

视线引导与自专家克隆融合强化学习的无人船路径跟踪

刘嘉辉, 赵一诺, 田丰, 齐光鹏, 李江涛, 刘驰

引用本文

刘嘉辉, 赵一诺, 田丰, 齐光鹏, 李江涛, 刘驰. [视线引导与自专家克隆融合强化学习的无人船路径跟踪](#)[J]. 计算机科学, 2025, 52(12): 239-251.

LIU Jiahui, ZHAO Yīnuo, TIAN Feng, QI Guangpeng, LI Jiangtao, LIU Chi. [Line of Sight Guided Self Expert Cloning with Reinforcement Learning for Unmanned Surface Vehicle Path Tracking](#) [J]. Computer Science, 2025, 52(12): 239-251.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于图卷积神经网络的多属性个性化航空行程推荐系统](#)

Personalized Multi-attribute Airline Itinerary Recommendation System by Graph Convolutional Neural Network

计算机科学, 2025, 52(11A): 250200088-6. <https://doi.org/10.11896/jsjcx.250200088>

[改进深度强化学习的多智能体联合导航策略研究](#)

Research on Multi-agent Joint Navigation Strategy Based on Improved Deep Reinforcement Learning

计算机科学, 2025, 52(11A): 250200095-7. <https://doi.org/10.11896/jsjcx.250200095>

[利用融合2-opt的强化学习算法求解TSP问题](#)

Hybrid Reinforcement Learning Algorithm Combined with 2-opt for Solving Traveling Salesman Problem

计算机科学, 2025, 52(11A): 250200121-8. <https://doi.org/10.11896/jsjcx.250200121>

[基于汤普森采样的自适应安卓程序测试方法](#)

Adaptive Android Program Test Method Based on Thompson Sampling

计算机科学, 2025, 52(11): 330-338. <https://doi.org/10.11896/jsjcx.240900150>

[基于卷积双延迟深度确定性策略梯度的卫星网络多路径路由算法](#)

Multipath Routing Algorithm for Satellite Networks Based on Convolutional Twin Delay Deep Deterministic Policy Gradient

计算机科学, 2025, 52(11): 280-288. <https://doi.org/10.11896/jsjcx.240800161>

视线引导与自专家克隆融合强化学习的无人船路径跟踪

刘嘉辉¹ 赵一诺¹ 田丰² 齐光鹏^{3,4} 李江涛² 刘驰¹

1 北京理工大学计算机学院 北京 100081

2 中国民航信息网络股份有限公司北京市民航大数据工程技术研究中心 北京 100318

3 浪潮集团有限公司 济南 250101

4 浪潮云洲工业互联网有限公司 济南 250098

(ljhjiayoua@126.com)

摘要 无人船路径跟踪对海上自主作业至关重要,然而,风、浪、流以及无人船自身的控制误差等因素会影响路径跟踪的性能。强化学习算法凭借在线交互与实时反馈的特点,能够主动适应动态环境,在无人船路径跟踪任务中展现出良好的应用前景。然而,其试错训练模式在实际应用中存在安全风险,且理想仿真场景与现实复杂环境之间的差距也进一步制约了强化学习在实际应用中的效果。针对这些挑战,提出了一种视线引导与自专家克隆融合强化学习的无人船路径跟踪算法 LECUP。LECUP 算法首先在静水环境中训练专家策略,随后通过自专家克隆将智能体迁移至更复杂的环境中。为了确保知识能够有效传递,LECUP 算法引入数据填充机制,将自专家在静水环境中积累的经验数据进行升维填充并存储,并以此初始化复杂环境中的智能体。之后,运用强化学习算法对智能体在复杂环境中进行微调,从而进一步适应复杂环境。此外,LECUP 算法结合视线算法计算目标航向,将路径跟踪控制与路径几何形状解耦,增强了无人船对不同路径形状的适应能力。该方法不仅能够在复杂环境中持续优化策略,还能缓解随机初始化带来的安全风险。大量实验结果表明,相较于基线方法,LECUP 算法能够更好地完成无人船路径跟踪任务。

关键词: 无人船;路径跟踪;强化学习;自专家克隆;视线算法

中图分类号 TP181;TP273;U664.82

Line of Sight Guided Self Expert Cloning with Reinforcement Learning for Unmanned Surface Vehicle Path Tracking

LIU Jiahui¹, ZHAO Yinuo¹, TIAN Feng², QI Guangpeng^{3,4}, LI Jiangtao² and LIU Chi¹

1 School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

2 Beijing Civil Aviation Big Data Engineering Technology Research Center, Travelsky Technology Limited, Beijing 100318, China

3 INSPUR Group Co., Ltd., Jinan 250101, China

4 INSPUR Yunzhou Industrial Internet Co., Ltd., Jinan 250098, China

Abstract Unmanned Surface Vehicle(USV) path tracking is crucial for marine autonomous operations, as environmental factors such as wind, waves, currents, and USV's control errors can affect tracking performance. Reinforcement learning(RL), with its online interaction and real-time feedback, offers a promising approach for actively adapting to dynamic environments. However, its trial-and-error training process poses safety risks in real-world applications, and the gap between ideal simulation environments and complex real-world conditions further limits its practical effectiveness. To address these challenges, this paper proposes LECUP(Line-of-sight-guided self-Expert Cloning for USV Path tracking), a new algorithm designed for complex marine environments. LECUP first trains an RL expert in a still water environment and then uses self-expert cloning to transfer the agent to a more complex environment. To ensure effective knowledge transfer, LECUP introduces a data filling mechanism, where the experiences accumulated by the self-expert in the still-water environment are dimensionally padded and stored for initializing the agent in the complex environment. Then, reinforcement learning is used to fine-tune the agent in the complex environment, further enabling adaptation to the complexities of the environment. Moreover, LECUP incorporates a line-of-sight guidance module to calculate the target heading, decoupling the path tracking control from the specific geometry of the path and enhancing the USV's

到稿日期:2025-02-17 返修日期:2025-05-19

基金项目:智能动态行程规划技术研究项目;国家自然科学基金(U23A20310)

This work was supported by the Intelligent Dynamic Journey Planning Technology Research Project and National Natural Science Foundation of China(U23A20310).

通信作者:李江涛(ljtao@travelsky.com.cn)

adaptability to various path shapes. This method enables ongoing policy refinement in complex environments while mitigating safety risks associated with random initialization. Extensive experimental results show that LECUP performs better than baseline methods in path tracking tasks, especially under challenging conditions.

Keywords Unmanned surface vehicle, Path tracking, Reinforcement learning, Self-expert cloning, Line of sight

1 引言

无人船(Unmanned Surface Vehicle, USV)在海洋作业中愈发重要,已被广泛应用于环境监测、海上安全以及海洋探测等诸多领域^[1-2]。USV 路径跟踪任务是 USV 领域的研究重点之一,该任务的目标是控制 USV 在复杂的海洋环境中沿着预定路线航行。然而,USV 路径跟踪控制问题依旧面临着诸多挑战,实际复杂水域环境里存在的风、浪、流等环境扰动会对 USV 路径跟踪性能产生显著影响。除此之外,USV 自身的控制误差进一步加大了路径跟踪任务的复杂程度。

为实现精准高效的路径跟踪,研究者们提多了种控制方法。基于 PID 设计的控制算法^[3]因简洁且易于实现而被广泛运用,但其在应对非线性系统以及外部干扰时,性能相对有限。模型预测控制^[4]能够在约束条件下求解 USV 路径跟踪问题,但其计算成本较高,且实时性存在一定限制。另外,反馈控制^[5]、模糊控制^[6]及滑模控制^[7]等方法也常用于路径跟踪,但每种方法都存在一定局限性,如适应动态环境的能力较弱,设计复杂度较高,可能会引发高频振荡等。

视线算法^[8](Line-of-Sight, LOS)是一种常用的路径跟踪辅助算法,它的核心思路是通过计算 USV 当前位置与目标路径之间的视线角度,引导 USV 调整航向,使其朝着预定路径前行。视线算法在 USV 路径跟踪控制中具备显著优势,它能够实时依据 USV 的位置以及目标路径的变化动态调整航向,保障航行的灵活性与适应性。

图 1 展示了本文所考虑的 LOS 引导下的 USV 路径跟踪系统场景。

差问题,硬件偏差以及环境噪声等因素,导致期望控制信号与实际控制信号所产生的作用之间存在一定差异。这些因素共同加重了系统环境的复杂性,使得在实际环境中实现精确的路径跟踪成为挑战。

近年来,强化学习(Reinforcement Learning, RL)在计算机游戏^[9-10]、自动驾驶^[11-12]和机器人操纵^[13-14]领域取得了卓越成就。RL 利用在线学习的方式,能够在复杂环境中持续迭代更新,从而自动适应复杂环境,因此具有良好的应用前景。在 USV 路径跟踪控制方面,RL 智能体能够根据来自环境的实时反馈动态调整控制策略,使 USV 具备持续适应环境的变化与不确定性的能力^[15-16]。与传统方法相比,RL 在主动适应环境的同时,无需对环境进行显式的动态建模。

基于 RL 的路径跟踪控制研究可分为两类。一类是借助 RL 学习现有控制器的增益参数,以此调整控制器的行为。例如,调整纯追踪控制器中的前视视距,以减小误差^[17]。另一类则是通过 RL 直接生成转向命令,控制 USV 的路径跟踪行为。本研究属于第二类,其中 RL 用于直接控制转向命令,以便更灵活地适应复杂环境中的动态变化,并实现更为精确灵活的路径跟踪控制。

尽管 RL 在多个领域已经取得了显著成果,但是其在 USV 路径跟踪中的应用仍面临挑战。由于 RL 使用试错机制进行训练,为保障安全性,训练通常在理想化的仿真环境中进行,然而真实场景中的风、浪、流等扰动因素往往与仿真环境中的设置不完全一致,这会导致仿真环境中训练的策略难以适应实际的复杂环境。现有研究对这个问题的应对策略不尽相同。Wen 等^[15]主要关注策略设计等方面,没有主要针对环境扰动的影响进行研究;Wang 等^[16]在训练环境中建模扰动,以增强策略对复杂扰动环境的适应性,但这种方法对场景建模的要求较高,在实际应用中实现相对复杂。因此,如何在保证算法对实际环境适应性的同时,降低对环境精确建模的依赖,仍是需要解决的问题。

为降低对复杂环境精确建模的依赖,引入行为克隆是一种可行且有效的策略。Fan 等^[18]提出了一种自专家克隆方法,该方法在训练环境中预训练自专家策略并迁移至复杂场景中进行微调,提高了策略在实际环境中的适应性。这种方式以较优初始策略为基础,降低了从零探索带来的不稳定性风险。然而,在 USV 路径跟踪任务中,自专家策略预训练使用的环境与实际复杂环境的输入状态向量维度往往不一致,因此 Fan 等的方法无法直接适用于本场景,需要根据 USV 路径跟踪任务的特点调整自专家克隆机制,以实现有效的策略迁移。此外,路径形状的多样性和巨大的状态-动作空间也提升了问题的复杂度,对算法的鲁棒性提出了更高的要求。

