控制与决策
Control and Decision

面向多动态目标基于拍卖机制与MASAC的AUV协同围捕

谢地杰, 李敏, 曾祥光, 任文哲, 张滔, 彭倍

引用本文:
谢地杰, 李敏, 曾祥光, 等. 面向多动态目标基于拍卖机制与MASAC的AUV协同围捕[J]. 控制与决策, 2026, 41(5): 1229-1241.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2025.0710

面向多动态目标基于拍卖机制与 MASAC 的 AUV 协同围捕

谢地杰[1]，李 敏[1†]，曾祥光[1]，任文哲[1]，张 滔[1]，彭 倍[2]

(1. 西南交通大学 机械工程学院，成都 610031；2. 电子科技大学 机械与电气工程学院，成都 611731)

摘 要: 针对多动态目标的自主水下航行器集群协同围捕决策与控制问题, 提出一种融合拍卖机制与多智能体深度强化学习的围捕算法. 该方法将围捕任务分解为目标分配和运动控制两个阶段: 首先, 基于最优控制理论中的配点法, 综合考虑围捕态势、最短时间和最低能耗等优化目标, 生成训练数据与竞标值标签, 并利用监督学习训练拍卖神经网络, 实现了自主水下航行器的实时目标分配; 接着, 构建分配后的个体状态空间, 设计多目标围捕奖励函数, 采用多智能体柔性演员-评论家算法, 优化了围捕策略. 高效、自适应的拍卖算法确保了动态复杂环境下的快速目标分配, 多智能体强化学习则提升了群体的协同控制快速响应能力. 最后, 开展不同场景中的围捕实验. 实验结果表明, 所提方法能够显著提高围捕策略的表现效果, 在应对 2、3 和 4 个动态目标时, 平均围捕成功率分别为 79.04%、89.78% 和 90.43%, 相较于基线方法, 分别提升了 48.41%、54.00% 和 53.93%, 即所提算法在处理不同规模围捕任务时均具有更好的效果.

关键词: 自主水下航行器；协同围捕；多智能体深度强化学习；多动态目标；拍卖算法；目标分配；运动控制

中图分类号: TP249 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2025.0710

引用格式: 谢地杰, 李敏, 曾祥光, 等. 面向多动态目标基于拍卖机制与 MASAC 的 AUV 协同围捕 [J]. 控制与决策, 2026, 41(5): 1229-1241.

AUV cooperative hunting based on auction mechanism and MASAC for multiple dynamic targets

XIE Di-jie[1]，*LI Min*[1†]，*ZENG Xiang-guang*[1]，*REN Wen-zhe*[1]，*ZHANG Tao*[1]，*PENG Bei*[2]

(1. School of Mechanical Engineering， Southwest Jiaotong University， Chengdu 610031， China；2. School of Mechanical and Electrical Engineering， University of Electronic Science and Technology of China， Chengdu 611731，China)

Abstract: To address the decision-making and control problem of collaborative hunting of autonomous underwater vehicle (AUV) swarms with multiple dynamic targets, this paper proposes a hunting algorithm integrating auction mechanisms and multi-agent deep reinforcement learning. The method decomposes the hunting task into two stages: target allocation and motion control. Firstly, based on the point-matching method from optimal control theory, training data and bid value labels are generated, taking into account optimization objectives such as hunting posture, minimum time, and minimum energy consumption. The auction neural network is trained using supervised learning, achieving real-time target allocation for the AUVs. Next, the allocated individual state space is constructed, a multi-target hunting reward function is designed, and a multi-agent soft actor-critic algorithm is employed to optimize the hunting strategy. The efficient and adaptive auction algorithm ensures rapid target allocation in dynamic and complex environments, while multi-agent reinforcement learning enhances the swarm's rapid response capability in collaborative control. Finally, hunting experiments are conducted in various scenarios. Experimental results show that the proposed method can significantly improve the performance of the hunting strategy. When dealing with 2, 3 and 4 dynamic targets, the average roundup success rates are 79.04%, 89.78% and 90.43%, respectively. Compared with the baseline method, they are increased by 48.41%, 54.00% and 53.93%, respectively. In other words, the proposed algorithm has better performance in handling hunting tasks of different scales.

收稿日期: 2025-07-05；**录用日期:** 2025-12-04.

基金项目: 四川省科技厅重点研发计划项目 (2023YFG0285)；国家自然科学基金项目 (52075456).

责任编委: 关新平.

†通信作者. E-mail: liminfish008@163.com.

Keywords: AUV； collaborative hunting； multi-agent deep reinforcement learning； multiple dynamic targets； auction algorithm；target allocation；motion control

0 引 言

近年来, 我国的海洋权益受到日益严重的威胁, 周边国家不断在东海、南海上因岛屿问题挑衅我国权威, 试图削弱我国的海域管控能力, 导致局势日益复杂. 受限于自然环境与人体机能, 当前海洋探测技术仍待突破. 其中, 自主水下航行器 (AUV) 作为深海探测的关键装备, 可高效执行科考与军事任务. 但单一 AUV 难以完成复杂任务, 如对周边敌对潜艇进行侦察、拦截、占位攻击、包围攻击[1], 因此需构建多 AUV 协同对抗博弈系统. 通过模块化设计协同导航, 利用多机协作扩展作业范围, 能显著提升复杂任务执行效率, 其研究已引起全球学术界的广泛关注[2]. 然而, 对于协同围捕问题, 现有的研究大多停留在单一目标, 这是远远不够的. 综上, 对围捕多动态目标问题的研究是必要的.

部分学者已经意识到围捕多动态目标问题的重要性, 并展开了相关的研究. Wang 等[3] 提出了一种基于分布式拍卖的多 AUV 系统自适应目标分配算法, 在保证收益的同时显著降低了计算复杂度. 李海峰等[4] 结合 MAPPO 与任务重分配网络实现了多无人机动态任务分配, 提升了高对抗场景下的协同效率. Dong 等[5] 提出了一种基于改进K-means 和拍卖算法的多目标动态狩猎策略, 将系统分解为多个单目标围捕子系统, 并使用拍卖算法进行任务分配, 通过仿真验证了其有效性. 对于大规模 UAV 群体, Wang 等[6] 提出了一种两阶段贪心拍卖算法, 用于目标分配, 适用于实时性要求高的场景. 白小山等[7] 基于改进最小边际代价算法实现了多 USV 多 AUV 任务分配, 优化了总旅行距离. Okumura 等[8] 提出了一种 TSWAP 算法, 用于解决多智能体路径规划中的目标分配和路径规划问题. 该算法支持离线和在线场景, 离线时具有高效性, 在线时具有延迟容忍性. 然而, 上述方法主要针对静态任务分配, 且较少考虑到围捕双方的速度, 我方消耗的时间、能量等, 控制器固定, 无法实现实时适应环境分配, 难以适应多动态目标的协同围捕.

在人工智能快速发展背景下, 深度强化学习作为新兴范式, 在自主潜航器运动控制中展现出强大能力[9], 而多智能体强化学习 (MARL) 为集群围捕运动控制提供了新方法[10-12].

部分学者已经将深度强化学习运用到协同围捕当中. Han 等[13] 针对多智能体多目标追逐问题, 提出了一种改进的算法 DAO-MATD3, 并通过实验验证了其在捕获多个移动目标时能提升任务效率. Xia 等[14] 提出了一种基于 PPO 的围捕算法, 为了处理动态变化的观测特征维度, 设计了一个结合列最大池化和列平均池化两种特征压缩方法的特征嵌入块, 并采用集中训练、分散执行的框架进行训练, 验证了其算法的优越性. Awheda 等[15] 提出了一种结合卡尔曼滤波的模糊 Actor-Critic 算法, 可对逃逸者位置进行预测, 并依据预测信息自适应调整算法参数, 使方法可适用于当前环境和训练环境具有较大差异的围捕场景. Wang 等[16] 采用分布式多智能体近端策略优化方法, 解决了分布式自主水下航行器网络在目标追捕与环境搜索中的多目标调度问题. Cao 等[17] 提出了一种基于模糊势场的分层强化学习围捕方法, 将任务分解为搜索与捕获子任务, 利用势场引导路径规划, 引入模糊算法来提高围捕机器人轨迹的平滑度. 然而, 以上方法有些依赖预定义的分配策略, 或是规定逃逸者速度比围捕者速度低, 亦或是只适应于特定场景, 泛化能力较弱.

综上所述, 在多 AUV 协同围捕多动态目标的现有研究方面仍存在不足, 主要体现在以下几个方面: 1) 大多数研究聚焦于单一目标围捕, 缺乏对多目标围捕的研究; 2) 目标分配策略多基于简单距离或静态规则, 难以适应复杂动态的环境; 3) 部分方法假设围捕者速度高于逃逸者, 限制了算法的通用性; 4) 缺乏结合实时分配与协同运动控制的综合优化框架.

