采用多智能体强化学习的多传感器协同控制策略

张哲源[1]，冯文晖[1]，连峰[1]，刘霞[1]，阴志鹏[2]
（1. 西安交通大学电信学部，710049，西安；
2. 河北新天科创新能源技术有限公司，075000，河北张家口）

摘要：针对复杂对抗环境下多传感器系统在任务执行与威胁规避过程中的协同决策问题，提出一种采用多智能体强化学习的异构传感器协同控制策略。首先，构建包含主传感器、次传感器、威胁目标和固定监视目标的二维对抗场景，建立传感器机动、辐射控制、目标探测以及威胁目标锁定与命中机制模型，并将多传感器协同管理问题建模为多智能体部分可观测马尔可夫决策过程。其次，在集中训练、分散执行框架下，采用多智能体近端策略优化（MAPPO）算法实现协同决策，其中主传感器通过联合优化平台机动策略与辐射参数（功率与波束指向）完成穿越与监视任务，次传感器通过机动协同与辐射诱引分担主传感器威胁风险。最后，针对强对抗和稀疏反馈条件，设计兼顾任务收益与生存风险的奖励函数，并采用三阶段课程学习策略提升策略训练效率。在包含 1 个主传感器、2 个次传感器和 3 个威胁目标的仿真场景中进行 50 回合对比评估，结果显示：在覆盖阈值为 0.75 时，所提方法的任务成功率达到 82%，高于威胁感知启发式策略（66%）、独立近端策略优化策略（68%）和随机策略（0）；同时主传感器存活率达到 82%，平均覆盖率达到 83.5%。所提方法能够有效协调监视收益与辐射暴露风险，提高复杂对抗环境下多传感器系统的任务完成能力与生存能力。

关键词：多智能体强化学习；传感器系统；对抗环境；多传感器协同；目标探测
中图分类号：TP212 **文献标志码**：A
DOI：10.7652/xjtuxb202611007 **文章编号**：0253-987X(2026)11-0071-13

Multi-Sensor Cooperative Control Strategy Based on Multi-Agent Reinforcement Learning

ZHANG Zheyuan[1], FENG Wenhui[1], LIAN Feng[1], LIU Xia[1], Yin Zhipeng[2]
(1. Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. HeBei Suntien New Energy Technology Co., Ltd., Zhangjiakou, Hebei 075000, China)

Abstract: To address the cooperative decision-making problem of multi-sensor systems during mission execution and threat avoidance in complex adversarial environments, this paper proposes a heterogeneous sensor cooperative control strategy based on multi-agent reinforcement learning. First, a two-dimensional adversarial scenario is constructed, which includes primary sensors, secondary sensors, threat targets, and stationary surveillance targets. Models for sensor maneuvering, emission control, target detection, and threat lock-on and engagement mechanisms are established. Subsequently, the multi-sensor cooperative management problem is formulated as a multi-agent partially observable Markov decision process. Under the framework of centralized training with decentralized execution

收稿日期：2026-03-30。 作者简介：张哲源（2001—），男，硕士生；连峰（通信作者），男，教授，博士生导师。 基金项目：国家自然科学基金资助项目（62173266）。
网络出版时间：2026-08-12 网络出版地址：https://link.cnki.net/urlid/61.1069.T.20260811.1421.002

(CTDE), the multi-agent proximal policy optimization (MAPPO) algorithm is employed for cooperative decision-making. In this framework, the primary sensor accomplishes traversal and surveillance tasks by jointly optimizing its platform maneuver strategy and emission parameters (power and beam pointing), while the secondary sensors share the threat risk of the primary sensor through maneuver coordination and emission-based decoying. To cope with highly contested conditions and sparse feedback, a reward function is designed to balance mission reward and survival risk, and a three-stage curriculum learning strategy is adopted to improve training efficiency and policy convergence. Comparative evaluations were conducted over 50 test episodes in a simulation scenario containing one primary sensor, two secondary sensors, and three threat targets. Results show that, at a coverage threshold of 0.75, the proposed method achieved a mission success rate of 82%, outperforming the threat-aware heuristic policy (66%), the independent proximal policy optimization (IPPO) policy (68%), and the random policy (0). Meanwhile, the survival rate of the primary sensor reached 82%, and the average coverage rate reached 83.5%. The proposed method effectively balances surveillance utility against emission exposure risk, enhancing both the mission completion capability and the survivability of multi-sensor systems in complex adversarial environments.

Keywords: multi-agent reinforcement learning; sensor system; adversarial environment; multi-sensor cooperation; target tracking

在现代信息化对抗任务中,态势感知能力成为决定任务成败的核心要素,而传感器系统作为获取态势信息的重要手段,在态势感知、目标探测与跟踪等任务中发挥着关键作用。随着智能化无人平台的发展,场景目标呈现出高机动、强对抗、低可观测性等特征,传统的单一固定传感器已难以满足复杂对抗环境下的感知需求。多传感器协同感知通过融合多源观测信息,可有效提高空间覆盖能力和目标跟踪精度,因此成为当前的重要研究方向[1-3]。

在多传感器目标跟踪系统中,如何在有限资源条件下合理安排各传感器的观测任务,是影响系统性能的重要因素。针对这一问题,学术界提出了传感器管理的概念,其核心思想是通过对传感器的工作模式、观测方向以及资源分配策略进行优化控制,从而最大化系统的信息获取能力或目标跟踪性能[4-5]。

围绕上述问题,国内外学者提出了多种传感器管理方法。其中,一类典型方法是基于信息论指标的传感器控制方法,通过最大化信息增益、互信息或信息散度等指标来选择最优观测动作。例如,Hoang 等在随机有限集理论框架下利用多 Bernoulli 滤波器和 Rényi 散度构建传感器控制策略[6];Jiang 等提出基于 Cauchy-Schwarz 散度的信息驱动多传感器控制方法[7];相关研究还将互信息引入多目标跟踪中的传感器选择与波束管理,有效提高了目标跟踪性能[8-9]。

另一类方法将传感器管理建模为部分可观测马尔可夫决策过程(POMDP),通过动态规划或近似优化获得观测策略[10]。Krishnamurthy 等提出多变量 POMDP 阈值策略用于多功能雷达调度,实现观测误差最小化[11]。Schöpe 等针对多目标跟踪中的雷达资源管理问题,提出受限 POMDP 框架实现跟踪性能优化[12]。有些学者也将 POMDP 应用于雷达资源调度和多传感器长期调度优化等问题[13-14]。

上述研究为多传感器系统的任务分配与观测控制提供了重要理论基础。然而,传统基于信息论或优化理论的方法通常依赖于较为精确的系统模型,并且随着传感器数及系统状态维度的增加,其计算复杂度会显著提高。在复杂动态环境或对抗环境中,系统状态的不确定性和决策空间的快速扩展使得传统优化方法难以实现实时决策。因此,如何在复杂环境下实现高效、自适应的传感器控制策略,仍然是一个具有挑战性的研究问题。

近年来,深度强化学习凭借无需精确模型、能够通过环境交互学习最优策略等优势,在复杂序贯决策问题中得到广泛应用,并逐渐用于资源调度、无人系统控制及传感器管理等领域[15]。在单智能体强化学习框架下,已有研究利用强化学习实现雷达资源管理、自主传感器控制以及扩展目标跟踪等任务[16-19],验证了强化学习在动态环境中的有效性。

在多传感器协同场景中，各传感器节点之间既存在信息共享关系，又需要在任务执行过程中进行协同决策。传统单智能体强化学习方法难以有效刻画这种多主体之间的协作关系。针对这一问题，多智能体强化学习（MARL）为多主体系统的协同决策提供了一种有效框架[20]。近年来，基于集中训练-分散执行（CTDE）的 MARL 方法在协作决策任务中得到了广泛应用，包括多智能体深度确定性策略梯度算法（MADDPG）[21]、Q 值混合算法（QMIX）[22] 和多智能体近端策略优化算法（MAPPO）等[23]。

