

专题：低空智联网

面向卫星无人机协同网络的NOMA辅助流水线并行推理研究

杨萌琪[1]，吴义曼[1]，周任飞[2]，吴鹏[3]，潘鹏[1]，郑长亮[1]
（1. 杭州电子科技大学通信工程学院，浙江 杭州 310018；
2. 中国人民解放军31007部队，北京 100079；
3. 中国人民解放军军事科学院国防科技创新研究院，北京 100071）

摘 要：低轨卫星与无人机协同推理为实时智能感知任务提供了新的解决途径。针对现有协同计算方法中任务卸载或非流水线模型切分易造成通信与计算相互等待、资源利用率较低的问题，提出一种结合模型切分与流水线并行的卫星无人机协同推理方案。该方案对深度神经网络进行切分，其中头部和尾部部署于无人机侧，计算密集的中间子网络部署于卫星侧，并通过中间特征传递实现完整推理功能。在此基础上，所提方案引入微批次流水线并行机制与非正交多址接入（non-orthogonal multiple access，NOMA）技术，使不同微批次的计算和传输阶段并行执行。进一步地，以端到端协同推理时延最小化为目标，联合优化神经网络部署结构、微批次划分粒度及上下行传输功率。仿真结果表明，所提方案能够有效降低协同推理时延，提升系统资源利用率。

关键词：模型切分；分布式推理；流水线并行；NOMA
中图分类号：TN927
文献标志码：A
doi: 10.11959/j.issn.1000−0801.DXKX260294

NOMA-assisted pipeline parallel inference for satellite-UAV collaborative networks

Yang Mengqi[1], Wu Yiman[1], Zhou Renfei[2], Wu Peng[3], Pan Peng[1], Zheng Changliang[1]
1. School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
2. Unit 31007, People's Liberation Army, Beijing 100079, China
3. National Innovation Institute of Defense Technology, Academy of Military Science, Beijing 100071, China

Abstract: Collaborative inference between low-earth-orbit (LEO) satellites and unmanned aerial vehicles (UAVs) provides a new solution for real-time intelligent sensing tasks. Existing collaborative computing methods based on task offloading or non-pipelined model partitioning tend to cause mutual waiting between communication and computation, resulting in low resource efficiency. To address this problem, a satellite-UAV collaborative inference scheme combining model splitting and pipeline parallelism was proposed. In this scheme, the deep neural network was parti-

收稿日期：2026−05−09；修回日期：2026−05−31
通信作者：吴鹏，hitwupeng@outlook.com
基金项目：国家自然科学基金资助项目（No. 62401612）
Foundation Item: The National Natural Science Foundation of China (No. 62401612)

tioned, with the head and tail subnetworks deployed on the UAVs and the computation-intensive intermediate subnetwork deployed on the satellite. Complete inference was achieved through intermediate feature transmission. On this basis, the micro-batch pipeline parallel mechanism and non-orthogonal transmission were employed, allowing the computation and transmission stages of different micro-batches to be executed in parallel. Furthermore, with the objective of minimizing the end-to-end collaborative inference delay, the neural network deployment structure, micro-batch partitioning granularity, and uplink and downlink transmission power were jointly optimized. Simulation results show that the proposed scheme can effectively reduce the collaborative inference delay and improve system resource utilization.

Key words: model splitting, distributed inference, pipeline parallelism, non-orthogonal multiple access

0　引言

随着空天地一体化网络的发展，低轨卫星与无人机节点正逐渐从单一通信或感知平台向通信、计算与智能处理融合的网络平台演进[1-3]，在人工智能技术的支持下具备目标检测、图像分析等重要能力[4-5]。然而，无人机的计算、存储资源显著受限，难以独立支撑复杂模型的部署与持续计算。因此，无人机与卫星协同计算成为一种极具潜力的解决方案[6-7]。

作为一种新兴的协同计算范式，计算模型切分技术将神经网络的不同组成部分分别部署在卫星和无人机上，通过交换模型推理的中间特征值或训练过程的中间梯度来实现节点间的协同计算。由于该范式不交换原始数据，相比传统计算任务卸载的协同计算方法具有保护数据隐私的关键优势[8-11]。文献[12]通过模块化切分实现卫星辅助的多无人机协同推理。文献[13]优化了模型层在卫星与地面节点间的部署以降低能耗。文献[14]利用无人机的中继与计算能力构建空天地多跳协同路径。文献[15]则将模型切分扩展至多星场景，借助星间链路实现分布式协同推理。

虽然模型切分具有提升数据安全性、缓解单节点模型缓存与计算压力的优势，但是也存在亟待解决的关键问题。由于神经网络模型通常具有较强的结构性，模型切分后各个模块的计算过程通常具有严格的处理顺序。同时，协作节点间的数据推理、转发、继续推理过程也存在因果性。因此，模型切分这一分布式计算范式展现出较强的顺序执行特性，在不同设备之间以及同一任务的计算与通信阶段之间，均存在显著的相互等待时间，限制了系统的工作效率。

作为上述问题的潜在解决方案，流水线并行推理机制是一种在模块化顺序推理框架下提升系统计算效率的关键技术手段。传统流水线并行推理主要面向具有稳定有线连接的数据中心场景，其核心思想是使不同的神经网络模块并行处理不同的数据单元，通过跨计算单元的并行计算减少模块间等待。针对数据中心场景，现有研究将顺序流水线[16]扩展到图结构流水线[17]，并结合调度优化提升并行效率[18]。文献[19]在模型切分基础上通过将流水线机制引入地面无线边端协同计算网络，实现多终端中间特征的分阶段执行。在此基础上，文献[20]进一步通过优化模型切分与调度机制来缓解流水线执行中的气泡问题。

将流水线并行推理机制引入空天地一体化网络的相关工作尚处于起步阶段，现有研究主要面向卫星间的流水线协同推理展开[21-24]。文献[21]提出在星群层进行任务分配和模型切分，并将部分任务卸载至地面站执行，从而降低任务响应时间。文献[22]针对遥感目标检测任务，构建基于链式星座的流水线分布式推理架构，使遥感

数据在向地面传输过程中同步完成分阶段推理。文献[23]将大语言模型划分为多个子模型部署在不同卫星上，并利用流水线机制重叠子模型计算与中间激活值传输，降低端到端推理时延。文献[24]提出面向低轨星座的PipeSIA流水线空间推理方法，通过联合优化流水线中的节点顺序与层分配策略来适应星间拓扑的动态性。

尽管上述工作为空天地一体化网络的节点间协同推理奠定了重要研究基础，但面向无人机与卫星协同推理的流水线并行机制尚未得到充分研究，仍存在以下研究空白与关键挑战。

（1）现有流水线协同推理研究[21-24]主要面向数据中心或卫星间协作场景，通常假设为高速点对点通信链路。而在无人机与低轨卫星协同推理场景，通信过程依赖共享的星地无线信道，不同无人机之间存在通信资源竞争，使流水线性能不仅受计算资源影响，还受到无线传输能力的制约。

（2）现有无人机卫星协同推理研究[12-15]尚未充分考虑流水线并行机制，且通常以完整批次或任务为基本执行单元，缺乏微批次级别的细粒度划分与调度。同时，这些研究多采用正交多址接入方式，不同设备需串行占用信道资源。这些因素造成显著的通信与计算间等待。

