华为研究

Communications of HUAWEI RESEARCH

内部资料　免费交流
准印证号：（粤B）L0240054
2024年11月
第 2 期（总第7期）

HUAWEI

跨越人联、物联，
迈向万物智联

编者按

早在 2018 年，华为就率先提出“智能互联（Connected Intelligence）”的 6G 愿景。今天，随着 AI 革命加速使能各行各业，我们的使命是要构建下一代移动平台，随时随地为每个人、每个家庭、每个组织提供智能服务。

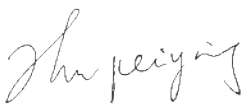
2023 年 11 月，国际电信联盟 ITU-R 就 6G 框架建议达成全球共识，将 AI 和通信一体化列为 6G 的支柱之一。这意味着，AI for Network 和 Network for AI 已经成为 6G 的标志性代际技术，终端和超算中心的全面 AI 化也在以前所未有的速度迅猛发展。古滕堡的文字印刷术开启了人类的信息革命，后来的工业革命使机器在体力上超越了人类，而现今的 AI 技术则会让机器在智力上超越人类。放眼未来，6G 移动网络将成为 AI 和通信一体化的基础平台，把超级智能服务带给每一个人、每一个物。

移动产业的规律在于以技术创新驱动市场演进。6G 的市场窗口期将在 2035 到 2045 年，因此 6G 网络和 6G 终端必须适应未来消费者和垂直行业的需求。回首 20 年前，互联网是最新技术的使能者，移动通信因拥抱互联网而获得了巨大的商业成功。今天，AI 成为了最新技术的使能者，并且正在成数量级地快速跃迁，带领我们走进超级智能的数字世界。6G 移动通信与 AI 的结合，将打造移动网络的新范式，成为数字世界与物理世界的接口，全面使能通用人工智能（AGI）和具身智能（Embodied AI）。6G 网络不应局限于生成式 AI，而应将 AI 贯穿到终端、无线网络和核心网的感知、推理、判决以及实体操作的全流程中。在 AI 和通信一体化的加持下，6G 网络将具备通感一体化特性及能力，实现原生的网络智能。

在 6G 终端演进方面，需要 AI 来驱动革命性突破，引领全产业链成功。在后移动宽带业务时代，终端技术的突破将是移动产业演进的关键。6G 移动终端也将迎来新的革命，向 Full AI 演进，通过终端能力的跃迁实现 6G 网络升级，最终引领全产业链成功。

在架构设计方面，6G 需要超越 SBA（基于服务的架构），基于 Agentic AI 技术进行重构，从而走向 ADGN（应用驱动的生成网络）。与 5G 相比，6G 不单单是技术和架构的演进，而是要带动移动产业以及各行各业进行广泛的技术革新。

围绕 AI 在无线通信领域的应用，本期《华为研究》荟集了多方研究成果，不仅有华为专家的贡献，还有学术界和同行合作伙伴的愿景分享与成果展示。希望这些研究成果能为 6G“智能互联”愿景的实现做出积极贡献。



童文博士
华为Fellow

朱佩英博士
华为Fellow

华为研究
内部资料，免费交流
准印证号：（粤B）L0240054

主编：
廖恒

本期责任主编：
童文，朱佩英

编委会：
廖恒，童文，肖新华，胡邦红，周慧慧，
鲍丰，Jeff Xu，陈海波，陆品燕，
王建兵，李瑞华，白博

索阅、投稿、建议和意见反馈，请联系：
HWResearch@huawei.com

印刷数量：4000本
印刷单位：雅昌文化（集团）有限公司
印刷地址：深圳市南山区深云路19号
印刷日期：2024年11月25日

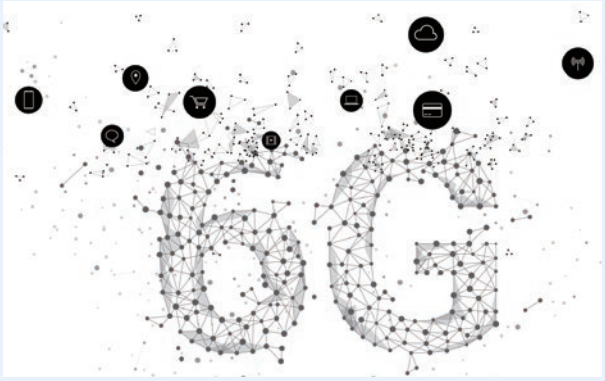
编印单位：华为技术有限公司
发送对象：本行业、本系统、本单位

目 录

CONTENTS

展望

技术



实践

面向通信的机器学习：通向智能传输与处理之路

王世雄，李烨

摘要

在人工智能与大数据时代到来之前，无线通信研究主要遵循传统路线，包含问题分析、模型构建与校准、算法设计与调优，以及整体测试和经验验证等环节。然而，在处理大规模复杂问题和管理动态海量数据时，这种方法往往存在局限性，导致传统通信系统和技术效率低下、性能受限。因此，借助人工智能和机器学习的革命浪潮，无线通信领域开发出高度自适应、更高效且更智能的系统和算法。这一技术革新为信息传输与处理的智能化铺平了道路。本文将探讨机器学习在智能无线通信中的典型作用，及其特点、挑战与实际考虑因素。

关键词

机器学习，智能传输，智能处理

1 引言

自 19 世纪以来，无线电通信开启了人类社会信息传输的新纪元。在摩尔斯电码和电报机等早期无线电时代，传输技术严重依赖人工操作，限制了信息交换的效率和可靠性。为了实现信息传输与处理的自动化，第一代“智能传输与处理”概念应运而生。到了 20 世纪，模块化通信系统取得了重大发展。模块化通信系统由信源编码、信道编码、调制、发射波束赋形、无线传输、接收波束赋形、解调、信号检测、信道解码和信源解码等基本模块组成 [1–3]。从方法论上来看，模块化无线通信与信号处理遵循系统化研究路径，包括问题分析、模型开发与校准、算法设计与调优以及经验验证、反馈与改进，每一步都需要人类投入大量的脑力来完成。

进入 21 世纪后，无线通信系统需传输音频、视频、文本等各种形式的海量数据，并需保证低时延、高速率和高可靠性。此外，物联网和无人机中继网络等新型网络拓扑，以及通感一体化（Integrated Sensing and Communications, ISAC）和算通一体化（Integrated Computing and Communications, ICAC）等尖端技术的出现，增加了现代通信系统设计的复杂性，这主要体现在以下三个方面：

- 系统和单模块在建模时存在各种不确定性；
- 用户设备和基站产生各种形式的大数据；
- 网络实现中面临各种算法挑战。

传统设计方法高度依赖人类智力，难以处理大规模复杂问题和动态海量数据，导致信息传输处理效率低下、性能受限。在此背景下，无线通信与信号处理开始利用人工智能和机器学习进行革新 [4]。关于机器学习在通信领域的应用研究，请参见 [5–11]。这一技术和方法转变推动业界开发出高度自适应、高效、鲁棒且智能的系统和算法，催生了第二代“智能传输与处理”概念，以期大幅减少对人类智力的依赖，并提升通信系统的整体性能。

图 1 智能传输与处理的内涵
（图片来源：CLEANPNG.com 和 FLATICON.com）

图 1 展示了智能传输与处理的内涵。图 2 展示了基于机器学习的智能传输与处理系统的端到端架构。

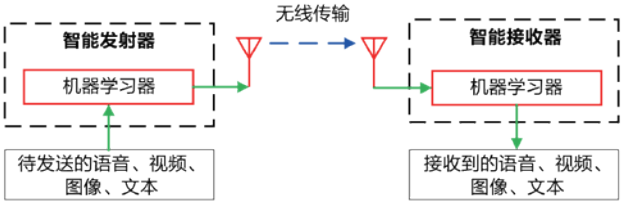


图 2 智能传输与处理系统的端到端架构。智能发射和接收模块作为端到端信息处理器f，均通过数据习得，并能实时自适应无线信道。智能化主要体现在不再需要人类研究大规模、动态和不确定的信息传递机制和处理方案。