基于上述方法，研究人员开始将 MARL 方法应用于多传感器协同感知与资源调度问题。Lee 等将深度 MARL 应用于雷达-通信联合调度问题，在资源受限条件下通过多智能体协作实现了高效的频谱动态分配[24]。Li 等针对机载雷达协同导引搜索任务，提出了基于 MARL 的多无人机智能决策算法，通过协同波束调度提升了目标搜索效率[25]。Beatty 等在无人机群传感器管理问题上提出基于信息论奖励的 MARL 算法，通过多智能体协作最大化预期信息增益，在对抗性地面目标跟踪任务中取得了优于传统方法的性能[26]。Yu 等将 MARL 算法应用于传感器系统控制，通过智能调节固定传感器指向，显著提高了多目标跟踪性能[27]。Cai 等在传感器的主动防护对抗场景中，将电子干扰器建模为多个智能体，使用 QMIX 算法处理干扰器之间的协同，减少了目标被探测时间，同时优化了干扰资源消耗[28]。针对雷达对抗场景，Feng 等提出基于 MARL 的多雷达协同抗干扰优化方法，有效提升了干扰环境下系统的鲁棒性[29]。Zhang 等进一步提出基于分层 MARL 的协同干扰策略决策方法，实现了对雷达威胁的实时智能响应[30]。多智能体强化学习已成为处理多功能雷达多目标调度、非短视优化和异构组网的关键技术之一[31]。

尽管现有研究在传感器控制方面取得了大量成果，但仍然存在一些不足。首先，大多数工作仍主要关注目标跟踪精度提升、信息获取效率优化或资源调度问题，对于复杂对抗环境下传感器系统的生存性与风险规避问题考虑较少。其次，现有研究中的多智能体系统通常以同构协同为主，对于主传感器与辅助传感器功能差异显著的异构协同机制研究仍较为有限。尤其是在外部威胁持续存在的条件下，如何通过辅助传感器主动诱导威胁、分担风险并保护关键传感器节点，仍缺乏系统性的研究与建模方法。此外，传感器辐射功率不仅影响目标探测与状态估计性能，同时还会增加被威胁目标锁定的风险，如何在感知收益与生存代价之间实现动态权衡，仍然是一个具有挑战性的研究问题。

基于上述问题，本文针对复杂对抗环境下的多传感器协同管理问题，重点研究主次传感器协同条件下的任务执行与威胁规避机制，并构建基于 MAPPO 的多智能体协同控制框架，实现任务收益与系统生存能力之间的动态平衡。为解决上述问题，本文开展了如下研究工作。

（1）针对传统传感器管理方法难以描述复杂对抗环境的问题，构建包含主传感器、次传感器、威胁目标以及固定目标的多传感器协同对抗场景，并统一建模传感器观测机制、威胁目标锁定规则以及任务完成条件。

（2）针对复杂威胁环境下关键传感器生存能力不足的问题，提出主次传感器协同的威胁规避机制，通过次传感器主动诱引威胁目标，实现风险分担与主传感器保护。

（3）针对传统单智能体强化学习难以处理异构协同决策的问题，构建基于 MAPPO 的多智能体协同控制框架。将主传感器与次传感器建模为异构智能体，并结合集中训练、分散执行机制实现多传感器协同决策，从而在任务执行效率与系统生存能力之间实现动态平衡。

（4）通过仿真实验验证所提方法的有效性。实验结果表明，该方法在保证目标监视任务完成的同时能够有效降低主传感器遭受威胁的风险，从而提升系统整体任务成功率。

综上，本文重点研究复杂对抗环境下异构多传感器系统的协同感知与生存优化问题。与传统主要关注目标跟踪精度或信息获取效率的传感器管理方法不同，本文进一步考虑了威胁目标持续存在条件下关键传感器节点的生存需求，并将主传感器与辅助传感器建模为具有不同功能分工与任务目标的异构智能体。通过引入主次传感器协同机制，实现任务执行、威胁规避与资源控制之间的联合优化。本文所构建的基于 MAPPO 的多智能体协同控制框架，为复杂对抗环境下异构多传感器系统的自主协同决策提供了一种新的研究思路。

1　问题描述与系统建模

1.1　场景与任务描述

本文研究对抗环境下的多传感器协同管理问题。仿真场景为二维平面区域，以直角坐标系描述

位置，x 轴表示横向，y 轴表示穿越方向（向北为正），区域空间范围为 $x \in [0, W]$，$y \in [0, H]$，其中 W 和 H 分别表示仿真区域的宽度和高度。区域沿 y 轴方向依次划分为起始区、威胁区与目的区：$y \leqslant y_{ez}$ 时为起始区，$y_{ex} < y < y_g$ 时为威胁区，$y \geqslant y_g$ 时为目的区。其中，y_{ez} 为起始区与威胁区的分界坐标，表示传感器进入威胁区的临界位置；y_g 为威胁区与目的区的分界坐标，表示传感器抵达目的区的临界位置。

在该场景中，系统存在以下关键元素。

（1）主传感器是系统中的关键节点，全局仅有一个，用编号 0 表示，负责完成穿越威胁区并对固定目标实施监视任务。

（2）次传感器为若干可损耗的辅助节点，编号为 $1, 2, \cdots, N-1$，N 为传感器总数。主要作用包括威胁目标观测与主动诱引威胁目标，以保护主传感器。

（3）威胁目标共有 N_m 个，部署在威胁区内，对己方传感器构成威胁。威胁目标依据传感器辐射强度选择锁定目标，进入距传感器 R_c 范围内后双方同时失效。

（4）场景中存在若干固定监视目标，其数量用 N_s 表示。主传感器需要在穿越威胁区过程中利用其雷达波束对这些目标进行观测，并在满足信噪比条件时完成目标检测。

仿真场景设置与各元素分布如图 1 所示。

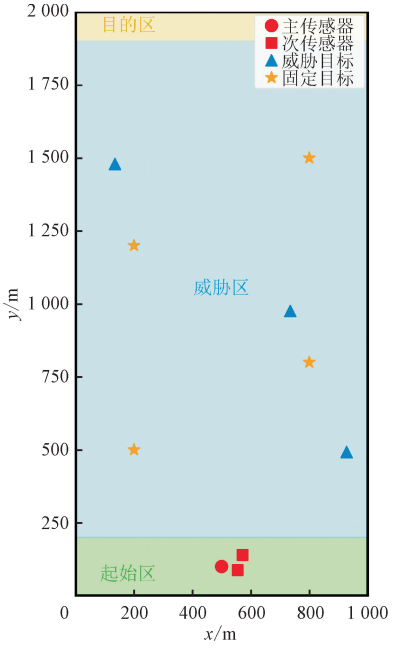


图 1 仿真场景与元素分布示意图

Fig. 1 Illustration of simulation scene and element layout

1.2 传感器运动模型

主传感器和次传感器均可在二维仿真环境中运动，其在 t 时刻的位置状态可表示为 $\boldsymbol{p}_{t,i} = [x_{t,i}, y_{t,i}]$，其中 $i = 0, 1, \cdots, N-1$ 为传感器编号，$x_{t,i}$ 和 $y_{t,i}$ 分别为传感器 i 在 t 时刻沿 x 轴（横向）和 y 轴（行进方向）的坐标分量。

主传感器可以沿 y 方向向威胁区推进，同时可以进行横向机动，因此其动作包含横移与前进两个分量，分别用横向速度 v_x 和前进速度 v_y 表示。在离散时间步长 Δt 下，主传感器的运动模型如下式

$$\boldsymbol{p}_{t+1,0} = \begin{bmatrix} x_{t+1,0} \\ y_{t+1,0} \end{bmatrix} = \begin{bmatrix} x_{t,0} + v_x \Delta t \\ y_{t,0} + v_y \Delta t \end{bmatrix} \tag{1}$$

次传感器承担辅助与诱导任务，可选择在某一方向上，以固定速度 $\boldsymbol{v}_d$ 进行机动。次传感器的运动模型同样采用离散时间步长 Δt 描述，如下式

$$\boldsymbol{p}_{t+1,i} = \boldsymbol{p}_{t,i} + \boldsymbol{v}_d \Delta t, \quad i = 1, 2, \cdots, N-1 \tag{2}$$

式中：$\boldsymbol{v}_d$ 为次传感器速度向量。

1.3 传感器观测与探测模型

在本任务场景中，主传感器雷达采用定向波束扫描，通过控制波束指向获取观测信息；次传感器采用全向辐射，不具备波束指向控制。下面从几何约束、信噪比（SNR）机制、威胁目标量测特性以及固定目标探测概率 4 个方面建立观测与探测模型。

