专题：水下通信网络技术

# 基于分布式强化学习的AUV水下三维洋流目标跟踪控制算法

赵剑楠[1]，覃琪琪[1]，李云[2,3,4,5]，苏毅珊[6]

（1. 广西大学电气工程学院，广西 南宁 530003；
2. 广西民族大学物理与电子信息学院，广西 南宁 530006；
3. 多模态信息智能感知处理与应用广西高校工程研究中心，广西 南宁 530006；
4. 广西智语人形机器人重点实验室，广西 南宁 530006；
5. 广西智能视觉协作机器人工程研究中心，广西 南宁 530006；
6. 天津大学电气自动化与信息工程学院，天津 300072）

**摘　要：**针对水下自主航行器（autonomous underwater vehicle，AUV）在复杂三维洋流环境中目标跟踪的高维、动态干扰和稀疏回报挑战，提出了一种基于分布式强化学习的水下自主航行器水下三维洋流目标跟踪控制算法。首先，引入真实三维洋流数据，设计动态目标跟踪场景，以准确描述AUV的运动过程；其次，结合对抗深度强化学习网络（dueling deep Q-network，Dueling DQN）结构与分位数回归方法，针对三维洋流环境可能导致Q值过高估计的问题，构建分布强化学习框架，以量化Q值的不确定性，提升策略对动态干扰的适应能力；最后，引入优先经验回放机制，设计约束条件下的奖励函数，优化数据采样策略，加速模型收敛。实验结果表明，相较于深度Q网络（deep Q-network，DQN）、双深度Q网络（double deep Q-network，DDQN）和Dueling DQN，所提算法在复杂洋流环境中表现更优，在收敛速度、目标跟踪精度和鲁棒性方面均取得了显著的进展。

**关键词：**强化学习；水下自主航行器；目标跟踪；分位数回归

**中图分类号：**TN911.7；TP18；U664.82

**文献标志码：**A

**doi:** 10.11959/j.issn.1000−0801.2025209

# Distributed reinforcement learning-based AUV 3D underwater current target tracking control algorithm

ZHAO Jiannan[1], QIN Qiqi[1], LI Yun[2,3,4,5], SU Yishan[6]

1. School of Electrical Engineering, Guangxi University, Nanning 530003, China
2. School of Physics and Electronic Information, Guangxi Minzu University, Nanning 530006, China
3. Engineering Research Center of Multi-Modal Information Intelligent Sensing,

**收稿日期：**2025−03−27；**修回日期：**2025−06−10

**通信作者：**李云，20240165@gxmzu.edu.cn

**基金项目：**国家自然科学基金资助项目（No.62361002，No.62206065）；桂林市重点研发计划项目（No. 20221B074250）

**Foundation Items:** The National Natural Science Foundation of China (No.62361002, No.62206065), The Guilin Key Research and Development Program (No. 20221B074250)

Processing and Application, Guangxi Universities, Nanning 530006, China
4. Guangxi Key Laboratory of ZHIYU Humanoid Robots, Nanning 530006, China
5. Guangxi Engineering Research Center of Intelligent Vision and Collaborative Robotics, Nanning 530006, China
6. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

**Abstract:** To address the challenges of high dimensionality, dynamic disturbances, and sparse rewards in autonomous underwater vehicle (AUV) target tracking within complex three-dimensional ocean current environments, a distributed reinforcement learning-based control algorithm was proposed. Firstly, realistic 3D ocean current data was incorporated to design dynamic target tracking scenarios, accurately modeling the AUV's motion dynamics. Secondly, a distributed reinforcement learning framework was constructed by integrating the dueling deep Q-network (Dueling DQN) architecture with quantile regression methods. This framework quantified the uncertainty of Q-values to mitigate overestimation risks in 3D turbulent environments, enhancing the policy's adaptability to dynamic disturbances. Finally, the prioritized experience replay mechanism was introduced, along with a constraint-optimized reward function, to optimize data sampling strategies and accelerate model convergence. Experimental results demonstrate superior performance of the proposed algorithm compared with the baseline methods, such as deep Q-network (DQN), double deep Q-network (DDQN), and Dueling DQN, in complex current conditions, achieving significant improvements in convergence speed, target tracking accuracy, and robustness.

**Key words:** reinforcement learning, AUV, target tracking, quantile regression

## 0　引言

水下自主航行器（autonomous underwater vehicles，AUV）作为一种高效、智能且环保的水下勘探工具，在海洋科学研究、资源开发、海洋设施监测等领域中发挥着越来越重要的作用[1]。AUV 目标跟踪任务要求智能体在动态干扰条件下，持续感知并预测目标的运动轨迹，同时实时调整自身路径，以实现高精度跟踪控制[2-3]。这不仅需要对复杂环境变化保持高度的适应性，还要求控制策略具备较强的决策灵活性和长时域规划能力。在水下环境中，如洋流、障碍物和其他水文变化等因素，都会对 AUV 的运动产生影响，使目标跟踪任务变得更加复杂[4]。

近年来，针对 AUV 目标跟踪控制问题，研究者们提出了基于模糊控制[5]、自适应控制[6]、模型预测控制[7]、神经网络控制[8]等多种方法。现有控制方法的优缺点见表 1。尽管这些方法在小规模、确定性环境中表现良好，但在复杂的动态环境中，增加的系统复杂性导致性能下降[9]。特别是在洋流动态变化、水下障碍多样且目标运动模式不确定的情况下，传统的基于模型或规则的方法难以实时、有效地完成高精度跟踪任务。

强化学习（reinforcement learning，RL）作为一种基于数据驱动的方法，在处理动态环境中的复杂控制问题时展现出巨大潜力[10]。与传统控

**表 1　现有控制方法的优缺点**

| 方法 | 优点 | 缺点 |
| --- | --- | --- |
| 模糊控制 | 无需精确系统模型，适用于不确定性环境初步建模 | 规则库依赖人工设计，难以动态适应目标运动变化，容易陷入局部最优 |
| 自适应控制 | 能实时调整控制参数，适应部分系统变化 | 计算复杂度高，收敛性和稳定性受限，依赖系统初始估计 |
| 模型预测控制 | 适合处理多变量约束优化问题，短期轨迹预测效果良好 | 实时性差，计算开销大，对模型精度和外部扰动敏感 |
| 基于神经网络的控制 | 具有较强的非线性建模与自适应能力 | 训练数据需求大，控制策略可解释性差，易出现过拟合问题 |

制方法不同，强化学习通过智能体与环境的交互，自主学习最优策略，并能够有效地适应环境的变化，已应用于自动驾驶[11]、机器人操作控制[12]。在强化学习方法中，深度Q网络（deep Q-network，DQN）是一种经典的值迭代算法，它通过深度神经网络逼近Q值，并利用经验回放和目标网络等技术提高训练的稳定性。文献[13]研究了网络诱导时延环境中的智能体目标跟踪控制问题，基于DQN优化控制策略，在长期能耗与跟踪误差之间取得了更优的平衡。为最大化飞行时间内的上行吞吐量，文献[14]提出了一种基于Safe-DQN的智能体轨迹设计算法，使智能体在合理策略集中选择最优行动。然而，DQN在实际应用中仍然存在Q值过高估计的问题，可能导致策略的不稳定性和次优解。为此，双深度Q网络（double deep Q-network，DDQN）通过双网络架构有效地缓解了Q值过高估计的问题，并通过双输入层结构提升了对高维环境信息的适应能力。文献[15]研究了智能体的平滑跟踪问题，提出了基于DDQN的路径平滑和跟踪控制方法，为控制方法提供了新思路。文献[16]提出了一种基于DDQN的深度强化学习方法，增强了AUV在复杂洋流环境中的决策能力。文献[17]提出了基于DDQN的分层定位策略，在不依赖固定锚节点的情况下提高了定位精度，优化子策略分配，从而降低了从属智能体的定位误差。

