# 面向低空物联网的云-边协同演进模型与通信范式

于馨博[1]，张舒航[2]，张泓亮[3]
（1. 北京大学信息科学技术学院，北京 100871；2. 鹏城实验室，广东 深圳 518055；3. 北京大学电子学院，北京 100871）

**摘 要：**低空物联网基于空地一体化网络，集成通信和计算功能，在低空场景可以高效地收集、传输和分析数据，为低空经济的发展持续赋能。在这一网络中，无人机（UAV, unmanned aerial vehicle）等空中平台利用机载传感器收集多模态感知数据，并进行基于人工智能（AI, artificial intelligence）的数据处理计算，以支持各种低空场景下的应用，如农业监控和环境建模。执行多模态数据的推理和内容生成任务需要大型 AI 模型。为了满足这些任务的需求，无人机需要具备强大的计算资源和大量数据支持。这些要求使得高效的推理模型训练和优化变得至关重要。然而，这给现有的低空物联网带来了巨大挑战。为解决这一问题，提出空地一体化云-边模型协同演化架构。在此架构中，无人机作为边缘节点，负责数据采集和小型模型的计算。云服务器通过无线信道与无人机进行信息交互，提供大型模型计算和边缘无人机的模型更新服务，从而实现空地协作。在有限的无线通信带宽限制下，该架构面临着边缘无人机与云服务器之间信息交换调度设计的挑战。为此，提出任务分配、传输资源管理、传输数据量化设计和边缘模型更新的联合策略。该策略通过最大化系统的平均精度（mAP, mean average precision）来提高空地一体化云-边模型协同演化架构的推理准确性。基于边缘模型的平均精度和云模型的平均精度推导出了所提出架构的平均精度闭式下界，并相应地提出了平均精度最大化问题的优化方案。基于视觉分类实验结果的仿真表明，在不同通信带宽和数据量条件下，相比于集中式云模型架构和分布式边缘模型架构，低空物联网在所提出的空地一体化云-边模型协同演化架构下的平均精度均有所提升。

**关键词：**大模型；边缘智能；无人机
**中图分类号：**TN92
**文献标志码：**A
**doi：**10.11959/j.issn.2096−3750.2024.00425

# An edge-cloud collaborative model evolution and communication paradigm in Internet of low-altitude UAV

YU Xinbo[1], ZHANG Shuhang[2], ZHANG Hongliang[3]
1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
2. PengCheng Laboratory, Shenzhen 518055, China
3. School of Electronics, Peking University, Beijing 100871, China

**Abstract:** The low-altitude Internet of things (IoT), based on an air-ground integrated network, combines communication and computing functions. This allows it to efficiently collect, transmit, and analyze data in low-altitude scenarios, continuously empowering the development of the low-altitude economy. In this network, aerial platforms such as unmanned aerial vehicle (UAV) uses onboard sensors to gather multimodal perception data and perform AI-based data processing to support various low-altitude applications, such as agricultural monitoring and environmental modeling. Executing multimodal data inference and content generation tasks requires large AI models. To meet these demands, UAV needs powerful

**收稿日期：**2024−08−19；**修回日期：**2024−09−10
**通信作者：**张舒航，zhangshh01@pcl.ac.cn
**基金项目：**国家自然科学基金项目（No. 62401302，No. 62371011）；北京市自然科学基金—小米创新联合基金项目（No. L243002）
**Foundation Items:** The National Natural Science Foundation of China (No. 62401302, No. 62371011), The Natural Science Foundation of Beijing-Xiaomi Innovation Joint Fund (No. L243002)

computing resources and vast data support, making efficient model training and optimization essential. However, this poses significant challenges to the current low-altitude IoT network. To address this, an integrated air-ground edge-cloud collaborative framework was proposed, where UAV function as edge nodes, collecting data and performing small-scale computations. Through wireless channels, cloud servers provide large-scale computations and update models for the UAV, enabling efficient collaborations. Given limited wireless communication bandwidth, the framework faces challenges in scheduling information exchange between the UAV and the cloud servers. To solve this, joint optimizations for task allocation, transmission resource management, data quantization, and edge model updates were presented, to improve inference accuracy by maximizing the mean average precision (mAP) of the proposed framework. A closed-form lower bound for the mAP based on the performance of the edge and cloud models were derived and a solution to mAP maximization was proposed. Simulations, based on visual classification experiments, show that the mAP of proposed framework under IoL-oUA consistently outperforms centralized and distributed frameworks across various bandwidth and data conditions.

**Key words:** large model, edge intelligence, unmanned aerial vehicle

## 0　引言

随着无人机（UAV, unmanned aerial vehicle）技术的进步和低空飞行器的普及，低空经济预计将成为现代社会发展的重要组成部分，有望在物流、交通、农业、安防等多个关键应用领域发挥重要作用[1]。这些应用的高效实现，都依赖快速、实时的数据传输和计算支持，而低空物联网正是这一领域的核心技术[2]。通过无人机等低空飞行设备与地面服务器联合构建的网络系统，低空物联网可以完成多设备之间的连接与协同，从而实现在各种低空场景下的数据采集、传输和处理，以支持无人机配送、农业监控、灾害预警、安防巡逻等典型应用[3]。为了有效提高低空物联网的效率，上述应用需要人工智能（AI, artificial intelligence）技术帮助其进行数据处理和传输决策。同时，为了进一步减少无人机对中心服务器的依赖、提升其自主数据分析和处理能力，边缘无人机上需要部署含有数十亿个参数的大型AI推理模型[4]，以实现识别任务的高推理精度和高泛化性[5]。这种模式不仅能加速低空物联网的部署，还能为低空经济的进一步发展提供技术保障，解决实际应用中遇到的困难[6]。

一段时间以来，大量研究致力于将无人机作为边缘计算节点用于各种AI应用[7]。文献[8]针对基于无人机的合成孔径雷达任务定制的AI模块，提出了一个由深度神经网络驱动的全面测试平台，用于目标检测。文献[9-10]采用部署在边缘无人机上的卷积神经网络对视频中的目标进行识别，可实现持续的目标跟踪能力，并进一步提出了一种可扩展的空中计算优化方案，适用于不同质量等级的计算任务，这些任务对应不同的计算工作负载和不同性能的计算结果，以适应边缘无人机的硬件计算能力。文献[11]提出了一种云-边混合系统架构，其中边缘无人机负责处理AI任务，而云服务器负责数据存储、操作和可视化。

尽管无人机作为边缘AI处理器具有广阔的潜力，但由于无人机的机载计算能力不足以应对低空物联网所需的高要求应用，架构[8-11]无法直接支持无人机作为低空物联网中的边缘AI节点。现有研究[8-11]表明，无人机只能执行需要较低计算能力的推理任务，这些任务由仅包含数百万参数的模型驱动，例如YOLOv8[12]。然而，低空物联网预计将支持诸如灾难响应和环境构建等应用，这些应用需要具有数十亿参数的大型多模态模型[13]，例如SORA[14]和Gemini[15]。因此，需要有具备强大计算能力的地面云服务器与边缘无人机协同工作，从而可以利用云端的大型模型进行高效的数据处理与分析。在最新的工作中，文献[16]基于模糊神经网络搭建了云-边协同网络以支持目标检测任务，同时提出了对应的任务调度与优化算法。同时，文献[17]给出了在考虑时延情况下，以无人机为边缘节点的云-边协同计算系统最优的任务调度方案。

然而，已有工作的系统架构没有考虑边缘推理模型的更新问题[16-17]。由于边缘无人机的计算能力不足以进行模型训练，因此无法自主更新机载推理模型[18]，这将导致边缘AI服务的鲁棒性和准确性受限[19]。因此，机载推理模型的准确性在环境变化时会显著降低[20]。基于上述原因，需要一个新的架构，在支持边缘无人机与具备强大计算能力的地面云服务器之间协作的同时，为边缘无人机提供由大

模型驱动的模型更新服务[21-22]。

为了解决上述问题，本文提出了一种基于联合数据和模型通信范式的全新空地一体化云-边模型协同演化架构。在所提出的架构中，每个边缘无人机需要进行感知数据的收集、压缩特征的提取和本地数据的分析，并根据实际通信条件选择性向云服务器上传残差数据和特征进行大型模型分析，以最大化系统推理精度。为了提升边缘模型的推理性能，云服务器还会向边缘无人机传输模型更新数据。构建了一个空地一体化模型协作优化问题，以提高整个网络的推理精度。该问题的设计涵盖了边缘无人机与云服务器之间的任务分配，同时考虑了特征传输、残差映射数据传输和模型更新传输的开销，以最大化无人机和云服务器的平均精度（mAP, mean average precision）。

需要注意的是，在设计此空地一体化云-边模型协同演化架构时，有几个问题需要仔细考虑。首先，需要定义一个性能指标，这一指标将作为优化边缘无人机与云服务器之间任务和通信资源分配的基础。其次，上行的数据传输有助于云端模型进行计算并实现较高的平均精度，而下行的模型更新则可以有效提高边缘模型的平均精度。因此，在有限的无线通信带宽限制下，深入研究上行和下行资源分配之间的权衡对提高系统的整体推理性能至关重要。最后，鉴于有限的上行传输带宽限制，边缘无人机面临着如下选择：向云服务器传输低分辨率的特征数据以处理更多任务，或是传输高分辨率的残差映射数据以处理较少任务。因此，同样需要对特征传输与残差映射数据传输的分配问题进行深入地分析和权衡。

为解决上述挑战，本文工作创新点总结如下。

1) 提出空地一体化云-边模型协同演进的新架构

本文提出了一种新的空地一体化云-边模型协同演化架构，旨在帮助边缘无人机处理与分析收集到的感知数据，并在地面云服务器的协助下对机载边缘模型进行更新。该架构包含 3 个独立的数据传输流：特征流、数据流和模型流，每个流上传输的数据量可以根据无线网络的通信带宽进行动态调整。

2) 设计云-边协同的动态任务调度

提出了一种基于联合数据和模型通信的动态任务调度策略，针对不同的通信条件和任务需求，灵活分配边缘无人机与云服务器的任务负载。该策略通过对特征和残差映射数据，以及模型增量更新数据的动态传输控制，实现了在有限带宽下的最优任务分配，提高了边缘 AI 服务的鲁棒性和系统整体效率。

