基于深度强化学习的短波自主检测与定位方法

冯祺玥，唐涛，张昀普，王鼎，吴志东
（信息工程大学信息系统工程学院，河南 郑州 450001）

摘要：在短波信号实时检测与定位环节，往往存在数据关联与误差传递问题，容易导致检测信号与定位信号不匹配，且难以实时输出结果。针对上述问题，提出一种基于多智能体近端策略优化的短波自主检测与定位方法。通过设计基于短波信号时频图的强化学习环境，有效表征短波信号的频域特征。同时，设计具有混合动作空间的滤波窗智能体以自适应选择信号。另外，通过定位误差椭圆概率、信号匹配度与多智能体协作策略设计奖励函数，利用改进的多智能体近端策略优化算法探索最优策略，实现短波自主检测与定位。仿真测试结果表明，相较于基线算法，信号检测与定位的平均时间缩短了 0.12 s，平均检测与定位的准确率提高了 22%，为复杂电磁环境下自主协同目标检测与定位提供新思路。

关键词：短波信号检测与定位；多智能体近端策略优化；强化学习；混合动作空间；定位误差椭圆概率
中图分类号：TN97
文献标志码：A
doi：10.11959/j.issn.1000–436x.TXXB260042

Shortwave autonomous detection and localization method based on deep reinforcement learning

Feng Qiyue, Tang Tao, Zhang Yunpu, Wang Ding, Wu Zhidong
School of Information System Engineering, Information Engineering University, Zhengzhou 450001, China

Abstract: In the process of real-time shortwave signal detection and localization, problems such as data association and error propagation are frequently observed, by which a mismatch between detection and localization results is easily induced, and the timely output of outcomes is impeded. To address the above issues, an integrated shortwave signal detection and localization method based on multi-agent proximal policy optimization was proposed. By constructing a reinforcement learning environment using shortwave signal time-frequency diagrams, spectral characteristics of shortwave signals were effectively captured. A filtering-window agent with hybrid action space was designed for adaptive signal selection. Furthermore, localization error ellipse probability, signal matching degree, and multi-agent collaboration strategies were incorporated into the reward function design. Optimal policies were explored by the optimized multi-agent proximal policy optimization to achieve autonomous shortwave signal detection and localization. Simulation results demonstrate that compared with baseline algorithms, the proposed method reduces average recognition-localization time by 0.12 s while improving accuracy by 22%, thereby providing a novel solution for autonomous cooperative target detection and localization in complex electromagnetic environments.

Key words: shortwave signal detection and localization, multi-agent proximal policy optimization, reinforcement learning, hybrid action space, localization error ellipses probability

收稿日期：2026–01–17；修回日期：2026–04–22
通信作者：唐涛，13703820621@163.com
基金项目：国家自然科学基金资助项目（No.62171469）
Foundation Item：The National Natural Science Foundation of China (No.62171469)

0 引言

短波通信凭借其独特的超视距传输能力，在应急通信、侦察及远洋导航等领域具有不可替代的优势，但受限于多径效应、动态衰减等复杂信道特性，短波信号检测与定位较困难[1-2]。短波常采用多站测角定位机制，需要利用波达方向（direction of arrival，DOA）估计方法得到入射信号方位角和仰角，然后利用参数估计方法确定目标位置，该方法依赖精确的角度测量结果[3-4]。因此，经典信号检测与定位（detection and localization，DL）方法往往需要先检测发现信号，再进行信号类型检测与定位等工作，该过程需要大量人工参与信号判别与测定等操作。在实时变化的信号环境下，传统方法往往存在数据关联与误差传递问题，导致定位实时性较差，难以满足短波信号实时检测与定位的需求。

随着人工智能技术的快速发展，强化学习在无线通信领域的应用为信号自主检测与定位提供了新思路[5]。单智能体深度强化学习通过与环境交互自主优化策略，已在非动态场景的频谱检测与功率控制中取得进展。在信号自主检测与定位领域，强化学习通过模拟智能体与环境交互，自适应地学习图像信号特征并优化分类策略[6-9]。Caicedo等[7]提出基于深度Q网络（deep Q network，DQN）的图形信号检测分类任务。Li等[10]提出利用DQN算法实现无线信号定位的方法，设计一种从无线接收信号强度中提取信号位置的奖励机制，通过模型训练实现无线信号定位。Paul等[11]提出一种基于强化学习的DOA位置估计方法，利用最小二乘法估计信号源位置，以获得高精度的定位结果。

与传统方法相比，强化学习在动态干扰的低信噪比环境下表现出更强的鲁棒性和自适应能力[12]。基于强化学习的定位方法通过建立定位场景的马尔可夫观测过程，动态优化测向准确性，从而提升定位精度与效率，表现出良好的定位性能。然而，短波测向定位系统是一个多智能体协作系统，各站点根据局部观测信息实时协同决策，传统单智能体强化学习方法容易忽略智能体间的协作，导致定位效率受限。

多智能体强化学习（multi-agent reinforcement learning，MARL）作为强化学习的研究热点，已经应用于无人机、物联网等领域[13]。MARL通过模拟多个智能体之间的协作与竞争，能够有效解决复杂环境下的决策问题[14]。其中，多智能体近端策略优化（multi-agent proximal policy optimization，MAPPO）算法作为一种高效的MARL算法，通过策略优化和分布式学习，为多智能体系统的协同决策提供新思路[15]。MAPPO更适用于实时变化环境下的多目标信号分析任务，可以优化信号接收任务，具有较强的自主性和适应性。Wang等[16]提出一种基于多智能体深度强化学习和策略优化的方法，用于解决多源信号定位问题，并通过智能体之间的协作实现高精度定位。文献[17]利用基于值分解的强化学习与补偿滤波的多智能体协同定位算法解决非线性环境下的定位问题。Alagha等[18]提出使用MAPPO算法解决目标定位问题，但该算法仅适用于二维平面的定位问题，无法解决三维空间中信号的定位问题。

综上所述，现有基于MARL算法的信号检测与定位方法仍存在以下不足：（1）泛化能力不足，部分算法仅适用于特定目标的定位，无法在时变的信号环境中对目标进行定位；（2）短波信号先检测再定位的两步方法中处理环节较多，在信号实时检测与定位环节中可能存在检测信号与定位信号不匹配的问题；（3）部分算法仅研究基于MARL算法的信号检测或定位的单一方案，尚未综合考虑两者，因此无法同时实现信号的检测与定位。

针对上述问题，本文基于MAPPO算法提出一种短波信号自主检测与定位方法（MAPPO-DL）。首先，通过充分利用信号特征的时空分布特性，对阵列数据进行滤波处理得到时频图。然后，在时频图基础上构建多智能体强化学习马尔可夫决策过程，通过智能体对信号进行滤波，同时完成检测与定位工作，以避免检测与定位信号不匹配问题。最后，引入定位误差椭圆概率、信号匹配度和多智能体协作策略设计奖励函数，实现信号高效检测与实时定位一体化。

1 系统模型

短波信号检测与定位模型如图1所示，短波信号接收设备接收信号后，将入射信号时域数据进行快速傅里叶变换（fast Fourier transform，FFT），得到以时间为横坐标、频率为纵坐标的入射信号时频图。在时频图上通过滤波检测信号获得信号数据，

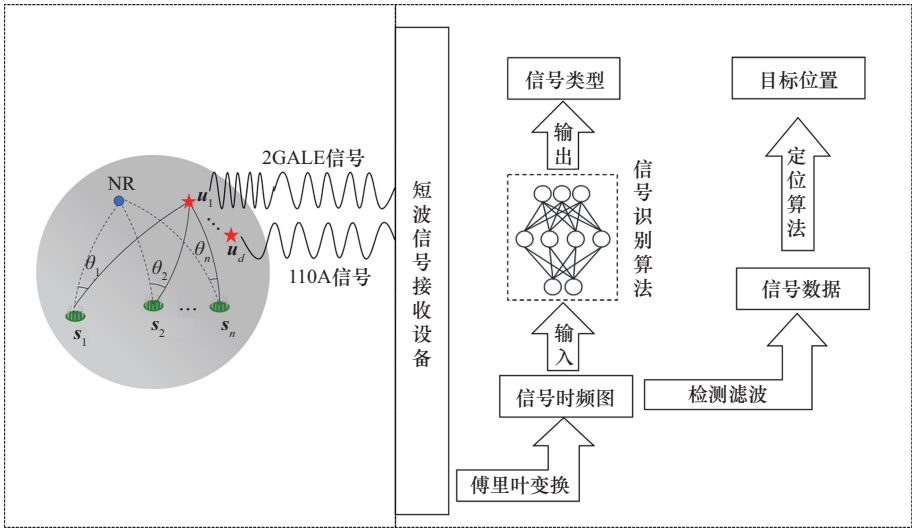


图1　短波信号检测与定位模型

再根据定位算法获得目标位置。另外，以时频图为输入，利用信号检测算法得到信号类型。

本文将地球建模成半径为等效地球半径的球体，并在该球体上建立一个忽略高度的地理坐标系。NR 表示北极，目标源和站点位于球体表面，其坐标用纬度 ρ 和经度 ω 表示。假设地球表面存在 D 个待定位目标，该目标源辐射短波信号。$\boldsymbol{u}=\left[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_D\right]^{\mathrm{T}}$ 表示 D 个未知的目标源位置，第 d 个目标源的经纬度为 $\boldsymbol{u}_d=\left[\omega_d,\rho_d\right]$，$d\in\{1,2,\cdots,D\}$，第 n 个测向站的经纬度为 $\boldsymbol{s}_n=\left[\omega_n,\rho_n\right]$，$n\in\{1,2,\cdots,N\}$。$\theta_n\in(-\pi,\pi]$ 表示信号源 $\boldsymbol{u}_d$ 到站点 $\boldsymbol{s}_n$ 的方位角，方位角在正北方向为0，顺时针方向为正。

每个测向站中包含 M 个阵元的阵列天线同步接收信号，设同时有 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 $\theta_1,\theta_2,\cdots\theta_D$ 入射该天线阵列，其中 θ_i 是第 i 个入射信号的方位角。窄带平面波信号的阵列输出数学模型矩阵形式为：

$$\boldsymbol{X}(t)=\boldsymbol{A}(\theta)\boldsymbol{S}(t)+\boldsymbol{N}(t) \tag{1}$$

其中，$\boldsymbol{X}(t)=[x_1(t),x_2(t),\cdots,x_M(t)]$ 为阵列输出矢量，$\boldsymbol{N}(t)=[n_1(t),n_2(t),\cdots,n_M(t)]$ 为阵列加性噪声矢量，$\boldsymbol{S}(t)=[s_1(t),s_2(t),\cdots s_D(t)]$ 为接收信源矢量，$\boldsymbol{A}(\theta)=\left[\boldsymbol{a}(\theta_1)\boldsymbol{a}(\theta_2)\cdots\boldsymbol{a}(\theta_D)\right]$ 为阵列流型矩阵，$\boldsymbol{a}(\theta_i)=[\mathrm{e}^{-\mathrm{j}\omega_0\tau_{1i}},\mathrm{e}^{-\mathrm{j}\omega_0\tau_{2i}},\cdots,\mathrm{e}^{-\mathrm{j}\omega_0\tau_{Mi}}]^{\mathrm{T}}$ 为阵列方向矢量，ω_i 为信号 i 的中心频率。

对于 D 个目标源信号的到达角，MUSIC 空间谱估计 $\hat{P}_{\mathrm{MUSIC}}(\theta)$ 可以用噪声子空间 $\boldsymbol{U}_M$ 的特征矢量 $\boldsymbol{\Phi}_i(i=D+1,\cdots,M)$ 表示为：

$$\hat{P}_{\mathrm{MUSIC}}(\theta)=\boldsymbol{a}^{\mathrm{H}}(\theta)\left(\sum_{i=D+1}^{N}\hat{\boldsymbol{\Phi}}_i\hat{\boldsymbol{\Phi}}_i^{\mathrm{H}}\right)\boldsymbol{a}(\theta) \tag{2}$$