为了展示机器学习在智能传输与处理中的重要作用，本文回顾了通信系统和方法中典型的机器学习应用，包括物理层通信 [6, 12]、语义通信 [13, 14]、通信资源分配 [15]、ICAC [16] 以及 ISAC [17]。受篇幅所限，本文并未穷尽列举该领域当前所有研究成果，仅阐明通往智能传输与处理的路径。

尽管机器学习有望革新无线通信理论和实践，基于机器学习的方法在带来机遇和优势的同时，也面临挑战和问题 [4]，如深度学习等“黑盒”学习方法缺乏可解释性导致的可靠性问题、训练数据有限及底层数据生成规律不平稳引发的泛化问题，以及在训练和存储大型机器学习模型（如深度学习）时资源不足的问题；参见图 3。除了上述主要挑战，还可能存在其他问题，如系统拓扑或硬件重配置（如添加或移除天线）引发的可扩展性问题（也可视为一种泛化问题），以及网络学习中的安全与隐私问题 [16]。在发展通信理论和系统时，切勿夸大机器学习的作用，机器学习（尤其是数据驱动的深度学习）可以助力通信，但不是必须遵循的绝对规则；问题分析与机制建模历来都至关重要。相关技术综述，请参见 [18–20]。下例展示了数据驱动的机器学习方法的缺陷。

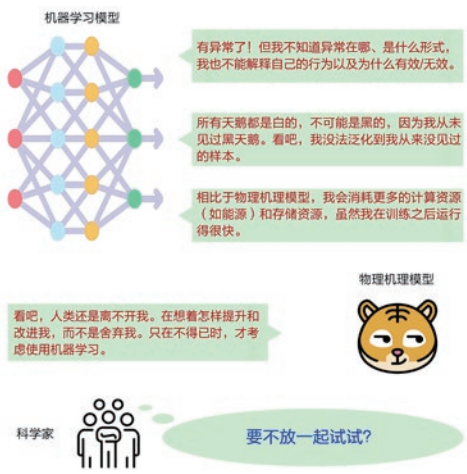


图 3 选择机器学习模型还是物理定律模型，这是个问题！我们需要选择最适合的方案！如果可能，将它们整合起来以提高整体系统性能。（图片来源：FLATICON.com）

示例（冰激凌销售与鲨鱼袭击）：历史数据的回归分析表明，冰淇淋销售与鲨鱼袭击存在正相关性，但这显然不合理。然而，这种相关性背后的主要因素是气温——气温升高导致冰淇淋销量和海滩游客增加，而海滩游客增多则引发更多的鲨鱼袭击 [21]。可见，机制建模至关重要。

在探讨通信领域机器学习的应用前，我们将在第 2 节简要回顾其概念和方法，特别是可信赖机器学习。这些资料有助于读者理解无线通信中使用机器学习时的主要考虑因素，包括哲学和技术因素。

2 机器学习的概念与方法

机器学习的首要任务是从数据中发现隐含信息和模式，其优势在于能够自动解读数据，无需人类研究数据底层生成机制。这一特点从根本上使得机器智能能够参与到通信实践，尤其是信息传输与处理方面 [5, 7, 8, 11, 22]。

根据任务特点，机器学习可分为四类：有监督学习、无监督学习、半监督学习和强化学习。从数学角度看，所有机器学习任务的关键在于习得一个函数f（亦称为“假设”），将观察数据映射到期望决策上。相关概念参见图 4，具体示例如下：

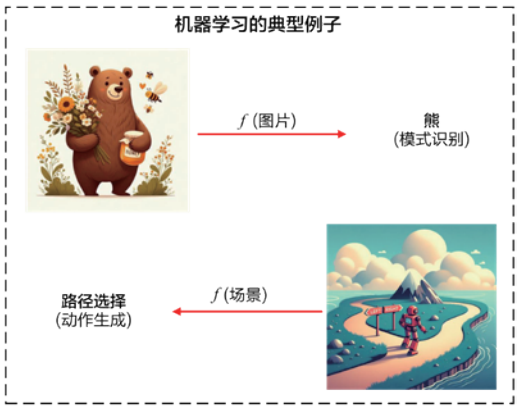


图 4 机器学习概念图。机器学习是一个寻找输入到决策的映射关系f的过程。上图是有监督学习，分类器f将图像识别为熊；下图是强化学习，动作生成器f结合当前场景为机器人提供路径选择建议。（图片均由 Microsoft Copilot 生成。）

- **有监督学习：**有监督学习挖掘了标记数据的隐含信息，可分为回归和分类两种任务类型。给定确定或随机变量对(x, s)，其中x表示特征向量，s表示连续值期望响应。**回归**的目的是找到从x到s的函数关系f，使得预测标记$f(x)$尽可能接近目标标记s。接近度由损失函数$L[s, f(x)]$来衡量，如均方误差$[s-f(x)]^H[s-f(x)]$。针对确定变量对(x, s)，可能存在函数f，使得每个(x, s)都有$f(x)=s$，即$L[s, f(x)] \equiv 0$。而对于随机变量对(x, s)，一般无法保证完全相等。相反，需在(x, s)的联合分布$\mathbb{P}_{x,\ s}$下计算损失，例如，损失的期望值$\mathbb{E}_{(x,\ s)\sim\mathbb{P}_{x,\ s}} L[\mathbf{s}, f(\mathbf{x})]$。当$s$为一维离散数值时，会出现**分类**问题，$L[s, f(x)]$由函数（如指示函数$\mathbb{I}\{s \neq f(x)\}$）决定。在这种情况下，$s$可表示不同的目标类，例如，对于二进制分类，有$s \in \{-1, 1\}$。
- **无监督学习：**在无监督学习中，收集的数据没有标记s，只有特征x。因此，无监督学习主要用于从数据变量x的实现中发现隐含信息。**聚类**是一种典型的无监督学习任务。聚类与分类不同，分类是通过具有已知标记的训练数据集预测新数据点x的分类标记s，而聚类则根据相似数据点x的特征将其进行分组，无需预定义分类标记。简而言之，聚类就是寻找函数f，将数据x映射到一个合适的组。**特征转换**也是一种常见的无监督学习任务。它通过映射函数f将原始特征x转换到另一个特征空间，即$y=f(x)$。例如，时域信号x及其傅里叶变换y就是一种特征转换。自编码器（Autoencoder）是一种人工神经网络，它的编解码操作也是一个很好的特征转换示例。另一个重要的无监督学习示例是**分布估计**，即估计出最能拟合（刻画）已收集数据的概率分布。在生成式任务中，分布估计尤为重要，例如根据已有样本生成新样本；具体比如给定一组猫的图像，利用拟合分布绘制出新的猫图像。
- **半监督学习：**半监督学习是监督学习的扩展，通过结合有标记数据(x, s)和无标记数据x'来提升模型性能。与完全依赖有标记数据的有监督学习不同，半监督学习利用无标记数据提高模型性能和泛化能力。当数据标记成本高耗时长，而无标记数据相对丰富且容易获取时，半监督学习就显得尤为有用。与单纯依赖标记数据的有监督学习相比，半监督学习能同时利用标记数据和无标记数据，习得的f能更准确地预测数据x的标记s。
- **强化学习：**强化学习关注的是动态、不确定环境中的决策问题。不同于依赖有标记数据的有监督学习或无标记数据的无监督学习，强化学习通过智能体与环境交互，接收奖励或惩罚反馈，从而习得各时间点的最优行为或策略。具体来说，智能体在当前状态s下执行动作a，通过自主学习决策获取最大化的累积奖励。因此，从数学上来看，需要习得从状态s到动作a的动作生成函数f。