3) 分析优化理论

基于本文所提架构，分析了在最优化任务调度要求下，数据量化位数的合理分配方法，并根据平均精度的数学定义，推导出了本文云-边协同架构联合平均精度的闭式下界表达式，极大地降低了优化算法的复杂度。

## 1　集成空地边缘云模型演进框架

空地一体化云-边模型协同演化架构保障了边缘计算和云计算的同时进行。所提出的空地一体化云-边模型协同演化架构如图 1 所示，框架由边缘节点（即无人机）和云节点（即云服务器）组成。为了便于说明，仅展示了一个边缘无人机和一个云节点。无人机配备了机载数据收集器（如摄像机）和边缘计算模块充当远程传感器和边缘服务器。地面的云服务器则作为增强分析和识别的中央节点。鉴于无线信道带宽存在不稳定性，空地一体化云-边模型协同演化架构需要灵活的通信范式设计，其中包括动态任务分配、按需的残差映射数据传输以及灵活的边缘模型更新。

具体来说，每个边缘无人机负责收集感知数据以进行后续的数据分析和识别。为简化说明，以视觉数据分类任务为例。对每一帧进行的分析和识别被称为一个任务。这些任务可以由无人机通过机载边缘模型执行，也可以由地面的云服务器通过云模型执行。

为了在云服务器上进行模型分析，边缘无人机首先使用机载的特征提取模型提取视觉数据特征，然后通过空中传输（OTA, over-the-air）将提取的视觉数据特征传输到云服务器，称为特征流。云服务器接收到上传的紧凑视觉特征，通过大型模型分析进行特征推理，以支持如图像分割或目标检测等任务。然而，尽管目前的特征提取技术可以极大地降低通信资源的开销，但基于特征的推理精度仍要劣于传统基于图像的计算机视觉分析，并且无法支持像图像重建一类的应用。因此，为了进一步增强云

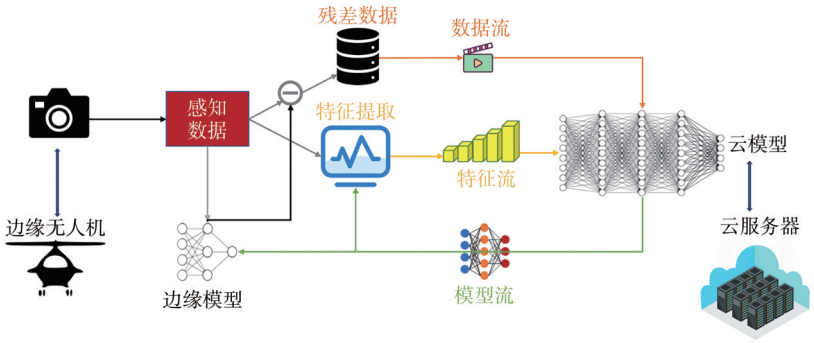



图1　空地一体化云-边模型协同演化架构示意图

服务器的推理性能，可以在空闲的OTA传输资源，传输比提取出的视觉数据特征信息更加详细，且可动态调整分辨率的压缩图像数据，这些数据是残差映射数据，称为数据流。数据流在特征流的基础上，向云服务器传输部分高质量压缩的原始图像数据，保障云服务器可以进行特征—图像的联合分析，从而有效地提升系统的推理精度。同时，为了处理来自各个领域的任务和数据，使系统可以快速地适应环境的变化，云服务器可以通过灵活地传输开销，将模型更新数据传输到无人机，以升级其特征提取和边缘推理模型，这被称为模型流。

最近，已经有几项支持性研究致力于在空地一体化云-边模型协同演化架构中实现上述3种数据流[21]。对于特征流，紧凑特征表示技术确保了高效的特征提取和数据压缩，使得特征流的开销降低到几kbit/s[22]。

对于数据流中的残差映射数据，智能编码技术促进了图像/视频的高效表示，能够将视频流动态编码到一个实际且合适的级别[23]。对于边缘模型更新，模型压缩和增量更新技术允许通过模型流动态更新模型，从而及时处理来自各个领域的任务和数据[24]。

上述研究已经证明了在空地一体化云-边模型协同演化架构中引入特征流、数据流和模型流的可行性。然而，在研究OTA通信时仍存在如下两个挑战。首先，考虑3个数据流的开销可以动态调整，因此需要研究系统性能与各个流通信数据量之间的关系；其次，需要联合研究这3种数据流的无线传输资源分配优化问题，以在有限的传输带宽限制下最大化系统的推理性能。这些问题的解决方案是实现空地一体化云-边模型协同演化架构的基础，并且需要深入研究。

借助这些支持性技术，所提出的空地一体化云-边模型协同演化架构能够显著扩展具备云模型支持的边缘AI在各种低空物联网场景中的应用，例如精准农业、目标搜索和灾区救援[25-26]。

## 2　系统模型

空地一体化云-边模型协同演化架构的基本系统模型包含一个地面云服务器和一个边缘无人机，可进行联合感知、边缘计算和通信，如图2所示。无人机配备了机载摄像头，负责捕捉视觉数据，随后与云服务器协作进行目标分类，以支持各种边缘AI应用，例如灾难响应和地理识别。

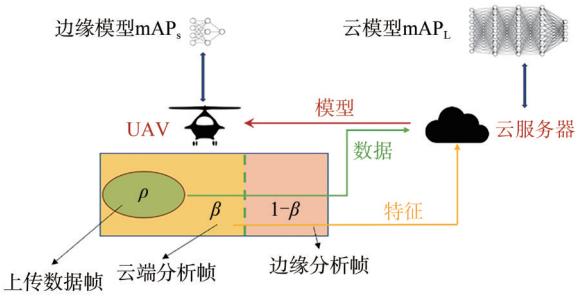



图2　空地一体化云-边模型协同演化架构的基本系统模型

由于受到能源和算力限制，无人机在使用机载边缘模型时难以独立执行高精度视觉分类任务。云服务器可以通过云模型计算和边缘模型更新协同支持无人机执行目标分类检测任务。云服务器连接到地面基站（BS, base station），并通过OTA网络，如新空口（NR, new radio）技术和长期演进（LTE, long-term evolution）技术与无人机通信。如第1节所述，无人机与云服务器之间通过OTA传输进行3种形式的数据交互，其中，模型流是下行传输，而特征流和数据流是上行传输。

假设无人机的机载摄像头以每秒$N$帧的频率捕捉图像，其中每帧包含$x$个像素，每个像素被量化

为$b$位。因此，无人机捕获视觉数据的速率为$D = N{\cdot}x{\cdot}b$。如图2所示，比例为$\beta$的帧通过OTA传输上传到云服务器进行视觉目标分类，而剩余比例为$1-\beta$的帧则由无人机通过机载边缘模型处理。将上传到云进行分析的帧的集合表示为$\Psi$，其中$|\Psi| = \beta N$。无人机首先提取$\Psi$中帧的特征，以便在云服务器上进行后续处理。设$\overline{F}$为每帧提取的特征的平均大小，则特征流的传输数据速率为

$$R_{\mathrm{F}} = \overline{F}{\cdot}\beta{\cdot}N \tag{1}$$

为了进一步提高云服务器的分类精度，可以将比例为$\rho$的帧的残差映射数据从无人机传输到云服务器。用$\Phi$表示被发送到云服务器的残差映射数据的帧的集合，其中$|\Phi| = \rho N$。由于残差映射数据需要与边缘无人机上提取的特征相结合以进行图像重构，因此集合$\Phi$仅包含在云服务器上分析的帧，即$\Phi \subseteq \Psi$。在有限的OTA传输带宽限制下，需要对残差映射数据进行适当量化，以充分利用上行的带宽资源。设$b = \{\hat{b}_i\}, \forall i \in \Phi$为残差映射数据的量化参数，其中$\hat{b}_i$表示帧$i$中每个像素的量化位数，该量化位数从一个离散值集合中选择，该集合记为$\Omega$。因此，数据流的传输数据速率可以表示为

$$R_{\mathrm{D}} = \sum_{i \in \Phi} x\hat{b}_i, \forall \hat{b}_i \in \Omega \tag{2}$$

特征流和数据流的数据传输速率之和不应超过无人机的上传容量，即

$$R_{\mathrm{F}} + R_{\mathrm{D}} \leqslant B_{\mathrm{u}}{\cdot}S_{\mathrm{u}} \tag{3}$$

其中，$B_{\mathrm{u}}$是无人机上行传输的带宽，$S_{\mathrm{u}}$是无人机上行传输的频谱效率，可以通过使用信道测量技术获得。

云服务器从边缘无人机处积累了大量的视觉特征数据和残差映射数据，随后可借此对无人机机载边缘模型进行更新，以提高其推理精度。假设模型更新数据以$M$ bit/s的平均开销传输到无人机，这代表了模型流的数据速率。需要注意的是，模型流的数据速率不能超过无人机的下载容量，即

$$M \leqslant B_{\mathrm{d}}{\cdot}S_{\mathrm{d}} \tag{4}$$

其中，$B_{\mathrm{d}}$是无人机下行传输的带宽，$S_{\mathrm{d}}$是无人机下行传输的频谱效率。同时，假设分配给无人机与基站通信的总带宽为$B$，并满足以下关系

$$B_{\mathrm{u}} + B_{\mathrm{d}} \leqslant B \tag{5}$$

本文中，上行传输的频谱效率$S_{\mathrm{u}}$和下行传输的频谱效率$S_{\mathrm{d}}$被视为具有任意值的常数。影响$S_{\mathrm{u}}$和$S_{\mathrm{d}}$的因素，如无人机轨迹、传输波束成形和干扰管理等，可以视为与本研究独立的设计。

