基于强化学习的多层卫星网络边缘安全决策方法

左珮良[1]，侯少龙[1,2]，郭超[1]，蒋华[1,2]，王文博[3]

（1. 北京电子科技学院电子与通信工程系，北京 100070；2. 西安电子科技大学通信工程学院，陕西 西安 710068；
3. 北京邮电大学信息与通信工程学院，北京 100876）

摘 要：多层卫星网络凭借其全球覆盖、高机动性、强抗毁性等优势，成为学术界和产业界的关注热点。然而，多层异构卫星网络如何依靠卫星节点的自主判决能力，发挥网络边缘场景中针对感知数据包含加解密和压缩在内的处理以及回传方面的任务协作相关问题亟待解决。鉴于此，确立了包含低轨卫星的多层卫星网络边缘模型，并提出了一种基于深度强化学习的数据压缩与加密回传决策方法，旨在以确保数据安全为前提，以低传输时延为目标，实现任务卫星在多层卫星网络架构中的边缘决策。仿真和分析表明，所提方法具备合理性和可行性，能够快速地实现收敛，且相比于多个启发式方法具备更优异的性能。

关键词：多层卫星网络；低轨卫星；边缘决策；强化学习；数据加密

中图分类号：TN92

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2022111

Security decision method for the edge of multi-layer satellite network based on reinforcement learning

ZUO Peiliang[1], HOU Shaolong[1,2], GUO Chao[1], JIANG Hua[1,2], WANG Wenbo[3]

1. Department of Electronics and Communication Engineering, Beijing Electronic Science and Technology Institute, Beijing 100070, China
2. School of Communication Engineering, Xidian University, Xi'an 710068, China
3. School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract: With the global coverage, high mobility, strong survivability and other advantages, multi-layer satellite network has become a hot spot in the academic and industrial circles. However, how to rely on the autonomous decision-making ability of nodes in multi-layer heterogeneous satellite networks to give full play to the task cooperation related to the processing of perceptual data, including encryption, decryption and compression, and the return in the network edge scene needs to be solved urgently. In view of this, the edge model of multi-layer satellite network including LEO satellite nodes was established, and a decision-making method of data compression and encryption backhaul on the basis of deep reinforcement learning was proposed, which aimed to realize the edge decision-making of mission satellites in multi-layer satellite network architecture on the premise of ensuring data security and with the goal of reducing the transmission delay. Simulation and analysis show that the proposed method is reasonable, and it can quickly achieve convergence, and has better performance than multiple heuristic methods.

Keywords: multi-layer satellite network, LEO satellite, edge decision, reinforcement learning, data encryption

收稿日期：2022−02−09；**修回日期：**2022−04−25

通信作者：侯少龙，hsl_1999@163.com

基金项目：国家自然科学基金资助项目（No.62001251, No.62001252）；北京高校“高精尖”学科建设基金资助项目（No.202100130401）；西安电子科技大学综合业务网理论及关键技术国家重点实验室基金资助项目（No.ISN22-13）

Foundation Items: The National Natural Science Foundation of China (No.62001251, No.62001252), “High-precision” Discipline Construction Project in Beijing Universities (No.202100130401), Xidian University Integrated Business Network Theory and Key Technology State Key Laboratory Project (No.ISN22-13)

0 引言

21 世纪以来，电子与通信行业发展迅猛，伴随着 5G 通信技术投入商用，全球诸多行业进入高速互联时代。卫星通信作为地面通信的有力补充，凭借着其覆盖范围广、不受地面地形影响的特点，为偏远地区以及广阔的海洋提供了基础通信保障[1]。然而，卫星通信所具备的价值却远不止如此，近几年来，多层卫星网络作为空天地一体化网络技术的空间构成，已成为学术界公认的下一代通信技术（6G）的重要组成部分[2-3]。

低轨卫星由于在卫星网络中具有对地服务时延小、轨道周期短和高机动的特点，成为地面通信网络服务的重要辅助者。目前，在轨运营的较知名的低轨卫星系统有铱星系统、OneWeb 和星链（Starlink）等[4]。由美国 SpaceX 公司主导的星链系统目前已部署在轨卫星 1 700 多颗，依靠较成熟的发射技术，该公司计划将星链打造成具备三层高度的低轨互联卫星系统，使其为所服务区域的用户提供能够与 4G 速度相媲美的网络服务。

低轨卫星的特征使其成为地面网络的重要补充[5]，一个典型的应用便是能够实现计算卸载和访问资源边缘化存储的低轨卫星边缘计算[6]。文献[7]对低轨星座通信网络边缘计算的架构开展了研究，并提出了一种依靠排队论和加权方式的计算节点选择策略。文献[8]考虑了星地多级边缘计算的场景，对卫星边缘计算网络和地面边缘计算网络混合模式下的负载调度策略进行了研究，并通过搭建仿真平台，验证了所提方案的可行性和优势性。文献[9]则考虑将低轨卫星和高空平台均视为边缘计算实体，提出了可靠的子问题转化方法，实现对星地融合网络场景下用户、多输入多输出天线预编码、计算任务和资源的联合划分。

对于融合高、中、低轨卫星的多层异构卫星网络场景，目前在理论分析和应用方面的相关研究相对较少。文献[10]在考虑卫星远程物联网的实际信道条件和太阳能摄取转换的前提下，依靠强化学习方法，解决了高、低轨卫星协同的联合资源划分和感知数据规划问题。文献[11]对多层卫星网络的容量水平进行了计算分析，并通过将多层异构卫星网络的特征纳入方法考量，实现了对算力和存储资源的合理规划。

与以上现有研究存在明显不同，本文关注于多层异构卫星系统内部的协同联动场景。低轨和中轨卫星与地面具有非同步性，且低轨卫星具备高速移动的特点，使多层卫星网络具备整网高动态和局部弱动态的特点[12]，虽然中、低轨卫星的运动相对地面具有周期性，但由局部多层卫星节点组成的区域卫星网络具备直接互联时间短、周期时间长的特征，基于以上特征，如何进行快速有效的决策，充分发挥边缘区域各层卫星节点及网络协同联动的潜力，成为一项极具挑战的研究任务。在本文的场景中，低轨卫星网络层节点主要负责对地观测任务，其所获得数据需要依靠卫星网络安全地传回给地面控制中心；中轨卫星节点由于具备较强的算力和存储能力，承担边缘计算任务；高轨卫星节点主要负责计算和数据转发。本文通过提出一种基于深度强化学习的边缘安全决策方法，实现观测数据的快速安全回传目标。

1 系统模型与待优化问题

本节首先对所关注的多层网络边缘决策模型进行介绍，然后对模型场景中存在的待优化数学问题进行描述总结。

1.1 系统模型

本文所考虑的多层卫星网络如图 1 所示，主要由低地球轨道（LEO, low earth orbit）卫星（又称低轨卫星）、中地球轨道（MEO, medium earth orbit）卫星（又称中轨卫星）以及地球静止轨道（GEO, geostationary earth orbit）卫星（又称高轨卫星）组成，其中实线代表层内卫星通信链路，虚线则代表层间卫星（地面）通信链路。

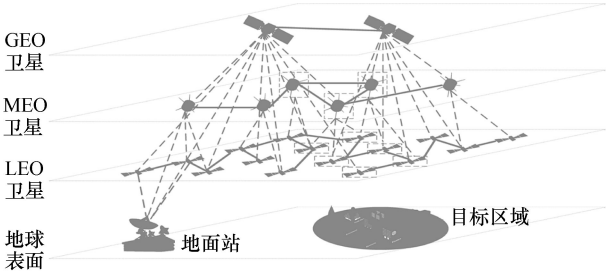


图 1 多层卫星网络

在图 1 所示的场景中，LEO 卫星节点负责观测侦察业务（如气象观测、地理侦测、情报侦察等），考虑到传统卫星网络空间电磁环境开放式的特点，本文设定观测卫星所获数据需要以加密的方式回传给地面站（地面控制中心）。需要说明的是，受限于国土资源在世界范围内的实际位置，与卫星网络进行信息交互的地面站一般数量有限、部署位置