受到上述文献的启发, 本文提出一种基于拍卖算法与多智能体强化学习的协同围捕方法. 拍卖算法通过监督学习所训练的神经网络产生的竞标值快速竞标, 实现了多目标的实时分配, 适应动态环境下的复杂态势. 多智能体柔性演员-评论家算法则通过集中式训练与分布式执行优化了 AUV 的协同运动控制策略. 最后通过交互式训练验证了所提出算法在多目标围捕任务中的高效性与可扩展性.

1 问题描述与建模

1.1 AUV 群围捕多目标问题描述

在大小为$L_1 \times L_2$的作战海域中, 我方拥有M艘 AUV, 敌方有N个动态活动目标, 双方在有界区域内开展围捕-反围捕对抗. 我方通过规划 AUV 群的控制策略在有界区域内有效地捕获所有动态目标, 而动态目标则需避免被捕获到.

首先对我方 AUV 群根据当前态势进行实时分

组, 然后由控制算法生成各 AUV 的动作开展围捕, 围捕场景如图 1 所示. AUV 群中的所有 AUV 是同构的, 且具有相同的运动学模型, 围捕目标时需要考虑围捕距离、安全距离和围捕态势等信息. 其中: $A_i(i=1,2,\ldots,M)$为第i艘 AUV; $T_j(j=1,2,\ldots,N)$为第j个动态目标; $d_{A_iA_k}$为第i艘 AUV 和第k艘 AUV 的距离; v_{A_i}和v_{T_j}分别为A_i和T_j的速度; $\theta_{A_iT_jA_k}$为第i艘和第k艘 AUV 相对于第j个动态目标所形成的包围角, 即两艘 AUV 相对于目标所形成的夹角; r_c和r_s分别为围捕半径和安全半径, 围捕半径是 AUV 与被围捕目标之间的最大围捕距离, 安全半径是 AUV 与队友或目标之间的最短安全距离. 在 AUV 群进行围捕时需避免相互碰撞以及与目标的碰撞.

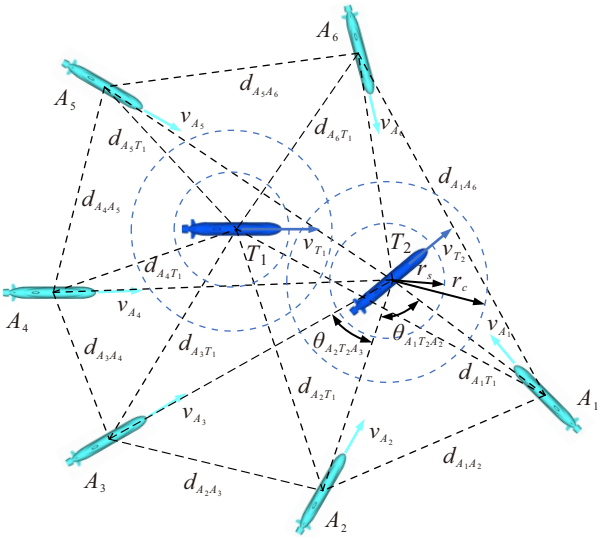

图1　AUV 群围捕多目标示意图

如图 2 所示, 当存在p艘以上 (包含p艘) AUV 与所围捕目标的距离小于围捕半径并大于安全半径, 且所形成分组内任意的包围角均不大于θ_{cap}时, 视为围捕成功.

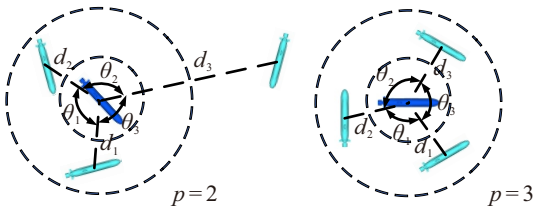

图2　AUV 群围捕成功示意图

此外, 在多 AUV 分配多动态目标时可能会出现多种分配结果, 从而导致不能尽量地多角度围捕多目标, 而导致目标逃离. 因此, 当M小于N时, 每个动态目标最多分配一个 AUV; 当M大于N时, 每个目标分配的 AUV 数量为M/N向下取整到M/N向上取整.

1.2　AUV 的运动学模型

如图 3 所示, 以具备轴向推进器和侧向推进器的 AUV 为运动学原型建立运动学方程, 其平面运动学模型如下所示:

$$\begin{cases} \dot{v}_x = a_\alpha \cos\varphi_i + a_\beta \sin\varphi_i, \\ \dot{v}_y = a_\alpha \sin\varphi_i + a_\beta \cos\varphi_i, \\ \dot{x}_i = v_x, \\ \dot{y}_i = v_y. \end{cases} \tag{1}$$

其中: a_α和a_β分别为轴向推进器和侧向推进器产生的轴向加速度和侧向加速度, φ_i为 AUV 的艏向角, (x_i, y_i)为 AUV 的位置坐标, (v_x, v_y)为 AUV 的速度矢量.

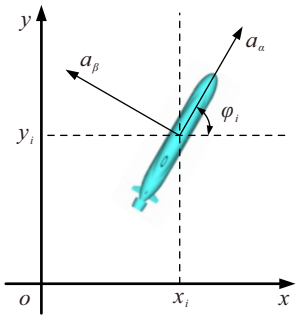

图3　AUV 各运动矢量示意图

1.3　动态目标的运动学模型

动态目标的逃逸方式采用人工势场法. 先计算动态目标j与 AUV 群、友方和边界的排斥加速度a_m, 然后算出动态目标的总加速度a_j, 如下所示:

$$\begin{aligned} a_m &= \sum_{i=1}^{N_m} \beta_m \mathrm{e}^{-\lambda_m d_{ij}} \cdot \frac{c_i - c_j}{d_{ij}}, \\ a_j &= a_1 + a_2 + a_3. \end{aligned} \tag{2}$$

其中: $m=1,2,3$分别代表 3 种不同的类别, 即 AUV 群、友方和边界, N_m是各类别的数量, β_m是排斥力系数, λ_m是指数衰减系数, $c_i=(x_i,y_i)$是 AUV$_i$的位置坐标, $c_j=(x_j,y_j)$是动态目标j的位置坐标, d_{ij}是 AUV$_i$与动态目标j的距离.

2　基于拍卖神经网络的多目标分配算法设计

在目标分配问题中, 应该综合考虑当前态势 (距离、速度)、时间成本以及能耗等方面的因素. 针对围捕问题, 传统的分配方法通常只考虑距离成本, 忽略了双方其他态势以及时间成本和能耗带来的影响, 故而在动态环境中的分配结果并不理想. 而拍卖机制有着计算高效和适应性强等特点, 神经网络具备强大的非线性拟合能力与多源信息融合能力. 因此, 本节设计一种基于拍卖神经网络的多目标分配算法, 在综合考虑各因素的同时实时高效地实现目标分配.

2.1 多目标分配问题的模型构建

多目标分配问题实际上是一个多目标优化问题, 对于协同围捕问题中的分配, 即将N个动态目标分配给M艘 AUV. 优化目标是最大化总收益, 且在保证公平性的同时, 满足同一目标不可分割以及M大于N时每个目标分配的 AUV 个数约束条件. 其目标函数为

$$\max \sum_{i\in M} b_{ij},\ j\in N. \tag{3}$$

其中b_{ij}是围捕者A_i对动态目标T_j的收益值, 由A_i和T_j的态势信息输入拍卖神经网络得出, 将在 2.2 节具体阐述.

所有 AUV 按图 4 所示流程进行分配.

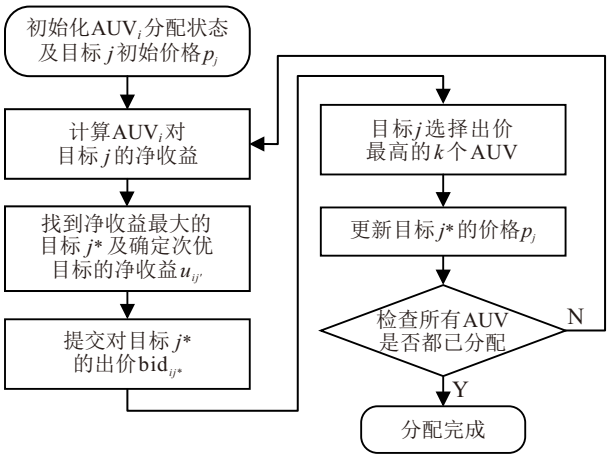

图4 多目标分配流程

首先规定每个目标初始的价格$p_j=0,\ j=1,2,\ldots,N$, 然后计算 AUV$_i$对每个目标j的净收益u_{ij}:

$$u_{ij}=b_{ij}-p_j. \tag{4}$$

找到净收益最大的目标j^*及确定次优目标j'的净收益$u_{ij'}$:

$$j^*=\arg\max_j u_{ij}, \tag{5}$$

$$u_{ij'}=\max_{j\neq j^*} u_{ij}. \tag{6}$$

然后提交对目标j^*的出价bid$_{ij^*}$:

$$\text{bid}_{ij^*}=u_{ij^*}-u_{ij'}+\varepsilon. \tag{7}$$

其中ε是一个小的正值, 用于确保价格的逐步调整. 利用bid$_{ij^*}$来描述对j^*的出价, 是为了确保 AUV 获得目标j^*相比获得其他目标的收益足够大时才去主动竞争. 而 AUV 获得每个目标的收益都差不多时先让其他 AUV 先竞争. 从而最终达成系统全局最优的分配方案[18].

