计算机科学
COMPUTER SCIENCE

# 基于改进近端策略优化的无人艇自主避障方法

孔超, 王维, 皇苏斌, 张义, 孟丹

引用本文

孔超, 王维, 皇苏斌, 张义, 孟丹. 基于改进近端策略优化的无人艇自主避障方法[J]. 计算机科学, 2025, 52(4): 40-48.

KONG Chao, WANG Wei, HUANG Subin, ZHANG Yi, MENG Dan. Autonomous Obstacle Avoidance Method for Unmanned Surface Vehicles Based on ImprovedProximal Policy Optimization [J]. Computer Science, 2025, 52(4): 40-48.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000084

# 基于改进近端策略优化的无人艇自主避障方法

孔 超[1] 王 维[1] 皇苏斌[1] 张 义[1] 孟 丹[2]
1 安徽工程大学计算机与信息学院 安徽 芜湖 241000
2 OPPO 研究院 广东 深圳 518000
(kongchao@ahpu.edu.cn)

**摘 要** 无人艇自主避障已成为其拓展应用场景的一项关键挑战。传统方法下无人艇避障主要依赖于对环境的精细建模,然而,复杂海洋环境下无人艇难以获取完整的感知状态,导致模型精度不足。针对上述问题,提出了一种改进近端策略优化的无人艇自主避障方法。首先,构建了基于马尔可夫决策过程的无人艇自主避障决策框架;然后,在近端策略优化算法中融合了循环神经网络的感知表征增强模块,提高无人艇对时序环境感知的记忆能力;最后,结合奖励重塑机制设计一套自主避障奖励函数,提升无人艇避障策略的优化速度。为了验证算法的有效性,在三维仿真平台下构建了典型无人艇自主避障算法的验证场景。实验结果表明,基于改进近端策略优化方法能够实现无人艇无碰撞自主航行,在模型收敛速度、碰撞率与超时率上均优于传统近端策略算法。

**关键词**:无人艇;自主避障;近端策略优化;时序决策;奖励重塑
**中图分类号** U664.82

# Autonomous Obstacle Avoidance Method for Unmanned Surface Vehicles Based on Improved Proximal Policy Optimization

KONG Chao[1],WANG Wei[1],HUANG Subin[1],ZHANG Yi[1] and MENG Dan[2]
1 School of Computer and Information,Anhui Polytechnic University,Wuhu,Anhui 241000,China
2 Oppo Research Institute,Shenzhen,Guangdong 518000,China

**Abstract** Autonomous obstacle avoidance has become a critical challenge for expanding the application scenarios of unmanned surface vehicles(USVs). Traditional methods for USVs obstacle avoidance mainly rely on fine-grained environmental modeling. However,in complex marine environments,USVs have difficulty obtaining complete perception states,leading to insufficient model accuracy. To address this issue,we propose an improved proximal policy optimization(PPO)-based autonomous obstacle avoidance method for USVs. First,a perception and decision framework for USVs based on Markov decision process is constructed. Then,a feature-sharing representation optimization module is designed by fusing recurrent neural networks to enhance the USV's memory ability for temporal environmental perception. Finally,an autonomous obstacle avoidance reward function is designed by combining reward reshaping mechanisms to improve the optimization speed of the USV obstacle avoidance strategy. To verify the effectiveness of the proposed algorithm,a typical USV autonomous obstacle avoidance algorithm verification scenario is constructed on a three-dimensional simulation platform. Experimental results show that the improved PPO-based method can achieve collision-free autonomous navigation for USVs and outperforms the traditional PPO algorithm in terms of model convergence speed,collision rate,and timeout rate.

**Keywords** Unmanned surface vehicles,Autonomous obstacle avoidance,Proximal policy optimization,Temporal perception,Reward shaping

---

到稿日期:2024-10-17 返修日期:2025-02-18
基金项目:安徽省高等学校科学研究项目(2023AH050914,2024AH052239);安徽省高等学校省级质量工程项目(2023zybj018);安徽省自然科学基金(2308085MF220);芜湖市科技计划项目(2023pt07,2023ly13);安徽工程大学本科教学质量提升计划项目(2022lzyybj02,2023jyxm15,2024jyxm76)
This work was supported by the Science Research Project of Anhui Higher Education Institutions(2023AH050914,2024AH052239),Quality Engineering Project of Anhui Higher Education Institutions(2023zybj018),Anhui Provincial Natural Science Foundation(2308085MF220),Science and Technology Project of Wuhu City(2023pt07,2023ly13) and Quality Improvement Program of Anhui Polytechnic University(2022lzyybj02,2023jyxm15,2024jyxm76).
通信作者:孟丹(mengdan90@163.com)

## 1　引言

无人艇是海上机器人系统，是海洋网络化无人系统中的重要节点。通过搭载各类传感设备，无人艇可广泛用于海洋运输与环境调查、水上搜救、情报搜集、警戒巡逻等领域[1]。无人艇是一项颠覆传统海战样式的系统，将催生全新的海洋装备体系，对海洋资源开发和国家海洋权益维护具有重要的意义[2]。

复杂海洋环境下的自主避障能力是当前无人艇拓展应用场景的需求之一[3]。传统无人艇避障方法依赖人工构建大量规则与专家知识库，通过预先精细化任务建模，来实现对特定任务的处理[4]。然而，在面对突发状况时，传统方法的预先建模往往难以覆盖所有场景，同时缺乏自主学习能力，泛化性、适应性较弱，应用范围有限。

近年来，嵌入式人工智能为智能无人系统的自主决策带来了新的范式[5]。基于嵌入式人工智能技术，无人系统能够在本地实时感知环境、分析数据、做出决策，从而实现更高程度的自主性和智能化[6]。嵌入式人工智能中，深度学习带来的感知与认知能力已经在诸多领域超越人类水平[7]，强化学习在深度学习的基础上取得了突破性进展。目前，深度强化学习算法已在电子竞技[8]、自动驾驶[9]、机器人控制[10]等领域取得诸多研究成果。相比于传统无人系统控制方法，基于近端策略优化(Proximal Policy Optimization，PPO)[11]的强化学习方法生成的策略拥有强大的表达能力，能够通过持续学习应对各种突发状况，从而支撑无人系统完成更多复杂任务，其泛化能力与自适应能力远超传统方法。

然而，在资源受限的嵌入式平台上实现实时、高效的避障决策仍然存在挑战。首先，在复杂海洋环境下，无人艇传感器观测区域分布的动态变化多样，环境状态难以准确完备感知，无人艇避障策略收敛依赖于对环境的充分感知，结构复杂的感知网络会给无人艇平台带来极大的计算压力。其次，近端策略优化算法的策略收敛速度依赖于良好的奖励函数设计，如果仅采用稀疏的回合奖励作为训练样本，则会极大地拉长整个训练周期，进而造成训练成本提高。