较为集中，且观测卫星所获数据一般无法直接传送给地面站，在此情况下，发挥多层卫星网络的协同处理能力，有助于观测数据更加安全高效地实现回传。此外，考虑到卫星星座的多样性和卫星网络技术高速发展的特点，本文所设定的由高、中、低轨卫星组成的多层卫星网络模型可以较直观地拓展至其他多层（可以为两层、三层或更多层）卫星网络场景，例如多层均为低轨卫星网络的场景或高低轨道、中低轨道混合多层卫星网络的场景，本文将在第 3 节中提出能够对这些场景具备较优适应性的智能决策算法，相比于单独依靠低轨卫星网络星间链路完成观测数据的回传，多层异构卫星网络在边缘运算能力和回传路径方面具备更突出的灵活性，因而能够为数据计算和回传业务提供更丰富的选择。考虑到不同层卫星的运行高度、覆盖范围、相互可见性[13]以及运算存储能力，即低轨卫星灵活机动性最好，但其地面覆盖范围和运算处理能力最弱；中轨卫星具备较大的地面覆盖范围和一定的机动性，运算处理能力较强；地球同步轨道卫星则具备最大的覆盖范围（一般 3 颗空间合理部署的 GEO 卫星可以服务整个地球）和最强的运算处理能力，本文将覆盖低轨观测卫星的中轨卫星视为边缘场景中的雾节点，并由其中一颗 MEO 卫星担任雾运算处理中心，负责规划观测数据的压缩处理和安全加密所在卫星节点以及数据回传的网络选择。具体来说，低轨卫星所观测数据的运算处理和回传路径分别有三项选择，对于数据运算处理来说，可以选择直接由低轨卫星加密后回传给地面站进行，也可以传送给中轨或者高轨卫星进行压缩加密处理；对于回传路径的选择，可以选择仅由低轨卫星网络进行传送，也可以由中轨卫星或者高轨卫星在内的多层网络完成数据传输。

1.2　待优化问题

图 2 更细化清晰地展示了本文所关注的多层卫星网络边缘决策模型，在该场景中，低轨卫星主要负责对地观测任务，而低轨、中轨和高轨卫星网络均能够与地面站进行通信连接，鉴于中轨卫星具备居中的空间位置以及较强的运算通信能力，本文设定边缘场景中的中轨卫星节点为边缘（雾）节点，且其中一个节点为边缘中心。考虑到不同卫星的轨道高度和覆盖范围情况，本文设定边缘场景中存在一颗高轨卫星、Z 颗中轨卫星以及 N 颗低轨卫星，其中低轨卫星为对地观测卫星。为了确保观测数据边缘处理和回传过程的安全保密性，防止攻击者对数据进行窃取，本文设定观测数据在星间的传输过程均为加密状态，需要说明的是，不同的卫星节点由于存在不同的密码算法库而具备不同的安全加密能力，且通过一定的合理配置可进一步提升对观测数据的安全保障，本文当前假设边缘场景中的卫星节点已经预先完成了密码算法选择和密钥协商，有关于融合二者的进一步综合决策将作为未来的研究内容。由于低轨卫星的运算处理能力较弱，本文假定低轨卫星仅具备数据加解密的能力，而不具备数据压缩处理的能力，而场景中的中轨和高轨卫星节点则同时具备此 2 种能力。

设定场景中低轨卫星所获取的数据量为 $\alpha_n, n=1,2,\cdots,N$，各低轨卫星的数据加密速度为 $\gamma_n^L, n=1,2,\cdots,N$，高轨卫星与中轨卫星的数据加密速度分别为 γ^{G} 和 $\gamma_z^{\mathrm{M}}, z=1,2,\cdots,Z$，解密速度分别为 φ^{G} 和 $\varphi_z^{\mathrm{M}}, z=1,2,\cdots,Z$，二者的数据压缩处理速度分别为 λ^{G} 及 $\lambda_z^{\mathrm{M}}, z=1,2,\cdots,Z$，且压缩比均为 $\kappa, 0<\kappa<1$。此外，低轨卫星节点与高轨卫星节点间的信噪比（SNR, signal-to-noise ratio）为

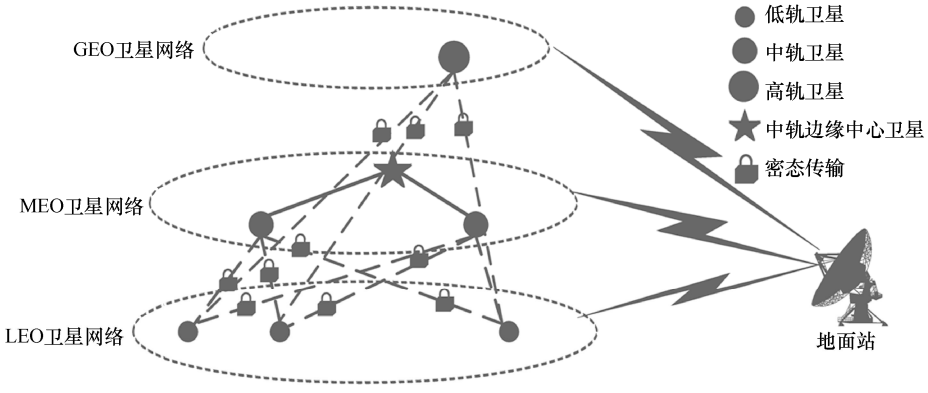


图 2　多层卫星网络边缘决策模型

$\eta_n^{\mathrm{LG}}, n=1,2,\cdots,N$，与中轨卫星节点间的信噪比为 $\eta_{n,z}^{\mathrm{LM}}, n=1,2,\cdots,N, z=1,2,\cdots,Z$，并设定所有信道连接的传输带宽均为 B MHz，设定低轨、中轨和高轨卫星网络内的传播时延分别为 $\beta^{\mathrm{L}},\beta^{\mathrm{M}},\beta^{\mathrm{G}}$，数据传输速度分别为 ϖ^{L}、ϖ^{M}、ϖ^{G}，低轨卫星至中轨卫星和高轨卫星的跨层传播时延分别为 $\psi_{n,z}^{\mathrm{LM}}$ 和 ψ_n^{LG}，$n=1,2,\cdots,N$, $z=1,2,\cdots,Z$。以上信息均可以通过预先计算、检测感知或是通信交互并最终由边缘中心节点获得。需要说明的是，受限于卫星的体积大小、发射入轨时间、载荷能力、所处状态等相关因素，不同轨道的不同卫星在对数据加解密的处理能力和支持的加解密算法方面可能存在不同，本文在系统模型和问题总结中确保相关数据在多层卫星网络的传输过程中始终保持密态，且将场景中各层卫星节点的加解密性能考虑在内，注意到，该设定兼容了中轨、高轨卫星使用安全强度更高的加密算法的情况，因此可以认定系统的安全性得到了保障。

本文设定加密后数据与加密前数据等长，且中轨和高轨卫星仅能够对明文数据进行压缩处理，对于低轨观测卫星的资源决策来说，不难总结并推算出存在以下几种情况。

1) 若低轨卫星节点 n 选择通过低轨卫星网络回传加密数据，则总时延为

$$\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\frac{\alpha_n}{\varpi^{\mathrm{L}}}+\beta^{\mathrm{L}} \tag{1}$$

式(1)中三项内容分别对应低轨卫星的明文数据加密时延、密文数据经由低轨卫星网络的传输时延和传播时延。

2) 若低轨卫星节点 n 选择中轨卫星节点 z 进行转发，并依靠中轨卫星网络将数据回传，则总时延为

$$\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_{n,z}^{\mathrm{LM}}+\frac{\alpha_n}{\mathrm{lb}(1+\eta_{n,z}^{\mathrm{LM}})}+\frac{\alpha_n}{\varpi^{\mathrm{M}}}+\beta^{\mathrm{M}} \tag{2}$$

