排灌机械工程学报
Journal of Drainage and Irrigation Machinery Engineering
2023 年 6 月 June 2023 | 第 41 卷 第 6 期 Vol.41 No.6

深度强化学习在自动控制领域研究现状与展望


曹凯

曹凯[1]，朱勇[1]，高强[1*]，刘金华[2]
（1. 江苏大学国家水泵及系统工程技术研究中心，江苏 镇江 212013；2. 共青科技职业学院 国际航运研究院，江西 九江 332020）

摘要： 针对深度强化学习在自动控制领域的成功应用以及其在泛化性、鲁棒性、可靠性等方面存在的不足，综述了深度强化学习算法的研究进展以及其在自动控制领域的应用概况.首先，简要阐述了深度强化学习算法的发展历程与基本原理；其次，将深度强化学习算法分为基于值函数和基于策略梯度 2 类算法，分别对这 2 类算法的基本原理、数学模型以及改进方法进行了系统的论述与分析；再次，综述了深度强化学习算法在无人机飞行控制、移动机器人轨迹控制、车辆自动驾驶控制以及液压伺服控制等控制领域的应用，在此基础上归纳总结了深度强化学习算法在不同控制问题中的优势与不足；最后，针对深度强化学习算法在自动控制领域面临的挑战与发展趋势进行了总结和展望，根据近年来自动控制领域前沿研究成果提出优化解决深度强化学习关键问题的研究思路并做出相应论证.

关键词： 人工智能；自动控制领域；深度强化学习
中图分类号： TP181　**文献标志码：** A　**文章编号：** 1674-8530(2023)06-0638-11
DOI：10.3969/j.issn.1674-8530.22.0108

曹凯，朱勇，高强，等.深度强化学习在自动控制领域研究现状与展望[J].排灌机械工程学报，2023，41(6)：638-648.
CAO Kai，ZHU Yong，GAO Qiang，et al. Research status and prospect of deep reinforcement learning in automatic control[J]. Journal of drainage and irrigation machinery engineering(JDIME)，2023，41(6)：638-648.(in Chinese)

Research status and prospect of deep reinforcement learning in automatic control

CAO Kai[1]，*ZHU Yong*[1]，*GAO Qiang*[1*]，*LIU Jinhua*[2]
(1. National Research Center of Pumps，Jiangsu University，Zhenjiang，Jiangsu 212013，China；2. International Shipping Research Institute，Gongqing Institute of Science and Technology，Jiujiang，Jiangxi 332020，China)

Abstract: Aiming at the successful application of deep reinforcement learning in automatic control field and its shortcomings in generalization，robustness and reliability，the research progress of deep reinforcement learning and its application in automatic control field were summarized. Firstly，the development process and basic principle of deep reinforcement learning were briefly described. Secondly，deep reinforcement learning was divided into two types of algorithms：value function-based algorithm and policy gradient-based algorithm. Then，the basic principles，mathematical models，and improved methods of these two kinds of algorithms were discussed and analyzed systematically. Moreover，the applications of deep reinforcement learning in automatic control fields such as UAV flight control，mobile robot trajectory control，vehicle automatic driving control and hydraulic servo control were summarized.

收稿日期： 2022-04-19；**修回日期：** 2022-08-08；**网络出版时间：** 2023-05-15
网络出版地址： https://kns.cnki.net/kcms/detail/32.1814.TH.20230512.1436.012.html
基金项目： 国家自然科学基金资助项目(52175052)；国家重点研发计划项目(2020YFC1512402)
第一作者简介： 曹凯(1995—)，男，江苏南通人，硕士研究生(2222011002@stmail.ujs.edu.cn)，主要从事流体传动与智能控制研究.
通信作者简介： 高强(1990—)，男，江苏镇江人，助理研究员，博士(gaoqiang116@ujs.edu.cn)，主要从事流体传动与智能控制研究.

On this basis, the advantages and disadvantages of the deep reinforcement learning algorithm were summarized. Finally, the challenges and development trends of deep reinforcement learning in the field of automatic control were summarized and prospected. According to the cutting-edge research achievements in the field of automatic control in recent years, the research ideas of optimizing and solving the key problems of deep reinforcement learning were put forward and demonstrated accordingly.

Key words: artificial intelligence; automatic control field; deep reinforcement learning

自动控制技术通过自动化装置代替人对仪器设备进行控制，使之达到预期的状态或性能指标，能够将操作人员从复杂、危险、繁琐的劳动环境中解放出来，对工业生产，尤其是对恶劣环境下的控制操作具有重要的应用价值.随着自动控制技术在机器人、航空航天、智能制造等高端装备领域的逐步应用，被控对象的结构逐渐趋于复杂化，尽管传统控制理论在可观测性与稳定性方面具有一定优势，但在面向复杂系统控制问题时仍需要人为干预，故无法满足未来自动控制领域提升自感知、自学习、自决策、自执行等方面的迫切需求.近年来，国内外研究人员将传统控制理论与模糊逻辑、神经网络、遗传算法等新技术相结合，以提高控制技术的自适应能力与初步模拟人类学习的能力，但由于该方法无法处理复杂环境中的高维数据以及易产生局部最优问题，导致其无法解决如建立精确数学模型一类的控制问题.因此，具有信息自主感知、环境交互与试错能力的人工智能学习算法已成为目前自动控制领域的研究热点.

与传统控制算法相比，强化学习算法在无需先验知识的条件下，通过直接与环境互动进行奖惩反馈以促使智能体不断学习，最终实现对决策问题的自适应处理.强化学习算法凭借环境交互与试错学习的特点，在解决动态与随机型问题时具有一定的优势[1].然而，传统的强化学习算法由于受到动作空间与样本空间维数的限制，在实际控制任务中极易出现维数灾难问题.同时，通过传统强化学习方法实现设备自动控制需要进行预训练，时间成本较高且无法保证训练成功率，因此在解决对时效性、准确性要求较高的自动控制问题时存在一定局限性.与强化学习算法相比，深度学习具有良好的信息感知能力、复杂环境适应性以及网络模型跨领域可移植性[2]，但缺乏一定决策能力.基于此，研究人员通过有效融合强化学习与深度学习的优缺点，最终构建了深度强化学习(deep reinforcement learning, DRL)算法[3].

