[引用格式] 赵少靖, 付松琛, 白乐天, 等. 基于自适应多目标优化的 UUV 全覆盖路径规划方法 [J]. 水下无人系统学报, 2025, 33(3): 459-472.

基于自适应多目标优化的 UUV 全覆盖路径规划方法

赵少靖[1,2], 付松琛[1,2], 白乐天[1,2], 郭雨桐[1], 黎 塔[1,2]

(1. 中国科学院声学研究所 语音与智能信息处理实验室, 北京, 100190; 2. 中国科学院大学, 北京, 100049)

摘 要: 全覆盖路径规划作为无人水下航行器(UUV)在未知水域环境中的一项关键任务,受环境不确定性、运动约束和能耗限制等因素影响, 传统路径规划方法难以适应复杂场景。文中提出了一种基于自适应多目标优化的 UUV 全覆盖路径规划方法, 结合近端优化强化学习算法与动态权重调节机制, 通过奖励目标的相关性分析与线性回归估计, 自适应调整不同优化目标的权重, 使 UUV 能够在未知障碍物和洋流环境中自主规划高效的覆盖路径。为验证方法的有效性, 构建了一个基于二维仿真环境的 UUV 运动与声呐探测模型, 其中 UUV 运动模型在 6 自由度刚体运动的基础上简化为平面运动, 并在多种障碍物分布与随机洋流条件下进行了对比实验分析。实验结果表明, 相较于传统方法, 该方法能够在提高覆盖率的同时优化任务完成率、轨迹长度、能耗与信息延时等关键指标。其中, 覆盖率提升 4.03%, 任务完成率提高 10%, 效用值提升 10.96%, 任务完成时间缩短 14.13%, 轨迹长度减少 26.85%, 能耗降低 10.3%, 信息延时减少 19.34%。结果证明该方法能够在复杂环境中显著提升 UUV 的适应性和鲁棒性, 为自主水下探测任务提供了新的优化策略参考。

关键词: 无人水下航行器; 全覆盖路径规划; 强化学习; 多目标优化; 自适应权重调节

中图分类号: TJ630; U674.941 **文献标识码:** A **文章编号:** 2096-3920(2025)03-0459-14

DOI: 10.11993/j.issn.2096-3920.2025-0031

Adaptive Multi-Objective Optimization-Based Coverage Path Planning Method for UUVs

ZHAO Shaojing[1,2], *FU Songchen*[1,2], *BAI Letian*[1,2], *GUO Yutong*[1], *LI Ta*[1,2]

(1. Laboratory of Speech and Intelligent Information Processing, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Coverage path planning for unmanned undersea vehicles(UUVs) in unknown aquatic environments is a critical task. However, due to environmental uncertainties, motion constraints, and energy limitations, traditional path planning methods struggle to adapt to complex scenarios. This paper proposed an adaptive multi-objective optimization-based coverage path planning method for UUVs, integrating proximal policy optimization(PPO) with a dynamic weight adjustment mechanism. By analyzing the correlation between reward objectives and employing linear regression estimation, the proposed approach adaptively adjusted the weights of different optimization objectives, enabling UUVs to autonomously plan efficient coverage paths in environments with unknown obstacles and ocean currents. To validate the effectiveness of the proposed method, a UUV motion and sonar detection model based on a two-dimensional simulation environment was constructed. Among them, the UUV motion model was simplified to a planar motion model on the basis of the six-degree-of-freedom rigid-body motion. Comparative experiments were conducted under various obstacle distributions and random ocean currents. Experimental results demonstrate that compared with traditional methods, the proposed approach improves coverage while optimizing mission completion rate, trajectory length, energy consumption, and information latency. Specifically, it increases

收稿日期: 2025-02-25; **修回日期:** 2025-03-19; **录用日期:** 2025-03-25.

作者简介: 赵少靖(1998-), 男, 在读博士, 主要研究方向为水下航行器的自主决策算法及其应用.

coverage by 4.03%, enhances mission completion rate by 10%, improves utility by 10.96%, reduces mission completion time by 14.13%, shortens trajectory length by 26.85%, lowers energy consumption by 10.3%, and decreases information latency by 19.34%. These results indicate that the proposed method significantly enhances the adaptability and robustness of UUVs in complex environments, providing a novel optimization strategy for autonomous underwater exploration tasks.

Keywords: unmanned undersea vehicle; coverage path planning; reinforcement learning; multi-objective optimization; adaptive weight adjustment

0 引言

无人水下航行器(unmanned undersea vehicle, UUV)在水下自主探测、环境监测和侦察等任务中发挥着重要作用[1]。然而, 其在未知水域中自主作业仍面临环境不确定性、运动约束、复杂任务需求及能耗限制等挑战, 使得传统路径规划方法难以有效适应[2]。

全覆盖路径规划是 UUV 自主作业的关键问题, 旨在确保 UUV 在有限水域内高效遍历所有自由空间, 完成探测任务。该技术广泛应用于海洋资源勘探、环境监测和水下侦察等领域, 例如海底地形扫描、污染监测和目标搜索[3]。然而, 未知障碍物[4]、洋流干扰[5] 及运动约束[6] 使路径规划复杂度显著增加, 需要在覆盖率、能耗、轨迹长度和任务效率等多个目标之间权衡[7]。

传统全覆盖路径规划方法主要包括随机探索法、几何法、栅格法和分解法[8]。随机探索法通过随机生成路径点扩展覆盖范围, 如快速扩展随机树(rapidly-exploring random tree, RRT)方法在空间中随机采样路径点并扩展路径树, 以提高搜索均匀性和效率[9]。几何法利用环境几何特征进行规划, 如扫描线法通过平行扫描线遍历目标区域[10]。栅格法将目标区域离散化为网格单元, 例如螺旋生成树法构建覆盖所有网格的树形结构[11]。分解法则将目标区域划分为多个子区域, 并在各子区域内应用几何方法规划路径后合并, 如梯形分解法通过扫描线遍历所有子区域[12]。

近年来, 强化学习为全覆盖路径规划提供了新的解决思路。不同于依赖先验信息和预设规则的传统方法, 强化学习可通过环境交互自主优化策略, 提高适应性和规划效率。相关研究已取得一定进展, 如 Kyaw 等[13] 结合循环神经网络(recurrent neural network, RNN))和策略梯度算法优化旅行商问题(travelling salesman problem, TSP), 降低计算复杂度; Heydari 等[14] 采用双重深度 Q 网络(double deep Q-network, DDQN)与优先经验回放(prioritized experience replay, PER)减少路径重复, 提高规划效率; Ai 等[15] 在海上搜索与救援任务中提出强化学习驱动的规划方法, 优先搜索高概率区域; Rückin 等[16] 结合强化学习与蒙特卡洛树搜索(Monte Carlo tree search, MCTS), 加速大动作空间中的路径选择; Zhao 等[17] 利用 DDQN 训练策略, 通过区域分割扩展至复杂环境, 实现全覆盖路径规划; Xing 等[18] 提出基于深度强化学习的无人水面船路径规划算法, 提高覆盖率并减少路径重复; Jonnarth 等[19] 基于柔性“演员-评论家”(soft actor-critic, SAC)算法, 直接输出控制信号, 实现端到端的路径规划。

尽管强化学习方法在全覆盖路径规划中展现出优越性, 但仍存在诸多挑战。首先, 多数方法采用网格离散化的动作空间, 难以准确刻画物理环境的连续性; 其次, 路径规划的多目标优化问题未得到充分建模, 难以在覆盖率、能耗以及轨迹优化等目标间自适应权衡; 此外, 部分研究对 UUV 运动模型考虑不足, 未能有效反映运动约束; 同时, 许多方法假设传感器覆盖范围为理想化圆形, 而未考虑 UUV 声呐探测区域的实际形态及洋流对航行轨迹的影响, 降低了路径规划的实用性。

针对上述问题, 提出了一种面向复杂水域环境的 UUV 全覆盖路径规划方法, 支持多目标优化。构建了二维连续空间与动作仿真环境, 结合 6 自由度 UUV 运动模型和声呐探测模型, 模拟未知障碍物与洋流影响, 以更贴近真实水域环境。在此基础上, 结合近端策略优化(proximal policy optimization, PPO)算法与多目标优化方法, 提出在线路径规划策略, 实现不同优化目标的自适应权衡, 提高任务整体效用值。关注的优化目标包括区域覆盖率、碰撞率、能耗、轨迹长度、信息延时、任务完成时间和任务完成率(见 1.6 节), 并通过效用值进行综合评估。文中研究的主要贡献如下:

1) 提出动态多目标权衡机制, 使强化学习能够优化未直接作用于奖励回报的效用函数, 从而提升 UUV 全覆盖路径规划任务的整体效用值;

2) 在仿真环境中引入简化的 UUV 6 自由度运动模型、声呐特性及洋流影响, 以提高仿真的真实性;

