信息与控制　2025 年　第 54 卷　第 4 期：545 ~ 555

**DOI**：10.13976/j.cnki.xk.2024.2122　　**CSTR**：32166.14.xk.2024.2122

# 基于深度强化学习的海上搜救覆盖路径规划算法应用

韩靖童[1,2]，余　倩[1,2]，刘　源[2]

1. 上海理工大学健康科学与工程学院，上海　200093；
2. 海军军医大学卫生勤务学系，上海　200433

基金项目：军队后勤科研重大项目（AHJ22C003）

通信作者：刘源，yawnlau@126.com　收稿/录用/修回：2024 - 07 - 06/2024 - 09 - 11/2024 - 12 - 27

**摘要**

目前海上搜救（SAR）辅助决策系统依旧采用传统的固定式搜寻模式，其存在效率低下、适应性弱等问题。为此，提出了一种基于深度强化学习的海上搜救覆盖路径规划模型。首先，将海上搜救覆盖路径规划问题转化为马尔可夫决策过程。然后，结合 DDQN（Double Deep Q-Network）、Prioritized DDQN、Distributional DQN 和 Noisy DQN，设计了适用于单搜救船只的海上搜救覆盖路径规划算法。最后，通过模拟实验验证了所提算法的可行性和有效性。对比实验结果表明，所提算法无论在路径规划质量还是搜寻效率上，均显著优于其他算法。

**关键词**

海上搜救
深度强化学习
覆盖路径规划

**中图法分类号**：TP391.9

**文献标志码**：A

# Application of Deep Reinforcement Learning-based Maritime Search and Rescue Coverage Path Planning Algorithm

HAN Jingtong[1,2], YU Qian[1,2], LIU Yuan[2]

1. *School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*;
2. *Faculty of Military Health Service, Naval Medical University, Shanghai* 200433, *China*

**Abstract**

Given that current maritime search and rescue (SAR) decision support systems still rely on traditional fixed search patterns, which are inefficient and lack adaptability, we propose a maritime SAR coverage path planning model based on deep reinforcement learning. First, we formulate the maritime SAR coverage path planning problem as a Markov decision process. Then, by integrating a double deep Q-network (DDQN), prioritized DDQN, distributional DQN, and noisy DQN, we design a coverage path planning algorithm tailored for a single rescue vessel. Finally, we validate the feasibility and effectiveness of the proposed algorithm through simulation experiments. Comparison results demonstrate that the proposed algorithm substantially outperforms existing methods in path planning quality and search efficiency.

**Keywords**

maritime search and rescue;
deep reinforcement learning;
coverage path planning

## 0　引言

海上搜救具有搜救目标存活时间短、待搜寻区域广、探测概率低、漂流轨迹难以预测等特征[1]，遇险者的数量和位置通常是不确定的，大大增加了搜救难度。区域覆盖路径规划作为一种路径规划的方式，不受落水人员漂泊不定的位置因素影响，确保了搜救区域内的每一部分都搜索到，从而最大程

度地提高搜救目标的发现概率。目前，海上搜救辅助决策系统仍然采用传统的搜寻模式，如平行线、扩展方形、扇形和横移线等方法。其效率较低且难以准确量化实时调整搜寻方案[2]。因此，研究一种可以高效覆盖区域的路径规划算法至关重要。

区域覆盖路径规划算法目前主要分为两类：经典算法和启发式算法[3]。经典算法包括随机游走算法[4]、生成树覆盖算法[5]和人工势场法[6]等，其只适用于简单环境，依赖搜救智能体的初始位置，而且规划的路径存在重复率高、效率低下等问题。启发式算法包括图搜索[7-9]、群智能算法[10-11]和神经网络算法[12]等，其存在收敛速度较慢、容易陷入局部最优解、需要大量数据进行训练、训练过程复杂且时间长、泛化能力不足等问题。强化学习作为一种前沿的机器学习方法，在区域覆盖路径规划中能够有效弥补这些算法的不足[13]。在强化学习中，智能体通过试错学习，逐步优化策略，最大化长期收益。然而在海洋这样未知的环境下，随着动作和状态空间不断膨胀，强化学习算法通过表格存储动作价值的方法难以承载这些数据。为了克服这些问题，深度强化学习应运而生。深度强化学习通过引入神经网络，将状态和动作的空间映射到神经网络的输入和输出空间中，通过神经网络近似值函数或策略函数，从而实现对复杂环境的建模和学习[14]。

深度强化学习理论在实践中的不断发展，使其能够应用于覆盖路径规划任务[15]。因此，本文将构建一个基于深度强化学习的单搜救船只的海上搜救覆盖路径规划模型。本文首先将海上搜救的覆盖路径规划问题转化为马尔可夫决策过程；然后通过设计和训练深度强化学习算法，引导搜救船只学会区域覆盖路径规划；最后，通过模拟实验和对比实验，验证了所提算法在海上搜救的覆盖路径规划问题上可满足优先搜索高概率区域，能够避免重复路径，尽可能地快速覆盖区域的目标，并且与其他算法相比拥有更卓越的性能。

本文的创新点主要包括：1）开发了基于深度强化学习的单搜救船只的海上搜救覆盖路径规划模型，设计了适用于海上搜救场景的状态空间、动作空间和奖励函数。2）所提算法结合了 DDQN 的双 Q 网络、Prioritized DDQN 的优先级经验回放池、Distributional DQN 的 Q 值的分布估计和 Noisy DQN 的噪声网络，形成一个更加强大的深度强化学习算法。3）所提算法在模拟实验下的多种评价指标都优于传统的平行线搜寻模式和其它深度强化学习算法。

## 1 海上搜救覆盖路径规划的马尔可夫决策过程

马尔可夫决策过程是描述强化学习问题的框架。马尔可夫性质是指当且仅当某时刻的状态只取决于上一时刻的状态时的一个随机过程[16]。在海上搜救中，搜救船只的未来状态仅依赖于当前的状态和动作。因此，可以将覆盖路径规划问题转化为马尔可夫决策过程，用状态空间 $S$、动作空间 $A$、状态转移函数 $P$、奖励函数 $R$ 和折扣因子 $\gamma$ 组成的元组$(S, A, P, P, \gamma)$来描述问题。

### 1.1 状态空间

将海上搜救区域划分为 $M \times N$ 个栅格，每个栅格根据包含概率(Probability Of Containment，POC)对应不同的颜色，如图 1 所示。红色代表 POC 值大的区域，淡红色代表 POC 值较大的区域，浅绿色代表 POC 值较小的区域，绿色代表 POC 值小的区域，白色代表 POC 值为 0 的区域。将区域按照 POC 值划分为 5 个等级$\{0, 1, 2, 3, 4\}$。

