GNSS拒止环境下面向覆盖质量的多无人机协同搜索方法

李婧[1]，于天琪[1]，羊箭锋[1]，胡剑凌[2,1]
（1. 苏州大学电子信息学院，江苏 苏州 215006；2. 无锡学院电子信息工程学院，江苏 无锡 214000）

摘　要：针对 GNSS 拒止环境下定位漂移、感知间歇失效及动态通信拓扑对多无人机协同搜索的影响，提出面向覆盖质量的多智能体近端策略优化算法 QC-MAPPO。构建包含随机感知噪声、累积漂移和间歇性失效的三维协同搜索模型，将感知可靠度融入累计质量覆盖评价与目标检测证据更新；采用通信拓扑掩码注意力 Critic 进行集中式价值评估，并结合邻机运动预测和渐进重连机制设计一步安全动作掩码。固定单次训练所得策略，在 100 组随机测试场景下进行评估。5 架无人机、20 个目标的实验结果表明，QC-MAPPO 的最终累计质量覆盖率为 91.76%±5.99%，平均检测目标数量为 19.89±0.37，累计断连步数为 10.54±12.99 步，较 MAPPO 降低 73.4%，300 步内任务成功率为 91%。消融及规模变化测试进一步表明，所提方法能够在所设基准场景下兼顾覆盖质量与通信保持，并具有一定的任务规模适应能力。

关键词：GNSS 拒止；多无人机；协同搜索；覆盖质量；多智能体强化学习；MAPPO
中图分类号：TN915.08
文献标志码：A
doi：10.11959/j.issn.2096−3750.WLW26180

Quality-aware coverage-oriented multi-UAV cooperative search in GNSS-denied environments

LI Jing[1], YU Tianqi[1], YANG Jianfeng[1], HU Jianling[2,1]
1. School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China
2. School of Electronic and Information Engineering, Wuxi University, Wuxi 214000, China

Abstract: To address the challenges of localization drift, intermittent sensing failures, and dynamic communication topology in multi-UAV cooperative search under GNSS-denied conditions, a quality-aware multi-agent proximal policy optimization algorithm, QC-MAPPO, is proposed. A three-dimensional cooperative search model incorporating random sensing noise, accumulated drift, and intermittent sensing failures is developed. Sensing reliability is integrated into cumulative quality-aware coverage evaluation and target detection evidence updates. A communication-topology-masked attention critic is employed for centralized value estimation, while a one-step safe action mask incorporating neighboring UAV motion prediction and progressive reconnection is designed for decentralized decision-making. A policy obtained from a single training run is evaluated across 100 randomly seeded test scenarios. With five UAVs and 20 targets, QC-MAPPO achieves a final cumulative quality coverage of 91.76% ± 5.99%, detects 19.89 ± 0.37 targets on average, and records 10.54 ± 12.99 cumulative disconnection steps, representing a 73.4% reduction compared with MAPPO. The task success rate within 300 steps is 91%. Ablation and task-scale variation tests further indicate that the proposed framework achieves a favorable balance between coverage quality and connectivity preservation under the considered simulation conditions, while exhibiting a certain degree of adaptability to task-scale changes.

Key words: GNSS-denied environment, multi-UAV, cooperative search, quality-aware coverage, multi-agent reinforcement learning, MAPPO

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

0　引言

随着无人机平台、机载传感器和边缘智能技术的发展，多无人机协同系统在灾害搜救、环境监测、复杂区域巡检和应急侦察等任务中展现出重要应用价值[1]。与单无人机相比，多无人机能够通过任务分工与并行搜索提高区域搜索效率和任务鲁棒性。然而，在地下空间、城市峡谷、密林遮挡及强干扰等场景中，GNSS 信号可能失效或严重退化，使依赖绝对位置信息的协同搜索与轨迹规划方法难以直接应用[2]。因此，如何在GNSS拒止条件下依靠相对感知、本地通信和局部环境信息实现多无人机自主协同搜索，是具有实际意义的研究问题。

现有多无人机搜索方法主要包括启发式搜索、集中式优化和多智能体强化学习等[3][4]。其中，基于集中式训练、分布式执行的MAPPO具有较好的训练稳定性和协同决策能力，已成为合作型多智能体任务中的重要基线方法[5]。围绕多无人机搜索中的观测质量、通信约束和安全决策，已有研究开展了相应探索：Papatheodorou 等[6]研究了协同视觉覆盖问题，说明不同观测条件下的覆盖质量存在差异；Güven 等[7]关注多无人机搜救中的连通保持问题；动作约束与掩码方法可用于减少强化学习训练中的无效动作[8]。传统覆盖路径规划研究为区域遍历提供了较系统的方法基础[9-11]，Hayat 等[12]进一步将搜索收益与通信链路质量结合进行规划；图注意力网络通过邻域掩码聚合节点信息，也为利用动态通信拓扑进行多智能体价值评估提供了参考[13]。

尽管已有研究在视觉质量覆盖、图注意力协同决策和安全动作约束等方面取得了一定进展，但GNSS拒止条件下的多无人机搜索仍面临感知不确定性与动态通信约束相互耦合的问题：首先，绝对定位缺失引起的估计漂移及间歇性感知失效，使区域被观测并不等同于获得可靠搜索信息；其次，无人机间通信关系随运动状态变化，影响局部信息交互和协同搜索决策；此外，基于不准确状态估计进行空间扩展，容易增加碰撞及通信断连风险，而已断连无人机又需要具备逐步恢复连通的能力。针对上述问题，本文以MAPPO为基础，提出面向覆盖质量的多无人机协同搜索算法QC-MAPPO，主要贡献如下。

（1）针对GNSS拒止条件下定位漂移与随机感知失效导致的有效观测质量下降问题，在已有视觉质量覆盖模型基础上，融合感知有效性、位置估计漂移和视场观测质量，结合累计检测证据表征任务进度，使搜索收益能同时反映区域覆盖范围与观测信息质量。

（2）针对无人机通信邻域动态变化及协同状态关联难以显式表达的问题，以实际通信邻接矩阵约束Critic的注意力聚合范围，融合节点状态、感知可靠度与全局任务进度，为共享Actor的协同策略更新提供与动态通信结构相关的价值评估信息。

（3）针对估计误差与多机同步运动引起的边界、碰撞及通信风险，结合估计状态、安全裕度和邻机一步运动预测筛选可行动作；对无法一步恢复通信的断连无人机，引入通信距离缺口驱动的渐进恢复规则，协调搜索扩张与通信保持。

上述设计将感知可靠度评价、通信拓扑感知价值学习与一步安全动作筛选相结合，形成面向GNSS拒止多无人机协同搜索的QC-MAPPO框架。通过基线对比、消融实验及不同任务规模测试，评价其搜索覆盖、目标检测和通信保持性能。

1　GNSS拒止下多无人机协同搜索系统模型与优化问题

1.1　多无人机协同搜索系统框架

GNSS拒止环境下的多无人机协同搜索与目标检测任务如图1所示。系统主要由无人机集群、地面基站和待搜索的地面目标组成。多架无人机在有限任务时长内对指定地面区域进行三维协同搜索，在满足飞行边界、避碰安全、通信连通和能量约束的条件下，提高区域有效覆盖并尽可能检测全部地面目标。本文主要关注搜救流程中的区域搜索与目标检测阶段，目标特征仅用于避免重复检测计数，不涉及目标身份确认、路径回溯和后续任务调度。

设无人机的集合为$\mathcal{M}=\{1,2,\cdots,M\}$，其中$M$为

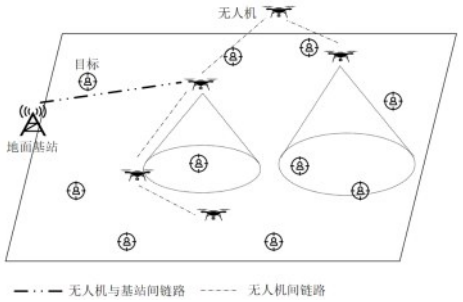


图1　多无人机协同搜索系统示意图

参与协同搜索的无人机总数。地面目标集合为$\mathcal{N}=\{1,2,\cdots,N\}$，其中$N$为任务区域内待检测的目标总数。设地面任务区域为二维有界区域$\mathcal{A}_g=[x_{\min},x_{\max}]\times[y_{\min},y_{\max}]$，无人机可用的三维飞行空间为$\mathcal{A}=\mathcal{A}_g\times[h_{\min},h_{\max}]$，其中$x_{\min}$、$x_{\max}$、$y_{\min}$和$y_{\max}$表示地面任务区域在两个水平方向的边界；$h_{\min}$和$h_{\max}$分别表示无人机最低和最高飞行高度。

将单次搜索任务划分为若干等长决策时隙，时隙集合为$\mathcal{T}=\{1,2,\cdots,T_{\max}\}$，其中$T_{\max}$为单次任务允许的最大决策步数，单个决策时隙长度为$\Delta t$。在时隙$t$，无人机$m$的真实位置、真实航向角和剩余能量分别表示为$\boldsymbol{p}_m(t)=[x_m(t),y_m(t),h_m(t)]^{\mathrm{T}}$、$\psi_m(t)$和$E_m(t)$。地面目标$n$的真实位置为$\mathbf{u}_n=[x_n^{\mathrm{tar}},y_n^{\mathrm{tar}},0]^{\mathrm{T}}$，地面基站位置为$\boldsymbol{p}_b=[x_b,y_b,h_b]^{\mathrm{T}}$。其中，$\boldsymbol{p}_m(t)$、$\psi_m(t)$和$\mathbf{u}_n$为仿真环境中的物理真实状态，不直接提供给无人机策略网络。GNSS拒止环境下，无人机主要依据机载传感器获得的自身状态、基站相对信息、邻机相对信息和局部目标观测信息来决策。

为突出GNSS绝对定位缺失、感知不确定性及动态通信拓扑对协同搜索决策的影响，考虑任务区域边界已知、待搜索目标数量由任务先验给定但其空间位置未知的有限目标搜索场景。目标在任务过程中保持静止，无人机在无遮挡、无外部风扰的三维空域内运动，无法直接获取GNSS绝对定位信息，主要依靠存在噪声、累积漂移及随机失效的局部感知进行决策。上述假设用于控制环境复杂因素，重点分析覆盖质量、通信连通与动作安全之间的耦合关系。