## 3　问题建模与分解

为第2节描述的空地一体化云-边模型协同演化架构制定联合云-边平均精度最大化问题，然后将该问题分解为两个子问题以进行进一步分析。

### 3.1　联合云-边平均精度最大化问题

空地一体化云-边模型协同演化架构的联合云-边平均精度，由3个因素决定：云服务器的大模型平均精度，记作$\mathrm{mAP_L}$，无人机边缘小模型的平均精度，记作$\mathrm{mAP_S}$，以及在云服务器上分析的目标分类帧的比例$\beta$。为简化表示，将表示联合云-边平均精度的函数记为$\mathrm{mAP} = f\left(\mathrm{mAP_L}, \mathrm{mAP_S}, \beta\right)$。函数$f(\cdot)$的函数表达式和特性将在第4节中进行研究。

假设云服务器上的模型经过充分训练，性能稳定，因此变量$\mathrm{mAP_L}$的大小由从无人机传输的特征和残差映射数据的质量决定[27]。如文献[28]所研究的，基于特征推理的平均精度在通信开销远小于残差映射数据大小时会收敛到一个稳定水平。因此，考虑对集合$\Psi$中的每个帧采用具有固定开销的特征提取方法。云模型的平均精度，即$\mathrm{mAP_L}$，是残差映射数据传输比例$\rho$和相应的量化位数$\hat{b}_i$的函数，简化表示为$\mathrm{mAP_L} = g\left(\rho, \hat{b}_i\right)$，其中$\forall i \in \Phi$。函数$g(\cdot)$的表达式可能因任务和模型的不同而有所变化，可以通过文献[29]相关实验数据进行拟合。

边缘无人机能够获取所有帧的无损数据，其平均精度表现受到机载边缘模型推理准确性的影响，而这意味着云服务器对机载边缘模型的更新将会影响无人机的平均精度。从这一角度出发，机载边缘模型的平均精度，即$\mathrm{mAP_S}$，可表示为$\mathrm{mAP_S} = h(M)$，其中$M_{\min} \leqslant M \leqslant M_{\max}$，$M_{\min}$和$M_{\max}$分别是最小和最大的模型更新开销。函数$h(\cdot)$的表达式可能因不同任务而异，可以通过文献[26]的相关实验数据进行拟合。

为了最大化空地一体化云-边模型协同演化架构的平均精度，有必要将优化云-边任务分配、上行—下行带宽分配、残差映射数据传输设计以及模型更新开销设计联合起来。该问题可以表述为以下优化问题

$$\max_{\beta,\rho,b,B_d,B_u,M} \mathrm{mAP} = f\left(\mathrm{mAP}_\mathrm{L}, \mathrm{mAP}_\mathrm{S}, \beta\right) \tag{6}$$

$$\text{s.t. } \mathrm{mAP}_\mathrm{L} = g\left(\rho, \hat{b}_i\right) \tag{7}$$

$$0 \leqslant \rho \leqslant \beta \leqslant 1 \tag{8}$$

$$\hat{b}_i \in \Omega, \forall i \in \Phi \tag{9}$$

$$\mathrm{mAP}_\mathrm{S} = h(M) \tag{10}$$

$$M_{\min} \leqslant M \leqslant M_{\max} \tag{11}$$

$$R_\mathrm{F} + R_\mathrm{D} \leqslant B_\mathrm{u} \cdot S_\mathrm{u} \tag{12}$$

$$M \leqslant B_\mathrm{d} \cdot S_\mathrm{d} \tag{13}$$

$$B_\mathrm{u} + B_\mathrm{d} \leqslant B \tag{14}$$

目标函数式(6)表示最大化空地一体化云-边模型协同演化架构的平均精度，这是关于变量$\mathrm{mAP}_\mathrm{L}$、$\mathrm{mAP}_\mathrm{S}$和$\beta$的函数。约束条件式(7)描述了云模型的平均精度函数。约束条件式(8)约束了传输残差映射数据的帧比例不得超过在云服务器上分析的帧比例。约束条件式(9)描述了残差映射数据传输的量化约束。约束条件式(10)表示了无人机机载边缘模型的平均精度函数，而约束条件式(11)则对边缘模型更新的开销施加了限制。最后，约束条件式(12)~式(14)分别涉及特征流、数据流和模型流的传输数据大小的限制。

式(6)在直接求解时面临显著挑战，主要有两个原因。首先，它是一个混合整数规划问题，包含离散变量$b$和连续变量$\beta$、$\rho$、$B_\mathrm{d}$、$B_\mathrm{u}$、$M$，这是一个NP难问题；其次，不能确保该问题的凸性，因为实验拟合函数$g(\cdot)$和$h(\cdot)$的凸性尚不确定。在接下来的分析中，将式(6)分解为两个子问题：数据流设计子问题和特征/模型流设计子问题，并依次分析这两个子问题。通过这样的分解，离散变量$b$可以与参数$\beta$、$B_\mathrm{d}$、$B_\mathrm{u}$、$M$分离，从而简化式(6)中的复杂表述，并且对函数$g(\cdot)$和$h(\cdot)$凸性的讨论可以解耦为两个独立的子问题分别进行讨论和解决。

### 3.2　问题分解

1) 数据流设计子问题

在数据流设计子问题中，考虑上行数据流，它影响云服务器上模型推理的平均精度。此问题包括残差映射数据传输的帧比例$\rho$，以及相应的残差映射数据的量化位数$b$的设计。与边缘模型更新、任务分配和传输资源分配相关的参数在此子问题中被视为固定值，不进行优化。本子问题的目标是通过优化具有残差映射数据传输的帧比例及其对应的量化位数，最大化云模型推理的平均精度。第1个子问题可以表述如下

$$\max_{\rho,b} \mathrm{mAP}_\mathrm{L} \tag{15}$$

$$\text{s.t. } \mathrm{mAP}_\mathrm{L} = g\left(\rho, \hat{b}_i\right), \forall i \in \Phi \tag{16}$$

$$0 \leqslant \rho \leqslant \beta \tag{17}$$

$$\hat{b}_i \in \Omega, \forall i \in \Phi \tag{18}$$

$$R_\mathrm{F} + R_\mathrm{D} \leqslant B_\mathrm{u} \cdot S_\mathrm{u} \tag{19}$$

约束条件式(16)~式(19)均与数据流相关，这些约束已经在第3.1节中介绍过。

2) 特征/模型流设计子问题

假设与数据流相关的参数已经通过子问题式(15)的求解得到优化，在这个子问题中，将对特征流和模型流的相关参数进行设计。第2个子问题的目标是最大化云模型和边缘模型的联合平均精度。为此，需要对分配给云服务器的帧比例、上行和下行传输的带宽分配以及边缘模型更新的开销进行优化来最大化联合云-边平均精度。第2个子问题可以表述如下

$$\max_{\beta,B_d,B_u,M} \mathrm{mAP} = f\left(\mathrm{mAP}_\mathrm{L}, \mathrm{mAP}_\mathrm{S}, \beta\right) \tag{20}$$

$$\text{s.t. } 0 \leqslant \rho \leqslant \beta \leqslant 1 \tag{21}$$

$$\mathrm{mAP}_\mathrm{S} = h(M) \tag{22}$$

$$M_{\min} \leqslant M \leqslant M_{\max} \tag{23}$$

$$R_\mathrm{F} + R_\mathrm{D} \leqslant B_\mathrm{u} \cdot S_\mathrm{u} \tag{24}$$

$$M \leqslant B_\mathrm{d} \cdot S_\mathrm{d} \tag{25}$$

$$B_\mathrm{u} + B_\mathrm{d} \leqslant B \tag{26}$$

其中，约束条件式(21)~式(26)均与云-边计算的联合优化相关，这些约束已经在第3.1节中介绍过。

## 4　空地一体化云-边模型协同演化架构联合云-边平均精度最大化问题的解决方案

本节解决式(6)中的联合云-边平均精度最大化问题。子问题式(15)和式(20)分别在第4.1节和第4.2节中解决。对问题式(6)的整体解决方案将在第4.3节中描述。

### 4.1　数据流设计子问题的解决方案

对数据流进行设计，并解决子问题式(7)。其中，与特征流和模型流相关的参数可视为常数。由于每帧的量化位数可以不同，云模型在分析不同帧时的平均精度也可能不同。用$\mathrm{mAP}_\mathrm{L}^i$表示云模型在分析第$i$帧时得到的平均精度（这里某帧的平均精度代表的是衡量推理模型准确性的一个指标，和式(13)中的定义不同），为了解决子问题式(15)，首先给出关于约束条件式(16)中函数$g(\cdot)$的函数特性

与假设。

**备注 1**　云模型对第 $i$ 帧的平均精度，即 $\mathrm{mAP}_{\mathrm{L}}^{i}$，随着其残差映射数据的量化位数 $\hat{b}_i$ 单调递增。

**假设 1**　云模型对第 $i$ 帧的平均精度，即 $\mathrm{mAP}_{\mathrm{L}}^{i}$，是量化位数 $\hat{b}_i$ 的凹函数，同时 $\hat{b}_i$ 被视为一个连续变量，且 $\hat{b}_i \in \Omega$。

**假设 2**　云模型对第 $i$ 帧的平均精度，即 $\mathrm{mAP}_{\mathrm{L}}^{i}$，是量化位数 $\hat{b}_i$ 的凹函数，同时 $\hat{b}_i$ 被视为一个连续变量，且 $\hat{b}_i \in \Omega \cup \{0\}$，其中 $\hat{b}_i=0$ 对应于没有残差映射数据传输的情况。

备注 1 强调了精确的残差映射数据有助于提高云服务器的平均精度，这在直观上是可以理解的。假设 1 是基于对多个数据集的各种实验[29-31]观察得出的，虽然缺乏理论证明，但在当前大多数研究中，假设 1 是成立的。因此，假设 1 对于现有绝大多数的基于视觉的分类任务来说是有效的。假设 2 是假设 1 的扩展陈述，涵盖了帧的残差映射数据未传输到云服务器的情况，此时可以视作帧对应的残差映射数据的量化位数为 0。在这种情况下，只有提取的视觉特征被发送到云服务器作为云模型的输入。

然而，需要特别指出的是，备注 1 和假设 1 并不能确保子问题式(15)的凸性，因为 $\mathrm{mAP}_{\mathrm{L}}^{i}$ 和 $\mathrm{mAP}_{\mathrm{L}}$ 并不等同。接下来，将进一步给出两个与 $\mathrm{mAP}_{\mathrm{L}}$ 相关的定理，为解决子问题式(7)提供理论基础。

