

DOI:10.3969/j.issn.1671-0673. 2026. 03. 002

基于注意力增强 DDPG 的短波协同测向与定位方法

冯祺玥, 唐 涛, 张昀普, 赵排航, 陈晓云

(信息工程大学, 河南 郑州 450001)

摘要:复杂场景下的短波定位在战场通信中发挥着至关重要的作用,而现有基于深度学习的短波智能测向方法主要依赖人工标注且优质短波信号样本较少。针对上述问题,在短波信号空时高分辨时频图基础上,提出了一种基于注意力增强深度确定性策略梯度(DDPG)算法的短波协同测向与定位方法。该方法首先设计了基于空时高分辨时频图的强化学习环境和具有混合动作空间的智能体,再依据估计定位结果设计奖励函数,以此建立可观测马尔可夫决策过程。其次,利用注意力机制增强深度强化学习的多层次非精细化评估,实现自动测向定位,可以减少人工标注,实现边工作边学习的自主进化,逐步提升短波信号的智能测向定位能力。实验结果表明,所提算法在保证定位精度与参数估计算法性能相当的情况下,使得测向定位时间缩短了 77.7%。

关键词:测向定位;马尔可夫决策过程;空时高分辨;混合动作空间;深度确定性策略梯度算法

中图分类号:TN974;TP391.4

文献标识码:A

文章编号:1671-0673(2026)03-0259-08

A Cooperative Direction-Finding and Localization Method for Shortwave Based on Attention-Enhanced DDPG

FENG Qiyue, TANG Tao, ZHANG Yunpu, ZHAO Paihang, CHEN Xiaoyun

(Information Engineering University, Zhengzhou 450001, China)

Abstract: Shortwave localization in complex scenarios plays a crucial role in battlefield communications. Existing deep-learning-based intelligent shortwave direction-finding methods mainly rely on manual annotation, and there is a scarcity of high-quality shortwave signal samples. To address the aforementioned challenges, a shortwave autonomous direction-finding and localization method is proposed, based on a space-time high-resolution time-frequency representation of shortwave signals and an attention-enhanced deep deterministic policy gradient (DDPG) algorithm. A reinforcement learning environment is first designed based on space-time high-resolution time-frequency representation, along with an agent possessing a hybrid action space. Then, a reward function is designed based on the estimated localization results, thereby establishing an observable Markov decision process. Subsequently, by enhancing deep reinforcement learning with an attention mechanism for multi-level non-fine evaluation, automatic direction finding and positioning are achieved. This design reduces the need for manual labeling and enables autonomous evolution through simultaneous work and learning, gradually improving the intelligent direction-finding and localization capabilities for shortwave signals. Experimental results show that, comparable localization accuracy and parameter estimation performance are maintained, while the direction-finding and localization time is improved by approximately 77.7%.

Key words: direction finding; Markov decision process; space-time high-resolution; hybrid action space; deep deterministic policy gradient algorithm

收稿日期:2026-01-15;修回日期:2026-04-09

作者简介:冯祺玥(2002—),女,硕士生,主要研究方向为阵列信号处理、强化学习。E-mail:fengqiyue2002@163.com

短波通信是远距离通信的主要方式,具有安装简单、通信距离远等优点,是军事上重要的通信方式^[1-2]。对短波信号进行测向定位,在态势分析方面具有重大意义。常用的短波信号测向方法通过多重信号分类(Multiple Signal Classification, MUSIC)算法获得仰角和方位角二维角度信息^[3]。目前,短波定位中最常用的方法是多站测角交会定位。该方法首先需要在每个观测站部署阵列天线,利用波达方向(Direction of Arrival, DOA)估算入射信号的方位角与仰角;进而,基于 DOA 参数确定目标位置^[4-5]。但传统的测向定位工作往往分为两步进行,即先对短波信号测向后定位,而且基于参数估计的定位方法计算量较大,因此难以实时测向与定位^[6-7]。

随着人工智能技术的快速发展,很多研究者对使用深度学习解决短波测向定位问题进行了深入研究,实现了复杂背景下短波信号的自动测向与定位^[8-9]。基于深度学习的信号目标检测与识别算法通过判断一些候选区域来达到检测和识别目标的目的^[10]。现有基于深度学习的自动测向方法基于实时空间谱测向结果聚类后形成若干个方向的数字波束合成数据,再利用深度学习对若干个方向的波束合成数据择优选取最优方向^[11],但深度学习需要大量信号样本和人工标注,且深度网络不能离线训练,自主智能程度不够。

强化学习作为人工智能领域另一个研究热点,已经应用于无人机、物联网等领域^[12-13]。其核心思想是智能体与环境交互获得的长期奖励最大化,持续学习达成目标的最优策略^[14]。因此,强化学习方法更侧重于问题解决策略的学习^[15],在智能博弈、资源调度和路径规划中有很应用潜力^[16]。

