文章编号:1000 - 4653(2025)03 - 0129 - 08

基于深度强化学习的无人艇航迹规划与控制

关 巍, 奚赵勇, 崔哲闻, 张显库
(大连海事大学 航海学院, 辽宁 大连 116026)

摘 要:本研究旨在运用强化学习方法解决无人艇航迹规划与控制问题。在航迹规划方面,采用 Q 学习(Q-learning)算法,针对真实水域进行航迹规划。在奖励函数设计中考虑了浅水区,并致力于减少航迹的转向点数量。在航迹控制方面,将柔性动作评价(SAC)算法与比例积分微分(PID)控制算法相结合,克服了传统 PID 控制器参数人工整定、调节困难的问题的同时,也规避了深度强化学习缺乏可解释性的缺点。通过与传统 PID 算法、遗传算法(GA)和深度确定性策略梯度(DDPG)进行对比试验,展现出所提出 SAC-PID 方法的优越性。仿真结果表明,所规划的航迹能够综合考虑航迹距离、浅水区、航路转向点数量等优化目标,所提出的 SAC-PID 方法能够很好实现航迹跟踪效果。

关键词:航迹规划; 航迹控制; 深度强化学习; 水面无人艇
中图分类号:U674.91 **文献标志码**:A **DOI**:10.3969/j.issn.1000 - 4653.2025.03.016

Trajectory planning and control for unmanned surface vehicle based on deep reinforcement learning

GUAN Wei, *XI Zhaoyong*, *CUI Zhewen*, *ZHANG Xianku*
(Navigation College, Dalian Maritime University, Dalian 116026, China)

Abstract: This study aims to apply deep reinforcement learning to address the challenges of trajectory planning and control for unmanned surface vehicles. In trajectory planning, the Q-learning algorithm is employed to generate trajectories in real-world aquatic environments. For the design of the reward function, factors such as shallow water areas are taken into account, with an emphasis on minimizing the number of turning points along the path. For trajectory tracking control, we integrate the Soft Actor-Critic (SAC) algorithm with the Proportional-Integral-Derivative (PID) control method to alleviate the difficulties of manual parameter tuning associated with conventional PID controllers. This hybrid approach also mitigates the interpretability limitations often found in pure deep reinforcement learning methods. Comparative experiments involving the traditional PID algorithm, Genetic Algorithm (GA), and Deep Deterministic Policy Gradient (DDPG) algorithm demonstrate the superiority of the proposed SAC-PID method. Simulation results show that the planned trajectories effectively incorporate multiple factors, including travel distance, shallow water regions, and number of turning point, the SAC-PID method achieves outstanding performance in trajectory tracking.

Key words: trajectory planning; trajectory control; deep reinforcement learning; unmanned surface vehicle

智能船舶航迹规划与控制是航海领域的重要研究方向。传统的路径规划方法和控制算法已很难达到人们的期望要求,面临越来越多新的挑战。深度强化学习的到来,为航海领域智能船舶航迹规划与控制带来了新的机遇与可能。

传统规划方法规划的路径存在不符合研究对象运动学模型的问题,夏雨奇等[1]基于经验分类提出了一种具有一定泛化能力的状态空间的深度 Q 网

收稿日期:2024-09-01
基金项目:国家自然科学基金项目(52171342)
通信作者:关 巍(1982—),男,博士,教授,博士生导师,研究方向为船舶智能航行决策,船舶运动控制。E-mail:gwwtxdy@dlmu.edu.cn
引用格式:关巍,奚赵勇,崔哲闻,等.基于深度强化学习的无人艇航迹规划与控制[J].中国航海,2025,48(3):129 - 136.
GUAN W, XI Z Y, CUI Z W, et al. Trajectory planning and control for unmanned surface vehicle based on deep reinforcement learning [J]. Navigation of China, 2025, 48(3): 129 - 136. (in Chinese)

络方法用于无人机路径规划。李敏等[2]针对四足机器人通过SAC算法训练得到了低能耗的越障策略与轨迹规划参数,解决了四足机器人在越障行进中存在的能耗高与关节振动问题。舒健生等[3]将SAC算法用于二维平面无人机航迹规划,在实时性和航迹平滑度上取得了较好的效果。在船舶路径规划方面,欧昌奎等[4]基于船舶的历史轨迹进行全局路径规划,利用DDPG算法设计局部路径避碰方法,利用深度Q网络设计局部路径回归方法,从而让船舶能够规避局部动态风险,实现安全航行。

航迹跟踪与控制是实现智能船舶安全航行的关键技术,PID控制算法和随后开发的模糊PID算法已经在实际船舶上得到验证,可以有效地实现无人艇的航迹控制[5]。针对存在未知系统不确定性和时变外部干扰的水下航行器轨迹跟踪问题,LIU等[6]提出了一种基于非线性扰动观测器的反演有限时间滑模控制方案,通过将反步法与滑模控制相结合来构造控制框架,保证了系统的有限时间稳定性。FAN等[7]结合定时扩展状态观测器和定时微分器,提出了一种定时滑模控制方法,在降低控制器设计难度的同时,缩短了误差的收敛时间。ZHAO等[8]基于深度Q网络和强化学习的平滑收敛方法,降低了无人艇(Unmanned Surface Vehicles, USV)模型路径跟踪控制律的复杂性,为深度强化学习在轨迹跟踪控制中的应用提供了范例。李诗杰等[9]将无模型自适应控制与自抗扰控制相结合,能够让船舶克服复杂环境下的干扰,实现航迹跟踪与智能航行。

PID控制算法因其简单性、有效性在工业系统中被广泛使用,在船舶控制领域也不例外。然而,它仍然存在明显的限制,包括需要手动调整PID增益和无法动态修改调整后的增益[10]。吴沁等[11]通过将粒子群算法、神经网络和PID控制相结合,实现智能参数整定的同时,大大提升了控制精度和稳定性。CARLUCHO等[12]提出了一种基于演员-评论家(Actor-Critic)框架的混合控制策略,能够成功地控制诸如水下航行器、地面移动机器人等具有不同动力学和几个转向点要求的多输入多输出系统。LEE等[13]采用DDPG算法开发了船舶动态定位系统的自适应PID控制器。该控制器不仅考虑了大型船舶响应慢的特点,并且对自适应控制器增益的范围进行限制。但该方法在不同环境下的泛化效果不够优秀。LAI等[14]提出了一种基于近端策略优化(Proximal Policy Optimization,PPO)的智能自适应PID控制器,以实现USV的航向保持。张梦杰等[15]将双向长短期记忆网络融合进双延迟深度确定性策略梯度(Twin Delayed Deep Deterministic Policy Gradient, TD3)算法中,从而使PID参数自整定更加精确,适应性更强。