1.2　GNSS拒止环境下相对感知模型

GNSS拒止环境是指无人机无法稳定获得全球绝对定位信息，或GNSS定位误差显著增大的任务环境。本文不针对具体硬件底层驱动和传感器融合算法展开研究，而是从协同搜索决策建模角度对感知能力进行抽象。首先，在队友相对感知方面，假设无人机搭载UWB模块，可获得邻近无人机之间的相对距离信息[14][15]，为通信连通保持和邻机避碰提供支撑。其次，在地面目标感知方面，假设无人机搭载单目视觉传感器，并结合几何反投影方法获得目标相对位置信息。最后，在高度测量方面，假设无人机配备气压计和下视激光测距仪，获得垂向状态信息。上述能力在本文中统一抽象为局部观测输入和高度状态输入[16]。

定义无人机m在时隙t的理想局部感知信息为

$$\boldsymbol{\vartheta}_m(t)=\left\{\boldsymbol{\vartheta}_m^{\mathrm{bs}}(t),\boldsymbol{\vartheta}_m^{\mathrm{nei}}(t),\boldsymbol{\vartheta}_m^{\mathrm{tar}}(t),\boldsymbol{\vartheta}_m^{\mathrm{self}}(t)\right\} \tag{1}$$

其中，$\boldsymbol{\vartheta}_m^{\mathrm{bs}}(t)$表示基站相对方向和相对距离等信息；$\boldsymbol{\vartheta}_m^{\mathrm{nei}}(t)$表示邻近无人机的相对位置、相对距离和剩余能量等信息；$\boldsymbol{\vartheta}_m^{\mathrm{tar}}(t)$表示局部目标相对观测和目标记忆信息；$\boldsymbol{\vartheta}_m^{\mathrm{self}}(t)$表示自身高度、剩余能量、任务进度及通信状态等信息。

考虑传感器测量噪声和状态估计漂移，无人机实际获得的局部感知信息统一表示为

$$\tilde{\boldsymbol{\vartheta}}_m(t)=\boldsymbol{\Gamma}_m(t)\left[\boldsymbol{\vartheta}_m(t)+\boldsymbol{v}_m(t)+\boldsymbol{\delta}_m(t)\right] \tag{2}$$

其中，$\boldsymbol{v}_m(t)$为随机感知噪声，用于抽象描述测距、视觉观测及高度测量等短时随机误差；仿真中采用零均值高斯噪声模型，不同观测分量根据其误差特征设置相应标准差，具体参数见表2。$\boldsymbol{\delta}_m(t)$为累积漂移项，用于统一描述GNSS定位信息缺失后，由惯性递推误差、视觉里程计漂移、传感器慢变偏差及多次状态更新所引起的累积误差。$\boldsymbol{\Gamma}_m(t)$为感知有效性对角矩阵，其对角元素取0或1。当某类感知信息在当前时隙有效时，相应元素取1；当发生随机感知失效或信息暂时无法更新时，取0。本文通过感知失效概率与恢复概率描述感知有效性的随机变化，尚未根据障碍物的空间几何关系显式建模视线遮挡。为避免数值0与真实测量值混淆，$\Gamma_m(t)$同时作为局部观测的一部分输入策略网络。

式（2）中的随机噪声、累积漂移和有效性矩阵主要作用于基站相对信息、邻机相对信息、目标相对观测以及高度和航向等传感器相关分量。对于剩余能量、时间进度和累计检测进度等由系统内部直接维护的状态，其对应噪声和漂移分量置为0，有效性标志恒取1。

需要说明的是，本文的感知失效主要用于描述基站相对观测、邻机相对观测、视觉观测及位姿相关感知信息的暂时不可用，其状态变化通过失效概率和恢复概率刻画。该机制不等同于无线通信链路中的随机丢包，通信链路可用性仍由第1.4节的距离门限模型决定。

累积漂移采用一阶递推模型表示为

$$\boldsymbol{\delta}_m(t+1)=\boldsymbol{\kappa}_m^{\mathrm{dr}}(t)\boldsymbol{\delta}_m(t)+\boldsymbol{\xi}_m^{\mathrm{dr}}(t) \tag{3}$$

其中，$\boldsymbol{\xi}_m^{\mathrm{dr}}(t)$为每个时隙新增的微小漂移扰动；

$\kappa_m^{\mathrm{dr}}(t)$ 为误差保持系数。当缺少有效外部观测时，取 $\kappa_m^{\mathrm{dr}}(t)=1$，已有误差不能得到修正并继续累积；当有效观测能够对状态估计进行修正时，令 $0 \leqslant \kappa_m^{\mathrm{dr}}(t)<1$，表示累积漂移得到部分抑制。

由此，无人机 m 的位置和航向估计分别表示为

$$\hat{\mathbf{p}}_m(t)=\mathbf{p}_m(t)+\grave{\mathbf{o}}_m^{p}(t) \tag{4}$$

$$\hat{\psi}_m(t)=\psi_m(t)+\epsilon_m^{\psi}(t) \tag{5}$$

其中，$\boldsymbol{\epsilon}_m^{p}(t)$ 和 $\epsilon_m^{\psi}(t)$ 分别为综合感知噪声与累积漂移作用下的位置和航向估计误差，它们可视为式（2）中相应分量的状态估计结果。

为表征当前感知结果的总体可信程度，引入综合感知可靠度：

$$\beta_m^{\mathrm{sen}}(t)=\mathrm{clip}\left[\bar{\gamma}_m(t)\exp\left(-\kappa_\beta\left\|\boldsymbol{\delta}_m^{p}(t)\right\|_2\right),0,1\right] \tag{6}$$

其中，$\bar{\gamma}_m(t)=\dfrac{1}{n_\Gamma}\mathrm{tr}\left[\boldsymbol{\Gamma}_m(t)\right]$ 为感知有效比例，其中 n_Γ 为有效性矩阵的维数。$\boldsymbol{\delta}_m^{p}(t)$ 为位置相关的漂移分量，$\kappa_\beta>0$ 为可靠度衰减系数。综合感知可靠度随感知有效比例降低和累积漂移增大而下降，并在有效观测恢复、状态估计误差得到修正后提高。

则无人机 m 用于分布式决策的感知输入表示为

$$\boldsymbol{o}_m^{\mathrm{sen}}(t)=\left\{\tilde{\boldsymbol{\vartheta}}_m(t),\boldsymbol{\Gamma}_m(t),\beta_m^{\mathrm{sen}}(t)\right\} \tag{7}$$

1.3　覆盖质量模型

设地面区域被划分为网格集合：$G=\{g_1,g_2,\cdots,g_L\}$，其中第 l 个网格 g_l 中心坐标为 $\boldsymbol{c}_l=\left[x_l,y_l,0\right]^{\mathrm{T}}$。传统覆盖方法仅根据网格是否进入无人机视场进行二值判定，难以区分不同观测条件下的有效信息质量。为描述“看得见”与“看得清”间的区别，本文引入质量覆盖思想[6]。设无人机 m 的下视单目视觉传感器视场半角为 α，无人机根据估计高度计算其地面投影视场半径：

$$\hat{r}_m(t)=\hat{h}_m(t)\tan\alpha \tag{8}$$

无人机估计位置到网格 g_l 的水平距离为

$$\hat{d}_{m,l}(t)=\sqrt{\left(\hat{x}_m(t)-x_l\right)^2+\left(\hat{y}_m(t)-y_l\right)^2} \tag{9}$$

定义无人机 m 对网格 g_l 的估计观测质量为

$$\begin{aligned}\hat{q}_{m,l}(t)=\gamma_m^{vis}(t)\mathbb{I}\left\{\hat{d}_{m,l}(t)\leqslant\hat{r}_m(t)\right\}\cdot q_h\left(\hat{h}_m(t)\right)\\ \cdot q_e\left(\hat{d}_{m,l}(t),\hat{r}_m(t)\right)\cdot\beta_m^{sen}(t)\end{aligned} \tag{10}$$

其中，$\gamma_m^{\mathrm{vis}}(t)\in\{0,1\}$ 为视觉观测有效性变量，$\mathbb{I}\{\cdot\}$ 为指示函数；$q_h(\cdot)$ 为高度分辨率因子；$q_e(\cdot)$ 为视场边缘衰减因子。高度分辨率因子定义为

$$q_h(h_m(t))=\mathrm{clip}\left[\left(\frac{h_{\mathrm{ref}}}{h_m(t)}\right)^{\alpha_h},q_{\min},q_{\max}\right] \tag{11}$$

其中，h_{ref} 是参考飞行高度，α_h 是高度分辨率指数，$q_{\min}$ 和 $q_{\max}$ 分别为质量截断下限和上限。边缘衰减因子定义为

$$q_e(d_{m,l}(t),r_m(t))=1-\eta_e\left(\frac{d_{m,l}(t)}{r_m(t)}\right)^{\alpha_e} \tag{12}$$

其中，η_e 为边缘衰减系数，α_e 为边缘衰减指数。

多无人机联合搜索质量取所有无人机中对该网格观测质量的最大值：

$$\hat{q}_l(t)=\max_{m\in\mathcal{M}}\hat{q}_{m,l}(t) \tag{13}$$

若 $q_l(t)\geqslant q_{\mathrm{thr}}$，则认为网格 g_l 在时隙 t 达到搜索级有效覆盖，其中 q_{thr} 为观测质量阈值。

无人机系统据此更新估计累计覆盖质量：

$$\hat{Q}_l(t)=\max\left\{\hat{Q}_l(t-1),\hat{q}_l(t)\right\},\quad \hat{Q}_l(0)=0 \tag{14}$$

根据无人机真实位置 $\boldsymbol{p}_m(t)$ 计算真实观测质量 $q_{m,l}(t)$，并据此得到真实累计质量覆盖率：

$$C_{\mathrm{cum}}(t)=\frac{1}{L}\sum_{l=1}^{L}\mathbb{I}\left\{Q_l(t)\geqslant q_{\mathrm{thr}}\right\} \tag{15}$$

真实累计质量覆盖率 $Q_l(t)$ 用于奖励计算与性能评价；各无人机依据自身估计观测维护局部累计质量记忆，并在通信连通分量内进行信息共享，供分布式搜索决策使用。仿真中另行维护全局估计累计质量覆盖率，用于评价估计地图与真实覆盖情况之间的差异。

1.4　通信模型与安全约束模型

地面基站与全部无人机共同构成时变通信图：$\mathcal{G}_c(t)=(\boldsymbol{v},\varepsilon(t))$，其中 $\boldsymbol{v}=\{b\}\cup\mathcal{M}$ 为通信节点集合，$\varepsilon(t)$ 表示时隙 t 内有效通信链路的集合。无人机 i 和无人机 j 之间的真实距离为 $d_{ij}(t)$，无人机 m 与地面基站之间的真实距离为 $d_{mb}(t)$；设无人机之间的最大通信距离为 R_u，地面基站与无人机之间的最大通信距离为 R_b。则无人机 i 与无人机 j 之间的通信链路状态定义为

$$a_{ij}(t)=\begin{cases}1, & \mathrm{d_{ij}(t)}\leqslant \mathrm{R_u}\\ 0, & \mathrm{d_{ij}(t)}>\mathrm{R_u}\end{cases} \tag{16}$$