DRL 算法同时具备深度学习算法的高维空间信息感知能力与强化学习算法的环境自主交互、试错学习等能力.经典 DRL 算法的性能已在 Atari2600 游戏中得到了验证，但在记忆、认知、推理等高级自学习能力方面尚有欠缺[4].由于被控对象需要与环境不断进行交互，导致 DRL 算法需要学习大量网络参数，因此在状态信息部分可观测、奖励延迟以及数学模型不确定等复杂任务中适用性不佳，无法满足智能控制技术发展需求.针对上述问题，研究人员对 DRL 算法从算法训练、网络结构、学习机制等方面进行深入研究，并提出相应的改进策略.

文中聚焦于 DRL 算法在自动控制领域中的研究历程和现状.首先，介绍 DRL 算法的基本原理与数学模型；其次，以无人机飞行控制、移动机器人轨迹控制、车辆自动驾驶控制以及液压伺服控制等 4 个自动控制领域典型应用场景为例，详细阐述 DRL 算法在国内外的研究现状；最后，总结国内外学者在 DRL 算法方面的研究进展，进一步对 DRL 算法在自动控制领域的发展趋势进行展望.

1　典型深度强化学习算法介绍

作为当前人工智能领域的研究热点，DRL 算法充分融合了深度学习对数据极强的特征表示能力以及强化学习的时序决策能力，同时能够直接通过原始输入数据进行动作选择，可以满足自动控制中最优决策问题所面临的场景复杂、非线性以及时变性的特殊需求.DRL 算法已经出现多种改进型，根据模型训练过程中是否对环境进行直接建模，DRL 算法可以分为基于模型和无模型 2 种类型[5]，如图 1 所示.基于模型的 DRL 算法尝试根据学习到的环境模型来选择最优策略，主要包括微调算法[6]、智能增强体算法[7]等；对于无模型 DRL 算法而言，最优策略是通过智能体直接与任务环境交互的试错方式来获得，主要包括基于值函数和基于策略梯度 2 种类型[8].在自动控制领域，由于被控对象一般具有非线性、时变性以及不确定性等特点，导致其高精度数学模型难以构建.相较于基于模型的 DRL 算

法,无模型 DRL 算法凭借泛化能力强与算法结构简单等优势,在自动控制领域仍占据主导地位,对此文中将重点予以阐述.

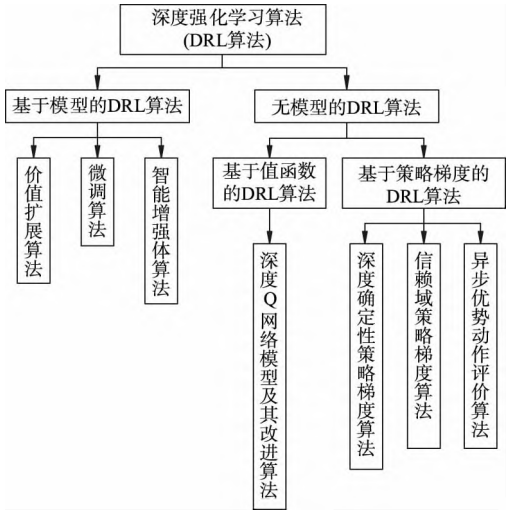


图 1 DRL 算法分类示意图

Fig.1 Schematic diagram of DRL algorithm classification

1.1 基于值函数的 DRL 算法

深度 Q 网络(deep Q network,DQN)是 MNIH 等[9]基于强化学习开发出的一种新型人工智能算法,该算法利用卷积神经网络代替传统 Q 学习的近似价值函数,以实现 DQN 算法任务的最优策略学习.

DQN 算法模型采用经验库代替传统 Q 学习表格以实现对智能体 Q 值的记录,此外将每个时间点的环境探索数据储存为记忆单元(s,a,r,s'),并利用卷积神经网络对价值函数进行近似.在 DQN 算法模型中,深度卷积神经网络权重参数 θ 随训练过程实时更新,并通过网络模型 $Q(s,a,\theta_i)$ 模拟动作值函数 $Q^{\pi}(s,a)$,即

$$Q(s,a,\theta_i)=Q^{\pi}(s,a), \tag{1}$$

式中:s 和 a 为当前时间步的状态和动作;下标 i 为迭代次数;$\theta^{\pi}(s,a)$ 为策略 π 在状态 s 下采取动作 a 的长期期望收益.

Q 网络每次迭代的优化目标值由单独目标网络产生,计算公式为

$$y=r+\gamma\max_{a'}Q(s',a',\theta_i^-), \tag{2}$$

式中:r 为当前状态的奖励值;s'和 a'为下一时间步的状态和动作;γ 为折扣因子;θ_i^- 为神经元连接权重.

利用梯度下降算法更新网络参数 θ,梯度公式为

$$\nabla_{\theta_i}L_i(\theta_i)=E_{(s,a,r,s')}\{[y-Q(s,a,\theta_i)]\cdot\nabla_{\theta_i}Q(s,a,\theta_i)\}. \tag{3}$$

根据式(1)—(3)可以看出,DQN 算法首先将环境状态 s 输入到当前神经网络,并在执行动作 a 后与环境进行交互学习,以更新状态 s'和动作 a';其次,将(s,a,r,s')储存到样本库中,并通过构造损失函数以对样本库中任意样本进行训练;最后,根据训练结果实时更新网络参数.通过上述框架设计可以实现 DRL 算法的端到端控制,并提高算法的更新效率与收敛速度.

1.2 基于策略梯度的 DRL 算法

基于策略梯度的 DRL 算法主要包括深度确定性策略梯度(deep deterministic policy gradient,DDPG)[10]、信赖域策略优化(trust region policy optimization,TRPO)[11]和异步优势演员-评论家(asynchronous advantage actor-critic,A3C)[12]等类型算法,如图 2 所示.

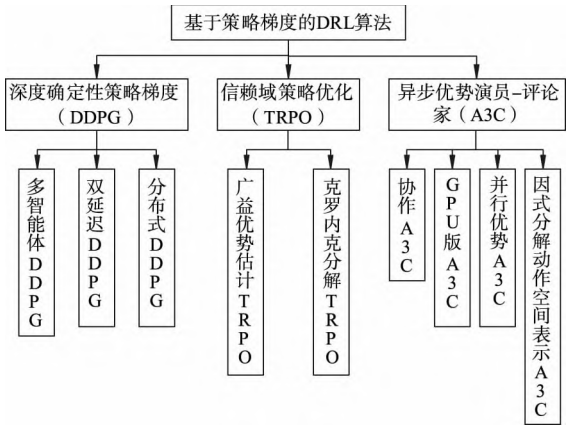


图 2 基于策略梯度的 DRL 算法分类

Fig.2 DRL algorithm classification based on policy gradient

策略梯度算法直接使用逼近器来近似表示和优化策略,通过不断计算策略期望总奖赏关于策略参数的梯度来更新策略参数,最终得到最优策略.一个片段内所获得的奖赏总和表示为

$$R=\sum_{t=0}^{T-1}r_t, \tag{4}$$

式中:r_t 为每个时间 t 内得到的奖励;T 为每个情节内算法模型需要优化的参数个数.

