基于强化学习的无人机海上搜寻路径规划

张楠[1],刘虎[1],田永亮[1],蒋佳炜[2],沈贝宁[1],路巽[1],汪志瑶[1]
(1.北京航空航天大学 航空科学与工程学院,北京 100083)
(2.中国特种飞行器研究所 飞行器高速水气耦合动力学学科与技术中心,荆门 448000)

摘　要: 海上应急救援是保障海上活动安全的重要组成部分,也是完善当前救援体系的重要环节。相较传统以有人直升机和船舶为主的搜救方式,无人机具有部署灵活、成本低、响应速度快等优势,可作为海上救援力量的重要补充。然而,受海洋动态环境影响,遇险目标位置预测存在不确定性,对高效开展海上搜寻任务提出了挑战。为此,提出一种基于强化学习的海上搜寻路径规划方法。首先,构建无人机智能体模型和海上搜寻任务的状态—动作空间,并设计综合考虑搜寻概率与探索激励的奖励函数;其次,基于PPO强化学习算法搭建算法架构,通过智能体与环境交互实现策略训练;最后,通过典型想定案例对算法进行仿真验证,并对关键参数进行优化,同时与其他路径规划方法进行对比。结果表明:所提方法能够在搜寻初期优先覆盖高概率目标区域,提高整体的搜寻效率,从而在目标位置不确定的情况下获得更优的搜寻路径规划结果。

关键词: 海上搜寻;无人机;动态规划;路径规划;强化学习
中图分类号: V19; TP39　　**文献标识码:** A

Reinforcement learning-based path planning for UAV maritime search

ZHANG Nan[1], LIU Hu[1], TIAN Yongliang[1], JIANG Jiawei[2], SHEN Beining[1],
LU Xun[1], WANG Zhiyao[1]
(1. School of Aeronautic Science and Engineering, Beihang University, Beijing 100083, China)
(2. Discipline and Technology Center for High Speed Water-Gas Coupling Dynamics of Aircraft, China Special Aircraft Research Institute, Jingmen 448000, China)

Abstract: Maritime emergency rescue is an important component of ensuring the safety of maritime activities and represents a key aspect of improving the current rescue system. Compared with traditional search and rescue operations mainly relying on manned helicopters and ships, unmanned aerial vehicles (UAVs) offer advantages such as flexible deployment, low cost, and rapid response, and can serve as an important supplement to maritime rescue forces. However, due to the dynamic marine environment, the predicted location of distress targets is subject to uncertainty, which poses challenges for conducting efficient maritime search operations. To address this issue, a reinforcement learning-based path planning method for maritime search is proposed. First, a UAV agent model and the state-action space of the maritime search task are constructed, and a reward function that comprehensively considers search probability and exploration incentives is designed. Second, an algorithmic framework based on the Proximal Policy Optimization (PPO) reinforcement learning algorithm is established, and the policy is trained through interactions between the agent and the environment. Finally, a typical scenario is employed to conduct simulation-based verification of the proposed algorithm, optimize key parameters, and perform comparisons with other path planning methods. The results demonstrate that the proposed method can prioritize coverage of high-probability target areas in the early stage of the search, thereby improving overall search efficiency and achieving superior path planning performance under conditions of uncertain target locations.

Key words: maritime search; unmanned aerial vehicle (UAV); dynamic programming; path planning; reinforcement Learning

收稿日期: 2025-10-28;　**修回日期:** 2026-04-14
基金项目: 复杂海况起降动力学建模及人机交互的虚拟仿真平台构建技术(23100002022105002)
通信作者: 田永亮(1985—),男,博士,副教授。E-mail: tianyl@buaa.edu.cn
引用格式: 张楠,刘虎,田永亮,等. 基于强化学习的无人机海上搜寻路径规划[J]. 航空工程进展.
ZHANG Nan, LIU Hu, TIAN Yongliang, et al. Reinforcement learning-based path planning for UAV maritime search[J]. Advances in Aeronautical Science and Engineering. (in Chinese)

0 引 言

我国拥有广袤的领海，根据《联合国海洋法公约》有关规定和我国的主张，我国管辖的海域面积约300万平方千米[1]。随着经济的不断发展，海上事业的安全是发展海洋强国的前提条件，而海上应急救援工作是海上安全的重要保障[2]。在目前海上应急救援体系的发展情况下，海上应急搜救任务的执行主要依靠救援直升机和救援船舶[3]。对于救援船舶而言，其航行速度约为12 kt(22 km/h)，主要不足之处在于救援的时效性和安全性，尤其是针对中远海搜救任务，船舶很难较快到达现场。因此需要航空力量在海上搜救中发挥重要作用，目前海上搜救主要使用的航空器为搜救直升机，但其也有一定的局限性，例如采用目视搜索的精度较低，搜寻覆盖区域受限，多机协同的难度较高。近年来无人机的快速发展展现了其在海上搜寻任务中有较大的应用潜力，能够填补直升机与船舶在海上搜救中的不足。无人机具有快速响应和广泛覆盖能力，能够在短时间启动响应，前往目标区域搜索；搭载多种传感器和相关设备，提供更多的数据和图像，能够更加快速、准确地定位搜索目标；所需的能源少，且多无人机可集成操作控制，具有一定的自主飞行能力，所需操作人员少，成本更低[4-7]。

在无人机搜寻路径规划研究方面，目前已经有较多的应用研究。例如，Ma Y等[8]针对恶劣天气的情况下采用改进的遗传算法进行海上多机救援任务规划，给出优先处理紧急任务的方案；卓星宇[9]针对山区搜寻问题，采用模拟退火算法进行无人机山区搜寻路径规划，给出了不同山区的搜寻路径方案；孙艺松等[10]针对海上漂移目标搜索问题改进蚁群算法，提高算法效率；许海涛等[11]针对传统人工势场法存在的目标点不可达及易陷入局部最小值的问题，对斥力函数进行了改进，有效提升了算法在路径规划中的可靠性。对于无人机搜寻规划，目前大多数的研究场景为陆地场景，例如城市[12]、山区等，涉及三维避障等问题；且启发式算法在路径规划问题中容易陷入局部最优，如何获得全局最优是较难解决的问题。

对于无人机海上搜寻问题，与陆地有所区别，海上没有地形坡度影响，可以简单抽象为二维平面搜索，无需考虑三维避障问题，只需考虑飞行器之间冲突问题以及区域恶劣环境。海上搜寻最大的特点是目标的不确定性，即漂移问题[13]。即便是有精确的初始坐标，随着时间的推移目标位置也会发生变化。目前已有相应的预测方法[14]，但是算法具有一定的误差，预测的精度会随着时间降低，因此搜寻的效率至关重要，如何提高无人机海上搜寻效率是一个需要解决的问题。