地面基站与无人机 m 间的通信链路状态定义为

$$a_{mb}(t)=\begin{cases}1, & \mathrm{d_{mb}(t)}\leqslant \mathrm{R_b}\\ 0, & \mathrm{d_{mb}(t)}>\mathrm{R_b}\end{cases} \tag{17}$$

由此可构造时隙 t 的通信邻接矩阵

$$\mathbf{A}_c(t)=\left[a_{ij}(t)\right] \tag{18}$$

规定$a_{ii}(t)=1$，以保留无人机自身状态信息。

当无人机不能直接与地面基站通信时，可通过其他无人机形成多跳中继链路。定义无人机m的基站根连通状态为

$$\chi_m(t)=\begin{cases}1,\text{通信图中存在由}\\ \quad\text{无人机m到基站b的可达路径}\\ 0,\textit{通信图中不存在由}\\ \quad\textit{无人机m到基站b的可达路径}\end{cases}\tag{19}$$

全体无人机的通信状态表示为

$$\chi(t)=\prod_{m=1}^{M}\chi_m(t)\tag{20}$$

当$\chi(t)=1$时，所有无人机均能够通过直接链路或多跳链路与地面基站保持通信；否则$\chi(t)=0$。该状态以抽象通信反馈标志的形式用于局部观测与动作安全判断，无人机策略不直接读取其他节点的真实位置。暂不考虑通信状态反馈的时延和错误。

采用基于节点间距离门限的二值通信模型，以描述无人机运动过程中通信邻接关系的动态变化。该模型假设通信链路的建立与否主要由节点间距离及预设通信半径决定，在距离门限内视为链路可用，超过门限则视为不可用。该抽象适用于链路状态能够及时、可靠获得的理想化通信条件，主要用于分析多跳连通关系与协同搜索行为之间的耦合，不涉及无线物理层信号传播及数据传输过程的详细建模。

为统计任务执行过程中的通信连通性，定义时隙t的断连指示量为

$$v_{dis}(t)=1-\chi(t)\tag{21}$$

一个搜索回合内的累计断连步数为

$$N_{dis}=\sum_{t=0}^{T-1}v_{dis}(t)\tag{22}$$

通信邻接矩阵$\mathrm{A}_c(\mathrm{t})$同时作为后续通信拓扑掩码注意力机制的拓扑依据。

设无人机m在时隙t选择候选动作a执行后得到的下一时隙候选位置为$\hat{\boldsymbol{p}}_m(t+1|a)$。为减小定位误差造成的越界、碰撞和通信临界失效风险，引入统一感知安全裕度$\delta_s>0$。

候选位置首先应满足任务空域边界约束，即

$$\begin{aligned}&x_{\min}+\delta_s\leqslant\hat{x}_m(t+1|a)\leqslant x_{\max}-\delta_s\\&y_{\min}+\delta_s\leqslant\hat{y}_m(t+1|a)\leqslant y_{\max}-\delta_s\\&h_{\min}+\delta_s\leqslant\hat{h}_m(t+1|a)\leqslant h_{\max}-\delta_s\end{aligned}\tag{23}$$

定义候选动作的边界可行性指示量为$\hat{\mathbb{I}}_m^{bou}(a,t)$。

设无人机m据最近一次有效邻机信息得到的无人机j的位置估计为$\hat{\boldsymbol{p}}_{j|m}(t)$，候选动作的避碰条件为

$$\|\hat{\boldsymbol{p}}_m(t+1|a)-\hat{\boldsymbol{p}}_{j|m}(t)\|_2\geqslant d_{\min}+\delta_s,\forall j\neq m\tag{24}$$

相应的避碰可行性指示量记为$\hat{\mathbb{I}}_m^{col}(a,t)$。

设无人机m执行动作a时的水平速度和垂直速度分别为$v_{m,xy}(a)$和$v_{m,h}(t)$，对于非悬停动作，其单步能量消耗[17]表示为

$$\begin{aligned}e_{\mathrm{m}}(a,t)=&[P_0+k_v\cdot v_{m,xy}^2(a)+kh\cdot|v_{m,h}(a)|+\\&P_{\mathrm{com}}]\Delta t\end{aligned}\tag{25}$$

其中，P_0为基础推进功率，k_v为水平运动功率系数，k_h为垂直运动功率系数，P_{com}为通信功率。

对于悬停动作，其单步能量消耗表示为

$$e_m^{\mathrm{hov}}(t)=(P_{\mathrm{hov}}+P_{\mathrm{com}})\Delta t\tag{26}$$

其中，P_{hov}为无人机悬停功率。

设无人机m在时隙t的剩余能量为$E_m(t)$，则候选动作应满足

$$E_m(t)\geqslant e_m(a,t)\tag{27}$$

候选动作的能量可行性指示量定义为$\mathbb{I}_m^{\mathrm{ene}}(a,t)$。

无人机执行动作后的剩余能量更新为

$$E_m(t+1)=E_m(t)-e_m(a,t)\tag{28}$$

仅依据当前时隙的通信状态无法避免无人机执行下一步动作后出现链路中断。因此，本文根据候选位置对动作的通信安全性进行一步预测。

设当前能够通过直接链路或多跳链路连接至地面基站的无人机集合为

$$\hat{\mathcal{C}}_{b,m}(t)=\{\,j\in\mathcal{M}\,\big|\,\hat{\chi}_{j|m}(t)=1\,\}\tag{29}$$

其中，$\hat{\chi}_{j|m}(t)$表示无人机m根据当前成功接收的邻机状态或路由联通标志得到的节点j根连通状态估计，$j\in\hat{\mathcal{C}}_{b,m}(t)$。

为降低运动误差和多机同步运动引起的临界断连风险，引入通信安全裕度δ_c，并定义无人机间和基站-无人机间的有效安全通信距离分别为

$$\hat{R}_u^s=R_u-\delta_{\mathrm{c}}-\delta_s,\hat{R}_b^s=R_b-\delta_{\mathrm{c}}-\delta_s\tag{30}$$

若候选位置能够直接接入地面基站，即满足

$$\|\hat{\boldsymbol{p}}_m(t+1|a)-\boldsymbol{p}_b\|_2\leqslant\hat{R}_b^s\tag{31}$$

或者能够与当前基站根连通集合中的至少一架无人机建立安全链路，即存在$j\in\mathcal{C}_b(t)$且$j\neq m$，使得

$$\|\hat{\boldsymbol{p}}_m(t+1|a)-\hat{\boldsymbol{p}}_{j|m}(t)\|_2\leqslant\hat{R}_u^s\tag{32}$$

则认为候选动作满足考虑感知误差的一步通信

安全条件。相应的通信可行性指示量记为$\hat{\mathbb{I}}_m^{\mathrm{com}}(a,t)$。

为减小多机同步运动造成的链路预测偏差，在判断无人机候选动作时，根据邻机当前运动状态预测其下一时隙位置，并据此评估候选联合位置下的通信连通性。若无人机当前处于断连状态且不存在能够一步恢复根连通的候选动作，则进入恢复模式，不强制要求候选动作立即满足根连通条件，而优先保留能够缩小其与当前根连通集合之间通信距离缺口的动作，使无人机逐步向可重连区域移动。

综合上述约束，候选动作的安全可行性表示为

$$\hat{\mathbb{I}}_m(a,t)=\hat{\mathbb{I}}_m^{\mathrm{bou}}(a,t)\cdot\hat{\mathbb{I}}_m^{\mathrm{col}}(a,t)\cdot\mathbb{I}_m^{\mathrm{ene}}(a,t)\cdot\hat{\mathbb{I}}_m^{\mathrm{com}}(a,t) \tag{33}$$

当$\mathbb{I}_m(a,t)=1$时，候选动作a可作为当前状态下的可执行动作，否则应在策略采样前予以屏蔽。

无人机m在时隙t的安全动作集合定义为

$$\hat{A}_m^{\mathrm{safe}}(t)=\{\,a\in\mathcal{A}_m \mid \hat{\mathbb{I}}_m(a,t)=1\,\} \tag{34}$$

该安全集合将在策略采样前用于一步安全动作掩码，形成“断连预防—渐进恢复”的闭环约束机制。

为避免极端状态下全部候选动作均被屏蔽，设置动作掩码的退化处理机制。当不存在同时满足全部安全约束的候选动作时，优先从满足边界、避碰及能量约束的动作中，保留通信距离缺口最小的至多10个候选动作；若该集合仍为空，则回退至悬停动作，以保证策略动作分布非空。该退化处理不保证所有情况下的通信连通性，主要用于避免无可执行动作导致的策略决策失效。

1.5　多约束协同搜索优化问题

基于上述感知、覆盖质量、通信和安全模型，本文将GNSS拒止环境下的多无人机协同搜索任务描述为一个有限时域多约束优化问题，优化变量为多无人机联合决策策略π。

区域累计质量覆盖率$C_{\mathrm{cum}}(t)$沿用第1.3节定义。

为避免单次低质量观测直接触发目标确认，采用累计检测证据机制。定义$\varepsilon_n(t)$为截至时隙t时目标n的累计检测证据，初始值为0。

设无人机m与目标n间的真实距离为$D_{m,n}(t)$。为区分目标的基础检测概率和实际用于证据累加的贡献量，定义无人机m对目标n的检测概率为：

$$p_{m,n}(t)=\mathrm{clip}\left(p_0 e^{-\lambda_d D_{m,n}(t)} q_{m,n}(t),0,0.95\right) \tag{35}$$

其中，p_0为基础检测概率系数，λ_d为距离衰减系数，$q_{m,n}(t)$为考虑感知有效性和可靠度的目标观测质量。$\mathrm{clip}(x,a,b)$表示将x截断至区间$[a,b]$。本文采用多时隙概率证据累加方式更新检测状态，不对各时隙检测事件额外进行伯努利随机采样

对于已经积累部分检测证据但尚未确认的目标，在合适高度与视场中心区域重访时，适当增大其证据贡献，将有效检测贡献表示为：

$$\tilde{p}_{m,n}(t)=\mathrm{clip}\left(p_{m,n}(t)\left[1+0.35I_{m,n}^{\mathrm{rev}}(t)\right],0,0.98\right) \tag{36}$$

