# 基于行为克隆的高通量卫星通信频谱资源分配

秦浩[1,2]，李双益[1]，赵迪[1]，孟昊炜[1]，宋彬[1,2]
（1.西安电子科技大学空天地一体化综合业务网全国重点实验室，陕西 西安 710071；
2.西安电子科技大学杭州研究院，浙江 杭州 311200）

**摘　要：**为应对在高通量多波束卫星系统中，随着波束数量和用户规模的扩大，频谱资源分配问题的维度急剧增加和求解复杂度呈指数级上升这一挑战，提出了一种结合行为克隆与深度强化学习的两阶段算法。第一阶段基于行为克隆，利用已有卫星资源分配决策数据对策略网络进行预训练，通过模仿专家行为减少盲目探索，加快算法收敛。第二阶段基于近端策略优化，进一步优化策略网络，并通过引入卷积注意力模块有效地提取用户业务状态特征，以提升算法整体性能。仿真结果表明，所提算法在收敛速度和算法稳定性方面均优于其他基准算法，并在系统时延、系统平均满意度和频谱效率等性能指标上表现更佳。

**关键词：**高通量卫星；行为克隆；深度强化学习；近端策略优化；卷积注意力模块
**中图分类号：**TN92
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2024100

# Spectrum resource allocation for high-throughput satellite communications based on behavior cloning

QIN Hao[1,2], LI Shuangyi[1], ZHAO Di[1], MENG Haowei[1], SONG Bin[1,2]
1. State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an 710071, China
2. Hangzhou Institute of Technology, Xidian University, Hangzhou 311200, China

**Abstract:** In high-throughput multi-beam satellite systems, the dimensionality of the spectrum resource allocation problem increased drastically with the number of satellite beams and service users, which caused an exponential rise in the complexity of the solution. To address the challenge, a two-stage algorithm that combined behavior cloning (BC) with deep reinforcement learning (DRL) was proposed. In the first stage, the strategy network was pretrained using existing decision data from satellite operation through behavior cloning, which mimicked expert behavior to reduce blind exploration and accelerate algorithm convergence. In the second stage, the strategy network was further optimized using the proximal policy optimization (PPO), and a convolutional block attention module (CBAM) was employed to better extract the user traffic features, thereby enhancing overall algorithm performance. Simulation results demonstrate that the proposed algorithm outperforms the benchmark algorithms in terms of convergence speed and algorithm stability, and also delivers superior performance in system delay, average system satisfaction, and spectrum efficiency.

**Keywords:** high-throughput satellite, behavior cloning, deep reinforcement learning, proximal policy optimization, convolutional block attention module

**收稿日期：**2024–02–05；**修回日期：**2024–05–07
**通信作者：**宋彬，bsong@mail.xidian.edu.cn
**基金项目：**国家自然科学基金资助项目（No.62071354, No.62201419）；陕西省重点研发计划基金资助项目（No.2022ZDLGY 05-08）
**Foundation Items:** The National Natural Science Foundation of China (No.62071354, No.62201419), The Key Research and Development Program of Shaanxi Province (No.2022ZDLGY05-08)

## 0 引言

尽管5G地面通信系统可以为陆地上的人、机和物提供宽带移动接入服务，却难以覆盖海洋、沙漠等偏远地区[1]，高通量卫星通信系统凭借其多点波束覆盖、频率复用等优点可作为地面通信的重要补充，二者相互配合可实现全球无缝覆盖。而在高通量卫星通信系统中，频率资源有限[2]，如何对频率资源进行合理分配以提升高通量卫星频谱效率是卫星通信系统进一步发展的关键。

近年来，在卫星通信系统的资源分配方面已有大量基于传统优化算法的研究工作[3-6]。文献[3]基于遗传算法提出了一种联合功率和带宽分配的方法，与分配功率相比，卫星通信系统效率显著提高。文献[4]提出了一种基于模拟退火（SA, simulated annealing）算法的资源分配方法，对卫星的频率和功率资源进行灵活动态分配，相比于传统的静态资源分配算法，其在保证公平性的同时也能够有效匹配时变的业务流量需求。在密集多波束组网的卫星通信场景下，文献[5]提出了一种以初始解集构造和额外信息素沉积为核心的改进蚁群优化算法，在多突发性业务和多调度任务数量的场景下，该算法具有更高的调度效率。文献[6]提出了一种结合SA算法和非支配排序遗传算法的改进算法，在缓存资源限制下实现了用户业务满意度和频谱效率的多目标优化。然而，随着用户业务需求及星载资源数量的不断增加，高通量卫星系统资源分配优化问题的规模和复杂性也随之增加[7]。在这种情况下，传统的资源分配算法面临着求解时间增加、难以应对动态变化的业务需求以及网络不确定性等挑战[8]。

新一代人工智能技术的快速发展为解决大规模卫星资源分配问题提供了新的机遇，深度强化学习（DRL, deep reinforcement learning）是一种连续交互的学习范式，能够通过不断的探索和反馈持续改进策略，特别适合用来求解动态环境下的资源分配问题。许多研究人员已经利用DRL算法来解决卫星资源分配问题[9-11]，并证明了其可行性。文献[9]提出了基于深度Q网络的动态资源分配算法，相较于传统的优化算法，提升了低轨卫星系统的多业务服务质量。文献[10]基于DRL提出了一种解决动态信道分配的算法，在提高频谱效率的同时有效降低了卫星系统的业务阻塞率。Ma等[11]提出了一种基于近端策略优化（PPO, proximal policy optimization）的动态带宽分配（DBA, dynamic bandwidth allocation）算法，即DBA-PPO算法。该算法能够将有限的卫星频率资源一次性分配给所有的波束，相比于模拟退火算法，同等性能下该算法复杂度更低，且在多波束卫星系统下具有较高的收益。然而，DBA-PPO算法仅考虑了波束层面的资源分配，如果将该算法应用于用户层面的细粒度资源分配，将面临动作空间规模激增以及算法复杂度显著增加等问题。针对上述问题，文献[12]基于PPO算法，将系统动作空间分解为单个波束资源分配的子动作，初步实现了动作空间降维。但是，PPO算法作为DRL算法只能“从零开始”通过不断试错的方式完成学习，随着问题规模的不断增大，即使缩小了动作空间，仍然存在训练时间长和收敛速度慢[13]的问题。

行为克隆（BC, behavior cloning）技术能够通过监督学习的方式准确地模仿专家数据[14]，有助于解决PPO算法“从零开始”带来的盲目探索问题，从而进一步提高了PPO算法的学习效率[15]。为此本文聚焦于用户级资源分配，针对动作空间大和探索效率低的问题，以文献[12]为基础，提出了一种基于BC和PPO的资源分配（BC-PPORA, resource allocation based on BC and PPO）算法。该算法采用BC与PPO算法相结合的两阶段框架：在第一阶段，充分利用人类在资源分配领域取得的优秀成果，选取已有决策数据作为专家轨迹，通过BC完成策略网络训练获得专家策略；在第二阶段，以专家策略作为PPO的起点指导智能体的探索方向，充分利用前人已有的成果，从而避免盲目地随机探索，大大加快了模型的收敛速度。进一步地，为了更好地获取不同用户业务需求之间的相对位置关系对资源分配的影响，BC-PPORA算法将环境状态表示为多通道状态矩阵，并加入卷积注意力模块（CBAM, convolutional block attention module）对状态特征进行加权处理，提升了算法的全局信息捕捉能力，有效地增加了算法的稳定性。仿真结果表明，相比于其他基准资源分配算法，本文所提算法具有更快的收敛速度和更稳定的收敛效果，在性能指标上拥有更低的系统时延、更高的频谱效率和系统平均满意度。

## 1 系统模型及问题建模

高通量卫星系统下行链路如图1所示。假设卫

星通过$N_b$个波束为其覆盖范围的$N_u$个地面用户提供服务，其中波束集合定义为$\mathcal{B}=\{1,2,\cdots,N_b\}$，用户集合定义为$\mathcal{U}=\{1,2,\cdots,N_u\}$，定义$U(b)$为由波束$b$提供服务的用户集合，且满足以下要求。

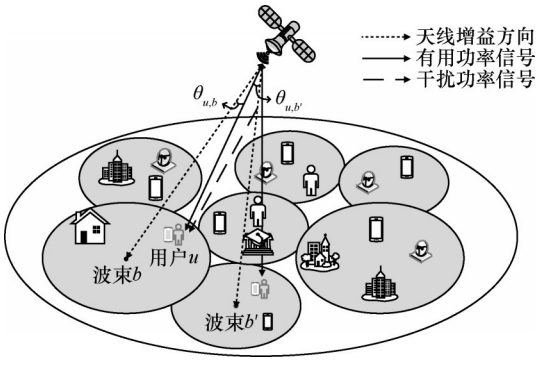


图1　高通量卫星系统下行链路

$$U(b_1)\cap U(b_2)=\varnothing,\forall b_1,b_2\in\mathcal{B}$$
$$\mathcal{U}=\bigcup_{b\in\mathcal{B}}U(b) \tag{1}$$

令总的下行带宽为$B$，共分为$M$个子信道，子信道资源集合记为$\Omega=\{1,2,\cdots,M\}$，每个子信道带宽均为$B_{\mathrm{sc}}=\dfrac{B}{M}$。为了最大化系统吞吐量，波束间采用全频率复用的形式，但是这种做法可能引起较为严重的波束间同频干扰问题[16]，如图1所示，当波束$b$在某个子信道上向用户$u$发送消息时，用户$u$会受到来自相邻波束$b'$在相同子信道上的干扰，当用户$u$正好处于2个波束交界处时，波束间同频干扰最为严重，其对系统吞吐量有着重要的影响，具体说明如下。