为应对这些挑战,本文提出了 USV 路径跟踪算法 LECUP(Line-of-sight-guided self-Expert Cloning for USV Path tracking)。该架构在使用 RL 进行自适应训练与更新的同

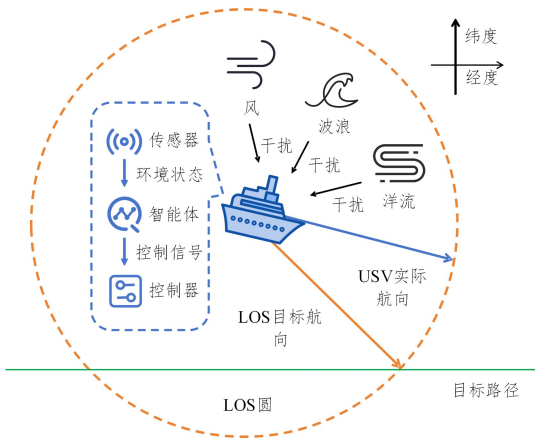


图 1 LOS 引导下复杂 USV 路径跟踪场景(电子版为彩图)

Fig. 1 LOS-guided complex USV path tracking scenario

该系统由 3 个关键组件构成:被控制的 USV、预先设定的 LOS 圆(以虚线圆表示),以及通过计算 LOS 圆与目标路径交点而得出的目标航向(以橙色箭头指示)。在复杂的水域环境中,此系统会受到风、浪、流等环境扰动因素的影响,这些因素会致使 USV 偏离目标航向。此外,该系统还面临控制误

时,提升了算法对实际复杂环境的适应性。具体而言,LECUP在基于RL的USV路径跟踪中引入了两项主要改进措施:1)使用自专家克隆算法,在静水环境中预训练自专家策略并积累经验数据集,之后将策略迁移至复杂扰动环境并进行微调,提升了算法对复杂扰动环境的适应性;2)引入了一个LOS引导模块,将USV路径跟踪问题转化为目标航向跟踪问题,该模块在降低问题的解空间规模的同时,增强了算法对各类USV目标路径几何形状的适应能力。其中,针对USV路径跟踪任务,LECUP对自专家克隆算法进行了调整与改进,设计了一种自专家克隆方法——数据填充机制。该机制首先对自专家策略在静水环境中收集的经验数据集进行升维填充与存储,随后利用这些数据对复杂场景下的RL智能体策略进行初始化。在此过程中,数据填充机制支持额外初始化智能体的价值网络,以进一步提升算法在复杂环境中微调初期的性能。

本文开展了大量实验,以评估LECUP算法在USV路径跟踪问题中的性能表现。首先,进行超参数调整实验,分析了超参数的影响;其次,通过消融实验分析了自专家克隆在提升算法微调初期表现的价值,以及使用RL在复杂环境中进行微调的必要性;接着,将LECUP与4种基准方法进行对比,这些方法涵盖了经典控制、模仿学习以及基础的RL算法,并进一步研究了不同自专家克隆策略和RL微调方法的表现;然后,对不同训练阶段以及不同路径复杂度下的USV路径跟踪控制效果进行了可视化分析,在进一步分析算法各模块价值的同时,展示了LOS引导下算法对不同路径几何形状的适应性;最后,对算法在长程USV路径跟踪中的表现进行了分析。

总体而言,本文的主要贡献如下:

1)针对复杂海洋环境USV路径跟踪控制任务,提出了一种基于视线引导与自专家克隆的强化学习USV路径跟踪控制算法LECUP。该算法中的视线算法和自专家克隆机制有效提升了基于RL的USV路径跟踪算法在复杂扰动任务环境下的适应性和性能。

2)为了改善RL在训练初期的性能表现,针对USV路径跟踪任务,设计了一种自专家克隆方法——数据填充机制。该机制通过实行升维填充以及对价值网络展开额外训练,在确保算法适配USV路径跟踪任务的同时,进一步提升了RL微调初期算法的性能。

3)进行了大量的实验,包括超参数调整实验、消融实验和对比实验等。与多种基准方法相比,LECUP在复杂环境条件下呈现出更为出色的路径跟踪效果。

本文第2章回顾相关工作,第3章介绍问题模型和定义,第4章描述LECUP算法框架,第5章展示实验结果,最后总结全文并展望未来。

2 相关工作

2.1 路径跟踪控制算法

在路径跟踪控制领域,研究人员已研究了多种路径跟踪算法。Xiong等^[19]从几何学、运动学和动力学等角度总结了无人驾驶车辆路径跟踪的经典算法。在经典路径跟踪控制

中,PID算法因简单有效而被广泛使用。Zhao等^[20]提出了一种模糊自适应调节PID算法,通过优化PID控制器的内部参数和增益输出,实现了机器人路径跟踪的精确控制。反馈控制通过实时调整控制输入,根据当前的状态与目标路径的偏差进行修正,确保系统能够适应外部扰动和动态变化。Yang等^[21]利用鲁棒状态反馈控制与模型预测控制方法解决车辆轨迹跟踪问题,并比较了二者在不同场景下的性能表现。滑模控制通过引入滑模面,实现对系统的不确定性和外部扰动的强鲁棒性。Abdillah等^[22]利用滑模控制解决了自动驾驶车辆控制中的系统不确定性及无法获取全部状态变量的问题。模糊控制通过模糊化控制规则,能够处理系统模型不明确或不完整的情况,从而灵活应对环境变化和外部扰动。Mancilla等^[23]提出了一种基于群体智能优化算法的模糊控制方法,结合后轮模糊逻辑控制器解决自动路径跟踪问题。模型预测控制(Model Predictive Control, MPC)算法通过实时优化控制输入来最小化未来状态偏差,该算法能够处理系统约束并适应动态变化。Zhang等^[24]通过结合Koopman算子与鲁棒MPC,解决了非线性系统在建模误差和外部扰动下的鲁棒路径跟踪问题。同时,LOS算法在导航任务中起到重要作用,通过计算当前航向与目标点方向之间的夹角,为路径调整提供实时方向指导。Fossen等^[25]提出了一种基于LOS的自适应路径跟踪控制器,用于补偿由海流、风和波浪引起的侧滑力。该方法适用于海洋船只、自动水下航行器等水平面机动的应用。

近年来,随着人工智能技术的发展,研究人员提出了多种基于学习的路径跟踪算法框架。行为克隆算法直接从专家示例中学习状态到动作的映射,在路径跟踪控制中能够快速生成有效的控制策略,但遇到未包含在专家示例中的新状态时,控制效果不理想。Azam等^[26]使用行为克隆设计了一个基于深度学习的控制器,利用手动驾驶数据训练神经网络,实现了自主车辆的路径跟踪控制。逆向强化学习算法则通过推断专家行为背后的潜在奖励函数,在路径跟踪控制中能够学习到更具通用性的策略。然而,其由于需要处理复杂的推断过程,计算复杂度较高。Wang等^[27]提出了一种基于自对比学习的逆强化学习框架,通过双层解耦训练机制优化奖励模型,以提升轨迹预测的泛化能力。生成对抗模仿学习通过结合对抗训练和模仿学习,使智能体能够高效模仿专家的路径跟踪行为,避免了对显式奖励信号的依赖,在复杂环境中能够更好地捕捉专家策略。然而,由于对抗训练过程的复杂性,生成对抗模仿学习在训练过程中可能需要较长的时间,且对网络结构和训练数据的依赖性较强,可能导致计算开销较大。Chen等^[28]提出了重要性重加权生成对抗模仿学习算法,解决了自主水下航行器控制中难以设计有效奖励函数的问题。

RL算法在应对动态和未知环境方面具有显著优势,已被广泛应用于无人车和无人机的路径跟踪控制领域,其核心优势在于能够通过与环境的互动进行自我学习,逐步优化控制策略。这些方法通常依赖大量数据进行训练或实施在线学习,使得控制系统能够灵活适应复杂多变的环境。Yang等^[29]使用RL开发了一种路径跟踪算法,确保了自主车辆在路径跟踪中有着高精度和高稳定性。Jiang等^[30]提出了一种

分层强化学习框架,该框架通过引入微分补偿器的 PPO-DC 算法提升了收敛速度和控制精度,并设计了累积奖励机制以解决多目标学习问题,从而有效解决了固定翼无人机路径跟踪中的控制和制导问题。

总体而言,研究人员已提出多种路径跟踪解决方案。经典算法实施简便,具有较强的可解释性,但均存在一定局限性。近年来,随着人工智能技术的发展,基于学习的路径跟踪框架得到了广泛关注,这些方法有效提升了系统的适应性和性能。然而,正如前文所述,这些方案仍面临挑战。在此背景下,RL 因在动态未知环境中的自我学习与适应能力,展现出了在路径跟踪中的应用潜力。然而,基于 RL 的算法目前仍存在一些挑战,例如训练过程需要大量的计算资源,并且在实际环境中进行训练时的试错过程可能带来安全隐患。因此,虽然 RL 为路径跟踪控制提供了强大的适应能力,但其高计算成本和安全风险仍然是需要解决的问题。

2.2 USV 路径跟踪算法

在 2.1 节提到的路径跟踪控制领域研究的背景下,针对 USV 路径跟踪问题的具体特点,研究人员开发了多种算法,以提高 USV 在复杂环境中的路径跟踪性能。PID 控制算法凭借其简洁性和易于实现的特点,被广泛用于 USV 路径跟踪任务。然而,PID 控制主要针对线性系统进行设计,对于复杂环境中的非线性动态和外部扰动的适应性相对受限。Alim 等^[3]将 PID 控制与滑模控制结合,解决了 USV 路径跟踪中的抖振问题。反馈控制算法通过分析 USV 与目标路径之间的几何关系,直接计算转向角和航向角以纠正航向偏差。这种方法计算简单,稳定性较高,但依赖几何关系建模,在非线性环境或强扰动条件下效果相对受限。Winursito 等^[5]提出了一种基于积分状态反馈控制的设计方法,用于解决 USV 的航向控制问题。模糊控制算法能够利用规则推导专家行为,对非线性问题具备一定的适应能力。然而,其规则设计过程复杂,尤其在面对高度不确定性场景时,需要人工大量调整规则,实际应用难度较高。He 等^[6]提出了一种基于不确定性与干扰估计器的同步模糊控制方法,解决了多船在外部时间变化干扰下的协调同步运动问题。MPC 算法能够在多步预测范围内优化控制输入,并实时调整控制策略,以应对动态环境变化。然而,其计算复杂度较高,尤其是在处理高维非线性动态系统时,系统实时性要求可能难以满足。Cai 等^[4]提出了一种基于 MPC 和 RL 相结合的方法,用于优化 USV 路径跟踪、自动停靠和它们之间的过渡过程的控制策略。滑模控制算法通过控制系统状态维持在滑模面上,有效地增强了控制的稳定性和鲁棒性。然而,该算法在滑模面附近可能出现高频抖振问题(即“颤振”),这不仅增加了执行模块的负担,还可能影响控制性能。Jiang 等^[7]通过滑模控制方法解决了 USV 网络在环境扰动下的队形控制问题,其设计的新型滑动面可用于 USV 路径跟踪控制。

