© *2023* 作者：本文的一个版本已获准在 *IEEE*《海洋工程学报》发表。

基于深度强化学习的自主水下航行器自适应编队运动规划与控制

Behnaz Hadi[1], Alireza Khosravi[*1]、Pouria Sarhadi[2]

[1] 伊朗巴博尔巴博尔·努希尔瓦尼理工大学电气与计算机工程系

[2] 英国哈特菲尔德赫特福德大学物理、工程与计算机科学学院

摘要

在未知和不确定环境中规划安全路径，是领航者—跟随者编队控制的一项挑战。在该架构中，领航者通过采取最优动作朝目标前进，跟随者则需在保持期望编队队形的同时避开障碍物。该领域的大多数研究都将编队控制和避障分开考察。本研究提出一种基于深度强化学习（DRL）的新方法，用于欠驱动自主水下航行器（AUV）的端到端运动规划与控制。研究旨在设计基于演员-评论家结构的最优自适应分布式控制器，以实现自主水下航行器的编队运动规划。该方法通过控制自主水下航行器的速度和航向来实现。在避障方面，采用了两种方法。第一种方法旨在为领航者和跟随者设计控制策略，使各航行器分别学习自己的无碰撞路径。此外，跟随者还遵循统一的编队保持策略。第二种方法中，只有领航者学习控制策略，并安全地引导整个群体朝目标前进。此时，跟随者的控制策略是保持预定距离和角度。研究表明，在存在洋流、通信延迟和感知误差的情况下，所提出的方法在受到现实扰动的环境中仍具有鲁棒性。通过多项计算机仿真，对算法的效率进行了评估并验证了其有效性。

关键词：多自主水下航行器、编队控制、运动规划、避障、深度强化学习（DRL）、演员-评论家网络

1. 引言

多水下机器人领域的一项重要研究课题是编队控制。与单独部署航行器相比，协同执行复杂任务能够提升性能、降低成本和任务时间、确保达到预期结果，并提高任务成功率[1, 2]。自主水下航行器是无人自主航行器，能够利用自身的控制与导航系统，在水下环境中连续数小时乃至数天执行特定任务。自主水下航行器可用于海床调查与测绘、勘探、识别与搜救任务、调查地质威胁，以及检查石油和天然气

* 电子邮箱：akhosravi@nit.ac.ir

管道、确定水体生物学参数、开展导航与控制系统研究、进行流体动力学计算等 [3-5]。自主水下航行器在各类应用中的发展，凸显了研究如何改进其群体运动控制的重要性。在水下机器人领域的多项研究中，打造完美、智能且具备决策能力的自主水下航行器是研究目标。

编队结构是指智能体在执行指定命令的同时保持预期队形。编队控制主要有三种架构：领航者-跟随者、虚拟结构和行为式，每种架构各有优缺点。在领航者-跟随者方法中，将一个或多个自主水下航行器设为领航者，其余自主水下航行器设为跟随者。领航者沿轨迹或预先设计的参考路径行进，而跟随者则通过保持预先设定的期望距离和角度来跟随领航者的状态，从而生成并维持编队形状。在虚拟结构方法中，所有智能体都基于一个虚拟点保持刚性几何关系，以形成期望编队。在行为式架构中，适当的行为被定义为多个控制目标的加权组合 [6, 7]。编队控制有多种应用，例如移动机器人 [8], 无人水面航行器（USV）[9], 自主水下航行器（AUV）[10], 无人机（UAV）[11], 卫星 [12], 和航天器 [13]。在上述领域中，由于系统动力学存在不确定性和耦合性，还受到海浪和洋流造成的环境扰动以及水下通信的影响，自主水下航行器的编队控制最具挑战性 [14]。该领域广泛采用的方法包括自适应控制 [15], 滑模控制 [16, 17], 预测控制 [18], 基于神经网络的控制 [19], 图论 [10], 等。在 [20], 中，采用动态面控制与扰动观测器相结合的方法，实现自主水下航行器的时变编队控制。[21] 采用拉格朗日方法控制水面船舶编队，并通过应用约束函数实现编队。[22]提出了一种基于李雅普诺夫理论与反步法相结合的欠驱动自主水下航行器领航者-跟随者编队控制方法。Bono 等人利用分布式预测控制模型，研究了一组自主水下航行器的编队控制与避障 [18]。Heshmati 等人提出了一种多水下航行器机械臂系统（UVMS）的领航者-跟随者架构 [23] ，以完成协同物体运输任务，并为跟随者提出了一种具有预设性能的轨迹估计方案。[6] 对大多数自主水下航行器编队控制技术进行了评估。

尽管经典控制方法具有理想的控制性能，但仍需作出系统动力学模型线性化以及精确辨识模型或系统参数等假设，而且大多数此类方法都需要复杂的解析计算；由于 AUV 系统具有高度非线性的特性和不确定系数，并且所处环境中存在未知扰动，因此将其应用于实际场景颇具挑战性。

作为机器学习的一个分支，强化学习（RL）已广泛应用于为连续和离散时间下的非线性不确定动态系统设计具有演员-评论家结构的最优自适应控制器，从而提升自主水下航行器的自主性水平 [24]。在强化学习中，一个或多个智能体与可能未知的环境进行实时交互，并根据获得的经验，学习选择优化策略以实现特定目标。该方法仅使用一个效率标量指标（称为强化信号或奖励信号），即可训练

智能体在复杂环境中以不确定且随机的方式、在无监督的情况下 [25]。强化学习与深度学习方法的结合称为深度强化学习，已成为解决自主系统复杂问题并适应不可预测且不确定的多维的强大工具，可降低控制系统的设计复杂度，同时提高其效率 [26]。

迄今为止，关于将强化学习应用于自主水下航行器编队控制的研究并不多。目前，与自主水下航行器相关的大多数文献主要集中于利用强化学习控制自主水下航行器，以及单航行器的路径规划和路径跟踪。在[27], 中，深度强化学习已被用于为自主水下航行器设计自适应低层控制器。在[28] 中，强化学习与模仿学习相结合，被用于控制无人水下航行器（UUV）。Sun 等人提出使用强化学习方法对自主水下航行器进行运动规划。近端策略优化（PPO）算法已被用于设计纵荡和偏航控制信号 [29]。在[30], 中，修正后的Q学习已被用于使自主水下航行器避开障碍物。当自主水下航行器进入预先定义的危险区域时，探索过程停止，并持续执行控制指令的利用过程，直到离开不安全区域。在 [31], 中，强化学习策略已被用于控制自主水下航行器的对接，随后将该方法与经典控制技术进行了比较。Zhang 等人将交互式深度强化学习应用于自主水下航行器的路径跟踪。在该方法中，深度Q网络（DQN）方法与交互式强化学习相结合，奖励函数同时包含环境奖励和人类奖励[32]。

由于深度强化学习的普及，研究人员已在无人水面航行器、无人机、移动机器人等领域开展了协同控制和编队控制研究。Zhang 等人提出了仿鱼机器人的环形编队控制方法 [33]。在 [34], 中，基于强化学习和模仿学习的编队控制与避障方法仅针对少数具有全驱动二阶动力学的领导者-跟随者系统提出。基于深度强化学习方法，在采用领航者-跟随者架构的无人水面航行器中，已分别开展了分散式异步编队控制和具有泛化能力的分布式编队研究 [35] 和 [36]。Lin 等人结合深度强化学习与长短期记忆网络（LSTM），解决了无人机编队中的动态频谱交互问题 [37]。在 [38], 中，提出了基于多目标强化学习的固定翼无人机群编队控制方法。在 [39], 中，基于带有优先经验回放缓冲区机制的深度强化学习，提出了移动机器人群体的协同运动控制方法。

受前述讨论启发，本文提出了一种基于深度强化学习的欠驱动自主水下航行器分布式自适应编队控制器，具备避障和自碰撞规避功能。基于演员-评论家架构的最优自适应控制器能够评估当前控制策略的性能，并利用这些数据更新控制器，从而降低编队运动规划误差。本文还提出了用于多自主水下航行器运动规划的马尔可夫决策过程（MDP）模型。 系统直接将自主水下航行器的状态信息及环境信息映射为对航向和速度的控制指令，实现信息的端到端处理。在整个探索过程中实现避障，对自主水下航行器而言是一项具有挑战性的任务。由于所提出的方法是一种无模型方法，因此无需依赖数学模型，并基于被控对象在未知环境中的行为构建了智能控制器。避障通过两种方法实现。第一种方法是为所有自主水下航行器配备避障模块。当障碍物进入各自主水下航行器的探测范围时，便将其纳入相应的马尔可夫决策过程模型，并据此调整方向和速度以避开障碍物。第二种方法则仅由领航者负责确保安全。当障碍物出现并

进入领航者的马尔可夫决策过程模型时，考虑到领航者能够掌握编队规模，便会选择一条路径，使整个群体都能避开障碍物。

本文结构安排如下：第 2 节介绍问题的基础知识与建模，包括自主水下航行器运动模型、深度强化学习基础知识以及编队控制架构。第 3 节介绍深度强化学习的结构，包括状态空间、动作空间和奖励方案的定义。第 4 节介绍用于多自主水下航行器编队控制和避障的深度强化学习算法细节。第 5 节展示各种场景下的仿真与实验。第 6 节总结全文。

2. 问题的基础知识与建模

2.1. 自主水下航行器运动模型

本节旨在描述自主水下航行器的运动学和动力学模型。为描述自主水下航行器的运动，根据图 1 定义了两个坐标系：体固定坐标系 {B} 和地固坐标系 {E}。水平面上的自主水下航行器编队适用于恒定深度作业，例如海床探测。带有表 1 所示符号说明的 3 自由度 REMUS自主水下航行器动力学与运动学模型可描述如下 [40]:

2 2 ( ) ( ) 螺旋桨推力 m X u my r m vr x r X u u X vr X r X g g vr rr u u u− · · − · − + = + + +))

2 2 ((( ) g r m Y v mx Y r m ur y r Y v v Y r r Y ur Y uv Y u δ g ur uv r v v r r v r uu

− · + − · + − = + + + + δ)) 2 ((( ) r I N r mx N v my u m x ur y vr N v v N r r N ur N uv N u

− · + − · − + + = + + + + δ δ z g g g g ur uv r v v r r r v uu (1) cos( )x · = ψsin( )u v− ψsin( ) cos( ) y u v

· = ψ + ψrψ =其中，[, ,]Tx yψ 是自主水下航行器在地固坐标系中的位置向量，其中包含 ( )x y 坐标和偏航角 [0), ψ ∈ 2π。自主水下航行器在体固定坐标系中的速度向量为 [, ,]Tu v r，其中 u 为纵荡速度，ν 为横荡速度，r 为偏航角速度。[,] propX rδ 是输入向量，包括螺旋桨推力（propX）和舵角偏转量（rδ）。

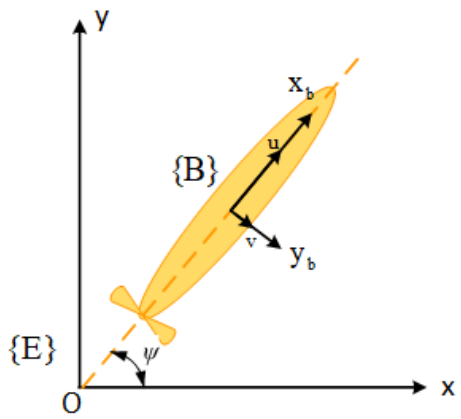

图 1. 自主水下航行器运动模型

表 1. 自主水下航行器系数定义

参数	Unit	描述	参数	Unit	描述
$X_{u\|u\|}$	/千克 米	横流阻力	Y_{ur}	/千克 弧度	附加质量交叉项和
$N_{v\|v\|}$	kg		N_{ur}	. /千克米 弧度	尾翼升力
$Y_{v\|v\|}$	/千克 米		N_{uv}	kg	艇体与鳍升力和芒克力矩
$Y_{r\|r\|}$	². /千克米 弧度		Y_{uv}	/千克 米	机身升力和尾翼升力
$N_{r\|r\|}$	² ². /千克米 弧度		$_{uu}Y_{\delta}^{'}$	/(.)千克 毫弧度	尾翼升力
$X_{\dot{u}}$	kg	附加质量	N_{δ}^{uu} $_{r}$	/千克 弧度	舵鳍升力矩
$N_{\dot{v}}$	$.kg\,m$		m	kg	车辆质量
$N_{\dot{r}}$	². /千克米 弧度		I_z	². 千克	车辆转动惯量绕z轴
X_{vr}	/千克 弧度	附加质量交叉项	,	米	x和y方向上的重心
X_{rr}	。/kg m 弧度		$_{g}$ 克克斯 y	M	相对于.....的方向浮心