在水下环境中，强化学习算法的应用面临着诸多挑战。AUV的运动稳定性受到多方面因素的影响，包括不确定洋流的扰动、目标运动的不规则性以及感知信息的不完备性[18]。然而，目前大量关于洋流模型的研究主要依赖数学函数拟合的方法[19-21]。这些方法在一定程度上能够模拟洋流环境，但是单纯依赖数学模型难以全面地刻画复杂性。因此，一些研究者尝试将数据驱动方法与物理建模相结合，以更准确地描述洋流环境。文献[22]基于温度场数据，以优化采样轨迹并考虑二维洋流影响。文献[23]将未知的二维流场表示为集成预报生成的“基流场”的线性组合。然而，实际水下环境中的洋流具有显著的三维特性，仅依赖二维流场描述难以满足AUV的精确控制需求，需要向三维洋流建模扩展。文献[24]考虑到三维洋流的垂直速度通常远小于经度和纬度方向的速度，为了简化计算，在建模时忽略了垂直速度的影响。文献[25]基于Lamb涡数值方程模拟了三维洋流环境，但该方法依赖参数化涡流结构。可见，现有方法在处理三维洋流环境的复杂性方面仍然存在局限性。

强化学习方法在AUV水下目标跟踪任务中的应用，依赖于高效的Q值估计策略。尽管DDQN通过双网络结构有效地缓解了Q值过高估计的问题，但在复杂的三维洋流环境中，仅基于动作–状态对的期望Q值进行策略更新，无法充分捕捉环境的不确定性带来的回报变化，从而引起价值估计误差，影响策略的稳定性与泛化能力。针对这一问题，分布式强化学习被提出来建模状态–动作对所对应的完整回报分布，而非单一均值，从而提高策略对环境动态变化的适应能力。文献[26]指出，基于回报分布的策略学习可以有效地减缓过高估计和策略振荡等问题。文献[27]通过分布式建模进一步提高了在动态环境中的样本效率和鲁棒性。

然而，单纯依赖分布式回报建模，在状态重要性判别能力不足、动作优势不明显的情况下，会出现训练收敛缓慢和模型学习冗余的问题[28]。为提升分布建模的效率与判别能力，本文进一步结合了Dueling DQN结构，通过分别估计状态值函数和动作优势函数，并在融合后对动作价值进行完整分布建模，从而在复杂环境中更高效地提升策略评估的判别能力与稳定性。基于此，本文提出了一种基于分布式强化学习的AUV目标跟踪方法（value-distributional dueling double deep Q-network with prioritized experience replay，

VD3QN-PER），并结合真实海洋洋流数据，研究单个AUV在三维洋流干扰下的稳定性及跟踪精度。本文的创新点和贡献如下。

（1）提出了一种基于分布式强化学习的AUV水下三维洋流目标跟踪控制算法（VD3QN-PER），以优化AUV的跟踪策略。不同于DQN、DDQN直接估计Q值期望，VD3QN-PER通过分位数回归建模Q值分布，细化不确定性估计。在Dueling结构基础上，解耦状态和动作的价值评估，同时引入优先经验回放机制，有效地提高了决策稳定性和收敛效率。

（2）引入真实三维洋流数据，并基于分布时间差分（temporal difference，TD）误差动态地调整样本采样概率，不同于仅基于简化洋流模型或忽略垂向流动的传统方法的处理方式。通过真实、复杂的环境数据驱动训练，VD3QN-PER能够更准确地描述AUV在复杂流场中的运动特性，提升了模型的泛化能力和实际应用的可行性。

（3）实验结果表明，在动态目标跟踪场景中，与DQN、DDQN、Dueling DQN等传统方法相比，VD3QN-PER在收敛速度、目标跟踪精度、环境鲁棒性等关键指标上均得到了更显著的提升，验证了所提方法的有效性与优越性。

## 1　系统模型

为准确描述AUV在复杂水下环境中的运动状态及其与环境的交互，本节建立了AUV的运动学模型，并描述了三维洋流环境的建模方法以及AUV在执行目标跟踪任务时所需满足的约束条件。

### 1.1　AUV运动学模型

AUV运动受到自身推进系统、操纵面控制力矩以及外部环境扰动等因素的影响。本文采用六自由度运动学模型来描述AUV在三维空间中的运动状态，该模型在大地坐标系和AUV体坐标系下分别定义了AUV的位姿与运动参数。六自由度的AUV运动学模型如图1所示，大地坐标系$O_E x_E y_E z_E$和体坐标系$O_B x_B y_B z_B$是AUV运动描述的两个主要参考系。大地坐标系用于描述AUV在全局环境中的位置和姿态$\boldsymbol{H}=[\xi, \eta, \zeta, \phi, \theta, \psi]^T$，体坐标系以AUV自身为参考，用于描述AUV的运动状态，包括线性速度和角速度$\boldsymbol{V}=[u, v, w, p, q, r]^T$。

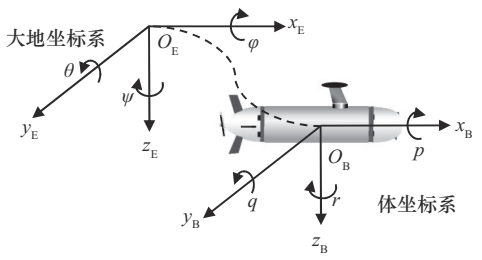



图1　六自由度的AUV运动学模型

为了更清晰地表示AUV的运动参数及其物理含义，AUV的运动参数及含义见表2。将AUV获取到体坐标系上的速度和力转化为地球坐标系上的值为[29]：

表2　AUV的运动参数及含义

| 符号 | 运动参数 | 含义 | 单位 |
|---|---|---|---|
| $\xi, \eta, \zeta$ | 纵向位置、横向位置、垂直位置 | AUV在三维空间中的位置坐标 | m |
| $\phi, \theta, \psi$ | 滚动角、俯仰角、偏航角 | AUV绕三轴的姿态角度 | rad |
| $u, v, w$ | 纵向速度、横向速度、垂直速度 | AUV沿三轴方向的线速度 | m/s |
| $p, q, r$ | 滚动速率、俯仰速率、偏航速率 | 姿态角在三轴方向的变化率 | rad/s |
| $X, Y, Z$ | 纵向力、横向力、垂直方向力 | AUV在三轴方向上的合力 | N |
| $K, M, N$ | 滚动力矩、俯仰力矩、偏航力矩 | AUV绕三轴的合力矩 | N·m |

$$\begin{cases}\dot{\xi}=u\cos\psi\cos\theta+v(\cos\psi\sin\theta\sin\phi-\sin\psi\cos\phi)+w(\cos\psi\sin\theta\cos\phi+\sin\phi\sin\psi),\\ \dot{\eta}=u\sin\psi\cos\theta+v(\sin\psi\sin\theta\sin\phi+\cos\psi\cos\phi)+w(\sin\psi\sin\theta\cos\phi-\sin\phi\cos\psi),\\ \dot{\zeta}=-u\cos\theta+v\cos\theta\sin\phi+w\cos\theta\cos\phi,\\ \dot{\varphi}=p+q\tan\theta\sin\phi+r\tan\theta\cos\phi,\\ \dot{\theta}=q\cos\phi-r\sin\phi,\\ \dot{\psi}=(q\sin\phi-r\cos\phi)/\cos\theta\end{cases}\tag{1}$$