值得一提的是,经典算法 LOS 可以增强系统的鲁棒性和对动态变化的适应能力,因此在 USV 路径跟踪任务中得到了广泛应用。Liu 等^[8]将 LOS 与非线性航向控制器结合,以解决复杂扰动水域环境下的 USV 路径跟踪问题;Wang 等^[31]通过引入有限时间正切漂角视线制导律,解决了欠驱动 USV 在

复杂环境下的路径跟踪与稳定性问题;Yang 等^[32]提出一种改进的 USV 航迹控制方法,通过结合模糊控制可变船长比的 LOS 算法和自抗扰航向控制器,解决了 USV 在复杂环境下因环境干扰出现的航迹偏离问题。

随着智能算法的快速发展,基于智能算法的 USV 路径跟踪技术也得到了广泛应用。Yang 等^[33]通过遗传算法优化基于 LOS 的 PID 控制器,解决了 USV 路径跟踪中的环境扰动和参数不确定问题,提高了系统动态性能和自适应能力;Zhang 等^[34]利用神经网络解决了 USV 系统在未知非线性动力学和不确定性影响下的稳定性和成本最小化问题;Zhu 等^[35]通过引入长短期记忆网络补偿 USV 模型不确定项和外界干扰,抑制了滑模控制中的抖动现象,提升了轨迹跟踪精度和抗干扰能力。与此同时,RL 算法在 USV 路径跟踪控制任务中也表现出良好的适应性,得到了广泛的关注。Wang 等^[16]提出了一种基于 RL 的控制方法,用于 USV 轨迹跟踪。该控制方法采用改进的 DDPG 算法在线调整 PID 控制参数,以增强对环境扰动的适应性,并提高路径跟踪精度。Wen 等^[15]通过递归视界强化学习实现 USV 的高精度轨迹跟踪控制。另外,RL 也被用于优化经典 USV 路径跟踪控制算法,如前文提到的 Cai 等^[4]利用 RL 优化 MPC 的参数。然而,正如引言中讨论的,基于 RL 的 USV 路径跟踪控制算法仍然面临挑战。

总体而言,各类路径跟踪算法在 USV 路径跟踪中均有所应用。经典算法如 PID 和 MPC 具备较高的可解释性,但在复杂环境下存在一定的局限性。LOS 算法可为 USV 路径跟踪提供引导,为多种控制算法提供支持,具有较高的应用价值。相比之下,RL 等智能算法在复杂场景中展现出强大的适应能力,但其应用仍面临引言中提到的挑战。因此,本文旨在优化基于强化学习的 USV 路径跟踪算法,以提升其在复杂环境中的适应性和性能。

2.3 强化学习算法

近年来,RL 的发展吸引了广泛关注,多种算法在不同应用中取得了显著成果, Yang 等^[36]对此进行了梳理。其中, Mnih 等^[9]提出的深度 Q 网络(Deep Q-Network, DQN)作为一种基于价值的 RL 方法,使用深度神经网络学习动作-价值函数,推动了 RL 研究的发展。Lillicrap 等^[37]提出了适用于连续动作空间的深度确定性策略梯度算法(Deep Deterministic Policy Gradient, DDPG),进一步扩展了 RL 算法的应用范围。Haarnoja 等^[38]提出的演员-评论家软更新算法(Soft Actor-Critic, SAC)通过最大化随机策略,显著提升了策略优化的鲁棒性;而 Schulman 等^[39]提出的近端策略优化(Proximal Policy Optimization, PPO)通过剪切技术,有效平衡了探索与开发的矛盾。此外,离线强化学习利用离线数据提高了样本效率和训练稳定性。Kostrikov 等^[40]提出的模仿 Q 学习是离线强化学习的代表算法之一;Nakamoto 等^[41]在保守 Q 学习的基础上引入“校准”机制,进一步提升了算法的性能。分层强化学习通过将复杂任务分解为子目标,学习不同层次的策略来高效解决长期目标问题。该算法能有效缓解稀疏奖励等问题,但仍面临子目标空间过大时,如何高效发现有用于子目标的问题。Luo 等^[42]提出了一种高层模型逼近算法,以增强分

层强化学习的训练稳定性和探索效率。多智能体强化学习通过协调多个智能体的行为来完成复杂的决策和协作任务。Yu等^[43]将经典RL算法PPO扩展至多智能体领域,在多个经典测试平台上表现优秀。为了提升RL算法的泛化性,Fan等^[18]提出了自专家克隆算法,使用数据增强的方法提高RL策略的泛化能力。在其设定中,智能体观察空间设置在弱增强和强增强条件下保持一致,然而本研究面临的复杂环境在静水环境的基础上引入了额外的噪声状态维度,因此存在维度不一致的问题,这使得文献^[18]中的自专家克隆算法在本场景下不直接适用。

综上所述,RL算法发展迅速,并已广泛应用于多个领域,但在USV路径跟踪任务中仍面临挑战。自专家克隆能够有效提升RL算法在训练初期的性能,然而,如前文所述,其经典方法并不直接适用于本文场景。因此,本文在现有研究的基础上,将自专家克隆与经典算法LOS适配并融入基于RL的USV路径跟踪算法中,以增强算法在复杂环境强化学习训练过程中(尤其是训练初期)的表现,以及其对复杂扰动场景USV路径跟踪任务的适应性性能。

3 问题建模

3.1 USV路径跟踪数学模型

在USV路径跟踪控制中,算法的任务是根据目标路径、USV状态和环境信息,给出USV控制信号,使得USV沿预定路径航行。在此过程中,需要对USV运动学模型和环境扰动模型进行建模。本节将给出USV路径跟踪运动学模型(三自由度模型)和复杂环境扰动模型(包括风、浪、流的扰动模型和USV控制误差模型)的建模过程和数学公式描述,其建模方式参考文献^[44]。

1) 三自由度模型

本文采用三自由度模型来描述USV的运动学模型。该模型考虑了USV的平面运动,包括位置、航向角和速度变化。具体而言,模型的动态变化通过式(1)描述:

$$\begin{cases} \dot{x} = u \cos(\phi) - v \sin(\phi) \\ \dot{y} = u \sin(\phi) + v \cos(\phi) \\ \dot{\phi} = \tau \end{cases} \quad (1)$$

其中, x 和 y 是USV在平面中的位置坐标, ϕ 是航向角, u 和 v 分别是前向速度和侧向速度, τ 是转角速度。另外,本节所有微分符号均默认为对时间求导。

2) 风、浪、流的扰动模型

本文将风、浪、流对USV路径跟踪的影响建模为对USV航行位移的扰动,具体包括持续位置偏移和周期性位置扰动。风和洋流对USV的影响通常是持续积累的,这种影响会导致USV的位置逐渐偏移;而波浪扰动具有周期性,会对USV运动产生周期性影响。此建模方法将风、浪、流的影响分别建模为持续位置偏移和周期性扰动,可以有效地捕捉这些外部因素对USV运动的整体影响,同时保持模型的简洁性和适用性。

风、浪、流引起的持续位置偏移可表示为:

$$\dot{\boldsymbol{p}}_{\text{con}} = L_w \cdot C_w \cdot \boldsymbol{v}_w + L_c \cdot C_c \cdot \boldsymbol{v}_c + L_s \cdot C_s \cdot \boldsymbol{v}_s \quad (2)$$

其中, $\dot{\boldsymbol{p}}_{\text{con}}$ 是一个向量,记录了风、浪、流引起的USV在 X 和 Y 方向上的位置偏移扰动的速度; L_w, L_c 和 L_s 分别表示风、浪、流的强度等级; C_w, C_c 和 C_s 是风、浪、流对USV的影响因数,该因子与USV的物理特性和响应特性有关; $\boldsymbol{v}_w, \boldsymbol{v}_c$ 和 \boldsymbol{v}_s 分别是风、浪、流扰动的方向向量。该方程表明,风、浪、流的影响是累积性的,随着时间的推移,USV的位置会持续发生偏移。同时,式(2)将波浪的持续扰动部分也考虑在内,在后续计算波浪扰动时只需关注其周期性扰动部分。

波浪引起的周期性位置干扰可以通过式(3)描述:

$$\dot{\boldsymbol{p}}_{\text{per}} = L_w \cdot C_p \cdot \sin(2\pi f_w t + \psi) \cdot \boldsymbol{v}_w \quad (3)$$

其中, $\dot{\boldsymbol{p}}_{\text{per}}$ 是一个向量,表示波浪干扰引起的水平和垂直坐标偏移的速度; L_w 表示波浪强度的等级; C_p 代表波浪对USV周期性干扰的影响因数。此处将波浪对USV的周期性影响近似为一个简谐波,使用 $\sin(2\pi f_w t + \psi)$ 描述波浪干扰随时间变化的周期函数,其中 f_w 是波浪的频率, ψ 是相位偏移。 \boldsymbol{v}_w 是波浪扰动的方向向量。该方程建模了波浪干扰的周期性变化。

综合式(2)与式(3),得出风、浪、流造成的USV位置偏移扰动公式:

$$\dot{\boldsymbol{p}}_{\text{dist}} = \dot{\boldsymbol{p}}_{\text{con}} + \dot{\boldsymbol{p}}_{\text{per}} \quad (4)$$

其中, $\dot{\boldsymbol{p}}_{\text{dist}}$ 为 Δt 时间内风、浪、流造成的USV位置偏移扰动的速度。

3) 控制误差

控制误差指控制信号造成的实际效果与其期望值之间的偏差,主要由硬件误差和环境噪声等因素引起,误差大小受到USV物理属性和环境干扰水平的影响。这些误差由多个独立的随机因素累积而成,因此使用高斯分布进行建模,其均值为 μ ,方差为 σ^2 。控制信号误差可以表示为:

$$\dot{\phi}_{\text{dist}} = (L_e \cdot C_e \cdot \boldsymbol{v}_e) \cdot \mathcal{N}(\mu, \sigma^2) \quad (5)$$