式(2)中五项分别对应低轨卫星明文数据加密时延、低轨卫星至中轨卫星的传播时延、密文数据由低轨卫星至中轨卫星的发送时延、密文数据在中轨卫星网络的传输时延和传播时延。

3) 若低轨卫星节点 n 选择中轨卫星节点 z 进行数据处理，并依靠中轨卫星网络将数据回传，则总时延为

$$\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_{n,z}^{\mathrm{LM}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_{n,z}^{\mathrm{LM}})}+\frac{\alpha_n}{\varphi_z^{\mathrm{M}}}+\frac{\alpha_n}{\lambda_z^{\mathrm{M}}}+\frac{\kappa\alpha_n}{\gamma_z^{\mathrm{M}}}+\frac{\kappa\alpha_n}{\varpi^{\mathrm{M}}}+\beta^{\mathrm{M}} \tag{3}$$

式(3)中八项分别对应低轨卫星明文数据加密时延、低轨卫星至中轨卫星的传播时延、密文数据由低轨卫星至中轨卫星的发送时延、密文数据解密时延、明文数据压缩时延、明文压缩数据加密时延、密文压缩数据经由中轨卫星网络的传输时延和传播时延。

4) 若低轨卫星节点 n 选择通过高轨卫星节点进行数据回传，则总时延为

$$\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_n^{\mathrm{LG}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_n^{\mathrm{LG}})}+\frac{\alpha_n}{\varpi^{\mathrm{G}}}+\beta^{\mathrm{G}} \tag{4}$$

式(4)中五项分别对应低轨卫星明文数据加密时延、低轨卫星至高轨卫星的传播时延、密文数据由低轨卫星至高轨卫星的发送时延、密文数据在高轨卫星网络的传输时延和传播时延。

5) 若低轨卫星节点 n 选择通过高轨卫星节点进行数据解加密并回传，则总时延为

$$\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_n^{\mathrm{LG}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_n^{\mathrm{LG}})}+\frac{\alpha_n}{\varphi^{\mathrm{G}}}+\frac{\alpha_n}{\lambda^{\mathrm{G}}}+\frac{\kappa\alpha_n}{\gamma^{\mathrm{G}}}+\frac{\kappa\alpha_n}{\varpi^{\mathrm{G}}}+\beta^{\mathrm{G}} \tag{5}$$

式(5)中八项分别对应低轨卫星明文数据加密时延、低轨卫星至高轨卫星的传播时延、密文数据由低轨卫星至高轨卫星的发送时延、密文数据在高轨卫星的解密时延、明文数据在高轨卫星的压缩时延、明文压缩数据在高轨卫星的加密时延、密文压缩数据经由高轨卫星网络的传输时延和传播时延。

设定 $\mu_n^{\mathrm{L}},\mu_{n,z}^{\mathrm{M}},\mu_n^{\mathrm{G}}=0$ 或 1, $n=1,2,\cdots,N, z=1,2,\cdots,Z$ 分别表示低轨卫星节点 n 由不同层网络进行数据处理并回传的指示参数，$\xi_{n,z}^{\mathrm{M}},\xi_n^{\mathrm{G}}=0$ 或 $1, n=1,2,\cdots,N$, $z=1,2,\cdots,Z$ 表示低轨卫星节点 n 分别由中轨卫星节点 z 和高轨卫星节点所对应的卫星网络实现数据转发的指示参数，以回传平均时延最小化为目标，边缘中心节点的决策问题可以描述为

$$\min_{\substack{\mu_n^{\mathrm{L}},\mu_{n,z}^{\mathrm{M}},\mu_n^{\mathrm{G}}\\ \xi_{n,z}^{\mathrm{M}},\xi_n^{\mathrm{G}}}} \frac{1}{N}\Bigg(\sum_{n=1}^{N}\mu_n^{\mathrm{L}}t_n^{\mathrm{L}}+\xi_n^{\mathrm{G}}t_n^{\mathrm{G}}+\mu_n^{\mathrm{G}}(t_n^{\mathrm{G}c}+\bar{G}_t^c)+\sum_{z=1}^{Z}(\xi_{n,z}^{\mathrm{M}}t_{n,z}^{\mathrm{M}})+\sum_{z=1}^{Z}\left(\mu_{n,z}^{\mathrm{M}}(t_{n,z}^{\mathrm{M}c}+\bar{M}_{t,z}^c)\right)\Bigg) \tag{6a}$$

$$\text{s.t.}\quad t_n^{\mathrm{L}}=\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\frac{\alpha_n}{\varpi^{\mathrm{L}}}+\beta^{\mathrm{L}} \tag{6b}$$

$$t_n^{\mathrm{G}}=\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_n^{\mathrm{LG}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_n^{\mathrm{LG}})}+\frac{\alpha_n}{\varpi^{\mathrm{G}}}+\beta^{\mathrm{G}} \tag{6c}$$

$$t_n^{\mathrm{Gc}}=\frac{\alpha_n}{\gamma_n^{L}}+\psi_n^{\mathrm{LG}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_n^{\mathrm{LG}})}+\frac{\alpha_n}{\varphi^{\mathrm{G}}}+\frac{\alpha_n}{\lambda^{\mathrm{G}}}+\frac{\kappa\alpha_n}{\gamma^{\mathrm{G}}}+\frac{\kappa\alpha_n}{\varpi^{\mathrm{G}}}+\beta^{\mathrm{G}} \tag{6d}$$

$$t_{n,z}^{\mathrm{M}}=\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_{n,z}^{\mathrm{LM}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_{n,z}^{\mathrm{LM}})}+\frac{\alpha_n}{\varpi^{\mathrm{M}}}+\beta^{\mathrm{M}} \tag{6e}$$

$$t_{n,z}^{\mathrm{Mc}}=\frac{\alpha_n}{\gamma_n^{\mathrm{L}}}+\psi_{n,z}^{\mathrm{LM}}+\frac{\alpha_n}{B\mathrm{lb}(1+\eta_{n,z}^{\mathrm{LM}})}+\frac{\alpha_n}{\varphi_z^{\mathrm{M}}}+\frac{\alpha_n}{\lambda_z^{\mathrm{M}}}+\frac{\kappa\alpha_n}{\gamma_z^{\mathrm{M}}}+\frac{\kappa\alpha_n}{\varpi^{\mathrm{M}}}+\beta^{\mathrm{M}} \tag{6f}$$

$$\mu_n^{\mathrm{L}},\mu_{n,z}^{\mathrm{M}},\mu_n^{\mathrm{G}},\xi_{n,z}^{\mathrm{M}},\xi_n^{\mathrm{G}}=0\text{或}1\text{ 且}$$

$$\mu_n^{\mathrm{L}}+\mu_n^{\mathrm{G}}+\xi_n^{\mathrm{G}}+\sum_{z=1}^{Z}(\mu_{n,z}^{\mathrm{M}}+\xi_{n,z}^{\mathrm{M}})=1 \tag{6g}$$

$$\bar{G}_t^c=\frac{1}{\sum_{n=1}^{N}\mu_n^{\mathrm{G}}}\sum_{n=1}^{N}\mu_n^{\mathrm{G}}\left(\frac{\alpha_n}{\varphi^{\mathrm{G}}}+\frac{\alpha_n}{\lambda^{\mathrm{G}}}+\frac{\kappa\alpha_n}{\gamma^{\mathrm{G}}}\right) \tag{6h}$$

$$\bar{M}_{t,z}^c=\frac{1}{\sum_{n=1}^{N}\mu_{n,z}^{\mathrm{M}}}\sum_{n=1}^{N}\mu_{n,z}^{\mathrm{M}}\left(\frac{\alpha_n}{\varphi_z^{\mathrm{M}}}+\frac{\alpha_n}{\lambda_z^{\mathrm{M}}}+\frac{\kappa\alpha_n}{\gamma_z^{\mathrm{M}}}\right) \tag{6i}$$

$$n=1,2,\cdots,N,z=1,2,\cdots,Z \tag{6j}$$

其中，$\bar{M}_{t,z}^c$ 与 $\bar{G}_t^c$ 分别表示中轨卫星和高轨卫星节点与计算过程对应的平均排队时延。式(6g)表示一个低轨卫星节点仅能同时与一种卫星网络（回传方式）建立关联。考虑到中、高轨卫星一般配备了多组收发天线，且具备较强的缓存能力，本文在待优化问题中忽略中、高轨卫星的数据收发排队时延。以上待优化问题的求解目标对应于获得指示参数的值，由于存在大量的指示参数的结果组合，该问题是具有 NP-hard 性质的 0-1 规划问题，传统的基于优化的启发式方法较难求解，本文在第 3 节提出使用基于深度强化学习的方法对该问题进行求解。