关于这四类机器学习在无线通信中的具体应用，在 [8, 23] 中有具体介绍。

数据驱动学习与模型驱动学习：根据人类智力和领域知识的参与程度，机器学习可分为两类：数据驱动的机器学习和模型驱动的机器学习。数据驱动的机器学习完全依赖历史数据，不分析底层生成机制。而模型驱动的机器学习则在不同程度上结合了底层物理机制与数据生成模型，通过通信系统建模与大数据挖掘的协作，可以提升智能信息传输与处理的整体性能 [12, 23]。以信号检测为例，假设有T组导频数据$\{(s_1, x_1), (s_2, x_2), \ldots, (s_T, x_T)\}$，式中，$x_i$表示接收信

号，s_i表示传输符号（$i=1,2,\ldots,T$）。数据驱动的机器学习直接利用所有数据（x 到 s）训练出检测器f。而模型驱动的机器学习则首先考虑信号传输模型$x=Hs+v$（式中，H表示信道矩阵，v表示噪声），然后基于该底层数据生成机制找到检测器f。详细的技术处理和讨论，请参见 [19, 20, 24, 25]。

假设空间和深度学习：为了找到最佳决策函数f，需要指定一个候选函数空间$\mathcal{H}$，即所谓的假设空间。有监督统计机器学习可以表述为：

$$\min_{f\in\mathcal{H}} \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}} L(\mathbf{s}, f(\mathbf{x})),$$

式中，未知的联合分布$\mathbb{P}_{x,\,s}$可通过历史数据（如经验分布）估计。信号检测问题可以如上表述，其中f表示检测器，x表示天线接收的信号，s表示传输的符号（例如星座点）[20]，损失函数L可以是均方误差或误符号率。典型的假设空间$\mathcal{H}$有以下几种：

- 线性函数空间：$\mathcal{H}$只包含输入x的线性变换。在信号检测中，$\mathcal{H}$只包含线性检测器。
- 再生核希尔伯特空间：基于特征映射函数φ，$\mathcal{H}$包含原始特征x的非线性提升特征$\varphi(x)$的所有线性变换组合。实质上，$\mathcal{H}$包含了输入x的一些特定类型的非线性函数。
- 神经网络函数空间：$\mathcal{H}$由神经网络表征，如多层感知机、递归神经网络、卷积神经网络（Convolutional Neural Network，CNN）、径向基函数神经网络、自编码器和 Transformer 等。每个神经网络都定义了一个特定类型的函数空间$\mathcal{H}$。当使用的神经网络具有多个隐藏层时，$\mathcal{H}$表示深度神经网络的函数空间。基于深度神经网络的机器学习又称为**深度学习**。

针对特定领域问题，可以引入领域知识和专家设计，调整或定制假设空间 [12, 18, 23]。因此，模型驱动的机器学习通过利用已知问题特征和数据生成机制来设计一个专门且结构化的候选空间$\mathcal{H}$。

可解释性、可靠性和可持续性：机器学习需要研究各种复杂问题，包括模型的可解释性、可靠性和可持续性 [26, 27]。可解释性的目标是通过特征工程和物理建模等技术，使学习模型透明、可解释、可纠错 [28]。基于底层物理数据生成机制的模型驱动机器学习就是这样的实现方式 [12, 18]。可靠性旨在构建鲁棒且准确的学习模型，能够很好地泛化到新数据（即训练未用到的数据），以解决过拟合、泛化、知识迁移、小样本学习等问题 [20, 29–31]。可持续性旨在开发对环境和社会负面影响最小的学习模型，解决能效、隐私安全、公平与偏见等问题 [16, 32]。对于智能信息传输与处理而言，可解释性、可靠性和可持续性尤为重要，在开发基于机器学习的无线通信方案时，这些都是首要考虑因素。

集中式学习与分布式学习：根据数据分布结构和计算架构，可以采用不同方法训练机器学习模型，主要有集中式学习和分布式学习两种 [33, 34]。集中式学习将所有训练数据统一收集并存储在一个位置（如数据中心或云服务器），机器学习模型在此基础上进行训练。而分布式学习则是在多个设备或节点上分别训练机器学习模型，每个设备或节点均存有部分数据。**联邦学习**是典型的分布式学习模式。联邦学习中，多个客户端（如智能手机、物联网设备或不同组织）在不共享其本地数据的情况下共同训练模型，各客户端利用其本地数据进行训练，仅与中心节点共享模型更新（梯度或权重），由中心节点聚合这些更新，形成新的全局模型。集中式和分布式学习方法适用于不同类型的现代通信网络，有助于推动通信系统的进一步演进。

3 物理层通信

物理层通信旨在通过物理介质可靠地传输原始数据流，如二进制比特。图 5 展示了传统的无线通信架构，其功能模块由科学家根据基本数学和物理原理精心设计而来。这种模块化架构与图 2 所示的数据驱动的机器学习架构不同，后者利用端到端操作取代了互连的功能模块。

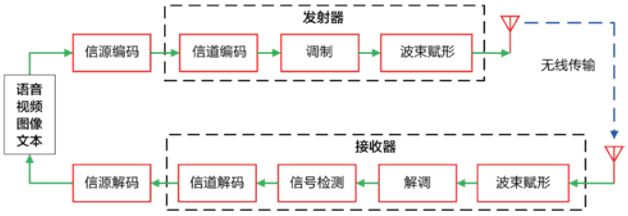


图 5 传统传输与处理系统的模块化架构。每个模块作为信息处理器f，由科学家根据底层物理机制和数学定律精心设计而来。

除了图 2 中高度集成化（即高度智能化）的架构外，基于机器学习的传输与处理系统也可以实现部分智能化。例如，在某些情况下，仅由机器学习管理信道编码或解码模块，即信道编解码方案是由机器而非信息科学家设计的。再如，机器学习仅用于发射波束赋形器或接收波束赋形器，如图 6 所示。

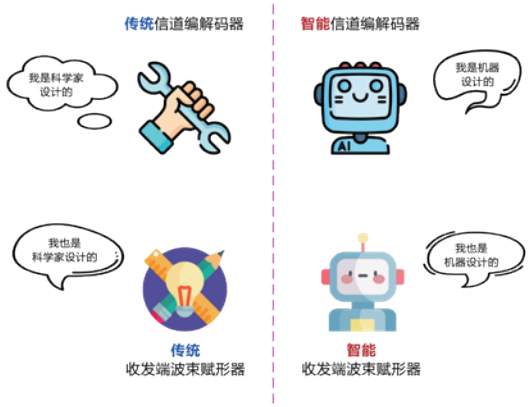


图 6 与图 2 所示的高度集成架构相比，图 5 中的每个模块都可以通过机器学习增强（图片来源：FLATICON.com）

技术上，机器学习在物理层通信中的应用包括整体端到端系统设计 [35, 36]（图 2）和单模块设计（图 6）。其中单模块设计包括：

- 编解码技术，如信源编码 [37]、信道编码 [38, 39] 以及联合信源信道编码（Joint Source–Channel Coding，JSCC）[40, 41]
- 信号调制与检测 [25, 42]
- 发射和接收波束赋形 [20, 43–46]，如波束对齐和波束跟踪 [47–50]
- 信道估计与反馈 [51, 52]

