引用格式：*熊韫文，李毅，魏才盛.面向地面移动目标观测的多星成像在线调度方法[J].飞控与探测，2025，8(5)：34-43.*
Citation：*XIONG Y W，LI Y，WEI C S. Online imaging scheduling method of multiple satellites for ground moving target observation[J]. Flight Control & Detection，2025，8(5)：34-43.*

面向地面移动目标观测的多星成像在线调度方法

熊韫文[1]，李　毅[2]，魏才盛[1*]

（1. 中南大学 自动化学院·长沙·410083；
2. 中国星网网络应用研究院有限公司·北京·100001）

摘　要：针对地面移动目标信息不可观的多星成像协同规划问题，开展了基于改进深度强化学习的在线成像调度方法研究。首先，基于卫星成像覆盖计算方法设计了一种成像条带划分算法；其次，基于可见时间窗口、条带、卫星轨道信息，建立了基于部分可观马尔可夫决策过程（Partially Observable Markov Decision Process，POMDP）的多星协同观测任务规划模型；然后，提出了一种融合动作掩码与优势度归一化机制的在线近端策略优化强化学习算法，提升了算法对求解部分条带覆盖任务区域调度问题的收敛速率；最后，通过 3 组仿真验证了所提出算法对在线求解该问题的正确性与优越性。

关键词：移动目标；成像卫星调度；部分可观马尔可夫决策过程（POMDP）；深度强化学习
中图分类号：V448.2　　**文献标志码**：A　　**文章编号**：2096-5974(2025)05-0034-10
DOI：10.20249/j.cnki.2096-5974.2025.05.004

Online Imaging Scheduling Method of Multiple Satellites for Ground Moving Target Observation

XIONG Yunwen[1]，LI Yi[2]，WEI Caisheng[1]

（1. School of Automation，Central South University，Changsha 410083；
2. China Satellite Network Application Co.，Ltd.，Beijing 100001）

Abstract：Aiming at the problem of multiple satellites collaborative planning with unobservable ground moving targets，this paper studies the online imaging scheduling method based on improved deep reinforcement learning algorithms. First，based on the satellite imaging coverage calculation method，an imaging strip partitioning algorithm is proposed. Second，based on the visible time window，strip and satellite orbit information，a multiple satellites cooperative observation mission planning model based on the partially observable Markov decision process（POMDP）is established. Then，a proximal policy optimization algorithm with an action mask and advantage normalization mechanism is proposed，which improves the convergence rate of the algorithm for solving the partial strip coverage task area scheduling problem. Finally，the correctness and superiority of the proposed algorithm are verified by three sets of simulations.

基金项目：国家自然科学基金（62373379）
作者简介：熊韫文，男，博士生。
*通信作者简介：魏才盛，男，博士，教授，博士生导师。

Keywords：moving target；imaging satellite scheduling；partially observable Markov decision process（POMDP）；deep reinforcement learning

0　引言

随着空间遥感技术的飞速发展，成像卫星以成像载荷覆盖范围大、工作时间长和不受空域限制等优点而备受关注，其应用逐渐成为气象灾害监测、环境保护以及国土安全防护的重要手段之一[1]。成像卫星主要分为光学成像卫星和雷达成像卫星两大类，其中，光学成像卫星的成像分辨率高，但易受光照环境和天气因素影响；相反，雷达成像卫星可在较差的天气与光照条件下执行观测任务，具备全天候的对地观测优势[2]。但雷达成像卫星具有成像分辨率与立体观察效果较差等缺点。因此，如何实现在轨光学、雷达等多种类型成像卫星的协同调度是提升成像卫星对地观测效益的关键。

现有解决成像卫星协同调度问题的算法可归纳为精确搜索算法与近似算法。例如，文献［3］采用穷举搜索的方法，设计了一个适用于单个卫星和少量观测的成像任务规划需求的任务规划方案。文献［4］采用迭代算法将观测需求分组排序，随后使用完全搜索技术获得最佳任务规划方案。然而，精确搜索算法需要穷尽寻优，无法快速求解出大规模的卫星成像规划问题的最优解。为克服精确搜索算法的局限，近似算法因其高效的搜索效率得到广泛关注。例如，文献［5］针对海洋移动目标成像侦测任务规划问题，在位置先验信息下动态构造了目标潜在区域与运动预测模型，并设计了一种基于模拟退火的改进遗传算法对问题进行求解，有效提升了求解速度。文献［6］针对遗传算法求解敏捷卫星成像调度时的编码问题，提出了一种二进制与实数杂合的编码方式，并将量子优化机制与遗传算法相结合，有效提高了搜索效率。文献［7］基于分支定界与两种裁剪枝规则设计了敏捷卫星任务优化调度算法，同时推导了卫星动中成像姿态规划算法，在最大化观测目标数量的基础上将成像质量调整到最优。文献［8］针对成像卫星任务中相邻目标间转换方式的问题，在转换时间、存储与能量的约束下构建了基于动态拓扑图结构的任务规划模型，并提出了一种改进的动态路径搜索算法对问题进行求解，提高了任务规划结果的准确性。

