DOI：10.13245/j.hust.269997

分布式强化学习驱动的UAVs-USV自主协同决策

崔哲闻 关 巍 张显库
(大连海事大学航海学院，辽宁 大连 116026)

摘要 面向复杂动态海洋环境下无人机(UAVs)与无人船(USV)异构系统难以同时兼顾自主避障与稳定协同编队的问题，提出一种基于分布式强化学习的UAVs-USV自主协同决策方法．针对异构系统在运动特性、任务目标与环境感知能力上的差异，将协同决策任务解耦为USV自主航行子系统与UAVs编队协同子系统，分别采用近端策略优化(PPO)与多智能体深度确定性策略梯度(MADDPG)算法进行训练，实现局部自主避障与全局协同控制的协同优化．进一步地，构建融合视觉信息与状态信息的多模态神经网络；同时，引入基于速度障碍(RVO)的奖励约束机制，提升协同避碰与编队稳定性．最后，通过仿真与对比实验验证了本文方法在复杂动态环境下能够有效实现USV自主避障、UAVs稳定编队及异构协同决策，具有较好的协同精度、动态避障能力与泛化性．

关键词 无人船；无人机；分布式强化学习；协同决策；多模态神经网络
中图分类号 U675.73 **文献标志码** A **文章编号** 1671-4512(2026)10-0081-08

Autonomous cooperative decision-making for UAVs-USV system driven by distributed reinforcement learning

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

Abstract Aiming at the problem that the heterogeneous unmanned aerial vehicle (UAVs)-unmanned surface vehicle (USV) system is difficult to simultaneously achieve autonomous obstacle avoidance and stable cooperative formation in complex dynamic marine environments，a distributed reinforcement learning-based autonomous cooperative decision-making method for UAVs-USV was proposed. Considering the differences in motion characteristics，task objectives，and environmental perception capabilities of heterogeneous systems，the cooperative decision-making task was decoupled into a USV autonomous navigation subsystem and a UAVs formation cooperation subsystem，which were trained by the proximal policy optimization (PPO) algorithm and the multi-agent deep deterministic policy gradient (MADDPG) algorithm，respectively，to achieve the collaborative optimization of local autonomous obstacle avoidance and global cooperative control. Furthermore，a multimodal neural network integrating visual information and state information was constructed. Meanwhile，a reward constraint mechanism based on the reciprocal velocity obstacle (RVO) was introduced to improve cooperative collision avoidance and formation stability. Finally，simulation and comparative experiments were conducted to verify that the proposed method can effectively achieve USV autonomous obstacle avoidance，stable UAVs formation，and heterogeneous cooperative decision-making in complex dynamic environments，with high cooperative accuracy，dynamic obstacle avoidance capability，and generalization performance.

Key words unmanned surface vessel； unmanned aerial vehicle； distributed reinforcement learning； cooperative decision-making；multimodal neural network

随着海洋资源开发、环境监测与海上救援任务 日益复杂，无人系统集群因低成本、高灵活性与分

收稿日期 2026-03-17.
作者简介 崔哲闻 (1999—)，男，博士研究生；关 巍(通信作者)，教授，E-mail：gwwtxdy@163.com.
基金项目 国家自然科学基金资助项目(52171342).

布式协同能力，成为海上交通工程的重要研究方向．无人机(unmanned aerial vehicle，UAV)具备广域感知与快速任务规划能力，无人船(unmanned surface vehicle，USV)则适用于长时间水面作业与精细探测．UAVs-USV 协同决策，作为实现“空-海”跨域任务的基础，不仅能够提高海上任务效率和安全性，而且也是一个迫切需要解决的技术挑战．

现有UAVs-USV协同方法多基于模型预测控制(model predictive control，MPC)、事件触发控制(event-triggered control，ETC)与视觉感知技术．相关研究通过轨迹规划、视觉伺服及通信拓扑优化，实现了复杂环境下的协同控制与目标跟踪[1-3]．然而，此类方法通常依赖精确动力学模型与实时环境感知，且在动态环境下易受到模型误差、通信延迟与多智能体耦合效应影响，难以兼顾协同稳定性与实时决策能力[4]．

强化学习(reinforcement learning，RL)通过与环境交互自主学习最优策略，为复杂海洋环境下的协同决策提供了新思路．已有研究表明，RL能够有效实现无人系统避障、路径规划与自主控制[5-8]．同时，将长短期记忆网络(LSTM)与门控循环单元(GRU)等时序建模结构引入强化学习框架，可进一步提升动态环境中的状态感知与策略预测能力[9-11]．然而，当前多数UAVs-USV协同研究仍采用集中式整体建模方式，忽略了UAVs与USV在运动特性、任务目标及环境感知能力上的差异，导致策略耦合严重、学习效率较低，难以满足复杂海洋场景下的实时协同需求．

针对上述问题，本研究提出一种基于分布式强化学习的UAVs-USV自主协同决策方法，主要贡献如下．**a**．提出UAVs-USV分布式协同解耦框架，将任务划分为USV自主航行与UAVs编队协同两个子系统，从系统层面降低策略耦合与环境非平稳性．**b**．构建结合多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient，MADDPG)与近端策略优化(proximal policy optimization，PPO)的异构协同强化学习方法，分别针对USV自主避障与UAVs高耦合协同特性进行训练，实现集中式价值评估与稳定策略更新的协同互补．**c**．融合GRU多模态网络与基于RVO的奖励约束，并采用高层动作输出方式，提升系统协同精度、动态避障能力与泛化性能．