3) 通过多组实验验证, 证明了多目标显式建模与自适应权重调节方法在优化路径规划效果方面的有效性。

1　研究基础

1.1　多目标强化学习

多目标与单目标强化学习均可建模为马尔可夫决策过程(Markov decision process, MDP), 表示为五元组: $\mathcal{M}=\langle \mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma\rangle$。其中: $\mathcal{S}$为状态空间; $\mathcal{A}$为动作空间; $\mathcal{P}(s'|s,a)$为状态转移概率; $\mathcal{R}(s,a)$为奖励函数; $\gamma\in[0,1]$为折扣因子。

强化学习目标是寻找最优策略π^*, 使累积奖励最大化。单个目标的累积奖励可通过价值函数$V^{\pi}(s)$或 Actor-Critic 函数$Q^{\pi}(s,a)$估计, 即

$$V^{\pi}(s)=\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0=s,\pi\right] \tag{1}$$

$$\begin{aligned} Q^{\pi}(s,a)=&\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0=s,a_0=a,\pi\right]= \\ &\mathbb{E}\left[r_0+\gamma V^{\pi}(s_1)\mid s_0=s,a_0=a,\pi\right] \end{aligned} \tag{2}$$

最优策略可以表示为$\pi^*=\arg\max\limits_{\pi} V^{\pi}$, 同时定义动作优势值$A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s)$衡量动作相较于状态值的优势, 其中$s$和$a$分别表示特定状态和动作, 而下标$t$表示特定时刻。

在连续状态与动作空间中, 常采用 Actor-Critic 框架[20]。值函数模型的优化目标基于贝尔曼前向公式的时序差分误差, 即

$$J_V(\psi)=\mathbb{E}_{\pi_\theta}\left[\frac{1}{2}\left(\hat{V}(s')-V_\psi(s)\right)^2\right] \tag{3}$$

式中, $\hat{V}(s)=r+\gamma V_\psi(s')$为目标值, 其中, r为单步奖励值。策略模型的优化目标基于策略梯度, 即

$$J_\pi(\theta)=\mathbb{E}_{\pi_\theta}\left[-A(s,a)\log\pi_\theta(a|s)\right] \tag{4}$$

该目标增加优势值为正的动作选择概率, 减少优势值为负的动作选择概率。

多目标强化学习中, 单步奖励$\boldsymbol{r}_t$是一个包含多个目标奖励值的向量, 通常可以通过多个价值函数来估计其累计奖励期望。文中主要研究效用值视角下的单策略方法, 根据标量化和策略优化方式的不同, 可分为以下 3 类。

1) 加权和法(weighted sum reward, WSR): 对m个目标奖励加权求和, 奖励i的权重为ω_i, 形成t时刻的标量奖励r_t, 并优化单一价值函数, 即

$$r_t=\sum_{i=1}^{m}\omega_i\mathcal{R}_i(s_t,a_t) \tag{5}$$

2) 标量化期望回报(scalarized expected return, SER)[21]: 分别估计各目标回报, 并最大化多目标回报的期望标量化值, 即

$$\pi^*=\arg\max_{\pi} u\left(\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\boldsymbol{r}_t|\pi,s_0\right]\right) \tag{6}$$

3) 期望标量化回报(expected scalarized return, ESR)[22]: 分别估计各目标回报, 并最大化标量化效用值的期望, 即

$$\pi^*=\arg\max_{\pi}\mathbb{E}\left[u\left(\sum_{t=0}^{\infty}\gamma^t\boldsymbol{r}_t\right)\Big|\pi,s_0\right] \tag{7}$$

其中, $u:\mathbb{R}^m\to\mathbb{R}$为标量化函数。WSR 和 SER 可在任务执行过程中借助价值函数计算效用值, 适用于线性场景; 而 ESR 需在任务完成后计算, 适用于非线性场景。此外, SER 和 ESR 依赖直接作用于奖励回报的标量化函数, 以实现从回报值到奖励值的映射。

在全覆盖路径规划问题中, 效用值无法直接从回报映射得出, 因为它涉及覆盖率、能耗、路径长度和碰撞率等多维指标, 因此 SER 和 ESR 方法均不适用。为此, 文中结合 SER 和 ESR 方法, 在训练过程中自适应调整权重, 以动态平衡各目标的影响, 最大化任务的总体效用值。

1.2　UUV 运动模型

采用 6 自由度刚体运动模型来描述 UUV 在水下环境中的运动状态, 其中, 局部坐标系用于表示 UUV 相对于自身的姿态和运动, 而全局坐标系用于表征其绝对位置[23]。6 个自由度包括局部坐标系中的 3 个平移自由度(沿 x、y、z 轴的位移)以及绕 3 个坐标轴的旋转自由度(横滚角、俯仰角和偏航角)。为简化环境仿真, 并基于 UUV 在多数任务中维持相对固定工作深度的经验, 假设 UUV 处

于固定深度, 即约束 6 自由度中的 z 轴位移和俯仰角为零。

UUV 在包括螺旋桨、舵板和水流等作用力影响的运动方程可以表示为

$$\begin{aligned}\dot{\boldsymbol{\eta}} = \boldsymbol{J}(\boldsymbol{\eta})(\boldsymbol{v}_r+\boldsymbol{v}_c)\boldsymbol{M}\dot{\boldsymbol{v}}_r+\boldsymbol{C}(\boldsymbol{v}_r)\boldsymbol{v}_r+\\ \boldsymbol{D}(\boldsymbol{v}_r)\boldsymbol{v}_r+\boldsymbol{g}(\boldsymbol{\eta})+\boldsymbol{g}_o=\boldsymbol{\tau}\end{aligned} \tag{8}$$

式中: $\boldsymbol{\eta}=[x^n,y^n,z^n,\phi,\theta,\psi]^{\mathrm{T}}$为 UUV 在全局坐标系下的位置和欧拉角; $\boldsymbol{v}_r=[\boldsymbol{u}_r,\boldsymbol{v}_r,\boldsymbol{w}_r,\boldsymbol{p},\boldsymbol{q},\boldsymbol{r}]^{\mathrm{T}}$为 UUV 在局部坐标系下对水的速度和角速度; $\boldsymbol{v}_c$为水流在局部坐标系下的速度和角速度; $\boldsymbol{J}(\boldsymbol{\eta})=\begin{bmatrix}\boldsymbol{R}(\boldsymbol{\eta}) & \boldsymbol{0}_{3\times3}\\ \boldsymbol{0}_{3\times3} & \boldsymbol{T}(\boldsymbol{\eta})\end{bmatrix}$为状态向量和速度向量之间的雅可比矩阵, 由旋转矩阵$\boldsymbol{R}(\boldsymbol{\eta})$和角速度转换矩阵$\boldsymbol{T}(\boldsymbol{\eta})$构成; $\boldsymbol{M}$为 UUV 的质量矩阵; $\boldsymbol{C}(\boldsymbol{v}_r)$为科里奥利力矩阵, 描述非惯性力的影响; $\boldsymbol{D}(\boldsymbol{v}_r)$为阻力矩阵, 表示流体阻力对运动的影响; $\boldsymbol{g}(\boldsymbol{\eta})$ 为重力和浮力引起的恢复力矩阵; $\boldsymbol{g}_o$为其他外力的影响; $\boldsymbol{\tau}$为由螺旋桨、舵板产生的控制力矩, 由三部分构成, 具体为

$$\boldsymbol{\tau}=\boldsymbol{M}_T+\boldsymbol{M}_\psi+\boldsymbol{M}_\theta \tag{9}$$

式中: $\boldsymbol{M}_T=\boldsymbol{k}_T\cdot\omega^2$为螺旋桨提供的推进力矩, $\boldsymbol{k}_T$为推力常数矩阵, ω为螺旋桨转速; $\boldsymbol{M}_\psi=\boldsymbol{C}_\psi\cdot\delta_\psi$为偏航舵板产生的偏航力矩, $\boldsymbol{C}_\psi$为力矩增益, δ_ψ为舵板的偏转角; $\boldsymbol{M}_\theta=C_\theta\cdot\delta_\theta$为俯仰舵板产生的俯仰力矩。

UUV 的控制输入变量为螺旋桨转速ω和偏航舵板的偏转角δ_ψ, 其中螺旋桨转速的取值范围为 500~3 050 r/min, 偏转角的取值范围为−15°~15°。UUV 长度为 1.6 m, 直径 0.19 m, 螺旋桨转速限制下的最大速度约 10 kn(5.2 m/s), 最小速度约 1.5 kn (0.8 m/s), 舵板偏转角限制下的最小转弯半径约 50 m。