2.2 深度强化学习方法

强化学习（RL）是一种决策框架，智能体通过与环境直接交互来学习期望的行为或策略（图2）。在每个时间步，智能体在状态s下执行动作a。随后，它进入新状态s'，并从环境中获得奖励r。采用马尔可夫决策链对强化学习环境进行建模，可表示为一个由4个元素组成的元组[, , ,]$sarp$。其中，s表示环境的动态行为，a表示智能体执行的动作，r表示环境给出的奖励，p表示状态转移概率函数。经过与环境的长期交互，智能体学会了能使总体期望回报最大化的最优策略。期望回报定义为未来奖励的折扣和，如下所示:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+(k+1)} \tag{2}$$

其中，$\gamma \in [0\ \ 1]$是折扣因子，用于确定未来奖励的当前价值。折扣期望状态价值函数定义如下：从状态s开始并遵循π 策略：

$$V^{\pi}(s) = E_{\pi}[R_t \mid s_t = s] = E_{\pi}[\sum_{k=0}^{\infty} \gamma^k r_{t+(k+1)} \mid s_t = s] \tag{3}$$

类似地，动作价值函数表示在π 策略下，状态s中动作a的价值，定义如下：

$$Q^{\pi}(s,a) = E_{\pi}[R_t \mid s_t = s, a_t = a] = E_{\pi}[\sum_{k=0}^{\infty} \gamma^k r_{t+(k+1)} \mid s_t = s, a_t = a] \tag{4}$$

最优状态价值函数给出智能体从状态 s 开始并执行最优策略时所能获得的最大折扣奖励。

$$V^*(s) = \max_a E[r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a] \tag{5}$$

上述方程称为贝尔曼最优方程，它指出，最优策略下某个状态的价值应等于该状态下最佳动作的期望回报。类似地，动作价值函数的贝尔曼方程如下：

$$Q^*(s,a) = E[r_{t+1} + \gamma \max_a Q^*(s_{t+1}, a') \mid s_t = s, a_t = a] \tag{6}$$

其中，对于所有 s，$V^*(s) = \max_a Q^*(s,a)$ 是。通过交互确定Q^* 后，可直接根据以下方程得到最优a策略：

$$\pi^*(s) = \arg\max_a Q^*(s,a) \tag{7}$$

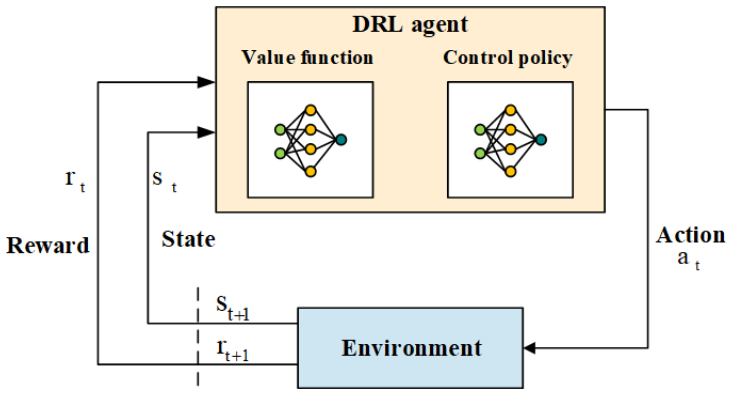

图2. 马尔可夫决策过程中深度智能体与环境的交互

标准强化学习方法基于离散状态与动作空间。因此，状态价值或动作价值的数值都记录在表格中。这类方法的一项主要优势是，总能找到精确优化的价值函数和最优策略 [25]。基于表格的强化学习算法在空间规模较大或连续的问题中会遭遇维数灾难。为解决这一问题，人们采用深度等函数逼近器的组合

神经网络。深度神经网络由多个顺序层构成，可将高维输入转换为维度更低的输出特征；网络参数则通过监督式方法进行训练，该方法采用反向传播算法中的梯度下降法等训练算法。因此，无需保留包含状态-动作价值函数数据的表格，也就不需要大量内存或关于环境的精确信息。尽管深度Q网络（DQN）算法能够解决高维观测空间问题，但它只能处理规模较小的离散动作空间 [41]。由于大多数物理控制任务具有连续且高维的动作空间，因此引入了采用演员-评论家结构的深度确定性策略梯度（DDPG），以学习高维连续动作空间中的策略 [42]。DDPG算法使用两个深度神经网络来逼近状态-动作价值函数和策略，并结合回放缓冲区和目标网络的思路。回放缓冲区用于打破采样数据之间的马尔可夫特性，两个目标网络则用于保证训练过程的稳定性。尽管DDPG算法表现理想，但由于当前策略过拟合，以及超参数和其他设置的影响，它在复杂环境中的表现仍不尽如人意。因此，提出了双延迟深度确定性策略（TD3）算法，通过采用双评论家网络、平滑目标策略以及策略更新延迟来提升学习速度并改善性能 [43]。

*2.2.1 TD3*算法

TD3算法是一种基于DDPG算法的在线强化学习算法。该方法采用类似于双重DQN的方法，减少函数逼近中的过拟合，并延迟演员网络的更新，同时通过向目标演员网络添加噪声来消除DDPG网络的敏感性和不稳定性[43]。该算法的结构框图如图3所示。

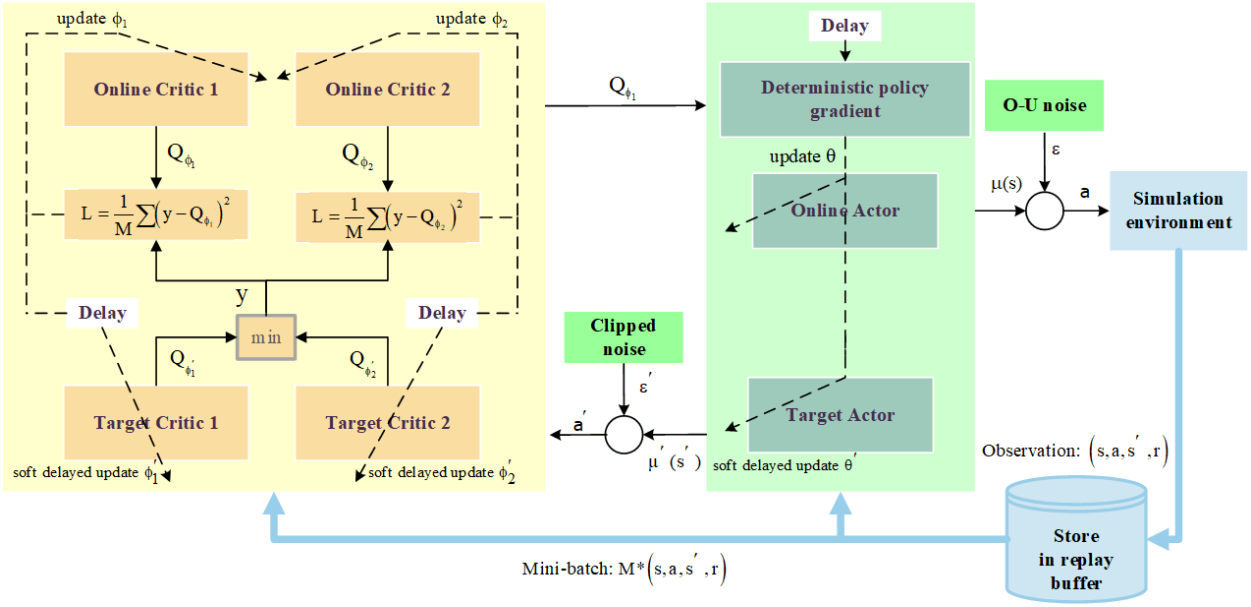

图3. 基于演员-评论家结构的TD3框架

由图可知，该算法基于演员-评论家架构，并具有双评论家网络。通过选择动作并将其应用于环境，将更新网络所需的状态转移存储在缓冲区中。当缓冲区中存储了一定数量的状态转移后，从中采样一个小批量的状态转移以寻找策略。评论家网络利用这些缓冲区样本，基于两个评论家中较小的目标价值函数，通过梯度下降法最小化贝尔曼误差的均方值，计算在线评论家网络的状态-动作价值函数，从而更新网络参数。随后，演员网络利用计算得到的价值函数，基于确定性梯度定理，以延迟更新的方式更新策略网络参数，实现期望回报优化。重复上述过程，直至获得该环境下的最优策略。

2.3 多*AUV*编队控制与避障架构

如图4所示，本文介绍的领导者-跟随者编队中的自主水下航行器可分为三组，包括：领导者自主水下航行器，它不跟随任何其他自主水下航行器，并尝试学习一种策略，帮助其在尽可能短的距离 内到达目的地，同时避开沿途可能遇到的障碍物或其他自主水下航行器；以及位于领导者左侧或右侧的跟随者，它们需要找到一种最优策略，以与领导者保持预定距离和角度，并在整个任务过程中维持这些数值；如果路径上出现障碍物，它们会避开障碍物，也会避免彼此碰撞。每个跟随者都可以充当新加入编队的其他自主水下航行器的领导者，以扩大编队规模。自主水下航行器可以配备各种机载传感器。可考虑配备深度传感器、惯性测量单元（IMU）、罗盘、多普勒测速仪（DVL）、声呐传感器、通信模块，以及用于水面校正的全球定位系统（GPS）。IMU可测量方位角、三轴角速度和三轴加速度。罗盘可指示基本地理方向。DVL用于测定水下速度。声呐传感器利用声学探测障碍物。为实现编队控制，自主水下航行器应通过无线通信交换必要信息。跟随者需要跟踪领导者。因此，它们不需要配备领导者所用的高级传感器。本文探讨了多自主水下航行器领导者-跟随者编队控制所面临的挑战。我们提出了一种新型编队控制算法，第一种运动规划方法采用单向信息流，第二种则采用双向信息流。在第一种方法中，跟随者从领导者自主水下航行器获取信息，但领导者不会反向获取信息。为避开障碍物，所有自主水下航行器均配备前视声呐。领导者自主水下航行器将持续向两台跟随者自主水下航行器传送其实时状态，以使它们与领导者保持期望距离和方向。在第二种方法中，领导者和跟随者自主水下航行器彼此交换信息。只有领导者配备声呐传感器来探测障碍物，跟随者只需跟随领导者。

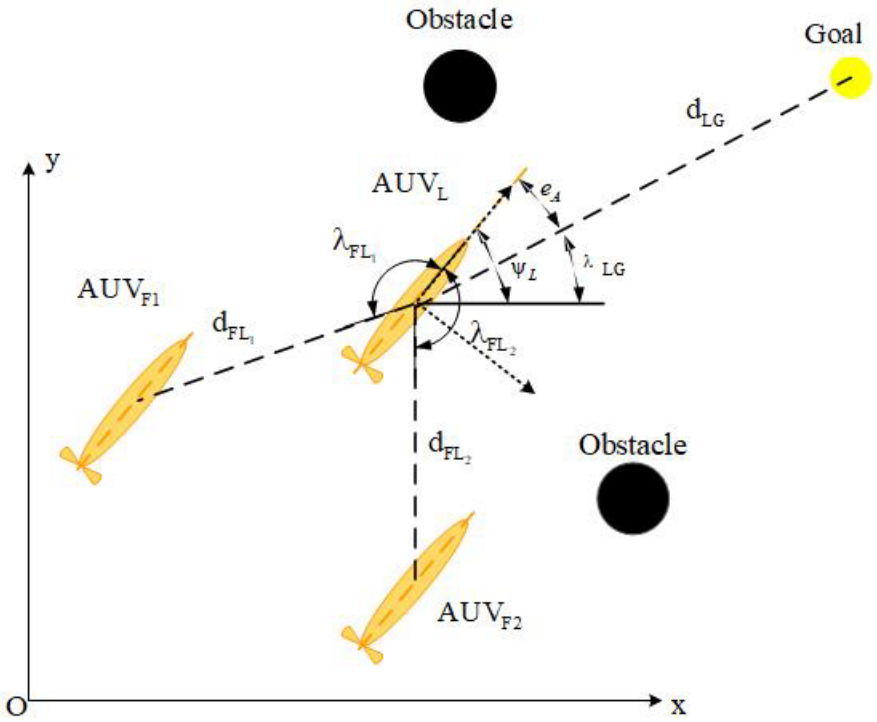

图4. 多自主水下航行器编队控制的领航者-跟随者架构

3. 多自主式水下航行器编队运动规划的深度强化学习结构

图5展示了基于TD3的分布式编队模型，该模型用于欠驱动自主水下航行器的编队控制。编队控制的训练模型是一个与环境反复交互的过程。通过观察环境状态()s，每台自主水下航行器都可以作出