### 1.2 洋流模型

AUV的运动不仅受自身控制输入的影响，还受到复杂洋流环境的扰动。传统方法大多基于简化的数学模型或正弦函数模型进行洋流模拟，计算量较低，但缺乏真实海洋数据的场景。即使有真实的洋流数据，大多研究集中考虑了二维洋流的影响，而忽略了三维空间中垂直方向上的洋流流动变化，垂直分量的洋流参数对航行器在下潜上仰过程中影响较大。

针对上述问题，本文构建了一个基于实际海洋数据的三维洋流模型，以真实地模拟水下环境中的洋流情况。数据来源于国家海洋科学数据中心和哥白尼海事服务中心[30]，涵盖了东经118.5°至120.0°、北纬20.75°至22.25°的海域，研究区域海面以下至2 000 m深的三维水下环境。三维洋流速度场分布如图2所示，展示了研究区域内三维洋流速度场的分布特征。$z$ = 32 m时平面洋流矢量如图3所示，呈现了该平面内的洋流矢量图，箭头表示洋流方向。

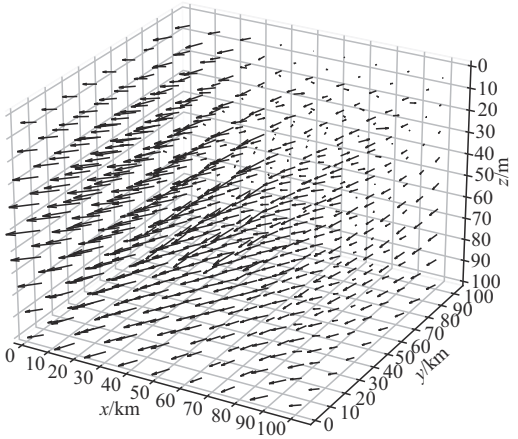



图2 三维洋流速度场分布

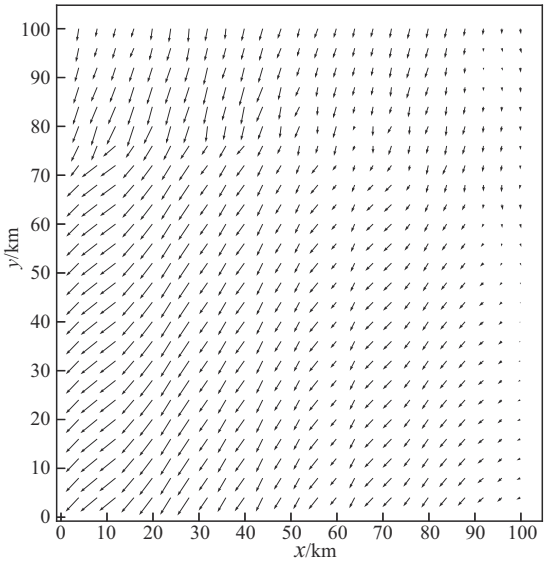



图3 $z$ = 32 m时平面洋流矢量

### 1.3 环境交互与约束条件

AUV在水下环境中的运动受到复杂洋流、环境阻力以及自身推进力等因素的综合影响，动力学特性通常采用六自由度模型进行描述。本文中，AUV的状态变量包括位置、速度等信息。洋流对AUV的运动会产生显著的影响，AUV的实际速度由自身推进速度与环境洋流速度的叠加决定。同时，AUV的运动须满足最大速度限制、深度范围约束。其中，AUV推进系统的功率限制决定了最大航速，水下作业的环境条件规定了可运行的深度范围。

## 2 基于强化学习的AUV目标跟踪

AUV作为智能体，通过与环境的交互不断地优化决策策略，以提升目标跟踪的稳定性和效率。针对水下动态目标跟踪任务，本文构建了合

理的状态空间和动作空间，并设计了优化的奖励函数，以引导AUV在复杂洋流环境中实现高效跟踪。

在复杂的水下环境中，AUV的目标跟踪任务涉及高维状态空间、非平稳动态干扰及稀疏奖励问题，传统路径规划方法难以在不确定性洋流条件下实现自适应决策。因此，本文采用强化学习框架，以数据驱动的方式优化AUV的决策策略，并在无监督环境中学习高效的目标跟踪方案。在强化学习框架下，AUV的运动控制问题可建模为一个马尔可夫决策过程（Markov decision process, MDP），其核心要素包括状态空间$S$、动作空间$A$和奖励函数$R$。因此，在三维网格水下洋流模型的基础上，设计的RL环境主要包括两个基本模块：状态转移和奖励函数。

### 2.1　状态转移

本文构建了一个包含环境信息与AUV自身状态信息的连续状态向量，以全面地刻画AUV的运动及其周围环境。其中，包括障碍物位置的实时检测数据、动态目标位置信息以及影响航行决策的洋流速度等。用数学符号表示为：

$$s_t = (x, y, z, u, v, w, u_{\text{cur}}, v_{\text{cur}}, w_{\text{cur}}, \Delta P_x, \Delta P_y, \Delta P_z, \text{obs}_x, \text{obs}_y, \text{obs}_z) \tag{2}$$

其中，$(x, y, z, u, v, w)$表示AUV在三维空间中的位置及速度，$(u_{\text{cur}}, v_{\text{cur}}, w_{\text{cur}})$表示当前AUV所在位置的洋流速度，$(\Delta P_x, \Delta P_y, \Delta P_z)$为与目标的相对位置。$(\text{obs}_x, \text{obs}_y, \text{obs}_z)$为AUV预警范围内检测到障碍物的信息。

在动作空间构建上，本文采用六维离散动作集，分别对应AUV在水下环境中的上、下、前、后、左、右6个方向的移动，以有限且明确的指令集有效地引导AUV的运动决策，灵活地应对复杂环境的挑战。6个动作分别对应AUV体坐标系中的纵向、横向和垂直位置变化自由度，通过对这3个方向上的正负方向移动进行离散建模，实现对AUV位置的灵活调整。通过离散化动作空间，降低了强化学习算法的计算复杂度，并提高了决策的稳定性。考虑到AUV需根据环境状态选择动作以调整自身运动，本文采用$\varepsilon$-greedy策略，在训练初期以较高的概率探索不同动作以增强环境探索能力，随着训练的推进，逐步降低随机探索概率，从而基于经验选择最优动作，不断优化决策策略，使AUV能够在复杂动态的水下环境中实现路径自适应调整，提升目标跟踪的稳定性与效率。