1　模型构建

1.1　UAVs-USV协同运动数学模型

为保证RL协同决策模型具备UAV与USV运动特性，采用了欧拉-拉格朗日形式的UAVs-USV混合维非线性运动模型[1,12]，如图1所示，图中：N和G分别为正北方向和USV重心；X_{OG}，Y_{OG}，Z_{OG}分别为坐标系三个坐标轴方向；u_s，v_s，r_s分别为USV纵向速度、横向速度及艏摇角速度；对于UAV系统，w_{1i}～w_{4i}分别为第i个UAV四个旋翼的角速度，F_{1i}～F_{4i}为对应旋翼产生的升力；α_x，α_y，α_z分别为UAV在机体坐标系下三个方向的线加速度，$r_{\phi i}$，$r_{\theta i}$，$r_{\psi i}$分别为第i个UAV的滚转、俯仰和偏航角速度；x_s和y_s为USV的位置坐标；ψ_s为USV的偏航角．

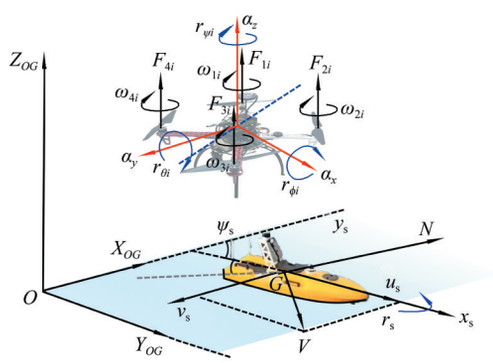


图1　UAVs-USV运动数学模型

具体公式如下所示．

$$\dot{x}_s = u_s\cos\psi_s - v_s\sin\psi_s;$$

$$\dot{y}_s = u_s\sin\psi_s + v_s\cos\psi_s;\ \dot{\psi}_s = r_s;$$

$$m_{su}\dot{u}_s = f_u(\boldsymbol{v}_s) + \tau_u + d_{wu};\ m_{sv}\dot{v}_s = f_v(\boldsymbol{v}_s) + d_{wv};$$

$$m_{sr}\dot{r}_s = f_r(\boldsymbol{v}_s) + \tau_r + d_{wr};$$

$$m_a\dot{u}_{axi} = f_{xi}(\boldsymbol{v}_{ai}) + R_{xi}\tau_{fi} + d_{xi};$$

$$m_a\dot{u}_{ayi} = f_{yi}(\boldsymbol{v}_{ai}) + R_{yi}\tau_{fi} + d_{yi};$$

$$m_a\dot{u}_{azi} = f_{zi}(\boldsymbol{v}_{ai}) - g_a + R_{zi}\tau_{fi} + d_{zi};$$

$$\dot{r}_{\phi i} = f_{\phi i}(\boldsymbol{v}_{ai}) + d\tau_{\phi i}/I_{xx} + d_{\phi i};$$

$$\dot{r}_{\theta i} = f_{\theta i}(\boldsymbol{v}_{ai}) + d\tau_{\theta i}/I_{yy} + d_{\theta i};$$

$$\dot{r}_{\psi i} = f_{\psi i}(\boldsymbol{v}_{ai}) + d\tau_{\psi i}/I_{zz} + d_{\psi i};$$

$$R_{xi} = \cos\phi_{ai}\sin\theta_{ai}\cos\psi_{ai} + \sin\theta_{ai}\sin\psi_{ai};$$

$$R_{yi} = \cos\phi_{ai}\sin\theta_{ai}\cos\psi_{ai} - \sin\theta_{ai}\sin\psi_{ai};$$

$$R_{zi} = \cos\phi_{ai}\cos\psi_{ai},$$

式中：$\boldsymbol{v}_s = [u_s,\ v_s,\ r_s]^{\mathrm{T}}$为USV在三个自由度的合速度；$d_{wu}$，$d_{wv}$，$d_{wr}$为USV的风流海浪干扰项；$f_u(\boldsymbol{v}_s)$，$f_v(\boldsymbol{v}_s)$，$f_r(\boldsymbol{v}_s)$分别为USV在三个自由度的速度不确定项；$x_{ai}$，$y_{ai}$，$z_{ai}$，$\phi_{ai}$，$\theta_{ai}$，$\psi_{ai}$分别为第$i$个UAV在$x$，$y$，$z$方向上的位置坐标、偏航角、滚

动角和俯仰角；u_{axi}，u_{ayi}，u_{azi}，$r_{\phi i}$，$r_{\theta i}$，$r_{\psi i}$分别为第i个UAV的速度状态；u_{axi}，u_{ayi}，u_{azi}分别为三个方向的线速度；$f_{xi}(\boldsymbol{v}_{ai})$，$f_{yi}(\boldsymbol{v}_{ai})$，$f_{zi}(\boldsymbol{v}_{ai})$，$f_{\phi i}(\boldsymbol{v}_{ai})$，$f_{\theta i}(\boldsymbol{v}_{ai})$，$f_{\psi i}(\boldsymbol{v}_{ai})$分别为第$i$个UAV在六个自由度的速度不确定项；$d_{xi}$，$d_{yi}$，$d_{zi}$，$d_{\phi i}$，$d_{\theta i}$，$d_{\psi i}$分别为第$i$个UAV的干扰项；$m_{su}$，$m_{sv}$，$m_{sr}$分别为USV在纵荡、横荡和艏摇三个自由度下的等效惯性参数；m_a为UAV的质量；I_{xx}，I_{yy}，I_{zz}分别为UAV绕机体坐标系三个坐标轴的转动惯量；$[\tau_u,\ 0,\ \tau_r]^{\mathrm{T}}$为USV的控制力矩，$[\tau_f,\ \tau_\phi,\ \tau_\theta,\ \tau_\psi]^{\mathrm{T}}$为UAVs的控制力矩，其可以分别映射到USV的螺旋桨速度和舵角，以及UAVs的四轴螺旋桨转速和转向角；d为UAV受到的外部扰动量．本研究假定，在协同航行过程中，所有UAVs均保持同一飞行高度，且具备一致的运动学特性．