海上搜寻是一个动态环境问题，对于这类问题强化学习方法有着较好的应用价值[15]。Wu C X等[16]以设计自主寻找目标路径的无人机为目的，采用DQN进行训练，实现无人机自主搜救和调度；杨清清等[17]面对海战场搜救问题，采用Rainbow方法进行训练，给出海战场搜救方案；邹良骥[18]采用QMIX方法结合AC算法进行多智能体搜寻训练，给出多无人机的搜寻方案。但随着算法的发展，DQN方法相对低效，容易丢失相关信息。Rainbow方法虽是吸取不同方法优势的算法，但其较为复杂，收敛难度较大，稳定性较差，对于不同案例常常需要精细调整超参数。PPO方法结合了策略梯度的高效性和信任域方法的稳定性，通过限制更新幅度避免策略崩溃，实现较为简单并且更加稳定，且对于离散动作空间和连续动作空间均适用，易于拓展到多智能体MAPPO方法，适合用于无人机海上搜寻问题。

本文面向海上搜寻问题，结合无人机应用可行性及现实辅助决策方案需求，研究无人机海上搜寻路径规划，建立环境模型及智能体单元模型，通过强化学习方法进行智能体搜寻路径规划，为无人机海上搜寻提供新的算法方案，提高海上搜寻效率。

1 无人机搜寻的强化学习模型

1.1 智能体模型

无人机可以分为固定翼与多旋翼，其中多旋翼无人机操作灵活但航程较短，固定翼无人机机动性相对受限但航程较长，且载荷能力更强。考虑到海上搜寻任务具有范围广、时间长等特点，本文选择固定翼无人机作为研究对象，并构建其智能体模型。

在二维路径规划问题中，为降低问题复杂度，并使强化学习模型更加专注于搜索路径策略优化，本文在构建无人机智能体模型时对其飞行动

力学进行适当简化。考虑到在海上搜寻任务中,无人机通常以稳定巡航状态进行大范围目标搜索,且相邻搜索航段之间的方向变化较小,因此可作如下假设:

1) 无人机在任务执行过程中以恒定巡航速度飞行,并忽略最小转弯半径约束;

2) 无人机续航能力能够覆盖任务区域,不考虑燃油或电量约束;

3) 无人机在执行搜寻任务过程中保持飞行高度和扫视宽度不变。

在上述假设条件下,将无人机抽象为质点智能体模型,不考虑复杂的飞行动力学与控制约束,仅保留与搜寻任务密切相关的关键参数,如无人机位置及扫视宽度等。根据不同搜索目标类型及飞行高度,无人机对应的扫视宽度参数如表1所示。

表1 固定翼航空器扫视宽度[19]

Table 1 Sweep width of fixed-wing aircraft[19]

搜索目标	高度/m		
	150 m	300 m	600 m
水中人员	370	370	—
4人救生筏	7 600	8 000	8 000
船只<5 m	6 100	6 800	7 600
船只12 m	35 700	35 700	39 800

在实际搜寻任务中,无人机的有效扫视宽度还需根据环境条件(如风速、浪高、能见度等)进行修正(如表2~表3所示)。当天气条件良好时,可将无人机扫视宽度设定为300 m。

表2 天气修正系数

Table 2 Weather correction coefficient

搜索目标		目标	
风速(km/h)	浪高(m)	水中人员	救生筏
0—28	0—1	1.0	1.0
28—46	1—1.5	0.5	0.9
>46	>1.5	0.25	0.6

表3 能见度修正系数

Table 3 Visibility correction coefficient

能见度/km	系数	能见度/km	系数
6	0.4	28	0.9
9	0.6	>37	1.0
19	0.8		

1.2 状态空间

依据海上漂移预测数据(如国家海上搜救环境保障服务平台提供的漂移数据),对目标漂移的时空预测信息进行筛选,并提取所需的漂移预测散点。在此基础上构建POC(Probability of Containment)矩阵,依据无人机的扫视宽度等参数对目标海域进行栅格化处理,划分为多个子区域,得到所需的二维栅格地图[20]。将任务区域划分为个$M\times N$栅格,每个栅格的中心点坐标即为该网格的位置坐标,记为(m,n),其中m、n的取值范围如式(1):

$$\begin{cases}1\leqslant m\leqslant M\\ 1\leqslant n\leqslant N\end{cases} \tag{1}$$

同时,基于栅格内的漂移预测散点,统计栅格内的散点数量计算每个栅格的搜寻概率,并进行归一化处理,得到每个栅格(m,n)对应一个初始的POC值,即先验概率P_{mn}。

在与智能体交互过程中,需要对各栅格当前状态进行实时更新。为避免重复搜索同一栅格区域或多智能体同时处于同一栅格,使用一个表格实时记录并维护每个栅格的搜寻状态值,用$visited_{mn}(t)$表示。$visited_{mn}(t)=0$为在当前时间下栅格(m,n)未被搜寻过,$visited_{mn}(t)=-1$为在当前时间下栅格(m,n)已被搜寻过,$visited_{mn}(t)=1$为在当前时间下栅格(m,n)正在被搜寻。

对状态空间进行可视化处理,依据栅格概率大小进行颜色映射处理,并且将已被搜寻过的栅格进行遮罩处理,同时将上述两种状态进行融合得到状态空间,如图1所示。

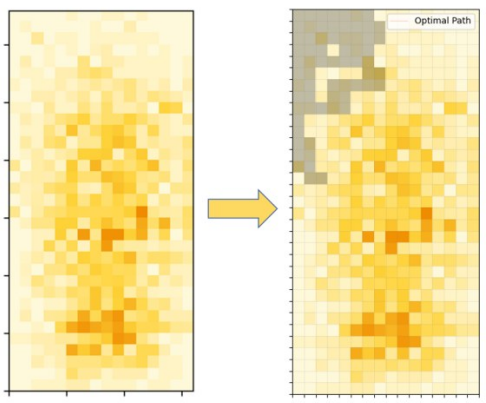
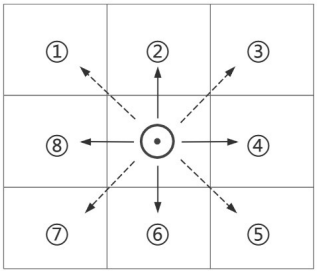


图1 可视化状态空间

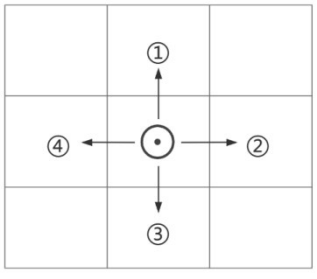
Fig. 1 Visualization of the state space

1.3 动作空间

动作空间的定义直接影响无人机路径规划结果。在栅格化地图中,每个栅格周围存在8个相邻栅格,在每个决策时间点,无人机可以选择任一相邻栅格作为下一位置,如图2所示。然而,根据《国际航空和海上搜寻救助手册》(IAMSAR Manual)[21]的搜寻理论,在实际搜寻任务中需要保证对目标区域的连续覆盖,而当无人机沿对角方向移动时,无法满足搜寻完整栅格的要求。此外,栅格尺度的选取与扫视宽度相匹配,若采用上、下、左、右四个方向移动,可以保证无人机在相邻栅格之间形成连续覆盖。因此,在动作空间设计中仅保留上、下、左、右四个动作选择,如图2(b)所示。为了方便进行模型建立,t时刻无人机智能体的动作决策变量可表示为$a_t=\{1,2,3,4\}$,分别表示上、右、下、左[22]。