### 2.2　奖励函数

本文在奖励函数设计上，结合洋流环境的动态特性，采用融合离散奖励与连续奖励的混合机制，从而激励AUV高效跟踪目标，增强对洋流干扰的适应能力，并提升任务执行效率。离散奖励用于回合终止时的任务评估，当AUV成功跟踪目标点时，将获得正向奖励，当超出环境边界、丢失目标、与障碍物发生碰撞以及在最大时间步内未能完成任务时，则通过正负奖励的引导，纠正AUV的无效或危险行为。为进一步优化跟踪过程中的决策表现，连续奖励在每一步提供即时反馈，主要包括目标跟踪奖励、洋流利用奖励、运动能耗奖励和避障奖励，从而促使AUV在动态环境中持续调整和优化行为。目标跟踪奖励用于衡量AUV与目标点之间的接近程度，数学表达式为：

$$r_{\text{dis}} = \begin{cases} +1, & \text{Dis(AUV, target)} \leqslant \mu \\ -\text{Dis(AUV, target)}/\mu, & \text{Dis(AUV, target)} > \mu \end{cases} \tag{3}$$

其中，$\mu$表示跟踪阈值。$r_{\text{dis}}$用于激励AUV每个时间步上跟踪上动态目标点。洋流利用奖励用于衡量AUV运动方向与洋流方向的一致性，定义为：

$$r_{\text{cur}} = v_{\text{AUV}} \cdot v_{\text{cur}} \tag{4}$$

其中，$r_{\text{cur}}$用于鼓励AUV顺流运动以减少能耗，并在逆流运动时优化其航向。为进一步控制能耗，在奖励函数中引入动作成本项$r_{\text{cost}}$，该项与一轮跟踪过程中AUV执行的时间步数成正比，从而鼓励AUV以更少的步骤完成跟踪任务。因此将奖励函数表示为：

$$r = l_1 r_{\text{dis}} + l_2 r_{\text{cur}} + l_3 r_{\text{cost}} + \boldsymbol{L} r_{\text{done}} \quad (5)$$

其中，$l_i$表示权重因子，$\boldsymbol{L}$表示一个维度为5的终局奖励向量，每个元素的值为0或1，表示对应的回合终结条件是否被满足，1代表条件满足，0代表条件未满足，从而影响最终的回合奖励。

### 2.3 Dueling deep Q network网络

强化学习的核心任务是学习一个最优策略，以最大化智能体在环境中的长期回报。传统Q-learning由于依赖表格存储Q值，难以扩展到高维环境。因此，引入深度神经网络来近似Q值函数，从而提出了深度Q网络。DDQN通过引入两个网络来分离动作选择和动作值评估，从而避免了Q值的过高估计。在DDQN中，目标Q值的计算方式为：

$$Q^{-} = r + \gamma Q(s',\ \arg\max_{a'} Q(s',\ a';\ \theta^{-});\ \theta) \quad (6)$$

其中，$Q^{-}$为目标Q值，表示在当前状态下采取动作的期望回报。$r$是从环境中获得的即时奖励，$\gamma$是折扣因子，决定了未来奖励的重要性。$s'$是下一状态，$a'$是动作。$\theta$和$\theta^{-}$分别表示当前网络和目标网络的参数。DDQN是通过目标网络计算出的最大Q值来更新当前网络的Q值估计，从而避免了Q值的过高估计。相应的损失函数为：

$$\text{Loss}_{\text{DQN}}(\theta) = \text{E}_{(s,a,r,s')\sim D}\left[\left(Q_\theta(s,a)\right) - \left(r + \gamma Q\left(s',\ \arg\max_{a'} Q(s',\ a';\ \theta);\ \theta^{-}\right)\right)^2\right] \quad (7)$$

其中，$D$是经验池，包含了智能体与环境交互的历史样本$(s,a,r,s')$。Dueling DQN进一步优化了Q值的计算方式，将Q值拆分为状态价值函数$V(s)$和动作优势函数$A(s,a)$。这样，Q值函数的计算可以表示为：

$$\begin{aligned} &Q(s,a) = V(s) + A(s,a) - A(s,a).\text{mean}(\text{dim} = 1), \\ &A(s,a).\text{mean} = \frac{1}{|A|}\sum\nolimits_{a'} A(s,a') \end{aligned} \quad (8)$$

其中，$|A|$表示动作空间维度。通过将状态和动作的价值分离，Dueling DQN能够更有效地评估动作的相对重要性，特别是在状态价值相对固定，而动作选择差异较大时，能更精确地计算Q值。

## 3 基于VD3QN-PER算法的AUV目标跟踪

为了提升Dueling DQN在复杂环境中的表现，本文提出了一种基于分布式强化学习的AUV水下目标跟踪算法VD3QN-PER。该算法结合了Dueling DQN的状态–动作解耦思想和分位数回归模型，通过引入分位数回归，能够在分布式Q值框架下建模回报的概率分布，从而捕捉环境中的不确定性，提高决策的稳定性。此外，结合优先经验回放机制，使经验样本的采样更高效，提升训练收敛速度。

### 3.1 分位数回归建模

在传统的Q学习方法中，Q学习仅预测状态–动作对的Q值，无法充分地刻画回报的不确定性。而分布强化学习通过学习Q值的分布，智能体能够基于回报的不确定性进行更稳健的决策[31]。定义目标分布为通过离散化后的$N$个分位点，设这些分位数为：

$$\tau_i = \frac{i - 0.5}{N},\ i = 1, 2, \cdots, N \quad (9)$$

其中，$\tau_i$表示第$i$个分位数。基于此，状态–动作对$(s,\ a)$的回报分布函数$Z(s,\ a)$可近似为：

$$Z(s,\ a) \approx \sum\nolimits_{i=1}^{N} Q_{\tau_i}(s,\ a) \cdot \delta(\tau_i) \quad (10)$$

其中，$Q_{\tau_i}(s,\ a)$为每个分位数下的Q值，$\delta(\tau_i)$为Dirac函数，表示在该分位数点上的概率质量。回报分布的更新通过分布贝尔曼方程来实现，如下：

$$Q'_{\tau_i}(s,\ a) = r + \gamma Q_{\tau_i}\left(s',\ \arg\max_{a'} \sum\nolimits_{j=1}^{N} Q_{\tau_j}(s',\ a')\right) \quad (11)$$

为了训练所提出的网络，本文使用了分位数

Huber 损失，其目标是最小化预测分位数与目标分位数之间的 Wasserstein 距离。该损失函数不仅结合了 Huber 损失的鲁棒性（对离群点具有较强的抵抗力），还利用了分位数回归的优点，从而能够处理 Q 值的分布而非单一的期望值。损失函数的形式为[31]：

$$\mathrm{Loss}_{\mathrm{QR}}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\rho_{\tau_i}^{k}\left(Q_{\tau_i}-Q'_{\tau_j}\right) \tag{12}$$

其中，$\rho_{\tau_i}^{k}(\cdot)$是分位数加权函数，函数定义式为[31]：

$$\rho_{\tau_i}^{k}(m)=\left|\tau-\boldsymbol{I}(m<0)\right|\cdot\mathrm{Loss}_k(m)$$

$$\mathrm{Loss}_k(m)=\begin{cases}\dfrac{1}{2}m^2, & |m|\leqslant k\\ k\left|m-\dfrac{1}{2}k\right|, & \text{其他情况}\end{cases} \tag{13}$$