1.2　视觉定位模型

为实现UAV对USV的协同跟踪，建立UAV与USV之间的视觉定位模型，如图2所示，图中x_{im}与y_{im}为UAV经过投影映射后的坐标位置．该模型通过坐标系转换与透视投影，为后续协同误差计算与奖励函数设计提供位置信息．根据透视投影原理，可建立三维空间到二维图像平面的映射关系

$$\begin{cases} x_{im} = (f_{\mathrm{cam}} x_{\mathrm{c}}/z_{\mathrm{c}})/\rho_u + c_u; \\ y_{im} = (f_{\mathrm{cam}} y_{\mathrm{c}}/z_{\mathrm{c}})/\rho_v + c_v, \end{cases}$$

式中：f_{cam}为焦距；ρ_u与ρ_v分别为相机成像传感器沿图像横向和纵向的像元尺寸；c_u与c_v为以像素为单位的光学坐标中心；x_{c}，y_{c}，z_{c}为UAV相机坐标系位置．

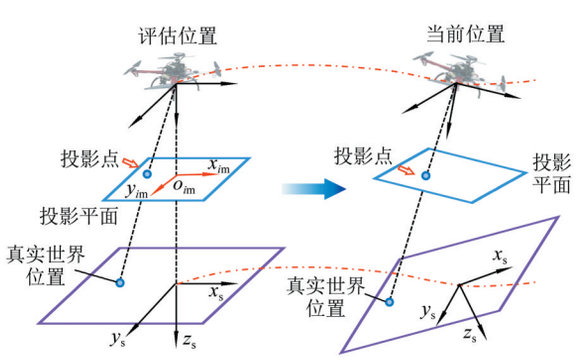


图2　UAVs-USV视觉定位模型

2　USV自主航行训练框架构建

本研究将系统解耦为UAVs集群协同与USV自主航行两个分布式子系统．针对USV局部感知与自主避障需求，采用单智能体PPO算法构建自主航行决策模型，并结合局部环境感知信息与多维奖励约束实现自主避障决策，其整体框架如图3所示，图中：s_t为状态信息，包括相对目标信息、雷达观测信息等；R为奖励函数，包括目标引导奖励、航向偏差奖励、安全距离奖励；a_t为动作输出，包括舵角控制指令；θ为策略网络参数；φ为价值网络参数．

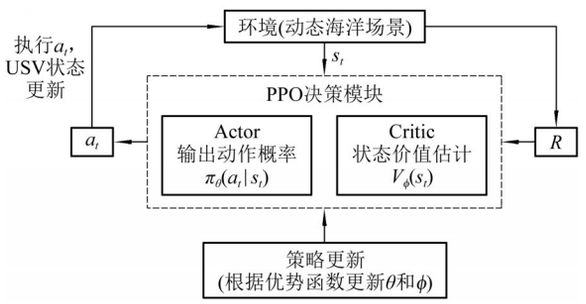


图3　USV自主航行训练框架下的PPO算法与环境交互过程

2.1　PPO算法框架设计

PPO算法基于Actor-Critic(策略-价值)训练框架，通过在目标函数中加入概率比值截断项，确保新策略与旧策略的差异不超过阈值，实现稳定且高效的策略更新，其优势函数可以表示为

$$E_1\left[\sum_{t=0}^{\infty}\gamma^t A_1(s_t,a_t)\right] = \sum_{t=0}^{\infty}\gamma^t E_2 E_3[A_1(s_t,a_t)],$$

式中：E_1为轨迹策略期望；E_2为状态分布期望；E_3为动作概率期望；$A_1(s_t,\ a_t)$为USV在状态s_t下执行动作a_t的优势函数；γ为折扣项系数；t为时间步．本研究进一步引入广义优势估计策略(GAE)，以平衡偏差与方差，具体表示为

$$\hat{G}_t = \sum_{k=0}^{\infty}(\gamma\lambda)^k \delta_{t+1}^{\mathrm{V}},$$

式中：$\lambda \in [0,\ 1]$为GAE算法中的超参数；$\delta_{t+1}^{\mathrm{V}}$为$t$+1时刻的时间差分残差．此时，PPO算法的策略迭代目标函数可以表示为

$$J(\theta) = E_4\left[\min\left(\frac{\pi(a_t|s_t)}{\pi_{\mathrm{old}}(a_t|s_t)}\hat{G}^t(a_t|s_t),\right.\right.$$

$$\left.\left.\mathrm{clip}\left(\frac{\pi(a_t|s_t)}{\pi_{\mathrm{old}}(a_t|s_t)},\ 1-\varepsilon,\ 1+\varepsilon\right)\hat{G}^t(a_t|s_t)\right)\right],$$

式中：π为当前策略；π_{old}为先前策略；clip(·)为剪切函数；ε为裁剪超参数；E_4为目标函数在先前策略π_{old}下产生的状态和动作分布．

2.2　USV状态与动作空间设计

本研究考虑将USV的雷达观测数据构建状态空间，进而降低对全局信息的依赖性．USV的状态空间可以设计为

$$\mathcal{S}^{\mathrm{s}} = \{\, s_t^{\mathrm{s}} :\, = (\boldsymbol{p}_{\mathrm{s}},\ \boldsymbol{g}_{\mathrm{USV}},\ o_1)\,\},$$

式中：$\boldsymbol{g}_{\mathrm{USV}}$为USV的目标位置；$o_1 \in \mathbf{R}^{1\times n_1}$为雷达观测数据；$\boldsymbol{p}_{\mathrm{s}}$为USV当前位置．此外，本研究将PPO算法的输出定义为USV的指令舵角，其动作空间定义为

$$\mathcal{A}^s = \{\, a_t^s :=[(\delta^s_{\mathrm{port}} \sim \delta^s_{\mathrm{starb}})]\,\},$$