针对上述问题，本文面向无人机与卫星协同推理，提出一种非正交多址接入（non-orthogonal multiple access，NOMA）辅助的微批次流水线并行推理方案。区别于现有流水线推理研究多面向高速点对点通信链路，本文考虑多无人机共享星地无线信道下的通信资源竞争，将NOMA传输机制纳入流水线执行过程，使无线传输阶段与计算阶段共同构成端到端流水线。基于此，本文联合优化模型切分、微批次划分和NOMA功率分配，以降低协同推理端到端时延。本文主要贡献总结如下3点。

（1）设计了卫星与无人机流水线协同推理方案，将神经网络以U型结构切分部署在无人机和卫星侧，其中，网络的头、尾部分部署在无人机侧，计算密集的中间部分部署在卫星侧，在充分利用卫星计算能力的同时保护无人机数据隐私。同时，通过将推理的输入数据划分为多个微批次，实现不同通信、计算阶段间的并行处理，提升端到端推理效率。

（2）针对多无人机推理任务的并发特性，引入NOMA通信机制，使多个无人机能够在同一时频资源上并行传输中间特征值，利用功率域复用与串行干扰消除（successive interference cancellation，SIC）技术提高频谱利用率，缩短通信阶段持续时间，缓解通信瓶颈对流水线性能的限制。

（3）建模端到端推理时延最小化问题，对微批次划分粒度、模型切分位置、NOMA功率分配进行联合优化。由于该优化问题属于混合整数非凸规划，本文设计了基于交替优化（alternating optimization，AO）的联合求解算法。仿真结果表明，合理优化网络参数能够显著提升流水线各阶段之间的平衡性，从而最大化流水线的并行程度、降低系统整体推理时延。

1　系统模型与流水线协同推理方案

所考虑的协同智能计算系统由多架无人机和一个低轨卫星节点组成，其中，无人机集合为$\mathcal{U}=\{1,2,\cdots,I\}$，$I$表示无人机数量。每架无人机采集本地图像数据并利用智能模型进行推理，识别图像内容。同时，无人机会将部分计算任务卸载到低轨卫星，由算力更强的卫星节点提供协同计算服务。

本文提出一种卫星和无人机间U型切分网络的微批次推理方案，系统模型如图1所示。总体而言，无人机和卫星节点共同维护一个神经网络，其中卫星负责承载网络中间层，无人机则承载神经网络的头部和尾部。采用U型切分结构的

原因在于，该结构能够同时满足数据隐私保护、卫星算力利用和无人机本地任务闭环三方面需求。首先，头部子网络部署在无人机侧，使无人机无须上传原始图像数据，而是上传中间激活值，从而降低原始数据暴露风险。其次，神经网络中间层通常包含较高的计算负载，将该部分部署在卫星侧能够利用卫星较强的计算能力，缓解无人机端算力受限问题。最后，尾部子网络仍部署在无人机侧，使最终结果在无人机本地生成，有利于后续目标跟踪等的本地闭环操作。

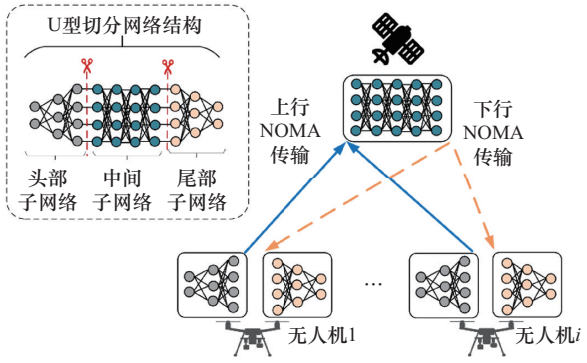

图1　系统模型

在推理过程中，每架无人机将本地图像数据的推理批量先拆分为微批次，以微批次为单位进行推理计算。整个协同推理过程依次执行以下5个步骤：（1）各无人机并行执行本地头部子网络推理；（2）各无人机将头部子网络输出的中间特征值通过上行NOMA的方式传输至低轨卫星；（3）卫星执行中间子网络的推理；（4）卫星将中间输出特征值通过下行NOMA的方式回传给对应无人机；（5）各无人机本地执行尾部子网络推理并输出结果。可以看出，整个流程包括3个计算过程和2个传输过程。流水线通信与计算过程如图2所示，这些过程以流水线方式并行执行，提高系统推理效率。后文将具体展开介绍计算过程、传输过程以及流水线流程的设计与建模。

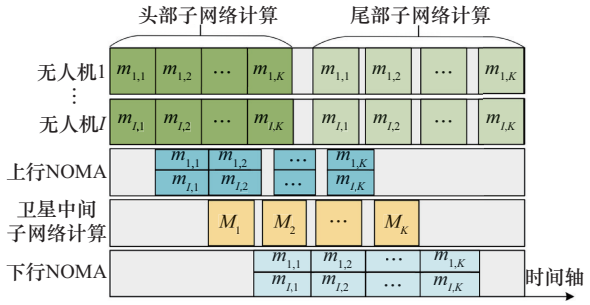

图2　流水线通信与计算过程

注：$m_{i,k}$表示无人机i第k个微批次，M_k表示卫星处的聚合批次

1.1　计算模型

对于网络中的任一无人机i，假设B_i为该无人机一轮推理中需要处理的批量数据大小。为提高通信与计算之间的重叠度，每架无人机将批量数据均匀划分为K个微批次。相应地，无人机i的每个微批次大小可以计算为：

$$b_i = B_i / K \tag{1}$$

针对无人机和卫星共同维护的神经网络，假设该神经网络包含L个顺序层，每个层的索引记为$l \in \{1, 2, \cdots, L\}$。为了方便表示无人机和卫星的神经网络承载情况，引入s和s'表示网络层的两个切分点，且满足$1 \leqslant s < s' < L$。相应地，第1层至第s层构成部署在无人机侧的头部子网络，第$s+1$层至第s'层构成部署在卫星侧的中间子网络，第$s'+1$层至第L层构成部署在无人机侧的尾部子网络。

令c_l表示第l层处理一个数据样本所需的计算量。因此，对于给定切分点(s, s')，无人机i处理一个微批次的头部子网络计算量可以写为：

$$C_i^h = b_i \sum_{l=1}^{s} c_l \tag{2}$$

当各架无人机并行完成头部子网络的微批次本地推理后，通过上行NOMA将输出的特征值发送给卫星。为了充分利用卫星节点更强的存储和并行计算能力，卫星将所有无人机上传的一轮微批次中间特征值沿批量维度进行拼接，形成一个聚合的推理批量。该批量大小可以表示为：

$$b_{\mathrm{ag}}=\sum_{i\in\mathcal{U}}b_i \tag{3}$$

因此，无人机上传一轮微批次后，卫星侧中间子网络的计算量为：

$$C^m=b_{\mathrm{ag}}\sum_{l=s+1}^{s'}c_l \tag{4}$$

当卫星完成中间子网络推理后，通过下行NOMA将各架无人机所需的中间输出特征值同时反馈给各无人机。各无人机通过尾部子网络完成计算，得到图像识别结果。相应地，无人机i处理一个微批次的尾部子网络推理所需的计算量为：

$$C_i^t=b_i\sum_{l=s'+1}^{L}c_l \tag{5}$$

假设无人机i的计算能力为f_i^u，卫星的计算能力为f^s，各子网络的计算时延可以分别写为：

$$T_i^h=\frac{C_i^h}{f_i^u},T^m=\frac{C^m}{f^s},T_i^t=\frac{C_i^t}{f_i^u} \tag{6}$$