其中，$I_{m,n}^{\mathrm{rev}}(t)$为重访增强指示量。当目标已存在未确认的检测证据、无人机高度处于60～82 m，且目标位于归一化视场中心距离不大于0.35的区域时取1，否则取0。多机联合检测概率为

$$\tilde{P}_n(t)=1-\prod_{m=1}^{M}\left[1-\tilde{p}_{m,n}(t)\right] \tag{37}$$

对尚未确认的目标，其候选累计检测证据为

$$\mathcal{E}_n^{+}(t+1)=\min\left\{1.5\theta_{\mathrm{det}},\mathcal{E}_n(t)+\tilde{P}_n(t)\right\} \tag{38}$$

其中，$\mathcal{E}_n^{+}(t+1)$为本时隙更新后的候选累计检测证据，$\tilde{P}_n(t)$为多机联合检测证据贡献，$\theta_{\mathrm{det}}=1.0$为检测确认证据阈值，上限$1.5\theta_{\mathrm{det}}$用于限制累计证据的数值范围。当$\mathcal{E}_n^{+}(t+1)\geqslant\theta_{\mathrm{det}}$时，目标达到候选确认条件，但还需要进行重复检测判别。

定义当前候选目标的归一化观测特征为$\tilde{\mathbf{f}}_n(t)$，已确认目标的特征集合为$\mathcal{F}(t)$，则最大相似度为

$$s_n(t)=\max_{\mathbf{f}\in\mathcal{F}(t)}\tilde{\mathbf{f}}_n(t)^{\top}\mathbf{f} \tag{39}$$

当集合为空时，不执行重复匹配检查，直接按新目标处理。设置阈值$\theta_{\mathrm{id}}=0.92$，当$s_n(t)\geqslant\theta_{\mathrm{id}}$时，将候选观测视为与历史已确认目标重复，不新增检测目标计数，并将该候选目标的检测证据调整为：$\mathcal{E}_n(t+1)=0.85\theta_{\mathrm{det}}$以保留部分检测证据供后续补检。当证据达到确认阈值且未匹配到已有目标时，才新增一次确认检测。确认后的目标不再重复参与检测证据更新。

当目标检测证据达到阈值且通过重复匹配判别时，目标n被判定为已检测，其累计检测状态为$z_n(t)=1$；否则$z_n(t)=0$。截至时隙t的累计检测目标数量和累计检测率分别为

$$D(t)=\sum_{n=1}^{N}z_n(t) \tag{40}$$

$$\rho_D(t)=\frac{D(t)}{N} \tag{41}$$

采用目标特征进行重复检测匹配，通过累计检

测证据与特征相似度判别减少低质量观测直接确认及目标重复计数。由于未显式模拟背景杂波或非目标物体产生的虚警事件，本文的检测指标主要评价所设仿真条件下的目标确认能力，不涉及复杂视觉背景中的虚警抑制性能。

任务结束时的单机平均剩余能量比例为

$$E_{\mathrm{rem}}(T_e)=\frac{1}{M}\sum_{m=1}^{M}\frac{E_m(T_e)}{E_{\max}} \tag{42}$$

其中，$E_{\max}$为单架无人机最大可用能量。定义平均重叠覆盖率为

$$O_{\mathrm{avg}}=\frac{1}{T_e}\sum_{t=0}^{T_e-1}O(t) \tag{43}$$

其中，$O(t)$表示时隙t内由两架及以上无人机重叠覆盖的区域比例。根据第1.4节定义的断连指示量$v_{dis}(t)$，归一化累计断连步数表示为

$$V_{\mathrm{dis}}=\frac{1}{T_e}\sum_{t=0}^{T_e-1}v_{\mathrm{dis}}(t)=\frac{N_{\mathrm{dis}}}{T_e} \tag{44}$$

为评价GNSS拒止条件下算法对定位漂移的适应能力，定义平均位置估计误差指标：

$$E_{\mathrm{loc}}=\frac{1}{MT_e}\sum_{t=0}^{T_e-1}\sum_{m=1}^{M}\left\|\hat{\mathbf{p}}_m(t)-\mathbf{p}_m(t)\right\|_2 \tag{45}$$

本文的多无人机协同搜索优化问题可表示为

$$\max_{\pi}J(\pi)=\mathbb{E}_{\pi}\left[\eta_c C_{\mathrm{cum}}(T_e)+\eta_d\rho_D(T_e)+\eta_e E_{\mathrm{rem}}(T_e)-\eta_t\frac{T_e}{T_{\max}}-\eta_o O_{\mathrm{avg}}-\eta_r V_{\mathrm{dis}}\right] \tag{46}$$

约束条件为：

$$\mathbf{p}_m(t)\in\mathcal{A},\quad \forall m\in\mathcal{M},0\leqslant t\leqslant T_e \tag{46a}$$

$$\left\|\mathbf{p}_m(t)-\mathbf{p}_j(t)\right\|_2\geqslant d_{\min},\quad \forall m\neq j,0\leqslant t\leqslant T_e \tag{46b}$$

$$0\leqslant E_m(t)\leqslant E_{\max},\quad \forall m\in\mathcal{M},0\leqslant t\leqslant T_e \tag{46c}$$

$$a_m(t)\in\hat{\mathcal{A}}_m^{\mathrm{safe}}(t),\quad \forall m\in\mathcal{M},0\leqslant t<T_e \tag{46d}$$

$$s(t+1)\sim P\left[s(t+1)|s(t),\boldsymbol{a}(t),\boldsymbol{v}(t),\boldsymbol{\delta}(t),\boldsymbol{\Gamma}(t)\right] \tag{46e}$$

其中，$J(\pi)$为联合策略π对应的期望综合任务收益；η_c、η_d、η_e、η_t、η_o和η_r为非负权重系数，分别用于调节累计覆盖质量、累计检测率、剩余能量、任务完成时间、重叠覆盖和通信断连代价在优化目标中的相对重要性。式（46a）—式（46e）分别表示飞行空域边界约束、最小安全距离约束、剩余能量约束、安全动作集合约束和状态转移约束。上述优化目标表明，本文希望在满足飞行边界、避碰安全、剩余能量和动作可行性约束的条件下，提高区域有效覆盖水平和目标检测比例，同时降低任务完成时间、重复覆盖程度和通信断连风险。

2　协同搜索任务建模与QC-MAPPO算法

基于第1章建立的GNSS拒止环境下多无人机协同搜索系统模型与多约束优化问题，本文进一步将考虑随机感知噪声、累积定位漂移和感知信息缺失的协同搜索任务建模为分散式部分可观测马尔可夫决策过程，并提出面向覆盖质量的多智能体近端策略优化算法QC-MAPPO进行求解。该算法以MAPPO为基础框架，采用集中式训练、分布式执行机制。在此基础上，本文引入通信拓扑掩码注意力Critic和基于估计状态的一步安全动作掩码，使无人机集群在定位信息持续缺失和相对感知不确定的条件下，兼顾覆盖质量、目标检测、通信连通、飞行安全和能量消耗。

2.1　面向覆盖质量的多无人机协同搜索DEC-POMDP建模

GNSS拒止环境下，多无人机协同搜索具有典型的多智能体、部分可观测、状态动态演化和长期收益耦合特征。单架无人机无法直接获得全局绝对状态，只能依据自身估计状态、与基站及邻机的相对信息、局部目标观测和搜索记忆等信息进行分布式决策；同时，多架无人机的联合动作共同影响真实覆盖质量、目标检测进度、通信拓扑、定位漂移和系统能耗。因此，本文将该协同搜索过程建模为分散式部分可观测马尔可夫决策过程，表示为

$$<\mathcal{S},\{\mathcal{O}_m\}_{m=1}^{M},\{\mathcal{A}_m\}_{m=1}^{M},P,Z,R,\gamma> \tag{47}$$

其中，$\mathcal{S}$为系统状态空间，$\mathcal{O}_m$为第m架无人机的局部观测空间，$\mathcal{A}_m$为第m架无人机的动作空间，P为状态转移函数，Z为随机观测函数,R为奖励函数，$\gamma\in(0,1)$为折扣因子。

在时隙t，系统真实状态表示为

$$\begin{aligned}s(t)=\{&x_m^{phy}(t),x_m^{est}(t),\boldsymbol{\delta}_m(t),\boldsymbol{\Gamma}_m(t),\beta_m^{sen}(t),\\&z_n(t),Q_l(t),\hat{Q}_l(t)|m\in\mathcal{M},n\in\mathcal{N},l\in\mathcal{G}\}\end{aligned} \tag{48}$$

其中，$x_m^{\mathrm{phy}}(t)=\left[\mathbf{P}_m^{\mathrm{T}}(t),\psi_m(t),E_m(t)\right]^{\mathrm{T}}$和$x_m^{est}(t)=\left[\hat{\mathbf{P}}_m^{\mathrm{T}}(t),\hat{\psi}_m(t)\right]^{\mathrm{T}}$分别表示无人机$m$的物理状态和估计状态，$\boldsymbol{\delta}_m(t)$为累积漂移状态，$\boldsymbol{\Gamma}_m(t)$为感知有效性矩阵，$\beta_m^{sen}(t)$为综合感知可靠度，$z_n(t)\in\{0,1\}$表示第$n$个目标是否已被检测，$Q_l(t)$和$\hat{Q}_l(t)$分别表示网格$g_l$的真实累计观测质量和估计累计观测

质量。

为降低价值网络输入维度，算法实现中采用压缩全局状态作为Critic输入：

$$\begin{aligned} s_c(t) = [\,&\tilde{\mathbf{P}}_1(t),\cdots,\tilde{\mathbf{P}}_M(t),\tilde{\tilde{\mathbf{P}}}_1(t),\cdots,\tilde{\tilde{\mathbf{P}}}_M(t),\tilde{E}_1(t),\cdots,\tilde{E}_M(t),\\ &\beta_1^{sen}(t),\cdots,\beta_M^{sen}(t),\bar{\gamma}_1(t),\cdots,\bar{\gamma}_m(t),\frac{t}{T},\frac{D(t)}{N},C_{cum}(t),\widehat{C}_{cum}(t)\,] \end{aligned} \tag{49}$$

其中，$\tilde{\mathbf{P}}_m(t)$、$\tilde{\tilde{\mathbf{P}}}_m(t)$和$\tilde{E}_m(t)$分别表示归一化真实位置、归一化估计位置和归一化剩余能量，$\bar{\gamma}_m(t)$表示当前有效感知信息比例，$D(t)$表示累计检测目标数，$C_{cum}(t)$和$\widehat{C}_{cum}(t)$分别为真实和估计累计质量覆盖率。压缩全局状态由每架无人机的10维状态信息及4维全局任务信息组成，其维度为$d_s = 10M + 4$。