策略期望总奖赏在算法迭代中不断优化,任意策略 π 带来的期望总奖赏表示为

$$\max_{\theta}E[R\mid\pi_{\theta}]. \tag{5}$$

当状态为 s 时,若动作 a 符合参数为 θ 的某个概率分布,表现为随机性策略 $\pi_{\theta}(a\mid s)=P[a\mid s;\theta]$,此时相同状态可以对应不同动作,随机策略梯度公

式为

$$\nabla_\theta L(\pi_\theta) = E_{s\sim\rho^\pi, a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s)\, Q^\pi(s,a)]. \tag{6}$$

对于确定性策略 $a=\mu_\theta(s)$，每个状态与动作唯一且对应.确定性策略梯度公式为

$$\nabla_\theta L(\mu_\theta) = E_{s\sim\rho^\mu}[\nabla_\theta \mu_\theta(s)\ \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}]. \tag{7}$$

相较于基于值函数的 DRL 算法，基于策略梯度的 DRL 算法以端到端方式在策略空间中直接搜索最优策略，结构较为简单，因此更适用于处理连续和高维动作空间等问题.

2　深度强化学习在自动控制领域的应用

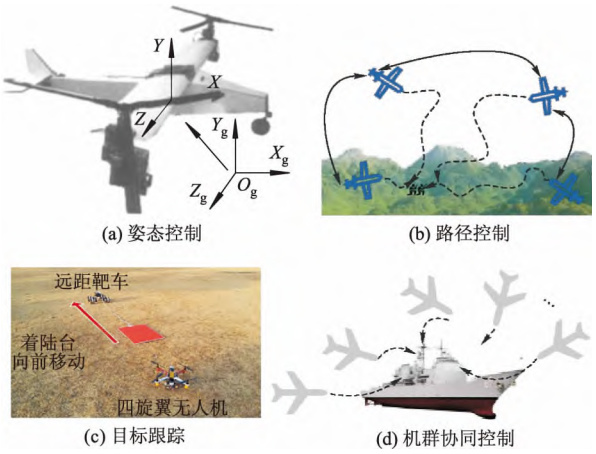
自动控制技术能够在无人类直接干预的情况下采用自动化仪器或自动控制装置对设备进行控制，使之完成预定任务，自 20 世纪 40 年代出现以来一直与工业发展联系紧密.随着现代科技的发展，控制对象、控制器与控制任务日益复杂，基于时域法的传统控制理论难以满足需求.20 世纪 70 年代，普渡大学 FU [13] 首次将人工智能的自学习与自适应能力应用于控制系统，自此智能控制成了国内外学术界的研究热点.

DRL 算法作为深度学习领域快速发展的一个分支，能够为计算机从感知到决策提供解决方案，从而实现智能控制.近年来，研究人员针对传统 DRL 算法及其模型做出诸多改进，如在感知与学习阶段引入新的学习算法，以改进样本采集与处理过程；在决策控制阶段，通过阶段优化网络模型结构来减少行为策略的计算量与样本数，最终提高算法的收敛速度及鲁棒性.可以看出，对传统 DRL 算法的改进可以提高其在自动控制领域的适应能力.目前 DRL 算法已逐步应用于无人机飞行控制、移动机器人轨迹控制、车辆自动驾驶控制以及液压伺服控制等典型自动控制领域.

2.1　深度强化学习在无人机飞行控制领域的应用

近年来，无人机凭借机动灵活、成本低廉、安全性高等优点，广泛应用于军用和民用领域，如遥感测绘、野外救援、基础设施检查和环境监测等.由于工作环境的复杂性，必然对无人机自主飞行控制能力提出了更高要求.

目前，无人机飞行控制一般采用诸如反馈线性化控制 [14]、模型预测控制 [15]、滑模控制 [16]、反步控制 [17] 等非线性控制算法 [18]，但上述控制方法均需要建立精确的飞行器模型，实际应用难度较大.相对于传统非线性控制方法，DRL 算法在无法获得飞行器精确模型的情况下仍能实现有效控制，故已成为解决无人机路径规划、导航和控制等问题的研究热点.DRL 算法在无人机飞行控制领域的应用如图 3 所示.



(a) 姿态控制　(b) 路径控制　(c) 目标跟踪　(d) 机群协同控制

图 3　DRL 在无人机飞行控制领域的应用

Fig.3　Application of DRL in UAV flight control field

由图 3 可以看出，DRL 算法主要用于解决无人机姿态控制、路径控制、目标跟踪以及机群协同控制等问题.在不同任务场景中，DRL 算法的鲁棒性和实时性将直接影响无人机飞行控制的综合性能.2021 年，HU 等 [19] 提出了一种优先经验回放机制和 DDPG 算法相结合的模型来提高学习收敛速度，并在三维复杂模拟环境中进行了验证.KOCH 等 [20] 在 DDPG 算法基础上结合演员-评论家框架，并设置经验重放缓冲区，通过均匀小批次采样提高数据利用率，最终提升了算法的收敛性与稳定性.周毅等 [21] 在 DQN 框架中引入能效奖励函数和惩罚动作，使无人机能够自主学习最优覆盖部署策略，从而实现自主部署和能效优化等功能.

为了更好地规划飞行路线，强化学习还可以与各种优化器结合使用，例如模仿了狼群在狩猎时行为的灰狼优化算法 [22].灰狼优化算法属于元启发式算法中较为新颖的研究成果，其核心特征在于模拟了灰狼群的领导权等级以及包围和攻击猎物的过程，具有可灵活实现的仿生结构.QU 等 [23] 针对无人机最优路径规划问题，提出了一种无人机路径规划新算法，该算法通过结合强化学习和灰狼优化算法的优点，以实现对无人机勘探、姿态调整与路径规

划的优化控制.

DRL 算法同样可被用于无人机横向与纵向控制,以及无人机群之间的协同控制,例如在固定翼无人机、混合无人机、四旋翼无人机等不同类型无人机中,近端策略优化算法(proximal policy optimization,PPO)对其均能实现快速精确的姿态控制[24-25].控制固定翼无人机编队时,相晓嘉等[26]提出了一种双 DQN 和竞争架构 Q 网络相结合的算法来实现编队协调控制.