深度强化学习(Deep Reinforcement Learning, DRL)是深度学习与强化学习相结合的方法,以端到端方式训练智能体与环境交互^[17]。Mnih 等^[18]提出的深度 Q 网络(Deep Q Network, DQN)有效解决了状态空间爆炸问题,但存在估值过高的缺陷。为此,文献[19]开发了双重 DQN(Double DQN, DDQN)予以改进。基于行动者-评论家(Actor-Critic, AC)算法框架,Lillicrap 等^[20]针对连续动作空间问题提出深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)算法,该算法通过引入目标网络和软更新方法,在价值网络与策略网络的深度模型构建中发挥了关键作用。Bellver 等^[21]提出了基于强化学习的目标检测方法,证实了强化学习算法在目标检测与定位中的巨大潜力。Wang 等^[22]利

用 DRL 解决移动站点的调度和能量平衡的问题,适用于大规模站点对目标定位的场景。可见,深度强化学习在目标检测与定位等领域具有广阔的应用前景。

由上述研究可以看出,当前短波信号的协同测向定位研究较少,深度学习方法存在样本依赖强、在线自主学习能力不足的短板,经典强化学习算法的网络结构则缺乏动态筛选和聚焦的能力。注意力机制在处理任务时,可以根据重要性为其他元素分配不同的权重。因此,本文考虑将注意力机制融入强化学习,进一步提升协同定位的智能化水平。基于此,本文提出了基于注意力增强 DDPG 的短波协同测向定位(Direction-Finding and Localization Based on DDPG, DL-DDPG)方法。首先,引入对阵列数据进行空时高分辨处理的方法^[11],得到空时高分辨时频图,以此为基础设计了具有混合动作空间的强化学习测向环境;其次,设计可变矩形滤波窗为智能体,对所选信号利用 MUSIC 算法测得方位角和仰角进行 DOA 估计,根据定位结果的准确度设计奖励函数评价定位精度;最后,利用 DL-DDPG 训练智能测向定位环境,实现自动测向定位,可以减少人工标注,实现边工作边学习的自主进化,逐步提升短波信号的智能测向定位能力。

1 系统模型与问题描述

1.1 短波信号空时高分辨处理

短波信号测向模型如图 1 所示。

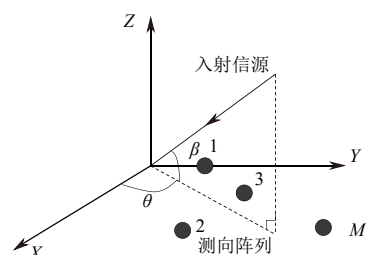


图 1 短波信号测向模型

假设任意几何位置排列的 M 个阵元的阵列天线同步接收信号,且同时有 D 个具有相同中心频率 ω_0 、波长为 λ 的空间窄带平面波($M > D$)分别以来向角 $\Theta_1, \Theta_2, \dots, \Theta_D$ 入射该天线阵列。其中: $\Theta_i = (\theta_i, \varphi_i)$, $i = 1, 2, \dots, D$; θ_i 和 φ_i 分别是第 i 个入射信号的方位角和仰角,并且满足 $0^\circ \leq \theta_i \leq 360^\circ$, $0^\circ \leq \varphi_i \leq 90^\circ$ 。根据 MUSIC 算法计算得到的空间谱采用预设的谱峰搜索信号来向 $\Theta_1, \Theta_2, \dots, \Theta_D$ 。

为得到空时高分辨时频图,将入射信号时域

阵列数据分为 L 段,对每段时域数据作傅里叶变换,得到入射信号的以时间为横坐标、以频率为纵坐标、以颜色为功率大小的时频谱图和以时间为横坐标、以方位角为纵坐标、以颜色为功率大小的时频向图。

首先,对每个时间片的数据依据频域模型进行高分辨测向处理,记录下每个时频点对应的多个测向结果,从而得到每个时间片的空谱测向结果。随后,测向设备基于多个时间片的空谱测向结果,对入射信号进行聚类分析,得到若干个角度聚类点。将这些聚类点作为基准角度,将每个时频点的测向结果与各基准角度进行比较,最终将判定存在有效测向结果的时频点,其对应的时频图输出为测向时频图。

1.2 短波信号DOA定位模型

在空间直角坐标系下,可利用 N 个观测站对单个静止目标进行定位。假设目标位置为 $\mathbf{u}^o = [u_x^o, u_y^o, u_z^o]^T$,第 n 个测向站的位置为 $\mathbf{u}_n = [u_{x,n}, u_{y,n}, u_{z,n}]^T, 1 \leq n \leq N$,则观测站与目标之间的方位角和仰角分别为:

$$\theta_n = \arctan\left(\frac{u_y^o - u_{y,n}}{u_x^o - u_{x,n}}\right) \quad (1)$$

$$\varphi_n = \arctan\left[\frac{u_z^o - u_{z,n}}{\sqrt{(u_x^o - u_{x,n})^2 + (u_y^o - u_{y,n})^2}}\right] \quad (2)$$