作出合理的决策。结合任务目标与状态观测，可以评估动作选择。根据评估值，可以训练多台自主水下航行器的编队运动规划。最终，模型在相应更新后的环境中执行所选动作，并重新获取观测状态。重复上述过程，直到编队运动规划模型选出预期动作。通过这种方式，模型能够学习如何为多台自主水下航行器的编队控制作出理想的控制决策。如图5所示，决策网络()a 的输出是各自主水下航行器的动作。每台自主水下航行器执行动作后都会转移到新状态。随后，通过评估自主水下航行器的行为生成奖励()r。最后，将每台自主水下航行器的数据存入相应的经验池，以便提取经验数据来构建训练数据集。经验池用于存储每台自主水下航行器的导航经验记录。单次执行中的状态数据、决策网络的动作以及相应奖励共同构成一条经验记录。每台自主水下航行器都可以通过决策网络进行训练，以规划编队路径并避开障碍物。

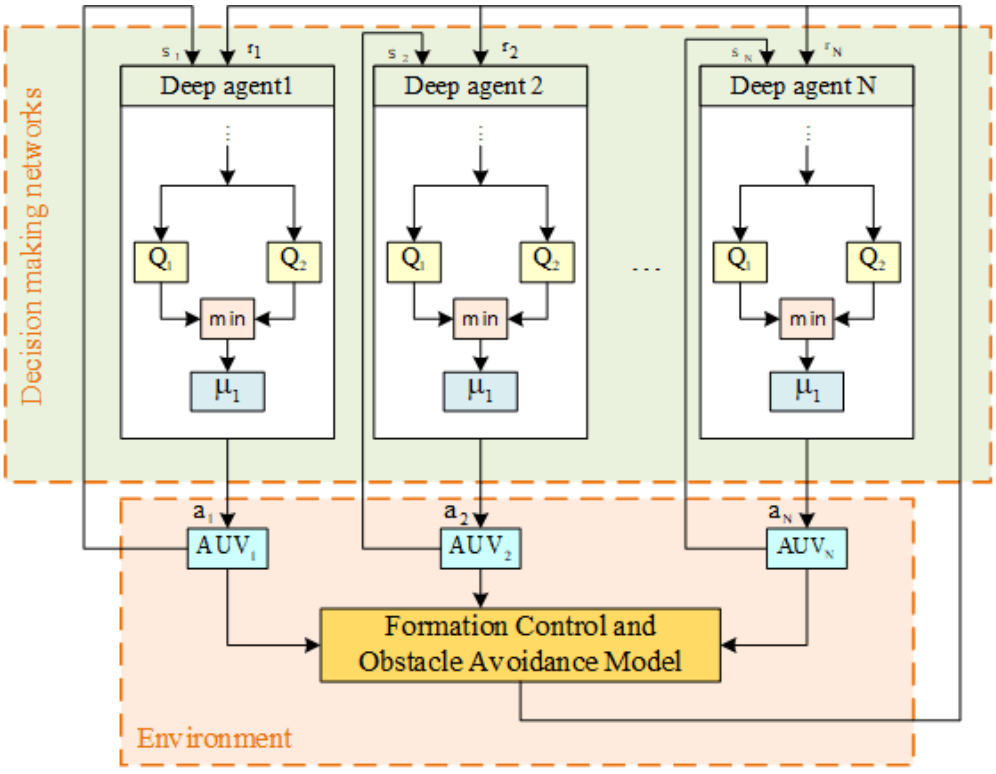

图5. 用于编队控制与多自主水下航行器避障的分布式深度强化学习结构 s

在避障过程中，采用两种方法。第一种方法是由编队中的每台自主水下航行器分别作出避障决策。根据所设计的奖励函数，跟随者自主水下航行器学习在保持编队的同时避开障碍物。遇到障碍物时，避免与其碰撞的优先级高于编队保持。第二种方法是仅由领航者负责作出与障碍物相关的决策。领航者识别障碍物，并在了解编队规模的基础上，学习适用于使整个群体远离障碍物的策略。根据上述两种方法，面对障碍物时会考虑不同的马尔可夫决策过程 。

3.1 设置状态空间和动作空间

为了基于领航者-跟随者架构实现欠驱动自主水下航行器的编队控制与避障，首先应介绍状态空间和动作空间。为每个自主水下航行器定义一个基于领航者-跟随者架构的马尔可夫决策过程。根据各自主水下航行器在编队架构中的角色，设计相应的状态空间。利用自主水下航行器的机载传感器识别障碍物。根据自主水下航行器传感器的探测范围，将与已识别障碍物之间的距离加入每个智能体的马尔可夫决策过程中。

3.1.1 为领航者设计状态空间

领航者的目标是沿尽可能短的路径安全抵达目的地。因此，状态空间基于领航者相对于目标的方向误差、与已识别障碍物的距离以及速度向量。

$$S_{Leader}=\left[e_A,\nu^T,\frac{d_{LO}-r_{det}}{r_{det}} \right], \tag{8}$$

$d_{LO}-r_{det}$ 值（如果障碍物在自主水下航行器的传感器范围内，则进行检测）$=0$ 否则 r_{det}

其中，$e_A=\psi_{LG}-\lambda_L$ 表示领航者朝向目标的方向误差，而 λ_{LG} 表示相对于x轴的目标角（图4）。ψ_L 表示领导者AUV的航向角。AUV的速度向量为 $\nu=[u,v,r]^T$。d_{LO} 表示从已识别障碍物到该处的距离 $1\times N$ 向量，N 表示障碍物的数量。r_{det}为AUV传感器的最大探测范围。领航者的两种避障方法具有相同的状态空间。

3.1.2 为跟随者设计状态空间

跟随者旨在学习一种策略，以便在面对障碍物时避开障碍物，同时与领航者保持预定距离和角度。为此，跟随者的状态空间定义如下：

$$S_{Follower}=\left[e_d,e_A,\psi_L,\psi_F,\frac{V_L-V_d}{V_d},\frac{V_F-V_d}{V_d},\frac{d_{LO}-r_{det}}{r_{det}} \right] \tag{9}$$

$d_{LO}-r_{det}$ 如果障碍物处于AUV的传感器探测范围内，则值为 0；否则为 r_{det}

其中 $e_d=\overline{d_{FL}-d_{desired}}$ 表示距离缩放误差，d_{FL} 和 $d_{desired}$ 分别表示领航者与跟随者之间的距离以及期望距离。$e_A=\lambda_{FL}-\lambda_{desired}$ 表示编队角度误差，其中 λ_{FL} 是领航者与跟随者之间连线相对于领航者运动方向的夹角，$\lambda_{desired}$ 是编队期望角度。ψ_L 和 ψ_F 分别是领航者和跟随者的航向角。$V_{iLF}=\sqrt{u_i^2+v_i^2}$ 表示合速度，而 V_d 表示期望编队速度。d_{LO} 表示与已识别障碍物距离的 $1\times N$ 向量，N表示障碍物数量。r_{det}表示自主水下航行器传感器的最大探测范围。在状态空间中进行避障的第二种方法会从马尔可夫决策过程中去除障碍物元素，因为障碍物识别完全由领航者执行。

3.1.4 领航者与跟随者的动作空间

如前所述，领航者旨在安全抵达目的地，同时使跟随者根据领航者保持编队，并在整个导航过程中避开障碍物，这需要进行速度和方向控制。通过控制螺旋桨推力$_{prop}X$ 和舵角$_{r}\delta$来实现这一点。

$$A=\left[X_{\text{prop}},\delta_{r} \right] \tag{10}$$

3.2 自主水下航行器编队的奖励塑造

自主水下航行器编队与避障的奖励函数采用两种不同的方法进行设计。奖励函数的基本作用是鼓励预期行为并惩罚非预期行为。为领航者和跟随者定义了三类奖励函数。领航者的奖励函数单独设计，以便在避开障碍物并避免与其他自主水下航行器发生碰撞的同时，找到通往目标的最优路径。对于跟随者，奖励函数用于塑造并维持编队、避开障碍物，以及避免与其他自主水下航行器发生碰撞。在第二种避障方法中，跟随者奖励函数移除了避障部分。

3.2.1 设计领航者的奖励

1）目标奖励：

该奖励确保领航者能够到达目的地。

$$r_{A}=-\left|\lambda_{LG}-\psi_{L}\right| \tag{11}$$

其中$_{LG}\lambda$ 是与领航者之间的视线角目标，ψ_{L} 是领航者的航向角。

2）障碍物避障奖励：

方法1： 在该方法中，每个自主水下航行器学习一种用于避障的策略。如图6a所示，首先为自主水下航行器模型构造一个圆形外壳。外壳与障碍物发生重叠表示发生碰撞。将传感器探测范围内所有障碍物的缩放距离输入第3.1.1节中相应自主水下航行器的马尔可夫决策过程模型，因此，奖励函数定义如下：

$$r_{OA1}=\sum_{i=1}^{\ell}r_{i},\qquad r_{i}=\begin{cases}0 & \text{if } d_{AO}>d_{avoid}\\ -\left|d_{avoid}-d_{AO}\right| & \text{otherwise}\end{cases} \tag{12}$$

其中，ℓ 表示在自主水下航行器传感器的探测范围内检测到的障碍物数量。$_{AO}d$ 是自主水下航行器与障碍物之间的距离，$=\quad+\quad+_{avoid\,A\,safe\,ob}drdr$，其中 $_{A}r$ 是自主水下航行器的外壳半径，$_{safe}d$ 是自主水下航行器外壳到安全距离边界的宽度，$_{ob}r$ 是障碍物半径。

如果自主水下航行器与障碍物之间的距离大于d_{avoid}，则奖励为零；否则，将受到惩罚。

方法二：在此方法中，编队运动规划完全由领航者执行。跟随者的任务是与领航者保持期望距离和角度。根据图6b，考虑一个三角形，其顶点为领航者以及左侧和右侧跟随者的位置。当障碍物进入领航者的探测范围时，根据领航者和跟随者的位置计算该三角形外接圆的圆心。利用外接圆圆心到障碍物的距离，奖励函数定义如下：

$$r_{OA2}=\sum_{i=1}^{\ell} r_i, \qquad r_i=\begin{cases} 0 & \text{if } d_{CO}>d_{avoid} \\ -\left|d_{avoid}-d_{CO}\right| & \text{otherwise} \end{cases} \tag{13}$$

其中，d_{CO} 表示外接圆圆心到障碍物的距离，$d_{avoid}=r_{cir}+d_{safe}+r_{ob}$，其中 r_{cir} 为外接圆半径，d_{safe} 为自主水下航行器重心到安全距离边界的宽度，r_{ob} 为障碍物半径。

备注：在第一种方法中，所有自主水下航行器都配备了声纳探测传感器，以便在遇到物体时能够做出正确决策。除了保持期望距离和方向外，跟随者还必须避开障碍物，并在穿越危险区域后恢复到期望编队。每个自主水下航行器都配备了避障模块，因此这种策略可能更加安全。而在第二种方法中，只有领航者配备了障碍物检测传感器，跟随者则保持期望距离和方向。当领航者检测到障碍物时，如果根据跟随者的位置判断队伍有可能与障碍物发生碰撞，领航者就会改变队伍的运动路径，并尝试带领整个队伍离开危险区域。在这两种方法中，跟随者都需要领航者的状态信息来形成恰当的编队。必要的数据通过安装在自主水下航行器上的低成本无线调制解调器传输给跟随者。在第一种方法中，领航者无需了解跟随者的状态，因此信息以单向方式传输。在第二种策略中，领航者还需要获取跟随者的位置信息，双方以这种双向方式交换数据。这种方法也更适用于集中式编队框架。总体而言，这两种方法相似，但实现方式不同，可根据系统设计选择其中一种。仿真部分表明，所提出的两种避障方法均能正常工作。