其中，T_i^h、T^m和T_i^t分别表示无人机i的头部子网络计算时延、卫星侧共享中间子网络计算时延和无人机i的尾部子网络计算时延。

1.2 通信模型

在所提方案下，每个微批次协同推理过程均经历一次上行传输和一次下行传输。为提高传输并行度，上行和下行通信均采用NOMA方式。

（1）上行NOMA传输模型：在第k轮微批次上行传输中，各无人机通过NOMA方式上传其第k个微批次的头部子网络输出特征值。需要说明的是，本文主要关注给定信道状态下的模型切分、NOMA传输资源分配与微批次流水线调度联合优化问题，采用准静态信道假设，即在一个调度周期或一个推理批次执行周期内，信道状态信息可被估计并近似保持不变。当信道状态发生明显变化时，系统可在后续调度周期内基于更新后的信道状态重新计算传输速率、功率分配和流水线调度方案。结合物理层信道估计误差补偿、抗多普勒、SIC误差传播补偿的跨层流水线鲁棒调度问题将作为后续研究方向。不失一般性地，假设上行信道排序为$h_{1,k}^u<h_{2,k}^u<\cdots<h_{I,k}^u$，其中，$h_{i,k}^u$表示无人机$i$到卫星的上行信道增益。根据NOMA原理，卫星按照从强用户到弱用户的顺序执行SIC。相应地，无人机i的上行传输速率可写为：

$$R_{i,k}^u=W^u\mathrm{lb}\left(1+\frac{p_{i,k}^u h_{i,k}^u}{\sum_{j=1}^{i-1}p_{j,k}^u h_{j,k}^u+\sigma^2}\right) \tag{7}$$

其中，W^u表示上行信道带宽，$p_{i,k}^u$表示无人机i在第k轮微批次推理中的上行发射功率，σ^2表示噪声功率。

为了便于表达，用$a^u(s)$表示每个样本的头部子网络输出特征值大小，用$a^d(s,s')$表示每个样本的中间子网络输出特征值大小。相应地，对于无人机i，一个微批次的上行和下行传输数据量可分别计算为：

$$D_i^u=a^u(s)b_i,\ D_i^d=a^d(s,s')b_i \tag{8}$$

由于卫星需要在接收各无人机的特征值后进行聚合推理，上行传输时延由所有无人机所需传输时间的最大值决定，即：

$$\tau_k^u=\max_{i\in\mathcal{U}}\left\{\frac{D_i^u}{R_{i,k}^u}\right\} \tag{9}$$

（2）下行NOMA传输模型：卫星完成第k轮微批次的中间子网络推理后，将对应不同无人机的输出特征值通过下行NOMA广播发送。假设下行信道排序为$h_{1,k}^d\leqslant h_{2,k}^d\leqslant\cdots\leqslant h_{I,k}^d$。根据下行NOMA的原理，弱信道用户通常被分配更大的功率，从而保证信道较差用户的通信质量需求。相应地，强信道用户可通过SIC依次解码并消除弱用户信号。因此，无人机i解码自身信号时的下行可达速率为：

$$R_{i,k}^d=W^d\mathrm{lb}\left(1+\frac{p_{i,k}^d h_{i,k}^d}{\sum_{j=i+1}^{I}p_{j,k}^d h_{i,k}^d+\sigma^2}\right) \tag{10}$$

其中，W^d表示下行带宽，$h_{i,k}^d$表示卫星到无人机i的下行信道增益，$p_{i,k}^d$表示卫星分配给无人机i的下行功率。

对于无人机i，只需将自身所需的中间特征值接收完成，即可立即启动本地尾部子网络的推理过程。因此，无人机i下行传输时延计算为：

$$\tau_{i,k}^d = \frac{D_i^d}{R_{i,k}^d}, \forall i \in \mathcal{U} \tag{11}$$

1.3　通信与计算流水线模型

为刻画微批次流水线协同推理过程，本节用各计算和传输阶段的开始时间与结束时间描述微批次的执行时序。

（1）头部子网络计算阶段：对于无人机i，不同微批次的计算需按顺序执行，且第1个微批次的计算开始时间为0。对应递推关系可写为：

$$\begin{cases} S_{i,1}^h = 0 \\ S_{i,k}^h = F_{i,k-1}^h, k \geqslant 2 \\ F_{i,k}^h = S_{i,k}^h + T_i^h, \ \forall i \in \mathcal{U} \end{cases} \tag{12}$$

其中，$S_{i,k}^h$和$F_{i,k}^h$分别表示无人机i第k个微批次头部子网络计算的开始时间和完成时间。

（2）上行传输阶段：第k个微批次的上行传输需等待头部子网络计算完成，且上一个微批次上行传输完成，即链路空闲。对应递推关系可写为：

$$\begin{cases} S_k^u = \max\left\{\max_{i \in \mathcal{U}} F_{i,k}^h, F_{k-1}^u\right\} \\ F_k^u = S_k^u + \tau_k^u \end{cases} \tag{13}$$

其中，S_k^u和F_k^u分别表示第k个微批次上行传输的开始时间和完成时间。

（3）中间子网络计算阶段：中间子网络计算需要等待上行传输完成，且需前一个微批次中间子网络计算完成。即：

$$\begin{cases} S_k^m = \max\left\{F_k^u, F_{k-1}^m\right\} \\ F_k^m = S_k^m + T^m \end{cases} \tag{14}$$

其中，S_k^m和F_k^m分别表示第k个微批次中间子网络计算的开始时间和完成时间。

（4）下行传输阶段：下行传输需等待中间子网络计算完成，同时需等待下行链路空闲，即：

$$\begin{cases} S_k^d = \max\{F_k^m, F_{k-1}^d\} \\ F_{i,k}^d = S_k^d + \tau_{i,k}^d \\ F_k^d = \max_{i \in \mathcal{U}} F_{i,k}^d \end{cases} \tag{15}$$

其中，S_k^d和$F_{i,k}^d$分别表示第k个微批次下行传输的开始时间和结束时间，F_k^d表示第k个微批次下行阶段的整体完成时间。

（5）尾部子网络计算阶段：尾部子网络计算需等待下行传输完成，且上一个微批次计算完成，即：

$$\begin{cases} S_{i,k}^t = \max\left\{F_{i,k}^d, F_{i,k-1}^t\right\} \\ F_{i,k}^t = S_{i,k}^t + T_i^t \end{cases} \tag{16}$$