其中, $\dot{\phi}_{\text{dist}}$ 表示控制信号(转角速度)的控制误差; L_e 是误差等级,用于指定误差的强度; C_e 是控制误差的影响因数;向量 \boldsymbol{v}_e 表示误差方向(对转向角度控制信号来说,是顺时针或逆时针)。 $\mathcal{N}(\mu, \sigma^2)$ 表示一个高斯分布,其中均值 μ 捕捉误差中的系统性偏倚,通常与硬件缺陷或环境条件相关,而方差 σ^2 量化了误差分布的不确定性。该模型能够建模控制误差的随机性和潜在偏倚。

综上,记 $\dot{\boldsymbol{p}}_{\text{dist}} = (\dot{x}_{\text{dist}}, \dot{y}_{\text{dist}})$,并将其与 $\dot{\phi}_{\text{dist}}$ 引入三自由度模型(式(1)),得到复杂扰动场景下USV路径跟踪建模总公式:

$$\begin{cases} \dot{x} = u \cos(\varphi) - v \sin(\varphi) + \dot{x}_{\text{dist}} \\ \dot{y} = u \sin(\varphi) + v \cos(\varphi) + \dot{y}_{\text{dist}} \\ \dot{\varphi} = \tau + \dot{\phi}_{\text{dist}} \end{cases} \quad (6)$$

3.2 USV路径跟踪问题建模

本节针对USV路径跟踪控制问题进行建模。具体而言,本文设计了两种任务环境:静水环境和复杂环境。静水环境中,仅考虑USV的位置、航向以及待跟踪路径。而复杂环境则在静水环境的基础上,进一步加入了风、浪、流等外部扰动以及USV的控制误差。静水环境对应仿真环境,在该环境下,算法可以预训练路径跟踪策略并积累经验。与此相对,复杂环境则反映了真实世界的复杂条件,其中引入了外部扰动

和控制误差,突出了仿真与现实环境之间的差异。在复杂环境中,除了路径跟踪任务基础设置外,还考虑了风、浪、流的扰动和 USV 自身的控制误差对路径跟踪控制的影响。需要指出的是,算法在复杂环境中并不能直接利用这些扰动的具体建模,而是通过获取相关影响参数(如各种扰动的等级和方向)来进行决策,以模拟实际环境中扰动影响效果的不确定性。

本文主要关注算法在复杂环境中的路径跟踪控制表现,并额外关注算法在复杂环境微调初期的表现,以评估自专家克隆算法在提升微调初期效果方面的优势,并以此分析其在提升 RL 算法在实际 USV 路径跟踪控制任务中的可用性和稳定性方面的优势。

据此,以下定义了系统输入、系统输出和优化目标,并建立了一系列随机化设定。

1)系统输入:在静水环境中,系统的输入包括 USV 的当前位置坐标和目标路径信息。在复杂环境中,除了包含静水环境中的信息外,还增加了环境因素,如风速、风向、波浪频率、波浪高度、流速和流向等。

2)系统输出:系统的输出为控制信号,具体选取为 USV 的角速度。

3)优化目标:系统的优化目标是在复杂环境的扰动条件下,控制 USV 的实际航向尽可能与 LOS 算法计算得到的目标航向保持一致。RL 的优化目标通过马尔可夫决策过程中的奖励函数来定义,相关的奖励设计将在第 4 章中详细介绍。而评估指标的具体内容将在第 5 章中进一步阐述。

4)随机化设定:为了增强训练场景的多样性,在静水环境和复杂环境中均随机初始化 USV 状态。静水环境中,随机化 USV 初始航向与目标路径之间的夹角;在复杂环境中,除了随机化 USV 的初始航向外,风速、风向、波浪频率、波浪高度、流速和流向等参数也进行了随机初始化。这些随机化设定,对算法的鲁棒性提出了进一步的要求。

4 解决方案:LECUP

针对复杂扰动场景 USV 路径跟踪控制任务,本文提出了一种视线算法与自专家克隆融合强化学习的 USV 路径跟踪算法框架 LECUP (Line-of-sight-guided self-Expert Cloning for USV Path tracking),如图 2 所示。LECUP 有 3 个主要组成部分:LOS 算法引导模块、自专家克隆模块和 RL 微调模块。

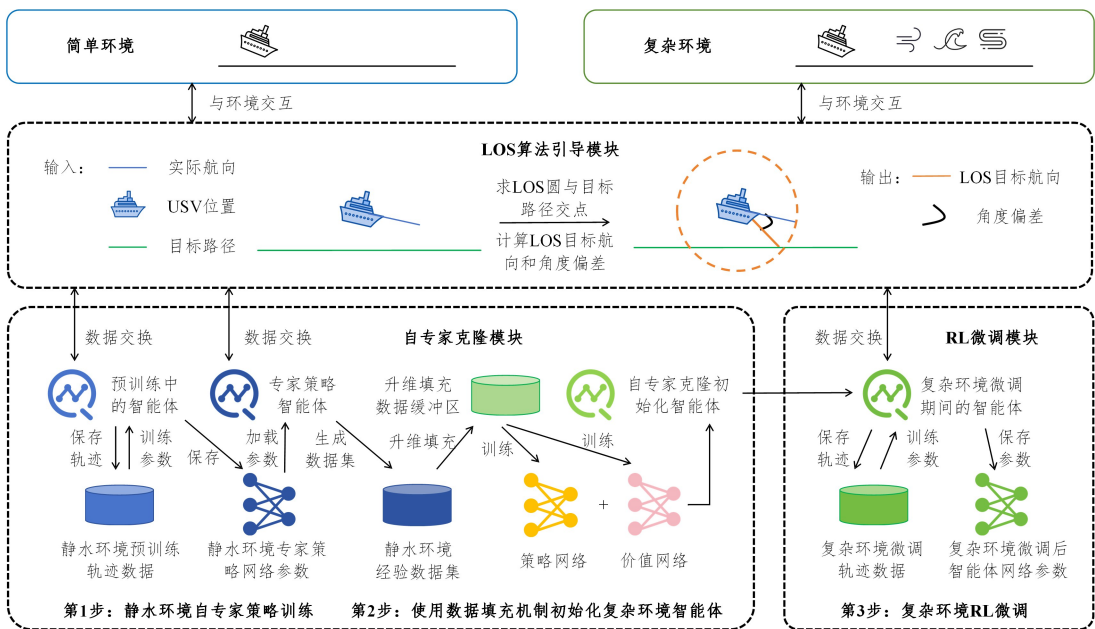


图 2 LECUP 算法框架

Fig. 2 Framework of LECUP algorithm

LOS 引导模块在自专家克隆和 RL 微调阶段中都发挥着关键作用。该模块根据当前 USV 位置与目标路径之间的几何关系得出目标航向 θ ,从而为智能体的导航提供了有效的指导。LOS 模块将目标航向计算与 RL 控制策略优化解耦,将目标路径跟踪问题转化为目标航向跟踪控制,使得算法无需直接处理复杂的路径几何问题,从而提高了智能体在不同路径形状下的适应性。

自专家克隆模块用于初始化复杂环境智能体。在此过程中,首先在静水环境中训练自专家策略智能体,以学习基本的 USV 路径跟踪策略。训练完成后,使用自专家智能体与静水环境交互并收集经验数据集,为后续复杂环境智能体初始化做准备。然后使用经验数据集进行自专家克隆。本文针对

USV 路径跟踪任务提出了一种自专家克隆机制,称为数据填充机制。在数据填充机制中,静水环境下收集到的经验数据会首先进行升维填充,确保其数据维度与复杂环境一致,以保证数据在初始化复杂环境智能体策略时的可用性。另外,与传统的行为克隆方法主要考虑训练策略网络不同,数据填充机制在训练智能体策略网络的同时额外训练价值网络,这种方法进一步提升了 RL 在微调初期的表现,增强了 RL 微调的稳定性和训练效率。

完成自专家克隆后进行 RL 微调,智能体将在复杂环境中进一步使用 RL 算法微调训练,以适应环境中的动态变化。该阶段利用 RL 算法自适应迭代更新的优势,使智能体进一步适应复杂扰动场景。同时,与随机初始化的传统 RL 算法

相比,自专家克隆初始化后的 RL 训练过程更加安全且高效。

接下来将详细讨论 LECUP 中视线算法模块、自专家克隆模块和强化学习微调模块的设计与实现。

4.1 LOS 算法引导模块

LOS 模块通过定义 LOS 圆来计算目标航向^[8]。首先,定义 USV 在时刻 t 的位置为 $(x_t^{\text{USV}}, y_t^{\text{USV}})$,并将以该位置为圆心、半径为 R_t 的圆定义为 LOS 圆。LOS 圆的半径 R_t 计算式如式(7)所示:

$$R_t = R_0 + D_t \quad (7)$$

其中, D_t 是 t 时刻 USV 与目标路径之间的距离, R_0 是根据 USV 的特性和速度选定的常数。目标航向通过求解 LOS 圆与目标路径的交点来确定。离路径终点最近的交点被定义为目标点 $(x_t^{\text{tar}}, y_t^{\text{tar}})$ 。

在该框架中,LOS 模块执行两个主要功能。首先,它提供目标航向作为 RL 的状态输入,其计算式如下:

$$\theta_t = \begin{cases} \arctan\left(\frac{y_t^{\text{tar}} - y_t^{\text{USV}}}{x_t^{\text{tar}} - x_t^{\text{USV}}}\right), & \text{if } x_t^{\text{tar}} \neq x_t^{\text{USV}} \\ 90, & \text{if } x_t^{\text{tar}} = x_t^{\text{USV}} \text{ and } y_t^{\text{tar}} > y_t^{\text{USV}} \\ -90, & \text{if } x_t^{\text{tar}} = x_t^{\text{USV}} \text{ and } y_t^{\text{tar}} < y_t^{\text{USV}} \end{cases} \quad (8)$$

其中,通过 \arctan 函数输出的角度使用角度制,用于计算 USV 实时位置 $(x_t^{\text{USV}}, y_t^{\text{USV}})$ 指向 LOS 目标点 $(x_t^{\text{tar}}, y_t^{\text{tar}})$ 的方向向量的方向角,该方向角即为 LOS 算法引导模块计算的目标航向 θ_t ;同时,该式给出了 $x_t^{\text{tar}} = x_t^{\text{USV}}$ 这个特殊情况下的目标航向 θ_t ,此目标航向 θ_t 直接作为 RL 训练的状态输入之一。

LOS 模块的第二个功能是计算目标航向 θ_t 和实际航向 ϕ_t 在 t 时刻的差值,并以此计算奖励函数,以推动 RL 算法更新。其具体设计见静水场景与复杂场景 MDP 建模。

总之,在 LECUP 框架中,LOS 模块不仅为 RL 提供方向性状态输入,还通过计算目标航向和实际航向之间的差异来辅助奖励函数的设计,这有助于提高智能体的路径跟踪能力。具体而言,LOS 模块将 USV 的位置坐标与路径信息转换为目标航向,使路径跟踪能够独立于路径的具体几何形状,提高了算法的泛化能力。进一步地,该模块相当于在策略网络输入中引入了一种先验信息,使得策略学习不再完全依赖于数据驱动,从而有效提升了算法的可解释性和最终效果。