令$S_b(m)$表示波束$b$在子信道$m$上的发射功率，则用户$u$在该信道上的信号接收功率$P_{u,b}(m)$可以表示为

$$P_{u,b}(m)=S_b(m)G_b^{\mathrm{tx}}(u)L(f,d_{b,u})G_u^{\mathrm{rx}}(b) \tag{2}$$

其中，$G_u^{\mathrm{rx}}(b)$为用户$u$在波束$b$方向上的接收天线增益，$L(f,d_{b,u})$为信号传播损耗，$f$为载波频率，$d_{b,u}$为波束$b$到用户$u$的直线距离，$G_b^{\mathrm{tx}}(u)$为波束$b$在用户$u$方向上的发射天线增益，本文采用文献[17]的天线设计，从而$G_b^{\mathrm{tx}}(u)$可以表示为

$$G_b^{\mathrm{tx}}(u)=G_{\max}-12\eta\left(\frac{\theta_{u,b}}{\theta_{3\,\mathrm{dB},b}}\right)^2 \tag{3}$$

其中，$\eta$表示天线孔径效率，$\theta_{u,b}$表示用户$u$偏离波束$b$主轴方向的偏轴角，在波束主轴（即$\theta_{u,b}=0°$）方向上波束天线具有最大发射增益$G_{\max}$，在$\theta_{3\,\mathrm{dB},b}$方向上波束$b$的发射增益下降3 dB，$G_b^{\mathrm{tx}}(u)$和$G_{\max}$的单位为dBi。

接下来，基于接收功率$P_{u,b}(m)$讨论接收信干噪比（SINR, signal to interference plus noise ratio），假设调度间隔为$\Delta t$，也就是说系统以$\Delta t$为单位，周期性地基于用户的实时业务需求分配频率资源，下文描述的系统模型均基于调度时刻$t$，为简洁起见，所有相关变量的记号中均省略了$t$。令函数$\mathrm{bc}(u)$为用户$u$提供服务的波束，则$t$时刻用户$u$在子信道$m$上的接收信干噪比可以表示为

$$\mathrm{SINR}_u(m)=\frac{P_{u,\mathrm{bc}(u)}(m)\mathcal{I}_{y_{\mathrm{bc}(u),m}=u}}{N_0B_{\mathrm{sc}}+I_u(m)} \tag{4}$$

其中，$N_0$为噪声功率谱密度，$P_{u,\mathrm{bc}(u)}(m)$为用户$u$在子信道$m$上接收到来自其服务波束$\mathrm{bc}(u)$的信号功率，$\mathcal{I}_{y_{\mathrm{bc}(u),m}=u}$为指示函数，当等式$y_{\mathrm{bc}(u),m}=u$成立时取1，不成立时取0，资源分配变量$y_{b,m}\in U(b)\cup\{0\}$表示$t$时刻波束$b$的子信道$m$被分配给了用户，$y_{b,m}=0$时表示子信道$m$未被占用。因此$t$时刻所有$N_b$个波束的频率资源分配可表示为

$$\varsigma=\begin{bmatrix}y_{1,1}&\cdots&y_{1,M}\\ \vdots&\ddots&\vdots\\ y_{N_b,1}&\cdots&y_{N_b,M}\end{bmatrix} \tag{5}$$

式(4)中$I_u(m)$表示用户$u$在子信道$m$上接收到来自其他波束干扰信号的总功率，可具体表示为

$$I_u(m)=\sum_{b'\in\mathcal{B},b'\neq b}P_{u,b'}(m)\mathcal{I}_{y_{b',m}\neq 0} \tag{6}$$

其中，$P_{u,b'}(m)$表示用户$u$在子信道$m$上接收到来自其他波束$b'$的同频干扰信号功率。

根据式(4)和香农公式，给定$t$时刻波束$b$的频率资源分配决策为$\{y_{b,1},y_{b,2},\cdots,y_{b,M}\}$，则波束$b$在调度间隔$\Delta t$内能够传输的比特数为

$$T_b^{\mathrm{sup}}=B_{\mathrm{sc}}\Delta t\sum_{u\in U(b)}\sum_{m=1}^{M}\mathrm{lb}\left(1+\mathrm{SINR}_u(m)\right) \tag{7}$$

$t$时刻波束$b$内总的业务需求$T_b^{\mathrm{need}}$可以表示为

$$T_b^{\mathrm{need}}=\sum_{u\in U(b)}\mathrm{TH}_u^{\mathrm{need}} \tag{8}$$

其中，$\mathrm{TH}_u^{\mathrm{need}}$表示用户$u$在当前调度周期内等待传输的业务数据，单位为bit。由此，定义波束$b$下用户满意度指数（SI, satisfaction index）为

$$\mathrm{SI}_b=\begin{cases}\min\left(\dfrac{T_b^{\mathrm{sup}}}{T_b^{\mathrm{need}}},1\right), & T_b^{\mathrm{need}}\leqslant T_b^{\max}\\ \dfrac{T_b^{\mathrm{sup}}}{T_b^{\max}}, & T_b^{\mathrm{need}}>T_b^{\max}\end{cases} \tag{9}$$

其中，$T_b^{\max}$ 表示波束 $b$ 内所能达到的最大传输容量，即式(4)中干扰 $I_u(m)=0$ 时波束 $b$ 所能提供给所有用户的吞吐量。本文旨在通过优化系统频率资源分配方案来最大化高通量卫星通信系统中的平均用户满意度 $\mathcal{P}$，该优化问题被表示为

$$\begin{aligned}\max_{\varsigma}\ &=\frac{1}{N_b}\sum_{b=1}^{N_b}\mathrm{SI}_b\\ \text{s.t. }\ &\mathrm{C1}: y_{b,m}\in U(b)\cup\{0\},\forall m\in\Omega,\forall b\in\mathcal{B}\\ &\mathrm{C2}: \mathrm{TH}_{y_{b,m}}^{\mathrm{need}}>0,\forall m\in\Omega,\forall b\in\mathcal{B}\\ &\mathrm{C3}: \theta_{u,\mathrm{bc}(u)}\leqslant\theta_{3\,\mathrm{dB},\mathrm{bc}(u)},\forall u\in\mathcal{U}\end{aligned}\tag{10}$$

其中，约束C1表示同一波束内的一个子信道只能被分配给一个用户；约束C2表示子信道 $m$ 只能被分配给由当前波束提供服务的有业务需求的用户，避免信道资源浪费；约束C3表示每个用户都在其服务波束的3 dB发射增益覆盖范围内。

## 2 BC-PPORA算法

目前多数基于策略的DRL算法，如PPO算法[18]，其通常做法是随机初始化一个策略网络，然后通过随机动作探索与环境交互获得大量关于状态、动作和奖励的样本，进而利用这些样本改进策略网络，通过不断迭代获取样本和策略改进2个步骤，持续改进策略网络直到最优。但是在大规模高通量卫星资源分配场景中，上述做法将会面临巨大的动作空间，通过随机初始化策略网络进行随机动作探索，智能体将不得不执行大量劣质动作，获得大量低质量样本，这将需要极长的时间才能学习到优化分配策略，学习效率极低。事实上探索与利用之间如何权衡是所有DRL算法面临的一个重要问题，一方面，DRL算法必须保持足够的探索，充分地尝试未曾执行过的动作，以便准确评估不同动作的价值，避免陷入局部最优[19]；另一方面，DRL算法还应该充分利用已有经验，避免不必要的盲目探索，以便尽快找到最优策略。

为了避免大规模高通量卫星资源分配场景中大量无意义的探索，降低DRL算法复杂度，本文提出了两阶段的BC-PPORA算法，能够有效地将BC与PPO算法结合在一起，其整体框架如图2所示。在第一阶段，基于BC思想，充分利用前人已有成果来获得专家轨迹，之所以说是前人已有成果，是因为这些专家轨迹是卫星网络历史运行过程中自然产生的较好的决策经验（这些经验可能使用了模拟退火、比例公平[20]（PF, proportional fairness）等优秀的资源分配算法），决策样本的获取只需付出较低的存储成本。通过对专家轨迹采用监督学习的方式来训练专家策略 $\pi_{\theta_{\mathrm{bc}}}$，为DRL算法提供足够的“利用”基础；在第二阶段，使用 $\theta_{\mathrm{bc}}$ 来初始化策略网络参数作为PPO算法的起始策略，换言之，以专家策略指导智能体的探索方向，在“利用”的基础上保持充分“探索”，通过PPO算法对策略网络进行反复训练和学习，持续改进策略网络，尽可能地获得最大收益和最优策略。由于本文算法2个阶段共用同一个策略网络，以下首先说明策略网络的输入输出及结构设计，随后分别讨论了该算法的2个阶段。

### 2.1 策略网络

策略网络是资源分配的核心模块，也是连接本文算法2个阶段的纽带，其输入为环境状态，具体包括卫星波束覆盖范围信息、业务需求信息、资源占用信息等；输出为要执行的动作分布函数，即给定环境状态下资源分配方案的概率分布矩阵，本文算法将依据策略网络输出的概率分布进行随机采样以获得资源分配决策。