综上,与传统控制方法相比,基于 DRL 框架开发的飞行控制方法具有更高的控制精度与更快的响应速度,同时,DRL 算法训练的控制策略可以引导无人机自主空间探索,使无人机在面对未知环境时具备一定的决策能力.使用 DRL 算法标志着无人机的应用领域和控制精度得到进一步发展,因此基于 DRL 算法的无人机飞行控制策略设计流程将会向规范化与模块化发展,最终形成一套成熟的控制器设计流程.

2.2 深度强化学习在移动机器人轨迹控制领域的应用

移动机器人具有运动规划、自主导航、主动避障以及快速部署等功能,因而在野外救援、安全巡防、空间探测等领域发挥着重要作用,同时被广泛应用于农林、航天、物流、应急救援等行业.随着人工智能以及信息融合技术的快速发展,移动机器人的应用越来越智能化,这就需要移动机器人具有一定的理解能力,即在无人干预的情况下,移动机器人能够在未知环境中独立移动,并自主完成任务.

近年来,DRL 算法正逐步取代经典控制技术被应用到移动机器人的轨迹控制领域,在仿真环境和真实场景中均取得了较好的效果,显著提高了移动机器人在非结构环境中的自适应性和复杂任务中的轨迹规划能力.开发并持续完善面向移动机器人轨迹控制领域的 DRL 算法已成为国内外研究热点.DRL 算法在机器人轨迹控制领域的应用如图 4 所示.

在机器人对任务环境的分析和任务目标的引导过程中,准确有效的自主控制策略是实现机器人智能化的关键.为验证 DRL 算法的可行性,TAI 等[27]使用深度图像作为 Q 网络的输入,并在仿真环境中实现了移动机器人躲避多种障碍物的运动规划.CARLUCHO 等[28]使用面向目标的确定策略控制结构,并在演员-评论家网络中输入原始感官信息,最终取得了较好的运动规划效果,但生成的控制策略依赖于学习模型,适应性较差.CARLUCHO 等[29]改进了 A3C 控制结构,该结构只接受底层动态信息作为输入,同时估计多个参数或控制器的增益,解决了不确定环境下运动规划的无监督、无模型问题.XIONG 等[30]利用具有 2 个隐含层的全连接 DDPG 神经网络分别处理多个子任务,并通过并行运算来提高算法鲁棒性.

(a) 运动规划 (b) 自主导航 (c) 自主避障 (d) 快速部署

图 4 DRL 在机器人轨迹控制领域的应用

Fig.4 Application of DRL in robot trajectory control field

在多机器人协同控制研究方面,WANG 等[31]提出了一种多机器人运动规划策略,该策略使用机器人为中心的顶视图与第一人称视图的状态信息作为 DQN 算法的输入变量,并利用多机器人协同、协调以及协作等方式共同执行任务.EHO 等[32]基于 DQN 框架的学习算法来模拟多机器人之间的复杂交互与合作过程,进一步在单机器人预先训练策略的基础上,通过借助对象传输技术来学习多机器人协助策略.

针对避障问题,FENG 等[33]利用 Gazebo 仿真平台模拟了狭小紧凑空间中虚拟飞行器的运动轨迹,仿真结果表明,基于 DRL 算法的避障策略优于势场法和动态窗口法等传统避障策略.LIN 等[34]提出了一种递归神经网络记忆强化学习算法,该算法将包含障碍物信息的原始检测图像数据映射到转向命令,成功实现了无碰撞路径预测控制.

作为实现机器人自主移动的核心难点之一,机器人轨迹控制研究长期受到国内外学者的重点关注,而 DRL 算法的发展为解决移动机器人在复杂环境中的轨迹控制问题提供了新的思路和方向.就研究现状而言,基于 DRL 算法开发的轨迹控制策略仍处于试验模拟阶段,在现实环境应用中,由于受到不确定性因素影响,控制效果难以达到模拟水平.因

此，提高移动机器人轨迹控制方法的鲁棒性和泛化能力是亟待解决的问题之一.

2.3　深度强化学习在车辆自动驾驶控制领域的应用

自动驾驶汽车能够改善交通安全，提高燃油效率，减少拥堵，被认为是构成未来智能交通系统的关键要素.实现车辆自动驾驶的智能控制系统主要由感知子系统、决策子系统和执行子系统组成.随着计算机视觉技术、图像处理技术和智能控制技术的快速发展，自动驾驶系统的环境识别能力和车辆控制能力得到了显著提升，然而有关自动驾驶策略方面的研究仍处于起步阶段.

传统自动驾驶策略是通过模仿人类驾驶经验或建立车辆行驶参数与道路图像数学模型进行学习，但由于泛化能力较弱，故无法适应复杂交通环境.基于 DRL 算法的自动驾驶决策系统使车辆具备自主学习操作驾驶的能力，为真正实现车辆驾驶智能化提供了技术支撑.DRL 算法在车辆自动驾驶控制领域的应用如图 5 所示.

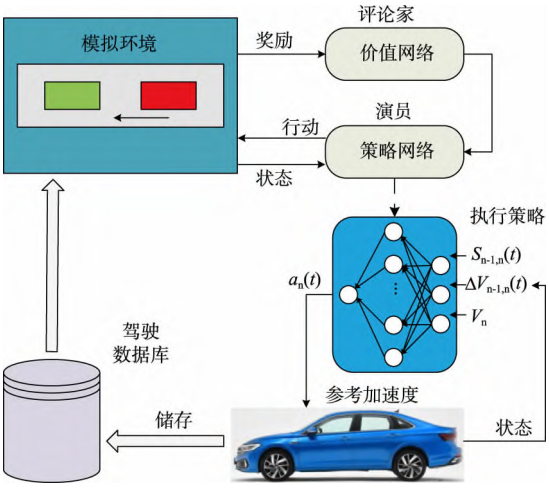


图 5　DRL 在车辆自动驾驶控制领域的应用

Fig.5　Application of DRL in vehicle autopilot control field

由图 5 可以看出，如何实现 DRL 算法与模拟环境的交互学习是实现车辆自动驾驶控制的重点.而对提高控制系统的性能而言，关键在于如何优化 DRL 算法对环境信息的感知与处理能力.ZHU 等[35]提出了一种基于 DDPG 算法的自主跟车策略学习框架，在模拟环境中借助自动驾驶数据与专家经验数据的偏差程度来通过奖励函数从试错交互中构建跟车模型，该模型能够模拟人类进行自动驾驶，并随着驾驶数据的输入进行实时优化.JIN 等[36]提出了一种基于卷积神经网络注意力机制的端到端 DDPG 算法，该算法的演员网络和评论家网络具有相同的对称结构，同时引入注意力机制辅助模型收集环境信息，试验结果表明，该方法在缺少人工决策经验的情况下仍具有较好的性能.YE 等[37]提出了一种将 DDPG 训练算法与高保真虚拟仿真环境相结合的决策训练与学习框架，并通过该框架进行跟车行为训练，结果表明，在保证舒适度的情况下，自动驾驶系统训练效率得到了显著提升.