1.3.1 主传感器定向观测的几何约束

主传感器波束在方位上被离散为 K 个扇区，在 t 时刻仅有一个扇区被观测。记 t 时刻波束指向角为 φ_t。对任意观测对象（威胁目标或固定目标），记其位置为 $\boldsymbol{p} = [p_x, p_y]$，其中 p_x 和 p_y 分别为该对象在 x 轴和 y 轴上的坐标分量。定义其相对主传感器的方位角 $\psi_t(\boldsymbol{p})$ 如下式

$$\psi_t(\boldsymbol{p}) = \arctan(p_y - y_{t,0}, p_x - x_{t,0}) \tag{3}$$

当该方位角与波束指向角之差（取 $(-\pi, \pi]$ 主值）满足下式时

$$|\psi_t(\boldsymbol{p}) - \varphi_t| \leqslant \frac{\pi}{K} \tag{4}$$

认为该对象落入当前波束覆盖范围，可能产生有效观测。其中 π/K 即单个扇区的半角宽。

1.3.2 SNR 驱动的观测质量机制

为描述辐射功率-观测质量-生存风险的耦合关系，本文引入由简化雷达方程驱动的回波信噪比模型描述威胁目标量测误差与固定目标探测概率。设传感器 i 在 t 时刻的辐射功率为 $P_{t,i}$，则对距离传感器 i 为 $\| \boldsymbol{p} - \boldsymbol{p}_{t,i} \|$ 的目标，其回波信噪比 ρ 近似为

$$\rho = \kappa \frac{P_{t,i}}{\| \boldsymbol{p} - \boldsymbol{p}_{t,i} \|^4} \tag{5}$$

式中：κ 为雷达常数，与系统参数相关。该模型表明：

辐射功率越大、距离越近，观测质量越高。需要指出的是，辐射功率 $P_{t,i}$ 同时也是威胁目标锁定模型中的关键变量，因此功率控制同时影响观测质量和生存风险。

1.3.3　威胁目标量测特性

传感器对威胁目标的量测采用极坐标形式，量测值在真实值上叠加零均值高斯噪声，噪声标准差随该目标回波信噪比增大而减小，即信噪比越高、量测越精确。为向决策提供威胁目标的位置、速度估计及其不确定度，本文采用无迹卡尔曼滤波（UKF）对威胁目标状态进行跟踪滤波，量测噪声水平随式(5)的信噪比自适应设定。

1.3.4　固定目标探测概率与覆盖更新

固定目标监视任务采用概率检测模型描述。设第 j 个固定目标位置为 $\boldsymbol{p}_j$，$j=1,2,\cdots,N_s$。当其落入主传感器当前波束扇区且辐射功率非零（非静默）时，其回波信噪比 $\rho_{t,j}$ 为

$$\rho_{t,j}=\kappa\frac{P_{t,0}}{\|\boldsymbol{p}_j-\boldsymbol{p}_{t,0}\|^4} \tag{6}$$

并将检测概率建模为信噪比的单调增函数

$$P_{\mathrm{d},j}(t)=1-\exp\left(-\frac{\rho_{t,j}}{\rho_{\mathrm{ref}}}\right) \tag{7}$$

式中：$P_{\mathrm{d},j}(t)$ 为 t 时刻对固定目标 j 的检测概率；ρ_{ref} 为参考信噪比。基于 $P_{\mathrm{d},j}(t)$ 进行伯努利试验，若检测成功则将该目标标记为已覆盖，并在后续时刻保持。

1.4　威胁目标锁定与碰撞模型

在对抗环境下，威胁目标根据被动接收到的传感器辐射强度来锁定目标。令威胁目标 m 在 t 时刻的位置与速度分别为 $\boldsymbol{p}_{t,m}$、$\boldsymbol{v}_{t,m}$，则传感器 i 对该威胁目标产生的辐射强度 $I_{t,m,i}$ 与自身功率成正比、与距离平方成反比，即

$$I_{t,m,i}\propto\frac{P_{t,i}}{\max(1,\|\boldsymbol{p}_{t,m}-\boldsymbol{p}_{t,i}\|)^2} \tag{8}$$

式中：$\max(1,\|\cdot\|)$ 用于避免极近距离下的数值奇异。由于辐射方式不同，主传感器仅在波束覆盖且非静默时产生有效辐射；次传感器采用全向辐射，辐射强度仅由功率与距离决定。

在目标选择与锁定规则方面，本文采用“阈值触发+最大强度优先”的锁定机制。令威胁目标 m 在 t 时刻的锁定目标索引为 $b_{t,m}\in\{-1,0,1,\cdots,N-1\}$，其中 $b_{t,m}=-1$ 表示未锁定任一传感器。给定辐射强度锁定门限 $I_{\min}$，若存在传感器使得辐射强度超过门限，则威胁目标选择辐射强度最大的传感器作为锁定目标，表达式如下

$$b_{t,m}=\begin{cases}\underset{i\in\{0,\cdots,N-1\}}{\operatorname{argmax}}\ I_{t,m,i}, & \underset{i\in\{0,\cdots,N-1\}}{\max}\ I_{t,m,i}>I_{\min}\\ -1, & \underset{i\in\{0,\cdots,N-1\}}{\max}\ I_{t,m,i}\leqslant I_{\min}\end{cases} \tag{9}$$

本文采用离散时间的追踪运动模型描述威胁目标运动。当威胁目标 m 已锁定传感器（$b_{t,m}\neq-1$）时，其速度方向指向被锁定传感器（$\boldsymbol{p}_{t,b_{t,m}}$ 是下标为 $b_{t,m}$ 的传感器的位置），模长为定值 v_m，即

$$\boldsymbol{v}_{t,m}=v_{\mathrm{m}}\frac{\boldsymbol{p}_{t,b_{t,m}}-\boldsymbol{p}_{t,m}}{\|\boldsymbol{p}_{t,b_{t,m}}-\boldsymbol{p}_{t,m}\|} \tag{10}$$

当威胁目标未锁定任一传感器（$b_{t,m}=-1$）时，为描述威胁的不确定性，威胁目标按随机游荡方式机动（例如以一定概率发生随机转向）。威胁目标位置按离散时间步长更新，表达式如下

$$\boldsymbol{p}_{t+1,m}=\boldsymbol{p}_{t,m}+\boldsymbol{v}_{t,m}\Delta t \tag{11}$$

为提高仿真场景真实性，若威胁目标越出威胁区域边界进入非威胁区域，则判定其失效并从后续交互中移除。

威胁目标碰撞判定采用距离阈值模型。当任一威胁目标与任一传感器之间的距离小于判定半径 R_c 时，认为发生碰撞，威胁目标与传感器同时失效。该机制使次传感器的损耗能够在策略层面转化为对主传感器的风险分担。

1.5　任务目标成功判定

在上述运动模型、观测模型与威胁模型的基础上，本节给出任务目标的判定规则及回合终止条件。本文任务需同时满足穿越约束与覆盖约束，可概括为：主传感器穿越威胁区并完成固定目标监视覆盖，同时避免与威胁目标发生碰撞。

(1)穿越约束。当主传感器到达或越过终点线，即满足 $y_{t,0}\geqslant y_{\mathrm{g}}$，且途中未与威胁目标发生碰撞，认为穿越完成。

(2)固定目标覆盖约束。固定目标一旦被成功检测，即标记为已覆盖，并在后续时刻保持。定义 t 时刻的覆盖率 η_t 为已覆盖固定目标数与固定目标总数 N_s 之比，取值范围为[0,1]。当 η_t 达到给定阈值 $\eta_{\min}\in(0,1]$ 时，认为满足覆盖约束。

综上所述，定义主传感器存活指示 $\delta_{t,0}\in\{0,1\}$（$\delta_{t,0}=1$ 表示存活），则任务成功事件定义为 $\{y_{t,0}\geqslant y_{\mathrm{g}},\ \eta_t\geqslant\eta_{\min},\ \delta_{t,0}=1\}$。

当主传感器与威胁目标发生碰撞、完成穿越或达到最大决策步数 $T_{\max}$ 时，当前回合结束。若主传感器在到达终点前与威胁目标发生碰撞，或达到最大步数仍未满足成功条件，则任务失败。

2 基于 MAPPO 的传感器协同控制方法

在系统建模的基础上，本章将多传感器协同控制问题建模为部分可观测序贯决策问题，并采用多智能体近端策略优化算法进行求解。

2.1 传感器协同管理问题的 Dec-POMDP 建模

在本文描述的对抗环境中，由于威胁目标机动行为具有随机性、目标检测结果具有概率性，且传感器观测受到波束指向和探测距离限制，多传感器系统无法获得完整的环境状态信息。同时，各传感器节点需要通过协同决策共同完成任务。因此，本文将该问题建模为多智能体部分可观测马尔可夫决策过程。在该框架下，系统可表示为如下元组

$$\mathcal{M}=\langle \mathcal{I},\ \mathcal{S},\ \mathcal{A},\ \mathcal{O},\ \mathcal{P},\ \mathcal{R},\ \gamma\rangle \tag{12}$$