无人机m的局部观测定义为：

$$\begin{aligned} o_m(t) = \{\,&\tilde{o}_m^{\mathrm{bs}}(t),\tilde{o}_m^{\mathrm{nei}}(t),\tilde{o}_m^{\mathrm{tar}}(t),\\ &\hat{o}_m^{\mathrm{map}}(t),o_m^{\mathrm{self}}(t),\boldsymbol{\Gamma}_m(t),\beta_m^{\mathrm{sen}}(t)\,\} \end{aligned} \tag{50}$$

其中，$\tilde{o}_m^{\mathrm{bs}}(t)$表示带有随机噪声、累积漂移和有效性标志的基站相对方向及相对距离信息，$\tilde{o}_m^{\mathrm{nei}}(t)$表示当前能获得的邻机相对位置、相对距离、剩余能量和根连通状态等信息，$\tilde{o}_m^{tar}(t)$表示带误差的局部目标相对观测和候选目标记忆，$\tilde{o}_m^{map}(t)$表示由估计累计质量图得到的局部搜索记忆摘要，$o_m^{\mathrm{self}}(t)$表示自身高度、时间进度、检测进度、通信紧张度和剩余能量比例等状态，$\boldsymbol{\Gamma}_m(t)$表示各类感知信息的有效性，$\beta_m^{\mathrm{sen}}(t)$表示当前综合感知可靠度。

为使共享策略网络适应不同无人机规模，本文采用固定维度的局部观测表示。对于邻机相对观测，首先根据当前通信拓扑筛选可通信邻机，并在邻机感知有效时按照估计相对距离由近及远排序，保留距离最近的至多$M_{\max}$架邻机。每架邻机采用5维特征描述，包括机体坐标系下的三维相对位置、归一化剩余能量及基站根连通状态。若可观测邻机数量不足$M_{\max}$架，则以全零向量填充；若超过$M_{\max}$架，则仅保留排序靠前的$M_{\max}$架。邻机感知失效时，该部分观测置零。

结合基站相对观测、局部目标观测、候选目标记忆、搜索记忆及自身状态，单架无人机的局部观测维度固定为$63 + 5M_{\max}$。所有无人机共用同一组Actor网络参数，且单机动作空间维度不随集群规模变化。因此，当无人机数量发生变化时，仅需按上述规则重新构造各无人机的局部观测，无须调整Actor网络结构或重新训练策略参数。

无人机动作采用离散化三维机动方式表示。第m架无人机在时隙t的动作定义为

$$a_m(t) = (v_m(t),\Delta\psi_m(t),v_{m,h}(t)) \tag{51}$$

其中，$v_m(t)$是水平速度档位{5,10,15}m/s，并包含悬停动作；$\Delta\psi_m(t)$是相对航向增量{0°,±45°,±90°,±135°,180°}；$v_{m,h}(t)$是垂向速度档位{−2.5,0,2.5}m/s。因此，单机离散动作空间共包含75种动作，其维度不随无人机数量变化。

状态转移由无人机运动更新、通信拓扑更新、能量更新、覆盖质量更新和目标检测状态更新共同决定。在每个决策时隙t，首先依据当前估计状态构造一步安全动作掩码并采样联合动作；环境执行动作后更新真实运动状态、通信拓扑、能量、覆盖与检测状态，并生成下一时隙包含噪声、漂移和信息缺失的局部观测。各无人机根据当前有效感知信息更新自身状态估计和局部搜索记忆，并基于估计状态完成通信拓扑判断、安全动作筛选、覆盖质量和目标检测状态更新。

为采用强化学习方法求解第1.5节的多约束优化问题，将累计覆盖质量、目标检测收益、能耗、重叠覆盖、安全违规和通信断连等目标与约束项标量化为即时奖励。时隙t的全局即时奖励为

$$\begin{aligned} r(t) = {}&\omega_c\Delta C_{\mathrm{cum}}(t) + \omega_d\Delta D(t)\\ &- \omega_e E_{\mathrm{sys}}(t) - \omega_o O(t) - \omega_s S(t) - \omega_r R_{\mathrm{dis}}(t) \end{aligned} \tag{52}$$

其中，$\Delta C_{\mathrm{cum}}(t) = C_{\mathrm{cum}}(t) - C_{\mathrm{cum}}(t-1)$表示新增累计质量覆盖增量；$\Delta D(t) = D(t) - D(t-1)$表示新增检测目标数量；$E_{\mathrm{sys}}(t)$表示系统能耗；$O(t)$表示重叠覆盖惩罚；$S(t)$表示碰撞和空转惩罚；$R_{\mathrm{dis}}(t)$表示通信连通性违规惩罚。根据算法实现，奖励函数进一步细化为覆盖收益、目标检测收益、能耗惩罚、重叠覆盖惩罚、安全惩罚、通信断连惩罚和终止奖励等部分，主要权重设置如表1所示。

能耗惩罚在实现中采用归一化能耗形式，其有效惩罚系数为$\omega_e k_{ene} = 0.20 \times 0.20 = 0.04$。奖励权重按照任务优先级与数值尺度平衡的原则设置。覆盖增量、目标检测和能耗均首先进行归一化处理，其中覆盖、检测和能量主权重分别取0.35、0.45和

表1　奖励函数主要权重设置

奖励类型	取值	奖励类型	取值
覆盖收益总权重ω_c	0.35	重叠覆盖惩罚系数k_{ov}	0.10
检测收益总权重ω_{d}	0.45	碰撞惩罚系数k_{col}	0.50
能耗惩罚总权重ω_e	0.20	通信断连惩罚系数k_{conn}	0.50
恢复进度奖励系数	0.80	断连持续惩罚系数	0.35
重连奖励系数	1.50	剩余目标惩罚系数k_{rem}	0.45
边界违规惩罚系数	0.35	全检测完成奖励k_{full}	40.0
提前完成奖励系数k_{time}	8.0	未完成目标惩罚系数k_{miss}	18.0

0.20，使目标发现作为主要任务目标，同时保持对区域搜索和能量消耗的约束。碰撞和通信断连系数均设置为0.50，高于重叠覆盖和悬停惩罚的0.10，以优先抑制安全违规和拓扑失联行为；终止阶段采用40.0的全目标检测奖励和18.0的未完成目标惩罚，以增强稀疏任务完成信号。相关参数在上述原则下通过小规模预实验进行有限调整，使各奖励分量处于相近有效量级，避免单一奖励项长期主导策略更新。**2.2** QC-MAPPO算法设计与训练流程

MAPPO采用CTDE框架。设策略网络参数为θ，价值网络参数为ϕ，时间差分误差为

$$\delta_t = r(t) + \gamma V_\phi(s_c(t+1),\mathbf{A}_c(t+1)) - V_\phi(s_c(t),\mathbf{A}_c(t)) \tag{53}$$

其中，$\mathbf{A}_c(t)$为时隙t的真实通信邻接矩阵。

基于广义优势估计（generalized advantage estimation, GAE），优势函数为

$$\hat{A}_t = \sum_{k=0}^{T-t-1}(\gamma\lambda)^k\delta_{t+k} \tag{54}$$

其中，λ为平滑因子。由于动作采样前使用了安全动作掩码，策略概率比应基于掩码后的新旧策略分布计算。对于无人机m，定义新旧策略概率比为

$$\rho_{m,t}(\theta) = \frac{\pi_\theta^{\mathrm{mask}}[a_m(t)|o_m(t),\mathbf{I}_m^{\mathrm{mask}}(t)]}{\pi_{\theta_{\mathrm{old}}}^{\mathrm{mask}}[a_m(t)|o_m(t),\mathbf{I}_m^{\mathrm{mask}}(t)]} \tag{55}$$

其中，$\mathbf{I}_m^{\mathrm{mask}}(t)$为采样该动作时实际使用的安全动作掩码。在策略更新阶段必须使用经验缓存中保存的原始掩码，不能根据已变化的环境状态重新计算。

PPO裁剪策略目标函数[18]为

$$L^{\mathrm{clip}}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t,\mathrm{clip}(\rho_t(\theta),1-\epsilon,1+\epsilon)\hat{A}_t\right)\right] \tag{56}$$

价值网络采用均方误差损失更新：

$$L_V(\phi) = \mathbb{E}_t\left[\left(V_\phi[s_c(t),\mathbf{A}_c(t)] - \hat{R}_t\right)^2\right] \tag{57}$$

其中，$\hat{R}_t$为基于GAE得到的回报目标。

基础MAPPO通常直接利用全局状态输入多层感知机进行价值估计，难以显式表达通信拓扑对协同行为的影响。为此，QC-MAPPO在Critic中引入通信拓扑掩码注意力机制[19]。

构造无人机m的节点特征：

$$x_m(t) = \mathrm{Concat}[\,\overline{\mathbf{P}}_m(t),\overline{\hat{\mathbf{P}}}_m(t),\bar{E}_m(t),\beta_m(t),\bar{\gamma}_m(t),\chi_m(t),\mathbf{z}(t)\,] \tag{58}$$

其中，$\overline{\mathbf{P}}_m(t)$和$\overline{\hat{\mathbf{P}}}_m(t)$表示归一化真实与估计三维位置；$\bar{E}_m(t)$为归一化剩余能量，$\beta_m(t)$为综合感知可靠度，$\bar{\gamma}_m(t)$为有效感知比例，$\chi_m(t)$为基站根连通状态，$\mathbf{z}(t)$为全局任务特征。

设第i架无人机的嵌入特征为$\boldsymbol{e}_i$，其Query、Key和Value向量[20]分别为

$$\boldsymbol{Q}_i = \boldsymbol{W}_Q\boldsymbol{e}_i,\quad \boldsymbol{K}_i = \boldsymbol{W}_K\boldsymbol{e}_i,\quad \boldsymbol{V}_i = \boldsymbol{W}_V\boldsymbol{e}_i \tag{59}$$

定义通信拓扑掩码为

$$\mu_{ij}(t) = \begin{cases}1, a_{ij}^c(t) = 1 \text{ or } i = j\\ 0, \text{otherwise}\end{cases} \tag{60}$$

其中，$a_{ij}^c(t)$为实际通信链路状态。注意力权重为

$$\alpha_{ij} = \frac{\mu_{ij}(t)\exp(\boldsymbol{Q}_i\boldsymbol{K}_j^{\mathrm{T}}/\sqrt{d_k})}{\sum_{j'=1}^{M}\mu_{ij}(t)\exp(\boldsymbol{Q}_i\boldsymbol{K}_{j'}^{\mathrm{T}}/\sqrt{d_k})} \tag{61}$$

其中d_k为Key向量维度。

定义4维全局任务特征：

$$\mathbf{z}(t) = \left[\frac{t}{T_{\max}}, \frac{D(t)}{N}, C_{\mathrm{cum}}(t), \hat{C}_{\mathrm{cum}}(t)\right] \tag{62}$$