4.2 自专家克隆模块

自专家克隆算法由两个主要步骤组成:第一步是在静水环境中训练专家智能体策略,并使用智能体与静水环境进行交互以收集经验数据集;第二步是使用本文提出的数据填充机制,首先对经验数据集进行升维填充,然后将其用于复杂环境智能体策略的初始化。

4.2.1 专家策略预训练

专家策略预训练的目标是在静水环境中预训练一个专家策略,为后续在更复杂环境中的训练提供支持。在此阶段,USV 路径跟踪问题被形式化为一个马尔可夫决策过程(Markov Decision Process, MDP),其主要要素包括状态、动作和奖励,相关定义如下。

1)状态:在静水环境中,状态空间主要由目标航向(通过 LOS 计算获得)和 USV 的实际运动方向组成。状态空间可表示为:

$$s_t = [\theta_t, \phi_t] \quad (9)$$

其中, θ_t 表示目标航向, ϕ_t 是 USV 的实际航向。

2)动作:智能体的动作空间对应于 USV 的控制信号,具体为 USV 的角速度。

3)奖励:由于系统的优化目标是降低目标航向与实际运动方向之间的差异,因此奖励函数设计为该角度差的相反数,智能体通过最小化角度差来实现更准确的路径跟踪性能。奖励可以表示为:

$$r_t = -|\theta_t - \phi_t| \quad (10)$$

在专家策略预训练阶段,智能体通过 PPO 算法^[39]与环境进行交互,利用裁剪策略平衡探索与开发,并不断更新价值网络与策略网络以提高其性能。PPO 的详细训练过程与损失函数将在第 4.3 节中展开说明。

在 PPO 训练完毕后,专家智能体与静水环境进行交互以生成经验数据集。该数据集供随后的数据填充机制使用,为初始化复杂环境智能体的网络参数提供了基础。

4.2.2 数据填充机制

为了适配 USV 路径跟踪任务并实现更优的自专家克隆初始化性能,在自专家克隆算法^[18]的基础上,提出了一种数据填充机制。该机制首先对来自静水环境的经验数据进行升维填充,以使其与复杂环境维度匹配;然后在训练策略网络的同时额外训练价值网络,以进一步提升算法在 RL 微调早期的性能。

在本文的问题设置中,与静水环境相比,复杂环境引入了风、浪、流的方向与等级等状态维度,状态维度更高,因此本文引入了升维填充操作以确保经验数据集在复杂环境中的兼容性。具体而言,需要将静水环境中维度为 d_s 的状态扩展至复杂环境所需的维度 d_c 。升维填充后的状态表示为:

$$\tilde{s}_t = [s_t, 0, 0, \dots, 0] \quad (11)$$

其中, $d_c - d_s$ 个 0 被附加到状态后面,以使状态的维度与复杂环境所需的维度保持一致。

在数据升维填充完成后,即进行复杂环境智能体初始化。在此过程中,首先随机初始化一个与复杂场景设置匹配的 PPO 智能体,然后从升维填充后的经验数据集中提取状态-动作-奖励三元组,将其存入复杂环境 PPO 智能体的回放缓冲区中,并调用 PPO 的网络更新模块进行智能体的初始化。与经典的模仿学习方法主要训练策略网络不同,数据填充机制利用 PPO 的更新模式额外预训练了价值网络,能有效提升智能体在复杂环境 RL 微调初期探索阶段对状态与动作未来价值判断的准确度,从而为后续的优化过程和策略更新提供更加稳健的指导。以上做法可行的原因是,尽管复杂环境中包含了额外的与扰动相关的状态信息,但静水环境和复杂环境中的场景任务和优化目标(奖励函数)总体是相同的。这种一致性确保了转移的数据(状态-动作-奖励)在复杂环境中依然具有很高的参考价值。

综上,自专家克隆作为一种预训练方法,其核心优势在于将策略探索过程限制在较优的初始策略分布附近,以降低强化学习在复杂环境中的探索难度,从而提升策略的泛化能力与鲁棒性。为实现静水环境与复杂环境之间经验数据的无缝迁移,算法使用数据填充机制,通过升维填充保持状态输入分

布的一致性。同时,该机制结合 PPO 的更新模式,在初始化策略网络的同时对价值网络进行预训练,有效缓解了策略更新中的高方差问题,进一步提升了强化学习微调初期的训练效果。整体而言,数据填充机制与自专家克隆的结合可视为一种约束优化过程,在保持输入一致性的基础上提升了策略性能与训练过程的稳定性。

4.3 RL 微调模块

在 RL 微调模块中,复杂环境下的 USV 路径跟踪任务同样被建模为 MDP。下文给出状态、动作和奖励的定义。

除了目标航向和 USV 实际运动方向外,复杂环境中的状态还包括风、浪、流等外部扰动因素信息。这些扰动因素信息统一记为 ω_t ,它们会影响 USV 路径跟踪的性能,这对智能体对动态环境变化的适应性提出了更高的要求。复杂环境中的状态空间可以表示为:

$$s_t = [\theta_t, \phi_t, \omega_t] \quad (12)$$

其中, θ_t 表示目标航向, ϕ_t 是 USV 的实际航向。

在复杂环境中,动作空间和奖励函数与静水环境中的设计是一致的。动作对应于 USV 角速度,允许智能体调整其运动方向以跟踪目标航向。奖励函数根据目标航向与实际运动方向之间的差异设计,具体定义同式(10)。

在 RL 微调阶段,同样采用 PPO 算法^[39]来优化智能体性能。具体来说,主要是进行价值网络和策略网络的训练。策略网络根据当前状态输出动作,而价值网络则用于估计当前策略下执行动作的期望回报。在训练过程中,智能体与复杂环境交互,收集状态、动作和奖励数据,并利用这些数据更新网络参数。其中,策略网络更新的损失函数如下:

$$L_{\text{clip}}(\Theta) = E_t [\min(r_t(\Theta) \hat{A}_t, \text{clip}(r_t(\Theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)] \quad (13)$$

该损失函数通过限制策略更新的幅度来提高训练的稳定性。其中, $r_t(\Theta)$ 是当前策略概率与旧策略概率的比值; \hat{A}_t 为优势估计函数,表示动作相对期望回报的优劣。该损失函数使用 clip 函数将 $r_t(\Theta)$ 限制在 $[1 - \epsilon, 1 + \epsilon]$ 之间,防止策略更新过快,保持策略更新的平衡性。这种裁剪机制有助于在探索和开发之间取得平衡,确保训练过程中的稳定性和收敛性。

另一方面,价值网络的更新目标是 minimized 估计价值 $V(s_t)$ 和实际累计奖励 R_t 之间的均方误差。损失函数表示为:

$$L_V = E_t [(V(s_t) - R_t)^2] \quad (14)$$

在以上策略优化过程中,算法在每次更新中限制新策略与旧策略之间的偏移量,确保策略不会发生过大的变化,从而提高训练的稳定性。同时,利用多步更新来充分利用样本信息,并通过优势估计减少梯度计算的方差,从而提升优化效率。在此基础上,自专家克隆提供了高质量的初始策略,使优化过程在初始阶段围绕更优的策略分布更新,减少无效探索,从而缓解强化学习算法探索-开发困境的负担,提高训练效率,并降低策略训练初期的不确定性风险。

5 实验与分析

本章首先介绍实验的系统参数设置和评估指标;其次通过超参数调整实验分析了超参数对算法性能的影响;接着通

过消融实验和对比实验分析了算法的性能;然后通过可视化 USV 路径跟踪效果进一步分析了 LECUP 算法的表现;最后分析了算法在长程 USV 路径跟踪中的表现。

5.1 实验设置

5.1.1 系统设置

针对 USV 路径跟踪问题,本文设计了两个在线训练环境(一个是静水环境,另一个是复杂扰动环境),并在这两个环境中模拟了 USV 对目标路径的路径跟踪过程,其中静水环境用于自专家克隆中预训练专家策略,复杂环境用于 RL 微调,算法效果以复杂环境中 USV 路径跟踪的表现为准。参考文献[25]的数据设置训练测试环境,将 USV 的静水速度设置为 3 m/s,最大角速度设置为 $30^\circ/\text{s}$ 。RL 智能体每隔 0.2 s(一个时间步长)决策一次,每个训练回合持续 200 个时间步。为了考虑不同的初始条件,最大初始夹角设置为 90° ,即 USV 与目标路径之间的初始夹角从 $[-90^\circ, 90^\circ]$ 范围内随机选取。另外,本文算法采用了 LOS 算法进行航向引导,并将 LOS 圆的半径常数 R_0 设置为 6 m。复杂环境中引入风、浪、流和控制误差等扰动因素,其扰动等级在 0~10 之间随机选择,扰动方向等其他设定在每个训练回合开始前随机初始化,以确保训练测试场景的多样性,默认情况下其设定在一个训练回合中保持不变。在最高扰动级别 10 的情况下,环境扰动造成的 USV 的位置偏移最大可达到 USV 在静水中航行位移的 50%,这代表了严峻的航行条件。后续实验中,智能体在静水环境中交互生成自专家策略预训练所需的数据集;自专家克隆使用自专家策略在静水环境中积累的经验数据集进行训练;复杂环境下,智能体使用与环境实时交互过程收集的数据集进行算法训练;其他算法的训练数据集则基于表现较好的策略(如 MPC)在环境中的表现进行整理。

5.1.2 评估指标

计算目标航向 θ_t 与实际运动方向 ϕ_t 之间的角度偏差 $\delta_t = |\theta_t - \phi_t|$,并以此定义一个回合内 δ_t 的平均值为整体均值偏差(Overall Mean Deviation, OMD)指标,以反映算法在整个路径跟踪过程中的表现。较低的 OMD 值,表示较好的路径跟踪性能。

为了进一步分析算法性能,定义了以下评估指标:路径跟踪效率(Path Tracking Efficiency, PTE)和路径跟踪精度(Path Tracking Accuracy, PTA)。具体来说,首先引入归一化偏差 δ_t^{norm} :

$$\delta_t^{\text{norm}} = 1 - \frac{\delta_t}{180} \quad (15)$$

其中, δ_t 是角度偏差。 δ_t^{norm} 的值越大,表示目标航向与实际运动方向之间的偏差越小,反映了更好的优化性能。回合前半段的 δ_t^{norm} 平均值定义为 PTE 指标,代表了 USV 前期转向过程中调整航向的效率;而回合后半段 δ_t^{norm} 的平均值则定义为 PTA 指标,用于评估直线跟踪的跟踪精度。这两个指标用于进一步分析算法控制 USV 转向与直行的表现。

5.2 超参数调整实验

本实验研究了以下两个超参数对算法性能的影响:自专家克隆阶段的训练轮数(Epoch)和 RL 算法 PPO 中的 epsilon-clip 值。对比了在不同自专家克隆训练轮数下算法微调

