

基于多智能体深度强化学习的测运控一体化资源调度方法

成思玥, 李浩然, 白卫岗, 周 笛, 朱 彦
(西安电子科技大学通信工程学院, 陕西 西安 710071)

摘 要: 随着卫星通信技术的发展, 星座规模的不断扩大, 测运控一体化成为主流趋势。星座规模大、调度对象多、复杂操作联合控制给卫星网络测运控一体化资源调度带来巨大的挑战。受制于调度算法求解效率低、约束复杂等问题, 传统的测运控资源调度技术采用提前上注测控指令, 按照固定部署执行任务, 难以满足突发事件与紧急任务的调度需求。因此, 提出一种基于多智能体演员-评判家确定性策略梯度算法的测运控一体化资源调度方法, 采用集中式训练和分布式执行的方法, 建立测运控一体化任务的多智能体模型, 通过分析邻居智能体局部信息计算调度策略, 提高任务的响应速度。依据测运控一体化资源调度问题中的模型和约束, 选择影响意义大、可解释的约束, 建立多智能体资源调度强化学习模型, 并进行仿真测试。测试结果显示, 该方法的任务收益较传统方法提高22%。

关键词: 测运控一体化; 大规模星座系统; 资源调度; 多智能体深度强化学习; 任务收益

中图分类号: V19

文献标志码: A

doi: 10.11959/j.issn.2096-8930.2023002

Resource Scheduling Method for Integration of TT&C and Observation Based on Multi-Agent Deep Reinforcement Learning

CHENG Siyue, LI Haoran, BAI Weigang, ZHOU Di, ZHU Yan
School of Telecommunications Engineering, Xidian University, Xi'an 710071, China

Abstract: With the development of satellite communication technology and the continuous expansion of the constellation scale, the integration of TT&C and observation technology has become the mainstream trend. The large constellation scale, many scheduling objects and complex operation joint control bring great challenges to the integrated resource scheduling of satellite network TT&C and observation. Subject to the low solution efficiency and complex constraints of scheduling algorithms, the traditional TT&C resource scheduling technology adopts the advance injection TT&C instructions to perform tasks according to the fixed deployment, which is difficult to meet the scheduling needs of emergencies and emergency tasks. Therefore, a kind of resource scheduling method based on multi-agent actor-Agent Actor-Critic Deterministic Policy Gradient Algorithms (MADDPG) was presented. With centralized training and distributed execution, the multi-agent model of integrated task of TT&C and observation was established. By analyzed the scheduling strategy of neighbor agent, the response speed of local information was improved. According to the model and constraints in the integrated resource scheduling problem of TT&C and observation, selected significant and interpretable constraints, then established the multi-agent resource scheduling reinforcement learning model, and carried on the simulation test. The simulation results showed that the task benefit of this method was 22% higher than the traditional method.

Keywords: integration of TT&C and observation, large-scale constellation system, resources scheduling, multi-agent deep reinforcement learning, tasks reward

收稿日期: 2022-11-15; **修回日期:** 2023-03-01

通信作者: 李浩然, hrl@xidian.edu.cn

基金项目: 国家重点研发计划资助项目 (No.2020YFB1806100); 国家自然科学基金青年项目 (No.62101410); 秦创原引用高层次创新创业人才项目 (No.QCYRCXM-2022-228)

Foundation Items: National Key Research&Development Program of China (No.2020YFB1806100), National Natural Science Foundation of China (No.62101410), The Foundation of Shaanxi Province(No.QCYRCXM-2022-228)

0 引言

随着卫星通信技术的发展,在轨卫星数量不断增加,星座规模不断扩大^[1],卫星测控资源逐渐紧张,不同需求的观测任务和测控任务之间资源使用冲突加剧。因此,如何在有限的资源条件下最大化资源的调度以满足巨型星座系统的高强度任务需求,同时提升测控资源利用率是目前关注的重点。

当前的资源调度技术多针对单一任务,如对测控任务或观测任务进行规划,没有考虑实际中测控任务和观测任务同时存在的情况。在仅考虑单一测控任务或观测任务的场景下,任务的需求较为统一,测控任务的需求一般为任务执行时间,观测任务的需求一般为任务执行时间、缓存和相关的拍摄约束条件。测运控一体化的资源调度场景较单一类型任务场景的需求更加复杂和多样化,对可视弧段的合理分配提出了更高的要求。因此,沿着测运控一体化的发展方向,同时考虑测控任务和观测任务之间的冲突和耦合关系,如何设计一种面向大规模星座的测运控一体化的高效资源调度算法是目前亟须解决的难点。

已有的测控和观测资源调度算法可分为数学规划方法、启发式算法以及人工智能算法3类。在测控资源调度问题中,通常将地面测控中心或中继卫星等测控设备与目标卫星过境期间产生的可视弧段作为待调度的测控资源,采用不同优化算法以得到调度的最优策略。其中数学规划方法已经发展成熟,在测控资源调度领域应用时间较长。GABREL等^[2]将成像侦察卫星的任务规划建立成0-1整数规划模型,采用二进制变量0-1来表示传感器是否选择该目标,以多个侦察目标间的时间互斥性作为约束。王海波等^[3]使用0-1型整数变量表示任务是否选择该卫星,使用连续型变量表示时间,建立了LTAP问题的混合整数线性规划模型。王沛^[4]采用完全分支定价和近似分支定价算法,以堆积收益为目标对成像卫星与地面站之间的测控资源调度问题进行整数规划建模并求解。徐小辉等^[5]依据航天器的可用窗口是否被利用的实时状态,采用分支定界算法对航天器测控任务进行调度。此外,还有基于图论的方法。王远振等^[6]采用时间约束Petri网和着色Petri网相结合的方法,分析卫星的任务需求类型、地面站设备类型和调度规则,建立多颗卫星与地面站之间的模型。上述基于数学规划的方法对于测控问题的约束进行了较为精确的建模,但没有考虑实际应用下求解最优调度策略的复杂性,为问题的求解带来较大的难度。

启发式算法在测控和观测资源调度问题中的应用较为广泛。ZHANG等^[7-9]采用蚁群算法解决卫星测控资源调度问题,使用混合遗传和蚁群算法相结合的方法,建立冲突构造图模型,但其考虑的约束较为简单。张天骄等^[9]采用两种不

