi-objective optimization.
Energy management in hydrogen-electric systems[40]	Energy engineering	DQN	Energy optimization using DQN-based management.
Microgrid energy scheduling[41]	Energy engineering	SARSA	Iterative scheduling with state-action-reward-state-action (SARSA) online reinforcement learning.
Offshore wind and solar power optimization[42]	Energy engineering	DDPG	Multi-objective optimization for efficiency improvement and oscillation suppression.
Wind turbine power scheduling[43]	Energy engineering	DQN	Combines deep learning and reinforcement learning to adjust turbine speed and blade angles.
Task scheduling in cloud manufacturing[44]	Logistics	DT	Offline DRL for task scheduling, reducing training costs and improving generalization.
Real-time demand response in energy management[45]	Manufacturing	DMADQN	Explainable multi-agent DRL framework for reducing energy consumption while meeting production demands.
Robotic assembly planning in prefab construction[46]	Construction	DQN, DDQN, A2C, PPO	Building information modeling (BIM)-based optimization for efficiency improvement in dynamic environments.
Unmanned surface vehicle control[47]	Marine engineering	MBRL	Filtered probabilistic model predictive control (MPC) to enhance robustness in complex marine environments.

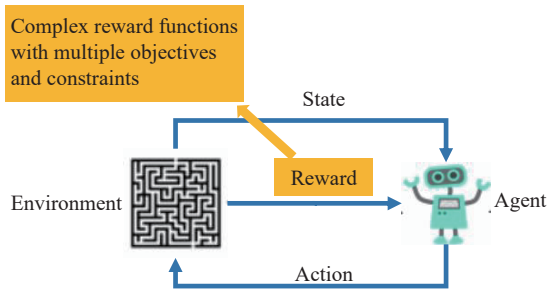

图 3　强化学习框架中多目标约束条件的问题

Fig.3　Problems with multi-objective constraints in the reinforcement learning framework

扰动性. 而 Tang 等[49] 则采用多目标深度强化学习(MODRL)框架, 在优化锌回收率和减少碳排放之间实现了平衡. Che 等[50] 针对钢铁厂的氧气生产系统在降低能耗和生产成本之间平衡问题, 提出了一种结合 DRL 和多目标进化算法(MOEA)的调度优化方法, 解决了钢铁生产过程中氧气系统的高能耗问题, 减少了操作模式切换次数, 同时灵活应对频繁变化的电价和需求. Neto 等[51] 开发了一种基于双深度 Q 网络(DDQN)的维护优化策略, 应用于废钢生产线的关键设备, 通过动态调整维护计划来减少停机时间和维护成本, 在设备退化不确定性、生产率波动和维护成本之间找到最佳平衡, 显著提高了生产线的可靠性. Liu 等[52] 将多智能体强化学习(MARL)用于铸造过程中基于图像处理的缺陷检测, 每个智能体只负责观测图像的一小部分, 并通过协同工作实现整体的分类判断, 在减少计算负担的同时保证高精度检测. Ma 等[53] 提出了参考向量强化学习(RV–RL)方法, 通过动态调整优化方向, 解决了铜料配料系统中的多目标优化问题, 在生产成本、铜品位、能耗和污染物排放等冲突目标之间实现了平衡. 这与锌冶炼过程的多目标优化相似, 表明 DRL 在复杂生产过程中能够有效应对多个目标和约束条件. Canales 等[54] 研究了堆浸过程控制问题. 由于堆浸过程的复杂性和变化多样的外部条件, 传统的控制方法效率较低, 无法实现精准控制. 其采用 TD3 算法, 设计了经济收益导向的奖励函数, 通过优化酸和水的使用, 显著提高了铜的回收率并减少了资源浪费. Zheng 等[55] 在金氰化浸出过程进行了探索. 金氰化浸出过程需要在确保安全的前提下提高产出率, 提出了一种安全强化学习(SRL)确保氰化物浓度在复杂多约束条件下控制在安全范围内, 并最大化了金的浸出率. Wang 等[56] 通过物理启发的安全强化学习方法, 将电力和气体网络的物理约束集成到多能源微电网的能量管理中. 这一研究不仅优化了多能源系统的经济运行, 还通过约束马尔科夫决策过程(CMDP)框架确保了系统的安全性和稳定性. Qin 等[57] 的研究通过结合 D3QN 和 PR-DQN 算法, 优化了近零能耗建筑的供暖、通风和空调(HVAC)系统控制, 在保证室内舒适度的同时, 大幅减少了能源消耗. Ruan 等[58] 针对冷热电联产系统(CCHP), 采用了一种改进的 TD3 算法, 结合光伏发电和储能系统进行多目标优化, 通过分季节训练模型, 提高了系统的操作效率并降低了运行成本. Queiroz 等[59] 探究了自然风味分子的研发, 他们通过 DRL 结合生成模型, 优化了自然风味分子的设计, 该研究设计出了能够被应用于实际工业生产中的风味分子, 同时最大化它们的合成可行性和自然产品相似性. 表 3 总结了多目标和约束条件下的决策的相关研究.