**定理 1**　在没有离散量化位数约束条件式(18)的情况下，最大化 $\mathrm{mAP}_{\mathrm{L}}$ 的解满足 $\hat{b}_1=\hat{b}_2=\cdots=\hat{b}_i$，其中 $\forall i \in \Phi$。

**证明**　见附录第 7.1 节。

**定理 2**　当满足假设 2 时，集合 $\Psi$ 中所有帧的残差映射数据都应以相同的量化位数发送到云服务器，即 $\hat{b}_1=\hat{b}_2=\cdots=\hat{b}_i$，其中 $\forall i \in \Psi$。

**证明**　见附录第 7.2 节。

根据定理 1 和定理 2，子问题式(15)可以按以下步骤求解。在式(15)中，变量 $R_{\mathrm{F}}$、$B_{\mathrm{u}}$ 和 $S_{\mathrm{u}}$ 是已知的，因此约束条件式(19)可以转换为

$$\sum\nolimits_{i \in \Phi} x\hat{b}_i \leqslant B_{\mathrm{u}} \cdot S_{\mathrm{u}} - R_{\mathrm{F}}$$

当满足假设 2 时，首先设定 $\rho=\beta$ 且 $\hat{b}_1^{\mathrm{opt}}=\hat{b}_2^{\mathrm{opt}}=\cdots=\hat{b}_i^{\mathrm{opt}}=\dfrac{B_{\mathrm{u}} \cdot S_{\mathrm{u}} - R_{\mathrm{F}}}{|\Psi|}, \forall i \in \Psi$。如果 $\hat{b}_1^{\mathrm{opt}}$ 的值不满足约束条件式(18)，则从 $\hat{b}_{\mathrm{l}}$ 和 $\hat{b}_{\mathrm{u}}$ 中，根据实际情况按比例选择 $b$ 的元素值。其中 $\hat{b}_{\mathrm{l}}$ 和 $\hat{b}_{\mathrm{u}}$ 是最接近 $\hat{b}_1^{\mathrm{opt}}$ 的两个值，满足 $\hat{b}_{\mathrm{l}} < \hat{b}_1^{\mathrm{opt}} < \hat{b}_{\mathrm{u}}$ 且 $\hat{b}_{\mathrm{u}}$、$\hat{b}_{\mathrm{l}} \in \Omega \cup \{0\}$。

接下来，按比例分配量化位数：比例为 $\left\lceil \dfrac{\hat{b}_{\mathrm{opt}} - \hat{b}_{\mathrm{l}}}{\hat{b}_{\mathrm{u}} - \hat{b}_{\mathrm{l}}} \right\rceil$ 的帧使用 $\hat{b}_{\mathrm{u}}$ 位数对残差映射数据进行量化，而比例为 $\left\lceil \dfrac{\hat{b}_{\mathrm{u}} - \hat{b}_{\mathrm{opt}}}{\hat{b}_{\mathrm{u}} - \hat{b}_{\mathrm{l}}} \right\rceil$ 的帧使用 $\hat{b}_{\mathrm{l}}$ 位数进行量化，其中 $\lceil \cdot \rceil$ 是获取最接近的整数的函数。

在不满足假设 2 的情况下，提出了一种基于启发式的方法来解决子问题式(15)。这种启发式方法需要计算残差映射数据传输的最大数据速率，记为 $B_{\mathrm{u}} \cdot S_{\mathrm{u}} - R_{\mathrm{F}}$。同时，引入了平均精度增量效率的概念，用来表示每个帧在数据量化位数单位增量下的平均精度增加量。该策略优先将剩余的通信资源分配给平均精度增量效率最高的帧。此分配迭代过程将一直持续，直到所有的通信资源都得到有效分配为止。

### 4.2　特征/模型流设计子问题的解决方案

优化与特征流和模型流相关的参数，以解决子问题式(20)。为了帮助解决优化问题，引入一个定理，该定理概述了涉及云模型和边缘模型的联合云-边平均精度。

**定理 3**　云模型和边缘模型的联合云-边平均精度是云模型和边缘模型的召回率—精度对的函数（召回率—精度对及相关参数的含义在附录第 7.1 节对定理 1 的证明中有详细解释），可以表示为

$$\mathrm{mAP}=\frac{1}{2}\cdot\sum\nolimits_{k=1}^{K}\left(\frac{1}{\dfrac{\beta}{r_{\mathrm{L}}^{k}}+\dfrac{1-\beta}{r_{\mathrm{S}}^{k}}}-\frac{1}{\dfrac{\beta}{r_{\mathrm{L}}^{k-1}}+\dfrac{1-\beta}{r_{\mathrm{S}}^{k-1}}}\right)\cdot\left(\frac{1}{\dfrac{\beta}{p_{\mathrm{L}}^{k}}+\dfrac{1-\beta}{p_{\mathrm{S}}^{k}}}+\frac{1}{\dfrac{\beta}{p_{\mathrm{L}}^{k-1}}+\dfrac{1-\beta}{p_{\mathrm{S}}^{k-1}}}\right) \tag{27}$$

其中，$r_{\mathrm{L}}^{k}$ 和 $p_{\mathrm{L}}^{k}$ 分别是云模型在第 $k$ 个交并比（IoU，intersection over union）下的召回率和精度值，$r_{\mathrm{S}}^{k}$ 和 $p_{\mathrm{S}}^{k}$ 分别是边缘模型在第 $k$ 个交并比下的召回率和精度值。

**证明**　见附录第 7.3 节。

定理 3 证明了联合云-边平均精度与精度—召回值之间的关系。然而，式(20)中的优化变量并未直接与精度—召回值相关联。在随后的讨论中，将基于定理 3，深入探讨 mAP、$\mathrm{mAP}_{\mathrm{L}}$ 和 $\mathrm{mAP}_{\mathrm{S}}$ 之间的关系，以进一步深化这些变量之间的联系。

**定理 4**　云模型和边缘模型的平均精度与系统

的联合云-边平均精度满足以下关系

$$\mathrm{mAP} \geqslant \frac{\mathrm{mAP_L}\cdot\mathrm{mAP_S}}{(1-\beta)\mathrm{mAP_L}+\beta\mathrm{mAP_S}} \tag{28}$$

**证明**　见附录第7.4节。

在定理4中，推导了mAP的下界，作为$\mathrm{mAP_S}$、$\mathrm{mAP_L}$和$\beta$的函数。为了进一步解决优化问题，本文目标是在特定条件下建立mAP与$\mathrm{mAP_S}$、$\mathrm{mAP_L}$和$\beta$之间的闭式关系。接下来将在云模型和边缘模型的平均精度性能处于相同量级的场景下[29-31]，给出mAP的闭形式表达式。

**定理5**　当约束条件$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$和$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$对于所有$1 \leqslant k \leqslant K$始终成立时，本架构的联合云-边平均精度可以近似表示为

$$\mathrm{mAP} \approx \frac{\mathrm{mAP_L}\cdot\mathrm{mAP_S}}{(1-\beta)\mathrm{mAP_L}+\beta\mathrm{mAP_S}} \tag{29}$$

**证明**　见附录第7.5节。

即使在不严格满足约束条件$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$，$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$，$\forall 1 \leqslant k \leqslant K$的情况下，式(29)仍然可以作为平均精度的下界，用于解决优化问题式(20)。通过计算其Hessian矩阵$\boldsymbol{H}$，可以确定式(29)关于$\mathrm{mAP_L}$和$\mathrm{mAP_S}$的凸性，即

$$\boldsymbol{H}=\begin{bmatrix} \dfrac{\partial^2(\mathrm{mAP})}{\partial(\mathrm{mAP_L})^2} & \dfrac{\partial^2(\mathrm{mAP})}{\partial(\mathrm{mAP_L})\partial(\mathrm{mAP_S})} \\ \dfrac{\partial^2(\mathrm{mAP})}{\partial(\mathrm{mAP_S})\partial(\mathrm{mAP_L})} & \dfrac{\partial^2(\mathrm{mAP})}{\partial(\mathrm{mAP_S})^2} \end{bmatrix}= \frac{2\beta(1-\beta)}{\left((1-\beta)\mathrm{mAP_L}+\beta\mathrm{mAP_S}\right)^3}\times \begin{bmatrix} -(\mathrm{mAP_S})^2 & \mathrm{mAP_S}\cdot\mathrm{mAP_L} \\ \mathrm{mAP_L}\cdot\mathrm{mAP_S} & -(\mathrm{mAP_L})^2 \end{bmatrix} \tag{30}$$

如式(30)所示，Hessian矩阵的一阶和二阶主子式均为非正值。因此，式(29)关于$\mathrm{mAP_L}$和$\mathrm{mAP_S}$是一个凹函数。这意味着在优化问题式(20)中，联合云-边平均精度是关于云模型和边缘模型的平均精度，即$\mathrm{mAP_L}$和$\mathrm{mAP_S}$的凹函数，因此可以采用凸优化方法求解该问题。

在分析了$f(\cdot)$的凸性之后，进一步研究$\mathrm{mAP_L}$相对于$R_{\mathrm{F}}+R_{\mathrm{D}}$的凸性，以考察子问题式(20)的凸性。正如在第4.1节所研究的那样，$\mathrm{mAP_L}$是关于$R_{\mathrm{D}}$的凹函数。由于$R_{\mathrm{F}}$的值不会影响$\mathrm{mAP_L}$，因此$\mathrm{mAP_L}$可以被视为关于$R_{\mathrm{F}}+R_{\mathrm{D}}$的凹函数。考虑上行传输数据大小$R_{\mathrm{F}}+R_{\mathrm{D}}$是上行传输带宽$B_{\mathrm{u}}$的线性函数，因此$\mathrm{mAP_L}$相对于$B_{\mathrm{u}}$也是凹函数。

