DOI:10.11918/202410030

局部风信息启发的 AVW-PPO 室内气源定位算法

李世钰[1],袁 杰[2],谢霖伟[1],郭 旭[1],张宁宁[1]

(1.新疆大学 电气工程学院,乌鲁木齐 830017; 2.新疆大学 智能科学与技术学院,乌鲁木齐 830017)

摘 要:为解决当前复杂、动态室内羽流环境中气源定位(OSL)效率低下和成功率不足的问题,尤其在湍流条件下机器人难以准确感知环境并实现有效导航的挑战,提出了一种基于深度强化学习的辅助价值与风导向的近端策略优化(AVW-PPO)算法。首先,在原始 PPO 算法的基础上引入辅助价值网络,以减少单一值网络的估计偏差,从而提升策略更新的稳定性与预测精度。其次,设计了一种风导向策略,将局部环境风场信息融入强化学习框架中的状态空间与奖励函数,使机器人能够更敏锐地感知羽流环境的动态变化,优化其决策路径,从而有效提高气源定位的效率。最后,通过构建二维环境中的气体扩散模型,在3种不同的湍流条件下对所提算法进行了测试。结果表明:相同环境条件下,AVW-PPO 算法在平均搜索步数和成功率两个指标上均优于其他同类算法,且定位成功率超过99%。其中,风导向策略在提升搜索效率方面表现尤为突出,有助于减少机器人完成任务所需的时间。本研究为解决室内复杂湍流环境下的气源定位问题提供了新思路和新方法。

关键词:气源定位;深度强化学习;近端策略优化(PPO);辅助价值网络;风导向策略

中图分类号:TP242.6 **文献标志码**:A **文章编号**:0367-6234(2025)08-0057-12

Local wind information-inspired AVW-PPO indoor odor source localization algorithm

LI Shiyu[1], YUAN Jie[2], XIE Linwei[1], GUO Xu[1], ZHANG Ningning[1]

(1. School of Electrical Engineering, Xinjiang University, Urumqi 830017, China;
2. School of Intelligence Science and Technology, Xinjiang University, Urumqi 830017, China)

Abstract: To address the challenges of low efficiency and insufficient success rates in odor source localization (OSL) within complex and dynamic indoor plume environments, particularly where robots struggle to accurately perceive the environment and navigate effectively under turbulent conditions, this paper proposes an auxiliary value and wind-guided proximal policy optimization (AVW-PPO) algorithm based on deep reinforcement learning. First, an auxiliary value network is introduced into the original PPO framework to reduce the estimation bias of a single value network, thereby improving prediction accuracy and stabilizing policy updates. Next, a wind-guided strategy is designed to integrate local wind field information into the state space and reward function of the reinforcement learning framework, enabling the robot to better perceive dynamic changes in the plume environment and optimize its decision-making path, thus significantly improving the efficiency of odor source localization. Finally, a gas diffusion model in a two-dimensional environment is constructed to test the proposed algorithm under three different turbulence conditions. Experimental results demonstrate that, under identical environmental conditions, the AVW-PPO algorithm outperforms other comparable algorithms in terms of average search steps and success rates, achieving a localization success rate of over 99%. Notably, the wind-guided strategy significantly boosts search efficiency, helping to reduce the time required for the robot to complete tasks. This study provides new insights and methodologies for addressing odor source localization problems in complex turbulent indoor environments.

Keywords: odor source localization; deep reinforcement learning; proximal policy optimization (PPO); auxiliary value network; wind-guided strategy

气源定位是机器人技术中的一个关键问题[1],在环境监测、搜索与救援以及安全检测等领域具有广泛的应用前景。这些任务要求机器人能够在未知或动态变化的环境中准确地追踪到化学物质的源头。尽管已有多种算法被提出来解决这些问题,但大多数传统方法依赖于严格的环境假设或预设的行

收稿日期:2024-10-14;**录用日期**:2025-01-04;**网络首发日期**:2025-07-07
网络首发地址:https://link.cnki.net/urlid/23.1235.t.20250707.1218.002
基金项目:国家自然科学基金(62263031);新疆维吾尔自治区自然科学基金(2022D01C53)
作者简介:李世钰(1998—),男,硕士研究生;袁 杰(1975—),男,教授,博士生导师
通信作者:李世钰,lsy534066742@163.com;袁 杰,yuanjie@xju.edu.cn

为模式,这限制了它们在复杂环境中的适用性。

早期定位气源主要是依靠人工或静态传感器节点来完成的[2],但传统依赖于警犬的人工搜索方式存在一些局限性,比如会损害动物及其人类操作者的安全。而相较于静态传感器网络的方法,机器人解决方案需要更少的传感器节点。因此,利用机器人进行气源定位有以下优势:能够根据不同环境条件进行自我调整,从而在执行长期任务时无需休息,显示出更高的灵活性和效率[3]。

迄今为止,机器人气源定位(odor source localization, OSL)的研究已经发展出多种不同的算法,主要可分为以下:梯度爬升算法、仿生算法、基于概率的算法和机器学习的方法[4]。梯度爬升算法[5-7]作为早期的研究成果,通过机器人追踪气味浓度梯度来定位气源,这种方法简单且直观,但由于气流的湍流特性,使得气味路径不是平滑的,常导致机器人在搜索区域内徘徊,难以快速定位气源。仿生算法[8-10]则是受到生物寻找食物或伴侣行为的启发,设计了一系列简单且计算成本低的搜索策略。尽管这些算法易于实现,但它们在搜索效率上通常不如预期。基于概率的算法[11-13]通过将气源位置建模为概率分布,并通过对搜索区域内不同位置的连续观测来迭代更新位置估计。这类算法在性能上通常优于梯度爬升和仿生算法,但实现难度大且计算成本较高。机器学习的方法[14-17]通过学习历史数据,可以更好地利用环境信息和气流模式,从而提高气源定位的准确性和效率。

近年来,深度强化学习(deep reinforcement learning, DRL)为应对 OSL 问题提供了新的视角。OSL 本质上是一个顺序决策问题[18],机器人需要在每一步采取适当的行动,以最小的成本帮助机器人找到气源。作为机器学习的典型范式和方法之一,DRL 算法可用于解决顺序决策问题[19]。Loisy 等[20]提出一种近似方法解决了嗅觉搜索任务中的大型部分可观察马尔可夫决策过程(partially observable Markov decision process, POMDP),通过智能体学习策略来指导其行为,以最大化累积奖励或通过与环境的交互来实现特定目标。Alagha 等[21]提出了两种用于复杂环境中目标定位的多智能体深度强化学习(multiagent deep reinforcement learning, MDRL)模型,使用卷积神经网络(convolutional neural network, CNN)对 PPO(proximal policy optimization)算法的 Actor-Critic 结构进行优化学习,该模型在机器人定位时间和成本方面具有优势。Li 等[22]将门控循环单元网络应用于 PPO 算法的 Actor-Critic 框架,从历史数据中提取时间特征,并以端到端方式生成最优决策,提高了源定位的成功率。因此,上述研究表明,PPO 算法在执行 OSL 任务时具有较好的稳定性和适应性。然而,PPO 算法中的值函数网络在面对高度动态和不确定的羽流分布环境时,可能因为过高估计偏差而导致策略更新的效率和效果不佳。为了解决这一问题,本文提出了增强型的近端策略优化(auxiliary value and wind-guided proximal policy optimization, AVW-PPO)算法,引入一个辅助价值网络,旨在降低值函数网络的预测偏差,提供更准确地价值估计,帮助算法更好地调整策略。特别地,现有的 DRL 气源定位研究大多依赖于羽流浓度梯度,且风场信息多被视为影响气味空间分布的一个因素,鲜少将环境风信息集成到机器人系统中进行实际应用。基于此,本文设计了一种风导向策略用来提高机器人的 OSL 效率。在气源扩散的初期阶段,机器人可能远离源,需要较长时间才能捕捉到气味信号。为了加快 OSL 过程,在机器人的状态空间中增加风向信息,同时结合风速信息引入新的奖励机制,使得机器人逆风方向移动,以鼓励机器人在泄漏初期迅速靠近高浓度区域。