其中，$D(t)$为累计确认检测目标数量，N为地面目标总数，$C_{\mathrm{cum}}(t)$和$\hat{C}_{\mathrm{cum}}(t)$分别为真实和估计累计质量覆盖率。

通过通信拓扑掩码注意力得到的聚合特征为：

$$\mathbf{h}_m(t) = \sum_{j=1}^{M}\alpha_{m,j}(t)\mathbf{V}_j(t) \tag{63}$$

该节点经过拓扑注意力聚合及后续特征映射得到的上下文特征为：

$$\mathbf{c}_m(t) = \sigma(W_c\mathrm{Concat}[\,\mathbf{e}_m(t),\mathbf{h}_m(t)\,] + \mathbf{b}_c) \tag{64}$$

定义无人机m的价值网络输入为

$$g_m(t) = \mathrm{Concat}[\,\boldsymbol{c}_m(t),\mathbf{z}(t)\,] \tag{65}$$

对应价值输出为

$$V_\phi^m(t) = f_\phi(g_m(t)) \tag{66}$$

其中W_c和$\mathbf{b}_c$为共享特征映射层的参数，σ为ReLU激活函数。f_ϕ为各节点共享参数的价值输出网络。

在通信拓扑注意力聚合后，各节点上下文特征

与固定维度全局任务特征分别拼接，通过共享价值输出网络得到各无人机的状态价值估计。该结构采用逐节点共享特征映射，不对全部节点特征进行展平拼接。当无人机数量变化时，仅需相应调整节点数量和通信邻接矩阵阶数，网络可训练参数维度保持不变。

为减少训练早期无效探索，QC-MAPPO 在 Actor 动作采样前引入一步安全动作掩码[21][22]。对候选动作 $a \in \mathcal{A}_m$，其可行性掩码采用第 1.4 节定义的综合安全可行性式 $\mathrm{I}_m^{\mathrm{mask}}(a,t)$，掩码后的动作分布为

$$\pi_\theta^{\mathrm{mask}}(a|o_m(t)) = \frac{\exp\left(f_\theta(o_m(t),a)\right)\mathrm{I}_m^{mask}(a,t)}{\sum_{a'\in\mathcal{A}_m}\exp\left(f_\theta(o_m(t),a')\right)\mathrm{I}_m^{mask}(a',t)} \tag{67}$$

其中，$f_\theta(o_m(t),a)$ 为 Actor 输出的动作 logit。该机制能够在动作采样前过滤明显不可行动作，也包含已断连状态下的恢复动作筛选，使动作约束具有“断连预防—断连恢复”的闭环特性。

基于上述设计，QC-MAPPO 训练流程如算法 1 所示。

算法 1 QC-MAPPO 训练流程

1)：**输入：**环境参数、无人机数量、最大训练轮数、最大决策步数、感知噪声强度、漂移强度、感知失效概率、统一感知安全裕度、折扣因子、GAE 参数、PPO 裁剪系数。
2)：**输出：**训练后的策略网络参数。
3)：初始化 Actor、拓扑注意力 Critic 和经验缓存；
4)：**for** 训练轮次 $k = 1,2,\cdots,K_{\max}$ **do**
5)：　重置环境，获得各无人机初始局部观测；
6)：　**for** 每个环境交互步 **do**
7)：　　根据局部观测计算动作 logits；
8)：　　利用联合运动预测和安全约束构造动作掩码并采样联合动作；
9)：　环境执行联合动作，更新环境状态；
10)：　将状态、观测、动作、奖励、价值估计和旧策略概率存入经验缓存；
11)：　**end for**
12)：　基于 GAE 计算优势函数和回归目标
13)：　使用 PPO 裁剪目标更新 Actor
14)：　使用均方误差损失更新拓扑注意力 Critic
15)：　衰减策略熵系数
16)：**end for**

2.3　算法计算复杂度分析

设无人机数量为 M，单机局部观测维为 d_o，全局状态维度为 d_s，单个无人机的 Critic 节点特征维度为 d_f，全局辅助特征维度为 d_g，离散动作数为 A，隐藏层维数为 H。QC-MAPPO 与 MAPPO 采用相同的参数共享 Actor，单架无人机 Actor 一次前向传播的计算复杂度均为 $O\left[(d_oH + H^2 + HA)\right]$，因此改进方法未增加 Actor 网络的计算复杂度。

基础 MAPPO 采用两层 MLP 集中式 Critic，相应的前向传播复杂度为 $O\left[d_sH + H^2 + HM\right]$。QC-MAPPO 在集中式 Critic 中引入通信拓扑掩码机制。首先将 M 个无人机的节点特征映射至 H 维隐空间，再通过 Query、Key 和 Value 映射计算节点间注意力，并利用通信邻接矩阵屏蔽不可通信节点，最后融合全局任务特征完成价值估计。因此其主要计算复杂度可表示为 $O\left(Md_fH + MH^2 + M^2H + Md_gH\right)$。其中 MH^2 项主要来源于节点间注意力相关性及加权聚合计算。与 MAPPO 相比，QC-MAPPO 增加的计算量主要集中于节点特征映射和通信拓扑注意力计算[23-25]。

虽然通信拓扑注意力 Critic 增加了集中训练阶段的计算开销，但分布式执行阶段仅需保留共享 Actor 及非学习型安全动作筛选机制，无须部署 Critic。与基础 MAPPO 相比，QC-MAPPO 的 Actor 网络结构和可训练参数规模保持不变；一步安全动作掩码虽然不增加可训练参数，但会产生额外的候选动作预测与可行性判断开销。因此，本文的轻量化主要体现在执行端神经网络参数规模未增加[25]。

3　性能评估

3.1　仿真设置

本文在三维空间开展仿真实验，地面搜索区域范围为 800m×800m，无人机飞行高度为 50～100m。系统包含 5 架无人机、20 个随机分布地面目标和 1 个地面基站，地面区域采用 10m×10m 的网格进行离散。主要环境参数如表 2 所示。

本文采用静态目标、无障碍物及无外部风扰的基准仿真环境，所有算法在相同任务边界、目标生成规则、通信参数及感知误差设置下进行训练与测试。该设置旨在减少地形遮挡、目标运动和飞行动力学扰动等因素的耦合影响，便于分析覆盖质量评价、通信拓扑感知价值评估及一步安全动作掩码对协同搜索性能的作用。因此，现有实验结果主要反映上述统一基准条件下的算法性能，暂未讨论复杂地形、动态目标或风扰环境等因素的影响。

训练采用 TensorFlow 实现，实验平台为 Intel

Core i7-11800H CPU 和 NVIDIA GeForce RTX 3090 GPU。训练超参数如表3所示。

表2　多无人机协同搜索仿真实验环境参数设置

参数	参数值	参数	参数值
无人机数量M	5架	地面目标数N	20个
水平空间范围	[0,800]m	地面基站位置	(0,400,2)m
飞行高度范围	[50,100]m	初始飞行高度	70m
时隙长度	1s	最大决策步数	300步
网格边长	10m	传感器视场半角	45°
参考飞行高度	70m	高度分辨率指数	1.45
观测质量下限截断值	0.35	观测质量上限截断值	1.15
感知安全裕度	8m	联合运动安全裕度	8m
水平瞬时位置噪声标准差	1.5m	垂向瞬时位置噪声标准差	0.525m
瞬时航向噪声标准差	2°	水平单步漂移标准差	0.30m
垂向单步漂移标准差	0.09m	单步航向漂移标准差	0.25°
感知失效概率	0.04	感知恢复概率	0.35
漂移修正系数	0.85	可靠度衰减系数	0.025
视场边缘衰减系数	0.55	观测质量阈值	0.42
基础检测概率系数	0.82	距离衰减系数	0.018
无人机间通信半径	300m	最小安全距离	10m
基站-无人机通信半径	600m	动作掩码连通安全缓冲	40m
基础推进功率	99.66W	悬停功率	100W
水平气动阻力系数	0.01	垂向功率权重	6.0
通信功率	5W	单机最大可用能量	40000J
最多可观测邻机数	4架		

表3　模型训练超参数设置

训练参数	参数值	训练参数	参数值
训练总轮数	1500	每个epoch采样步数	4000
每回合最大时间步	300	策略网络更新次数	10
价值网络更新次数	30	评估间隔	50
每次评估回合数	40	训练随机种子	0
PPO裁剪系数	0.15	折扣因子	0.99
GAE平滑因子	0.95	初始熵系数	0.008
最小熵系数	0.0015	隐藏层规模	(128,128)
Buffer容量	4000	更新方式	整批更新
Actor学习率	0.0001	Critic学习率	0.0003

强化学习参数以PPO和MAPPO的常用配置为基础，并结合本任务训练稳定性进行有限调整。折扣因子和GAE参数分别取0.99和0.95，以兼顾长期搜索收益和优势估计方差；PPO裁剪系数设置为0.15，相较常用的0.2采用稍保守的策略更新范围，以减小多智能体联合策略快速变化造成的训练波动。Actor和Critic隐藏层宽度均取128，在模型表达能力和计算开销之间进行折中。熵系数由0.008逐步衰减至0.0015，使策略在训练前期保持充分探索，并在后期逐渐提高决策确定性。

各学习算法采用训练随机种子0进行一次训练，训练总轮数为1500轮。训练过程中每隔50轮在固定验证场景上进行40回合评估，并依据覆盖率、目标检测、通信断连、任务时间及能耗构成的综合评价分数选取最优checkpoint。验证场景与正式测试场景相互独立，测试数据不参与模型选择。正式测试阶段固定选定的网络参数，不进行额外训练或微调。

本文设置QC-MAPPO、w/o TopoCritic、w/o SafetyMask、MAPPO、IPPO和Random Policy共6种方法进行比较。w/o TopoCritic采用集中式MLP Critic替代完整的拓扑注意力Critic，保留共享Actor及一步安全动作掩码，用于评价拓扑感知价值网络的整体作用。w/o SafetyMask保留原有Actor-Critic结构，移除一步安全动作筛选机制，用于评价其对搜索与通信性能的影响。MAPPO和IPPO作为强化学习基线，Random Policy作为非学习基准。

本文各组基线对比及消融实验采用统一的通信和感知参数设置，重点验证所提方法在上述基准配置下的性能，不涉及通信半径及感知失效概率的系

统性敏感性评估。

3.2　训练收敛性分析

图2给出了三种强化学习算法在固定训练随机种子下的平均回合奖励变化。随着训练推进，各算法的回合奖励总体呈上升趋势，并在后期逐渐趋于平稳，表明在当前训练配置下均能够逐步学习有效搜索策略。