根据定理5可以观察到，当满足$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$，$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$，$\forall 1 \leqslant k \leqslant K$时，mAP的表达式相对于$\mathrm{mAP_L}$和$\mathrm{mAP_S}$是凹的。此外，约束条件式(22)中的函数$h(\cdot)$已在现有研究中拟合为凹函数[24]。在这些条件下，子问题式(20)是一个关于变量$\beta$、$B_{\mathrm{d}}$、$B_{\mathrm{u}}$和$M$的凹函数优化问题，可以使用凸优化方法来解决。即使在$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$，$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$，$\forall 1 \leqslant k \leqslant K$这一限制条件不被满足的情况下，也可以通过式(29)来对函数$f(\cdot)$进行近似处理，从而获得一个下界解。

### 4.3　总体算法与复杂度分析

首先总结用于解决空地一体化云-边模型协同演化架构设计问题式(6)的总体算法。

解决联合云-边平均精度最大化问题式(6)的方法在算法1中概述。首先，对于可行的$B_{\mathrm{u}}$值进行遍历，通过求解子问题式(15)，得到与$B_{\mathrm{u}}$相关的$\mathrm{mAP_L}$的最优解。随后，对于每个子问题式(15)对应的最优解，解决子问题式(20)并比较得出变量$\beta$、$B_{\mathrm{d}}$、$B_{\mathrm{u}}$和$M$的最优解。然后将对应的$B_{\mathrm{u}}$代入子问题式(15)，得出$\rho$的最终解。当条件$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$，$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$，$\forall 1 \leqslant k \leqslant K$成立时，定理5适用，并可以基于此得到最优解。如若条件不满足，可以考虑通过$\mathrm{mAP} \approx \frac{\mathrm{mAP_L}\cdot\mathrm{mAP_S}}{(1-\beta)\mathrm{mAP_L}+\beta\mathrm{mAP_S}}$获得次优解。根据定理4，真实的mAP值不会低于$\frac{\mathrm{mAP_L}\cdot\mathrm{mAP_S}}{(1-\beta)\mathrm{mAP_L}+\beta\mathrm{mAP_S}}$，并且通过算法1获得的解可以作为问题式(6)的解的一个下界。

**算法1**　空地一体化云-边模型协同演化架构的云/边模型设计

**输入：**变量$B, S_{\mathrm{d}}, \Omega, M_{\min}, M_{\max}$，函数$g(\cdot), h(\cdot)$；

遍历并解子问题式(7)以获得关于变量$B_{\mathrm{u}}$的最大$\mathrm{mAP_L}$的函数表达式；

**if** $r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$，$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_L^k$ $\forall 1 \leqslant k \leqslant K$；

解凸优化问题式(8)并遍历解以获得最优的$\beta$、$B_{\mathrm{d}}$、$B_{\mathrm{u}}$、$M$；

**else**

以mAP的下界解出次优的$\beta$、$B_{\mathrm{d}}$、$B_{\mathrm{u}}$、$M$；

**end if**

通过子问题式(8)中得到的$B_{\mathrm{u}}$，解出对应的$\rho$；

**输出：**任务分配比例$\beta$,数据量化变量$\rho$、$b$，通

信变量$B_{\mathrm{d}}$、$B_{\mathrm{u}}$，模型更新变量$M$。

对给出的算法复杂度进行分析，以说明本系统在大规模低空物联网场景下应用的可行性。

1) 时间复杂度

对于子问题式(15)，假设其求解复杂度为$O(f(n))$，$n$代表优化变量的维度。根据算法1，求解过程需要对每个可行的$B_{\mathrm{u}}$给出最优情况下的$\mathrm{mAP}_{\mathrm{L}}$。假设求解过程中对$B_{\mathrm{u}}$的遍历次数为$m$，则总的遍历与求解子问题式(15)的时间复杂度可以表示为

$$T_7 = O(m \cdot f(n))$$

由于在第4.1节中的定理2给出了在给定$B_{\mathrm{u}}$时最大化$\mathrm{mAP}_{\mathrm{L}}$的闭式函数表达式，因此可以直接求解子问题式(15)，无须占用更多计算资源。换言之，可以认为$f(n)=1$成立，从而使得

$$T_7 = O(m \cdot f(n)) = O(m)$$

在每个子问题式(15)解决后，对应的最优解将继续输入子问题式(20)中完成最终的求解。假设单个子问题式(20)优化求解的时间复杂度为$O(g(n))$，则总的优化复杂度表示为

$$T_8 = O(m \cdot g(n))$$

在实践中，一般采取内点法求解优化问题式(20)，在大规模数据的情况下，$g(n)$的求解复杂度为$O(n^{3.5} \cdot \log(1/s))$，$s$代表目标的优化精度。同时，为了得到唯一的最优解，需要对每个遍历初始值对应的最终解进行比较，这也会增加$O(m)$的复杂度。综上，求解的总时间复杂度为

$$T = O(m) + O(mn^{3.5}\log(1/s)) + O(m) = O(mn^{3.5}\log(1/s))$$

由于在大规模低空物联网场景下，$n$的取值会快速增长，因此有限值$m$和$\log(1/s)$对系统时间开销的影响会逐渐减小，保证了优化问题在有限多项式时间内可以解决。

2) 空间复杂度

子问题式(15)需要对每个遍历值存储对应的最优解，对应的空间开销为$O(m)$；子问题式(20)在遍历的基础上，还需要对优化过程中产生的临时变量和解进行存储，空间复杂度为$O(mn)$。

因此，总的空间复杂度可以表示为

$$S = O(mn) + O(m) = O(mn)$$

在大规模低空物联网场景下，算法的空间开销与变量规模呈线性关系，对有关应用的存储需求提供了保障。

## 5　仿真结果

本节评估了所提出的空地一体化云-边模型协同演化架构的性能，该架构包括联合任务分配、传输资源分配、传输数据量化位数优化和边缘模型更新设计。为了进行比较，将所提出的架构与4种基线架构进行对比：集中式云模型架构、分布式边缘模型架构、对空地一体化云-边模型协同演化架构的穷尽搜索方法以及基于带宽分配的云-边模型协同演化架构。

1) 集中式云模型架构：在此架构中，边缘无人机没有分类能力。它将所有帧的提取特征和量化的残差映射数据传输到云服务器进行目标分类。帧的量化位数由OTA上行传输的带宽和频谱效率决定。

2) 分布式边缘模型架构：在此架构中，边缘无人机在本地执行分类任务。云服务器仅根据OTA传输能力向边缘无人机传输模型更新。

3) 穷尽搜索：此架构采用了所提出的空地一体化云-边模型协同演化架构。任务分配、传输资源分配、传输数据量化位数优化和边缘模型更新设计是通过对$10^8$种候选变量组合进行枚举来选择的，从而实现联合云-边平均精度最大化。该性能可以视为所提出架构的上限。

4) 动态带宽云-边模型协同演化架构：此架构采用了所提出的空地一体化云-边模型协同演化架构。其中上行和下行的带宽分配比例和实际上、下行信道的频谱效率成反比关系，其余的优化变量均通过算法1进行优化求解得到。

在本次仿真中，以基于无人机视觉的目标检测任务为例，相关仿真参数见表1。实验数据是从基于VisDrone数据集的目标检测任务上获取的[32]。边缘无人机上的模型为YOLOv8s，云服务器上的模型为YOLOv8x2[12]。边缘无人机的训练与模型更新遵循文献[33]中提出的方法进行。表示$\mathrm{mAP}_{\mathrm{L}}$与量化位数关系的函数$g(\cdot)$是根据文献[29]中提出的算法与实验结果拟合而成的。同样，表示$\mathrm{mAP}_{\mathrm{S}}$与模型更新开销关系的函数$h(\cdot)$是根据文献[33]中提出的算法与实验结果拟合而成的。需要注意的是，本文所提出的系统架构及其相应的优化方案可以应用于具有不同模型和数据集的各种任务。

**表1　　仿真参数**

| 参数 | 取值 |
|---|---|
| 每秒生成的感知帧数目$N$ | 10 |
| 每帧的像素数目$x$ | $10^7$ |
| 提取特征的平均大小$F$/(kbit·s$^{-1}$) | 0.86 |
| 空中下载带宽$B$/MHz | 10 |
| 上行频谱效率$S_u$/(bit/(s·Hz)) | 2.55 |
| 下行频谱效率$S_d$/(bit/(s·Hz)) | 5 bit |
| 最大模型更新开销$M_{max}$/(Mbit·s$^{-1}$) | 40 |
| 最小模型更新开销$M_{min}$/(kbit·s$^{-1}$) | 40 |

在图3中，展示了不同总传输带宽下空地一体化云-边模型协同演化架构的联合云-边平均精度，即mAP。结果显示，随着传输带宽的增加，平均精度也随之增加，并在总带宽超过60 MHz时收敛到一个稳定值。收敛的平均精度值表明，在带宽足够大的情况下，所有帧都可以上传到云服务器进行分析。当带宽小于5 MHz时，所提出的架构的性能与分布式边缘模型架构相当，此时由于数据传输能力有限，大部分任务在边缘模型上执行。当带宽大于20 MHz时，所提出架构的联合云-边平均精度与云模型架构的平均精度接近，大部分残差映射数据被发送到云服务器进行分析。作为集中式云模型计算和分布式边缘模型计算的动态组合，所提出的空地一体化云-边模型协同演化架构通过动态调整通信资源以获得最佳的$\mathrm{mAP_L}$和$\mathrm{mAP_S}$，其联合云-边平均精度始终优于集中式云模型架构和分布式边缘模型架构。实验结果显示，本架构与穷尽搜索方法之间的平均精度性能差距在所有情况下均小于0.5%。