#### 2.1.1 状态和动作设计

综上所述，环境状态包括卫星波束覆盖信息、业务需求信息、资源占用信息等，多数基于DRL的研究通常会将环境状态表示为一维向量[11]，但是这种做法难以反映不同用户业务需求之间的相对位置关系。事实上，不同用户之间的相对位置关系对资源分配具有重要的影响，如果2个用户分属2个不同波束，相距很近且处于波束覆盖重叠区，应分配不同的子信道以免造成强烈的同频干扰；如果2个用户各自位于其波束覆盖范围中心，相距较远，则可以分配相同的子信道。基于上述考虑，本文将环境状态设计为多通道状态矩阵，以便于直观地反映用户位置与业务需求之间的关联，通过在策略网络中引入卷积注意力模块学习每个通道的通道注意力和空间注意力，有效提取状态矩阵中不同相对位置用户业务的空间特征，实现高效动态频率资源分配。

具体来说，如图3所示，首先将卫星覆盖区域近似为正方形，边长为 $H$，表示卫星总覆盖区域在经纬度上的跨度值，然后将其平均划分为 $\phi\times\phi$ 个边长为 $l$ 的小正方形网格，其中，$\phi\in\mathbb{N}^+$ 为正整数，$l=\dfrac{H}{\phi}$。基于这种区域划分方法，可以将环境状态

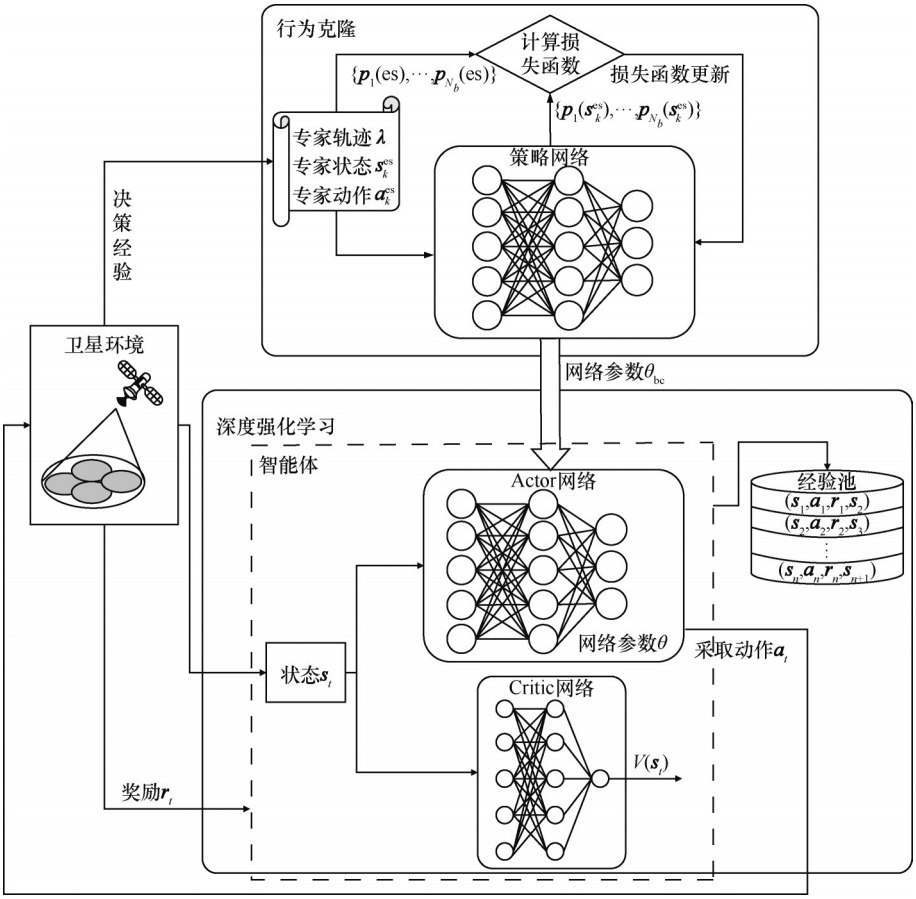


图2　BC-PPORA 算法整体框架

表示为[$\phi \times \phi$]矩阵$\boldsymbol{s}$，每个元素$\boldsymbol{s}_{i,j}$均为$D$维向量，表示图3中对应网格编号为$(i,j)$内用户的$D$维业务特征，这些特征包括当前时刻用户等待传输字节数、波束增益向量以及上一时刻已传输字节数，其中，传输数据信息反映用户服务体验，天线传输增益表示不同波束到用户的天线发射增益大小，以便反映可能对用户造成严重干扰的潜在波束，帮助系统进行干扰控制和资源优化。

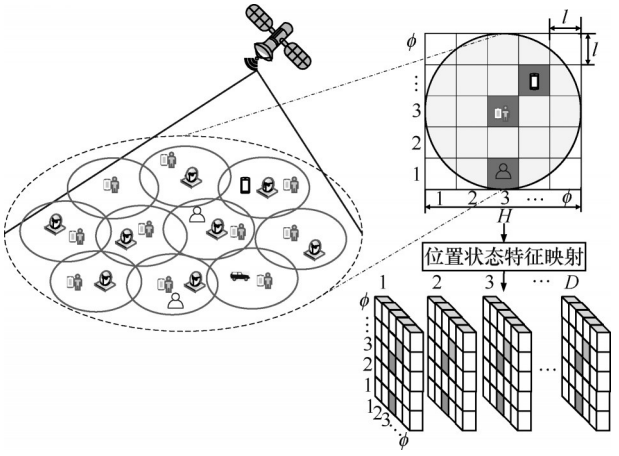


图3　卫星区域划分

动作方面，本文采用式(5)所示的动作设计，即每个波束的每个子信道分别分配给了哪个用户。需要强调的是，基于本文的动作设计，动作空间的大小为$(N_u+1)^{N_b M}$，也就是说，在大规模高通量卫星资源分配场景下，随着波束、用户和子信道数量的增加，动作空间的规模将会急剧增加，下一节将讨论如何解决这个问题。

### 2.1.2　策略网络结构

本文策略网络结构如图4所示。在网络输入端，为了从$D$通道状态矩阵中进一步提取用户终端的空间位置信息和潜在的干扰特征，采用了轻量级通用混合注意力模型CBAM[21]，不同于传统卷积运算将跨通道和空间信息混合在一起来提取信息特征的做法，CBAM顺序使用通道注意力和空间注意力模块，通过学习强调重要信息的表征，抑制不必要的特征。通过CBAM能够更加准确地提取出与用户终端空间位置信息密切相关的特征。

具体来说，首先使用卷积核大小为3 × 3的卷积网络con 1对输入的多通道状态矩阵$\boldsymbol{s}$进行初步信

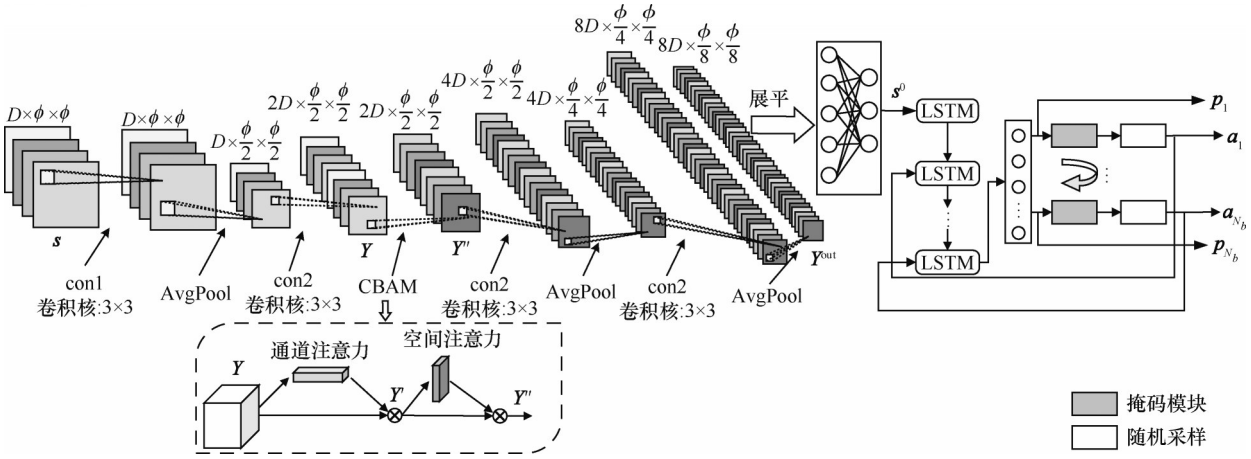


图4　本文策略网络结构

息提取，为提取主要特征减小计算量，再利用平均池化层AvgPool采样压缩状态矩阵的空间信息，这可能会导致一些业务特征细节信息丢失，为了弥补这些信息丢失所产生的影响，进而采用卷积核大小为3 × 3的卷积网络con 2使通道数变为原来的2倍，引入新的抽象特征，避免特征过度损失。con 2的输出可以表示为

$$\boldsymbol{Y}=f_{\text{con}2}(f_{\text{avg}}(f_{\text{con}1}(\boldsymbol{s}))) \tag{11}$$

其中，$\boldsymbol{Y}\in\mathbb{R}^{2D\times\frac{\phi}{2}\times\frac{\phi}{2}}$，之后采用CBAM进一步对状态矩阵进行加权调整。

$$\boldsymbol{Y}''=M_s(M_c(\boldsymbol{Y})\otimes\boldsymbol{Y})\otimes M_c(\boldsymbol{Y})\otimes\boldsymbol{Y} \tag{12}$$