1 问题建模与设计

OSL 任务通常被视为羽流追踪与源声明的过程,可细分为 3 个子任务:羽流发现、羽流追踪和气源定位。羽流发现阶段是在环境中搜索以便发现气味羽流;羽流追踪阶段则是沿着气味羽流的方向进行搜索;气源定位就是到达气源附近并确定其确切位置。由于羽流的复杂性和分散性,本文将羽流发现和羽流追踪这两个阶段考虑作为一个任务进行实现,在定位气源时,规定在机器人到达源附近(机器人最小步长范围内)时即为成功。

1.1 气源模型

在气源定位的研究中,准确模拟室内空气动力学对于开发和验证新算法至关重要。本文采用 Fluent 软件构建二维湍流环境气源扩散模型,模拟环境为一个简化的室内二维空间,尺寸为 10 m × 10 m,固体障碍物可在模型中设置。通风条件假设为单一进风口和出风口,分别位于室内两端,以模拟室内湍流环境。合理改变进风口的风速,确保有足够的空气流动以模拟真实情况。选择 Fluent 的压力-速度耦合求解器,设置适当的时间步长(1 s/步)以及迭代次数(每个时间步 20 次迭代)。由于气体发生泄漏时,需要机器人能够快速定位气源,长泄漏时间的研究没有意义。因此将整个模拟过程设定持

续 120 s。气体(本文采用 CO 替代)在 $t=0$ s 时刻开始释放,同时送风口开始送风,在此期间监控各位置的气体质量分数和风速变化,用于后续的数据分析和算法验证。部分时间下的气体扩散过程见图 1。

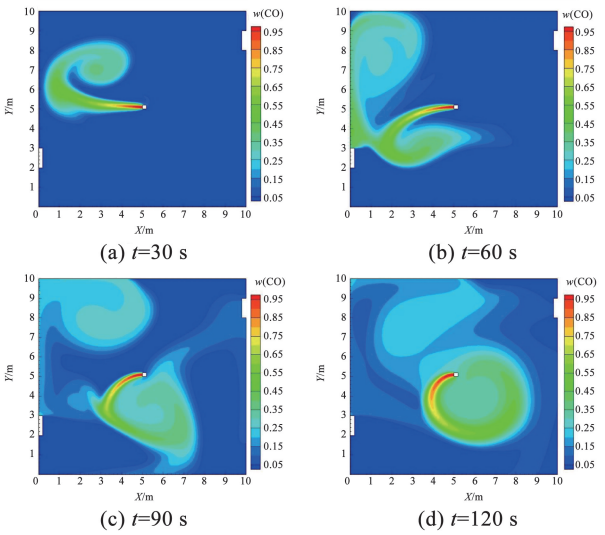


图 1　不同时间下 CO 的质量分数分布图

Fig. 1　CO concentration distribution maps at different time points

为确保模拟结果的稳定性和可靠性,本文采用了足够高的网格分辨率[23]。过细的网格可提高模型精度,但会大幅增加网格数量和计算成本;相反,较粗的网格可降低计算时间,但可能牺牲模型准确性。因此,通过比较不同网格分辨率下泄漏口的平均质量分数,确定了合适的网格数目。经测试,0.05 m 的网格分辨率可兼顾精度和计算复杂度,最终得到 40 240 个网格单元。这一网格设置不仅保证了模型的精度,也优化了计算成本,为进一步的扩散分析和气源定位提供了可靠的数值模拟基础。具体参数设置见表 1。

1.2　构建马尔可夫决策过程

由于气体泄漏和羽流环境的不确定性,气体的扩散受风速、风向、环境温度、障碍物等多种因素影响,这些因素的变化使得环境的状态转移概率难以精确预测。在气源定位任务中,机器人需要根据当前的感知信息(如气体质量分数、风向传感器数据等)实时做出决策,以最有效地逼近气源。基于此,本文将气源定位问题建模为具有未知转移概率的马尔可夫决策过程(Markov decision process, MDP)。MDP 提供了一个框架,允许在每个状态下基于当前信息做出最优决策。这种方法解决了影响气体扩散的环境条件的固有不确定性和动态复杂性,利用强化学习技术,所提出的模型使机器人能够自适应地改进其策略,以实现高效的源定位。

表 1　Fluent 参数设置及边界条件

Tab. 1　Fluent parameter settings and boundary conditions

项目	设置
计算域/m	10×10
网格数	40 240
湍流模型	标准 $k-\varepsilon$ 模型
速度-压力耦合	压力耦合方程组半隐式方法
对流项离散格式	二阶迎风格式
扩散项离散格式	二阶迎风格式
近壁面处理	标准壁面函数[24]
残差	1×10^{-6}
进风口	速度入口
污染源	速度入口
出风口	自由出流
墙壁	无滑移壁面[24]
障碍物	无滑移壁面[24]

典型的 MDP 包含以下 5 个要素:状态空间(S)、动作空间(A)、状态转移概率(P)、奖励函数(R)以及折扣因子(γ)。在每个时间步,环境处于某个状态 $s_t\in S$,机器人采取一个动作 $a_t\in A$ 与环境交互,在执行动作 a_t 后,环境以概率 $P(s_{t+1}\mid s_t,a_t)$ 转移到下一个状态 s_{t+1},机器人随后获得一个奖励 $r(t)=R(s,a)$,该奖励是对当前事件的数值评估。这个过程重复进行,直到触发特定的终止条件。机器人的目标是通过调整策略,以最大化期望累积奖励 $E[\sum_{t=0}^{\infty}\gamma^t r_t]$,其中 r_t 为时间步 t 获得的奖励,折扣因子 γ^t 决定了即时奖励相对于远期奖励的重要性。

1.2.1　状态空间

在湍流 OSL 任务中,气体浓度分布会因释放速率的变化和湍流的影响而呈现动态变化。考虑到环境状态信息随时间变化,地面移动机器人主要通过传感器来探测其周围有限的环境信息。因此,本文定义时刻 $t(t\in[0,120]$ s)的状态空间包括机器人的当前位置的风向信息及其周围 8 个方向网格单元的实际位置和浓度信息。若机器人在气体泄漏后的 120 s 内未能成功定位到气源,则认为定位失败。将机器人当前位置的风向信息纳入机器人的状态空间,有助于机器人更有效地应对环境的动态变化,从而提高定位的准确性和效率。假设机器人 t 时刻所处位置的物质的量浓度为 $c_{(x,y)}$,风向为 $d_{(x,y)}$,则机器人的部分状态空间表示为

$$s_t=\begin{pmatrix} c_{(x-1,y+1)} & c_{(x,y+1)} & c_{(x+1,y+1)} \\ c_{(x-1,y)} & d_{(x,y)} & c_{(x+1,y)} \\ c_{(x-1,y-1)} & c_{(x,y-1)} & c_{(x+1,y-1)} \end{pmatrix} \tag{1}$$

1.2.2 动作空间

动作空间定义了机器人可以执行的所有可能动作。在本研究中,动作空间 A 包含 8 个可选动作 $\{\uparrow,\downarrow,\leftarrow,\rightarrow,\nwarrow,\swarrow,\searrow,\nearrow\}$,即向北、南、西、东 4 个正向及西北、西南、东南和东北移动。这样的动作空间设计使得机器人拥有更多潜在的动作方向,灵活应对复杂的环境变化。将 Fluent 建立的 100 m^2 的室内环境定义为一个二维网格空间[25],每个网格单元表示一个固定的空间位置,机器人在这些网格单元之间移动。机器人在网格单元内的位置见图 2。

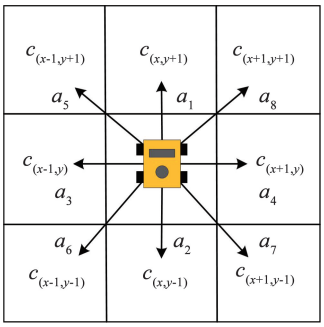


图 2 机器人在网格单元的状态与动作表示

Fig. 2 Representation of robot states and actions in grid cells

机器人可选择 8 个动作之一进入周围的网格单元,而动作的目的单元就是 t 时刻下的浓度位置,因此动作空间可表示为

$$a_t = \{c_{(x,y+1)}, c_{(x,y-1)}, \cdots, c_{(x+1,y+1)}\} \tag{2}$$

1.2.3 奖励机制

奖励机制是强化学习中用于指导机器人学习最优策略的关键因素。在本文中,奖励机制基于机器人对气源的接近程度来设计。考虑到在气体泄漏初期,机器人可能无法及时捕捉到气味信息,因此需要更加有效的探索策略。具体设计原理见图 3。