3.3　复杂系统的性能增强

随着工业系统变得越来越复杂, 传统的控制方法往往难以应对系统的非线性、强耦合和高时延等困难. 合理设计 DRL 算法能提升这些复杂系统的性能, 包括设备控制、过程优化等. DRL 在提高工业系统稳定性、响应速度和整体性能方面已有大量研究. 图 4 直观展示了在复杂系统下强化学习框架中的问题.

Ai 等[60]、Zheng 等[61] 和 Jiang 等[62] 聚焦于泡沫浮选过程的控制优化. Ai 等[60] 利用历史工业数据训练离线保守双 Q 学习算法, 避免了与系统进行交互, 而 Zheng 等[61] 通过结合物理模型与数据驱动的混合模型, 提升了强化学习的样本效率, 实现了泡沫浮选过程的精确控制. Jiang 等[62] 开发了一种称之为交错学习的强化学习技术, 克服了建立精确数学模型的困难, 在线计算操作最优控制解决方案, 实时学习最优控制. 他们的研究都表明, DRL 在应对工业流程中的复杂物理化学现象方面有着广阔的应用前景. Yuan 等[63] 的研究聚焦于浓密机底流浓度的优化控制问题. 他们提出了基于强化学习的双网架构, 并通过引入短期经验回放机制, 在面对浓密机这种非线性和高时滞的工业系统时, 能够实现精确的控制和快速的响应. Yuan 等[64] 针对浓密机底流浓度控制问题进行了进一步的研究, 提出了基于 TBCQ 的新型离线强化学习算法, 引入具有时序特征的状态表示增强了 DRL 的控制性能. Liu 等[65] 研究了 DRL 在电解液温度控制中的应用. 电解液温度的变量波动大, 传统的控制方法难以在不建立复杂物理模型的情

表 3　多目标和约束条件下的决策研究总结

Table 3　Summary of research on decision making under multiple objectives and constraints

Research problem	Field	Algorithm	Algorithm innovation
Zinc roasting temperature optimization[48]	Metallurgical engineering	Q-learning	Combines convolutional Q-learning (CQLN) with computational fluid dynamics (CFD) modeling to optimize temperature distribution.
Zinc recovery and carbon emission optimization[49]	Metallurgical engineering	CMODRL	Constrained multi-objective DRL (CMODRL) for balancing zinc recovery and carbon emissions.
Oxygen system scheduling in steel plants[50]	Metallurgical engineering	PPO	Integrates DRL with multi-objective evolutionary algorithm (MOEA) for energy consumption optimization.
Maintenance optimization in scrap production lines[51]	Metallurgical engineering	DDQN	Uses double deep Q-network (DDQN) to dynamically adjust maintenance plans.
Defect detection in casting processes[52]	Metallurgical engineering	MARL	Multi-agent reinforcement learning (MARL) for efficient and precise defect detection.
Copper burdening system optimization[53]	Metallurgical engineering	RV–RL	Reference vector reinforcement learning (RV–-RL) for dynamic adjustment of optimization directions.
Acid and water usage optimization in heap leaching[54]	Metallurgical engineering	TD3	Model-free twin-delayed deep deterministic (TD3) policy gradient algorithm for optimizing acid and water consumption.
Cyanide concentration control in gold leaching[55]	Metallurgical engineering	SRL	Safe reinforcement learning (SRL) ensures safe cyanide levels and maximizes gold leaching rates.
Energy management in multi-energy microgrids[56]	Energy engineering	PPO	Physics-informed safe RL for integrating electrical and gas network constraints.
HVAC system optimization in nearly zero-energy buildings[57]	Energy engineering	D3QN and PR–DQN	Combines dueling double deep Q-network (D3QN) and probabilistic (PR)–DQN to optimize heating, ventilation, and air conditioning systems.
Multi-objective optimization in CCHP systems[58]	Energy engineering	TD3	TD3 combined with photovoltaic (PV) and storage systems for seasonal operation strategies.
Flavor molecule design optimization[59]	Food engineering	Q-learning	Combines generative models with DRL to optimize flavor molecule design.