DQN 及其改进算法凭借结构简单与控制效率较高等优势，被广泛应用于智能驾驶决策系统.为验证 DQN 学习体系在车辆智能驾驶中的可行性，YANG 等[38]分别在二维 Python 和三维 Unity 环境中进行了验证，结果表明，DQN 算法在 2 种不同仿真环境中均构建了最佳自动驾驶策略.为进一步提高复杂城市环境下传统 DQN 算法的收敛速度，KOH 等[39]提出了一种将多层神经网络作为 Q 值状态映射函数逼近器的双 DQN 算法，并同时引入竞争构架 Q 网络结构以解决学习过程中过高估计 Q 值以及无法收敛等问题，仿真结果表明，该算法能够使车辆与复杂城市环境实时交互.

DQN 算法能够分别输出各种动作对应的 Q 值，在处理离散动作任务时具有一定的优势，常被应用于感知和执行系统.ZHAO 等[40]利用基于卷积神经网络的车道检测器输出车道的初步位置，进一步借助基于 DQN 算法开发的车道定位器实现车道的准确检测与定位.在提高驾驶安全性方面，LI 等[41]在信号输入阶段增加储存非均匀样本的重放缓冲区，以简化 DQN 算法的学习过程，并缩短了紧急避障的反应时间.现阶段，大部分自动驾驶系统仅能将原始像素输入至决策模型中，因而驾驶性能难以提高，PENG 等[42]将融合相机图像和车速矢量的混合状态信息作为竞争架构双深度 Q 网络的输入变量，并通过引入对决神经网络结构来降低方差，不仅提高了采样效率，同时实现了端到端车道保持自动驾驶.

DRL 算法的应用使得自动驾驶技术在面对复杂多变的交通环境时仍能取得良好的控制效果，但输入数据与算法框架复杂等问题的存在，会导致 DRL 算法学习策略的收敛速度逐渐变慢以及收敛效果逐渐变差.可以预见，基于 DRL 算法的智能驾驶控制策略正逐步向提高环境探索效率、加快训练速度、降低决策行为异常随机性等方向发展.

2.4　深度强化学习在液压伺服控制领域的应用

液压伺服系统凭借高精度和高频响的优势，目前在航空航天、重型装备、舰船以及深海探测等领

域得到了广泛应用.随着工业智能化的发展,对液压伺服系统的控制精度、鲁棒性以及可靠性提出了更高的要求.然而将现有液压系统中的控制阀全部替换为电液伺服阀成本较高且难度较大,因此需要在现有液压伺服系统基础上研究先进控制策略以代替传统控制方法.由于液压伺服系统存在诸如伺服阀压力流量和摩擦等非线性因素,以及随温度而变化的油液弹性模量和黏性系数等参数不确定性问题,基于线性理论的经典控制方法难以满足液压伺服系统强抗干扰和强鲁棒等高品质跟踪需求.深度强化学习能够模拟人类学习,并具有自我更新优化的能力,在应对复杂非线性控制问题时具有独特的优势,因此逐渐成为液压伺服控制领域的研究热点.DRL 算法在液压伺服控制领域的应用如图 6 所示.

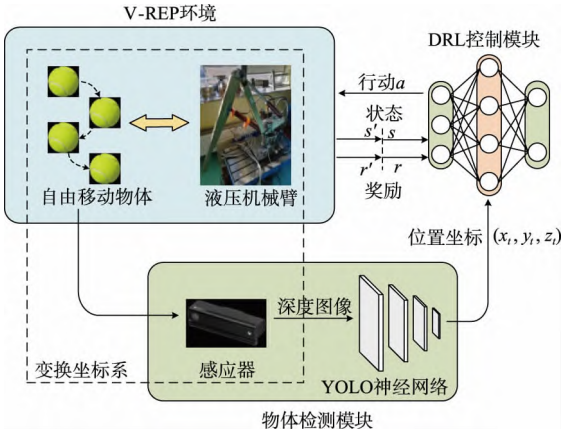


图 6 DRL 在液压伺服控制领域的应用

Fig.6 Application of DRL in hydraulic servo control field

可以看出,DRL 控制模块通过传感设备与被控系统所处任务环境进行交互学习,可以有效减小控制误差,并使被控系统在未知工况下具备自适应能力.为验证 DRL 算法的控制效果,在仿真环境下,BECSI 等[43]分别将蒙特卡洛树搜索算法、监督学习代理算法、规划代理算法以及 DRL 算法应用于车辆气动执行变速器位置控制系统,结果表明,基于 DRL 算法开发的控制策略在减少排放、提高燃油效率等方面具有显著的优势.高正杰[44]利用拟牛顿法训练值函数网络,并基于模糊原理的探索策略设计了一种改进 DDPG 控制方法,以提高液压驱动单元位置控制与环境交互的自学习能力,最终提升了液压驱动型足式机器人在复杂工况下的控制性能.EGLI 等[45]提出了一种基于机器学习的机械臂速度跟踪控制策略,以高度非线性液压挖掘机动臂作为研究对象,并使用强化学习近似策略优化算法对机器运行期间收集的测量数据进行训练,解决了 PID 控制器参数调整方式在工况变化情况下适应性较差的问题.张子扬[46]对 PPO 算法奖励计算公式进行了改进以获得更快的收敛速度,并将其应用于水下液压机械臂目标物抓取,结果表明,该算法可以取得良好的控制效果.由于 PPO 算法在液压伺服控制领域的优异表现以及 DDPG 算法处理连续动作学习问题的有效性,WYRWAL 等[47]在为液压缸设定最优控制策略时引入了一种控制参数整定方法,该方法通过结合 PPO 算法和 DDPG 算法的优点,克服了液压缸运行过程中力和油压扰动问题.