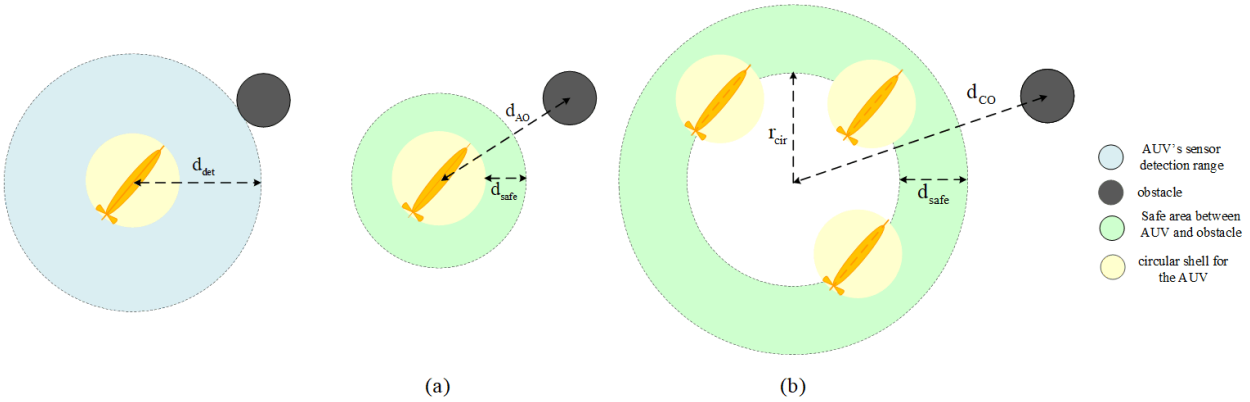

图 6. 两种避障方法：(a) 方法 1：每个自主水下航行器都能在传感器的检测范围内感知障碍物，并获取与障碍物的相对距离。(b) 方法 2：只有领航者能感知障碍物

3）自碰撞规避奖励：

$$r_{CA}=\begin{cases}0 & \text{if } d_{AA}>d_{avoid}\\ -\left|d_{avoid}-d_{AA}\right| & \text{否则}\end{cases} \tag{14}$$

其中，d_{AA} 表示两个自主水下航行器之间的距离（图 7）。该奖励用于鼓励自主水下航行器彼此避免碰撞。2( ) *avoid* *A* *safe* $d\ r=$ + 。

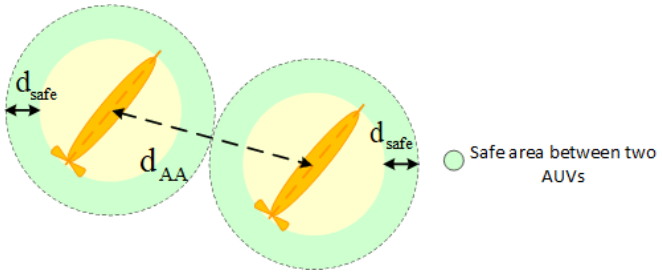

图 7。 碰撞规避：如果两个自主水下航行器之间的距离小于安全距离，它们将受到负奖励

4）降低控制努力：

自主车辆控制器设计中的一项实际挑战是执行器饱和 [44]。因此，为了考虑实际约束，采用以下项来降低总体控制努力：

$$r_E=[-\left|X_{prop}\right|,\ -\left|\delta_r\right|] \tag{15}$$

其中 螺旋桨X 和 $r\delta$ 分别表示螺旋桨推力和舵角。

5）总奖励：

总奖励函数定义为上述所有奖励函数的加权和：

$$r_L = w_A r_A + w_{OA} r_{OA} + w_{CA} r_{CA} + [w_{E1},\ w_{E2}] r_E^T \tag{16}$$

其中Aw、OAw、CAw、$1Ew$ 和$2Ew$ 为正定常数。上述每个增益表示各部分在控制策略中的重要性。

3.2.2 跟随者奖励函数设计

对于跟随者而言，其目标是学习一种策略，使其在跟踪领航者的同时，保持与领航者之间的期望距离和角度，并避开障碍物及其他自主水下航行器。避开障碍物、避免与其他自主水下航行器发生碰撞以及限制控制信号幅值所对应的奖励，分别根据第3.2.1节中的公式（13）、公式（15）和公式（16）计算。因此，编队奖励函数如下：

1）编队距离奖励：

$$r_{Fd} = -\left| \frac{d_{FL} - d_{desired}}{d_{desired}} \right| \tag{17}$$

领航者与跟随者距离 表示跟随者与相应领航者之间的距离。期望编队距离表示为期望距离 。

2) 编队角度奖励：

$$r_{FA} = -\left| \lambda_{FL} - \lambda_{desired} \right| \tag{18}$$

领航者与跟随者λ 表示编队角度（连接领航者与跟随者的直线和领航者方向之间的夹角）（图5），期望 λ 表示期望编队角度。

3) 总奖励：

总 奖励函数定义为所有奖励函数的加权和 s:

$$r_F = w_{Fd} r_{Fd} + w_{FA} r_{FA} + w_{OA} r_{OA} + w_{CA} r_{CA} + [w_{E1},\ w_{E2}] r_E^T \tag{19}$$

其中，Fdw 、FAw 、OAw 、CAw 、$1Ew$和$2Ew$ 均为正常数。上述各权重分别表示控制策略中各部分的重要性。在第二种避障方法中，移除了OAr 项。

根据第 3.1.1 节和第 3.1.2 节收集观测。然后，根据所提出的编队控制模型选择动作，为自主水下航行器的

执行器生成控制指令并执行。相应地，评估领航者的目标到达误差以及跟随者为实现期望控制目标所产生的编队误差。重复这些步骤，直到自主水下航行器以期望的编队安全抵达目的地。在控制周期内，自主水下航行器依次执行环境感知、决策和执行过程。通过环境感知，可以观察编队中自主水下航行器的状态，并据此更新编队中自主水下航行器的情况。

4. 深度强化学习算法：多自主水下航行器的编队控制与避障 **s**

所提出的编队控制与避障模型使用 TD3 算法为每个智能体寻求最优策略。在回合开始之前，完成所有设置和编队配置，包括确定领航者和跟随者自主水下航行器、期望角度和距离、训练区域的尺寸、训练率、演员网络的探索噪声、目标网络的平滑噪声等。在每个回合中，速度向量变量和自主水下航行器的位置都会重置。随机选择一个目标区域，并在训练区域内随机分布障碍物。在该算法中，评论家包含四个网络，其中两个结构相似，即参数为 ϕ_1 和 ϕ_2 的在线网络，以及参数为 ϕ_1' 和 ϕ_2' 的目标网络。演员包含两个结构相似的在线网络和目标网络，其参数分别为 θ 和 $'\theta$。为了在选择策略时对状态空间和动作空间进行适当探索，噪声 N_t 按如下方式添加：

$$a_t = \mu_\theta(s_t) + N_t \tag{20}$$

在每个时间步，通过对自主水下航行器施加动作，根据公式（9）获得新的状态 $_{t+1}$ s，并根据为控制目的而设计的奖励函数获得即时奖励 $_t$r；该奖励与所施加的动作及系统当前状态一同存储在经验缓冲区（D）中。在线演员网络的训练过程根据从缓冲区中随机采样的数据 $()_{iiii+1}$ M * s , a , r ,s，按照以下公式进行更新，其依据是确定性策略梯度，且更新频率低于评论家网络。这样，在用于训练目标网络之前，评论家网络会更加稳定，并且误差也会减小。

$$\nabla_\theta J(\theta) = \frac{1}{M}\sum_{s\in M}\nabla_\theta Q_{\phi_1}(s, \mu_\theta(s)) \tag{21}$$

其中，M 为小批量的大小。$_1Q_\phi$是参数为ϕ_1的在线评论家网络的价值函数。通过最小化基于较小目标价值函数的贝尔曼误差均方值来更新ϕ_1 参数。

$$L(\phi_k) = \frac{1}{M}\sum_{i=1}^{M}\left(y_i - Q(s_i, a_i \mid \phi_k)\right)^2 \quad , k = 1, 2 \tag{22}$$

其中，y_i是由目标深度网络生成的较小目标状态-动作值，目标深度网络为$Q_{\phi_1'}$ 和$Q_{\phi_2'}$ ，其参数分别为ϕ_1' 和ϕ_2'，其中包括：

$$y_i = r + \gamma \min_{j=1,2} Q_{\phi_j'}(s_{i+1}, a'(s_{i+1})) \tag{23}$$

将平滑正则化策略应用于目标策略，并向目标演员网络针对状态 s_{i+1} 的确定性输出 a' 添加截断噪声，可防止更新评论家网络时出现高方差。在线评论家网络的参数 ϕ_1 和 ϕ_2采用随机梯度下降方法更新。按照与演员策略函数更新频率相同的频率，对目标演员网络的参数θ' 以及目标评论家网络参数 ϕ_1'和 ϕ_2'进行软更新，以提高学习过程的稳定性。每个回合都重复此过程，直到领航者自主水下航行器到达目的地，或回合步数达到上限。因此，本文使用 TD3 算法实现多自主水下航行器的编队运动规划，具体见算法1。为估计深度强化学习算法中演员网络和评论家网络的价值函数（状态-动作）及策略，本文采用了具有两个隐藏层的架构。

5. 仿真结果

为验证所提方法的有效性，在多种精心设计的场景中对欠驱动自主水下航行器的运动规划和编队控制进行了评估。为此，考虑了航路点跟踪中的编队保持以及存在障碍物时的编队保持。此外，还评估了海流、可变通信延迟和导航误差存在时算法的性能和鲁棒性。本节将讨论所获得的仿真结果。

5.1. 仿真设置

5.1.1. 仿真环境设置

采用 AMD Ryzen 7 3800XT 8 核 3.89 GHz CPU 处理器和 NVIDIA GeForce RTX 2060 GPU，在 MATLAB 中运行算法1，以用于三自由度REMUS模型自主式水下航行器。模型参数的取值如表2所示。训练区域为 500 × 500 米^2 正方形。自主水下航行器置于该正方形的中心区域。目标是训练区域边界处半径为3米的圆形区域。每个长度为1米的自主水下航行器均内接于半径为0.5米的圆形中。仿真中使用的动力学模型参数$_i$自主水下航行器（$i = 1, 2, 3$，，）取自 [40]。控制信号的允许范围为$[3 \quad]$、$_{prop} \in \quad 13X$（N）和$\delta_r \in [-20 \quad 20]$（度）。期望编队速度为1.5米/秒。障碍物在训练区域内半径为14米的范围内随机分布。

5.1.2. 机器学习算法训练与参数设置

演员-评论家网络的结构相似，即采用了具有 400 个和 300 个神经元的双层全连接网络，并使用 ReLU 激活函数。这些取值是根据状态和动作空间的维度确定的，足以

表 2. REMUS自主水下航行器系数的取值 [40]

参数	值	Unit	参数	值	Unit
$X_{u\|u\|}$	-1.62e+000	/千克·米	X_{rr}	-1.93e+000	. /kgm 弧度
$N_{v\|v\|}$	-3.18e+000	kg	Y_{ur}	5.22e+000	/千克 弧度
$Y_{v\|v\|}$	-1.31e+000	/千克 米	N_{ur}	-2.00e+000	. /千克米 弧度
$Y_{r\|r\|}$	6.32e-001	². /千克米 弧度	N_{uv}	-2.40e+001	kg
$N_{r\|r\|}$	-9.40e+001	² ². /kgm rad	Y_{uv}	-2.86e+001	/千克·米
$X_{\dot{u}}$	-9.40e-001	kg	$N_{u\|u\|\delta}$	9.64e+000	/(.)千克·米弧度
$Y_{\dot{v}}$	-3.55e+001	kg	$N_{uu\delta}$	-6.15e+000	/千克 弧度
$N_{\dot{v}}$	1.93e+000	.kg m	m	30.51e+000	kg
$N_{\dot{r}}$	-4.88e+000	². /kgm rad	I_z	3.45e+000	²。千克·米
X_{vr}	3.55e+001	/千克·弧度	，g 克乘以 y	0.0e+000	M