同融合策略的新型遗传蚁群优化方法对天基和地基测控资源的一体化调度问题进行求解。陶孙杰等^[10]提出一种融合遗传算法和启发式站网调度算法的方法,针对测控数据传输一体化网络的地面站资源和时间资源进行调度。金光等^[11-12]采用地面站资源冲突消解策略的启发式调度算法来判断地面站服务的强度。这些方法较数学规划方法在求解效率上有了有一定提高,但所求得解的质量往往无法达到最优。

近年来,机器学习、神经网络等智能算法迅速发展,在多个领域具有较好的应用效果,在测控、观测任务资源调度方面的应用也逐渐增多。刘嵩等^[13]采用单隐含层节点的反向传播BP神经网络集成技术求解成像任务可调度性问题,但因其采用单隐含层结构,训练效果准确度不高。杜红梅等^[14]采用实体映射方法将测控资源调度系统划分为任务智能体、网管智能体和资源智能体三大类,对应急任务等建立了多智能体模型并进行任务规划。BADALONI等^[15]采用模糊事件推理系统解决卫星任务调度问题,将卫星调度问题描述为模糊性和不确定性下的定性和定量时间信息集成模型,然后通过给任务分配偏好度以及引入约束的可满足性的渐进概念求解。WANG等^[16]采用深度强化学习构建调度网络,对建模为动态随机背包问题的测控任务资源调度问题进行求解,可完成对遥感卫星的即时、高效响应。目前应用机器学习进行任务资源调度的研究多集中于神经网络,针对的是较为单一的测控或观测任务,没有考虑测运控多种任务的复杂场景以及在紧急任务时的快速响应问题,缺乏针对测运控一体化的资源调度研究。

为了解决上述问题,本文提出了基于多智能体深度强化学习的基于多智能体演员-评判家确定性策略梯度(Multi-Agent Actor-Critic Deterministic Policy Gradient, MADDPG)算法的测运控一体化资源调度算法,面向测控和观测任务的需求多样化和冲突严重的问题,采用多智能体强化学习和深度神经网络方法,实现最大化任务总体收益的目标,同时提高任务完成率,并最终通过仿真验证算法的有效性。本文的主要贡献如下。

(1) 建立大规模星座场景下的测运控一体化资源调度问题模型。针对大规模星座场景中存在的地面测控设备有限且全球部署困难的问题,考虑地面测控中心和中继卫星混合的天地一体化测控资源,同时面向测控和观测任务具有的不同需求建立测运控一体化的目标和约束。

(2) 设计大规模测运控一体化资源调度的多智能体强化学习算法。针对大规模星座场景下启发式算法或集中式算法所需计算和传输开销大,同时算法运行时间长、无法及时响应等问题,采用多智能体深度强化学习算法,该算法具有运行时间短、响应速度快以及计算和

传输开销小等特点。

(3) 设计大规模测控一体化资源调度场景下智能体局部特征信息。针对大规模星座场景下全局信息冗余且收集开销大的问题,对关键信息进行提取,设计每个智能体输入网络的局部特征信息,减少算法对冗余信息的关注,提升算法对大规模测控一体化资源调度场景的适应性。

1 测控一体化资源调度模型

1.1 场景

测控一体化资源调度问题是以最大化测控任务完成收益和资源利用率为目标,在有限的地面站数量与测控资源的条件下,将卫星的星间链路和卫星与地面站的星地链路,以及卫星天线数量联合考虑协同调度的问题^[17]。以最大化满足测控任务所需资源和要求为基础,为测控和观测任务安排时间、星地/星间连接和可视弧段资源,从而达到任务需求和测控资源分配的平衡关系,减少任务间对资源使用的冲突,增加总体任务收益,提高资源利用率,降低任务失败率的目标。

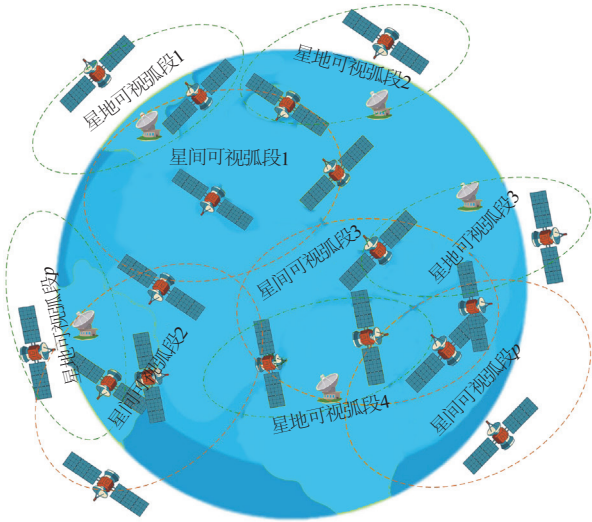


图1 大规模星座测控一体化资源调度场景

大规模星座测控一体化资源调度场景如图1所示,在网络中需要进行测控任务服务的卫星为 $\{S_1, S_2, \dots, S_n\}$,而可执行测控任务的设备资源包括地面测控中心 $\{G_1, G_2, \dots, G_p\}$ 和中继卫星 $\{S_1, S_2, \dots, S_q\}$ 两种^[18],二者总数为 M 。当目标卫星处于测控设备的可视范围时,两者间产生可用于执行测控任务的连接弧段,地面测控中心与目标卫星形成星地可视弧段,中继卫星与目标卫星形成星间可视弧段。在大规模场景下弧段间的

重叠和冲突问题较为严重,即同一测控设备同时间段内与多颗卫星存在可用弧段,但由于测控任务约束,同一时间该设备只能执行一个任务,且执行两个相邻任务间存在切换时间,若不同卫星的任务同时选择该设备执行则会产生冲突,导致任务失败。资源调度技术就是对弧段进行调度,从而在卫星与设备多对多的关系中确定满足任务需求且最大化资源效益的方案。

1.2 问题模型

卫星集合 $S = \{s_1, s_2, \dots, s_i, \dots, s_n\}$, $\forall s_i = \{S_{id_i}, S_{status_i}, S_{stor_i}, S_{location_i}\}$, 卫星总数为 N ,对每颗卫星有 S_{id_i} 为卫星 i 的标识符, S_{status_i} 为卫星 i 的状态, S_{stor_i} 为卫星 i 的缓存, $S_{location_i}$ 为卫星 i 当前所处位置。