式中: θ_n 表示方位角; φ_n 表示仰角。

在多站交会定位中,各观测站可获取目标的方位角与仰角信息。据此,可推导出如下矩阵形式的等式:

$$\mathbf{G} = \mathbf{H}\mathbf{u}^o \quad (3)$$

$$\mathbf{H} = \begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) & 0 \\ \sin(\theta_1)\sin(\varphi_1) & \cos(\theta_1)\sin(\varphi_1) & -\cos(\varphi_1) \\ \vdots & \vdots & \vdots \\ \cos(\theta_N) & -\sin(\theta_N) & 0 \\ \sin(\theta_N)\sin(\varphi_N) & \cos(\theta_N)\sin(\varphi_N) & -\cos(\varphi_N) \end{bmatrix} \quad (4)$$

$$\mathbf{G} = \begin{bmatrix} u_{x,1}\cos(\theta_1) - u_{y,1}\sin(\theta_1), \\ u_{x,1}\sin(\theta_1)\sin(\varphi_1) + \\ u_{y,1}\cos(\theta_1)\sin(\varphi_1) - u_{z,1}\cos(\varphi_1), \\ \cdots, \\ u_{x,N}\cos(\theta_N) - u_{y,N}\sin(\theta_N), \\ u_{x,N}\sin(\theta_N)\sin(\varphi_N) + \\ u_{y,N}\cos(\theta_N)\sin(\varphi_N) - u_{z,N}\cos(\varphi_N) \end{bmatrix}^T \quad (5)$$

当存在误差时,式(3)转换为

$$\boldsymbol{\sigma} = \mathbf{G} - \mathbf{H}\mathbf{u}^o \quad (6)$$

针对上述模型,基于最小二乘准则(Least Squares)得到目标估计结果:

$$\mathbf{u}_{ls}^o = (\mathbf{H}^T \boldsymbol{\Omega}^{-1} \mathbf{H})^{-1} \mathbf{H}^T \boldsymbol{\Omega}^{-1} \mathbf{G} \quad (7)$$

式中,观测误差 $\boldsymbol{\sigma}$ 协方差矩阵 $\boldsymbol{\Omega} = \mathbf{E}[\boldsymbol{\sigma}\boldsymbol{\sigma}^T]$ 。

2 短波协同测向定位方案设计

2.1 强化学习环境建模

为实现短波无线电协同测向定位,利用深度强化学习机制将其建模为可观测马尔可夫决策过程,以减少人工标注,实现边工作边学习的自主进化。表1为基于强化学习的智能测向定位基本要素。

表1 强化学习模型概述

类型	描述
智能体	单个矩形框
动作	选中不同区域的时频图
状态	多节点被选中的时频图
奖励	各节点选中状态的交会定位结果
目标	由智能体对时频图进行自主框选并定位

图2展示了短波智能测向定位的总体研究方案。蓝色部分表示对阵列数据进行空时高分辨处理得到空时高分辨时频图,以此设计深度强化学习测向环境;黄色部分表示深度强化学习自主测向定位环节,智能体框选时频图得到测向结果,深度强化学习算法根据当前状态预测智能体动作。然后根据测向结果进行多站定位计算奖励,指导DRL算法预测动作。在基于空时高分辨时频图的测向定位环境中,多个测向站点同步接收信号。

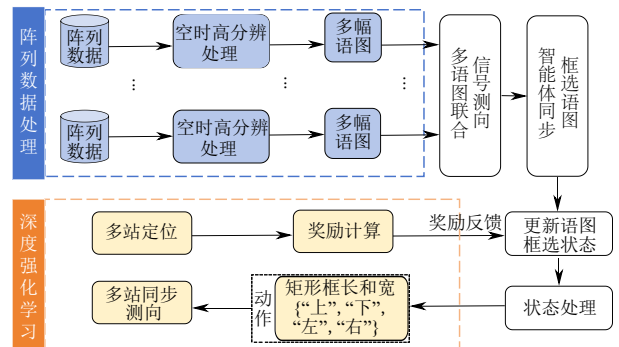


图2 短波协同测向定位研究方案

图3展示了协同测向定位的马尔可夫决策过程。强化学习的数学基础是马尔可夫决策过程,它本质上是一个序贯决策过程。其核心目标是寻找一个最优策略,以使智能体在与环境交互中获

得最大的累积奖励,即实现价值最大化。

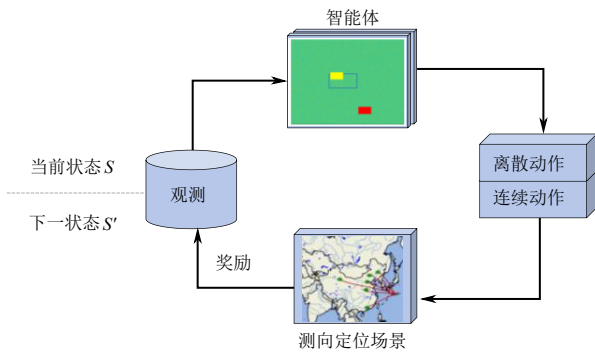


图3 智能测向与定位马尔科夫决策过程

2.2 状态空间

在所提协同测向定位环境中,观测空间 S 为短波信号时频图。对输入信号采样作快速傅立叶变换(FFT),得到信号时频图。为便于深度强化学习神经网络处理图像,将信号时频图转换为(400, 300, 3)的时频图RGB图像,以便训练感知矩形框智能体和目标信号的相对状态。

2.3 动作空间

动作 a 是指智能体根据当前状态为达到目标做出的决策。为设计可变矩形框,使得智能体能够快速准确选择信号并测向定位,本文考虑设计离散动作和连续动作相结合的混合动作。对于矩形框智能体,设计离散动作为{“上”,“下”,“左”,“右”},连续动作空间为矩形框的长 w 和宽 h ,取值范围分别为 $(0, w_{\max}]$ 和 $(0, h_{\max}]$,如表2所示。