其中，$S_{i,k}^t$和$F_{i,k}^t$分别表示无人机i第k个微批次尾部子网络计算的开始时间和完成时间。

最后，当所有无人机完成第K个微批次的尾部子网络计算后，单个批量的推理结束。相应地，系统单轮批量推理时延可写为：

$$T_{\text{round}} = \max_{i \in \mathcal{U}} F_{i,K}^t \tag{17}$$

式（12）~式（17）建立的时序递推关系刻画了各流水线阶段之间的依赖与重叠关系，因此最终推理时延不仅反映了各阶段的计算与传输时延，同时也综合体现了流水线利用率、阶段重叠度、气泡开销以及瓶颈阶段等内在因素。当流水线重叠度提高或气泡开销减少时，各阶段等待时间降低，从而使端到端推理时延减小；反之，当某一阶段成为系统瓶颈时，其持续时间将直接限制流水线推进速度并增加总体推理时延。因此，本文采用端到端推理时延作为统一性能指标，以综合评价所提流水线协同推理机制的性能。

1.4　问题建模

基于上述系统模型，本文通过联合优化功率分配、微批次划分粒度和模型切分点，实现推理时延的最小化。该问题可建模为：

$$\begin{aligned} \boldsymbol{P}1: \quad & \min_{\boldsymbol{p}^u,\boldsymbol{p}^d,K,s,s'} \quad T^{\text{round}} \\ \text{s.t.} \quad & 0 \leqslant p_{i,k}^u \leqslant P_{i,\max}^u \\ & \sum_{i\in\mathcal{U}} p_{i,k}^d \leqslant P_{\max}^d \\ & K \in \mathbb{Z}_+,\ 1 \leqslant K \leqslant K_{\max} \\ & s,s' \in \mathbb{Z},\ 1 \leqslant s < s' < L \end{aligned} \tag{18}$$

其中，T^{round}由递推式（12）~式（17）决定，$\boldsymbol{p}^u$和$\boldsymbol{p}^d$分别表示上下行功率分配$p_{i,k}^u$和$p_{i,k}^d$的集合，$P_{i,\max}^u$表示无人机i的最大发射功率，$P_{\max}^d$表示卫星下行传输最大发射功率，第3条约束表示K需为正整数，且不超过设定范围$K_{\max}$，第4条约束表示s与s'为整数层索引，且满足先后顺序。

2 问题求解

问题$\boldsymbol{P}1$同时包含连续变量和离散变量，且目标函数T^{round}由流水线递推关系决定，同时上下行速率表达式使传输时延与功率变量之间呈现非凸耦合关系。因此，问题$\boldsymbol{P}1$是一个混合整数非凸优化问题，难以直接获得全局最优解。

为解决上述问题，本文提出基于AO的求解算法，将原问题分解为功率分配优化子问题和推理参数优化子问题，进一步采用连续凸近似（successive convex approximation，SCA）和投影梯度下降（projected gradient descent，PGD）算法对子问题进行求解。具体地，先通过引入辅助变量η将原问题转化为目标函数为单一参量的优化问题：

$$\begin{aligned} \boldsymbol{P}2: \quad & \min_{\eta,\{\boldsymbol{p}^u,\boldsymbol{p}^d\},K,s,s'} \quad \eta \\ \text{s.t.} \quad & F_{i,K}^t \leqslant \eta,\ \forall i \in \mathcal{U} \\ & 0 \leqslant p_{i,k}^u \leqslant P_{i,\max}^u \\ & \sum_{i\in\mathcal{U}} p_{i,k}^d \leqslant P_{\max}^d \\ & K \in \mathbb{Z}_+,\ 1 \leqslant K \leqslant K_{\max} \\ & s,s' \in \mathbb{Z},\ 1 \leqslant s < s' < L \end{aligned} \tag{19}$$

其中，上界约束$F_{i,K}^t \leqslant \eta$必然是紧的，否则可以通过进一步减小η来降低目标函数值，从而得到更优解。因此，问题$\boldsymbol{P}2$与原问题在最优解处等价。

进一步地，对离散变量进行连续松弛。针对微批次数量K，引入连续变量$q=1/K$。在获得连续优化解$q^\star$后，需要通过取整恢复成整数变量。需要指出的是，连续松弛和整数恢复可能引入一定性能损失，因此本文不声称恢复后的整数解保持松弛问题的最优性。相应地，无人机i的微批次大小可改写为$b_i=B_i q$。针对离散切分点(s,s')，引入层分配变量$x_l^h,x_l^m,x_l^t \in \{0,1\}$，其中，$x_l^h=1$、$x_l^m=1$和$x_l^t=1$分别表示第$l$层被分配给头部、中间和尾部子网络，且满足$x_l^h+x_l^m+x_l^t=1$。此外，由于子网络结构上必须连续，且不能为空，$x_l^h,x_l^m,x_l^t$还应满足$x_l^r \geqslant x_{l+1}^r$和$\sum_{l=1}^{L} x_l^r \geqslant 1$，其中，$r \in \{h,m,t\}$。进一步地，将$x_l^h,x_l^m,x_l^t$松弛为连续变量，即$0 \leqslant x_l^h,x_l^m,x_l^t \leqslant 1$。

为了让x_l^h、x_l^m、x_l^t的优化结果逼近二进制整数，引入罚项：

$$\Phi(\boldsymbol{x}) = \sum_{l=1}^{L} \sum_{r\in\{h,m,t\}} x_l^r (1-x_l^r) \tag{20}$$

从式（20）中可以看出，当$x_l^r=0$或1对于任意的l和r成立时，罚项$\Phi(\boldsymbol{x})$取值为0；而在其他情况下，$\Phi(\boldsymbol{x})$取正值。因此，在最小化η的过程中，该罚项会使解向0或1靠近。

综合上述连续松弛，并施加惩罚后的问题可写为：

$$\begin{aligned} \boldsymbol{P}3: \quad & \min_{\eta,\boldsymbol{x},q,\boldsymbol{p}^u,\boldsymbol{p}^d} \quad \eta + \lambda\Phi(\boldsymbol{x}) \\ \text{s.t.} \quad & F_{i,K}^t \leqslant \eta,\ \forall i \in \mathcal{U} \\ & 0 \leqslant x_l^r \leqslant 1,\ \forall l,r \\ & \sum_{r\in\{h,m,t\}} x_l^r = 1,\ \forall l \\ & x_l^h \geqslant x_{l+1}^h,\ l=1,\cdots,L-1 \\ & x_l^t \leqslant x_{l+1}^t,\ l=1,\cdots,L-1 \\ & \sum_{l=1}^{L} x_l^r \geqslant 1,\ r \in \{h,m,t\} \\ & 0 \leqslant p_{i,k}^u \leqslant P_{i,\max}^u \\ & \sum_{i\in\mathcal{U}} p_{i,k}^d \leqslant P_{\max}^d \end{aligned} \tag{21}$$

其中，λ为惩罚因子。

问题 $\boldsymbol{P}3$ 的进一步求解采用 AO 框架，将变量划分为功率分配 $\boldsymbol{p}^u,\boldsymbol{p}^d$ 优化与推理参数 $q,\boldsymbol{x}$ 优化两个子问题。为保证 AO 迭代稳定性，所提算法仅接受使目标函数不增加的变量更新。因此，AO 迭代产生的目标函数序列单调非增。由于 $\eta+\lambda\Phi(\boldsymbol{x})$ 为非负，目标函数序列存在下界，所提算法在目标函数意义下收敛。然而，由于原问题为混合整数非凸优化问题，所提算法难以保证获得全局最优解。