由于缺乏实际运行情况下的液压伺服系统训练数据,目前多采用启发式 DRL 方法训练控制策略以驱动设备完成控制任务.JOHANSSON 等[48]利用递归确定性策略梯度算法,实现了载货汽车前轮液压驱动系统的自适应无模型控制.WU 等[49]首次提出将强化学习算法应用于隔管工具能源回收液压系统,并利用径向基函数非线性映射能力以及较快的网络训练速度精确预测液压泵和蓄能器的压力,此外,通过结合 Q 网络学习算法处理外部环境数据,实时调整液压泵的流量,最终提高节能效率.BECSI 等[50]基于蒙特卡洛树搜索算法提出了一种用于汽车自动变速器的闭环控制策略,该策略通过对浮动活塞双作用气缸的电磁阀进行离散控制,最终实现重型车辆的瞬时平稳换挡.

相较于传统 PID 控制方法,基于 DRL 算法开发的控制策略可以克服液压伺服系统运行过程中产生的超调、振荡和稳态误差,并且能够在未知工况下通过与被控系统实时交互来更新控制参数,具备良好的自适应能力.但对于液压伺服系统而言,环境不确定性会引起控制量波动,导致学习数据收敛性差,因此,为提高基于 DRL 算法开发的控制策略的综合性能,后续研究将聚焦于如何提高样本数据质量以及优化探索策略.

3 关键问题及展望

通过近年来 DRL 算法在自动控制领域广泛应用的回顾可以看出,与常规自动控制方法相比,基于 DRL 算法开发的自动控制方法在复杂工况下表现更优,并且具备较强的容错能力.但随着控制对象的日益复杂化,利用 DRL 算法实现设备自主操控时仍存在许多亟待解决的共性技术问题.本节针对若干关键问题进行了展开分析.

3.1　自动控制系统通用性研究

由于仿真环境与实际应用场景的差异，将基于仿真环境训练的控制策略应用到真实环境中，极易出现控制性能衰减的问题，并且不同任务间的控制策略难以在设备间互相迁移.例如，学习自动驾驶策略时，仿真平台复现并泛化出真实场景时存在真实性损失问题，影响学习策略的置信度，这极大地限制了基于 DRL 算法开发的自动控制系统应用范围与场景，同时也增加了系统开发成本.

通过构建高保真仿真环境可以提高算法对新样本数据的适应能力.目前，虚拟现实、增强现实和数字孪生等技术发展迅速，其中，数字孪生技术侧重于映射对象的模型构建及其数据描述，能够生成将工况状态进行数字化描述的镜像实体，在 DRL 算法领域应用前景广阔.WANG 等[51] 将数字孪生技术与深度学习技术相结合，并应用于电网调度系统中，提高了电网整体控制效率.后续研究可以将数字孪生技术与 DRL 算法相融合，借助该技术强大的交互性、保真性，构建模拟仿真环境，在模拟环境中建立近似于真实场景的物理约束，减少环境差异.

3.2　自动控制设备计算速度与可靠性研究

DRL 算法的生产力由算力支撑，因此计算机的计算能力直接影响算法的性能.例如，无人设备或自动驾驶车辆高速状态下的紧急避障过程中，设备高速运动要求控制系统必须尽可能缩短从感知到决策的处理时间.目前，英伟达 Orin X 自动驾驶芯片单颗算力为 254 TOPS，而自动驾驶系统计算量已达到 1 000 TOPS 级别，硬件性能无法满足控制系统的算力需求.随着中国超算平台的快速发展，基于超算“云化”的形式共享计算能力可以有效解决算力不足问题.例如依托“天河二号”超算构建的超算+无人机遥感网数字化运营平台的初步部署，使无人机遥感网空地联动识别的峰值计算速度能够达到每秒 10.07 兆次.为提高大数据计算能力，后续研究可以融合超算+大数据+人工智能的运算策略，借助云计算技术与网络大数据处理技术来实现 DRL 算法的任务在线分配与处理，利用超算强大的计算能力提升算法的综合性能.

3.3　深度强化学习算法鲁棒性研究

在复杂工况下的自动控制问题中，由于输入数据维度过高，训练样本无法覆盖所有区域，会产生对抗样本，而深度神经网络会受对抗样本影响而产生错误输出.在测试与部署阶段，对抗样本造成的控制系统脆弱性会严重影响无人设备的安全性.但现阶段受到外部计算力及神经网络非线性、大规模特点的限制，仍无法实现深度学习模型的全局鲁棒性.LIN 等[52] 认为大多数现有研究都集中在利用复杂数学模型增强算法鲁棒性，这无法从根本上解决欠鲁棒性问题，而基于传统抗鲁棒性研究提出的模型视觉鲁棒性这一概念，尝试设计出与人类神经系统相似的模型，提高算法抽象能力，最终在鲁棒性研究中达到“超越人类”的效果.后续可以在模型视觉鲁棒性研究的基础上通过对比人类视觉系统与神经网络模型的一致性，从抽象能力出发，通过改进神经网络的深度与结构，缩小人类与模型的差异.

3.4　深度强化学习算法奖励机制研究

奖励函数作为 DRL 算法训练的支撑，设计合理有效的奖励函数可以有效避免算法在学习过程中失去目标的问题.目前大量采用稀疏奖励评价控制系统是否完成目标任务，但在无人设备多步操作过程中，奖励的稀疏性使得智能体需要频繁地通过硬件设备与环境交互，学习速率降低，并且产生高昂的学习成本.对于奖励稀疏或难以定义的任务，目前多采用逆强化学习、经验回放机制、辅助任务设计或多目标学习等方法.其中，多目标学习方法可以同时适用于在策略与离策略的强化学习，并且实现方式直接，成为目前最常用的解决方法.后续研究可以针对多目标学习方法存在目标选择局限性的问题，优化样本采样方法，并将多目标学习方法与元学习、层次化强化学习等方法相融合，最终强化算法求解最优策略的能力，提高学习到的最优策略的水平.

4　结　论

1) 得益于对复杂高维环境强大的表征能力，DRL 算法已经成为当前人工智能领域的研究热点.文中首先对基于值函数与基于策略梯度两类 DRL 算法的基本原理与模型结构进行了深入分析，并阐述了这 2 类算法的特点.其次，详细介绍了基于值函数与基于策略梯度 2 类 DRL 算法的主要改进方法，并分析了各改进方法的主要功能.