其中，$\boldsymbol{I}$是指示函数，表示条件$m$<0是否成立。若条件成立，则指示函数为1，否则为0。$\mathrm{Loss}_k(m)$是平滑的 Huber 损失。通过最小化该损失函数，可确保预测的分位数更稳定，从而提高 AUV 在复杂洋流环境中的决策精度。

### 3.2　Dueling 架构的状态–动作解耦

Dueling DQN 通过状态价值与动作优势的分离建模，提高了高维状态空间下的学习效率。然而，传统 Dueling 架构仅针对标量 Q 值进行建模，无法有效地结合分布强化学习框架。因此，本文在 Dueling 架构的基础上，引入分位数回归，使状态价值和动作优势在不同分位数层面进行建模。分位数形式的 Q 值计算为：

$$Q_{\tau_i}(s,\,a)=V_{\tau_i}(s,\,\tau_i)+\left(A_{\tau_i}(s,\,a,\,\tau_i)-\frac{1}{|A|}\sum_{a'}A_{\tau_i}(s,\,a',\,\tau_i)\right) \tag{14}$$

其中，$V_{\tau_i}(s,\tau_i)=[V_{\tau_1}(s,\tau_1),V_{\tau_2}(s,\tau_2),\cdots,V_{\tau_N}(s,\tau_N)]^{\mathrm{T}}$为状态价值分量，动作优势分量为$A_{\tau_i}(s,\,a,\,\tau_i)=[A_{\tau_1}(s,\,a,\,\tau_1),A_{\tau_2}(s,\,a,\,\tau_2),\cdots,A_{\tau_N}(s,\,a,\,\tau_N)]^{\mathrm{T}}$。此架构能减少动作价值估计的方差，提高策略学习的稳定性，同时结合分位数信息，更精准地建模回报分布，使 AUV 在洋流环境中的行为决策更鲁棒。

### 3.3　优先经验回放（PER）与分布 TD 误差

在传统经验回放机制中，样本的采样通常采用均匀分布，无法突出重要经验的作用。而优先经验回放（PER）机制通过引入经验优先级，使学习过程更高效。本文基于分布式强化学习的特点，提出基于分布 TD 误差的优先经验回放方法，以更准确地衡量经验的重要性。目标分位数$q'_{\tau_i}(s,\,a)$和当前 Q 值 $q_{\tau_i}(s,\,a)$的差异为 TD 误差$\delta_{\tau_i}(s,a)$，定义为：

$$\delta_{\tau_i}(s,\,a)=r+\gamma q'_{\tau_i}\left(s',\,a^{*};\theta^{-}\right)-q_{\tau_i}(s,\,a;\theta) \tag{15}$$

每个经验$n$的优先级可以通过分位数 TD 误差的$L_1$范数决定，计算式为：

$$p_n=\left(\frac{1}{N}\sum_{k=1}^{N}\left|\delta_{\tau_i}^{(i)}\right|+\xi\right)^{\alpha} \tag{16}$$

其中，$\xi$是一个小常数，用于避免零误差。为校正 PER 带来的偏差，采用重要性采样权重进行修正，计算式为：

$$w_n=\left(|B|\cdot p_n\right)^{-\beta}\cdot\max_{j}\ w_j \tag{17}$$

其中，$\beta$控制重要性采样权重的修正力度。$\beta$值越大对非均匀采样的偏差校正越强。优化目标函数结合 PER 与分位数损失，可表示为：

$$\mathrm{Loss}=\frac{1}{N^2}\sum_{n=1}^{N}w_n\cdot\sum_{k=1}^{N}\rho_{\tau_k}^{k}\left(\delta_{\tau_k}^{(n)}\right) \tag{18}$$

基于分布 TD 误差的优先经验回放方法有效地增强了训练过程中对关键经验的利用率，提高了 AUV 策略的收敛速度与性能稳定性。

### 3.4　算法整体流程

VD3QN-PER 算法以 AUV 的当前状态（位置、速度、洋流信息、障碍物信息）作为输入，采用分位数 Dueling DQN 进行动作决策，并结合分布 TD 误差的优先经验回放机制进行高效训练。AUV 的前进方向由算法控制，VD3QN-PER 算法的网络框架如图 4 所示，Dueling DQN 与值分布框架结构如图 5 所示，VD3QN-PER 算法如算法 1 所示。