式中δ^s_{port}和$\delta^s_{\mathrm{starb}}$分别为USV向左舷和右舷转动舵角的边界，USV向左舷改变航向定义为负值，向右舷改变航向定义为正值．

2.3　USV自主航行模块奖励函数设计

本研究将USV奖励函数设计为制导奖励、偏航修正奖励与安全奖励三部分，以实现USV在未知环境中有效规避静态障碍与动态船舶．

制导奖励函数旨在引导USV驶向目标位置，其奖励值随USV与目标间距离的减小而单调递增，可以表示为

$$R_{\mathrm{g}} = \begin{cases} -k_1 \quad (\text{发生碰撞}); \\ -w_{\mathrm{g}} \ln\left\| (\boldsymbol{g}_{\mathrm{usv}} - \boldsymbol{p}_{\mathrm{s}}) + 1 \right\| \quad (\text{其他}), \end{cases}$$

式中：k_1为碰撞惩罚项；w_{g}为制导权重．

偏差角修正奖励用于引导USV在避障决策结束后回到其原有航线上，具体可以表示为

$$R_{\mathrm{y}} = e_{\mathrm{tr}} w_{\mathrm{y}} \exp\left[(\varepsilon_{\mathrm{y}} \theta_{\mathrm{d}})^2 \left\| \boldsymbol{g}_{\mathrm{USV}} - \boldsymbol{p}_{\mathrm{s}} \right\| \right],$$

式中：θ_{d}为偏差角；e_{tr}为误差系数；w_{y}为偏差角修正奖励的权重参数；ε_{y}为距离变化参数．

针对PPO算法因依概率选择动作而限制探索能力的问题，本研究在雷达探测范围内设定安全距离，当障碍物处于雷达探测范围内但位于安全距离外时，给予持续小幅度正奖励，以激励USV对未知环境的主动探索，安全奖励可以表示为

$$R_{\mathrm{s}} = \begin{cases} k_2 \quad (\eta_3 < \eta_1 < \eta_2); \\ -k_3 \quad ((\eta_1 < \eta_2) \cap (\eta_1 < \eta_3)), \end{cases}$$

式中：η_1为USV与障碍物或其他船舶之间的距离；η_2为雷达探测距离；η_3为安全距离；k_2为探索奖励项；k_3为探索惩罚项．

3　UAVs-USV协同训练框架构建

在UAVs-USV协同框架中，UAVs须在分布式架构下自主追踪USV的视觉定位信息，实现动态协同．针对多智能体交互与编队协同需求，本研究采用MADDPG算法进行训练．

3.1　UAVs编队模型

本文构建虚拟引导器作为编队位置参考[13]，如图4所示，图中：l_i为编队几何中心与不同UAV之间的距离；β_i为编队成员与虚拟引导器切线方向的误差角度；$\boldsymbol{p}^i_{\mathrm{im}}$为第$i$个UAV实际位置在二维图像平面上的投影坐标．

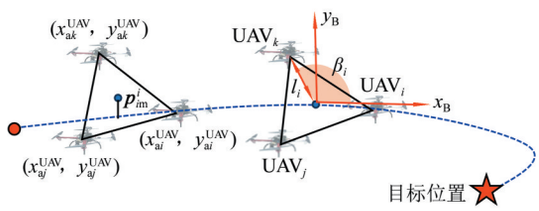


图4　UAV编队模型设计

根据虚拟引导器在二维图像平面中的投影坐标，可计算UAV的期望位置

$$\begin{cases} x_{pi}^{\mathrm{UAV}} = x_{im} + l_i \cos(\arctan(\Delta y/\Delta x) + \beta_i); \\ y_{pi}^{\mathrm{UAV}} = y_{im} + l_i \sin(\arctan(\Delta y/\Delta x) + \beta_i), \end{cases}$$

式中Δx和Δy分别为虚拟引导器在图像平面坐标系中的横向和纵向位置变化量．编队中第i个UAV的实际编队与预期编队之间的误差e_i可以由二者间的距离衡量

$$e_i = \sqrt{(x_{pi}^{\mathrm{UAV}} - x_{ai}^{\mathrm{UAV}})^2 + (y_{pi}^{\mathrm{UAV}} - y_{ai}^{\mathrm{UAV}})^2},$$

式中$(x_{ai}^{\mathrm{UAV}},\ y_{ai}^{\mathrm{UAV}})$为第$i$个UAV的实际位置在二维图像平面上的投影坐标．编队误差可以通过对每架UAV的误差进行求和或取平均来计算．

假设编队中存在ξ架UAV，则编队平均误差可以表示为

$$e_{\mathrm{e}} = \frac{\sum_i^{i+\xi} \sqrt{(x_{pi}^{\mathrm{UAV}} - x_{ai}^{\mathrm{UAV}})^2 + (y_{pi}^{\mathrm{UAV}} - y_{ai}^{\mathrm{UAV}})^2}}{\xi}.$$

3.2　MADDPG算法框架设计

多智能体同时更新策略会导致环境非平稳性．MADDPG结合Actor-Critic框架与集中式训练、分散式执行机制，可有效缓解该问题，其交互过程如图5所示，图中：o为当前时刻的联合观测；o'为下一时刻的联合观测；N为智能体个数；R_i为MADDPG算法的当前奖励．在训练阶段，各智能体通过Critic网络利用全局状态与动作信息，并基于时序差分误差更新集中式Q函数，其损失函数为

$$L(\varphi_i) = E_5[(Q_i^{\mathrm{u}}(o,\ a_1,\ a_2,\ \cdots,\ a_N) - y_i)^2],$$