(a) 相邻栅格选择



(b) 动作选择

图2 动作空间示意图

Fig. 2 Action space illustration

1.4 奖励函数设计

在强化学习中,通过奖励函数对无人机智能体动作进行定量评价,引导学习最优海上搜寻策略。算法目标为最大化累计奖励,从而获得最优搜寻路径[23]。奖励函数的设计对最终结果有非常重要的影响,因此设计合理的奖励函数对强化学习效果具有重要意义。

本文依据海上搜寻的实际情况,基于Koopman等[24-26]的最优搜寻理论的研究成果,建立奖励函数,并考虑到稀疏奖励问题,建立相关激励机制。因此,将奖励分为即时奖励与回合奖励,同时建立探索激励机制。

即时奖励为每一步智能体探索栅格得到的奖励,记录为$r_t=\{r_1,r_2,r_3,\cdots,r_L\}$,表示无人机在每一个回合内得到的即时奖励集合,每个回合的步长为L步,得到L个即时奖励。为了鼓励无人机探索高概率区域,并避免重复探索相同区域,设置栅格先验奖励与重复探索惩罚,依据栅格概率大小进行加权奖励。依据海上搜寻现实情况,随着时间的增加搜寻的可能性会降低,因此对概率采取衰减机制,栅格内的概率将会以比例进行衰减。其中每个栅格内的即时奖励如下式(2)所示:

$$r_{mn}(t)=\begin{cases}P_{mn}\cdot w\cdot g^t, visited_{mn}(t)=0\\ r_{punish}, visited_{mn}(t)=-1\end{cases},\ 0\leqslant t\leqslant L \tag{2}$$

式中:$r_{mn}(t)$,为当前时间步长t栅格(m,n)的即时奖励;P_{mn}为栅格(m,n)的先验概率;w为当前栅格的奖励权重,用于调节概率收益在奖励函数中的贡献程度,对高概率栅格赋予更大的w值;g为衰减比例,取值范围为[0,1],当$g=1$时不进行衰减;r_{punish}为重复搜寻的惩罚;$visited(t)$为当前栅格的状态,表示是否已经被探索;L为回合步长。

其中,奖励权重w、衰减比例g及重复搜索惩罚项r_{punish}的取值通过多组仿真实验进行调节,在保证训练稳定性的前提下选择能够获得较优搜索效率的参数组合。

由当前的栅格状态计算得到即时奖励$r_t=r_{mn}(t)$后,为避免奖励数值尺度差异对强化学习训练稳定性产生影响,对即时奖励进行归一化处理:

$$\tilde{r}_t=\frac{r_t-r_{\min}}{r_{\max}-r_{\min}} \tag{3}$$

其中$r_{\min}$和$r_{\max}$分别为奖励函数在训练过程中的最小值与最大值,$\tilde{r}_t$为归一化的结果(后文为书写方便,将归一化结果任写作r_t)。

在得到即时奖励之后,对整个回合中的动作进行重新奖励分配,利用折扣因子,通过反向时序差分方法计算回合的累积折扣奖励,如下式(4)所示:

$$R_t = r_t + \gamma \cdot r_{t+1}(1-d_t), \gamma \in [0,1] \quad (4)$$

其中γ为折扣因子,用于计算未来奖励的当前价值,若$\gamma \to 0$,表示采取短视策略,即优先最大化即时奖励,忽略长期收益;若$\gamma \to 1$,表示采取长期策略,平衡即时和未来奖励,鼓励长期最优;若$\gamma=1$,则为未来权重与即时奖励权重相同;在PPO算法中,通常取值[0.9,0.99]。d_t为当前时间步的终止标志,取值为0(未终止)或1(终止)。

为鼓励无人机探索更多区域,在回合结束时增加路径效率奖励,记录为r_e,仅在回合结束时计算。路径效率奖励依据已被探索栅格个数与总步长比例计算,如下式(5)所示:

$$\begin{cases} r_e = \dfrac{E}{L} \cdot B \\ \dfrac{E}{L} \geqslant 0.6 \end{cases} \quad (5)$$

其中E为回合探索栅格数量,L为回合总步长,B为基础奖励值。当且仅当探索比例超过60%的阈值才给予路径效率奖励,以鼓励形成更高效的搜索路径。该阈值参考海上搜救任务中常用的区域覆盖评价标准设置。

综合上述设计,每回合智能体得到的总奖励为:

$$G_t = \sum_{0}^{L} R_t + r_e \quad (6)$$

2 基于PPO的海上无人机搜寻算法

PPO是由Open AI团队于2017年提出的近端策略优化强化学习算法,属于策略梯度(Policy Gradient)方法的改进版本。相对于传统策略优化方法(如TRPO),PPO具有信任区域策略优化的一些好处,复杂性更低,实现较为简单并且更通用,同时保持训练稳定性和样本效率[27]。

2.1 状态向量

根据搜寻算法的基本环境参数及智能体与环境交互需求,当前状态包含栅格化区域(概率矩阵)、栅格被搜寻情况、智能体状态,为了便于神经网络输入输出,需要将状态信息通过独热编码等方法转化为张量形式,如表4所示。

表4 状态向量

Table 4 State vector

要素	数据类型	长度
栅格化区域	数值	$M \times N$
栅格被搜寻情况	布尔值	$M \times N$
智能体状态	数值	6

其中栅格化区域向量为每个栅格先验概率P_{mn}数值的拼接;栅格被搜寻情况向量为每个栅格搜寻情况的拼接,布尔值为TRUE或FALSE;智能体状态为智能体位置、当前动作、当前栅格概率、当前栅格是否被搜索的拼接,具体向量可表示为下式(7)所示:

$$s_t = (x, y, a_{t-1}, P_{mn}, visited_{mn}(t)) \quad (7)$$

其中,x、y为智能体当前位置坐标,a_{t-1}为上一时刻动作。

2.2 初始化阶段

初始化阶段包括网络初始化与超参数设置,首先构建AC网络架构,定义隐藏层大小。Actor网络输出的动作概率,用于决策;Critic网络输出当前状态的评估值,用于策略优化。Actor网络包括全连接层、激活函数、输出动作维度及概率归一化等。输入为状态向量s_t,即所需的观测值,输出为动作概率分布$\pi_\theta(a_t|s_t)$,其中θ为Action网络的参数,π_θ表示参数为θ的策略函数。Critic网络包括全连接层、激活函数、状态价值输出等。输入为状态向量s_t,输出为价值估计$V_\varphi(s_t)$,其中φ为Critic网络的参数。神经网络为Pytorch工作流,采用ReLU激活函数,使用Softmax函数将输出转换为概率分布,通过forward函数接收输入数据并执行前向传播和返回结果。