测控设备资源集合 $E = \{e_1, e_2, \dots, e_i, \dots, e_m\}$, $\forall e_i = \{E_{id_i}, E_{type_i}, E_{status_i}, E_{location_i}\}$, 设备总量为 M ,对每个设备, E_{id_i} 为设备 i 的标识符, E_{type_i} 为设备 i 的类型(包括中继卫星和地面测控站), E_{status_i} 为设备 i 的状态, $E_{location_i}$ 为设备 i 当前所处位置。

测控弧段集合 $A = \{a_1, a_2, \dots, a_i, \dots, a_k\}$, $\forall a_i = \{A_{id_i}, A_{sat_i}, A_{equ_i}, A_{t_i^s}, A_{t_i^e}, A_{sg_i}\}$, 弧段总数为 K ;对每个弧段, A_{id_i} 为弧段 i 标识符, A_{sat_i} 为弧段 i 的目标卫星, A_{equ_i} 为弧段 i 的测控设备, $A_{t_i^s}$ 和 $A_{t_i^e}$ 为弧段 i 的开始时间和结束时间, A_{sg_i} 用来标识弧段 i 是星间链路或星地链路。

测控任务 $T = \{t_1, t_2, \dots, t_i, \dots, t_l\}$, $\forall t_i = \{T_{id_i}, T_{req_i}, T_{pro_i}, T_{status_i}\}$, 任务总数为 L ,对每个任务有 T_{id_i} 为任务 i 标识符, T_{req_i} 为任务 i 所需资源,对于测控任务其所需资源为任务持续时间 T_{dura_i} ,对于观测任务其所需资源为观测任务持续时间 T_{dura_i} 和任务占用卫星缓存 T_{stor_i} , T_{pro_i} 为任务收益, T_{status_i} 为任务状态。任务状态又包括任务开始时间 T_{start_i} 、结束时间 T_{end_i} ,任务类型 T_{type_i} 。

1.3 目标与约束

由于巨型星座系统测控一体化资源调度问题具有高复杂度,在保持场景与实际情况相符的前提下简化问题从而便于分析,因此做出如下假设^[18]:

- (1) 不考虑卫星和测控设备的故障概率;
- (2) 卫星采用一星四链的模式,一颗卫星最多配置4副天线;
- (3) 每颗卫星既可以作为目标卫星,也可以作为中继卫星;
- (4) 由于卫星间多跳路由产生的时延较大,且与路由算法关系密切,暂不考虑中继卫星到目标卫星的星间链路为多跳的情况。

在上述假设条件基础上,测控一体化资源调度问

题具有下列约束：

(1) 每个测控设备只能同时执行一个测控或观测任务；

(2) 服务同一卫星的不同任务之间需要有时间间隔，间隔大小由卫星特性决定；

(3) 同一个测控设备执行不同测控任务间需要有时间间隔，时间间隔必须大于或等于设备的切换时间；

(4) 每个测控或观测任务只可执行一次，且只能选择一个弧段执行。

测运控一体化资源调度问题的目标是最大化利用有限的测控资源，完成大规模星座的测运控需求，充分发挥资源的最大效益。上述假设和约束及问题目标的公式表示如下

$$\sum_{i=1}^L z_{ij}^k \leq 1, \forall j \in \{1, 2, \dots, M\}, k \in \{1, 2, \dots, K\} \quad (1)$$

$$\sum_{j=1}^M \sum_{k=1}^K z_{ij}^k \leq 1, \forall i \in \{1, 2, \dots, L\} \quad (2)$$

$$T_{\text{start}_{l+1} \in \text{Satellite}_l} - T_{\text{end}_{l \in \text{Satellite}_l}} \geq \beta, \forall i \in \{1, 2, \dots, N\}, l \in \{1, 2, \dots, L\} \quad (3)$$

$$[Ta_{s_{ik}}, Ta_{e_{ik}}] \geq Ta_{d_i} + t_{\text{transfer}}, \forall i \in \{1, 2, \dots, L\}, k \in \{1, 2, \dots, K\} \quad (4)$$

其中， z_{ij}^k 取值为 $\{0, 1\}$ ，表示任务 i 是否在设备 j 的 k 弧段执行， $T_{\text{start}_{l+1} \in \text{Satellite}_l}$ 和 $T_{\text{end}_{l \in \text{Satellite}_l}}$ 分别表示卫星 l 上的任务 i 实际开始执行的时间和实际结束的时间， $[Ta_{s_{ik}}, Ta_{e_{ik}}]$ 为任务 i 在弧段 k 上的有效执行时间， Ta_{d_i} 为任务 i 的持续时间， t_{transfer} 为设备的切换时间。式(1)表示在同一时刻设备上最多只能执行1个任务，式(2)表示在所有设备的所有弧段上任务 i 只执行一次，式(3)表示在同一卫星上的相邻两任务间时间间隔必须大于或等于 β ， β 为常数，式(4)表示任务 i 在弧段上的有效执行时间要大于等于任务 i 的持续时间和设备的切换时间之和。

综上所述，测运控一体化资源调度模型为

$$\begin{aligned} \max F(X) &= \sum_{i=1}^L x_i \cdot T_{\text{pro}_i} \\ \text{s.t. } x_i &\in \{0, 1\} \\ \sum_{i=1}^L z_{ij}^k &\leq 1, \forall j \in \{1, 2, \dots, M\}, k \in \{1, 2, \dots, K\} \\ \sum_{j=1}^M \sum_{k=1}^K z_{ij}^k &\leq 1, \forall i \in \{1, 2, \dots, L\} \\ T_{\text{start}_{l+1} \in \text{Satellite}_l} - T_{\text{end}_{l \in \text{Satellite}_l}} &\geq \beta, \forall i \in \{1, 2, \dots, N\}, l \in \{1, 2, \dots, L\} \\ [Ta_{s_{ik}}, Ta_{e_{ik}}] &\geq Ta_{d_i} + t_{\text{transfer}}, \forall i \in \{1, 2, \dots, L\}, k \in \{1, 2, \dots, K\} \end{aligned} \quad (5)$$

式(5)中资源调度问题的目标函数为 $\max F(X)$ ，表示调度任务的总收益最大化； x_i 为布尔变量，为1时表示任务 i 成功调度，为0则表示任务 i 调度失败； T_{pro_i} 为任务 i 的调