1.3　声呐模型

UUV 搭载前置主动声呐, 其探测性能受发射频率、声速、水深和环境噪声等因素影响。为简化研究, 假设声呐为理想单波束指向性声呐, 水下环境均匀, 无温跃层、多路径效应或湍流影响。同时, 忽略目标的反射特性, 假设所有目标的声学反射能力相同, 仅考虑方位角对探测距离的影响。主动声呐的探测能力在主瓣方向(正前方)最强, 随角度偏移逐渐减弱, 通常可用高斯波束近似描述: $D(\theta)=\mathrm{e}^{-\beta\theta^2}$。假设探测距离主要由波束指向性决定, 即 $r_{\mathrm{sonar}}(\theta)=r_{\max}\cdot D(\theta)=\alpha\mathrm{e}^{-\beta\theta^2}$。其中: β为控制波束的角度衰减, 值越大则探测范围越窄; α为最大探测距离, 取决于声呐系统性能; r_{sonar}为探测距离; θ为目标相对 UUV 的方位角(与前进方向的夹角)。在此假设下, r_{sonar}与θ之间的关系可以近似表示为

$$r_{\mathrm{sonar}}=\begin{cases}\alpha\cdot\mathrm{e}^{-\beta\cdot\theta^2} & \theta\in\left[-\dfrac{\pi}{2},\dfrac{\pi}{2}\right]\\ 0 & \text{else}\end{cases} \tag{10}$$

为简化仿真, 假设$\alpha=2$ km, $\beta=1$, 即 UUV 正前方探测距离为 2 km, 方位角约为 47°时衰减至 1 km。文中模型旨在提高计算效率并聚焦于方法验证, 因此不考虑复杂声传播效应及目标材质、形状对回波强度的影响。

1.4　任务场景

针对全覆盖路径规划问题, 设定的目标区域为一个边长为 10 km 的正方形, 总面积为 100 km^2。根据障碍物信息的已知性和形状特征的不同, 任务场景被划分为 4 种类型。场景 1: 障碍物已知, 形状为矩形; 场景 2: 障碍物未知, 形状为矩形; 场景 3: 障碍物已知, 形状为不规则多边形; 场景 4: 障碍物未知, 形状为不规则多边形。在所有场景中, 障碍物数量随机分布在 1~5 之间, 且每个障碍物的大小和形状均随机生成。

在障碍物已知的情况下, UUV 可以在初始时刻获取全局地图信息; 而在障碍物未知的情况下, UUV 初始时仅能获取空白地图, 并需要在任务执行过程中实时探测并构建环境地图。4 种任务场景均引入了随机生成的洋流, 以模拟现实环境中的水流影响。假设洋流由多个涡旋源产生, 每个涡旋源产生的速度场可以表示为[24]

$$\boldsymbol{V}(\boldsymbol{p})=\begin{bmatrix}\boldsymbol{V}_x(\boldsymbol{p})\\ \boldsymbol{V}_y(\boldsymbol{p})\end{bmatrix} \tag{11}$$

其中, $\boldsymbol{p}=[x,y]$为当前位置。

涡旋源位置为$\boldsymbol{P}_t=[x_0,y_0]$, 涡旋强度为$\Omega$, 并且涡旋的影响范围由半径$R$控制。涡旋场可以用下述方程描述:

$$\begin{cases}\boldsymbol{V}_x(\boldsymbol{p})=-\dfrac{\Omega}{2\pi}\cdot\dfrac{y-y_0}{\|\boldsymbol{p}-\boldsymbol{P}_t\|^2}\cdot\left(1-\exp\left(-\left(\dfrac{\|\boldsymbol{p}-\boldsymbol{P}_t\|}{R}\right)^2\right)\right)\\ \boldsymbol{V}_y(\boldsymbol{p})=\dfrac{\Omega}{2\pi}\cdot\dfrac{x-x_0}{\|\boldsymbol{p}-\boldsymbol{P}_t\|^2}\cdot\left(1-\exp\left(-\left(\dfrac{\|\boldsymbol{p}-\boldsymbol{P}_t\|}{R}\right)^2\right)\right)\end{cases} \tag{12}$$

式中, $\|\boldsymbol{p}-\boldsymbol{P}_t\|$为当前位置到涡旋源的欧氏距离。对于任务场景中的任意位置$\boldsymbol{p}$, 其由洋流产生的速度为多个涡旋速度的叠加, 即

$$\boldsymbol{V}=\sum_{t=1}^{N}\boldsymbol{V}_t(\boldsymbol{p}) \tag{13}$$

在生成随机洋流速度场时, 假定涡旋强度$\Omega=5$, 涡旋半径R在 200 m~2 km 间取随机值, 涡旋源的数量$N=20$。根据 20 m 的采样间距, 生成 500×500 的洋流速度场矩阵, 并通过归一化将速度场的最大水流速度限制为 1 kn(约 0.514 m/s)。在仿真过程中, UUV 所处位置的洋流速度取自最近采样点的速度值。

1.5　仿真环境

基于上述运动模型、声呐模型及任务场景, 设计了一个针对全覆盖路径规划问题的特定模拟仿真环境。

在时间维度上, UUV 在仿真环境中按时间步进行观测与决策, 每步的实际时间 30 s。仿真环境内部, 在给出 UUV 控制量后, 假设该控制量在接下来的 30 s 内保持不变, 并以 0.2 s 的离散时间步长推演 UUV 的位置、姿态和运动状态。UUV 启用声呐时, 会获取 30 s 步长开始时的环境信息, 因此声呐的探测结果存在 30 s 的滞后。

在空间维度上, 仿真环境模拟 UUV 在连续空间中的运动, 即 UUV 的位置坐标是连续值。根据障碍物是否已知及声呐探测结果, 仿真环境按 1∶200 的比例尺构建全局地图。当 UUV 执行主动声呐探测时, 依据声呐模型提供的探测距离更新地图, 并将相应位置标记为已探测区域。遇到障碍物时, 仿真环境会模拟遮挡效应, 并将障碍物边缘所在的区域标记为“有障碍”。如图 1 左侧所示, 当 UUV 启用主动声呐时, 浅红色的声呐覆盖区域内存在一黑色不规则多边形障碍物。图 1 右侧展示了生成的地图, 其中: 白色区域表示未探测区域; 红色区域表示 UUV 当前位置; 绿色区域表示已探测且无障碍物的区域; 黑色区域表示已探测到障碍物的区域。由于障碍物的遮挡效应, 其后方的区域仍处于未探测状态。

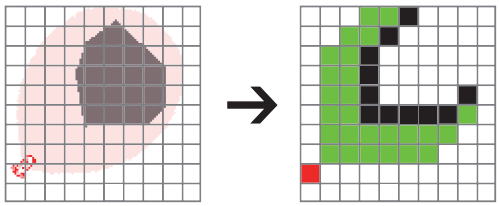


图 1　仿真环境地图构建过程

Fig. 1　Building process of simulation environment map

在碰撞检测方面, 仿真环境将 UUV 视作宽度为 10 m、长度为 20 m 的矩形(检测范围大于 UUV 的实际尺寸, 以确保安全), 并对 UUV 与障碍物之间进行多边形碰撞检测, 若检测到交叉, 则判定为碰撞发生, 并提前终止仿真任务。

1.6　评价指标

针对全覆盖路径规划问题, 通过多维度评价指标量化评估 UUV 性能, 并将多个指标映射为可比较的任务效用值, 以衡量策略优劣。具体指标如下。

1) 覆盖率r_{coverage}: 任务区域内自由空间被声呐探测到的网格数占总自由空间网格数的比值。

2) 碰撞率r_{crash}: UUV 在多任务场景下的碰撞实验次数占总实验次数的比值。

3) 轨迹长度l_{path}: 考虑覆盖率与轨迹长度的相关性, 为公平比较, 仅统计 UUV 从任务开始到覆盖率达 90% 时的运动轨迹长度, 未达 90% 覆盖率的实验不计入统计。

4) 能耗率r_{cost}: 与轨迹长度及声呐启用次数相关, 同样以达到 90% 覆盖率为基准, 计算 UUV 能耗占初始能量的比值, 未达 90% 覆盖率的实验不计入统计。

5) 信息延时t_{delay}: 衡量区域信息的新旧程度。统计 UUV 最后一次探测某网格至覆盖率达 90% 间的时间差。例如, 某网格在 0、25、60 时刻被探测, UUV 在 100 时刻达到 90% 覆盖率, 则该网格信息延时为 100−60 = 40 s。未探测网格不计入统计, 最终取所有探测网格的平均信息延时。

6) 任务完成率r_{finished}: 覆盖率达到 90% 的实验次数占总实验次数的比值。

7) 任务完成时间t_{finished}: 从任务开始至覆盖率达 90% 所需的时间。

在不使用声呐的情况下, UUV 的续航约为 150 km。全覆盖路径规划中, 覆盖率和碰撞率是关键指标。为提升对高覆盖率(接近 1)和低碰撞率(接近 0)区域的分辨率, 采用指数变换归一化函数进行非线性调整, 可得

$$f(x,T)=\frac{\mathrm{e}^{T\cdot x}-1}{\mathrm{e}^{T}-1} \tag{14}$$

其中, T控制非线性程度, 以增强高覆盖率和低碰撞率下的刻画精度。任务效用值U计算如下:

$$U=\begin{pmatrix} f\left(r_{\text{coverage}},3\right)\times 0.7+\left(1-\dfrac{l_{\text{path}}}{150\ \text{km}}\right)\times \\ 0.2+(1-r_{\text{cost}})\times 0.1 \end{pmatrix}\times f(1-r_{\text{crashed}},3) \tag{15}$$