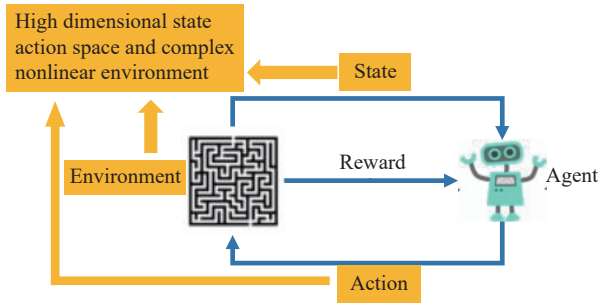

图 4　强化学习框架中复杂系统的问题

Fig.4　Problems with complex systems in reinforcement learning framework

况下实现精确的温度控制, Liu 等[65] 通过时间因果网络与基于 DQN 的多控制器集成方法相结合, 实现了锌电解过程中电解液温度的精确控制, 提升了电流效率并降低了能耗. Yang 等[66] 针对能源传输受限的复杂岛屿群能源管理问题提出了混合策略强化学习(HPRL)方法, 通过离散–连续混合动作空间的优化, 显著提高了岛屿间的能源传输效率和能源利用率. Jendoubi 和 Bouffard[67] 则通过多智能体分层强化学习(MARL), 用于协调不同设备的运行, 每个智能体独立决策, 避免通信瓶颈同时保持较高的协调性, 实现了分布式电力系统中多个独立设备的高效协同控制, 解决了传统的集中式能源管理方法难以适应高维状态空间、多目标优化和不完全观测等挑战. Wang 等[68] 的研究重点在于极端灾后情况下的微电网恢复能力, 通过 MARL 调度移动电源和维修队伍, 在不依赖中央控制的情况下, 实现了负荷恢复的最大化, 提高了系统的恢复效率. 这种多智能体的分布式技术为未来复杂智能电网的恢复控制提供了重要的技术支持. Xu 等[69] 在电动汽车的能源管理中, 针对混合能源存储系统(HESS)在能量流管理的复杂问题上, 利用 SAC 算法进行了优化, 解决了传统方法中能量损失和系统收敛速度慢的问题. 通过并行计算和动态规划嵌入, SAC 算法提高了系统的适应性和收敛效率, 降低了能量损失, 提升了能源管理的整体性能. Zhu 等[70] 的研究针对具有高维状态空间的复杂化工过程, 开发了因子化快速核动态策略编程(FFDPP)算法, 用于解决高维状态空间下的乙酸乙烯酯(VAM)生产控制问题, 实现了更高的控制精度和稳定性. Croll 等[71] 在过程复杂且非线性的废水处理领域, 通过对 DQN、TD3、PPO 等强化学习算法的系统评估, 显著优化了能源消耗并保持了处理效果的稳定性. Choi 等[72] 则在具有

高维、连续状态空间且任务复杂的航空航天领域应用模块化强化学习和课程学习, 成功提升了无人机自主飞行控制的效率和任务完成率. Hu 等[73]研究了耗时且成本高昂的药物发现过程, 现有的算法在生成分子多样性方面存在不足, 导致药物候选分子缺乏多样性和新颖性, 影响药物开发的成功率. Hu 等[73]提出了基于 GPT 模型的 MARL 框架, 用于生成多样性高且具有理想生物活性的药物分子, 特别是在抗 SARS-CoV-2 药物开发中表现出色. Dong 等[74]研究了大规模电动汽车与电网的高效整合的调度机制, 提出了 MARL 模型, 优化了电动汽车在电网中的放电调度, 有效削减了电网高峰负荷, 避免了局部放电高峰的产生. 表 4 总结了复杂系统的性能增强的相关研究.