其次设定超参数,包括折扣因子、PPO-Clip范围、学习率、每轮数据复用次数、总回合数、每回合步长等参数。

2.3 数据采样阶段

在数据采样阶段,策略网络输出动作概率分布并从中采样动作,通过输出策略的随机性实现智能体的探索。智能体依据策略网络采样动作与环境进行交互,计算并储存相关数据,更新智能体状态。动作采样方法如下式(8)所示:

$$a_t \sim \pi_\theta(a_t|s_t), \pi_\theta(a_t|s_t) = \mathrm{Softmax}(f_\theta(s_t)) \quad (8)$$

其中，$f_\theta(s_t)$是策略网络的原始输出，Softmax函数确保概率归一化，即所有动作概率和为1。

同时计算对数概率，用于后续策略梯度的计算：

$$\log \pi_\theta(a_t|s_t)=\log \frac{e^{f_{a_t}(s_t)}}{\sum_a e^{f_a(s_t)}} \tag{9}$$

2.4 优势估计阶段

在优势估计阶段，需要计算每个时间步的优势值A_t和回报R_t。

对于时间优势函数A_t，单步优势函数即TD Error(时序差分)方法特点为低方差、高偏差，容易忽略多步奖励依赖，而纯蒙特卡洛方法特点为无偏差、高方差，会导致策略梯度估计不稳定，因此采用GAE(广义优势估计)方法。GAE通过平衡方差与偏差，能够提升训练的稳定性和效率，是对传统优势估计方法的通用化扩展，能够兼容多种策略梯度算法[28]。GAE的本质是多步TD Error的指数加权平均，计算方法如下式(10)和(11)所示：

$$\delta_t=r_t+\gamma\cdot V_\varphi(s_{t+1})\cdot(1-d_t)-V_\varphi(s_t) \tag{10}$$

$$A_t^{\mathrm{GAE}(\gamma,\lambda)}=\sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k} \tag{11}$$

其中，δ_t为t时刻的TD Error，r_t为t时刻的即时奖励，γ为折扣因子，d_t为终止状态(1表示终止，0表示未终止)；λ为GAE方法中的调节参数，通过调节λ在蒙特卡洛和时序差分之间平滑过渡，当$\lambda=0$时GAE退化为时序差分方法，当$\lambda=1$时GAE退化为蒙特卡洛方法，通常取$\lambda\in[0.9,0.99]$，在PPO中常取0.95。

之后计算回报(Return)，用于值函数优化，计算方法如下式(11)所示：

$$R_t=A_t^{\mathrm{GAE}}+V_\varphi(s_t) \tag{12}$$

2.5 策略优化阶段

策略优化阶段是PPO算法的核心部分，通过策略梯度更新及特有的策略更新范围限制确保训练的稳定性和更新效率，通过多次小批量更新优化策略和值函数，策略优化流程如下：

1) 首先计算新旧策略的重要性采样比率，用于衡量新旧策略在动作选择上的差异，计算方法如下式(13)所示：

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta old}(a_t|s_t)}=\exp(\log\pi_\theta(a_t|s_t)-\log\pi_{\theta old}(a_t|s_t)) \tag{13}$$

2) 通过核心计算方法Clipped替代目标函数计算目标函数损失，限制新旧策略的差异，以防止更新步长过大，计算方法如下式所示：

$$L^{CLIP}(\theta)=\mathrm{E}_t[\min(r_t(\theta)\cdot A_t^{\mathrm{GAE}},\mathrm{clip}(r_t(\theta),1-\varepsilon,1+\varepsilon)\cdot A_t^{\mathrm{GAE}}] \tag{14}$$

其中，ε为Clip的范围，即梯度裁剪范围，在超参数设定时已经完成数值设定，一般取值范围为$\varepsilon\in[0.1,0.2]$，强制策略变化范围在$[1-\varepsilon,1+\varepsilon]$区间内。

3) 为了优化Critic网络，需要计算值函数损失，使得Critic网络更加准确地预测状态的期望累积回报，PPO算法中值函数损失通常采用均方误差(MSE)或Huber损失，本文采用MSE方法，计算方法如下式(15)所示：

$$L^{VF}(\varphi)=\mathrm{E}_t[(V_\varphi(s_t)-R_t)^2] \tag{15}$$

4) 为了鼓励策略探索，在PPO中加入熵项(Entropy Bonus)，提升算法的性能和鲁棒性。熵项通过最大化策略熵$H(\pi)$使动作分布更加均匀，能够避免策略过早固定。并且熵项能够防止策略崩溃，尤其是稀疏奖励任务，熵正则化强制策略保留一定的动作随机性，避免策略退化到仅选择单一动作。对于离散动作，熵项计算方法如下式(16)所示：

$$H(\pi_\theta)=-\sum\pi_\theta(a|s)\log\pi_\theta(a|s) \tag{16}$$

5) 完成Clipped策略损失(Policy Loss)、值函数损失(Value Loss)及策略熵(Entropy Bonus)的计算之后，计算总损失函数，如下式(17)所示：

$$L^{TOTAL}=L^{CLIP}(\theta)+c_1L^{VF}(\varphi)-c_2H(\pi_\theta) \tag{17}$$

其中，c_1为值函数权重，能够平衡值函数损失的贡献，通常取0.5～1.0，在高维状态空间时，通常适当减小，避免值函数主导训练；c_2为熵系数，控制探索强度，通常取0.01～0.05，熵系数越大，探索强度越大，在稀疏奖励任务中通常取较大值。

6) 梯度更新方法是强化学习算法高效稳定训练的核心，通过backward反向传播，自动进行梯度计算，计算方法如下式(18)所示：

$$\begin{cases}\nabla_\theta L^{CLIP}=\mathrm{E}_t[\nabla_\theta \min] \\ \nabla_\theta L^{VF}=\mathrm{E}_t[(V_\theta(s_t)-R_t)\nabla_\theta V_\theta(s_t)] \\ \nabla_\theta H(\pi_\theta)=-\mathrm{E}_t[\nabla_\theta \sum \pi_\theta(a|s)\log \pi_\theta(a|s)]\end{cases} \tag{18}$$

为了防止梯度爆炸,选择梯度裁剪,提升训练的稳定性,裁剪方法如式(19)所示:

$$\mathrm{grad} \leftarrow \mathrm{clip}(\mathrm{grad}, -\max\_\mathrm{norm}, \max\_\mathrm{norm}) \tag{19}$$