近似本研究中的状态-动作函数和策略函数。利用奥恩斯坦-乌伦贝克过程噪声选择动作，以充分探索状态和动作空间。噪声过程定义如下：

$$N_{k+1} = N_k + \alpha_{OU}(\mu_{OU} - N_k) + \sigma_{OU} N_G(0,1) \tag{24}$$

其中，N_k 和 N_{k+1}分别表示奥恩斯坦-乌伦贝克过程在时刻 k 和 k+1的取值。(N_G 0,1) 是均值为零、标准差为 1 的高斯噪声。$_{OU}\alpha$ 、$_{OU}\mu$和 $_{OU}\sigma$ 是奥恩斯坦-乌伦贝克过程的参数。$_{OU}\alpha$ 是一个常数，用于确定噪声输出趋近均值的速度。$_{OU}\mu$ 和 $_{OU}\sigma$ 分别是噪声模型的均值和方差。基于编队运动规划和自主水下航行器控制的 TD3 参数取值列于表 3。

自主水下航行器在每个回合中最多可移动 300 步。自主水下航行器的动作是纵荡力和偏航力矩，因此自主水下航行器的航向和位置根据公式（1）和公式（2）更新。当领航者 AUV到达目标区域或每个回合的全部时间步均已结束时，回合终止。

表 3：网络参数

参数	值
演员网络学习率	0.001
评论家网络学习率	0.0001
折扣因子	0.99
记忆容量	1e6
平滑更新	0.005
策略与目标延迟更新	2
目标策略噪声方差	0.2
探索方差	0.1
采样时间	0.1
μ_{OU}	0
α_{OU}	0.15

Alg法1：Ada p基于深度强化学习的自适应编队控制与避障

1. 配置期望编队、初始位姿 [0， 0,ψ 0] x y 、回合数 T、每个回合的步数 N、经验回放缓冲区大小 D 、决策网络的学习率 γ2. 初始化在线评论家网络 1Qϕ、2Qϕ 和在线演员网络 θμ ′ 3. 初始化目标评 Q 、2Qϕ′ 和目标演员网络 ′θμ 4. 对于回合 = 1，执行 T 次 5. 初始化一个用于动作探索的随 I 6. 接收初始观测状态（公式（8）和公式（9））7. 对于 t = 1，执行 N 次 8. 根据当前策 声选择动作 t t a (s |) N= μ θ + 9. 在自主水下航行器上执行动作 ta，观测新状态 s_{t+1}（公式（1）），并获得奖励 r_t（公式（16）和公式（19））10. 将状态转移 ( ) t t t t +1 s , a , r , s 存入经验回放缓冲区 D 11. 从经验回放缓冲区 D 中采样一个随机小批量 (( ) i i i i 1M * s , a , r ,s +) 12. 对于 i =1，执行 M 次 13. 计算 ( ) i 1 i 1 a s | 截断噪声 + + ′ ′ ′=μ θ +14. 使用公式 ( )γ =1,2 ϕ′ +1 +1 | ϕmin , ′ ′ = + j i i i i j jy r Q s a 计算目标状态-动作值 15. 结束循环 16. 通过最小化贝尔曼方程误差的代价函数更新评论家网络 ( ) ( )

$\phi_1 = + \phi_1 \quad 1 \sum_{=1}^{M} - \phi_1 \quad | \phi_1$ arg min , i i i i i r y Q s aM ( ) ( )

$\phi_2 = + \phi_2 \quad 1 \sum_{=1}^{M} - \phi_1 \quad | \phi_2$ arg min , i i i i i r y Q s aM 17. 如果 t mod d 成立，则 18. 使用采样确定性策略梯度 ( ) ( )( ) 11 ,$\nabla_\theta \quad \theta = \sum_{s \in M} \nabla_\theta \quad \phi \quad {}_\theta$J Q s sM μ 更新在线演员策略 (θ)，更新频率低于评论家网络；软更新目标网络 ( )l ′ , 1,2′ ′ ′ = + − =i i i，其中 i ϕ τϕ τ ϕ( ) l′ ′θ =τθ + −τ θ19. 结束条件判断 20. 结束循环 21. 结束循环

—

—

—

5.2. 运动规划与编队控制：方法***1***

针对三角形编队，考虑三个自主水下航行器。领航者的初始位置（L自主水下航行器）位于坐标（250,250），而跟随者自主水下航行器（F1自主水下航行器）和（F2自主水下航行器）分别位于坐标（220,220）和（280,220），并将自主水下航行器的航向角设为零。障碍物随机分布在训练区域内。目标是训练自主水下航行器，使其以25米的期望距离和150度（2.618弧度）的期望角度组成编队，并安全到达目标点。每个自主水下航行器均配备一个避障

模块。所有图中的（L）副标题表示领航者，而（$_1$F）和（$_2$F）副标题分别表示左侧跟随者和右侧跟随者。图8展示了三个智能体的训练示意图。学习曲线表示每个回合中的累积奖励以及所有回合的平均值。从图中可以看出，平均奖励随着训练回合的增加而逐渐提高，且奖励值趋于稳定。

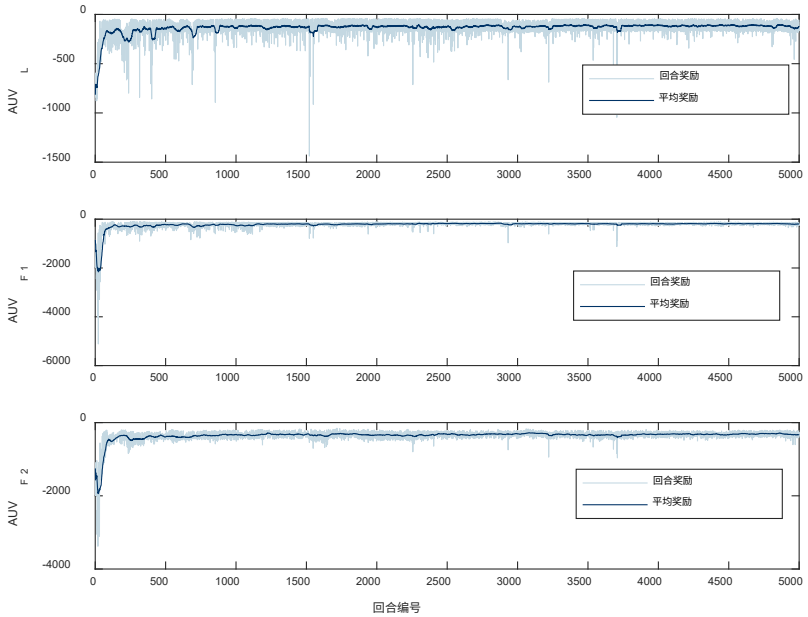

图 8。 训练结果：第一种方法的奖励及每回合平均奖励

5.2.1. 无障碍编队运动规划结果

领航自主水下航行器负责到达目标，跟随自主水下航行器则尝试通过与领航者保持特定距离和角度来形成预定义的编队结构。图 9 和图 10 展示了在无障碍物情况下到达不同目标的各种场景。场景1中展示了距离和角度误差信号。根据图示，在不同的运动场景中均实现了期望编队。

5.2.2. 存在障碍物时的运动规划与编队控制

本节评估存在障碍物时的编队运动规划。图11展示了不同场景下自主水下航行器编队的轨迹及避障情况。在场景1中，根据目标和障碍物的位置，领航者和左侧跟随者可能发生碰撞。它们检测到障碍物后，通过改变运动方向来避免与障碍物碰撞，并在通过危险区域后返回期望的编队模式。在第二个场景中，根据目标位置、障碍物以及编队的尺寸，自主水下航行器的跟随者仅略微改变方向便到达目标。在第三个场景中，障碍物位于预定构型内部。当领航者和左侧跟随者都发现障碍物时，左侧跟随者绕过障碍物，而领航者通过改变方向来避免与障碍物碰撞，随后返回期望的编队形式。场景4是编队面对

两个间距小于编队尺寸的障碍物。如图所示，两个跟随者在通过危险区域时都改变了方向，并返回编队路径。例如，图12和图13分别展示了场景1和场景2中的控制信号以及系统状态变量。