每个目标j收集对其竞标的 AUV 的出价, 选择出价最高的k个 AUV(k为分配给目标的 AUV 数量), 并做出分配; 然后更新目标j的价格p_j:

$$p_j=\min_{i=1,2,\ldots,k}(\text{bid}_{ij}). \tag{8}$$

检查是否所有 AUV 都已被分配, 如果条件满足, 则算法终止; 否则, 继续下一轮竞标, 直到满足所有围捕者都被成功分配到目标, 且价格没有进一步变化.

2.2 拍卖神经网络设计

拍卖神经网络旨在求出上述模型构建中的收益值b_{ij}. 若通过直接利用最优控制理论计算b_{ij}, 其所消耗的时间过长, 而轻量级的神经网络具备高效的前向推理能力与极低的在线计算开销, 故通过神经网络进行预测是一种高效且可行的解决方案. 拍卖神经网络将M个围捕者和N个动态目标的位置与速度信息映射为$M\times N$维的收益值矩阵, 反映每个围捕者对各目标的出价偏好, 也是拍卖算法拍卖的重要依据. 拍卖神经网络采用监督学习进行训练, 训练数据标签由最优控制理论中的配点法, 利用 CasADi 优化框架构造非线性优化问题, 以最小化时间成本和能量消耗为目标, 利用 IPOPT 求解器计算出最少时间和最小能耗.

2.2.1 训练数据生成

为计算围捕者i围捕目标j的收益值, 这里将问题简化成围捕者追击目标的问题, 即使得围捕者与目标位置速度最终均一致. 采用 AUV 动力学模型描述围捕者的运动, 状态向量定义为$X=[x,y,v_x,v_y]^{\mathrm{T}}$, 包括位置坐标$(x,y)$和速度分量$(v_x,v_y)$. 控制输入为加速度$U=[a_x,a_y]^{\mathrm{T}}$. 系统的动力学方程为$\dot{X}=[v_x,v_y,a_x,a_y]^{\mathrm{T}}$, 目标的运动假设为匀速直线运动. 优化目标为最小化围捕时间T和能量消耗E, 采用加权和形式表达如下:

$$J=w_1T+w_2E. \tag{9}$$

其中: w_1为时间惩罚权重, w_2为能量项的权重, 消耗的能量$E=\int_0^T(a_x^2+a_y^2)\mathrm{d}x$.

求解收益值标签时围捕者的初始状态即是当前的位置和速度$X_{t=0}=[x_i,y_i,v_{x_i},v_{y_i}]^{\mathrm{T}}$. 追击成功条件即围捕者在时间$t=T$时与目标位置和速度均一致. 约束条件为加速度和速度的绝对值小于$a_{\max}$和$v_{\max}$. 然后采用四阶龙格-库塔方法离散化动力学方程, 通过使用 IPOPT 求解器计算出最优时间T^*和能量消耗E^*, 若求解失败则返回保守估计. 再将得到的T^*和E^*代入下式即可得到围捕者i围捕目标j的收益值标签:

$$b_{ij}^{*}=\frac{1}{1+\eta T^{*}}+\gamma \mathrm{e}^{-\delta E^{*}}. \tag{10}$$

其中: η和δ分别为时间和能量权重系数, γ为指数衰减系数.

2.2.2　监督学习训练拍卖神经网络

拍卖神经网络采用深度神经网络架构, 旨在捕捉围捕者与目标之间的复杂非线性关系. 如图 5 所示, 其由围捕者特征编码器、目标特征编码器和收益值预测头组成. 计算围捕者i围捕目标j的收益值时, 将i的当前状态向量$[x_i, y_i, v_{x_i}, v_{y_i}]^{\mathrm{T}}$和$j$的当前状态向量$[x_j, y_j, v_{x_j}, v_{y_j}]^{\mathrm{T}}$分别输入围捕者和目标编码器, 将两者输出拼接并输入给收益值预测头, 最后输出i对j的预测收益值b_{ij}.

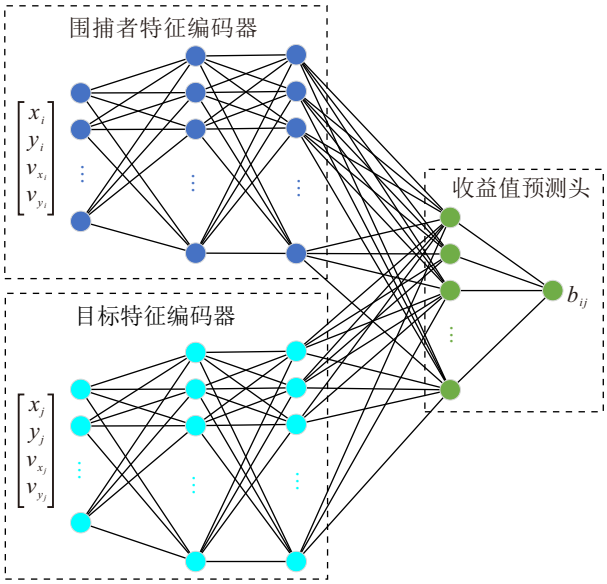

图5　拍卖神经网络框架

用该拍卖神经网络将同一时刻M个围捕者对N个目标的预测收益值全部算出, 再计算预测收益值与收益值标签的均方误差, 然后将均方误差反向传播, 从而更新拍卖神经网络参数.

本文总共利用 100 万条数据通过以上方式不断更新拍卖神经网络参数, 直至预测的收益值矩阵与收益值标签矩阵在拍卖算法中的分配结果相一致.

3　多智能体强化学习围捕算法

在协同围捕问题中, 多个智能体需要同时做出动作来协作完成对目标的围捕, 并且当前状态包含了对未来决策所需的全部信息, 还具有连续时间序列决策特征、复杂的状态空间. 因此, 可以将协同围捕问题建模为马尔可夫决策过程 (MDPs). 针对多个智能体的马尔可夫博弈定义S为所有智能体的状态集合, $O_1, O_2, \ldots, O_M$为每一个智能体的观测状态, $A=A_1\times A_2\times\ldots\times A_M$为所有智能体的联合动作集, $R_1, R_2, \ldots, R_M$为每一个智能体的奖励函数, P: $S\times A\times S\to[0,1]$表示状态转移概率.

多智能体强化学习的目标是寻找M个最优策略$\pi_1^{*}, \pi_2^{*}, \ldots, \pi_M^{*}$来使期望累计奖励最大化, 而本文采用的是多智能体柔性动作-评价 (MASAC) 算法, 故在最大化累计折扣奖励的同时还需极大化策略在每一个状态下动作分布的熵, 故目标函数为

$$J(\pi_i)=\mathrm{E}\Big[\sum_{k=0}^{\infty}\gamma^{k} r_{i,t+k}^{\mathrm{soft}}(s_{i,t+k}, a_{i,t+k})\Big]. \tag{11}$$

其中: γ为奖励折扣因子, π_i为智能体i的策略, $r_{i,t+k}^{\mathrm{soft}}$表示为

$$r_{i,t+k}^{\mathrm{soft}}(s_{i,t}, a_{i,t})=r(s_{i,t}, a_{i,t})+\alpha \mathrm{E}_{s_{i,t+1}\sim p}[H(\pi_i(\cdot|s_{i,t+1}))]. \tag{12}$$

其中: α为熵温度系数, 决定了对熵最大化的重视程度; $H(\pi_i(\cdot|s_{i,t+1}))$为智能体$i$策略的熵.

3.1　MASAC 强化学习算法及原理

在多智能体系统中, 环境的非平稳性以及状态空间维度高等问题使得传统的单智能体强化学习方法往往难以取得理想效果. 多智能体强化学习通过引入联合动作值函数、集中式训练等机制, 能够应对上述挑战. 而 MASAC 还引入了最大熵原则, 鼓励策略在保证奖励的同时保留探索性, 为连续动作空间任务提供了强有力的优化手段. 因此, 本文选用 MASAC 作为多智能体协同围捕问题的控制方法.

MASAC 的核心思想是将 SAC 的最大熵框架引入多智能体系统, 通过引入熵正则化项鼓励智能体在优化奖励的同时保持策略的随机性, 从而增强探索能力[19]. MASAC 使用集中式训练与分布式执行 (CTDE) 框架, 即在训练时利用全局信息, 而在执行时每个智能体仅依赖局部观测. 传统的 MASAC 对于每个智能体而言, 其主要是训练一个策略网络π_θ、两个价值网络Q_{φ_1}、Q_{φ_2}及两个目标价值网络$Q_{\hat{\varphi}_1}$、$Q_{\hat{\varphi}_2}$的参数. 在多智能体场景中, 环境的动态性不仅来自环境本身的变化, 更来自其他智能体行为的不可预测性, 从而导致环境非平稳性的问题. 本文引入一个目标策略网络$\pi_{\hat{\theta}}$, 在更新价值网络参数时使用更新较慢的目标策略网络, 使得价值网络的更新能够更平稳, 在一定程度上缓解非平稳性问题.