需要指出的是，QC-MAPPO的训练回报低于IPPO，并在部分阶段低于MAPPO。这可能与其采用的一步安全动作掩码有关。该机制在动作采样前筛除部分具有较高安全或通信风险的候选动作，可能使策略采取相对保守的搜索行为，并产生一定的时间与能耗代价。此外，训练回报受到随机动作采样和回合持续时间的共同影响，不能直接等同于最终确定性策略的搜索性能。

因此，图2主要用于反映当前训练过程中的策略学习趋势，不同算法的最终性能仍以固定已训练策略在独立测试场景中的评价结果为依据。

3.3　综合性能统计分析

正式测试阶段固定各算法在训练阶段选定的最优策略参数，不再进行训练或微调。采用相同的100组测试随机种子（20000～20099）生成测试场景，并执行确定性策略。各指标以“均值±标准差”表示，其中标准差反映固定已训练策略在不同测试场景下的性能波动，不代表不同训练随机种子之间的差异，统计结果如表4所示。

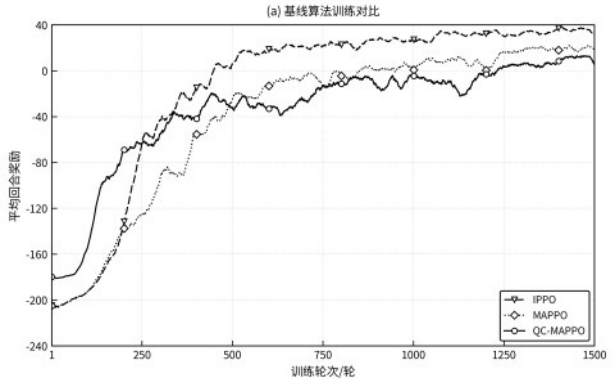


图2　基线算法的训练奖励曲线对比图

由表4可见，Random Policy在区域覆盖、目标检测和任务完成效率上均明显低于强化学习方法，表明随机运动难以形成有效的协同搜索策略。各强化学习方法均能够完成较高水平的区域搜索和目标检测，其中目标检测数量已接近任务上限，算法间差异主要体现在搜索完整性、通信稳定性及任务效率等方面。

与MAPPO和IPPO相比，QC-MAPPO有更高的最终累计覆盖率和更低的重复覆盖，明显减少通信断连，说明所提方法能够在扩大有效搜索范围的同时改善集群协同与通信保持。其完成步数和能量消耗并非最优，表明通信约束与安全动作筛选会限制部分具有较高空间扩展速度但通信风险较大的动作，使策略在任务效率和通信可靠性之间采取相对保守的折中。

消融结果中，去除一步安全动作掩码后，任务完成速度有所提高，但覆盖率下降，通信断连和重复覆盖明显增加，说明该模块能够在动作执行前有效抑制高风险行为，是改善通信稳定性的主要直接机制。与采用集中式MLP Critic的w/o TopoCritic相比，完整QC-MAPPO取得了更高的平均累计质量覆盖率和更低的平均重叠覆盖率，但累计断连步数有所增加。由于该消融同时改变了Critic的特征组织、注意力结构及网络容量，相关结果主要反映完整拓扑感知价值网络的整体效果，具体权衡见第3.4节。

表4　6种方法在测试场景的主要统计指标（均值±标准差）

算法	最终累计覆盖率/%	最终检测数量/个	平均回合终止步数/步	平均最终剩余能量/J	累计断连步数/步	平均重叠覆盖率/%
QC -MAPPO	**91.76±5.99**	19.89±0.37	185.65±57.73	19075.6±6476.5	10.54±12.99	**0.72±0.70**
w/o TopoCritic	90.11±5.19	19.75±0.54	183.32±67.21	19376.3±7455.8	7.29±9.66	1.12±1.08
w/o SafetyMask	89.80±5.14	19.86±0.37	168.12±61.99	21322.9±6815.7	33.68±17.75	1.54±1.03
MAPPO	88.78±6.42	19.81±0.44	160.44±70.34	21621.7±8059.7	39.62±24.03	1.34±1.25
IPPO	87.57±5.58	**19.92±0.37**	**132.61±51.71**	**21573.7±5559.1**	32.32±20.60	1.24±1.20
Random Policy	20.40±4.55	4.04±2.20	300.00±0.00	5458.7±87.4	**0.03±0.30**	1.99±1.29

考虑到不同算法的任务完成时间存在差异，进一步统计相同100组测试场景下各算法在固定时间步的累计质量覆盖率，结果如表5所示。对于提前终止的回合，其覆盖率保持终止值不变。QC-MAPPO在搜索前期的覆盖率增长相对较慢，但中后期逐渐取得优势。结合表4可知，QC-MAPPO较高的最终覆盖率伴随着更长的任务执行时间，该算法在前期搜索效率与后期覆盖完整性之间存在一定权衡，其优势主要体现在中后期的累计覆盖水平。

表5　固定时间步下的累计质量覆盖率（均值±标准差，%）

固定步数	QC-MAPPO	MAPPO	IPPO
100	74.21±5.87	78.69±5.75	81.30±4.60
150	86.57±5.29	87.71±6.00	86.84±4.97
200	90.85±5.50	88.49±6.35	87.35±5.35
300	91.76±5.99	88.78±6.42	87.57±5.58

另外，进一步统计QC-MAPPO的任务完成情况。以300步内检测全部20个目标作为成功条件，100组测试中有91个回合成功，成功率为91%；其余9个回合至300步仍未完成，按右删失处理。成功回合的平均完成时间为174.34±47.34步，而将超时回合按300步计入后，全部回合的平均终止步数为185.65步。完成时间累积分布表明，截至150、200和250步的累计成功率分别为31%、66%和85%，说明QC-MAPPO在所设任务时限内具有较高的任务成功率，同时避免了将超时回合与成功回合混合统计所造成的完成时间解释偏差。

综合来看，QC-MAPPO在较高区域覆盖、较低重复搜索和较稳定通信之间取得了较为均衡的综合性能，更符合GNSS拒止条件下多无人机协同搜索对搜索质量与通信可靠性的共同要求。

3.4　最终累计覆盖率与通信断连权衡分析

区域覆盖和通信连通是本文协同搜索任务中的两个重要且需要权衡（trade-off）的指标，仅根据单一覆盖率或累计断连步数难以完整反映算法的性能[26][27]，因此，以100次独立测试的最终累计覆盖率均值为横坐标，以累计断连步数均值为纵坐标绘制5种学习算法的双目标性能图，如图3所示。横坐标越大表示搜索区域完成程度越高，纵坐标越小表示通信拓扑保持能力越强。

由图3可见，MAPPO、IPPO以及w/o SafetyMask均表现出相对较高的通信断连水平，而QC-MAPPO在提高覆盖率的同时明显降低了断连，说明一步安全动作约束能够有效避免因过度空间扩张导致的通信失效。去除SafetyMask后，算法在覆盖率和通信稳定性两个方向均发生退化，进一步验证了该模块对安全协同搜索的重要作用。

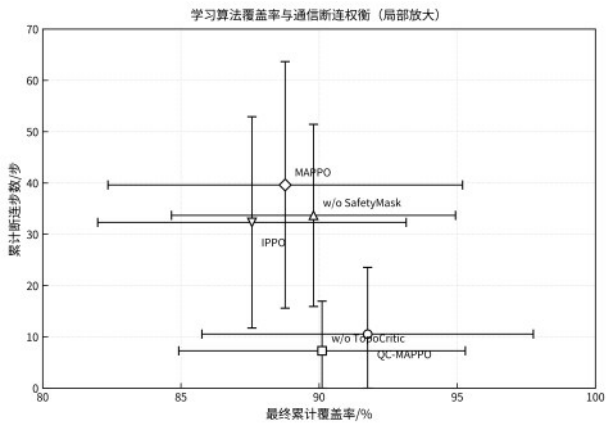


图3　学习算法覆盖率与通信断连权衡（局部放大）

与采用集中式MLP Critic的w/o TopoCritic相比，完整QC-MAPPO的最终累计质量覆盖率提高1.65个百分点，平均重叠覆盖率降低约35.7%，但累计断连步数增加3.25步，体现出覆盖完整性与通信保持之间的权衡。上述结果说明，完整拓扑感知价值网络在当前场景下取得了较好的搜索覆盖表现。由于该消融同时改变了Critic的特征组织、注意力结构及网络容量，尚不能将性能差异单独归因于通信拓扑掩码，仍需通过严格控制变量的实验进一步验证。

3.5　代表性回合空间搜索行为分析

从100次测试中选取累计覆盖率和累计断连步数接近总体均值的典型回合绘制真实、估计轨迹及累计质量覆盖分布图，如图4所示。5架无人机向不同方向扩展搜索，形成较明显的空间分工，最终有效覆盖区域较为连续。真实与估计轨迹整体趋势一致，但局部存在偏差。在此条件下，无人机仍能保持有效搜索并完成全部目标检测，较低的累计断连表明算法能够在扩大搜索范围的过程中兼顾通信拓扑稳定性，并保持较低的累计断连步数，直观展示了QC-MAPPO在所设感知不确定性条件下的协同搜索行为。

3.6　泛化性能分析

为评价QC-MAPPO对任务规模变化的适应能力，固定原始5架无人机、20个目标场景下训练所得的策略参数，不进行额外训练或微调，分别改变

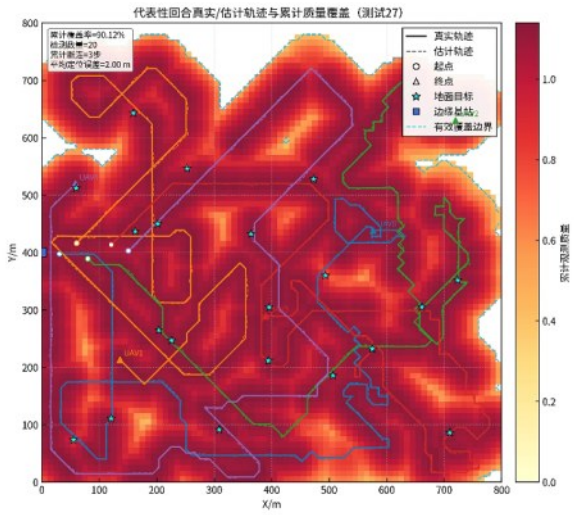


图4　代表性测试回合真实/估计轨迹与累计质量覆盖分布

测试阶段的目标数量和无人机数量，每种配置独立测试100个回合。

保持无人机数量为5，将目标数量设置为10、15、20、25和30，结果如图5所示。随着目标数量增加，最终累计覆盖率整体提高，目标检测率始终保持较高水平，而完成步数逐渐增加。其主要原因是目标数量增加后，达到全目标检测条件需要更长的搜索过程，使无人机能够进一步扩展有效搜索区域。与此同时，断连时间比例仅在较小范围内波动，未随任务规模扩大而持续增加，说明所提策略在不同目标密度下仍能够保持较稳定的通信状态。