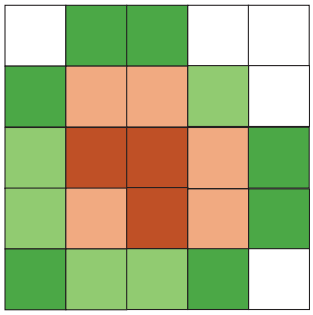


图 1 海上搜救区域的 POC 矩阵示意图

Fig.1 Schematic diagram of the POC matrix for maritime search and rescue area

覆盖路径规划问题通常涉及在给定区域内找到一条路径，使得路径能够覆盖到指定的目标或区域。为了定义状态空间，本文考虑了几个要素：

1）位置状态：海上搜救区域划分为 $M \times N$ 个栅格，记录在当前位置执行一个移动方向之后所在的新的栅格位置。这是一个 2 维数组，包含横坐标和纵坐标，将其转为 1 维数组，记录位置的新坐标。

2）POC 状态：记录每个位置的 $P$。当搜救船只覆盖过某个栅格时，其 $P$ 转为 0。

3）局部视野：记录搜救船只在新位置下周围的 $P$ 情况，通过一个局部视野来感知周围环境。

4）动作历史：记录搜救船只最近动作的历史

记录。可以帮助搜救船只更好地理解环境和当前情况。

状态空间描述如表 1 所示。

表 1　状态空间描述

Tab.1　State space description

| 要素 | 大小 |
|---|---|
| 位置状态 | 1 |
| POC 状态 | $M \times N$ |
| 局部视野 | 8 |
| 动作历史 | 25 |

### 1.2　动作空间

设计合理的动作空间对于搜救船只覆盖路径规划的有效性至关重要[17]。在每个决策时刻，搜救船只可以运动到相邻栅格，具有上、下、左、右四种基本运动状态。同时，搜救船只在搜索区域的边缘活动时，为了防止船只意外越界导致脱离目标区域或进入潜在的危险地带，其动作空间被限制为只能采取不越界的 2 ~ 3 个方向的动作；在非边缘区域，船只可以在 4 个基本方向自由移动。对动作空间的定义如图 2 所示。

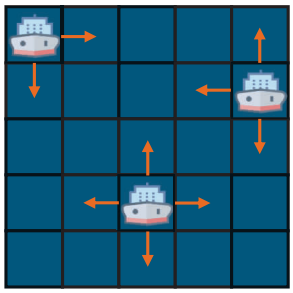


图 2　动作空间示意图

Fig.2　Schematic diagram of the action space

### 1.3　奖励函数

深度强化学习中的奖励函数用于评估智能体在环境中执行动作后的反馈信号，它直接影响智能体所学习到的策略和行为。因此，设计合适的奖励函数对于训练一个有效的深度强化学习算法至关重要。

在海上搜救覆盖路径规划中，奖励函数的设计需要综合考虑 4 个目标：一是优先搜索 POC 值高的区域；二是避免搜索 POC 值为 0 的区域；三是避免重复路径；四是尽可能快速地覆盖区域。

为了避免奖励稀疏产生的“高原问题”[18]，本文设计的奖励函数结合了即时奖励和回合奖励：即时奖励用来引导搜救船只在每一步选择最优的动作，以使得未来累积奖励最大化；回合奖励是搜救船只在完成覆盖任务或者执行最大步数后获得的奖励总和。

搜救船只每执行一个动作获得即时奖励 $R_{\mathrm{p}}$，如式(1)所示，并且在执行动作后将当前区域的 $P$ 重设为 0。

$$R_{\mathrm{p}} = \frac{P}{u} \tag{1}$$

式中，$u$ 为搜救船只在当前状态下经历的步数。随着步数的增加，$R_{\mathrm{p}}$ 减小，从而引导搜救船只优先覆盖概率高的区域。

若搜救船只完全覆盖某个 $P$ 等级的区域，获得对应的覆盖奖励 $R_{\mathrm{C}}$：

$$R_{\mathrm{C}} = \frac{10P}{u} \tag{2}$$

密集的奖励函数可以提供更精确的反馈，使得搜救船只能够更快速地了解其行为的好坏。因此，覆盖完全某个等级的区域再获得对应等级奖励，可进一步引导搜救船只优先覆盖概率高的区域。

若搜救船只探索至新区域，则给予奖励；若重复至已搜索过的区域，则给予惩罚：

$$R_{\mathrm{a}} = \begin{cases} 0.1, & \text{此区域未被覆盖过} \\ -0.1, & \text{此区域已被覆盖过} \end{cases} \tag{3}$$

$R_{\mathrm{a}}$ 的奖惩机制可以使搜救船只进一步地了解在当前状态下的最优动作，引导搜救船只避免重复覆盖区域。

当搜救船只完全覆盖区域或已执行了最大回合步数时，获得结束奖励 $R_{\mathrm{d}}$：

$$R_{\mathrm{d}} = \begin{cases} \dfrac{10(u_{\max} - u)}{u_{\max}}, & \text{完全覆盖区域} \\ -10, & \text{执行最大步数} \end{cases} \tag{4}$$

式中，$u_{\max}$为最大回合步数。当智能体达到任务关键点时给予奖励，这样可以使智能体更快地学习到重要的决策。因此，搜救船只在完全覆盖区域时，给予丰富的奖励；若执行了最大回合步数仍未完成，则给予严重的惩罚。这样可以有效地引导搜救船只尽可能快速地覆盖区域。

搜救船只执行动作时，还会获得额外的累加奖励，用于最后的回合奖励：

$$R_{\mathrm{r}}^{i} = R + \gamma R_{\mathrm{r}}^{i-1} \tag{5}$$

式中，$R_{\mathrm{r}}^{i}$ 为第 $i$ 步的回合奖励，$\gamma(0 < \gamma < 1)$ 为折扣因子。这种奖励模式使得越靠近回合结束时的阶段动作对回合奖励的贡献越大。

最终，奖励函数的定义为

$$R = R_{\mathrm{p}} + R_{\mathrm{c}} + R_{\mathrm{a}} + (R_{\mathrm{r}} + R_{\mathrm{d}})(1 - d) \tag{6}$$

式中，$d$ 为是否结束此回合的标记，当回合结束时加上回合奖励和结束奖励。

## 2 基于深度强化学习的海上搜救覆盖路径规划算法

海上搜救任务因其环境的不确定性和复杂性，对路径规划提出了极高的要求。DQN[19] 作为深度强化学习的里程碑算法，尽管在许多应用中表现优异，但在处理如海上搜救等复杂任务时，仍存在若干不足之处。为了提高搜救任务中的路径规划质量和搜寻效率，本文结合了 DQN 的 4 个扩展算法：DDQN[20]、Prioritized DDQN[21]、Distributional DQN[22] 和 Noisy DQN[23]。DDQN 通过引入双重估计机制缓解了 $Q$ 值过估计问题，降低了选择风险较高的路径概率。Prioritized DDQN 对经验回放中的样本分配优先级，重点学习对策略改进影响较大的经验，提高在复杂环境中的搜救效率。Distributional DQN 通过捕捉 Q 值的全概率分布，更好地评估不同行动的潜在风险和回报，制定更加稳健的搜救路径。Noisy DQN 在神经网络中引入噪声，有效地探索广阔复杂的海域，同时避免局部最优解。本文结合这 4 个扩展算法对海上搜救的覆盖路径规划问题进行建模，使其无论是在路径规划质量还是搜寻效率方面，都实现了远优于其它算法性能。