价值网络损失函数为

$$L_i^{Q}(\varphi_j)=\mathbb{E}_{(s_{i,t}, a_{i,t}, r_{i,t}, s_{i,t+1})\sim D}[Q_{\varphi_j}(s_{i,t}, a_{i,t})-\hat{Q}(s_{i,t}, a_{i,t})]^2,\quad i=1,2,\ldots,M,\ j=1,2. \tag{13}$$

目标价值$\hat{Q}(s_{i,t}, a_{i,t})$表示为

$$\hat{Q}(s_{i,t},a_{i,t})=r_{t,i}+\gamma\mathbb{E}_{s_{i,t+1}\sim p}[V(s_{i,t+1})],$$
$$V(s_{i,t+1})=\mathbb{E}_{a_{i,t+1}\sim\pi_{\hat{\theta}}}[\min_{\hat{\varphi}_j}Q_{\hat{\varphi}_j}(s_{i,t+1},a_{i,t+1})-\alpha\log(\pi_{\hat{\theta}}(a_{i,t+1}|s_{i,t+1}))],\tag{14}$$

其中$V(s_{i,t+1})$为软状态价值函数.

策略网络损失函数为

$$L_i^{\pi}(\theta)=\mathbb{E}_{s_{i,t}\sim D}[\mathbb{E}_{a_{i,t}\sim\pi_\theta}[\alpha\log\pi_\theta(a_{i,t}|s_{i,t})-\min_{\varphi_j}Q_{\varphi_j}(s_{i,t},a_{i,t})]],\quad i=1,2,\dots,M,\ j=1,2.\tag{15}$$

熵温度系数的损失函数为

$$L(\alpha)=\alpha\mathbb{E}_{s_{i,t}\sim D}[H(\pi(\cdot|s_{i,t}))-\hat{H}],\quad i=1,2,\dots,M,\tag{16}$$

其中$\hat{H}=-\dim(A)$为目标熵.

目标策略网络和两个目标价值网络的参数均以软更新的方式进行训练

$$\hat{\theta}_i\leftarrow\tau\theta_i+(1-\tau)\hat{\theta}_i,$$
$$\hat{\varphi}_{i,1}\leftarrow\tau\varphi_{i,1}+(1-\tau)\hat{\varphi}_{i,1},$$
$$\hat{\varphi}_{i,2}\leftarrow\tau\varphi_{i,2}+(1-\tau)\hat{\varphi}_{i,2},\quad i=1,2,\dots,M,\tag{17}$$

其中τ为软更新系数.

通过以上损失函数分别求出对应的网络参数和熵温度系数梯度, 再通过梯度下降法不断进行更新, 即可实现网络参数的迭代优化, 并最终使策略收敛到一个高回报且具有充分探索性的最优解.

本文通过引入拍卖神经网络进行前置目标分配, 有效地减小了每个智能体的状态空间. 这一设计从机制上降低了策略学习的复杂度. 同时, 在网络设计方面, 引入了一个目标策略网络, 使得更新价值网络参数时使用目标策略网络, 而非直接使用当前快速变化的策略网络, 从而提高了训练过程的稳定性. 前置目标分配和目标策略网络的引入, 能够在一定程度上缓解非平稳性问题.

本文通过系统性的实验验证了无模型强化学习算法的有效收敛性. 在不同规模的围捕场景中, 平均步长奖励随训练轮次增加呈现出稳步上升并逐渐收敛的趋势, 最终围捕胜率稳定在较高水平, 且多次实验结果的置信区间紧凑, 这些实证证据共同构成了对无模型强化学习算法在实际问题中收敛性的有力支撑.

3.2 状态空间设计

在协同围捕任务中, M个 AUV 智能体需要在水下环境中协作围捕N个动态目标. 为支持 MASAC 算法的集中式训练与分布式执行 (CTDE) 框架, 基于全局状态空间来实现, 以捕捉任务所需的所有关键信息, 同时保持状态表示的紧凑性和充分性.

传统的基于全局可观状态的问题基本上是将几乎所有的状态信息都输入给智能体, 这样会引起多智能体强化学习中的“维度爆炸”问题, 而本文基于分配算法将状态空间进行分割, 让每个智能体只接受组内成员以及相应目标和最近智能体的状态信息, 在一定程度上缓解了“维度爆炸”的问题, 并且能使智能体决策效率更高, 加快算法的收敛.

如图 6 所示, 设分配算法将智能体根据目标数量分为N组, 智能体A_1、A_2、A_3为第k组G_k, T_1为该组的动态目标.

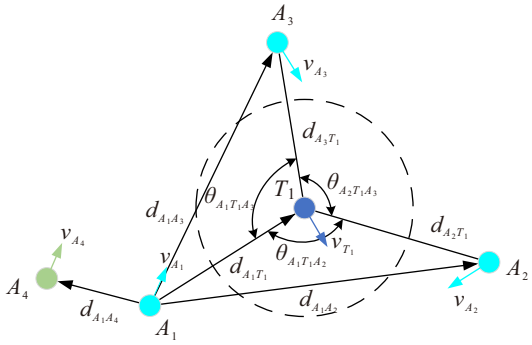

图6 状态信息示意图

以智能体A_1为例, 智能体A_4为除G_k组智能体外距离A_1最近的智能体, 计$o_{A_1}^{\text{ex}}$为针对智能体A_1的组外状态信息, $o_{A_1}^{\text{in}}$为针对智能体A_1的组内状态信息. 智能体A_1所接受的状态信息$o_{A_1}=\{o_{A_1}^{\text{ex}},o_{A_1}^{\text{in}}\}$.

$o_{A_1}^{\text{ex}}$由智能体A_4与A_1的相对位置向量$d_{A_1A_4}$以及智能体A_4的速度向量v_{A_4}所组成. $o_{A_1}^{\text{in}}$由智能体A_1指向组内其他智能体A_2、A_3以及目标T_1的相对位置向量$d_{A_1A_2}$、$d_{A_1A_3}$、$d_{A_1T_1}$, 智能体A_1与组内相邻的智能体相对于目标所形成的夹角$\theta_{A_1T_1A_2}$、$\theta_{A_1T_1A_3}$, 自身的位置坐标(x_{A_1},y_{A_1}), 以及组成所有智能体和目标的速度向量v_{A_1}、v_{A_2}、v_{A_3}、v_{T_1}所构成.

综上, 围捕目标T_k的G_k组智能体A_i的状态信息可表示为

$$o_{A_i}=\{o_{A_i}^{\text{ex}},o_{A_i}^{\text{in}}\},\ i=1,2,\dots,M;$$
$$o_{A_i}^{\text{ex}}=\{d_{A_iA_{\text{close}}},v_{A_{\text{close}}}\},$$
$$o_{A_i}^{\text{in}}=\{x_{A_i},y_{A_i},v_{A_i},d_{A_iT_k},v_{T_k},\theta_{A_iT_kA_{\text{close1}}^k},\theta_{A_iT_1A_{\text{close2}}^k}\}+\sum_{j\in G_k,j\neq i}\{d_{A_iA_j},v_{A_j}\}.\tag{18}$$

其中: A_{close}的索引$\text{close}=\arg\min_{j=1,2,\dots,M,j\neq i}(d_{A_iA_j})$, A_{close1}^k和A_{close2}^k为G_k组中距离A_i最近的两个智能体. 当各组智能体数量不相等时, 用数字 0 补齐状态信息维度不一致的情况.

3.3　动作空间设计

在多 AUV 协同围捕多动态目标的任务中, 每个 AUV 智能体通过策略网络生成动作来控制其推进器的推力, 以调整自身的位置和速度, 从而实现围捕目标. 动作空间的设计旨在提供足够的控制自由度, 同时符合物理约束和任务需求. 为了更符合水下 AUV 的运动学规律, 本文采用二阶系统来描述, 考虑到 AUV 在水下环境的运动特性, 本文以轴向加速度a_α和侧向加速度a_β作为 MASAC 算法策略网络的动作输出, 且考虑到加速度应设有最大值, 故$a_\alpha, a_\beta \in [-1, 1]$. 并且在获得连续动作时, 通过对高斯分布采用重参数采样的技巧, 以实现梯度可微, 提高训练稳定性.

3.4　奖励函数设计

奖励函数是指导M个 AUV 智能体协作完成围捕任务的关键. 每个智能体A_i被分配到特定目标T_j, 并与其他智能体组成围捕小组G_k. 需综合考虑距离、角度均匀性、任务完成和碰撞惩罚, 旨在引导多智能体协作形成均匀且紧凑的包围圈, 同时避免碰撞. 智能体A_i所得到的奖励定义如下:

$$r_i = r_{\text{dist}} + r_{\text{angle}} + r_{\text{done}} + r_{\text{col}}, \tag{19}$$

其中r_{dist}、r_{angle}、r_{done}和r_{col}分别表示距离奖励、角度均匀性奖励、任务完成奖励和碰撞惩罚.