在上述研究成果中,可以发现很多路径规划和控制成果为非船舶领域成果。此外,大多数规划算法并没有结合实际水域和航线要求进行路径规划。航迹控制方面,传统控制方法已难以适应当下对于智能船舶的控制性能要求,深度强化学习方法又因其缺乏可解释性[16,17]和安全性[18]而迟迟不能广泛应用于实船。因此,本文首先利用Q学习算法综合考虑实际水域的深水区、浅水区以及航路转向点数量,进行航迹规划。接着将SAC算法与PID控制算法相结合,提出SAC-PID方法用于智能船舶航迹跟踪与控制,最终获得了出色的控制效果与泛化能力。

1 背景知识

1.1 PID控制算法

PID控制器由比例、积分和微分单元组成,具有原理简单、易于实现、适用性广等优点。其数学表达式如下:

$$u(t) = k_p e(t) + k_i \int e(t)\mathrm{d}t + k_d \frac{\mathrm{d}e(t)}{\mathrm{d}t} \quad (1)$$

式中:$u(t)$为控制器的输出,$e(t)$表示在时刻t的误差,k_p、k_i、k_d分别是PID控制器的比例、积分、微分系数。

后续随着计算机的发展,又出现了位置式PID和增量式PID。本文采用增量式PID[19],其表达式如下:

$$\Delta u(t) = k_p[e(t) - e(t-1)] + k_i e(t) + k_d[e(t) - 2e(t-1) + e(t-2)] \quad (2)$$

式中:$\Delta u(t)$为控制增量。$e(t)$,$e(t-1)$,$e(t-2)$为误差在t、$t-1$、$t-2$时刻的采样。

1.2 船舶运动数学模型

船舶运动数学模型是研究船舶运动控制的基础,被广泛应用于船舶运动控制器的设计和船舶运动模拟器的开发。船舶运动数学模型,如图1所示。

图1中$O_n - X_n Y_n Z_n$为惯性坐标系,O_n表示起始点,一般选在船舶重心所在位置,$O_n X_n$轴指向正北,$O_n Y_n$轴指向正东,$O_n Z_n$轴垂直于水平面并指向船底。$O_b - x_b y_b z_b$为附体坐标系,O_b一般选择船舶重心处,$O_b x_b$轴沿船舶中心线指向船头,$O_b y_b$指向右舷,$O_b z_b$指向地心。ψ表示航向角。u,v,r分别表示前进速度,横漂速度和艏摇角速度。本文沿用的船舶运动方程[20]如下所示:

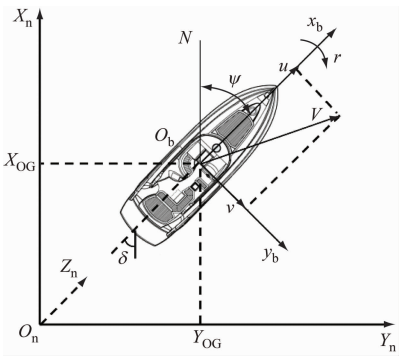


图 1　船舶运动数学模型

Fig. 1　The ship motion mathematical model

$$\dot{\boldsymbol{\eta}}=\boldsymbol{R}(\psi)\boldsymbol{V}\boldsymbol{M}\dot{\boldsymbol{V}}+\boldsymbol{C}(\boldsymbol{V})\boldsymbol{V}+\boldsymbol{N}(\boldsymbol{V})\boldsymbol{V}+g(\boldsymbol{V})=\boldsymbol{\tau}_f+\boldsymbol{\tau}_w \tag{3}$$

式中:$\boldsymbol{\eta}=[x \quad y \quad \psi]^{\mathrm{T}}$ 代表无人艇的位置坐标以及航向。$\boldsymbol{R}(\psi)$ 为无人艇的旋转矩阵,$\boldsymbol{V}=[u \quad v \quad r]^{\mathrm{T}}$ 表示无人艇的运动速度矢量。$\boldsymbol{M}$ 表示系统惯性矩阵,$\boldsymbol{C}(\boldsymbol{V})$ 由向心矩阵和流体动力学 Coriolis 矩阵组成,$\boldsymbol{N}(\boldsymbol{V})$ 为非线性阻尼矩阵,$g(\boldsymbol{V})=[g_u \quad g_v \quad g_r]^{\mathrm{T}}$ 代表未建模的动力学模型。$\boldsymbol{\tau}_f$ 表示推进力和艏摇力矩,该部分主要通过控制器控制船舶的舵角和螺旋桨转速来实现,$\boldsymbol{\tau}_w$ 为环境干扰外力的总和。

1.3　深度强化学习

强化学习任务通常用马尔可夫决策过程来描述。智能体与环境进行交互,每次接收到一个状态 $\boldsymbol{s}_t$,再根据策略 $\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)$ 选择动作 $\boldsymbol{a}_t$,凭借该动作在环境中的表现,智能体会收到奖励 r_t,之后进行状态迁移。强化学习便是通过不断地试错来探索添加折扣因子 γ 后的最大总累积奖励 $G_t=\sum_{n=0}^{\infty}\gamma^n r_{t+n+1}$,从而找到最优策略的方法。

当遇到连续和复杂的状态集合时,传统的强化学习方法将面临挑战。深度学习技术的引入为强化学习提供了新的可能性。通过神经网络来近似表达价值函数和策略函数,如图 2 所示。演员-评论家框架中,二者对应的网络参数分别为 θ 和 ω,前者通过策略梯度进行更新,后者则往往通过均方误差损失函数进行参数更新。

