# 基于多智能体深度强化学习的低轨星座跳波束资源调度研究

张晨[1,2]，徐阳威[1,2]，李宛静[1,2]，王威[1,2]，张更新[1,2]
（1. 南京邮电大学通信与信息工程学院，江苏 南京 210003；2. 南京邮电大学通信与网络技术国家工程研究中心，江苏 南京 210003）

**摘　要：**针对低轨星座跳波束资源调度的需求，提出一种基于多智能体深度强化学习的低轨星座跳波束资源调度方法。通过多目标优化的选星接入方式，建立卫星与服务区域之间的映射关系。在此基础上，根据业务类型和QoS需求的多样性，采用混合专家模型的方法，构建一个资源调度多智能体，用于星上资源与跳波束图案的实时决策调度。仿真结果表明，与传统方法相比，所提资源调度方法不仅能满足不同业务对时延和吞吐量的性能需求，还能有效平衡算法的复杂度，适应多样化业务的融合传输需求，应对业务流量的时空分布不均和动态变化，具有较强的泛化能力。

**关键词：**低轨卫星；跳波束；深度强化学习；资源调度
**中图分类号：**TN927
**文献标志码：**A
**DOI：**10.11959/j.issn.1000−436x.2025009

# Research on low earth orbit constellation beam hopping resource scheduling based on multi-agent deep reinforcement learning

ZHANG Chen[1,2], XU Yangwei[1,2], LI Wanjing[1,2], WANG Wei[1,2], ZHANG Gengxin[1,2]
1. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China
2. National Engineering Research Center of Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

**Abstract:** A low earth orbit constellation beam hopping resource scheduling method based on multi-agent deep reinforcement learning was proposed to meet the requirements of low earth orbit constellation beam hopping resource scheduling. The mapping relationship between the satellite and the service area was established by optimizing the access of multi-target satellite selection. On this basis, according to the diversity of service types and QoS requirements, based on the concept of mixture of experts, a resource scheduling multi-agent was constructed to carry out real-time decision scheduling of on-board resources and beam hopping patterns. The simulation results show that compared with the traditional methods, the proposed resource scheduling method can not only meet the performance requirements of different services on delay and throughput, but also effectively balance the algorithm complexity. At the same time, the algorithm can adapt to the converged transmission requirements of diversified services, cope with the uneven spatiotemporal distribution and dynamic changes of traffic and have strong generalization ability.

**Keywords:** low earth orbit satellite, beam hopping, deep reinforcement learning, resource scheduling

## 0　引言

低轨卫星通信系统因其覆盖范围广、时延低和部署快等特点，已成为未来空天地一体化通信系统的关键部分[1]。然而，由于业务需求在时空上的不均匀分布和动态变化，传统的多波束低轨卫星系统采用固定资源分配策略，这种方法缺乏

收稿日期：2024−08−01；修回日期：2024−12−12
基金项目：国家重点研发计划基金资助项目（No.2022YFB2902600）
**Foundation Item：**The National Key Research and Development Program of China (No.2022YFB2902600)

灵活性，导致资源浪费[2]。而跳波束技术采用时间分片技术，在同一时隙中只激活部分波束，从而显著提高卫星的资源利用率并解决资源碎片化配置问题[3]，因此其被视为下一代低轨星座通信系统的关键技术。

国内外学者针对低轨跳波束的资源调度开展了大量研究，文献[4]和文献[5-6]分别使用遗传算法和启发式算法进行跳波束系统资源分配，以提高系统的吞吐量。另外，以降低系统业务时延为目标，文献[7]采用梯度理论推导出数据包排队时延的闭式解，有效减少了业务等待时间。文献[8]通过将波束位置划分问题转化为p-center问题，实现了用最少的波束位置覆盖所有用户，从而降低数据排队时延。尽管上述传统优化方法在一定程度上提升了系统性能，但它们存在时效性低、优化目标单一且难以适应高动态性的低轨卫星场景等问题。因此，借助机器学习方法赋能资源调度成为研究热点，而深度强化学习方法由于在序贯决策问题上的优异性能，被应用于卫星资源调度算法。文献[9]提出一种以最大化吞吐量为目标的深度强化学习算法，用于优化跳波束系统容量。文献[10-11]建立了最小化排队时延模型，并利用深度强化学习方法对波束进行灵活分配，从而降低数据传输时延。此外，综合考虑多目标优化问题，文献[12]提出一种用于波束管理和资源配置协作的深度强化学习方案。文献[13]通过改进的多目标优化深度强化学习（DRL-MOP, deep reinforcement learning multi-objective optimization）算法显著提升了系统性能。文献[14]针对基于DVB-S2X标准的跳波束卫星，提出一种无模型多目标深度强化学习方法，结合双环学习（DLL, double-loop learning）的多动作选择策略，能够根据用户需求和信道条件动态分配资源。文献[15]将每颗低轨卫星视为一个智能体，在星地协同的场景下，通过多星多智能体深度强化学习进行跳波束资源智能调度。

综上所述，虽然传统优化方法在系统场景和业务需求发生变化时能够重新进行优化迭代，但其灵活性不足，不适用于动态变化的低轨卫星资源管理。此外，现有的基于深度强化学习的资源调度方法存在以下不足。首先，适用场景和通信体制受限，部分方法只适用于DVB-S2X体制的高轨卫星，无法完全适用于低轨跳波束卫星。其次，缺乏泛化性，对单颗卫星在静态或半静态环境下训练的神经网络难以推广到整个星座的其他低轨卫星，也未能充分体现低轨卫星运动导致的接续服务特性。最后，现有方法在优化目标和智能体构建上存在限制，一些方法只考虑单目标的资源调度，或者在进行多目标资源调度优化时没有考虑用户层面的服务质量（QoS, quality of service）需求；同时，构建的多智能体架构将每颗卫星视作一个智能体，在训练过程中需要多个智能体之间进行交互，对于大规模低轨星座可视范围内卫星数量较多的情况，可能导致训练开销大幅增加。

因此，本文旨在构建一个可解释、轻量化、强泛化性的低轨跳波束资源智能调度策略，提出一种基于多智能体深度强化学习的低轨跳波束资源调度方法。首先，综合考虑低轨卫星的高动态性、业务需求的多样性与时空不均、有限的星上处理能力及全频谱复用导致的同频干扰问题，建立低轨跳波束资源调度系统模型。利用低轨星座的多重覆盖和密集波束特性，通过多目标的选星优化接入方式，完成星间切换与接续传输，构建低轨卫星与服务区域的时变映射关系。然后，借鉴混合专家模型（MoE, mixture of expert）的概念，并根据3GPP协议对不同业务的QoS需求，以最小化实时数据时延、最大化非实时数据吞吐量和最大化服务小区的业务满意度为优化目标，提出一种基于MoE的多智能体模型深度强化学习算法。仿真结果表明，本文所提出的资源调度方法能够有效应对业务流量的时空不均与动态变化，具有较强的泛化性，并在保证算法复杂度可控的同时，满足不同业务对时延和吞吐量等性能的需求，灵活应用于多样化业务混合传输的实际系统场景。

区别于现有的基于深度强化学习的低轨卫星跳波束资源调度算法，本文的主要贡献和创新点如下。

1) 资源调度架构与维度方面

本文在大时空尺度上，充分考虑低轨卫星的高动态性，利用低轨星座的多重覆盖和密集波束特性，通过多目标选星优化接入，完成星间切换与接续传输，建立低轨卫星与服务区域的时变映射关系。在资源调度智能体的管理和部署上，采用信关站训练与低轨卫星星上决策相结合的模式，有效减少星上的计算和训练开销。在小时空尺度上，每颗

低轨卫星在各自的过顶服务期内，根据不同业务的QoS需求，利用资源调度多智能体完成星上时隙、功率资源的实时调度、频谱资源的干扰规避以及跳波束图案的实时决策。

2) 智能体的模型设计方面

本文引入MoE架构，创建了一个用于处理大规模数据集上复杂任务的神经网络模型。该模型能自适应地组合多个专家网络来处理不同的数据子集，实现多维资源的同时调度，以满足实时和非实时业务对时延、吞吐量等不同的QoS需求。此外，该模型具有较强的泛化性，能有效应对业务需求的时空二维动态变化，并且具有较低的推理开销，适合在星上有限的计算与处理能力环境中使用。

3) 基于深度强化学习的调度算法方面

在目标函数设计方面，根据3GPP标准对不同业务类型的QoS指标进行定义，分别以最小化数据时延、最大化非实时数据吞吐量和最大化服务小区业务满意度为训练目标。在奖励函数和神经网络结构方面，考虑目标函数的复杂性，采用多样化的$Q$网络结构，有效降低算法的复杂度；在训练方法方面，采用去中心化的多智能体训练模式，不仅有效提升了训练速度，也降低了智能体之间的交互成本。

## 1　系统模型

### 1.1　低轨星座的跳波束模型

跳波束技术采用时间分片原理进行大时空尺度上的资源调度，其中卫星上的功率、带宽等资源以跳波束时隙为颗粒度分配给各波束，并且各波束可共享星上的无线资源[16]。因此，构建高效的时间分片结构和明确关键参数的定义是实现跳波束卫星系统资源分配的首要任务。针对低轨跳波束卫星系统中多星、多波束、多业务融合传输、多维资源调度的需求，本文以高效通信为目标，拟采用如图1所示的低轨卫星跳波束时隙分片立体架构。

在图1中，低轨跳波束时间分片的立体架构由低轨卫星、波束、跳波束时隙3个维度构成。$X$轴为跳波束时隙，一个时隙的持续时间为$T_s$，表示最小的时间资源尺度；$Y$轴为波束序号，表示某个低轨卫星的波束；$Z$轴为低轨星座中的卫星序号。综上所述，有$T_s(i,j,n)=N$成立，表示第$n$号卫星的第$j$个波束，在第$i$个跳波束时隙指向第$N$号波位，与波位内的用户进行通信。由于本文考虑的低轨卫星跳波束系统具有星上处理和转发的能力，星上调度器可以根据缓存的数据包、业务的QoS需求和系统可用资源等，实时进行跳波束资源调度，从而产生如图1所示的跳波束时隙表。这与传统基于星上透明转发的跳波束系统不同，后者根据用户的业务申请，提前生成下一个调度周期的跳波束时间规划表（BHTP, beam hopping time plan），因此缺乏调度的灵活性。