表2 动作空间描述

类型	描述	取值
离散	向上、向下、向左、向右	{“上”,“下”,“左”,“右”}
连续	矩形框的长 h	$(0, h_{\max}]$
连续	矩形框的宽 w	$(0, w_{\max}]$

2.4 奖励函数设计

在模拟空间谱测向过程中,测向的准确度与所选区域的信噪比成反比。即信噪比越小,测向的精度越高;信噪比越大,测向的精度越低。当智能体选中信号时,智能体与信号的重叠面积 Y_{over} 与智能体矩形框的面积 Y_{agent} 的比值记为测向误差标准差 σ ,即:

$$\sigma = \frac{Y_{\text{over}}}{Y_{\text{agent}}} \in [0, 1] \quad (8)$$

因此,在已知测向站经纬度坐标的情况下,需要单个或多个目标的真实位置经纬度,以便计算目标相对于各个测向站的真实方位角 θ_i 和真实仰

角 φ_i ,得到真实角度矩阵 $\alpha_r = [\theta_r, \varphi_r]$,可以模拟实际的测向工作。各个测向站相对目标的测向角度 $\alpha = [\theta, \varphi]$ 为各个测向站相对目标的真实角度加上测向误差。方位角观测误差 σ_θ 和仰角观测误差 σ_φ 均为 $(N \times 1)$ 维的列向量,服从均值为0、标准差为 σ 的高斯分布,即:

$$\alpha = \alpha_r + \sigma \quad (9)$$

式中, σ 表示观测误差,为 $(2N \times 1)$ 维的列向量,并且协方差矩阵为 $\Omega = E[\sigma\sigma^T]$ 。

在目标解算阶段,为进一步提高定位精度。利用加权最小二乘准则(Weighted Least Squares),求解目标得:

$$u^o = (H^T W^{-1} H)^{-1} H^T W^{-1} G \quad (10)$$

式中, $W = A_w^T \Omega A_w$ 表示加权(Weighted)矩阵,计算方式为

$$A_w = C_w - [B_{w,11} u_{wls}^o, B_{w,12} u_{wls}^o, \dots, B_{w,M1} u_{wls}^o, B_{w,N2} u_{wls}^o] \quad (11)$$

式中: $B_{w,i1} = \partial G_w / \partial \theta_i$; $B_{w,i2} = \partial G_w / \partial \varphi_i$; $C_w = \partial h_w / \partial \alpha^T$ 。

为评估定位精度,考虑将第 i 次目标估计值作为第 $(i+1)$ 次估计的初始值,即设计基于误差相对值的归一化的奖励函数,有:

$$R = 100 \times \left(1 - \frac{E}{E_{\max}} \right) \quad (12)$$

$$E = \sqrt{\|u_i^o - u_{i+1}^o\|_2^2} \quad (13)$$

式中:设置奖励函数取值范围为 $[0, 100]$; E 为第 i 次目标估计值作为第 $(i+1)$ 次估计的初始值的误差; $E_{\max}$ 为最大允许估计误差,根据定位精度要求,在定位误差大于100 km时定位失效,因此设置 $E_{\max} = 10^4$ 。随着定位误差的增大,奖励函数值不断减小,为保证定位精度,将终止状态设置为

$$F_{\text{done}} = \begin{cases} 1, R \geq \lambda; \\ 0, \text{其他}. \end{cases} \quad (14)$$

因此,当智能体在空时高分辨时频图环境中一直向目标信号的位置移动时,定位精度越来越高,奖励回报不断增加。

3 深度强化学习算法设计

3.1 状态空间处理

由于观测状态为 400×300 的图像,若直接输入网络进行训练,参数规模过大,会导致神经网络训练速度慢且效果不佳。因此,首先将维度为(400, 300, 3)的RGB图像的观测状态进行灰度化处理;然后调整图像形状为(10, 10);再将二维向量展平为一维向量,与智能体的坐标 (x, y) 、长 w 和宽 h 拼

接为一维观测向量,如表3所示。

表3 观测空间构成

类型	描述	维度
环境观测	(10, 10)灰度图	100
智能体观测	矩形框的长 w	1
	矩形框的宽 h	1
	矩形框的横坐标 x	1
	矩形框的纵坐标 y	1
总观测		104

3.2 DL-DDPG 算法原理

所提DL-DDPG算法是基于注意力增强DDPG的深度强化学习改进算法。在深度强化学习中,智能体通过与环境的持续交互,记录其观察到的状态 s 、所执行的动作 a 以及获得的即时奖励,并利用这些积累的经验数据来学习策略函数 $\pi(a|s)$ 。在DL-DDPG算法中,既要考虑价值函数,也要考虑策略近似,学习近似策略的网络被称为策略网络(Actor),学习近似价值函数的网络被称为价值网络(Critic)。

DL-DDPG的策略网络不同于Actor-Critic的方法的策略网络,Actor-Critic的方法的策略网络 $\pi_{\varphi}(a|s)$ 是一个概率密度函数,输出的是概率值,而DL-DDPG的确定性策略网络 $\mu_{\varphi}(s)$ 的输出是 d 维的动作。因此确定性策略梯度定理推导为

$$\nabla_{\varphi} J(\varphi) = \mathbb{E}_{s \sim d^{\pi}} \left[\nabla_{\varphi} \mu_{\varphi}(s) \nabla_a Q_{\theta}^{\mu_{\varphi}}(s, a) \Big|_{a = \mu_{\varphi}(s)} \right] \quad (15)$$