因此，如何设计基于近端策略优化算法的轻量级无人艇感知表征增强机制，实现对复杂海洋场景中的不完备感知到较完备的感知映射，以及对无人艇自主避障任务设计出较为合适的奖励函数，是实现无人艇自主避障策略生成的关键问题。

为了解决上述问题，本文提出基于改进近端策略优化的无人艇自主避障方法。本文的主要贡献如下：

1)设计了一种融合循环神经网络的轻量级感知表征增强模块，降低近端策略优化算法对状态价值评估的偏差，加强无人艇避障过程对环境的时序记忆能力；

2)设计了一套适用无人艇自主避障任务的瞬时奖励函数，从而有效克服无人艇避障过程中的奖励稀疏问题，提高模型收敛速度；

3)通过仿真实验验证了所提方法的有效性，以及在任务成功率、碰撞率和超时率上的性能优势。

本文第 2 章阐述了国内外相关工作；第 3 章介绍了本文提出的改进近端策略优化算法；第 4 章开展了对算法的实验验证；最后总结全文。

## 2　相关工作

无人艇的自主避障是实现其自主航行的一项关键技术，具体避障方法可分为两大类，即基于规则与先验知识的方法，以及基于学习的方法。基于规则与先验知识的方法主要通过预先设定一系列规则和模型，使无人艇能够在复杂的水域环境中安全高效地航行。Guan 等[12]提出了一种结合改进型 $A^*$ 和动态窗口方法，避免陷入局部优化，实现了无人艇在复杂场景下的自主避障。Bai 等[13]基于植物生长的趋光性原理，设计了无人艇导航与避障算法，实验表明，该算法能成功避开静态障碍物，同时在搜索时间和效率上有着一定优势。Yu 等[14]通过改进路径代价函数，以及减少节点扩展范围，提出了一种基于改进 D * LITE 算法的无人艇路径规划算法。与传统的 D * LITE 算法相比，改进算法实现了无人艇路径长度最短，规划时间最短，路径最平滑。Ouyang 等[15]针对突发障碍物与非严格保形规划点碰撞问题，提出了基于改进快速搜索随机树算法的无人艇编队路径规划技术，通过在碰撞检测环节提出可调节避碰圆区域与障碍物修正向量，使无人艇安全避碰并最大程度地保持队形稳定。然而，上述方法主要依赖于精确的环境模型和无人艇动力学模型，通过数学建模和优化算法来规划出一条安全的避障路径，如果环境不确定性增加，那么传统方法容易陷入局部最优。同时，建立精确的数学模型和求解优化问题需要大量的计算资源，这对于实时性要求高的无人艇系统来说是一个极大的挑战。

基于学习的无人艇自主避障方法不依赖于人工设计的特征，无需设计人员手动指定决策规则，而是能够基于训练样本自主学习和优化策略，因此具有更强的泛化能力[16]。Wu 等[17]提出了一种基于深度强化学习的无人艇自主避障方法，通过量身定制的状态和行动空间设计，实现了在复杂环境下的自主航行。Xu 等[18]提出了一种跟踪当前网络权重更新目标网络权重的深度强化学习方法，提高了算法学习最优策略的稳定性，确保无人艇在遵循国际避障规则避开障碍物的同时导航到目标。Gan 等[19]采用深度强化学习方法，为无人艇自主导航任务精心设计了观测空间、行动空间、奖励函数，使得无人艇仅依靠本地传感器实现无人艇无地图自主导航。Wang 等[20]采用近端策略优化算法，结合奖励重塑机制，为无人艇训练过程提供单步奖励，最终实现基于视觉的无人艇自主避障。综上，深度强化学习在无人艇自主避障领域展现出了巨大的潜力，但仍存在一些挑战和局限性，如模型训练时间长，对环境状态变化敏感性强，以及奖励信号稀疏导致策略优化难的问题等。

## 3　本文方法

本章首先对无人艇自主避障任务进行马尔可夫决策过程建模，其次设计了无人艇感知表征增强模块，然后采用奖励重塑机制对任务进行了奖励函数设计，最后将上述相关改进融合到近端策略优化算法决策框架中。

### 3.1 无人艇自主避障 MDP 建模

本文采用深度强化学习方法来解决无人艇自主避障问题，因此在本节中，首先采用马尔可夫决策过程(Markov Decision Process，MDP)对任务进行建模。MDP 框架主要由〈$S$，$A$，$P$，$R$，$\gamma$〉五元组构成。其中，$S$ 代表环境状态集合，$A$ 代表无人艇的动作集合，$R$ 为奖励函数，$\gamma$ 用于权衡当前奖励与未来奖励重要性的贴现因子。无人艇与环境在 MDP 框架内的交互过程如图 1 所示。

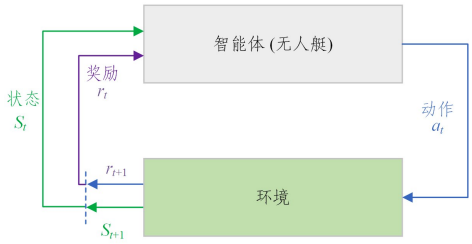



图 1 马尔可夫决策过程

Fig. 1 Markov decision process

首先，无人艇在第 $t$ 时刻通过传感器感知周围环境状态 $s_t$，随后根据当前策略执行动作 $a_t$，此时环境会向无人艇反馈奖励 $r_t$。至此，无人艇会进入 $t+1$ 时间步，并继续感知环境、执行动作以及收集奖励，直至本次回合交互结束。在本文的设定中，当无人艇发生碰撞或无人艇达到目标位置时，该回合的交互立即结束；此外还设定了最大回合步数，当无人艇与环境的交互超过了最大回合步数，那么该回合的交互也会立即结束。

复杂海洋环境包括各种障碍物和预定的目的地点。本研究着重考虑了无人艇与障碍物的相对距离，以及无人艇和目标点位置的相对距离，故单一时刻下的环境状态 $\boldsymbol{s}_t$ 包含了无人艇的自身状态信息，以及无人艇利用传感器感知周围障碍物的距离信息，如式(1)所示：

$$\boldsymbol{s}_t=\{USV_{\text{state}},USV_{\text{sensor}}\} \tag{1}$$

其中，自身状态信息 $USV_{\text{state}}$ 包含无人艇与目标点的相对距离、无人艇艏向角、速度以及上一时刻的动作。

无人艇的动作由舵角和油门两部分构成，其中舵角会直接影响到无人艇的艏向角的变化，油门将直接影响无人艇速度的变化。如式(2)所示，在时刻 $t$ 下，无人艇舵角可以在 $-60°$到 $60°$之间直接调节，油门的值在 0 到 1 之间调节。

$$a_t=\begin{cases}rudder\in\left[-\dfrac{\pi}{3},\dfrac{\pi}{3}\right]\\ \text{throttle}\in[0,1]\end{cases} \tag{2}$$