PPO通过多步更新提高数据利用率,将数据划分为小批量(mini-batch),多次更新,常采用的范围为3~10次/轮,本文设定4次每轮,即update_epochs=4。同时使用Adam优化器实现自适应学习率,平衡训练的收敛速度和稳定性。Adam优化器参数更新方法如下式(20)所示:

$$\theta \leftarrow \theta - \alpha \nabla_\theta L^{TOTAL} \tag{20}$$

式中:α为初始学习率,可根据实际情况调整。

2.6 主训练循环阶段

主训练循环整合上述阶段,实现完整的训练流程。设定训练总回合数,每个回合初始化环境并清空轨迹缓存列表,轨迹列表数据包括状态序列、动作序列、动作对数概率、即时奖励、终止标志和Critic网络预测的状态价值。在每个回合内,智能体通过采样动作与环境交互得到轨迹数据,并储存在轨迹列表中。回合结束之后,对终止状态补充价值估计,并采用GAE方法计算每个时间步的折扣回报,执行PPO算法的梯度更新。完成训练之后,对回合数据进行日志记录及可视化。

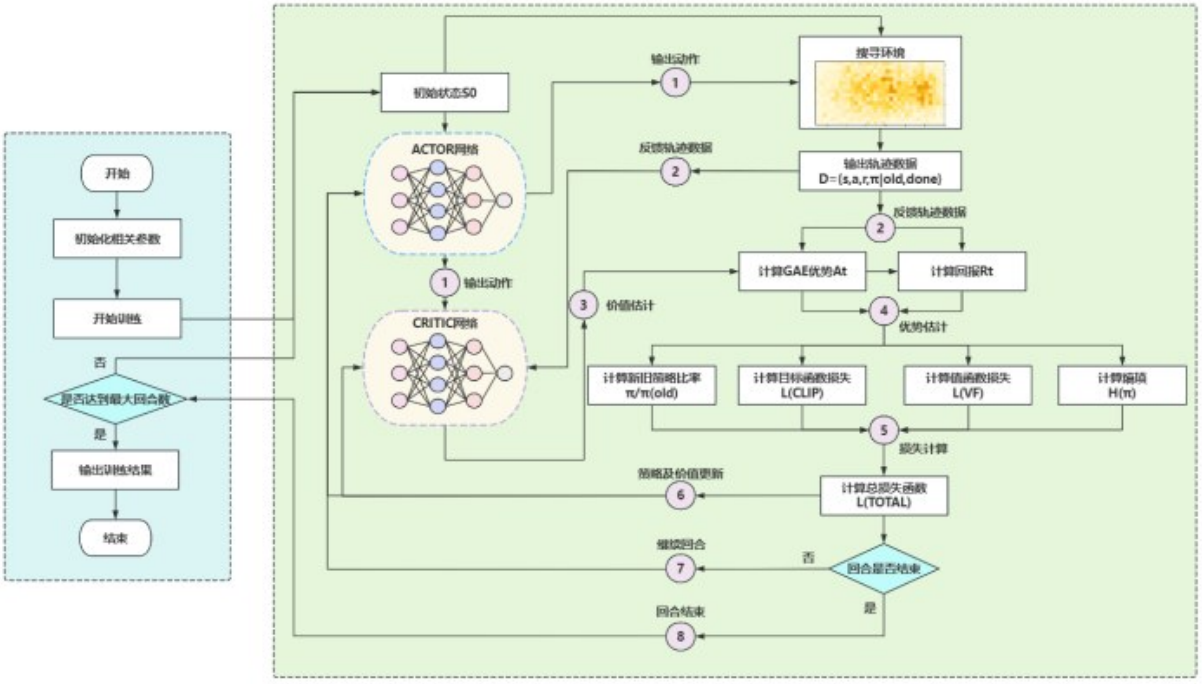


图3 PPO算法训练流程图

Fig. 3 Training flowchart of the PPO algorithm

3 仿真实验

3.1 仿真案例背景及初始化

本文以一次假设的海上事故为例,进行算法仿真实验验证。假定2025年1月8日上午9点,在东海海域(25.80 °N,119.79 °E)某渔船发生倾覆事故,事发地点天气良好,拟采用无人机快速响应并首先进行搜寻工作,与救援直升机、救援船舶协同完成任务。

数据来源采用国家海上搜救环境保障服务平台提供的预测数据,并对漂移预测散点进行筛选处理,计算得到最优搜寻区域,将区域进行二维栅格化处理,得到归一化之后的概率矩阵,对概率矩阵进行可视化,得到概率矩阵的二维热力图,如图4所示。

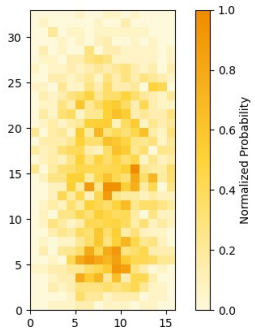


图4 概率矩阵热力图

Fig. 4 Heatmap of the probability matrix

3.2 参数设置

超参数的设置会影响强化学习算法的学习效果，为了取得更为高效的参数组合，本文通过研究不同参数大小对算法的影响，来优化参数选取。

折扣因子影响算法对未来奖励的重要程度，折扣因子较大代表更加重视未来奖励，较小的折扣因子代表更加重视即时奖励。取不同大小的折扣因子进行对比实验，为了方便展示，截取前5 000回合的算法结果，如图5所示，当折扣因子过小时会使得算法发生震荡，过大的折扣因子会使训练时间延长，从图中可以看出取0.995时效果最好。

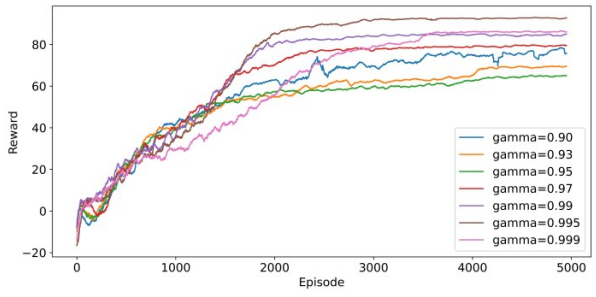


图5 不同折扣因子的算法性能对比

Fig. 5 Comparison of algorithm performance under different discount factors

学习率影响模型更新频率，是强化学习中最重要的超参数之一。学习率过大会使模型过早收敛，可能错过最优策略；过小则会使训练时间过长。取不同学习率进行对比实验，设置回合数为20 000，算法结果如图6所示，可以看出：当学习率取3×10^{-5}时，在20 000个回合内算法仍未收敛；而当学习率取8×10^{-5}时，算法收敛较快但最终的策略并没有最优。综合考虑收敛速度与策略择优，学习率取值为6×10^{-5}时效果最好。