其中，$M_c$和$M_s$分别表示通道注意力和空间注意力计算函数[21]，计算式分别为

$$\begin{aligned}&M_c(\boldsymbol{Y})=\\&\sigma\left(\text{MLP}\left(\text{AvgPool}(\boldsymbol{Y})\right)+\text{MLP}\left(\text{MaxPool}(\boldsymbol{Y})\right)\right)\\&M_s(\boldsymbol{Y}')=\\&\sigma\left(f^{7\times7}\left(\text{AvgPool}(\boldsymbol{Y}');\text{MaxPool}(\boldsymbol{Y}')\right)\right)\end{aligned} \tag{13}$$

其中，$\sigma(\cdot)$表示sigmoid函数，MLP$(\cdot)$表示多层感知器，MaxPool表示最大池化，$f^{7\times7}$表示卷积核大小为7 × 7的卷积操作。将CBAM输出$\boldsymbol{Y}''$再经过2次卷积网络con 2和平均池化层AvgPool，进一步提取特征汇聚业务特征的全局信息，最终输出可以表示为$\boldsymbol{Y}^{\text{out}}\in\mathbb{R}^{8D\times\frac{\phi}{8}\times\frac{\phi}{8}}$，进而将$\boldsymbol{Y}^{\text{out}}$展平为一维向量并送入全连接网络中进行信息汇总得到$\boldsymbol{s}^0$。

为应对动作空间规模激增带来的挑战，并考虑系统波束间频率资源分配的密切相关性，本文参考作者在文献[12]中采用的递归式动作设计模块，将多波束高通量卫星资源分配问题分解为多个子问题，每个子问题负责单个波束的资源分配，并将分配结果作为先验知识优化后序波束的动作选择，最大程度地避免了同信道干扰，通过逐波束迭代求解得到系统的总体资源分配方案，从而将动作空间从$(N_u+1)^{N_bM}$缩减为$(N_u+1)^M$。具体来说，在第1次递归时通过对$\boldsymbol{s}^0$进行处理，长短期记忆（LSTM, long-short term memory）网络输出波束1的动作概率分布矩阵$\boldsymbol{p}_1$为

$$\boldsymbol{p}_1=\begin{bmatrix}p_{1,1}^0 & p_{1,1}^1 & \cdots & p_{1,1}^{N_u}\\ \vdots & \vdots & \ddots & \vdots\\ p_{1,M}^0 & p_{1,M}^1 & \cdots & p_{1,M}^{N_u}\end{bmatrix},\ \sum_{u=0}^{N_u}p_{1,m}^u=1 \tag{14}$$

其中，$p_{1,m}^u$表示波束1的子信道$m$被分配给用户$u$的概率，$p_{1,m}^0$表示波束1的子信道$m$不进行分配的概率。为提高策略网络的探索效率，本文采用掩码模块（如图4中灰色矩形块所示）将$\boldsymbol{p}_1$中非法动作（如将子信道分配给没有业务请求的用户或者分配给其他波束的用户）的概率置为0并归一化得到$\boldsymbol{p}_1^{\text{mask}}$

$$\boldsymbol{p}_1^{\text{mask}}=\text{mask}\left(\boldsymbol{p}_1\right) \tag{15}$$

然后依据$\boldsymbol{p}_1^{\text{mask}}$的每一行，将每个子信道的分配概率向量进行随机采样，获得波束1的资源分配决策$\boldsymbol{a}_1$

$$\boldsymbol{a}_1=\left[y_{1,1}\ \cdots\ y_{1,M}\right] \tag{16}$$

随后将$\boldsymbol{s}^1=(\boldsymbol{s}^0,\boldsymbol{a}_1,0,\cdots,0)$作为新的状态特征输入LSTM网络得到波束2的资源分配决策$\boldsymbol{a}_2$，依次类推，在$N_b-1$次递归后状态$\boldsymbol{s}^{N_b-1}=(\boldsymbol{s}^0,\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{N_b-1})$，第$N_b$次递归后得到波束$N_b$的子动作$\boldsymbol{a}_{N_b}=[\,y_{N_b,1}\ \cdots\ y_{N_b,M}\,]$，从而得到所有波束的分配决策$\boldsymbol{a}$

$$\boldsymbol{a}=\{\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_{N_b}\} \tag{17}$$

### 2.2　第一阶段:行为克隆

BC 是模仿学习中的一类主流方法[22]，通过模仿专家在面对特定环境下的行为范例，学习相关的经验和技巧，从而提升智能体的决策能力。这里所说的专家可以是人或者是一些已知的决策经验，专家的决策能够为智能体提供正确的搜索方向[23-24]。

本文将前人已有的决策经验视为专家，智能体通过观察和学习专家在不同环境状态下的资源分配决策，模仿专家行为，从而为算法的第二阶段指明探索方向，帮助智能体缩小需要探索的动作空间，使算法能够更快地找到最优策略。

将前人已有的 $k$ 个决策经验作为状态动作对 $(\boldsymbol{s}_k^{\mathrm{es}},\boldsymbol{a}_k^{\mathrm{es}})$ 放入经验池中作为专家轨迹 $\boldsymbol{\lambda}$

$$\boldsymbol{\lambda}=\left\{\left(\boldsymbol{s}_1^{\mathrm{es}},\boldsymbol{a}_1^{\mathrm{es}}\right)\ \left(\boldsymbol{s}_2^{\mathrm{es}},\boldsymbol{a}_2^{\mathrm{es}}\right)\ \cdots\ \left(\boldsymbol{s}_k^{\mathrm{es}},\boldsymbol{a}_k^{\mathrm{es}}\right)\right\} \tag{18}$$

如图 2 所示，本文采用以下方法进行 BC 训练。首先，从专家轨迹 $\boldsymbol{\lambda}$ 中随机抽取一批状态动作对作为样本，以状态 $\boldsymbol{s}_k^{\mathrm{es}}$ 作为模型输入，对应的动作 $\boldsymbol{a}_k^{\mathrm{es}}$ 作为标签。具体地，将状态 $\boldsymbol{s}_k^{\mathrm{es}}$ 输入策略网络，得到该状态下所有波束的概率分布矩阵 $\boldsymbol{p}_1(\boldsymbol{s}_k^{\mathrm{es}}),\boldsymbol{p}_2(\boldsymbol{s}_k^{\mathrm{es}}),\cdots,\boldsymbol{p}_{N_b}(\boldsymbol{s}_k^{\mathrm{es}})$；将标签动作 $\boldsymbol{a}_k^{\mathrm{es}}$ 对应的概率设置为1，其余动作概率设为0，构成与 $\boldsymbol{p}_b(\pi)$ 形状相同的标签概率矩阵 $\boldsymbol{p}_1(\mathrm{es}),\boldsymbol{p}_2(\mathrm{es}),\cdots,\boldsymbol{p}_{N_b}(\mathrm{es})$，从而评估各个专家在同一状态下的决策差异，更好地融合不同专家知识。然后，使用交叉熵损失函数训练策略网络，使其预测概率尽可能地逼近标签概率，经过迭代优化，策略网络可以模仿专家的决策。

$$\mathrm{loss}=-\frac{1}{\mathcal{L}}\sum_{\mathcal{L}\in\boldsymbol{\lambda}}\sum_{b=1}^{N_b}\sum_{m=1}^{M}\sum_{u=0}^{N_u}p_{b,m}^{u}(\mathrm{es})\log\left(p_{b,m}^{u}(\boldsymbol{s}_k^{\mathrm{es}})\right) \tag{19}$$

其中，$\mathcal{L}$ 表示每次训练的样本个数，接下来，采用梯度下降的方法更新策略网络的参数。

$$\theta_{t+1}=\theta_t-\alpha\frac{\partial\,\mathrm{loss}}{\partial\theta}\Big|_{\theta=\theta_t} \tag{20}$$

其中，$\theta$ 为策略网络的参数，$\alpha$ 为学习率。将训练收敛后的策略网络参数保存为 $\theta_{\mathrm{bc}}$ 以供本文算法第二阶段使用。

### 2.3　第二阶段:行为改进

PPO 算法[18]是由 OpenAI 所提出的一种基于策略（Actor）网络结构和价值（Critic）网络结构的 DRL 方法，它结合了基于价值的 Critic 网络和基于策略的 Actor 网络，能够处理离散或连续的动作空间，并且具有较高的采样效率，已被广泛应用于机器人、游戏和通信领域[25-27]。基于上述优点，本文使用 PPO 算法作为第二阶段完成式(10)给出的下行链路频率资源分配问题的求解。首先，对状态、动作、奖励等算法关键要素定义如下。

1) 状态 $\boldsymbol{s}_t$：采用 2.1.1 节给出的多通道状态矩阵作为状态 $\boldsymbol{s}_t=\left[\boldsymbol{s}_{i,j}^{t}\right]_{\phi\times\phi\times D}$。

2) 动作 $\boldsymbol{a}_t$：采用式(17)定义的递归式动作 $\boldsymbol{a}_t=\{\boldsymbol{a}_1^t,\cdots,\boldsymbol{a}_b^t,\cdots,\boldsymbol{a}_{N_b}^t\}$。

3) 奖励 $r_t$：智能体在状态 $\boldsymbol{s}_t$ 执行动作 $\boldsymbol{a}_t$ 获得的奖励，该奖励用于反馈智能体当前动作决策的好坏，帮助智能体对策略进行更新。本文的奖励计算与式(10)中的优化目标相同，定义为

$$r_t=\frac{1}{N_b}\sum_{b=1}^{N_b}\mathrm{SI}_b \tag{21}$$

PPO 算法的 Actor 网络结构和 BC 策略网络结构完全相同。Critic 网络作为价值函数用于评估状态价值，其卷积网络和注意力部分的构造顺序与 Actor 网络相同，输出端 $V(\boldsymbol{s}_t)$ 表示对当前状态价值的估计。Actor-Critic 网络结构如图 5 所示。