初期(前 50 回合)的 OMD 值和不同 epsilon-clip 值下算法微调的最终 OMD 值。实验结果如表 1 和表 2 所列。

表 1 自专家克隆训练 epoch 数调参结果

Table 1 Self-expert clone training epoch number tuning results

epochs	OMD
10	17.04
50	13.46
100	11.63
500	11.91
1000	12.70

表 2 PPO epsilon-clip 参数调参结果

Table 2 PPO epsilon-clip parameter tuning results

eps-clip	OMD
0.10	7.79
0.15	7.72
0.20	7.68
0.25	7.80
0.30	8.27

从表 1 中可以看出,随着自专家克隆训练轮数的增加,强化学习微调初期的 OMD 逐步降低,并在 100 轮时达到最佳表现。然而,超过 100 轮后,奖励值未出现明显改善,表明训练 100 轮是自专家克隆的最优选择。

epsilon-clip 值通常推荐在 0.1 至 0.3 之间。从表 2 中结果来看,最终奖励对 epsilon-clip 值的敏感性较低,算法性能在这一参数范围内较为稳定。算法最佳性能出现在 epsilon-clip 值为 0.2 时,达到 OMD 最低值 7.72。因此,选取 0.2 为 PPO 算法的 epsilon-clip 值。

综上,在自专家克隆阶段,训练 100 轮时可获得最佳效果;在 PPO 微调过程中,epsilon-clip 值设为 0.2 可使算法达到最优。因此,在后续实验中,自专家克隆阶段的训练轮数设置为 100,PPO 微调阶段的 epsilon-clip 值设置为 0.2,以达到算法最优效果。另外,后文涉及的所有算法默认采取最优的超参数。

5.3 消融实验

本实验探讨了 LECUP 框架中的自专家克隆机制和强化学习微调对复杂场景训练过程中性能的贡献。如图 3 所示,若去除自专家克隆机制(如图中绿线所示),算法在复杂环境训练初期 OMD 为 73.34,约为 LECUP 初期表现的 7.29 倍,这将给训练过程带来严重的安全风险,凸显了自专家克隆机

制在保证算法早期表现中的重要性。另一方面,在没有强化学习微调的情况下(如图中蓝线所示),算法的 OMD 保持在 10 左右,未能进一步适应复杂场景;而 LECUP 在经过微调后,能够持续适应复杂环境,OMD 最终减少到 7.68,减少了 24.84%(如图中红线所示),这体现了强化学习微调在适应复杂场景时的重要作用。

消融实验的结果体现了自专家克隆和强化学习微调在 LECUP 框架中的关键作用,两者的结合使得复杂环境的智能体在优良初始条件下,通过训练提升了对复杂场景的适应能力。

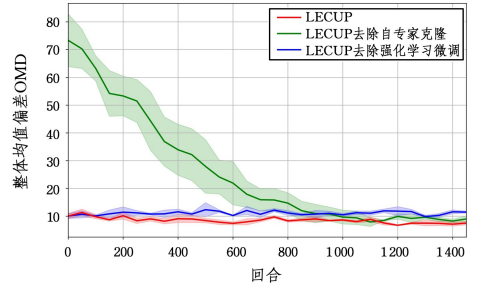


图 3 消融实验结果(电子版为彩图)

Fig. 3 Ablation study results

5.4 对比实验

5.4.1 LECUP 与基线算法对比

本节从经典控制、模仿学习和强化学习中选择了 4 个基准方法进行对比。

- 1)比例-积分-微分控制器(PID)^[3]:使用经典的 PID 控制器完成 USV 路径跟踪任务。
- 2)模型预测控制(MPC)^[4]:利用系统模型优化控制动作并预测未来状态,以实现路径跟踪。
- 3)行为克隆(BC)^[26]:利用从 MPC 收集的数据进行模仿学习,学习路径跟踪策略。
- 4)近端策略优化算法(PPO)^[39]:一种在线强化学习算法,从头开始学习路径跟踪策略,不依赖于离线示范数据。

此外,本节在复杂环境下对不同最大扰动等级、最大初始夹角和 USV 最大角速度下的所有对比算法进行了性能比较。表 3 中列出了在不同最大扰动等级、最大初始夹角、USV 最大角速度设定的复杂环境下,LECUP 与基线算法的 OMD 指标对比结果。

表 3 LECUP 与基线算法的 OMD 指标对比结果

Table 3 Comparison of OMD metric between LECUP and baseline algorithms

	最大扰动等级					最大初始夹角/(°)					USV 最大角速度/(°/s)				
	2	5	10	15	20	2	5	10	15	20	2	5	10	15	20
LECUP	5.70	5.87	7.68	17.91	35.15	4.41	5.67	7.68	10.55	16.90	12.23	10.53	7.68	6.81	6.75
PID ^[3]	6.37	10.43	14.77	39.14	70.98	5.06	8.69	14.77	39.14	70.98	18.53	16.72	14.77	13.51	11.49
MPC ^[4]	<u>6.12</u>	<u>6.64</u>	<u>8.61</u>	<u>19.87</u>	<u>39.27</u>	<u>4.82</u>	<u>6.19</u>	<u>8.61</u>	12.34	19.09	<u>13.76</u>	<u>12.13</u>	<u>8.61</u>	<u>7.90</u>	<u>7.74</u>
BC ^[26]	6.67	7.36	9.93	22.65	51.20	5.40	6.72	9.93	13.52	22.21	15.82	13.47	9.93	8.62	8.20
PPO ^[39]	6.45	6.80	9.05	21.36	41.51	5.27	6.36	9.05	<u>11.90</u>	<u>18.99</u>	14.45	13.60	9.05	8.27	7.87

表中数值是 5 个随机种子下的平均结果,每个环境下最佳性能用粗体标出,次佳性能用下划线标出。可以看出,LECUP 展现了最佳的整体性能,凸显了算法在各种复杂环境设置下的适应性。例如,在默认设置(最大扰动等级为 10,最大

初始夹角为 90°,USV 最大角速度为 30°/s)下,LECUP 的 OMD 值为 7.68,而次优的 MPC 算法的结果为 8.61,说明其平均路径跟踪角度偏差比 LECUP 大 12.11%。从算法比较的角度来说,PID 算法在没有额外机制处理环境干扰的情况

下,在复杂场景中的表现受限。尽管 MPC 的表现较好,但该算法的计算复杂性较高,这给实际应用带来了挑战。相比之下,LECUP 通过在线交互学习控制策略,无需显式建模环境。BC 算法也具有较好的抗干扰能力,但其表现受到数据集质量的限制,其适应变化环境的能力也比较有限。例如,当最大干扰等级增加到 20 时,它的表现急剧下降,OMD 上升到 51.20。基于 RL 的 PPO 方法也能通过试错训练适应复杂环境,但其性能逊色于带有自专家克隆的 LECUP 方法。另外,消融实验也展示了直接使用 RL 进行训练的其他弊端。

效率指标 PTE 和精度指标 PTA 的进一步对比:为了进一步比较 4 种较为有效的方法 (LECUP, MPC, BC, PPO) 在路径跟踪效率和精度上的表现,本实验提供了基于 PTE 和 PTA 指标的比较结果,如图 4 和图 5 所示。可以观察到,LECUP 在所有条件下均有着最高的效率和精度。

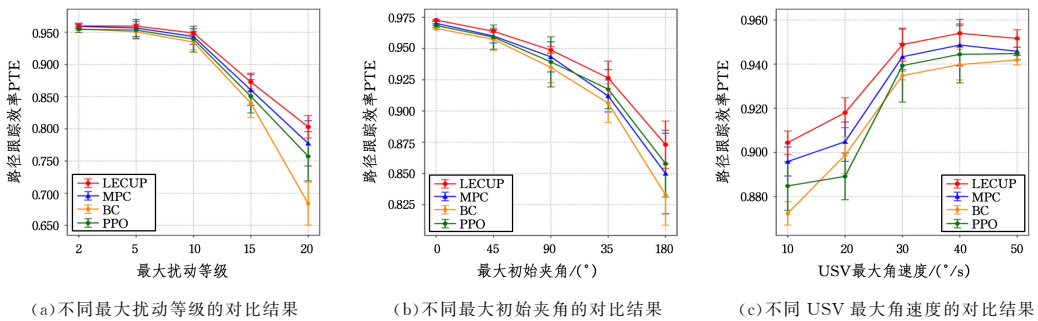


图 4 算法在路径跟踪效率指标 PTE 上的比较

Fig. 4 Comparison of algorithms on the PTE metric in path tracking efficiency

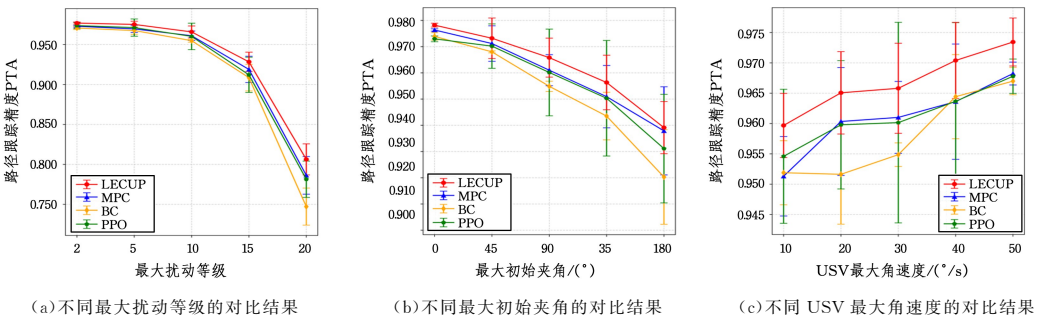


图 5 算法在路径跟踪精度指标 PTA 上的比较

Fig. 5 Comparison of algorithms on the PTA metric in path tracking accuracy

5.4.2 自专家克隆与强化微调算法选型对比

LECUP 算法使用数据填充机制 (Data Filling Mechanism, DFM) 进行自专家克隆,并使用强化学习算法 PPO 进行微调。以下将这两部分的实现分别替换为其他方法并进行效果对比分析。

1) 针对自专家克隆策略选型的对比研究。本实验将 LECUP 算法与两种替代的自专家克隆策略进行比较: 1) PPO+交叉熵 (Cross Entropy, CE), 该方法将 PPO 与交叉熵损失函数结合,用于初始化 USV 的控制策略; 2) PPO+均方误差 (Mean Squared Error, MSE), 该方法将 PPO 微调与基于均方误差的模仿损失结合,用于初始化 USV 的控制策略。实验结果如图 6 所示。可以看出,使用数据填充机制的 LECUP 在所有环境设置下均优于其他方法。例如,图 6(b)中最大初始