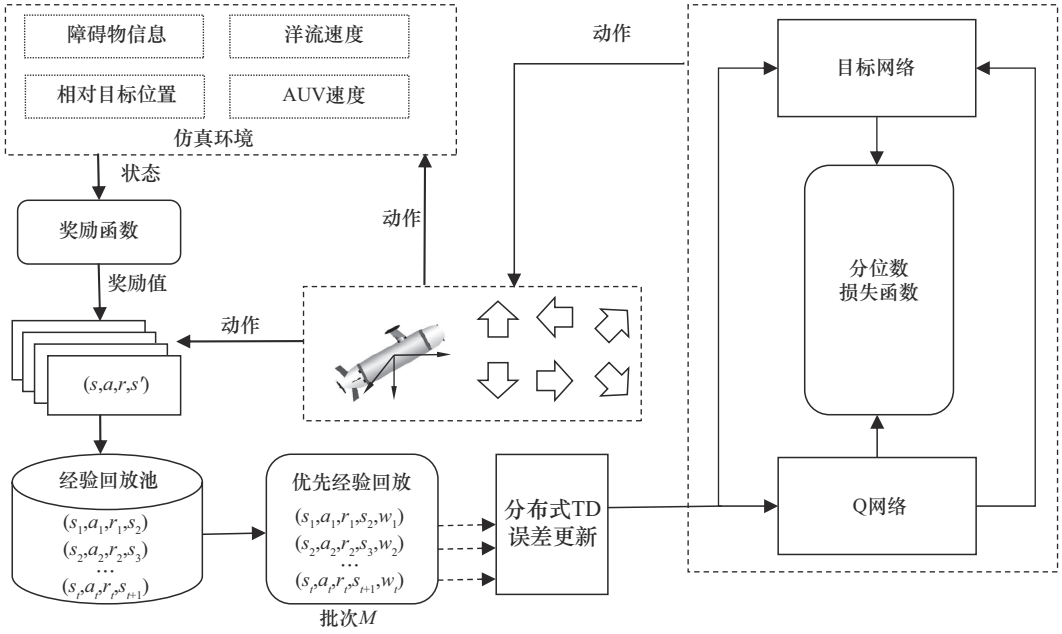



图 4　VD3QN-PER 算法的网络框架

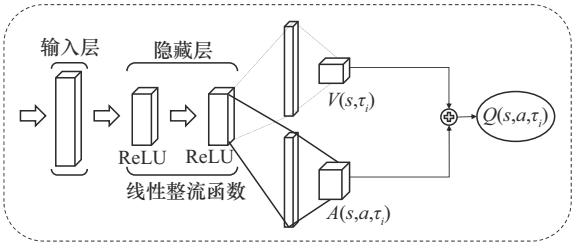



图 5　Dueling DQN 与值分布框架结构

**算法 1**　VD3QN-PER 算法

**初始化**　当前网络 $Q(s, a; \theta)$ 和目标网络 $Q'(s, a; \theta^{-}) \leftarrow Q(s, a; \theta)$，优先级回放缓冲区 $B$

**输入**　分位数数量 $N$，PER 参数 $(\alpha, \beta)$，折扣因子 $\gamma$，优先级偏移 $\xi$，更新间隔 $K$

**for** episode = 1 to MaxEpisodes **do**

　重置环境，获得初始状态 $s$；

　**for** $t$ = 1 to MaxSteps **do**

　　根据 $\varepsilon$-greedy 策略选择动作 $a$；

　　执行动作 $a$，获得即时奖励 $r$ 并观察下一个状态 $s'$；

　　计算分位数 TD 误差；

　　存储环境数据经验 $(s, a, r, s', d)$ 到 $B$，设置优先级，式（16）；

　　经验回放与网络更新；

　　**if** $t$ mod $K == 0$ **then**

　　依据优先级从 $B$ 中采样 batch_size 数据元 $\{(s_i, a_i, r_i, s_i', d_i)\}$；

　　计算重要性权重，式（17）；

　　计算分位数 Huber 损失，式（18）；

　　更新 $\theta$ 以最小化损失 Loss，更新 $\theta^{-} \leftarrow \theta$；

　　**end if**

　更新状态 $s^{-} \leftarrow s$；

　**if** $s$ 为终止状态 **then**

　　终止循环；

　**end if**

　**end for**

**end for**

## 4　仿真实验与结果分析

### 4.1　实验参数设置

为了验证所提的 VD3QN-PER 算法在单个动态目标跟踪任务中的有效性，本文构建了一个具有动态洋流干扰的仿真实验环境，并在复杂洋流

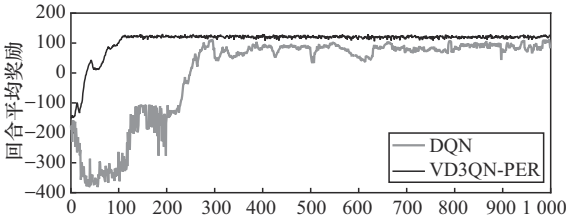
场景中进行算法性能评估。AUV 在每个时间步内按照当前决策动作方向执行 2 m/s 速度运动。水平洋流速度范围为–1.0 ~ 1.0 m/s，垂直洋流速度为–0.001 ~ 0.001 m/s，动态目标是在洋流基础上随流漂移，并叠加幅值不超过±0.2 m/s 的小范围随机扰动，实现连续位置变化。本实验在高性能服务器上进行，配置如下：Intel Core i9-11900K 处理器、64 GB 内存，操作系统为 64 位，实验程序基于 Python 3.8 实现。仿真参数设置见表 3。实验采用回合平均奖励、回合平均步数以及平均跟踪误差作为核心评价指标，并选取 DQN、DDQN 和 Dueling DQN 作为对比基准，以分析 VD3QN-PER 在强化学习框架下的性能优势。

**表 3　仿真参数设置**

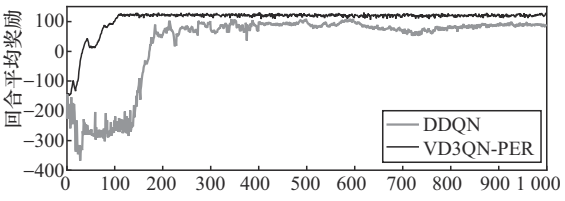
| 参数 | 参数值 |
|---|---|
| 学习率 | 10 000 |
| 折扣因子 | 0.99 |
| 经验池大小/条 | 1 000 000 |
| 目标网络更新频率/步 | 1 000 |
| 训练回合数/回合 | 1 000 |
| 每回合最大步数/步 | 300 |

## 4.2　实验结果与分析

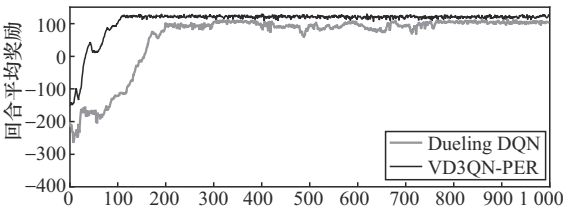
各算法在仿真实验中的回合平均奖励曲线对比如图 6 所示。从结果来看，VD3QN-PER 在训练初期即可展现出更快的收敛速度，并在回合奖励上优于 DQN、DDQN 和 Dueling DQN。VD3QN-PER 在约 100 回合后即可进入稳定收敛阶段，最终平均回合奖励为 120.80，而 DQN 需 300 回合收敛，Dueling DQN 需 220 回合收敛，相比之下，VD3QN-PER 的收敛速度分别提升 66.67%（相较 DQN）和 54.55%（相较 Dueling DQN）。主要原因在于，DQN 和 DDQN 受限于 Q 值高估问题，在复杂环境中难以兼顾奖励最大化与决策稳定性，导致收敛缓慢甚至策略震荡。Dueling DQN 仅在一定程度上缓解了状态值估计偏差，但缺乏对动态干扰因素的精细建模能力，使其在复杂环境中的适应性受限。VD3QN-PER 结合值分布建模与优先经验回放，能够有效地量化洋流扰动、障碍物影响等环境不确定性，从而优化策略学习过程。分位数回归建模确保了算法在复杂场景中的稳定收敛，同时优先经验回放机制避免了对低效经验的过度依赖，增强了决策的稳健性。




（a）VD3QN-PER和DQN回合平均奖励对比




（b）VD3QN-PER和DDQN回合平均奖励对比




（c）VD3QN-PER和Dueling DQN回合平均奖励对比

图 6　各算法在仿真实验中的回合平均奖励曲线对比

为了验证不同强化学习算法在不同洋流强度中的鲁棒性，不同洋流强度下各个算法的平均奖励和平均步数如图 7 所示，不同洋流强度下各个算法的平均跟踪精度如图 8 所示，随着洋流强度的增加，所有算法的平均奖励均有所下降，表明外部环境的不确定性增强，使 AUV 精确执行目标跟踪的难度加大。VD3QN-PER 与 Dueling DQN 在较强洋流干扰下仍能保持相对稳定的平均奖励，表明其具有较强的环境适应能力。其中，VD3QN-

PER的平均奖励波动最小，跟踪误差最优，进一步验证了值分布建模对复杂环境的适应性。Dueling DQN误差随洋流强度增加而有所上升，表明其缺乏对复杂动态环境不确定性的建模能力，在强洋流条件下策略表现不稳定。DQN和DDQN在高洋流干扰环境中表现不佳，奖励值下降幅度较大，且回合步数增加，说明其无法有效应对复杂水下环境的不确定性。因此，VD3QN-PER结合分布强化学习的优势，在洋流强度增加的情况下仍能保持稳定决策，表现出良好的鲁棒性。