如图 2 下侧所示，首先使用 $\theta_{\mathrm{bc}}$ 来初始化 Actor 网络参数 $\theta$，将经过 BC 训练得到的专家策略作为强化学习的初始策略，指导智能体在较小的动作空间中搜索可能的最优策略。然后与环境进行交互，根据用户业务信息对频率资源进行调度获得一系列关于状态、动作和奖励的决策轨迹。

$$\boldsymbol{\chi}=\{\boldsymbol{s}_1,\boldsymbol{a}_1,r_1,\boldsymbol{s}_2,\boldsymbol{a}_2,r_2,\cdots\} \tag{22}$$

将交互的决策轨迹依次存入经验池 $\mathcal{J}=\{\boldsymbol{\chi}_1,\boldsymbol{\chi}_2,\cdots\}$，当经验池收集到足够多的轨迹信息时，可以从中随机抽取一批数据对 Actor 网络进行改进更新，Actor 网络的更新目标可以表示为

$$\theta'=\arg\max_{\theta}\frac{1}{|\mathcal{J}|}\frac{1}{T}\sum_{\chi\in\mathcal{J}}\sum_{t=1}^{T}L(\boldsymbol{s}_t,\boldsymbol{a}_t,\theta,\theta') \tag{23}$$

其中，$\theta'$ 和 $\theta$ 分别表示新旧 2 个策略参数，$|\mathcal{J}|$ 表示经验池中所有决策轨迹的数量，$T$ 表示轨迹的总步长，$L(\boldsymbol{s}_t,\boldsymbol{a}_t,\theta,\theta')$ 被定义为

$$L(\boldsymbol{s}_t,\boldsymbol{a}_t,\theta,\theta')=\min\left(\frac{\pi_{\theta'}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{s}_t)}A_{\theta}(\boldsymbol{s}_t,\boldsymbol{a}_t),\mathrm{clip}\left(\frac{\pi_{\theta'}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{s}_t)},1-\varepsilon,1+\varepsilon\right)A_{\theta}(\boldsymbol{s}_t,\boldsymbol{a}_t)\right) \tag{24}$$

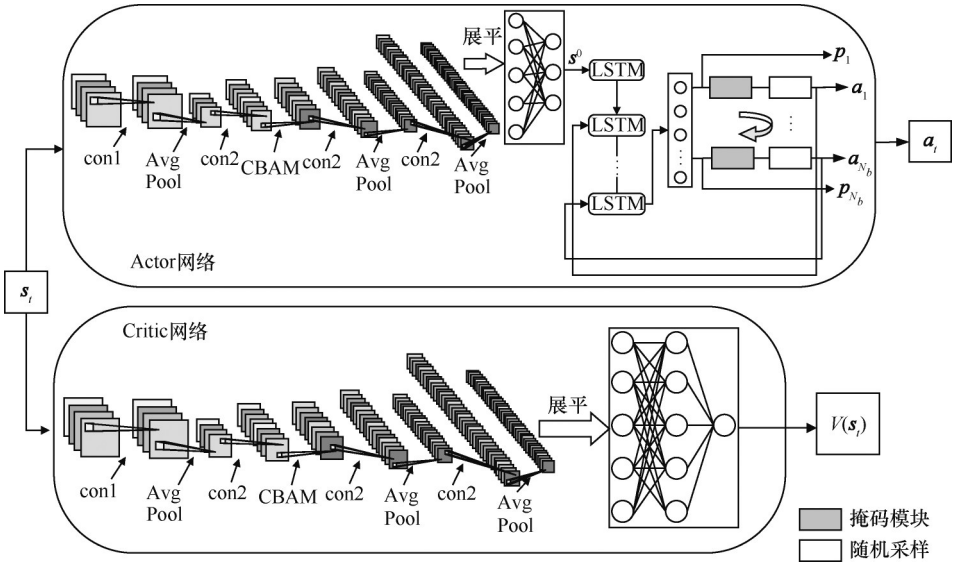


图5 Actor-Critic网络结构

其中，$\pi_{\theta'}(\boldsymbol{a}_t|\boldsymbol{s}_t)$和$\pi_{\theta}(\boldsymbol{a}_t|\boldsymbol{s}_t)$分别表示训练过程中更新的新策略和旧策略；$A_{\theta}(\boldsymbol{s}_t,\boldsymbol{a}_t)$表示Critic网络输出的优势函数，用于表示当前动作相对于平均动作而言的优势；clip函数是将新旧策略的比值约束在$[1-\varepsilon,1+\varepsilon]$，限制了策略更新幅度；$\varepsilon$是一个小数值的超参数，用于控制clip函数的裁剪幅度。

为了使Critic网络的估计尽可能地贴近真实环境中的累计奖励值，Critic网络的更新目标函数表示为

$$\delta' = \arg\min_{\delta} \frac{1}{|\mathcal{J}|}\frac{1}{T}\sum_{\chi\in\mathcal{J}}\sum_{t=1}^{T}\left(V_{\delta}(\boldsymbol{s}_t)-\sum_{t'>t}\gamma^{t'-t}r_{t'}\right)^2 \tag{25}$$

其中，$V_{\delta}(\boldsymbol{s}_t)$表示网络更新的状态价值函数，$\gamma$表示奖励折扣因子，$r_t$表示根据当前时刻得到的即时反馈奖励，$\delta'$表示更新后的Critic网络参数。

在专家策略的指导下，智能体不断与环境进行交互，收集数据，通过上述Actor网络和Critic网络的更新对专家策略进行改进，最大化累计收益，并获得最优策略。

### 2.4 BC-PPORA算法流程和步骤

结合图2所示的算法整体框架，BC-PPORA算法的第一阶段基于BC的专家策略模仿如算法1 Part1所示；第二阶段基于PPO算法的行为改进如算法1 Part2所示。首先，将已有的决策数据作为专家轨迹，通过策略网络学习专家轨迹得到专家策略，保存策略网络参数。然后，基于PPO算法进行行为改进，初始化卫星环境和算法参数，使用BC训练的参数初始化深度强化学习的Actor网络，并随机初始化Critic网络参数。最后，通过智能体与环境交互获得轨迹存入经验池中，并利用轨迹信息更新网络参数。

**算法1** BC-PPORA

**Part1** 基于BC的专家策略模仿

选择$k$个前人已有的决策经验放入经验池中，定义训练总回合数$\mathcal{D}$

1) 从经验池中依次取出不同环境状态下的动作决策作为状态动作对$\left(\boldsymbol{s}_k^{\mathrm{es}},\boldsymbol{a}_k^{\mathrm{es}}\right)$形成专家轨迹$\boldsymbol{\lambda}$
2) 设置学习率$\alpha$并初始化策略网络参数$\theta$
3) for 0 to $\mathcal{D}$ do
4) 从专家轨迹中随机取出批次为$\mathcal{L}$的数据$\left(\boldsymbol{s}_k^{\mathrm{es}},\boldsymbol{a}_k^{\mathrm{es}}\right)$
5) 将状态$\boldsymbol{s}_k^{\mathrm{es}}$输入策略网络中输出动作概率$\boldsymbol{p}(\boldsymbol{s}_k^{\mathrm{es}})$
6) 将专家决策$\boldsymbol{a}_k^{\mathrm{es}}$转换为标签概率矩阵$\boldsymbol{p}(\mathrm{es})$
7) 根据式(19)计算标签$\boldsymbol{p}(\boldsymbol{s}_k^{\mathrm{es}})$和$\boldsymbol{p}(\mathrm{es})$的损失函数loss
8) 根据式(20)以及loss更新策略网络参数$\theta$
9) end for
10) 保存神经网络参数$\theta_{\mathrm{bc}}$

**Part2** 基于PPO算法的行为改进

初始化卫星环境，使用$\theta_{\mathrm{bc}}$来初始化Actor网络参数$\theta$并初始化Critic网络参数，定义经验池最大容量$|\mathcal{J}|$、每回合交互轨迹步长$|\chi|$以及迭代总回合数

1) for episode to end do

2)　更新用户分布和业务请求情况
3)　清空经验池，将经验池容量重置为0
4)　for 0 to $|\chi|-1$ do
5)　　初始化轨迹长度为0
6)　　for 0 to $|\mathcal{J}|-1$ do
7)　　收集当前时刻的卫星环境的状态$\boldsymbol{s}_t$
8)　　智能体将$\boldsymbol{s}_t$输入Actor网络，依据策略$\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$得到动作$\boldsymbol{a}_t$，并基于式(21)得到即时奖励$r_t$
9)　　经验池储存交互信息，经验池大小加1，卫星环境基于转移概率将状态$\boldsymbol{s}_t$更新为状态$\boldsymbol{s}_{t+1}$
10)　　end for
11)　　初始化卫星环境，开始新的轨迹交互
12)　end for
13)　根据式(23)计算Actor网络的损失函数，更新$\theta$
14)　根据式(25)计算Critic网络的损失函数，更新$\delta$
15) end for

## 3　仿真设置

本节给出了卫星仿真环境参数及算法仿真参数，在Ubuntu 20.04、Python3.6.13、GeForce RTX 3090实验环境下，对本文所提BC-PPORA算法的可行性进行分析。