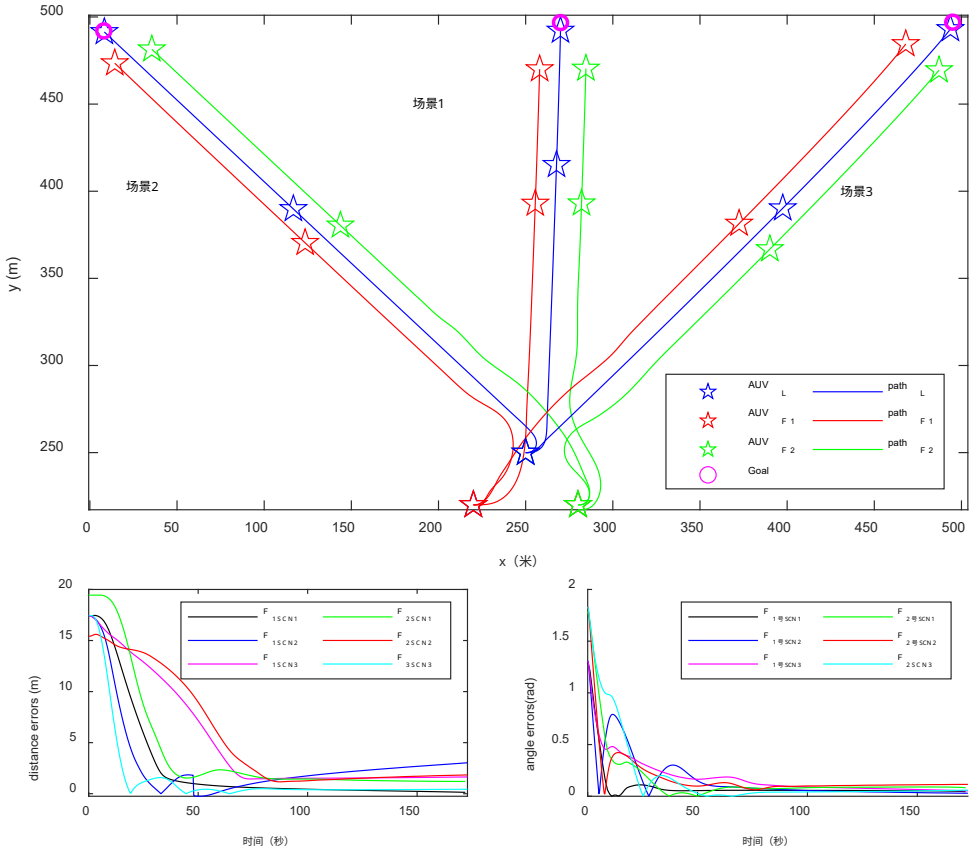

图9。 5.2.1节所有无障碍场景中的自主水下航行器轨迹、距离和角度编队误差（目标位于第一和第二象限）

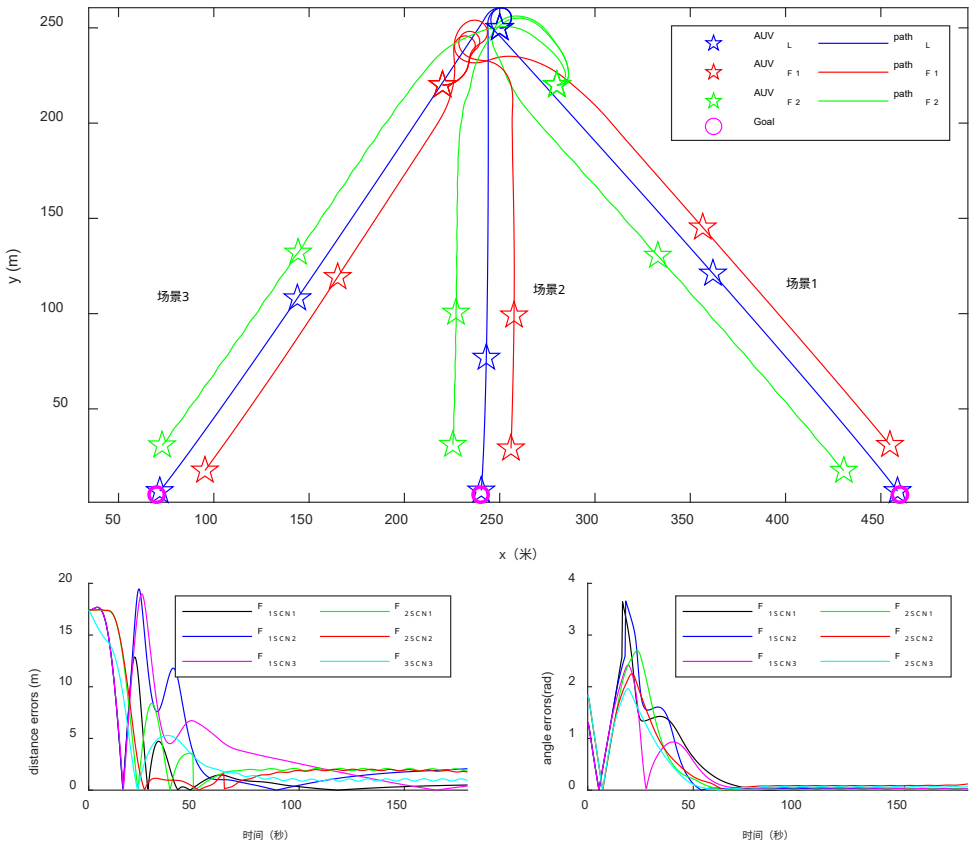

图10。5.2.1节所有无障碍场景下自主水下航行器的轨迹、距离和角度编队误差（目标位于第三和第四象限）

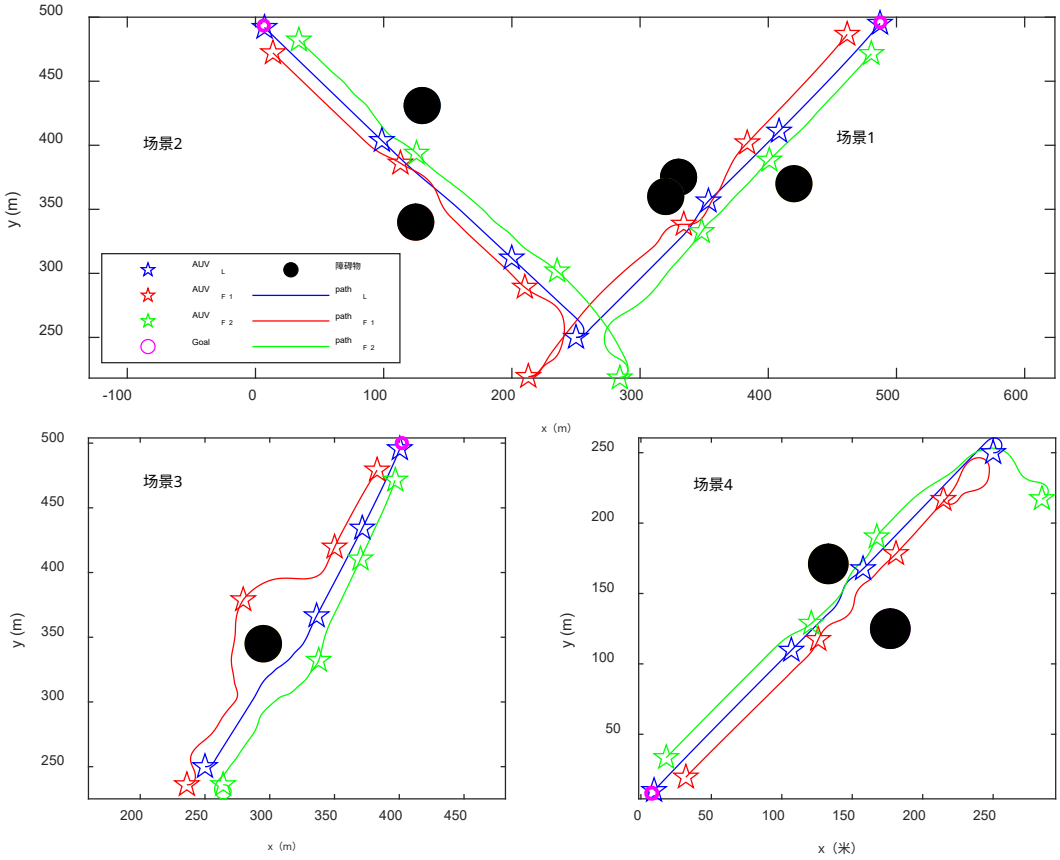

图 11. 第一种方法下自主水下航行器编队与避障的轨迹

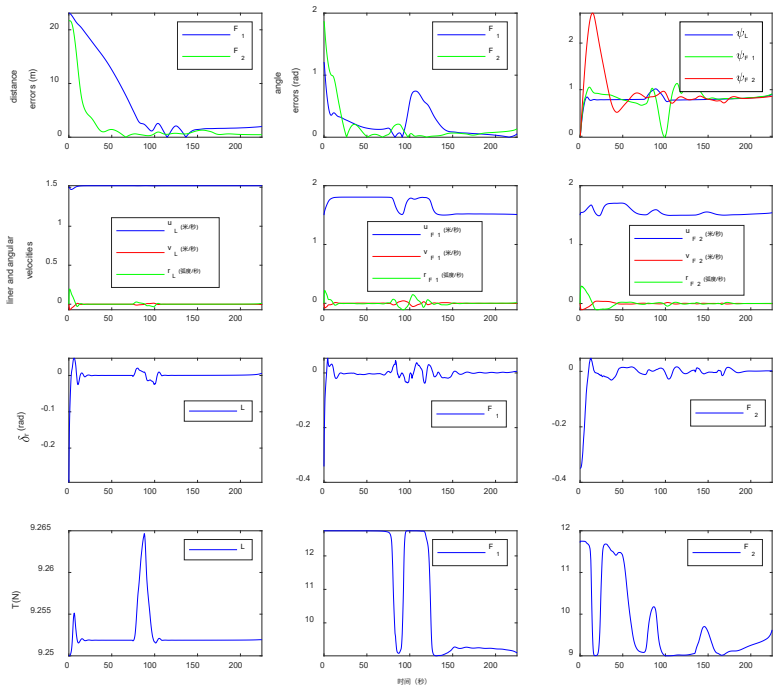

图12. 场景1中自主水下航行器的状态变量和控制信号

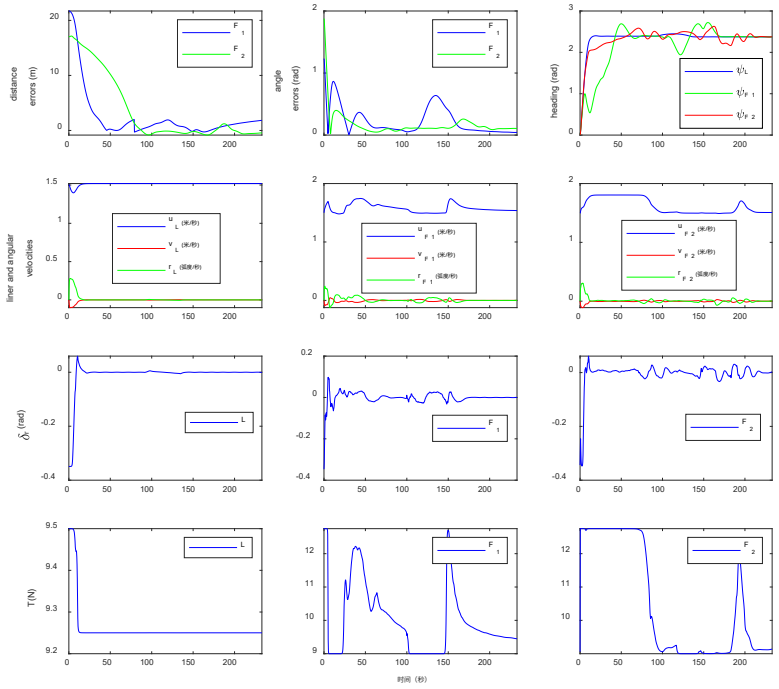

图13. 场景2中自主水下航行器的状态变量和控制信号

5.3. 运动规划与编队控制：方法二

本节介绍采用第3.2.1节方法进行的避障训练。此处的环境设置与前几节相同，只有领航者配备了避障模块。在这种方法中，有效避障取决于期望编队的规模。因此，训练阶段从创建期望编队开始，随后在训练阶段加入障碍物。网络训练示意图见图14。如图15所示，领航者选择一条路径，以安全地引领整个群体前往目的地。系统的控制信号和状态空间见图16。

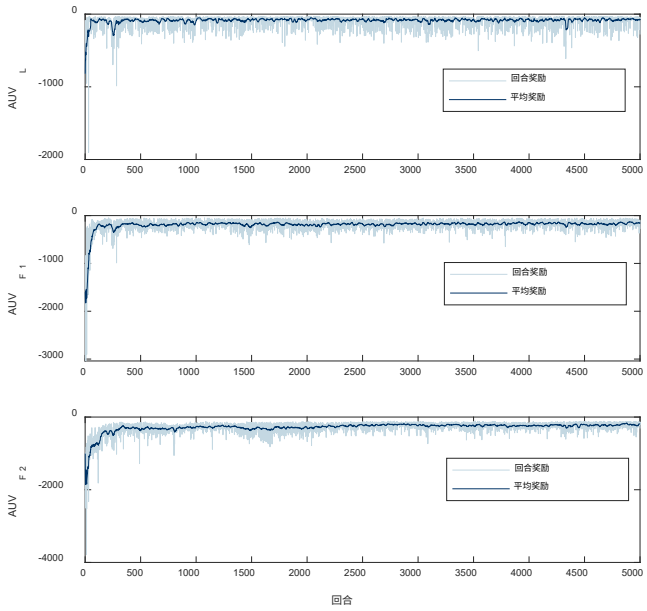

图14. 训练结果：第二种方法的回合奖励和每回合平均奖励

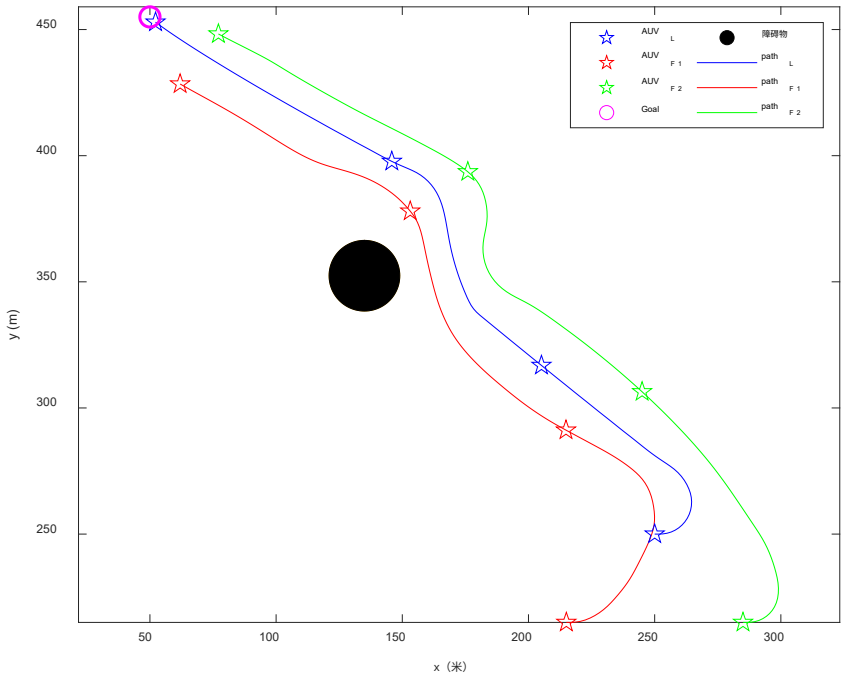

图 15. 第二种方法下的自主水下航行器编队与障碍物避障轨迹

根据仿真结果，每种避障方法在存在障碍物的情况下都表现良好。所提出的两种方法均采用分布式控制架构，可根据具体应用加以采用。

5.4. 存在海流时的仿真

海流是海洋水体中水平和垂直的环流系统，由风摩擦、重力等多种因素在海洋不同区域形成。为纳入海流及其对自主水下航行器运动的影响，可用相对速度 [45] 表示运动方程：