式中：$\mathcal{I}$为智能体集合；$\mathcal{S}$为系统状态空间；$\mathcal{A}$为联合动作空间；$\mathcal{O}$为联合观测空间；$\mathcal{P}$为受联合动作驱动的状态转移函数；$\mathcal{R}$为奖励函数；$\gamma\in(0,1)$为折扣因子。POMDP 系统循环过程如图 2 所示。

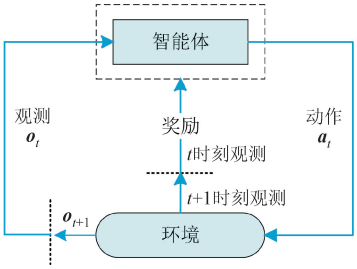


图 2 POMDP 系统循环过程

Fig. 2 POMDP system loop process

系统中每个传感器节点均作为一个智能体，各智能体在每个离散时间步根据自身观测信息选择动作，并通过联合动作共同影响环境状态的演化。

系统状态 $\boldsymbol{s}_t$ 描述 t 时刻的完整环境信息，包括传感器状态、威胁目标状态以及固定目标覆盖状态等变量。

由于传感器观测能力受到波束指向、探测距离以及辐射功率等因素限制，各智能体无法直接获取完整的系统状态，只能根据自身传感器可获取的局部环境观测信息进行决策。设智能体 i 在 t 时刻获得的观测为 $\boldsymbol{o}_{t,i}$，则系统联合观测为

$$\boldsymbol{o}_t=[\boldsymbol{o}_{t,0},\ \boldsymbol{o}_{t,1},\ \cdots,\ \boldsymbol{o}_{t,N-1}] \tag{13}$$

设智能体 i 在 t 时刻的动作记为 $\boldsymbol{a}_{t,i}$，则系统联合动作为

$$\boldsymbol{a}_t=[\boldsymbol{a}_{t,0},\ \boldsymbol{a}_{t,1},\ \cdots,\ \boldsymbol{a}_{t,N-1}] \tag{14}$$

主传感器的动作主要包括机动控制、雷达波束指向选择以及辐射功率调节等；次传感器的动作则主要包括机动方向选择及辐射功率控制。系统允许各传感器之间共享部分威胁目标跟踪信息，从而提升整体态势感知能力。

环境状态的演化由系统动力学模型决定，其中状态转移函数由传感器运动、威胁目标机动、目标检测概率以及碰撞判定共同决定。由于威胁目标锁定与目标检测均具有随机性，因此系统状态转移过程具有随机特征。

为引导智能体学习有效协同策略，本文通过奖励函数评价任务执行效果，使多传感器系统在目标监视收益与威胁规避之间实现权衡。

2.2 状态空间、动作空间与奖励函数设计

根据系统任务需求，本节分别对多智能体强化学习中的状态空间、动作空间和奖励函数进行建模。为简洁计，下文以 $\{a_i\}_{i=l}^{u}$ 表示向量 $[a_l,\ \cdots,\ a_u]$ 的有序拼接。

2.2.1 状态空间设计

本文采用集中训练、分散执行的多智能体强化学习框架。在训练阶段，集中式价值网络利用全局状态信息进行价值评估，提高策略学习的稳定性；在策略执行阶段，各智能体仅依据自身局部观测进行决策。

在训练阶段，集中式价值网络以系统全局状态为输入进行价值评估，包含传感器、威胁目标与覆盖状态等完整信息。t 时刻全局状态 $\boldsymbol{s}_t$ 为

$$\begin{aligned}\boldsymbol{s}_t=[&\varphi_t,\ \{\boldsymbol{p}_{t,i},\ P_{t,i},\ \delta_{t,i}\}_{i=0}^{N-1},\\ &\{\boldsymbol{p}_{t,m},\ \boldsymbol{v}_{t,m},\ h_{t,m},\ b_{t,m}\}_{m=1}^{N_m},\ \boldsymbol{c}_t]\end{aligned} \tag{15}$$

式中：$\boldsymbol{p}_{t,i}$、$P_{t,i}$、$\delta_{t,i}$ 分别为传感器位置、辐射功率与存活标记；$\boldsymbol{p}_{t,m}$、$\boldsymbol{v}_{t,m}$、$h_{t,m}$、$b_{t,m}$ 分别为威胁目标位置、速度、存活标记与锁定索引；$\boldsymbol{c}_t=[c_{t,1},\ \cdots,\ c_{t,N_s}]$ 为固定目标覆盖标记，$c_{t,j}=1$ 表示目标 j 已被覆盖。需要指出的是，$\boldsymbol{s}_t$ 仅在训练阶段用于价值估计，执行阶段各智能体决策不依赖该状态。

在执行阶段，每个智能体仅获得局部观测 $\boldsymbol{o}_{t,i}$。本文将局部观测统一表示为“自身状态信息”与“态势信息”的拼接。由于主/次传感器的任务角色与可获取信息不同，其观测结构有所差异。

主传感器观测包含自身位置、剩余穿越距离、自身动作记忆(功率、波束指向)、次传感器状态、威胁目标态势估计与固定目标覆盖情况，即

$$\begin{aligned}\boldsymbol{o}_{t,0}=[&\boldsymbol{p}_{t,0},\ d_t,\ P_{t,0},\ \varphi_t,\ \{\boldsymbol{p}_{t,i},\ \delta_{t,i}\}_{i=1}^{N-1},\\ &\{\hat{\boldsymbol{p}}_{t,m},\ \hat{\boldsymbol{v}}_{t,m}\}_{m=1}^{N_m},\ \boldsymbol{c}_t]\end{aligned} \tag{16}$$

式中：d_t 为 t 时刻主传感器到终点的剩余距离；$\hat{\boldsymbol{p}}_{t,m}$、$\hat{\boldsymbol{v}}_{t,m}$ 为威胁目标位置、速度的滤波估计，由 UKF 跟

踪模块给出。

第 i 个次传感器在 t 时刻观测 $\boldsymbol{o}_{t,i}$ 包含自身位置与辐射功率、主传感器位置、其他次传感器状态、威胁目标局部跟踪估计等，即

$$\boldsymbol{o}_{t,i}=\left[\boldsymbol{p}_{t,i},\ P_{t,i},\ \boldsymbol{p}_{t,0},\ \{\boldsymbol{p}_{t,k},\ \delta_{t,k}\}_{k\neq i},\ \{\hat{\boldsymbol{p}}_{t,i,m},\ \hat{\boldsymbol{v}}_{t,i,m},\ h_{t,m}\}_{m=1}^{N_{\mathrm{m}}}\right] \tag{17}$$

式中：$\hat{\boldsymbol{p}}_{t,i,m}$、$\hat{\boldsymbol{v}}_{t,i,m}$ 为第 i 个次传感器对威胁目标 m 的位置、速度滤波估计。

2.2.2　动作空间设计

考虑到实际传感器系统控制方式的离散化特征和强化学习训练稳定性，本文将各类控制变量离散化处理。主传感器动作包括横向机动、前进速度、发射功率以及波束指向；次传感器动作包括平面机动方向和发射功率。

主传感器在时刻 t 的动作向量 $\boldsymbol{a}_{t,0}$ 定义为

$$\boldsymbol{a}_{t,0}=\left[u_{t,\mathrm{lat}},\ u_{t,\mathrm{fwd}},\ u_{t,\mathrm{pow}},\ u_{t,\mathrm{beam}}\right] \tag{18}$$

式中：$u_{t,\mathrm{lat}}\in\{0,1,2,3,4\}$ 为横向机动档位，对应横向速度集合 $\{-v_{\mathrm{lat,max}},-\frac{1}{2}v_{\mathrm{lat,max}},0,\frac{1}{2}v_{\mathrm{lat,max}},v_{\mathrm{lat,max}}\}$；$u_{t,\mathrm{fwd}}\in\{0,1,2\}$ 为前进速度档位，对应前进速度集合 $\{0,v_{\mathrm{fwd,nom}},v_{\mathrm{fwd,fast}}\}$；$u_{t,\mathrm{pow}}\in\{0,1,\cdots,L_P-1\}$ 为辐射功率等级，对应 $P_{t,0}=\frac{u_{t,\mathrm{pow}}}{L_P-1}P_{\max}$；$u_{t,\mathrm{beam}}\in\{0,1,\cdots,K-1\}$ 为波束扇区索引，对应 $\varphi_t=\frac{2\pi}{K}u_{t,\mathrm{beam}}$。