例如,在图 4(c)中,当 USV 最大角速度为 $30^\circ/\text{s}$ (最大扰动等级为 10, 最大初始夹角为 90°) 时,LECUP 的 PTE 得分为 0.949, 而第二名的 MPC 得分为 0.943, 这意味着 MPC 在路径跟踪包含转向的前半段的平均 δ_t 比 LECUP 大 11.76%, 说明 LECUP 在 USV 路径跟踪中可以更高效地实现转向控制。如图 5(c) 所示,当最大角速度为 $30^\circ/\text{s}$ (最大扰动等级为 10, 最大初始夹角为 90°) 时,LECUP 的 PTA 得分为 0.966, 而第二名的 MPC 得分为 0.961, 这意味着 MPC 在路径跟踪后半段直线跟踪阶段的平均 δ_t 比 LECUP 大 14.71%, 说明 LECUP 在 USV 路径跟踪中可以更精确地跟踪目标航向。总而言之,图 4 和图 5 中的结果体现了 LECUP 在转向效率和直线跟踪精度上的表现均比基线算法更优,即算法在应对初始航向变化和后续路径跟踪时均保持着出色的性能表现。

夹角为 90° 时,LECUP 的 OMD 值为 10.06, 比 PPO+CE 和 PPO+MSE 的 OMD 值 15.39 与 14.27 分别小 34.63% 和 29.50%。此外可以观察到,在环境设定变得严苛时,数据填充机制体现出了更好的适应能力,例如图 6(b) 中初始角度为 180° 时,PPO+CE 和 PPO+MSE 的 OMD 值增大到 34.67 和 35.54, 但 LECUP 的 OMD 值维持在了 25.68。这一优势的产生,是因为 LECUP 利用数据填充机制额外训练了价值网络,从而进一步提高了演员-评论家架构中智能体的起始表现。因此可以看出,LECUP 在强化学习微调初期的表现中受到恶劣环境条件的影响相对较小;相比之下,PPO+CE 和 PPO+MSE 仅模仿控制策略,使得它们在更具挑战性的环境条件下,算法表现受到的影响更大。

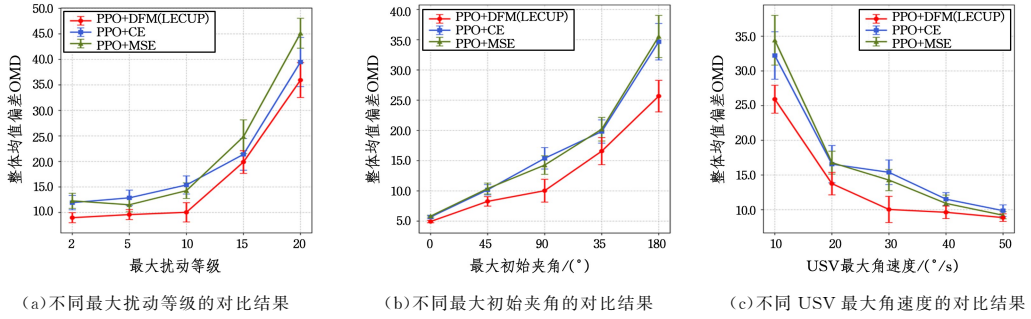


图 6 不同自专家克隆策略在强化学习微调前 50 回合的 OMD 指标对比

Fig. 6 Comparison of OMD metric among different self-expert cloning strategies in the first 50 episodes of RL fine-tuning

2) 针对强化学习算法选型的对比研究。本实验选择了 4 种广泛使用的强化学习算法进行比较,并评估它们在 USV 路径跟踪中的 OMD 指标。LECUP 使用 PPO 算法进行在线微调,而其他 3 种选择分别是:SAC^[38],DDPG^[37],DQN^[9]。

实验结果如图 7 所示。可以看出,在这 4 种强化学习微调算法中,PPO 的表现优于 SAC,DDPG 和 DQN,这主要得益于其基于策略的学习方法和裁剪机制。PPO 通过限制策

略更新的幅度,有效避免了过大的策略变化,从而提高了训练过程的稳定性。同时,PPO 能够在较少的样本下实现高效的策略优化,尤其是在面对复杂环境时,能够更好地平衡探索和利用,降低收敛到次优解的风险。因此,PPO 在复杂环境下具有更强的适应性,能够更快地适应风、浪、流等外部干扰,提升算法的路径跟踪性能。因此,LECUP 使用 PPO 算法进行 RL 微调。

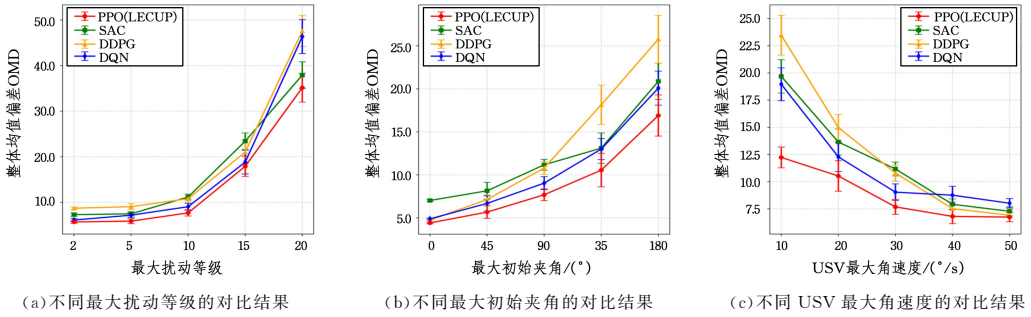


图 7 不同强化学习微调算法在微调完成时的 OMD 指标对比

Fig. 7 Comparison of OMD metric among different RL fine-tuning algorithms at the completion of fine-tuning

5.5 USV 路径跟踪效果可视化

图 8 展示了随机初始化、自专家克隆初始化和强化学习微调后 USV 路径跟踪的实际表现。如图 8(a)所示,在随机策略初始化下,USV 实际路径与目标路径偏差较大,这表明如果没有进行适当的预训练,强化学习训练的初期表现就会很差,从而带来安全风险。相比之下,如图 8(b)

所示,使用自专家克隆初始化能够显著改善强化学习算法的初始表现,为强化学习训练提供更安全的起点,但是此时 USV 的路径跟踪效果没有达到最优。在完成强化学习微调后,如图 8(c)所示,算法能精确控制 USV 跟踪目标路径,并在动态扰动下保持稳定,验证了该算法在实际复杂场景中的有效性。

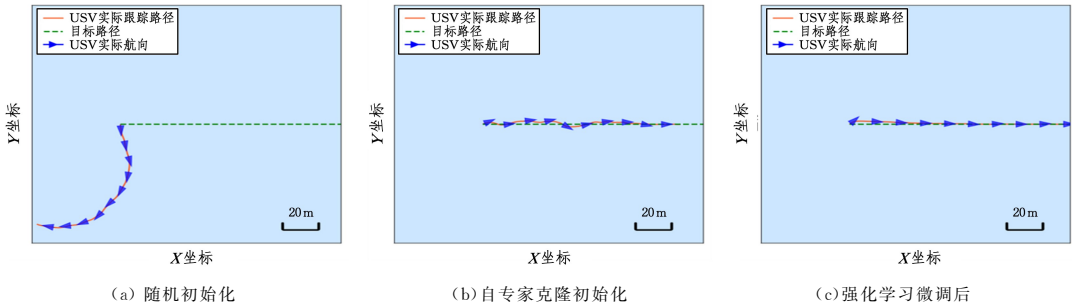


图 8 随机初始化、自专家克隆初始化和强化学习微调后 USV 路径跟踪的表现

Fig. 8 USV path tracking performance after random initialization, expert clone initialization, and RL fine-tuning

图 9 中可视化了 LECUP 算法在不同目标路径复杂度下的路径跟踪表现。通过引入 LOS 模块,LECUP 将路径跟踪转化为目标航向跟踪,提高了算法在不同形状路径下的适应性。对于图 9(a)中的开阔水域直线航行场景,LECUP 实现

了精确的路径跟踪,即使面对更具挑战性的目标路径时(图 9(b)的转弯避障航行场景和图 9(c)的复杂水域航行场景),算法仍然表现出较好的路径跟踪效果,这说明 LOS 引导下的 LECUP 算法对不同目标路径的鲁棒性较高。

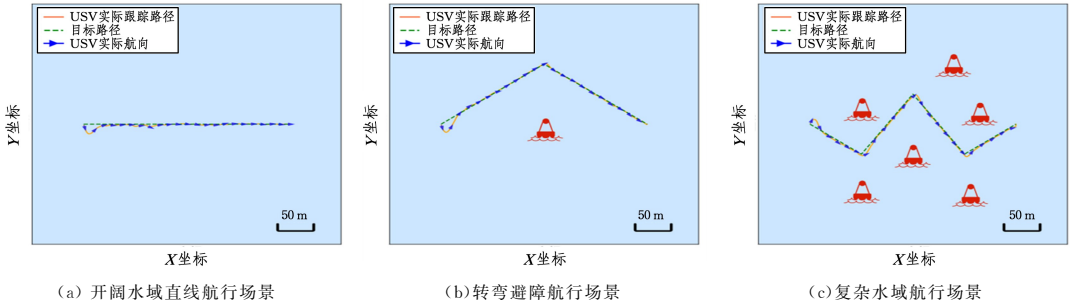


图9 不同目标路径复杂度下 USV 的路径跟踪表现

Fig. 9 USV path tracking performance under different target path complexities

5.6 长程 USV 路径跟踪表现研究

本实验研究了 LECUP 算法在长程 USV 路径跟踪任务中的表现。本节在默认实验设置的基础上,设置了 100, 200, 500, 1000, 5000 时间步的路径跟踪任务,并在实验过程中每 100m 左右设置目标路径的随机转向,其中 5000 时间步的任务设置中 USV 的目标路径约 3000m,以此验证算法在长程 USV 路径跟踪任务中的表现,结果如表 4 所列。

表 4 不同任务长度下 LECUP 的表现

Table 4 Performance of LECUP under different task lengths

Steps	OMD
100	8.01
200	7.68
500	7.92
1000	7.75
5000	7.82

从表 4 中结果可以看出,算法在长程 USV 路径跟踪任务中仍然保持性能不降。产生这样结果的原因是,LECUP 算法使用基于强化学习的决策架构,算法根据当前状态实时调整策略,能有效避免复杂扰动场景下误差的长期积累,从而保证算法在长程 USV 路径跟踪任务中的适应性。