度收益，是任务 i 的特性，约束即为满足上述4个约束条件。

1.4 理论分析

根据大规模星座测运控一体化资源调度问题优化目标式(5)可知，资源调度问题的目标即为最大化总任务收益，该目标与成功任务数以及成功任务的单任务收益有关。其中，单任务收益是每个任务自身的性质，与任务特性和任务优先级有关。成功任务数是资源调度的结果，与任务执行时间是否冲突以及资源是否满足任务需求有关。在系统总任务数和测运控资源一定且理想的情况下，资源调度的最优结果是任务完成率百分之百，获得最大总任务收益。如式(6)所示，理想情况下的最大总任务收益为所有任务收益之和。

$$F(X)_{\max} = \left(\sum_{i=1}^L x_i \cdot T_{\text{pro}_i} \right)_{\max} = \sum_{i=1}^L T_{\text{pro}_i} \quad (6)$$

由于实际场景中存在紧急突发任务，任务的执行时间多存在冲突，且实际中不同测控资源间能力存在差异，因此资源调度理想化的最优结果无法达到。如图2所示，对于同一个测运控设备，在 T_0 到 T_5 内与3颗卫星间存在相互重叠的可视弧段，且3颗卫星均发起了任务请求，因此会产生在 T_1 到 T_2 间任务1和任务3冲突，在 T_2 到 T_3 间任务1、任务2、任务3冲突，在 T_3 到 T_4 间任务2和任务3冲突的情况，在这种情况下，只能够执行其中的部分任务。

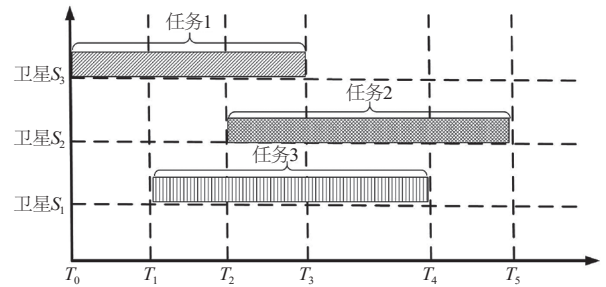


图2 测运控资源冲突示意

因此需要利用资源调度算法合理分配可视弧段给有不同收益的任务，从而尽可能避免任务冲突。在实际约束条件下，对于收益高的任务具有高任务完成率的同时，对系统资源高效利用以使总任务完成率最大化，即最优资源调度。本文算法采用多智能体强化学习和深度神经网络方法致力于在实际场景约束下努力解决任务间的冲突问题，提高任务完成率和任务收益。

2 算法设计

2.1 MADDPG算法

MADDPG算法是在2017年由OpenAI提出的一种多

智能体强化学习算法,通过对Actor-Critic算法进行改进,从而解决多智能体复杂场景下的问题。针对大规模星座场景下的资源调度问题,传统的数学规划方法和集中式的机器学习方法,在核心网进行算法计算后将结果传递给各颗卫星,在巨型星座场景下将会使得时延大大增加,且信令开销过大。而多智能体强化学习算法使得每个智能体可以分布式进行决策,减少了网络中信令传输的开销,可对任务做出实时反应,降低了时延。

MADDPG算法相对于深度 Q 网络(Deep Q Network, DQN)算法和深度策略梯度(Deep Policy Gradient, DPG)算法做出了以下改进。

(1) 采用集中式训练和分布式执行相结合的模式:对每个智能体采用Actor-Critic结构,训练时收集全局信息和其他智能体的动作和奖励信息从而进行策略估计,而执行时只需知道自身的局部信息即可。

(2) 改进了DQN经验回放机制中记录的数据,原经验回放中存储的数据为 (s_i, a_i, r_i, s'_i) ,分别表示智能体 i 的当前状态、当前动作、动作的奖励以及下一状态信息。改进后变为 $(s, a_1, \dots, a_n, r_1, \dots, r_n, s')$,这里 s 和 s' 表示全局状态信息, $a_1, \dots, a_n$ 及 $r_1, \dots, r_n$ 表示所有智能体的动作及对应奖励,便于进行集中式训练。

(3) 采用策略集合进行效果优化,即每个智能体的策略是由多个子策略组成的策略集合,每个训练episode只采用其中一个子策略,最终达到最大化策略集合的总体奖励的目标。

MADDPG算法结构如图3所示,共有 N 个智能体,每个智能体具有一个Actor网络和一个Critic网络,算法训练过程中Actor网络输入观测信息observation(o),输出所选动作action(a),生成策略 π_i ,Critic网络输入全局信息即所有智能体的观测信息和当前动作,输出对当前动

作及生成策略的评估,生成价值函数 Q_i 。在算法实际执行过程中只需要Actor网络和本智能体的局部信息即可输出动作、进行调度。

2.2 算法流程

MADDPG算法的测运控一体化资源调度框架设计如图4所示。将MADDPG算法中的每个智能体对应为一颗待执行测运控任务的卫星,因此每颗卫星作为一个单独的智能体可以通过策略网络独立自主地执行自身的决策动作。然而每颗卫星选择的资源调度策略会影响其余卫星的资源调度,所以在训练阶段采用集中式的模式将整体环境状态以及所有卫星的调度策略作为价值网络的输入,对策略网络的参数进行调整优化,使得卫星网络在考虑不同卫星间测运控资源的竞争合作关系后达到网络整体最优化的调度策略。具体设计是将每颗卫星的待测运控任务的优先级、时间窗口以及当前备选可用资源抽象为状态矩阵,将每颗卫星当前测运控任务的接受情况作为决策动作,将网络中总的任务收益作为测运控资源调度性能的评价指标以及算法的决策回报,从而达到卫星网络中调度任务总收益最大化的目标^[25]。

采用MADDPG算法解决测运控一体化资源调度问题的基本思想是:首先将待服务卫星作为智能体,依据测运控任务的目标卫星不同将其分配到不同的智能体队列中分别进行资源调度;对每一个智能体依据任务开始时间的先后依次进行资源调度,生成当前可调度的测运控资源即弧段集合,根据其是否满足任务约束做出决策;依据所做决策进行环境状态的更新,直至所有服务该卫星的任务均已做出决策,则该智能体满足结束条件;所有智能体均满足算法结束条件则已为所有任务完成资源调度,输出调度结果。MADDPG算法的测运控一体化资源调度流程如图5所示。