第 i 个次传感器在时刻 t 的动作向量 $\boldsymbol{a}_{t,i}$ 定义为

$$\boldsymbol{a}_{t,i}=\left[u_{t,\mathrm{dir}},\ u_{t,\mathrm{pow}}\right],\ i=1,2,\cdots,N-1 \tag{19}$$

式中：$u_{t,\mathrm{dir}}\in\{0,1,\cdots,8\}$ 表示平面机动方向（含静止），对应八方向单位向量与零向量的集合；$u_{t,\mathrm{pow}}\in\{0,1,\cdots,L_P-1\}$ 为辐射功率等级，映射方式与主传感器一致。

通过上述离散化动作设计，策略网络能够在有限动作集合上实现机动控制与辐射控制的联合决策。

2.2.3　奖励函数设计

奖励函数用于衡量智能体在执行任务过程中的行为优劣，并为策略优化提供学习信号。根据任务目标，本文采用分量叠加的奖励设计，包括主传感器前进奖励、监视覆盖奖励、存活/失效惩罚、次传感器损耗与诱导奖励，以及终局奖励与超时惩罚。记第 i 个智能体在 t 时刻的单步奖励为 $r_{t,i}$，各分量定义如下，$w(\cdot)$ 为对应权重系数（取值见实验设置）。

（1）主传感器前进奖励。当主传感器本步纵向坐标较上一步增加（$y_{t,0}-y_{t-1,0}>0$）且尚未到达终点（$y_{t,0}<y_{\mathrm{g}}$）时，给予前进奖励，否则该项为 0，即

$$r_{t,\mathrm{fwd},0}=w_{\mathrm{fwd}} \tag{20}$$

（2）监视覆盖奖励。固定目标覆盖收益由全体智能体共享，按本步新增覆盖数计

$$r_{t,\mathrm{cov},i}=w_{\mathrm{cov}}n_{t,\mathrm{new}},\ i\in\{0,1,\cdots,N-1\} \tag{21}$$

式中：$n_{t,\mathrm{new}}$ 为本步新增覆盖的固定目标数。

（3）主传感器生存与失效惩罚。若主传感器发生碰撞（$\delta_{t,0}=0$），则对全体智能体施加惩罚 $r_{t,\mathrm{death},i}$；否则给予主传感器每步存活奖励 $r_{t,\mathrm{surv},0}$，如下式

$$r_{t,\mathrm{death},i}=-w_{\mathrm{death}}(1-\delta_{t,0}) \tag{22}$$

$$r_{t,\mathrm{surv},0}=w_{\mathrm{surv}}\delta_{t,0} \tag{23}$$

（4）次传感器损耗与诱导奖励。当次传感器在主传感器未发生碰撞的条件下发生碰撞时，视为有效损耗，给予该次传感器奖励 $r_{t,\mathrm{sac},i}$，并向主传感器回传部分收益 $r_{t,\mathrm{sac},0}$，即

$$r_{t,\mathrm{sac},i}=w_{\mathrm{sac}}(1-\delta_{t,i})\delta_{t,0} \tag{24}$$

$$r_{t,\mathrm{sac},0}=\frac{1}{2}w_{\mathrm{sac}}\sum_{i=1}^{N-1}(1-\delta_{t,i})\delta_{t,0} \tag{25}$$

当次传感器处于辐射状态（$P_{t,i}>0$）且有存活威胁目标在本步锁定该次传感器时，给予诱导奖励，否则该项为 0；锁定数越多，诱导奖励越大，即

$$r_{t,\mathrm{att},i}=w_{\mathrm{att}}n_{t,\mathrm{att},i} \tag{26}$$

$$r_{t,\mathrm{att},0}=\frac{1}{2}w_{\mathrm{att}}\sum_{i=1}^{N-1}n_{t,\mathrm{att},i} \tag{27}$$

式中：$n_{t,\mathrm{att},i}$ 为本步锁定次传感器 i 的存活威胁目标数（次传感器未辐射时视 $n_{t,\mathrm{att},i}$ 为 0）。

（5）终局奖励与超时惩罚。当主传感器到达终点（$y_{t,0}\geqslant y_{\mathrm{g}}$）时回合结束，并视覆盖情况给予一次性终局奖励或惩罚；主传感器未到达终点时终局奖励为 0，即

$$r_{t,\mathrm{term},i}=\begin{cases}w_{\mathrm{mission}}(0.5+0.5\eta_t), & \eta_t\geqslant\eta_{\min}\\ -w_{\mathrm{fail}}(1+2(\eta_{\min}-\eta_t)), & \eta_t<\eta_{\min}\end{cases} \tag{28}$$

式(28)中的两种取值分别对应覆盖达标与未达标两种情形。

若达到最大决策步数 $T_{\max}$ 时主传感器仍未到达终点（$y_{t,0}<y_{\mathrm{g}}$），则额外施加超时惩罚，即

$$r_{t,\mathrm{timeout},i}=-\frac{1}{2}w_{\mathrm{fail}} \tag{29}$$

综上，主、次传感器的单步奖励如下

$$r_{t,i}=\begin{cases}r_{t,\mathrm{fwd},0}+r_{t,\mathrm{cov},i}+r_{t,\mathrm{death},i}+r_{t,\mathrm{surv},0}+r_{t,\mathrm{sac},i}+r_{t,\mathrm{att},i}+r_{t,\mathrm{term},i}+r_{t,\mathrm{timeout},i}, & i=0\\ r_{t,\mathrm{cov},i}+r_{t,\mathrm{death},i}+r_{t,\mathrm{sac},i}+r_{t,\mathrm{att},i}+r_{t,\mathrm{term},i}+r_{t,\mathrm{timeout},i}, & i=1,\cdots,N-1\end{cases} \tag{30}$$

式中：前进奖励 $r_{t,\mathrm{fwd},0}$ 与存活奖励 $r_{t,\mathrm{surv},0}$ 为主传感器专属，仅计入 $i=0$ 项；$r_{t,\mathrm{sac},0}$、$r_{t,\mathrm{att},0}$ 为次传感器损耗/诱引行为向主传感器的回传收益，计入主传感器，对应的 $r_{t,\mathrm{sac},i}$、$r_{t,\mathrm{att},i}(i\geqslant 1)$ 为次传感器自身获得的损耗/诱引奖励。

2.3 基于 MAPPO 的协同策略学习算法与训练框架

考虑到本文问题具有多智能体协作、局部可观测与强耦合决策等特征，本文采用多智能体近端策略优化算法(MAPPO)进行策略学习。

在 CTDE 框架下，第 i 个智能体的策略网络(actor)以局部观测 $\boldsymbol{o}_{t,i}$ 为输入，输出动作分布 $\pi_{\theta_i}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})$；集中式价值网络(critic)以全局状态 $\boldsymbol{s}_t$ 为输入，为每个智能体 i 输出其个体状态价值 $V_{\phi_i}(\boldsymbol{s}_t)$，其中 θ_i、ϕ_i 分别为第 i 个智能体策略网络与价值网络的参数。

2.3.1 多维离散动作的因子分解策略

本文将主/次传感器的控制量离散化，并采用多维离散动作(MultiDiscrete)形式建模。设智能体 i 的动作由 D_i 个离散维度构成，写作 $\boldsymbol{a}_{t,i}=(a_{t,i,1},\cdots,a_{t,i,D_i})$，本文采用按动作维度因子分解的策略 $\pi_{\theta_i}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})$ 表示如下

$$\pi_{\theta_i}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})=\prod_{d=1}^{D_i}\pi_{\theta_i,d}(a_{t,i,d}\mid\boldsymbol{o}_{t,i}) \tag{31}$$

其中每一维动作向量分量 $\pi_{\theta_i,d}(\cdot\mid\boldsymbol{o}_{t,i})$ 均为分类分布(Categorical)。相应地，动作对数概率为各维对数概率之和，表达式如下

$$\lg\pi_{\theta_i}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})=\sum_{d=1}^{D_i}\lg\pi_{\theta_i,d}(a_{t,i,d}\mid\boldsymbol{o}_{t,i}) \tag{32}$$

该方法能够自然适配主/次传感器在动作维度与离散粒度上的差异，便于实现稳定的策略梯度更新。

2.3.2 同类次传感器的参数共享

考虑到各次传感器在动力学与控制接口上同构，角色差异主要来自所处位置与威胁态势，本文对次传感器智能体采用网络参数共享以提高样本利用效率并降低参数规模，如下式

$$\theta_i=\theta_{\mathrm{d}},\quad i\in\{1,2,\cdots,N-1\} \tag{33}$$

式中：θ_i 为第 i 个次传感器的网络参数；θ_{d} 为共享网络参数。参数共享不改变执行阶段的分布式决策形式：各次传感器仍基于自身局部观测独立决策，仅在训练中通过共享参数实现经验汇聚。