2　准备工作

本节对本文所使用的强化学习的相关知识进行介绍，首先简介了强化学习的基本概念，并在此基础上描述了深度强化学习的知识内涵。

2.1　强化学习

在强化学习中，智能体通过与环境交互获得不同状态下所能采取动作的奖励值情况。详细来说，当所处时间点为 t、环境状态为 s_t 时，智能体采取了动作 a_t，然后智能体获得了一个数值奖励 r_t，且环境状态转化为 s_{t+1}。随着循环进行，智能体与环境持续交互得到了经验序列 $\{(s_t,a_t,r_t,s_{t+1}),\cdots\}$。进而，基于该经验序列，智能体能够对其策略 $\pi_t(s,a)$ 进行更新，该策略定义为状态为 $s_t=s$ 时采取动作 $a_t=a$ 的概率。在强化学习中，智能体的目标是最大化其未来能够接收的折扣奖励和，即 $R_t=\sum_{k=0}^{\infty}\vartheta^k r_{t+k}$，其中 $\vartheta\in[0,1]$ 为折扣率。

在众多的强化学习算法中，Q-learning 是较常用的一个，在该算法中，智能体与环境进行交互以便更新 Q 值，即在策略 π 前提和状态 s 条件下采取动作 a 所具备的效用值[14]

$$Q^{\pi}(s,a)=E\left[R_t\mid s_t=s,a_t=a\right] \tag{7}$$

定义最优的动作值函数为 $Q^*(s,a)=\max_{\pi}Q^{\pi}(s,a)$，依据贝尔曼最优性方程，$Q^*(s,a)$ 可以表示为

$$Q^*(s,a)=E\left[r_{t+1}+\vartheta\max_{a'}Q^*(s',a')\mid s_t=s,a_t=a\right] \tag{8}$$

其中，s' 是采取动作 a 后的新状态。Q-learning 的本质思路是最优的动作值函数 $Q^*(s,a)$ 可以通过与环境交互所获得的经验序列进行持续更新。令 $q(s_t,a_t)$ 为迭代过程中所估计的 Q 值，则 Q-learning 更新过程可表示为

$$q(s_t,a_t)\leftarrow q(s_t,a_t)+\xi\left(r_t+\vartheta\max_a q(s_{t+1},a_{t+1})-q(s_t,a_t)\right) \tag{9}$$

其中，$\xi\in[0,1]$ 为学习速率。为了学习到最优的 Q 值，智能体需要在探索与利用之间取得平衡，因为若完全按照当前未更新到位的策略进行动作的选择（即利用过程），其奖励情况极有可能无法达到最大，一个广为应用的平衡方法为使用 ε 贪心算法，该算法可以用以下概率描述所采取的动作

$$a=\begin{cases}\arg\max_a q(s,a), & \text{以}1-\varepsilon\text{的概率}\\ \text{随机动作}, & \text{以}\varepsilon\text{的概率}\end{cases} \tag{10}$$

2.2　深度强化学习

在本文所关注的模型中，状态和动作的数量均随着卫星节点的数量呈指数增加，毫无疑问，在这种情况下，强化学习的状态–动作空间将会变得异常庞大，传统的强化学习方法由于状态很少被遍历学习或者所需构建的记录表过大而导致效率低下，

鉴于此，本文考虑应用基于神经网络的深度强化学习技术作为动作值的近似器学习最优策略。需要说明的是，深度强化学习方法沿袭了传统强化学习方法的工作模式，但却依靠深度神经网络来代替传统方法的记录表，由于深度神经网络最大的优势在于能够通过简单的线性与非线性映射，实现对任意复杂参数关联关系的拟合，因此使用深度神经网络也能够较好地拟合不同状态与不同动作之间的关联关系，对于复杂或者庞大的状态空间或者动作空间，深度强化学习可以通过简单地增加神经元数量或者网络深度（层数）去应对，避免了传统方法记录表的复杂度出现超线性或者指数增加的困境。

具体来说，神经网络的输入为状态 s，而输出则为动作空间中每个动作的 Q 值，给定状态 s 和动作 a，输出 $q(s,a|\theta)$ 仅由深度神经网络的权重（即 θ）所决定，该权重通过学习过程的反向传播进行更新。特别地，本文将 3 种关键技术应用到所提的深度强化学习方法中：一是经验回放，智能体所获得的经验序列被存放于经验池中，进而从中随机取出小批量经验用于神经网络的学习过程，该方法打破了训练序列间的关联性，提升了训练的收敛速度；二是固定目标网络[15]，该方法固定了用于训练的主网，同时设定了一个目标网络用于目标训练值的更新，2 个网络的结构完全一致，且目标网络的权重依据主网的参数进行周期性的更新，这种方式也加快了收敛的速度；三是动作选择与平均的解耦[16]，目标网络生成 Q 值以便用于训练过程中计算损失情况，而主网的 Q 值则用于指导在下一状态下所应当采取的最优动作，通过将动作选择与评价过程相解耦，Q 值过拟合的危险被大为缓解。

3 基于强化学习的智能决策方法

本节对本文所提出的基于强化学习的多层卫星网络边缘安全决策（DQN-ESD, edge security decision based on deep Q network）方法进行详细介绍，需要说明的是，深度强化学习方法运行的主要支撑元素是状态、动作和奖励，而神经网络是深度强化学习方法的主要组成构件，以下分别对这些相关内容进行介绍。

3.1 方法参数设定

1) 状态设置

在本文所关注的待优化问题中，优化的目标为最小化平均回传时延，虽然影响时延的因素有很多，包含链路信噪比、不同节点加解密和压缩处理的计算速度、分层网络的传播时延和传输速度等，但对于深度强化学习网络来说，最直观、效率最高的用于方法判定 Q 值的参考因素是计算得到的各链路所对应的时延值，若所提方法状态空间由影响时延的因素组成，虽然主网络能够通过学习的过程掌握各因素与优化目标的对应关系，但该学习过程无疑会影响方法的收敛速度，进而影响方法的效能。鉴于此，本文设定状态空间 $S=(t_n^{\mathrm{L}},t_n^{\mathrm{G}},t_n^{\mathrm{Gc}},t_{n,z}^{\mathrm{M}},t_{n,z}^{\mathrm{Mc}})_{1\times N(2Z+3)}$，$\forall n,z$。其中 t_n^{L}、t_n^{G}、t_n^{Gc}、$t_{n,z}^{\mathrm{M}}$、$t_{n,z}^{\mathrm{Mc}}$ 的定义如式 6(b)~式 6(f)所示，分别对应低轨卫星 n 在边缘场景中所面临的卫星节点和网络环境条件下的时延情况。