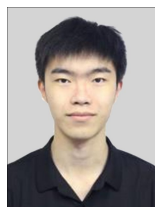
结束语 在风、浪、流等环境干扰以及 USV 自身控制误差的双重影响下,USV 路径跟踪依旧是一项颇具挑战性的任务。尽管 RL 算法具备在复杂环境中自适应学习路径跟踪策略的能力,但其试错训练机制不仅导致样本效率较低,还可能引发潜在的安全风险。为解决这些问题,本文提出了 LECUP 算法,将视线引导与自专家克隆融入 RL,以实现 USV 路径跟踪。LECUP 框架首先在无噪声的静水环境中训练自专家策略,并收集经验数据;随后在复杂环境中,通过数据填充机制对所收集的经验数据进行升维填充与存储,并用处理后的数据来初始化强化学习智能体的价值网络与策略网络。在此基础上,使用 PPO 算法进行强化学习微调,以进一步提升算法性能。此外,为增强算法对不同路径形状的适应能力,LECUP 还引入了 LOS 算法引导模块,该模块能够依据路径几何形状以及 USV 当前位置计算目标航向。大量的实验及分析表明,LECUP 在解决 USV 路径跟踪问题时展现出了优秀的性能与鲁棒性。

值得一提的是,本文主要聚焦于优化 RL 在复杂场景 USV 路径跟踪中的应用模式。在未来的工作中,可以将 LECUP 算法与路径规划算法或实际任务中基于 USV 动力学的控制方法等结合,以提升算法对实际任务场景的适应性。另外,未来可以考虑结合一些多智能体强化学习算法,以探索 LECUP 算法在 USV 编队控制中的应用价值。

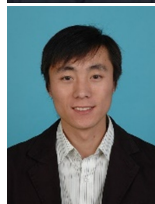
参考文献

- [1] ERM J, MA C, LIU T, et al. Intelligent Motion Control of Unmanned Surface Vehicles: A Critical Review[J]. Ocean Engineering, 2023, 280: 114562.
- [2] ZHANG W S, LI Z, ZHENG Y. The Current Status and Research and Development Trend of Unmanned Ships at Home and Abroad[J]. Ship Science and Technology, 2024, 46(15): 79-83.
- [3] ALIM M F A, KADIR R E A, GAMAYANTI N, et al. Autopilot System Design on Monohull USV-LSS01 Using PID-Based Sliding Mode Control Method[C] // IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2021: 012058.
- [4] CAI W, KORDABAD A B, ESFAHANI H N, et al. MPC-Based Reinforcement Learning for a Simplified Freight Mission of Autonomous Surface Vehicles[C] // 2021 60th IEEE Conference on Decision and Control (CDC). IEEE, 2021: 2990-2995.
- [5] WINURSITO A, DHEWA O A, NASUHA A, et al. Integral State Feedback Controller with Coefficient Diagram Method for USV Heading Control[C] // 2022 5th International Conference on Information and Communications Technology (ICOIACT). IEEE, 2022: 295-300.
- [6] HE S, DAI S L, ZHAO Z, et al. Uncertainty and Disturbance Estimator-Based Distributed Synchronization Control for Multiple Marine Surface Vehicles with Prescribed Performance[J]. Ocean Engineering, 2022, 261: 111867.
- [7] JIANG X, XIA G. Sliding Mode Formation Control of Leaderless Unmanned Surface Vehicles with Environmental Disturbances[J]. Ocean Engineering, 2022, 244: 110301.
- [8] LIU Z, YU L, XIANG Q, et al. Research on USV Trajectory Tracking Method Based on LOS Algorithm[C] // 2021 14th International Symposium on Computational Intelligence and Design (ISCID). IEEE, 2021: 408-411.
- [9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-Level Control Through Deep Reinforcement Learning[J]. Nature, 2015, 518(7540): 529-533.
- [10] PEROLAT J, DE VYLDER B, HENNES D, et al. Mastering the Game of Stratego with Model-Free Multiagent Reinforcement Learning[J]. Science, 2022, 378(6623): 990-996.
- [11] CHEMIN J, HILL A, LUCET E, et al. A Study of Reinforcement Learning Techniques for Path Tracking in Autonomous Vehicles[C] // 2024 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2024: 1442-1449.
- [12] DAI S S, LIU Q. Action Constrained Deep Reinforcement Learning Based Safe Automatic Driving Method [J]. Computer Science, 2021, 48(9): 235-243.
- [13] QIN Y, HUANG B, YIN Z H, et al. Dexpoint: Generalizable Point Cloud Reinforcement Learning for Sim-to-Real Dexterous

- Manipulation[C]// Conference on Robot Learning. PMLR, 2023:594-605.
- [14] HAN D,MULYANA B,STANKOVIC V,et al. A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation[J]. *Sensors*,2023,23(7):3762.
- [15] WEN Y,CHEN Y,GUO X. USV Trajectory Tracking Control Based on Receding Horizon Reinforcement Learning[J]. *Sensors*,2024,24(9):2771.
- [16] WANG X,HONG Y,XU J,et al. PID Controller Based on Improved DDPG for Trajectory Tracking Control of USV[J]. *Journal of Marine Science and Engineering*,2024,12(10):1771.
- [17] GOEL A,CHAUHAN S. Adaptive Look-Ahead Distance for Pure Pursuit Controller with Deep Reinforcement Learning Techniques[C]//Proceedings of the 2021 5th International Conference on Advances in Robotics. 2021:1-5.
- [18] FAN L,WANG G,HUANG D A,et al. SECANT: Self-Expert Cloning for Zero-Shot Generalization of Visual Policies[C]// International Conference on Machine Learning. PMLR, 2021: 3088-3099.
- [19] XIONG L,YANG X,ZHUO G R,et al. Review on Motion Control of Autonomous Vehicles[J]. *Journal of Mechanical Engineering*,2020,56(10):127-143.
- [20] ZHAO H M. Method for Robot Path Tracking Based on Fuzzy Adaptive Tuning PID Control[J]. *Computer Measurement and Control*,2024,32(12):146-152.
- [21] YANG K,TANG X,QIN Y,et al. Comparative Study of Trajectory Tracking Control for Automated Vehicles via Model Predictive Control and Robust H-Infinity State Feedback Control [J]. *Chinese Journal of Mechanical Engineering*,2021,34:1-14.
- [22] ABDILLAH M,MELLOULI E M. A New Adaptive Second-Order Non-Singular Terminal Sliding Mode Lateral Control Combined with Neural Networks for Autonomous Vehicle[J]. *International Journal of Vehicle Performance*,2024,10(1):50-72.
- [23] MANCILLA A,GARCÍA-VALDEZ M,CASTILLO O,et al. Optimal Fuzzy Controller Design for Autonomous Robot Path Tracking Using Population-Based Metaheuristics [J]. *Symmetry*,2022,14(2):202.
- [24] ZHANG X,PAN W,SCATTOLINI R,et al. Robust Tube-Based Model Predictive Control with Koopman Operators[J]. *Automatica*,2022,137:110114.
- [25] FOSSENT I,PETTERSEN K Y,GALEAZZI R. Line-of-Sight Path Following for Dubins Paths with Adaptive Sideslip Compensation of Drift Forces[J]. *IEEE Transactions on Control Systems Technology*,2014,23(2):820-827.
- [26] AZAM S,MUNIR F,RAFIQUE M A,et al. N 2 C: Neural Network Controller Design Using Behavioral Cloning [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2021, 22(7):4744-4756.
- [27] WANG S,CHEN Z,ZHAO Z,et al. EscIRL: Evolving Self-Contrastive IRL for Trajectory Prediction in Autonomous Driving [C]//8th Annual Conference on Robot Learning. 2024.
- [28] CHEN T,ZHANG Z,FANG Z,et al. Imitation Learning from Imperfect Demonstrations for AUV Path Tracking and Obstacle Avoidance[J]. *Ocean Engineering*,2024:298:117287.
- [29] YANGS G,CHO E H,KIM J,et al. Deep Reinforcement Learning-Based Path-Tracking for Unmanned Vehicle Navigation Enhancement[C]// 2024 International Conference on Electronics, Information, and Communication(ICEIC). IEEE,2024:1-4.
- [30] JIANG T M,TAN T,LI H,et al. Path Following of 6-DOF Fixed-Wing UAV Based on Hierarchical Deep Reinforcement Learning[J/OL]. <https://doi.org/10.19678/j.issn.1000-3428.0070197>.
- [31] WANG N,JIA W,WU H J. Path Following of Underactuated Marine Vehicles: A Finite-Time Sideslip-Tangent LOS Guidance Approach[J]. *Control and Decision*,2025,40(1):187-195.
- [32] YANG Z K,ZHONG W B,FENG Y B,et al. Unmanned Surface Vehicle Track Control Based on Improved LOS and AD-RC[J]. *Chinese Journal of Ship Research*,2021,16(1):121-127,135.
- [33] YANG C,JIANG X,BAI B,et al. Path Following Control of PID Controller Parameters Optimized by Genetic Algorithm [J]. *Manufacturing Automation*,2022,44(5):78-81.
- [34] ZHANG J,ZHANG W,TONG S. Adaptive Neural Optimal Tracking Control for Uncertain Unmanned Surface Vehicle[J]. *Ocean Engineering*,2024,312:119031.
- [35] ZHU D,TAO R N,CHEN W,et al. LSTM-Based Sliding Mode Trajectory Tracking Control Algorithm for Unmanned Surface Vehicles[J]. *Electronic Measurement Technology*,2024,47(7): 61-68.
- [36] YANG S M,SHAN Z,DING Y,et al. Survey of Research on Deep Reinforcement Learning[J]. *Computer Engineering*,2021, 47(12):19-29.
- [37] LILLICRAP T P. Continuous Control with Deep Reinforcement Learning[J]. *arXiv*:1509.02971,2015.
- [38] HAARNOJA T,ZHOU A,ABBEEL P,et al. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor[C]// International Conference on Machine Learning. PMLR,2018:1861-1870.
- [39] SCHULMAN J,WOLSKI F,DHARIWAL P,et al. Proximal Policy Optimization Algorithms[J]. *arXiv*:1707.06347,2017.
- [40] KOSTRIKOV I,NAIR A,LEVINE S. Offline Reinforcement Learning with Implicit Q-Learning [J]. *arXiv*: 2110.06169, 2021.
- [41] NAKAMOTO M,ZHAI S,SINGH A,et al. Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning[J]. *Advances in Neural Information Processing Systems*,2023,36: 62244-62269.
- [42] LUO Y,JI T,SUN F,et al. Goal-Conditioned Hierarchical Reinforcement Learning with High-Level Model Approximation[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024,36(2):2705-2719.
- [43] YU C,VELU A,VINITSKY E,et al. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games[J]. *Advances in Neural Information Processing Systems*,2022,35:24611-24624.
- [44] KARIMIH R,LU Y. Guidance and Control Methodologies for Marine Vehicles: A Survey[J]. *Control Engineering Practice*, 2021,111:104785.



LIU Jiahui, born in 1999, postgraduate. His main research interests include deep reinforcement learning and path tracking of unmanned surface vehicles.



LI Jiantao, born in 1982, senior engineer. His main research interest is ICT for China aviation industries.