文献 [6, 53, 54] 等已提供最新、最全的综述，本文不再赘述。

编解码技术在数字通信中至关重要，可确保数据的高效可靠传输。近年来，机器学习被广泛应用于增强信源编码、信道编码和 JSCC 等编解码技术。传统信源编码（即数据压缩）通过减少数据冗余实现高效传输和存储，而基于神经网络的自编码器等机器学习技术已彻底改变了这一方式。自编码器通过将数据编码到低维空间并重构来习得高效的数据表征，从而在信息损失最小的情况下实现高压缩率 [55]。信道编码通过为数据添加冗余来检测并纠正由噪声信道引起的传输错误。伴随着机器学习模型（尤其是深度学习）模型的应用，一些新型纠错码开始涌现。例如，用神经解码器解码复杂方案，如低密度奇偶校验（Low–Density Parity Check，LDPC）码 [56] 和 Turbo 码 [57]，性能优于传统编码方式，尤其在噪声大的环境中表现更佳。JSCC 同时集成了信源编码与信道编码，提升了系统整体性能。变分自编码器（Variational Autoencoder，VAE）[58]、CNN [59] 和生成式对抗网络（Generative Adversarial Network，GAN）[60] 等机器学习模型可用于联合学习表征和纠错码，它们能够同时适应信源和信道特征，压缩和容错性能远胜过传统方式。总体而言，基于机器学习的编解码技术极大地促进了数字通信的发展，通过利用其预测与自适应功能，这些技术增强了数据压缩、纠错及传输效率，为构建更鲁棒、高效的通信系统奠定了基础。

在数字通信中，信号调制与检测是数据传输和解码的基本步骤。近年来，机器学习技术被应用于信道调制与检测，提高了传输效率和可靠性。调制通过改变载波信号的幅度、频率或相位等属性来编码信息。传统调制方案包括调幅（Amplitude Modulation，AM）、调频（Frequency Modulation，FM）和相移键控（Phase Shift Keying，PSK）等技术。机器学习（尤其是深度学习）模型正应用于自适应调制方案设计，它们可以根据信道条件动态调整调制参数，实时优化性能。例如，神经网络可以学习复杂的调制模式，实现数据吞吐率最大化和误码率最小化 [61]。检测通过解调接收信号来恢复传输信息。传统方法使用预设算法估计传输数据，但通常是基于特定信道特征的假设。全连接深度神经网络 [42] 和迁移学习 [62] 等机器学习方法的出现，使得信号检测大大增强。这些模型能够从数据中学习，在复杂多变的信道条件下准确检测信号，提高系统的抗噪声 / 干扰能力。总体而言，机器学习技术与信号调制检测的融合，是通信技术领域的重大进步，提高了数据传输效率、抗噪能力和系统整体性能。

在无线通信系统中，发射和接收波束赋形技术对增强信号质量和提高数据吞吐率至关重要。该技术通过天线阵列定向发射或接收信号，提高信号强度并减少干扰。近年来，机器学习显著提升了波束赋形性能，从而带来更广阔的应用场景。传统的发射波束赋形方法，如相控阵列系统，通过预设算法调整多天线信号的相位和幅度。而引入机器学习（尤其是深度学习）模型后，系统可从环境数据中自学习、自优化。例如，强化学习能够根据通信环境的反馈实时调整波束赋形模式，从而在复杂多变的场景下提升性能 [63, 64]。传统的接收波束赋形方法，如最小方差无失真响应（Minimum Variance Distortionless Response，MVDR）和最大比合并（Maximal Ratio Combining，MRC），依赖于信号环境统计模型。而 CNN 等机器学习方法则通过直接从数据中习得最佳波束赋形权重，能在各种动态环境中实现准确、稳健地接收信号 [65, 66]。波束对准与跟踪是波束赋形的关键技术，在毫米波（Millimeter Wave，mmWave）和太赫兹通信等高频通信中尤为重要。它们能确保收发端始终保持最佳对准状态，从而实现信号强度和数据吞吐率的最大化 [49, 50, 67]。传统方法依赖穷举搜索或迭代算法，不仅耗时，而且对计算资源的需求很大。有监督学习、多臂老虎机、强化学习等机器学习方法可基于历史数据预测最优波束方向，提供更高效的解决方案，大幅缩小搜索空间。即使收发端移动或环境产生变化的情况下，波束也能始终保持对准。机器学习（尤其是深度学习）模型通过实时预测波束方向的变化来增强跟踪能力，特别是能捕捉时间依赖性的循环神经网络（Recurrent Neural Network，RNN）和长短期记忆（Long Short–Term Memory，LSTM）网络。在基于机器学习的波束对准与跟踪方面，[47–50, 67] 等文献提供了翔实的技术细节，此处不再赘述。总之，机器学习与波束赋形（包括波束对准与跟踪）技术的融合，通过预测和自适应能力提升信号质量、减少干扰并优化系统性能，对 5G 等下一代网络至关重要。

信道估计与反馈技术用于测量信道特性并向发射端提供必要反馈，确保准确表征信道特征并高效传输数据，在无线通信系统中至关重要。近年来，随着机器学习技术的引入，信道估计的准确性和效率显著提升 [68, 69]。信道估计通过预测信道状态优化信号收发。传统方法如最小均方差（Minimum Mean Square Error，MMSE）依赖统计模型，计算资源需求高。机器学习（尤其是深度学习）模型为提升信道估计精度及降低计算复杂度提供了新的思路。例如，CNN 可以基于接收信号数据自我学习并估计信道状态，为复杂场景提供了鲁棒性和自适应性更优的解决方案 [70]。LSTM 可以有效捕捉信道条件的时间依赖性，提高估计精度 [71]。反馈机制将信道状态信息（Channel State Information，CSI）从接收端传回发送端，便于实时调整发

射设置。传统反馈方法通常涉及 CSI 量化和编码，可能引入时延和误差。自编码器和 CNN 等机器学习技术通过压缩、重构 CSI，在尽量减少信息损失的同时提高反馈效率 [52]，从而更精确及时地调整传输策略。此外，机器学习模型可以同时处理信道估计和反馈，从而整体优化这两个过程 [69]，提升系统性能。

4 语义通信

与传统物理层通信不同，语义通信侧重于传输图像、文本、音频等数据中的语义信息，而非逐位传输原始数据。相比之下，语义通信能显著减少无线信道的传输负载，从而大幅提升信息传输速度和效率。在语义通信方面，[72–75] 等文献提供了最新研究，此处不再赘述。

语义通信的核心是从原始数据中提取语义信息，这可以通过精心设计的信源编解码策略或 JSCC 来实现。然而，原始数据中的语义信息通常仅限于特定任务（见图 7），目前尚未形成通用的数学分析、建模和计算框架。关于该方向的探索性研究，请参见 [76] 等文献。因此，对于特定通信任务，需要设计合适的语义编解码方案。在智能传输与处理系统中，语义通信由在高度集成的端到端收发器实现，参见图 2。

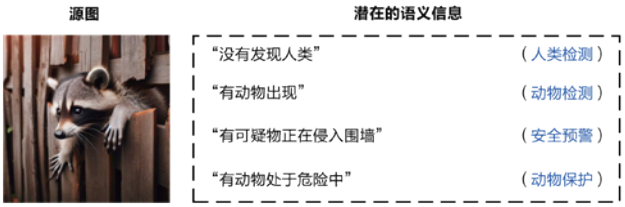


图 7 每个任务的原始数据包含不同的语义信息。无损传输高清图像既耗时又耗资源，而准确传输语义消息则相对简单且成本更低。（图片由 Microsoft Copilot 生成）