根据式(2)计算得到的空间谱，采用谱峰搜索获得信号来向 $\theta_1,\theta_2,\cdots,\theta_D$。根据方位角估计结果进行交叉定位，对所得交点取平均得到目标估计位置 $\hat{\boldsymbol{u}}_d$。

在信号检测任务中，以 2GALE、LINK4A、110A、CLOVER2000 和 LINK11 短波信号为例进行分析。如图2所示，经过下变频采样设备接收的短波信号在时频图上的特点各有不同：2GALE 和 LINK4A 采用 FSK 调制方式，其时频图呈现多条间隔出现的线段；110A 采用 PSK 调制方式，时频图中有效区域呈矩形；CLOVER2000 的时频图包含 8 条断断续续的斑点；LINK11 的特殊帧结构导致其时频图由哑铃状的矩形组成。因此，通过时频图即可区分短波特定信号。

在动态变化的信号环境下，基于深度学习的智能检测算法不仅依赖大规模数据集进行训练，而且信号间的相互干扰可能导致检测信号与定位信号不一致，同时还未能实现短波信号的测向与定位一体化。针对短波多信号动态性强、干扰大的问题，本文采用基于多智能体的深度强化学习对信号检测与定位问题进行马尔可夫决策过程（Markov decision process，MDP）建模，使智能体能够在信号环境中自主学习和探索，从而实现短波信号的实时测向与定位。

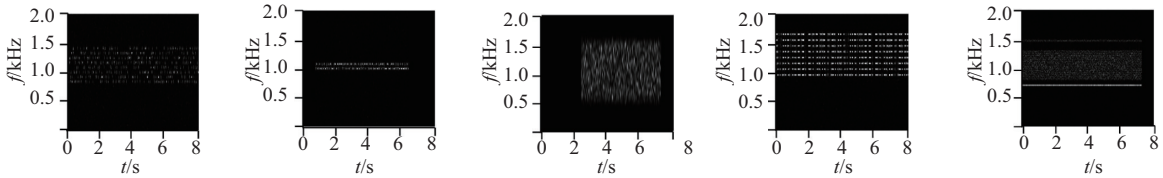


图2　5类短波信号时频图

2　MAPPO-DL

2.1　建立马尔可夫决策过程

本文所提MAPPO-DL方案利用深度强化学习机制，将短波信号检测与定位过程建模为可观测马尔可夫决策过程。智能体通过与环境的不断交互试错，寻找累积奖励最大的策略，从而实现边执行边学习的自主进化能力。为实现短波信号自主检测和定位，需要先对入射信号进行滤波处理得到空时时频图，MAPPO-DL框架如图3所示。

在图3中，对入射信号时域数据进行FFT，得到入射信号的空间谱时频图。其次，对每个时间片数据按频域模型进行MUSIC测向处理，记录每个时频点对应的测向结果，形成每个时间片数据的空间谱测向结果[19]。在基于时频图的多智能体信号检测与定位环境中，多个测向站点同步接收信号。以下分别对该强化学习马尔可夫过程的动作、状态与奖励设计进行阐述。

2.1.1　动作

动作指智能体根据当前状态为达到目标做出的决策。智能体指做动作的主体，在MAPPO-DL环境中自适应滤波窗即智能体，智能体的长度为w，宽度为h。为使智能体能够快速准确选择信号并完成检测与定位，考虑设计离散动作x_t^i和连续动作$\boldsymbol{y}_t^i$相结合的混合动作，则多智能体环境中的动作空间为：

$$\boldsymbol{A}_t=\left[\boldsymbol{a}_t^1,\boldsymbol{a}_t^2,\cdots,\boldsymbol{a}_t^n\right] \tag{3}$$

其中，$\boldsymbol{a}_t^i=\left[x_t^i,\boldsymbol{y}_t^i\right]^{\mathrm{T}}$，离散动作$x_t^i$为滤波窗移动方向，主要包括上、下、左、右、左上、左下、右上和右下8个方向，左上、左下、右上和右下均为45°方向；连续动作$\boldsymbol{y}_t^i=\left[w_t^i,h_t^i\right]$为自适应滤波窗的尺度变换。

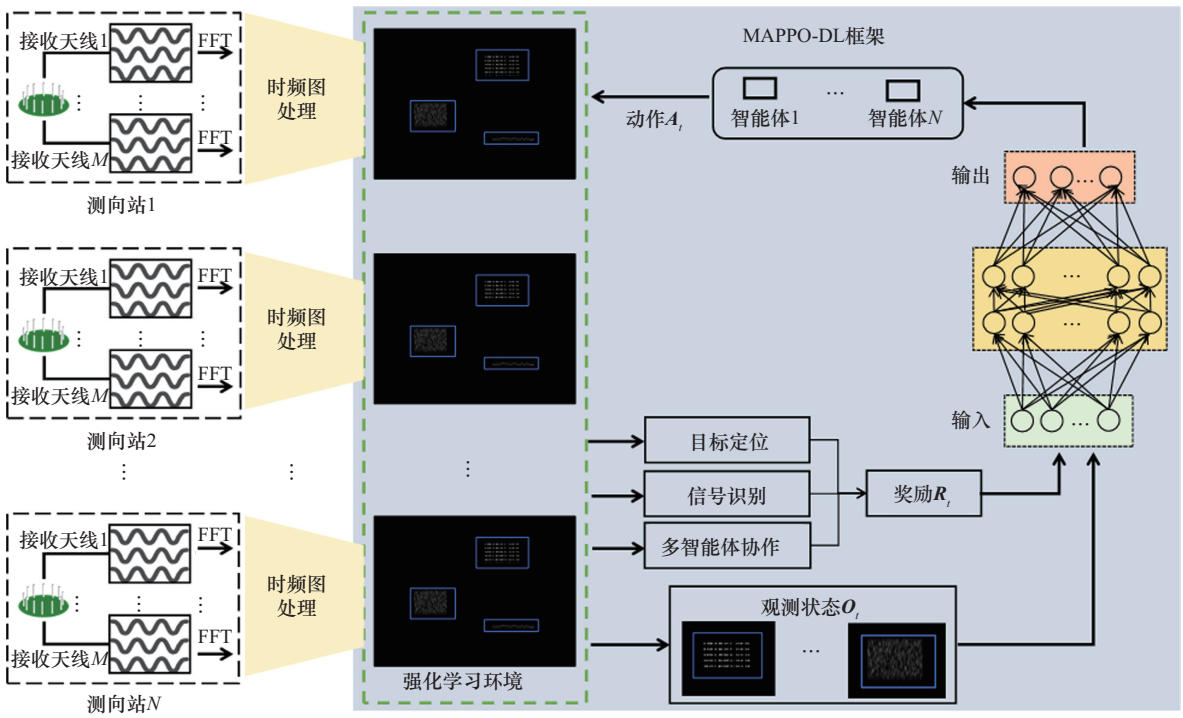


图3　MAPPO-DL框架

2.1.2 状态

在信号智能检测与定位环境中，t时刻的全局状态s_t为接收信号的时频图。智能体i的局部观测状态为其周围$(2w_{\max},2h_{\max})$大小的图像o_t^i，使滤波窗智能体能更好地检测执行动作后的环境变化。因此，总体观测状态空间为：

$$\boldsymbol{O}_t=\left[o_t^1,o_t^2,\cdots,o_t^n\right] \tag{4}$$

2.1.3 奖励

（1）定位奖励

在短波空间谱测向过程中，测向的准确度与智能体所选信号的信噪比呈正相关，即信噪比越高，测向精度越高；反之越低。在任意多个观测站条件下同步测向交会定位时，将测向站经纬度转换为地心地固坐标。考虑利用N个测向站对地球表面的短波辐射源进行定位，第n个测向站的经纬度为(ω_n,ρ_n)，则该测向站的地心地固坐标为：

$$\boldsymbol{s}_n=\begin{bmatrix}\mu_n\cos(\rho_n)\cos(\omega_n)\\ \mu_n\cos(\rho_n)\sin(\omega_n)\\ \mu_n\sin(\rho_n)\end{bmatrix} \tag{5}$$

其中，$1\leqslant n\leqslant N$，$\mu_n=\dfrac{R_{\mathrm{e}}}{\sqrt{1-(\sin(\rho_n))^2}}$，$R_{\mathrm{e}}=6\ 378.160$ km为地球等效半径。该测向站对应的坐标转换矩阵为：

$$\boldsymbol{T}_n=\begin{bmatrix}-\sin(\omega_n) & \cos(\omega_n) & 0\\ -\cos(\omega_n)\sin(\rho_n) & -\sin(\omega_n)\sin(\rho_n) & \cos(\rho_n)\\ \cos(\omega_n)\cos(\rho_n) & \sin(\omega_n)\cos(\rho_n) & \sin(\rho_n)\end{bmatrix}=\begin{bmatrix}\boldsymbol{a}_n^{\mathrm{T}}\\ \boldsymbol{b}_n^{\mathrm{T}}\\ \boldsymbol{c}_n^{\mathrm{T}}\end{bmatrix} \tag{6}$$

其中，$1\leqslant n\leqslant N$。假设第d个短波信号源的经纬度为(ω_d,ρ_d)，则该短波辐射源的地心地固坐标为：

$$\boldsymbol{u}_d=\begin{bmatrix}\mu_d\cos(\rho_d)\cos(\omega_d)\\ \mu_d\cos(\rho_d)\sin(\omega_d)\\ \mu_d\sin(\rho_d)\end{bmatrix} \tag{7}$$

其中，$\mu_d=\dfrac{R_{\mathrm{e}}}{\sqrt{1-\mathrm{e}^2(\sin(\rho_d))^2}}$。

根据各测向站对目标测得的测向角度θ_i，求得N条沿地球表面的曲线，交叉定位形成C_N^2个目标估计点$\hat{\boldsymbol{u}}_d=\left[\hat{\boldsymbol{u}}_d^1,\hat{\boldsymbol{u}}_d^2,\cdots,\hat{\boldsymbol{u}}_d^{C_N^2}\right]$。假设目标估计误差服从零均值高斯分布，协方差矩阵为$\boldsymbol{R}_{uu}=\mathrm{E}\left[(\hat{\boldsymbol{u}}_d-\boldsymbol{u}_d)(\hat{\boldsymbol{u}}_d-\boldsymbol{u}_d)^{\mathrm{T}}\right]$，则目标估计的概率密度函数为：

$$p_{\hat{\boldsymbol{u}}_d}(\boldsymbol{\xi})=(2\pi)^{-\frac{n}{2}}(\det[\boldsymbol{R}_{uu}])^{-\frac{1}{2}}\cdot\exp\left\{-\frac{(\boldsymbol{\xi}-\hat{\boldsymbol{u}}_d)^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}(\boldsymbol{\xi}-\hat{\boldsymbol{u}}_d)}{2}\right\} \tag{8}$$

其中，n表示$\boldsymbol{u}_d$的维数。概率密度的等值曲线可描述为：

$$(\boldsymbol{\xi}-\boldsymbol{u}_d)^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}(\boldsymbol{\xi}-\boldsymbol{u}_d)=\kappa \tag{9}$$

其中，κ为任意正常数，由此可以确定曲线表面所包围的n维区域大小。本文设置n=3，其表面为椭圆体。

目标估计值$\hat{\boldsymbol{u}}_d$位于椭圆体内部的概率为：

$$\Pr(\kappa)=\iint_{\Omega}\cdots\int p_{\hat{\boldsymbol{u}}_d}(\boldsymbol{\xi})\mathrm{d}\xi_1\,\mathrm{d}\xi_2\cdots\mathrm{d}\xi_n \tag{10}$$