应当补充说明的是，若依靠当前设定的时延作为状态，则状态空间的大小为 $N(2Z+3)$；若完全以前述各影响时延的参数为状态，则状态空间大小为 $2(2N+NZ+3Z+4)$。一方面，后者相比于前者来说，状态数量增加了 $N+6Z+8$，这无疑会明显影响训练的收敛速度；另一方面，更重要的是，若使用后者作为状态，强化学习所使用的深度神经网络则需要额外通过训练过程学习掌握各相关参数与时延或奖励之间复杂的非线性关系，进而达到较好的决策性能，而这会耗费很长的时间，严重拖慢算法的收敛速度。

2) 动作设置

所提方法的最终目的为由边缘（雾）中心节点通过合理的链路规划，达到低轨卫星节点回传数据的时延最小化，换言之，对于本文所提方法来说，即依靠状态情况，合理地对低轨卫星节点的回传链路进行选择，用数学语言描述动作空间，即 $A=(\mu_n^{\mathrm{L}},\mu_n^{\mathrm{G}},\mu_{n,z}^{\mathrm{M}},\xi_{n,z}^{\mathrm{M}},\xi_n^{\mathrm{G}})_{1\times N(2Z+3)}$，$\mu_n^{\mathrm{L}},\mu_{n,z}^{\mathrm{M}},\mu_n^{\mathrm{G}},\xi_{n,z}^{\mathrm{M}},\xi_n^{\mathrm{G}}=0$ 或 1，且 $\mu_n^{\mathrm{L}}+\mu_n^{\mathrm{G}}+\xi_n^{\mathrm{G}}+\sum_{z=1}^{Z}(\mu_{n,z}^{\mathrm{M}}+\xi_{n,z}^{\mathrm{M}})=1$，$\forall n,z$。

3) 奖励函数设置

与状态空间和动作空间设置过程所遵循的原则一样，奖励函数的设定寻求能够直接反映某状态情况下所提方法进行动作选择并执行后的效果，考虑到链路选择后的直观效果即低轨卫星节点数据回传的时延情况，本文基于回传平均时延对奖励函数进行设置，参照待优化问题的求解目标，具体为 $r=-\frac{1}{N}\sum_{n=1}^{N}\Big(\mu_n^{\mathrm{L}}t_n^{\mathrm{L}}+\xi_n^{\mathrm{G}}t_n^{\mathrm{G}}+\mu_n^{\mathrm{G}}\ (t_n^{\mathrm{Gc}}+\bar{G}_t^c)+\sum_{z=1}^{Z}(\xi_{n,z}^{\mathrm{M}}t_{n,z}^{\mathrm{M}})+\sum_{z=1}^{Z}\Big(\mu_{n,z}^{\mathrm{M}}(t_{n,z}^{\mathrm{Mc}}+\bar{M}_{t,z}^c)\Big)\Big)$，不难看出，当平均回传时延较大时，奖励值较小，此种设定有利于引导强化学习方

法在不同状态条件下选择具备低时延的动作，进而提升所提方法的性能。

4) 神经网络及其他设置

鉴于残差网络（ResNet）能够很好地避免传统网络结构容易出现的退化问题，本文所提方法中主网络和目标网络采用八层结构的 ResNet 对 Q 值进行估计，同时采用 Adam 优化器和 ReLU 激活函数[17]，网络的输入和输出则分别与状态空间和动作空间的维度相对应。

3.2 智能边缘决策方法

本文通过使用深度神经网络（即残差网络）对 Q 值（也称深度 Q 网络，即 DQN）进行估计，从数学上来讲，该估计过程可以描述为 $Q^*(s,a)\approx Q(s,a|\theta)$，其中的权重 θ 可以通过式(11)的过程进行更新。

$$L(\theta,\theta')=\mathrm{E}\left[\left(r(s,a)+\gamma\max_{a'}Q(s',a'|\theta')-Q(s,a|\theta)\right)^2\right] \tag{11}$$

其中，θ 与 θ' 分别为主网络和目标网络的权重。

最终，算法 1 描述了本文所提出的基于深度 Q 网络的边缘安全决策（DQN-ESD, edge security decision based on deep Q network）方法，该方法通过设置历史回放库 Γ 来随机进行小批量的网络训练，以避免网络陷入过拟合的状态，对于每个更新后的贪心门限值 ε，所提方法仅在学习阈值达到后开展网络的训练过程，这样能够向回放库存放足量的历史经验数据，同时也避免了频繁学习操作，此外，所提方法还设定了合理的目标网络更新频率 ϕ，防止主网络学习过程中的过拟合，增加了训练过程的收敛速度。

算法 1　DQN-ESD 方法

输入　状态空间 S，动作空间 A，折扣率 ϑ，学习率 ξ，目标网络的更新频率 ϕ

输出　主网络权重 θ，与状态所对应的动作

初始化　回放库 Γ，随机权重的主网络 $Q(s,a|\theta)$，相同权重和架构的目标网络 $Q(s,a|\theta')$，$\varepsilon,\varepsilon_{\text{delay}},\varepsilon_{\min}$，学习阈值 L_t，循环次数 I

while $\varepsilon>\varepsilon_{\min}$ do

　$\varepsilon\leftarrow\varepsilon\varepsilon_{\text{delay}}$

　for $i\leftarrow 1,\cdots,I$ do

　　随机产生 0~1 的数值，并执行式(5)；

　　执行所选动作 a，计算奖励 r，并将四元组 (s,a,r,s') 存入 Γ；

　　if $i>L_t$ then 从回放库 Γ 中随机取一批量经验，使用式(6)计算 $L(\theta,\theta')$，并更新权重 θ，$j=j+1$；

　　if $j \bmod \phi=0$ then 令 $\theta'\leftarrow\theta, s\leftarrow s'$；

　　end if

　end if

　end for

end while

3.3 方法泛化能力说明

需要补充说明的是，本文所提方法具备针对卫星数量变化的前向兼容性，详细来说，本文所提方法所训练的模型能够适用于比当前模型所对应卫星数量更少的场景，举例来说，若所训练模型对应的多层卫星配置（高轨卫星数–中轨卫星数–低轨卫星数）为 1-3-8，则该模型能够兼容（适用）1-2-8、1-3-7 等多层卫星场景，因为对于卫星数量较少的场景，本文所提方法可以通过自动地把相应位置参数进行填补来保持正常运行，若低轨卫星数量减少，则可以通过把待传数据量（或者时延参数）设置成 0 进行填补；若中轨或者高轨卫星数量减少，则可以通过把对应的时延参数调至较大，以确保本文所提方法的运行。由于所做改动仅涉及状态输入的简单调整，因此本文所提方法具备较好的前向兼容性，但本文所提方法不适用于卫星数量更多的场景，鉴于此，可以通过预先训练得到卫星数量较多的场景下的算法模型，以确保本文所提方法的泛化能力。此外，由于本文所提方法在应用时主要涉及状态和动作等相关参数的设定，因此其也能够兼容具备多层特征的卫星星座场景，而不需要严格要求高轨–中轨–低轨的卫星网络层次关系。

4 性能仿真与分析

4.1 仿真设置

为了验证本文所提 DQN-ESD 方法的性能，本节采用 Keras 作为深度强化学习的仿真平台。在仿真实验中，本节以多层卫星网络中的某一区域作为仿真对象，设定该区域低轨卫星数量为 8 颗，中轨卫星数量为 3 颗（其中一颗为边缘决策中心，不参与数据处理和转发业务），高轨卫星数量为 1 颗。需要说明的是，本文设定当前中、低轨卫星的数量只是为了验证和展现本文所提方法的性能，目前有较多区域内低轨卫星节点相对密集的系统，如 Starlink 系统、Kuiper 系统等，由于本文所提方法具备前向兼容性，因此其并不仅限于该数量配置。此