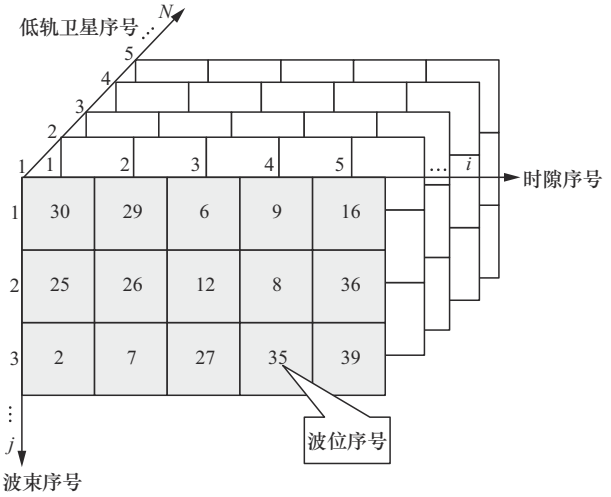



图1　低轨卫星跳波束时隙分片立体架构

此外，低轨星座跳波束不是同步地球轨道（GEO, geosynchronous earth orbit）跳波束的复制延伸或简单移植。采用低轨卫星实现跳波束具有一些特殊性，首先，区别于GEO跳波束系统的单颗卫星且对地静止，低轨卫星因高速移动、过顶时间短及服务区域不断变化，并且叠加业务需求的变化，使得低轨星座中的每颗卫星的接入连接关系和服务用户呈快速动态变化，需要进行星间切换以完成业务的接续传输。其次，低轨卫星星座表现出单层节点数量多、多层立体覆盖的特性，导致同一地面波位可以被同轨或异轨的多颗卫星覆盖，即在满足最小仰角的条件下，用户终端存在多颗候选服务卫星。最后，在波位设计方面，采用固定式波位（波位相对地面固定，不随卫星移动），利用星载相控阵天线的优势，波束指向可以快速灵活地调整，以服务于所指向的波位。更重要的是，固定式波位在低轨星座多重覆盖的条件下，避免了移动式波位（波位随卫星运动而移动）各卫星之间需要协调规划的问题。

如图2所示，深色区域P表示重叠区域，卫星C与卫星D表示同轨卫星，而卫星A、卫星B处于不同轨道高度，代表异轨卫星。区域P处于卫星A、B、C和D的多重覆盖区域内。

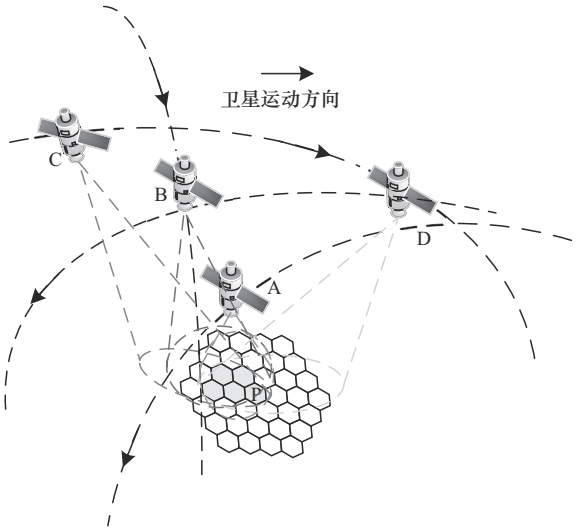



图2　多星覆盖下的接入场景示意

因此，在面向低轨星座的跳波束系统进行星间切换时，多重覆盖区域内的地面波位需要从多个候选卫星中选择合适的卫星进行业务的接续传输。针对上述问题，本文综合考虑星地距离、卫星可服务的时长等因素，采用加权策略进行优化选星接入[17]，以选择效用函数值$p_i$最大的卫星进行服务，如式(1)所示。

$$p_i = \alpha \frac{t_i}{T} - \beta \frac{d_i}{D} \tag{1}$$

其中，$\alpha$、$\beta$是2个归一化的参数，分别代表各优化目标的权重系数，可以根据不同的业务类型进行赋权[17]；变量$t_i$和$d_i$分别为卫星$i$的剩余服务时间和星地传输距离，$T$和$D$则分别代表在可视卫星集合中的最长剩余服务时间和最长星地传输距离。由于星地距离$d_i$越大，对通信传输的影响越不利，但$\frac{d_i}{D}$的取值反而越大，从而对效用函数值$p_i$起到了消极作用，因此式(1)中采用的是减法运算。综上所述，通过式(1)建立了跳波束服务卫星与服务区域随时间变化的映射关系[18-19]。值得注意的是，多重覆盖是单层星座和多层星座都具有的特性，因此，本文提出的多目标选星优化方法同样适用于这2种情况，只是在多层星座中，多重覆盖的程度更大，候选的服务卫星集合中的卫星数量也更多。

### 1.2　链路建模

低轨卫星跳波束通信系统前向链路模型如图3所示，包括信关站、网络控制中心、低轨卫星以及地面波位等。低轨卫星采用星上处理转发（再生）模式进行跳波束资源调度，同时工作的波束数量为$\mathcal{K} = \{k|k = 1,2,3,\cdots,K\}$，$K < N$，$N$为地面波位总数，表示为$\mathcal{N} = \{i|i = 1,2,3,\cdots,N\}$。

低轨卫星跳波束系统采用全频复用方法，各波束可以使用卫星的全部带宽资源$B$，以提高系统资源的利用率。然而，由于采用了全频复用方法，波束间的共信道干扰成为不可忽视的问题。对上述干扰问题进行建模[17]，如图4所示，其中卫星星下点

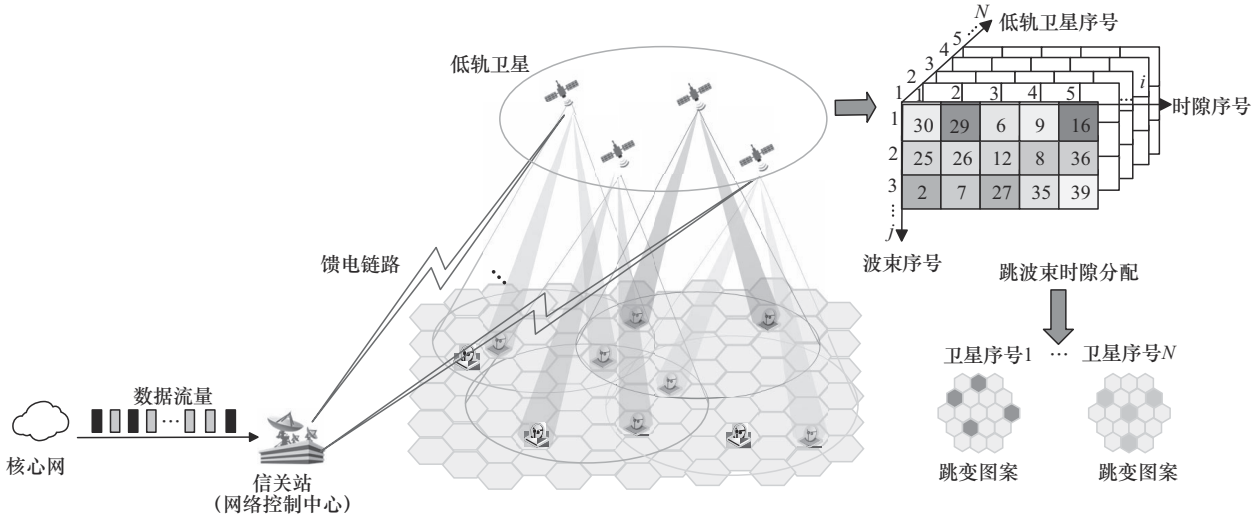



图3　低轨卫星跳波束通信系统前向链路模型

为$o$，卫星轨道高度为$h$，卫星通过波束$b_n$向波位$n$发送期望信号，传输距离为$d_n$，同时卫星通过波束$b_m$为波位$m$服务，传输距离为$d_m$，则波束$b_n$受到干扰波束$b_m$的影响。

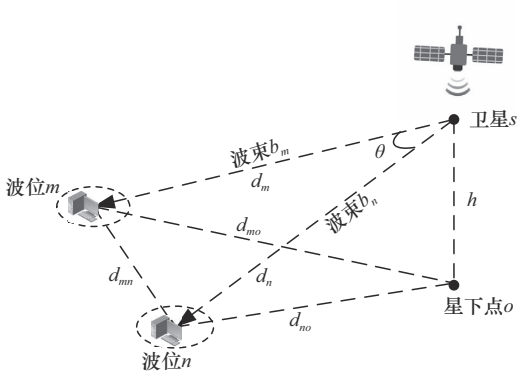



图4　波束间干扰示意

如果波束$b_n$受到干扰波束$b_m$的干扰值为$I_{mn}$[20]，则$I_{mn}$可通过式(2)计算。

$$I_{mn}=\frac{g_m P_m G\left(\theta_{mn}\right)\lambda^2}{(4\pi d_{mn})^2} \tag{2}$$

其中，$g_m$为卫星到波位$m$的天线增益，$P_m$为卫星对波位$m$的发射功率，$G(\theta_{mn})$是波束$b_m$对波位$n$的天线增益，由式(3)计算可得[21]

$$G\left(\theta_{mn}\right)=G_T\left[\frac{J_1\left(u_{mn}\right)}{2u_i}+36\frac{J_3\left(u_{mn}\right)}{{u_i}^3}\right]^2 \tag{3}$$

其中，$J_1(\cdot)$和$J_3(\cdot)$分别为一阶和三阶贝塞尔函数[22-23]。$u_{mn}$由式(4)计算可得

$$u_{mn}=2.071\,23\frac{\sin\theta_{mn}}{\sin\theta_{3\,\mathrm{dB}}} \tag{4}$$

其中，$\theta_{mn}$为波束$b_n$和波束$b_m$之间的夹角，由式(5)计算可得

$$\theta_{mn}=\arccos\frac{d_{mo}^2+d_{no}^2+2h^2-d_{mn}^2}{2\sqrt{(d_{mo}^2+h^2)(d_{no}^2+h^2)}} \tag{5}$$

由上述分析可知，在某一时刻，波位与星下点$o$的距离$d_{mo}$和$d_{no}$确定后，干扰$I_{mn}$主要与干扰波位$n$与受干扰波位$m$的距离$d_{mn}$有关。