2) 以无人机飞行控制、移动机器人轨迹控制、车辆自动驾驶控制以及液压伺服控制等 4 个自动控制领域典型应用场景为例，详细阐述 DRL 算法的国内外研究现状.例如：在无人机集群控制方面，多智能体的 DRL 模型能够学习到有效的协同策略并得到最优控制系统；在机器人控制领域，将共享样本

库嵌入到 DRL 框架中使得学习到的控制策略可以适应新设备的操作要求;在车辆自动驾驶领域,DRL 算法与环境的交互学习使得训练的控制策略可以主动感知道路交通环境并自主决策避障,实现安全驾驶;在液压伺服控制领域中,基于 DRL 算法训练的控制策略在复杂的任务环境中能够实现自主探索.可以预见,随着理论研究与实际应用的不断发展,DRL 算法逐渐向通用、灵活、智能化的方向发展,在未来自动控制技术智能化过程中担任重要角色,并且有望被应用于解决各类复杂控制问题.

参考文献(References)

[1] FARIAS G, GARCIA G, ZAMORA M G, et al. Reinforcement learning for position control problem of a mobile robot [J]. IEEE access, 2020, 8: 152941-152951.

[2] LIU S, WANG Y, YANG X, et al. Deep learning in medical ultrasound analysis: a review [J]. Engineering, 2019, 5(2): 261-275.

[3] ARULKUMARAN K, DEISENROTH P M, BRUNDAGE M, et al. Deep reinforcement learning: a brief survey [J]. IEEE signal processing magazine, 2017, 34(6): 26-38.

[4] BELLEMARE G M, NADDAF Y, VENESS J, et al. The arcade learning environment: an evaluation platform for general agents [J]. Artificial intelligence review, 2013, 47: 253-279.

[5] 张峻伟,吕帅,张正昊,等. 基于样本效率优化的深度强化学习方法研究综述 [J].软件学报,2021,33(11):4217-4238.
ZHANG Junwei, LV Shuai, ZHANG Zhenghao, et al. Survey on deep reinforcement learning methods based on sample efficiency optimization [J]. Journal of software, 2022,33(11):4217-4238. (in Chinese)

[6] 罗磊,李路,阳睿,等. 矢量天调邻域搜索微调算法 [J].通信技术,2019,52(7): 1800-1803.
LUO Lei, LI Lu, YANG Rui, et al. Vector antenna neighborhood search tuning algorithm [J]. Communications technology, 2019, 52(7): 1800-1803. (in Chinese)

[7] HERRERA E M, CALVET L, GHORBANI E, et al. Enhancing carsharing experiences for Barcelona Citizens with data analytics and intelligent algorithms [J]. Computers, 2023, 12(2): 33.

[8] 刘建伟,高峰,罗雄麟. 基于值函数和策略梯度的深度强化学习综述 [J]. 计算机学报,2019,42(6): 1406-1438.
LIU Jianwei, GAO Feng, LUO Xionglin. Survey of deep reinforcement learning based on value function and policy gradient [J]. Chinese journal of computers, 2019, 42(6): 1406-1438. (in Chinese)

[9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518: 529-533.

[10] TIAN S S, LI Y X, ZHANG X. Fast UAV path planning in urban environments based on three-step experience buffer sampling DDPG [J/OL]. Digital communications and networks. [2023-03-09]. https://doi.org/10.1016/j.dcan.2023.02.016.

[11] LUO X F, WANG Y H. PMA-DRL: a parallel model-augmented framework for deep reinforcement learning algorithms [J]. Neurocomputing, 2020, 403: 109-120.

[12] SUN F Y, KONG X Y, WU J Z, et al. DSM pricing method based on A3C and LSTM under cloud-edge environment [J]. Applied energy, 2022, 315: 118853.

[13] FU K S. Learning control systems-review and outlook [J]. IEEE transactions on automatic control, 1986, 8(3): 327-342.

[14] 张文清,徐雪松,刘瑞. 基于反馈线性化的四旋翼无人机姿态控制研究 [J]. 计算机仿真,2019,36(1): 87-91,242.
ZHANG Wenqing, XU Xuesong, LIU Rui. Research on attitude control system of quadrotor UAV based on feedback linearization [J]. Computer simulation, 2019, 36(1): 87-91,242. (in Chinese)

[15] PETKAR J S, GUPTA A A, KETKAR D V, et al. Robust model predictive control of PVTOL aircraft [J]. IFAC-PapersOnLine, 2016, 49(1): 760-765.

[16] LI S, WANG Y, TAN J, et al. Adaptive RBFNNs/integral sliding mode control for a quadrotor aircraft [J]. Neurocomputing, 2016, 216: 126-134.

[17] 虞棐雄,王永超,曹立佳,等. 基于误差逼近器的巡航飞行器反步控制 [J]. 航天控制,2017,35(4): 26-32.
YU Feixiong, WANG Yongchao, CAO Lijia, et al. Backstepping control for cruise aircraft based on error estimation [J]. Aerospace control, 2017, 35(4): 26-32. (in Chinese)

[18] SEDLMAIR N, THEIS J, THIELECKE F. Experimental comparison of nonlinear guidance laws for unmanned aircraft [J]. IFAC-PapersOnLine, 2020, 53(2): 14805-14810.

[19] HU Z, GAO X, WAN K, et al. Relevant experience learning: a deep reinforcement learning method for UAV autonomous motion planning in complex unknown envi-

ronments [J]. Chinese journal of aeronautics, 2021, 34: 187-204.

[20] KOCH W, MANCUSO R, WEST R, et al. Reinforcement learning for UAV attitude control [J]. ACM transactions on cyber-physical systems, 2019, 3(2): 1-21.

[21] 周毅，马晓勇，郜富晓，等. 基于深度强化学习的无人机自主部署及能效优化策略 [J]. 物联网学报, 2019, 3(2): 47-55.
ZHOU Yi, MA Xiaoyong, GAO Fuxiao, et al. Autonomous deployment and energy efficiency optimization strategy of UAV based on deep reinforcement learning [J]. Chinese journal on internet of things, 2019, 3(2): 47-55. (in Chinese)

[22] DHARGUPHT S, GHOSH M, MIRJALILI S, et al. Selective opposition based grey wolf optimization [J]. Expert systems with applications, 2020, 151: 113389.

[23] QU C, GAI W, ZHONG M, et al. A novel reinforcement learning based grey wolf optimizer algorithm for unmanned aerial vehicles (UAVs) path planning [J]. Applied soft computing, 2020, 89: 106099.

[24] XU J, DU T, FOSHEY M, et al. Learning to fly: computational controller design for hybrid UAVs with reinforcement learning [J]. ACM transactions on graphics, 2019, 38(4): 1-12.