$$v_r = v - v_c \tag{25}$$

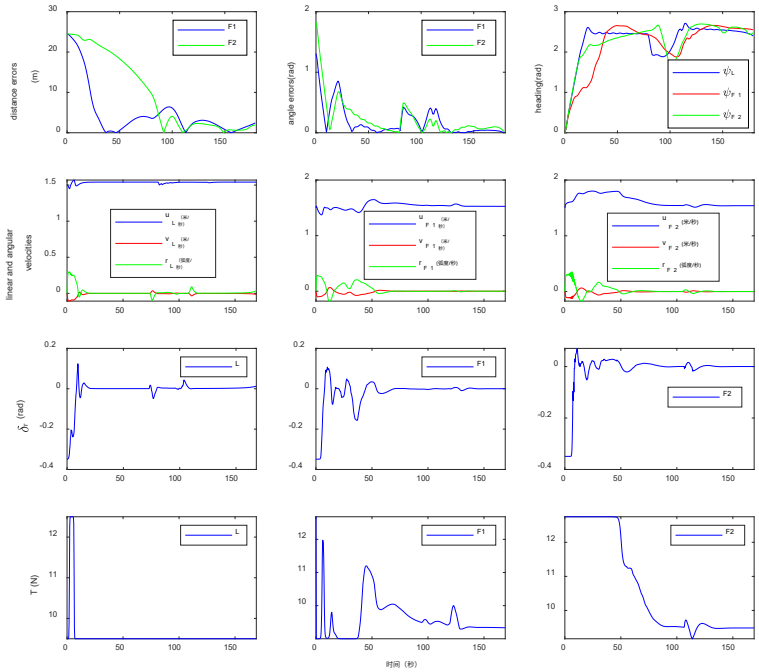

图16. 误差信号、状态变量和控制信号

其中，=[,υ 0]$Tvucc$ 表示体固定坐标系中的海流， =[, ,]$rvuvr$ 表示自主水下航行器在体固定坐标系中的线速度和角速度。为纳入海流，可将式（1）表示为相对速度的形式，如下 [45]:

$$(m-X_{\dot{u}})\dot{u}-my_g\dot{r}-m(v_r r+x_g r^2)=X_{u|u|}\left|u_r\right|u_r+X_{vr}v_r r_r+X_{rr}rr+X_{prop} \tag{26}$$

$$(m-Y_{\dot{v}})\dot{v}+(mx_g-Y_{\dot{r}})\dot{r}+m(u_r r_r-y_g r^2)=Y_{v|v|}\left|v_r\right|v_r+Y_{r|r|}\left|r\right|r+Y_{ur}u_r r+$$
$$Y_{uv}u_r v_r+Y_{uu\delta_r}u_r{}^2\delta_r$$

$$(I_z-N_{\dot{r}})\dot{r}+(mx_g-N_{\dot{v}})\dot{v}-my_g\dot{u}+m(x_g u_r r+y_g v_r r)=N_{v|v|}\left|v_r\right|v_r+N_{r|r|}\left|r\right|r+$$
$$N_{ur}u_r r+N_{uv}u_r v_r+N_{uu\delta_r}u_r{}^2\delta_r$$

其中，β_c 和 v_c 分别表示洋流的方向和速度。为评估洋流的影响及系统对洋流的响应，本文考虑了自主水下航行器的不同运动方向以及随机海流变化。两种编队方法的表现相似；我们仅添加了方法1。因此，根据图17展示了两个不同的场景，其中编队分别在有洋流和无洋流的情况下前往目的地。为模拟随机海流，采用了马尔可夫移动平均技术。在这两种场景中，洋流的速度和方向均随机变化，范围分别为0至0.3

米/秒和80至140度。结果表明，尽管自主水下航行器的路径和编队误差略有偏差，该群组仍能以编队形式安全抵达目的地。因此，基于深度强化学习的编队运动规划系统具有良好的适应性和鲁棒性。

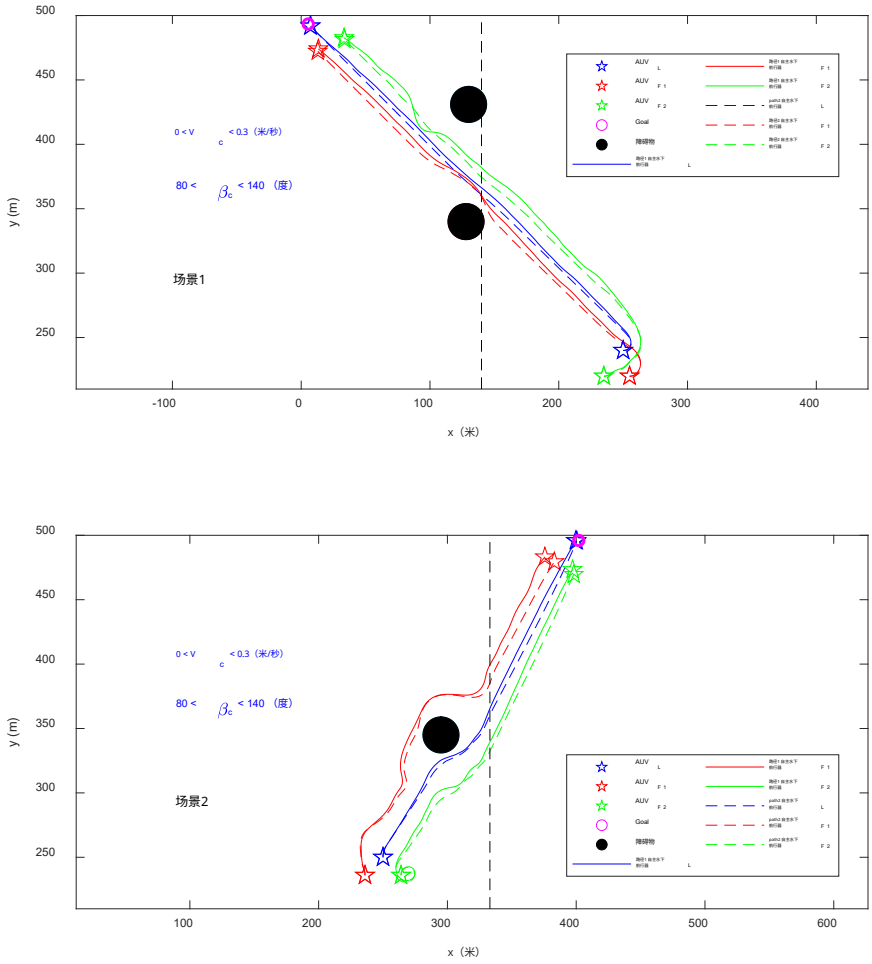

图 17. 两种不同场景下自主水下航行器在洋流中的轨迹，虚线：无洋流时自主水下航行器的轨迹，实线：有洋流时自主水下航行器的轨迹

5.5. 存在通信延迟和导航误差时的运动规划与编队控制

在任何自主水下航行器编队控制系统中，还存在通信延迟和导航（感知）误差等实际挑战。尽管本文主要研究运动规划和控制技术，但在某处的

算法设计中应考虑这些导航和控制误差。为此，本节评估了所提技术在存在未知通信延迟和导航误差时的韧性。

水下通信通常通过声学调制解调器建立，这可能成为误差来源，主要是可变通信延迟 [46, 47]。为模拟水下通信行为，本文考虑具有瑞利分布的时变延迟。可变延迟的峰值设为 0.1s，并在 1.2 s 后逐渐衰减，这对任何控制与规划系统而言都颇具挑战。自主水下航行器的导航还通过多种换能器组合实现，例如多普勒测速仪（DVL）、姿态与航向参考系统（AHRS）、深度传感器等；这些设备在数据融合后可能出现未知误差。本文使用马尔可夫移动平均过程对随机导航误差进行建模。添加到自主水下航行器的位置和航向状态中的导航误差信号如图 18 所示：

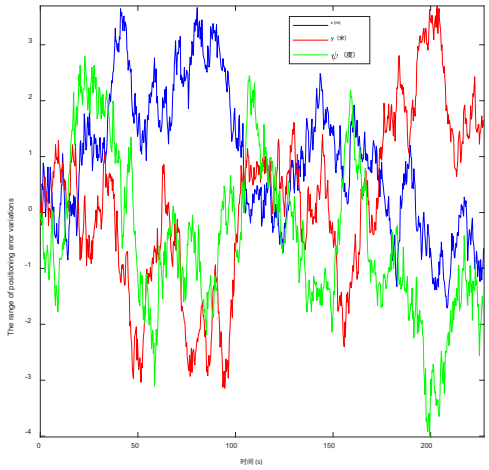

图 18. 自主水下航行器 x-y 位置和 ψ 角度的导航误差示例

需要指出的是，机器学习训练过程中并未考虑通信延迟和导航误差，因此这些因素能够在扰动条件下对所提出的技术构成适当挑战。图19展示了在目标场景中存在随机时变延迟和随机导航误差时，所提出系统的性能。由图可见，在这些扰动存在的情况下，所提出的算法仍能保持其性能和有界误差。在类似场景中，这些结果同样具有可重复性。所得性能表明，所提出的智能系统具有良好的鲁棒性和泛化能力。

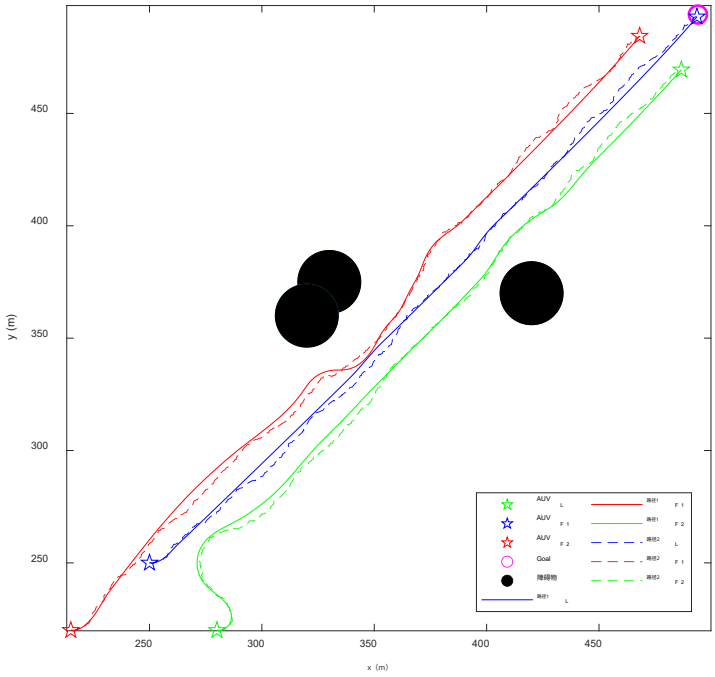

图 19. 自主水下航行器的队形 存在导航和通信误差的路径以虚线表示，而不存在这些问题的路径以实线表示。

实现说明： DRL算法的实时实现是一个充满活力且不断发展的研究主题。近年来，该领域的突破主要由NVIDIA及其他领先的人工智能 硬件公司 [48-53]推动。由于所提出的方法为每个智能体部署了两个评论家网络和四个评论家网络（均包含两个隐藏层），因此可以使用成本相对较低的商用现货产品（COTS），例如NVIDIA Jetson Nano和TX2 [49, 54, 55]来实现。尽管文献中已有成功的实现案例，但未来几年这一趋势可能会呈指数级增长 [49-52]。例如，在 [51], 中，采用DRL控制X型舵自主水下航行器的运动和自救。训练过程在Linux系统上进行，使用Python 2.7和TensorFlow 1 7 78008 0 64 . .，系统配备GB内存和Intel i X处理器。Nvidia Jetson Nano开发板用于实现。在 [49], 中，提出了一个用于自主敏捷四旋翼飞行的硬件和软件框架。该平台采用功能强大的Nvidia Jetson TX2开发板。Zhu等人 [52] 提出了一个基于深度强化学习的端到端控制与避障框架，用于视觉水下航行器。该框架集成了水下视觉感知和自动控制决策。该模型采用PyTorch构建，并在GeForce GTX 2080 Ti上进行训练。实际测试使用了BlueROV2模型和Jetson TX2控制器。可参考 [49] ，以比较该领域现有的解决方案。因此，所提出算法的实时实现是可行的，并预计将作为补充本研究的下一步工作。