在跳波束时隙$t$，如果波位$n$被波束$i$服务，则系统可提供的容量（理想吞吐量）表示为[24]

$$C_i=B\mathrm{lb}(1+\mathrm{SINR}_i) \tag{6}$$

其中，$\mathrm{SINR}_i$表示波位$i$的信干噪比，如式(7)所示。

$$\mathrm{SINR}_i=\frac{P_i G_i^{\mathrm{T}} G_k^{\mathrm{R}} H_i}{L_{\mathrm{SL}}\left(\sum\limits_{m\in\varphi} I_{mn}+N_0\right)} \tag{7}$$

其中，$B$为全频复用带宽，$G_i^{\mathrm{T}}$和$P_i$分别为波束$i$的天线发射增益和发射功率，$G_k^{\mathrm{R}}$为用户$k$的接收天线增益，$H_i$为星地信道系数，$L_{\mathrm{SL}}$为自由空间传播损耗，$\sum\limits_{m\in\varphi} I_{mn}$为受到邻近波束的同频干扰功率之和，$N_0$为噪声功率[24]。

### 1.3　波位划分

本文将仿真区域划分为12个波位，每个波位的半径约为73 km，如图5所示。传统的高轨跳波束系统受限于星载天线的能力，通常采用波束分簇的设计方法，即若干个波位形成一簇，由同一个波束进行跳变服务。而低轨卫星跳波束系统大多采用相控阵波束技术，这种波束的指向性可以在卫星覆盖区域内灵活调度。因此，波束可以对卫星覆盖范围内的任意波位进行服务，这增加了资源调度的灵活性，并与前文提到的固定式波位设计相匹配。

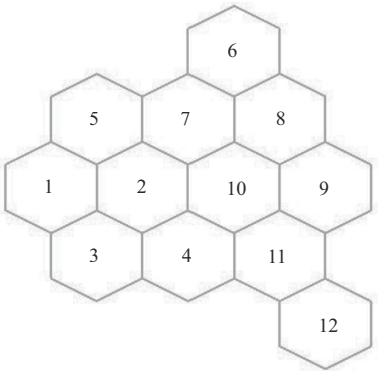



图5　波位划分示意

### 1.4　业务建模

基于跳波束的低轨卫星通信系统主要承载卫星宽带业务，这些业务主要涉及人与人之间的语音或数据通信。这类业务具有类型多样化、空间分布不均、随时间动态变化等特点，因此，本文从以下3个维度建立卫星宽带业务量模型。

1) 空间分布

根据文献[17]，本文综合考虑区域经济发展程度、卫星服务的普及程度以及卫星服务在通信业务中的市场占有率等因素，结合地理格栅的划分方法，可以建立反映不同区域业务量差异化分布的模型，如图6所示，各波位序号与图5的波位划分对应。

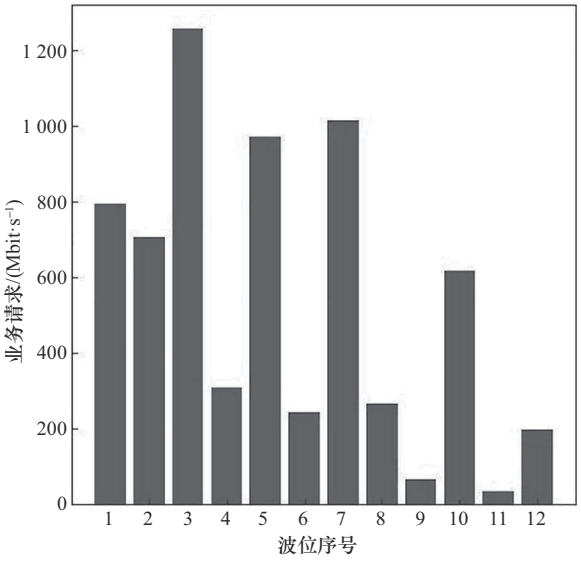



图6　空间业务量模型

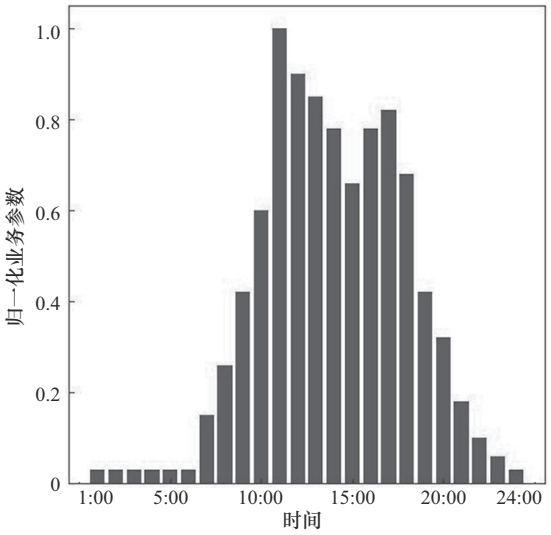



图7　时间业务量模型

为了描述业务空间分布的不均性，并定量分析业务分布不均对资源调度算法的影响，本文引入了业务空间分布离散系数$\zeta$，用以表征地面各波位业务量需求分布的不均程度[25]，$\zeta$的计算方式为

$$\zeta = \frac{S_{\mathrm{ATD}}}{\mathrm{Mean}_{\mathrm{ATD}}} \tag{8}$$

其中，$\mathrm{Mean}_{\mathrm{ATD}}$为地面波位业务量均值，$S_{\mathrm{ATD}}$为地面波位业务量标准差。

2) 时间周期性

卫星宽带业务量不仅具有空间分布的区域差异性，还会随时间的变化呈周期性变化。为了分析业务量变化受时间周期因素的影响，本文引入归一化的业务量时间加权因子[14]。此因子考虑日常活动规律，包括业务的波峰和波谷。在一天（24 h）的不同时间段，根据活动的高峰期和低谷期，给业务量分配不同的时间加权因子，如图7所示。通过将时间加权因子与各波位的业务需求峰值加权相乘，构建业务量时间变化模型。

3) 业务类型

根据3GPP R17标准[26-27]，针对不同的QoS需求以及对资源调度的不同要求，将语音会话、视频会话、流媒体视频、数据业务等多种业务划分为实时业务和非实时业务两大类。为了不失一般性，假设这两类业务在总业务量中所占的比例相同。根据文献[28]，采用泊松模型来描述实时和非实时业务数据包的到达过程。

### 1.5　最优化问题建模

在图8所示的业务QoS驱动的低轨卫星跳波束多维资源调度中，波位总数为$N$，每个波位的业务需求包括实时数据和非实时数据，卫星为每个波位提供有限长度的队列，用于存储对应波位的业务需求。地面波位的业务需求以数据包的形式存储在卫星提供的队列中，每个时隙内各波位的业务需求可以表示为泊松过程$\Lambda_t^i = \{\lambda_{1,t}^i, \lambda_{2,t}^i | i \in \mathcal{N}\}$，其中$\lambda_{1,t}^i$为在时隙$t$时波位$i$的实时数据包到达量，$\lambda_{2,t}^i$为在时隙$t$时波位$i$的非实时数据包到达量。在卫星与地面波位进行通信服务过程中，卫星各个波位队列中实际存在的实时数据包与非实时数据包的数量分别表示为$\Psi_{1,t} = \left\{\psi_{1,t}^1, \psi_{1,t}^2, \cdots, \psi_{1,t}^i, \cdots, \psi_{1,t}^N\right\}$和$\Psi_{2,t} = \left\{\psi_{2,t}^1, \psi_{2,t}^2, \cdots, \psi_{2,t}^i, \cdots, \psi_{2,t}^N\right\}$。其中，$\psi_{1,t}^i$表示在时隙$t$时波位$i$的实时数据包数量，$\psi_{2,t}^i$表示在时隙$t$时波位$i$的非实时数据包数量。

如前文所述，不同的业务类型对应不同的QoS需求。在实时的语音和视频会话等通信场景中，用户对时延较敏感；而在流媒体、数据传输等通信场景中，用户则更关注数据传输的速率。因此，亟须摆脱现有低轨卫星深度强化学习资源调度算法单一优化目标的局限，综合考虑实时业务的低时延需求、非实时业务的高吞吐量需求以及两者的公平性均衡。因此，本文以最小化实时业务的数据包平均排队时延[14,28]、最大化非实时业务的数据包吞吐量[28]、最大化服务波位的业务满意度为目标[13]，建立多目标的优化函数，具体

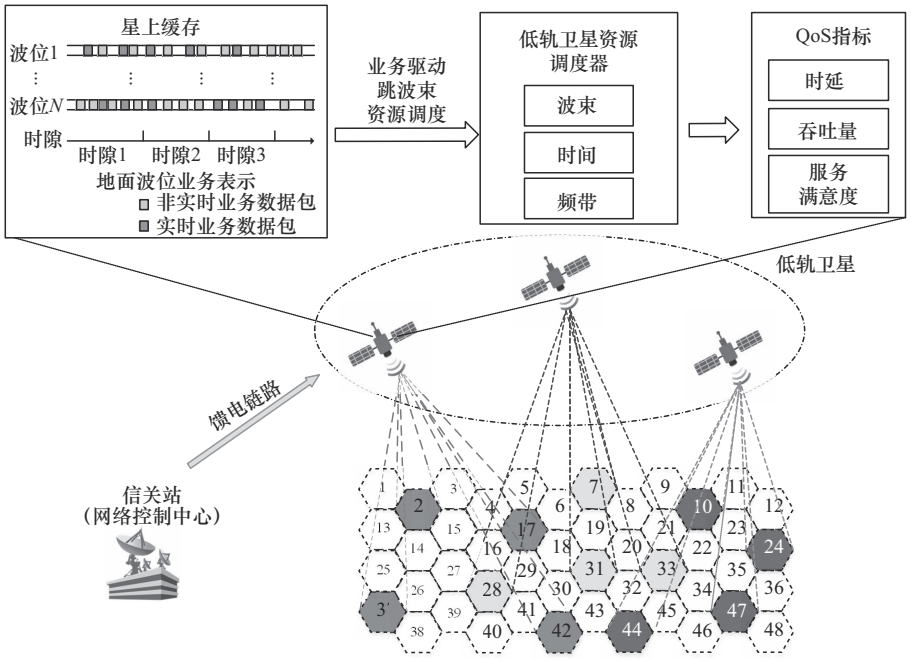



图8　业务QoS驱动的低轨卫星跳波束多维资源调度

计算方式为

$$\text{opt.}\quad p_1 = \min \sum_{t_j \in T} \sum_{i \in \mathcal{N}} \sum_{\text{pac}=1}^{S^i_{\psi^i_{1,t_j}}} \frac{1}{S^i_{\psi^i_{1,t_j}}} \left[ t^i_j(\text{pac}) - t^i_{\text{begin}}(\text{pac}) \right] \tag{9}$$

$$p_2 = \max \sum_{t_j \in T} \sum_{i \in \mathcal{N}} (\psi^i_{2,t_{j-1}} + \lambda^i_{2,t_j} - \psi^i_{2,t_j}) \tag{10}$$

$$p_3 = \max \sum_{t_j \in T} \sum_{i \in \mathcal{N}} \left( \frac{\sum_{t=1}^{t_j} S^i_{\psi^i_t}}{\sum_{t=1}^{t_j} \Lambda^i_t} \right) \tag{11}$$

$$p = -\omega_1 p_1' + \omega_2 p_2' + \omega_3 p_3', \omega_o \in [0,1], \sum_{o=1}^{3} \omega_o = 1 \tag{12}$$

$$\text{s.t.}\quad \sum_{i \in \mathcal{N}} x^i_{t_j} \leqslant K,\ x^i_{t_j} \in \{0,1\}, \forall i \in \mathcal{N}, t_j \in T \tag{13}$$

$$\sum_{i \in \mathcal{N}} P^i_{t_j} \leqslant P_{\text{tot}}, \forall t_j \in T \tag{14}$$

$$P^i_{t_j} \leqslant P_b, \forall i \in \mathcal{N}, t_j \in T \tag{15}$$

$$t^i_j(\text{pac}) - t^i_{\text{begin}}(\text{pac}) \leqslant T_{\text{th}}, \forall i \in \mathcal{N} \tag{16}$$