2.1　功率分配优化子问题求解

由于不同轮微批次推理在功率约束上相互独立，可以针对第 k 轮微批次推理构建功率分配子问题。基于上行传输持续时间由最慢用户决定的特性，定义辅助变量 ρ_k^u 如下：

$$\rho_k^u \leqslant \frac{R_{i,k}^u}{D_i^u},\ \forall i \in \mathcal{U} \tag{22}$$

相应地，最大化 ρ_k^u 等价于最大化所有用户中最小的归一化速率，也即等价于最小化上行传输时间 τ_k^u。因此，上行传输功率分配问题可写为：

$$\begin{aligned} \max_{\boldsymbol{p}_k^u,\rho_k^u}\quad & \rho_k^u \\ \text{s.t.}\quad & R_{i,k}^u \geqslant D_i^u \rho_k^u,\ \forall i \in \mathcal{U} \\ & 0 \leqslant p_{i,k}^u \leqslant P_{i,\max}^u \end{aligned} \tag{23}$$

由于 NOMA 速率表达式（7）中包含分式形式，上述问题是非凸的。为便于执行凸近似，首先将上行速率改写为对数差形式：

$$R_{i,k}^u = W^u\left[\mathrm{lb}\left(S_{i,k}^u\right) - \mathrm{lb}\left(I_{i,k}^u\right)\right] \tag{24}$$

其中，$S_{i,k}^u = \sum_{j=1}^{i} p_{j,k}^u h_{j,k}^u + \sigma^2$，$I_{i,k}^u = \sum_{j=1}^{i-1} p_{j,k}^u h_{j,k}^u + \sigma^2$。由于 $\mathrm{lb}(I_{i,k}^u)$ 为凹函数，其一阶泰勒展开构成全局上界，即满足：

$$\mathrm{lb}\left(I_{i,k}^u\right) \leqslant \mathrm{lb}\left(I_{i,k}^{u,(r)}\right) + \frac{1}{\ln 2}\frac{I_{i,k}^u - I_{i,k}^{u,(r)}}{I_{i,k}^{u,(r)}} \tag{25}$$

其中，$I_{i,k}^{u,(r)}$ 是将第 r 次 SCA 迭代值 $\boldsymbol{p}_k^{u,(r)}$ 代入 $I_{i,k}^u$ 计算得到的结果。应用该不等式关系，可以得到速率的凹下界近似：

$$\tilde{R}_{i,k}^u = W^u\left[\mathrm{lb}\left(S_{i,k}^u\right) - \mathrm{lb}\left(I_{i,k}^{u,(r)}\right) - \frac{1}{\ln 2}\frac{I_{i,k}^u - I_{i,k}^{u,(r)}}{I_{i,k}^{u,(r)}}\right] \tag{26}$$

其中，在迭代点处与原函数保持函数值一致和一阶梯度一致，同时由下界近似构造的速率约束是原约束的保守近似。因此，所采用的 SCA 近似满足局部紧性、一阶一致性和可行性保持条件。

基于式（26），上行功率分配可转化为如下凸优化问题：

$$\begin{aligned} \max_{\boldsymbol{p}_k^u,\rho_k^u}\quad & \rho_k^u \\ \text{s.t.}\quad & \tilde{R}_{i,k}^u \geqslant D_i^u \rho_k^u,\ \forall i \in \mathcal{U} \\ & 0 \leqslant p_{i,k}^u \leqslant P_{i,\max}^u \end{aligned} \tag{27}$$

该子问题可通过 CVX 等工具高效求解。此外，下行功率分配同样采用上述方法类似求解。

2.2　推理参数优化子问题求解

在固定发射功率 $\{\boldsymbol{p}^u,\boldsymbol{p}^d\}$ 下，$\boldsymbol{P}3$ 可写为：

$$\begin{aligned} \min_{q,\boldsymbol{x}}\quad & \eta\left(q,\boldsymbol{x}\right) + \lambda\Phi\left(\boldsymbol{x}\right) \\ \text{s.t.}\quad & F_{i,K}^t\left(q,\boldsymbol{x}\right) \leqslant \eta,\ \forall i \in \mathcal{U} \\ & 1/K_{\max} \leqslant q \leqslant 1 \\ & 0 \leqslant x_l^r \leqslant 1,\ \forall l,r \\ & \sum_{r \in \{h,m,t\}} x_l^r = 1,\ \forall l \end{aligned} \tag{28}$$

由于目标函数中包含最大值运算，且 $\boldsymbol{x}$ 和 q 以非线性方式耦合，该推理参数优化子问题非凸。为了解决最大值运算带来的函数非光滑问题，采用平滑最大值函数对最大值计算进行近似，即：

$$\mathrm{LSE}_\beta(a,b) = \frac{1}{\beta}\mathrm{lb}\left(\exp(\beta a) + \exp(\beta b)\right) \tag{29}$$

其中，$\beta>0$ 为平滑参数，当 β 足够大时，平滑函数逼近最大值函数。通过用平滑函数替代流水线

递推计算式中的最大值运算，推理时延可以近似为：

$$\tilde{\eta}(q,\boldsymbol{x})=\frac{1}{\beta}\mathrm{lb}\left(\sum_{i\in\mathcal{U}}\exp\left(\beta\tilde{F}_{i,K}^{t}\right)\right) \tag{30}$$

相应地，式（28）中的目标函数可以近似为：

$$\tilde{\mathcal{L}}(q,\boldsymbol{x})=\tilde{\eta}(q,\boldsymbol{x})+\lambda\Phi(\boldsymbol{x}) \tag{31}$$

进一步地，采用PGD方法更新优化参数。具体而言，在第r次迭代中，首先沿负梯度方向更新：

$$\tilde{q}^{(r+1)}=q^{(r)}-\alpha_q\nabla_q\tilde{\mathcal{L}}\left(q^{(r)},\boldsymbol{x}^{(r)}\right) \tag{32}$$

$$\tilde{\boldsymbol{x}}^{(r+1)}=\boldsymbol{x}^{(r)}-\alpha_x\nabla_x\tilde{\mathcal{L}}\left(q^{(r)},\boldsymbol{x}^{(r)}\right) \tag{33}$$

其中，α_q和α_x分别为q和$\boldsymbol{x}$的更新步长。由于梯度更新后的变量可能不再满足可行性约束，需要将其投影回可行域。对于q，采用区间投影：

$$q^{(r+1)}=\min\left\{q_{\max},\max\left\{q_{\min},\tilde{q}^{(r+1)}\right\}\right\} \tag{34}$$

其中，$q_{\min}=1/K_{\max}$，$q_{\max}=1$。对于层分配变量$\boldsymbol{x}_l=[x_l^h,x_l^m,x_l^t]$，将其投影到：

$$\Delta=\left\{\boldsymbol{z}\left|z_r\geqslant 0,\sum_{r\in\{h,m,t\}}z_r=1\right.\right\} \tag{35}$$

即：

$$\boldsymbol{x}_l^{(r+1)}=\Pi_\Delta\left(\tilde{\boldsymbol{x}}_l^{(r+1)}\right),\ \forall l \tag{36}$$