其中，积分区域$\Omega=\left\{\boldsymbol{\xi}\middle|(\boldsymbol{\xi}-\boldsymbol{u}_d)^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}(\boldsymbol{\xi}-\boldsymbol{u}_d)\leqslant\kappa\right\}$。为将多重积分转换为单重积分，引入向量$\boldsymbol{v}=\boldsymbol{\xi}-\boldsymbol{u}_d$，则式(10)转换为：

$$\Pr(\kappa)=\eta\cdot\iint_{\Omega_1}\cdots\int\exp\left\{-\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}\boldsymbol{v}}{2}\right\}\mathrm{d}v_1\,\mathrm{d}v_2\cdots\mathrm{d}v_n \tag{11}$$

其中，$\eta=(2\pi)^{-\frac{n}{2}}(\det[R_{uu}])^{-\frac{1}{2}}$，积分区域$\Omega_1=\left\{\boldsymbol{v}\middle|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}\boldsymbol{v}\leqslant\kappa\right\}$。由于$\boldsymbol{R}_{uu}^{-1}$是正定矩阵，那么一定存在正交矩阵$\boldsymbol{U}$满足：

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}\boldsymbol{U}=\boldsymbol{\Sigma}^{-1}\Leftrightarrow\boldsymbol{R}_{uu}^{-1}=\boldsymbol{U}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}} \tag{12}$$

根据椭圆体体积证明过程[20]，得到：

$$\Pr(\kappa)=\mathrm{erf}\left(\sqrt{\frac{\kappa}{2}}\right)-\sqrt{\frac{2\kappa}{\pi}}\cdot\exp\left\{-\frac{\kappa}{2}\right\} \tag{13}$$

其中，$\mathrm{erf}(\cdot)$为误差函数，定义为$\mathrm{erf}(x)=\dfrac{2}{\sqrt{\pi}}\cdot\int_0^x\exp(-t^2)\mathrm{d}t$。

下面进行定位误差椭圆概率有效性分析，假设5个测向站的经纬度坐标分别为(116.47°E, 39.72°N)(125.52° E, 44.24° N)(113.55° E, 23.23° N)(119.49° E, 39.38°N)和(121.53°E, 31.46°N)，目标位置(134°E, 28°N)。测向站测向误差σ_θ服从零均值高斯分布，设σ_θ分别为0.3°、0.7°和1.0°，采用交叉定位方式

求得目标估计位置，将交点平均值$\bar{\boldsymbol{u}}_d$替代目标真实值$\boldsymbol{u}_d$进行2 000次蒙特卡罗仿真实验，则误差椭圆概率为：

$$\Pr(\kappa)=\Pr\left[(\hat{\boldsymbol{u}}_d-\bar{\boldsymbol{u}}_d)^{\mathrm{T}}\boldsymbol{R}_{uu}^{-1}(\hat{\boldsymbol{u}}_d-\bar{\boldsymbol{u}}_d)\right] \tag{14}$$

如图4所示，误差椭圆概率随着参数κ的增大而增大，仿真值与理论值具有良好的一致性，从而验证理论分析的有效性。由此可以得出，误差椭圆概率越大，则定位精度越低；误差椭圆概率越小，则定位精度越高。因此，t时刻第i个智能体对信号定位的奖励为：

$$r_{t1}^{i}=100\times\left(1-\Pr(\kappa)\right) \tag{15}$$

（2）检测奖励

在信号检测任务中，考虑2GALE、110A、CLOVER2000、LINK4A和LINK11短波信号。需要说明的是，本文构造的检测奖励模型可根据信号类型的增加进行扩展，为便于表述，此处以5类短波信号为例。如图5所示，采用双塔网络计算信号类型匹配度。该网络由两个相同的子网组成，两个网络共享权重，以确保提取的特征具有一致性。第一个子网的输入为智能体当前状态；第二个子网的输入为5类短波信号的时频图，经过卷积层、池化层等特征提取模块后，得到特征向量$\boldsymbol{v}_1$和$\boldsymbol{v}_2^i$。随后，计算两个向量的余弦相似度作为匹配度，并取匹配度的最大值作为智能体i的检测奖励r_{t2}^{i}，具体为：

$$r_{t2}^{i}=\max\frac{\boldsymbol{v}_1\boldsymbol{v}_2^{i}}{|\boldsymbol{v}_1|\cdot|\boldsymbol{v}_2^{i}|} \tag{16}$$

（3）协作奖励

为避免智能体对信号的重复检测与定位，设t时刻信号j被标记为检测定位已完成的状态为$H_j(t)=1$，此时智能体i对信号j不再进行信号检测与定位，否则$H_j(t)=0$。为实现多智能体协作，智能体i在确认信号j未被其他智能体检测时，可以获得额外协作奖励r_{t3}^{i}。

另外，为保证智能体移动的有效性，对智能体i的每次移动增加移动惩罚r_{t4}^{i}。综上所述，t时刻第i个智能体奖励为：

$$r_t^{i}=\sum_{i=1}^{N}\delta r_{t1}^{i}+(1-\delta)r_{t2}^{i}+r_{t3}^{i}+r_{t4}^{i} \tag{17}$$

其中，$\delta\in[0,1]$为定位奖励权重。因此，多智能体环境中的总奖励为$\boldsymbol{R}_t=\left[r_t^1,r_t^2,\cdots,r_t^n\right]$。

2.2　算法结构

MAPPO-DL通过扩展PPO的框架，结合集中训练与分布执行模式，有效解决多智能体协作或竞争问题。如图6所示，在MAPPO-DL算法的实现

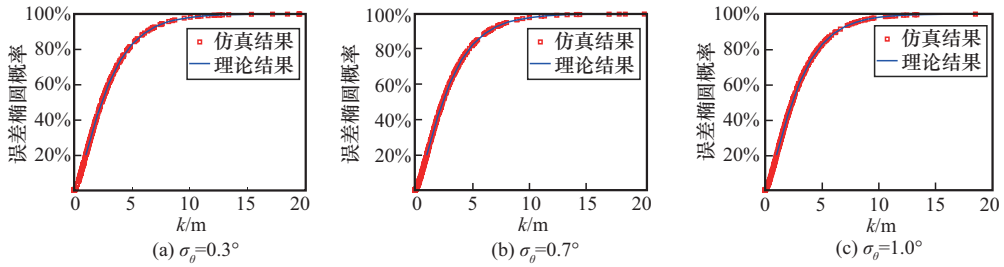


图4　误差椭圆概率仿真结果与理论结果对比

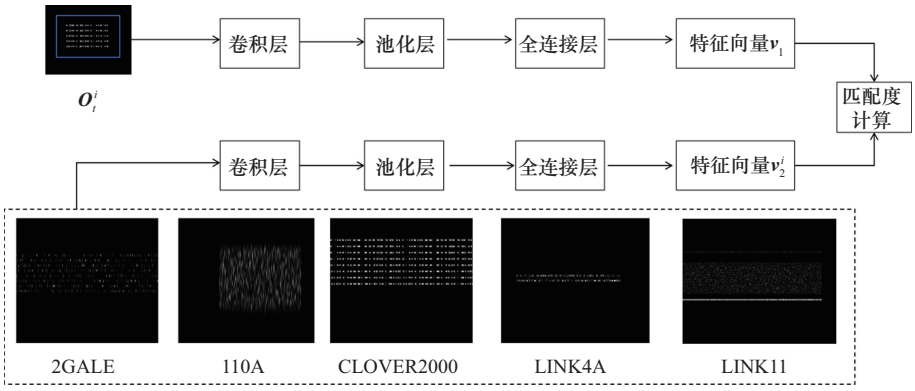


图5　信号匹配度计算过程

中，每个智能体需要训练两个网络，即Actor网络和Critic网络。Actor网络根据贝尔曼方程得到动作价值函数$Q(o_t^i,a_t^i)$，并学习一个从观测o_t^i到动作a_t^i的映射函数，记为：

$$Q(o_t^i,a_t^i)=\mathbb{E}\left[r_t^i+\gamma\max_{a_t^i}Q(o_{t+1}^i,a_{t+1}^i)|o_t^i,a_t^i\right] \quad (18)$$

其中，γ为折扣回报率。Critic网络的目标是学习一个从观测空间$\boldsymbol{O}_t$到价值的映射函数$V_\phi(\boldsymbol{O}_t)$，记为：

$$V_\phi(\boldsymbol{O}_t)=\mathbb{E}\left[r_t^i+\gamma V_\phi(\boldsymbol{O}_{t+1})|\boldsymbol{O}_t\right] \quad (19)$$

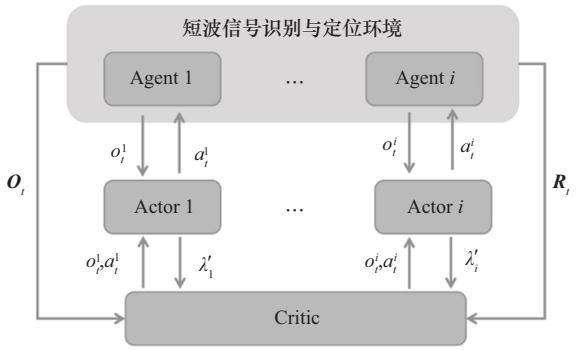


图6　MAPPO-DL算法网络结构

在MAPPO-DL中，每个智能体i的策略由Actor网络决定，Actor网络更新基于以下目标函数，记为：

$$L_i(\lambda_i)=\mathbb{E}\left[\min\left(b_t^i(\lambda_i)A_t^i,\mathrm{clip}(b_t^i(\lambda_i),1-\varepsilon,1+\varepsilon)A_t^i\right)\right] \quad (20)$$

其中，$b_t^i(\lambda_i)=\dfrac{\pi_{\lambda_i}(a_t^i|o_t^i)}{\pi_{\lambda_{i,\mathrm{old}}}(a_t^i|o_t^i)}$，$\varepsilon$为裁剪率，$\mathrm{clip}(x,l,r)=\max(\min(x,r),l)$，即把$x$限制在$[l,r]$内。在短波检测与定位这一特定场景下，策略梯度更新幅度过大，会导致学习过程不稳定。通过引入裁剪机制，控制新旧参数之间的差异大小，在信任域内搜索最优策略，从而提升训练稳定性。A_t^i由集中式Critic网络计算，Critic网络输入为全局状态$\boldsymbol{O}_t$，输出为全局优势值，记为：

$$A_t^i=Q(o_t^i,a_t^i)-V_\phi(\boldsymbol{O}_t),a_t\sim\pi(a_t,o_t),\quad o_{t+1}\sim P(o_{t+1}|o_t,a_t) \quad (21)$$

其中，$P(o_{t+1}|o_t,a_t)$为状态转移概率。另外，为防止智能体陷入次优策略，在Actor网络的损失函数$L_i(\lambda_i)$中加入策略熵项，并乘以系数μ，记为：

$$L'(\lambda_i)=L(\lambda_i)-\mu\sum_{a_t}\pi_{\lambda_i}(a_t^i|o_t^i)\lg(\pi_{\lambda_i}(a_t^i|o_t^i)) \quad (22)$$

Critic网络V_ϕ通过最小化均方误差更新目标函数，具体为：

$$L(\phi)=\mathbb{E}\left[(r_t^i+\gamma V_\phi(\boldsymbol{O}_{t+1})-V_\phi(\boldsymbol{O}_t))^2\right] \quad (23)$$

MAPPO-DL使用卷积网络处理时频图像数据，在时间维度上保留动态特征。MAPPO- DL算法的Actor网络基于局部观测o_t^i生成动作a_t^i，如图7所示；Critic网络利用全局状态s_t优化价值估计V_ϕ，如图8所示。在执行时，各智能体仅依赖局部观测独立决策。连续动作部分输出均值向量，离散动作部分输出多个动作的概率分布。