距离奖励鼓励智能体A_i接近其目标T_j, 同时避免过于靠近目标而造成碰撞. 距离奖励r_{dist}设计为

$$r_{\text{dist}} = \begin{cases} 1 - \mathrm{e}^{||d_{A_iT_j}||}, & ||d_{A_iT_j}|| > \dfrac{r_c + r_s}{2}; \\ \mathrm{e}^{||d_{A_iT_j}||} - 1 - \dfrac{r_c + r_s}{2}, & ||d_{A_iT_j}|| \leqslant \dfrac{r_c + r_s}{2}. \end{cases} \tag{20}$$

其中r_c和r_s分别为此前提到的围捕半径和安全半径.

角度均匀性奖励旨在激励围捕小组中的智能体围绕目标形成均匀分布的包围圈, 提升围捕效率. 设智能体A_i相对于目标T_j的方向角为$\theta_{A_iT_j}$, 如图 7 所示.

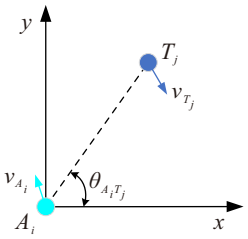

图7　方向角示意图

对于围捕小组G_k, 计算出组内所有智能体的方向角$\{\theta_{A_iT_j} | A_i \in G_k\}$, 并按升序排列为$\{\theta_1, \theta_2, \ldots, \theta_{|G_k|}\}$, 则角度差$\Delta\theta_l$为

$$\Delta\theta_l = \begin{cases} \theta_{l+1} - \theta_l, & l \neq |G_k|; \\ 2\pi - (\theta_{|G_k|} - \theta_1), & l = |G_k|. \end{cases} \tag{21}$$

用角度差的标准差$\sigma_{\Delta\theta}$来衡量角度分布的均匀性, 则均匀性奖励为

$$r_{\text{angle}} = w_{\text{angle}}(1 - \min(1, \sigma_{\Delta\theta})), \tag{22}$$

其中w_{angle}为权重系数.

碰撞惩罚用于避免智能体与目标、其他智能体或环境边界发生碰撞. 若智能体A_i与任一目标T_j、其他智能体或边界发生碰撞, 则奖励为

$$r_{\text{col}} = r_{\text{col}}^{\text{target}} + r_{\text{col}}^{\text{other}}. \tag{23}$$

其中: $r_{\text{col}}^{\text{target}}$为目标碰撞惩罚值, $r_{\text{col}}^{\text{other}}$为碰撞除目标以外物体的惩罚值.

任务完成奖励用于判断围捕任务是否成功, 属于团队奖励. 对于围捕小组G_k, 计算每个智能体A_i到目标T_j的距离并按升序排列为$\{d_1, d_2, \ldots, d_{|G_k|}\}$. 若满足

$$\begin{cases} r_s \leqslant d_l \leqslant r_c, \ l = 1, 2, \ldots, p; \\ \max(\Delta\theta_l) < \theta_{\text{cap}}, \ l = 1, 2, \ldots, |G_k|. \end{cases} \tag{24}$$

则视为围捕完成, 其中p为设定的围捕者数量. 将围捕小组G_k标记为完成状态, 并给予奖励

$$r_{\text{done}} = r_{\text{success}}, \tag{25}$$

否则$r_{\text{done}} = 0$.

3.5　协同围捕算法框架设计

利用监督学习训练拍卖神经网络的流程如图 8 所示. 整个过程分为训练和测试两个阶段.

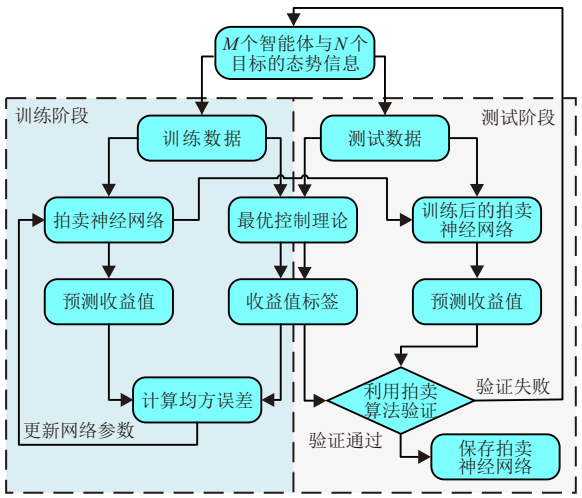

图8　拍卖神经网络训练流程

在训练阶段, 先计算拍卖神经网络预测的收益值与最优控制理论得出的收益值标签的均方误差, 然后对拍卖神经网络参数进行反向传播更新, 如此往复, 直到均方误差足够小. 在测试阶段, 对训练后的拍卖神经网络进行验证. 先计算训练后的拍卖神

经网络预测的收益值与收益值标签; 然后通过拍卖算法计算出两者的分配结果; 再以收益值标签的分配结果为标准, 计算预测收益值的分配准确率, 直到准确率足够高, 否则重新进行训练. 验证通过后将拍卖神经网络进行保存.

在训练好拍卖神经网络之后, 进行策略网络的训练. 基于拍卖神经网络的 MASAC 算法流程如图 9 所示. 首先对各智能体价值网络和策略网络参数以及熵温度系数进行初始化, 环境给出各智能体与目标的态势信息, 拍卖神经网络根据态势信息把目标分配给 M 个智能体, 形成若干围捕小组. 根据分组, 将环境所反馈的态势信息进行分割, 各策略网络根据分割之后的态势信息进行决策, 输出联合动作给环境, 环境返回下一态势信息和奖励信息, 将这一次完整交互信息保存在经验回放池, 即完成一次智能体与环境的交互.

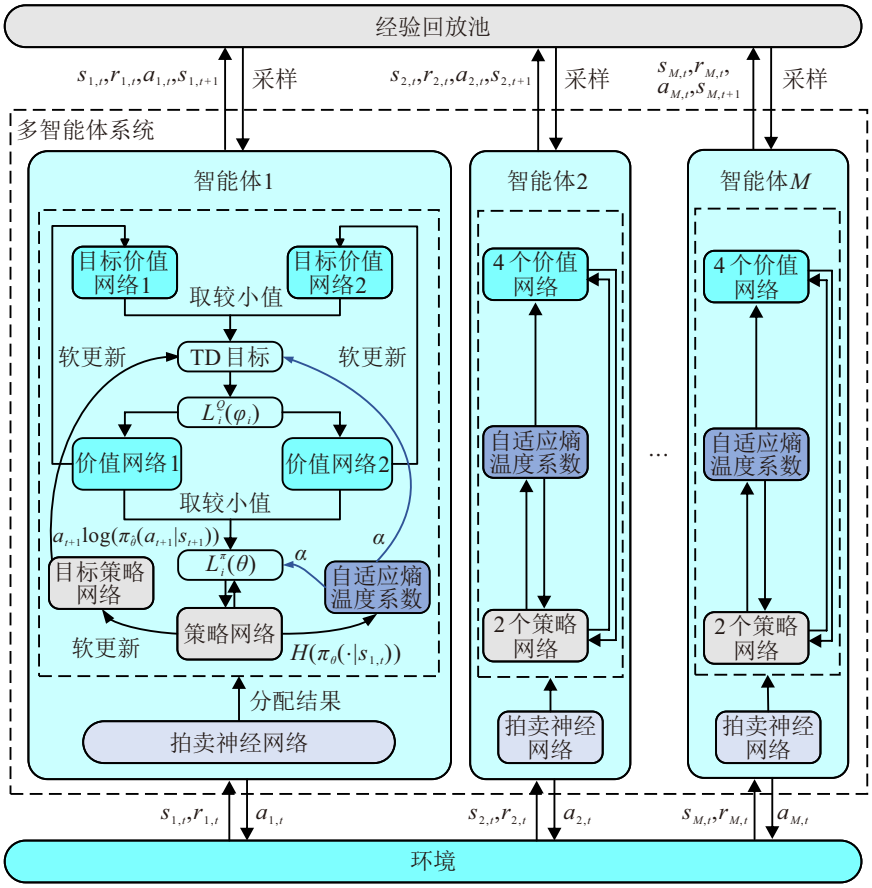

图9 基于拍卖神经网络的 MASAC 算法流程

当交互次数达到更新频率 f 时, 各智能体在经验回放池进行采样, 利用集中式训练与分布式执行框架计算价值网络损失 $L_i^Q(\varphi_j)$(见式 (13))、策略网络损失 $L_i^\pi(\theta)$(见式 (15)) 和熵温度系数损失 $L(\alpha)$(见式 (16)), 分别对其参数求导, 得出梯度信息; 其次进行反向传播更新价值网络参数 φ_j、策略网络参数 θ 以及熵温度系数 α; 再次对目标价值网络参数 $\hat{\varphi}_j$ 和目标策略网络参数 $\hat{\theta}$ 进行软更新 (见式 (17)); 最后重新累计交互次数, 如此往复, 直至策略网络输出的动作能够完成所设定的围捕任务.

4 多智能体协同围捕实验设计及分析

为验证基于拍卖神经网络的 MASAC 算法在多 AUV 协同围捕多动态目标任务的有效性, 设计了一系列实验来评估算法性能, 包括实验环境的构建、超参数的设置以及与其他算法性能的对比分析.