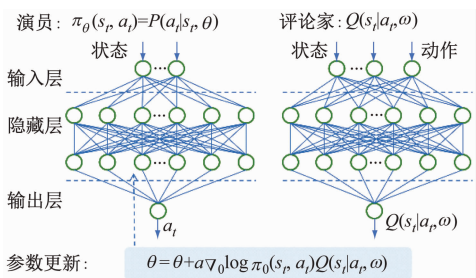


图 2　深度强化学习网络结构示意

Fig. 2　The network structure of DRL

2　船舶航迹规划与控制方法

2.1　整体框架

航迹规划与控制的整体框架,如图 3 所示。首先,对真实水域进行建模。考虑到采用 Q 学习算法完全可以满足我们的要求,我们通过 Q 学习将动作和奖励构建成一张 Q 表来储存 Q 值,进而根据 Q 值来选取获得最大收益的动作,最终规划出航迹。接着进行航迹控制部分,将规划好的航迹信息输入到航迹控制模块中。在航迹控制模块中,SAC 算法被用于 PID 参数整定,其内部有 5 个网络,网络 Q_1 和 Q_2 用于减少 Q 值计算中的过估计问题,网络 V_1 和 V_2 则通过软更新的方式进行参数更新,能够很好地兼顾算法稳定性与速度。智能体从环境中获取位置误差和偏航误差等作为状态输入,进而输出动作(即控制器参数)赋予 PID 控制器,接着控制器通过计算输出舵角和速度控制量对无人艇进行航迹控制。通过对无人艇航迹跟踪效果收集和评价,并给予相应奖励反馈给智能体,以用于智能体对该动作进行评价,实现策略寻优。当网络参数稳定后,将会获得较高的平均奖励,并使算法收敛至最优策略。

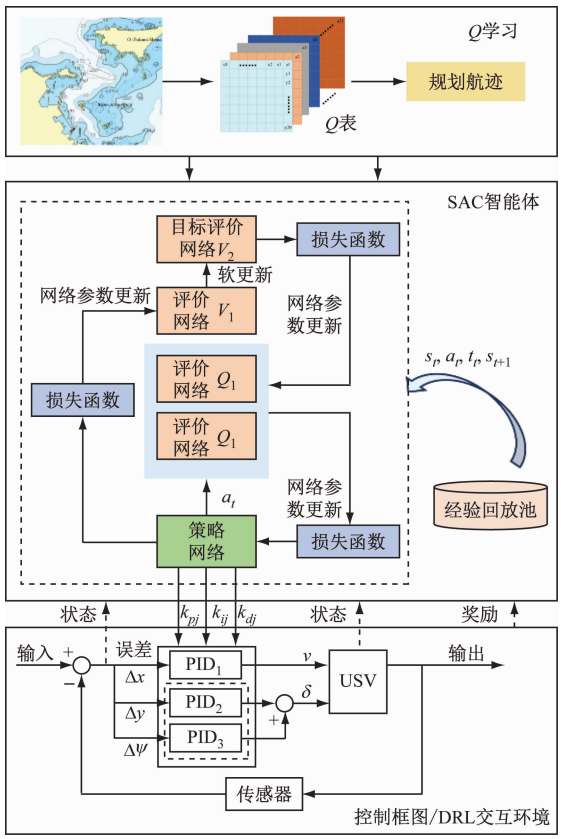


图 3　USV 航迹规划与控制整体框架示意图

Fig. 3　Schematic diagram of the overall framework for USV trajectory planning and control

2.2　航迹规划

航迹规划部分使用 Q 学习方法[21],它是强化学

习中基于值的方法,由于使用了时序差分法,因此是一种离线方法。将状态与动作构建成 Q 表进而通过贪心策略选取 Q 值对应动作找到最优策略。Q 表依据下式进行更新:

$$Q(\boldsymbol{s},\boldsymbol{a})\leftarrow(1-\alpha)Q(\boldsymbol{s},\boldsymbol{a})+\alpha[r(\boldsymbol{s},\boldsymbol{a})+\gamma\max_{\boldsymbol{a}'}Q(\boldsymbol{s}',\boldsymbol{a}')] \tag{4}$$

式中:α 为学习速率,γ 为折扣因子。

状态空间为栅格化后的信息,鉴于实际航迹较为简单,因此动作空间设计为前后左右以及斜角的 8 个方向,即[正北、正南、正东、正西、北偏东 45°、北偏西 45°、南偏东 45°、南偏西 45°]8 个维度。

奖励函数设计部分,通常需要考虑抵达目标点、碰撞、和行进一步所获奖励。为了满足实际水域和航迹要求,我们对奖励函数进行改进,增加了深水区、浅水区的奖励函数以及控制航路转向点数量的奖励函数,从而使船舶尽可能以较少的航路转向点的规划航迹抵达终点,在此过程中不会发生与静态障碍物的碰撞,且综合考虑了对浅水区的穿越和路径长短(即大大减少路径长度,同时允许船舶短暂经过浅水区)。具体奖励函数设计如下。

(1)碰撞与到达奖励函数:

$$G_1=\begin{cases}p, & \forall P_{i,t}=P_{\text{destination}}\\ -p, & P_{i,t}\in D_{\text{obstacle}}\end{cases} \tag{5}$$

式中:p 为抵达奖励系数。$P_{i,t}$表示在第 i 次迭代的 t 时刻本船位置坐标,$P_{\text{destination}}$表示目的地位置坐标,D_{obstacle}为障碍物范围的位置坐标集合。

(2)最优路线奖励函数:

$$G_2=-\rho \tag{6}$$

式中:ρ 为距离优化系数,船舶每行进一步得到一个较小的负的奖励,从而路径越短,在此函数中奖励越大。

(3)深水区浅水区奖励函数:

$$G_3=\begin{cases}0, & \text{深水区}\\ -r_{\text{shallow}}, & \text{浅水区}\end{cases} \tag{7}$$

每当船舶在浅水区范围走一步便会得到一个较小的负的奖励 r_{shallow},旨在引导船舶尽量寻找深水区最优航迹。

(4)控制航路转向点数量的奖励函数:

$$G_4=-r_{\text{turn}},\ F_t\neq F_{t-1} \tag{8}$$

式中:r_{turn}为航路转向点奖励系数,F_t 表示 t 时刻船舶的航行方向;F_{t-1}表示 $t-1$ 时刻船舶的航行方向。当航路出现转向点,即动作选择出现变化时,智能体会得到一个轻微的负奖励,该奖励函数旨在控制航路转向点数量尽可能少,降低航迹的复杂性。