Clip范围是PPO算法的核心机制之一，其通过限制更新的幅度确保训练过程的稳定性。Clip控制了策略更新的信任域范围，平衡稳定性和学习速度。更大的Clip范围会允许更大的策略更新范围，可能会有更快的学习速度，但也可能会让稳定性降低，出现震荡情况；更小的Clip范围会限制更小的策略更小范围，采取更保守的更新，有更稳定的训练效果，但是收敛速度较慢，并可能陷入局部最优。取不同的Clip范围进行对比实验，如图7所示，为平衡收敛速度及更优的结果，Clip范围取0.2。

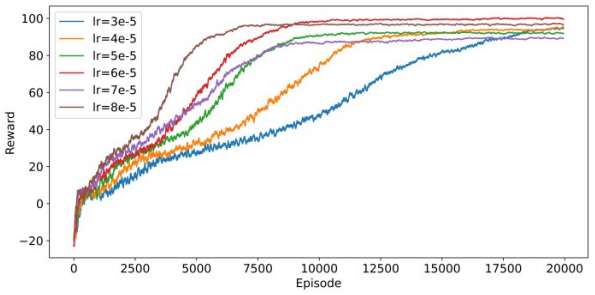


图6 不同学习率的算法性能对比

Fig. 6 Comparison of algorithm performance under different learning rates

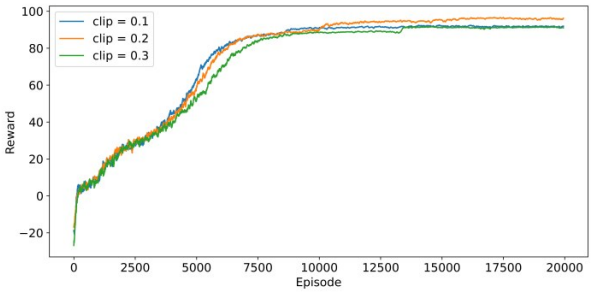


图7 不同Clip范围的算法性能对比

Fig. 7 Comparison of algorithm performance under different clip ranges

基于上述对比实验及环境动作空间，PPO算法采用的各项参数如表5所示。

表5 超参数设定

Table 5 Hyperparameter setting

超参数	参数值
折扣因子	0.995
Clip范围	0.2
学习率	6×10^{-5}
每轮数据复用次数	4
总回合数	20 000
每回合步长	120

3.3 实验结果与对比分析

依据上一节所得到的参数设定,训练得到最终的路径规划结果。奖励与损失曲线是判定算法是否收敛的重要依据,根据图8可以看出,在一万回合之前训练效果稳步提高,并且提升效率也较高,在一万步之后效果提升较慢并趋近于收敛,同时损失函数也在不断下降并趋近于0,证明算法收敛效果。

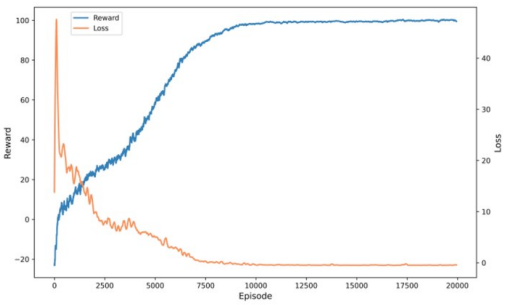


图8 奖励与损失曲线

Fig. 8 Reward and loss curves

为了验证算法的有效性,以算法给出的路径方案的搜寻效率作为评价,同时,以现在常用的平行线搜寻方法及遗传算法作为对比,计算方法如下式(21)所示。

$$E_f=\frac{\sum_{i=1}^{L}p_i}{L} \tag{21}$$

其中,L代表当前总步长,即代表搜寻时间,p_i代表每一个栅格的先验概率。搜寻效率反映了路径方案选择高概率区域的能力,当搜寻效率较高时,证明算法能够更好地覆盖优先区域。对于同一个案例,计算不同方法内不同步长的搜寻效率,如表6所示,同时绘制相应的曲线,如图9所示。

表6 不同方法搜索效率对比

Table 6 Comparison of search efficiency among different methods

方法	步长		
	20	40	60
PPO	0.467 9	0.492 9	0.507 5
GA	0.354 6	0.449 3	0.408 6
平行线	0.049 2	0.087 8	0.156 8
方法	步长		
	80	100	120
PPO	0.498 8	0.483 8	0.451 2
GA	0.401 5	0.391 9	0.394 1
平行线	0.175 6	0.211 2	0.238 3

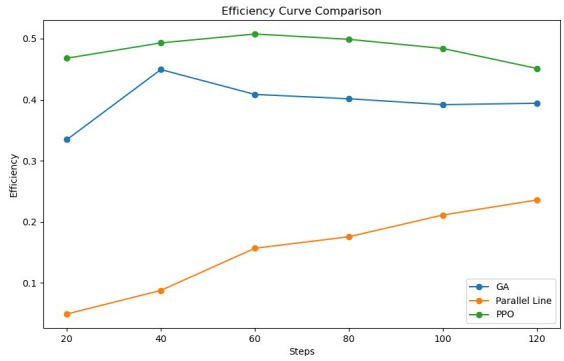


图9 不同方法搜索效率变化曲线

Fig. 9 Variation of search efficiency for different methods

图中绿色折线为基于本文算法的搜寻方案结果,蓝色折线为基于遗传算法的搜寻方案结果,橙色折线为基于IAMSAR标准规定的平行线搜寻方法结果。从图表中可以看出:本文算法的搜寻效率随着步数的增加先升高后降低,说明算法判断高概率区域的能力较强,能够在搜寻的早期优先覆盖关键区域,拥有较强的时效性。而目前实践中常用的平行线搜寻方法搜寻效率逐渐提高,说明在搜寻的早期区域并不是特别重要,随着搜寻的进展,才逐渐搜寻到较为关键的区域。

同时计算不同方法的累计概率,如图10所示。可以看出:PPO算法在搜寻效率上整体优于遗传算法与实际中常用的平行线搜索算法,能够得到更高的累计概率。说明算法给出的方案能够覆盖更多的关键区域,有更好的高概率区域判断能力。