在无人艇避障任务期间，当油门给定值为 0 时，无人艇在摩擦力和惯性作用下速度逐渐降为 0，此时舵角无法控制其艏向角改变。如果给定的油门值为 1，那么无人艇将加速至全速。无人艇的主要组件如图 2 所示。

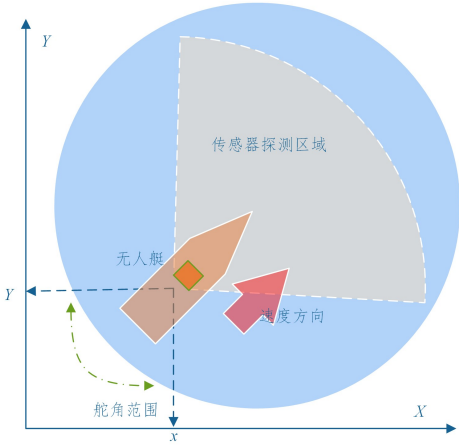



图 2 无人艇主要组件

Fig. 2 Major components of USV

### 3.2 感知表征增强模块

为了提升无人艇在复杂环境下的感知能力，本文引入轻量级感知表征增强模块，以提升近端策略优化算法的时序决策能力。如图 3 所示，将用于无人艇自主避障的网络模型分为两部分：感知增强模块以及近端策略优化算法中的 Actor-Critic 模块。

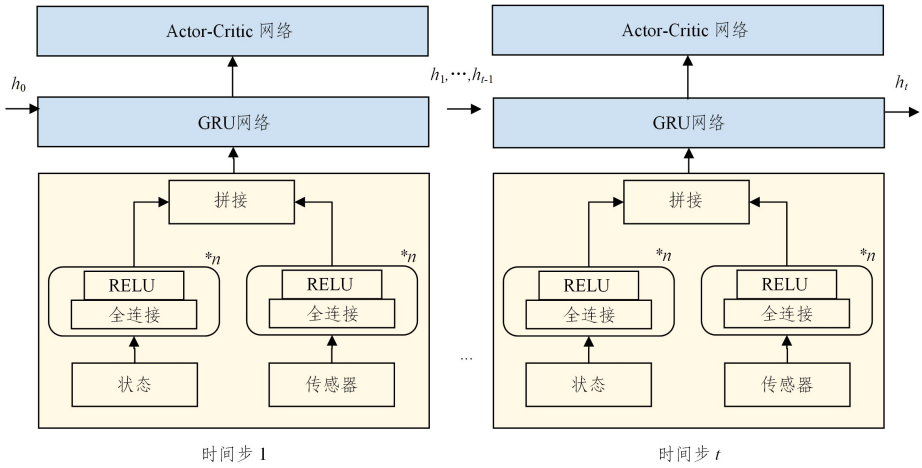



图 3 感知表征增强模块

Fig. 3 Perception representation enhancement module

如图 3 所示，感知表征增强模块的处理流程包含以下步骤。首先，在时间步 $t$ 中，采用具有 RELU[21] 激活函数的多层感知机分别从无人艇状态信息和传感器信息中进行特征提取，并将提取的特征进行连接，以获得在时间步 $t$ 中融合的当前环境特征信息。此时无人艇的一条航迹信息表示如下：

$$\tau=\{\boldsymbol{o}_0,\boldsymbol{a}_0,\boldsymbol{r}_0,\cdots,\boldsymbol{o}_t,\boldsymbol{a}_t,\boldsymbol{r}_t,\cdots\} \tag{3}$$

其中，$\boldsymbol{o}_t$代表在时间步 $t$ 中融合后的感知信息，$\boldsymbol{a}_t$为 $t$ 时刻下无人艇执行的动作，$\boldsymbol{r}_t$为 $t$ 时刻下无人艇获取到的奖励。其次，将融合后的$o_t$输入到循环神经网络中。本文采用门控神经单元(Gated Recurrent Unit，GRU)[22]，实现对经过特征提取状态值的进一步处理，以形成一个能更好地代表当前环境状态的表征。GRU 使用重置门$\boldsymbol{r}_t$和更新门$\boldsymbol{z}_t$，重置门作用于上一时刻的隐藏状态，决定需要遗忘多少历史信息；更新门用于当前时刻和上一时刻的隐藏单元，决定需要向下传递多少有用信息。具体计算过程如下：

$$\boldsymbol{r}_t=\sigma(Linear(\boldsymbol{o}_t,\boldsymbol{h}_{t-1})) \tag{4}$$

$$\boldsymbol{z}_t=\sigma(Linear(\boldsymbol{o}_t,\boldsymbol{h}_{t-1})) \tag{5}$$

其中，$\sigma$ 为 sigmoid 激活函数，Linear 为线性神经网络。$\boldsymbol{r}_t$和$\boldsymbol{z}_t$用于控制需要重置和更新多少上一时刻的候选隐藏状态$\boldsymbol{h}_{t-1}$。在完成对 GRU 门控信号的处理后，要进一步对重置门$\boldsymbol{r}_t$与前一时刻的隐藏状态$\boldsymbol{h}_{t-1}$和$\boldsymbol{o}_t$进行处理，通过式(6)获得时间步 $t$ 的候选隐状态信息$\boldsymbol{h}_t'$。

$$\boldsymbol{h}_t'=\mathrm{Tanh}(Linear(\boldsymbol{o}_t,\boldsymbol{h}_{t-1}\odot\boldsymbol{r}_t)) \tag{6}$$

其中，$\odot$为元素乘积计算，Tanh 为双曲正切激活函数。候选隐藏状态$\boldsymbol{h}_t'$是更新门和重置门共同作用的结果，它表示当前时刻的隐藏状态的更新情况，同时也参与到隐状态$\boldsymbol{h}_t$的计算中。隐状态$\boldsymbol{h}_t$的计算过程如下：

$$\boldsymbol{h}_t=\boldsymbol{z}_t\odot\boldsymbol{h}_{t-1}+(1-\boldsymbol{z}_t)\odot\boldsymbol{h}_t' \tag{7}$$

隐状态$\boldsymbol{h}_t$是候选隐状态和前一时间步的隐状态的线性组合，并由更新门控制更新幅度。综上所述，在引入 GRU 网络后，近端策略优化算法使得无人艇时序感知数据时能更好地捕捉长期依赖关系，同时保持相对较低的计算复杂度。此时，无人艇的一条航迹信息如下：

$$\tau=\{\boldsymbol{h}_0,\boldsymbol{a}_0,\boldsymbol{r}_0,\cdots,\boldsymbol{h}_t,\boldsymbol{a}_t,\boldsymbol{r}_t,\cdots\} \tag{8}$$