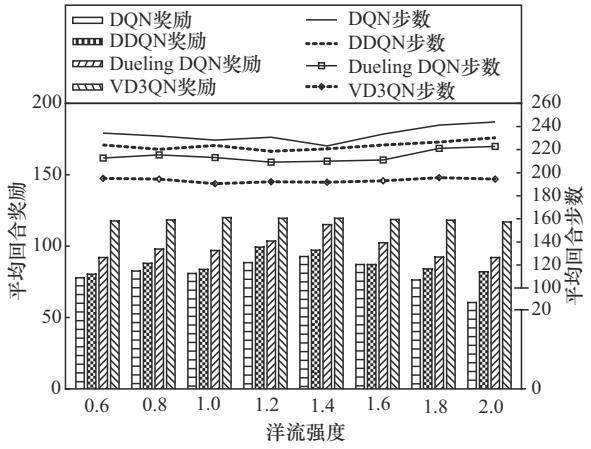



图7　不同洋流强度下各个算法的平均奖励和平均步数

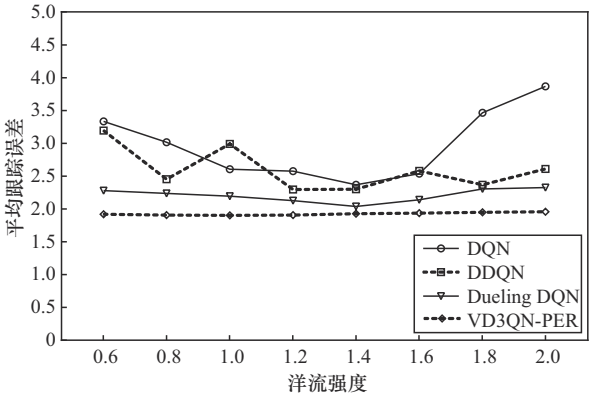
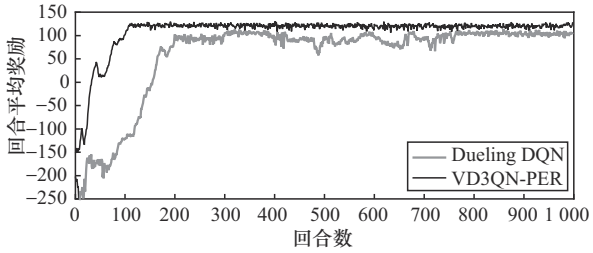
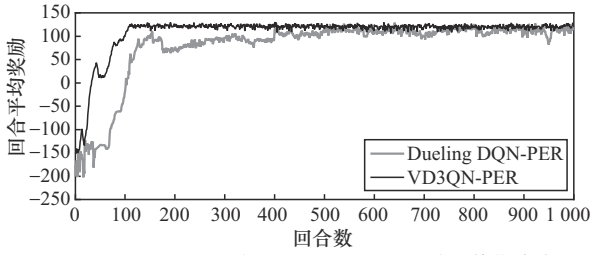



图8　不同洋流强度下各个算法的平均跟踪精度

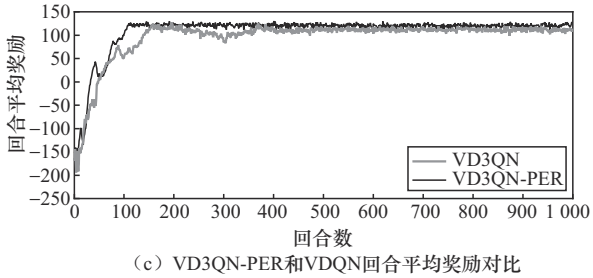
为验证各模块对VD3QN-PER最终训练结果的贡献，设定洋流强度为1.0，对本文进行消融实验，VD3QN-PER在不同模块下回合奖励函数曲线对比如图9所示，VD3QN-PER在不同模块下回合跟踪误差对比如图10所示。实验分析表明：（1）与VD3QN相比，VD3QN-PER的收敛速度提升，表明PER机制在数据利用效率上起到了关键作用。通过优先采样高TD误差样本，PER有效地提升了经验重放的有效性，从而加快了训练进程。（2）与Dueling DQN-PER相比，VD3QN-PER的最终平均奖励更高，同时误差精度更优，这说明值分布建模能够更全面地刻画回报的不确定性，减少价值估计的偏差，提高策略稳定性。（3）与Dueling DQN相比，VD3QN和VD3QN-PER在收敛速度和回合奖励值上均有显著提升，误差精度优势尤为明显。这表明Dueling架构虽能提升Q值估计的效率，但结合值分布建模和PER后，策略对复杂环境的适应能力进一步增强。




（a）VD3QN-PER和Dueling DQN回合平均奖励对比




（b）VD3QN-PER和Dueling DQN-PER回合平均奖励对比




（c）VD3QN-PER和VDQN回合平均奖励对比

图9　VD3QN-PER在不同模块下回合奖励函数曲线对比

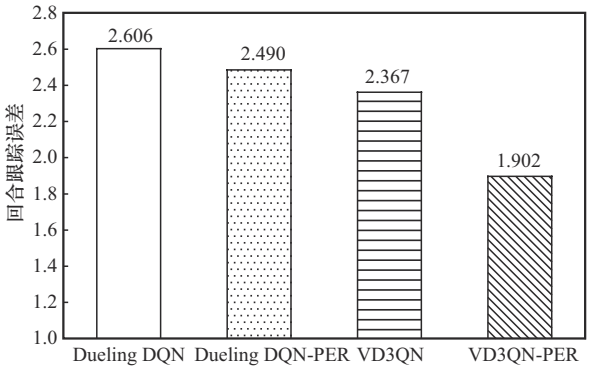



图 10　VD3QN-PER 在不同模块下回合跟踪误差对比

## 5　结束语

本文基于真实三维洋流数据构建了动态目标跟踪环境，模拟 AUV 在水下三维洋流条件下的跟踪任务，提出了一种基于分布式强化学习的 AUV 水下三维洋流目标跟踪控制算法，以优化 AUV 的跟踪策略。该算法通过引入值分布建模，改善 Q 值估计的准确性。采用 Dueling 结构，解耦状态和动作的价值评估，从而增强策略的稳定性。引入优先经验回放机制，加速关键经验的学习过程。实验结果表明，VD3QN-PER 在不同洋流强度下均展现出更快的收敛速度、更高的跟踪准确率和更强的鲁棒性，显著优于 DQN、DDQN、Dueling DQN 等基准算法。

然而，当前研究主要集中于提升目标跟踪精度和决策稳定性，未充分考虑如何在有限的能量约束下优化跟踪过程中的能量消耗。未来的研究可以进一步探索如何在目标跟踪任务中融入能量效率的优化策略，从而在保证性能的同时降低能耗。

## 参考文献：

[1] 李锦江, 向先波, 刘传,等. 基于预设性能制导律的欠驱动 AUV 海底地形鲁棒时滞跟踪控制[J]. 上海交通大学学报, 2022, 56(7): 944-952.
LI J J, XIANG X B, LIU C, et al. Robust seabed terrain following control of underactuated AUV with prescribed performance guidance law under time delay of actuator[J]. Journal of Shanghai Jiaotong University, 2022, 56(7): 944-952.

[2] SHEN C, SHI Y, BUCKHAM B. Integrated path planning and tracking control of an AUV: a unified receding horizon optimization approach[J]. IEEE/ASME Transactions on Mechatronics, 2016, 22(3): 1163-1173.