DRL 在多个领域展现了强大的问题解决能力, 能够应对不同工业应用场景中的复杂任务. 通过现有研究能分析出将 DRL 应用于工业场景的关键是深入挖掘现实工业场景的实体问题, 将实体问题抽象封装在 MDP 框架中, 在这个基础上再针对 MDP 框架不完善的部分选择合适的 DRL 算法或者提出创新的解决思路. 不同的 DRL 算法均有应用, 这是由于不同算法适应于不同的工业问题, 需结合现场情况和实验效果综合分析. 对于算法评估来说, 部分工业场景允许进行 DRL 算法的在线评估, 这无疑对于现场智能化水平的提高有极大助力. 但也存在一些工业场景由于安全等因素无法为 DRL 算法提供在线评估的环境, 只能通过不准确的仿真模型来进行评估, 这极大限制了工业现场智能化水平的发展. 未来如何提升仿真模型的精度, 或者挖掘如何在保障绝对安全的情况下进行算法评估, 还需要更多的探索. 总的来说, 面向工业场景的 DRL 通过智能调度、能效优化和系统控制, 实现了生产效率的提升、能源消耗的减少以及系统鲁棒性的增强. 这些研究为未来工业

表 4 复杂系统的性能增强研究总结

Table 4 Summary of research on performance enhancement of complex systems

Research problem	Field	Algorithm	Algorithm innovation
Flotation process control[60]	Mining engineering	CDQL	Two-layer control structure using DRL, with feature extraction via deep learning.
Flotation process control[61]	Mining engineering	Hybrid Model-Based RL	Combines physical models with data-driven approaches for higher sample efficiency and precise control.
Flotation process control[62]	Mining engineering	Interleaved Learning RL	New interleaved learning RL outperforming standard policy and value iteration methods.
Thickener underflow concentration control[63]	Mining engineering	Heuristic Critic Network	Dual-network architecture with short-term experience replay for enhanced adaptability.
Thickener underflow concentration control[64]	Mining engineering	TBCQ	Time-series state representation based on batched-constrained deep-Q learning (BCQ), suitable for partially observable Markov decision processes (MDPs).
Electrolyte temperature control[65]	Metallurgical engineering	DQN	Combines temporal causal networks with RL for precise control, improves current efficiency, and reduces energy consumption.
Energy transmission in islands[66]	Energy engineering	Hybrid Policy-Based RL	Optimizes discrete-continuous action spaces to improve energy transmission efficiency.
Distributed power coordination[67]	Energy engineering	MARL	Multi-agent hierarchical RL for efficient coordination and multi-objective scheduling.
Post-disaster microgrid recovery[68]	Energy engineering	MARL	Multi-agent RL for scheduling mobile power sources and repair teams, maximizing load recovery.
Hybrid energy storage in EVs[69]	Energy engineering	SAC	Parallel computation and dynamic programming to enhance adaptability, efficiency, and reduce energy loss.
VAM production control[70]	Chemical engineering	FFDPP	Factorized fast kernel dynamic programming (FFDPP) for improved control precision and stability.
Wastewater treatment energy optimization[71]	Environmental engineering	TD3	Evaluates RL algorithms to optimize energy use while maintaining treatment stability.
Autonomous UAV flight control[72]	Aerospace	SAC	Modular RL to improve unmanned aerial vehicle (UAV) autonomy, flight control efficiency, and task completion rates.
Drug molecule generation[73]	Pharmaceutical design	MolRL–MGPT	Multi-agent RL with generative pretraining transformer (GPT) for generating diverse bioactive drug molecules.
EV discharge scheduling[74]	Transportation	Actor–Critic	Multi-agent collaboration for optimizing electric vehicle (EV) discharge scheduling.

和环境中的智能控制与优化提供了广阔的应用前景.