### 2.1 神经网络结构设计

本文算法的神经网络结构主要由输入层、全连接层和输出层组成。输出层采用了 Distributional DQN 的思想，能够输出每个动作的 $Q$ 值分布。在传统的 DQN 网络中，通常只输出对动作价值 $Q$ 的单一估计值，这种方式在面对海上搜救任务中复杂的环境时可能不足以捕捉细微的风险和奖励差异。而 Distributional DQN 通过估计动作价值 $Q$ 的分布，能够更全面地反映海上不同区域的风险等级和搜救优先级。具体而言，它将价值 $Q$ 限定在之间[$V_{\min}$，$V_{\max}$]，选择 51 个等距的采样点，通过神经网络输出这些采样点的概率分布。这使得搜救船只能够理解在不同海况的条件下，执行各类搜救动作的潜在风险与回报，从而更有效地制定搜救策略。分布式 $Q$ 函数如图 3 所示。

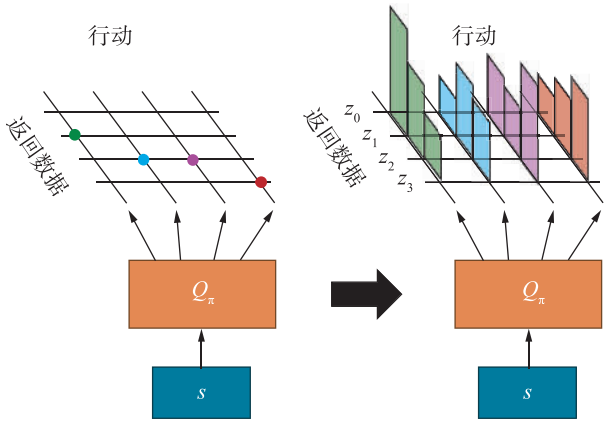



图 3 分布式 $Q$ 函数

Fig.3 Distributional $Q$-function

此外，在复杂多变的海上环境中，确保探索的有效性和策略的稳定性尤为重要。为了平衡探索与策略，本文在神经网络的全连接层中引入了 Noisy DQN 的噪声网络。Noisy DQN 通过向神经网络的权重和偏置项中添加随机噪声，使得每次计算 $Q$ 值时都带有微小的扰动。这种变化是小幅度的且是连续的，使得网络在每一步中的探索是渐近的，避免了过大的波动，保持了训练的稳定性。这种自适应噪声能够在早期阶段提供足够的随机性，有助于搜救船只应对未知的海上情况。当策略逐渐稳定时，噪声强度会自动减弱，从而减少不必要的扰动，确保搜救决策的稳定性。通过这种方式，网络能够在保持稳定性的同时，更加全面地探索不同的搜救策略，从而在广泛的海域中实现高效、准确的覆盖。

本文通过结合 Distributional DQN 和 Noisy DQN 的特性，设计出一个适用于海上搜救场景的强大且稳健的神经网络结构。算法的神经网络结构如图 4 所示。

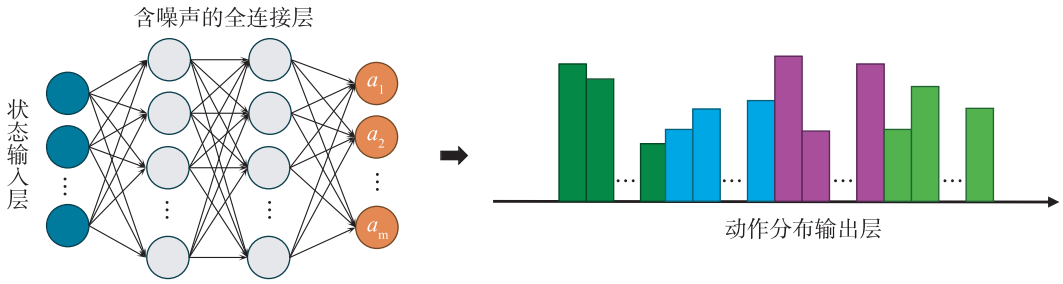



图 4 神经网络结构示意图

Fig.4 Schematic diagram of neural network architecture

### 2.2　动作选择策略

在深度强化学习中，智能体需要在利用已知信息来最大化奖励和探索未知状态以发现新的奖励之间进行权衡，这一点在海上搜救任务中尤为关键。为了避免搜救船只陷入“探索-利用窘境”[24]，本文采用 $\varepsilon$-greedy 策略来平衡探索与利用问题。针对海上搜救中的实际需求，恒定的 $\varepsilon$ 难以保证算法在复杂多变的环境中稳定收敛。为此，本文在迭代计算中动态调整了 $\varepsilon$，使算法能够在初期阶段进行广泛的环境探索，逐步向利用已知信息过渡，从而更好地应对海上各种不确定性状况。通过逐渐减小 $\varepsilon$，搜救船只可以在训练过程中逐步从探索向利用过渡。这种调整机制使得搜救船只在初期能以较大的 $\varepsilon$ 探索广阔的海域和多变的环境，发现潜在的高回报搜救路径。而随着训练的深入，搜救船只将逐渐收敛于已知的高效路径，以较小的 $\varepsilon$ 进行精确的策略执行，确保搜救任务的高效完成。$\varepsilon$ 取值为

$$\varepsilon = \varepsilon_{\mathrm{e}} + (\varepsilon_{\mathrm{s}} - \varepsilon_{\mathrm{e}})\mathrm{e}^{-\frac{c}{\varepsilon_{\mathrm{d}}}} \tag{7}$$

式中，$\varepsilon_{\mathrm{s}}$ 为 $\varepsilon$ 的初始值，$\varepsilon_{\mathrm{e}}$ 为 $\varepsilon$ 的最终值，$\varepsilon_{\mathrm{d}}$ 为 $\varepsilon$ 的衰减率，$c$ 为采取行动的次数。

在动作选择过程中，通过 Distributional DQN 的分布式 $Q$ 函数选择当前已知的最佳动作。具体而言，智能体通过神经网络计算各个动作的分布，对每个动作的分布进行求和，并选择求和值中的最大值作为最佳动作。这种基于动作值分布的方法，能够让搜救船只不仅估计出动作的期望价值，还能够提供关于不确定性的额外信息，从而更好地理解复杂海况并做出更精准的搜救决策。最终动作选择策略 $\pi(a|s)$ 为

$$a_t = \begin{cases} \xi_{\mathrm{rand}} = \{1, 2, 3, 4\}, & \text{采样概率为 } \varepsilon \\ \underset{a}{\operatorname{argmax}}\, Q(s, a), & \text{采样概率为 } 1-\varepsilon \end{cases} \tag{8}$$

式中，$\xi_{\mathrm{rand}}$ 为一个随机数。通过这种策略，本文算法在应对海上搜救中的复杂环境时，能够有效地平衡探索与利用，从而提高搜救任务的成功率和效率。