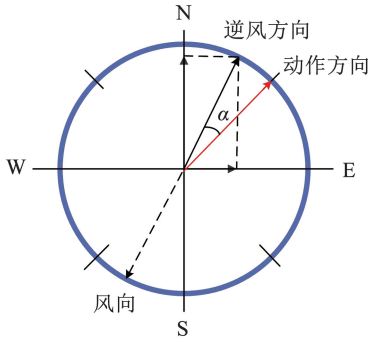


图 3 逆风方向奖励设计图

Fig. 3 Upwind direction reward design chart

机器人由 t 时刻到下一个 $t+1$ 时刻的过程中,环境信息会变更,通过传感器可以感知 8 个方向上的浓度差,使得机器人向附近浓度最大位置移动。考虑到风速与风向影响,具体奖励设计如下:

$$R_1 = \begin{cases} \omega\cos\alpha + \rho_{\max}, & \text{if } \rho_{\max} < 0.1 \\ \rho_{\max} - \rho_{\min}, & \text{otherwise} \\ 2(\rho_{\max} - 0.2) & \text{if } \rho_{\max} > 0.2 \end{cases} \tag{3}$$

$$R_2 = \begin{cases} -0.5, & \text{步数惩罚} \\ -100.0, & \text{碰撞惩罚} \\ +100.0, & \text{目标奖励} \end{cases} \tag{4}$$

$$R = \sum_k R_1^k + R_2^k \tag{5}$$

式中:ω 为风速,α 为逆风方向与机器人移动方向的夹角,$\rho_{\max}$、$\rho_{\min}$ 分别为机器人周围 8 个方向上的浓度最大值与最小值,R_1 为机器人在接近目标中的过程奖励,当 CO 质量分数小于一定的阈值 0.1 时,给予逆风与周围最大质量分数的双重奖励,超过阈值则采用质量分数差值奖励,当机器人越过 0.2 时,放大奖励倍数鼓励机器人向更大浓度方向前进;R_2 为机器人触发特定事件的奖惩,为避免盲目搜索,会给予每步 0.5 的惩罚,当接触到障碍物或者墙壁时会有 100.0 的负奖励,搜寻到源头位置时视为完成目标,赋予 100.0 的正奖励;R 为机器人移动 k 步的总奖励。通过这种设计,一方面可以克服奖励的高稀疏性,另一方面使得 AVW-PPO 算法能够有效地利用环境信息,优化机器人的策略,使其在泄漏初期快速找到高浓度区域,提高气源定位效率。

2 算法设计

2.1 原始 PPO 算法

PPO 算法是一种在策略空间中进行优化的方法,它提供一个裁剪的代理目标函数来减少策略更新后与原策略的偏差[26]。该代理目标函数定义为

$$L^{\text{clip}}(\theta) = \hat{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{6}$$

式中:$r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ 为动作概率比,$\hat{A}_t$ 为状态 - 动作对的优势函数估计,ϵ 为剪裁比例超参数。PPO 算法的核心是利用这个裁剪的概率比率来限制策略更新步骤中的变化幅度,从而避免过大的策略更新导致性能下降,实现稳定学习。

2.2 AVW-PPO 算法

尽管 PPO 在许多强化学习任务中表现出色,但在处理复杂状态空间或需要快速适应新环境的任务时,其性能仍有待提升。值函数估计作为强化学习的一个关键部分,对算法的策略更新有着很大的影响,而单一的价值网络可能会受到噪声和不稳定性等因素干扰,从而导致在值函数预测时可能存在较大的估计偏差。辅助价值网络的加入,主要用于为每个状态 - 动作对提供不同的值估计,进而选择偏

差较小的网络输出更准确的值估计,以此优化策略更新,提高学习的稳定性与鲁棒性。

在 AVW-PPO 算法中,由于数据在采样阶段不会更新网络参数,主价值网络和辅助价值网络可以分别对收集到的数据轨迹进行价值预测,选择值估计较小的网络作为值函数输出,即

$$V_\theta(s_t)=\min\{V_{\theta_1}(s_t),V_{\theta_2}(s_t)\} \tag{7}$$

式中:$V_{\theta_1}(s_t)$为主价值网络输出的值,$V_{\theta_2}(s_t)$为辅助价值网络输出的值。R_t 为作为算法采用的总回报,表示对 t 时刻后的未来奖励折扣求和。利用优势估计 $\hat{A}_t(s_t,a_t)$,回报与优势函数分别定义为:

$$R_t=\sum_{k=0}^{T-t-1}\gamma^k r_{t+k}+\gamma^{T-t}V_\theta(s_t) \tag{8}$$

$$\begin{cases}\hat{A}_t(s_t,a_t)=\sum_{k=t}^{T-1}(\gamma\lambda)^{k-t}\delta_k\\ \delta_t=r_t+\gamma V_\theta(s_{t+1})-V_\theta(s_t)\end{cases} \tag{9}$$

式中:δ_t 为时序差分误差,γ、λ 分别为折扣因子和衰减因子,$V_\theta(s_t)$为估计偏差较小的网络输出的状态价值。两个价值网络通过最小化损失函数更新各自网络,其目标函数可分别定义为:

$$L_{\text{critic}}(\theta_1)=E_t[(R_t-V_{\theta_1}(s_t))^2] \tag{10}$$

$$L_{\text{aux}}(\theta_2)=E_t[(R_t-V_{\theta_2}(s_t))^2] \tag{11}$$

算法中加入了熵正则化项,用于鼓励探索,防止过早收敛到次优策略,即

$$S_e[\pi_\theta](s_t)=-\sum[\pi_\theta(a_t|s_t)\log\pi_\theta(a_t|s_t)] \tag{12}$$

通过最大化目标函数的近似值来更新优化参数 θ 为

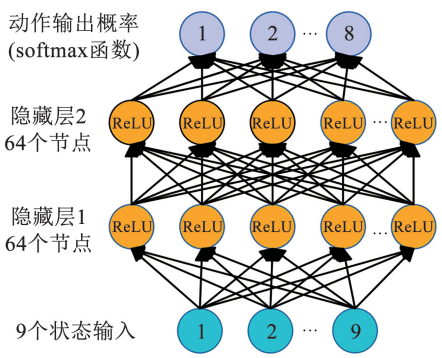
$$\arg\max_\theta L(\theta_k)=L^{\text{clip}}(\theta_k)-c_1L^{\text{vf}}(\theta)+c_2E_t[S_e[\pi_\theta](s_t)] \tag{13}$$

式中:c_1、c_2 为常系数,用来调整网络目标函数中的各部分权重;$L^{\text{vf}}(\theta)$为 $L_{\text{critic}}(\theta_1)$与 $L_{\text{aux}}(\theta_2)$中值估计较小的目标函数。