与未处理感知信息前相比，此时 $t$ 时刻下无人艇环境感知信息$\boldsymbol{o}_t$转变为$\boldsymbol{h}_t$，无人艇已完成对感知的表征增强；最后，用$\boldsymbol{h}_t$替代$\boldsymbol{o}_t$输入到近端策略算法中的 Actor-Critic 网络中，其中 Actor 网络生成无人艇的决策(油门和舵角)，Critic 网络评估状态价值。本文采用 Actor 网络与 Critic 网络共享特征提取网络，减少整个网络模型的参数规模。

### 3.3　奖励函数设计

传统回合奖励设计会导致无人艇只能在回合结束时获取到实质奖励值，这会导致高价值场景的奖励稀疏情况，使得无人艇避障策略难以在短期内实现收敛。因此，在无人艇自主避障任务中，我们使用奖励重塑机制[23]确保无人艇在单步行动中都能获得实质性奖励，从而提高近端策略优化算法的收敛速度。具体来说，我们设计了两类单步化奖励函数。一类是 USV 的分步导航奖励函数，主要包括距离奖励、碰撞奖励和停滞奖励。式(9)为距离奖励，其中，$P_t$为时刻 $t$ 下无人艇与任务点之间的距离，$P_{t-1}$为时刻 $t-1$ 下无人艇与任务点之间的距离，两者之间的差值意味着无人艇在一个时间步内与任务点之间距离的变化情况，距离越近则会产生正奖励，距离越远则产生负奖励。$\lambda_{\mathrm{distacne}}$为距离奖励权重，设置为 0.01。

$$r_{\mathrm{distance}}=-\lambda_{\mathrm{distacne}}(P_t-P_{t-1}) \tag{9}$$

式(10)为碰撞奖励，当无人艇与障碍物发生碰撞后，则会给予值为$-1$的负奖励，此时回合也会直接结束，其他情况下碰撞奖励一直为 0。为了防止在训练过程中无人艇出现停滞来规避碰撞奖励，在每一时间步引入了较小的停滞负奖励。在本文中，无人艇在每一时间步中都会被赋予$-0.001$的奖励值，如果无人艇一直保持停滞或无意义游走，则会在一个回合下累积到较大的负奖励。引入停滞负奖励会促使无人艇学习自主避障策略。

$$r_{\mathrm{collide}}=\begin{cases}-1, & \mathrm{collide}\\ 0, & \mathrm{otherwise}\end{cases} \tag{10}$$

另一类为无人艇的动作稳定约束奖励函数。在该奖励约束下，无人艇的自主避障模型能够生成较为柔顺的动作轨迹，从而降低无人艇硬件磨损。式(11)计算了相邻时刻下无人艇动作变化的幅度，其中$a_t^i$为无人艇在 $t$ 时刻下第 $i$ 个动作的具体数值。在完成了无人艇连续时刻之间动作幅度计算后，还需对幅度进行归一化操作，从而获得更加有效的奖励值。

$$S(a_t,a_{t+1})=\sqrt{\sum_{i=1}^{n}(a_t^i-a_{t+1}^i)^2} \tag{11}$$

在式(12)中，将动作幅度进行了归一化操作，从而可以得到无人艇每一时刻的柔顺奖励。

$$r_{\mathrm{smooth}}=-\lambda_{\mathrm{smooth}}\frac{S(a_t,a_{t+1})}{\max(S(a_t,a_{t+1}))} \tag{12}$$

其中，$\lambda_{\mathrm{smooth}}$为柔顺奖励权重。为了避免无人艇过度追求柔顺而不考虑避障和航行，本文将其权重设置为 0.001。

### 3.4　改进近端策略优化算法决策框架

本文基于近端策略优化算法构建无人艇自主避障最优学习策略。近端策略优化算法是一种典型的 Actor-Critic 架构算法，与其他 AC 框架算法相比，近端策略优化算法的优点在于引入优势函数来评估当前无人艇动作的收益情况，其计算式如式(13)所示：

$$\boldsymbol{A}_t=\sum_{t'>t}\boldsymbol{\gamma}^{t'-t}\boldsymbol{r}_{t'}-\boldsymbol{V}_{\phi}(\boldsymbol{s}_t) \tag{13}$$

其中，$\boldsymbol{r}_{t'}$为无人艇在$t'$时刻获得的瞬时奖励，$\boldsymbol{\gamma}$为奖励折扣因子，$\boldsymbol{V}_{\phi}(\boldsymbol{s}_t)$为状态价值函数，$\phi$为 Critic 网络的参数。优势函数用于计算在无人艇状态$\boldsymbol{s}_t$中执行动作$\boldsymbol{a}_t$所获得的奖励与期望状态奖励相比的优势。在改进近端策略优化算法中，状态$s_t$经过表征增强处理，因此能够更加准确地描述当前真实环境。同时，在每一时刻的瞬时奖励都经过了奖励重塑处理，从而避免了奖励稀疏的情况，为后续网络优化提供了支撑。状态价值函数$\boldsymbol{V}_{\phi}(\boldsymbol{s}_t)$的计算过程如式(14)所示：

$$\boldsymbol{V}_{\phi}(\boldsymbol{s}_t)=\boldsymbol{r}_t+\boldsymbol{\gamma}\boldsymbol{r}_{t+1}+\cdots+\boldsymbol{\gamma}^n\boldsymbol{V}_{\phi}(\boldsymbol{s}_{t+n}) \tag{14}$$

状态价值函数需要大量经验进行优化，这需要无人艇不断与环境进行交互来产生足够的真实状态值，从而达到拟合效果。式(15)为 Critic 网络的损失函数，通过不断优化损失函数可以使得状态价值函数趋向于真实，这也为后续 Actor 网络的更新提供了支撑。

$$\boldsymbol{Loss}(\phi)=-\sum_{t=1}^{T}(\sum_{t'>t}\boldsymbol{\gamma}^{t'-t}\boldsymbol{r}_{t'}-\boldsymbol{V}_{\phi}(\boldsymbol{s}_t))^2 \tag{15}$$

此外，近端策略优化算法允许在策略更新过程中重复使用旧策略的样本，这极大提高了无人艇样本利用效率。实现该机制的原理在于近端策略优化算法引入了重要性采样，

具体计算过程如式(16)所示：

$$\boldsymbol{r}_t(\theta)=\boldsymbol{\pi}_\theta(\boldsymbol{a}_t\mid\boldsymbol{s}_t)/\boldsymbol{\pi}_{\theta_{\text{old}}}(\boldsymbol{a}_t\mid\boldsymbol{s}_t) \tag{16}$$