[25] 张耀中，许佳林，姚康佳，等. 基于 DDPG 算法的无人机集群追击任务 [J]. 航空学报, 2020, 41(10): 324000.
ZHANG Yaozhong, XU Jialin, YAO Kangjia, et al. Pursuit missions for UAV swarms based on DDPG algorithm [J]. Acta aeronautica et astronautica sinica, 2020, 41(10): 324000. (in Chinese)

[26] 相晓嘉，闫超，王菖，等. 基于深度强化学习的固定翼无人机编队协调控制方法 [J]. 航空学报, 2021, 42(4): 524009.
XIANG Xiaojia, YAN Chao, WANG Chang, et al. Coordination control method for fixed-wing UAV formation through deep reinforcement learning [J]. Acta aeronautica et astronautica sinica, 2021, 42(4): 524009. (in Chinese)

[27] TAI L, LIU M. Mobile robots exploration through CNN-based reinforcement learning [J]. Robotics and biomimetics, 2016, 3: 1-8.

[28] CARLUCHO I, PAULA D M, WANG S, et al. Adaptive low-level control of autonomous underwater vehicles using deep reinforcement learning [J]. Robotics and autonomous systems, 2018, 107: 71-86.

[29] CARLUCHO I, PAULA D M, ACOSTA G G. An adaptive deep reinforcement learning approach for MIMO PID control of mobile robots [J]. ISA transactions, 2020, 102: 280-294.

[30] XIONG H, MA T, ZHANG L, et al. Comparison of end-to-end and hybrid deep reinforcement learning strategies for controlling cable-driven parallel robots [J]. Neurocomputing, 2020, 377: 73-84.

[31] WANG D, DENG H. Multirobot coordination with deep reinforcement learning in complex environments [J]. Expert systems with applications, 2021, 180: 115128.

[32] EHO G, PARK H T. Cooperative object transportation using curriculum-based deep reinforcement learning [J]. Sensor, 2021, 21(14): 21144780.

[33] FENG S, SEBASTIAN B, PINHAS B T. A collision avoidance method based on deep reinforcement learning [J]. Robotics, 2021, 10(2): 10020073.

[34] LIN G, ZHU L, LI J, et al. Collision-free path planning for a guava-harvesting robot based on recurrent deep reinforcement learning [J]. Computers and electronics in agriculture, 2021, 188: 106350.

[35] ZHU M, WANG X, WANG Y, et al. Human-like autonomous car-following model with deep reinforcement learning [J]. Transportation research part C: emerging technologies, 2018, 97: 348-368.

[36] JIN Y, LIU Q, SHEN L, et al. Deep deterministic policy gradient algorithm based on convolutional block attention for autonomous driving [J]. Symmetry, 2021, 13: 13061061.

[37] YE Y, ZHANG X, SUN J, et al. Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment [J]. Transportation research part C: emerging technologies, 2019, 107: 155-170.

[38] YANG T K, LI L K, NGIAP T K, et al. Deep Q-network implementation for simulated autonomous vehicle control [J]. IET intelligent transport systems, 2021, 15(7): 875-885.

[39] KOH S S, ZHOU B, FANG H, et al. Real-time deep reinforcement learning based vehicle navigation [J]. Applied soft computing, 2020, 96: 106694.

[40] ZHAO Z Y, WANG Q, LI X, et al. Deep reinforcement learning based lane detection and localization [J]. Neurocomputing, 2020, 413: 328-338.

[41] LI J X, YAO L, XU X, et al. Deep reinforcement learning for pedestrian collision avoidance and human-machine cooperative driving [J]. Information sciences, 2020, 532: 110-124.

[42] PENG B, SUN Q, LI S, et al. End-to-end autonomous

driving through dueling double deep Q-network [J]. Automotive innovation, 2021, 4: 328-337.

[43] BECSI T, ARADI S, SZABO A, et al. Policy gradient based reinforcement learning control design of an electro-pneumatic gearbox actuator [J]. IFAC-Papers OnLine, 2018, 51(22): 405-411.

[44] 高正杰. 基于深度强化学习的液压驱动单元位置控制 [D]. 秦皇岛: 燕山大学, 2019.

[45] EGLI P, HUTTER M. A general approach for the automation of hydraulic excavator arms using reinforcement learning [J]. IEEE robotics and automation letters, 2021, 20: 475773.

[46] 张子扬. 基于深度强化学习的水下机械臂抓取研究 [D].合肥: 中国科学技术大学, 2020.

[47] WYRWAL D, LINDNER T, NOWAK P, et al. Control strategy of hydraulic cylinder based on Deep Reinforcement Learning [C] // 2020 International Conference Mechatronic Systems and Materials (MSM), Bialystok, 2020: 1-5.

[48] JOHANSSON O, LUNDGREN B. Adaptive model-free control applied to truck front wheel drive: real time control with reinforcement learning utilising recurrent deterministic policy gradient [D]. Gothenburg: Chalmers University of Technology, 2021.

[49] WU T, ZHAO H, GAO B, et al. Energy-saving for a velocity control system of a pipe isolation tool based on a reinforcement learning method [J]. International journal of precision engineering and manufacturing-green technology, 2020, 9: 225-240.

[50] BECSI T, SZABO A, KOVARI B, et al. Reinforcement learning based control design for a floating piston pneumatic gearbox actuator [J]. IEEE access, 2020, 8: 147295-147312.

[51] WANG Xingzhi, ZHAI Haibao, YAN Yaqin, et al. Predispatching method of new generation dispatching and control system based on digital twin and deep learning [J]. Journal of Shanghai Jiaotong University, 2021, 55(S2): 37-41.

[52] 林点,潘理,易平.面向图像识别的卷积神经网络鲁棒性研究进展 [J].网络与信息安全学报,2022,8(3): 111-122.
LIN Dian, PAN Li, YI Ping. Research on the robustness of convolutional neural networks in image recognition [J]. Chinese journal of network and information security, 2022,8(3): 111-122.(in Chinese)

(责任编辑 朱漪云)

(上接第 623 页)

[53] RAHAL C, STERLING M, COULBECK B. Parameter tuning for simulation models of water distribution networks [J]. Proceedings of the Institution of Civil Engineers, 1980,69(3): 751-762.

[54] KOTOWSKI J, OLESIAK M. G4.3: the optimization of the energy wastes in the complex water-supply system [J]. IFAC proceedings volumes, 1980,13(9): 389-395.

(责任编辑 徐云峰)