2.3.3 PPO 截断目标与 GAE 优势估计

在策略优化方面，本文采用 PPO 的截断代理目标限制策略更新幅度。对第 i 个智能体，记重要性采样比率 $\mathcal{G}_{t,i}(\theta_i)$ 为

$$\mathcal{G}_{t,i}(\theta_i)=\frac{\pi_{\theta_i}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})}{\pi_{\theta_i^{\mathrm{old}}}(\boldsymbol{a}_{t,i}\mid\boldsymbol{o}_{t,i})} \tag{34}$$

式中：θ_i^{old} 为更新前的策略参数，其策略更新目标 $\mathcal{L}_i(\theta_i)$ 写为

$$\mathcal{L}_i(\theta_i)=E[\min(\mathcal{G}_{t,i}A_{t,i},\ \mathrm{clip}(\mathcal{G}_{t,i},1-\varepsilon,1+\varepsilon)A_{t,i})] \tag{35}$$

式中：$A_{t,i}$ 为优势函数；ε 为截断系数；clip(·)为截断函数(将比值限制在$[1-\varepsilon,1+\varepsilon]$内)。

优势函数采用广义优势估计(GAE)。为使集中式价值网络稳定地刻画各智能体的长期收益，第 i 个智能体以其个体即时奖励 $r_{t,i}$ 构造价值回报信号，其时序差分误差 $e_{t,i}$ 为

$$e_{t,i}=r_{t,i}+\gamma V_{\phi_i}(\boldsymbol{s}_{t+1})-V_{\phi_i}(\boldsymbol{s}_t) \tag{36}$$

则 GAE 形式的回报估计为

$$\hat{G}_{t,i}=V_{\phi_i}(\boldsymbol{s}_t)+\sum_{n=0}^{\infty}(\gamma\lambda)^n e_{t+n,i} \tag{37}$$

式中：λ 为 GAE 系数。

个体优势函数为回报估计与状态价值之差，即

$$A_{t,i}=\hat{G}_{t,i}-V_{\phi_i}(\boldsymbol{s}_t) \tag{38}$$

集中式价值网络通过最小化各智能体价值 $V_{\phi_i}(\boldsymbol{s}_t)$ 与其回报估计 $\hat{G}_{t,i}$ 之间的均方误差之和进行更新，各策略网络(actor)基于各自优势信号 $A_{t,i}$ 进行更新。

2.3.4 训练框架与控制耦合流程

本文采用“并行采样-批量更新”的在线训练流程：训练过程中通过并行环境采样轨迹，并利用 GAE 计算优势函数，随后采用 PPO 目标分别批量更新主传感器策略、共享次传感器策略以及集中式 critic 网络。基于 MAPPO 的传感器协同管理策略学习框架如图 3 所示。

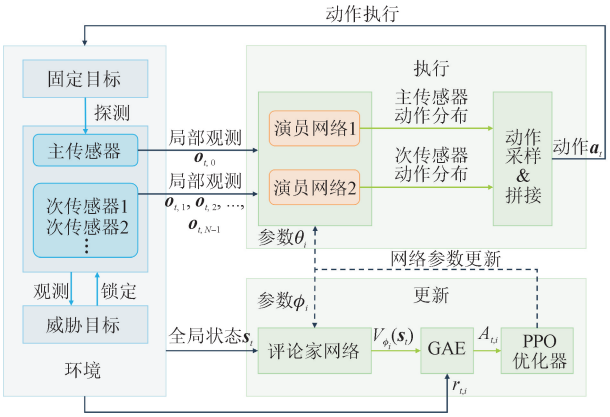


图 3 传感器协同管理策略算法框架

Fig. 3 Collaborative sensor management strategy algorithm framework

3 实验与结果分析

为验证本文传感器协同管理方法的有效性，本

节在对抗环境仿真平台上开展实验。实验围绕任务完成(穿越＋覆盖)-生存风险-协同代价 3 类目标，分别从任务成功率、主传感器存活率与监视覆盖率等维度对策略性能进行评估，并与 3 种基线策略进行对比。

3.1　实验设置

3.1.1　仿真场景与任务配置

仿真场景尺寸为 1 000 m×2 000 m，终点线为 $y = y_g = 1\,900$ m，威胁区起始边界为 $y = y_{ez} = 200$ m。系统包含 1 个主传感器与 2 个次传感器，主传感器最大横移速度 $v_{lat,max}$ 为 10 m/s，分为 5 档；最大前进速度 $v_{fwd,fast}$ 为 15 m/s，常规前进速度 $v_{fwd,nom}$ 为 10 m/s。次传感器移动速度 v_d 为 15 m/s。敌方威胁目标数 N_m 为 3，移动速度 v_m 为 25 m/s。传感器与威胁目标的速度设定保证威胁目标速度大于传感器最大移动速度，增大了传感器任务完成难度，更符合实际应用场景。固定监视目标数 N_s 为 4，其位置为(200 m，500 m)、(800 m，800 m)、(200 m，1 200 m)、(800 m，1 500 m)。每回合最大步数 $T_{max}=200$，时间步长 $\Delta t=1$ s。对抗环境关键参数设置如表 1 所示。

表 1　对抗环境关键参数设置

Table 1　Adversarial environment key parameters

参数	数值
主传感器波束扇区数 K	6
雷达最大发射功率 P_{max}/W	1 000
雷达功率档位数	5
雷达常数 κ	3×10^8
雷达信噪比参考尺度常数 ρ_{ref}	10
威胁目标锁定辐射强度门限 I_{min}/(W·m^{-2})	4×10^5
威胁目标命中半径 R_c	20 m

任务成功判定沿用之前定义，当主传感器安全到达终点且固定目标覆盖率满足 $\eta_t \geqslant \eta_{min} = 0.75$ 时，判定任务成功。

3.1.2　强化学习训练参数

本文 MAPPO 训练与评估的关键超参数见表 2。为提升在稀疏成功反馈与强威胁对抗条件下的学习稳定性，训练采用三阶段课程学习：阶段 0 降低威胁程度并放宽覆盖阈值以学习基本“前进＋覆盖”行为；阶段 1 逐步提高威胁与覆盖要求以学习初步协同；阶段 2 恢复完整难度进行最终策略收敛。课程学习仅调整威胁目标速度、锁定阈值与覆盖阈值等难度参数，不改变威胁目标数、监视目标数与智能体数，因此网络结构与观测/动作维度在整个训练过程中保持一致。奖励函数各分量权重见表 3。

表 2　MAPPO 超参数设置

Table 2　MAPPO hyperparameter settings

参数	数值
Actor/Critic 网络隐藏层维度	256
MLP 网络层数	2
演员/评论家学习率	3×10^{-4}
折扣因子 γ	0.99
采样步长	2 048
每批数据更新轮数	10
minibatch 批大小	512
训练总交互步数	5×10^6

表 3　奖励函数各分量权重设置

Table 3　Weight settings for each component of the reward function

参数	数值
前进奖励权重系数 w_{fwd}	0.5
目标覆盖奖励权重系数 w_{cov}	8
主传感器每步存活奖励权重系数 w_{surv}	0.5
主传感器被击中惩罚权重系数 w_{death}	120
任务成功奖励权重系数 $w_{mission}$	50
超时惩罚权重系数 w_{fail}	15
次传感器诱导奖励权重系数 w_{att}	1.5
次传感器防护贡献奖励权重系数 w_{sac}	8

3.2　强化学习训练结果

为验证策略学习的稳定性，训练过程中记录了团队平均回报函数曲线，如图 4 所示。由于本场景中每个训练回合的步数不固定，因此曲线横坐标用训练步数表示。从图 4 中可以看到，随着训练步数增加，团队平均回报整体呈上升趋势，并在训练约 2×10^6 步后逐步趋于稳定，表明 MAPPO 策略能够在训练过程中持续获得更优策略。