4.1 实验环境和超参数配置

实验计算机的配置为 12th Gen Intel(R) Core (TM) i7-12700F 2.10 GHz 型号的 CPU、32.0 GB 的 RAM、NVIDIA GoForce RTX 3060 型号的 GPU, 使用的深度学习框架为 PyTorch 2.0, 强化学习训练环境框架为 OpenAI Gym.

图 10 所示为 6 个围捕者围捕两个动态目标时各时间点的围捕状态, 仿真场景为一个 200×200 m^2 的正方形区域, 围捕者和动态目标的初始位置在仿真区域内随机生成. 所有围捕者是同构的, 并且围捕者间能够实时通信并共享动态目标的位置. 拍卖神经网络根据 6 个围捕者和 2 个动态目标的当前位置和速度将围捕者分为青色和绿色两个队伍, 每个队

伍对各自目标进行围捕.

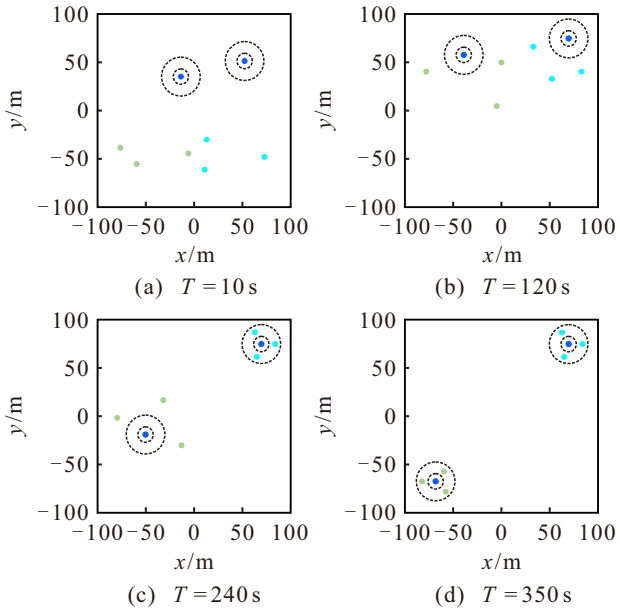

图10　仿真场景示意图

围捕者和动态目标的最大速度均为 3 节, 最大加速度均为 1 节/s. 为模拟水下环境, 设置阻尼系数为 0.25. 任务的最大时间为 400 s. 蓝色圆点为动态目标, 围捕半径r_c和安全半径r_s分别为 20 m 和 7.5 m. 当达到围捕成功条件时, 动态目标失去移动能力, 如图 10(c) 的右上角目标. 待全部目标都被围捕成功, 任务完成.

首先通过监督学习训练好拍卖神经网络, 在生成监督数据时, $w_1=0.1$, $w_2=1$, $\eta=0.5$, $\delta=0.1$, $\gamma=0.3$. 动态目标的运动学模型中, $\beta_1=0.3$, $\beta_2=0.5$, $\beta_3=0.5$, $\lambda_1=3$, $\lambda_2=1$, $\lambda_3=1$. 在训练策略网络时, 算法 ANN-MASAC 各超参数的具体取值如表 1 所示.

表1　训练时算法超参数设置

参数名称	参数取值
最大训练步数	4×10^5
回合最大步数	400
经验池容量	1×10^5
批训练大小	1 024
随机步数	2 000
随机数种子	42
奖励折扣因子(γ)	0.99
更新频率(f)	50
策略网络学习率	1×10^{-4}
价值网络学习率	3×10^{-4}
熵温度系数学习率	4×10^{-4}
熵温度系数初始值	0.2
熵目标($\hat{H}$)	−2
软更新系数(τ)	0.005

4.2　算法性能对比分析

为验证本文提出的基于拍卖神经网络的 MASAC(ANN-MASAC) 算法在协同围捕任务中的性能, 本文将 MAPPO、MATD3 以及 QMIX 算法作为对比算法. 其中: 作为多智能体协作任务基线算法的 MAPPO, 通过引入剪切概率比和集中式价值函数, 确保了策略更新的稳定性, 适合需要稳定学习和大规模多智能体协作的动态围捕任务[20]; MATD3 利用双重Q网络和延迟更新机制减少Q值过估计, 结合噪声增强的探索能力, 能够在连续动作空间中生成平滑且鲁棒的单智能体决策, 适应动态目标逃逸的复杂场景[21]; QMIX 通过单调值函数分解, 将全局动作值函数分解为各智能体的局部值函数, 利用单调性约束确保一致性, 提高训练稳定性[22]. 利用传统的最小化各智能体与各目标之间距离的平均分组方式给各智能体分组[23], 奖励函数根据分组给予各智能体奖励.

在训练的过程中, 每个算法都使用相同的随机数种子, 将每个回合的平均步长奖励记录下来并做平滑处理, 在 6 个围捕者围捕 2 个动态目标时, 各算法平均步长奖励如图 11 所示, 其横轴表示算法与环境的交互次数.

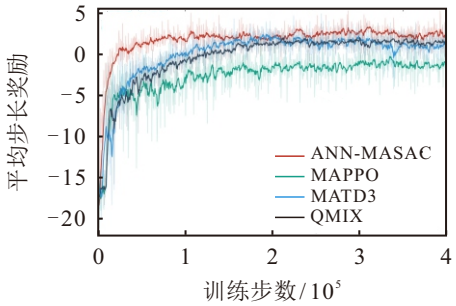

图11　各算法平均步长奖励图

从图 11 可以看出, 本文算法初期快速提升, 在与环境交互约 8 万次后收敛, 保持稳定, 且波动最小. MATD3 算法表现效果次之, 初期提升较快, 但马上就进入了缓慢上升的过程, 在约 19 万次交互后奖励进入稳定状态, 但波动仍不小, 说明 MATD3 在应对多动态目标环境时探索能力很强, 但样本利用能力较弱. QMIX 算法与 MATD3 算法的表现效果比较相近, 收敛速度两者几乎一样, 但其收敛后更稳定. 基线算法 MAPPO 表现则较差, 并且随着步数增加, 也有着不小的波动, 表明其应对此多动态目标问题不太理想.

为了进一步验证算法的可信度和稳定性, 每个算法分别在 12 个随机数种子上进行实验. 分别记录训练初期、中期、后期各 100 回合的步长奖励, 然后

表2 各算法在各阶段的奖励均值及置信区间

采样阶段	ANN-MASAC	MAPPO	MATD3	QMIX
初期100回合	−1.984 9 ± 0.152 5	−9.163 8 ± 0.471 0	−7.041 9 ± 0.674 6	−7.724 0 ± 0.167 7
中期100回合	2.083 9 ± 0.408 4	−2.053 1 ± 0.329 2	1.093 8 ± 0.775 8	1.173 5 ± 0.127 8
后期100回合	2.727 6 ± 0.179 5	−1.267 2 ± 0.244 5	0.436 6 ± 0.795 0	1.364 6 ± 0.101 7

求均值, 再基于 12 次实验数据求出置信水平为 95% 的置信区间, 其结果如表 2 所示.

ANN-MASAC 算法在训练初期即表现出优于其他算法的性能, 并随着训练进程持续提升, 在最后 100 回合达到最高平均奖励, 且置信水平为 95% 的置信区间范围也比较紧凑, 因而证实了本算法的可信度和稳定性.

4.3 可扩展性实验与分析

为验证本文所提算法的可扩展性, 将围捕者数量与动态目标的数量分别进行调整, 增加 8 对 2、10 对 2、9 对 3、12 对 3、15 对 3、12 对 4、16 对 4 等实验, 其实验结果如图 12 所示.