该效用函数通过指数变换提高对关键区域的敏感度, 使高覆盖率、低碰撞率的方案获得更优评价。在单次实验中, 若 UUV 未发生碰撞, 则碰撞率设为 0; 否则, 设为 1。

2 研究方法

基于 PPO 框架, 构建多目标 Critic 架构, 并创新性地设计动态自适应权重估计机制, 提出自适应权重多目标 PPO(adaptive weighting multi-objective PPO, AdaptiveW-MO-PPO)算法。详细介绍该算法的关键组成部分, 包括观测空间、动作空间、奖励函数、多目标优化方法及模型结构。

为验证算法的有效性, 实验设计涵盖多层次对比分析: 以经典 PPO 作为基线方法, 并构建多个对比算法, 包括平均权重多目标 PPO(AveragedW-MO-PPO)、等权重多目标 PPO(EqualW-MO-PPO)及平均权重单目标 PPO(adaptive weighting single-objective-PPO, AveragedW-SO-PPO)。

2.1 观测空间

由于研究的全覆盖路径规划问题对于 UUV 是部分可观测的, 因此采用观测空间代替状态空间, 作为策略模型和价值函数模型的输入。观测空间由两部分组成, 即二维网格地图信息和 UUV 状态信息。

基于仿真环境构建的网格地图, 对地图信息进行扩展, 使用与地图相同大小的附加网格。二维网格地图信息包含以下 6 部分: 1) UUV 声呐探测结果绘制的地图; 2) UUV 的历史运动轨迹; 3) 每个网格中心点到 UUV 的距离; 4) 每个网格中心点相对于 UUV 方位角的正弦值; 5) 每个网格中心点相对于 UUV 方位角的余弦值; 6) UUV 已知障碍物分布地图, 由声呐探测结果和任务提供的已知信息绘制。

以上 6 部分构成$6\times 50\times 50$的张量矩阵$\boldsymbol{O}_{\text{map}}$, 作为 UUV 的二维网格地图信息观测。

UUV 当前时刻的状态信息由一个 17 维向量描述, 具体包括: 螺旋桨、偏航舵板和声呐控制量; 声呐覆盖区域内最近障碍物的距离及其方位角的正弦值和余弦值(无障碍物时填充−1); 正前方距离最近障碍物的距离及其方位角的正弦值和余弦值(无障碍物时填充−1); UUV 的速度、偏航角的正弦和余弦值; 剩余能量比率; 全局 x 轴和 y 轴坐标; 声呐覆盖区域内未探测网格的比率, 反映障碍物遮挡情况; 声呐覆盖区域位于任务区域内部分网格的数量占声呐最大覆盖网格数的比率, 反映 UUV 在任务区域边界的情况。

将历史的 4 个 17 维状态向量与当前的状态向量堆叠成5×17的状态矩阵$\boldsymbol{O}_{\text{state}}$, 作为 UUV 的状态信息观测。

2.2 动作空间

采用策略模型直接预测 UUV 的控制信号, 包括螺旋桨转速、偏航舵板的偏转角度和声呐使能信号, 从而使 UUV 的路径更加符合真实场景。在控制信号的处理过程中, 螺旋桨转速和舵板偏转角度的取值范围(见 1.2 节)被归一化到$[-1,1]$。当声呐使能信号的值小于或等于 0 时, 声呐被关闭; 当信号值大于 0 时, 声呐被启用。

2.3 奖励函数

由于评价指标需要在全覆盖路径规划任务结束后进行计算, 因此必须设计中间过程的反馈奖励值。针对多个评价指标, 从不同维度设计了 7 个奖励函数, 生成 7 维的即时向量化奖励值。具体如下。

1) 时间: $r_{\text{time}}=-0.05$, 每一步固定奖励值。

2) 声呐: $r_{\text{sonar}}=\begin{cases}-0.1 & \text{声呐启用}\\ 0 & \text{声呐禁用}\end{cases}$。

3) 探测: 分为“探索”和“完成”两部分。“探索”部分为

$$r_{\text{explored}}^{(1)}=\left(C_{\text{cur}}-C_{\text{prev}}\right)\Big/13+T_{\text{delay}}\Big/500 \tag{16}$$

“完成”部分为

$$r_{\text{explored}}^{(2)}=\begin{cases}200 & r_{\text{coverage}}\geqslant 0.9\\ 0 & r_{\text{coverage}}<0.9\end{cases} \tag{17}$$

总奖励为

$$r_{\text{explored}}=r_{\text{explored}}^{(1)}+r_{\text{explored}}^{(2)} \tag{18}$$

其中: C_{cur}为累计至当前被声呐探测过的网格总数;

C_{prev}为累计至上一步被声呐探测过的网格总数, UUV 一次声呐最多能覆盖大约 52 个网格; T_{delay}为当前被声呐覆盖的网格的平均信息延时, 若未启用声呐或者首次探测到的网格, 则将其信息延时置零; r_{coverage}为任务结束时的覆盖率。

该奖励鼓励 UUV 探索新区域以及信息较旧的区域, 从而尽可能覆盖更多区域。

4) 边界:

$$r_{\text{boundary}}=\begin{cases}0 & C_{\text{sonar}}\geqslant 26\\ C_{\text{sonar}}/26-2 & 13\leqslant C_{\text{sonar}}<26\\ 5/26C_{\text{sonar}}-4 & C_{\text{sonar}}<13\end{cases}\tag{19}$$

其中, C_{sonar}为位于声呐覆盖区域内的网格数量。

5) 障碍物: 分为“避障”和“碰撞”两部分。“避障”部分为

$$r_{\text{obstacle}}^{(1)}=\begin{cases}0.8\cdot\Delta D_{\text{front}}/20-4 & D_{\text{front}}<500\\ 0.4\cdot\Delta D_{\text{front}}/20-2 & 500\leqslant D_{\text{front}}<1\,000\\ 0 & D_{\text{front}}\geqslant 1\,000\end{cases}\tag{20}$$

“碰撞”部分为

$$r_{\text{obstacle}}^{(2)}=\begin{cases}200 & \text{uncrashed}\\ -200 & \text{crashed}\end{cases}\tag{21}$$

总奖励为

$$r_{\text{obstacle}}=r_{\text{obstacle}}^{(1)}+r_{\text{obstacle}}^{(2)}\tag{22}$$

其中: ΔD_{front}表示位于 UUV 正前方障碍物的距离在当前步和上一步的差值, 若距离变小则差值小于 0, 产生负奖励信号, 当距离增大且大于一定幅度, 则产生正奖励信号; D_{front}表示上一步位于 UUV 正前方障碍物的距离, 根据 1.3 节可知, 声呐最远探测距离为 2 km。

该奖励鼓励 UUV 通过转向避免碰撞, 同时在任务顺利结束时给予大额正反馈, 因碰撞而结束时给予大额负反馈。

6) 速度: $r_{\text{speed}}=v/v_{\max}-0.5$。其中: v为 UUV 速度; $v_{\max}$为 UUV 最大速度。这一奖励是鼓励 UUV 以较大的速度航行。

7) 转向: $r_{\text{yaw}}=\begin{cases}-0.05 & |C_{\text{rudder}}|\geqslant 2\\ 0 & |C_{\text{rudder}}|<2\end{cases}$。其中, C_{rudder}为累计的偏航舵板偏转角度, 左右 2 个方向的偏转角度会抵消, 并且在舵板偏转角小于 0.5°时, 会将累计偏转值置零。该奖励旨在抑制 UUV 连续向同一侧转向。

2.4　多目标优化

以单目标优化和加权 SER 方法作为基础, 结合 Actor-Critic 框架与 PPO 算法, 设计了一种多目标建模与优化方法。

1) 单目标方法

通过特定的权重将 7 维的即时奖励信号加权求和, 得到标量的单步奖励r, 形成标量奖励信号, 文中使用了 2 组不同的权重。

a. 等权重:

$$r=\begin{pmatrix}r_{\text{time}}+r_{\text{sonar}}+r_{\text{explored}}+r_{\text{boundary}}+\\ r_{\text{obstacle}}+r_{\text{speed}}+r_{\text{yaw}}\end{pmatrix}\Big/7\tag{23}$$

b. 平均权重: 依赖于自适应多目标优化方法, 将该方法训练过程中的权重计算均值$\bar{\omega}_i(i=1,\cdots,7)$作为奖励$i$的权重, 即

$$r=\begin{pmatrix}\bar{\omega}_1 r_{\text{time}}+\bar{\omega}_2 r_{\text{sonar}}+\bar{\omega}_3 r_{\text{explored}}+\bar{\omega}_4 r_{\text{boundary}}+\\ \bar{\omega}_5 r_{\text{obstacle}}+\bar{\omega}_6 r_{\text{speed}}+\bar{\omega}_7 r_{\text{yaw}}\end{pmatrix}\tag{24}$$