然而,直接使用式(15)会导致策略梯度的方差过高,原因在于它直接依赖于奖励。为解决这一问题,考虑在DL-DDPG算法的梯度中添加长期回报:

$$\nabla_{\varphi} J(\varphi) = \mathbb{E}_{s \sim d^{\pi}} \left[\nabla_{\varphi} \mu_{\varphi}(s) \nabla_a Q_{\theta}^{\mu_{\varphi}}(s, a) \Big|_{a = \mu_{\varphi}(s)} - \eta l_{\varphi}(s) \right] \quad (16)$$

$$l_{\varphi}(s) = \mathbb{E}_{s \sim d^{\pi}} [Q_{\pi}(s_t, a_t)] = \sum_{a \in A} \pi_{\varphi}(a|s_t) \cdot Q_{\pi}(s_t, a) \quad (17)$$

式中, η 为缩放因子。公式(17)中的 $l_{\varphi}(s)$ 可以使智能体能够通过策略网络更积极地探索动作空间。

对于价值网络而言,其所代表的状态—动作值函数用以评估策略网络在状态 s 下所做出动作 a 的价值,通过环境反馈的立即回报建立时序差分误差更新神经网络参数 θ :

$$J_Q(\theta) =$$

$$\mathbb{E}_{(s, a, r, s') \sim D} \left[\frac{1}{2} \left(Q_{\theta}(s, a) - \left(r(s, a) + \gamma E_{s' \sim p_{\pi}(s)} [Q_{\theta}(s', a')] \right) \right)^2 \right] \quad (18)$$

式中, $Q_{\theta}(s', a')$ 为目标网络。全连接(Fully Connected, FC)网络用于计算查询(Query, Q)、键(Key, K)和值(Value, V),并将它们分别组织成矩阵 Q 、 K 和 V 。假设编码维度为 d ,输入源数量为 n 。在多头注意力机制中,不同的注意力头可以捕捉不同的特征。为了计算最终的 Q 值,各个头投影后的特征结果会被拼接起来,并输入到一个线性层中,该线性层的表达式为

$$V_{\pi}(s_t) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V W^0 \quad (19)$$

式中, W^0 表示输出层的权重矩阵。针对目标网络DL-DDPG算法采用软更新的方式更新网络参数,即每一个时间步都更新 $\theta^- \leftarrow \tau \theta + (1 - \tau) \theta^-$ 和 $\varphi^- \leftarrow \tau \varphi + (1 - \tau) \varphi^-$,并且 $0 < \tau \ll 1$,因此也能够稳定地利用时序差分误差对策略参数进行更新。

3.3 DL-DDPG 算法结构

如图4所示,DL-DDPG算法策略网络输入层为104维的观测向量,价值网络的输入为104维观测状态以及策略网络输出的2维矩形框的长和宽向量,因此总输入为106维的观测向量。策略网络输出连续动作空间2维向量 a 。Critic网络输出离散动作每个维度的 Q 值,算法网络的参数如表4所示。

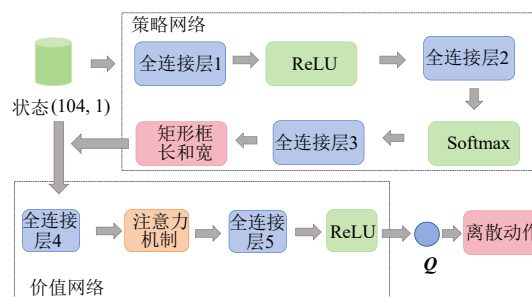


图4 DL-DDPG网络结构

表4 神经网络构成参数

名称	性质	输出大小
全连接层1	线性	104×128
全连接层2	线性	128×32
全连接层3	线性	32×2
全连接层4	线性	106×128
全连接层5	线性	128×1

3.4 DL-DDPG 算法步骤

DL-DDPG算法训练时,首先将观测状态传入策略网络,用于确定连续动作空间的参数,输出是连续参数的值。然后将观测状态和连续参数拼接

后输入价值网络得到 Q 值,通过最大化 Q 值来选择离散行动类型。最后将离散动作与连续动作拼接为 3 维动作向量 $\mathbf{a}$ 。DL-DDPG 算法执行算法 1 的步骤更新价值网络和策略网络。