外，不失一般性地，仿真假定低轨卫星的星座为常见的 Walker 星座，考虑到 Walker 星座对于不同的纬度具备不同数量的可见卫星，这也与本文所关注的多层卫星网络边缘场景的特征以及本文所提方法的兼容性相吻合。鉴于本文所提方法的元素设定过程并未对卫星星座提出严格要求，因此其对于非 Walker 星座也具有一定的适用性。在这种参数设定下，不难算出，边缘决策中心所面临的动作空间高达 $7^8=5\,764\,801$ 个，若使用普通的遍历方法或强化学习方法，则计算量过大，且过于耗时。此外，设定深度强化学习过程的折扣因子为 0.9，ε 贪心算法的探索因子 $\varepsilon\in[0.005,0.900]$，且其衰减率为 0.995，学习速率 ξ 为 0.01，经验回放库的大小 Γ =500，且经验回放库小批次容量大小为 32，目标网络的更新频率 ϕ=500。需要额外说明的是，多层卫星网络是相对来说较复杂的网络系统，相关参数的取值受卫星节点的空间位置、综合能力、节点间相互关系影响较大，例如对于低轨卫星网络来说，当低轨卫星恰好在地面站上空时，其传播时延为理论上最小，而当其处于地面站所在位置的对立位置时，其传播时延很大，因此本文设定低轨卫星网络的传播时延从一定的范围中取值，且设定所考虑的场景具备快照的性质，即相关参数和节点与网络的逻辑关系在所认定的时长内保持恒定不变，且在仿真过程中，低轨卫星待传数据量、卫星的数据处理能力、链路信噪比等参数均在一定范围内随机取值，具体的仿真参数设置如表 1 所示，本节通过对大量的快照进行实验得到仿真结果。

表 1　　仿真参数设置

超参数	设置值
低轨卫星、中轨卫星、高轨卫星数量/颗	8、3、1
折扣因子 γ 、探索因子 ε	0.9、0.005～0.900
ε 衰减速率、学习率 ξ	0.995、0.01
经验回放库 Γ 、小批次容量	500、32
观测数据量 $\alpha_n, n=1,2,\cdots,N$ /MB	3～20
$\gamma_n^{\mathrm{L}}, n=1,2,\cdots,N$ / (KB·s^{-1})	400～1 200
γ^{G} , $\gamma_z^{\mathrm{M}}, z=1,2,\cdots,Z$	4 MB/s，1 500～3 600 KB/s
φ^{G} , $\varphi_z^{\mathrm{M}}, z=1,2,\cdots,Z$	3 MB/s，800～2 400 KB/s
λ^{G} , $\lambda_z^{\mathrm{M}}, z=1,2,\cdots,Z$	6 MB/s，3 000～5 500 KB/s
压缩比 κ 、带宽 B /MHz	0.4、2
信噪比 $\eta_n^{\mathrm{LG}}, n=1,2,\cdots,N$ /dB	5～20
$\eta_{n,z}^{\mathrm{LM}}, n=1,2,\cdots,N, z=1,2,\cdots,Z$ /dB	25～40
$\beta^{\mathrm{L}}, \beta^{\mathrm{M}}, \beta^{\mathrm{G}}$ /ms	10～400，1 000，1 500
$\psi_{n,z}^{\mathrm{LM}}, \psi_n^{\mathrm{LG}}, n=1,2,\cdots,N,\ z=1,2,\cdots,Z$ /ms	10～15，100～110
$\varpi^{\mathrm{L}}, \varpi^{\mathrm{M}}, \varpi^{\mathrm{G}}$ / (MB·s^{-1})	2，4，8

为了充分体现本文所提方法的优势性，本节共采用 4 种方法进行性能对比，介绍如下。

1) 最优边缘安全决策（O-ESD, optimal edge security decision）。通过在考虑的场景中遍历决策结果来找到最优解，该方法能够表征本文所提方法的性能，但由于复杂度很高（例如在仿真中，该方法针对每一快照需要通过遍历 7^8 个组合来得到最优解），在实际应用中几乎不具备可行性。

2) 随机边缘安全决策（R-ESD, random edge security decision）。通过令每一低轨观测卫星随机选择数据处理卫星节点和回传网络，得到回传时延性能。

3) 以信噪比参数为导向的边缘安全决策（S-ESD, SNR-edge security decision）。设定每一低轨卫星在中、高轨卫星节点中选择与其之间链路信噪比最高的节点进行数据处理和回传，得到回传时延性能。

4) 本文所提方法在“*”网络中的执行情况（DQN-ESD*）。“*”可以为“L”“M”“G”以及三者的混合，三者分别对应于低、中、高轨卫星网络，在仿真中，考虑到可选节点的多样性，设定该类方法主要包含 DQN-ESD$^{\mathrm{M}}$ 、DQN-ESD$^{\mathrm{LM}}$ 、DQN-ESD$^{\mathrm{LG}}$ 以及 DQN-ESD$^{\mathrm{MG}}$，此外，本文所提方法 DQN-ESD 等同于 DQN-ESD$^{\mathrm{LMG}}$ 。

4.2　性能与分析

本节首先对本文所提方法的收敛性能进行仿真验证，对于一个随机快照，设定低轨卫星网络传播时延 β^{L} 为 400 ms，本文所提方法的收敛过程如图 3 所示，其中，DQN-ESD$^{\mathrm{M}}$ 表示本文所提方法在仅有中轨卫星节点可供选择的情况，此种情况类似于地面通信网络的边缘计算场景。考虑到不同快照的状态以及所采用方法的性能差异可能很大，本节在仿真结果的呈现中使用归一化时延来进行性能表征。

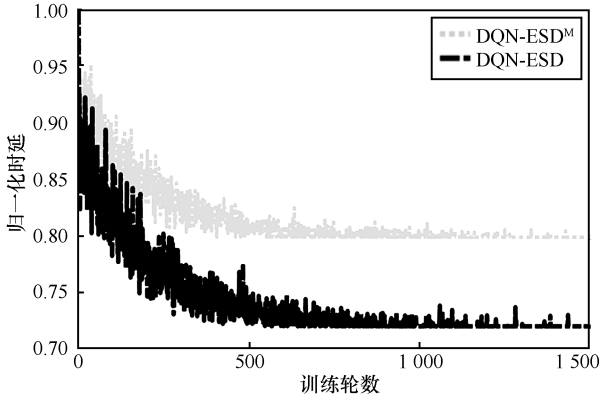


图 3　本文所提方法的训练收敛过程

从图 3 中可以清晰地看出，2 种方法的归一化时延均随训练轮数的增加而逐步收敛，本文所提方法在 500 轮的训练时即可基本收敛，此外，考虑到单独依靠低轨卫星网络和高轨卫星网络进行数据回传几乎不存在迭代收敛过程，二者在此情况下的归一化时延分别为 0.825 和 0.916，可以看出本文所提方法的性能明显优于 3 种单一的网络，这是因为后者可供选择的卫星节点更少，这也进一步印证了多层卫星网络相比于单层卫星网络在数据处理和回传方面具有优势。

此外，考虑到本文所提方法具备前向兼容性，训练卫星数量更多的模型有助于提升方法的泛化能力，为了进一步体现卫星数量不同情况下方法的性能，本节对卫星数量不同情况下各方法针对随机快照的收敛性能进行仿真，结果如图 4 所示。图 4 中展示了 3 种卫星数量情况，分别为情况 A（高轨、中轨和低轨卫星数量分别为 1、2、12）、情况 B（高轨、中轨和低轨卫星数量分别为 1、2、10）和情况 C（高轨、中轨和低轨卫星数量分别为 1、2、8）。需要说明的是，由于卫星数量的变化直接影响了快照（状态）的参数，因此各时延值不具备互相对比的可行性，但从图 4 中可以直观地看出，1) 本文所提方法针对不同卫星数量的情况均表现出了收敛趋势，这表明本文所提方法能够很好地适用于更加复杂的多层卫星网络场景；2) 随着卫星数量的增加，本文所提方法达到收敛所需要的训练次数明显增加，情况 A、情况 B 和情况 C 初步收敛的轮数分别为 500 轮、2 800 轮和 6 000 轮，这是合理的，卫星数量的增加大幅提升了方法动作空间的大小（即指示参数组合的数量或者解空间的大小），情况 B 和情况 A 的解空间分别高达 7^{10} 和 7^{12}，如此庞大的解空间对于遍历等方法来说几乎是不可行的，这进一步体现了本文所提方法的优势。