上述PGD更新过程迭代执行，直至子问题的目标函数变化小于预设误差。

2.3 算法复杂度分析

对于功率分配子问题，上行与下行功率分配均采用SCA求解。每次凸优化采用内点法的复杂度为$O(I^{3.5})$。设SCA内迭代次数为N_{SCA}，则该子问题复杂度为$O(KN_{\mathrm{SCA}}\cdot I^{3.5})$。对于推理参数子问题，采用PGD优化标量$q$与层分配变量$\boldsymbol{x}\in\mathbb{R}^{3L}$。每轮PGD迭代中，递推计算流水线时延的复杂度为$O(KI)$，投影操作复杂度为$O(L)$。设PGD迭代次数为$N_{\mathrm{PGD}}$，则复杂度为$O(N_{\mathrm{PGD}}(KI+L))$。综上，设AO主循环次数为$N_{\mathrm{AO}}$，算法总复杂度为$O\left(N_{\mathrm{AO}}\left(KN_{\mathrm{SCA}}I^{3.5}+N_{\mathrm{PGD}}(KI+L)\right)\right)$。

3 仿真结果

除非另行说明，仿真参数设置为1个卫星节点，3架无人机，每架无人机的推理批量大小设为128，无人机推理计算能力设为2×10^{10} FLOPS，卫星计算能力设为2.5×10^{11} FLOPS，无人机上行最大发射功率设为0.5 W，下行卫星发射功率设为3 W，上下行链路带宽均设为20 MHz，噪声功率谱密度为−174 dBm/Hz，信道增益在$[10^{-8},10^{-6}]$内随机生成，最大微批次数设为$K_{\max}=8$。卫星和无人机采用ResNet18对CIFAR-10图像集进行分类。根据ResNet18的结构，网络被划分为$L=6$层，每层的计算量和输出的特征值大小通过对ResNet18进行实验和统计采样获得，ResNet18模型统计量见表1。此外，PGD更新步长分别设置为$\alpha_q=0.01$和$\alpha_x=0.2$。仿真中以平均单轮推理时延为性能指标，其中，单轮推理时延指3架无人机在卫星协同下各自完成128个样本推理所需的时间。

表1 ResNet18模型统计量

分层	FLOPS/样本	特征值大小/样本
输入层	$1.966\,08\times10^{6}$	262 144 Byte
残差层1	$1.517\,81\times10^{8}$	262 144 Byte
残差层2	$1.346\,76\times10^{8}$	131 072 Byte
残差层3	$1.344\,47\times10^{8}$	65 536 Byte
残差层4	$1.343\,32\times10^{8}$	32 768 Byte
分类器	$1.332\,2\times10^{4}$	—

通信链路带宽对平均推理时延的影响如图3所示，给出了不同通信链路带宽下所提方案与基准方案的平均单轮推理时延。从图3中可以看出，随着通信链路带宽的增大，本地推理方案的平均推理时延不发生变化，3种切分推理方案的时延

均随带宽增大而下降。相比之下，传统切分推理取得的时延最高，所提方案取得最小平均推理时延。这一现象的原因是，基于时分多址接入（time division multiple access，TDMA）的流水线切分方案采用并行计算范式，相比传统切分方案的无人机间串行计算减少了计算等待时间，因此取得更小的推理时延。而所提方案利用NOMA的并行传输和更高的吞吐量优势，进一步将传输并行化，减少了节点间传输等待和信息传输所用时间，因此能够取得相比其他对比方案更小的平均推理时延。

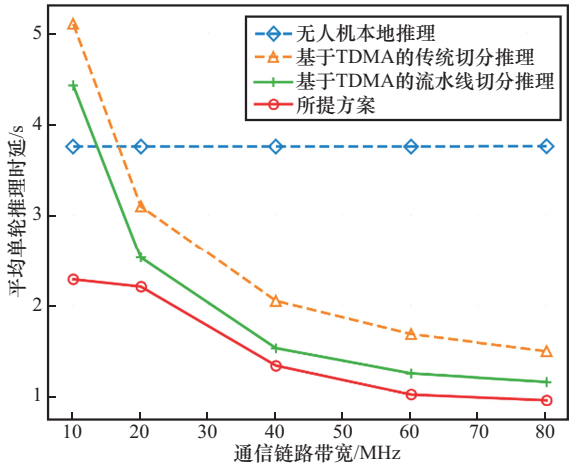

图3　通信链路带宽对平均推理时延的影响

卫星计算能力对平均推理时延的影响如图4所示，给出了不同卫星计算能力下各个方案所取得的平均单轮推理时延。从图4中可以看到，无人机本地推理方案的时延不随卫星计算能力的变化而改变，其他3种协同推理方案所取得的时延随卫星计算能力的提升而下降，说明卫星侧的中间层计算是切分推理时延的重要组成部分。同时，随着卫星计算能力的不断提升，3种协同推理方案的推理时延下降速度都逐渐平缓。这是由于，当卫星算力增大到一定程度，推理时延将受限于无人机计算能力和网络通信能力。此外，相比所有对比方案，所提方案始终取得最低推理时延，验证了所提方案的有效性。

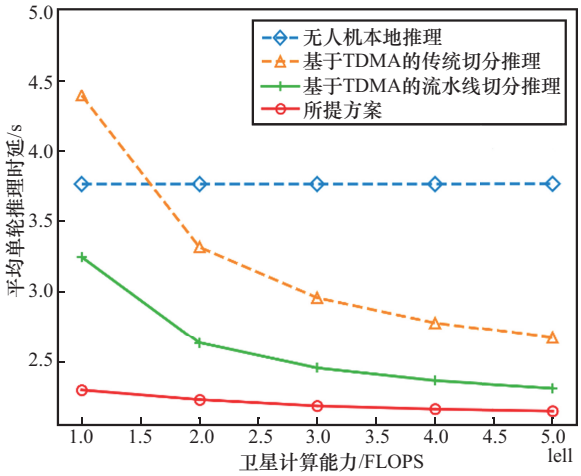

图4　卫星计算能力对平均推理时延的影响

切分点选择对平均推理时延的影响如图5所示，将所提方案下联合优化$\{p_{i,k}^u, p_{i,k}^d, K, s, s'\}$与固定推理参数$\{K, s, s'\}$，但使用相同方法优化功率$p_{i,k}^u$和$p_{i,k}^d$所取得的仿真结果进行对比。具体地，图5中除联合优化外，其余结果均采用$K=8$的微批次划分粒度。从图5中可以看出，在不同的网络切分点下，系统所取得的平均推理时延存在显著差异。这是由于，网络切分点不仅决定了无人机和卫星分别承担的计算量，也决定了上下行通信数据量。以带宽10 MHz为例，此时切分点$(s, s')=(3, 5)$取得了比其他固定切分点更小的平均推理时延。这是由于，在该情况下通信带宽较小，系统推理时延受限于通信速率。而如表1所示，ResNet18越靠后的分层所输出的特征值数据量越低，对通信速率的要求也相应降低。对于带宽为60 MHz的情况，$(s, s')=(1, 5)$取得了更小的推理时延，意味着此时更优的决策是将更多计算任务部署到卫星。此外，联合优化$\{p_{i,k}^u, p_{i,k}^d, K, s, s'\}$能够取得比固定推理参数$\{K, s, s'\}$更小的平均推理时延。