近年来，深度强化学习技术在组合优化领域得到广泛应用，其具有很强的泛化性和高速的求解速度，为成像卫星任务规划问题的解决提供了新的思路和方法[9]。深度强化学习可以描述为智能体在与环境交互的过程中，通过不断探索试验学习获得的最大累计回报策略。基于此，文献［10］提出了一种基于强化学习的敏捷卫星调度问题的通用解决方案，建立了具有连续状态空间和离散动作空间的有限马尔可夫决策过程，利用深度 Q 网络在经验数据上建立了一个在线价值函数，经过训练的 Q 网络能够有效地处理未知敏捷卫星的调度数据。文献［11］针对敏捷卫星成像任务规划问题，在深度强化学习神经网络设计中引入循环神经网络和注意力机制，使策略获得的奖励得到了一定的提升。文献［12］针对大规模任务下卫星任务调度问题，提出了一种基于图论的最小团划分算法，用于任务聚类预处理，然后采用深度确定性策略梯度算法来解决一个时间连续的卫星任务调度问题。文献［13］针对敏捷卫星多目标调度规划问题，利用加权法将多目标问题分解成子问题并分别建立了子问题的马尔可夫决策过程，然后采用深度强化学习训练得到每个子问题的解。文献［14］针对卫星观测任务规划问题约束复杂、求解空间大和输入任务序列长度不固定的问题，提出了一种多头注意力机制，对指针网络进行改进，然后通过动作评价深度强化学习算法对指针网络进行训练。然而，在处理大规模复杂成像调度问题时，现有的强化学习方法无法在可接受的时间范围内找到满足实际物理约束的解决方案，难以满足多星在线实时优化调度的任务要求。

基于以上分析，本文针对成像条带划分与地面位置不确定条件下移动目标的多星成像调度问题，开展基于改进深度强化学习的在线成像调度方法研究。首先，设计了一种成像条带划分方法，建立了该问题的部分可观马尔可夫决策过程；其次，对深度强化学习的动作选择与更新方式进行了改进；最后，通过不同初始概率与转移概率的算例验证了提出算法的正确性与优越性。

1 面向移动目标观测的多星成像调度问题描述

考虑移动目标在任务区域内机动，首先，计算得到卫星在任务时间内对任务区域的所有可见时间窗口 w_1，w_2，…，w_n；其次，根据条带划分方法计算得出每颗卫星过境时的所有条带和每个条带所包含的网格编号；然后，基于 POMDP 框架对多星成像调度问题进行建模；最后，基于所设计的算法在每个时间窗口决策出卫星应该选择哪个条带以实现概率奖励最大化。

根据移动目标成像任务规划问题的需求和特点，下面将给出该问题存在的主要约束和一些简化假设：

1）每次观测活动中卫星只能选择一个条带且不允许切换至其他条带。

2）每颗卫星只搭载一个成像载荷。

3）任务时间内移动目标在任务区域的网格内机动。

1.1 卫星覆盖计算与条带划分算法

卫星对地覆盖角 d 以及卫星对地中心角 α 分别可由以下公式计算得出

$$d=\arccos\left(\frac{R_e}{R_e+h}\right) \tag{1}$$

$$\alpha=\frac{\pi}{2}-\arccos\left(\frac{R_e}{R_e+h}\right)=\arcsin\left(\frac{R_e}{R_e+h}\right) \tag{2}$$

其中，R_e 为地球半径，h 为卫星瞬时高度。

图 1 给出了覆盖角与卫星成像角的几何关系，在直角三角形 O_eZQ 中，$\angle QO_eZ=\sigma$，可得

$$L_{O_eZ}=R_e\cos\sigma \tag{3}$$

其中，σ 为观测角。在直角三角形 O_eZS 中，有

$$L_{O_eZ}=(R_e+h)\sin\alpha_\sigma \tag{4}$$

$$\alpha_\sigma+d_\sigma+\sigma=\frac{\pi}{2} \tag{5}$$

而卫星成像角度 α_σ 已知，由式(3)、式(4)与式(5)可以计算出覆盖角 d_σ

$$\sigma=\arccos[(R_e+h)\sin\alpha_\sigma/R_e] \tag{6}$$

$$d_\sigma=\frac{\pi}{2}-\alpha_\sigma-\arccos[(R_e+h)\sin\alpha_\sigma/R_e] \tag{7}$$

A，B 两点对应的成像角度可由任意时刻卫星侧视角 ψ_t 以及卫星视场角 ϕ 计算得出，即

$$\begin{cases}\alpha_\sigma^A=\psi_t+\dfrac{\phi}{2}\\ \alpha_\sigma^B=\psi_t-\dfrac{\phi}{2}\end{cases} \tag{8}$$

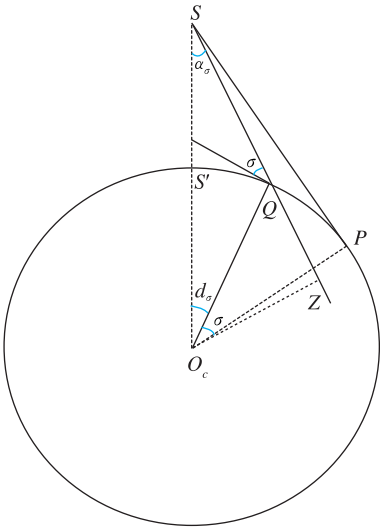


图 1 卫星对地覆盖示意图

Fig. 1 Diagram of satellite ground coverage

如图 2 所示，A，B 两点在卫星轨道左上方区域，已知卫星 S 的轨道倾角为 i，t 时刻星下点经纬度为(λ_{S_t}，φ_{S_t})，那么 A，B 点的经纬度可根据下式计算得到

$$\begin{cases}\lambda_A=\lambda_{S_t}-d_\sigma^A\sin i\\ \varphi_A=\varphi_{S_t}+d_\sigma^A\cos i\end{cases} \tag{9}$$

$$\begin{cases}\lambda_B=\lambda_{S_t}+d_\sigma^B\sin i\\ \varphi_B=\varphi_{S_t}-d_\sigma^B\cos i\end{cases} \tag{10}$$