保持目标数量为20，将无人机数量设置为3、4、5、6和7，结果如图6所示。随着无人机数量增加，目标检测率提高、完成步数明显下降，而累计覆盖率在达到训练规模后逐渐趋于稳定。其原因在于更多无人机能够并行承担不同区域的搜索任务，从而提高任务完成效率；当集群规模进一步扩大后，受有限任务区域和覆盖重叠影响，继续增加无人机所带来的边际覆盖收益逐渐减小。不同无人机规模下断连时间比例总体保持稳定，表明策略未因集群规模变化出现明显通信性能退化。参数共享策略能够复用同一Actor决策规则，也是其适应同构无人机数量变化的重要基础。

综上，固定训练所得策略在目标数量和无人机规模变化时仍表现出一定的任务适应能力，表明共享Actor与固定维度观测能够支持所设范围内的规模泛化。需要指出的是，上述实验针对目标数量作为任务先验给定的有限目标搜索场景，对于目标总

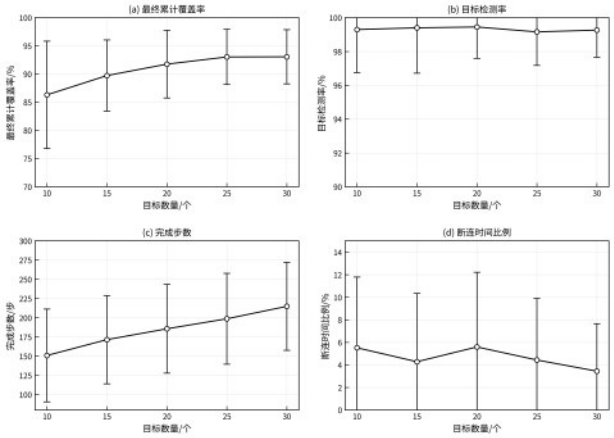


图5　不同目标数量下QC-MAPPO的泛化性能对比

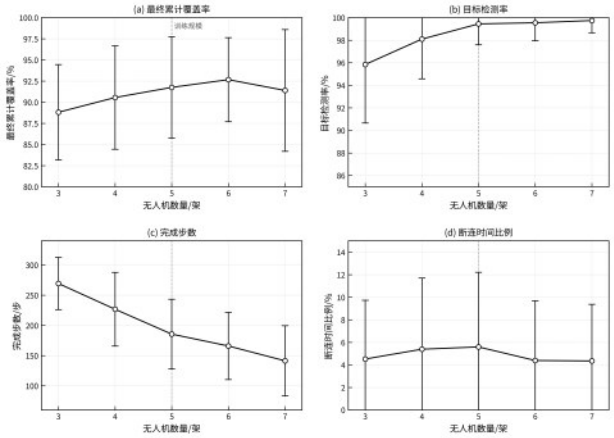


图6　不同无人机数量下QC-MAPPO的泛化性能对比

数未知的开放式搜索任务，仍需进一步研究任务进度表征与自主终止判定机制[28-31]。

4　结束语

针对GNSS拒止条件下感知误差累积、信息间歇失效及动态通信拓扑对多无人机协同搜索的影响，本文提出面向覆盖质量的多智能体近端策略优化算法QC-MAPPO。在已有视觉质量覆盖、图注意力和动作掩码方法的基础上，构建考虑感知可靠度的质量覆盖评价模型，引入通信拓扑掩码注意力Critic，并设计包含联合运动预测和断连渐进恢复机制的一步安全动作掩码，以协调区域搜索、目标检测与通信连通保持。仿真结果表明，在所设GNSS拒止搜索场景下，QC-MAPPO能够取得较高的累计质量覆盖率和目标检测水平，并有效降低通信断连与重复覆盖。消融及不同任务规模测试进一步展示了一步安全动作掩码的作用、完整拓扑感知价值网络的整体性能表现，以及策略在所设范围内

的任务规模适应能力。

本文仍存在一定局限性。现有研究采用静态目标、无遮挡及无外部风扰的简化环境，通信链路由距离门限进行二值判定，尚未考虑复杂地形遮挡、随机丢包、传输时延及信道波动，也未开展通信参数敏感性分析。此外，当前性能统计主要基于单次训练所得策略，拓扑注意力Critic的消融采用整体结构替换方式，尚不能充分评价训练随机性及拓扑掩码的独立贡献。后续将逐步扩展复杂环境与通信模型，并通过多训练随机种子、通信参数敏感性及参数匹配控制实验，进一步验证算法的适用性与关键机制的作用。

参考文献：

[1] LYU M Y, ZHAO Y B, HUANG C, et al. Unmanned aerial vehicles for search and rescue: a survey[J]. Remote Sensing, 2023, 15(13): 3266.

[2] JARRAYA I, AL-BATATI A, KADRI M B, et al. GNSS-denied unmanned aerial vehicle navigation: analyzing computational complexity, sensor fusion, and localization methodologies[J]. Satellite Navigation, 2025, 6: 9.

[3] SHEN G Q, LEI L, ZHANG X M, et al. Multi-UAV cooperative search based on reinforcement learning with a digital twin driven training framework[J]. IEEE Transactions on Vehicular Technology, 2023, 72(7): 8354-8368.

[4] LIU Y F, LI X S, WANG J, et al. Reinforcement-learning-based multi-UAV cooperative search for moving targets in 3D scenarios [J]. Drones, 2024, 8(8): 378.

[5] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2022, 35: 24611-24624.

[6] PAPATHEODOROU S, TZES A, STERGIOPOULOS Y. Collaborative visual area coverage[J]. Robotics and Autonomous Systems, 2017, 92: 126-138.

[7] GÜVEN I, YANMAZ E. Maintaining connectivity for multi-UAV multi-target search using reinforcement learning[C]//Proceedings of the ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications. New York: ACM, 2023: 109-114.

[8] LIAO G, WANG J, YANG D, et al. Multi-UAV escape target search: a multi-agent reinforcement learning method[J]. Sensors, 2024, 24(21): 6859.

[9] CHOSET H. Coverage for robotics—a survey of recent results[J]. Annals of Mathematics and Artificial Intelligence, 2001, 31: 113-126.

[10] GALCERAN E, CARRERAS M. A survey on coverage path planning for robotics[J]. Robotics and Autonomous Systems, 2013, 61 (12): 1258-1276.

[11] CABREIRA T M, BRISOLARA L B, FERREIRA P R. Survey on coverage path planning with unmanned aerial vehicles[J]. Drones, 2019, 3(1): 4.

[12] HAYAT S, YANMAZ E, BETTSTETTER C, et al. Multi-objective drone path planning for search and rescue with quality-of-service requirements[J]. Autonomous Robots, 2020, 44(7): 1183-1198.

[13] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//International Conference on Learning Representations. 2018.

[14] WANG S, MARTÍNEZ ALMANSA C, PEÑA QUERALTA J, et al. UWB-based localization for multi-UAV systems and collaborative heterogeneous multi-robot systems[J]. Procedia Computer Science, 2020, 175: 357-364.

[15] PEÑA QUERALTA J, MARTÍNEZ ALMANSA C, SCHIANO F, et al. UWB-based system for UAV localization in GNSS-denied environments: characterization and dataset[C]//Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway: IEEE Press, 2020: 4521-4528.

[16] ZHANG Z, XU X, CUI J, et al. Multi-UAV area coverage based on relative localization: algorithms and optimal UAV placement [J]. Sensors, 2021, 21(7): 2400.

[17] ZENG Y, XU J, ZHANG R. Energy minimization for wireless communication with rotary-wing UAV[J]. IEEE Transactions on Wireless Communications, 2019, 18(4): 2329-2345.

[18] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. arXiv:1707.06347, 2017.

[19] ZHAO B, HUO M, LI Z, et al. Graph-based multi-agent reinforcement learning for collaborative search and tracking of multiple UAVs[J]. Chinese Journal of Aeronautics, 2025, 38(3): 103214.

[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2017, 30: 5998-6008.

[21] HUANG S, ONTAÑÓN S. A closer look at invalid action masking in policy gradient algorithms[C]//Proceedings of the Thirty-Fifth International Florida Artificial Intelligence Research Society Conference. Hutchinson Island: LibraryPress@UF, 2022, 35.

[22] RIZVI D, BOYLE D. Multi-agent reinforcement learning with action masking for UAV-enabled mobile communications[J]. IEEE Transactions on Machine Learning in Communications and Networking, 2025, 3: 117-132.

[23] ZHAO B, HUO M, LI Z, et al. Graph-based multi-agent reinforcement learning for large-scale UAVs swarm system control[J]. Aerospace Science and Technology, 2024, 150: 109166.

[24] TOLSTAYA E, GAMA F, PAULOS J, et al. Learning decentralized controllers for robot swarms with graph neural networks[C]// Proceedings of the Conference on Robot Learning. PMLR, 2020, 100: 671-682.

[25] KHAN A, TOLSTAYA E, RIBEIRO A, et al. Graph policy gradi-

ents for large scale robot control[C]//Proceedings of the Conference on Robot Learning. PMLR, 2020, 100: 823-834.

[26] LIN J, YANG X, ZHENG P, et al. Connectivity guaranteed multi-robot navigation via deep reinforcement learning[C]//Proceedings of the Conference on Robot Learning. PMLR, 2020, 100: 661-670.

[27] HE Y, XIE J, HU G, et al. Joint optimization of communication and mission performance for multi-UAV collaboration network: a multi-agent reinforcement learning method[J]. Ad Hoc Networks, 2024, 164: 103602.

[28] WANG W, CHEN Y, ZHANG Y, et al. Collaborative search algorithm for multi-UAVs under interference conditions: a multi-agent deep reinforcement learning approach[J]. Drones, 2025, 9(6): 445.

[29] YANG X X, YAO W Q, ZHANG Y, et al. Multi-UAV cooperative search in partially observable low-altitude environments based on deep reinforcement learning[J]. Drones, 2025, 9(12): 825.

[30] QIU J, CAI Y, LI H, et al. A high- and low-level decoupled reinforcement learning method for multi-UAV cooperative search[J]. Drones, 2026, 10(7): 483.

[31] LIU M, XIE L, XIAO L, et al. Self-attention-enhanced multiagent deep reinforcement learning for UAVs target search in obstacle-dense environments[J]. IEEE Transactions on Aerospace and Electronic Systems, 2026, 62: 7533-7545.