算法 1 基于 DDPG 算法的短波信号协同测向与定位算法

输入:接收信号的时空高分辨时频图
输出:目标信号的位置 $\mathbf{u}^o$
初始化经验回放池 D ,价值网络记为 $q(s, \mathbf{a}; \boldsymbol{\omega})$,DDPG 算法的确定性策略为 $\boldsymbol{\mu}(s; \boldsymbol{\theta})$,神经网络的当前参数记作 $\boldsymbol{\omega}_{\text{now}}$ 和 $\boldsymbol{\theta}_{\text{now}}$
for episode = 1:K do
 矩形框和信号随机初始位置,获得当前状态 s_t ,
 for step = 1:K do
 矩形框根据当前状态 s_t ,依据策略网络执行动作 $\mathbf{a}_t$,
 依据选中时频图测向获得测向误差,计算奖励 r_t ,并获得新状态 s_{t+1} 。
 策略网络做预测 $\hat{\mathbf{a}}_{t+1} = \boldsymbol{\mu}(s_{t+1}; \boldsymbol{\theta}_{\text{now}})$
 价值网络做预测 $\hat{q}_{t+1} = q(s_{t+1}, \hat{\mathbf{a}}_{t+1}; \boldsymbol{\omega}_{\text{now}})$
 根据预测结果计算误差 $\hat{y}_t = r_t + \gamma \cdot \hat{q}_{t+1}, \delta_t = \hat{q}_t - \hat{y}_t$,
 更新价值网络
 $\boldsymbol{\omega}_{\text{new}} \leftarrow \boldsymbol{\omega}_{\text{now}} - \alpha \cdot \delta_t \cdot \nabla_{\boldsymbol{\omega}} q(s_t, \mathbf{a}_t; \boldsymbol{\omega}_{\text{now}})$
 更新策略网络
 $\boldsymbol{\theta}_{\text{new}} \leftarrow \boldsymbol{\theta}_{\text{now}} + \beta \cdot \hat{q}_t \cdot \nabla_{\boldsymbol{\theta}} \boldsymbol{\mu}(s_t; \boldsymbol{\theta}_{\text{now}}) \cdot \nabla_{\boldsymbol{\alpha}} q(s_t, \hat{\mathbf{a}}_t; \boldsymbol{\omega}_{\text{now}})$
 end for
end for

4 仿真实验与性能分析

4.1 实验仿真条件

本文考虑设计 6 个测向站和一个目标,测向站经纬度如表 5 所示。分别将算法训练 2 000 回合,每一回合随机初始智能体和信号位置,不同信号位置代表不同的目标源,设置每一回合最大步长为 50。

表 5 测向站经纬度坐标

测向站	经度(E)	纬度(N)
1	115.4°	35.6°
2	104.41°	24.64°
3	124.52°	31.57°
4	116.27°	23.69°
5	124.32°	42.61°
6	108.51°	30.43°

在短波协同测向定位环境中,分别采用 AC、DQN、Double-DQN、DL-DDPG 算法训练强化学习环

境,其中 AC、DQN、Double-DQN 算法为对比算法。DL-DDPG 算法的参数设置如表 6 所示。

表 6 训练过程的参数设置

符号	描述	DL-DDPG 参数
$\varepsilon_{\text{start}}$	策略初始概率	0.9
ε_{end}	策略截止概率	0.01
γ	折扣回报率	0.96
α	价值网络学习率	0.000 1
β	策略网络学习率	0.000 2
episode	回合数	2 000
step	每回合最大步长	50

4.2 算法训练结果分析

图 5 和图 6 分别展示了 AC、DQN、DDQN、DL-DDPG 算法训练的平均奖励和平均步长的变化曲线,其中:横坐标表示训练回合数;纵坐标分别表示每 20 回合的平均奖励或平均步长;阴影部分表示平均奖励和平均步长的标准差。

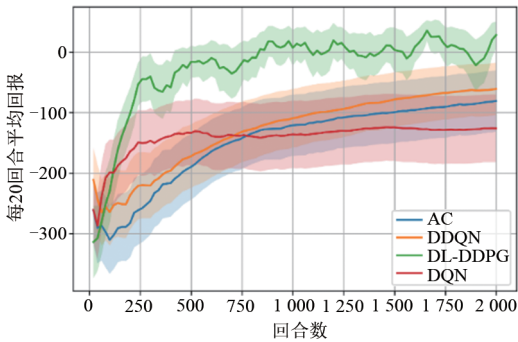


图 5 不同算法平均回报对比曲线

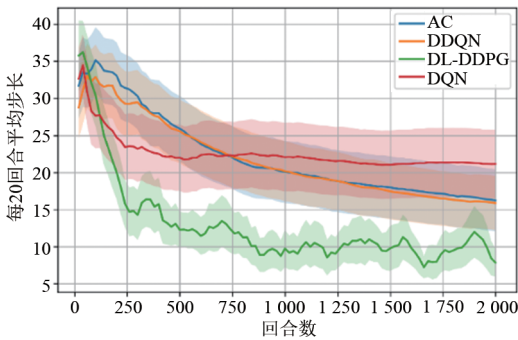


图 6 不同算法平均步长对比曲线

由图 5 和图 6 可以看出,采用混合动作的 DL-DDPG 算法相比于 AC、DQN、DDQN 算法,收敛速度快,并且平均回报高于 AC、DQN、DDQN 算法,但训练过程存在波动。所提的 DL-DDPG 算法的平均步长可收敛至 10 步以内,相比于 AC、DQN、DDQN 算法收敛速度快,即以较短的路径达到较优的目标。

4.3 关于定位时间的结果分析

将 DL-DDPG 算法训练模型保存后测试,随机

测试500次,记录每次测试的步长,并计算每个目标的测向定位时间。

图7表示本文算法与传统两步测向定位算法^[2]的平均运行时间对比,可以看出,传统两步测向定位算法的平均运行时间均大于本文算法。实验结果表明,本文算法的平均测向定位时间为0.08 s,传统两步测向定位算法的平均测向定位时间为0.36 s。本文算法的处理速度达到了传统算法的4.5倍,定位时间缩短了77.8%,显著提升了测向定位实时性。