4　结论与研究展望

本文详细探讨了 DRL 的基本原理和算法, 并讨论了在不同工业领域的应用. DRL 在工业应用中展现了巨大的潜力, 但在实际场景中也面临许多挑战. 未来的研究可以在以下几个方面展开, 以推动 DRL 在工业智能控制中的应用:

(1) 提高数据质量[75]: 开发数据增强和合成技术, 以扩展和增强现有数据集, 提高模型训练效果. 或者研究半监督和自监督学习方法, 利用未标注的数据提高模型性能, 减少对高质量标注数据的依赖.

(2) 提高样本效率[76] 与高维空间处理[14] 能力: 设计高效的探索策略和基于模型的强化学习方法, 以提高样本效率, 减少实际操作中的实验成本. 研究有效的特征提取和降维技术, 开发新的算法以改进高维状态和动作空间中的策略优化能力.

(3) 提高环境建模与模拟[77] 能力: 开发高保真度的工业过程模拟器, 确保训练和验证的可靠性. 研究在线环境建模和更新技术, 使模拟环境能够动态反映真实世界的变化.

(4) 提高模型泛化能力[78]: DRL 模型的泛化能力是另一个关键问题. 在工业控制中, 环境条件可能会发生变化, 如设备老化、外部干扰等. 这要求 DRL 模型能够在各种情况下保持良好的性能. 如何提高模型的泛化能力, 以适应不同的操作条件和环境变化, 是研究的重要方向.

(5) 提高安全性、稳定性[20] 与可解释性[79]: 在 DRL 算法中引入安全约束, 确保控制策略的安全性和稳定性. 开发鲁棒性训练技术, 提高模型在不同环境变化中的适应性. 设计决策过程可视化工具, 帮助用户理解和分析 DRL 模型的行为和决策依据.

总之, DRL 在工业智能控制中的应用具有广阔的前景, 但要实现其潜力, 需要解决现有挑战并进行持续的研究和创新. 通过跨学科的合作和技术的不断进步, DRL 有望在未来的工业智能控制中发挥越来越重要的作用.

参　考　文　献

[1] Le N, Rathour V S, Yamazaki K, et al. Deep reinforcement learning in computer vision: A comprehensive survey. *Artif Intell Rev*, 2022, 55(4): 2733

[2] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: a survey. *Artif Intell Rev*, 2022, 55(2): 895

[3] Chen W H, Qiu X Y, Cai T, et al. Deep reinforcement learning for Internet of Things: A comprehensive survey. *IEEE Commun Surv Tutor*, 2021, 23(3): 1659

[4] Mosavi A, Faghan Y, Ghamisi P, et al. Comprehensive review of deep reinforcement learning methods and applications in economics. *Mathematics*, 2020, 8(10): 1640

[5] Wang X, Wang S, Liang X X, et al. Deep reinforcement learning: A survey. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(4): 5064

[6] Ladosz P, Weng L L, Kim M, et al. Exploration in deep reinforcement learning: A survey. *Inf Fusion*, 2022, 85: 1

[7] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. *Nature*, 2017, 550(7676): 354

[8] Vinyals O, Babuschkin I, Czarnecki W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019, 575(7782): 350

[9] Levine S, Kumar A, Tucker G, et al. Offline reinforcement learning: Tutorial, review, and perspectives on open problems [J/OL]. *arXiv preprint* (2020–05–04) [2024–10–29]. https://arxiv.org/abs/2005.01643

[10] Sutton Richard S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press, 1998.

[11] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, 521(7553): 436

[12] Li X R, Ban X J, Yuan Z L, et al. Review on deep learning models for time series forecasting in industry. *Chin J Eng*, 2022, 44(4): 757
(李潇睿, 班晓娟, 袁兆麟, 等. 工业场景下基于深度学习的时序预测方法及应用. 工程科学学报, 2022, 44(4): 757)

[13] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518(7540): 529

[14] Sutton R S, McAllester D, Singh S, et al. Policy gradient methods for reinforcement learning with function approximation // *Proceedings of the* 12*th International Conference on Neural Information Processing Systems*. Denver, 1999: 1057

[15] Konda V R, Tsitsiklis J N. Actor-critic algorithms // *Proceedings of the* 12*th International Conference on Neural Information Processing Systems*. Denver, 1999: 1008