其中，$\boldsymbol{\pi}_\theta$为当前优化更新的 Actor 策略，而$\boldsymbol{\pi}_{\theta_{\text{old}}}$为旧策略参数，不参与更新，$\theta$为 Actor 网络参数。由于重要性采样使用了过去策略采样数据，因此梯度方差更小，训练更加稳定；同时算法可以重复使用过去策略采样数据，采样效率更高。在有了重要性比重$\boldsymbol{r}_t(\theta)$和优势函数$\boldsymbol{A}_t$后，要确保无人艇策略朝着优势函数更大的方向优化，最终可以得到无人艇策略优化目标函数，即 Actor 网络优化函数，具体如式(17)所示：

$$\boldsymbol{J}(\boldsymbol{\theta})=\sum_{t=1}^{T}\min(\boldsymbol{r}_t(\boldsymbol{\theta})\boldsymbol{A}_t, clip(\boldsymbol{r}_t(\boldsymbol{\theta}),1-\epsilon,1+\epsilon)\boldsymbol{A}_t) \tag{17}$$

其中，$clip$为梯度裁剪函数，该函数通过引入一个超参数$\epsilon$来限制无人艇避障策略更新的幅度，这使得重要性采样比率不会过大，避免了避障策略更新过度导致训练不稳定。综上可得改进近端策略优化算法决策框架图，如图 4 所示。该框架图分为两大部分，即改进近端策略优化算法的主体结构以及环境部分。

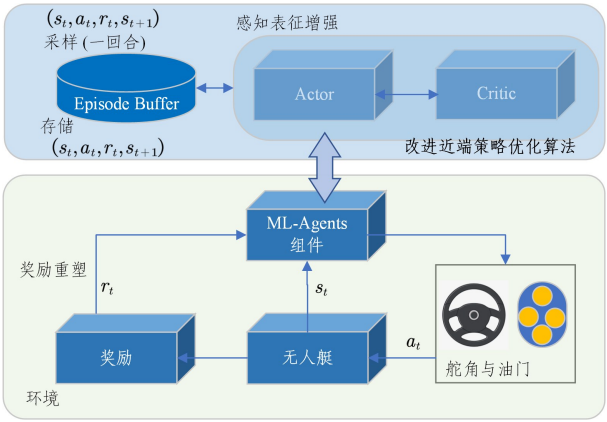



图 4 改进近端策略优化算法框架

Fig. 4 Framework for Improved-PPO

主体结构包括 Actor-Critic 架构以及回合经验缓冲区。算法在优化过程中会进行回合经验采样，其中采样的数据会经过感知表征增强处理，经处理后的样本会更加准确地评估当前状态价值。环境部分则用于产生训练数据，其中奖励模块中的奖励重塑机制可以在单步中产生有价值的瞬时奖励，避免无人艇在单步交互中出现稀疏奖励的情况。图 4 中，ML-Agents[24]组件为算法训练提供环境信息交互的作用。通过该组件，能够设定各种实验条件，如不同布局的障碍物，包括稀疏和密集的障碍物场景，使无人艇在这些模拟环境中进行训练和学习；同时，它可以传递训练数据，从而支撑算法的训练。

图 5 为基于改进近端策略优化的无人艇自主避障方法流程图。左侧区域为算法驱动的无人艇探索-执行部分，具体流程为，无人艇通过传感设备感知周围环境，并将感知信息传输至感知表征增强网络中处理，处理完毕后由 Actor-Critic 网络分别产生动作和对状态进行评估。右侧区域为算法的训练部分，即无人艇将相关交互信息进行存储，当回合结束后进行参数优化，当满足训练结束条件后退出训练并保存模型。

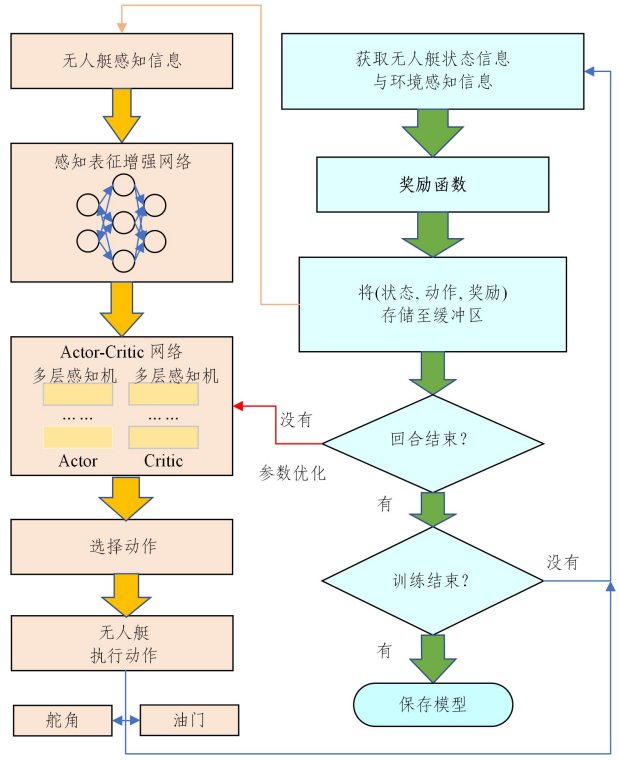



图 5 基于改进近端策略优化的无人艇自主避障方法流程图

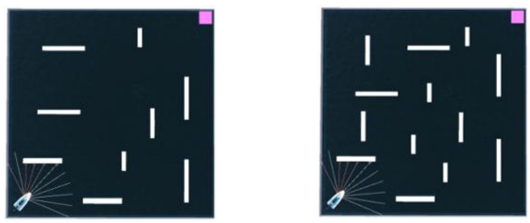
Fig. 5 Flowchart of autonomous obstacle avoidance method for USV based on Improved-PPO

## 4 实验与性能分析

本章主要对改进近端策略优化算法进行无人艇自主避障实验验证。首先，设计了无人艇自主避障实验场景和改进近端策略优化算法参数。随后，在实验场景中对算法的性能进行评估与分析。

### 4.1 实验设置

本文基于虚拟场景和 ML-Agents 组件构建了典型的两个无人艇避障任务场景。图 6(a)为较为稀疏障碍物布局的任务场景，图 6(b)为较为密集障碍物布局的任务场景。



(a) (b)

图 6 实验环境(电子版为彩图)

Fig. 6 Experimental environment

图 6 中，右上方红色标记处为目标位置，白色方块为障碍物。无人艇被限制在一个方形环境内，需要从场景的左下方自主出发，通过传感器感知周围环境，并结合算法进行实时避障操作并最终达到目标位置。本次实验基于 16 核 CPU 的 Windows 平台，64 GB DDR4 DRAM 以及英伟达 3090TI GPU 实现。实验算法的超参数设置如表 1 所列。