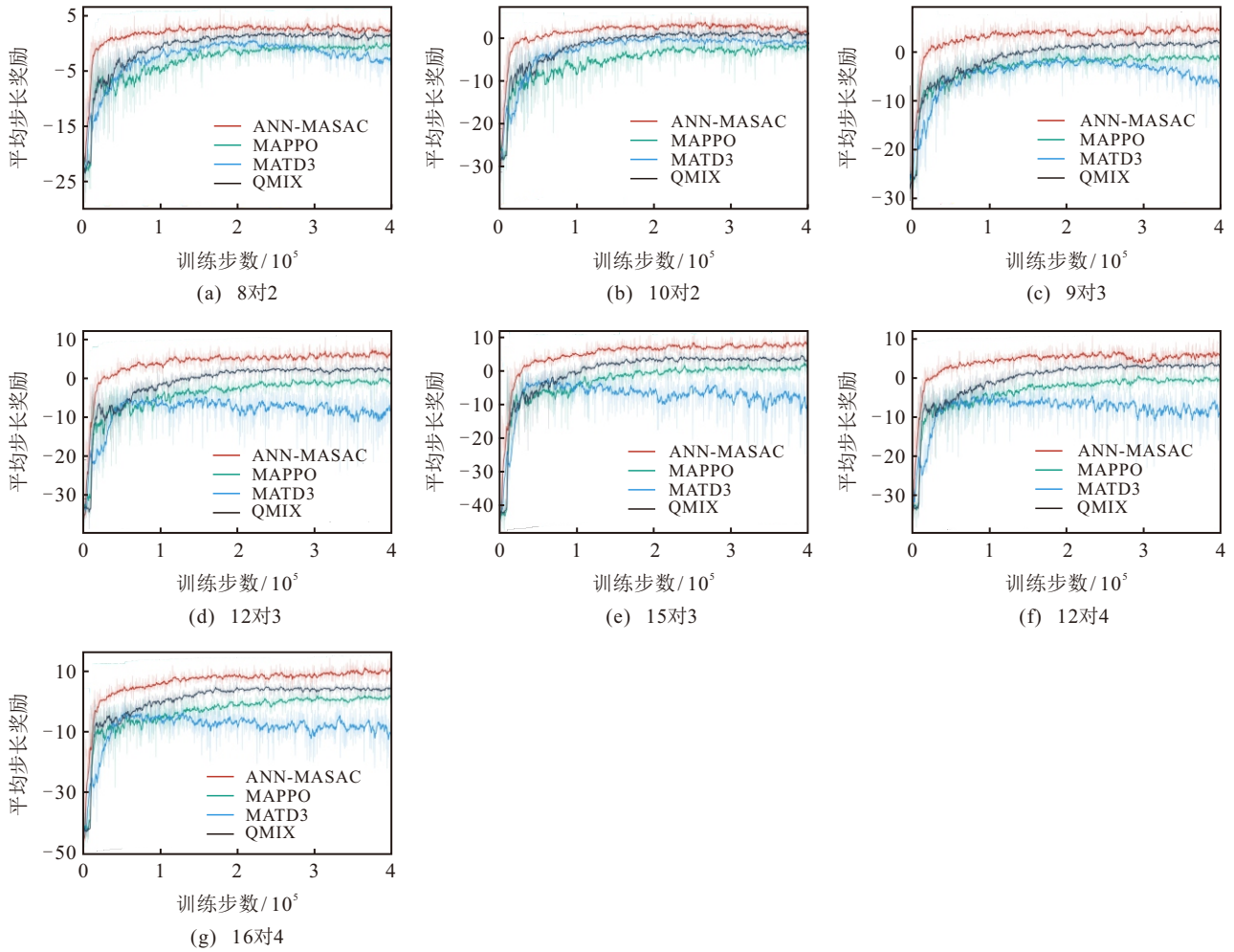

图12 不同场景下各算法平均步长对比

从奖励大小来看: ANN-MASAC 在所有场景中均获得最高奖励, 表明其围捕策略更优、捕获效率更高; QMIX 次之, MATD3 在应对两个目标时与 QMIX 相近, 但在 3、4 个目标时奖励显著下降; MAPPO 在应对两个目标时奖励较低, 但在目标增多时优于 MATD3.

收敛速度上, ANN-MASAC 优势明显, 通常在 10 万至 15 万步内快速收敛, 且围捕者数量增加时仍能保持; MATD3 和 MAPPO 则需 15 万至 20 万步, 其中 MATD3 收敛最慢, 反映出了各算法在样本利用与策略探索上的差异.

稳定性方面, ANN-MASAC 奖励曲线平滑、波动小, 鲁棒性突出; MATD3 在多目标场景 (如 12 对 3、15 对 3 等) 中波动较大, 稳定性随任务复杂度上升而下降.

总体而言, ANN-MASAC 综合性能最优, 适用

于复杂多智能体任务; MAPPO、MATD3 和 QMIX 在低复杂度场景中仍具价值, 但需进一步优化以应对高维挑战.

本文还增加了围捕者数量小于等于目标数量以及无法均匀分配时的实验验证, 包括 3 对 5、4 对 5、5 对 5 以及 5 对 2 等实验, 其实验结果如图 13 所示. 可以看出, 无论是在奖励大小、收敛速度还是稳定性方面, ANN-MASAC 均表现优异.

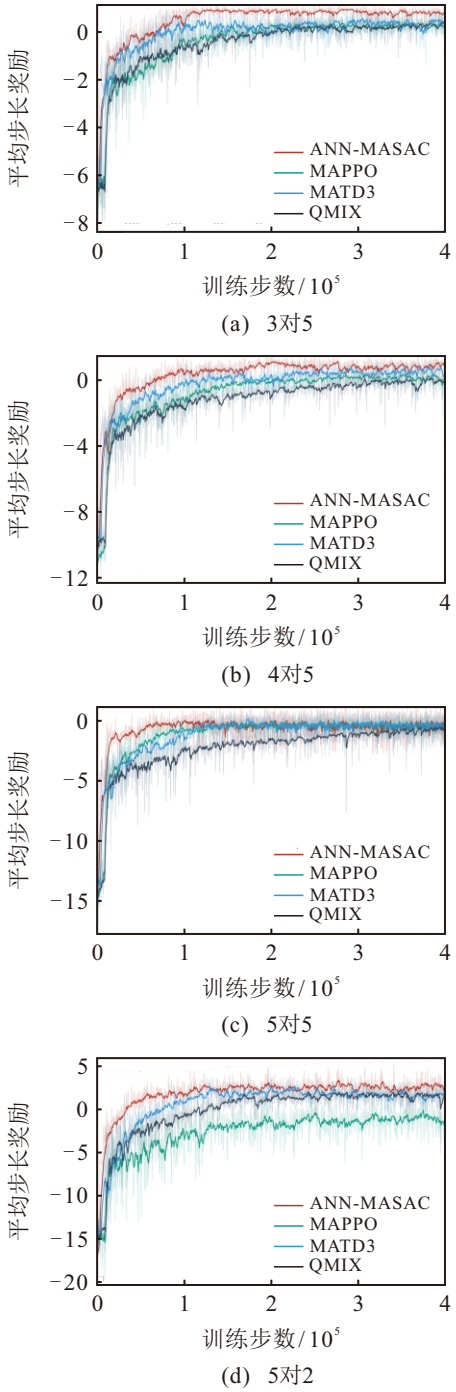

图13　不同场景下各算法平均步长对比

本文所提算法通过结合拍卖神经网络, 综合考量目标位置、速度与围捕的最短时间, 在多智能体围捕任务中展现出独特优势. 相较于 MAPPO、MATD3 与 QMIX, 其收敛速度更快, 主要在于拍卖机制能够高效分配任务并动态调整协作策略, 从而减少冗余探索并提升学习效率; 稳定性更强, 主要源于对速度与时间的实时优化, 使智能体更适应环境变化, 降低协调失误与环境噪声带来的波动; 其奖励更高, 则因拍卖神经网络能更精准匹配围捕者与目标的速度特性, 缩短围捕时间并优化全局策略. 虽然 QMIX 也取得较高奖励, 但 ANN-MASAC 在应对复杂动态目标时表现出更强的适应性与优化能力.

为了进一步评估算法性能, 本文通过调整成功条件中到达动态目标包围圈内的围捕者数量p来测试各算法的具体表现, 其胜率如表 3 ～ 表 5 所示.

表3　6 对 2、9 对 3 和 12 对 4 的胜率　%

场景	算法	$p=1$	$p=2$	$p=3$
6对2	ANN-MASAC	96.00	85.50	52.50
	MAPPO	28.50	11.00	2.50
	MATD3	87.50	66.50	23.50
	QMIX	92.50	59.50	12.00
9对3	ANN-MASAC	94.67	91.00	71.67
	MAPPO	44.00	32.33	11.00
	MATD3	63.33	33.67	5.33
	QMIX	96.00	70.67	14.33
12对4	ANN-MASAC	94.75	91.00	78.00
	MAPPO	48.75	32.25	14.25
	MATD3	55.00	17.50	1.75
	QMIX	94.75	78.00	14.50

表4　8 对 2、12 对 3 和 16 对 4 的胜率　%

场景	算法	$p=1$	$p=2$	$p=3$	$p=4$
8对2	ANN-MASAC	95.00	92.50	80.00	44.00
	MAPPO	51.50	41.00	28.00	9.00
	MATD3	84.50	68.00	29.50	5.50
	QMIX	95.50	74.00	41.50	5.50
12对3	ANN-MASAC	97.00	96.33	91.67	78.67
	MAPPO	56.00	44.00	31.33	11.33
	MATD3	72.00	27.67	3.67	0
	QMIX	97.33	90.67	62.00	12.33
16对4	ANN-MASAC	96.75	96.00	93.00	83.50
	MAPPO	61.00	48.75	35.50	15.00
	MATD3	76.00	34.75	7.75	1.00
	QMIX	97.25	93.25	59.00	12.25

表5　10 对 2 和 15 对 3 的胜率　%

场景	算法	$p=1$	$p=2$	$p=3$	$p=4$	$p=5$
10对2	ANN-MASAC	99.50	96.00	94.00	75.50	38.00
	MAPPO	72.50	59.50	44.50	17.00	2.50
	MATD3	86.00	86.50	60.00	16.50	3.00
	QMIX	96.50	89.50	69.50	30.50	8.50
15对3	ANN-MASAC	96.67	95.67	95.67	90.67	77.67
	MAPPO	60.67	50.00	41.67	30.00	17.00
	MATD3	81.67	42.67	9.67	2.00	0
	QMIX	99.67	94.67	82.33	43.67	9.33

从表中可见, 随着围捕难度增大 (即围捕者数量p增加), 各算法胜率均呈下降趋势. ANN-MASAC 虽有所下降, 但波动明显小于其他算法. MATD3 与 MAPPO 在应对两个目标时胜率相近; 而在应对 3、4 个目标时, MATD3 波动最大, 在 12 对 3、15 对 3、12 对 4、16 对 4 等场景中胜率甚至接近为零. 本文所提算法在所有场景中均保持最优胜率, 且随着围捕双方数量的增加, 尤其在全员围捕 (即各规模中p为最大值) 时, 胜率持续提升. 这表明引入拍卖神经网络能更高效地分配任务, 优化多智能体间的协作与决策, 尤其在复杂场景中表现出更强的适应性.