[16] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning [J/OL]. *arXiv preprint* (2015–09–09) [2024–10–29]. https://arxiv.org/abs/1509.02971

[17] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization // *Proceedings of the* 32*nd International Conference on International Conference on Machine Learning - Volume* 37. Lille, 2015: 1889

[18] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms [J/OL]. *arXiv preprint* (2017–07–20) [2024–10–29]. https://doi.org/10.48550/arXiv.1707.06347

[19] Hiraoka T, Imagawa T, Hashimoto T, et al. Dropout Q-functions for doubly efficient reinforcement learning [J/OL]. *arXiv preprint*

(2021–10–05) [2024–10–29]. https://arxiv.org/abs/2110.02034

[20] Bhatt A, Palenicek D, Belousov B, et al. CrossQ: Batch normalization in deep reinforcement learning for greater sample efficiency and simplicity [J/OL]. *arXiv preprint* (2024–03–25) [2024–10–29]. https://arxiv.org/abs/1902.05605

[21] Fujimoto S, Chang W D, Smith E, et al. For sale: State-action representation learning for deep reinforcement learning // 37*th Conference on Neural Information Processing System*. New Orfeans, 2023

[22] Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration // *Proceedings of the* 36*th International Conference on Machine Learning*. Long Beach, 2019: 2052

[23] Kumar A, Fu J, Tucker G, et al. Stabilizing off-policy Q-learning *via* bootstrapping error reduction [J/OL]. *arXiv preprint* (2019–06–03) [2024–10–29]. https://arxiv.org/abs/1906.00949

[24] Wu Y F, Tucker G, Nachum O. Behavior regularized offline reinforcement learning [J/OL]. *arXiv preprint* (2019–11–26) [2024–10–29]. https://arxiv.org/abs/1911.11361

[25] Fujimoto S, Gu S S, Fujimoto S, et al. A minimalist approach to offline reinforcement learning // *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. Sydney, 2021: 20132

[26] Kumar A, Zhou A, Tucker G, et al. Conservative Q-learning for offline reinforcement learning // *Proceedings of the* 33*th International Conference on Neural Information Processing Systems*. Vancouver, 2020: 1179

[27] Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning [J/OL]. *arXiv preprint* (2021–10–12) [2024–10–29]. https://arxiv.org/abs/2110.06169

[28] Stouffer K, Falco J, Scarfone K. Guide to industrial control systems (ICS) security (final draft). *NIST special publica*, 2011, 800(82): 16

[29] Johnson M A, Moradi M H. *PID Control*. London: Springer-Verlag, 2005

[30] Driankov D. *An Introduction to Fuzzy Control*. Berlin: Springer Science & Business Media, 2013

[31] Waterman D A. *A Guide to Expert Systems*. Reading: Addison-Wesley Longman Publishing Co. , Inc. , 1985

[32] Kumar A, Dimitrakopoulos R. Production scheduling in industrial mining complexes with incoming new information using tree search and deep reinforcement learning. *Appl Soft Comput*, 2021, 110: 107644

[33] Levinson Z, Dimitrakopoulos R, Keutchayan J. Simultaneous stochastic optimization of an open-pit mining complex with preconcentration using reinforcement learning. *Appl Soft Comput*, 2023, 138: 110180

[34] Huo D, Sari Y A, Kealey R, et al. Reinforcement learning-based fleet dispatching for greenhouse gas emission reduction in open-pit mining operations. *Resour Conserv Recycl*, 2023, 188: 106664

[35] Noriega R, Pourrahimian Y, Askari-Nasab H. Deep Reinforcement Learning based real-time open-pit mining truck dispatching system. *Comput Oper Res*, 2025, 173: 106815

[36] Cao P, Liu J, Wang Y, et al. Inversion of mine ventilation resistance coefficients enhanced by deep reinforcement learning. *Process Saf Environ Prot*, 2024, 182: 387

[37] Kumar A, Dimitrakopoulos R. Updating geostatistically simulated models of mineral deposits in real-time with incoming new information using actor-critic reinforcement learning. *Comput Geosci*, 2022, 158: 104962