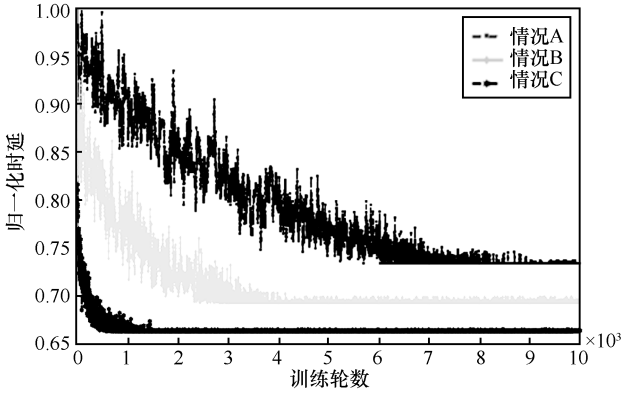


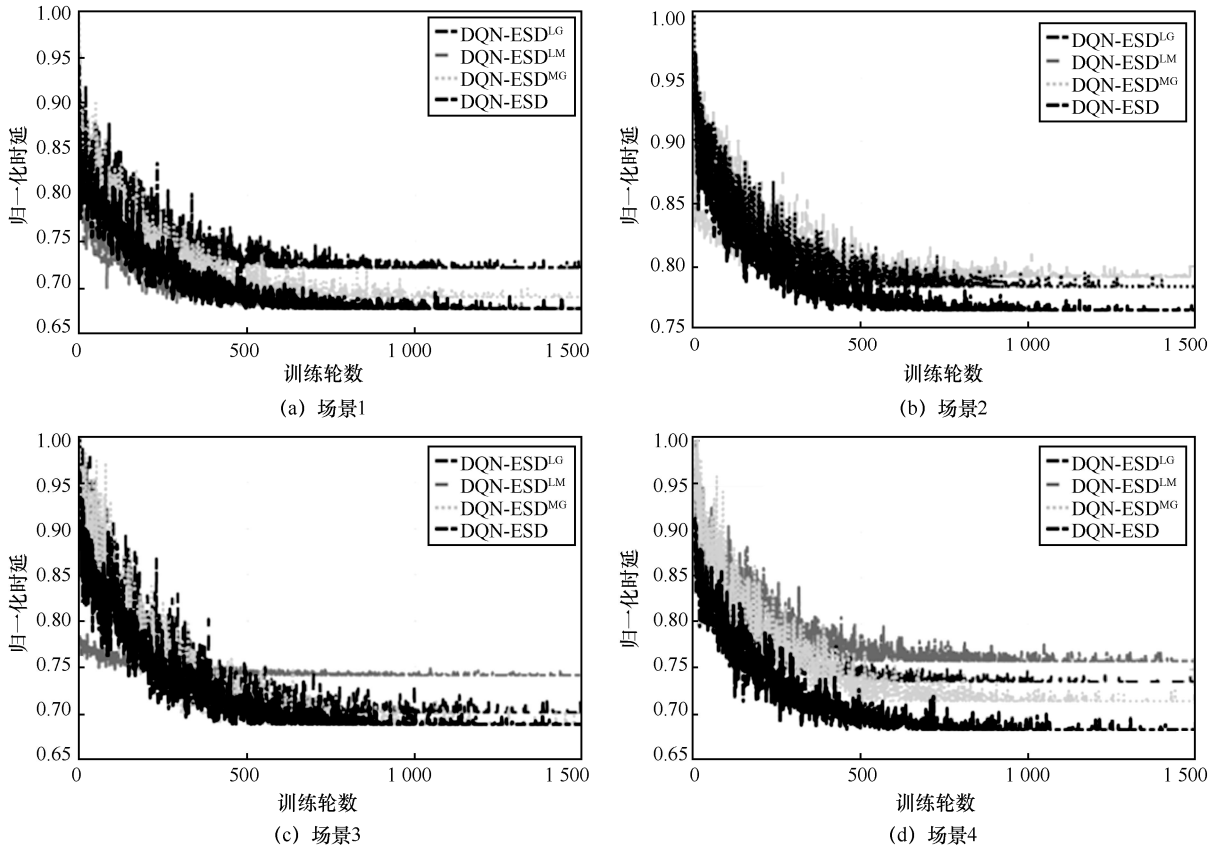
图 4　不同卫星数量情况下本文所提方法的收敛过程

其次，本节随机选取了 4 个不同的快照（其中均设定 β^{L}=400 ms）对本文所提方法在不同网络构型条件下的性能进行了对比，仿真结果如图 5 所示。从图 5 中可以看出：1) 本文所提方法在所有 4 种不同构型条件下均具有良好的收敛性能，基本上能够在 500 轮时完成收敛；2) 虽然在不同快照下本文所提方法具备不同的性能表现，但本文所提方法在高、中、低轨多层网络下的时延性能最优，这再一次印证了多层卫星节点为低轨卫星数据的处理和回传提供了更丰富的选择；3) 在部分快照（图 5(b)和图 5(c)）中，DQN-ESD$^{\mathrm{LG}}$ 的起始性能非常优异，但随着训练的进行，其收敛后的性能却相对较差，这是因为 DQN-ESD$^{\mathrm{LG}}$ 所对应的网络构型为低轨卫星网络协同高轨卫星网络，任一低轨卫星的链路选择仅有 2 个，即低轨卫星网络或者单一高轨卫星节点所在的高轨卫星网络，这限制了本文所提方法的性能。

接着，本文对不同低轨卫星网络传播时延情况下各方法的归一化时延性能进行了验证，结果如表 2 所示。从表 2 可以看出，当低轨卫星传播时延非常小时，本文所提方法相较 DQN-ESD$^{\mathrm{MG}}$ 收敛时延相差最大，这是由于在这种情况下，各低轨卫星会优先选择直接回传地面，而随着低轨传播时延的逐渐增大，低轨卫星直接回传所需时延增大，此时，本文所提方法能够自适应地选择中轨或高轨卫星进行数据回传，而这使该方法整体的时延性能依然保持最优。DQN-ESD$^{\mathrm{MG}}$ 由于不选择低轨卫星网络进行数据回传，因而其时延性能不受低轨卫星网络传播时延的影响。

与 DQN-ESD$^{\mathrm{LM}}$ 相比，本文所提方法始终更理想，随着低轨传播时延的增加，本文所提方法的时延与 DQN-ESD$^{\mathrm{LM}}$ 的时延的差值也逐步增大，这是因为本文所提方法存在高轨卫星网络选项，当经由高轨卫星处理或转发的时延更短时，本文所提方法能够智能地选择高轨卫星网络。同时可以观察到，当低轨卫星网络传播时延大到一定程度时，本文所提方法时延与 DQN-ESD$^{\mathrm{MG}}$ 时延趋于相同，这是因为随着低轨传播时延的增大，各低轨卫星不再选择通过低轨卫星网络直接回传，而是都经由中轨或高轨卫星进行数据回传。

最后，本节采用测试集对本文所提方法与对比方法的性能进行仿真验证，图 6 展示了各方法在 20 个



(a) 场景1　(b) 场景2　(c) 场景3　(d) 场景4

图 5　不同快照条件下 4 种方法在不同网络构型条件下的性能对比

表 2　不同低轨卫星网络传播时延下各方法的归一化时延性能

方法	10 ms	20 ms	30 ms	40 ms	50 ms	60 ms
DQN-ESD	0.793	0.834	0.861	0.888	0.901	0.901
DQN-ESD^{MG}	0.901	0.901	0.901	0.901	0.901	0.901
DQN-ESD^{LM}	0.811	0.855	0.895	0.924	0.951	0.958
DQN-ESD^{LG}	0.858	0.926	0.994	1.000	1.000	1.000