### 2.3　经验回放池设计

本文算法的经验回放池机制采用了 Prioritized DDQN 的思想，即对数据按照优先级进行采样，这些数据通常包括当前状态 $S_t$、动作 $a_t$、奖励 $r_t$、下一状态 $S_{t+1}$ 和结束标志 $d_t$。在传统的 DQN 中，经验池中的数据采样通常是随机的，这样可能会忽略掉一些关键的经验数据。然而，在海上搜救任务中，不同环境下的经验可能具有不同的重要性，随机采样可能导致搜救策略在关键情境下的训练不足。为了解决这一问题，本文采用了 Prioritized DDQN 中的优先级经验回放机制。该机制通过对采样数据的时序差分误差（TD-error）的大小来确定数据的重要性，从而为不同经验分配不同的优先级。当 TD-error 越大时，代表数据越不好训练，因此给予它更大的概率被采样到[25]。优先级经验回放池会根据 TD-error 动态更新池中样本的优先度 $p$ 和采样权重 $\omega$：

$$\delta = r_t + \hat{Q}(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{9}$$

$$p(i) = \frac{|\delta_i|^{\alpha}}{\sum_j |\delta_j|^{\alpha}} \tag{10}$$

$$\omega_j = \frac{(N \cdot p(j))^{-\beta}}{\max_i(\omega_i)} \tag{11}$$

式中，$\delta$ 为 TD-error，是基于当前状态的策略网络 $Q$ 与基于下一状态的目标网络 $\hat{Q}$ 和 $r_t$ 对当前 $Q$ 值估计的偏差程度；$|\delta_i|$ 是第 $i$ 个经验的 TD-error 的绝对值，$\sum_j |\delta_j|^{\alpha}$ 是经验池中所有经验的优先级的归一化因子，使得优先级的总和为 1，$\alpha$ 是一个超参数；$\omega_j$ 是第 $j$ 个经验的采样权重；$N$ 为经验池中的总样本数；$p(j)$ 是第 $j$ 个经验的优先级；$\beta$ 是一个超参数；$\max_i(\omega_i)$ 是所有采样权重中的最大值，通常用于归一化采样权重，使得它们的最大值为 1，从而保持稳定的学习过程。通过引入优先级经验回放机制，确保了搜救船只能够更频繁地接触到这些关键数据，从而在复杂的海况下形成更稳健的搜救决策。优先级经验回放示意图如图 5 所示。

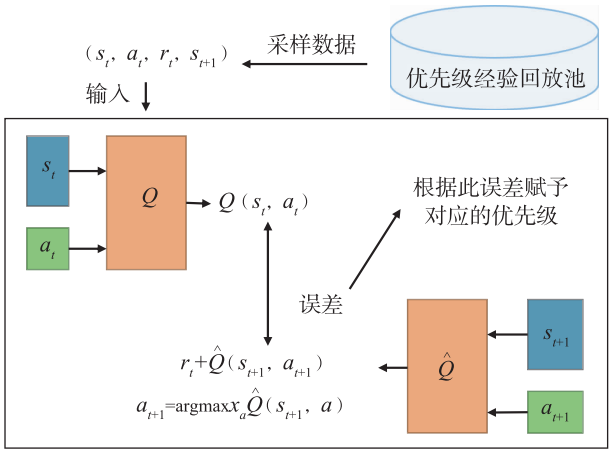



图 5　优先级经验回放示意图

Fig.5　Schematic diagram of prioritized experience replay

### 2.4　算法框架和流程

基于深度强化学习的覆盖路径规划流程如图 6 所示。具体流程为：

**步骤 1** 搜救船只根据当前状态 $S_t$ 选择一个动作 $a_t$。

**步骤 2** 搜救船只在海洋环境中执行这个动作，得到反馈信息，即下一个状态 $S_{t+1}$、奖励 $r_t$ 和结束标志 $d_t$。

**步骤 3** 使用策略网络 $Q_\pi$ 预测当前状态下选择某个动作的 $Q$ 值。使用目标网络 $\hat{Q}_\pi$ 预测下一状态下的最大 $Q$ 值。计算 TD 误差，如果是结束状态，即 $d_t$ 为 True，则只计算当前奖励和预测的 $Q$ 值之间的差异，否则还要考虑折扣后的未来奖励。