2.3 算法详细结构

2.3.1 神经网络结构

每个智能体的Actor网络和Critic网络结构如图6所示,采用4层神经网络结构,Actor网络输入为本智能体的观测信息即状态信息,输出为本智能体的动作;而Critic网络输入为全局状态信息以及所有智能体当前的动作集合,输出为当前动作的效用值。神经网络复杂性通常依据算法特性以及实际场景的样本数量来共同决定,对于在较大规模卫星的测运控一体化资源调度场景下,单个智能体的观测信息较为冗杂,样本数量较多,采用4层的深度神经网络相较于采用2~3层的神经网络学习到的策略具有更高的收益;且由于问题不是关于图像的处

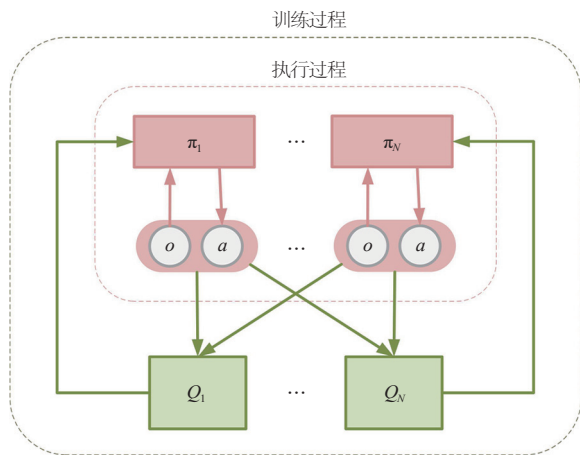


图3 MADDPG算法结构

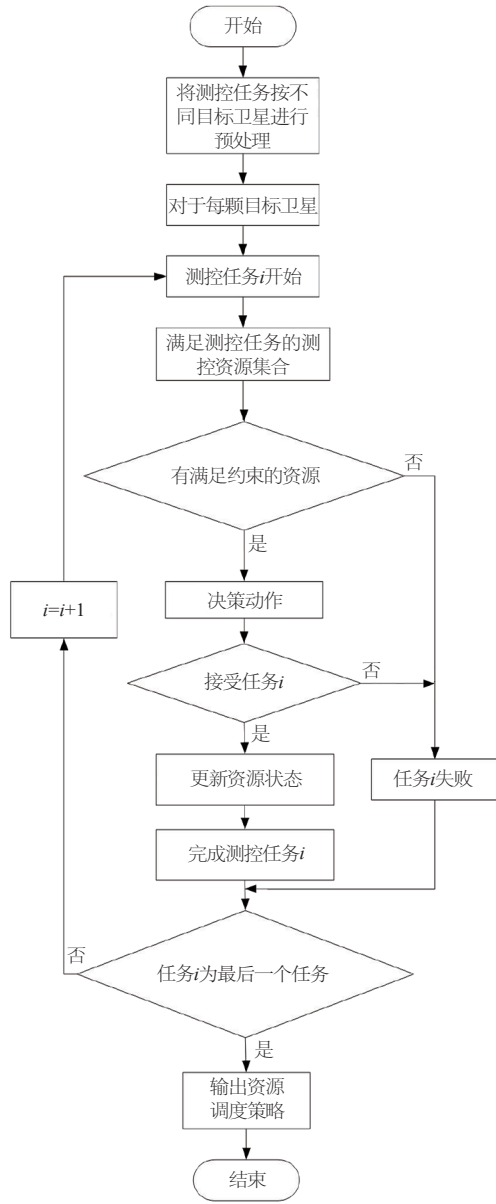


图5 MADDPG算法的测运控一体化资源调度流程

根据测运控一体化资源调度问题的目标函数设计 MADDPG算法的奖励函数如下

$$r_{i,t} = \sum_{l=1}^{L_i} x_{il} \cdot T_{pro_{il}}, \forall i \in \{1, 2, \dots, N\} \quad (7)$$

式(7)中 $r_{i,t}$ 表示智能体 i 的奖励函数, 由于 MADDPG算法每个智能体都有其相互独立的奖励函数, 所以每个智能体的奖励函数只与自身成功完成的任务数及其收益有关。其中 x_{il} 为决策变量, 表示智能体 i 的任务 l 是否成功完成, $T_{pro_{il}}$ 为智能体 i 的任务 l 的收益, L_i 为智能体 i 的总任务数。

2.3.4 策略更新

用 $\theta = \{\theta_1, \theta_2, \dots, \theta_N\}$ 分别表示每个智能体的策略网络参数, 用 $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$ 分别表示每个智能体的策略, 则每个智能体的累计期望奖励如下

$$J(\theta_i) = E_{s \sim \rho^\pi, a_i \sim \pi_{\theta_i}} \left[\sum_{t=0}^{\infty} \gamma^t r_{i,t} \right] \quad (8)$$

式(8)为每个智能体对每时隙奖励值乘上时隙权重后对时间和取期望。为了获得最大化收益, MADDPG算法按式(9)进行策略梯度更新

$$\nabla_{\theta_i} J(\mu_i) = E_{x, a \sim D} [\nabla_{\theta_i \mu_i} (a_i | o_i) \nabla_{a_i} Q_i^\mu(x, a_1, a_2, \dots, a_N) |_{a_i = \mu_i(o_i)}] \quad (9)$$

式(9)中 a_i 和 o_i 为第 i 个智能体的动作和观测信息, $x = [o_1, o_2, \dots, o_N]$ 为 N 个智能体的观测向量, 即状态信息。 $Q_i^\mu(x, a_1, a_2, \dots, a_N)$ 为智能体 i 的集中式状态-动作函数, 与全局状态及所有智能体动作有关, 每个智能体独立学习自己的状态-动作函数。 D 表示经验回放机制中存储经验的经验池, 每条经验为 $(s, a_1, \dots, a_n, r_1, \dots, r_n, s')$ 。状态动作函数, 即 Q 值函数的更新公式如下

$$L(\theta_i) = E_{x, a, r, x'} [(Q_i^\mu(x, a_1, \dots, a_N) - (r_i + \gamma Q_i^{\mu'}(x', a_1', \dots, a_N')) |_{a_i = \mu_i(o_i)})^2] \quad (10)$$

式(10)中 $Q_i^{\mu'}$ 为目标网络的 Q 值函数, $\mu' = [\mu'_1, \mu'_2, \dots, \mu'_N]$ 表示目标策略网络中滞后的网络参数, 采用软更新的方法对目标动作网络和目标策略网络