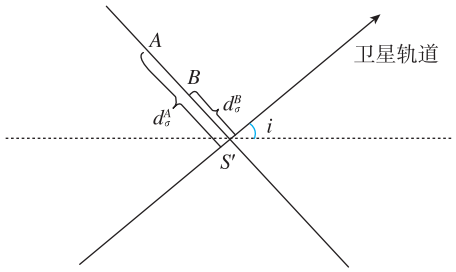


图 2 成像条带顶点计算图

Fig. 2 Diagram of imaging strip vertex calculation

由于卫星载荷的成像幅宽有限，不能完全覆盖较大的任务区域，需要对卫星成像区域进行条带划分。本文基于上述覆盖计算公式，设计了一种成像条带划分算法。

预处理：计算得到卫星过境任务区域的开始时间窗口 t_s、结束时间窗口 t_e、开始时刻星下点经纬度($\lambda_{S_{ts}}$，$\varphi_{S_{ts}}$)、结束时刻星下点经纬度($\lambda_{S_{te}}$，$\varphi_{S_{te}}$)、开始时刻轨道高度 h_s、结束时刻轨道高度 h_e。

步骤 1：根据侧视角范围 $[\psi_{\min}, \psi_{\max}]$ 和侧视角离散值 $\Delta\psi$ 计算得到离散的侧视角序列 $[\psi_{\min} : \Delta\psi : \psi_{\max}]$。

步骤 2：根据侧视角序列和视场角计算得到每个条带的两条边的角度。对于侧视角度 ψ_i，其条带对应边的角度为 $\psi_i + \phi/2$ 和 $\psi_i - \phi/2$，直至求解完所有条带。

步骤 3：将条带两条边的角度、星下点经纬度、轨道高度代入式（6）～式（12），即可得到当前条带 4 个顶点的经纬度，直至循环求解完所有条带。

通过上述算法可以求解出卫星在时间窗口 $[t_s, t_e]$ 过境任务区域时的所有成像条带，记作 $A_s = \{b_1, b_2, \cdots, b_{N_s}\}$，如图 3 所示。

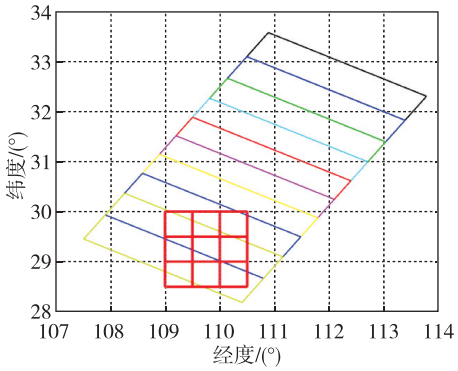


图 3　卫星成像条带划分图

Fig. 3　Diagram of strip division

1.2　POMDP 框架

马尔可夫决策过程（Markov Decision Process，MDP）框架描述了智能体与环境交互的过程。然而，本文所研究的移动目标的位置信息并不可观。因此，本文引入 POMDP 来对该问题进行建模。POMDP 框架由 $\{\boldsymbol{S}, \boldsymbol{A}, \boldsymbol{T}, \boldsymbol{R}, \boldsymbol{\Omega}, O\}$ 六元组组成：$\boldsymbol{S}$，$\boldsymbol{A}$，$\boldsymbol{R}$ 与 MDP 框架中的定义一致，分别为状态空间、动作空间、奖励；$\boldsymbol{T}$ 为状态间的转移函数集合，即在智能体在状态 $\boldsymbol{s}$ 选择动作 $\boldsymbol{a}$ 转移到状态 $\boldsymbol{s}'$ 的条件概率 $\boldsymbol{T}(\boldsymbol{s}' \mid \boldsymbol{s}, \boldsymbol{a})$；$\boldsymbol{\Omega}$ 为智能体观测到的部分信息的有限集合；O 为观察函数，即智能体在选择动作 $\boldsymbol{a}$ 后转移到状态 $\boldsymbol{s}'$ 获得观察 o 的条件概率[15]。下面将给出移动目标成像任务规划问题的 POMDP 框架。

（1）条件转移概率

在多星成像调度问题中，智能体在每个可见时间窗口 w_1，w_2，…，w_n 进行决策，状态间的转移是确定的，与智能体当前的状态与选择的动作无关。因此，所有卫星在任务时间内过境任务区域的可见时间窗口有限集 $\boldsymbol{T}$ 可表达为

$$\boldsymbol{T} = \{w_1, w_2, \cdots, w_n\} \tag{11}$$

（2）动作空间

在组网卫星中每颗卫星过境任务区域的覆盖情况不同，每颗卫星过境任务区域具有多个条带，并且每个条带所覆盖的网格也各不相同。卫星在可见时间窗口过境任务区域的条带集合 $\boldsymbol{A}_s$ 为

$$\boldsymbol{A}_s = \{b_1, b_2, \cdots, b_{N_s}\} \tag{12}$$

当选择条带 b_i 时，定义其覆盖的网格集合为 $\boldsymbol{G}_{\text{stripe}_i} = \{g_1, g_2, \cdots, g_{N_g}\}$。所有卫星的条带构成了该问题的动作空间 $\boldsymbol{A}$，即

$$\boldsymbol{A} = \{\boldsymbol{A}_{s_1}, \boldsymbol{A}_{s_2}, \cdots, \boldsymbol{A}_{s_{N_{\text{sat}}}}\} \tag{13}$$

（3）信念状态

信念状态为状态空间中所有状态的概率分布。在该问题中，将移动目标任务区域划分为网格，并利用贝叶斯概率来量化移动目标在各网格的概率信息。因此，信念状态可设计为移动目标在各网格的离散分布概率。用先验概率描述移动目标在机动过后产生的分布概率

$$\hat{\boldsymbol{P}}(t_n) = [\hat{p}_1(t_n), \hat{p}_2(t_n), \cdots, \hat{p}_{|r|}(t_n)] \tag{14}$$