机器学习使系统可以更准确地理解、处理和传达语义，在语义通信中起着关键作用。深度学习模型（如 Transformer、CNN 和 RNN）已广泛应用于分析和预测数据的语义相关性，从而提高带宽利用率并通信效率。具体而言，自然语言处理（Natural Language Processing，NLP）算法可以从文本和音频中提取并解释语义内容，更高效地压缩与传输数据 [77, 78]；计算机视觉方法则可从图像和视频中提取并解释语义信息 [79]。

最近研究表明，基于机器学习的语义通信应用潜力巨大。例如，通过设计收发器神经网络直接传输文本语义 [13]，可显著减少通信资源需求并提升传输性能；开发高效视频会议系统 [80]，可提高传输效率。大多数语义通信研究致力于通过 JSCC 来节省资源，但这些成果需要改变现有基础设施，不利于实际应用。在此背景下，[14] 提出一个务实方法，仅修改现有基础设施中的部分模块来实现无线语义传输。为确保语义传输的可靠性和通信效率，有学者研究了语义域中的频谱效率及基于语义感知的资源分配问题 [81]。此外，语义通信与物联网（Internet of Things，IoT）[82]、边缘计算 [83] 等新兴技术协作，为智能化、上下文感知通信系统创造了新机遇。基于分布式机器学习模型，语义通信系统可以动态适应环境变化和用户需求，确保稳健高效的信息交换 [84]。

总之，语义通信依托先进的机器学习技术，为通信系统指明了发展方向，有望革新信息传输和解码方式，将会对未来的通信效率和效能产生深远影响。

5 通信资源分配

无线通信中的资源分配旨在高效管理和利用频谱、功率、计算、空间和时间等资源，以提高网络性能，实现更高的吞吐率、更低的时延、更广的覆盖范围和更高的可靠性 [15, 85, 86]。典型应用包括链路调度、消息路由、功率分配、信道选择、波束赋形、频谱接入与管理以及信道划分协议设计（即时分、频分等）等。从数学规划角度来看，资源分配可表示为优化问题。从运筹学角度，分配和调度是关键技术——分配解决静态资源分配问题，而调度则解决动态资源分配问题（静态和动态是相对时间而言的）。从计算和算法角度，主流解决方案框架包括：

- 连续优化、离散（如组合、整数）优化和混合优化
- 单目标优化和多目标优化
- 线性规划和非线性规划
- 凸优化与非凸优化
- 光滑优化与非光滑优化
- Min−Max 优化（如博弈论、最坏情况鲁棒性分析）
- 确定性规划和随机规划（即是否涉及随机变量；如涉及，考虑相应的概率分布）
- 单阶段优化（即静态规划）和多阶段优化（即动态规划）
- 启发式优化（如遗传算法、粒子群优化和模拟退火）
- 代理优化，也称黑盒优化（如贝叶斯优化）
- 基于机器学习的优化（如基于强化学习和深度学习的求解方法）

无线通信中的典型资源分配应用及解决方案框架如图 8 所示，更多信息请参考 [85, 87, 88]。

随着 ISAC 的出现 [89]，对资源分配的要求有所改变，即便是最佳的通信资源分配方案，也未必同样适用于感知业务（详见 [90]）。例如，由于通信和感知功能设计偏好存在差异甚至冲突，二者的最优波形也有所不同 [91, 92]。因此，应细化空口、计算、功率、时间、波束等资源的分配方案，满足通信和感知的性能要求。ICAC（如边缘计算 [93] 和网络控制* [94]）也面临同样的困境，需在计算与通信之间合理分配有限的资源（详见 [95, 96]）。

* 控制器本质上是信息处理器，因此属于特定用途的计算模块。

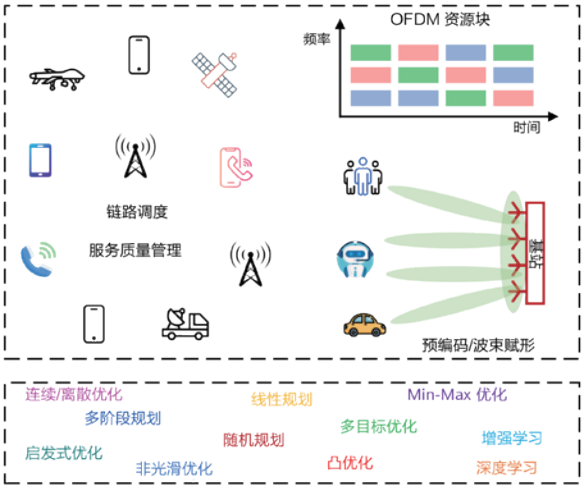


图 8 无线通信中的典型资源分配应用及解决方案框架。OFDM：正交频分复用。（图片来源：FLATICON.com）

机器学习技术已成为解决无线通信中资源分配问题的有力工具。最近研究表明，机器学习在各类资源分配任务中具有巨大潜力。例如，深度强化学习已应用于优化异构网络中的频谱分配、功率控制和用户关联等问题，与传统方法相比，优化效果显著提升 [22, 97–102]。同样，监督学习算法已成功应用于解决混合整数非线性规划（Mixed Integer Nonlinear Programming，MINLP）等资源分配的复杂优化问题 [103]。此外，无监督学习技术还可用于解决资源分配问题，如图嵌入技术在链路调度中的应用 [104]。

传统资源分配方法在很大程度上依赖人类智力构建精确模型、开发特定方案，但在大规模动态复杂场景中可能表现欠佳甚至无法适应。机器学习，尤其是深度学习和强化学习等技术，可以对环境中的复杂交互进行建模，预测未来通信网络状态，并实时优化决策，从而提升无线网络性能和适应性 [15]。例如，现网信道条件会不断变化，但我们无法预知信道会如何随时间而变化，因此难以用数学手段处理这种不确定性。此外，数学规划模型计算复杂，很难高效解决问题。而机器学习则可以利用现实数据，发现数据中隐含知识和模式，并自动做出令人满意的资源分配决策。从技术角度来看，机器学习可以将资源的分配优化问题视为参数–决策的映射问题，从而帮助解决复杂的优化问题。这种映射依赖于深度神经网络强大的有监督学习函数拟合能力，标注数据–决策对由性能好的人工解法生成。此外，机器学习还可将资源分配问题中的效用函数作为训练阶段中的损失函数，无需依赖传统人工算法即可生成高质量的资源分配决策。除了这两种方案外，还有一种“算法展开”技术 [18]，即使用神经网络来展开现有高效迭代算法。具体来说，每层神经网络充当迭代算法的一个迭代步骤，通过多层级联模拟该算法的迭代过程。这种由算法指导的方法也称为“模型驱动的深度学习”[12]，根据领域知识定制深度神经网络架构，以提高网络泛化能力，并减少训练所需的数据量。强化学习是第四种有前景的方案。强化学习通过探索未知且难以建模的环境（如复杂信道条件）并与其动态交互来获得智能资源分配方案。图 9 总结了机器学习在资源分配中的四个典型职能。机器学习方法的第一个优势是在运行阶段计算速度极快——尽管训练阶段的计算量可能较大（相比前三种方案）；第二个优势在于其不依赖理建模即可应对动态、不确定甚至未知的环境（相比第四种方案，即强化学习方案）。