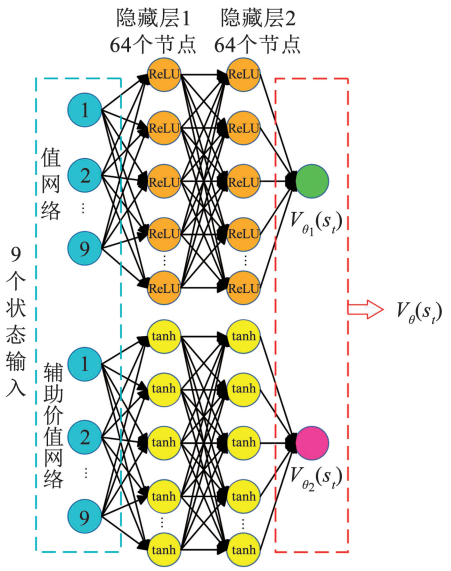
2.3　算法网络模型与训练

在 PPO 算法中,网络基于 Actor-Critic 框架,由策略网络(Actor)和值函数网络(Critic)构成,两者都是深度神经网络[27]。AVW-PPO 算法的网络架构见图 4。AVW-PPO 算法结合了 PPO 中梯度策略和价值函数的优势,通过引入辅助价值网络,使用三重网络结构来分别估计策略和价值。具体而言,策略网络接收 9 个状态输入(包括机器人周围 8 个方向上的浓度信息和当前时刻的风向信息),通过两个隐藏层,每层 64 个节点,并使用 ReLU 激活函数,最终通过 softmax 函数输出动作的概率分布,其决定了机器人在 8 个方向上的移动概率。两个值函数网络在结构上保持一致,但激活函数不同,主价值网络使用 ReLU 激活函数,而辅助价值网络则采用 tanh 激活函数,由此两者能够捕捉不同的特征和模式。为提高价值估计的精度,比较两个值网络预测的偏差,选择值估计较小的网络作为最终的值估计输出,确保结果的稳定性和可靠性。

算法还通过状态、奖励归一化、经验回放机制等一系列技巧来优化训练过程和提高算法的稳定性及效率。基于 AVW-PPO 算法的机器人 OSL 导航决策的整体结构见图 5。



(a) 策略网络



(b) 双价值网络

图 4　AVW-PPO 算法网络架构

Fig. 4　Network architecture of AVW-PPO algorithm

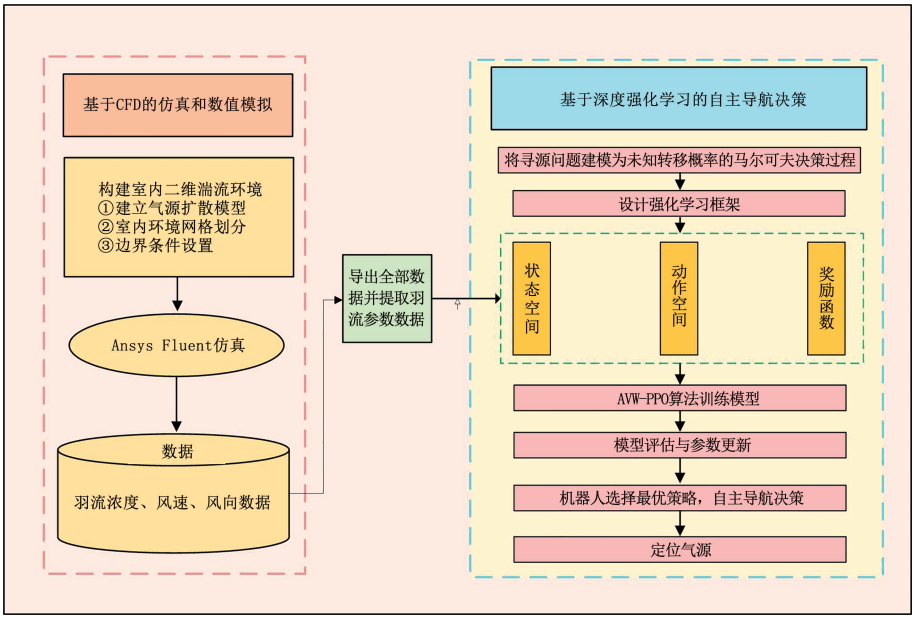


图 5 基于 AVW-PPO 算法的机器人 OSL 导航决策

Fig. 5 Robot OSL navigation decision based on the AVW-PPO algorithm

AVW-PPO 算法训练过程伪代码见算法 1。

算法 1 AVW-PPO 算法训练过程

输入：初始策略网络参数 θ

输出：优化的策略参数 θ_{k+1}

for $i = 1,2,\cdots$,do

  for $n = 1,2,\cdots,N$ do

  初始化状态 s_0

    for $t = 0,1,2,\cdots,T$ do

      输入状态 s_t 到策略网络，根据 π_θ 选择动作 a_t

      计算奖励 r_t 更新到下一个状态 s_{t+1}

      存储 s_t, a_t, r_t, s_{t+1}, done, $\log \pi_\theta(a_t|s_t)$ 到经验池

    end for

    对于 $t = 0,1,2,\cdots,T$，选择 $V_\theta(s_t)$，计算 R_t 与 $\hat{A}_t$，利用 $L_{\text{critic}}(\theta_1)$ 和 $L_{\text{aux}}(\theta_2)$ 更新两个价值网络

  end for

  for $k = 0,1,2,\cdots,K$ do

    更新策略参数 $\theta_{k+1} \leftarrow \underset{\theta}{\operatorname{argmax}} L(\theta_k)$

  end for

end for

3 结果与分析

在本文中，设置不同的模拟环境，通过改变机器人的初始位置口的风速（W）来评估算法性能及策略的有效性。采用两个指标评估 AVW-PPO 算法的性能：1）成功率。反映机器人在一定实验次数内到达气源的成功次数。2）平均搜索步数。反映算法效率。实验设备采用 13th Gen Intel(R) Core(TM) i5 - 13490 F/2. 50 GHz, NVIDIA RTX 4060, 16. 0 GB RAM，基于 Windows 系统下 Python 3.9 平台实现。

首先，在模拟的 100 m^2 区域内进行 10 000 次迭代训练。设置气源位置为（5 m，5 m）处，释放速率为 0.5 m/s，进风口风速为 0.3 m/s，机器人在正向移动与对角线移动时步长固定为 0.20、0.28 m，响应时间分别为 2.0、2.8 s。机器人的响应时间定义为从开始检测到成功定位气源的时间，本文旨在实现气源的快速与准确定位。因此从气源扩散开始，机器人就需要尽快感知羽流并定位气源位置，待羽流分布稳定时再去定位气源没有实际意义。在每次迭代中，机器人的初始位置是随机的。在每一集的训练中，机器人根据当前策略选择动作，环境根据动作反馈新的状态和奖励。当机器人达到气源、碰到障碍物或墙壁，移动步数超过 120 步时，该集训练结束。算法的超参数见表 2，超参数由 ElegantRL 强化学习库推荐设置，经一系列实验进行调优确定。

表 2 AVW-PPO 算法超参数

Tab. 2 Hyperparameters of the AVW-PPO algorithm

超参数	值
策略网络学习率	0.001
价值网络学习率	0.001
辅助价值网络学习率	0.001
批量大小	32
隐藏层	64
GAE 折扣因子 γ	0.99
GAE 衰减因子 λ	0.98
clip	0.2
K-epochs	10
价值损失权重	0.50
策略熵权重	0.01

为了证明分别引入辅助价值网络和风策略的有

效性,本文区分成两种类型的算法。1) AV-PPO 算法,它只包含一个辅助价值网络。2) AVW-PPO 算法,包含辅助价值网络与风导向策略。算法收敛过程如图 6 所示,阴影部分为实际训练过程,实线为拟合后的结果。