定义后验概率为移动目标被观测后在网格中的概率为

$$\boldsymbol{P}(t_n) = [p_1(t_n), p_2(t_n), \cdots, p_{|r|}(t_n)] \tag{15}$$

先验概率与后验概率之间的转化公式可参考文献［16］，本文由于篇幅原因不再赘述。

（4）状态空间

设计状态为上述信念状态以及卫星覆盖情况，卫星覆盖情况包括卫星过境任务区域的经纬度 λ_{sat_i}，ϕ_{sat_i}，卫星的轨道倾角 θ_{sat_i}，所以状态可表示为

$$\boldsymbol{s} = [P(t_n), \lambda_{\text{sat}_i}, \phi_{\text{sat}_i}, \theta_{\text{sat}_i}] \tag{16}$$

每个可见时间窗口对应了一个状态，所有的状态组成了该问题的状态空间。

（5）奖励函数

定义奖励函数为条带 b_i 对应的网格集合 $\boldsymbol{G}_{b_i}$ 内，所有网格的观测收益之和再乘以该卫星的观测持续时间，即

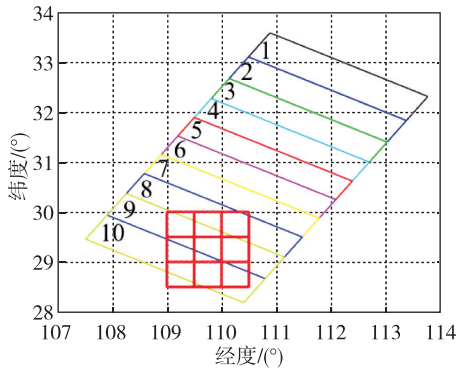
$$r(\boldsymbol{s}, b_i) = \left(\sum_{k=1}^{|\boldsymbol{G}_{b_i}|} \hat{p}_k(t_n) p_d\right)(w_e - w_s) \tag{17}$$

其中：$\boldsymbol{G}_{b_i}$ 为条带所包含的网格集合；$\hat{p}_k(t_n)$ 为网格 k 的先验概率；p_d 为发现概率；w_e，w_s 分别为

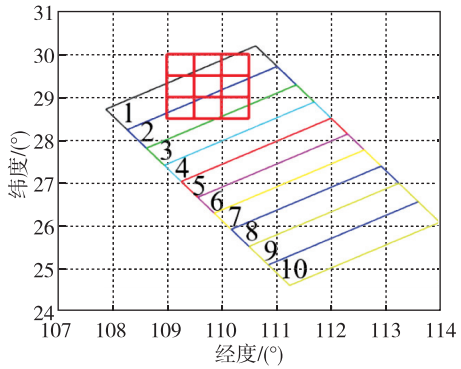
卫星可见时间窗口的开始时间与结束时间。

2 改进 PPO 算法设计

移动卫星成像任务规划问题中，每颗卫星过境任务区域的覆盖情况不同，这就会导致卫星各条带的覆盖情况不同。如图 4 所示，对条带从上至下进行顺序编号，可以看出图 4（a）中只有编号为 8，9，10 的条带覆盖了任务区域，图 4（b）中只有编号为 1，2，3 的条带覆盖了任务区域。因此，对于不同的状态，执行某些动作是没有奖励的。传统的解决方法是对奖励函数进行设计：当智能体选择到无效的动作或者边界外的动作就给予一个较差的收益。但是，这样的“软约束”无法保证智能体在训练过程中都能满足约束，从而无法确保最终学习到满足约束的最优策略。



（a）卫星 1 条带覆盖图



（b）卫星 2 条带覆盖图

图 4 不同卫星条带覆盖图

Fig. 4 Coverage of different satellite bands

为了解决上述问题，本文在 PPO 算法中加入动作掩码机制，直接在动作网络输出时屏蔽无效的动作。在 PPO 算法中，动作网络 π_θ 使用 softmax 函数，将所有动作的非标准化概率 l_{a_1}，l_{a_2}，…，l_{a_n} 转化为归一化动作概率 $\pi_\theta(a_1 \mid \boldsymbol{s})$，…，$\pi_\theta(a_n \mid \boldsymbol{s})$[17]，即

$$\begin{aligned}\pi_\theta(\cdot \mid \boldsymbol{s}) &= [\pi_\theta(a_1 \mid \boldsymbol{s}), \cdots, \pi_\theta(a_n \mid \boldsymbol{s})] \\ &= \mathrm{softmax}([l_{a_1}, \cdots, l_{a_n}])\end{aligned} \tag{18}$$

其中，softmax 函数公式为

$$\mathrm{softmax}(l_{a_i}) = \frac{\mathrm{e}^{l_{a_i}}}{\sum_j \mathrm{e}^{l_{a_j}}} \tag{19}$$

当 $l_{a_i} \to -\infty$ 时，有

$$\lim_{l_{a_i} \to -\infty} \mathrm{e}^{l_{a_i}} = 0 \tag{20}$$

因此，对于状态 $\boldsymbol{s}$ 的无效动作，可让其未归一化概率等于一个非常大的负数，那么经过 softmax 函数映射后，其概率将趋近于 0。同时，为了确保加入动作掩码机制后的损失函数可以求偏导，必须保证掩码函数可微，因此，掩码函数如下

$$\mathrm{mask}(l_{a_i}) = \begin{cases} l_{a_i}, & a_i \text{ 是有效动作} \\ M, & \text{其他} \end{cases} \tag{21}$$

其中，M 为较大的负数。那么最终得到的概率为

$$\pi'_\theta(\cdot \mid \boldsymbol{s}) = \mathrm{softmax}(\mathrm{mask}(l(\boldsymbol{s}))) \tag{22}$$