式中：y_i为目标值；E_5为经验回放池采样期望；$Q_i^{\mathrm{u}}(o|a_1|a_2|\cdots|a_N)$为第$i$个智能体Critic网络根据当前时刻的联合观测和所有智能体动作a_1，a_2，…，a_N估计的动作价值函数．在经验回放区中采样一小批经验值来估计期望值，并通过梯度下降来最小化损失．Critic网络的梯度下降更新可以表示为

$$\nabla_{\varphi_i} L(\varphi_i) = E_5[2(Q_i^{\mathrm{u}}(o,\ a) - y_i)\nabla_{\varphi_i} Q_i^{\mathrm{u}}(o,\ a)],$$

式中∇_{φ_i}为损失函数关于Critic网络的梯度．

在执行阶段，每个智能体通过Actor网络，根

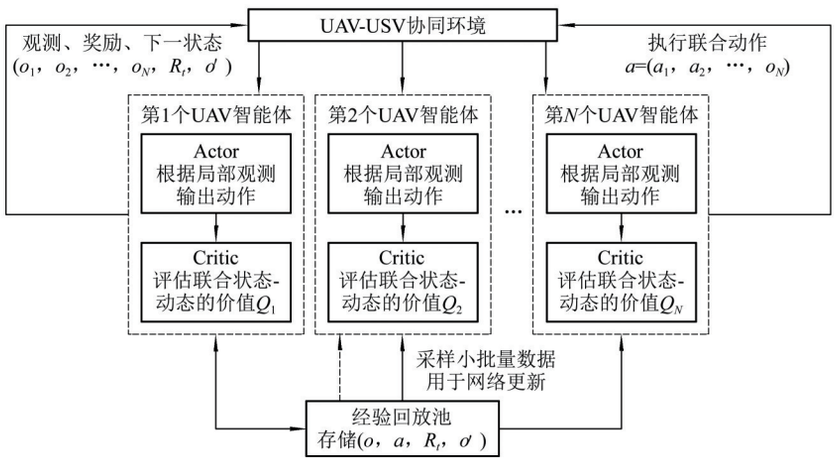


图5　UAVs-USV编队协同训练框架下的MADDPG算法与环境交互过程

据自己的局部观测，独立地选择动作．Actor网络的更新目标是最大化期望回报，根据策略梯度定理，其梯度可以表示为

$$\nabla_{\theta_i} J(\mu_i) \approx E_6[\nabla_{\theta_i}\mu_i(a_i|o_i)\nabla_{a_i} Q_i^{\mathrm{u}}(o, a_1, a_2, \cdots, a_N)|_{a_i=\mu_i(o_i)}],$$

式中：$\nabla_{\theta_i}\mu_i(a_i|o_i)$为动作对参数的梯度；$\nabla_{a_i} Q_i^{\mathrm{u}}(o, a_1, a_2, \cdots, a_N)$为$Q$值对第$i$个智能体动作的梯度；$E_6$为观测和动作期望．

3.3　UAVs编队状态与动作空间设计

考虑到UAV具有通过相机采集的连续图像数据的能力，将UAVs编队的状态空间设计为

$$\mathcal{S}_i^{\mathrm{a}} = \{ s_t^{\mathrm{a}} : = (\boldsymbol{p}_{\mathrm{im}}^i, \boldsymbol{p}_{\mathrm{ai}}^{\mathrm{UAV}}, v_i^{\mathrm{a}}, f_i^{\mathrm{c}}) \};$$

$$\mathcal{S}^{\mathrm{a}} = \{(\mathcal{S}_1^a, \mathcal{S}_2^a, \cdots, \mathcal{S}_N^a), \boldsymbol{p}_{\mathrm{s}}, \boldsymbol{g}_{\mathrm{UAV}}\},$$

式中：$\boldsymbol{p}_{\mathrm{ai}}^{\mathrm{UAV}}$为第$i$个UAV期望位置在二维图像平面上的投影坐标；$v_i^{\mathrm{a}}$为第$i$个UAV当前时刻速度；$f_i^{\mathrm{c}} \in \mathbf{R}^{1\times n_{\mathrm{c}}}$为第$i$个UAV相机采集的连续图像数据；$\boldsymbol{g}_{\mathrm{UAV}}$为编队的飞行目标位置．本研究将MADDPG算法的输出定义为每架UAV的转向角度与速度，其动作空间定义为

$$\mathcal{A}_i^{\mathrm{a}} = \{ a_t : = [(\delta_{\mathrm{r}}^{\mathrm{a}} \sim \delta_{\mathrm{l}}^{\mathrm{a}}), (v_{\min}^{\mathrm{a}} \sim v_{\max}^{\mathrm{a}})] \};$$

$$\mathcal{A}^{\mathrm{a}} = \{ \mathcal{A}_1^{\mathrm{a}}, \mathcal{A}_2^{\mathrm{a}}, \cdots, \mathcal{A}_N^{\mathrm{a}} \},$$

式中：$\delta_{\mathrm{r}}^{\mathrm{a}}$与$\delta_{\mathrm{l}}^{\mathrm{a}}$分别为第$i$个UAV向右和向左的转向角度边界；$v_{\min}^{\mathrm{a}}$与$v_{\max}^{\mathrm{a}}$分别为第$i$个UAV的最小与最大速度．