将奖励转化为标量后, 使用单一价值函数模型估计给定观测下的期望回报。为了优化该模型, 采用式(3)作为价值函数模型的目标函数, 式(4)作为策略模型的目标函数。

2) 加权 SER 方法

在加权 SER 方法中, 使用 7 个结构相同的价值函数模型, 分别估计在给定状态下 7 个不同奖励的期望回报, 并使用式(3)作为每个价值函数模型的目标函数。多个价值函数模型的并行训练, 能够更好地学习和建模各个目标的回报。在优化策略模型时, 首先计算每个价值函数模型的动作优势值, 然后通过加权和将多个优势值合并为一个标量优势值$A(s,a)$(即式(6)中的标量化函数采用线性加权形式)。最终, 使用式(4)作为策略模型的目标函数, 同样使用了 2 组不同的权重。

a. 等权重:

$$A(s,a)=\begin{pmatrix}A_{\text{time}}+A_{\text{sonar}}+A_{\text{explored}}+A_{\text{boundary}}+\\ A_{\text{obstacle}}+A_{\text{speed}}+A_{\text{yaw}}\end{pmatrix}\Big/7\tag{25}$$

其中不同下标的A对应于不同奖励目标的动作优势。

b. 平均权重: 依赖于自适应多目标优化方法, 将该方法训练过程中的权重计算均值作为权重。

$$A(s,a)=\begin{pmatrix}\bar{\omega}_1 A_{\text{time}}+\bar{\omega}_2 A_{\text{sonar}}+\bar{\omega}_3 A_{\text{explored}}+\\ \bar{\omega}_4 A_{\text{boundary}}+\bar{\omega}_5 A_{\text{obstacle}}+\\ \bar{\omega}_6 A_{\text{speed}}+\bar{\omega}_7 A_{\text{yaw}}\end{pmatrix}\tag{26}$$

3) AdaptiveW-MO-PPO

PPO 算法的训练过程由采样与模型更新交替进行。AdaptiveW-MO-PPO 在每次采样后、更新前, 对多目标动作优势值的权重进行估计与调整。假设策略的小幅更新不会跳出当前解空间的局部区域, 从而使任务效用值与奖励回报之间的关系保持稳定。因此, 可采用局部线性近似, 并结合训练过程中自适应权重的调整, 以更好地逼近解空间的整体非线性关系。训练初始时, 各目标权重相等, 随后依据以下步骤自适应调整。

a. 收集任务完成评估结果

在训练过程中, 收集最近 128 次任务完成后的效用值和对应的多目标回报, 并将其存储在缓存中。如果缓存中的任务数不足 128 次, 则跳过本次多目标权重的更新; 如果超过 128 次, 则仅保留最近的 128 个任务评估结果。

b. 最小二乘法计算权重

在每次样本采集完成后, 将多目标回报作为自变量, 效用值作为因变量, 采用最小二乘法进行线性回归, 得到每个目标的回归系数作为对应的权重。为了消除尺度和偏置的影响, 回归前对多目标回报和效用值进行标准化处理。此处的权重可能为负值, 即允许策略朝着某些目标的反方向优化。线性回归的形式为

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\omega}^{\text{new}} + \varepsilon \tag{27}$$

式中: $\boldsymbol{y}$为标准化后的效用值; $\boldsymbol{X}$为标准化后的多目标回报矩阵; $\boldsymbol{\omega}^{\text{new}}$为回归系数(即目标的权重); ε为误差项。

c. 计算权重尺度系数

使用缓存中 128 个评估结果, 计算不同目标回报与效用值之间的线性相关系数r_i。然后, 利用 Softmax 函数对这些相关系数进行归一化, 得到不同目标的权重系数c_i, 具体计算公式为

$$\begin{cases} r_i = \dfrac{\sum\limits_{j=1}^{128}\left(x_i^{(j)} - \bar{x}_i\right)}{\sqrt{\sum\limits_{j=1}^{128}\left(x_i^{(j)} - \bar{x}_i\right)^2}\sqrt{\sum\limits_{j=1}^{128}\left(y^{(j)} - \bar{y}\right)^2}} \\ c_i = \dfrac{\mathrm{e}^{r_i}}{\sum\limits_{i=1}^{7}\mathrm{e}^{r_i}}, i = 1,2,\cdots,7 \end{cases} \tag{28}$$

此处利用 Softmax 的指数放大效应和平滑特性, 提高对高相关性目标的区分度, 同时避免低相关性目标的权重被削减至零, 从而减少优化过程中的剧烈波动。其中: $\bar{x}_i$表示 128 个样本的标准化后目标i回报的均值; $x_i^{(j)}$表示第j个样本的目标i的回报; $\bar{y}$表示标准化后效用的均值; $y^{(j)}$表示第j个样本的效用值。

d. 更新多目标权重

文中方法使用权重系数调整各目标的权重尺度, 使相关性较低或负相关的目标权重适当缩小, 从而减少其对优化过程的干扰, 使策略更聚焦于高相关性目标, 以提升优化稳定性和效率。同时, 为了使权重更新更加平滑, 避免剧烈波动, 采用指数滑动平均方法。权重ω_i更新的具体公式为

$$\omega_i \leftarrow c_i \cdot \omega_i^{\text{new}} \times 0.25 + \omega_i^{\text{old}} \times 0.75, i = 1,2,\cdots,7 \tag{29}$$

在每次采集样本后, 都会对目标的权重进行重新估计和更新, 使策略能够自适应地权衡不同的目标, 以最大化效用值。

2.5　模型结构

采用 Actor-Critic 框架, 包括策略模型和价值函数模型, 二者前端的特征提取器结构相同, 整体结构如图 2 所示。单目标方法使用 1 个价值函数模型, 多目标方法则使用 7 个。推理时, 仅计算策略模型及其特征提取器, 多目标方法的多个价值函数模型不会增加计算量和延时。实测在树莓派 4 开发板上单线程推理延时约 15 ms, 满足 UUV 控制的实时性要求。

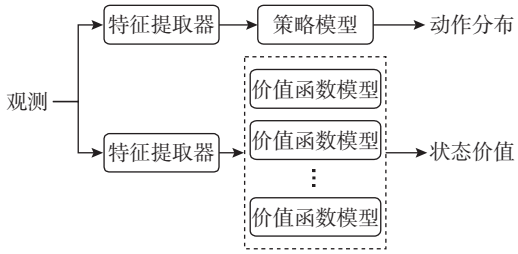


图 2　模型整体结构

Fig. 2　Overall structure of the model

特征提取器结构如图 3 所示。地图特征模块通过卷积层扩展通道并压缩空间维度, 状态特征模块则利用全连接层进行变换。随后, 两者经注意力机制融合, 并与状态特征输出及当前状态向量拼接。

策略模型结构如图 4 所示。输入特征经全连接层变换, 输出 3 维动作空间的高斯分布期望, 标

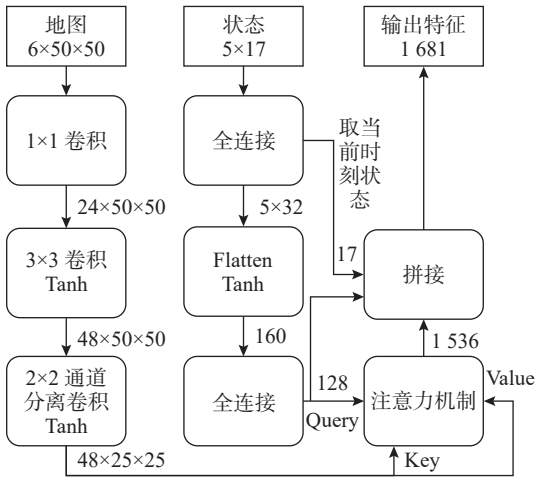


图 3　特征提取器结构

Fig. 3　Structure of the feature extractor

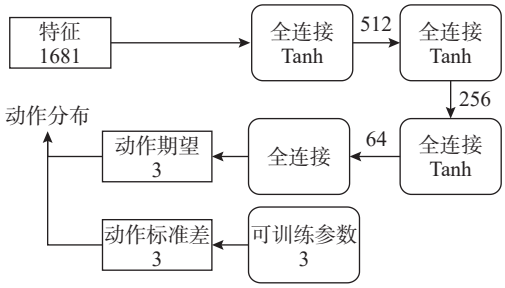


图 4　策略模型结构

Fig. 4　Structure of the policy model

准差为可训练参数(初始值 0.5)。为增强探索, 目标函数加入熵正则项, 防止标准差过快衰减。训练时, 从该分布随机采样动作; 测试时, 直接采用期望作为动作输出。