(5)总奖励函数:

$$R_{\text{total}}=\eta_1G_1+\eta_2G_2+\eta_3G_3+\eta_4G_4 \tag{9}$$

最终的奖励函数便是以上四个分支奖励函数加权后的总和,η_1、η_2、η_3、η_4 为加权系数。

2.3 航迹控制

本文通过 SAC 算法整定 PID 控制器参数,进而实现航迹控制。SAC 算法[22]是一种基于最大熵的无模型深度强化学习算法,引入最大熵的目的在于增加算法的探索能力,使其动作输出不集中在一个最优动作上。

控制部分,本文使用由 3 个 PID 控制器组成的控制器组对无人艇的速度、舵角进行控制,具体控制框图如图 3 所示。用于整定 PID 控制器参数的 SAC 算法的状态空间定义如下:

$$\boldsymbol{s}_t=[\Delta x_i,\ \Delta y_i,\ \Delta\psi_i,u,r] \tag{10}$$

式中:Δx_i、Δy_i、$\Delta\psi_i(i=1,2,3)$分别表示以本船为基准向前采样的 3 个点与本船当前位置的纵、横坐标误差和航向偏差,u 表示无人艇的速度,r 表示无人艇的艏摇角速度。

动作空间定义如下:

$$\boldsymbol{a}_t=[k_{p1},k_{i1},k_{d1},k_{p2},k_{i2},k_{d2},k_{p3},k_{i3},k_{d3}] \tag{11}$$

式中:k_{pj}、k_{ij}、$k_{dj}(j=1,2,3)$表示 3 个 PID 控制器的比例系数、积分系数、微分系数。PID_1 控制器接收误差 Δx,输出航速 u 给 USV;PID_2 控制器和 PID_3 控制器进行组合,分别接收误差 Δy 和 $\Delta\psi$,二者输出相加对 USV 舵角进行控制。

最大熵强化学习的目标函数如下所示:

$$J(\pi)=\sum_{t=0}^{T}E_{(\boldsymbol{s}_t,\boldsymbol{a}_t)\sim\rho_\pi}[r(\boldsymbol{s}_t,\boldsymbol{a}_t)+\mu H(\pi(\boldsymbol{a}_t|\boldsymbol{s}_t))] \tag{12}$$

式中:ρ_π 表示在策略 π 下的状态动作分布,$H(\pi(\boldsymbol{a}_t|\boldsymbol{s}_t))$表示熵,$\mu$ 为权重,用来衡量熵对目标函数的重要性,即最优策略的随机程度。

SAC 算法中,动作价值和状态价值被重新定义如下:

$$Q(\boldsymbol{s}_t,\boldsymbol{a}_t)=E_{\boldsymbol{s}_{t+1}\sim\boldsymbol{D}}[r(\boldsymbol{s}_t,\boldsymbol{a}_t)+\gamma V(\boldsymbol{s}_{t+1})] \tag{13}$$

$$V(\boldsymbol{s}_t)=E_{\boldsymbol{a}_t\sim\pi}[Q(\boldsymbol{s}_t,\boldsymbol{a}_t)+\mu H(\pi(\boldsymbol{a}_t|\boldsymbol{s}_t))]=E_{\boldsymbol{a}_t\sim\pi}[Q(\boldsymbol{s}_t,\boldsymbol{a}_t)-\mu\log\pi(\boldsymbol{a}_t|\boldsymbol{s}_t)] \tag{14}$$

其中,SAC 算法包含 5 个网络,分别是 4 个评价网络和 1 个策略网络,如图 3 所示,$\boldsymbol{D}$ 表示经验回放池。

评价网络 V_1 的目标函数如下所示:

$$J_V(\zeta)=E_{\boldsymbol{s}_t\sim\boldsymbol{D}}\left[\frac{1}{2}(V_\zeta(\boldsymbol{s}_t)-E_{\boldsymbol{a}_t\sim\pi_f}[\min_{i=1,2}Q_{\theta_i}(\boldsymbol{s}_t,\boldsymbol{a}_t)-\mu\log\pi_f(\boldsymbol{a}_t|\boldsymbol{s}_t)])^2\right] \tag{15}$$

式中:ζ 为评价网络 V_1 的网络参数,$\theta_i(i=1,2)$ 表示评价网络 Q_1 和 Q_2 的网络参数。此处借鉴双 Q 网络的做法,取两个 Q 值中的较小值以减少过估计问题。

目标评价网络 V_2 的参数通过软更新实现:

$$\bar{\zeta} = \tau\zeta + (1-\tau)\bar{\zeta} \tag{16}$$

式中:$\bar{\zeta}$ 为评价网络 V_2 的网络参数,τ 为软更新系数,通常取较小的正数,从而使得更新不急促。恰当的 τ 值有利于算法的稳定并可加快收敛速度。

评价网络 Q_1 和 Q_2 具有相同的结构和网络参数。其目标函数如下所示:

$$J_Q(\theta) = E_{(s_t,a_t)\sim D}\left[\frac{1}{2}(Q_\theta(\boldsymbol{s}_t,\boldsymbol{a}_t) - \hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t))^2\right] \tag{17}$$

$$\hat{Q}(\boldsymbol{s}_t,\boldsymbol{a}_t) = r(\boldsymbol{s}_t,\boldsymbol{a}_t) + \gamma E_{s_{t+1}\sim D}[V_{\bar{\zeta}}(\boldsymbol{s}_{t+1})] \tag{18}$$

策略网络的目标函数定义如下:

$$J_\pi(\phi) = E_{s_t\sim D,\varepsilon_t\sim N}[\log \pi_\phi(\tilde{\boldsymbol{a}}_t|\boldsymbol{s}_t) - \min_{i=1,2} Q_\theta(\boldsymbol{s}_t,\boldsymbol{a}_t)] \tag{19}$$

$$\tilde{\boldsymbol{a}}_t = f_\phi(\varepsilon_t;\boldsymbol{s}_t) \tag{20}$$

式中:$\tilde{\boldsymbol{a}}_t$ 通过神经网络重参数化得到,ε_t 为输入的噪声。

状态空间包含了位置误差以及航向误差,动作空间则是 PID 控制器增益。奖励函数设置如下:

$$r_t = \begin{cases} \rho_d d(i) + \rho_v v(i) + l, & \text{到达目标点} \\ \rho_d d(i) + \rho_v v(i) - l, & \text{偏航} \\ \rho_e e_d(i) + 1, & \text{其他} \end{cases} \tag{21}$$