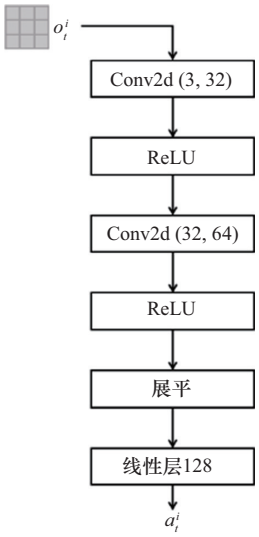


图7　Actor网络结构

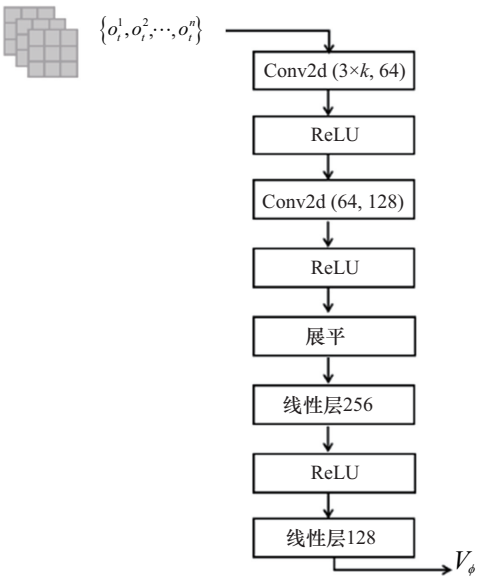


图8　Critic网络结构

2.3 算法流程

算法 1 给出了 MAPPO-DL 算法流程。具体来说，MAPPO-DL 算法先对 Actor 网络和 Critic 网络的参数进行初始化，以缓解训练初期可能出现的梯度消失或爆炸等问题。

算法 1 MAPPO-DL

输入 短波信号检测与定位时频图环境

输出 信号估计类型 Z，估计位置 $\hat{\boldsymbol{u}}_d$

1) 初始化经验回放池 B，初始化 Critic 网络 $V_\phi(a_1,a_2,\cdots,a_N)$ 的参数 ϕ，初始化每个智能体 Actor 网络 $\pi_{\lambda_i}(o_t^i)$ 的参数 λ_i，初始化目标 Critic 网络 $V'_\phi(a_1,a_2,\cdots,a_N)$ 的参数 ϕ。设置裁剪率 ε，Actor 网络学习率 α，Critic 网络学习率 β

2) for EP =1: E do

3) 滤波窗智能体和目标信号随机初始位置，获得初始观测状态 s_0 和各智能体观测 $\{o_0^1,o_0^2,\cdots,o_0^N\}$

4) for SP =1: T do

5) 每个智能体根据局部观测状态 o_t^i，依据 Actor 网络 $\pi_{\lambda_i}(o_t^i)$ 选择动作 a_t^i

6) 执行动作 $a_t=\{a_t^1,a_t^2,\cdots,a_t^N\}$

7) 依据选中时频图 MUSIC 测向结果 θ_i 计算定位奖励 r_{t1}^i，并得到信号估计位置 $\hat{\boldsymbol{u}}_d$

8) 根据信号类型检测结果 Z 计算 r_{t2}^i，得到总奖励 r_t^i，并获得新的智能体观测 o_{t+1}^i

9) 计算广义优势估计 $A_t^i=R_t+\gamma V_\phi(s_{t+1})-V_\phi(s_t)$

10) 将 $(\boldsymbol{O}_t,\boldsymbol{A}_t,\boldsymbol{R}_t,\boldsymbol{O}_{t+1})$ 存储至经验回放池 B

11) 计算 Critic 网络损失 $L(\phi)=\mathbb{E}\left[(\boldsymbol{R}_t+\gamma V_\phi(s_{t+1})-V_\phi(s_t))^2\right]$

12) 计算 Actor 网络损失 $L_i(\lambda_i)=\left[\mathbb{E}\min(b_{i,t}(\lambda_i)A_{i,t},\mathrm{clip}(b_{i,t}(\lambda_i),1-\varepsilon,1+\varepsilon)A_{i,t})\right]$

13) 更新 Actor 网络参数 $\lambda_i\leftarrow\mathrm{Adam}(\lambda_i,\nabla_{\lambda_i}L_i(\lambda_i))$

14) 更新 Critic 网络参数 $\phi\leftarrow\mathrm{Adam}(\phi,\nabla_\phi L(\phi))$

15) 更新目标 Critic 网络 $\phi'\leftarrow\tau\times\phi+(1-\tau)\times\phi$

16) end for

17) end for

在训练阶段，N 个智能体在每回合迭代中收集长度为 T 的步长。智能体 i 根据测向时频图数据获取局部观测 o_t^i，依据 Actor 网络选择动作 a_t^i，实现智能体与环境的交互。随后，每个智能体在交互过程中获取并存储观测状态、动作、即时奖励、下一状态，记作 $(\boldsymbol{O}_t,\boldsymbol{A}_t,\boldsymbol{R}_t,\boldsymbol{O}_{t+1})$。

基于收集的经验回放数据集，采用广义优势估计方法估算优势函数 A_t^i，并利用值归一化方法计算归一化的值函数 V_ϕ。最后，Actor 网络与 Critic 网络均采用 Adam 优化器进行参数更新，同时配合梯度裁剪与衰减学习率，其中梯度更新的幅度由参数 τ 限定，以避免梯度爆炸问题。训练过程持续进行，直至完成预定的训练轮次 E。

2.4 算法性能分析

在 MAPPO 算法中，采用裁剪操作限制策略梯度更新幅度。由于采用裁剪操作，MAPPO 区别于传统信赖域策略优化中使用的库尔贝克–莱布勒（Kullback-Leibler，KL）散度约束，从而显著提升算法的灵活性[21-22]。设 π_λ 表示一个随机策略，MAPPO 通过裁剪 A_{π_λ} 来保证 $\pi_{\lambda_{\mathrm{old}}}$ 和 $\pi_{\lambda_{\mathrm{new}}}$ 之间的差异在合理范围内。因此，MAPPO-DL 的目标函数为：

$$J(\lambda)=\mathbb{E}_{\pi_{\lambda_{\mathrm{old}}}}\left[\min\left\{\frac{\pi_{\lambda_{\mathrm{new}}}}{\pi_{\lambda_{\mathrm{old}}}}A_{\lambda_{\mathrm{old}}},\mathrm{clip}_{1-\varepsilon}^{1+\varepsilon}\left(\frac{\pi_{\lambda_{\mathrm{new}}}}{\pi_{\lambda_{\mathrm{old}}}}\right)A_{\lambda_{\mathrm{old}}}\right\}\right] \tag{24}$$

（1）收敛性分析

验证 MAPPO-DL 的学习收敛性能，即验证当满足裁剪操作中的分布约束时，MAPPO 中策略更新的结果与初始策略梯度方法的结果之间存在约束变化，表示为：

$$1-\varepsilon\leqslant\frac{\pi_{\lambda_{\mathrm{new}}}}{\pi_{\lambda_{\mathrm{old}}}}\leqslant 1+\varepsilon \tag{25}$$

设 P_π 为折扣访问频率，记为：

$$P_\pi(s)=P(s_0=s)+\gamma P(s_1=s)+\gamma^2P(s_2=s)+\cdots \tag{26}$$

其中，$P(s_i=s)$ 表示在采样轨迹的第 i 个时刻遇到状态 s 的概率。为了表示新的策略 $\pi_{\lambda_{\mathrm{new}}}$ 预期价值收益相对旧的策略 $\pi_{\lambda_{\mathrm{old}}}$ 优势，则有：

$$V(\pi_{\lambda_{\text{new}}}) = V(\pi_{\lambda_{\text{old}}}) + \mathbb{E}_{s_0,a_0,\cdots,\pi_{\lambda_{\text{new}}}}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi_{\lambda_{\text{old}}}}(s_t,a_t)\right] = V(\pi_{\lambda_{\text{old}}}) + \sum_s P_{\pi_{\lambda_{\text{new}}}}(s)\sum_a \pi_{\lambda_{\text{new}}}(a|s)A_{\pi_{\lambda_{\text{old}}}}(s,a) \tag{27}$$

综上所述，如果策略更新$\pi_{\lambda_{\text{old}}} \to \pi_{\lambda_{\text{new}}}$在每个状态$s$下都能带来正向的价值期望，即$\sum\pi_{\lambda_{\text{new}}}(a|s)A_{\pi_{\lambda_{\text{old}}}}(s,a) \geqslant 0$，则策略价值函数$V_\phi$不会下降。然而，在近似过程中，由于存在估计和近似误差，通常不可避免地存在某些状态s，使得这些状态的价值期望为负，即$\sum\pi_{\lambda_{\text{new}}}(a|s)A_{\pi_{\lambda_{\text{old}}}}(s,a) < 0$。由于$P_{\pi_\lambda}(s)$对$\pi_{\lambda_{new}}$的相互依赖，直接优化式(24)较困难。因此，考虑引入局部近似方法：

$$G_{\pi_{\lambda_{\text{old}}}}(\pi_{\lambda_{\text{new}}}) = V(\pi_{\lambda_{\text{old}}}) + \sum_s P_{\pi_{\lambda_{\text{old}}}}(s)\sum_a \pi_{\lambda_{\text{new}}}(a|s)A_{\pi_{\lambda_{\text{old}}}}(s,a) \tag{28}$$

其中，使用$\pi_{\lambda_{old}}$代替$\pi_{\lambda_{new}}$。在此基础上，定理1给出策略价值函数V_ϕ改进的下界。

定理1　通过调整MAPPO策略更新过程中$\pi_{\lambda_{\text{new}}}$和$\pi_{\lambda_{\text{old}}}$之间的差异，改进的$V_\phi$下界表示为：

$$G_{\pi_{\lambda_{\text{old}}}}(\pi_{\lambda_{\text{new}}}) - V(\pi_{\lambda_{\text{old}}}) \geqslant \frac{4\gamma\rho}{(1-\gamma)^2}\alpha^2 \tag{29}$$

其中，α为分歧的概率，$\rho = \max\left|A_{\pi_\lambda}(s,a)\right|$，具体证明如附录1所示。

（2）计算复杂度分析

设环境中存在N个智能体滤波窗，每个智能体的Actor网络和Critic网络均为卷积神经网络（convolutional neural network，CNN），网络层数为L，卷积核大小固定，每层输出通道数为C_l，输入时频图尺寸为$H \times W$。CNN的浮点运算次数约为$O(\sum_{l=1}^{L} C_{l-1}\cdot k^2\cdot H_l\cdot W_l)$，其中$k$为卷积核大小，$H_l, W_l$为第$l$层特征图尺寸。由于采用全卷积结构，计算量与输入图像尺寸近似呈线性关系，记为$O(\mathcal{F}(H,W))$。

每个智能体需完成一次Actor前向、Critic前向、优势估计、Critic反向传播。其中，Critic反向传播计算量为Critic前向的2~3倍，因此单智能体单步训练时间复杂度为$O(\mathcal{F}(H,W))$。

由于MAPPO-DL中多智能体共享网络参数，N个智能体在同一环境步内的计算可并行，故单步总时间复杂度仍为$O(\mathcal{F}(H,W))$，与N无关。训练需经历E个回合，每回合最大步长为T，总时间复杂度为$O(E\cdot T\cdot\mathcal{F}(H,W))$。

3　仿真实验及结果分析

3.1　仿真环境设置

本文实验通过模拟产生5类短波信号，并添加噪声，然后进行FFT处理得到时频图。在信号产生阶段，设定时频图持续时间为4 s，信号持续时间为100~500 ms，符号速率f_b为2 400~4 800 Bd，采样频率f_s为120 kHz，FFT点数N_{FFT}为1 024，载波频率f_c为25 kHz。为抑制码间串扰并提高频谱利用率，采用平方根升余弦滚降滤波器进行脉冲成型，余弦滚降系数α为0.8，信号带宽B为4.32~8.64 kHz，信噪比范围为5~25 dB。