3.4　UAVs-USV协同模块神经网络结构设计

针对UAVs-USV编队协同问题，设计了一种融合GRU的神经网络架构，如图6所示．

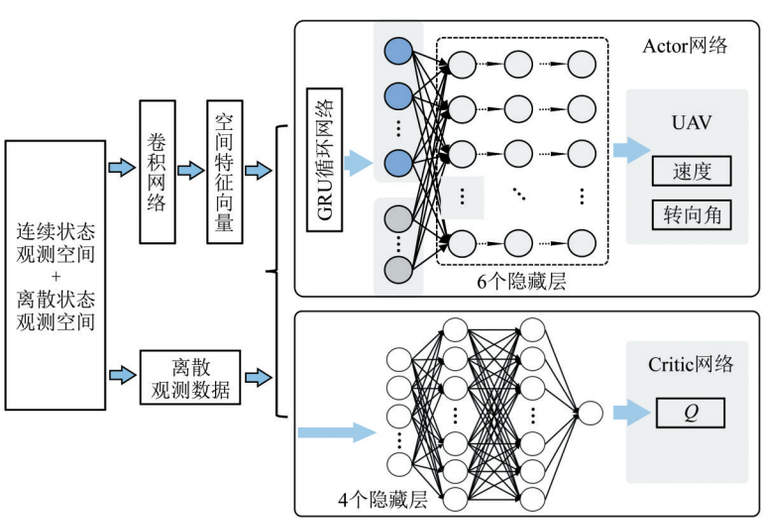


图6　协同框架下基于GRU的多模态神经网络结构设计

Actor采用卷积神经网络(CNN)提取连续图像的空间特征，并与UAVs状态向量进行融合，通过GRU对时序特征进行建模，以增强对动态环境与轨迹变化的感知能力．相比传统前馈网络，该结构能够更有效地处理编队飞行中的时序依赖关系．Critic网络采用简化结构，对全局状态与所有UAVs动作进行统一编码，学习多智能体间的协同关系，并输出对应Q值，为策略优化提供梯度信息．

3.5　UAVs-USV协同模块奖励函数设计

针对UAVs-USV协同任务需求，MADDPG的奖励函数包含三个部分：UAVs-USV协同奖励，UAVs编队误差奖励，速度奖励．UAVs-USV协同奖励旨在实现UAVs与USV之间的异构协同，有

$$R_{\mathrm{c}}^{\mathrm{a}}=-w_1\sum_1^i\left\|\boldsymbol{p}_{ai}^{\mathrm{UAV}}-\boldsymbol{p}_{im}^{i}\right\|^2-w_2\left\|\boldsymbol{p}_{\mathrm{s}}-\boldsymbol{p}_1\right\|^2,$$

式中：w_1与w_2为协同权重参数；$\boldsymbol{p}_1$为虚拟引导器预期跟踪位置在二维图像平面上的投影位置．

UAVs编队误差奖励用于确保UAVs编队按照预设编队形状进行飞行，并且当UAVs编队到达目标位置时给予正向奖励，可以表示为

$$R_{\mathrm{e}}^{\mathrm{a}}=\begin{cases}k_4 & (\text{到达目标位置});\\ -w_3\exp\left(e_{\mathrm{e}}\left\|\boldsymbol{g}_{\mathrm{UAV}}-\boldsymbol{p}_1\right\|\right) & (\text{其他}),\end{cases}$$

式中w_3为编队误差权重参数．

速度奖励设计通过引入RVO算法，能够有效引导UAV在维持编队构型的同时，实现动态避障与多机速度协同，具体表达式为

$$R_{\mathrm{RVO}}^{\mathrm{c}}=-\sum_{j\neq i}\max\left(0,\frac{l_{\mathrm{safe}}-\left\|\boldsymbol{p}_{ai}^{\mathrm{UAV}}-\boldsymbol{p}_{aj}^{\mathrm{UAV}}\right\|^2}{l_{\mathrm{safe}}}\right)\cdot$$

$$I_{\mathrm{con}}(v_i^{\mathrm{a}},v_j^{\mathrm{a}});$$

$$R_{\mathrm{RVO}}^{\mathrm{sd}}=-\sum_1^i\left|\mathrm{v}_{\mathrm{i}}^{\mathrm{a}}-v_i^{\mathrm{d}}\right|^2,$$

式中：l_{safe}为安全距离阈值；I_{con}为指示函数，若速度方向在RVO冲突锥内则为1，否则为0；v_i^{d}为第i个UAV的期望速度；$R_{\mathrm{RVO}}^{\mathrm{c}}$为RVO碰撞规避奖励；$R_{\mathrm{RVO}}^{\mathrm{sd}}$为RVO速度协同奖励．

4　实验分析

本研究通过仿真与对比实验验证所提分布式强化学习UAVs-USV协同决策方法的有效性．实验平台采用Intel Core i9-11900H处理器、NVIDIA RTX 3090显卡，运行Ubuntu 20.04与ROS Noetic系统．USV与UAV动力学参数参考文献[7]．

4.1　UAVs-USV协同决策模型训练结果与分析

考虑到系统的分布式特性，先独立完成USV自主航行策略训练，待其具备稳定的局部避障与航线跟踪能力后，再启动UAVs编队协同训练，通过异构系统的分步优化确保整体协同精度与效率．

4.1.1　USV自主航行训练结果与分析

USV自主航行训练以平均奖励收敛特性作为主要评价指标，并选取DDPG与传统Actor-Critic (AC)算法作为对比方法，实验结果如图7所示．

PPO算法能够更快完成策略收敛，并获得更高且更稳定的平均奖励；同时，最大奖励值高于DDPG与AC算法．相比于DDPG对Critic网络连续动作价值估计的强依赖，PPO通过概率比值裁剪机制限制策略更新幅度，有效降低了训练过程中的梯度震荡与策略发散问题，从而提升了复杂环境中的训练稳定性．此外，PPO结合GAE对时间回报进行平滑处理，可在短时避障行为与长期航向保持之间实现有效平衡，使USV在完成动态避障后能够快速恢复至目标航线．