表 1　算法超参数设置

Table 1　Algorithm hyperparameter settings

| 超参数 | 值 |
|---|---|
| 折扣因子 | 0.99 |
| 回合最大步数 | 1 000 |
| 优化器 | Adm |
| Clip | 0.2 |
| 激活函数 | Tanh |
| 表征层 | (128,128,128) |
| GRU 单元 | 256 |
| Actor 隐藏层 | (256,256) |
| Critic 隐藏层 | (256,256) |
| Actor 网络学习率 | 0.000 2 |
| Critic 网络学习率 | 0.000 1 |

本文设计了 3 种无人艇自主避障任务结束条件。首先，将每回合的最大步数限制为 1 000 步，如果无人艇在该回合的行动步数超过了这一限制，则该回合立即结束。其次，如果无人艇与障碍物相撞，则该回合立即结束。最后，如果无人艇无碰撞地航行至目标位置，则该回合立即结束。一旦回合结束，那么无人艇将会重新回到初始位置，并进行新一轮的训练。

## 4.2　结果分析

本文通过评估算法的相关典型性能指标来验证其在无人艇自主避障任务下的性能：首先，通过回合奖励来验证算法的收敛速度；其次，通过统计训练期间无人艇的回合步数和最大行驶距离来评估算法在避障任务中的基础性能；最后，通过无人艇在训练过程中的动作差异来分析单步奖励函数设计的有效性。本文采用 3 种基准算法与改进近端策略优化算法进行对比。

1）PPO

PPO 算法是一种在强化学习中广泛应用的策略优化算法。其通过限制新旧策略之间的差异，在保证模型稳定性的同时提高学习效率，尤其在处理连续动作空间的问题时表现优异。

2）PPO_no_smooth_reward（No_smooth_reward）

PPO_no_smooth_reward 算法，在本文中也称为 No_smooth_reward 算法，即在 PPO 算法基础上移除了无人艇自主避障任务重的动作柔顺奖励设置。

3）Deep deterministic policy gradient（DDPG）[25]

DDPG 算法是一种用于解决连续动作空间强化学习问题的深度确定性策略梯度算法，它结合了值函数和策略梯度方法，通过 Actor-Critic 结构，实现了在连续动作空间中的策略学习。

### 4.2.1　回合累计奖励走势

回合累计奖励指强化学习算法中智能体在单一回合内所产生的所有奖励之和，这也是强化学习算法中最核心的性能指标。因此，本文首先分析不同算法的回合累计奖励走势，其直接体现无人艇自主避障策略的优化情况。

图 7 给出了在障碍物稀疏场景下无人艇自主避障训练过程中的奖励趋势，其中横轴和纵轴分别代表训练步数和回合数奖励值。图中，改进近端策略算法在 100 万步的训练中，奖励趋势呈现出持续上升的走势，在训练初始阶段，无人艇获得的累计奖励为负值，原因在于训练初始阶段无人艇还未掌握任何避障策略，因此比较容易发生碰撞，当发生碰撞后会立刻结束回合且返回一个较大的负奖励。

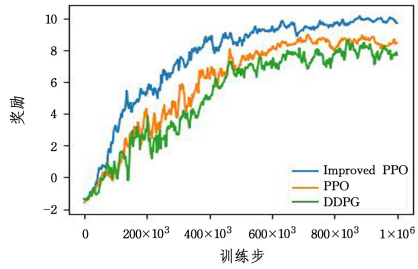



图 7　稀疏场景下算法奖励趋势

Fig. 7　Algorithmic reward trends in sparse scenario

随着训练的进行，无人艇自主避障策略也在不断优化，同时回合累计奖励走势也在持续不断上升，最终在 100 万步训练下无人艇自主避障模型达到收敛，改进近端策略优化算法能够获得约＋9 的奖励。相比之下，对比算法在相同训练步数下所获的回合累计奖励要更低，同时整体奖励走势出现了更大的抖动。其中，PPO 算法的奖励走势比 DDPG 算法略好，原因在于 PPO 通过限制策略更新的幅度来提高稳定性，减少了方差，因此训练过程中训练曲线抖动幅度更小，而 DDPG 受到高方差影响，需要更多的经验实现无人艇避障策略的收敛。综上所述，在固定训练步数情况下，改进近端策略优化呈现出的算法稳定性和避障性能要优于对比算法，核心原因在于改进近端策略优化算法在感知表征增强模块的加持下，能够更精准地评估无人艇单步感知状态的价值，从而强化无人艇时序决策能力。

图 8 给出了在密集障碍物场景无人艇自主避障过程中的奖励趋势。与稀疏场景中的实验结果类似，相比于基准近端策略优化算法，改进近端策略优化算法表现出更稳定的奖励曲线，并且在训练达到 100 万步后也获得了更高的奖励；同时，密集场景要求无人艇具备更强的避障能力，因此在相同训练阶段下算法的抖动程度比稀疏场景更高。

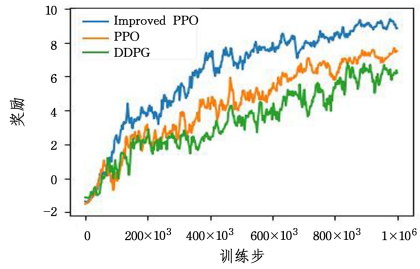



图 8　稀疏场景下算法奖励趋势

Fig. 8　Algorithmic reward trends in dense scenario

### 4.2.2　回合步数与航行距离统计

为了进一步评估算法的性能，本文还对算法训练过程中的回合长度进行了统计。回合长度指标代表无人艇在一回合中采取了多少步的动作，当无人艇具备自主避障能力后，回合长度越小，算法性能越好。本文在不同训练时间节点对不同算法进行了 3 次推演，并取均值作为最终统计结果。具体

统计如图 9 所示。

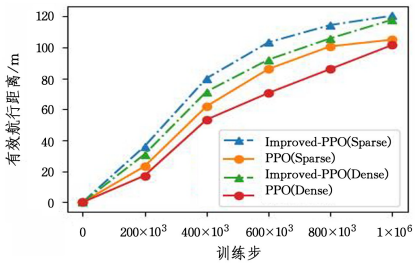



图 9 回合长度统计

Fig. 9 Episode length statistics

图 9 中，横轴和纵轴分别代表训练步数和回合长度。可以明显看出，在稀疏场景和密集场景下，无人艇自主避障的回合长度都呈现出相同趋势，即先上升后下降。出现该趋势的原因在于，训练早期阶段无人艇自主避障决策能力不足且极易发生碰撞，因此回合结束较快，导致回合长度值较小；随着训练的进行，无人艇也逐步具备避障能力，因此回合长度也在增加；在第 60 万步，改进近端策略优化算法和传统近端策略优化算法在稀疏和密集场景下的回合长度都为 800 左右，此时无人艇在当前避障策略辅助下实现了对环境的充分探索；在 60 万步后，无人艇的回合长度逐渐减少，这也意味着无人艇已逐步掌握避障策略，并开始在奖励函数约束下优化自身动作。