强化学习环境为随机信噪比的时频图，并在其上添加2~6个随机分布的短波信号。每张时频图中短波信号的类型、信噪比、带宽、幅度和符号速率等参数均在特定范围内随机变化。时频图大小固定为400×300，利用pygame创建时频图强化学习环境，并随机生成矩形滤波窗作为智能体。信号参数和强化学习环境设置如表1所示。

站点与目标均在(10°N, 100°E)×(40°N, 140°E)的地理区域内选取。环境依赖关系根据第3节所述进行MDP建模。本文实验采用torch1.13.0，计算机显卡配置为NVIDIA GeForce RTX A5000。在训练时，每回合随机初始智能体和信号位置，时频图上的不同信号代表不同的目标。

表1　信号参数和强化学习环境设置

参数	取值	参数	取值
信号符号速率f_b/Bd	2 400~4 800	最大信号持续时间t_s/ms	100~500
采样频率f_s/kHz	120	智能体的最大宽度$w_{\max}$	100
余弦滚降系数α	0.8	智能体的最小宽度$w_{\min}$	10
信号带宽B/kHz	4.32~8.64	智能体的最大高度$h_{\max}$	100
信号信噪比SNR/dB	5~25	智能体的最小高度$h_{\min}$	5

另外，需要对奖励函数中的超参数δ进行取值测试，以保证本文算法的最佳检测效果与定位性能。图9展示了奖励值随δ的变化情况。从图9可以看出，当δ=0.7时，定位奖励、检测奖励、协作奖励和综合奖励值最大。因此，在后续实验中，为获得更好的训练效果，考虑δ取值为0.7。

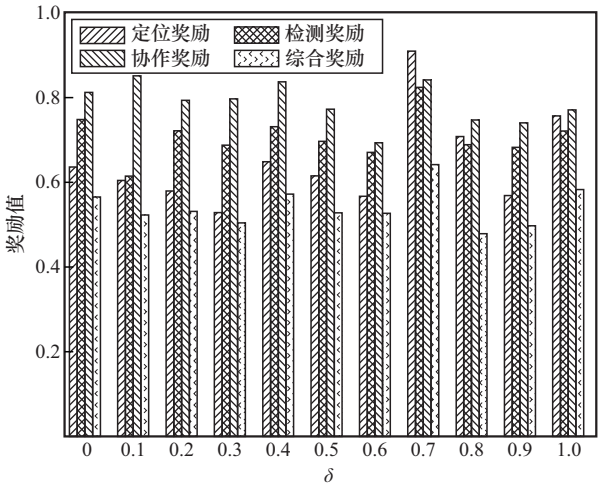


图9　各奖励值随δ变化情况

3.2　基线算法

为验证本文MAPPO-DL算法与所设计奖励函数的有效性，本文分别采用4种MARL基线算法进行对比，分别是多智能体深度确定性策略梯度（multi-agent deep deterministic policy gradient，MADDPG）算法[23]、基于注意力机制的多智能体演员评论家（multi-actor-attention-critic，MAAC）算法[24]、查询–策略对齐（query-policy alignment，QPA）算法[25]和近端策略探索（proximal policy exploration，PPE）算法[26]。

为进一步验证本文所提MAPPO-DL算法对短波信号检测与定位的有效性，设计3种算法对检测与定位模块进行测试，具体如下。

（1）基于目标检测的多智能体深度强化学习（MAPPO-D）：奖励函数只考虑目标检测结果，即定位奖励权重$\delta=0$。

（2）基于目标定位的多智能体深度强化学习（MAPPO-L）：奖励函数只考虑目标信号定位结果，即定位奖励权重$\delta=1$。

（3）基于目标检测与定位的单智能体深度强化学习（PPO-DL）：算法中只有一个智能体，奖励函数同时考虑目标信号检测和定位情况。

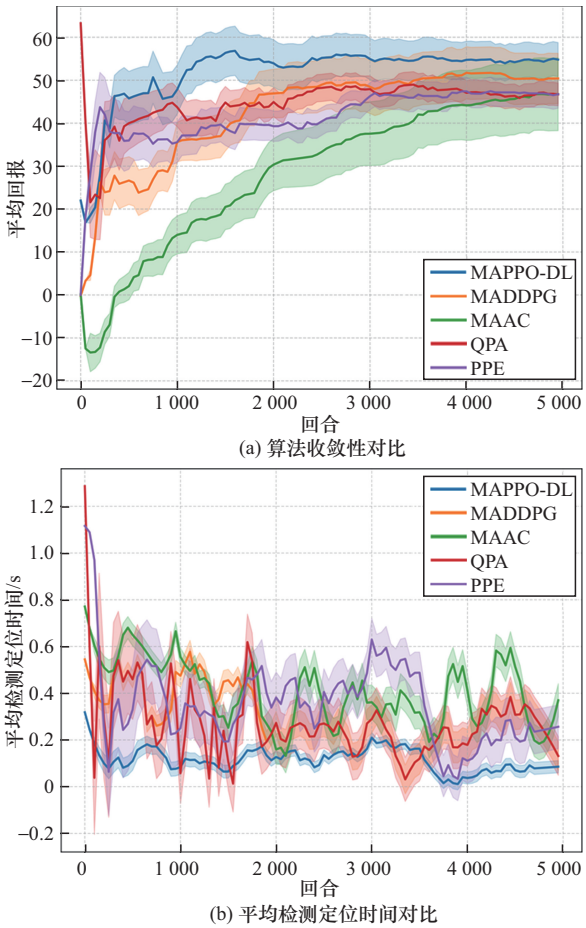
本文算法与基线算法训练过程的参数设置如表2所示。

表2　　强化学习训练参数设置

参数	取值
折扣回报率γ	0.95
裁剪率ε	0.2
价值网络学习率α	0.000 1
策略网络学习率β	0.000 3
经验回放池大小B	1 000
回合数E	5 000
执行最大步长T	50
协作奖励r_{t3}^{i}	20
移动惩罚r_{t4}^{i}	−10

3.3　训练结果对比

图10对比了MAPPO-DL算法与其他MARL算法的收敛性曲线和平均检测定位时间曲线，其中阴影部分表示训练结果的标准差。



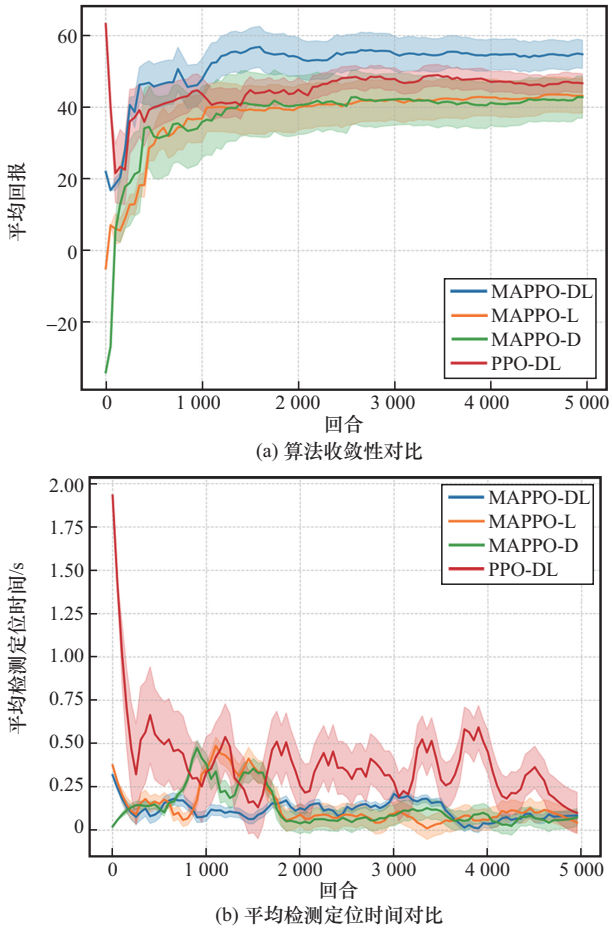
(a) 算法收敛性对比

(b) 平均检测定位时间对比

图10　不同MARL算法训练结果对比

从图 10(a)可以看出，MAPPO-DL 算法收敛后的平均回报最高，且收敛速度较快。由图 10(b)可以看出，MAPPO-DL 算法收敛后的平均检测定位时间最短，且训练过程更稳定。MADDPG 和 MAAC 都是离线更新算法，在多智能体信号检测与定位环境训练中收敛速度较慢，收敛后的平均奖励也低于本文算法。QPA 算法更专注于在当前策略分布内学习奖励，但依赖于人类反馈。PPE 算法虽然可扩展在缺乏采样的策略近端区域的探索范围，但它对经验缓冲区的要求更高。

图11展示了MAPPO-DL、MAPPO-D、MAPPO-L 和 PPO-DL 算法的收敛性曲线和平均检测定位时间曲线。



(a) 算法收敛性对比

(b) 平均检测定位时间对比

图 11　不同算法在不同模块的训练结果对比

从图 11(a)可以看出，4 种强化学习算法均能在多次训练后收敛。从收敛后平均回报来看，MAPPO-DL 算法收敛后的平均回报最大，其次是 PPO-DL 算法，原因在于 PPO-DL 算法的奖励函数考虑了目标信号检测和定位情况；MAPPO-D 和 MAPPO-L 收敛后的平均回报基本一致。从收敛速度来看，4 种算法的收敛速度相当，但仍可看出 MAPPO-D 和 MAPPO-L 收敛速度较快，MAPPO-DL 收敛速度较慢，原因在于多智能体之间的协作影响收敛速度。同时，裁剪操作保证策略更新，使奖励曲线较平滑，避免出现振荡。

从图 11(b)可以看出，多智能体算法检测定位时间均比单智能体算法短，MAPPO 算法的波动较小。PPO-DL 算法的检测定位时间较长且波动较大，原因在于单智能体强化学习在处理多目标问题时，需要在时频图上逐一寻找信号，再完成检测与定位工作。

多智能体强化学习方法中的协作策略可以有效共享智能体之间的信息，进一步协调智能体的动作。因此，MAPPO-DL 能有效克服单智能体深度强化学习算法的局限。

4　模型性能测试

在完成深度强化学习算法训练后，选择 MAPPO-DL 和基线算法第 5 000 步训练得到的策略网络参数作为测试初始模型，在不同多目标场景下进行 500 次信号检测与定位，对比分析目标的定位时间、目标检测率和定位误差。

4.1　关于信号检测与定位时间的性能测试

图 12 展示不同算法测试时间和测试平均奖励的变化。在测试集上运行 500 次，记录从输入时频图到输出检测定位结果的平均时间。如表 3 所示，MAPPO-DL 由于采用并行计算与参数共享，单步训练时间低于 MADDPG、MAAC、QPA、PPE 等独立更新算法，收敛总时间虽然略高于 MAPPO-D 和 MAPPO-L，但测试时间明显低于其他算法。

表 3　　不同算法信号检测与定位时间

算法	单步训练时间/ms	测试时间/ms
MAPPO-DL	12.1	0.02
MAPPO-L	10.2	0.12
MAPPO-D	10.3	0.10
PPO-DL	9.7	0.14
MADDPG	17.4	0.08
MAPPO	16.2	0.06
QPA	17.8	0.08
PPE	19.6	0.09