**步骤 4** 将这个误差和对应的经验存入优先级经验($S_t$, $a_t$, $r_t$, $S_{t+1}$, $d_t$)回放缓冲区，用于训练时的样本选择。

**步骤 5** 更新当前状态为 $S_{t+1}$，并进行算法更新。

重复上述过程，直到完成所有的训练回合。这个循环逻辑使搜救船只在多个训练回合中不断与海洋环境互动、积累经验，并通过不断更新算法来提升策略的效果。

算法更新的过程如图 7 所示。具体流程为：

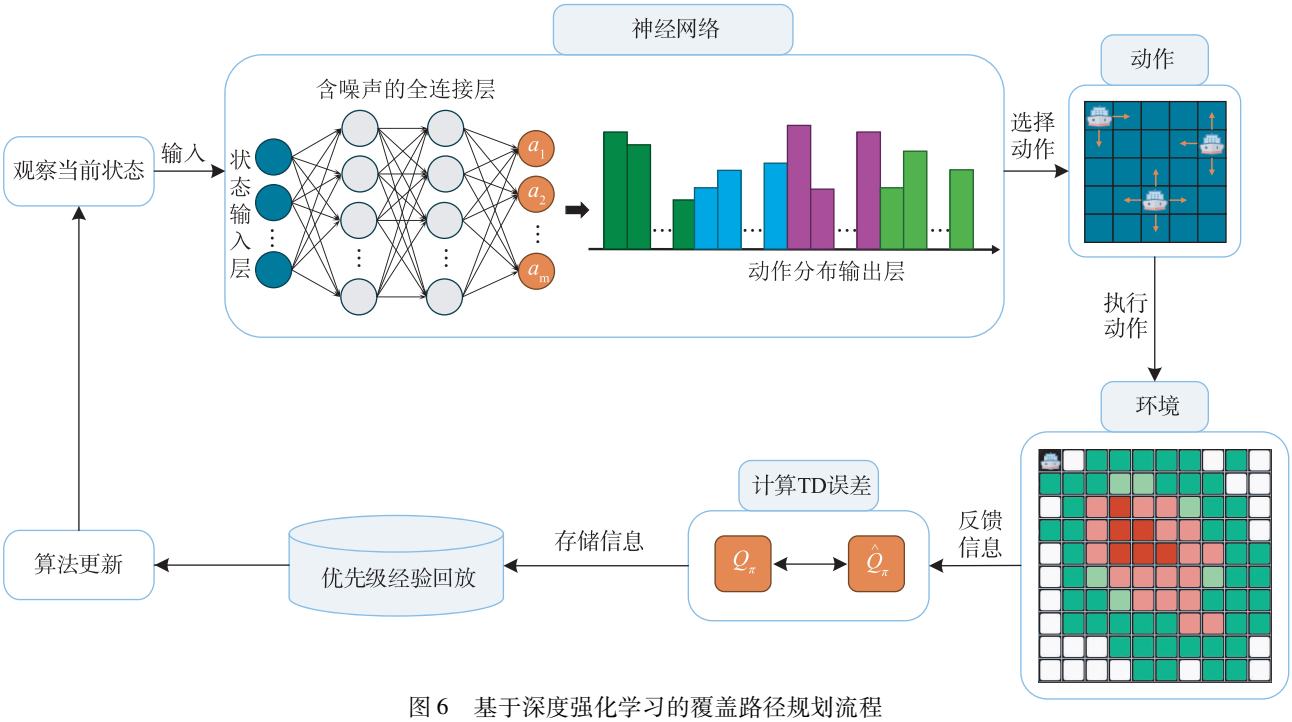



图 6 基于深度强化学习的覆盖路径规划流程

Fig.6 Process for coverage path planning based on deep reinforcement learning

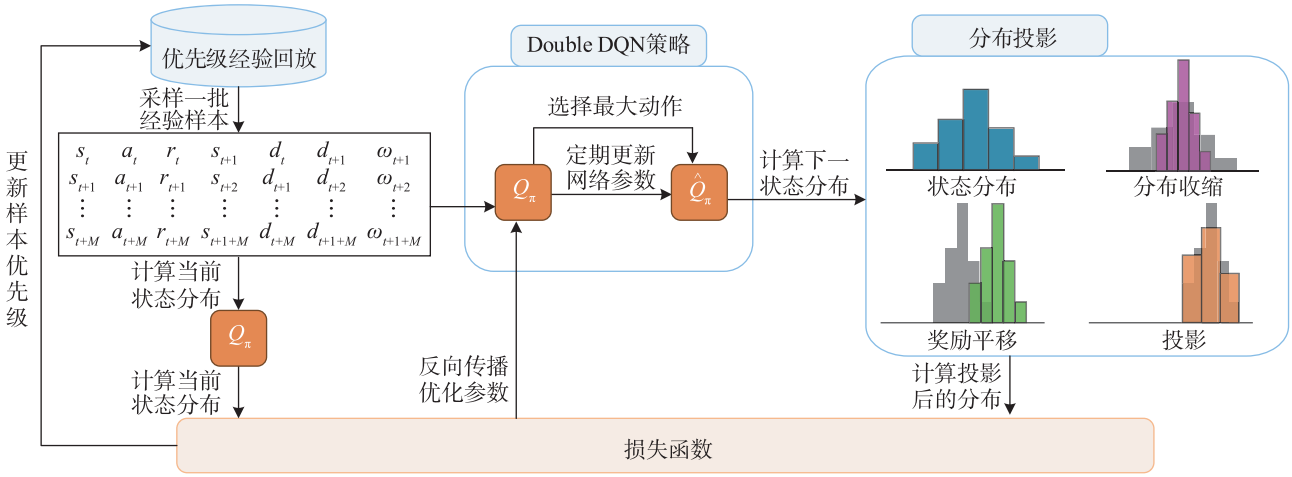



图 7 算法更新流程

Fig.7 Algorithm update process

**步骤 1**　从优先级经验回放池中采样一批经验样本，获取状态、动作、奖励、下一状态、是否结束标志，以及相应的样本索引和权重，即（$S_t$，$a_t$，$r_t$，$S_{t+1}$，$d_t$，$id_t$，$\omega_t$）。

**步骤 2**　通过 Double DQN 策略，先使用 $Q_\pi$ 计算下一状态的分布，并选择使 $Q$ 值最大的动作。再使用 $\hat{Q}_\pi$ 计算这个动作对应的分布，得到下一状态下选择的动作的分布。

**步骤 3**　计算分布投影。先计算目标分布的值域 $T_z$，具体来说，是将奖励加上折扣后的未来分布的值域。然后对后的范围进行限制，确保其在支持范围内。接下来，将 $T_z$ 映射到离散的分布上，计算投影后的分布位置的上下界，并根据这些界限对相应的概率分布进行加权累加。

**步骤 4**　使用 $Q_\pi$ 计算当前状态下的动作分布，并获取对应动作的分布概率。

**步骤 5**　使用交叉熵损失来衡量分布之间的差异，目标是最小化投影分布与当前分布之间的差异。

**步骤 6**　使用经验采样时的权重对损失进行加权，执行反向传播更新策略网络的参数，并进行梯度裁剪以防止梯度爆炸。

**步骤 7**　根据计算的损失更新经验池中相应样本的优先级，这样在下次采样时，高误差样本更容易被选择。

**步骤 8**　每隔一段时间，将 $Q_\pi$ 的参数复制到 $\hat{Q}_\pi$ 中，以确保目标网络的稳定性。

**步骤 9**　重置 $Q_\pi$ 和 $\hat{Q}_\pi$ 中的噪声参数，这在使用 Noisy Net 时很重要，用于探索。

通过不断更新 $Q_\pi$ 的参数，使得搜救船只能够更准确地评估状态的价值分布，从而提升决策的效果。

## 3　算法应用研究

### 3.1　示例描述

本文通过国家海上搜救环境保障平台（marine-sar.cn），在渤海的事故易发地区，长岛县东北侧（121°7′E，38°7′N）模拟一起海上事故。利用蒙特卡罗随机粒子法对 25 位落水人员进行 24 h 的漂移轨迹预测，如图 8 所示，时间线为 1 月 17 日 14 时至 1 月 18 日 12 时。由于此处海域在 1 月份的理论存活时间仅为 3 h，且考虑到应急响应的时间，因此预测 17 日 15 时至 17 日 18 时的搜救区域，最后生成一个 10×10（1.72 km×1.57 km）的 POC 矩阵地图，如图 9 所示。