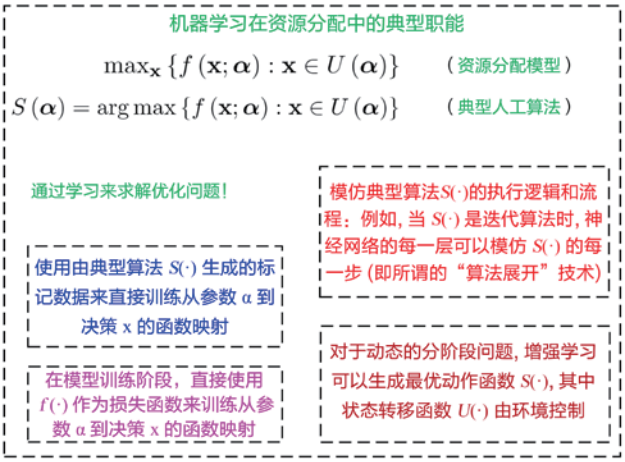


图 9 机器学习在资源分配中的四个典型职能（以学习的方式求解优化问题）

6 不仅仅是数据传输：感知与计算

随着对低时延高速连接、高性能通信辅助感知（如用户定位和跟踪）及协同计算设备需求的增加，无线通信系统正在经历一场变革，这推动了 ISAC [89, 105] 和 ICAC [106, 107] 等新型系统范式的发展，使得传统上分离的功能得到了统一，从而提高资源利用率、降低硬件成本并增强系统能力。例如，通过波束管理和资源分配，利用环境和用户传感数据提升通信性能；跨网络节点的传感数据共享实现实时网络监控与态势感知，从而提高感知精度并扩大覆盖范围。再如，边缘本地数据处理可减少实时通信业务的时延，而高速通信则为大规模数据分析提供高效的分布式计算。如前所述，机器学习，尤其是深度学习，在现代通信系统中至关重要。它们利用算法从海量数据中学习，优化网络资源分配、信号处理和故障检测等性能。这些优势同样适用于新兴的 ISAC 和 ICAC 系统。在 ISAC 系统中，CNN 和 Transformer 等深度学习模型可以提升感知精度与鲁棒性 [108]，并实现语义信息传输 [109]。在 ICAC 系统中，联邦学习等机器学习算法能够保护用户数据隐私并优化计算任务，提高数据处理与通信效率 [110, 111]。总之，机器学习 / 深度学习与新兴的一体化范式相结合，将使通信系统变得更加智能、自适应且高效。

6.1 通感一体化

ISAC 是一种将感知与通信融合为一体的系统，感知与通信共享基础设施和频谱资源。该系统对自动驾驶、智慧城市和先进监控系统等应用至关重要。通过同步数据采集与通信，ISAC 提升系统效率和性能，降低硬件成本并缓解频谱拥塞。然而，ISAC 使通信波形设计、系统和硬件资源分配、干扰管理及整体网络操作更加复杂 [112, 17]。这些挑战推动了方法创新，有利于释放 ISAC 在现实应用中的潜力。因此，机器学习和深度学习技术为 ISAC 提供了先进的数据处理与决策能力，其重要性不言而喻。如欲系统了解机器学习在 ISAC 的作用，请参见 [17] 和 [112]。

6.2 算通一体化

ICAC 是一种将计算和通信融为一体的技术。为满足应用日益增长的计算需求，同时保持高效稳健的通信性能，在源端就近处理大规模数据任务，减少时延并提升边缘计算效率，实现连接设备智能化，ICAC 应运而生。该技术有助于实时处理和分析数据，对于工业自动化、虚拟现实以及物联网等应用至关重要。ICAC 典型应用包括边缘计算、联邦学习、普适计算、雾计算、物联网 / 车联网及自治系统等。机器学习和深度学习是 ICAC 的重要组成部分，能够实现动态资源分配、系统配置自适应和实时信息分析。这些技术确保计算与通信资源得到有效利用，从而提升性能和响应能力。关于机器学习在 ICAC 中的具体作用，请参阅 [16]、[113] 和 [114]。需要注意的是，群体智能和网络控制 [94] 也与 ICAC 密切相关，因为控制器本质上是将系统状态信号映射到系统控制输入信号的信息处理器，属于特定用途的计算模块。

7 讨论和结论

本文探讨了机器学习在物理层通信、语义通信、资源分配、ISAC 及 ICAC（如联邦学习、边缘计算）等关键无线应用中的变革潜力。从信号处理算法到整网管理，机器学习都有着广泛的影响。尽管如此，我们不能过分夸大机器学习的作用，在实际应用中机器学习仍会面临一些挑战。模型的可解释性、故障处理、大规模训练数据集需求以及训练和部署所需的计算资源（如电力、处理速度）等问题必须解决。此外，还需关注机器学习系统的可靠性和安全性，尤其是在涉及数据隐私（如联邦学习 [115]）、数据时效性（如小样本学习 [116, 117]）及实时决策（如自动驾驶）等场景中。为应对这些挑战，本文建议将传统物理定律模型与数据驱动的机器学习模型相结合，这样既能利用物理机制的可靠性和可解释性，又能发挥机器学习的自适应性和学习能力，从而提升整体系统性能。关于智能传输与处理的特征、挑战及未来考量，详见图 10。

在所有可预见的挑战中，以下三项是实现基于机器学习通信系统的最低要求，也是实际应用中的关键挑战：

- 如何解读机器学习模型的性能增益与故障，并在系统宕机时排查和修复故障，以提高整体可靠性？从这个角度看，图 6 中的范式比图 2 中的更可靠、更易管理。
- 如何利用有限数据获得更好的泛化能力？如何整合新的可用数据来提高泛化能力 [20, 31]，包括如何将已习得模型快速适配到新数据上（如当环境中的数据生成规则随时间变化时 [62, 117]）？在机器学习领域，数据时效性、样本效率和数据分布鲁棒性都与此问题密切相关。
- 如何构建基于领域知识的机器学习模型（而非多层感知等通用深度神经网络）？如何设计计算高效的训练算法（而非随机梯度下降法）以缩短响应时间、降低功耗 [12, 108, 118, 119]？如何缩小模型（尤其是深度神经网络）大小以节省存储空间 [120]？对于嵌入式设备和边缘设备，这三点考虑尤为重要。

简言之，机器学习与通信系统的融合是重大的技术进步，为建立更加智能、高效、可靠的通信网络提供了可能。

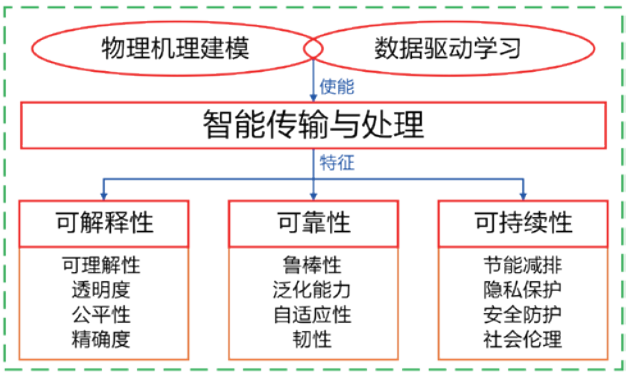


图 10 智能传输与处理的特点、挑战和注意事项（对照图 3）。尽管数据驱动的机器学习很强大，但机理建模（包括发现物理或数学定律）始终对提高可解释性、可靠性和可持续性至关重要。

参考文献

[1] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2006.

[3] G. L. Stüber and G. L. Steuber, *Principles of Mobile Communication*, 4th ed. Springer, 2017.

[4] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.

[5] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, vol. 24, no. 2, pp. 98–105, 2016.

[6] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 93–99, 2019.

[7] D. Gündüz, P. De Kerret, N. D. Sidiropoulos, D. Gesbert, C. R. Murthy, and M. Van der Schaar, "Machine learning in the air," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2184–2199, 2019.