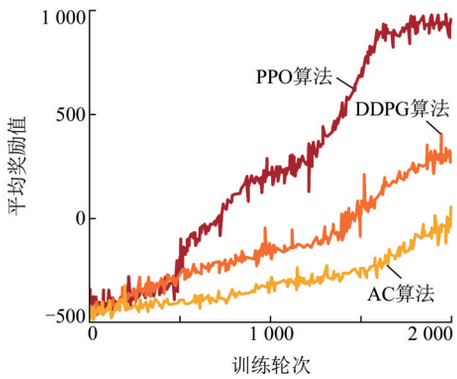


图7　USV自主航行训练结果

4.1.2　UAVs-USV编队协同训练结果与分析

UAVs-USV队协同训练以平均奖励收敛特性、UAVs编队误差及异构协同误差为评价指标，并与multi-agent PPO (MAPPO)、multi-agent AC (MAAC)和Q-MIX算法进行对比，结果如图8所示．可以看出：MADDPG在收敛速度、最大奖励值及协同与

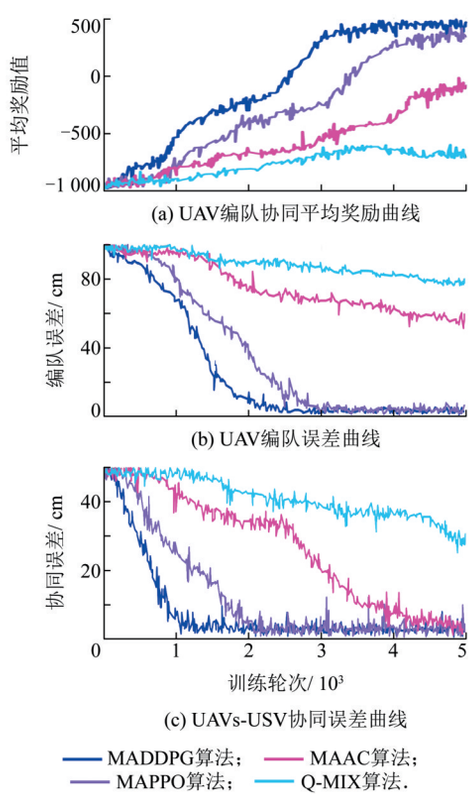


图8　UAVs-USV编队协同训练结果

编队误差控制方面均表现最优．

该结果与本研究采用的系统解耦与集中式价值评估协同框架密切相关．USV由PPO预训练并在协同阶段保持策略固定，使UAVs训练时的环境动态主要来源于USV轨迹而非策略变化，从而有效降低多智能体学习中的环境非平稳性．在此基础上，MADDPG的Critic网络可同时获取所有UAVs的状态与动作信息，使策略更新更准确地作用于编队与协同误差的联合优化．相比之下，MAPPO缺乏集中式价值评估，Q-MIX受可加性约束限制，MAAC在高维状态下易产生冗余注意力，均难以有效刻画强耦合协同行为．

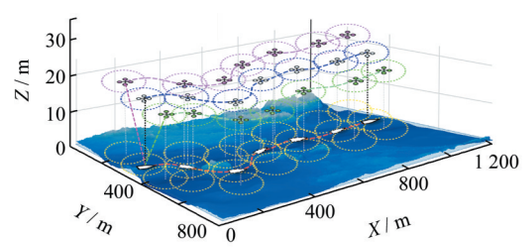
4.2　UAVs-USV协同决策泛化性仿真验证

为验证本文模型在复杂海洋环境中的有效性，分别在静态与动态水域环境下开展UAVs-USV协同决策仿真实验．实验场景均由1艘USV与3架UAVs组成，其中动态环境进一步引入3艘动态目标船作为时变干扰源．

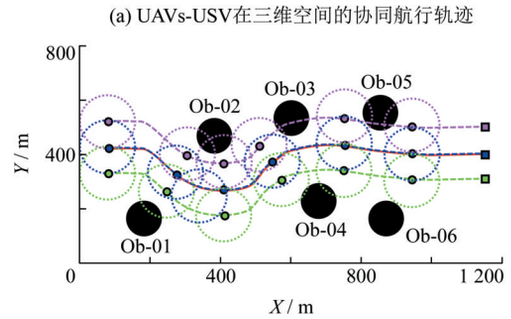
静态与动态环境下的UAVs-USV协同航迹分别如图9和图10所示，对应性能指标如表1和表2所示．图9和10中：Ob-01～Ob-06为静态障碍物；TS-01～TS-03为目标船舶(动态障碍物)．