[38] Jiang K, Jiang Z H, Jiang X D, et al. Reinforcement learning for blast furnace ironmaking operation with safety and partial observation considerations. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(3): 3077

[39] Liu C, Tang L X, Zhao C C. A novel dynamic operation optimization method based on multiobjective deep reinforcement learning for steelmaking process. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(3): 3325

[40] Shi T, Xu C, Dong W H, et al. Research on energy management of hydrogen electric coupling system based on deep reinforcement learning. *Energy*, 2023, 282: 128174

[41] Meng Q L, Hussain S, Luo F Z, et al. An online reinforcement learning-based energy management strategy for microgrids with centralized control. *IEEE Trans Ind Appl*, 2024, 61(1): 1

[42] Yin X X, Lei M Z. Jointly improving energy efficiency and smoothing power oscillations of integrated offshore wind and photovoltaic power: A deep reinforcement learning approach. *Prot Control Mod Power Syst*, 2023, 8(2): 1

[43] Li T H, Yang J, Ioannou A. Data-driven control of wind turbine under online power strategy *via* deep learning and reinforcement learning. *Renewable Energy*, 2024, 234: 121265

[44] Wang X H, Zhang L, Liu Y K, et al. Logistics-involved task scheduling in cloud manufacturing with offline deep reinforcement learning. *J Ind Inf Integr*, 2023, 34: 100471

[45] Yun L X, Wang D, Li L. Explainable multi-agent deep reinforcement learning for real-time demand response towards sustainable manufacturing. *Appl Energy*, 2023, 347: 121324

[46] Zhu A Y, Dai T H, Xu G Y, et al. Deep reinforcement learning for real-time assembly planning in robot-based prefabricated construction. *IEEE Trans Autom Sci Eng*, 2023, 20(3): 1515

[47] Cui Y D, Peng L, Li H Y. Filtered probabilistic model predictive control-based reinforcement learning for unmanned surface vehicles. *IEEE Trans Ind inf*, 2022, 18(10): 6950

[48] Liang H P, Yang C H, Lv M J, et al. Zinc roasting temperature field control with CFD model and reinforcement learning. *Adv Eng inf*, 2024, 59: 102332

[49] Tang F R, Feng Z X, Li Y G, et al. A constrained multi-objective deep reinforcement learning approach for temperature field optimization of zinc oxide rotary volatile kiln. *Adv Eng inf*, 2023, 58: 102197

[50] Che G, Zhang Y Y, Tang L X, et al. A deep reinforcement learning based multi-objective optimization for the scheduling of oxygen

production system in integrated iron and steel plants. *Appl Energy*, 2023, 345: 121332

[51] Neto W A F, Cavalcante C A V, Do P. Deep reinforcement learning for maintenance optimization of a scrap-based steel production line. *Reliab Eng Syst Saf*, 2024, 249: 110199

[52] Liu C Y, Zhang Y L, Mao S J. Image classification method based on multi-agent reinforcement learning for defects detection for casting. *Sensors*, 2022, 22(14): 5143

[53] Ma L B, Li N, Guo Y N, et al. Learning to optimize: Reference vector reinforcement learning adaption to constrained many-objective optimization of industrial copper burdening system. *IEEE Trans Cybern*, 2022, 52(12): 12698

[54] Canales C, Díaz-Quezada S, Leiva F, et al. Control of heap leach piles using deep reinforcement learning. *Miner Eng*, 2024, 212: 108707

[55] Zheng J, Jia R D, Liu S N, et al. Safe reinforcement learning for industrial optimal control: A case study from metallurgical industry. *Inf Sci*, 2023, 649: 119684

[56] Wang Y, Qiu D W, Sun M Y, et al. Secure energy management of multi-energy microgrid: A physical-informed safe reinforcement learning approach. *Appl Energy*, 2023, 335: 120759

[57] Qin H S, Yu Z, Li T L, et al. Energy-efficient heating control for nearly zero energy residential buildings with deep reinforcement learning. *Energy*, 2023, 264: 126209

[58] Ruan Y J, Liang Z Y, Qian F Y, et al. Operation strategy optimization of combined cooling, heating, and power systems with energy storage and renewable energy based on deep reinforcement learning. *J Build Eng*, 2023, 65: 105682