通过式(9)计算最小化实时数据的平均排队时延，$T$为从起始时隙到当前时隙$t_j$的集合，$\mathcal{N}$为卫星所服务的地面波位集合，$S^i_{\psi^i_{1,t_j}}$为在时隙$t_j$向波位$i$传输的实时数据包数量，其计算方式为

$$S^i_{\psi^i_{1,t_j}} = \min\left( \frac{C_i t_{\text{slot}}}{\text{Size}_{\text{pac}}}, \sum_{t \in t_j} \psi^i_{1,t} \right) \tag{17}$$

其中，$C_i$为通过式(6)计算得到的波位$i$的通信容量，$t_{\text{slot}}$为时隙长度，$\text{Size}_{\text{pac}}$为数据包大小，$\psi^i_{1,t}$为时隙$t$时波位$i$中的实时数据包数量。显然，$S^i_{\psi^i_{1,t_j}}$不能超过该波位队列中存在的实时数据包总和，因此取两者中的较小值。$t^i_j(\text{pac})$为实时数据包pac离开队列的时隙，$t^i_{\text{begin}}(\text{pac})$为实时数据包pac初始到达队列的时隙。

通过式(10)计算最大化非实时数据的吞吐量。$\psi^i_{2,t_{j-1}}$为队列开始时隙到时隙$t_{j-1}$存在于波位$i$的非实时数据包数量总和，$\lambda^i_{2,t_j}$为时隙$t_j$时波位$i$的非实时数据包到达量，$\psi^i_{2,t_j}$为队列开始时隙到时隙$t_j$存在于波位$i$的非实时数据包数量总和。

通过式(11)计算最大化服务波位的满意度[13]，定义为系统实际服务的业务量与用户的业务需求量的比值。$S^i_{\psi^i_t}$为从起始时隙到当前时隙$t_j$向波位$i$传输的数据包总量，$\sum_{t=1}^{t_j} \Lambda^i_t$为从起始时隙到当前时隙$t_j$波位$i$的总业务需求。

通过式(12)计算本文的目标函数，利用最大化目标函数实现降低实时数据平均时延、提高非实时数据平均吞吐量和服务波位业务满意度的目的。首先，对$p_1$、$p_2$、$p_3$目标函数下的$Q$值进行L2范数归一化处理，然后由其加权和构成最终的目标函数值。区别于传统的固定多目标加权参数方案[13]，本文提出的多目标加权参数$\omega_o$会根据不同业务的QoS需求，在训练过程中动态变化，逐渐收敛到使得奖励值最大的情况。

在约束条件中，通过式(13)计算当前时隙被服务的波位数量不能超过卫星的最大波束数量$K$，通过式(14)计算当前时隙所有波位的功率之和不能超过卫星的星上可用总功率$P_{\text{tot}}$。在式(15)中，分配给某波位的功率$P_{t_j}^i$小于单波束的最大功率$P_b$。此外，根据当前服务波位的总业务量需求和业务数据包对应的平均时延，定义波位$i$在时隙$t_j$的功率权重$\text{weight}_i(t_j)$为两者的乘积，从而得到波位$i$分配的功率为

$$P_{t_j}^i=\frac{\text{weight}_i(t_j)P_{\text{tot}}}{\sum_{i=1}^{K}\text{weight}_i(t_j)} \tag{18}$$

其中，分母表示当前所有服务波位的功率权重之和。式(18)计算结果表明，当前波位需要传输的业务总量（数据包总个数）越多或数据包排队时延越长，功率分配的权重就会越大，从而提高了波束当前时刻的信道速率，使得单位时间内能够传输更多的数据包，最终满足业务量需求、降低排队时延。最后，由于星载资源有限，当数据包的驻留时间超过时延阈值$T_{\text{th}}$时，该数据包将被丢弃[14,28]。

虽然在约束关系中并未显式地表示干扰对目标函数的影响，但是根据1.2节同频波束干扰与波束容量$C$关系的模型可知，同频工作的波束之间的距离、相对夹角等对干扰强度的影响，并由SINR得到与波束容量$C$的关系。从式(9)~式(11)、式(17)可以得到，速率、时延、满意度等优化目标都与一定时间内信道实际可传输数据包的数量直接相关，也就是与波束容量$C$相关，因此需要考虑波束间同频干扰对目标函数的影响。

此外，还需要关注跳波束对业务连续性的影响，主要是由波束跳跃的间断性造成的。传统方法通过设置波束重访间隔时间[24]来降低业务数据包的排队调度时延，从而保障业务的连续性。然而，这种方法不够灵活，也不区分业务类型，对资源的利用率也有待提高。因此，本文针对实时业务（时延敏感型业务）建立了最小化排队调度时延的目标，确保训练过程中当波束跳跃离开某个波位后，资源调度智能体能够在满足实时业务的QoS时延要求条件下，再次执行服务该波位的重访服务动作，以避免波束间断时间过长影响业务连续性。

综上，本文构建了一种最优化问题模型，以最小化实时数据时延、最大化非实时数据吞吐量和最大化服务小区的业务满意度为优化目标，根据星上资源的限制、不同业务类型对吞吐量、时延不同的QoS需求，优化问题模型的主要输出（即资源调度智能体的执行动作），具体如下。1) 对星上缓存中的实时和非实时业务数据包进行时间排队调度，同时确定当前跳波束时隙需要服务的波位，并调度波束进行按需覆盖和服务；2) 根据当前服务波位的总业务量和数据包对应的平均排队时延，合理分配各波束的功率资源；3) 在上述输出动作的基础上，根据各波位的总业务量完成各波位的波束驻留时间分配，从而构成波束跳跃图案。

## 2　基于多智能体的深度强化学习资源调度算法

### 2.1　资源调度总体架构

本文提出低轨星座跳波束资源调度总体架构，如图9所示，并提出一种可行的实施方案。首先，利用地面信关站的网络控制功能，收集业务流量、业务类型、系统无线资源、星历等数据，作为训练和验证数据；然后，借助信关站较强的计算和处理能力，对资源调度模型进行离线训练和验证；最后，通过测控链路将训练后的模型参数上传到低轨卫星。星上不进行模型训练，仅执行推理操作，根据不同的业务QoS需求，完成星上无线空口资源与跳波束图案的实时决策调度。由于本文提出的资源调度智能体模型具有较强的泛化性，当卫星任务、星座构型、覆盖区域等系统关键参数发生重大变化时，信关站会监测到当前的资源调度策略不满足用户的QoS需求或系统性能下降。此时，信关站根据当前的系统数据，对模型进行重新训练或微调，并重复参数上注过程。此外，

低轨卫星之间采用接力传输的方式，当发生星间切换时，根据多目标的选星策略选取合适的卫星接入并提供服务。

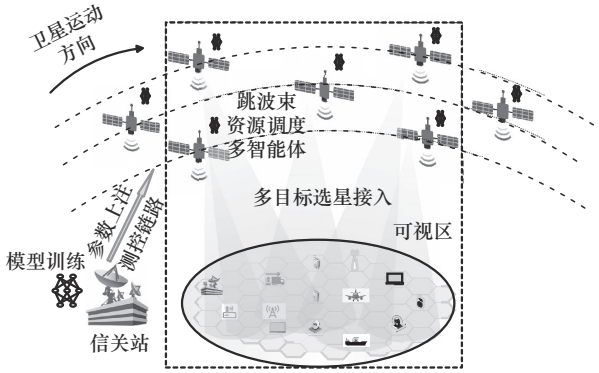



图9　低轨星座跳波束资源调度总体架构

## 2.2　深度强化学习算法设计

在资源调度总体架构的基础上，借鉴MoE的核心设计理念，本文提出多智能体深度强化学习模型结构，如图10所示。该模型以业务类型、系统资源、同频干扰为输入，设计了3个不同优化目标的智能体模型，分别对应最小化实时数据时延、最大化非实时数据吞吐量和最大化各波位服务满意度，并通过底层DQN方法支撑资源调度算法。


图10　多智能体模型结构

进一步地，将低轨卫星跳波束资源调度器建模为基于MoE架构的多智能体，各个波位的业务状态建模为环境。不同目标函数对应不同的奖励，分别根据实时数据的时延、非实时数据的吞吐量和各波位的满意度来决定奖励的大小。相应地，每个目标函数均有自身独特的$Q$网络1~$N$、目标网络1~$N$和经验池1~$M$。在归一化$Q$值后，选择加权$Q$值最大的波束调度作为动作。将低轨卫星系统建模为离散时间事件系统，并且该系统受业务驱动，因此上述优化问题可建模为马尔可夫决策过程。在训练周期内，智能体通过与环境的不断交互，学习最优策略，在实时数据的时延、非实时数据吞吐量和波位满意度之间找到平衡，从而逼近全局最优解，提高系统性能，算法结构如图11所示。

### 2.2.1　状态设计

假设实时数据包与非实时数据包大小相同，本文通过卫星服务区域内各波位的数据包数量来描述其业务请求。由于目标函数包括时延、吞吐量和服务波位的满意度，因此时隙$t_j$时的环境状态矩阵$\boldsymbol{s}_{t_j}$由数据包数量矩阵（包括实时数据包数量和非实时数据包数量）和服务波位的满意度矩阵组成，如式(19)所示

$$\boldsymbol{s}_{t_j}=\left\{\boldsymbol{\Psi},\boldsymbol{\eta}_{t_j}\right\} \tag{19}$$