价值函数模型结构如图 5 所示, 与策略模型类似, 但最终输出期望回报。

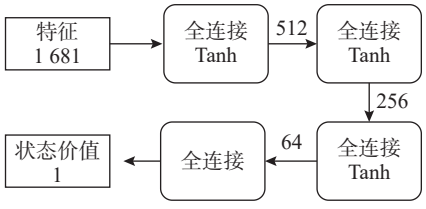


图 5　价值函数模型结构

Fig. 5　Structure of the value function model

3　实验验证

将 2.4 节中的 3 类方法与传统 A*方法在 1.4 节描述的 4 种任务场景中进行了测试, 涵盖多目标与单目标、固定权重与自适应权重的多组实验, 并对实验结果进行分析。同时, 对不同方法的路径规划实验结果进行示例分析, 并展示了 AdaptiveW-MO-PPO 中权重变化的曲线。

3.1　训练参数

在训练模型时, 使用了指数衰减的学习率, 其中初始学习率为1×10^{-4}, 最终学习率为 5×10^{-6}。强化学习中的奖励折扣系数设置为 $\gamma=0.99$, 采用广义优势估计(generalized advantage estimation, GAE)来计算动作的优势值, 参数$\lambda=0.95$。训练过程中, 熵正则项的系数设定为 0.01, 以促进策略的探索性。在样本采集阶段, 采用 32 个并行仿真环境, 每个环境采集 4 096 个时间步的样本。每次采样完成后, 对策略模型进行 10 轮训练, 批大小为 64。

实验首先对 AdaptiveW-MO-PPO 进行训练, 并统计训练完成后的均值, 具体结果如表 1 所示。对于采用平均权重的 AveragedW-MO-PPO 和 AveragedW-SO-PPO, 其时间、声呐、转向和边界相关的优化目标权重数值极低, 训练过程中几乎未被考虑, 而探测、障碍物及速度相关的优化目标因权重较大, 成为主导因素。

表 1　各目标平均权重

Table 1　Average weight of each objective

目标	平均权重
时间	−0.013 359
声呐	−0.002 343
探测	0.301 260
边界	0.020 247
障碍物	0.307 023
速度	0.194 401
转向	−0.007 376

3.2　训练结果

模型在随机生成的矩形障碍物场景中进行训练。为提高训练效率, 采用课程学习策略进行分阶段训练。在第 1 阶段, 场景设定较为简单, 仅包含 1 个障碍物。当 UUV 的效用值首次达到 0.4 时, 表明其已初步掌握了全覆盖路径规划能力, 此时进入第 2 阶段, 提高任务难度, 将障碍物数量随机设置为 1~5 个。通过这种两阶段训练策略, 模型能够在复杂度逐步提升的过程中更有效地学习全覆盖路径规划策略。

在整个训练过程中, 为模拟已知和未知障碍物环境, 设置了 50% 的概率, 使 UUV 在初始化时获

得障碍物的形状和位置信息, 并据此构建相应的环境地图, 从而增强模型在不同环境不确定性条件下的适应能力。

图 6 展示了 5 种方法在训练过程中多个评价指标及效用值的变化趋势, 文中图表中上下箭头分别表示数值越大或越小时性能更优(下文同)。横轴表示训练过程中 UUV 与仿真环境交互的总时间步, 纵轴为每次模型更新后, 通过 32 次随机测试统计得到的评价指标和效用值的平均值。曲线背景的颜色填充区域表示数值的波动区间。由于 UUV 在训练初期尚未具备全覆盖路径规划能力, 因此仅记录了覆盖率、碰撞率、轨迹长度、能耗率、信息延时和效用值的变化曲线, 且在统计时未对 90% 覆盖率设置限制。训练过程中, 策略的动作通过从输出的动作分布中随机采样生成, 以促进策略多样性探索。

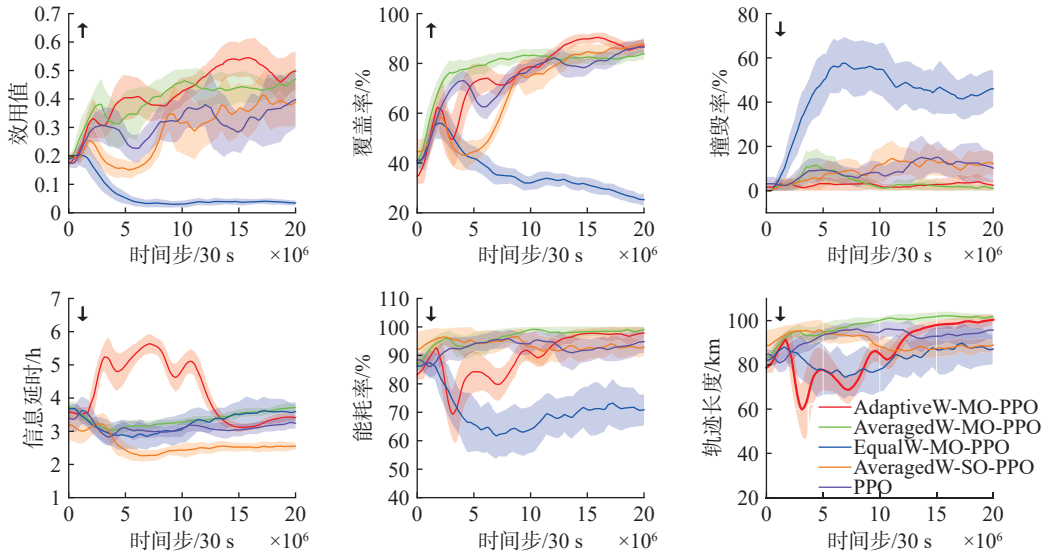


图 6　训练中各指标变化曲线

Fig. 6　Curves of indicators during training

实验结果表明, 在训练过程中, 文中所提的 AdaptiveW-MO-PPO 的关键评价指标表现出显著优势, 其效用值始终保持在相对最高水平, 同时覆盖率和碰撞率等核心指标的表现也优于其他方法。

值得注意的是, AveragedW-MO-PPO 通过提取 AdaptiveW-MO-PPO 训练过程中动态权重的均值, 作为多目标优化的固定权重, 展现出次优的多目标平衡能力。这说明自适应权重机制能够有效刻画不同优化目标间的相互关系, 实现更合理的权重分配。在单目标框架下, AveragedW-SO-PPO 和 PPO 的实验性能差异不显著, 推测原因可能在于多目标奖励的直接线性叠加会导致量纲较大的目标项占据主导地位, 从而削弱了权重对策略优化的作用。相比之下, EqualW-MO-PPO 由于对所有目标采用固定且相等的权重, 导致在训练初期过度关注次要的优化目标, 无法有效地探索, 使策略难以有效完成全覆盖路径规划任务, 效用值显著低于其他方法。

通过比较训练过程的效用值曲线可知, 各方法的实验效果呈现以下层级关系: AdaptiveW-MO-PPO > AveragedW-MO-PPO > AveragedW-SO-PPO ≈ PPO > EqualW-MO-PPO。该排序结果证实了动态权重估计机制与多目标建模策略在复杂路径规划任务中具有显著的协同优化效果。

3.3　测试结果

在测试阶段, 构建了 40 个初始任务场景, 每个场景包含 1~5 个障碍物。其中, 20 个场景的障碍物为规则矩形, 其余场景障碍物为不规则多边形。基于障碍物信息的已知或未知情况, 进一步将任务场景扩展至 80 个, 并分别对应 1.4 节中定义的 4 类场景(每类 20 个)。任务场景示例见图 7 和图 8。

为对比传统方法, 基于 1.5 节的网格地图, 进一步缩小尺寸, 并采用 A*全覆盖路径搜索算法离线计算完整路径, 再通过简单的 UUV 导航控制按

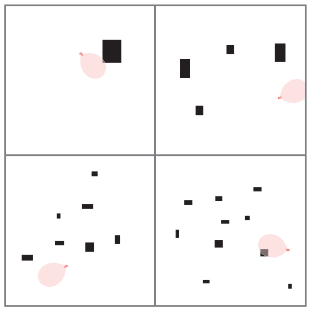


图 7　矩形障碍物场景示例

Fig. 7　Examples of rectangular obstacle scenarios

图 8　不规则障碍物场景示例

Fig. 8　Examples of irregular obstacle scenarios

预设轨迹执行任务。由于该方法无法适应未知障碍物场景, 因此未在障碍物信息未知的环境中测试。同时, 对训练完成的 5 种强化学习方法(Adaptive W-MO-PPO、AveragedW-MO-PPO、EqualW-MO-PPO、AveragedW-SO-PPO 和标准 PPO)进行了系统测试, 并在覆盖率达到 99% 时视为任务完成, 结束测试。