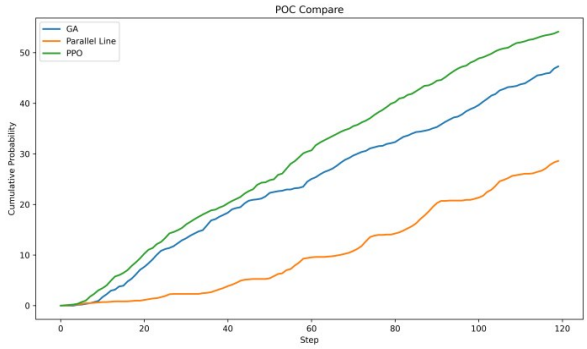
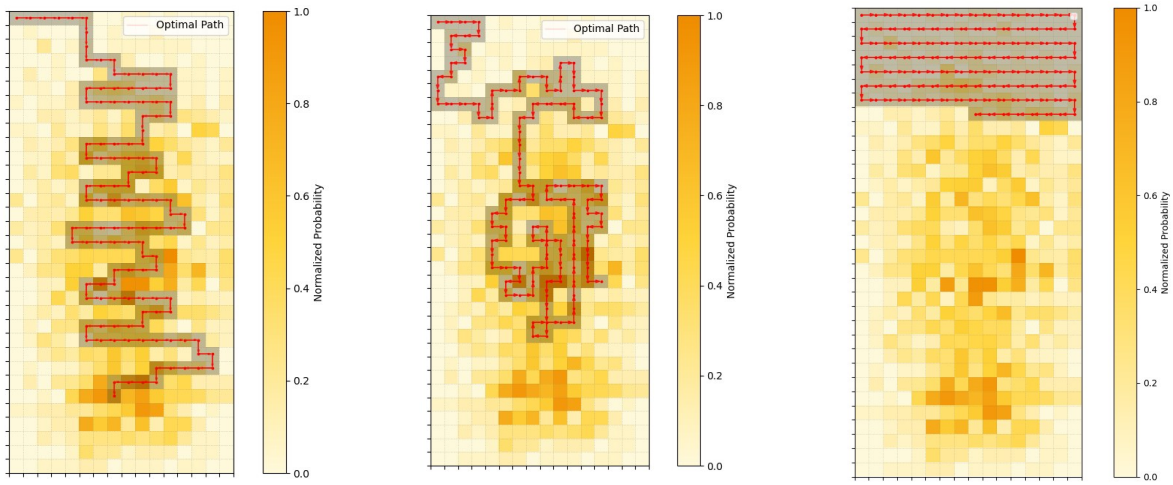


图10 不同搜寻方法的概率累计对比

Fig. 10 Comparison of cumulative probability for different search methods

不同方法得出的搜寻路径规划结果如图11所示。从路径图中可以更直观地验证上述结论,即本文算法得到的路径方案能够高效地覆盖大部分

高概率区域，更加高效，而遗传算法得到的路径相对混乱，并且会重复进行搜索，难以覆盖更关键的区域。基于强化学习的算法能够基于动态的地图进行策略更新，更好地给出路径方案。



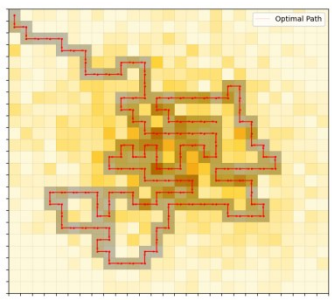
（a）基于PPO强化学习算法　（b）基于遗传算法　（c）平行线搜寻方法

图11　不同方法的搜寻路径规划结果对比

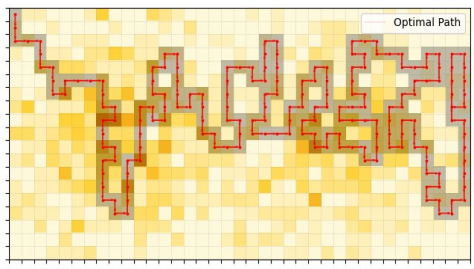
Fig. 11　Comparison of search path planning Results for different methods

3.4　算法鲁棒性分析

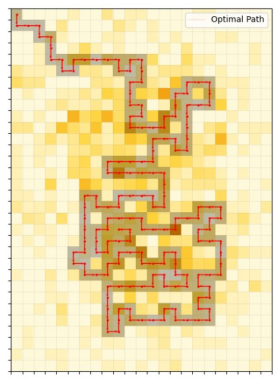
搜寻路径规划算法应具有良好的适应性，能够在不同条件下获得较优结果。为验证算法的鲁棒性，针对不同想定案例（包括不同区域规模及不同概率分布）进行路径规划实验，通过训练智能体获得最终搜寻结果，如图12所示。



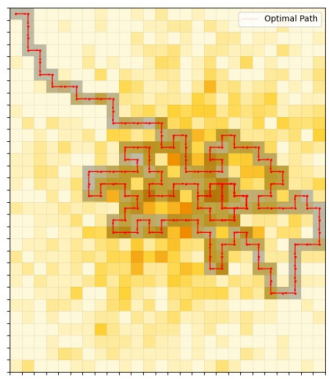
（a）想定案例一



（b）想定案例二



（c）想定案例三



（b）想定案例四

图12　不同概率地图的搜寻路径规划结果对比

Fig. 12　Comparison of search path planning results under different probability maps

为比较不同概率地图下的搜寻效率,需要对结果进行归一化处理,其计算方法如下:

$$E_{f,k}=\frac{\sum_{i=1}^{L}p_i}{\sum_{i=1}^{M\cdot N}p_i\cdot L}\cdot\eta \tag{22}$$

其中,$\sum_{i=1}^{M\cdot N}p_i$是地图所有概率之和,η是常数项,用于调节搜寻效率的数值规模,避免结果过小。

为评估算法稳定性,对仿真结果计算变异系数(CV)和四分位距(IQR),其计算公式分别如下:

$$CV=\frac{\sigma}{\mu}\times 100\%=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}}{\frac{1}{N}\sum_{i=1}^{N}x_i^{\,2}}\times 100\% \tag{23}$$

$$IQR=Q_3-Q_1=(n+1)\times 0.75-(n+1)\times 0.25 \tag{24}$$

其中,CV 用于衡量数据的相对离散程度,当CV<15%时,表明稳定性很好;15%<CV<30%表明稳定性一般;CV>30%则表明稳定性较差。IQR能够反映中间50%数据的分布范围,可直观反映数据的离散程度,相关计算结果如表7所示。

表7 搜寻效率统计结果

Table 7 Search efficiency statistical results

数据量	均值	标准差	变异系数CV	IQR
50	0.333 6	0.041 2	12.35%	0.049 3

从表7可以看出:变异系数CV小于15%,说明算法在不同概率地图条件下表现稳定,能够持续生成较优的搜寻路径方案,具有较好的鲁棒性。

4 结 论

1) 无人机在海上搜救中有较好的应用前景,相对于传统海上搜救手段有一定的优势,能够对海上搜救体系起到补充完善的作用。

2) 依据海上环境及无人机特点进行场景构建,将搜寻环境抽象为二维概率栅格,并将无人机动作简化为四个方向,设计PPO强化学习算法进行路径规划。

3) 在案例中通过智能体与环境交互进行学习训练,验证算法能够在一定回合后收敛,同时对不同参数设置进行结果对比,得到最优的参数组合。

4) 将本文算法与遗传算法和传统的平行线搜寻方法进行结果对比,验证本文算法有较好的搜寻效率,得到更高的概率累计,能够优先搜索高概率的关键区域,得到更优的路径方案。

参 考 文 献

[1] 中华人民共和国国务院. 国家海洋事业发展“十二五”规划[EB/OL]. (2014-09-02)[2025-10-28]. https://www.gov.cn/guoqing/2014-09-02/content_2744175_2.htm.
State Council of the People's Republic of China. The 12th Five-Year Plan for National Marine Career Development [EB/OL]. (2014-09-02)[2025-10-28]. https://www.gov.cn/guoqing/2014-09-02/content_2744175_2.htm. (in Chinese)

[2] 程明远. 建设海洋强国背景下我国海上应急救援工作发展建议[J]. 水运管理, 2021, 43(2): 14-15, 19.
Cheng Mingyuan. Development suggestions of maritime emergency rescue work under background of maritime power construction in China[J]. Shipping Management, 2021, 43(2): 14-15, 19. (in Chinese)

[3] Solberg K E, Jensen J E, Barane E, et al. Time to rescue for different paths to survival following a marine incident[J]. Journal of Marine Science and Engineering, 2020, 8(12): 997.