可以看出：在两种环境下，USV均能够完成自主避障并稳定到达目标位置，3架UAVs始终保持稳定编队结构，实现对USV的协同跟踪．

编队误差与协同误差变化曲线如图11所示．结果表明：在静态与动态环境下，编队误差与协同



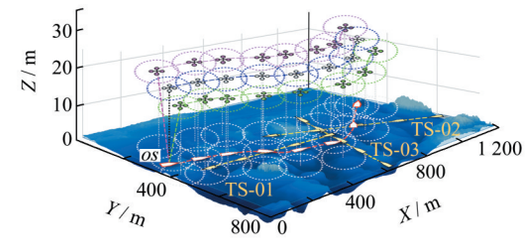
(a) UAVs-USV在三维空间的协同航行轨迹



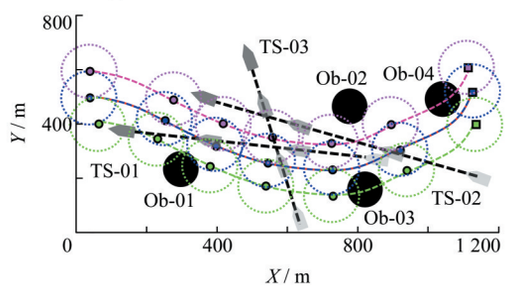
(b) UAVs-USV在二维空间的协同航行轨迹投影

---USV航行轨迹；---UAV-01飞行轨迹；
---UAV-02飞行轨迹；---UAV-03飞行轨迹(下同)．

图9　静态环境中UAVs-USV编队协同航行轨迹



(a) UAVs-USV在三维空间的协同航行轨迹



(b) UAVs-USV在二维空间的协同航行轨迹投影

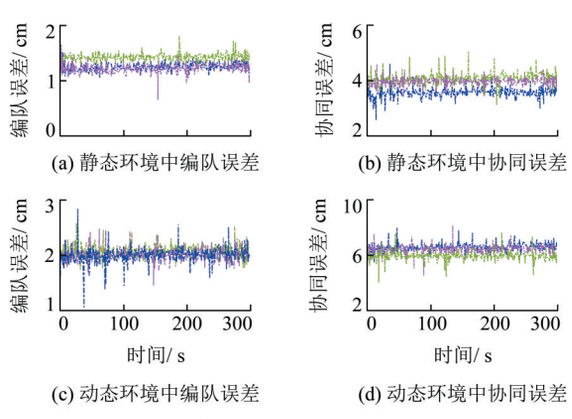
图10　动态环境中UAVs-USV编队协同航行轨迹

表1　静态环境中UAVs-USV编队协同性能指标值

	平均编队误差/cm	平均协同误差/cm	计算效率/s
UAV-01	1.38	3.44	0.129
UAV-02	1.43	4.17	0.117
UAV-03	1.22	3.89	0.125

表2　动态环境中UAVs-USV编队协同性能指标值

	平均编队误差/cm	平均协同误差/cm	计算效率/s
UAV-01	1.96	6.58	0.158
UAV-02	2.07	5.99	0.173
UAV-03	2.13	6.21	0.164



(a) 静态环境中编队误差　(b) 静态环境中协同误差

(c) 动态环境中编队误差　(d) 动态环境中协同误差

---UAV-01；---UAV-02；---UAV-03．

图11　静态与动态环境中UAVs-USV编队误差与协同误差随时间变化曲线

误差均能够快速收敛并稳定在较低水平．即使在动态避障过程中，UAVs编队仍能及时适应USV航向变化，并在避障结束后快速恢复稳定编队结构，未出现明显误差发散现象，验证了本文方法在复杂动态环境中的协同稳定．

5　结语

本研究提出一种基于分布式强化学习的UAVs-USV协同决策方法．通过将系统解耦为USV自主航行子系统与UAVs编队协同子系统，分别采用PPO与MADDPG进行训练，实现了局部自主避障与全局协同控制的协同优化；同时，结合GRU多模态网络与基于RVO的奖励约束机制，提升了系统对动态环境的感知能力及协同避碰性能．仿真与对比实验结果表明：本文方法能够在复杂动态环境下实现稳定编队保持与自主避障，具有较好的协同精度与泛化能力．未来将进一步考虑通信延迟、空域障碍物及动力学约束等实际因素，以提升系统的工程应用能力．

参　考　文　献

[1] LI J Q，ZHANG G Q，ZHANG X K，et al．Integrating dynamic event-triggered and sensor-tolerant control：application to USV-UAVs cooperative formation system for maritime parallel search[J]．IEEE Transactions on Intelligent Transportation Systems，2024，25(5)：3986-3998．

[2] LI S L，ZHU Y K，GUO G，et al．A separation and rendezvous control method for the UAV-USV system based on distributed NMPC[J]．IEEE Transactions on Intelligent Vehicles，2024，9(11)：7251-7263．

[3] 赵恩娇，莫小艳，王洪志，等.面向水面目标搜索的无人机-无人艇集群编队控制方法[J/OL]．[2026-01-17]．https://link.cnki.net/urlid/10.1567.TJ.20260414.1804.002．

[4] 张国庆，徐轶晖，李纪强，等．基于异步搜寻制导的机/船协同事件触发控制[J]．系统工程与电子技术，2025，47(9)：3058-3065．

[5] CUI Z W，GUAN W，ZHANG X K，et al．Autonomous collision avoidance decision-making method for USV based on ATL-TD3 algorithm[J]．Ocean Engineering，2024，312：119297．

[6] ZHU W，HAYASHIBE M．A hierarchical deep reinforcement learning framework with high efficiency and generalization for fast and safe navigation[J]．IEEE Transactions on Industrial Electronics，2023，70(5)：4962-4971．

[7] CUI Z W，GUAN W，HAO S H，et al．Heterogeneous agent proximal policy optimization for UAVs-USV formation navigation with multi-modal perception[J]．Ocean Engineering，2026，353：124690．

[8] CAO J J，WANG Y J，LIU Y，et al．Multi-robot learning dynamic obstacle avoidance in formation with information-directed exploration[J]．IEEE Transactions on Emerging Topics in Computational Intelligence，2022，6(6)：1357-1367．

[9] LOU M M，YANG X F，HU J B，et al．Design and field test of collision avoidance method with prediction for USVs：a deep deterministic policy gradient approach[J]．IEEE Internet of Things Journal，2025，12(3)：3363-3372．

[10] 梁成庆，李蕾，刘磊．基于CNN-LSTM融合网络的深度强化学习多无人机自主避碰[J/OL]．[2026-01-17]．https://doi.org/10.12194/j.ntu.20241112003．

[11] 邓明星，李爽，曾晨，等．基于规则操作性PPO的无人船避碰方法研究[J]．华中科技大学学报(自然科学版)，2026，54(4)：94-101．

[12] LABBADI M，CHERKAOUI M．Robust adaptive backstepping fast terminal sliding mode controller for uncertain quadrotor UAV[J]．Aerospace Science and Technology，2019，93：105306．

[13] ZHAO Y J，MA Y，HU S L．USV formation and path-following control via deep reinforcement learning with random braking[J]．IEEE Transactions on Neural Networks and Learning Systems，2021，32(12)：5468-5478．