为确保不同方法在性能指标上的公平性, 测试阶段采用与训练阶段不同的统计策略: 当覆盖率达到 90% 时, 记录轨迹长度、能耗率、信息延时和任务完成时间等性能指标。各项评价指标的具体定义详见 1.6 节。

测试结果如表 2 和表 3 所示, 展示了不同方法在矩形障碍物和不规则多边形障碍物场景下的性

表 2　矩形障碍物场景下测试结果

Table 2　Test results in rectangular obstacle scenarios

方法	障碍物是否已知	覆盖率/%↑	碰撞率/%↓	轨迹长度/km↓	能耗率/%↓	信息延时/h↓	效用值↑	任务完成率/%↑	任务完成时间/h↓
AdaptiveW-MO-PPO	是	**96.66**	0	**65.11**	**70.23**	1.73	**0.71**	**100**	**4.20**
	否	**97.15**	0	**62.42**	**67.4**	**1.69**	**0.72**	**100**	**4.02**
AveragedW-MO-PPO	是	94.30	10	67.12	71.45	**1.72**	0.62	85	4.32
	否	94.74	10	67.37	71.89	1.75	0.61	85	4.38
EqualW-MO-PPO	是	55.18	45	—	—	—	0.15	0	—
	否	40.86	65	—	—	—	0.09	0	—
AveragedW-SO-PPO	是	79.55	10	78.36	93.46	2.38	0.44	15	6.40
	否	76.66	15	75.30	89.37	2.41	0.43	30	6.06
PPO	是	82.30	0	69.61	80.01	2.00	0.50	40	5.11
	否	85.42	5	73.11	84.10	2.05	0.51	45	5.20
A*	是	93.20	0	90.61	79.99	2.12	0.65	95	5.00

表 3　不规则障碍物场景下测试结果

Table 3　Test results in irregular obstacle scenarios

方法	障碍物是否已知	覆盖率/%↑	碰撞率/%↓	轨迹长度/km↓	能耗率/%↓	信息延时/h↓	效用值↑	任务完成率/%↑	任务完成时间/h↓
AdaptiveW-MO-PPO	是	**96.69**	0	**65.92**	**71.62**	**1.65**	**0.71**	**100**	**4.29**
	否	**96.84**	0	**68.47**	73.70	1.81	**0.71**	**100**	**4.46**
AveragedW-MO-PPO	是	94.49	0	69.65	74.43	1.83	0.67	90	4.53
	否	95.02	0	68.73	**73.27**	**1.79**	0.68	90	4.48
EqualW-MO-PPO	是	56.11	25	—	—	—	0.19	0	—
	否	38.17	75	—	—	—	0.06	0	—
AveragedW-SO-PPO	是	73.96	0	83.21	98.16	2.45	0.43	5	6.58
	否	74.17	5	73.33	86.84	2.47	0.42	20	5.86
PPO	是	76.88	0	73.80	84.99	2.04	0.44	15	5.54
	否	79.92	0	73.07	83.95	2.12	0.47	15	5.77
A*	是	92.09	0	88.55	78.18	2.07	0.63	85	4.89

能评价。最优值已用加粗标注。需要注意的是, EqualW-MO-PPO 在所有测试中均未能达到 90% 的覆盖率, 按照统计规则, 其在轨迹长度、能耗率、信息延时和任务完成时间等指标上的结果为空值。

实验结果表明, A*方法在面对不规则障碍物场景时, 相较于矩形障碍物场景, 其效用值平均下降 3%, 且在覆盖率、任务完成率等关键指标上均有所损失。相比之下, 文中提出的 AdaptiveW-MO-PPO 在两类场景以及障碍物信息已知和未知的任务设置下, 效用值、覆盖率、碰撞率和任务完成率等指标几乎保持稳定, 表明其相较传统方法具有更强的适应性和鲁棒性。与基线方法 PPO 相比, AdaptiveW-MO-PPO 在所有指标上均表现出显著优势。

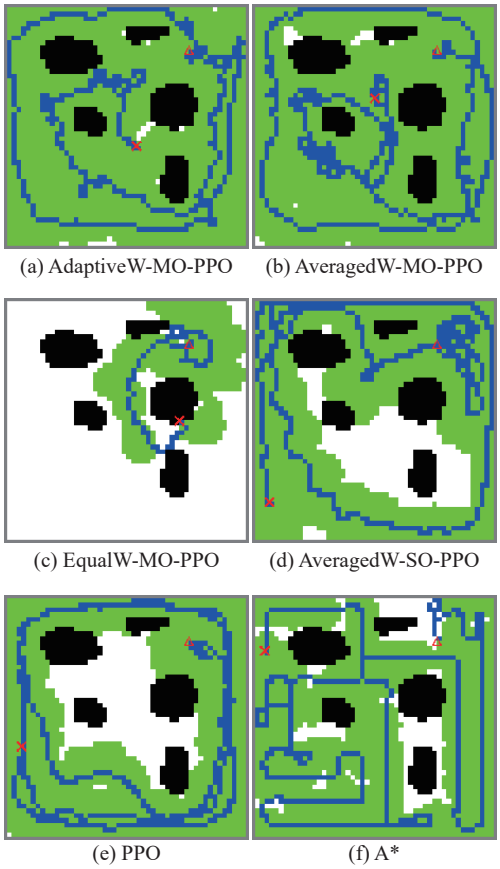
进一步对比 AveragedW-MO-PPO、AveragedW-SO-PPO 和 PPO 的实验结果, 验证了多目标建模在复杂全覆盖路径规划任务中的积极作用。通过独立建模不同目标的奖励回报分布, UUV 能够更准确地估计在不同观测条件下的多目标回报, 从而实现更精确的策略优化。

针对不同多目标权重机制的对比实验, AdaptiveW-MO-PPO、AveragedW-MO-PPO 和 EqualW-MO-PPO 的效用值呈现依次递减趋势, 进一步验证了文中提出的动态权重估计与更新机制的有效性。该机制能够在训练过程中自适应地平衡不同优化目标的相对重要性及其相互关系, 从而显著提升 UUV 的整体性能。

综合所有方法的对比结果, AdaptiveW-MO-PPO 在覆盖率、碰撞率、轨迹长度、能耗率、信息延时、效用值、任务完成率和任务完成时间等多个关键评价指标上均优于基线方法, 展现出显著的综合性能优势。

3.4　实验示例

图 9 展示了从 20 个不规则障碍物场景中选取的 1 组 UUV 路径轨迹及探测覆盖结果。其中, 绿色区域表示已探测覆盖区域, 白色为未探测的自由空间, 黑色为障碍物, 蓝色线条为 UUV 轨迹, 红三角和红叉分别表示起点和终点。AdaptiveW-MO-PPO 生成的轨迹较平滑, 重复较少; AveragedW-MO-PPO 与其相似, 但在中部存在较多重复绕行; EqualW-MO-PPO 发生碰撞, 未能有效避障; AveragedW-SO-PPO“过度避障”, 导致障碍物密集区域探测不足, 并在右上角出现重复绕行; PPO 方法表现类似, 因其同样直接加和奖励; A*算法生成的轨迹较规整, 但由于搜索网格构建方式, 部分自由空间未被覆盖, 并且依赖 UUV 的导航控制, 在障碍物附近以及路径拐弯处存在较多未探测区域。



(a) AdaptiveW-MO-PPO　(b) AveragedW-MO-PPO

(c) EqualW-MO-PPO　(d) AveragedW-SO-PPO

(e) PPO　(f) A*

图 9　路径规划结果示例

Fig. 9　Examples of path planning results

表 4 列出了该示例中各方法的性能指标。仅 AdaptiveW-MO-PPO 和 AveragedW-MO-PPO 的覆

表 4　示例的性能指标

Table 4　Performance indicators of the example

方法	覆盖率/%↑	是否碰撞	轨迹长度/km↓	能耗率/%↓	信息延时/h↓	效用值↑	任务完成时间/h↓
AdaptiveW-MO-PPO	**99.06**	否	**76.69**	**71.16**	**1.43**	**0.79**	**4.28**
AveragedW-MO-PPO	98.83	否	94.87	83.49	2.24	0.75	5.14
EqualW-MO-PPO	15.38	是	—	—	—	0	—
AveragedW-SO-PPO	80.62	否	84.36	—	—	0.46	—
PPO	76.62	否	87.40	—	—	0.41	—
A*	86.38	否	79.69	—	—	0.58	—

盖率达到 90%, 成功完成任务, 且前者在各指标上最优, 其余方法则因能量耗尽或路径结束前未能完成任务。

3.5　权重变化分析

图 10 展示了训练过程中自适应多目标方法的动态权重变化曲线。从中可以观察到, 在训练初期, UUV 尚未掌握全覆盖路径规划能力, 此时时间、声呐、转向和边界相关的奖励优化目标权重为负, 而探测、障碍物和速度等与覆盖率高度相关的奖励优化目标权重为正, 从而促进任务覆盖率的快速提升。