可以看出，mask 函数是一个可微的函数，那么 π'_θ 也是可微的，所以 $\partial \pi'_\theta(\cdot \mid \boldsymbol{s}) / \partial \theta$ 是存在的。

在加入动作掩码机制后，动作网络的输出层不需要激活函数，直接输出所有动作的未归一化概率，然后将其代入式（21）、式（22）就可以得到动作概率。同时，在该问题中，每颗卫星的条带覆盖情况不同，所以每个状态对应的 mask 函数都不同，因此需要将每个状态对应的 mask 函数加入到经验池中，以便训练策略在更新过程中可根据经验池中的状态得到正确的动作概率，那么改进算法的经验池由 $\{s_t, a_t, \ln p(s_t \mid a_t), r_t, v(s_t)$，done，mask$\}$ 七元组组成。其中，done 表示是否为终止状态，直接用于算法训练。改进 PPO 算法的算法框架如图 5 所示。

同时为了提升算法的收敛效率，消除更新过程中通过奖励计算得到的优势度的均值与标准差，对优势度进行归一化处理，即

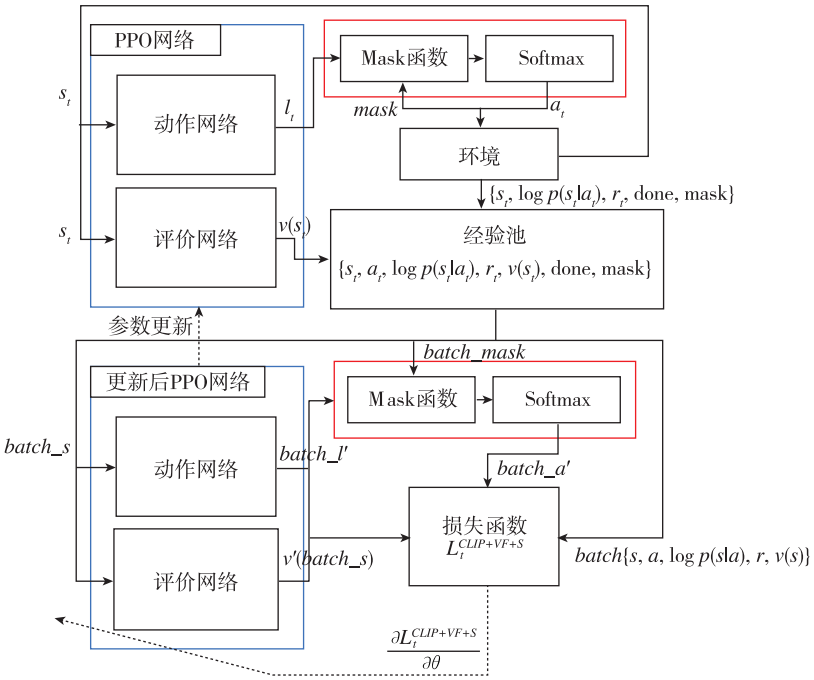


图 5　改进 PPO 算法流程图

Fig. 5　Flow chart of improved PPO algorithm

$$\hat{A}' = \frac{\hat{A} - \text{mean}(\hat{A})}{\text{std}(\hat{A}) + 10^{-5}} \tag{23}$$

其中：mean 函数为均值函数；std 函数为标准差函数，加上 10^{-5}是防止除以 0。

3　仿真结果与分析

为了验证所提出方法在求解卫星成像调度问题时的正确性与优越性，首先给定卫星可见时间窗口、成像载荷参数以及改进 PPO 算法参数设定，其次在同一参数设定下给出不同 PPO 算法的对比仿真，然后针对不同初始位置概率与目标转移概率算例给出相应的仿真结果。

（1）任务时间、任务区域以及网格设置

在仿真中，设置任务为 UTC 时间［2022/12/12　04：00：00，2022/12/13　04：00：00］，任务区域为｛(109°，30°)，(110.5°，28.5°)｝的矩形区域，以 0.5°为网格粒度将任务区域划分成 9 个网格。

（2）可见时间窗口

基于设计的部分组网卫星以及 STK 软件，可得到如表 1 所示的可见时间窗口。

表 1　可见时间窗口表

Tab. 1　Table of visible time windows

序号	开始时间	结束时间	卫星
1	2022/12/12 07：22：45	2022/12/12 07：23：31	Sat1
2	2022/12/13 01：44：22	2022/12/13 01：45：08	Sat10
3	2022/12/13 03：13：20	2022/12/13 03：14：07	Sat11
4	2022/12/12 05：03：58	2022/12/12 05：04：45	Sat12
5	2022/12/12 06：32：56	2022/12/12 06：33：43	Sat13
6	2022/12/12 23：30：38	2022/12/12 23：31：23	Sat13
7	2022/12/12 08：01：54	2022/12/12 08：02：41	Sat14
8	2022/12/13 00：59：36	2022/12/13 01：00：23	Sat14
9	2022/12/13 02：28：34	2022/12/13 02：29：21	Sat15
10	2022/12/12 04：19：11	2022/12/12 04：19：57	Sat16
11	2022/12/13 03：57：58	2022/12/13 03：58：18	Sat16
12	2022/12/12 08：51：40	2022/12/12 08：52：28	Sat2
13	2022/12/12 10：20：42	2022/12/12 10：21：09	Sat3
14	2022/12/12 21：17：26	2022/12/12 21：18：08	Sat7
15	2022/12/12 22：46：25	2022/12/12 22：47：11	Sat8
16	2022/12/13 00：15：23	2022/12/13 00：16：10	Sat9