其中，卫星缓存区中数据包数量矩阵$\boldsymbol{\Psi}$为

$$\boldsymbol{\Psi}=\begin{bmatrix}\boldsymbol{\Psi}_{\delta+1}^{1} & \boldsymbol{\Psi}_{\delta+2}^{1} & \cdots & \boldsymbol{\Psi}_{t_j}^{1}\\ \boldsymbol{\Psi}_{\delta+1}^{2} & \boldsymbol{\Psi}_{\delta+2}^{2} & \cdots & \boldsymbol{\Psi}_{t_j}^{2}\\ \vdots & \vdots & & \vdots\\ \boldsymbol{\Psi}_{\delta+1}^{N} & \boldsymbol{\Psi}_{\delta+2}^{N} & \cdots & \boldsymbol{\Psi}_{t_j}^{N}\end{bmatrix} \tag{20}$$

$$\boldsymbol{\Psi}_{t_j}^{N}=\left[\psi_{1,t_j}^{N},\psi_{2,t_j}^{N}\right]^{\mathrm{T}} \tag{21}$$

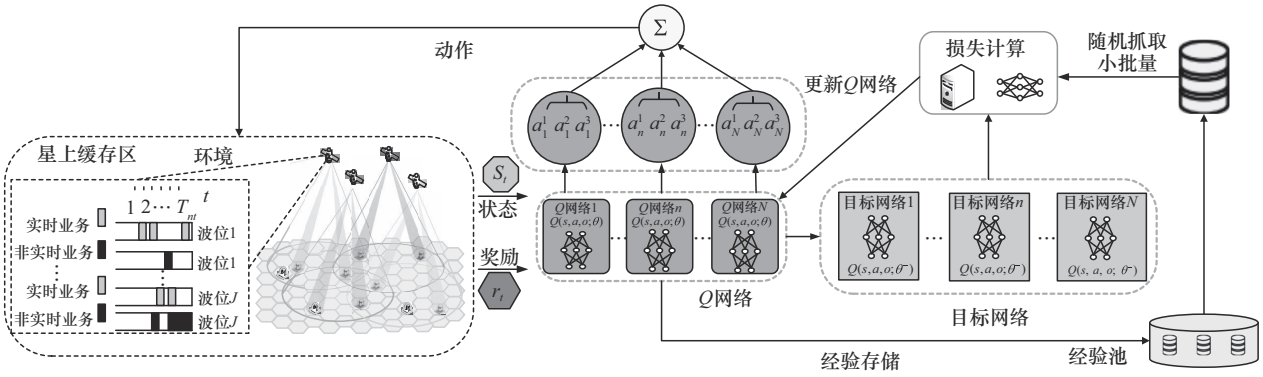



图11　基于多目标优化的深度强化学习算法结构

由于星载资源有限，矩阵$\boldsymbol{\Psi}$的列数小于时延阈值$T_{\text{th}}$，因此，$\delta$的取值为

$$\delta=\begin{cases}0, & t_j\leqslant T_{\text{th}}\\ t_j-T_{\text{th}}, & t_j>T_{\text{th}}\end{cases} \tag{22}$$

$\boldsymbol{\eta}_{t_j}$为业务满意度矩阵，表示为

$$\boldsymbol{\eta}_{t_j}=[\ \eta_{t_j}^1,\eta_{t_j}^2,\cdots,\eta_{t_j}^i,\cdots,\eta_{t_j}^N\ ]^{\mathrm{T}} \tag{23}$$

$$\eta_{t_j}^i=\frac{\sum_{t=1}^{t_j}S_{\psi_t^i}^i}{\sum_{t=1}^{t_j}A_t^i},\quad i\in\mathcal{N} \tag{24}$$

### 2.2.2　动作设计

在本文提出的深度强化学习算法中，每个目标函数对应一个DQN，每个网络需要分别存储和更新其状态、动作以及奖励值。如图12所示，该智能体由3个独立的DQN组成，第一个DQN负责最小化实时业务的数据包时延，为计算方便，对该网络输出的结果取负值，结果越大表示性能越差；第二个DQN负责最大化非实时业务的数据包吞吐量；第三个DQN负责最大化服务波位的满意度。每个DQN生成一组$Q$值，通过L2范数归一化后，智能体根据式(24)中的线性标量化函数（LS, linear scalar）[29]确定最大$Q$值输出对应的动作[30]，其中，$p_o$为单个目标函数。

$$\mathrm{LS}(x)=\sum_{o=1}^{3}\omega_o p_o\ ,\ \omega_o\in[0,1]\ ,\ \sum_{o=1}^{3}\omega_o=1 \tag{25}$$

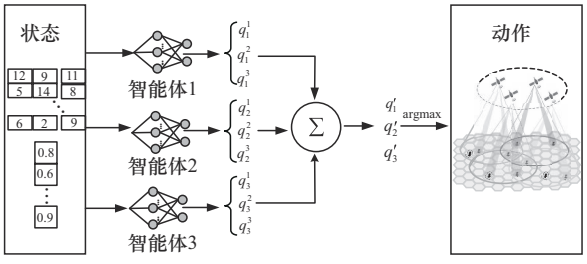



图12　基于多目标DQN的动作选择

对于动作的具体解释如下。将环境状态$s_t$输入$Q$网络后，输出多组$Q$值，智能体以最大$Q$值对应的波位输出为动作。为了避免上述动作选择策略陷入局部最优解，在动作选择中引入贪婪算法。通过贪婪算法，智能体以$\varepsilon$概率随机选择动作；否则，以1−$\varepsilon$概率选择最大$Q$值对应的波位作为输出动作，表示为

$$\boldsymbol{A}_{t_j}=\{\ x_{t_j}^1,\cdots,x_{t_j}^i,\cdots,x_{t_j}^N\mid\sum_{i=1}^{N}x_{t_j}^i=K,x_{t_j}^i\in\{\ 0,1\ \}\ ,\forall i\in\mathcal{N}\ \} \tag{26}$$

其中，$x_{t_j}^i=1$表示在$t_j$时隙智能体点亮服务序号为$i$的地面波位。

### 2.2.3　奖励设计

本文目标函数的优化目标为最小化实时业务的数据包平均排队时延，最大化非实时数据的吞吐量和服务波位的满意度，因此其奖励值共有3个指标。当智能体跳波束资源分配不合理导致缓冲队列中数据包的平均排队时延较大时，该动作的奖励值应被设置为较小的数值。如果智能体跳波束资源分配合理，使得数据包的平均排队时延较小，此时动作的奖励值应被设置为较大的数值。因此，最小化实时业务的数据包平均排队时延的奖励值可表示为

$$r_{1,t_j}=-\sum_{i\in\mathcal{N}}\sum_{\mathrm{pac}=1}^{S_{\psi_{1,t_j}^i}^i}\frac{1}{S_{\psi_{1,t_j}^i}^i}\left[t_j^i(\mathrm{pac})-t_{\mathrm{begin}}^i(\mathrm{pac})\right] \tag{27}$$

其次，在时隙$t_j$时，非实时数据的吞吐量越大，其智能体动作所对应的奖励值应设置得越大；否则，奖励值应设置得越小。因此，最大化非实时业务的吞吐量的奖励值表示为

$$r_{2,t_j}=\sum_{i\in\mathcal{N}}(\psi_{2,t_{j-1}}^i+\lambda_{2,t_j}^i-\psi_{2,t_j}^i) \tag{28}$$

最后，在时隙$t_j$时，服务波位的满意度总和越小，其智能体动作所对应的奖励值应设置得越小；否则，奖励值应设置得越大。因此，最大化服务波位满意度的奖励值表示为

$$r_{3,t_j}=\sum_{i\in\mathcal{N}}(\eta_{t_j}^i) \tag{29}$$

综上所述，智能体在时隙$t_j$的奖励最终取值为

$$r_{t_j}=\chi_1 r_{1,t_j}+\chi_2 r_{2,t_j}+\chi_3 r_{3,t_j} \tag{30}$$

其中，$\chi_i\in[0,1],\sum_{i=1}^{3}\chi_i=1$，权重参数$\chi_i$可根据业务类型、不同的系统场景进行设置。为了兼顾公平性与可行性，$\chi_i$在本文中取值为$[\chi_1,\chi_2,\chi_3]=\left[\frac{1}{3},\frac{1}{3},\frac{1}{3}\right]$。

值得注意的是，如前文所述，波束间的同频干扰会降低信道容量，从而影响优化目标函数。因此，当智能体执行的动作中包含相邻波位时，产生的共信道干扰会影响信道容量，导致波束的通信能

力下降，进而增加实时数据包的平均排队时延、降低非实时数据包的吞吐量以及服务波位的满意度，从而导致奖励值下降。在训练过程中，智能体会自我更新学习，逐渐调整策略，以增大奖励值为目标，从而避免执行相邻波位同时被跳波束服务的动作。

#### 2.2.4　$Q$网络设计

实时数据的平均时延、非实时数据的吞吐量的$Q$网络结构如图13所示，业务满意度的$Q$网络结构如图14所示。图13中的网络将卷积网络与神经网络相结合，即通过卷积网络提取状态矩阵的特征，然后通过神经网络拟合输入状态到输出状态的非线性映射。由于业务满意度的输入矩阵较简单，图14中的网络直接通过神经网络拟合输入状态到输出状态的非线性映射。