### 3.1　参数设置

为考虑满足真实卫星波束数量[28]，在仿真中将高通量卫星通信系统的波束个数设置为27个，324个用户随机分布在卫星的覆盖范围内，并且所有用户服从随机游走模型在卫星覆盖范围内以恒定的速度随机移动，每个时隙更新用户的位置。为了评估不同用户分布模式对系统性能的影响，本文设置了均匀分布和非均匀分布2种业务分布。

1) 均匀分布。绝对均匀的流量分布在现实生活场景中很少，仅用于评估系统在各个区域的平均性能。本文假设每个波束的业务速率为$2.8\times10^5$ bit/ms。

2) 非均匀分布。在非均匀分布下，不同波束间的业务分布差别较大，更符合实际应用场景，用于评估系统在重点覆盖区域的性能。本文假设每个波束的业务速率在$[2.5\times10^5, 4.4\times10^5]$ bit/ms随机分布。

以上2种业务分布均服从ON/OFF模型，具体来说，用户业务到达服从泊松分布过程，用户业务请求持续时间服从指数分布过程。其他主要的环境参数设置如表1所示，BC-PPORA算法参数设置如表2所示。

**表1　　参数设置**

| 参数 | 值 |
|---|---|
| 卫星高度/km | 35 786 |
| 卫星波束半径/km | 380 |
| 波束个数$N_b$/个 | 27 |
| 用户数量$N_u$/个 | 324 |
| 波束子信道个数$M$/个 | 12 |
| 系统带宽$B$/MHz | 500 |
| 子信道带宽$B_{sc}$/MHz | 41.66 |
| 调度间隔$\Delta t$/ms | 1 |
| 波束的发射功率$S_b$/W | 100 |
| 载波频率$f$/GHz | 20 |
| 噪声功率谱密度$N_0$/(dBW·MHz$^{-1}$) | −199.6 |
| 卫星天线孔径效率$\eta$ | 0.5 |
| 最大天线发射增益$G_{max}$/dBi | 50 |
| 用户接收天线增益$G_u^{rx}$/dBi | 40 |
| 用户移动速度/(m·s$^{-1}$) | 30 |
| 卫星区域划分粒度$\phi$ | 40 |
| 路径损耗$L(f,d)$ | 自由空间损耗模型 |

**表2　　BC-PPORA算法参数设置**

| 参数 | 值 |
|---|---|
| 行为克隆学习率 | $1\times10^{-4}$ |
| 行为克隆训练批次大小 | 32 |
| 行为克隆训练回合数/回合 | 1 000 |
| 强化学习训练回合数/回合 | 5 000 |
| 折扣因子 | 0.99 |
| 轨迹步长 | 50 |
| 经验池大小 | 200 |
| clip裁剪参数 | 0.2 |
| 每回合更新次数/次 | 80 |
| Actor学习率 | $3\times10^{-4}$ |
| Critic学习率 | $1\times10^{-4}$ |

### 3.2　收敛性分析

为了验证所提BC-PPORA算法的可行性，本文采用3.1节中所给出的参数进行仿真模拟。图6和图7分别给出了均匀分布和非均匀分布2种情况下BC策略网络的loss收敛曲线。

从图6和图7中可以看出，在2种分布下，loss收敛曲线均在训练开始约50回合时开始收敛，并在训练达到800回合时基本收敛，这表明BC-

PPORA算法能够按照期望方向模仿专家决策，并证实了该算法第一阶段中BC的有效性。

接下来，将BC-PPORA算法与基于LSTM递归式PPO资源分配（PPO-LOOPRA）[12]算法、基于PPO资源分配（PPO-RA）算法、基于行为克隆资源分配（BC-RA）算法以及随机资源分配（Random）算法进行了对比，验证本文所提算法的收敛性能。

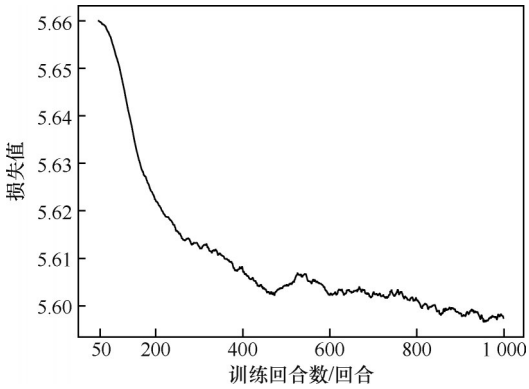


图6　均匀分布下BC策略网络的loss收敛曲线

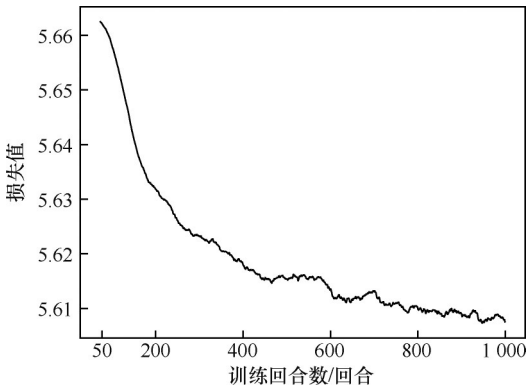


图7　非均匀分布下BC策略网络的loss收敛曲线

1) BC-PPORA：本文所提算法，状态采用多通道状态矩阵；动作为所有波束资源分配矩阵；初始策略网络经过了BC训练，输入端采用CBAM提取用户状态信息，输出端采用LSTM网络递归式输出波束动作概率矩阵。

2) PPO-LOOPRA：状态采用一维状态向量；动作为所有波束资源分配矩阵；初始策略网络为随机网络，输入端采用全连接网络，输出端采用LSTM网络递归式输出波束动作概率矩阵。

3) PPO-RA：状态采用一维状态向量；动作为所有波束资源分配矩阵；初始策略网络为随机网络，输入端采用全连接网络提取状态信息，输出端一次性输出所有波束动作概率矩阵。

4) BC-RA：经过行为克隆预训练后的策略网络直接进行行为决策。

5) Random：随机进行资源分配决策。

图8和图9为上述5种算法在均匀分布和非均匀分布2种情况下的累计奖励收敛曲线，从图8和图9中可以得到以下结论。

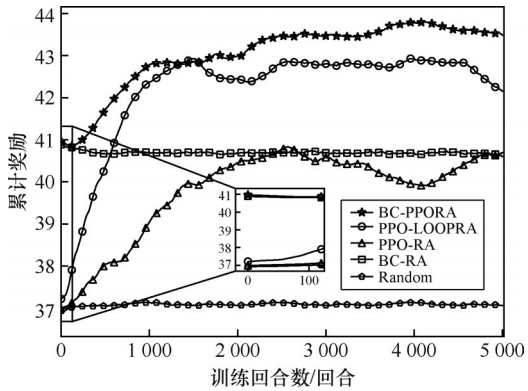


图8　均匀分布下不同算法的累计奖励收敛曲线

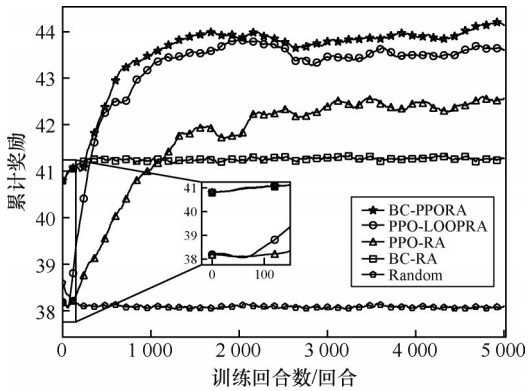


图9　非均匀分布下不同算法的累计奖励收敛曲线

1) BC-PPORA：在均匀分布下，初始累计奖励在41.0左右，在训练达到800回合后趋于稳定，并且在训练2 000回合后和3 500回合后仍有小幅上升，最终累计奖励收敛在43.8左右；在非均匀分布下，初始累计奖励在40.9左右，在训练大约1 700回合后趋于稳定，最终累计奖励收敛在44左右。

2) PPO-LOOPRA：在均匀分布下，初始累计奖励在37.2左右，在训练达到1 300回合后趋于稳定，但1 600~2 200回合和4 500~5 000回合存在剧烈波动，收敛效果始终不稳定；在非均匀分布下，初始累计奖励在38.1左右，训练大约1 100回合后开始收敛，并在训练3 800回合后趋于稳定，最终累计奖励收敛在43.5左右，可得非均匀分布的效果要明显优于均匀分布。

3) PPO-RA：在均匀分布和非均匀分布下，初

始累计奖励分别在37.0和38.1左右，均在训练大约1 500回合后累计奖励曲线趋于平缓，但在训练过程中始终存在波动。

4) BC-RA：经过行为克隆预训练后的策略网络，在均匀分布和非均匀分布下，初始累计奖励分别为40.7和41.1左右，且奖励曲线始终稳定没有出现明显的波动。

5) Random：在2种分布下，随着训练回合数增加，算法性能始终没有提升。

由上述分析可知，本文所提算法在初始累计奖励、收敛速度、最终收敛效果和稳定性方面都优于其他4种算法，这证明了所提出的行为克隆方法充分利用已有专家的决策信息，有效缩小了智能体的初始探索空间，提高了算法初始累计奖励，加速了算法收敛时间。另外，本文将环境状态设计为多通道状态矩阵，并利用CBAM进一步提取用户空间特征，提升算法对环境状态的感知能力，使算法性能更加稳定。