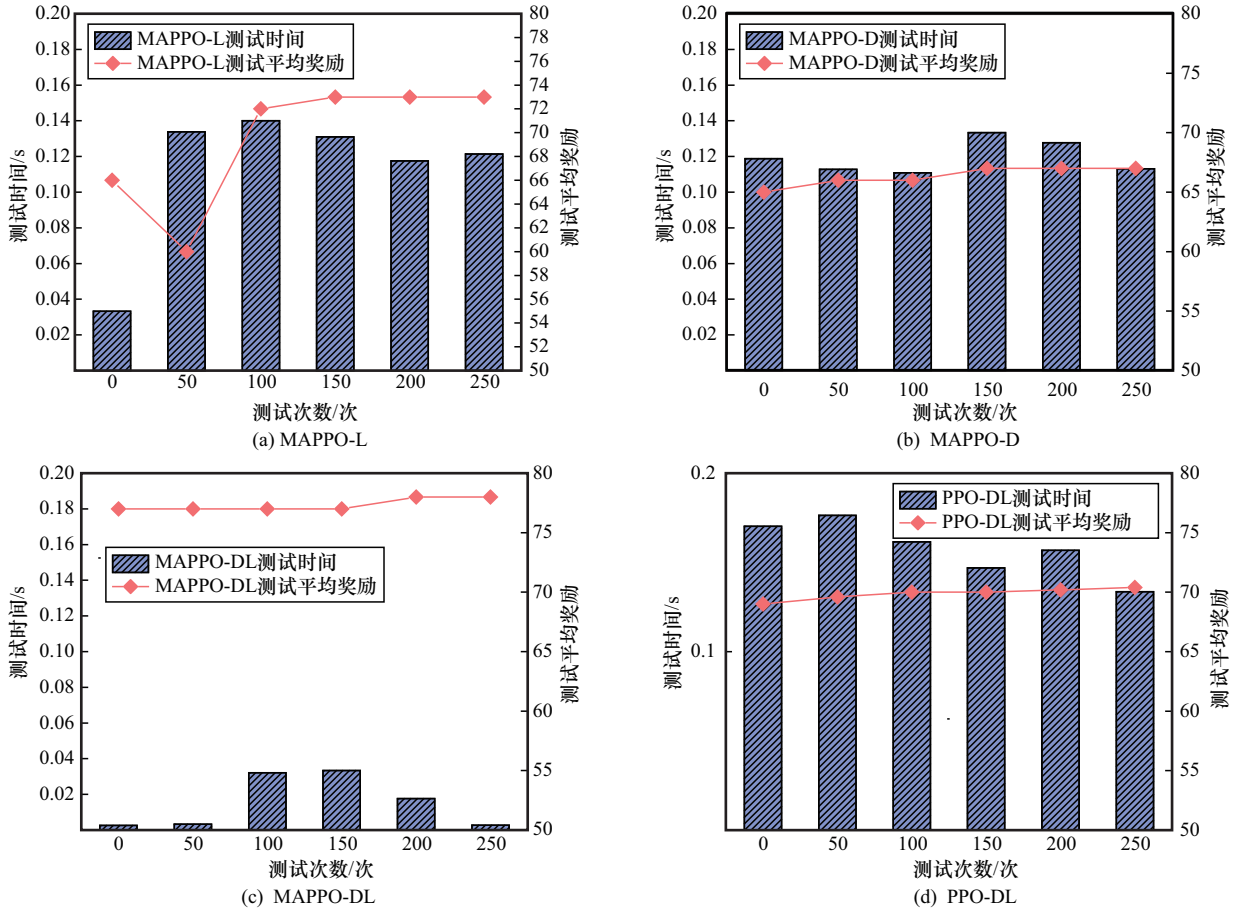


图 12　不同算法测试时间和测试平均奖励

从测试时间可以看出，图 12(c)中 MAPPO-DL 算法的测试时间最短，平均测试时间为 0.02 s；图 12(d)中 PPO-DL 算法的测试时间最长；图 12(a)中 MAPPO-L 算法与图 12(b)中 MAPPO-D 算法的定位时间相当，平均测试时间为 0.13 s。

同样，单智能体的 PPO-DL 算法在处理多信号检测与定位问题时，需要逐一进行测试，导致信号检测与定位的时间较长，平均测试时间为 0.15 s。因此，MAPPO-DL 平均测试时间约为 0.02 s，该算法将信号检测与定位时间平均缩短了 0.12 s。相较于其他多智能体算法，MAPPO-DL 算法的性能更优，满足短波信号实时检测需求。

从测试平均奖励来看，4 种算法的测试平均奖励均保持相对稳定，这说明保存的算法模型已得到稳定训练。相较于单智能体强化学习方法，多智能体强化学习方法无须逐一检测信号，通过智能体之间的协作可完成多目标的实时检测与定位工作。

4.2　关于定位误差的性能测试

在进行定位误差测试时，考虑使用的测向站点与目标经纬度坐标如表 4 所示。测向站与目标相对位置如图 13 所示，假设使用 3 个场景测试定位误差，具体如下。

表 4　　测向站点与目标经纬度坐标

站点和目标	经度	纬度
站点 1	117.00°	36.30°
站点 2	106.82°	14.75°
站点 3	104.95°	43.10°
站点 4	118.18°	25.54°
站点 5	124.17°	43.05°
目标 1	103.40°	34.69°
目标 2	97.40°	35.42°
目标 3	101.40°	32.69°
目标 4	95.70°	33.32°
目标 5	98.57°	30.26°
目标 6	92.82°	31.54°
目标 7	96.50°	28.45°

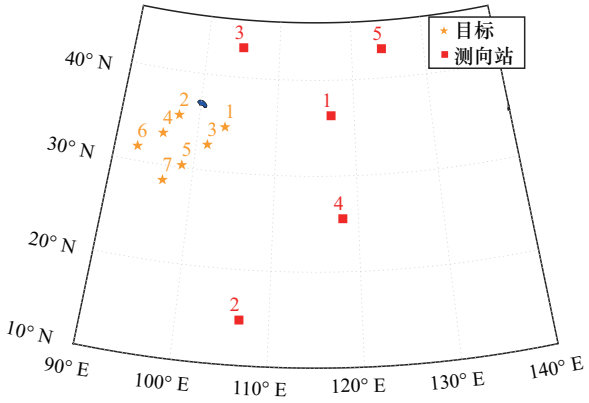


图 13　测向站与目标相对位置

（1）场景 1：5 个测向站和 2 个目标，测向站为 1、2、3、4 和 5，目标选择目标 1 和目标 2。

（2）场景 2：5 个测向站和 2 个目标，测向站为 1、2、3、4 和 5，目标选择目标 3 和目标 4。

（3）场景 3：5 个测向站和 3 个目标，测向站为 1、2、3、4 和 5，目标选择目标 5、目标 6 和目标 7。

为进一步突出多智能体的协作优势，4.2 节仅使用 MAPPO-DL 网络模型在 3 种场景下进行定位性能测试。由于目标定位结果以各个测向站交点平均值作为目标估计值，因此目标定位结果依赖于 MUSIC 算法的测向精度。进一步地，目标定位精度与信号信噪比成正比。因此，采用均方根误差（root mean square error，RMSE）来对比算法对目标的定位性能，计算式为：

$$\mathrm{RMSE}(\boldsymbol{u}_k)=\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left\|\hat{\boldsymbol{u}}_k-\boldsymbol{u}_k\right\|_2^2} \tag{30}$$

其中，K=2 000 为蒙特卡罗次数，$\hat{\boldsymbol{u}}_k$ 为第 k 次测试中目标位置的估计值，$\boldsymbol{u}_k$ 为目标位置的真实值。

图 14 展示了 3 种场景下各目标的定位误差随信噪比的变化。从图 14 可以看出，随着信噪比的增加，定位误差逐渐减小。在图 14(a)所示的场景 1 中，低信噪比时目标 1 的定位误差大于目标 2，高信噪比时两者的定位误差相当。在图 14(b)所示的场景 2 中，低信噪比时目标 3 的定位误差小于目标 4。在图 14(c)所示的场景 3 中，低信噪比时目标 6 的定位误差最大，目标 5 的定位误差最小，高信噪比时三者的定位误差相当。

因此，MAPPO-DL 算法能适应动态信号环境，具有较好的稳健性。通过以上分析可以得出，奖励函数中定位误差椭圆概率的设计可以有效引导智能体学习较高定位精度的最优策略。

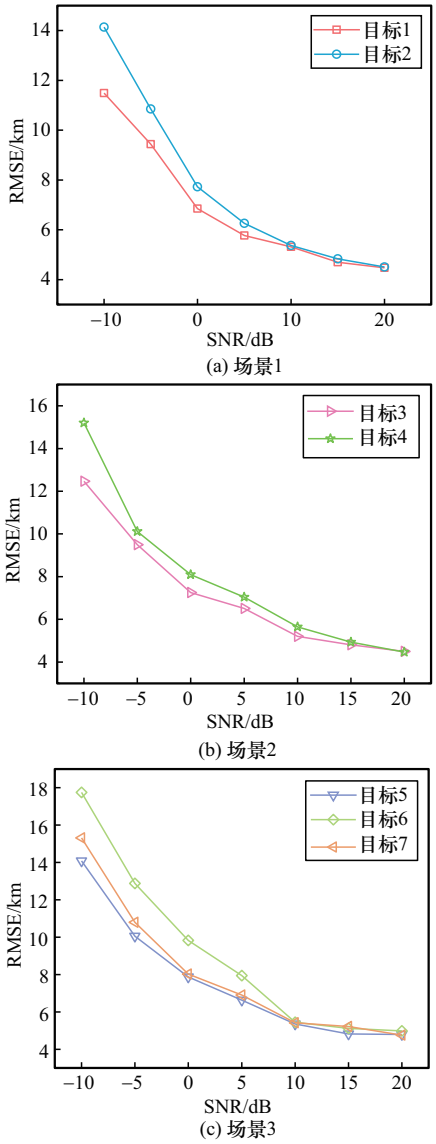


图 14　不同场景定位误差随信噪比变化

进一步分析目标与测向站的平均距离对 MAPPO-DL 算法的影响，图 15 展示了 7 个目标与 5 个测向站之间的平均距离，目标 1、3、5 和 7 与测向站之间的平均距离较大，目标 2、4 和 6 与测向站之间的平均距离较小。另外，图 15 还对比了 7 个目标的 RMSE，可以看出，当信噪比为−10 dB 时，目标 2、4 和 6 的 RMSE 明显大于目标 1、3、5、7；当信噪比为 10 dB 时，目标 2、4 和 6 的 RMSE 与目标 1、3、5、7 相差较小。随着目标与测向站的平均距离增大，RMSE 逐渐增大，且低信噪比时的变化趋势比高信噪比时明显。因此，在低信噪比时，MAPPO-DL 的定位精度对目标与测向站的远近较敏感。

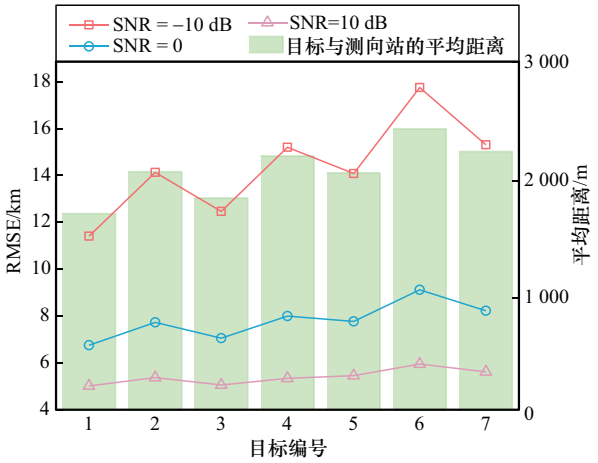


图 15　目标在不同信噪比下定位误差变化

假设目标 1、3、5、7 的位置点构成某一移动目标 A 的轨迹，目标 2、4 和 6 构成另一移动目标 B 的轨迹，图 16(a)和图 16(b)分别展示目标 A 和目标 B 的轨迹在不同测向误差下本文算法与 TWLS 算法[1]、ICWLS 算法[27]的 RMSE 以及克拉美罗下界（Cramer-Rao lower bound，CRLB）的变化。从图 16 可以看出，本文所提 MAPPO-DL 算法的 RMSE 比其他两种传统方法更小，与 CRLB 更接近。通过上述实验，验证了本文算法在信号时变条件下的检测与定位鲁棒性。