式中:$d(i)$ 为无人艇第 i 轮行驶过的距离,$v(i)$ 为无人艇的平均速度,$e_d(i)$ 为无人艇与下一个采样点的距离,l 为惩罚项,即到达目标点给正的奖励值,偏航则给出负的奖励值,ρ_d、ρ_v、ρ_e 为相关系数。

3　仿真试验与分析

3.1　规划航迹对比试验

规划航迹试验中,公式(9)的权值对试验结果有很大影响,本文分别取 $\eta_1=0.7$,$\eta_2=0.1$,$\eta_3=0.1$,$\eta_4=0.1$。权重大小主要与分支奖励函数重要程度有关,碰撞与到达奖励函数 G_1 直接决定了训练的成败,因此分配的权重最高。Q 学习算法的平均奖励曲线,如图 4 所示。从中可以看出,Q 学习奖励函数曲线在 800 轮附近收敛。

选取日本伊势湾和答志岛附近的真实水域作为训练环境,航迹规划对比结果如图 5 所示。其中,五角星表示终点,黑色区域表示陆地,视为障碍物。灰色区域表示浅水区,白色区域表示深水区。实线表示使用本文前述奖励函数的 Q 学习方法所规划的航线,点画线为将除起始点范围之外的浅水区视为障碍物时的 A * 算法所规划的航线,记作 A * 规划航迹 1。虚线为不考虑浅水区时的 A * 算法所规划的航迹,记作 A * 规划航迹 2。点划线表示不考虑浅水区时人为规划的最简航线。此处应当注意该人所规划的最简航迹基于两个前提,首先是规划时设置的无人艇的动作空间为[正北、正南、正东、正西、北偏东 45°、北偏西 45°、南偏东 45°、南偏西 45°]八个维度;其次是此处规划的航迹仅为航迹规划者结合航海图书资料,为无碰撞抵达目的地而规划的最简航迹。

同样是考虑浅水区,Q 学习通过奖励函数给予惩罚,而 A * 是将浅水区视作障碍物。对比 Q 学习规划航迹和 A * 规划航迹 1 可知,Q 学习规划航迹拥有更短的航行距离,且仅短暂经过浅水区。而 A * 规划航迹 2 则更多的经过浅水区,航行过程容易受浅水效应影响,可能会给船舶安全航行带来威胁。同样地,人为规划航迹尽管航路转向点最少,但长时间经过浅水区,不利于船舶安全航行。因此,采用 Q 学习算法所规划的航线为最优。

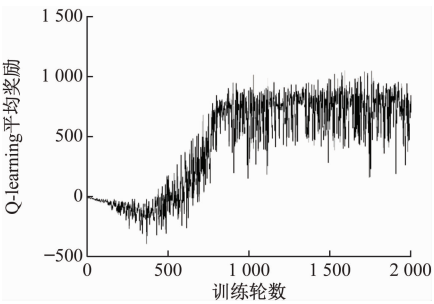


图 4　Q 学习算法平均奖励值曲线

Fig. 4　The average reward value of Q-learning algorithm

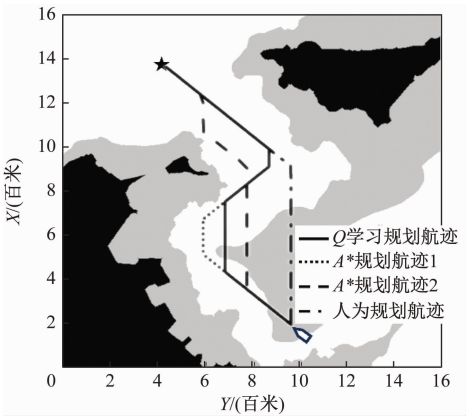


图 5　真实水域中的规划航迹对比

Fig. 5　Comparison of planning trajectories in the real water

3.2　SAC-PID 参数整定试验

本文仿真试验中用于航迹控制器设计和仿真试

验对比所使用的船舶模型的操纵性指数分别为 $K=0.524\ 48$，$T=0.169$。同时，公式(3)所描述的船舶运动数学模型用于构建 PyBullet 平台下所使用的船舶运动物理仿真模型，以进行后续 SAC-PID 控制模型的训练和航迹控制仿真研究[23]。

SAC 算法训练过程的平均奖励值曲线，如图 6 所示。从中可以看出，一开始智能体未能寻找到好的策略，甚至出现航迹偏离而无法抵达目标点，因此奖励函数值为负。SAC 智能体在训练 100 轮左右有了一个明显的上升趋势，这可能意味着其找到了一个不错的策略，但随着后续的探索，奖励值骤降，直至 300 轮以后 SAC 智能体才逐渐找到了最优策略，平均奖励值曲线开始上升，并在大约 400 轮时收敛，波动在 90 左右。

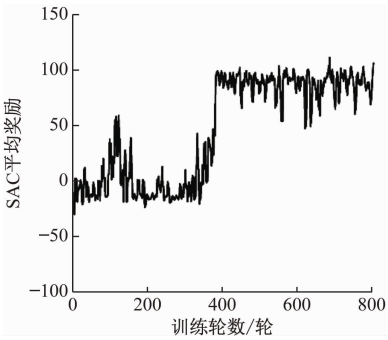


图 6 SAC 算法平均奖励值曲线

Fig. 6 The average reward value of SAC algorithm

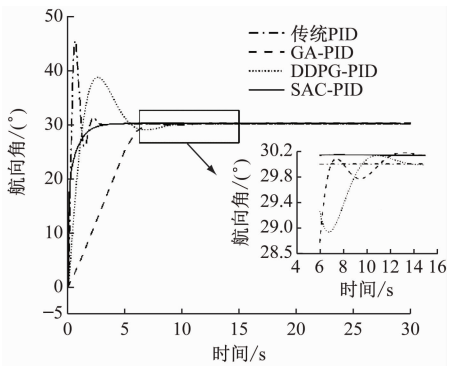


图 7 不同 PID 整定方法下的响应曲线对比

Fig. 7 Comparison of response curves under different PID tuning methods

接着，将 SAC-PID 方法与其他 PID 参数整定方法进行对比。传统 PID 方法中，由于等幅振荡法不能直接应用于二阶系统，而响应曲线法则不适用于含有积分环节的系统，因此，引入东北大学薛定宇教授提出的近似模型方法，对船舶模型近似表达后，再使用等幅振荡法进行参数整定。本文选择遗传算法作为启发式算法的代表，用于 PID 整定效果对比。此外，还将 SAC 算法与同为深度强化学习方法的 DDPG 算法进行横向对比。期望的舵角输出设置为 30 度，响应曲线对比结果，如图 7 所示。