通过对比不同算法的回合长度，可以发现无论是在稀疏场景还是在密集场景下，改进近端策略优化算法的回合长度都比常规近端策略优化算法要短，这意味着无人艇能够以更少的动作来完成避障任务。更少的动作意味着无人艇对硬件的磨损更少，同时停滞奖励也更少，无人艇获取到的正向奖励也更多。在训练了 100 万步后，改进近端策略优化算法已达到收敛，此时回合长度降低到 400 左右，而传统近端策略优化算法的回合长度为 480 左右。

此外，无人艇在训练过程中的有效航行距离也可以反映不同算法的性能。在本文中，有效航行距离指无人艇从初始位置到回合结束位置时，两者与目标位置之间的距离差值。本文在不同训练时间节点对不同算法模型进行了 3 次推演，并取均值作为最终统计结果。图 10 给出了不同算法在不同场景下无人艇有效航行距离的统计。

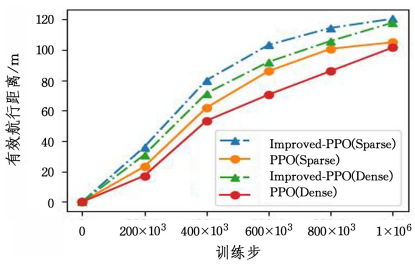



图 10 有效航行距离统计

Fig. 10 Valid distance travelled statistics

图 10 中，横轴和纵轴分别代表训练步数和无人艇有效航行距离。可以看出，在最初的 20 万步训练阶段中，改进近端策略优化算法能使无人艇有效航行超过 30 m，而对比算法只能达到 20 m。在 40 万步训练阶段，改进近端策略优化算法能够使无人艇有效航超 65 m，对比算法只能到达约 60 m。后续整体统计走势与上述一致，即改进近端策略优化算法在不同场景下都展现出了更好的航行能力。

#### 4.2.3 无人艇动作柔顺度评估

本文根据奖励重塑设计了无人艇自主避障任务单步奖励函数。其中，为了提高无人艇避障过程中的动作柔顺度，本文引入了动作稳定约束奖励函数。在本小节通过评估训练过程中的回合动作方差来验证该奖励函数的实际性能。具体实验结果如图 11、图 12 所示。

图 11 给出了稀疏障碍物场景下的实验结果，图中横轴和纵轴分别代表训练步数和回合动作方差。从图中可以看出，动作稳定约束奖励函数的算法在训练初始阶段，无人艇方向舵和油门的回合动作方差在 0.75。随着训练过程的推进，方向舵和油门的回合动作方差逐渐减小，经过 100 万步训练后，分别降至约 0.2 和 0.1。由于对比算法中没有引入动作稳定约束奖励函数，因此在整个训练过程中的回合动作方差出现较为剧烈的波动。经过 100 万步训练后，方向舵的回合动作方差会减小到约 0.35，油门回合动作方差减小到约 0.15，总体表现较差。

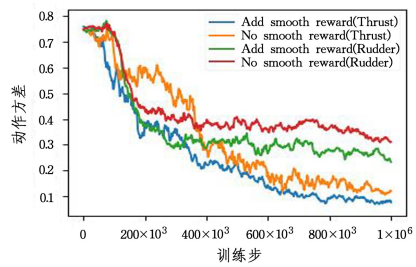



图 11 稀疏场景下无人艇动作方差变化趋势

Fig. 11 Trend of USV action variance in sparse scenario

图 12 给出了密集障碍物场景下的实验结果。与稀疏场景中的结果类似，引入动作稳定约束奖励函数的算法在整个训练过程中能更快地降低无人艇回合行动方差，使得无人艇在自主避障过程中表现出更稳定的性能。

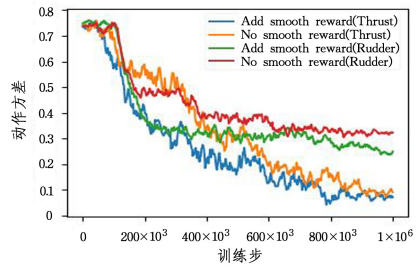


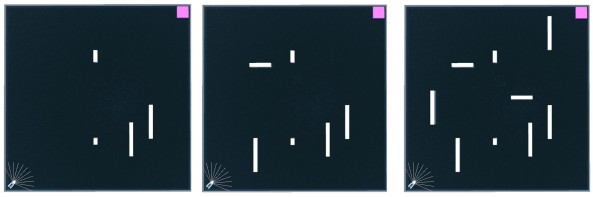

图 12 密集场景下无人艇动作方差变化趋势

Fig. 12 Trend of USV action variance in dense scenario

综上，无论是在稀疏场景还是在密集场景，添加动作稳定约束奖励函数都能够考虑连续两个时间步的动作幅度。当动作幅度过大时，会给予无人艇较大的负奖励，这使得无人艇自主避障策略在多次训练后会综合考虑动作幅度这一影响奖励走势的因素。

4.2.4　扩展场景下无人艇避障能力评估

为了进一步验证算法的避障能力，本文对实验场景规模进行了扩展，并形成如图 13 所示的无人艇自主避障测试环境。与原始实验场景相比，图 13 中的场景面积扩展了 4 倍，并根据障碍物的密集程度设置了稀疏、中等、密集 3 个子场景。



(a)　(b)　(c)

图 13　不同障碍物密集程度的无人艇避障测试场景

Fig. 13　Obstacle avoidance test scenarios for USV with different obstacle densities

在相关实验设置和算法参数保持不变的情况下，对所有算法进行了固定 150 万步的训练，在完成训练后，对模型进行 100 轮推演，并在表 2 中统计了不同算法的任务成功率和期望航行距离。

表 2　扩展场景下不同算法的无人艇自主导航性能统计

Table 2　Autonomous navigation performance statistics of USV with different algorithms in extended scenarios

| 算法 | 场景 | 成功率/% | 期望有效航行距离/m |
|---|---|---|---|
| Improved-PPO | a | 98±2 | 230 |
| | b | 90±3 | 215 |
| | c | 82±3 | 175 |
| PPO | a | 95±2 | 222 |
| | b | 85±4 | 180 |
| | c | 70±3 | 150 |
| DDPG | a | 86±5 | 200 |
| | b | 72±5 | 160 |
| | c | 55±5 | 140 |
| DQN | — | — | — |