随机多层卫星网络边缘快照状态下的性能结果。由于随机进行节点和网络的选择，所做决策不具备收敛特点，随机边缘安全决策 R-ESD 方法的性能在所有 4 种方法中表现最差，相比较而言，由信噪比参数为导向的边缘安全决策 S-ESD 方法则表现出了明显更优的性能，因为链路的信噪比特性一般能够在很大程度上影响回传链路的时延性能。与此同时，从图 6 中可以看出，本文所提方法在时延性能上与最优 O-ESD 方法几乎完全一致，这表明本文所提方法通过一定的模型训练，已经具备了较优的自主决策能力。

需要补充说明的是，本节主要以观测数据回传的时延性能对本文所提方法和对比方法进行了仿真和呈现，而有关数据的安全方面则未有结果进行直接展现，这是因为多层卫星网络场景中各节点的密码算法选定和加解密处理性能已在系统模型与仿真参数中进行了设定，这意味着观测数据在边缘场景和整个网络中均以密文形式进行传输。本文所提方法的主要能力体现在确保观测数据以安全密态形式回传的前提下，通过对场景中节点状态的合理把握，智能地提供具备低时延特点的边缘决策动作。有关进一步地展现本文所构建模型或所提方法在回传数据安全强度方面的量化性能，由于需要更加全面地考虑场景中各节点密码算法库维度的信息，笔者将此工作留待下一步进行针对性解决。

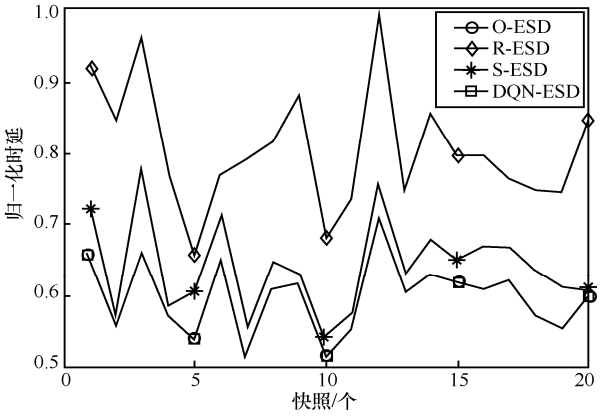


图 6　4 种方法在测试集快照上的时延性能对比

5　结束语

本文关注于高、中、低轨多层卫星网络中的边缘安全决策场景，针对场景中为低轨卫星进行多层卫星节点的链路选择问题，提出一种基于深度强化学习的数据压缩与加密回传决策方法。通过结合场景合理地设计方法的状态、动作、奖励以及训练网络等相关参数，所提方法能够以低传输时延为目标进行智能的边缘决策，大量的仿真表明，所提方法相比于多个对比方法具备明显较优的性能。

参考文献：

[1] 王丽娜, 王兵, 周贤伟, 等. 卫星通信系统[M]. 北京: 国防工业出版社, 2006.
WANG L N, WANG B, ZHOU X W, et al. Satellite communication system[M]. Beijing: National Defense Industry Press, 2006.

[2] YOU X H, WANG C X, HUANG J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts[J]. Science China Information Sciences, 2020, 64(1): 1-74.

[3] TATARIA H, SHAFI M, MOLISCH A F, et al. 6G wireless systems: vision, requirements, challenges, insights, and opportunities[J]. Proceedings of the IEEE, 2021, 109(7): 1166-1199.

[4] ZUO P L, WANG C, YAO Z, et al. An intelligent routing algorithm for LEO satellites based on deep reinforcement learning[C]//Proceedings of 2021 IEEE 94th Vehicular Technology Conference. Piscataway: IEEE Press, 2021: 1-5.

[5] DI B Y, ZHANG H L, SONG L Y, et al. Ultra-dense LEO: integrating terrestrial-satellite networks into 5G and beyond for data offloading[J]. IEEE Transactions on Wireless Communications, 2019, 18(1): 47-62.

[6] 夏士超, 姚枝秀, 鲜永菊, 等. 移动边缘计算中分布式异构任务卸载算法[J]. 电子与信息学报, 2020, 42(12): 2891-2898.
XIA S C, YAO Z X, XIAN Y J, et al. A distributed heterogeneous task offloading methodology for mobile edge computing[J]. Journal of Electronics & Information Technology, 2020, 42(12): 2891-2898.

[7] 钟磊. 低轨星座通信网络边缘计算架构研究[D]. 成都: 电子科技大学, 2020.
ZHONG L. Research on edge computing architecture of LEO constellation communication network[D]. Chengdu: University of Electronic Science and Technology of China, 2020.

[8] 王元君. 星地混合网络中的计算资源分配和负载均衡[D]. 北京: 北京邮电大学, 2020.
WANG Y J. Computing resource allocation and load balancing in hybrid satellite-terrestrial network[D]. Beijing: Beijing University of Posts and Telecommunications, 2020.

[9] DING C F, WANG J B, ZHANG H, et al. Joint optimization of transmission and computation resources for satellite and high altitude platform assisted edge computing[J]. IEEE Transactions on Wireless Communications, 2022, 21(2): 1362-1377.

[10] ZHOU D, SHENG M, WANG Y X, et al. Machine learning-based resource allocation in satellite networks supporting Internet of remote things[J]. IEEE Transactions on Wireless Communications, 2021, 20(10): 6606-6621.

[11] JIANG C X, ZHU X M. Reinforcement learning based capacity management in multi-layer satellite networks[J]. IEEE Transactions on Wireless Communications, 2020, 19(7): 4685-4699.

[12] 闵士权, 刘光明, 陈兵,等. 天地一体化信息网络[M]. 北京: 电子工业出版社, 2020.
MIN S Q, LIU G M , CHEN B, et al. Space-ground integrated information network[M]. Beijing: Electronic Industry Press, 2020.

[13] 黄娟. 基于MATLAB/STK的卫星通信场景仿真设计与实现[D]. 合肥: 安徽大学, 2016.
HUANG J. The design and implementation of simulation for satellite communication scene based on MATLAB/STK[D]. Hefei: Anhui University, 2016.

[14] GU B, ZHANG X, LIN Z Q, et al. Deep multiagent reinforcement-learning-based resource allocation for Internet of controllable things[J]. IEEE Internet of Things Journal, 2021, 8(5): 3066-3074.

[15] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning[J]. arXiv Preprint, arXiv: 1312.5602, 2013.

[16] HASSELT H V, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2016: 2094-2100.

[17] YU Y D, WANG T T, LIEW S C. Deep-reinforcement learning multiple access for heterogeneous wireless networks[C]//Proceedings of IEEE Journal on Selected Areas in Communications. Piscataway: IEEE Press, 2019: 1277-1290.

[作者简介]

左珮良（1991− ），男，山东烟台人，博士，北京电子科技学院讲师，主要研究方向为卫星通信、认知无线电、物联网、信息安全、软件定义网络。

侯少龙（1999− ），男，山西原平人，西安电子科技大学硕士生，主要研究方向为卫星通信、人工智能。

郭超（1987− ），女，江西九江人，北京电子科技学院讲师，主要研究方向为卫星通信、应急通信、传输控制、网络负载均衡、信息安全、物联网。

蒋华（1962− ），男，山西大同人，北京电子科技学院教授，主要研究方向为通信安全、应急通信、物联网、下一代网络。

王文博（1965− ），男，河北安国人，博士，北京邮电大学教授，主要研究方向为无线通信、3G/4G/5G/6G 通信、卫星通信、认知无线电、物联网、信息安全、软件定义网络。