[8] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1472–1514, 2020.

[9] W. Yu, F. Sohrabi, and T. Jiang, "Role of deep learning in wireless communications," *IEEE BITS the Information Theory Magazine*, vol. 2, no. 2, pp. 56–72, 2022.

[10] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges," *International Journal of Intelligent Systems*, vol. 2024, no. 1, p. 8845070, 2024.

[11] A. Celik and A. M. Eltawil, "At the dawn of generative AI era: A tutorial-cum-survey on new frontiers in 6G wireless intelligence," *IEEE Open Journal of the Communications Society*, 2024.

[12] H. He, S. Jin, C.-K. Wen, F. Gao, G. Y. Li, and Z. Xu, "Modeldriven deep learning for physical layer communications," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 77–83, 2019.

[13] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[14] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic transmission via revising modules in conventional communications," *IEEE Wireless Communications*, vol. 30, no. 3, pp. 28–34, 2023.

[15] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 341–356, 2019.

[16] M. A. Ferrag, O. Friha, B. Kantarci, N. Tihanyi, L. Cordeiro, M. Debbah, D. Hamouda, M. Al-Hawawreh, and K.-K. R. Choo, "Edge learning for 6G-enabled Internet of things: A comprehensive survey of vulnerabilities, datasets, and defenses," *IEEE Communications Surveys & Tutorials*, 2023.

[17] S. Lu, F. Liu, Y. Li, K. Zhang, H. Huang, J. Zou, X. Li, Y. Dong, F. Dong, J. Zhu, *et al*., "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet of Things Journal*, 2024.

[18] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Processing Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[19] N. Shlezinger and T. Routtenberg, "Discriminative and generative learning for the linear estimation of random signals [lecture notes]," *IEEE Signal Processing Magazine*, vol. 40, no. 6, pp. 75–82, 2023.

[20] S. Wang, W. Dai, and G. Y. Li, "Distributionally robust receive beamforming," *arXiv preprint arXiv:2401.12345*, 2024.

[21] G. James, D. Witten, T. Hastie, R. Tibshirani, *et al*., *An Introduction to Statistical Learning*, 2nd ed. Springer, 2021.

[22] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072–3108, 2019.

[23] Y. C. Eldar, A. Goldsmith, D. Gündüz, and H. V. Poor, *Machine Learning and Wireless Communications*. Cambridge University Press, 2022.

[24] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[25] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1702–1715, 2020.

[26] B. Thuraisingham, "Trustworthy machine learning," *IEEE Intelligent Systems*, vol. 37, no. 1, pp. 21–24, 2022.

[27] K. R. Varshney, *Trustworthy Machine Learning*. Chappaqua, NY, USA: Independently Published, 2022.

[28] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Leanpub, 2020.

[29] K. Kawaguchi, Z. Deng, K. Luh, and J. Huang, "Robustness implies generalization via data-dependent generalization bounds," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 866–10 894.

[30] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and S. Y. Philip, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.

[31] S. Wang and H. Wang, "Distributional robustness bounds generalization errors," *arXiv preprint arXiv:2212.09962*, 2024.

[32] A. Van Wynsberghe, "Sustainable AI: AI for sustainability and the sustainability of AI," *AI and Ethics*, vol. 1, no. 3, pp. 213–218, 2021.

[33] S. AbdulRahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi, and M. Guizani, "A survey on federated learning: The journey from centralized to distributed on-site learning and beyond," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5476–5497, 2020.

[34] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, "Distributed learning in wireless networks: Recent progress and future challenges," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 12, pp. 3579–3605, 2021.

[35] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3133–3143, 2020.

[36] H. Ye, G. Y. Li, and B.-H. Juang, "Deep learning-based end-to-end wireless communication systems without pilots." *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 3, pp. 702–714, 2021.

[37] S. Manouchehri, J. Haghighat, M. Eslami, and W. Hamouda, "A delay-efficient deep learning approach for lossless turbo source coding," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 6704–6709, 2022.

[38] H. Ye, L. Liang, and G. Y. Li, "Circular convolutional autoencoder for channel coding," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2019, pp. 1–5.

[39] Y. Zhang, H. Wu, and M. Coates, "On the design of channel coding autoencoders with arbitrary rates for ISI channels," *IEEE Wireless Communications Letters*, vol. 11, no. 2, pp. 426–430, 2021.

[40] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Deep joint source-channel coding for wireless image retrieval," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5070–5074.

[41] M. Yang, C. Bian, and H.-S. Kim, "Deep joint source-channel coding for wireless image transmission with OFDM," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[42] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 114–117, 2017.

[43] S. Mohammadzadeh, V. H. Nascimento, R. C. de Lamare, and N. Hajarolasvadi, "Robust beamforming based on complexvalued convolutional neural networks for sensor arrays," *IEEE Signal Processing Letters*, vol. 29, pp. 2108–2112, 2022.

[44] A. M. Elbir, K. V. Mishra, M. R. B. Shankar, and B. Ottersten, "A family of deep learning architectures for channel estimation and hybrid beamforming in multi-carrier mm-wave massive MIMO," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 642–656, 2022.

[45] D. d. S. Brilhante, J. C. Manjarres, R. Moreira, L. de Oliveira Veiga, J. F. de Rezende, F. Müller, A. Klautau, L. Leonel Mendes, and F. A. P. de Figueiredo, "A literature survey on AI-aided beamforming and beam management for 5G and 6G systems," *Sensors*, vol. 23, no. 9, p. 4359, 2023.

[46] N. Shlezinger, M. Ma, O. Lavi, N. T. Nguyen, Y. C. Eldar, and M. Juntti, "Artificial intelligence-empowered hybrid multiple-input/multiple-output beamforming: Learning to optimize for high-throughput scalable MIMO," *IEEE Vehicular Technology Magazine*, 2024.

[47] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, "Deep learning-based beam tracking for millimeter-wave communications under mobility," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7458–7469, 2021.

[48] F. Sohrabi, T. Jiang, W. Cui, and W. Yu, "Active sensing for communications by learning," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1780–1794, 2022.

[49] Y. Wei, Z. Zhong, and V. Y. Tan, "Fast beam alignment via pure exploration in multi-armed bandits," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3264–3279, 2022.

[50] W. Yi, W. Zhiqing, and F. Zhiyong, "Beam training and tracking in mmWave communication: A survey," *China Communications*, 2024.

[51] Q. Hu, F. Gao, H. Zhang, S. Jin, and G. Y. Li, "Deep learning for channel estimation: Interpretation, performance, and comparison," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2398–2412, 2020.

[52] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Convolutional neural network-based multiple-rate compressive sensing for massive MIMO CSI feedback: Design, simulation, and analysis," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2827–2840, 2020.

[53] B. Ozpoyraz, A. T. Dogukan, Y. Gevez, U. Altun, and E. Basar, "Deep learning-aided 6G wireless networks: A comprehensive survey of revolutionary PHY architectures," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1749–1809, 2022.

[54] N. Ye, S. Miao, J. Pan, Q. Ouyang, X. Li, and X. Hou, "Artificial intelligence for wireless physical-layer technologies (AI4PHY): A comprehensive survey," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[55] Y. Yang, S. Mandt, L. Theis *et al*., "An introduction to neural data compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023.

[56] S. Han, J. Oh, K. Oh, and J. Ha, "Deep-learning for breaking the trapping sets in low-density parity-check codes," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 2909–2923, 2022.