如表 2 所列，本文提出的 Improved-PPO 算法在 3 个扩展子场景下的任务成功率分别达到 98±2%，90±3% 和 82±3%，期望有效航行距离分别达到 230m，215m 和 175m。与 PPO 算法和 DDPG 算法相比，任务成功率最高，期望有效航行距离最远，DQN 算法只能支持离散动作，因此无法支持连续动作的无人艇自主避障任务。实验结果表明，本文提出的算法能够适应不同规模的任务场景。

**结束语**　本文提出了一种改进近端策略优化算法，实现了复杂场景下无人艇自主避障。与基于先验知识驱动方法不同，本文算法实现了一种端到端的自主避障控制策略，无人艇直接从环境感知映射到行为动作，无需依赖任何规则和经验。同时，与传统近端策略优化算法相比，本文算法的优势在于：1)引入循环神经网络，形成了感知表征增强模块，实现了近端策略优化算法的高效特征抽取，降低了模型整体参数量，提高了无人艇在时序决策过程中对环境感知的记忆能力；2)基于奖励重塑机制，为近端策略优化算法引入了一套自主避障奖励函数，提升了无人艇避障策略的优化速度和动作柔顺度。为了验证算法的有效性，本文设计了典型无人艇自主避障虚拟实验场景，实验结果表明，与传统近端策略优化算法相比，本文方法在无人艇自主避障策略收敛速度上优势明显，且在无人艇动作柔顺度上更优于传统近端策略优化算法。

然而，本文算法仅在虚拟场景下进行了测试，真实环境中存在着更多的不稳定与不确定性，包括复杂海况、海浪、海流、潮汐以及多样的障碍物。此外，传感器易受干扰，这给算法的迁移带来极大挑战。因此，在未来工作中，我们需要继续完善本文提出的改进近端策略优化算法，包括在感知端融合更多传感器技术，提升硬件性能，引入先进传感器和优化硬件架构，以及在算法端引入更加有效的避碰和柔顺控制策略，从而提高算法在真实环境中的适应性和可靠性。

## 参 考 文 献


[1] BARRERA C, PADRON I, LUIS F S, et al. Trends and challenges in unmanned surface vehicles(USV): From survey to shipping[J]. TransNav: International Journal on Marine Navigation and Safety of Sea Transportation, 2021, 15(1): 135-142.

[2] YAN R, PANG S, SUN H, et al. Development and missions of unmanned surface vehicle[J]. Journal of Marine Science and Application, 2010, 9: 451-457.

[3] POLVARA R, SHARMA S, WAN J, et al. Obstacle avoidance approaches for autonomous navigation of unmanned surface vehicles[J]. The Journal of Navigation, 2018, 71(1): 241-256.

[4] GUAN W, WANG K. Autonomous collision avoidance of unmanned surface vehicles based on improved A-star and dynamic window approach algorithms[J]. IEEE Intelligent Transportation Systems Magazine, 2023, 15(3): 36-50.

[5] ZHANG T, LI Q, ZHANG C, et al. Current trends in the development of intelligent unmanned autonomous systems[J]. Frontiers of information technology & electronic engineering, 2017, 18: 68-85.

[6] MA Y, WANG Z, YANG H, et al. Artificial intelligence applications in the development of autonomous vehicles: A survey[J]. IEEE/CAA Journal of Automatica Sinica, 2020, 7(2): 315-329.

[7] DONG S, WANG P, ABBAS K. A survey on deep learning and its applications[J]. Computer Science Review, 2021, 40: 100379.

[8] YE D, LIU Z, SUN M, et al. Mastering complex control in moba games with deep reinforcement learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 6672-6679.

[9] LU J, HAN L, WEI Q, et al. Event-triggered deep reinforcement learning using parallel control: A case study in autonomous driving[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(4): 2821-2831.

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

[11] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv: 1707.06347, 2017.

[12] GUAN W, WANG K. Autonomous collision avoidance of unmanned surface vehicles based on improved A-star and dynamic window approach algorithms[J]. IEEE Intelligent Transportation Systems Magazine, 2023, 15(3): 36-50.

[13] BAI X, LI B, XU X, et al. USV path planning algorithm based

on plant growth[J]. Ocean Engineering, 2023, 273: 113965.
[14] YU J, YANG M, ZHAO Z, et al. Path planning of unmanned surface vessel in an unknown environment based on improved D* Lite algorithm[J]. Ocean Engineering, 2022, 266: 112873.
[15] OUYANG Z, WANG H, HUANG Y, et al. Path planning technologies for USV formation based on improved RRT[J]. Chinese Journal of Ship Research, 2020, 15(3): 18-24.
[16] ZHAO Y, MA Y, HU S. 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.
[17] WU X, CHEN H, CHEN C, et al. The autonomous navigation and obstacle avoidance for USVs with ANOA deep reinforcement learning method [J]. Knowledge-Based Systems, 2020, 196: 105201.
[18] XU X, LU Y, LIU X, et al. Intelligent collision avoidance algorithms for USVs via deep reinforcement learning under COLREGs[J]. Ocean Engineering, 2020, 217: 107704.
[19] GAN W, QU X, SONG D, et al. Multi-usv cooperative chasing strategy based on obstacles assistance and deep reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2023, 21(4): : 5895-5910.
[20] WANG W, LUO X, LI Y, et al. Unmanned surface vessel obstacle avoidance with prior knowledge - based reward shaping[J]. Concurrency and Computation: Practice and Experience, 2021, 33(9): e6110.
[21] RAMACHANDRAN P, ZOPH B, LE Q V. Searching for activation functions[J]. arXiv: 1710.05941, 2017.
[22] PHANICHRAKSAPHONG V, TSAI W H. An Empirical Generation Technique on Background Music Using Gated Recurrent Neural Networks[C]// 2023 International Conference on Consumer Electronics-Taiwan. IEEE, 2023: 691-692.
[23] NG A Y, HARADA D, RUSSELL S. Policy invariance under reward transformations: Theory and application to reward shaping [C]// Proceedings of the Sixteenth International Conference on Machine Learning. 1999: 278-287.
[24] ALMÓN-MANZANO L, PASTOR-VARGAS R, TRONCOSO J M C. Deep reinforcement learning in agents' training: Unity ML-agents[C]// International Work-Conference on the Interplay Between Natural and Artificial Computation. Cham: Springer International Publishing, 2022: 391-400.
[25] LILLICRAP T P. Continuous control with deep reinforcement learning[J]. arXiv: 1509.02971, 2015.

**KONG Chao**, born in 1986, Ph.D, professor. His main research interests include massive data mining, smart education, and recommender systems.

**MENG Dan**, born in 1990, Ph.D, senior research expert. Her main research interests include multimodal machine learning, trustworthy AI, federated learning, and cloud-edge-IoT.

(责任编辑:何杨)