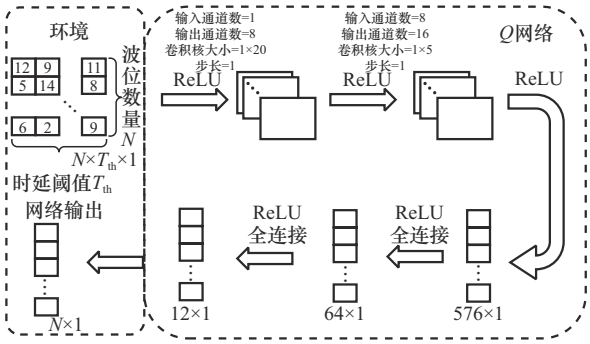



图13　平均时延、吞吐量的$Q$网络结构

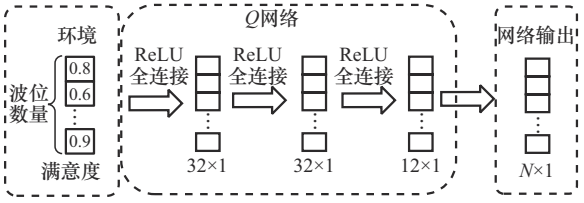



图14　业务满意度的$Q$网络结构

### 2.3　训练流程

为提高算法的稳定性，本文引入经验池和目标网络[31-32]。在智能体与环境交互过程中，生成多个四元组$(s_t,a_t,r_t,s_{t+1})$，并通过经验池存储和回放这些四元组，以打破神经网络训练过程中数据之间的相关性，即神经网络对独立同分布的数据进行学习。此外，本文引入目标网络以更新主网络，目标网络与主网络结构相同，但参数不同，目标网络的参数是主网络参数的慢速复制，即目标网络参数每$G$步从主网络复制更新，以降低目标$Q$值和当前$Q$值的相关性。定义$Q(s,a;\theta)$表示当前网络输出，$Q(s,a;\theta^-)$表示目标网络输出，则损失函数[33]为

$$L(\theta)=E\left[(y_t-Q(s,a;\theta))^2\right] \tag{31}$$

其中，$\theta$表示主网络的参数，$\theta^-$表示目标网络的参数，标签值$y_t$为

$$y_t=\begin{cases} r_t, & a_{t+1}=\varnothing \\ r_t+\gamma\max\limits_{a\in A}Q(s_{t+1},a_{t+1};\theta^-), & \text{其他} \end{cases} \tag{32}$$

其中，$\gamma$为未来奖励的折扣因子。

本文算法主要由场景环境的初始化和$Q$网络训练决策两部分组成。其中，场景环境的初始化主要包括低轨卫星跳波束通信场景、$Q$网络相关参数等；$Q$网络训练决策主要包括在智能体与卫星的交互训练中的学习最优策略和逼近全局最优解，具体步骤如算法1所示。

**算法1**　基于多智能体的深度强化学习低轨跳波束资源调度算法

1) 初始化

2) 初始化低轨卫星跳波束通信场景

3) 初始化经验池容量、卫星数据包缓冲队列

4) 构建主网络并初始化主网络$Q(s,a;\theta)$的权重参数

5) 构建目标网络并初始化目标网络$Q(s,a;\theta^-)$的权重参数

6) 设置LOOPS为训练周期，LOOPS_TIME_SLOT为每周期的时隙数

7) $Q$网络训练决策

8) for loop = 1, ⋯ , LOOPS do

9) 初始化本周期$Q$网络的状态$s$、动作$a$和奖励$r$

10) for $t$= 1, ⋯ , LOOPS_TIME_SLOT do

11) 在$t$时隙，根据卫星数据包缓冲队列中的$\psi_{1,t}$和$\psi_{2,t}$更新状态，智能体按照各自的$Q$网络遵从贪婪算法输出动作。然后根据动作输出一组$Q$值，并将$Q$值进行归一化，然后按照式(24)确定最大$Q$值。最后执行加权$Q$值最大的波束调度为动作$a_t$

12) 智能体执行动作$a_t$，此时状态从$s_t$变为$s_{t+1}$，计算该动作的奖励值$r_t$

13) 将四元组$(s_t,a_t,r_t,s_{t+1})$存入各自的DQN经验池，若经验池已满，则抛弃最早的四元组

14) 从各自的DQN经验池中随机选取batch大

小的四元组样本，根据式(30)和式(31)计算损失函数与标签值

15) 利用Adam算法更新各自主网络$Q(s,a;\theta)$的权重参数

16) 每$G$步从各自主网络复制更新目标网络$Q(s,a;\theta^{-})$的权重参数

17) end for

18) end for

## 3　仿真与分析

### 3.1　仿真参数设置

根据上文所述的算法对低轨卫星跳波束资源分配进行仿真与分析。首先，对场景参数进行归纳说明，如表1所示。然后，利用Python搭建深度强化学习模型，其相关参数如表2所示。训练数据集主要由两方面的公开来源组成：1) 根据3GPP协议[26-27]规定的业务模型生成的业务数据；2) 卫星轨道参数和无线资源配置采用实际星座系统的部分公开数据。最后，将本文算法MA-DRL分别与多波束多色复用算法Multi-Beam、基于遗传算法的跳波束资源分配算法GA、基于吞吐量最大化的跳波束深度强化学习算法Throughput-DRL和基于时延最小化的跳波束深度强化学习算法Latency-DRL进行对比。为了保证对比方案的公平性和有效性，Throughput-DRL和Latency-DRL分别以最大化吞吐量和最小化系统时延为目标函数，且均采用与本文算法相同的数据集进行训练。

### 3.2　仿真结果与分析

地面波位的业务需求随空间和时间的分布分别如图6和图7所示。相关仿真参数如表1和表2所示，从实时数据的平均时延、非实时数据的吞吐量和服务波位的满意度3个方面对本文算法进行性能评估，评估结果如图15所示。通过分析图15可知，在训练周期内，实时数据的平均排队时延、非实时数据的吞吐量和服务波位的满意度均已收敛至稳定值。

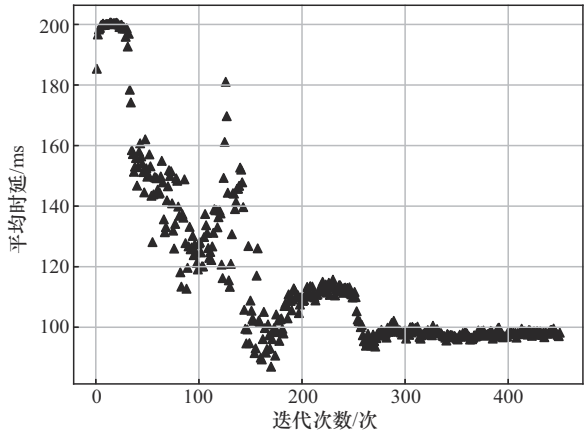
基于上述低轨卫星星座场景，本文对训练好的资源调度模型进行时空二维验证，以检测其泛化性。在时间维度上，选取9:00—14:00时段进行仿真，业务量时变形成波峰波谷的时间不均性；同时，多颗卫星在仿真持续期间交替完成通信服务。在空间维度上，通过离散度来定义各业务分布的空间不均匀性。综合以上因素，验证本文所提智能模型的泛化性，并评估其在平均时延、吞吐量、业务满意度等性能指标上的表现，时变环境下算法性能对比如图16所示。

**表1　　星座和无线空口仿真参数**

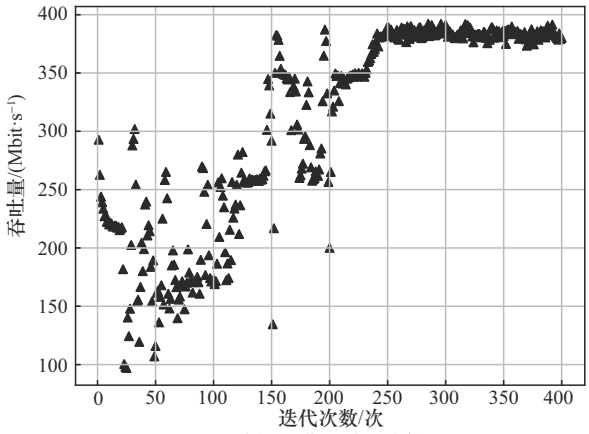
| 参数 | 数值 |
|---|---|
| 轨道数/个 | 36 |
| 单轨卫星数/个 | 20 |
| 轨道高度/km | 570 |
| 轨道倾角 | 70° |
| 卫星总数/个 | 720 |
| 波位数/个 | 12 |
| 波束数/个 | 4 |
| 载波频率/GHz | 20 |
| 波束带宽/MHz | 200 |
| 星上总功率/W | 120 |
| 最大波束功率/W | 60 |
| 卫星发射天线增益/dB | 40 |
| 用户接收天线增益/dB | 50 |
| 跳波束时隙长度/ms | 10 |
| 排队时延阈值/ms | 400 |
| 数据包大小/kbit | 10 |
| 噪声温度/K | 300 |
| 玻尔兹曼常数/dB | −228.6 |

**表2　　$Q$网络训练参数**

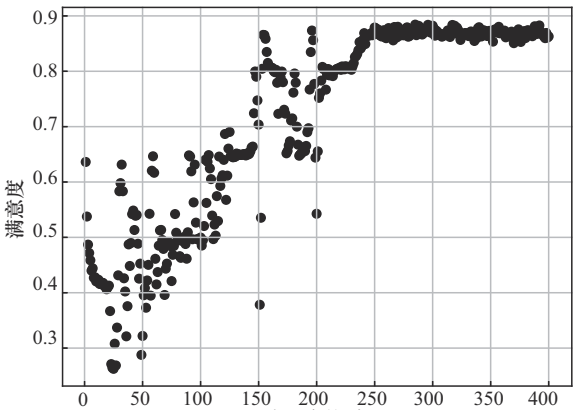
| 参数 | 数值 |
|---|---|
| 训练周期 | 450 |
| 每周期时隙数 | 1 000 |
| 主网络训练步长 | 4 |
| 目标网络更新步长 | 100 |
| 经验池容量 | 3 000 |
| batch训练样本数量 | 8 |
| 学习率 | $10^{-5}$ |
| 折扣因子 | 0.9 |
| 初始探索概率 | 0.5 |
| 终止探索概率 | 0.01 |
| 激活函数 | ReLU |