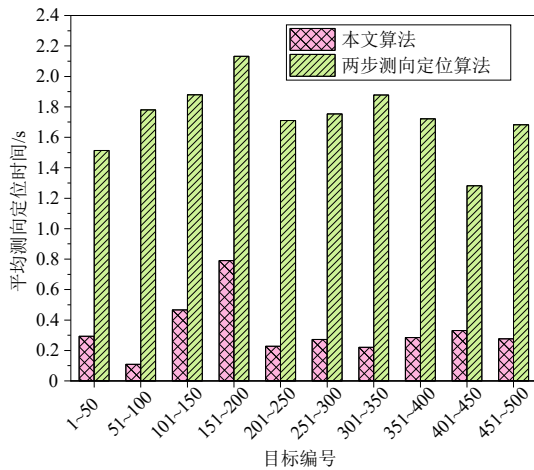
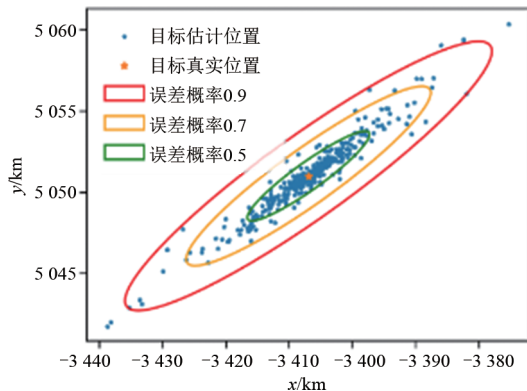


图7 本文算法与传统两步测向定位算法平均运行时间对比

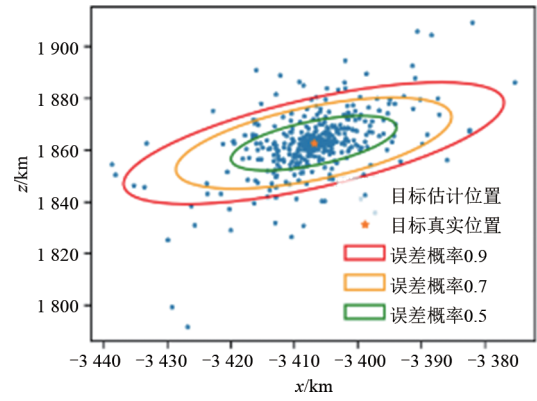
4.4 关于定位误差的实验分析

为进一步分析所提测向定位算法的定位性能,本节进行定位误差分析,选取目标经纬度为(126°E, 24°N),测向误差服从高斯分布。测试时分别使用DL-DDPG保存后模型进行500次蒙特卡洛实验。

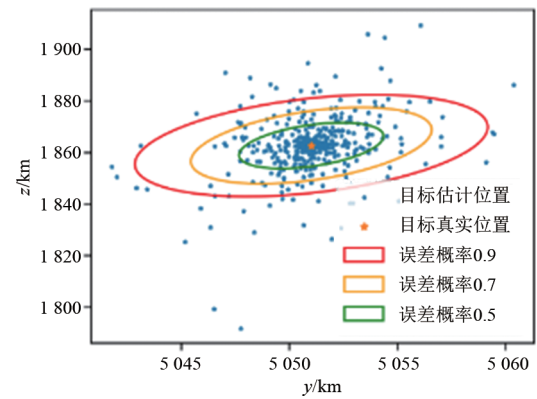
图8(a)、图8(b)和图8(c)分别展示了目标在x-y平面、x-z平面及y-z平面上的定位结果散点图及其对应的定位误差椭圆。实验定位结果的散点分布形态与误差椭圆的形状高度吻合,验证了定位结果的误差分布与理论预期置信区间具有良好的 consistency。



(a) 目标定位x-y平面结果



(b) 目标定位x-z平面结果



(c) 目标定位y-z平面结果

图8 目标定位平面结果

为进一步对比本文算法的定位性能,采用均方根误差(RMSE)来评判算法的定位性能:

$$E_{\text{RMS}} = \sqrt{\frac{1}{K} \sum_{k=1}^K \|\hat{\mathbf{u}}_k - \mathbf{u}_k\|_2^2} \quad (20)$$

式中: K 为蒙特卡洛次数; $\hat{\mathbf{u}}_k$ 为第 k 次测试中目标位置的估计值; $\mathbf{u}_k$ 为目标位置的真实值。

图9展示了不同信噪比(Signal-to-Noise Ratio, SNR)情况下,本文算法与两步测向定位算法^[2]和克拉美罗界(Cramer-Rao Bound, CRLB)的对比。可以看出,在该定位场景下,本文算法更接近于CRLB。由于两步测向定位算法在构建伪线性方程时引入了与目标位置相关的辅助变量,因此该算法分为两个阶段。第1阶段假设辅助变量与目标无关,先对方程进行求解;第2阶段则根据目标与辅助变量之间的关系重新建立方程,并利用第1阶段的估计结果进行求解,从而提升第1阶段的定位精度。该算法属于闭式解,具有计算复杂度低、无需先验初值等优点。然而,该算法存在解模糊的问题,最终结果需要根据目标的大致范围进行判定。本文所提出的短波信号自主测向定位方法是在短波信号空时高分辨时频图的基础上进行处理,即将短波信号频率搬移后再做处理,因此本方