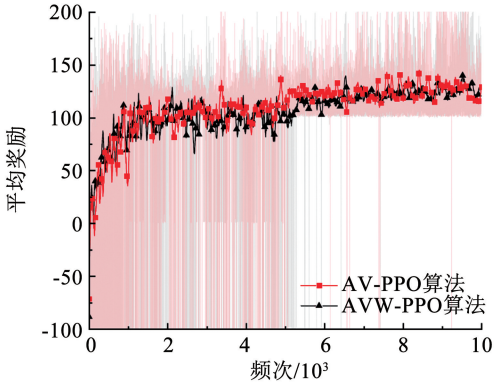


图 6　算法随机初始位置训练过程

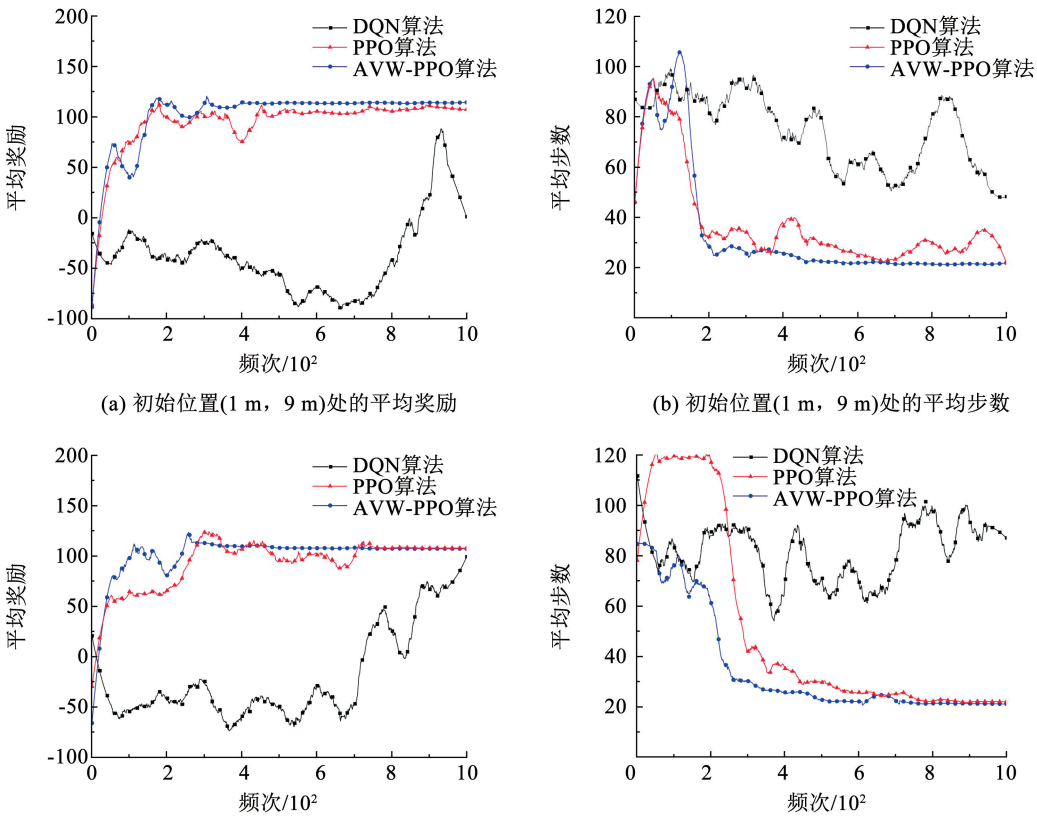
Fig. 6　Training process of random initial positions for the AVW-PPO algorithm

在训练过程中,AV-PPO 算法和 AVW-PPO 算法在初始阶段(前 2 000 次迭代)都表现出相当大的波动,主要因为算法在初期需要平衡探索与利用,并且网络模型也不够精确。随着训练的进行,两种算法的平均奖励逐渐增加。大约在第 4 000 次迭代后,平均奖励的波动显著减少,表明机器人的策略正在收敛。随着训练的进行(约超过 6 000 次迭代),两种算法的平均奖励都会稳定下来。AVW-PPO 算法表现出较少的变化,表明稳定性更高。由于初始放置的随机性,机器人有时可能会从墙壁或障碍物附近开始,即使在收敛后也可能出现定位失败的情况。然而,总体结果表明,AV-PPO 算法和 AVW-PPO 算法在大多数实验中能够成功定位羽流源,验证了所提出的方法和策略的有效性。

3.1　算法性能对比

为验证 AVW-PPO 算法的有效性,在 100 m^2 的无障碍物湍流环境中,保持羽流源位置不变,释放速率为 0.8 m/s,进风口风速为 0.4 m/s。初始位置分别设定在(1 m,9 m)和(9 m,1 m)处。选取原始 PPO 算法、深度 Q 学习网络(deep Q-network,DQN)算法与 AVW-PPO 算法作为对比,观察不同初始位置下算法的性能,得到的实验结果见图 7。

图 7 中描绘了 3 种算法在无障碍环境中的收敛过程。可以看出 DQN 算法在 1 000 次迭代结束时才开始有收敛趋势,且平均搜索步数与成功率表现较差,说明 DQN 算法在迭代过程中没有找到较好的策略指导机器人定位气源,搜索性能上远不如其他两种算法。



(a) 初始位置(1 m，9 m)处的平均奖励　(b) 初始位置(1 m，9 m)处的平均步数

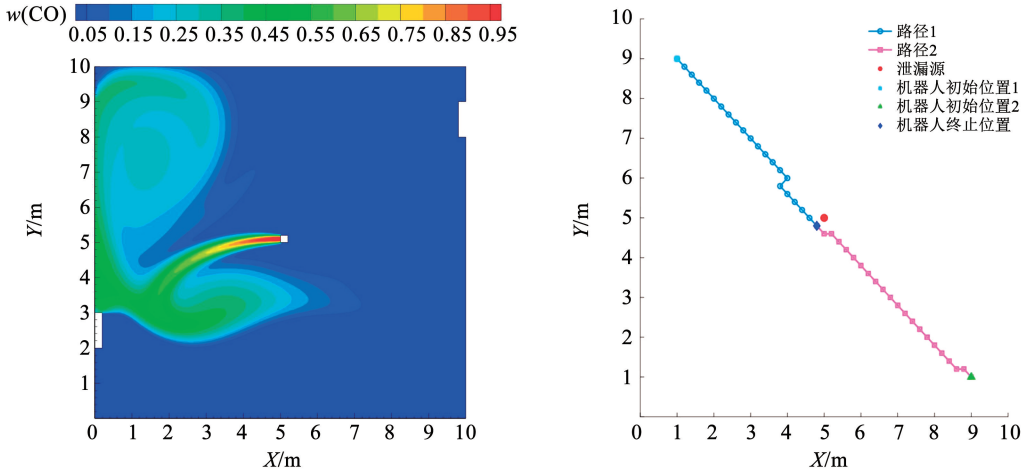
(c) 初始位置(9 m，1 m)处的平均奖励　(d) 初始位置(9 m，1 m)处的平均步数

图 7　各算法迭代过程图

Fig. 7　Iterative process for each algorithm

无障碍环境中,机器人在 AVW-PPO 算法的最优策略下,从两个初始位置到源头的最少搜索步数均为 21 步,响应时间均在 58 s 左右,机器人最优策略下的搜索过程可以用图 8 描绘。

从图 7、8 可知,在(1 m,9 m)位置,机器人能够较早得感知到羽流的浓度变化,是因为气源泄漏时羽流首先扩散到该位置附近,所以机器人在搜索初期就能获得较高的浓度信息,从而能够迅速调整策略,朝向浓度更高的区域移动,获得更高的奖励。而在(9 m,1 m)处,该位置距离气源较远,羽流扩散到该区域需要一定时间,机器人刚开始未能捕捉到羽流,说明在初期需要进行更多的探索,以捕捉羽流的方向和浓度梯度。尽管(1 m,9 m)附近的浓度较高,但在气源附近,浓度变化不明显,意味着机器人在接近气源时,浓度梯度较小,可能会导致策略选择的困难。但经过 400 次迭代后,AVW-PPO 算法已经能够学习到有效策略并趋于稳定,表明 AVW-PPO 算法在探索过程中能够有效利用环境信息,逐步优化策略。相比之下,PPO 算法需要更多的迭代次数才基本收敛。在(9 m,1 m)处,由于羽流扩散到此位置的过程中,羽流浓度梯度变化明显,所以 AVW-PPO 算法迭代不到 300 次就能稳定收敛,而 PPO 算法历经 700 次迭代后才趋于稳定。这表明 AVW-PPO 算法在策略学习和优化方面具有更高的效率。采取 500 次实验得到的计算结果见表 3。



(a) t=58 s时CO的质量分数羽流分布　　(b) 无障碍环境中最优策略下的搜索路径

图 8　羽流分布和机器人的搜索过程

Fig. 8　Plume distribution and robot search path

表 3　不同初始位置下 3 种算法搜索性能

Tab. 3　Search performance of three algorithms at different initial positions

初始位置	算法	平均奖励	平均搜索步数	成功率/%
(1 m,9 m)	DQN	−38.14	63.62	16.2
	PPO	106.44	27.44	99.4
	AVW-PPO	**113.71**	**21.68**	**100.0**
(9 m,1 m)	DQN	4.66	82.10	14.8
	PPO	102.97	24.15	97.2
	AVW-PPO	**107.47**	**21.98**	**100.0**