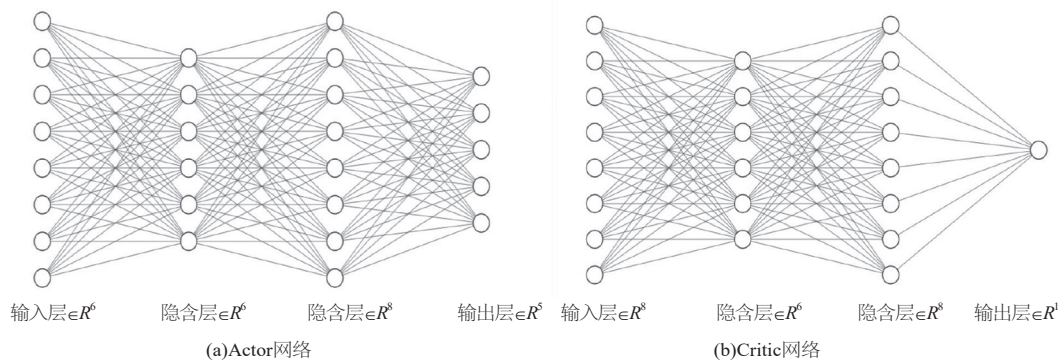


图6 每个智能体的Actor网络和Critic网络结构

的参数进行更新

$$\begin{aligned}\theta^{o'} &= \tau\theta^o + (1-\tau)\theta^{o'} \\ \theta^{\mu'} &= \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}\end{aligned}\quad (11)$$

式(11)中 $\theta^{o'}$ 和 $\theta^{\mu'}$ 分别为目标Actor和Critic网络的参数, τ 为常数, 一般取值为0.01。通过软更新的方法可使目标网络参数变化较小, 从而使算法稳定收敛。

3 仿真实例与结果分析

3.1 场景设计

本节设计场景为100颗低轨卫星, 15个地面站组成的调度场景, 其中每颗卫星既可以作为目标卫星同时也可以作为中继卫星, 测控资源为同时包括地面测控站和中继卫星的天地一体化测控资源。调度周期为一天, 所有任务在调度周期内随机产生, 测控资源切换时间为1 min。

3.2 参数设置

MADDPG算法仿真参数设置见表1, 运行总次数为1 000次, 每次仿真周期为24 h, 运行步数为8 400步, 运行50次后开始训练, 动作网络学习率为0.000 1, 策略网络学习率为0.001, 采用贪婪算法选择动作, 其参数 ε 为0.9, 计算损失的参数 γ 为0.95, 网络参数软更新 τ 等于0.01, 经验池大小为500 000条, 训练批次大小为256, 神经网络采用4层全连接结构, 神经元数量为256个, 均采用Relu激活函数, 并采用Adam优化器。

3.3 仿真结果

在100颗卫星和15个地面站场景下训练1 000次后, 结果如图7~图8所示。为衡量本文算法性能, 采用单智能体集中式的DQN算法和单智能体A3C (Asynchronous Advantage Actor-Critic) 算法作为对照算法, 比较3种算法在训练1 000次后调度结果并进行分析。

图7为3种算法训练1 000次的平均每步收益。可以看出训练前3种算法的起始收益接近, 经过1 000次的训练后MADDPG算法的平均收益相比于DQN和A3C算法有明显的提升, MADDPG算法的平均收益比DQN算法平均收益高182, 提高了约23%, 比A3C算法的平均收益高109, 提高了约13%。

图8为MADDPG算法、DQN算法和A3C算法的成功任务数。可以看出, 训练开始前3种算法成功任务数非常接近, 均在620个左右。随着训练的进行, MADDPG算法的成功任务数明显高于DQN和A3C算法, 经1 000次训练后MADDPG算法成功任务数较DQN算法多97个, 较A3C算法多68个, 成功任务数分别提高了14%和10%。

表1 MADDPG算法仿真参数

参数	值	参数	值
运行总次数/次	1 000	经验池大小	500 000条
运行步数/步	8 400	训练批次大小	256
开始训练的次数	50次	软更新参数 τ	0.01
动作网络学习率	0.000 1	动作选择参数 ε	0.9
策略网络学习率	0.001	计算损失参数 γ	0.95