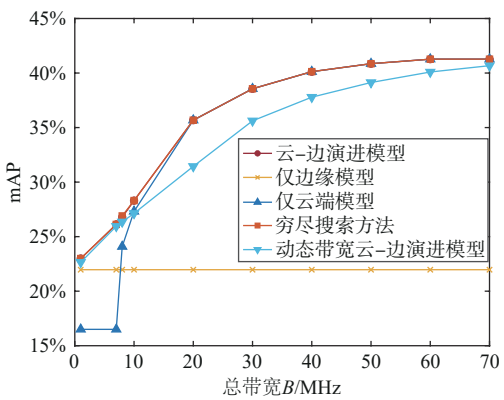



图3　不同总传输带宽下空地一体化云-边模型协同演化架构的联合云-边平均精度

在图4中，评估了在边缘无人机每秒捕获图片帧数不同的情况下，系统的联合云-边平均精度。对于固定的通信带宽，每秒捕获帧数越多，每帧的平均残差映射数据传输量就越少，从而导致平均精度下降。在分布式边缘模型架构中，平均精度仅由模型更新决定，与每秒捕获的帧数无关，因此在不同的$N$值下，其平均精度是一个常数。对于所提出的空地一体化云-边模型协同演化架构，尽管平均精度随着$N$值的增大而下降，但其下限是边缘模型的平均精度。实验结果显示，所提出架构的性能在不同的$N$值下始终优于集中式云模型架构、分布式边缘模型架构和动态带宽云-边演进模型架构，并且与穷尽搜索方法的平均精度性能差距始终小于0.65%。

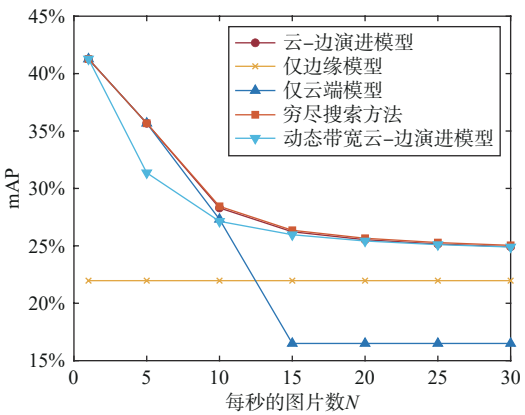



图4　每秒捕获图片帧数不同的情况下，系统的联合云-边平均精度

图5展示了上行和下行传输的频谱效率对空地一体化云-边模型协同演化架构的联合云-边平均精度的影响。上行频谱效率$S_u$和下行频谱效率$S_d$都与平均精度呈正相关。由于上行残差映射数据的开销比下行模型更新的开销更大，因此$S_u$对平均精度的影响比$S_d$更显著，如图5所示。当$S_u$超过10 bit/(s·Hz)时，平均精度不再受到$S_d$的影响。这是因为在这种情况下，所有帧都被发送到云服务器进行分析，使得机载边缘模型的更新变得不再必要。

图6展示了在不同总传输带宽下，所提出架构中的数据流、模型流和特征流的开销。由于特征流的开销显著低于数据流和模型流的开销，使得在较低的上行传输带宽下进行云模型计算成为可能。如图6(a)所示，当总带宽小于2 MHz时，模型流在OTA传输开销中占主导地位。这表明，在通信能力较差的情况下，大多数任务由机载边缘模型执行，强调了在这种情况下$\mathrm{mAP_S}$相对于$\mathrm{mAP_L}$的重要性。

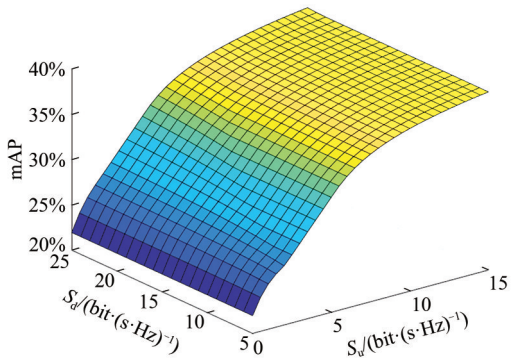
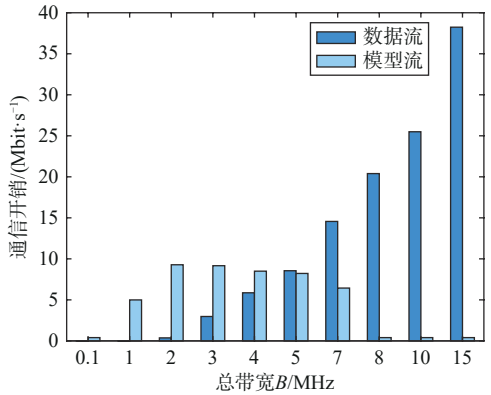
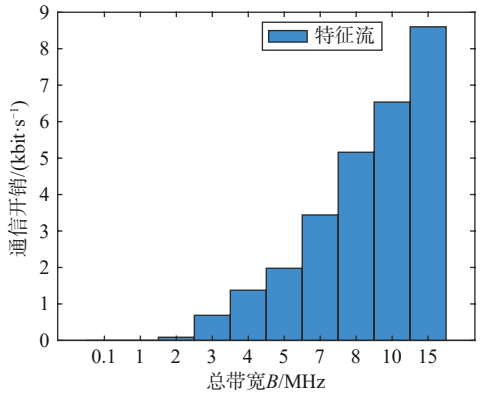



图5　上行和下行传输的频谱效率对平均精度的影响




(a) 数据/模型流的开销




(b) 特征流的开销

图6　在不同总传输带宽下各个流的通信开销

当带宽超过 2 MHz 时，数据流的开销显著增加，并且在带宽超过 8 MHz 时，数据流的开销变得远大于模型流的开销。这表明，随着带宽的增加，大多数目标检测任务将由云服务器处理。当带宽超过 5 MHz 时，由于在云服务器上分析的帧比例减少，模型流的开销开始减少。因此，$mAP_S$ 对平均精度的影响变得不如 $mAP_L$ 显著。此外，图 6(b)表明，由于传输更多帧的特征，特征流的开销随着总带宽 $B$ 的增加而稳步增长。

在图 7 中，研究了总带宽对在云服务器上分析的帧比例 $\beta$ 的影响。如第 4 节分析的那样，$\beta$ 的最优解受到云服务器和边缘无人机的平均精度的影响。因此，研究了 4 种不同的模型配置情况。为了进行这项分析，假设随着计算能力的提高，边缘无人机和云服务器都可以升级到更大的模型架构：边缘无人机使用 YOLOv8m 模型，云服务器使用 YOLOv8x3 模型。结果表明，在特定的总带宽下，云服务器更高的分类精度对应于更大的 $\beta$ 值。然而，当总带宽 $B$ 足够大（超过 40 MHz）时，只要云服务器的平均精度大于边缘无人机的平均精度，所有情况下的 $\beta$ 值都会接近 1。相反，当总带宽 $B$ 非常小时（不超过 1 MHz），所有情况下的 $\beta$ 值都趋向于 0。

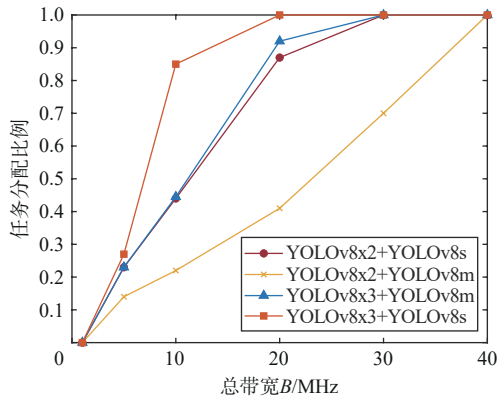



图7　总带宽对在云服务器上分析的帧比例 $\beta$ 的影响

在表 2 中，评估了在无人机每秒捕获图片帧数不同的情况下，对应得出的一些关键变量的数值。当 $N = 10$ 时，上行带宽是主导因素，所有任务此时都分配给云模型。随着每秒捕获帧数的增加，分配给下行传输的带宽逐渐增加，更多的分类任务被分配给边缘模型。这也符合变量 $\beta$ 的趋势，即随着

**表 2　不同 $N$ 值对应的变量取值**

| 每秒帧数 $N$ | 上行带宽 $B_u$/MHz | 下行带宽 $B_d$/MHz | 任务分配比例 $\beta$ | 模型更新开销 $M$/(Mbit·s$^{-1}$) | 残差数据平均量化位数 $\bar{b}$/(bit·pixel$^{-1}$) |
|---|---|---|---|---|---|
| 10 | 10 | 0 | 0.76 | $M_{min}$ | 0.335 4 |
| 15 | 8.13 | 1.87 | 0.37 | 9.35 | 0.373 5 |
| 20 | 7.70 | 2.30 | 0.26 | 11.5 | 0.377 5 |
| 25 | 7.45 | 2.55 | 0.20 | 12.75 | 0.379 9 |

$N$的增加，$\beta$逐渐减小。随着$N$的增加，模型更新的开销从最小值迅速增加到最大值，这与$\mathrm{mAP_S}$的较高值一致。残差映射数据的平均量化位数值随$N$的增加变化不大，表明在空地一体化云-边模型协同演化架构中，$\mathrm{mAP_L}$在不同的$N$值下基本保持稳定。

## 6　结束语

本文提出了一种新的空地一体化云-边模型协同演化架构，使得边缘模型和云模型能够并行进行数据分析，并可以通过地面云服务器辅助更新无人机边缘模型。本文推导了该架构的联合云-边平均精度下界的闭式表达式，并通过联合任务分配、上行—下行传输资源分配、传输数据量化位数和边缘模型更新设计解决了联合云-边平均精度最大化问题。仿真结果强调了该架构在不同通信带宽和数据规模下都具有优异的平均精度性能，优于集中式云模型架构和分布式边缘模型架构，总结如下。

1)所提架构的性能提升源于通过模型演进和数据上传，动态调整了边缘模型和云模型的平均精度，从而最大化整个架构的联合云-边平均精度。

2)在通信带宽较小且数据量较大的情况下，边缘模型处理大部分任务，此时大部分带宽被分配用于边缘模型的更新。

3)在通信带宽较大且数据量较小的情况下，云模型处理大部分任务，此时大部分带宽被分配用于残差映射数据的上传。

同时，本文仅对边缘存在单个用户的情况进行了讨论，而在大规模低空物联网场景下，往往会出现大量边缘用户。因此，未来仍有大量研究工作需要在此基础上进行，部分未来工作展望如下。