### 3.2　参数设置

本文通过网格搜索法，对参数进行多次调优，最终参数设置如表 2 所示。

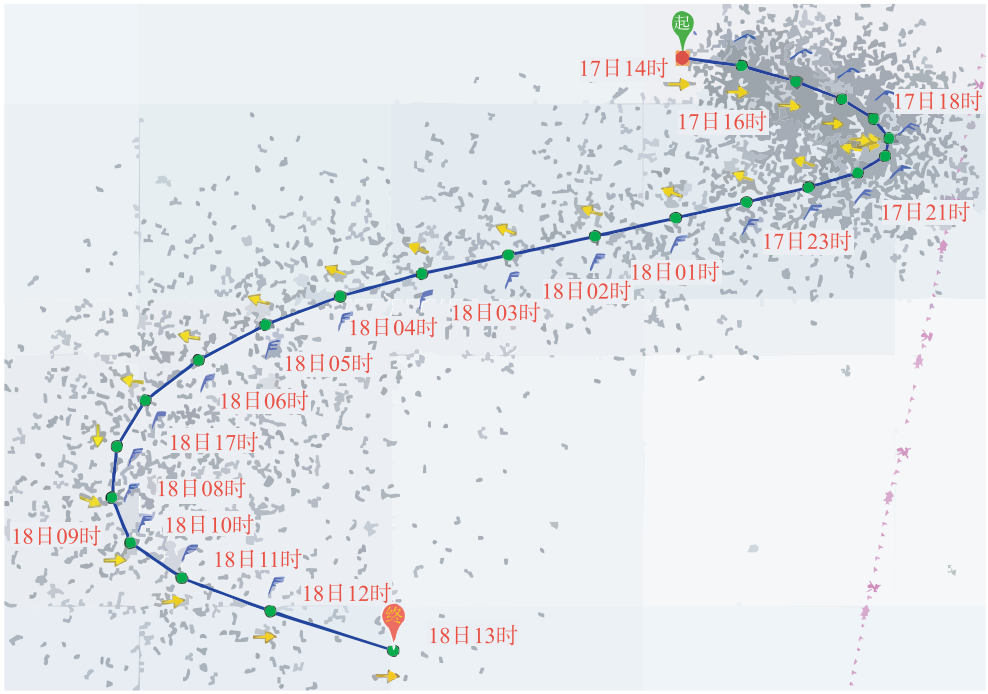



图 8　漂移轨迹预测路径

Fig.8　Drift trajectory prediction path

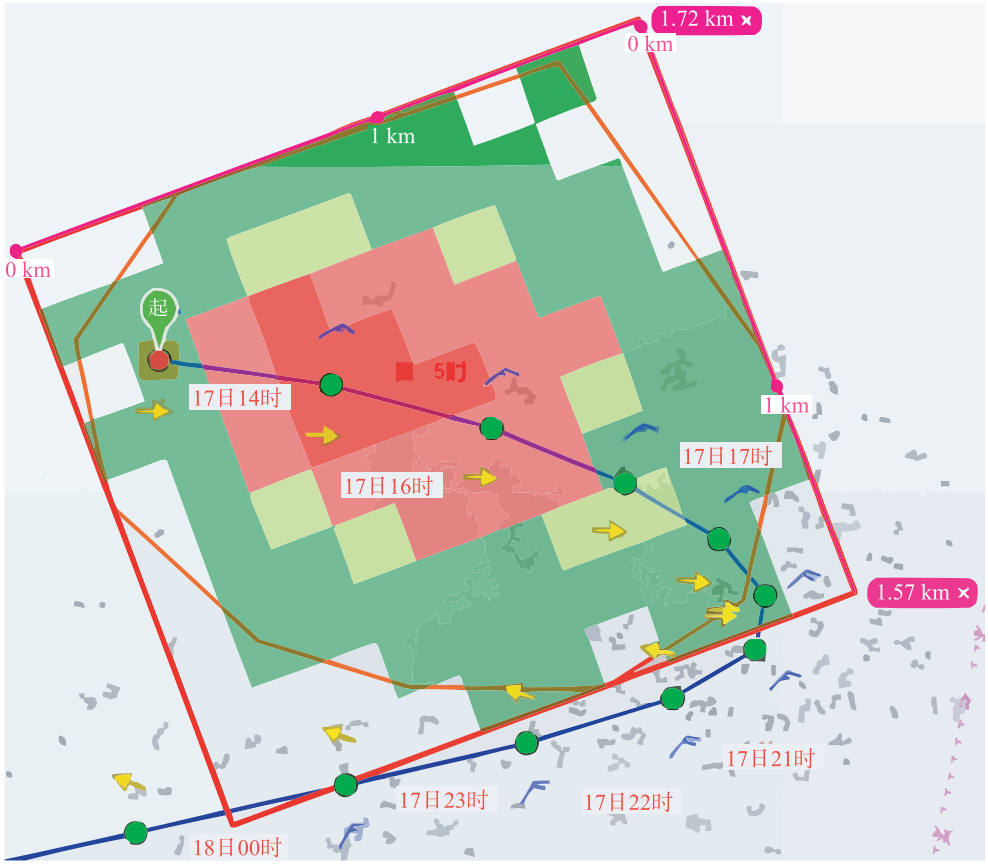



图 9 模拟海上事故的 POC 矩阵地图

Fig.9 POC matrix map for simulating maritime accidents

表 2 参数设置

Tab.2 Parameter setting

| 参数 | 值 |
|---|---|
| 训练回合数 | 4 000 |
| 每回合最大步数 | 350 |
| 学习率 | $5\times10^{-4}$ |
| 权重衰减 | $1\times10^{-6}$ |
| 折扣因子 | 0.7 |
| 初始值 | 0.99 |
| 最终值 | 0.1 |
| 衰减率 | 20 000 |
| 经验池容量 | $1\times10^{7}$ |
| 经验池采样批次 | 400 |
| 目标网络更新频率 | 300 |
| 价值采样点最大值 | 9 |
| 价值采样点最小值 | −9 |
| 第 1 层隐藏层 | 1 024 |
| 第 2 层隐藏层 | 512 |

### 3.3 实验验证及结果分析

#### 3.3.1 训练结果

绘制训练过程中智能体收到的奖励回报随时间的变化曲线是评估深度强化学习算法性能的一种常用方式。通过训练本文算法，得到的累积奖励值的变化趋势如图 10 所示，其中实线为累积奖励的平滑值，阴影区域表示累积奖励的实际值。由图 10 可以看出，仅仅训练了 500 步左右，就有显著的收敛效果，且随着训练次数的增加，累积奖励呈现整体上升的趋势，训练的波动趋于稳定，并逐渐收敛到最优值。