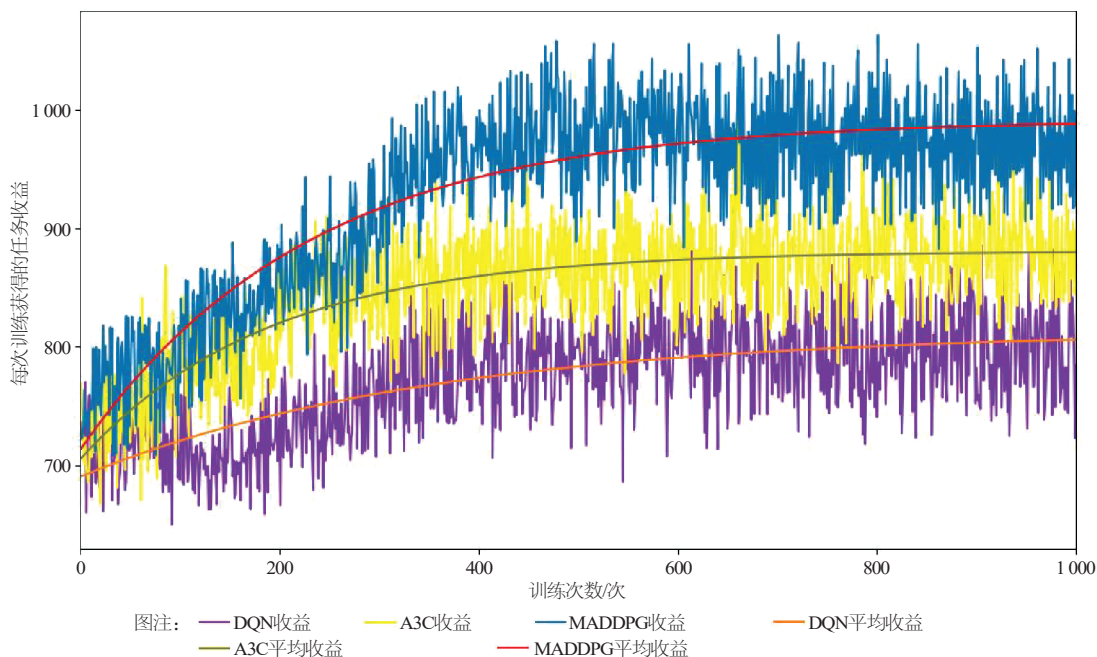


图7 MADDPG等3种算法训练1 000次的平均每步收益

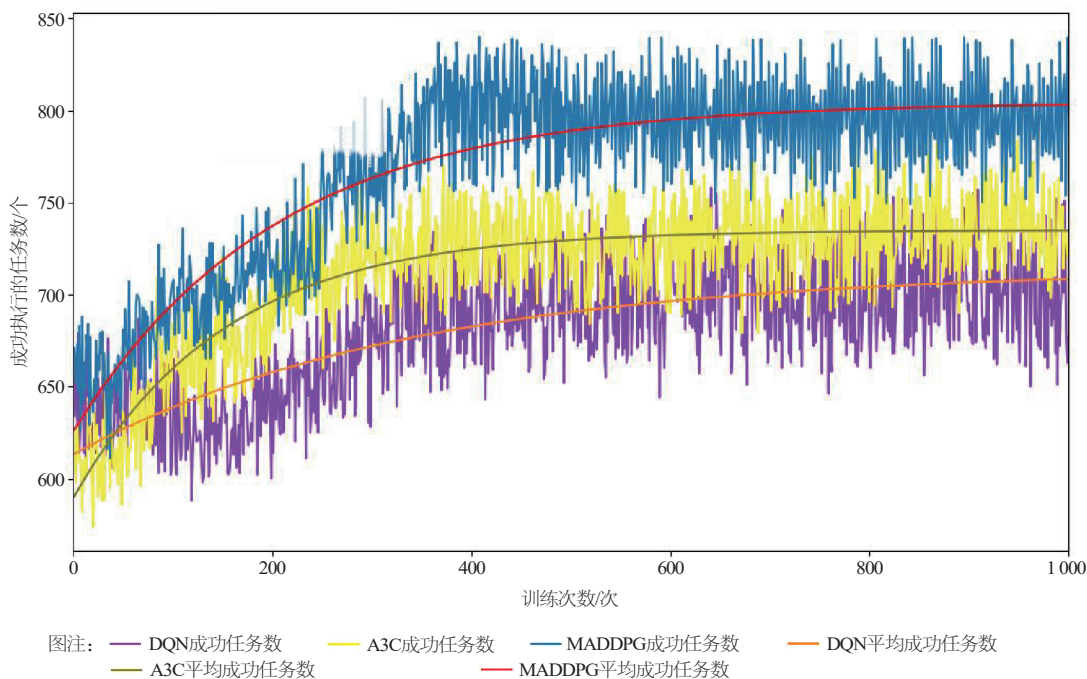


图8 MADDPG等3种算法成功任务数

从图7和图8中的仿真结果可以看出MADDPG算法相较于DQN算法和A3C算法在平均收益和平均成功任务数上均有显著的提升。此外MADDPG算法在运行效率上远优于单智能体的DQN算法和A3C算法。MADDPG算法的单次平均运行时长仅为61 s, 而单智能体DQN算法单次平均运行时长为274 s。MADDPG算法运行效率较DQN算法提高了78%。

4 结束语

为进一步提高所提算法在大规模星座系统测控一体化资源调度领域的多种场景应用适应度, 下一步将深入分析在大规模星座场景下为不同卫星进行测控的优先级对算法性能的影响, 以及对测控任务和观测任务进行详细的优先级划分, 改进场景的约束, 使算法更加具有现实场景应用的能力。同时, 对MADDPG算法进行优化, 进一步提升算法的性能, 提升测控资源的利用率和负载均衡。

参考文献:

- [1] 张威, 吴涛, 马宏, 等. 智能一体化航天测控网络发展探析[J]. 天地一体化信息网络, 2021, 2(2): 82-89.
ZHANG W, WU T, MA H, et al. Discussion on the development of integrated and intelligent space TTC & OC network[J]. Space-Integrated-Ground Information Networks, 2021, 2(2): 82-89.
- [2] GABREL V. Strengthened 0-1 linear formulation for the daily satellite mission planning[J]. Journal of Combinatorial Optimization, 2006, 11(3): 341-346.
- [3] 王海波, 徐敏强, 王日新, 等. 地对观测小卫星星座长期任务规划求解技术[J]. 系统工程与电子技术, 2011, 33(6): 1293-1298.
WANG H B, XU M Q, WANG R X, et al. Long-term acquisition plan method for small satellites constellation[J]. Systems Engineering and Electronics, 2011, 33(6): 1293-1298.
- [4] 王沛. 基于分支定价的多星多站集成调度方法研究[D]. 长沙: 国防科学技术大学, 2011.
WANG P. Research on branch-and-price based multi-satellite multi-station integrated scheduling method[D]. Changsha: National University of Defense Technology, 2011.
- [5] 徐小辉, 胡绍林, 郭小红, 等. 航天器测控资源调度模型及算法[J]. 中国空间科学技术, 2014, 34(3): 32-37.
XU X H, HU S L, GUO X H, et al. Spacecraft TT & C scheduling models and algorithms[J]. Chinese Space Science and Technology, 2014, 34(3): 32-37.
- [6] 王远振, 赵坚, 聂成. 多卫星一地面站系统的Petri网模型研究[J]. 空军工程大学学报(自然科学版), 2003, 4(2): 7-11.
WANG Y Z, ZHAO J, NIE C. Study on Petri net model for multi-satellites-ground station system[J]. Journal of Air Force Engineering University (Natural Science Edition), 2003, 4(2): 7-11.
- [7] ZHANG N, FENG Z R, FENG Y J, et al. An optimization model for multisatellite resources scheduling[C]//Proceedings of 2006