从图中可以看出，传统 PID 和 DDPG-PID 的超调量较大，分别为 50% 和 26.7%，且他们的输出峰值均超过了 35 度，不符合船舶舵角限制。此外，在所有方法中 GA-PID 的调节时间最长，SAC-PID 不仅超调量小而且响应快速。通过局部放大图可以看出 SAC-PID 的稳态误差仅为 0.4%，在可被允许的 2% ~5% 范围内。可以得出推论，这些误差对于 SAC 智能体的奖励函数评价体系影响较小，对于仿真过程中的船舶航迹跟踪效果影响甚微，可以忽略不计。图 8 为 SAC 动作输出变化曲线。横坐标表示对期望航迹进行采样的时刻。即在 SAC 参数整定完成后，我们得到的并不是一组参数，而是一系列经过整定的 PID 控制参数集合。因此，在航迹跟踪控制任务中，通过 SAC 智能体与环境的交互，船舶可以根据环境状态实现 PID 控制器参数的自适应整定。

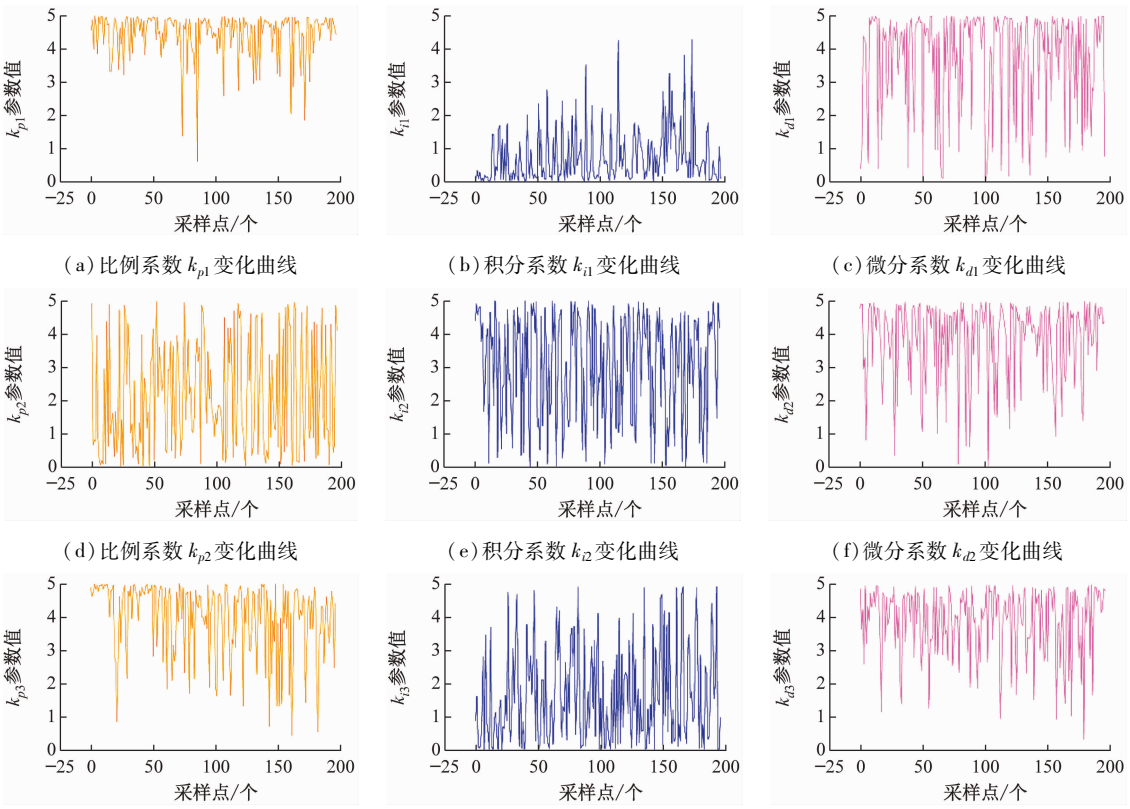
3.3 航迹跟踪仿真试验

针对日本伊势湾和答志岛附近的真实海域，采用 Q 学习方法规划的航迹，将其作为期望航迹，用于验证 SAC-PID 方法的航迹控制器的效果。不同方法的航迹跟踪曲线对比如图 9a 所示，小船表示起始位置，五角星标志着终点。图 9b 是局部放大图，可见在航路转向点处，航迹控制效果面临挑战。结合不同控制方法的平均航迹跟踪误差可知，传统 PID 方法的航迹跟踪误差为 44.292 4 米，GA-PID 的航迹跟踪误差为 47.653 1 米，SAC-PID 的航迹跟踪误差为 14.211 3 米，DDPG-PID 的航迹跟踪误差为 17.803 8 米。因此可以得出结论，基于深度强化学习和 PID 控制器的平均跟踪误差小于传统方法和启发式算法，其中 SAC-PID 的航迹跟踪误差最小。

为进一步探究航迹跟踪过程中的舵角变化情况，在图 10 中绘制出其变化曲线，四次大幅舵角变化对应四个航线转向点。由于传统 PID 与 GA-PID 的航迹跟踪误差较大，因此着重对比另外两种基于深度强化学习方法的舵角变化曲线。从图 10 中可以看出，DDPG-PID 的舵角存在频繁波动，不利于船舶安全航行。相比之下，SAC-PID 的舵角波动较为稳定。

综合上述试验可以看出，SAC-PID 很好地完成了航迹跟踪任务，成功抵达目标点，而且在此过程中最大限度降低了航迹误差，减少了舵角的非必要频繁振荡。同时该方法使得 PID 参数可以实时自适应调整。相比于其他 PID 参数整定方法，SAC-PID 方