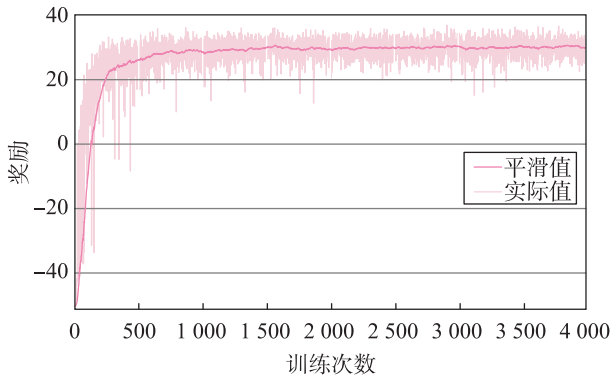



图 10 奖励函数收敛曲线

Fig.10 Reward function convergence curve

### 3.3.2　对比实验

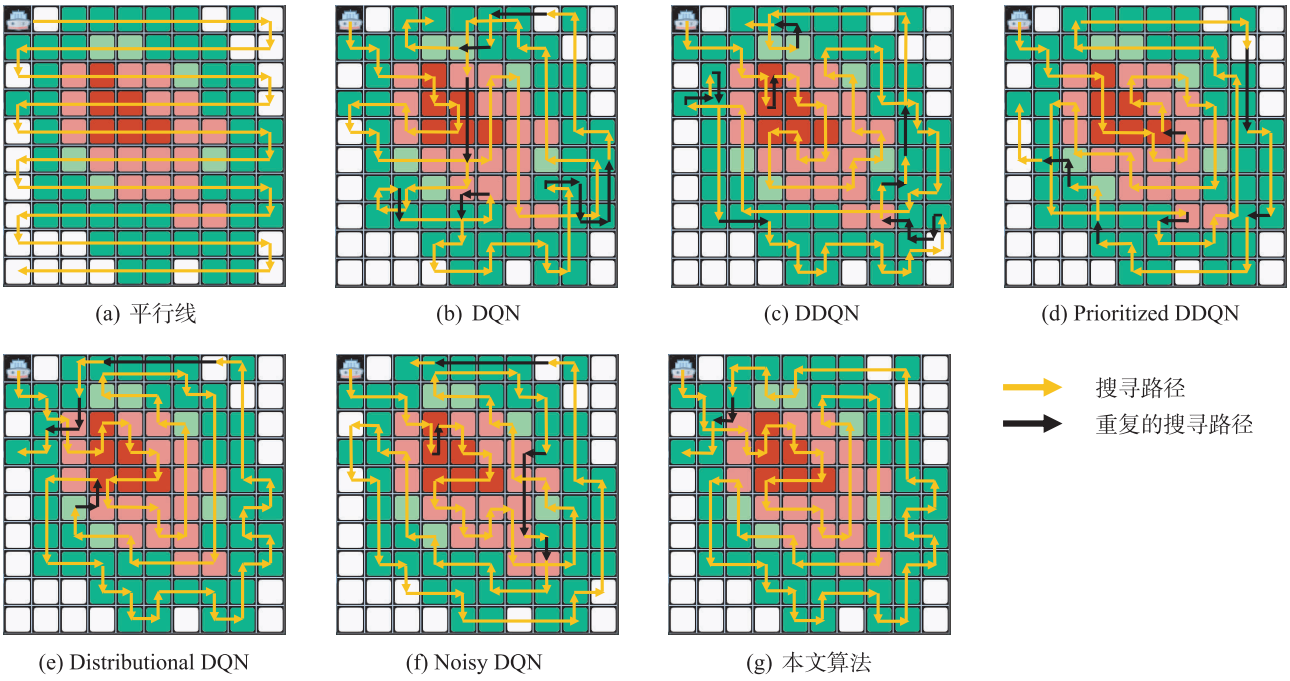
本文将所提算法与目前海上搜救辅助决策系统中常用的平行线搜寻模式[26]及其它深度强化学习算法，包括 DQN、DDQN、Prioritized DDQN、Distributional DQN 和 Noisy DQN，进行了比较。图 11 展示了这些方法生成的搜寻路径。

由图 11 可直观看出，平行线搜寻模式的路径较为规则，但难以迅速覆盖重点区域。而基于深度强化学习算法的搜寻路径则能够使搜救船只更快地搜寻高概率区域，并尽量避免零概率区域，从而提高搜寻效率。相比之下，本文算法规划的路径明显更加顺畅合理，极大降低了重复覆盖路径。虽然基于深度强化学习算法规划的搜寻路径都存在了重复路径，但路径质量不能仅仅通过是否存在重复路径来评判，搜寻效率同样是一个重要的标准，可以通过路径获得的奖励体现。因此，为了定量评估上述的方法，本文对覆盖路径执行的步数、重复率和获得的奖励值进行了统计，如表 3 所示。结果表明，本文算法生成的搜寻路径在覆盖步数、重复率和奖励值方面表现优异。

表 3　不同方法下路径规划结果的定量评价

Tab.3　Quantitative evaluation of path planning results under different methods

| 方法 | 覆盖步数 | 重复率 /% | 奖励值 |
|---|---|---|---|
| 平行线 | 100 | 0 | 27.51 |
| DQN | 97 | 14.43 | 33.49 |
| DDQN | 94 | 14.89 | 34.66 |
| Noisy DQN | 93 | 10.75 | 34.85 |
| Prioritized DDQN | 91 | 9.89 | 35.16 |
| Distribution DQN | 87 | 9.2 | 36.05 |
| 本文算法 | 81 | 2.47 | 37.26 |




(a) 平行线　(b) DQN　(c) DDQN　(d) Prioritized DDQN

(e) Distributional DQN　(f) Noisy DQN　(g) 本文算法

图 11　不同方法的路径规划结果

Fig.11　Path planning results of different methods

除了从路径规划结果的角度对深度强化学习算法进行评价外，本文还从深度强化学习算法的收敛速度和训练稳定性方面进行了定量评价。在相同的环境下，分别训练不同算法，得到的累积奖励值的变化趋势如图 12 所示，生成的覆盖路径所执行的步数曲线如图 13 所示。为了更明显地展示出算法的优劣差距，只展示了 1 000 次训练的奖励曲线和步长曲线。

通过观察训练曲线可直观看出，本文提出的算法在达到特定性能水平时所需的训练次数显著少于其他算法。与此同时，本文算法的训练曲线波动较小，表现出高度的平滑性，这表明训练过程极为稳定，远优于其它算法的表现。实验结果的对比进一步证明，无论是在路径规划效果还是在整体训练表

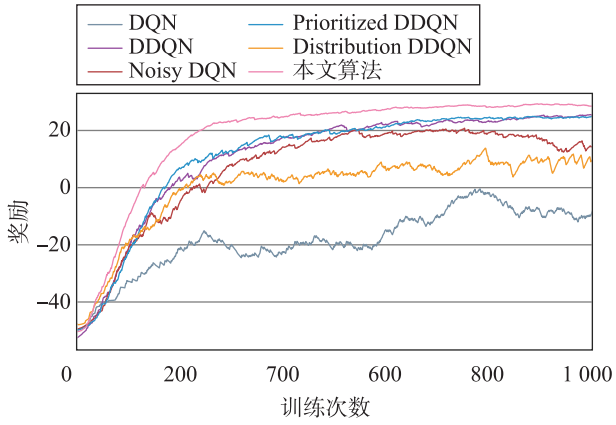



图 12 不同算法下的奖励曲线

Fig.12 Reward curves under different algorithms

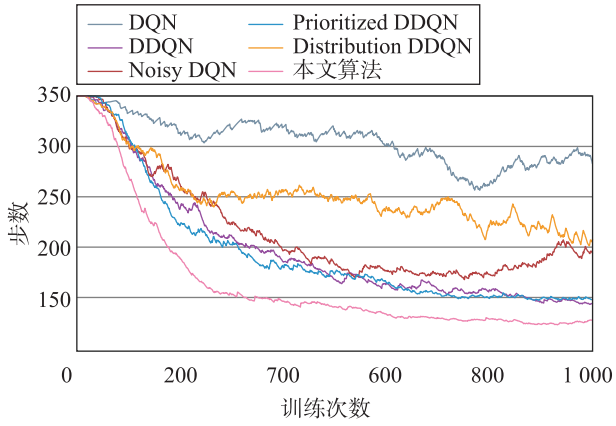



图 13 不同算法下的步长曲线

Fig.13 Step length curves under different algorithms

现方面，本文提出的算法均显著优于其它的搜寻模式，展现出更高的效率和可靠性。

## 4 结论

本研究针对海上搜救的覆盖路径规划问题，提出了一种基于深度强化学习的海上搜救覆盖路径规划模型。该模型算法结合了 DDQN、Prioritized DDQN、Distributional DQN 和 Noisy DQN，得到了一个更适用于海上搜救覆盖路径规划问题的深度强化学习算法。实验结果表明，训练后的算法能够在复杂的环境中规划出合理的路径来完成覆盖目标，可实现优先搜索 POC 值高的区域、避免搜索 POC 值为 0 的区域、尽量避免重复路径和尽可能快速覆盖区域的功能，有效地解决了在传统搜寻模式下固定的搜索路径带来的效率低、适应性弱等问题。对比实验验证了本文算法较其他算法生成的覆盖路径可执行更少的步长，更低的重复率和获取更高的奖励值，并且拥有更快的收敛和更稳定的训练效果。

但是，本研究存在一定的局限性，例如在路径规划过程中的环境变量是恒定的、模拟的搜救区域划分和搜救船只行动都是离散的、未考虑多搜救船只的协同搜救场景等，这些局限性对研究结果和应用的泛化能力产生影响。因此在未来的研究中，可以引入动态环境模拟、多船协同行动模拟、增加场景的复杂性和多样性等，进一步研究和验证。通过解决这些局限性，可以使研究结果更加全面和可信，并提高其在实际应用中的指导意义和可行性。

## 参考文献

[1] JIN Y, WANG N, SONG Y, et al. Optimization model and algorithm to locate rescue bases and allocate rescue vessels in remote oceans[J]. Soft Computing, 2020, 25(4): 1 – 18.