[59] Queiroz L P, Rebello C M, Costa E A, et al. A reinforcement learning framework to discover natural flavor molecules. *Foods*, 2023, 12(6): 1147

[60] Ai M X, Xie Y F, Tang Z H, et al. Deep learning feature-based setpoint generation and optimal control for flotation processes. *Inf Sci*, 2021, 578: 644

[61] Zheng J, Jia R D, Liu S N, et al. Sample-efficient reinforcement learning with knowledge-embedded hybrid model for optimal control of mining industry. *Expert Syst Appl*, 2024, 254: 124402

[62] Jiang Y, Fan J L, Chai T Y, et al. Data-driven flotation industrial process operational optimal control based on reinforcement learning. *IEEE Trans Ind inf*, 2018, 14(5): 1974

[63] Yuan Z L, He R Z, Yao C, et al. Online reinforcement learning control algorithm for concentration of thickener underflow. *Acta Auto Sini*, 2019, 45: 1

[64] Yuan Z L, Zhang Z X, Li X R, et al. Controlling partially observed industrial system based on offline reinforcement learning—a case study of paste thickener. *IEEE Trans Ind inf*, 2025, 21(1): 49

[65] Liu T H, Yang C H, Zhou C, et al. Integrated optimal control for electrolyte temperature with temporal causal network and reinforcement learning. *IEEE Trans Neural Netw Learn Syst*, 2024, 35(5): 5929

[66] Yang L X, Li X F, Sun M W, et al. Hybrid policy-based reinforcement learning of adaptive energy management for the energy transmission-constrained island group. *IEEE Trans Ind inf*, 2023, 19(11): 10751

[67] Jendoubi I, Bouffard F. Multi-agent hierarchical reinforcement learning for energy management. *Appl Energy*, 2023, 332: 120500

[68] Wang Y, Qiu D W, Teng F, et al. Towards microgrid resilience enhancement *via* mobile power sources and repair crews: A multi-agent reinforcement learning approach. *IEEE Trans Power Syst*, 2024, 39(1): 1329

[69] Xu D Z, Cui Y D, Ye J Y, et al. A soft actor-critic-based energy management strategy for electric vehicles with hybrid energy storage systems. *J Power Sources*, 2022, 524: 231099

[70] Zhu L W, Cui Y D, Takami G, et al. Scalable reinforcement learning for plant-wide control of vinyl acetate monomer process. *Control Eng Pract*, 2020, 97: 104331

[71] Croll H C, Ikuma K, Ong S K, et al. Systematic performance evaluation of reinforcement learning algorithms applied to wastewater treatment control optimization. *Environ Sci Technol*, 2023, 57(46): 18382

[72] Choi J, Kim H M, Hwang H J, et al. Modular reinforcement learning for autonomous UAV flight control. *Drones*, 2023, 7(7): 418

[73] Hu X Y, Liu G Q, Zhao Y, et al. *De novo* drug design using reinforcement learning with multiple GPT agents // *Proceedings of the* 37*th International Conference on Neural Information Processing Systems*. New Orleans, 2024: 7405

[74] Dong J W, Yassine A, Armitage A, et al. Multi-agent reinforcement learning for intelligent V2G integration in future transportation systems. *IEEE Trans Intell Transp Syst*, 2023, 24(12): 15974

[75] Feng W H, Han C Z, Lian F, et al. A data-efficient training method for deep reinforcement learning. *Electronics*, 2022, 11(24): 4205

[76] Xie H M, Xu X H, Li Y L, et al. Model predictive control guided reinforcement learning control scheme // 2020 *International Joint Conference on Neural Networks* (*IJCNN*). Glasgow, 2020: 1

[77] Kang K T, Belkhale S, Kahn G, et al. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight // 2019 *International Conference on Robotics and Automation* (*ICRA*). Montreal, 2019: 6008

[78] Swazinna P, Udluft S, Runkler T. Overcoming model bias for robust offline deep reinforcement learning. *Eng Appl Artif Intell*, 2021, 104: 104366

[79] Gupta C, Farahat A. Deep Learning for Industrial AI: Challenges, New Methods and Best Practices // *Proceedings of the* 26*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event, 2020: 3571