[4] 沈练高. 无人机在海洋救援中的应用分析[J]. 水上安全, 2023(8): 1-3.
Shen Liangao. Application analysis of UAV in ocean rescue [J]. Maritime Safety, 2023(8): 1-3. (in Chinese)

[5] 王帆. 无人机在海上救援中的应用[J]. 航海技术, 2022(5): 71-73.
Wang Fan. Application of drones in marine rescue operation [J]. Marine Technology, 2022(5): 71-73. (in Chinese)

[6] Lomonaco V, Trotta A, Ziosi M, et al. Intelligent drone swarm for search and rescue operations at sea[PP/OL]. V1. arXiv (2018-11-13)[2025-10-28]. https://doi.org/10.48550/arXiv.1811.05291.

[7] McRae J N, Gay C J, Nielsen B M, et al. Using an unmanned aircraft system (drone) to conduct a complex high altitude search and rescue operation: A case study[J]. Wilderness & Environmental Medicine, 2019, 30(3): 287-290.

[8] Ma Y, Li B, Huang W T, et al. An improved NSGA-II based on multi-task optimization for multi-UAV maritime search and rescue under severe weather[J]. Journal of Marine Science and Engineering, 2023, 11(4): 781.

[9] 卓星宇. 无人机山区搜寻方法研究[D]. 广汉: 中国民用航空飞行学院, 2017.
Zhuo Xingyu. The study on the mountain search method by unmanned aerial vehicles(UAV)[D]. Guanghan: Civil Aviation Flight University of China, 2017. (in Chinese)

[10] 孙艺松, 胡海军, 李乐, 等. 基于改进蚁群算法的海上目标搜索路径规划[J]. 传感器与微系统, 2024, 43(10): 160-164.

Sun Yisong, Hu Haijun, Li Le, et al. Maritime target search path planning based on improved ant colony algorithm [J]. Transducer and Microsystem Technologies, 2024, 43(10): 160-164. (in Chinese)

[11] 许海涛, 陈龙胜, 王宇翔. 改进势场法在无人机编队三维路径规划上的应用研究[J]. 航空工程进展, 2025, 16(4): 100-109.
Xu Haitao, Chen Longsheng, Wang Yuxiang. Application research on improved artificial potential field method in three-dimensional path planning for UAV formation[J]. Advances in Aeronautical Science and Engineering, 2025, 16(4): 100-109. (in Chinese)

[12] Liu Y X, Liu H, Tian Y L, et al. Reinforcement learning based two-level control framework of UAV swarm for cooperative persistent surveillance in an unknown urban area[J]. Aerospace Science and Technology, 2020, 98: 105671.

[13] Tamtare T, Dumont D, Chavanne C. The Stokes drift in ocean surface drift prediction [J]. Journal of Operational Oceanography, 2022, 15(3): 156-168.

[14] Yan S Y, Zhang J, Parvej M M, et al. Sea drift trajectory prediction based on quantum convolutional long short-term memory model[J]. Applied Sciences, 2023, 13(17): 9969.

[15] Arulkumaran K, Deisenroth M P, Brundage M, et al. Deep reinforcement learning: a brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[16] Wu C X, Ju B B, Wu Y, et al. UAV autonomous target search based on deep reinforcement learning in complex disaster scene[J]. IEEE Access, 2019, 7: 117227-117245.

[17] 杨清清, 高盈盈, 郭玮, 等. 基于深度强化学习的海战场目标搜寻路径规划[J]. 系统工程与电子技术, 2022, 44(11): 3486-3495.
Yang Qingqing, Gao Yingying, Guo Yu, et al. Target search path planning for naval battle field based on deep reinforcement learning[J]. Systems Engineering and Electronics, 2022, 44(11): 3486-3495. (in Chinese)

[18] 邹良骥. 基于强化学习的无人机协同区域搜索规划研究[D]. 武汉: 华中科技大学, 2023.
Zou Liangji. Research on UAV area search planning based on reinforcement learning[D]. Wuhan: Huazhong University of Science and Technology, 2023. (in Chinese)

[19] 王磊, 闫斯莹. 航空搜救范围与成功概率研究[J]. 指挥控制与仿真, 2023, 45(4): 52-56.
Wang Lei, Wen Siying. Research on the scope and successful probability of aerial SAR[J]. Command Control & Simulation, 2023, 45(4): 52-56. (in Chinese)

[20] Gallego A J, Pertusa A, Gil P, et al. Detection of bodies in maritime rescue operations using unmanned aerial vehicles with multispectral cameras[J]. Journal of Field Robotics, 2019, 36(4): 782-796.

[21] International Civil Aviation Organization. IAMSAR manual: organization and management[M]. 4th ed. Montreal: International Civil Aviation Organization, 2003.

[22] 疏利生, 李桂芳, 嵇胜. 基于强化学习的航空器机场智能静态路径规划[J]. 航空工程进展, 2021, 12(3): 65-70.
Shu Lisheng, Li Guifang, Ji Sheng. Aircraft AI static path planning on airport ground based on reinforcement learning [J]. Advances in Aeronautical Science and Engineering, 2021, 12(3): 65-70. (in Chinese)

[23] Siboo S, Bhattacharyya A, Naveen Raj R, et al. An empirical study of DDPG and PPO-based reinforcement learning algorithms for autonomous driving[J]. IEEE Access, 2023, 11: 125094-125108.

[24] Koopman B O. The theory of search. I. kinematic bases [J]. Operations Research, 1956, 4(3): 324-346.

[25] Koopman B O. The theory of search. II. target detection [J]. Operations Research, 1956, 4(5): 503-531.

[26] Koopman B O. The theory of search: III. The optimum distribution of searching effort[J]. Operations Research, 1957, 5(5): 613-626.

[27] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[PP/OL]. V2. arXiv (2017-08-28) [2025-10-28]. https://doi.org/10.48550/arXiv.1707.06347.

[28] Schulman J, Moritz P, Levine S, et al. High-dimensional continuous control using generalized advantage estimation [PP/OL]. V6. arXiv (2018-10-20) [2025-10-28]. https://doi.org/10.48550/arXiv.1506.02438.

(编辑:马文静)