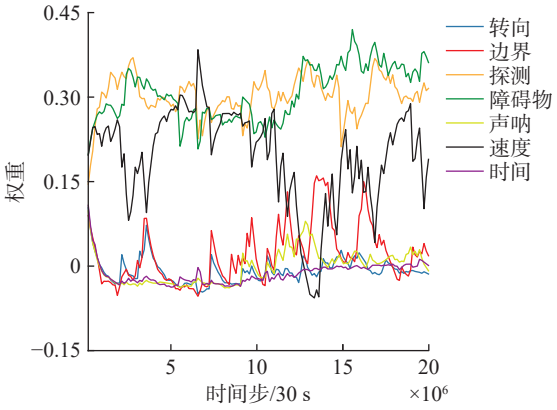


图 10　训练中自适应权重变化曲线

Fig. 10　Adaptive weight curves during training

当 UUV 逐步具备全覆盖路径规划能力后, 时间、声呐、边界及转向等奖励优化目标权重逐渐增加, 从而提升能耗、路径长度等次要指标, 以进一步增加任务效用值。这种动态权重调整机制使 UUV 在不同训练阶段优化的重点不断变化, 避免在早期过度关注次要指标而影响任务完成效果。此外, 在整个训练过程中, 覆盖率和碰撞率始终是关键指标, 因此与之相关的探测、障碍物奖励优化目标的权重始终保持较高水平。

4　结束语

文中提出了一种自适应多目标优化的 UUV 全覆盖路径规划方法, 并在包含障碍物和洋流等环境因素的二维仿真环境中进行验证。该方法结合了 PPO 强化学习算法与动态权重调节机制, 使 UUV 能够在训练过程中根据不同优化目标的重要性自适应调整权重, 以优化覆盖率、碰撞率、轨迹长度、能耗、信息延时、任务完成率及任务完成时间等关键指标。实验结果表明, 与标准 PPO 和传统 A*方法相比, 文中方法在多种测试环境下均表现出更优的规划性能, 尤其在两类障碍物已知场景中, 覆盖率提升约 4.03%, 任务完成率提高约 10%, 效用值相对提升约 10.96%, 任务完成时间相对减少约 14.13%, 轨迹长度相对减少 26.85%, 能耗相对减少 10.3%, 信息延时相对减少 19.34%。此外, 文中方法在不同障碍物形状、分布及随机洋流条件下均展现出较强的适应性和鲁棒性。

进一步的对比实验表明, 多目标优化建模能够有效提升 UUV 的路径规划能力, 使其在复杂环境下保持稳定性能。特别是, 相较于 AveragedW-MO-PPO、EqualW-SO-PPO 等方法, AdaptiveW-MO-PPO 在不同优化目标之间实现了更精准的动态权衡, 进一步验证了所提方法的有效性。

尽管文中所提方法在全覆盖路径规划方面取得了一定进展, 仍存在一些值得进一步研究的方向。首先, 现有研究基于二维仿真环境, 未来可进一步扩展至三维复杂水下环境, 考虑动态障碍物和非均匀流场等真实海洋条件, 以增强算法的适用性。其次, 可探索结合深度模仿学习或自监督学习, 以提升 UUV 的泛化能力, 使其在未见环境中具备更优的适应性。此外, 该方法可进一步推广至多 UUV 协作任务, 以提高任务效率和覆盖质量, 为实际水下探测与作业提供更具实用价值的路径规划解决方案。

参考文献:

[1] 陈昭, 丁一杰, 张治强. 无人潜航器发展历程及运用优势研究[J]. 舰船科学技术, 2024, 46(23): 98-102.
CHEN Z, DING Y J, ZHANG Z Q. Research on the development history and application advantages of unmanned underwater vehicle[J]. Ship Science and Technology, 2024, 46(23): 98-102.

[2] 延远航. 无人水下航行器运动控制研究[D]. 太原: 中北大学, 2024.

[3] 张翔鸢, 花吉. 国外超大型无人潜航器发展与运用研究[J]. 中国舰船研究, 2024, 19(5): 17-27.
ZHANG X Y, HUA J. Study on the development and application of foreign extra-largeunmanned underwater vehicles[J]. Chinese Journal of Ship Research, 2024, 19(5): 17-27.

[4] CHENG C, SHA Q, HE B, et al. Path planning and obstacle avoidance for AUV: A review[J]. Ocean Engineering, 2021, 235: 109355.

[5] ZENG Z, SAMMUT K, LIAN L, et al. A comparison of optimization techniques for AUV path planning in environments with ocean currents[J]. Robotics and Autonomous Systems, 2016, 82: 61-72.

[6] REPOULIAS F, PAPADOPOULOS E. Planar trajectory planning and tracking control design for underactuated AUVs[J]. Ocean Engineering, 2007, 34(11-12): 1650-1667.

[7] YU H, WANG Y. Multi-objective AUV path planning in large complex battlefield environments[C]//2014 Seventh International Symposium on Computational Intelligence and Design. Hangzhou, China: IEEE, 2014: 345-348.

[8] TAN C S, MOHD-MOKHTAR R, ARSHAD M R. A comprehensive review of coverage path planning in robotics using classical and heuristic algorithms[J]. IEEE Access, 2021, 9: 119310-42.

[9] GAMMELL J D, SRINIVASA S S, BARFOOT T D. Informed RRT*: Optimal sampling-based path planning focused via direct sampling of an admissible ellipsoidal heuristic[C]//2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. Chicago, USA: IEEE, 2014: 2997-3004.

[10] TORRES M, PELTA D A, VERDEGAY J L, et al. Coverage path planning with unmanned aerial vehicles for 3D terrain reconstruction[J]. Expert Systems with Applications, 2016, 55: 441-451.

[11] GABRIELY Y, RIMON E. Spanning-tree based coverage of continuous areas by a mobile robot[J]. Annals of Mathematics and Artificial Intelligence, 2001, 31: 77-98.

[12] HUANG W H. Optimal line-sweep-based decompositions for coverage algorithms[C]//Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation. Seoul, Korea(South): IEEE, 2001, 1: 27-32.

[13] KYAW P T, PAING A, THU T T, et al. Coverage path planning for decomposition reconfigurable grid-maps using deep reinforcement learning based travelling salesman problem[J]. IEEE Access, 2020, 8: 225945-56.

[14] HEYDARI J, SAHA O, GANAPATHY V. Reinforcement learning-based coverage path planning with implicit cellular decomposition[EB/OL]. [2025-4-14]. https://arxiv.org/abs/2110.09018.

[15] AI B, JIA M, XU H, et al. Coverage path planning for maritime search and rescue using reinforcement learning [J]. Ocean Engineering, 2021, 241: 110098.

[16] RÜCKIN J, JIN L, POPOVIĆ M. Adaptive informative path planning using deep reinforcement learning for UAV-based active sensing[C]//2022 International Conference on Robotics and Automation. Philadelphia, USA: IEEE, 2022: 4473-4479.

[17] ZHAO Y, SUN P, LIM C G. The simulation of adaptive coverage path planning policy for an underwater desilting robot using deep reinforcement learning[C]//International Conference on Robot Intelligence Technology and Applications. Cham, Switzerland: Springer International Publishing, 2022: 68-75.

[18] XING B, WANG X, YANG L, et al. An algorithm of complete coverage path planning for unmanned surface vehicle based on reinforcement learning[J]. Journal of Marine Science and Engineering, 2023, 11(3): 645.

[19] JONNARTH A, ZHAO J, FELSBERG M. Learning coverage paths in unknown environments with deep reinforcement learning[C]//International Conference on Machine Learning. Vienna, Austria: PMLR, 2024: 22491-508.

[20] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of Actor-Critic reinforcement learning: Standard and natural policy gradients[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C(Applications and Reviews), 2012, 42(6): 1291-307.

[21] VAN MOFFAERT K, DRUGAN M M, NOWÉ A. Scalarized multi-objective reinforcement learning: Novel design techniques[C]//2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. Singapore: IEEE, 2013: 191-199.

[22] REYMOND M, HAYES C F, STECKELMACHER D, et al. Actor-Critic multi-objective reinforcement learning for non-linear utility functions[J]. Autonomous Agents and Multi-Agent Systems, 2023, 37(2): 23.

[23] FOSSEN T I. Handbook of marine craft hydrodynamics and motion control[M]. Hoboken, USA: John Willy & Sons Ltd, 2011.

[24] WANG Z, DU J, JIANG C, et al. Task scheduling for distributed AUV network target hunting and searching: An energy-efficient AoI-aware DMAPPO approach[J]. IEEE Internet of Things Journal, 2022, 10(9): 8271-85.

(责任编辑: 许 妍)