(a) 训练周期时延收敛

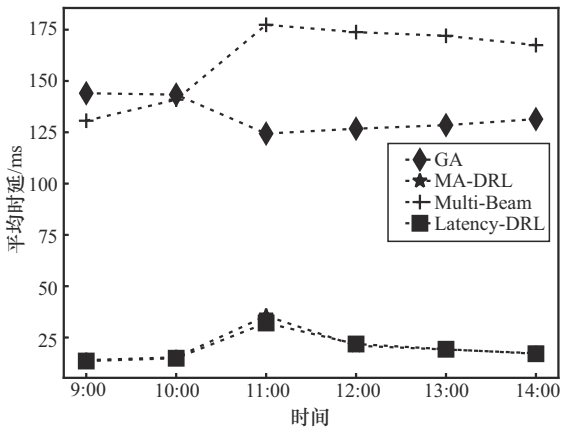


(b) 训练周期吞吐量收敛

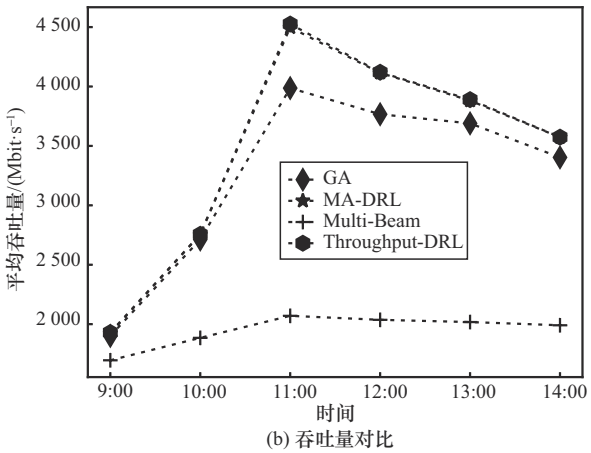


(c) 训练周期满意度收敛

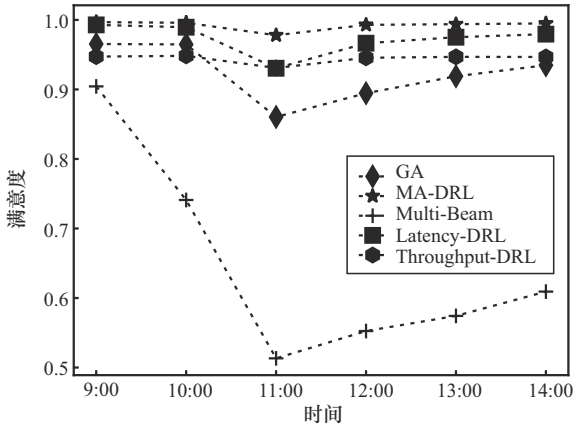
图15　训练周期收敛情况



(a) 时延对比



(b) 吞吐量对比



(c) 满意度对比

图16　时变环境下算法性能对比

通过分析图7中的时间业务量模型可得，在仿真时间内，波位业务量经历了先上升后下降的过程，且在11:00时业务量达到峰值，对资源调度算法的性能要求也达到最高。因此，基于峰值业务量对上述仿真结果进行分析。分析图16(a)可知，对于实时业务的数据包平均时延，本文算法与Latency-DRL算法性能相当，并且平均时延大幅低于GA算法和Multi-Beam算法，分别降低了71.42%和79.96%。分析图16(b)可知，对于非实时业务的数据包平均吞吐量，本文算法与Throughput-DRL算法性能相当，相较于GA算法和Multi-Beam算法，吞吐量分别提高了1.82%和117.27%。分析图16(c)可知，对于服务波位的业务满意度，本文算法的服务波位业务满意度为0.98，相较于Throughput-DRL

算法、Latency-DRL算法、GA算法和Multi-Beam算法，分别提高了5.07%、5.12%、1.64%和90.52%。综上所述，在业务需求动态时变方面，相较于GA算法和Multi-Beam算法，本文算法在3个目标函数指标上均有大幅提升；相较于单一目标的深度强化学习资源调度算法，本文提出的基于MoE架构的多智能体资源调度算法能够根据不同业务的QoS需求，调用不同的智能体完成相应的优化目标，既能满足实时业务低时延需求，又能满足非实时业务高吞吐量的需求。相比于单一目标的深度强化学习资源调度算法，本文算法可以更灵活地适应多样化业务混合传输的实际系统场景。综合以上，针对随时间变化的场景，本文提出的资源调度智能模型展现了较好的泛化性。

为进一步分析本文算法的泛化性，在波位业务需求空间分布不均的场景下开展仿真。根据前文定义的业务离散系数$\zeta$，对上述算法在不同离散系数下的性能进行进一步验证，不同离散系数下各算法性能对比如图17所示。

离散系数越大，业务需求的空间不均性越明显。对于实时业务的数据包平均时延，由于业务需求空间分布的不均性对波束调度时延的影响相比于业务时间动态变化更明显，因此本文算法性能略低于Latency-DRL算法，但性能降低的幅度较小。本文算法的性能显著优于GA算法和Multi-Beam算法，在离散系数为0.5时，其时延分别降低了77.78%和8.12%。对于非实时业务的数据包平均吞吐量，本文算法的系统吞吐量略低于Throughput-DRL算法，但整体性能相当。相较于GA算法和Multi-Beam算法，在离散系数为0.5时，吞吐量分别提高了1.53%和60.95%。对于服务波位的业务满意度，本文算法的业务满意度为0.97，相较于Throughput-DRL算法、Latency-DRL算法、GA算法和Multi-Beam算法，分别提高了4.55%、8.12%、1.52%和59.87%。综上所述，在应对业务需求空间分布不均性方面，相较于GA算法和Multi-Beam算法，本文算法在3个目标函数指标上均有大幅提升。与前文分析结果类似，相较于传统单一优化目标的深度强化学习资源调度算法，基于MoE架构的模型能够灵活调用不同任务的智能体，从而适应多样化的业务需求与时空分布不均性，智能地分配多维资源，满足不同业务类型的QoS需求。综合以上，针对业务空间分布变化的场景，本文提出的资源调度智能模型展现了较好的泛化性。

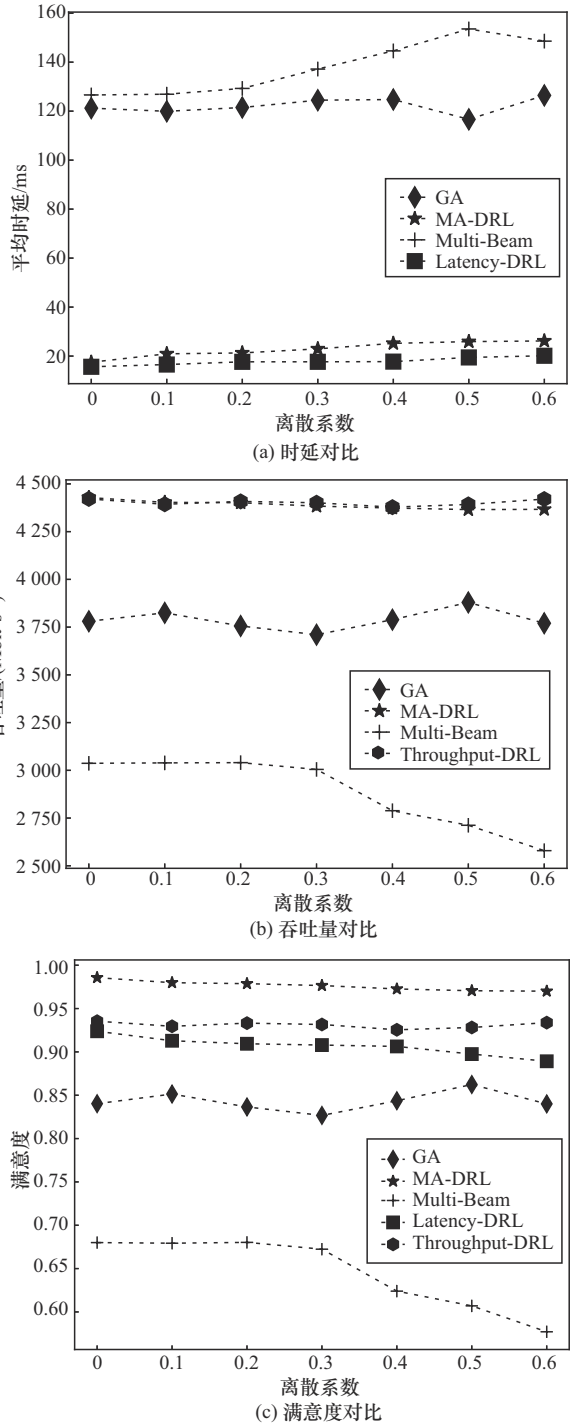



图17　不同离散系数下各算法性能对比

进一步，对算法复杂度进行分析。GA算法的复杂度与种群规模[13]（115个）和迭代次数（300次）相关，可以表示为

$$O(M_G N_G^2) = 300 \times 115^2 \approx 3.97 \times 10^6 \quad (33)$$

MA-DRL算法的复杂度主要与网络中的卷积层特征图维度、输入输出通道数、全连接层的输入输出维度相关[13]，可以表示为

$$O\Big(\sum_{m=1}^{L} K_{m_1}\times K_{m_2}\times C_{m_{\text{in}}}\times H_{m_{\text{out}}}\times W_{m_{\text{out}}}\times C_{m_{\text{out}}}+2\times\sum_{m=1}^{L'} X_m Y_m\Big)\approx 3.90\times 10^6 \tag{34}$$

其中，$L$为卷积层层数，$L'$为全连接层层数，$K_{m_1}\times K_{m_2}$为网络中卷积层$m$的卷积核大小，$C_{m_{\text{in}}}$为卷积层$m$的输入通道数，$H_{m_{\text{out}}}\times W_{m_{\text{out}}}$为卷积层$m$的特征图大小，$C_{m_{\text{out}}}$为卷积层$m$的输出通道数，$X_m$为全连接层$m$的输入尺寸，$Y_m$为全连接层$m$的输出尺寸。

类似地，根据非MoE架构下单智能体深度强化学习方法的网络卷积层、全连接层参数以及各层的输入输出通道数，得到对比方案Throughput-DRL和Latency-DRL算法的复杂度[14]为

$$O\Big(\sum_{m=1}^{L} K_{m_1}\times K_{m_2}\times C_{m_{\text{in}}}\times H_{m_{\text{out}}}\times W_{m_{\text{out}}}\times C_{m_{\text{out}}}+\sum_{m=1}^{L'} X_m Y_m\Big)\approx 2.1\times 10^6 \tag{35}$$

虽然本文算法在复杂度上高于单一优化目标的深度强化学习算法，但该多目标优化算法通过基于MoE架构的多智能体，能够同时优化时延、吞吐量、业务满意度等多个目标，从而满足实时和非实时业务的QoS需求，并能灵活适应多样化的业务混合传输场景。从式(34)可以看出，本文算法的复杂度与低轨卫星系统中经典的遗传算法相近，但其性能大幅优于遗传算法，表明其算法复杂度仍在可接受的范围内。

最后，关于模型参数上注的开销。经过估算，本文提出的智能体模型训练后的参数约占15 MB的上注开销，考虑现有卫星的测控链路和星载处理机的处理能力，本文方法的上注开销在可接受范围内。此外，即使系统关键参数发生重大变化需要重新更新模型并再次上注时，通常采用最小增量法，仅把变化的参数上注，不需要对所有参数进行更新，从而进一步减小了更新的上注开销。