### 3.3　算法复杂度分析

如表3所示，在同样的实验环境下，BC每回合平均迭代时间为1 484.456 9 ms，在800回合收敛；PPO算法每回合平均迭代时间为56 594.050 5 ms，在380回合左右达到专家策略奖励。因此在第一阶段花费较低的存储成本，利用已有决策数据集作为专家轨迹，通过行为克隆对专家轨迹进行有监督训练，避免智能体在未知环境中盲目探索，快速达到较高水平，为接下来DRL提供充分的“利用”基础，创造了一个高起点。相较于直接采用DRL与未知环境交互探索，BC-PPORA算法节省了交互探索和算法计算更新的过程。

**表3　　算法迭代时间分析**

| 算法 | 每回合平均迭代时间/ms | 达到专家策略奖励回合数/回合 | 达到专家策略奖励花费时间/s |
|---|---|---|---|
| BC | 1 484.456 9 | 800 | 1 187.56 |
| PPO | 56 594.050 5 | 380 | 21 505.74 |

在第二阶段相比其他强化学习基准算法，本文所提算法BC-PPORA的网络结构和训练方式都进行了改进。不同算法网络复杂度分析如表4所示。

由表4可知，本文所提算法采用卷积网络提取特征时，具有较高的每秒浮点操作数（FLOPS, floating-point operations per second），因此具有较高的时间复杂度。另外，本文所提算法的状态被重塑为多通道状态矩阵，利用卷积网络提取系统状态信息，相较于PPO-LOOPRA算法和PPO算法采用全连接网络提取状态信息，具有较低的网络参数量，其空间复杂度相对较小。根据图8和图9可知，本文所提算法在收敛速度、收敛效果以及算法的稳定性方面都要优于其他基准算法，因为在第一阶段牺牲较低的存储成本换取了训练时间的缩短，在第二阶段牺牲部分的时间复杂度换取了空间复杂度的降低，最终换来了本文所提算法整体性能的提升。

## 4　仿真结果分析

在卫星系统2种分布下，假设用户业务到达服从均值为$\omega$的泊松分布，用户业务请求持续时间服从均值为$\mu$的指数分布，其余参数采用3.1节所述，本文对上述4种算法从低到高定义9种不同业务强度（0.1~0.9）并进行了性能仿真，其中，用户的业务强度被定义为

$$\rho = \frac{\omega}{\mu} \tag{26}$$

### 4.1　系统时延对比

图10和图11给出了7种不同算法在2种分布下的系统时延对比。从图10和图11中可以看出，不同算法的系统时延随着业务强度的增加而增加，这是因为有限的频率资源难以处理高业务强度下用户需求。在业务强度较小（0.1~0.3）时，7种算法的系统时延都较低，这说明算法能够通过合理地调度可用信道来传输用户所产生的业务需求。在业务强度较大（0.4~0.9）时，相比于PPO-LOOPRA、PPO-RA、Random、BC-RA、PF和SA算法，在均匀分布

**表4　　网络复杂度分析**

| 算法 | 时间复杂度 | | 空间复杂度 | |
|---|---|---|---|---|
| | Actor网络FLOPS | Critic网络FLOPS | Actor网络参数量 | Critic网络参数量 |
| BC-PPORA | $13.01 \times 10^7$ | $3.89 \times 10^6$ | $6.89 \times 10^5$ | $1.95 \times 10^5$ |
| BC-RA | $13.01 \times 10^7$ | — | $6.89 \times 10^5$ | — |
| PPO-LOOPRA | $5.26 \times 10^7$ | $9.76 \times 10^6$ | $2.53 \times 10^6$ | $4.88 \times 10^6$ |
| PPO | $6.37 \times 10^7$ | $9.76 \times 10^6$ | $3.19 \times 10^7$ | $4.88 \times 10^6$ |

下，BC-PPORA的系统时延分别低26.23~132.23 ms、71.48~340.69 ms、100.32~388.23 ms、51.29~163.84 ms、91.65~231.03 ms和80.02~180.25 ms；在非均匀分布下，BC-PPORA的系统时延分别低20.84~292.08 ms、68.71~459.81 ms、266.44~573.86 ms、128.90~499.58 ms、194.79~551.62 ms和119.88~478.65 ms。

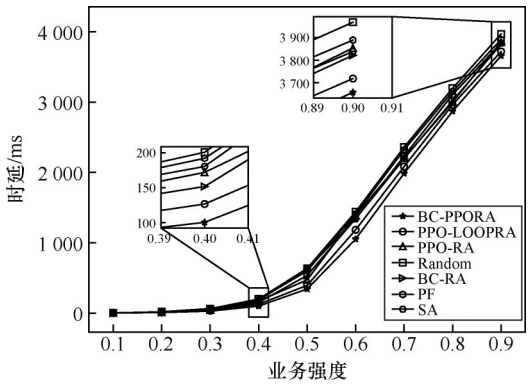

图10　均匀分布下系统时延对比

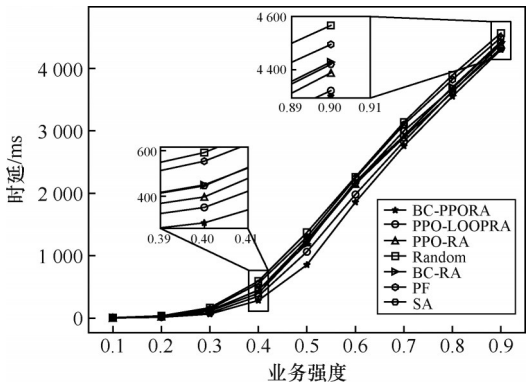

图11　非均匀分布下系统时延对比

## 4.2　频谱效率对比

频谱效率是通信系统中的常用指标，用于表示单位赫兹频带上每秒所传输的实际数据量。图12和图13给出了7种不同算法在2种分布下的频谱效率对比。

由图12和图13可以看出，不同算法随着业务强度的增加，频谱效率不断增长并趋于平缓，在业务强度较小（0.1~0.3）时，频率资源相对于业务需求来说较为充足，因此各算法都能较好地完成资源分配任务，频谱效率相差不大。在业务强度较大（0.4~0.9）时，BC-PPORA算法的频谱效率明显优于其他6种算法，相较于PPO-LOOPRA、PPO-RA、Random、BC-RA、PF和SA算法，在均匀分布下，频谱效率分别提升了0.02%~5.80%、0.41%~10.70%、11.30%~17.9%、1.01%~9.50%、3.50%~13.7%和2.05%~8.70%；在非均匀分布下，频谱效率分别提升了0.02%~4.30%、1.71%~11.8%、1.96%~18.6%、3.50%~13.40%、7.11%~16.23%和6.10%~18.00%。这是因为本文通过CBAM对环境状态中用户空间特征和潜在干扰进行分析，在动作决策时有效降低了波束间同频干扰，提升了频谱效率。

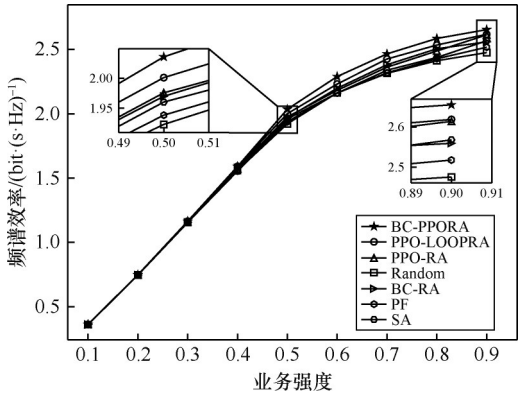

图12　均匀分布下频谱效率对比

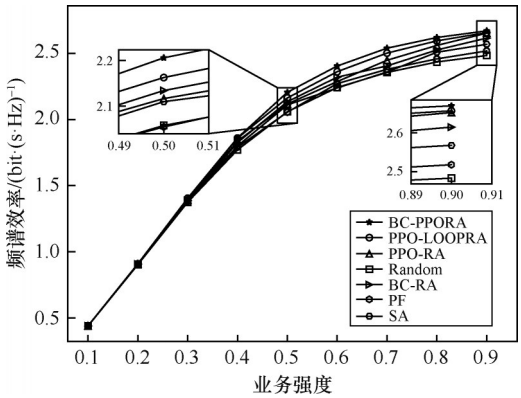

图13　非均匀分布下频谱效率对比

## 4.3　系统平均满意度对比

系统平均满意度用来衡量卫星各波束业务需求的匹配程度，图14和图15为7种算法在2种分布下的系统平均满意度对比。

从图14和图15中可以发现，在2种分布下，7种算法的系统平均满意度随着业务强度的增加呈现先下降后上升的趋势。这是因为当业务强度较小时，波束$b$总的业务需求小于波束$b$理论上能够传输的数据上限，根据式(9)波束$b$的系统满意度$\mathrm{SI}_b=\dfrac{T_b^{\mathrm{sup}}}{T_b^{\mathrm{need}}}$，随着业务强度的增加，系统波束间的同频干扰逐渐增大，波束$b$所能够提供的吞吐量的增长幅度小于波束$b$业务需求的增长幅度，因此波束$b$的满意度下降，而系统平均满意度是各个波束满意

度的加权平均值，故系统平均满意度也随之下降。随着业务强度的增加，波束 $b$ 的总业务需求不断增长直至超过波束 $b$ 理论传输的数据上限，此时满意度计算式变为 $\mathrm{SI}_b = \dfrac{T_b^{\mathrm{sup}}}{T_b^{\max}}$，目的是使波束在频率资源分配的同时尽可能地减少同频干扰，使 $T_b^{\mathrm{sup}}$ 尽可能地贴近理论上限 $T_b^{\max}$，此时由于分母固定，随着智能体策略的不断优化，波束 $b$ 所能提供的 $T_b^{\mathrm{sup}}$ 不断上升，系统平均满意度也随之上升。