（3）成像载荷参数设置

设置卫星载荷的侧视角范围、侧视角离散值、视场角等参数，如表 2 所示。

表 2 成像载荷参数设置表

Tab. 2 Imaging load parameters setting table

参数	设置值
侧视角范围	[−24°，24°]
侧视角离散值	6°
视场角	12°
发现概率	0.9
虚警概率	0.1

（4）移动目标初始概率与转移概率设置

将任务区域划分为 9 个网格，移动目标的初始概率设置为

$$\boldsymbol{Q}_{\text{init}}=[0.8,0.1,0.1,0,0,0,0,0,0] \tag{24}$$

移动目标更新时间间隔为 2 h，转移概率设置为

$$\boldsymbol{Q}_{\text{trans}}=\begin{bmatrix}0.1,0.9,0,0,0,0,0,0,0\\0,0.1,0.9,0,0,0,0,0,0\\0,0,0.1,0.9,0,0,0,0,0\\0,0,0,0.1,0.9,0,0,0,0\\0,0,0,0,0.1,0.9,0,0,0\\0,0,0,0,0,0.1,0.9,0,0\\0,0,0,0,0,0,0.1,0.9,0\\0,0,0,0,0,0,0,0.1,0.9\\0,0,0,0,0,0,0,0.3,0.7\end{bmatrix} \tag{25}$$

（5）网络结构与超参数设置

改进 PPO 算法的动作网络与评价网络的网络结构与超参数设置如表 3 所示，动作网络与评价网络两层隐含层的神经元个数分别为64和8，输出层的激活函数都为 Linear 函数，其他各层激活函数都为 Tanh 函数，动作网络的学习率为 0.000 3，评价网络学习率为 0.001，衰减因子为 0.99。对于该工况，设置最大迭代步数为 3 000。

表 3 网络结构与超参数设置

Tab. 3 Network structure and hyperparameter setting

参数	设置值
动作网络各层神经元个数	[6，64，8，9]
评价网络各层神经元个数	[6，64，8，1]
输出层的激活函数	Linear
其他各层神经元的激活函数	Tanh
动作网络学习率	0.000 3
评价网络学习率	0.001
衰减因子	0.99
最大迭代步数	3 000

经过训练后，得到每轮的平均奖励如图 6 所示，曲线为 3 次训练下奖励的均值，阴影为 3 次训练下平均奖励的范围。红色曲线为本文所提出的算法（PPO＋mask＋advNor），即融合动作掩码与优势度归一化的改进 PPO 算法，绿色曲线为标准 PPO 算法加上优势度归一化机制（PPO＋advNor），蓝色曲线为标准 PPO 算法（PPO）。从图中可以看出所设计的算法可以在 3 000 步内收敛，说明动作掩码机制可以有效地对无效动作进行约束，能让智能体快速地学习到满足约束的最优策略。同时，加入优势度归一化机制的 PPO 算法收敛效果也比标准 PPO 算法要好，在 500 步之前，两种算法得到的平均奖励相差不大，但在 500 步以后，加入优势度归一化机制的 PPO 算法也呈现出缓慢的收敛趋势。

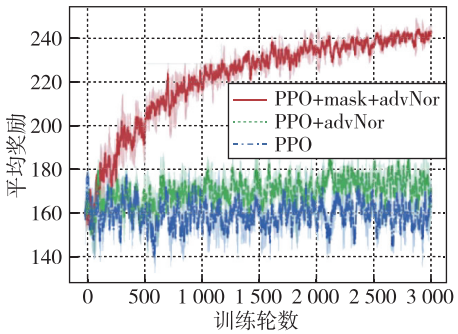
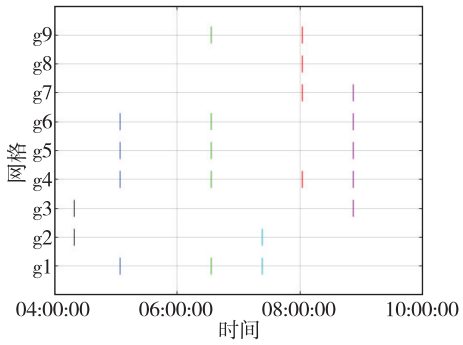


图 6 训练平均奖励图

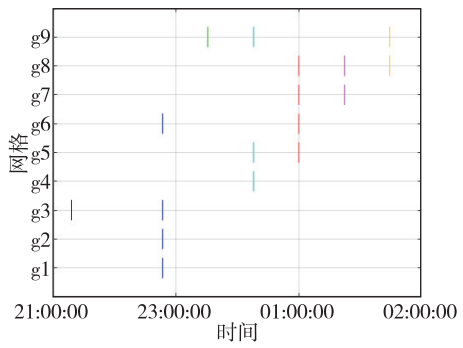
Fig. 6 Diagram of training average reward

利用训练好的策略进行测试，决策的条带结果如表 4 所示，可视化结果如图 7 所示。结果表明，训练好的策略可以选择最优的条带对移动目标进行观测，实现奖励函数最大化。在任务时域前期，智能体趋向于选择包含网格 1，2，3，4 的条带进行观测；在任务时域中期，智能体趋向于选择包含网格 5，6 的条带进行观测；在后期，智能体则选择包含网格 7，8，9 的条带进行观测，这与所设置的转移概率表征的路径大致相同。最优观测序列如图 7 所示，图中不同颜色线段代表不同卫星过境对目标进行观测，相同颜色线段集合表示当前过境选择的条带所包含的网格集合。

值得注意的是，在 2022/12/12 10：20：42，2022/12/12 21：17：26 时刻，过境任务区域的卫星 Sat3 以及 Sat7 只有一个条带覆盖了任务区域，所以这两颗卫星只能选择条带 9。