法可以不受短波信号频率的影响。

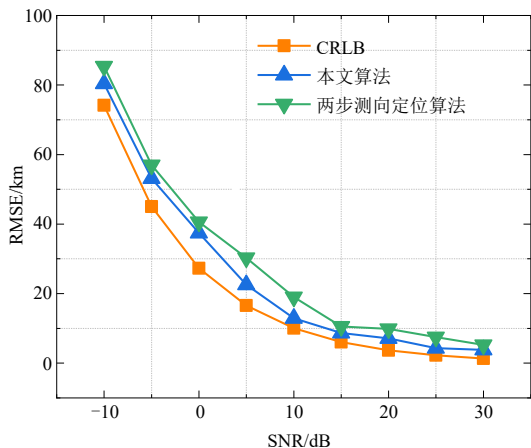


图9 本文算法与传统两步测向定位算法定位误差随信噪比变化

4.5 算法复杂度分析

DL-DDPG算法能够实现信号的同时测向与定位,从而显著减少处理时间并提升效率。算法架构包含一个Critic网络和一个Actor网络,算法网络由 N 个全连接层组成。全连接层的复杂度为 $O\left(\sum_l W_{l-1} W_l\right)$,其中 L 为全连接层数, W_{l-1} 和 W_l 分别表示第 l 层的输入和输出规模。因此,DL-DDPG算法复杂度为 $O\left(N \sum_l W_{l-1} W_l\right)$ 。

5 结束语

本文提出了一种基于注意力增强DDPG算法的协同测向定位技术,自主设计强化学习环境,并通过实验验证了该方法的有效性。设计了基于任务的奖励函数,能够较好地引导智能体快速完成测向定位任务。本文验证了基于深度强化学习的智能测向定位工作的可行性,不依赖大量人工标注,可以在线多回合自主学习,为智能测向定位工作提供了良好的技术参考。尽管本文所提算法也适用于多信号环境,但需要逐一寻找不同信号进行测向定位,效率较低。多智能体深度强化学习算法可以利用多个智能体同时处理多个信号,因此在未来将开展多智能体智能测向定位的研究。

参考文献:

[1] FENG Q Y, TANG T, ZHANG Y P, et al. Multistation target localization *via* direction-of-arrival: an efficient multiagent deep reinforcement learning approach [J].

IEEE Internet of Things Journal, 2025, 12(19): 40130-40145.

- [2] 王鼎,尹洁昕,朱中梁.针对超视距短波辐射源的测角与测时差协同定位方法[J].中国科学:信息科学, 2022, 52(11): 1942-1973.
- [3] XU X X, HUANG Q H. MD-DOA: a model-based deep learning DOA estimation architecture [J]. IEEE Sensors Journal, 2024, 24(12): 20240-20253.
- [4] ZHENG S L, YANG Z, SHEN W G, et al. Deep learning-based DOA estimation [J]. IEEE Transactions on Cognitive Communications and Networking, 2024, 10(3): 819-835.
- [5] ZHOU C W, GU Y J, SHI Z G, et al. Structured nyquist correlation reconstruction for DOA estimation with sparse arrays [J]. IEEE Transactions on Signal Processing, 2023, 71: 1849-1862.
- [6] 翟晨旭,余思雨,崔毓函,等.基于能量累积和双极性相关的ADS-B信号联合检测方法[J].信息工程大学学报, 2026, 27(1): 27-34.
- [7] ZHENG H, ZHOU C W, VOROBOYV S A, et al. Decomposed CNN for sub-nyquist tensor-based 2-D DOA estimation [J]. IEEE Signal Processing Letters, 2023, 30: 708-712.
- [8] CONG J Y, WANG X P, YAN C G, et al. CRB weighted source localization method based on deep neural networks in multi-UAV network [J]. IEEE Internet of Things Journal, 2023, 10(7): 5747-5759.
- [9] TOMIC S. Secure AoA-based localization in the presence of malicious nodes [J]. IEEE Internet of Things Journal, 2026, 13(7): 12800-12810.
- [10] ZHOU M, WANG R J, XIE C J, et al. ReinforceNet: a reinforcement learning embedded object detection framework with region selection network [J]. Neurocomputing, 2021, 443: 369-379.
- [11] 唐涛,王鼎,杨宾,等.一种基于高分辨测向语图智能识别的短波自动测向方法: CN110045322B [P]. 2021-01-22.
- [12] JIANG B Q, DU J, JIANG C X, et al. Underwater searching and multiround data collection *via* AUV swarms: an energy-efficient AoI-aware MAPPO approach [J]. IEEE Internet of Things Journal, 2024, 11(7): 12768-12782.
- [13] YU Y, LIU M. Deep reinforcement learning-based multi-sensor control for labeled multi-bernoulli filtering [J]. IEEE Transactions on Aerospace and Electronic Systems, 2025, 61(5): 13548-13564.
- [14] XU W R, YU J, MIAO Z J, et al. Spatio-temporal deep Q-networks for human activity localization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(9): 2984-2999.