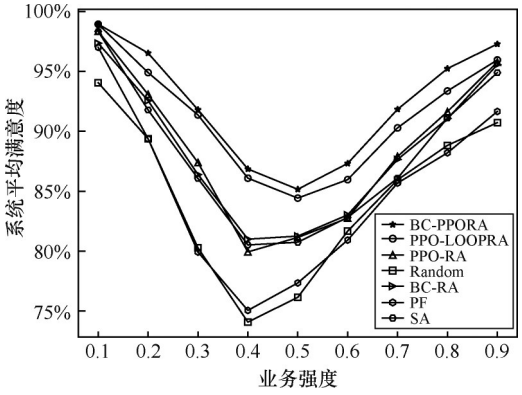


图 14　均匀分布下系统平均满意度对比

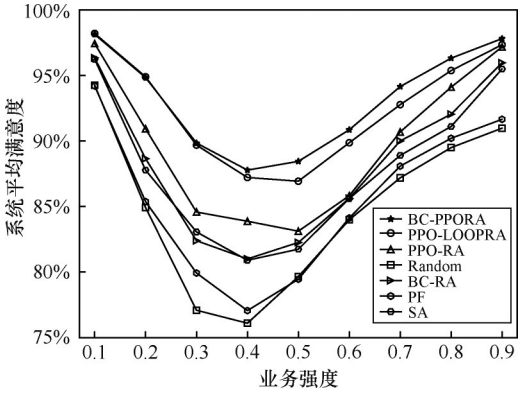


图 15　非均匀分布下系统平均满意度对比

在均匀分布下，当业务强度为 0.1 和 0.3 时，BC-PPORA 和 PPO-LOOPRA 系统平均满意度接近；当业务强度为 0.2 时，BC-PPORA 比 PPO-LOOPRA 提升了 1.63%，比 PPO-RA、Random、BC-RA、PF 和 SA 分别提升了 0.61%~4.41%、2.89%~11.52%、0.60%~4.41%、1.93%~11.87% 和 0.61%~4.41%；当业务强度为 0.4~0.9 时，BC-PPORA 相较于其他 6 种算法系统平均满意度分别提升了 0.74%~1.90%、1.52%~7.0%、5.6%~12.77%、1.72%~4.86%、5.63%~11.80% 和 1.50%~6.92%。在非均匀分布下，当业务强度为 0.1~0.3 时，BC-PPORA 和 PPO-LOOPRA 系统平均满意度接近，BC-PPORA 相较于 PPO-RA、Random、BC-RA、PF 和 SA 系统平均满意度分别提升了 0.70%~5.30%、3.87%~12.76%、1.79%~7.45%、3.93%~9.92% 和 1.90%~7.06%；当业务强度为 0.4~0.9 时，BC-PPORA 相较于其他 6 种算法系统平均满意度分别提升了 0.47%~1.52%、0.64%~5.33%、6.82%~11.70%、1.85%~6.77%、6.07%~10.71% 和 2.32%~6.88%。

## 5　结束语

本文提出了一种能够适用于大规模高通量卫星通信场景下的 BC-PPORA 二阶段算法。该算法在第一阶段通过行为克隆方法模仿专家决策，使策略网络获得专家决策；在第二阶段，将第一阶段学习到的专家策略作为深度强化学习起点，有效降低智能体前期无效探索次数，指导智能体更快地学习到最优决策。另外，本文算法加入 CBAM，使智能体能够从多通道状态矩阵更好地提取用户空间特征，减小同频干扰，使算法性能更加稳定。通过大量的仿真实验，验证了本文所提算法的有效性，仿真结果表明，本文算法在系统时延、频谱效率以及系统平均满意度方面均优于其他基准算法。

## 参考文献：

[1] WANG C X, YOU X H, GAO X Q, et al. On the road to 6G: visions, requirements, key technologies, and testbeds[J]. IEEE Communications Surveys & Tutorials, 2023, 25(2): 905-974.

[2] JIA M, ZHANG X M, SUN J T, et al. Intelligent resource management for satellite and terrestrial spectrum shared networking toward B5G[J]. IEEE Wireless Communications, 2020, 27(1): 54-61.

[3] PARIS A, DEL PORTILLO I, CAMERON B, et al. A genetic algorithm for joint power and bandwidth allocation in multibeam satellite systems[C]// Proceedings of the 2019 IEEE Aerospace Conference. Piscataway: IEEE Press, 2019: 1-15.

[4] COCCO G, DE COLA T, ANGELONE M, et al. Radio resource management optimization of flexible satellite payloads for DVB-S2 systems[J]. IEEE Transactions on Broadcasting, 2018, 64(2): 266-280.

[5] 何元智, 彭聪, 于季弘, 等. 面向密集多波束组网的卫星通信系统资源调度算法[J]. 通信学报, 2021, 42(4): 109-118.
HE Y Z, PENG C, YU J H, et al. Resource scheduling algorithm of satellite communication system for future multi-beam dense networking[J]. Journal on Communications, 2021, 42(4): 109-118.

[6] ZHANG P, WANG X H, MA Z G, et al. Joint optimization of satisfaction index and spectrum efficiency with cache restricted for resource allocation in multi-beam satellite systems[J]. China Communications, 2019, 16(2): 189-201.

[7] ORTIZ-GOMEZ F G, LEI L, LAGUNAS E, et al. Machine learning for radio resource management in multibeam GEO satellite systems[J].

Electronics, 2022, 11(7): 992.
[8] 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.
[9] LI Z W, XIE Z C, LIANG X W. Dynamic channel reservation strategy based on DQN algorithm for multi-service LEO satellite communication system[J]. IEEE Wireless Communications Letters, 2021, 10(4): 770-774.
[10] LIU S J, HU X, WANG W D. Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems[J]. IEEE Access, 2018, 6: 15733-15742.
[11] MA S J, HU X, LIAO X L, et al. Deep reinforcement learning for dynamic bandwidth allocation in multi-beam satellite systems[C]//Proceedings of the 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). Piscataway: IEEE Press, 2021: 955-959.
[12] MENG H W, XIN N, QIN H, et al. A recursive DRL-based resource allocation method for multibeam satellite communication systems[J]. Chinese Journal of Electronics, 2023, 33: 1-10.
[13] GUO W X, TIAN W H, YE Y F, et al. Cloud resource scheduling with deep reinforcement learning and imitation learning[J]. IEEE Internet of Things Journal, 2021, 8(5): 3576-3586.
[14] TORABI F, WARNELL G, STONE P. Behavioral cloning from observation[J]. arXiv Preprint, arXiv: 1805.01954, 2018.
[15] NAIR A, MCGREW B, ANDRYCHOWICZ M, et al. Overcoming exploration in reinforcement learning with demonstrations[C]//Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2018: 6292-6299.
[16] COTTATELLUCCI L, DEBBAH M, GALLINARO G, et al. Interference mitigation techniques for broadband satellite systems[C]//Proceedings of the 24th AIAA International Communications Satellite Systems Conference. Reston: AIAA, 2006: 1-13.
[17] MARAL G, BOUSQUET M, SUN Z L. Satellite communications systems: systems, techniques and technology[M]. New York: John Wiley & Sons, 2020.
[18] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint, arXiv: 1707.06347, 2017.
[19] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z C, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.
[20] NGUYEN T D, HAN Y. A proportional fairness algorithm with QoS provision in downlink OFDMA systems[J]. IEEE Communications Letters, 2006, 10(11): 760-762.
[21] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 3-19.
[22] WANG L G, FERNANDEZ C, STILLER C. High-level decision making for automated highway driving via behavior cloning[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(1): 923-935.
[23] BOJARSKI M, TESTA D D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv Preprint, arXiv: 1604.07316, 2016.
[24] PRICE B, BOUTILIER C. Accelerating reinforcement learning through implicit imitation[J]. Journal of Artificial Intelligence Research, 2003, 19: 569-629.
[25] LONG P X, FAN T X, LIAO X Y, et al. Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning[C]//Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2018: 6252-6259.
[26] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative multi-agent games[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM Press, 2022: 24611-24624.
[27] ZHANG H J, YANG N, HUANGFU W, et al. Power control based on deep reinforcement learning for spectrum sharing[J]. IEEE Transactions on Wireless Communications, 2020, 19(6): 4209-4219.
[28] 丁伟, 陶啸, 叶文熙, 等. 高轨道高通量卫星多波束天线技术研究进展[J]. 空间电子技术, 2019, 16(1): 62-69.
DING W, TAO X, YE W X, et al. Advances in research on multi-beam antenna techniques for GEO high throughput satellites[J]. Space Electronic Technology, 2019, 16(1): 62-69.

[作者简介]

**秦浩**（1976– ），男，陕西绥德人，博士，西安电子科技大学副教授、硕士生导师，主要研究方向为卫星通信星地通信体制、无线通信和卫星通信智能资源管控。

**李双益**（2001– ），男，河南许昌人，西安电子科技大学硕士生，主要研究方向为卫星通信和无线通信。

**赵迪**（1995– ），女，山东淄博人，西安电子科技大学博士生，主要研究方向为卫星通信、无线资源管理和机器学习。

**孟昊炜**（1998– ），男，河南周口人，西安电子科技大学硕士生，主要研究方向为卫星通信和无线通信。

**宋彬**（1973– ），男，河南郑州人，博士，西安电子科技大学教授、博士生导师，主要研究方向为多媒体通信、多模态数据融合与检索、基于图像内容的识别与机器学习、多模态知识图谱、强化学习、物联网、大数据和推荐系统。