[2] 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 Q Q, GAO Y Y, GUO Y, et al. Target search path planning for naval battle field based on deep reinforcement learning [J]. Journal of Systems Engineering and Electronics, 2022, 44(11): 3486 – 3495.

[3] TAN C S, MOHD-MOKHTAR R, ARSHAD M R. A comprehensive review of coverage path planning in robotics using classical and heuristic algorithms[J]. IEEE Access, 2021, 9: 119310 – 119342.

[4] PANG B, SONG Y, ZHANG C, et al. Effect of random walk methods on searching efficiency in swarm robots for area exploration[J]. Applied Intelligence, 2021, 51(7): 5189 – 5199.

[5] MOHAMMAD MINHAZ FALAKI P M, PADMAN A, NAIR V G, et al. Simultaneous exploration and coverage by a mobile robot[M]//Control Instrumentation Systems. Berlin, Germany: Springer-Verlag, 2020: 33 – 41.

[6] JIANG X, DENG Y. UAV track planning of electric tower pole inspection based on improved artificial potential field method [J]. Journal of Applied Science and Engineering, 2021, 24(2): 123 – 132.

[7] LE V A, PRABAKARAN V, SIVANANTHAM V, et al. Modified A-star algorithm for efficient coverage path planning in tetris inspired self-reconfigurable robot with integrated laser sensor[J/OL]. Sensors, 2018, 18(8)[2024 – 07 – 01]. https://www.mdpi.com/1424 – 8220/18/8/2585. DOI: 10.3390/s18082585.

[8] NG M, CHONG Y, KO K, et al. Adaptive path finding algorithm in dynamic environment for warehouse robot[J]. Neural Computing and Applications, 2020, 32(17): 13155 – 13171.

[9] CHOI S, LEE S, VIET H H, et al. B-Theta*: An efficient online coverage algorithm for autonomous cleaning robots[J]. Journal of Intelligent & Robotic Systems, 2017, 87(2): 265 – 290.

[10] SUN G, ZHOU R, DI B, et al. A novel cooperative path planning for multi-robot persistent coverage with obstacles and coverage period constraints[J/OL]. Sensors, 2019, 19(9)[2024 – 07 – 02]. https://www.mdpi.com/1424 – 8220/19/9/1994. DOI: 10.3390/s19091994.

[11] LE V A, KU P, TUN T T, et al. Realization energy optimization of complete path planning in differential drive based self-reconfigurable floor cleaning robot[J]. Energies, 2019, 12(6)[2024 – 06 – 15]. https://www.mdpi.com/1996 – 1073/12/6/1136. DOI: 10.3390/en12061136.

[12] KWON B, THANGAVELAUTHAM J. Autonomous coverage path planning using artificial neural tissue for aerospace applications [C/OL]//IEEE Aerospace Conference. Piscataway, USA: IEEE, 2020 [2024 – 07 – 04]. https://ieeexplore.ieee.org/abstract/document/9172556. DOI: 10.1109/AERO47225.2020.9172556.

[13] FU M C. Handbook of simulation optimization[M]. Berlin, Germany: Springer, 2015: 341 – 379.

[14] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529 – 533.

[15] LI S, CHEN X, ZHANG M, et al. A UAV coverage path planning algorithm based on double deep $Q$-network[J]. Journal of Physics: Conference Serie, 2022, 2216(1): 12 – 17.

[16] 张伟楠，沈键，俞勇. 动手学强化学习[M]. 北京：人民邮电出版社，2022：19.
ZHANG W N, SHEN J, YU Y. Hands-on reinforcement learning[M]. Beijing: People′s Posts and Telecommunications Press, 2022: 19.

[17] AI B, JIA M, XU H, et al. Coverage path planning for maritime search and rescue using reinforcement learning[J/OL]. Ocean Engineering, 2021, 241 [2024 – 07 – 01]. https://www.sciencedirect.com/science/article/pii/S0029801821014220. DOI: 10.1016/j.oceaneng.2021.110098.

[18] WU J, CHENG L, CHU S, et al. An autonomous coverage path planning algorithm for maritime search and rescue of persons-in-water based on deep reinforcement learning[J/OL]. Ocean Engineering, 2024, 291[2024 – 07 – 04]. https://www.sciencedirect.com/science/article/pii/S0029801823027877. DOI: 10.1016/j.oceaneng.2023.116403.

[19] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[C]//Workshops at the 26th Neural Information Processing Systems. New York, USA: ACM, 2013: 201 – 220.

[20] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double $Q$-learning[C]//AAAI Conference on Artificial Intelligence. Reston, USA: AIAA, 2016: 2094 – 2100.

[21] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[EB/OL]. (2015 – 11 – 18)[2024 – 07 – 08], https://arxiv.org/pdf/1511.05952.pdf. DOI: 10.48550/arXiv.1511.05952.

[22] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning[C]//International Conference on Machine Learning. New York, USA: PMLR, 2017: 449 – 458.

[23] FORTUNATO M, AZAR M G, PIOT B, et al. Noisy networks for exploration[EB/OL]. (2019 – 07 – 09)[2024 – 07 – 08]. https://arxiv.org/pdf/1706.10295.pdf. DOI: 10.48550/arXiv.1706.10295.

[24] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. 2nd ed. Cambridge, MA, USA: MIT, 2018: 26 – 27.

[25] 王琦，杨毅远，江季. Easy RL：强化学习教程[M]. 北京：人民邮电出版社，2022：165
WANG Q, YANG Y Y, JIANG J. Easy RL: Reinforcement learning course[M]. Beijing: People′s Posts and Telecommunications Press, 2022: 165

[26] JIN Y, WANG N, SONG Y, et al. Optimization model and algorithm to locate rescue bases and allocate rescue vessels in remote oceans[J]. Soft Computing, 2021, 25: 3317 – 3334.

## 作者简介

韩靖童(2000 – )，女，硕士生。研究领域为深度强化学习，智能体覆盖路径规划，海上搜救决策。

余　倩(1998 – )，女，硕士生。研究领域为救战伤员系统评估与预测。

刘　源(1980 – )，男，博士，教授。研究领域为应急医学救援组织，公共管理。