4.3　关于多目标检测与定位准确率的性能测试

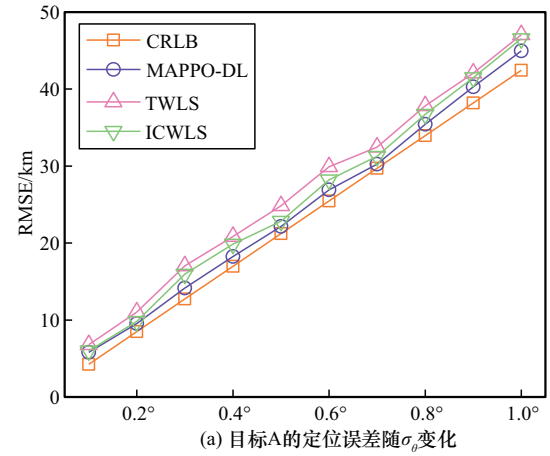
多目标检测与定位准确率可以反映算法的检测正确率和定位精度，目标检测与定位准确率 P_l 定义为：

$$P_l = \frac{N_{\mathrm{DL}}}{N} \tag{31}$$

其中，$N = 500$ 为测试样本数，N_{DL} 为检测与定位成功的样本数。

为全面评估 MAPPO-DL 算法在不同信噪比条件下的定位精度，固定 $\sigma_\theta = 0.5°$，采用平均精度均值（mean average precision，mAP）[28]评价目标检测与定位精度。设置多个定位误差阈值，并计算各阈值下的定位准确率，取其平均值作为综合定位精度指标。

如图 17 所示，当定位误差阈值大于 25 km 时，mAP 大于 0.8。因此，当目标定位误差小于 25 km 时判定为定位成功。需要说明的是，短波定位的相对误差通常小于 3%，即满足短波定位需求[4,29]。在本文实验场景中，测向站与目标之间的平均距离为 2 015.08 km，所采用的阈值为25 km，误差阈值在该距离上的相对误差为1.24%，满足短波定位需求。



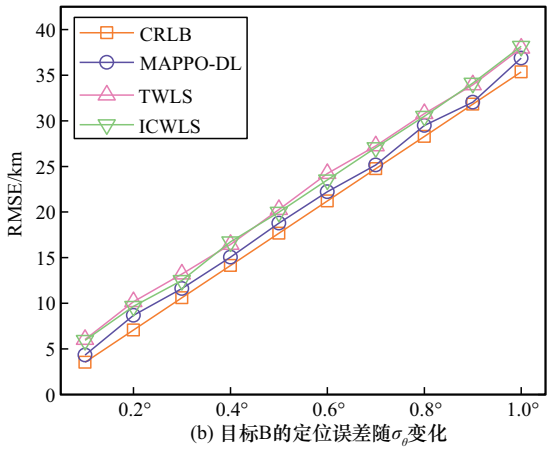


图 16　目标 A 和目标 B 的轨迹在不同测向误差 σ_θ 下定位误差变化

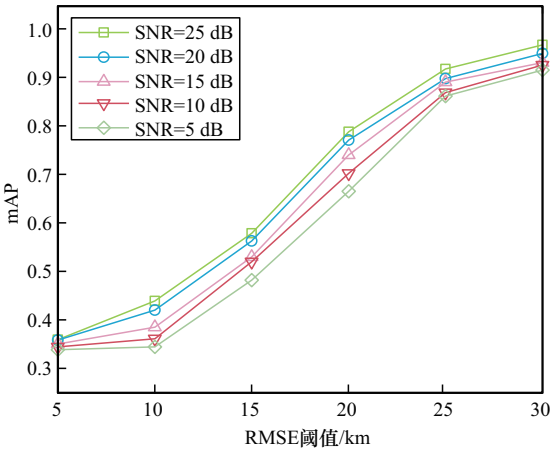


图 17　不同目标数量下的目标检测与定位准确率

图 18 对比了训练前后模型对不同目标数量下的目标检测与定位准确率。图 18(a)展示了训练前 4 种算法在不同目标数量下的检测率均较低，说明深度强化学习模型在与环境交互之前未学到策略，只能随机输出结果。图 18(b)展示了训练后模型对不同目标数量的目标检测与定位准确率，MAPPO-DL 的目

标检测与定位准确率最高，达到92%，其次是MAPPO-D的目标检测与定位准确率为70%，PPO-DL的目标检测与定位准确率为62%，MAPPO-L的目标检测与定位准确率仅为40%。主要原因在于MAPPO-L算法在训练时未考虑目标检测结果的约束，因此，测试时对信号检测的表现较差。

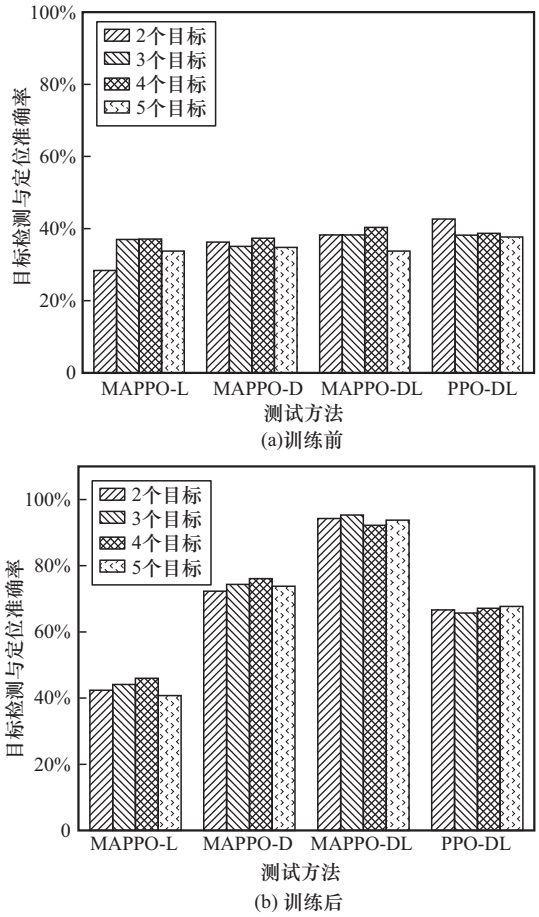


图18　训练前后不同目标数量下的目标检测与定位准确率

相较于目标检测与定位准确率最佳的基线算法，MAPPO-DL算法的目标检测与定位准确率提升了22%。目标检测与定位准确率的提升有效验证了信号匹配度奖励机制能够通过强化特征提取网络，实现对正确信号的检测。

4.4　关于多目标检测与定位可扩展性分析

本节旨在探究算法在更大规模协作场景下的扩展性，分别设置不同智能体数量，固定不同目标数量，训练至收敛，并测试模型的目标检测与定位准确率。如图19所示，由于采用集中训练与分布式执行，随着目标数量的增加，目标检测与定位准确率逐渐下降。本文所提MAPPO-DL算法的目标检测与定位准确率明显高于其他MARL算法，MAPPO-DL的策略熵能够防止训练进入次优策略，增加了算法的可扩展性。然而，当目标数量超过8个时，MADRL算法的性能会由于维度灾难而显著下降，主要原因是Critic网络直接拼接所有智能体观测的方式在高维输入下难以有效区分各智能体的贡献。

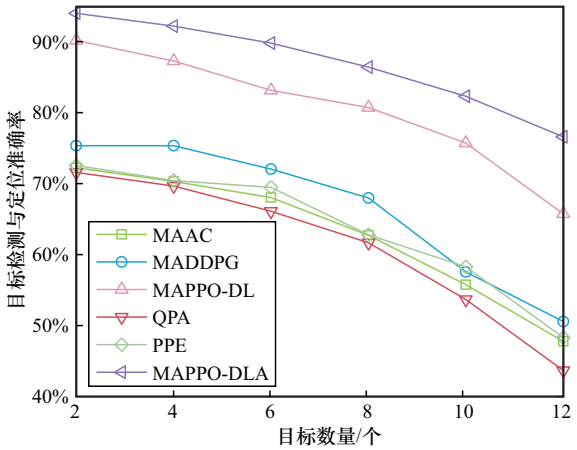


图19　不同目标数量下的目标检测与定位准确率

为缓解上述问题，考虑在MAPPO-DL的Critic网络中引入多头注意力机制，称为MAPPO-DLA（MAPPO-DL with attention mechanism）。通过自适应学习不同智能体观测的权重，更有效地提取协作信息，缓解维度灾难带来的性能瓶颈。而MAPPO-DLA在10个目标的场景下，比MAPPO-DL的准确率提升了约4%，可以有效抑制冗余信息干扰。因此，在基于强化学习的短波定位工作中引入注意力机制可以作为未来工作的重要方向之一。

5　结束语

本文提出一种基于MAPPO的短波信号自主检测与定位方法，通过设计多智能体强化学习环境，运用深度强化学习算法探索试错，模拟短波信号自主检测与定位工作。MAPPO-DL可以快速实现目标的测向定位，在保持高检测定位精度的同时，具备较低的计算开销与良好的可扩展性，对后续精准定位目标具有一定参考价值。在仿真阶段测试MAPPO-DL方案的实时性和有效性，该方法基于深度强化学习，不依赖大量人工标注，能够实现在线多回合自主学习。在未来工作中，将继续研究基于深度强化学习的目标跟踪与轨迹分选任务。

附录1 定理1证明

设 $D=\left(1+\gamma P_{\pi_{\lambda_{\text{old}}}}+\left(\gamma P_{\pi_{\lambda_{\text{old}}}}\right)^2+\cdots\right)=\left(1-\gamma P_{\pi_{\lambda_{\text{old}}}}\right)^{-1}$，$\tilde{D}=\left(1+\gamma P_{\pi_{\lambda_{\text{new}}}}+\left(\gamma P_{\pi_{\lambda_{\text{new}}}}\right)^2+\cdots\right)=\left(1-\gamma P_{\pi_{\lambda_{\text{new}}}}\right)^{-1}$，约定状态空间的密度$\boldsymbol{\chi}$是一个向量，状态空间上的奖励函数$\boldsymbol{r}$是一个对偶向量，即向量上的线性泛函数。因此，$\boldsymbol{r\chi}$是一个标量，表示在密度$\boldsymbol{\chi}$下的期望奖励，则 $V\left(\pi_{\lambda_{\text{old}}}\right)=\boldsymbol{r}D\boldsymbol{\chi}_0$，$V\left(\pi_{\lambda_{\text{new}}}\right)=c\tilde{D}\boldsymbol{\chi}_0$。设$\Delta=P_{\pi_{\lambda_{\text{new}}}}-P_{\pi_{\lambda_{\text{old}}}}$，$V\left(\pi_{\lambda_{\text{new}}}\right)-V\left(\pi_{\lambda_{\text{old}}}\right)=\boldsymbol{r}\left(\tilde{D}-D\right)\boldsymbol{\chi}_0$进行界定，需要从一些标准的扰动理论进行推导，得到：

$$D^{-1}-\tilde{D}^{-1}=\left(1-\gamma P_{\pi_{\lambda_{\text{old}}}}\right)-\left(1-\gamma P_{\pi_{\lambda_{\text{new}}}}\right)=\gamma\Delta \tag{32}$$

将式(32)左乘D，右乘$\tilde{D}$，得到：

$$\tilde{D}-D=\gamma D\Delta\tilde{D}\Rightarrow\tilde{D}=D+\gamma D\Delta\tilde{D} \tag{33}$$