- 6th World Congress on Intelligent Control and Automation. Piscataway: IEEE Press, 2006: 7400-7404.
- [8] ZHANG Z J, HU F N, ZHANG N. Ant colony algorithm for satellite control resource scheduling problem[J]. *Applied Intelligence*, 2018, 48(10): 3295-3305.
- [9] 张天骄, 李济生, 李晶, 等. 基于混合蚁群优化的天地一体化调度方法[J]. *系统工程与电子技术*, 2016, 38(7): 1555-1562.
- ZHANG T J, LI J S, LI J, et al. Space-ground integrated scheduling based on the hybrid ant colony optimization[J]. *Systems Engineering and Electronics*, 2016, 38(7): 1555-1562.
- [10] 陶孙杰, 宋竹. 一种测控数据传输一体化站网资源调度算法[J]. *电讯技术*, 2018, 58(7): 760-767.
- TAO S J, SONG Z. A ground-station resource scheduling algorithm based on integration of TT & C and data transmission[J]. *Telecommunication Engineering*, 2018, 58(7): 760-767.
- [11] 金光, 武小悦, 高卫斌. 卫星地面站资源调度优化模型及启发式算法[J]. *系统工程与电子技术*, 2004, 26(12): 1839-1841, 1875.
- JIN G, WU X Y, GAO W B. Ground Station resource scheduling optimization model and its heuristic algorithm[J]. *Systems Engineering and Electronics*, 2004, 26(12): 1839-1841, 1875.
- [12] 金光, 武小悦, 高卫斌. 基于冲突的卫星地面站系统资源调度与能力分析[J]. *小型微型计算机系统*, 2007, 28(2): 310-312.
- JIN G, WU X Y, GAO W B. Conflict based resource scheduling and capability analysis of satellite-ground station system[J]. *Journal of Chinese Computer Systems*, 2007, 28(2): 310-312.
- [13] 刘嵩, 白国庆, 陈英武. 地球观测网络成像任务可调度性预测方法[J]. *宇航学报*, 2015, 36(5): 583-588.
- LIU S, BAI G Q, CHEN Y W. Prediction method for imaging task schedulability of earth observation network[J]. *Journal of Astronautics*, 2015, 36(5): 583-588.
- [14] 杜红梅, 柯宏发. 基于多智能体技术的航天测控资源调度模型设计[C]//系统仿真技术及其应用. 2015(16): 78-81.
- DU H M, KE H F. Design of space TT&C resource scheduling model based on multi-agent technology [C]// *System simulation technology and its application*. 2015(16): 78-81.
- [15] BADALONI S, FALDA M, GIACOMIN M. Solving temporal over-constrained problems using fuzzy techniques[J]. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 2007, 18(3): 255-265.
- [16] WANG H J, YANG Z, ZHOU W G, et al. Online scheduling of image satellites based on neural networks and deep reinforcement learning[J]. *Chinese Journal of Aeronautics*, 2019, 32(4): 1011-1019.
- [17] ZHANG T J, KE L J, LI J S, et al. Fireworks algorithm for the satellite link scheduling problem in the navigation constellation[C]// *Proceedings of 2016 IEEE Congress on Evolutionary Computation (CEC)*. Piscataway: IEEE Press, 2016: 4029-4037.
- [18] 窦骄, 韩孟飞, 宁金枝, 等. 小卫星测控通信技术发展与趋势[J]. *航天器工程*, 2021, 30(6): 113-119.
- DOU J, HAN M F, NING J Z, et al. Development and trends of TT & C and communication technology for small satellite[J]. *Spacecraft Engineering*, 2021, 30(6): 113-119.
- [19] 陈峰, 武小悦. 天地测控资源一体化调度模型[J]. *宇航学报*, 2010, 31(5): 1405-1412.
- CHEN F, WU X Y. Space and ground TT & C resource integrated scheduling model[J]. *Journal of Astronautics*, 2010, 31(5): 1405-1412.
- [20] 孟欢. 测控资源调度预测与服务功能链优化映射研究[D]. 天津: 天津大学, 2019.
- MENG H. Research on measurement and control resource scheduling prediction and service function chain optimization mapping[D]. Tianjin: Tianjin University, 2019.
- [21] 安元元. 低轨卫星测控系统调度策略的设计与实现[D]. 西安: 中国科学院大学(中国科学院国家授时中心), 2021.
- AN Y Y. Design and realization of schedule strategy of the LEO satellite telemetry and telecontrol system[D]. Xi'an: National Time Service Center, Chinese Academy of Sciences, 2021.
- [22] 雷瀚, 杨帆, 刘建平, 等. 航天地基测运控资源调度多任务需求分析及统一的任务模型研究[C]//第三届体系工程学术会议论文集——复杂系统与体系工程. 2021: 100-110.
- LEI H, YANG F, LIU J P, et al. Multi-task demand analysis and unified task model research of space-based survey, transportation and control resource scheduling [C]// *Proceedings of the 3rd Symposium on Systems Engineering-Complex Systems and Systems Engineering Management*. 2021: 100-110.
- [23] 唐成圆. 面向测运控的卫星网络常规任务规划算法研究[D]. 西安: 西安电子科技大学, 2020.
- TANG C Y. Routine task planning algorithm for telemetry tracking and operation control center in satellite network[D]. Xi'an: Xidian University, 2020.
- [24] 武艺. 基于深度强化学习的多星测控资源调度方法研究[D]. 重庆: 重庆大学, 2020.
- WU Y. Research on A scheduling method of TT&C resources for multi-satellite based on deep reinforcement learning[D]. Chongqing: Chongqing University, 2020.
- [25] 李长德, 徐伟, 徐梁, 等. 基于深度神经网络的多星测控调度方法[J]. *中国空间科学技术*, 2022, 42(1): 65-72.
- LI C D, XU W, XU L, et al. Multi-satellite TT&C scheduling method based on deep neural network[J]. *Chinese Space Science and Technology*, 2022, 42(1): 65-72.

- [26] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[EB]. 2017.

[作者简介]



成思玥（1998-），女，西安电子科技大学综合业务网理论与关键技术国家重点实验室硕士生，主要研究方向为巨型星座系统宽带民航业务资源切片。



李浩然（1991-），男，西安电子科技大学综合业务网理论与关键技术国家重点实验室讲师，主要研究方向为大规模卫星星座资源分配与任务调度、网络的群智涌现与群体行为。



白卫岗（1987-），男，西安电子科技大学综合业务网理论与关键技术国家重点实验室副教授，主要研究方向为水声通信网络、卫星通信网络、空天地海一体化网络架构、组网协议及仿真系统。



周笛（1991-），女，西安电子科技大学综合业务网理论与关键技术国家重点实验室副教授，主要研究方向为空间信息网络任务规划及资源管理、卫星互联网资源管控技术等。



朱彦（1993-），男，西安电子科技大学综合业务网理论与关键技术国家重点实验室讲师，主要研究方向为端到端可靠传输、服务质量保障等。