[3] LI Y, CAI K, ZHANG Y, et al. Localization and tracking for AUVs in marine information networks: research directions, recent advances, and challenges[J]. IEEE Network, 2019, 33(6): 78-85.

[4] 郭银景, 侯佳辰, 吴琪, 等. AUV 全局路径规划环境建模算法研究进展[J]. 舰船科学技术, 2021, 43(17): 12-18.
GUO Y J, HOU J C, WU Q, et al. Research progress of AUV global path planning environment modeling algorithm[J]. Ship Science and Technology, 2021, 43(17): 12-18.

[5] 潘伟, 曾庆军, 姚金艺, 等. 全驱动自主水下机器人回收路径跟踪模糊滑模控制[J]. 船舶与海洋工程, 2022, 38(5): 45-50.
PAN W, ZENG Q J, YAO J Y, et al. A research on fuzzy sliding mode control for fully actuated AUV recovery path tracking[J]. Naval Architecture and Ocean Engineering, 2022, 38(5): 45-50.

[6] YANG J, NI J, XI M, et al. Intelligent path planning of underwater robot based on reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2022, 20(3): 1983-1996.

[7] HAO L Y, WANG R Z, SHEN C, et al. Trajectory tracking control of autonomous underwater vehicles using improved tube-based model predictive control approach[J]. IEEE Transactions on Industrial Informatics, 2024, 20(4-Part1): 11.

[8] 郭琳钰, 高剑, 焦慧锋, 等. 基于 RBF 神经网络的自主水下航行器模型预测路径跟踪控制[J]. 西北工业大学学报, 2023, 41(5): 871-877.
GUO L Y, GAO J, JIAO H F, et al. Model predictive path following control of underwater vehicle based on RBF neural network[J]. Journal of Northwestern Polytechnical University, 2023, 41(5): 871-877.

[9] MA D, CHEN X, MA W, et al. Neural network model-based reinforcement learning control for AUV 3-D path following[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 893-904.

[10] WANG J, WU Z, YAN S, et al. Real-time path planning and following of a gliding robotic dolphin within a hierarchical frame-

work[J]. IEEE Transactions on Vehicular Technology, 2021, 70(4): 3243-3255.

[11] 金立生, 韩广德, 谢宪毅, 等. 基于强化学习的自动驾驶决策研究综述[J]. 汽车工程, 2023, 45(4): 527-540.
JIN L S, HAN G D, XIE X Y, et al. Review of autonomous driving decision-making research based on reinforcement learning[J]. Automotive Engineering, 2023, 45(4): 527-540.

[12] SINGH B, KUMAR R, SINGH V P. Reinforcement learning in robotic applications: a comprehensive survey[J]. Artificial Intelligence Review, 2022, 55(2): 945-990.

[13] ZHANG M, WU S, JIAO J, et al. Energy- and cost-efficient transmission strategy for UAV trajectory tracking control: a deep reinforcement learning approach[J]. IEEE Internet of Things Journal, 2023, 10(10): 8958-8970.

[14] ZHANG T, LEI J, LIU Y, et al. Trajectory optimization for UAV emergency communication with limited user equipment energy: a safe-DQN approach[J]. IEEE Transactions on Green Communications and Networking, 2021, 5(3): 1236-1247.

[15] ZHANG W, GAI J, ZHANG Z, et al. Double-DQN based path smoothing and tracking control method for robotic vehicle navigation[J]. Computers and Electronics in Agriculture, 2019, 166: 104985.

[16] CHU Z, WANG F, LEI T, et al. Path planning based on deep reinforcement learning for autonomous underwater vehicles under ocean current disturbance[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1): 108-120.

[17] YUE Y, PAN Z, LI S, et al. Reinforcement learning based smart UUV-IoUT localization in underwater acoustic topology network[J]. IEEE Internet of Things Journal, 2025: 1.

[18] LI Y, HE X, LU Z, et al. Comprehensive ocean information-enabled AUV motion planning based on reinforcement learning[J]. Remote Sensing, 2023, 15(12): 3077.

[19] LIDTKE A K, RIJPKEMA D, DUZ B. General reinforcement learning control for AUV manoeuvring in turbulent flows[J]. Ocean Engineering, 2024, 309(2): 118538.

[20] LIU Z, HOU J, NING D, et al. Improving deep Q network based on marketing psychology for AUV path planning in unknown marine environments[J]. IEEE Internet of Things Journal, 2025, 12(5): 5476-5487.

[21] LI Y, HUANG H, ZHUANG Y, et al. An AUV-assisted data collection scheme for UWSNs based on reinforcement learning under the influence of ocean current[J]. IEEE Sensors Journal, 2023, 24(3): 3960-3972.

[22] YU Y, ZHENG H, XU W. Learning and sampling-based informative path planning for AUVs in ocean current fields[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2024.

[23] TO K Y C, KONG F H, LEE K M B, et al. Estimation of spatially-correlated ocean currents from ensemble forecasts and online measurements[C]//Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2021: 2301-2307.

[24] KONG F H, TO K Y C, BRASSINGTON G, et al. 3D ensemble-based online oceanic flow field estimation for underwater glider path planning[C]//Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2021: 4358-4365.

[25] LI X, YU S. Three-dimensional path planning for AUVs in ocean currents environment based on an improved compression factor particle swarm optimization algorithm[J]. Ocean Engineering, 2023, 280: 114610.

[26] DUAN J, GUAN Y, LI S E, et al. Distributional soft actor-critic: off-policy reinforcement learning for addressing value estimation errors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(11): 6584-6598.

[27] LUIS C E, BOTTERO A G, VINOGRADSKA J, et al. Value-distributional model-based reinforcement learning[J]. Journal of Machine Learning Research, 2024, 25(298): 1-42.

[28] XI M, YANG J, WEN J, et al. Comprehensive ocean information-enabled AUV path planning via reinforcement learning[J]. IEEE Internet of Things Journal, 2022, 9(18): 17440-17451.

[29] YANG J, NI J, XI M, et al. Intelligent path planning of underwater robot based on reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2022.

[30] Copernicus Marine Service. Global ocean physics reanalysis dataset [DB]. [2025-02-18].

[31] LIU X, YU M, YANG C, et al. Value distribution DDPG with dual-prioritized experience replay for coordinated control of coal-fired power generation systems[J]. IEEE Transactions on Industrial Informatics, 2024.

[作者简介]

**赵剑楠**（1994– ），男，博士，广西大学电气工程学院助理教授，主要研究方向为自主无人机、仿生视觉算法、人工智能等。

**覃琪琪**（1999– ），女，广西大学电气工程学院硕士生，主要研究方向为强化学习、水下传感器网络。

**李云**（1978– ），女，博士，广西民族大学物理与电子信息学院教授，多模态信息智能感知处理与应用广西高校工程研究中心、广西智语人形机器人重点实验室、广西智能视觉协作机器人工程研究中心核心成员，主要研究方向为水下传感器网络、大数据分析、人工智能等。

**苏毅珊**（1978– ），男，博士，天津大学电气与自动化工程学院副教授、博士生导师，主要研究方向为水声通信与海洋立体信息网络、计算机网络、传感器网络、物联网等。