从评价指标来看,AVW-PPO 算法在两个初始位置定位气源的成功率均为 100%,对于 PPO 算法分别提高了 1.6% 和 2.8%,相比 DQN 算法提高了 83.8% 和 85.2%。此外,AVW-PPO 算法相对于其他两种算法有更低的平均搜索步数,相应的机器人的气源定位时间也会更短。虽然平均奖励一般不作为气源定位中的性能指标,但从图 8 中可以看出其在一定程度上反映了算法的稳定性。不论是哪个初始位置,AVW-PPO 算法都有着相对较高的平均奖励,说明算法受风导向策略的影响,在提高搜索效率的同时也能够保持策略的稳定性。

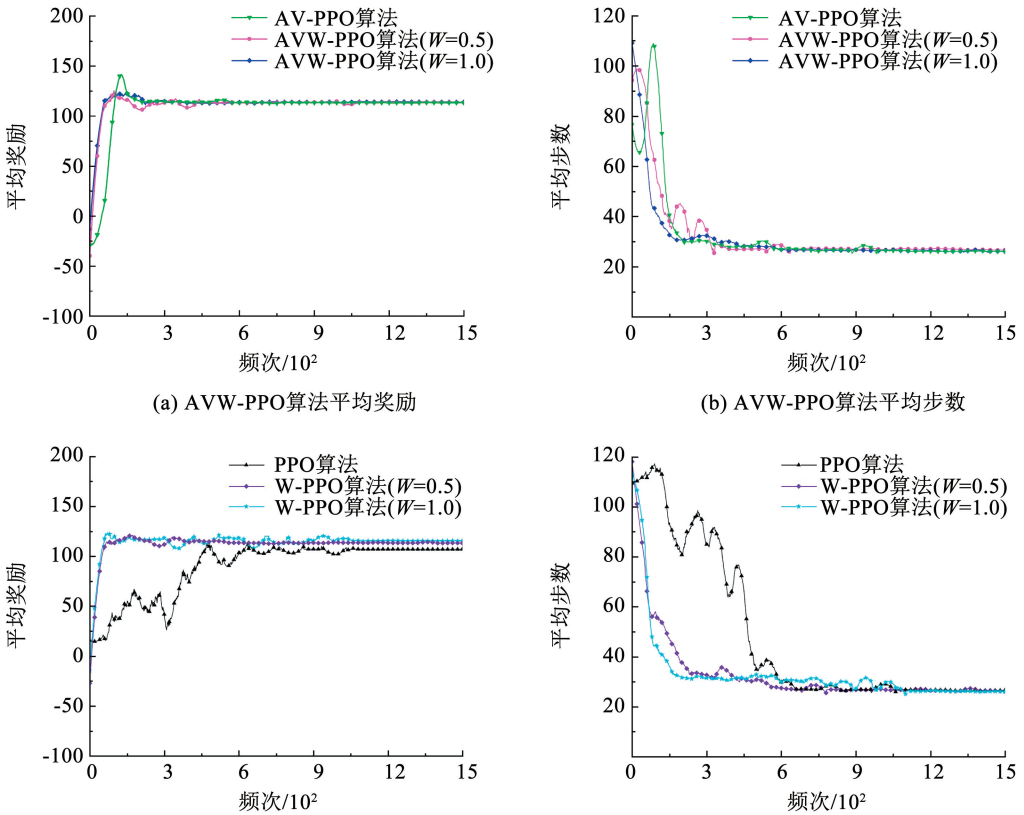
3.2　风导向策略测试

在有障碍物环境中,比较不同风速对算法策略的影响。羽流源位置与无障碍物环境中保持一致,源释放速率依然为 0.8 m/s,为保证策略的可行性,将机器人初始位置放在下风处,进风口速度分别设置为 0.5 m/s(W=0.5)及 1.0 m/s(W=1.0)。在不同风速下得到的实验结果见图 9,机器人的搜索过程可以用图 10 描绘。

第 2 组实验主要测试风导向策略的有效性,同时为验证策略的普适性,针对原始 PPO 算法也进行测试。因此,本组实验主要是用 W 风速值区分算法是否加入风信息。从图 9(a)、(b)观察到只优化网络结构的 AV-PPO 算法与 AVW-PPO 算法表现差异不大,且从表中也可得出,两者的平均搜索步数与定位成功率均有良好表现,说明算法在辅助网络的集

成下更能准确估计,学习到策略的稳定性。但平均奖励方面 AVW-PPO 算法($W=1.0$)略高,表明在风速相对较高时,算法能够更好地利用风信息,提高奖励值。横向对比来看,AVW-PPO 算法相较于 PPO 算法或是加入风策略的 W-PPO 算法均有更好的性能表现。

图 9(c)、(d)显示出风导向策略在 PPO 算法的效果更为明显。PPO 算法在训练初期的平均奖励值波动较大,且收敛速度较慢,迭代 700 次后才基本收敛,而结合风信息的 W-PPO 算法能够在羽流扩散初期指导机器人通过逆风迅速步入浓度较大区域,进而学习到有效策略,展示出了更快地收敛速度与稳定性。综合来看,风信息和辅助价值网络的结合,使得 AVW-PPO 算法在复杂环境中表现出色,能够更有效地引导机器人进行搜索,提高策略学习的效率和稳定性。

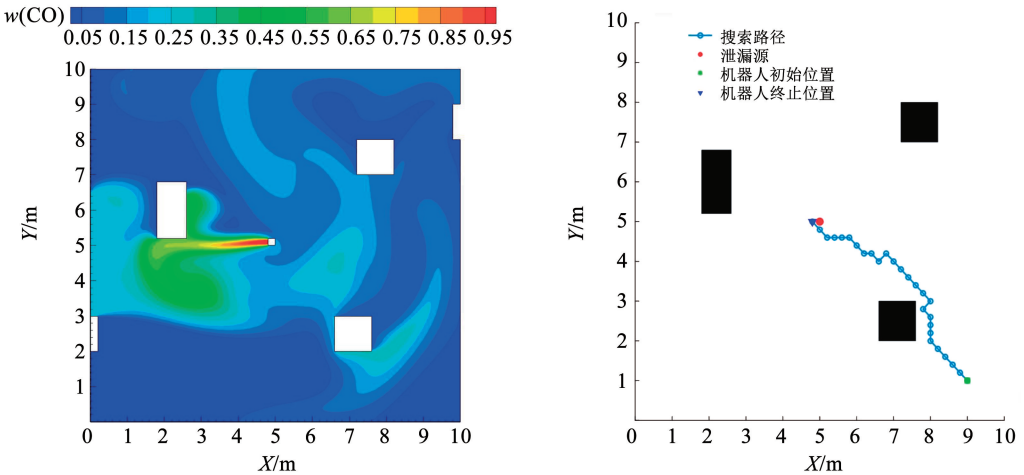


(a) AVW-PPO算法平均奖励　(b) AVW-PPO算法平均步数

(c) PPO算法平均奖励　(d) PPO算法平均步数

图 9　不同风速下算法迭代结果

Fig. 9　Algorithm iteration results under different wind speeds



(a) t=67 s时CO的质量分数羽流分布　(b) 障碍环境中最优策略下的搜索路径

图 10　羽流分布和机器人的搜索过程

Fig. 10　Plume distribution and robot search path

经过实验得出,机器人在最优策略下从初始位置到源头的最少搜索步数均为 26 步,响应时间均为 67.2 s,采取 800 次实验得到的计算结果见表 4。从表 4 中可以看出,风导向策略不仅有效减少了算法的平均搜索步数,还在一定程度上提高了 OSL 的成功率。因为该策略帮助机器人进一步优化了决策过程,这在 PPO 算法及改进算法中均有所体现。与原 PPO 算法相比,AVW-PPO 算法在平均搜索步数上最大减少了 2.89 步,有效提升了搜索效率,并在成功率上实现了最高 3.1% 的提升。