1)在边缘多用户的情况下，如何优化系统的资源分配和任务调度，以提高系统推理的综合性能。

2)在实际大规模用户场景下，如何减轻用户间的通信干扰，提升云-边协同工作的稳定性。

3)考虑实际的能耗限制，如何在用户的推理性能和能耗水平之间做出权衡。

## 7　附录—有关引理的证明

### 7.1　定理1的证明

根据平均精度的定义，其值对应于精确率—召回率曲线（PRC, precision rate-recall rate curve）下的面积，该曲线是从一组在不同交并比阈值下的精确率—召回率配对值得出的。精确率是通过真正例（TP, true positive）样本数与真正例和假正例（FP, false positive）样本数之和的比值来计算的，而召回率则是通过真正例样本数与真正例和假负例（FN, false negative）样本数之和的比值来确定的。

假设帧$i$和帧$j$在其残差映射数据传输中使用不同的量化位数，分别记为$\hat{b}_i$和$\hat{b}_j$。云服务器上接收到以$\hat{b}_i$和$\hat{b}_j$为量化位数的数据时，在此基础上执行分类任务的平均精度分别表示为$\mathrm{mAP_L}(\hat{b}_i)$和$\mathrm{mAP_L}(\hat{b}_j)$。以帧$i$为例，$\mathrm{mAP_L}(\hat{b}_i)$的值可以近似表示如下

$$\mathrm{mAP_L}(\hat{b}_i)\approx\sum_{k=1}^{K}\left(r_i^k-r_i^{k-1}\right)\left(p_i^k+p_i^{k-1}\right)/2 \tag{31}$$

其中，$r_i^k$和$p_i^k$分别是在第$k$个交并比阈值下的召回率和精确率，平均精度与PRC如图8所示。

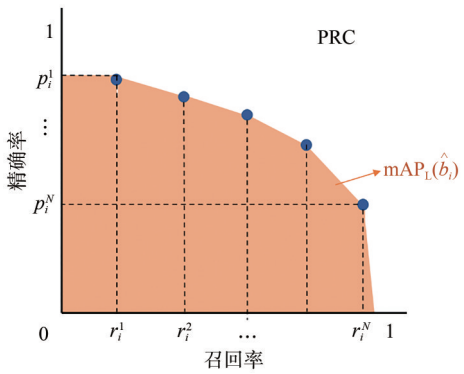



图8　平均精度与PRC曲线关系示意图

根据平均精度的定义，$r_i^k$和$p_i^k$可以分别表示为$r_i^k=\dfrac{\mathrm{TP}_i^k}{\mathrm{TP}_i^k+\mathrm{FN}_i^k}$和$p_i^k=\dfrac{\mathrm{TP}_i^k}{\mathrm{TP}_i^k+\mathrm{FP}_i^k}$，其中$\mathrm{TP}_i^k$是真正例样本的数量，$\mathrm{FP}_i^k$是假正例样本的数量，$\mathrm{FN}_i^k$是假负例样本的数量。通过定义$x_i^k=\dfrac{\mathrm{FP}_i^k}{\mathrm{TP}_i^k}$和$y_i^k=\dfrac{\mathrm{FN}_i^k}{\mathrm{TP}_i^k}$，可以得到$r_i^k=\dfrac{1}{1+y_i^k}$和$p_i^k=\dfrac{1}{1+x_i^k}$。类似地，对于帧$j$，变量$r_j^k$和$p_j^k$可以分别表示为$r_j^k=\dfrac{1}{1+y_j^k}$和$p_j^k=\dfrac{1}{1+x_j^k}$。

综合考虑帧$i$和帧$j$的平均精度性能，在第$k$个交并比阈值下的召回率值为$r^k=\dfrac{2}{2+y_i^k+y_j^k}$。接下来，将$r^k$的值与帧$i$和帧$j$的召回率平均值进行比较，具体如下

$$\frac{r_i^k + r_j^k}{2} - r^k = \frac{1}{1 + y_i^k} + \frac{1}{1 + y_j^k} - \frac{2}{2 + y_i^k + y_j^k} = \frac{\left(1 + y_i^k\right)^2 + \left(1 + y_j^k\right)^2}{\left(1 + y_i^k\right)\left(1 + y_j^k\right)\left(2 + y_i^k + y_j^k\right)} \geqslant 0 \tag{32}$$

式(32)中的关系表明，两帧的联合召回率值低于两帧召回率的平均值。同样，也可以得到以下关系

$$\frac{p_i^k + p_j^k}{2} - p^k \geqslant 0 \tag{33}$$

当将式(32)和式(33)代入式(13)后，得出结论：两帧的联合平均精度低于两帧平均精度的平均值。此外，假设1表明平均精度是关于残差映射数据量化位数的凹函数。因此，得到以下关系

$$\mathrm{mAP}\left(\hat{b}_i,\hat{b}_j\right) \leqslant \frac{\mathrm{mAP}\left(\hat{b}_i\right) + \mathrm{mAP}\left(\hat{b}_j\right)}{2} < \mathrm{mAP}\left(\frac{\hat{b}_i + \hat{b}_j}{2}\right) \tag{34}$$

其中，$\mathrm{mAP}\left(\hat{b}_i,\hat{b}_j\right)$是帧$i$和帧$j$的联合平均精度性能。式(34)表明，具有不同残差映射量化位数的两帧的联合平均精度低于具有相同量化位数的两帧的联合平均精度。因此，定理1成立。

### 7.2　定理2的证明

正如附录第7.1节中所证明的，集合$\Phi$中所有帧的残差映射数据的量化位数是相同的。比较以下两个满足定理1的情况。

(1) 情况1：比例为$\rho$（$0<\rho<1$）的帧的残差映射数据以量化位数$\hat{b}$传输到云服务器，而其余$1-\rho$比例的帧的残差映射数据不传输到云服务器。

(2) 情况2：所有帧的残差映射数据都以量化位数$\rho\hat{b}$传输到云服务器。

将情况1的$\Psi$中所有帧的联合平均精度表示为$\mathrm{mAP}\left(0_{|1-\rho},\hat{b}_{|\rho}\right)$。根据定理1的结果，当满足假设2时，这两种情况下的联合平均精度满足以下关系

$$\mathrm{mAP}\left(0_{|1-\rho},\hat{b}_{|\rho}\right) \leqslant (1-\rho)\,\mathrm{mAP}(0) + \rho\,\mathrm{mAP}\left(\hat{b}\right) < \mathrm{mAP}\left(\rho\hat{b}\right) \tag{35}$$

式(35)表明，情况2的联合平均精度大于情况1的联合平均精度。换句话说，$\rho=1$时的联合平均精度优于$0<\rho<1$的联合平均精度。因此，为了最大化云服务器的平均精度，集合$\Phi$中所有帧上传的残差映射数据应以相同的量化位数传输到云服务器，即$\rho=1$，因此定理2成立。

### 7.3　定理3的证明

如附录第7.1节中所讨论的那样，平均精度是不同交并比阈值下精确率—召回率对的函数。为了研究平均精度与两个模型的精确率—召回率对之间的关系，首先分析空地一体化云-边模型协同演化架构的联合精确率和召回率的表达式。将空地一体化云-边模型协同演化架构在第$k$个交并比阈值下的精确率和召回率分别表示为$r^k=\dfrac{\mathrm{TP}^k}{\mathrm{TP}^k + \mathrm{FN}^k}$和$p^k=\dfrac{\mathrm{TP}^k}{\mathrm{TP}^k + \mathrm{FP}^k}$。

以精确率值为例，如附录第7.1节中所分析的，云模型在第$k$个交并比阈值下的精确率值可以表示为

$$p_{\mathrm{L}}^k = \frac{1}{1 + x_{\mathrm{L}}^k} \tag{36}$$

其中，$x_{\mathrm{L}}^k=\dfrac{\mathrm{FP}_{\mathrm{L}}^k}{\mathrm{TP}_{\mathrm{L}}^k}$，边缘模型在第$k$个交并比阈值下的精确率值可以表示为

$$p_{\mathrm{S}}^k = \frac{1}{1 + x_{\mathrm{S}}^k} \tag{37}$$

其中，$x_{\mathrm{S}}^k=\dfrac{\mathrm{FP}_{\mathrm{S}}^k}{\mathrm{TP}_{\mathrm{S}}^k}$。式(36)和式(37)表明，云模型中的每个TP样本都伴随着$x_{\mathrm{L}}^k$个FP样本，而边缘模型中的每个TP样本都伴随着$x_{\mathrm{S}}^k$个FP样本。假设两个模型中的样本数量分别为$\beta$和$1-\beta$，则平均精确率值可以表示为

$$p^k = \frac{\mathrm{TP}^k}{\mathrm{TP}^k + \mathrm{FP}^k} = \frac{1}{1 + \beta x_{\mathrm{L}}^k + \left(1-\beta\right)x_{\mathrm{S}}^k} = \frac{1}{\dfrac{\beta}{p_{\mathrm{L}}^k} + \dfrac{1-\beta}{p_{\mathrm{S}}^k}} \tag{38}$$

同样，召回率值的关系可以表示为

$$r^k = \frac{1}{\dfrac{\beta}{r_{\mathrm{L}}^k} + \dfrac{1-\beta}{r_{\mathrm{S}}^k}} \tag{39}$$

通过将式(38)和式(39)代入式(31)，可以得到式(27)，从而证明了定理3。

### 7.4　定理4的证明

在式(27)中，联合云-边平均精度，即mAP，是不同交并比阈值下精确率—召回率对的函数的线性求和。可以通过分析特定交并比阈值下的精确率—召回率对来研究mAP、$\mathrm{mAP}_{\mathrm{L}}$和$\mathrm{mAP}_{\mathrm{S}}$之间的关系，并且该性质在线性变换后仍然成立。定义一个变量$\zeta^k=p^k r^k$，它仅与特定交并比阈值下的精确率—召回率对相关。相应地，云服务器上的云模型和边缘无