法展现出无可比拟的优越性。



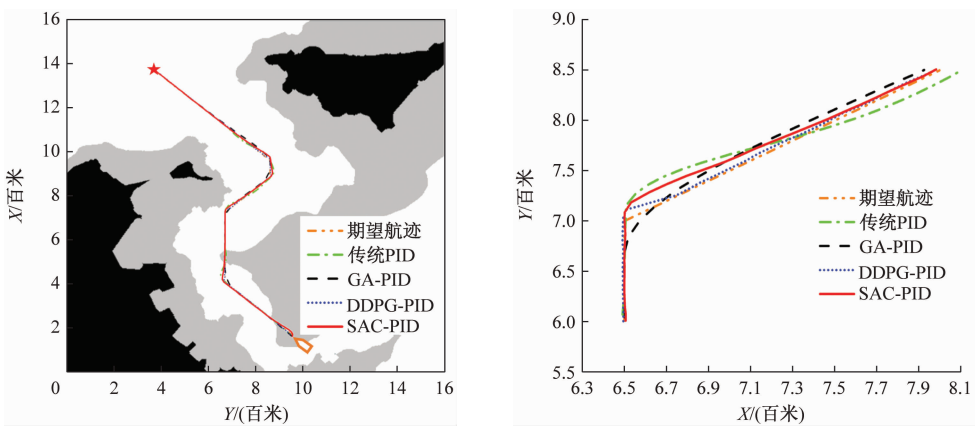
(a)比例系数 k_{p1} 变化曲线　(b)积分系数 k_{i1} 变化曲线　(c)微分系数 k_{d1} 变化曲线

(d)比例系数 k_{p2} 变化曲线　(e)积分系数 k_{i2} 变化曲线　(f)微分系数 k_{d2} 变化曲线

(g)比例系数 k_{p3} 变化曲线　(h)积分系数 k_{i3} 变化曲线　(i)微分系数 k_{d3} 变化曲线

图 8　动作输出曲线(PID 控制器组的比例系数、积分系数、微分系数变化曲线,采样间隔 4 秒)

Fig. 8　The output curves of actions (Proportional coefficient, integral coefficient, and differential coefficient variation curves of PID controller group, sampling interval is 4 seconds)



(a)真实海洋环境的航迹跟踪对比示意　(b)航迹局部放大示意

图 9　真实海洋环境中航迹跟踪控制效果对比

Fig. 9　Comparison of trajectory tracking control effects in the real marine environment

4　结　论

通过以上试验与分析可知,本文提出的基于深度强化学习算法的航迹规划与控制方法能够有效满足智能船舶自主航行要求。基于 Q 学习的航迹规划方法综合考虑了多个目标要求,包括航路安全性、航路距离、航路转向点数量以及浅水区等。而用于航迹控制的 SAC-PID 方法克服了传统 PID 参数整定需要手动调整和整定后难以修改的局限性,智能体只需要通过与环境的不断交互就可以寻找到更优的控制参数。此外,将深度强化学习与传统 PID 控制算法相融合也规避了深度强化学习在安全性和可

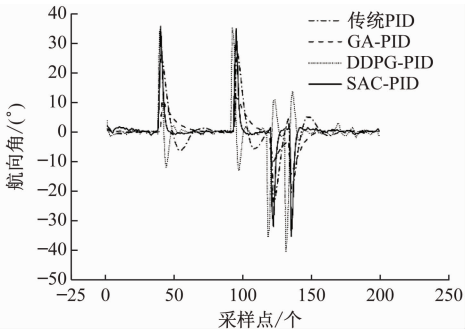


图10 不同控制方法下的船舶舵角变化曲线对比

Fig. 10 Comparison of ship rudder angle variation under the different control methods

解释性方面存在的问题。仿真试验结果表明,SAC-PID具有良好的控制性能,其平均航迹误差小于其他对比方法。

综上所述,本文对深度强化学习算法在海洋工程领域的应用具有积极意义。与陆地车辆行驶规划方法不同,本文增加了对海洋实际航行中浅水区和航路转向点数量的考量,以提升航线设计的质量,并帮助船舶规避浅水效应。同时,将SAC算法与PID控制方法的结合也有利于智能船舶针对不同的环境状态进行控制器控制参数的自适应整定,优化航迹控制的效果。未来研究将致力于提升所提出方法在更为复杂海洋工程应用场景中的抗干扰性能。

参 考 文 献

[1] 夏雨奇, 黄炎焱, 陈恰. 基于深度Q网络的无人车侦察路径规划[J]. 系统工程与电子技术, 2024,46(9):3070-3081.
XIA Y Q, HUANG Y Y, CHEN Q. Path planning for unmanned vehicle reconnaissance based on deep Q-network [J]. Systems Engineering and Electronics, 2024,46(9):3070-3081. (in Chinese)

[2] 李敏, 张森, 曾祥光, 等. 基于深度强化学习的四足机器人单腿越障轨迹规划[J]. 系统仿真学报, 2024(4):1-15.
LI M, ZHANG S, ZENG X G, et al. Trajectory planning of quadruped robot over obstacle with single leg based on deep reinforcement learning[J]. Journal of System Simulation, 2024:(4):1-15. (in Chinese)

[3] 舒健生, 周于翔, 郑晓龙, 等. 基于深度强化学习的无人机实时航迹规划[J]. 火力与指挥控制, 2023, 48(12): 133-141.
SHU J S, ZHOU Y X, ZHENG X L, et al. Deep Reinforcement Learning-based UAV Real-time Trajectory Planning[J]. Fire Control & Command Control, 2023, 48(12): 133-141 (in Chinese)

[4] 欧昌奎, 谢磊, 查天奇, 等. 基于深度强化学习和历史轨迹的船舶路径规划[J]. 中国航海, 2024, 47(1): 36-44.
OU C K, XIE L, ZHA T Q, et al. Ship path planning based on deep reinforcement learning and historical trajectories[J]. Navigation of China, 2024, 47(1): 36-44. (in Chinese)

[5] FAN Y S, SUN X J, WANG G F, et al. On fuzzy self-adaptive PID control for USV Course[C]. Proceeding of the 34th Chinese Control Conference. Hangzhou, China: IEEE, 2015:28-34.