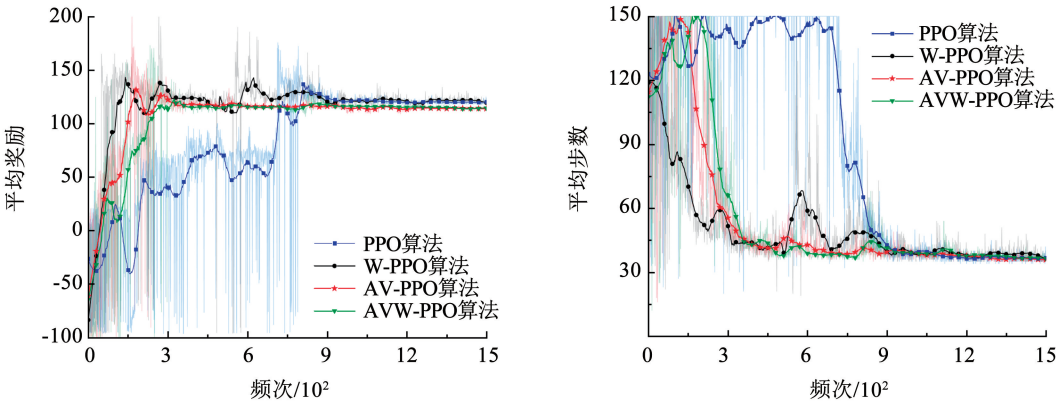
表 4 不同风速障碍环境下算法的搜索性能

Tab. 4 Search performance of algorithms in an obstructed environment with different wind speeds

算法	平均奖励	平均搜索步数	成功率/%	算法	平均奖励	平均搜索步数	成功率/%
PPO	105.40	28.93	96.9	AV-PPO	113.36	27.34	99.9
W-PPO($W=0.5$)	113.64	27.04	99.3	**AVW-PPO($W=0.5$)**	**113.57**	**26.56**	**100.0**
W-PPO($W=1.0$)	115.79	27.40	99.6	**AVW-PPO($W=1.0$)**	**113.96**	**26.04**	**100.0**

3.3 不同环境下算法性能分析

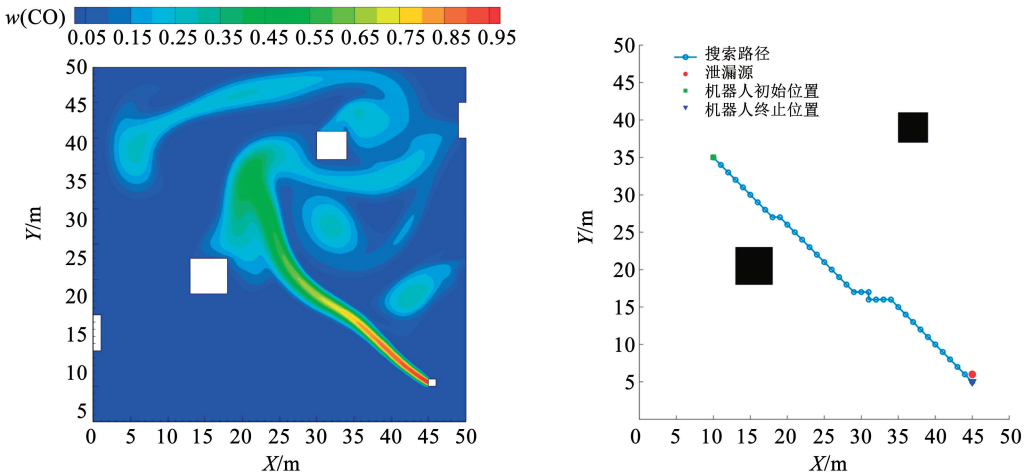
在实际应用中,机器人往往需要在大规模和复杂的环境中执行任务,因此验证算法及策略在大场景下的适用性至关重要。为全面地评估算法在不同环境条件下的性能,本组实验设置在有障碍物大场景下进行。将实验区域扩大到 50 m × 50 m,机器人的搜索步长相应地扩大 5 倍,即每步 1 m。羽流源初始位置设定为(45 m,6 m)处,增加源释放速率到 2 m/s,设置两个尺寸分别为 5 m 和 4 m 的障碍物,依然采用一进风口和一出风口,进风口风速设定为 1 m/s,其余各模型参数不变。得到的实验结果见图 11。在大场景环境中,机器人在最优策略下从初始位置到源头的最少步数为 38 步,响应时间约为 100 s,机器人的搜索过程可以用图 12 描绘。采取 1 000次实验计算得到的结果见表 5。



(a) 不同算法平均奖励 (b) 不同算法平均步数

图 11 不同算法迭代结果图

Fig. 11 Iteration results for different algorithms



(a) t=100 s时CO的质量分数羽流分布 (b) 大场景环境中最优策略下的搜索路径

图 12 羽流分布和机器人的搜索过程

Fig. 12 Plume distribution and robot search path

表 5　大场景环境下算法的搜索性能

Tab. 5　Search performance of algorithms in large scenario environments

算法	平均奖励	平均搜索步数	成功率/%
PPO	119.96	46.65	92.75
W-PPO	122.78	40.50	97.13
AV-PPO	114.97	38.79	98.75
AVW-PPO	**115.66**	**38.11**	**100.00**

观察图表,AV-PPO 算法与结合风信息的 AVW-PPO 算法在整个训练过程中表现都很出色,平均奖励值迅速上升并稳定在较高水平。AVW-PPO 算法虽然在训练初期的波动较大,但其收敛速度和稳定性均优于 PPO 算法和 W-PPO 算法。这表明,AVW-PPO 算法在没有风信息的情况下也能显著提升算法性能。

从迭代次数上发现,AVW-PPO 算法在 15 次迭代时就能够成功定位羽流源,机器人通过学习经验能够快速收敛,而 AV-PPO 算法在没有风信息的辅助时,迭代 200 次左右时才第 1 次成功定位气源。在足够大的空间,没有一定的策略指导机器人移动,仅通过浓度梯度不断学习经验,机器人无法快速步入浓度较高区域,因此局部风信息的利用可以更好地提升算法性能。相比之下,PPO 算法和 W-PPO 算法的平均奖励值在训练初期波动较大,且收敛速度较慢。虽然加入风信息后 PPO 算法也能够快速定位到羽流源,但由于算法本身缺乏对环境信息的精确估计,因此收敛稳定性较差,表明 PPO 算法在复杂环境下的效果不如 AVW-PPO 算法。

值得注意的是,W-PPO 算法相比于 PPO 算法提高了 4.38% 的气源定位成功率,减少了 6.15 的平均搜索步数,充分表明风导向策略在大环境中效果更为显著。而 AVW-PPO 算法相比于 PPO、W-PPO、AV-PPO 算法在平均搜索步数上分别缩短了 18.31%、5.90%、1.75%,在成功率方面分别提升了 7.25%、2.87%、1.25%,表明辅助价值网络与风导向策略均对算法有着不可忽视的影响。

4　结　论

1)本文提出了一种基于深度强化学习的 AVW-PPO 算法,旨在解决室内复杂环境下机器人气源定位效率低下和成功率不佳的问题。通过引入辅助价值网络,有效降低了策略训练过程中的估计偏差,加速了策略优化和模型收敛。算法充分结合主价值网络与辅助价值网络的优势,使模型在早期阶段即可获得更加精准、稳定的价值估计,为机器人高效执行 OSL 任务提供了重要支撑。

2)为提升机器人的 OSL 效率,设计了一种融合局部风信息的风导向策略。该策略将局部风场信息嵌入到机器人的状态空间和奖励函数中,增强了算法对局部环境的感知能力,使机器人能够更准确地感知气体扩散趋势。该策略整体改善了机器人的导航性能,尤其在大规模、复杂环境中展现了优异的定位效率及成功率。这种设计充分利用了环境中的额外信息,克服了仅依赖浓度信息的局限性,为提高算法的整体性能提供了新的思路。

3)在 3 种不同的湍流环境中对所提出的算法进行评估,证明了其在 OSL 的可行性和有效性。实验结果显示,AVW-PPO 算法在多种环境中表现出色,相对于同类算法有着更少的平均搜索步数,且气源定位成功率稳定在 99.00% 以上。该算法有效解决了传统方法在动态湍流环境下易受干扰、性能不稳定的局限性,为机器人在室内复杂环境的 OSL 研究提供了参考。