综上所述，本文算法不仅可以满足不同业务的QoS需求，且具有较强的泛化性，并在平衡算法复杂度的基础上，能够适应星上系统有限的处理能力。

## 4　结束语

本文以低轨星座跳波束多维资源调度面临的业务需求多样化、时空不均动态变化挑战为研究出发点，旨在构建一种可解释、轻量化、强泛化性的低轨跳波束资源智能调度策略，提出了基于多智能体深度强化学习的低轨跳波束资源调度方法。首先，综合考虑低轨卫星的高动态性、业务需求的多样性与时空不均性、星上处理能力的有限性、全频复用导致的同频干扰问题，建立了低轨跳波束资源调度系统模型。其次，基于低轨星座的多重覆盖和密集波束特性，通过多目标的选星优化接入，完成星间切换与接续传输，构建低轨卫星与服务区域的时变映射关系。最后，根据不同业务的QoS需求，以最小化实时数据时延、最大化非实时数据吞吐量和最大化服务小区的业务满意度为优化目标，提出基于MoE的多智能体模型深度强化学习算法。仿真结果表明，相较于传统算法，本文算法能够有效应对业务流量的时空不均与动态变化，体现出较强的泛化性，同时平衡了算法的复杂度，满足不同业务对时延、吞吐量等性能的需求。未来研究将进一步细化波束内资源调度的颗粒度，提升通信协议的适配性；同时，进一步优化多智能体资源调度方法与奖励权重参数的动态选择，降低算法的复杂度，提升工程可实现性。

## 参考文献：


[1] SHENG M, ZHOU D, BAI W G, et al. 6G service coverage with mega satellite constellations[J]. China Communications, 2022, 19(1): 64-76.

[2] 李聪, 何雯, 王一帆. 关于卫星跳波束系统的几点思考[J]. 空间电子技术, 2021, 18(1): 8-13.
LI C, HE W, WANG Y F. A study of beam-hopping technology in satellite systems[J]. Space Electronic Technology, 2021, 18(1): 8-13.

[3] ANZALCHI J, COUCHMAN A, GABELLINI P, et al. Beam hopping in multi-beam broadband satellite systems: system simulation and performance comparison with non-hopped systems[C]//Proceedings of the 2010 5th Advanced Satellite Multimedia Systems Conference and the 11th Signal Processing for Space Communications Workshop. Piscataway: IEEE Press, 2010: 248-255.

[4] ANGELETTI P, PRIM D F, RINALDO R. Beam hopping in multi-beam broadband satellite systems: system performance and payload architecture analysis[C]//Proceedings of the 24th AIAA International Communications Satellite Systems Conference. Reston: AIAA, 2006: 5376.

[5] ALEGRE R, ALAGHA N, VÁZQUEZ-CASTRO M Á. Heuristic algorithms for flexible resource allocation in beam hopping multi-beam sat-

ellite systems[C]//Proceedings of the 29th AIAA International Communications Satellite Systems Conference (ICSSC-2011). Reston: AIAA, 2011: 8001.

[6] ZHANG T, ZHANG L X, SHI D Y. Resource allocation in beam hopping communication system[C]//Proceedings of the 2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC). Piscataway: IEEE Press, 2018: 1-5.

[7] HAN H, ZHENG X Q, HUANG Q F, et al. QoS-equilibrium slot allocation for beam hopping in broadband satellite communication systems[J]. Wireless Networks, 2015, 21(8): 2617-2630.

[8] TANG J Y, BIAN D M, LI G X, et al. Optimization method of dynamic beam position for LEO beam-hopping satellite communication systems[J]. IEEE Access, 2021, 9: 57578-57588.

[9] XU G L, TAN F, ZHAO Y Y, et al. Joint beam-hopping scheduling and coverage control in multibeam satellite systems[J]. IEEE Wireless Communications Letters, 20, 12(2): 267-271.

[10] HU X, LIU S J, WANG Y P, et al. Deep reinforcement learning-based beam hopping algorithm in multibeam satellite systems[J]. IET Communications, 2019, 13(16): 2485-2491.

[11] HAN Y F, ZHANG C, ZHANG G X. Dynamic beam hopping resource allocation algorithm based on deep reinforcement learning in multi-beam satellite systems[C]//Proceedings of the 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). Piscataway: IEEE Press, 2021: 68-73.

[12] CAO Y, LIEN S Y, LIANG Y C, et al. Collaborative deep reinforcement learning for resource optimization in non-terrestrial networks[C]//Proceedings of the 2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). Piscataway: IEEE Press, 2023: 1-7.

[13] 张沛, 刘帅军, 马治国, 等. 基于深度增强学习和多目标优化改进的卫星资源分配算法[J]. 通信学报, 2020, 41(6): 51-60.
ZHANG P, LIU S J, MA Z G, et al. Improved satellite resource allocation algorithm based on DRL and MOP[J]. Journal on Communications, 2020, 41(6): 51-60.

[14] HU X, ZHANG Y C, LIAO X L, et al. Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems[J]. IEEE Transactions on Broadcasting, 2020, 66(3): 630-646.

[15] LIN Z Y, NI Z Y, KUANG L L, et al. Satellite-terrestrial coordinated multi-satellite beam hopping scheduling based on multi-agent deep reinforcement learning[J]. IEEE Transactions on Wireless Communications, 2024, 23(8): 10091-10103.

[16] 罗铭, 詹骐榜, 邱敏蓉. 面向V2I通信的异构跨域条件隐私保护环签密方案[J]. 信息网络安全, 2024, 24(7): 993-1005.
LUO M, ZHAN Q B, QIU M R. A heterogeneous cross-domain conditional privacy protection ring signcryption scheme for V2I communication[J]. Netinfo Security, 2024, 24(7): 993-1005.

[17] 彭明阳. 面向低轨星座的跳波束机制及资源分配研究[D]. 南京: 南京邮电大学, 2023.
PENG M Y. Research on beam hopping mechanism and resource allocation for LEO constellation[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2023.

[18] 黄飞. 低轨卫星通信接入与切换策略研究[D]. 成都: 电子科技大学, 2009.
HUANG F. Research on access and handover strategy of LEO satellite communication[D]. Chengdu: University of Electronic Science and Technology of China, 2009.

[19] 张凤磊. 多层卫星网络系统接入选择技术研究[D]. 西安: 西安电子科技大学, 2020.
ZHANG F L. Research on access selection technology of multi-layer satellite network system[D]. Xi'an: Xidian University, 2020.

[20] 潘成胜, 王羽夫, 杨力. 基于改进LSTM算法的天地一体化信息网络流量预测[J]. 天地一体化信息网络, 2020, 1(2): 57-65.
PAN C S, WANG Y F, YANG L. Traffic prediction of space-integrated-ground information network based on improved LSTM algorithm[J]. Space-Integrated-Ground Information Networks, 2020, 1(2): 57-65.

[21] CAINI C, CORAZZA G E, FALCIASECCA G, et al. A spectrum-and power-efficient EHF mobile satellite system to be integrated with terrestrial cellular systems[J]. IEEE Journal on Selected Areas in Communications, 2002, 10(8): 1315-1325.

[22] THYLWE K E, MCCABE P. On calculations of Legendre functions and associated Legendre functions of the first kind of complex degree[J]. Communications in Theoretical Physics, 2015, 64(7): 9-12.

[23] HERZ C S. Bessel functions of matrix argument[J]. The Annals of Mathematics, 1955, 61(3): 474.

[24] 张晨, 彭明阳, 张更新. 基于联合优化的高通量卫星跳波束图案设计研究[J]. 南京邮电大学学报(自然科学版), 2021, 41(3): 1-8.
ZHANG C, PENG M Y, ZHANG G X. Beam hopping pattern method for high-throughput satellite based on joint optimization[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2021, 41(3): 1-8.

[25] VIDAL F, LEGAY H, GOUSSETIS G, et al. A methodology to benchmark flexible payload architectures in a megaconstellation use case[J]. International Journal of Satellite Communications and Networking, 2021, 39(1): 29-46.

[26] 3GPP TS 38.214(v18.4.0). NR physical layer procedures for data[S]. 2024.

[27] 3GPP R1-070674. LTE physical layer framework for performance verification[S]. 2007.

[28] 张雨晨. 基于多目标深度强化学习的多波束卫星动态波束调度算法研究[D]. 北京: 北京邮电大学, 2020.
ZHANG Y C. Research on dynamic beam scheduling algorithm of multi-beam satellite based on multi-objective depth reinforcement learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.

[29] ZADEH L A. Optimality and non-scalar-valued performance criteria[J]. IEEE Transactions on Automatic Control, 1963, 8(1): 59-60.

[30] TAJMAJER T. Modular multi-objective deep reinforcement learning with decision values[C]//Proceedings of the 2018 Federated Conference on Computer Science and Information Systems. Piscataway: IEEE Press, 2018: 85-93.

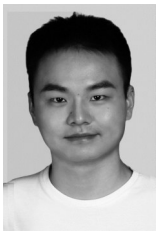
[31] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4(1): 237-285.

[32] 韩永锋. GEO卫星通信系统跳波束资源动态分配方法研究[D]. 南京: 南京邮电大学, 2022.
HAN Y F. Research on dynamic allocation method of beam hopping resources in GEO satellite communication system[D]. Nanjing: Nanjing

University of Posts and Telecommunications, 2022.

[33] 董豪, 丁子涵, 仉尚航, 等. 深度强化学习: 基础、研究与应用[M]. 北京: 电子工业出版社, 2021.
DONG H, DING Z H, ZHANG S H, et al. Deep reinforcement learning: fundamentals, research and applications[M]. Beijing: Publishing House of Electronics Industry, 2021.

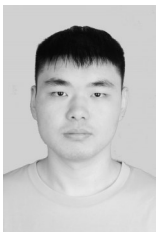
[作者简介]



**张晨**（1985– ），男，安徽淮南人，博士，南京邮电大学副研究员、高级工程师、硕士生导师，主要研究方向为天地一体化信息网络、卫星通信新体制。

**徐阳威**（1998– ），男，河南周口人，南京邮电大学硕士生，主要研究方向为卫星通信。

**李宛静**（2001– ），女，浙江台州人，南京邮电大学博士生，主要研究方向为卫星通信。



**王威**（2000– ），男，江苏宿迁人，南京邮电大学硕士生，主要研究方向为卫星通信。

**张更新**（1967– ），男，浙江平湖人，博士，南京邮电大学教授、博士生导师，主要研究方向为空间信息网络、卫星通信、深空通信、物联网、频谱监测。