6. 结论

本文提出了一种基于深度强化学习的新型编队控制与避障方案。在所提出的系统中，通过设计不同的奖励函数，提供了两种避障方案。第一种方法中，

所有群组成员均配备避障传感器，以进行安全的编队路径规划。第二种策略中，领航者仅负责识别障碍物，并引导整个群组安全地驶向目标。通过为领航者和跟随者设计不同的奖励函数，完成指定任务。领航者负责制定最优策略，在避开潜在障碍物的同时移动并引导群组前往目标。跟随者则负责保持期望编队。通过多种仿真场景验证了所提出的端到端运动规划与控制方法的性能。结果表明，在跟随航路点并避开障碍物时，领航者和跟随者能够保持其拓扑结构。此外，在存在洋流、通信延迟和感知误差的情况下，系统表现出的性能证明了所提出方案在受扰的现实条件下具有鲁棒性。所得结果验证了所提出算法的优点，表明其适合用于实际应用。

参考文献

1. Yang, E. 和 *D.J.I.A.t.o.m. Gu*，多自主水下航行器的非线性编队保持与系泊控制。《*IEEE/ASME*机电一体化学报》，*2007*。*12(2)*：第*164–178*页。 *2. Zeng, Z.* 等，多自主海洋航行器的会合路径规划。《*IEEE*海洋工程学报》，*2018*。*43(3)*：第*640–664*页。 *3. Carreras, M.* 等，*Sparus II*自主水下航行器——一种用于海床检查的悬停航行器。《*IEEE*海洋工程学报》，*2018*。*43(2)*：第*344–355*页。 *4. Sarhadi, P., A.R. Noei* 和 *A. Khosravi*，具有执行器饱和的自主水下航行器俯仰与偏航通道自适应积分反馈控制器。《*ISA*汇刊》，*2016*。*65*：第*284–295*页。 *5. Wynn, R.B.* 等，自主水下航行器（*AUV*）：过去、现在及其对海洋地球科学发展的未来贡献。《海洋地质学》，*2014*。*352*：第*451–468*页。 *6. Hadi, B., A. Khosravi* 和 *P. Sarhadi*，多自主水下航行器的路径规划与编队控制综述。《智能与机器人系统学报》，*2021*。*101(4)*：第*1–26*页。 *7. Oh, K.-K., M.-C. Park* 和 *H.-S.J.A. Ahn*，多智能体编队控制综述。《自动化学报》，*2015*。*53*：第*424–440*页。 *8. Wang, Q.* 等，切换网络中的分布式多机器人编队控制。《神经计算》，*2017*。*270*：第*4–10*页。 *9. Zhou, W.* 等，具有未知模型非线性和执行器饱和的无人水面航行器自适应模糊反步编队控制。《*IEEE*车辆技术汇刊》，*2020*。*69(12)*：第*14749–14764*页。 *10. Liu, H.* 等，受非线性因素和不确定性影响的多水下航行器鲁棒时变编队控制。《鲁棒非线性控制国际期刊》，*2019*。*29(9)*：第*2712–2724*页。 *11. Li, W., Q. Wen* 和 *H.J.T.A.J. Zhou*，考虑自动驾驶仪动态特性的多飞行器自适应滑模编队控制。《航空学报》，*2021*。*125(1290)*：第*1337–1357*页。 *12. Rouzegar, H.* 等，采用优化调谐开关型状态依赖黎卡提方程方法实现柔性卫星编队飞行的振动抑制与姿态控制。《测量与控制学会汇刊》，*2020*。*42(15)*：第*2984–3001*页。 *13. Rouzegar, H., A. Khosravi* 和 *P. Sarhadi*，采用开关型*SDRE*方法实现航天器在日地*L2*拉格朗日点附近的编队飞行控制。《空间研究进展》，*2021*。*67(7)*：第*2172–2184*页。

14. Emrani, S., A. Dirafzoon 和 *H.A. Talebi*，多自主水下航行器的自适应分布式编队控制。载于*2011*年*IEEE*控制应用国际会议（*CCA*）。*2011*。*IEEE*。 *15. Gao, Z., G.J.I.J.o.A.C. Guo* 和 *S. Processing*，具有模型不确定性的自主水下航行器自适应编队控制。《自适应控制与信号处理国际期刊》，*2018*。*32(7)*：第*1067–1080*页。 *16. Bian, J.* 和 *J.J.I.A. Xiang*，多自主水下航行器的三维协调控制。*IEEE Access*，*2019*。*7*：第*63913–63920*页。 *17. Wang, J.* 等，采用领导者-跟随者策略的欠驱动自主水下航行器滑模神经自适应编队控制。《应用海洋研究》，*2020*。*94*：*101971*。 *18. Bono, A., G. Fedele* 和 *G.J.I.T.o.C. Franzè*，用于不确定环境中自主车辆编队的基于群体的分布式模型预测控制方案。《*IEEE*控制论汇刊》，*2021*。 *19. Li, X.* 和 *D.J.I.T.o.I.E. Zhu*，一种用于一组自主水下航行器分布式编队控制的自适应*SOM*神经网络方法。《*IEEE*工业电子学汇刊》，*2018*。*65(10)*：第*8260–8270*页。 *20. Li, J., J. Du* 和 *W.-J.J.O.E. Chang*，输入饱和及扰动条件下欠驱动自主水下航行器的鲁棒时变编队控制。《海洋工程》，*2019*。*179*：第*180–188*页。 *21. Ihle, I.-A.F., J. Jouffroy* 和 *T.I.J.I.J.o.O.E. Fossen*，水面船舶编队控制：一种拉格朗日方法。《*IEEE*海洋工程学报》，*2006*。*31(4)*：第*922–934*页。 *22. Cui, R.* 等，欠驱动自主水下航行器的领导者–跟随者编队控制。《海洋工程》，*2010*。*37(17–18)*：第*1491–1502*页。 *23. Heshmati-Alamdari, S.* 等，精简通信条件下多个水下航行器机械臂系统的协作阻抗控制。《*IEEE*海洋工程学报》，*2020*。*46(2)*：第*447–465*页。 *24. Lewis, F.L., D. Vrabie* 和 *K.G. Vamvoudakis*，强化学习与反馈控制：运用自然决策方法设计最优自适应控制器。*IEEE*《控制系统杂志》，*2012*。*32(6)*：第*76–105*页。 *25. Sutton, R.S.* 和 *A.G. Barto*，《强化学习导论》。第*135*卷。*1998*：剑桥麻省理工学院出版社。 *26. Buşoniu, L.* 等，用于控制的强化学习：性能、稳定性与深度逼近器。《控制年度评论》，*2018*。*46*：第*8–28*页。 *27. Carlucho, I.* 等，利用深度强化学习实现自主水下航行器的自适应底层控制。《机器人学与自主系统》，*2018*。*107*：第*71–86*页。 *28. Chu, Z.* 等，通过深度模仿强化学习算法实现无人水下航行器的运动控制。《*IET*智能交通系统》，*2020*。*14(7)*：第*764–774*页。 *29. Sun, Y.* 等，利用基于策略梯度的深度强化学习实现自主水下航行器的无地图运动规划系统。《智能与机器人系统学报》，*2019*。*96(3–4)*：第*591–601*页。 *30. Bhopale, P., F. Kazi* 和 *N. Singh*，基于强化学习的自主水下航行器避障。《海洋科学与应用杂志》，*2019*。*18(2)*：第*228–238*页。 *31. Anderlini, E., G.G. Parker* 和 *G.J.A.S. Thomas*，利用强化学习实现自主水下航行器对接控制。《应用科学》，*2019*。*9(17)*：*3456*。 *32. Zhang, Q.* 等，利用深度交互强化学习实现自主水下航行器路径跟踪。*IEEE Access*，*2020*。*8*：第*24258–24268*页。 *33. Zhang, T.* 等，通过强化学习在真实世界中实现仿鱼机器人的分散式圆形编队控制。*arXiv*预印本 *arXiv:.05293*，*2021*。

34. Sui, Z., 等，利用模型引导示范的深度强化学习实现碰撞规避编队控制。*IEEE*神经网络与学习系统汇刊，*2020*。*32(6)*：第*2358-2372*页。*35. Xie, J.,* 等，面向多无人水面艇的基于强化学习的异步编队控制方案。《应用科学》，*2021*。*11(2)*：第*546*页。*36. Wang, S.,* 等，利用分布式深度强化学习对无人水面艇编队进行自适应可扩展控制。《应用海洋研究》，*2021*。*110*：第*102590*页。*37. Lin, Y.,* 等，具有优先级的无人机飞行编队通信动态频谱交互：一种深度强化学习方法。*IEEE*认知通信与网络汇刊，*2020*。*6(3)*：第*892-903*页。*38. Speck, C.* 和 *D.J. Bucci*。通过面向对象的多目标*SARSA*实现分布式无人机集群编队控制。发表于*2018*年美国控制会议年会（*ACC*）。*2018*。*IEEE*。*39. Zhu, P.,* 等，基于深度强化学习的多机器人集群控制。*IEEE Access*，*2020*。*8*：第*150397-150406*页。*40. Prestero, T.T.J.*，*REMUS*自主水下航行器六自由度仿真模型的验证。*2001*年，博士学位论文，麻省理工学院。*41. Mnih, V.,* 等，通过深度强化学习实现人类水平控制。《自然》，*2015*。*518(7540)*：第*529-533*页。*42. Lillicrap, T.P.,* 等，利用深度强化学习实现连续控制。*arXiv*预印本 *arXiv:.02971*，*2015*。*43. Fujimoto, S., H. Hoof,* 和 *D. Meger*。解决演员-评论家方法中的函数逼近误差。发表于国际机器学习会议。*2018*。*PMLR*。*44. Sarhadi, P., A.R. Noei,* 和 *A. Khosravi*，用于自主水下航行器的带抗积分饱和补偿器的模型参考自适应*PID*控制。《机器人学与自主系统》，*2016*。*83*：第*87-93*页。*45. Fossen, T.I.*，《海洋航行器的制导与控制》。*1999*年，挪威特隆赫姆大学：由约翰·威立父子出版公司印刷，英国奇切斯特，*ISBN*：*0 471 94113 1*，博士论文。*46. Wang, C.,* 等，多自主水下航行器编队的设计、建模、控制与实验。*IEEE*自动化科学与工程汇刊，*2021*。*19(4)*：第*2776-2787*页。*47. Li, L.,* 等，通信延迟、数据包离散性及丢包条件下的多自主水下航行器编队控制。《海洋科学与工程杂志》，*2022*。*10(7)*：第*920*页。*48. NVIDIA*。图形处理器与嵌入式系统。[引用于*2023*年；获取地址：*https://www.nvidia.com/*。*49. Foehn, P.,* 等，*Agilicious*：用于基于视觉的飞行的开源开放硬件灵活四旋翼飞行器。《科学·机器人学》，*2022*。*7(67)*：第*eabl6259*页。*50. Yan, C.,* 等，利用局部态势地图，通过深度强化学习为多架固定翼无人机学习无碰撞集群策略。*IEEE*工业信息学汇刊，*2021*。*18(2)*：第*1260-1270*页。*51. Fang, Y.,* 等，基于深度强化学习的自主水下航行器正常运动与主动自救控制策略。*AIP Advances*，*2022*。*12(1)*：第*015219*页。*52. Zhu, P.,* 等，视觉水下航行器的自主强化控制：基于计算机视觉的实时实验。*IEEE*车辆技术汇刊，*2022*。*71(8)*：第*8237-8250*页。*53. He, L., N. Aouf,* 和 *B. Song*，用于无人机自主路径规划的可解释深度强化学习。《航空航天科学与技术》，*2021*。*118*：第*107052*页。*54. NVIDIA*。*NVIDIA Jetson Nano*。[引用于*2023*年；获取地址：*https://www.nvidia.com/en-gb/autonomous-machines/embedded-systems/jetson-nano/*。*55. NVIDIA*。*Jetson TX2*模块。[引用于*2023*年；获取地址：*https://developer.nvidia.com/embedded/jetson-tx2*。