参考文献

[1] JIANG Mingrui, TONG Chengxin, LI Zhenfeng, et al. 3D multi-robot olfaction in naturally ventilated indoor environments: Locating a time-varying source at unknown heights[J]. Science of The Total Environment, 2024, 926: 171939. DOI: 10.1016/j.scitotenv.2024.171939

[2] SUDHAKAR S, VIJAYAKUMAR V, SATHIYA KUMAR C, et al. Unmanned Aerial Vehicle (UAV) based forest fire detection and monitoring for reducing false alarms in forest-fires[J]. Computer Communications, 2020, 149: 1. DOI: 10.1016/j.comcom.2019.10.007

[3] CHEN Xinxing, FU Chenglong, HUANG Jian. A Deep Q-Network for robotic odor/gas source localization: Modeling, measurement and comparative study[J]. Measurement, 2021, 183: 109725. DOI: 10.1016/j.measurement.2021.109725

[4] JING Tao, MENG Qinghao, ISHIDA H. Recent progress and trend of robot odor source localization[J]. IEEJ Transactions on Electrical and Electronic Engineering, 2021, 16(7): 938. DOI: 10.1002/tee.23364

[5] YANG Yibin, FENG Qilin, CAI Hao, et al. Experimental study on three single-robot active olfaction algorithms for locating contaminant sources in indoor environments with no strong airflow[J]. Building and Environment, 2019, 155: 320. DOI: 10.1016/j.buildenv.2019.03.043

[6] 赵攀, 袁杰, 王宏伟, 等. 基于决策树的羽流追踪机器人自主决策方法研究[J]. 计算机工程与应用, 2019, 55(14): 254
ZHAO Pan, YUAN Jie, WANG Hongwei, et al. Research on autonomous decision-making of plume tracking robots using decision tree[J]. Computer Engineering and Applications, 2019, 55(14): 254. DOI: 10.3778/j.issn.1002-8331.1805-0281

[7] JABEEN M, MENG Qinghao, JING Tao, et al. Robot odor source localization in indoor environments based on gradient adaptive extremum seeking search[J]. Building and Environment, 2023, 229: 109983. DOI: 10.1016/j.buildenv.2023.109983

[8] YANG Yibin, ZHANG Boyuan, FENG Qilin, et al. Towards locating time-varying indoor particle sources: Development of two multi-robot olfaction methods based on whale optimization algorithm [J]. Building and Environment, 2019, 166: 106413. DOI: 10.1016/j.buildenv.2019.106413

[9] 缪燕子, 王玥, 李元龙, 等. 融合学习策略与导向果蝇机制的气味源主动定位方法研究[J]. 控制理论与应用, 2023, 40(5): 913
MIAO Yanzi, WANG Yue, LI Yuanlong, et al. Study on active odor source localization method based on learning strategy and guided fruit fly mechanism[J]. Control Theory & Applications, 2023, 40(5): 913. DOI: 10.7641/CTA.2022.11078

[10] SHIGAKI S, YAMADA M, KURABAYASHI D, et al. Robust moth-inspired algorithm for odor source localization using multimodal information[J]. Sensors, 2023, 23(3): 1475. DOI: 10.3390/s23031475

[11] HUTCHINSON M, OH H, CHEN Wenhua. A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors[J]. Information Fusion, 2017, 36: 130. DOI: 10.1016/j.inffus.2016.11.010

[12] 陈欣星, 黄剑. 基于粒子滤波的烟雾羽流路径追踪算法[J]. 华中科技大学学报(自然科学版), 2020, 48(1): 66
CHEN Xinxing, HUANG Jian. Particle filter-based algorithm for smoke plume path tracking[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition), 2020, 48(1): 66. DOI: 10.13245/j.hust.200112

[13] LIU Shiqi, ZHANG Yan, FAN Shurui. Adaptive space-aware infotaxis II as a strategy for odor source localization[J]. Entropy, 2024, 26(4): 302. DOI: 10.3390/e26040302

[14] TRAN V P, GARRATT M A, KASMARIK K, et al. Multi-gas source localization and mapping by flocking robots[J]. Information Fusion, 2023, 91: 665. DOI: 10.1016/j.inffus.2022.11.001

[15] KIM H, PARK M, KIM C W, et al. Source localization for hazardous material release in an outdoor chemical plant via a combination of LSTM-RNN and CFD simulation[J]. Computers &Chemical Engineering, 2019, 125: 476. DOI: 10.1016/j.compchemeng.2019.03.012

[16] MA Shengshan, YUAN Jie, GUO Zhenyu, et al. Autonomous plume Near-Source search assisted by intermittent visible plume information using finite state Machine and YOLOv3-tiny[J]. Expert Systems with Applications, 2023, 228: 120350. DOI: 10.1016/j.eswa.2023.120350

[17] ZHAO Yong, CHEN Bin, WANG Xianghan, et al. A deep reinforcement learning based searching method for source localization[J]. Information Sciences, 2022, 588: 67. DOI: 10.1016/j.ins.2021.12.041

[18] NIU Lvyin, SONG Shiji, YOU Keyou. A plume-tracing strategy via continuous state-action reinforcement learning[C]//2017 Chinese Automation Congress (CAC). Jinan: IEEE, 2017: 759. DOI: 10.1109/cac.2017.8242868

[19] SAMI H, BENTAHAR J, MOURAD A, et al. Graph convolutional recurrent networks for reward shaping in reinforcement learning[J]. Information Sciences, 2022, 608: 63. DOI: 10.1016/j.ins.2022.06.050

[20] LOISY A, HEINONEN R A. Deep reinforcement learning for the olfactory search POMDP: a quantitative benchmark [J]. The European Physical Journal E, Soft Matter, 2023, 46(3): 17. DOI: 10.1140/epje/s10189-023-00277-8

[21] ALAGHA A, MIZOUNI R, BENTAHAR J, et al. Multiagent deep reinforcement learning with demonstration cloning for target localization[J]. IEEE Internet of Things Journal, 2023, 10(15): 13556. DOI: 10.1109/ JIOT.2023.3262663

[22] LI Hui, YUAN Jie, YUAN Hao. An active olfaction approach using deep reinforcement learning for indoor attenuation odor source localization[J]. IEEE Sensors Journal, 2024, 24(9): 14561. DOI: 10.1109/JSEN.2024.3373610

[23] 周晅毅, 王富玉, 杨流阔, 等. 基于 Nelder-Mead 算法的机器人主动嗅觉室内时变污染源定位[J]. 同济大学学报(自然科学版), 2022, 50(6): 812
ZHOU Xuanyi, WANG Fuyu, YANG Liukuo, et al. Locating indoor time-variant contaminant sources based on nelder-mead algorithm using robot active olfaction method[J]. Journal of Tongji University(Natural Science), 2022, 50(6): 812. DOI: 10.11908/j.issn.0253-374x.21481

[24] HANG Jian, LI Yuguo, CHING W. H., et al. Potential airborne transmission between two isolation cubicles through a shared anteroom[J]. Building and Environment, 2015, 89: 264. DOI: 10.1016/j.buildenv.2015.03.004

[25] WANG Lingxiao, PANG Shuo. Robotic odor source localization via adaptive bio-inspired navigation using fuzzy inference methods[J]. Robotics and Autonomous Systems, 2022, 147: 103914. DOI: 10.1016/j.robot.2021.103914

[26] 黄腾飞, 杜永文, 刘帅, 等. 边缘计算中时延敏感的启发式任务卸载方法[J/OL]. [2023-12-14]. https://link.cnki.net/urlid/23.1235.T.20231214.1310.013
HUANG Tengfei, DU Yongwen, LIU Shuai, et al. Latency-sensitive heuristic task offloading method in edge computing [J/OL]. [2023-12-14]. https://link.cnki.net/urlid/23.1235.T.20231214.1310.013

[27] WU Yuanqing, LIAO Siqin, LIU Xiang, et al. Deep reinforcement learning on autonomous driving policy with auxiliary critic network [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(7): 3680. DOI: 10.1109/TNNLS.2021.3116063

(编辑 张 红)