将式(33)右侧代入$\tilde{D}$得到：

$$\tilde{D}=D+\gamma D\Delta D+\gamma^2 D\Delta D\Delta\tilde{D} \tag{34}$$

最终得到：

$$V\left(\pi_{\lambda_{\text{new}}}\right)-V\left(\pi_{\lambda_{\text{old}}}\right)=\boldsymbol{r}\left(\tilde{D}-D\right)\boldsymbol{\chi}=$$
$$\gamma\boldsymbol{r}D\Delta D\boldsymbol{\chi}_0+\gamma^2\boldsymbol{r}D\Delta D\tilde{D}\boldsymbol{\chi}_0 \tag{35}$$

首先，考虑主导项$\gamma\boldsymbol{r}D\Delta D\boldsymbol{\chi}_0$。显然，$\boldsymbol{r}D=v$，即无穷远状态值函数，同时$D\boldsymbol{\chi}_0=\boldsymbol{\chi}_\pi$。因此，可以得出$\gamma cD\Delta D\boldsymbol{\chi}_0=\gamma v\Delta\boldsymbol{\chi}_\pi$。下面将证明该表达式等于期望的优势$L_\pi\left(\pi_{\lambda_{\text{new}}}\right)-L_\pi\left(\pi_{\lambda_{\text{old}}}\right)$，具体为：

$$\begin{aligned}
&L_\pi\left(\pi_{\lambda_{\text{new}}}\right)-L_\pi\left(\pi_{\lambda_{\text{old}}}\right)=\\
&\sum_s\boldsymbol{\chi}_\pi(s)\sum_a\left(\pi_{\lambda_{\text{new}}}(a|s)-\pi_{\lambda_{\text{old}}}(a|s)\right)A_\pi(s,a)=\\
&\sum_s\boldsymbol{\chi}_\pi(s)\sum_a\left(\pi_\theta(a|s)-\pi_{\tilde{\theta}}(a|s)\right)\cdot\\
&\left[r(s)+\sum_{s'}p(s'|s,a)\gamma v(s')-v(s)\right]=\\
&\sum_s\boldsymbol{\chi}_\pi(s)\sum_{s'}\sum_a\left(\pi_{\lambda_{\text{old}}}(a|s)-\pi_{\lambda_{\text{new}}}(a|s)\right)\\
&p(s'|s,a)\gamma v(s')=\\
&\sum_s\boldsymbol{\chi}_\pi(s)\sum_{s'}\left(p_\pi(s'|s)-p_{\pi_{\lambda_{\text{new}}}}(s'|s)\right)\gamma v(s')=\gamma v\Delta\boldsymbol{\chi}_\pi
\end{aligned} \tag{36}$$

然后，对$O\left(\Delta^2\right)$项$\gamma^2\boldsymbol{r}D\Delta D\Delta\pi_{\lambda_{\text{new}}}\boldsymbol{\chi}$进行界定。考虑$\gamma\boldsymbol{r}D\Delta=\gamma v\Delta$和对偶向量的分量$s$，则有：

$$\begin{aligned}
&\left|\left(\gamma v\Delta\right)_s\right|=\left|\sum_a\left(\pi_{\lambda_{\text{new}}}(s,a)-\pi_{\lambda_{\text{old}}}(s,a)\right)Q_\pi(s,a)\right|=\\
&\left|\sum_a\left(\pi_{\lambda_{\text{new}}}(s,a)-\pi_{\lambda_{\text{old}}}(s,a)\right)A_\pi(s,a)\right|\leqslant\\
&\sum_a\left|\pi_{\lambda_{\text{new}}}(s,a)-\pi_{\lambda_{\text{old}}}(s,a)\right|\cdot\max_a\left|A_\pi(s,a)\right|\leqslant 2\alpha\rho
\end{aligned} \tag{37}$$

其中，最后一步采用总变差散度的定义以及$\rho=\max\left|A_\pi(s,a)\right|$的定义，$\alpha$为不同意概率，通过$\ell_1$算子范数对另一部分$D\Delta\tilde{D}\boldsymbol{\chi}$进行有界处理，记为：

$$\|A\|_1=\sup_\rho\left\{\frac{\|A\boldsymbol{\chi}\|_1}{\|\boldsymbol{\chi}\|_1}\right\} \tag{38}$$

由于$\|D\|_1=\|\tilde{D}\|_1=\frac{1}{1-\gamma}$和$\|\Delta\|_1=2\alpha$，依此可得：

$$\|D\Delta\tilde{D}\boldsymbol{\chi}\|_1\leqslant\|D\|_1\|\Delta\|_1\|\tilde{D}\|_1\|\boldsymbol{\chi}\|_1=\frac{1}{1-\gamma}\cdot 2\alpha\cdot\frac{1}{1-\gamma}\cdot 1 \tag{39}$$

因此，最终证得：

$$\begin{aligned}
&\gamma^2\left|\boldsymbol{r}D\Delta D\Delta\tilde{D}\boldsymbol{\chi}\right|\leqslant\gamma\|\gamma\boldsymbol{r}D\Delta\|_\infty\|D\Delta\tilde{D}\boldsymbol{\chi}\|_1\leqslant\\
&\gamma\|v\Delta\|_\infty\|D\Delta\tilde{D}\boldsymbol{\chi}\|_1\leqslant\\
&\gamma\cdot 2\alpha\rho\cdot\frac{2\alpha}{\left(1-\gamma\right)^2}=\frac{4\gamma\rho}{\left(1-\gamma\right)^2}\cdot\alpha^2
\end{aligned} \tag{40}$$

参考文献：

[1] Wang D, Yin J X, Tang T, et al. A two-step weighted least-squares method for joint estimation of source and sensor locations: a general framework[J]. Chinese Journal of Aeronautics, 2019, 32(2): 417-443.

[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.
Wang D, Yin J X, Zhu Z L. Novel cooperative localization method of over-the-horizon shortwave emitters based on direction-of-arrival and time-difference-of-arrival measurements[J]. Scientia Sinica (Informationis), 2022, 52(11): 1942-1973.

[3] 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.

[4] Xia N, Xing B H. A direct localization method for HF source geolocation and experimental results[J]. IEEE Antennas and Wireless Propagation Letters, 2021, 20(5): 728-732.

[5] 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.

[6] Subramaniam A M, Gerogiannis A, Hare J Z, et al. Track-MDP: reinforcement learning for target tracking with controlled sensing[C]//Proceedings of the ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2025: 1-5.

[7] Caicedo J C, Lazebnik S. Active object localization with deep reinforcement learning[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 2488-2496.

[8] Bellver M, Giro X, Marques F, et al. Hierarchical object detection with deep reinforcement learning[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2016: 1-9.

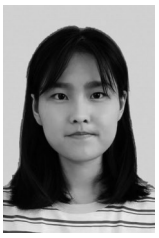
[9] Hao X Y, Yang S Y, Liu R Y, et al. VSLM: virtual signal large model for few-shot wideband signal detection and recognition[J]. IEEE Transactions on Wireless Communications, 2025, 24(2): 909-925.

[10] Li Y, Hu X, Zhuang Y, et al. Deep reinforcement learning (DRL): another perspective for unsupervised wireless localization[J]. IEEE Internet of Things Journal, 2020, 7(7): 6279-6287.

[11] Paul A, Singh K, Kaushik A, et al. Quantum-enhanced DRL optimization for DoA estimation and task offloading in ISAC systems[J]. IEEE Journal on Selected Areas in Communications, 2025, 43(1): 364-381.

[12] Zhao L, Liu X C, Shang T. Maximizing coverage in UAV-based emergency communication networks using deep reinforcement learning[J]. Signal Processing, 2025, 230: 109844.
[13] 邓钰洋, 芦天亮, 李知皓, 等. 融合 GAT 与可解释 DQN 的 SQL 注入攻击检测模型[J]. 信息网络安全, 2026, 26(1): 150-167.
Deng Y Y, Lu T L, Li Z H, et al. A SQL injection attack detection model integrating GAT and interpretable DQN[J]. Netinfo Security, 2026, 26(1): 150-167.
[14] Upadhyay P, Marriboina V, Goyal S J, et al. An improved deep reinforcement learning routing technique for collision-free VANET[J]. Scientific Reports, 2023, 13: 21796.
[15] Xue W T, Wu H X, Ye H, et al. An improved proximal policy optimization method for low-level control of a quadrotor[J]. Actuators, 2022, 11 (4): 105.
[16] Wang R, Xu C, Sun J, et al. Cooperative localization for multi-agents based on reinforcement learning compensated filter[J]. IEEE Journal on Selected Areas in Communications, 2024, 42(10): 2820-2831.
[17] Wang R, Sun J, Xu C, et al. Reinforcement learning compensated filter for multi-agents cooperative localization[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 101-105.
[18] Alagha A, Singh S, Mizouni R, et al. Target localization using multi-agent deep reinforcement learning with proximal policy optimization[J]. Future Generation Computer Systems, 2022, 136: 342-357.
[19] 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: CN110045322A[P]. 2019-03-21.
Tang T, Wang D, Yang B, et al. A short-wave automatic direction finding method based on intelligent recognition of high-resolution direction finding time-frequency spectrum: CN110045322A[P]. 2019-03-21.
[20] 王鼎, 吴瑛, 张莉. 无线电测向与定位理论及方法[M]. 北京: 国防工业出版社, 2016.
Wang D, Wu Y, Zhang L. Radio direction finding and localization[M]. Beijing: National Defense Industry Press, 2016.
[21] Schulman J, Levine S, Moritz P, et al. Trust region policy optimization[PP]. V5. (2017-04-20)[2026-01-03]. arXiv: arXiv.1502.05477.
[22] Gu Y, Cheng Y H, Philip C C L, et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7): 4600-4610.
[23] Lowe R, Wu Y, Tamar A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6382-6393.
[24] Iqbal S, Sha F. Actor-attention-critic for multi-agent reinforcement learning[PP]. V2. (2019-05-27)[2026-01-03]. arXiv: arXiv.1810.02912.
[25] Hu X, Li J X, Zhan X Y, et al. Query-policy misalignment in preference-based reinforcement learning[PP]. V3. (2024-07-05)[2026-01-03]. arXiv: arXiv.2305.17400.
[26] Zhu Y W, Liu J Y, Gu P J, et al. Improving reward models with proximal policy exploration for preference-based reinforcement learning[C]//Proceedings of the 39th Annual Conference on Neural Information Processing Systems. Massachusetts: MIT Press, 2025: 1-13.
[27] Yang Y, Zheng J B, Liu H W, et al. Optimal sensor placement and velocity configuration for TDOA-FDOA localization and tracking of a moving source[J]. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60(6): 8255-8272.
[28] Yang D F, Solihin M I, Zhao Y W, et al. Model compression for real-time object detection using rigorous gradation pruning[J]. iScience, 2025, 28(1): 111618.
[29] 张旭辉, 姜春华, 刘桐辛, 等. 电离层虚高对超视距雷达多站联合定位精度的影响[J]. 电波科学学报, 2022, 37(5): 761-767, 792.
Zhang X H, Jiang C H, Liu T X, et al. Effect of the ionospheric virtual height on the joint positioning accuracy of multi-station over-the-horizon radar system[J]. Chinese Journal of Radio Science, 2022, 37(5): 761-767, 792.

[作者简介]

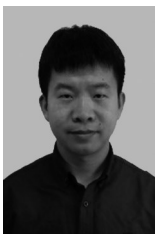


冯祺玥（2002– ），女，信息工程大学博士生，主要研究方向为强化学习、阵列信号处理和无源定位。

唐涛（1981– ），男，博士，信息工程大学教授、博士生导师，主要研究方向为阵列信号处理、无源定位。

张昀普（1995– ），男，博士，信息工程大学讲师，主要研究方向为强化学习、资源调度。

王鼎（1982– ），男，博士，信息工程大学教授、博士生导师，主要研究方向为阵列信号处理、无源定位。



吴志东（1986– ），男，博士，信息工程大学副教授、硕士生导师，主要研究方向为雷达信号处理、无源定位。