微批次划分数对平均推理时延的影响如图6所示，实验中采用固定切分点$(s, s')=(1, 5)$。从图6中可以看到，当$I=2$时，系统平均推理时延在$K=8$时取得最小值。当无人机数增加到$I=4$，系

统在$K=4$时取得最小平均推理时延，并随K的进一步增大而提升。该结果表明，微批次数的最优取值会依系统参数发生变化。原因是，增大K能够细化流水线通信与计算过程，提高计算与通信的重叠度，起到降低推理时延的作用。但是，随着用户数增加，卫星侧微批次聚合推理中的最慢用户传输瓶颈会得到增强，过大的微批次数会导致排队等待开销的累积，从而抵消流水线带来的增益。

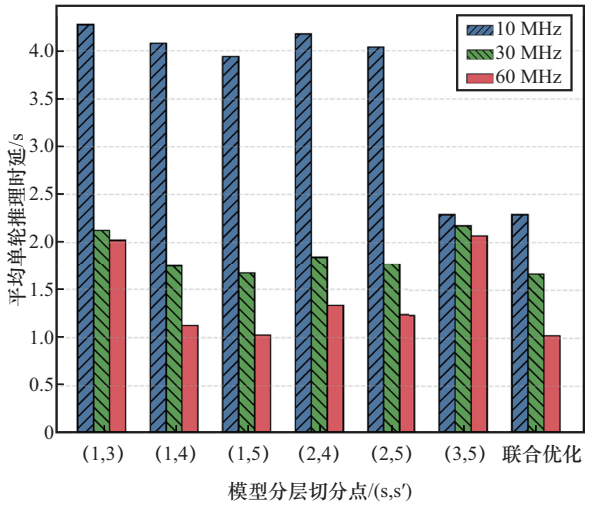

图5　切分点选择对平均推理时延的影响

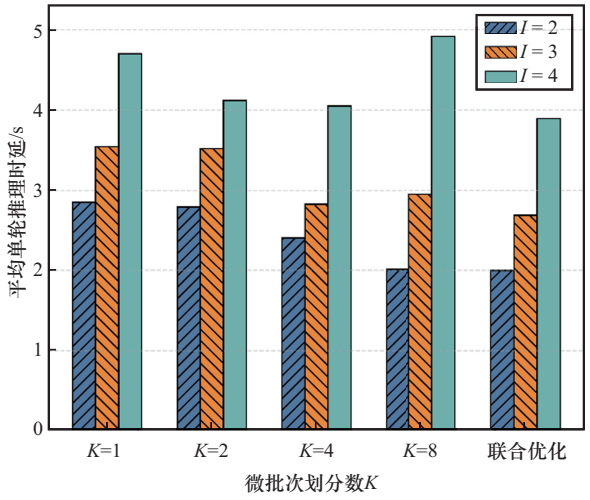

图6　微批次划分数对平均推理时延的影响

不同功率分配策略下通信链路带宽对平均推理时延的影响如图7所示。图7中参与比较的3种策略分别为所提基于AO的联合优化；最大功率/均分功率策略，该策略在上行传输中，各无人机均采用自身最大发射功率，而在下行传输时，卫星将最大发射功率对所有无人机均分；混合NOMA启发式功率策略，该策略根据上下行NOMA所采用的解码次序，在上行链路中采用信道正比分配，提高卫星接收不同信号在功率域上的差异性，即$p_i^u=P_{i,\max}^u h_i^u/\max_j h_j^u$，在下行链路中采用信道反比分配，保障弱链路无人机的通信速率，即$p_i^d=(P_{\max}^d/h_i^d)/\sum_j(1/h_i^d)$。对比结果表明，最大功率/均分功率策略取得的推理时延最高。这是由于，该策略无法保证NOMA所需的信号功率域差异。相比之下，所提功率分配优化策略能够在不同带宽下取得比另外两种对比策略更低的推理时延。

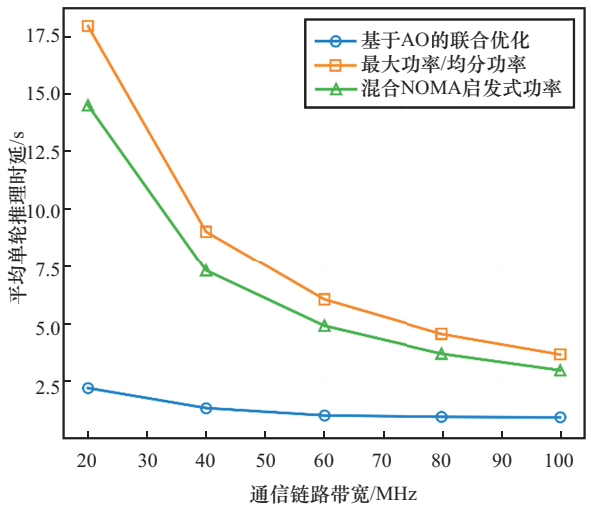

图7　不同功率分配策略下通信链路带宽对平均推理时延的影响

最后，仿真实验给出切分层特征值量化对推理精度和推理时延的影响。实验中对子网络输出采用均匀量化，用于压缩通信数据量。量化位宽对推理精度的影响如图8所示。从图8中可以看出，当量化位宽较低时，如2 bit和4 bit，推理精度相比32 bit基准显著降低。而当量化位宽超过8 bit后，推理能保持较好的性能。

不同带宽下量化位宽对平均推理时延的影响

如图9所示。从图9中可以看出，平均推理时延会随量化位宽的降低而下降。由于仿真采用均匀量化，量化与反量化计算复杂度与特征值数量线性相关。这部分开销相比卷积神经网络的推理计算与空口传输时延通常较小。因此，图9展示的时延变化趋势主要源于量化降低传输数据量所带来的通信时延下降。此外，传输带宽较小时，如10 MHz，量化带来的时延下降则更加明显，说明此时中间特征传输是端到端推理时延的主要瓶颈。

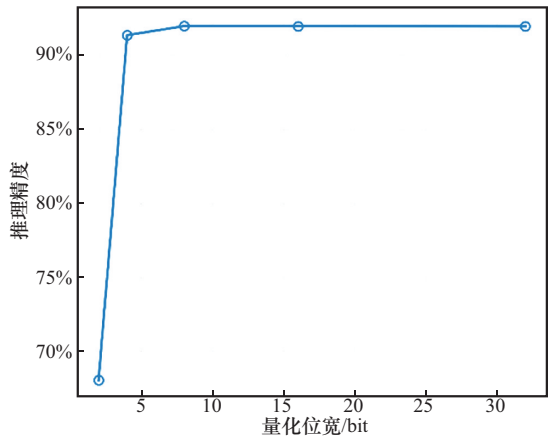

图8　量化位宽对推理精度的影响

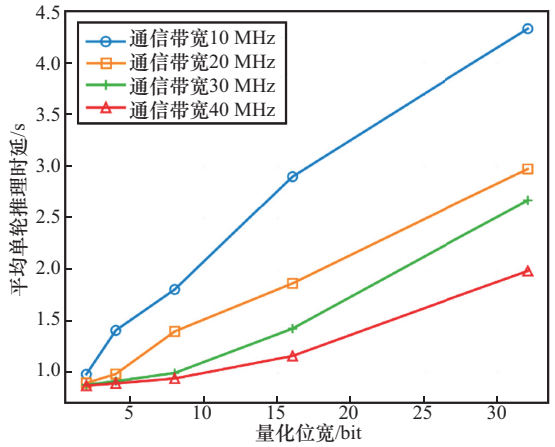

图9　不同带宽下量化位宽对平均推理时延的影响

4　结束语

本文面向卫星无人机协同推理场景，提出一种NOMA辅助的微批次流水线并行推理方案。该方案通过U型模型切分将计算密集的中间子网络部署于卫星侧，并利用微批次流水线和NOMA传输提高通信与计算阶段的重叠度。在此基础上，本文构建了端到端推理时延最小化问题，并联合优化了模型切分点、微批次数和上下行功率分配。仿真结果表明，所提方案能够有效降低系统推理时延，提升系统资源利用率。