人机上的边缘模型的变量可以分别表示为$\zeta_{\mathrm{L}}^k = p_{\mathrm{L}}^k r_{\mathrm{L}}^k$和$\zeta_{\mathrm{S}}^k = p_{\mathrm{S}}^k r_{\mathrm{S}}^k$。根据定理3，$\zeta^k$可以表示为

$$\zeta^k = p^k r^k = \frac{1}{\dfrac{\beta}{p_{\mathrm{L}}^k} + \dfrac{1-\beta}{p_{\mathrm{S}}^k}} \cdot \frac{1}{\dfrac{\beta}{r_{\mathrm{L}}^k} + \dfrac{1-\beta}{r_{\mathrm{S}}^k}} = \frac{\zeta_{\mathrm{L}}^k \zeta_{\mathrm{S}}^k}{(1-\beta)\zeta_{\mathrm{L}}^k + \beta\zeta_{\mathrm{S}}^k - \beta(1-\beta)\Delta} \tag{40}$$

其中，$\Delta=(p_{\mathrm{L}}^k - p_{\mathrm{S}}^k)\cdot(r_{\mathrm{L}}^k - r_{\mathrm{S}}^k)$表示云模型与边缘模型之间的推理能力差异。考虑云模型的推理能力一般优于边缘模型，因此合理的假设是$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k>0$和$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k> 0$。因此，关系$\Delta>0$成立，因此可以得到

$$\zeta^k > \frac{\zeta_{\mathrm{L}}^k \zeta_{\mathrm{S}}^k}{(1-\beta)\zeta_{\mathrm{L}}^k + \beta\zeta_{\mathrm{S}}^k} \tag{41}$$

由于mAP是多个$\zeta^k$的线性组合，因此式(41)同样适用于mAP，即

$$\mathrm{mAP} > \frac{\mathrm{mAP_L}\cdot\mathrm{mAP_S}}{(1-\beta)\mathrm{mAP_L} + \beta\mathrm{mAP_S}} \tag{42}$$

定理4故而得证。

### 7.5　定理5的证明

如附录第7.4节所证明的，变量$\zeta^k = p^k r^k$可以转换为

(43)

其中，$\Delta=(p_{\mathrm{L}}^k - p_{\mathrm{S}}^k)\cdot(r_{\mathrm{L}}^k - r_{\mathrm{S}}^k)$。当约束条件$r_{\mathrm{L}}^k - r_{\mathrm{S}}^k \ll r_{\mathrm{L}}^k$和$p_{\mathrm{L}}^k - p_{\mathrm{S}}^k \ll p_{\mathrm{L}}^k$得到满足时，有$\Delta \ll \zeta_{\mathrm{L}}^k \zeta^k = \dfrac{\zeta_{\mathrm{L}}^k \zeta_{\mathrm{S}}^k}{(1-\beta)\zeta_{\mathrm{L}}^k + \beta\zeta_{\mathrm{S}}^k - \beta(1-\beta)\Delta}$，因此

$$\zeta^k \approx \frac{\zeta_{\mathrm{L}}^k \zeta_{\mathrm{S}}^k}{(1-\beta)\zeta_{\mathrm{L}}^k + \beta\zeta_{\mathrm{S}}^k} \tag{44}$$

由于mAP是多个$\zeta^k$的线性组合，因此式(44)同样适用于mAP，式(29)也成立。

## 参考文献：


[1] WARGO C A, CHURCH G C, GLANEUESKI J, et al. Unmanned Aircraft Systems (UAS) research and future analysis[C]//Proceedings of the 2014 IEEE Aerospace Conference. Piscataway: IEEE Press, 2014: 1-16.

[2] HOSSEIN MOTLAGH N, TALEB T, AROUK O. Low-altitude unmanned aerial vehicles-based Internet of Things services: comprehensive survey and future perspectives[J]. IEEE Internet of Things Journal, 2016, 3(6): 899-922.

[3] LIU Z S, WANG L Y, LI B. Quality assessment of ecological environment based on google earth engine: a case study of the Zhoushan Islands[J]. Frontiers in Ecology and Evolution, 2022, 10: 918756.

[4] 孟婵媛, 熊轲, 高博, 等. 面向6G的生成对抗网络研究进展综述[J]. 物联网学报, 2024, 8(1): 1-16.
MENG C Y, XIONG K, GAO B, et al. Survey on the research progress of generative adversarial networks for 6G[J]. Chinese Journal on Internet of Things, 2024, 8(1): 1-16.

[5] SHEN Y F, SHAO J W, ZHANG X J, et al. Large language models empowered autonomous edge AI for connected intelligence[J]. IEEE Communications Magazine, 2024, 62(10): 140-146.

[6] 栾宁, 熊轲, 张煜, 等. 6G: 典型应用、关键技术与面临挑战[J]. 物联网学报, 2022, 6(1): 29-43.
LUAN N, XIONG K, ZHANG Y, et al. 6G: typical applications, key technologies and challenges[J]. Chinese Journal on Internet of Things, 2022, 6(1): 29-43.

[7] PAL O K, SHOVON M S H, MRIDHA M F, et al. A comprehensive review of AI-enabled unmanned aerial vehicle: trends, vision, and challenges[J]. arXiv preprint arXiv: 2310.16360 ,2023.

[8] BOTH C B, BORGES J, GONÇALVES L, et al. System intelligence for UAV-based mission critical with challenging 5G/B5G connectivity [J].arXiv preprint arXiv:2102.02318,2021.

[9] YANG B, CAO X L, YUEN C, et al. Offloading optimization in edge computing for deep-learning-enabled target tracking by Internet of UAVs[J]. IEEE Internet of Things Journal, 2021, 8(12): 9878-9893.

[10] LIU Z, ZHAN C, CUI Y, et al. Robust edge computing in UAV systems via scalable computing and cooperative computing[J]. IEEE Wireless Communications, 2021, 28(5): 36-42.

[11] KOUBAA A, AMMAR A, ABDELKADER M, et al. AERO: AI-enabled remote sensing observation with onboard edge computing in UAVs[J]. Remote Sensing, 2023, 15(7): 1873.

[12] VARGHESE R, SAMBATH M. YOLOv8: a novel object detection algorithm with enhanced performance and robustness[C]//Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Piscataway: IEEE Press, 2024: 1-6.

[13] AKKUS C, CHU L, DJAKOVIC V, et al. Multimodal deep learning[J].arXiv preprint arXiv: 2301.04856,2024.

[14] LIU Y, ZHANG K, LI Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models[J]. arXiv preprint arXiv: 2402.17177,2024.

[15] Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[J]. arXiv preprint arXiv:2403.05530,2024.

[16] YUAN Y Z, GAO S C, ZHANG Z T, et al. Edge-cloud collaborative UAV object detection: edge-embedded lightweight algorithm design and task offloading using fuzzy neural network[J]. IEEE Transactions on Cloud Computing, 2024, 12(1): 306-318.

[17] NARAYANA TINNALURI V S, VYANKATESH GHAMANDE M, SINGH S, et al. Edge-cloud computing systems for unmanned aerial vehicles capable of optimal work offloading with delay[C]// Proceedings of the 2023 Second International Conference on Electronics and Renewable Systems (ICEARS). Piscataway: IEEE Press, 2023: 844-849.

[18] XU M R, NIYATO D, ZHANG H L, et al. Sparks of generative pretrained transformers in edge intelligence for the metaverse: caching and inference for mobile artificial intelligence-generated content services[J]. IEEE Vehicular Technology Magazine, 2023, 18(4): 35-44.

[19] LI B, FEI Z S, ZHANG Y. UAV communications for 5G and beyond: recent advances and future trends[J]. IEEE Internet of Things Journal, 2019, 6(2): 2241-2263.

[20] YANG B, CAO X L, LI X F, et al. Lessons learned from accident of autonomous vehicle testing: an edge learning-aided offloading framework[J]. IEEE Wireless Communications Letters, 2020, 9(8): 1182-1186.

[21] ZHANG S H, ZHANG H L, SONG L Y. Beyond D2D: full dimension UAV-to-everything communications in 6G[J]. IEEE Transactions on Vehicular Technology, 2020, 69(6): 6592-6602.

[22] ZHANG S H, LIU Q Y, CHEN K, et al. Large models for aerial edges: an edge-cloud model evolution and communication paradigm[J]. IEEE Journal on Selected Areas in Communications，2024: 1-16.

[23] LI J, LI B, LU Y. Deep contextual video compression[J]. Advances in Neural Information Processing Systems, 2021, 34: 18114-18125.

[24] CHEN B, BAKHSHI A, BATISTA G, et al. Update compression for deep neural networks on the edge[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2022: 3075-3085.

[25] YANG G, TANG Y, WU Z J, et al. DMKD: improving feature-based knowledge distillation for object detection via dual masking augmentation[C]//Proceedings of the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2024: 3330-3334.

[26] MOYA OSORIO D P, AHMAD I, SÁNCHEZ J D V, et al. Towards 6G-enabled Internet of vehicles: security and privacy[J]. IEEE Open Journal of the Communications Society, 2022, 3: 82-105.

[27] REAL E, MOORE S, SELLA A, et al. Large scale evolution of image classifiers[C]//Proceedings of the International Conference on Machine Learning, Sydney, Australia: Association for Computing Machinery: 2017: 2902-2911.

[28] YANG W H, HUANG H F, HU Y Y, et al. Video coding for machines: compact visual representation compression for intelligent collaborative analytics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(7): 5174-5191.

[29] BALLÉ J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior[J]. arXiv preprint arXiv: 1802.01436,2018.

[30] LI W G, SUN W Y, ZHAO Y D, et al. Deep image compression with residual learning[J]. Applied Sciences, 2020, 10(11): 4023.

[31] CHENG Z X, SUN H M, TAKEUCHI M, et al. Learned image compression with discretized Gaussian mixture likelihoods and attention modules[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 7936-7945.

[32] DU D，ZHU P, WEN Let al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea: IEEE Press: 2019: 213-226.

[33] CHEN Z Q, DUAN L Y, WANG S Q, et al. Toward knowledge as a service over networks: a deep learning model communication paradigm[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1349-1363.

[作者简介]

**于馨博**（2005– ），男，北京大学信息科学技术学院在读，主要研究方向为无线通信。

**张舒航**（1993– ），男，鹏城实验室助理研究员、博士生导师，主要研究方向为无线网络、人工智能、空地一体化网络等。

**张泓亮**（1992– ），男，北京大学电子学院助理教授、博士生导师，主要研究方向为智能超表面、空地一体化网络等。