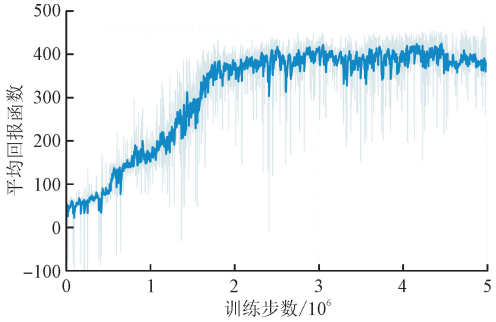
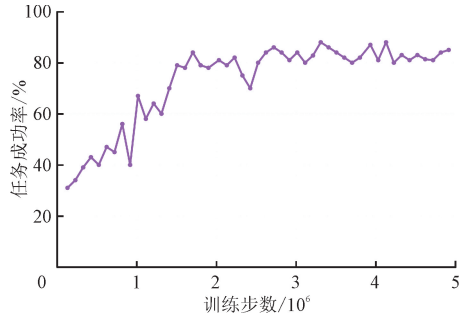


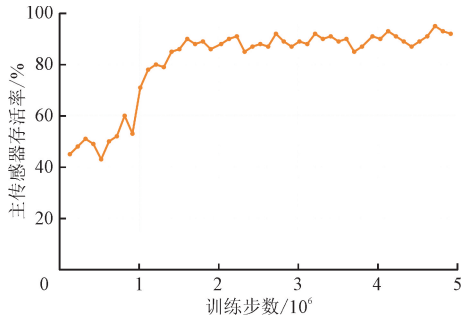
图 4　训练过程中团队平均回报函数

Fig. 4　Mean team reward curve during training

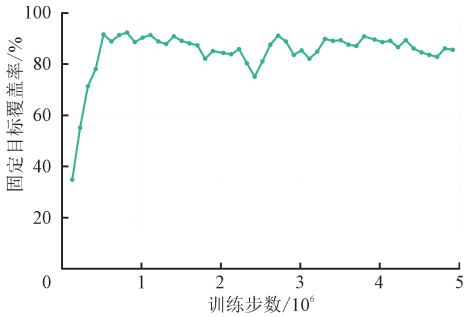
实验还记录了任务成功率、主传感器存活率与固定目标覆盖率等评估指标的学习曲线，如图 5 所示。任务成功率、主传感器存活率在训练 2×10^6 步后保持在 80%以上，固定目标覆盖率在训练 1×10^6 步后保持在 90%左右，表明策略在训练后期能够同时保持较高任务成功率、存活率和覆盖率。



(a)任务成功率



(b)主传感器存活率



(c)固定目标覆盖率

图 5 评估指标学习曲线

Fig. 5 Evaluation metric learning curves

3.3 对比实验设置与结果

为验证本文方法的有效性，设置如下对比方法。

(1)随机策略。各传感器在每个决策时刻从自身动作空间中均匀随机采样动作，不利用环境状态信息进行策略优化，用于衡量无策略指导条件下系统性能的下界。

(2)威胁感知启发式策略。该方法基于人工设计规则实现传感器协同控制。主传感器纵向保持恒定前进速度，横向根据最近威胁目标方位进行规避；波束始终指向最近的未覆盖目标；辐射功率依据最近威胁目标距离分级调节，当威胁目标距离较远时采用高功率保证目标监视，当威胁目标接近时降低功率以减少锁定风险。次传感器以距离主传感器最近的威胁目标为目标，机动至威胁目标与主传感器连线附近，并保持高功率辐射以吸引威胁目标，实现对主传感器的保护。

(3)独立近端策略优化策略(IPPO)。将各传感器节点视为相互独立的智能体，每个智能体采用独立的 PPO 算法进行策略学习，仅根据自身局部观测信息更新策略，不采用集中式价值网络进行全局信息融合，但保留与 MAPPO 相同的观测空间、动作空间以及网络结构。该方法用于分析集中训练机制以及多智能体协同建模对策略性能的影响。

各对比方法在相同环境参数下进行评估，并运行 $N_{ep}=50$ 个回合，统计任务成功率、主传感器存活率、目标覆盖率以及回合平均步数等指标。表 4 给出了 MAPPO 与 3 种基线策略在 50 回合评估下的关键指标均值对比。

表 4 不同策略性能对比($N_{ep}=50$)

Table 4 Performance comparison of different policies ($N_{ep}=50$)

策略	任务成功率/%	主传感器存活率/%	固定目标覆盖率/%	回合平均步数
MAPPO	**82**	**82**	**83.5**	**105.34**
IPPO	68	73	80.5	107.80
威胁感知启发式	66	70	79.0	110.26
随机策略	0	18	71.5	107.50

注：加粗数据为最优值。

由对比实验结果可以看出：

(1)随机策略几乎无法完成任务，任务成功率为 0，仅有 18%的主传感器能够存活至回合结束。虽然固定目标覆盖率达到 71.5%，但由于动作完全随机，主传感器无法有效规避威胁，绝大多数回合在到达终点前即失效，说明复杂对抗环境下仅依靠随机行为难以实现任务目标。

(2)威胁感知启发式策略能够根据威胁态势进行规则化决策，相比随机策略，任务成功率提高至 66%，主传感器存活率提升至 70%。这表明引入威胁规避机制能够有效降低碰撞风险。然而，该方法依赖固定规则，难以根据实时任务态势灵活调整波束控制、功率分配及主次传感器协同行为，因此覆盖率和整体任务完成能力仍然有限。

(3)IPPO 策略利用强化学习实现了各传感器的自主决策，任务成功率进一步提升至 68%，主传感器存活率达到 73%，策略固定目标覆盖率提高至 80.5%。相比启发式策略，IPPO 策略能够通过与环境交互学习更加合理的机动策略和资源控制策略。然而，由于各智能体独立学习，对每个智能体而言，训练过程中环境具有较强的非平稳性，且缺乏对其他智能体行为的有效建模，因此会导致主、次传感器之间的协同能力受到限制。

(4)相比上述方法，本文提出的 MAPPO 策略取得了最优性能。任务成功率达到 82%，较 IPPO 策略提高约 20%，较威胁感知启发式策略提高约 24%；主传感器存活率达到 82%，较 IPPO 策略提高约 12%，较威胁感知启发式策略提高约 17%，说明集中训练机制能够学习更加有效的威胁规避策略；固定目标覆盖率达到 83.5%，表明主传感器能够在保证生存性的同时完成更多监视任务。此外，MAPPO 策略的平均回合步数仅为 105.34，为所有策略中最小，说明所学习到的协同策略能够以更高的执行效率完成任务，而不是依赖长时间机动来获取覆盖收益。

需要指出的是，本文实验中任务成功率与主传感器存活率呈现相近趋势，主要原因在于当前实验场景下任务失败事件主要由主传感器与威胁目标发生碰撞引起。根据任务判定规则，任务成功需要同时满足“主传感器到达终点”和“固定目标覆盖率达到阈值”两个条件，因此任务成功事件是主传感器存活事件与任务完成事件的交集。在当前威胁配置和覆盖阈值设置下，能够规避威胁并完成穿越的策略通常也能够满足固定目标覆盖要求，因此两项指标统计结果较为接近。但任务成功率仍综合反映生存能力与任务执行能力，而覆盖率指标进一步刻画了不同策略在监视收益方面的差异。

综合以上结果可以看出，本文策略能够充分发挥主、次传感器之间的协同作用。一方面，主传感器能够根据任务态势动态调整机动、波束指向及辐射功率，在完成目标监视的同时降低自身暴露风险；另一方面，次传感器能够主动承担诱引威胁目标和分散威胁的任务。得益于 MAPPO 策略集中训练、分散执行框架对多智能体协同关系的有效建模，所提出的策略在任务完成能力、生存能力以及整体执行效率等方面均优于随机策略、启发式方法和独立强化学习策略，验证了其在复杂对抗环境下异构多传感器协同管理中的有效性。

4　结　论

本文针对对抗环境下多传感器协同管理中的任务收益与生存风险耦合问题，围绕主传感器穿越威胁区、完成固定目标监视并规避威胁的任务需求，开展了系统建模、协同控制方法设计及仿真验证研究，主要结论如下。

(1)建立了对抗环境下多传感器协同管理的统一仿真建模框架。通过引入主传感器定向波束几何约束、信噪比驱动的观测质量与检测概率模型，以及基于辐射强度的威胁目标锁定与碰撞判定机制，刻画了波束指向、功率控制、观测收益与锁定风险之间的耦合关系，为强化学习策略训练与性能评估提供了统一仿真环境基础。

(2)提出了一种基于 MAPPO 的主次传感器协同控制策略。通过将主传感器与次传感器建模为具有不同任务角色的智能体，并采用集中训练、分散执行机制，实现机动控制、辐射功率调节与波束调度的协同优化。同时，通过奖励机制设计，使策略能够在任务收益与生存风险之间进行动态权衡。