(a) 4 点到 10 点最优观测



(b) 21 点到次日 3 点最优观测

图 7　最优观测序列示意图

Fig. 7　Diagram of the optimal observation sequence

对不同初始概率与转移概率情况下的移动目标成像任务规划问题进行仿真验证。平均奖励如图 8 所示，蓝色曲线为概率相对集中条件下训练的平均奖励曲线，在该概率条件下移动目标相对容易被发现，因此得到的平均奖励也是最大的；绿色曲线为概率相对分散条件下训练的平均奖励曲线，在该概率条件下移动目标不容易被发现，因此得到的平均奖励较小；红色曲线给定的概率相对适中。从图中可以看出，3 种概率条件下的策略都能在 3 000 步内快速收敛，所设计的算法可以正确地求解移动目标成像任务规划问题。

为了实现在线对不同初始位置概率与目标转移概率情况下的成像任务规划问题进行求解，将动作网络和评价网络的第一层网络与第二层网络

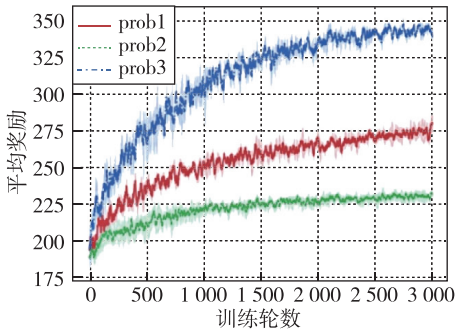


图 8　不同概率条件下训练平均奖励图

Fig. 8　Diagram of training average reward under different probability conditions

神经元个数分别设置为 400 与 200，并在每一轮训练开始随机初始化初始位置概率与转移概率。经过 15 000 轮训练，得到的平均奖励如图 9 所示。

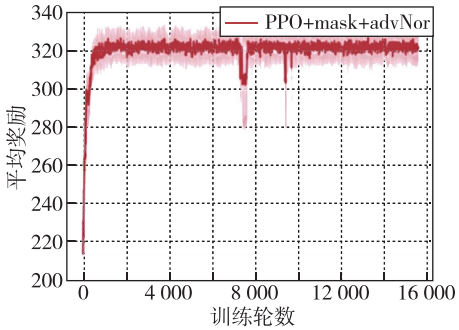


图 9　在线策略训练平均奖励图

Fig. 9　Diagram of online strategy training average reward

对训练好的策略进行 50 次在线测试，每次测试都随机初始化初始概率与转移概率。得到平均奖励如图 10 所示，在每次在线规划中，策略都能决策出正确观测条带，使得平均奖励最优。

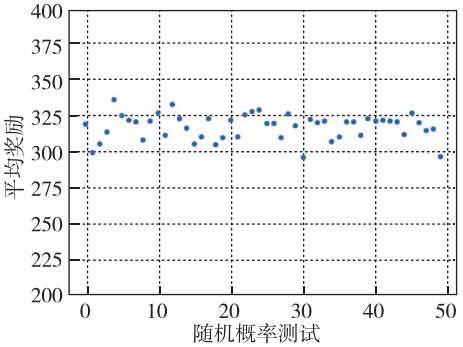


图 10　在线决策平均奖励图

Fig. 10　Diagram of online decision average reward

表 4 决策结果

Tab. 4 Decision results

序号	开始时间	结束时间	卫星	先验概率	条带与覆盖网格
1	2022/12/12 04：19：11	2022/12/12 04：19：57	Sat16	[0.8，0.1，0.1，0，0，0，0，0，0]	6：[2，3]
2	2022/12/12 05：03：58	2022/12/12 05：04：45	Sat12	[0.97，0.02，0.01，0，0，0，0，0，0]	2：[1，4，5，6]
3	2022/12/12 06：32：56	2022/12/12 06：33：43	Sat13	[0.01，0.97，0.02，0，0，0，0，0，0]	1：[1，4，5，6，9]
4	2022/12/12 07：22：45	2022/12/12 07：23：31	Sat1	[0，0.98，0.02，0，0，0，0，0，0]	7：[1，2]
5	2022/12/12 08：01：54	2022/12/12 08：02：41	Sat14	[0，0.01，0.96，0.03，0，0，0，0，0]	1：[4，7，8，9]
6	2022/12/12 08：51：40	2022/12/12 08：52：28	Sat2	[0，0.02，0.98，0，0，0，0，0，0]	9：[3，4，5，6，7]
7	2022/12/12 10：20：42	2022/12/12 10：21：09	Sat3	[0，0.01，0.11，0.86，0，0，0，0，0]	9：[1，2]
8	2022/12/12 21：17：26	2022/12/12 21：18：08	Sat7	[0，0，0，0.45，0.25，0.06，0.23，0，0]	9：[3]
9	2022/12/12 22：46：25	2022/12/12 22：47：11	Sat8	[0，0，0，0，0，0.04，0.17，0.42，0.37]	9：[1，2，3，6]
10	2022/12/12 23：30：38	2022/12/12 23：31：23	Sat13	[0，0，0，0，0，0，0.05，0.31，0.64]	1：[9]
11	2022/12/13 00：15：23	2022/12/13 00：16：10	Sat3	[0，0，0，0，0，0，0.12，0.71，0.16]	9：[4，5，9]
12	2022/12/13 00：59：36	2022/12/13 01：00：23	Sat14	[0，0，0，0，0，0，0.04，0.70，0.26]	1：[5，6，7，8]
13	2022/12/13 01：44：22	2022/12/13 01：45：08	Sat10	[0，0，0，0，0，0，0.05，0.90，0.05]	9：[7，8]
14	2022/12/13 02：28：34	2022/12/13 02：29：21	Sat15	[0，0，0，0，0，0.04，0.04，0.67，0.25]	3：[8，9]
15	2022/12/13 03：13：20	2022/12/13 03：14：07	Sat11	[0，0，0，0，0，0.03，0.28，0.13，0.56]	7：[4，5，9]
16	2022/12/13 03：57：58	2022/12/13 03：58：18	Sat16	[0，0，0，0，0，0.06，0.56，0.25，0.12]	4：[8，9]