ANN-MASAC、MAPPO、MATD3 和 QMIX 的总体平均胜率分别为 85.77%、33.95%、37.18% 和 61.19%, 这也充分展现了 ANN-MASAC 在解决该问题上的高效性和鲁棒性.

5 结 论

针对 AUV 协同围捕多动态目标的问题, 本文提出了一种结合拍卖神经网络的多智能体强化学习方法. 首先对 AUV 群围捕多动态目标的问题进行了问题建模, 并分别构建了同一深度下 AUV 和动态目标的运动学模型; 其次设计了基于拍卖神经网络的分配算法, 根据围捕者和目标双方的位置、速度、所需围捕时间及能量消耗来分配各自的目标; 然后设计了多智能体强化学习围捕算法, 根据分配算法的分配结果分别构建每个围捕者的状态空间, 围捕算法根据状态空间输出每个围捕者的轴向加速度和侧向加速度; 最后在不同围捕规模的仿真实验下对算法性能进行了测试和分析. 结果表明, 本文所提算法能够在各种围捕规模下高效完成协同围捕任务, 并且在总体胜率上相较于算法 QMIX、MATD3 以及基线算法 MAPPO 分别提高了 24.58%、47.97% 和 51.82%, 表明本文所提算法具有较好的表现力和泛化能力.

在未来的研究中, 将进一步挖掘多智能体强化学习强大的策略学习能力, 构建海洋环境下更精细化的 AUV 集群运动学或动力学模型, 研究在局部已知状态下的协同围捕问题, 探索在 AUV 间通信受限环境下的协作与控制方法, 并逐步向半实物仿真和实船测试过渡. 通过引入高保真仿真环境, 结合硬件在实际环境中测试, 验证算法在真实海洋环境中的鲁棒性与适应性, 为最终开展实船协同围捕实验奠定理论基础.

参考文献 (References)

[1] Cai W Y, Chen H, Zhang M Y. A survey on collaborative hunting with robotic swarm: Key technologies and application scenarios[J]. Neurocomputing, 2024, 598: 128008.

[2] 李一平, 许真珍. 多自主水下机器人协同控制[M]. 北京: 科学出版社, 2020: 11.
(Li Y P, Xu Z Z. Cooperative control of multiple autonomous underwater vehicles[M]. Beijing: Science Press, 2020.)

[3] Wang Y, Li H P, Yao Y. An adaptive distributed auction algorithm and its application to multi-AUV task assignment[J]. Science China Technological Sciences, 2023, 66(5): 1235-1244.

[4] 李海峰, 杨宏安, 盛梓茂, 等. 基于 MAPPO 的多无人机协同分布式动态任务分配[J]. 控制与决策, 2025, 40(5): 1429-1437.
(Li H F, Yang H A, Sheng Z M, et al. Multi-UAV collaborative distributed dynamic task allocation based on MAPPO[J]. Control and Decision, 2025, 40(5): 1429-1437.)

[5] Dong D B, Zhu Y H, Du Z Z, et al. Multi-target dynamic hunting strategy based on improved K-means and auction algorithm[J]. Information Sciences, 2023, 640: 119072.

[6] Wang G H, Wang F M, Wang J H, et al. Collaborative target assignment problem for large-scale UAV swarm based on two-stage greedy auction algorithm[J]. Aerospace Science and Technology, 2024, 149: 109146.

[7] 白小山, 佘桉奇, 郑心泉, 等. 基于改进最小边际代价算法的多 USV 多 AUV 任务分配[J]. 控制与决策, 2025, 40(1): 119-127.
(Bai X S, She A Q, Zheng X Q, et al. Task assignment for multiple USVs and AUVs based on improved minimum marginal cost algorithm[J]. Control and Decision, 2025, 40(1): 119-127.)

[8] Okumura K, Défago X. Solving simultaneous target assignment and path planning efficiently with time-independent execution[J]. Artificial Intelligence, 2023, 321: 103946.

[9] 潘云伟, 李敏, 曾祥光, 等. 基于人工势场和改进强化学习的自主式水下潜航器避障和航迹规划[J]. 兵工学报, 2025, 46(4): 72-83.
(Pan Y W, Li M, Zeng X G, et al. AUV obstacle avoidance and path planning based on artificial potential field and improved reinforcement learning[J]. Acta Armamentarii, 2025, 46(4): 72-83.)

[10] 周萌, 李建宇, 王昶, 等. 多机器人协同围捕方法综述[J]. 自动化学报, 2024, 50(12): 2325-2358.
(Zhou M, Li J Y, Wang C, et al. Multi-robot cooperative hunting: A survey[J]. Acta Automatica Sinica, 2024, 50(12): 2325-2358.)

[11] 郭戈, 康健. 具有复杂动力学的多智能体系统分布式优化综述[J]. 控制与决策, 2024, 39(7): 2113-2124.
(Guo G, Kang J. A survey on distributed optimization for multiagent systems with complex dynamics[J]. Control and Decision, 2024, 39(7): 2113-2124.)

[12] 夏家伟, 朱旭芳, 张建强, 等. 基于多智能体强化学习

的无人艇协同围捕方法[J]. 控制与决策, 2023, 38(5): 1438-1447.
(Xia J W, Zhu X F, Zhang J Q, et al. Research on cooperative hunting method of unmanned surface vehicle based on multi-agent reinforcement learning[J]. Control and Decision, 2023, 38(5): 1438-1447.)

[13] Han B Q, Shi L, Wang X Y, et al. Multi-agent multi-target pursuit with dynamic target allocation and actor network optimization[J]. Electronics, 2023, 12(22): 4613.

[14] Xia J W, Luo Y S, Liu Z K, et al. Cooperative multi-target hunting by unmanned surface vehicles based on multi-agent reinforcement learning[J]. Defence Technology, 2023, 29: 80-94.

[15] Awheda M D, Schwartz H M. A fuzzy reinforcement learning algorithm using a predictor for pursuit-evasion games[C]. 2016 Annual IEEE Systems Conference. Orlando, 2016: 1-8.

[16] Wang Z Y, Du J, Jiang C X, 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, 2023, 10(9): 8271-8285.

[17] Cao X, Zuo F. A fuzzy-based potential field hierarchical reinforcement learning approach for target hunting by multi-AUV in 3-D underwater environments[J]. International Journal of Control, 2021, 94(5): 1334-1343.

[18] Schneider E, Sklar E I, Parsons S, et al. Auction-based task allocation for multi-robot teams in dynamic environments[C]. Towards Autonomous Robotic Systems. Cham: Springer International Publishing, 2015: 246-257.

[19] Haarnoja T, Zhou A, Hartikainen K, et al. Soft actor-critic algorithms and applications[J/OL]. 2018, arXiv: 1812.05905.

[20] Yu C, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]. Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022: 24611-24624.

[21] Ackermann J, Gabler V, Osa T, et al. Reducing overestimation bias in multi-agent domains using double centralized critics[J/OL]. 2019, arXiv: 1910.01465.

[22] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(1): 7234-7284.

[23] Yu J J, Chung S J, Voulgaris P G. Target assignment in robotic networks: Distance optimality guarantees and hierarchical strategies[J]. IEEE Transactions on Automatic Control, 2015, 60(2): 327-341.

作者简介

谢地杰 (2001−), 男, 硕士生, 主要研究方向为 AUV 智能控制、多智能体强化学习, E-mail: 2010280844@qq.com;

李敏 (1981−), 男, 讲师, 博士, 主要研究方向为机器学习与智能控制, E-mail: liminfish008@163.com;

曾祥光 (1973−), 男, 副教授, 硕士, 主要研究方向为强化学习、智能控制与群体智能优化计算, E-mail: xgzeng@126.com;

任文哲 (2001−), 男, 硕士生, 主要研究方向为强化学习与 AUV 攻击占位, E-mail: 3177007826@qq.com;

张滔 (2000−), 男, 硕士生, 主要研究方向为强化学习与智能控制, E-mail: 2946177475@qq.com;

彭倍 (1976−), 男, 教授, 博士, 主要研究方向为微机电系统、微纳传感与检测、智能机器人与无人系统, E-mail: beipeng@uestc.edu.cn.