(3)仿真对比结果验证了所提策略的有效性。在与随机策略、独立 PPO 策略以及威胁感知启发式策略的对比实验中，本文 MAPPO 策略在 50 回合评估中取得了较高的任务成功率(82%)，同时获得 82%的主传感器存活率和 83.5%的平均覆盖率。实验结果表明，所提策略能够在前进效率、目标覆盖收益与辐射暴露风险之间实现较优平衡，提高满足“到达终点且覆盖达标”联合成功判据的概率。

尽管本文策略在所构建的对抗场景中取得了较好的效果，但仍存在进一步研究空间：一是引入更加复杂的对抗行为与威胁机制，如多类型威胁目标、协同追踪与欺骗干扰等，提升策略在强非平稳环境中的鲁棒性；二是考虑通信受限与信息延迟条件下的协同决策问题，分析不完美信息共享对系统性能的影响；三是扩展至更大规模异构传感器网络，并进一步考虑能量消耗等工程约束，以验证方法的可扩展性与实际应用潜力。

参考文献：

[1] Bier S G, Rothman P L, Manske R A. Intelligent sensor management for beyond visual range air-to-air combat [C]//Proceedings of the IEEE 1988 National Aerospace and Electronics Conference. Piscataway, NJ, USA: IEEE, 1988: 264-269.

[2] Ng G W, Ng K H. Sensor management-what, why and how [J]. Information Fusion, 2000, 1(2): 67-75.

[3] Bar-Shalom Y, Li Xiaorong. Multitarget-multisensor tracking principles and techniques [M]. Storrs, CT, USA: YBS Publishing, 1995.

[4] Hero A O, Castañón D A, Cochran D, et al. Foundations and applications of sensor management [M]. Boston, MA, USA: Springer US, 2008.

[5] Zuo Lei, Hu Juan, Sun Hao, et al. Resource allocation for target tracking in multiple radar architectures over lossy networks [J]. Signal Processing, 2023, 208: 108973.

[6] Hoang H G, Vo B T. Sensor management for multi-target tracking via multi-Bernoulli filtering [J]. Automatica, 2014, 50(4): 1135-1142.

[7] Jiang Meng, Yi Wei, Kong Lingjiang. Multi-sensor control for multi-target tracking using Cauchy-Schwarz divergence [C]//2016 19th International Conference on Information Fusion (FUSION). Piscataway, NJ, USA: IEEE, 2016: 2059-2066.

[8] Wang Ping, Ma Liang, Xue Kai. Multitarget tracking in sensor networks via efficient information-theoretic sensor selection [J]. International Journal of Advanced Robotic Systems, 2017, 14(5): 1729881417728466.

[9] 陈辉，韩崇昭. 机动多目标跟踪中的传感器控制策略的研究 [J]. 自动化学报，2016，42(4)：512-523.
Chen Hui, Han Chongzhao. Research on sensor control strategy for maneuvering multi-target tracking [J]. Acta Automatica Sinica, 2016, 42(5): 512-523.

[10] Castanon D A. Approximate dynamic programming for sensor management [C]//Proceedings of the 36th IEEE Conference on Decision and Control. Piscataway, NJ, USA: IEEE, 1997: 1202-1207.

[11] Krishnamurthy V, Djonin D V. Optimal threshold policies for multivariate POMDPs in radar resource management [J]. IEEE Transactions on Signal Processing, 2009, 57(10): 3954-3969.

[12] Schöpe M I, DriessenH, Yarovoy A. A constrained POMDP formulation and algorithmic solution for radar resource management in multi-target tracking [J]. Journal of Advances in Information Fusion, 2021, 16(1): 31-47.

[13] 王增福，杨广宇，金术玲. 考虑综合性能最优的非短视快速天基雷达多目标跟踪资源调度算法 [J]. 雷达学报，2024，13(1)：253-269.
Wang Zengfu, Yang Guangyu, Jin Shuling. A non-myopic and fast resource scheduling algorithm for multi-target tracking of space-based radar considering optimal integrated performance [J]. Journal of Radars, 2024, 13(1): 253-269.

[14] 徐公国，单甘霖，段修生. 采用马氏决策过程和后验克拉美罗下界的多被动式移动传感器长期调度方法 [J]. 西安交通大学学报，2019，53(6)：125-133.
Xu Gongguo, Shan Ganlin, Duan Xiusheng. Non-myopic scheduling method of multiple passive mobile sensors based on Markov decision process and posterior Cramér-Rao lower bound [J]. Journal of Xi'an Jiaotong University, 2019, 53(6): 125-133.

[15] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-533.

[16] Charlish A, Hoffmann F. Cognitive radar management [M]. Novel Radar Techniques and Applications Volume 2: Waveform Diversity and Cognitive Radar, and Target Tracking and Data Fusion. Stevenage, Hertfordshire, UK: IET, 2017: 157-193.

[17] Ravier R, Garagić D, Peskoe J, et al. Online reinforcement learning for autonomous sensor control [C]//2023 IEEE Aerospace Conference. Piscataway, NJ, USA: IEEE, 2023: 1-10.

[18] 张虹芸，陈辉，张文旭. 扩展目标跟踪中基于深度强化学习的传感器管理方法 [J]. 自动化学报，2024，50(7)：1417-1431.
Zhang Hongyun, Chen Hui, Zhang Wenxu. Sensor management method based on deep reinforcement learning in extended target tracking [J]. Acta Automatica Sinica, 2024, 50(7): 1417-1431.

[19] 闫实，贺静，王跃东，等. 基于强化学习的多机协同传感器管理 [J]. 系统工程与电子技术，2020，42(8)：1726-1733.
Yan Shi, He Jing, Wang Yuedong, et al. Multi-airborne cooperative sensor management based on reinforcement learning [J]. Systems Engineering and Electronics, 2020, 42(8): 1726-1733.

[20] Zhang Kaiqing, Yang Zhuoran, Başar T. Multi-agent reinforcement learning: a selective overview of theories and algorithms [M]. Handbook of Reinforcement Learning and Control. Cham, Switzerland: Springer International Publishing, 2021: 321-384.

[21] Lowe R, Wu Yi, 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. Red Hook, NY, USA: Curran Associates Inc., 2017: 6382-6393.

[22] Rashid T, Samvelyan M, De Witt C S, et al. Monotonic value function factorisation for deep multi-agent

reinforcement learning [J]. The Journal of Machine Learning Research, 2020, 21(1): 178.

[23] Yu Chao, Velu A, Vinitsky E, et al. The surprising effectiveness of PPO in cooperative multi-agent games [C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022: 24611-24624.

[24] Lee J, Niyato D, Guan Yongliang, et al. Learning to schedule joint radar-communication with deep multi-agent reinforcement learning [J]. IEEE Transactions on Vehicular Technology, 2022, 71(1): 406-422.

[25] Li Xiaoyang, Wang Teng, Wang Yongkun, et al. Intelligent decision-making algorithm for multi-UAV radar cooperative guided search task based on multiagent reinforcement learning [J]. IEEE Internet of Things Journal, 2025, 12(15): 31042-31063.

[26] Beatty E, Dong E. Multi-agent reinforcement learning for UAV sensor management [C]//Open Architecture/Open Business Model Net-Centric Systems and Defense Transformation 2023. Bellingham, WA, USA: SPIE, 2023: 125440H.

[27] Yu Yue, Liu Mei. 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.

[28] Cai Bingchen, Li Haoran, Zhang Naimin, et al. A cooperative jamming decision-making method based on multi-agent reinforcement learning [J]. Autonomous Intelligent Systems, 2025, 5(1): 3.

[29] Feng Cheng, Fu Xiongjun, Wang Ziyi, et al. An optimization method for collaborative radar antijamming based on multi-agent reinforcement learning [J]. Remote Sensing, 2023, 15(11): 2893.

[30] Zhang Wenxu, Zhao Tong, Zhao Zhongkai, et al. An intelligent strategy decision method for collaborative jamming based on hierarchical multi-agent reinforcement learning [J]. IEEE Transactions on Cognitive Communications and Networking, 2024, 10(4): 1467-1480.

[31] 黄洁瑜，谢军伟，杨子晴，等. 雷达资源管理技术发展研究综述 [J]. 现代雷达，2025，47(3)：1-13.
Huang Jieyu, Xie Junwei, Yang Ziqing, et al. A review of radar resource management technology development research [J]. Modern Radar, 2025, 47(3): 1-13.

（编辑　武红江）