4 结论

针对考虑成像条带划分及移动目标位置信息不可观下的多星协同观测调度问题，提出了基于改进 PPO 算法的成像任务规划方法。通过对不同初始概率与转移概率情况下的成像调度问题进行数值仿真，仿真结果表明，改进的 PPO 算法可以在 3 000 步内收敛，而标准 PPO 算法在 3 000 步内未呈现出收敛趋势，说明动作掩码机制可以有效地对无效动作进行约束剪枝，能让智能体快速地学习到能够满足约束的最优策略。因此，改进 PPO 算法可以在线正确求解移动目标成像任务规划问题。

参考文献(References)

[1] 总装备部. 卫星应用现状与发展[M]. 北京：中国科学技术出版社，2001.
The General Equipment Department. Current status and development of satellite applications[M]. Beijing：Science and Technology of China Press，2001 (in Chinese).

[2] 王永刚，刘玉文. 军事卫星及应用概论[M]. 北京：国防工业出版社，2003.
WANG Y G，LIU Y W. Introduction to military satellites and applications[M]. Beijing：National Defense Industry Press，2003(in Chinese).

[3] VERFAILLIE G，SCHIEX T. Solution reuse in dynamic constraint satisfaction problems[C]// Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence，Seattle：AAAI. 1994，94：307-312.

[4] CORDEAU J F，LAPORTE G. Maximizing the value of an Earth observation satellite orbit[J]. Journal of the Operational Research Society，2005，56(8)：962-968.

[5] 冉承新，王慧林，熊纲要，等. 基于改进遗传算法的移动目标成像侦测任务规划问题研究[J]. 宇航学报，2010，31(2)：457-465.
RAN C X，WANG H L，XIONG G Y，et al. Research on mission-planing of ocean moving targets imaging reconnaissance based on improved genetic algorithm[J]. Journal of Astronautics，2010，31(2)：457-465 (in Chinese).

[6] 王海蛟，贺欢，杨震. 敏捷成像卫星调度的改进量子遗传算法[J]. 宇航学报，2018，39(11)：1266-1274.
WANG H J，HE H，YANG Z. Scheduling of agile satellites based on an improved quantum genetic algorithm[J]. Journal of Astronautics，2018，39(11)：1266-1274(in Chinese).

[7] 陈雄姿，谢松，蔡熙，等. 敏捷卫星动中成像自主任务规划算法[J]. 宇航学报，2023，44(11)：1693-1705.

CHEN X Z, XIE S, CAI X,et al. Algorithms of autonomous mission planning for agile satellite active push-broom imaging[J]. Journal of Astronautics, 2023, 44(11):1693-1705(in Chinese).

[8] 王建江，邱涤珊，贺川，等. 考虑目标间不同转换方式的成像卫星调度[J]. 宇航学报，2012,33 (12)：1806-1814.
WANG J J, QIU D S, HE C, et al. Scheduling of imaging satellite with different transition modes between adjacent targets[J]. Journal of Astronautics, 2012, 33(12): 1806-1814(in Chinese).

[9] 李凯文，张涛，王锐，等. 基于深度强化学习的组合优化研究进展[J]. 自动化学报，2021，47(11)：2521-2537.
LI K W, ZHANG T, WANG R, et al. Research reviews of combinatorial optimization methods based on deep reinforcement learning[J]. Acta Automatica Sinica, 2021, 47(11): 2521-2537(in Chinese).

[10] HE Y, XING L, CHEN Y, et al. A generic Markov decision process model and reinforcement learning method for scheduling agile earth observation satellites[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 52(3): 1463-1474.

[11] CHEN M, CHEN Y, CHEN Y, et al. Deep reinforcement learning for agile satellite scheduling problem[C]// 2019 IEEE Symposium Series on Computational Intelligence (SSCI). Xiamen: IEEE, 2019: 126-132.

[12] HUANG Y, MU Z, WU S, et al. Revising the observation satellite scheduling problem based on deep reinforcement learning[J]. Remote Sensing, 2021, 13(12): 2377.

[13] WEI L, CHEN Y, CHEN M, et al. Deep reinforcement learning and parameter transfer based approach for the multi-objective agile earth observation satellite scheduling problem [J]. Applied Soft Computing, 2021, 110: 107607.

[14] 马一凡，赵凡宇，王鑫，等.基于改进指针网络的卫星对地观测任务规划方法[J].浙江大学学报(工学版)，2021,55(2):395-401.
MA Y F, ZHAO F Y, WANG X, et al. Satellite earth observation task planning method based on improved pointer networks[J]. Journal of Zhejiang University (Engineering Science), 2021, 55(2):395-401 (in Chinese).

[15] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains[J]. Artificial Intelligence, 1998, 101(1-2): 99-134.

[16] 慈元卓，贺仁杰，徐一帆，等.卫星搜索移动目标问题中的目标运动预测方法研究[J].控制与决策，2009，24(7):1007-1012.
CI Y Z, HE R J, XU Y F, et al. Method of target motion prediction for moving target search by satellite [J]. Control and Decision, 2009,24(7):1007-1012(in Chinese).

[17] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [J]. arXiv preprint arXiv:1707.06347, 2017. Available at: https://arxiv.org/abs/1707.06347.