[6] LIU S Y, LIU Y C, WANG N. Nonlinear disturbance observer-based backstepping finite-time sliding mode tracking control of underwater vehicles with system uncertainties and external disturbances[J]. Nonlinear Dynamics, 2017, 88(1): 465-476.

[7] FAN Y S, QIU B B, LEI L, et al. Global fixed-time trajectory tracking control of underactuated USV based on fixed-time extended state observer [J]. ISA transactions, 2022, 132: 267-277.

[8] ZHAO Y J, QI X, MA Y, et al. Path following optimization for an underactuated USV using smoothly-convergent deep reinforcement learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 99: 1-13.

[9] 李诗杰, 徐诚祺, 刘佳仑, 等. 船舶自抗扰无模型自适应航迹控制[J]. 中国舰船研究, 2024, 19(1): 280-289.
LI S J, XU C Q, LIU J L, et al. Tracking control of ships based on ADRC-MFAC[J]. Chinese Journal of Ship Research, 2024, 19(1): 280-289. (in Chinese)

[10] OGUZHAN D, KIRUBAKRAN V, FADI I, et al. Reinforcement learning approach to autonomous PID tuning[J]. Computers & Chemical Engineering, vol. 161: 1-17. 2022.

[11] 吴沁, 周顺仟, 王星联. 改进粒子群优化滚珠丝杠进给系统BP神经网络PID控制策略研究[J]. 西安交通大学学报, 2024,58(6):24-33.
WU Q, ZHOU S Q, WANG X L. Research on BP neural network PID control strategy for improving particle swarm optimization of ball screw feed system [J]. Journal of Xi′an Jiaotong University, 2024,58(6):24-33. (in Chinese)

[12] 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.

[13] Lee D, Lee J S, Yoon C S. Reinforcement learning-based adaptive PID controller for DPS [J]. Ocean Engineering, 2020, 216: 1-12.

(下转第146页)

ZHANG Y, MA J, CUI J W, et al. Rotation target detection algorithm for remote sensing image using attention mechanism [J]. Laser & Optoelectronics Progress, 2022,59(24): 192-200. (in Chinese)

[16] BO Z, LU Y Y. Improved YOLOv5 in remote sensing slender and rotating target detection [C] // 2022 IEEE International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). New York: IEEE, 2022: 918-923.

[17] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C] // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York: IEEE, 2023: 7464-7475.

[18] YANG X, YAN J. On the arbitrary-oriented object detection: Classification based approaches revisited[J]. International Journal of Computer Vision, 2022, 130(5): 1340-1365.

[19] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.

[20] YANG L, ZHANG R Y, LI L, et al. SimAM: a simple, parameter-free attention module for convolutional neural networks [C] // Proceedings of the 38 th International Conference on Machine Learning. New York: PMLR, 2021: 11863-11874.

[21] LIU Y, SHAO Z, HOFFMANN N. Global attention mechanism: Retain information to enhance channel-spatial interactions[DB/OL]. (2021-12-10)[2021-12-10]. https://arxiv.org/abs/2112.05561.

[22] XIA G S, BAI X, DING J, et al. DOTA: a large-scale dataset for object detection in aerial images[C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3974-3983.

[23] ZHU X, LV S, WANG X, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C] // Proceedings of the IEEE/CVF international conference on computer vision. New York: IEEE, 2021: 2778-2788.

(上接第136页)

[14] LAI P, LIU Y, ZHANG W, et al. Intelligent controller for unmanned surface vehicles by deep reinforcement learning[J]. Physics of Fluids, 2023, 35(3):1-11.

[15] 张梦杰, 陈姚节, 邓江. 改进TD3算法在电机PID控制器中的应用[J]. 计算机系统应用, 2024,33(5):262-270.

ZHANG M J, CHEN Y J, DENG J. Application of improved TD3 algorithm in motor PID controllers[J]. Computer Systems & Applications, 2024,33(5):262-270. (in Chinese)

[16] 吕金旭, 葛万成. 智能制造领域中深度强化学习的应用综述[J]. 信息与电脑(理论版), 2023, 35(5): 186-193.

LV J X, GE W C. A review of the application of deep reinforcement learning in the field of intelligent manufacturing[J]. Information & Computer, 2023, 35(5): 186-193. (in Chinese)

[17] 曹宏业, 刘潇, 董绍康, 等. 面向强化学习的可解释性研究综述[J]. 计算机学报, 2024, 47(8): 1853-1882.

CAO H Y, LIU X, DONG S K, et al. A survey of interpretability research methods for reinforcement learning[J]. Chinese Journal of Computers, 2024,47(8):1853-1882. (in Chinese)

[18] 刘潇, 刘书洋, 庄韫恺, 等. 强化学习可解释性基础问题探索和方法综述[J]. 软件学报, 2023, 34(5): 2300-2316.

LIU X, LIU S Y, ZHUANG Y K, et al. Explainable reinforcement learning: basic problems exploration and method survey[J]. Journal of Software, 2023, 34(5): 2300-2316. (in Chinese)

[19] 陈奇, 赵炳春, 尚明健, 等. 变频阻垢除垢装置增量式数字PID恒流控制[J]. 控制工程, 2015, 22(增刊1): 89-93.

CHEN Q, ZHAO B C, SHANG J M, et al. The constant current control of anti-scaling descaling device based on incremental digital PID controller [J]. Control Engineering of China, 2015, 22(Suppl.1): 89-93. (in Chinese)

[20] FOSSEN T. Handbook of marine craft hydrodynamics and motion control[M]. Sussex: John Wiley & Sons Ltd, 2011.

[21] 姚培源, 魏潇龙, 俞利新, 等. 基于Q-Learning算法的无人机空战机动决策研究[J]. 电光与控制, 2023, 30(5): 16-22.

YAO P Y, WEI X L, YU L X, et al. Research on UAV air combat maneuver decision based on Q-Learning algorithm[J]. Electronics Optics & Control, 2023, 30(5): 16-22. (in Chinese)

[22] 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. Stockholm, Swedn: PMLR, 2018:2976-2989.

[23] CUI Z, GUAN W, ZHANG X. Collision avoidance decision-making strategy for multiple USVs based on Deep Reinforcement Learning algorithm [J]. Ocean Engineering, 2024, 308: 118323.