参考文献：

[1] Baena E, Testolina P, Polese M, et al. Space-O-RAN: enabling intelligent, open, and interoperable non-terrestrial networks in 6G[J]. IEEE Communications Magazine, 2026, 64(2): 112-118.

[2] Fontanesi G, Ortiz F, Lagunas E, et al. Artificial intelligence for satellite communication: a survey[J]. IEEE Communications Surveys & Tutorials, 2026, 28: 1381-1435.

[3] Fan G T, Wu P, Yang M Q, et al. Internet of satellites (IoS) for intelligent satellite cluster: applications, methods, and challenges[J]. Engineering, 2025, 54: 155-170.

[4] Wen D Z, Zhang S, Zhu G X, et al. UAV-assisted edge inference with integrated sensing, communication, and computation[J]. IEEE Transactions on Wireless Communications, 2026, 25: 13336-13350.

[5] Huang J F, Lian L X, Wen D Z, et al. Dynamic UAV-assisted cooperative edge AI inference[J]. IEEE Transactions on Wireless Communications, 2025, 24(1): 615-628.

[6] Lin Y S, Feng W, Zhou T, et al. Integrating satellites and mobile edge computing for 6G wide-area edge intelligence: minimal structures and systematic thinking[J]. IEEE Network, 2023, 37(2): 14-21.

[7] Yuan S, Peng M G, Sun Y H. Satellite-terrestrial integrated fog networks: architecture, technologies, and challenges[J]. IEEE Wireless Communications, 2025, 32(4): 208-215.

[8] Fan W H, Meng Q C, Wang G, et al. Satellite edge intelligence: DRL-based resource management for task inference in LEO-based satellite-ground collaborative networks[J]. IEEE Transactions on Mobile Computing, 2025, 24(10): 10710-10728.

[9] Yao S, Lin Y Y, Wang M, et al. LEOEdge: a satellite-ground cooperation platform for the AI inference in large LEO constellation[J]. IEEE Journal on Selected Areas in Communications,

2025, 43(1): 36-50.

[10] Wang Y M, Zhao K, Zhang X X, et al. Towards space intelligence: adaptive scheduling of satellite-ground collaborative model inference with space edge computing[C]//Proceedings of the IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Piscataway: IEEE Press, 2024: 1-6.

[11] Qiao Y, Teng S Y, Luo J, et al. On-orbit DNN distributed inference for remote sensing images in satellite Internet of things[J]. IEEE Internet of Things Journal, 2025, 12(5): 5687-5703.

[12] Yin B L, Fang X M, Wang X B. Robust group target awareness inference in multi-UAV-enabled ISCC networks based on split deep reinforcement learning[J]. IEEE Transactions on Wireless Communications, 2025, 24(11): 9478-9492.

[13] Chen Y J, Zhang Q Y, Xing R L, et al. SLICE: energy-efficient satellite-ground co-inference via layer-wise scheduling optimization[J]. IEEE Transactions on Services Computing, 2025, 18(4): 2388-2402.

[14] Zhao H T, Huang S Y, Xu B, et al. Hastening co-inference in space-air-ground integrated networks: a submodular optimization based task offloading approach[J]. IEEE Transactions on Vehicular Technology, 2026.

[15] Xu Z Q, Zhang P Y, Li C C, et al. A collaborative inference algorithm in low-earth-orbit satellite network for unmanned aerial vehicle[J]. Drones, 2023, 7(9): 575.

[16] Fan S Q, Rong Y, Meng C, et al. DAPPLE: a pipelined data parallel approach for training large models[C]//Proceedings of the 26th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. New York: ACM Press, 2021: 431-445.

[17] Jeon B, Wu M D, Cao S Y, et al. GraphPipe: improving performance and scalability of DNN training with graph pipeline parallelism[C]//Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1. New York: ACM Press, 2025: 557-571.

[18] Xu G B, Le Z H, Chen Y H, et al. AutoCCL: automated collective communication tuning for accelerating distributed and parallel DNN training[C]//Proceedings of the 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI'25). Berkeley: USENIX Association, 2025: 667-683.

[19] Wang L Y, Dong Y J, Zhang L, et al. Wireless model splitting for communication-efficient personalized federated learning with pipeline parallelism[C]//Proceedings of the 2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC). Piscataway: IEEE Press, 2023: 421-425.

[20] Gao L Y, Liu J C, Xu H L, et al. Accelerating end-cloud collaborative inference via near bubble-free pipeline optimization[C]//Proceedings of the IEEE INFOCOM 2025-IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2025: 1-10.

[21] Chen Y, Jia X L, Wang H Q. Hierarchical offloading optimization for collaborative DNN inference in satellite edge computing networks[J]. IEEE Transactions on Mobile Computing, 2026, 25(8): 12128-12143.

[22] Wu C H, Guo C B, Yin Z S. A distributed-computing architecture for enabling in-orbit satellite adaptation to evolving AI models[J]. IEEE Internet of Things Journal, 2026, 13(2): 3513-3528.

[23] Zhang S G, Wu W, Wu S H, et al. Collaborative LLM inference over LEO satellite networks: model splitting and pipeline parallelism[C]//Proceedings of the 2025 Seventeenth International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2025: 1-6.

[24] Wu C H, Zheng P F, Jia Y H, et al. PipeSIA: an efficient pipeline-parallel distributed space inference approach for LEO constellations[C]//Proceedings of the 2025 IEEE International Conference on Communications. Piscataway: IEEE Press, 2025: 6976-6981.

[作者简介]



杨萌琪（1989− ），男，博士，杭州电子科技大学通信工程学院特聘副教授，主要研究方向为卫星通信、卫星计算等。



吴义曼（2001− ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为无线通信多址接入等。



周任飞（1983– ），男，中国人民解放军31007部队工程师，主要研究方向为电磁频谱管理、空间频轨管理等。



潘鹏（1983– ），男，博士，杭州电子科技大学通信工程学院教授，主要研究方向为低空经济、无线通信等。



吴鹏（1993– ），男，博士，中国人民解放军军事科学院国防科技创新研究院副研究员，主要研究方向为卫星通信、卫星计算等。



郑长亮（1980– ），男，博士，杭州电子科技大学通信工程学院讲师，主要研究方向为无线通信、卫星通信等。