[57] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[58] K. Choi, K. Tatwawadi, A. Grover, T. Weissman, and S. Ermon, "Neural joint source-channel coding," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1182–1192.

[59] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 567–579, 2019.

[60] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, "Generative joint source-channel coding for semantic image transmission," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 8, pp. 2645–2657, 2023.

[61] E. Bobrov, D. Kropotov, H. Lu, and D. Zaev, "Massive MIMO adaptive modulation and coding using online deep learning algorithm," *IEEE Communications Letters*, vol. 26, no. 4, pp. 818–822, 2021.

[62] N. Van Huynh and G. Y. Li, "Transfer learning for signal detection in wireless networks," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2325–2329, 2022.

[63] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1581–1592, 2019.

[64] M. Chu, A. Liu, V. K. Lau, C. Jiang, and T. Yang, "Deep reinforcement learning based end-to-end multiuser channel prediction and beamforming," *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 271–10 285, 2022.

[65] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1065–1069, 2019.

[66] P. Ramezanpour, M. J. Rezaei, and M. R. Mosavi, "Deep learning-based beamforming for rejecting interferences," *IET Signal Processing*, vol. 14, no. 7, pp. 467–473, 2020.

[67] K. Chen, C. Qi, C.-X. Wang, and G. Y. Li, "Beam training and tracking for extremely large-scale MIMO communications," *IEEE Transactions on Wireless Communications*, 2023.

[68] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Transactions on Communications*, vol. 70, no. 12, pp. 8017–8045, 2022.

[69] J. Guo, T. Chen, S. Jin, G. Y. Li, X. Wang, and X. Hou, "Deep learning for joint channel estimation and feedback in massive MIMO systems," *Digital Communications and Networks*, vol. 10, no. 1, pp. 83–93, 2024.

[70] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Dual CNN-based channel estimation for MIMO-OFDM systems," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5859–5872, 2021.

[71] R. Shankar, "Bi-directional LSTM based channel estimation in 5G massive MIMO OFDM systems over TDL-C model with Rayleigh fading distribution," *International Journal of Communication Systems*, vol. 36, no. 16, p. e5585, 2023.

[72] W. Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: Fundamentals, applications, and challenges," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 213–250, 2022.

[73] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[74] Z. Lu, R. Li, K. Lu, X. Chen, E. Hossain, Z. Zhao, and H. Zhang, "Semantics-empowered communications: A tutorial-cum-survey," *IEEE Communications Surveys & Tutorials*, 2023.

[75] C. Chaccour, W. Saad, M. Debbah, Z. Han, and H. V. Poor, "Less data, more knowledge: Building next generation semantic communication networks," *IEEE Communications Surveys & Tutorials*, 2024.

[76] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 5–41, 2022.

[77] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of Artificial Intelligence*, pp. 603–649, 2020.

[78] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023.

[79] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.

[80] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 1, pp. 230–244, 2022.

[81] L. Yan, Z. Qin, R. Zhang, Y. Li, and G. Y. Li, "Resource allocation for text semantic communications," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1394–1398, 2022.

[82] H. Xie and Z. Qin, "A lite distributed semantic communication system for Internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[83] W. Yang, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Cao, and K. B. Letaief, "Semantic communication meets edge intelligence," *IEEE Wireless Communications*, vol. 29, no. 5, pp. 28–35, 2022.

[84] H. Tong, Z. Yang, S. Wang, Y. Hu, O. Semiari, W. Saad, and C. Yin, "Federated learning for audio semantic communication," *Frontiers in Communications and Networks*, vol. 2, p. 734402, 2021.

[85] Z. Han and K. R. Liu, *Resource Allocation for Wireless Networks: Basics, Techniques, and Applications*. Cambridge University Press, 2008.

[86] Y. Teng, M. Liu, F. R. Yu, V. C. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2134–2168, 2018.

[87] R. Zheng and C. Hua, *Sequential Learning and Decision-Making in Wireless Resource Management*. Springer, 2016.

[88] E. Hossain, M. Rasti, and L. B. Le, *Radio Resource Management in Wireless Networks: An Engineering Approach*. Cambridge University Press, 2017.

[89] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

[90] F. Dong, F. Liu, Y. Cui, W. Wang, K. Han, and Z. Wang, "Sensing as a service in 6G perceptive networks: A unified framework for ISAC resource allocation," *IEEE Transactions on Wireless Communications*, vol. 22, no. 5, pp. 3522–3536, 2022.

[91] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu *et al.*, "A survey on fundamental limits of integrated sensing and communication," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 994–1034, 2022.

[92] S. Wang, W. Dai, H. Wang, and G. Y. Li, "Robust waveform design for integrated sensing and communication," *IEEE Transactions on Signal Processing*, 2024.

[93] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[94] X. Ge, F. Yang, and Q.-L. Han, "Distributed networked control systems: A brief overview," *Information Sciences*, vol. 380, pp. 117–131, 2017.

[95] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, "Resource scheduling in edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2131–2165, 2021.

[96] H. Djigal, J. Xu, L. Liu, and Y. Zhang, "Machine and deep learning for resource allocation in multi-access edge computing: A survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2449–2494, 2022.

[97] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *2017 IEEE International Conference on Communications (ICC)*. IEEE, 2017, pp. 1–6.

[98] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[99] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3163–3173, 2019.

[100] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141–5152, 2019.

[101] X. Xiong, K. Zheng, L. Lei, and L. Hou, "Resource allocation based on deep reinforcement learning in IoT edge computing," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 6, pp. 1133–1146, 2020.

[102] K. Xu, N. Van Huynh, and G. Y. Li, "Distributed-training-and-execution multi-agent reinforcement learning for power control in hetnet," *IEEE Transactions on Communications*, 2023.

[103] M. Lee, G. Yu, and G. Y. Li, "Learning to branch: Accelerating resource allocation in wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 958–970, 2019.

[104] ——, "Graph embedding-based wireless link scheduling with few training samples," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2282–2294, 2020.

[105] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3834–3862, 2020.

[106] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for B5G networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 1, pp. 9–39, 2023.

[107] D. Wen, X. Li, Y. Zhou, Y. Shi, S. Wu, and C. Jiang, "Integrated sensing-communication-computation for edge artificial intelligence," *IEEE Internet of Things Magazine*, vol. 7, no. 4, pp. 14–20, 2024.

[108] B. Zhang and G. Y. Li, "White-box 3D-OMP-transformer for ISAC," *arXiv preprint arXiv:2407.02251*, 2024.

[109] B. Zhang, Z. Qin, and G. Y. Li, "Compression ratio learning and semantic communications for video imaging," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[110] S. Zhou and G. Y. Li, "FedGiA: An efficient hybrid algorithm for federated learning," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1493–1508, 2023.

[111] ——, "Federated learning via inexact ADMM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9699–9708, 2023.

[112] U. Demirhan and A. Alkhateeb, "Integrated sensing and communication for 6G: Ten key machine learning roles," *IEEE Communications Magazine*, vol. 61, no. 5, pp. 113–119, 2023.

[113] S. H. Alsamhi, A. V. Shvetsov, S. Kumar, J. Hassan, M. A. Alhartomi, S. V. Shvetsova, R. Sahal, and A. Hawbani, "Computing in the sky: A survey on intelligent ubiquitous computing for UAV-assisted 6G networks and Industry 4.0/5.0," *Drones*, vol. 6, no. 7, p. 177, 2022.

[114] V. A. Nugroho and B. M. Lee, "A survey of federated learning for mmwave massive MIMO," *IEEE Internet of Things Journal*, 2024.

[115] H. Ye, L. Liang, and G. Y. Li, "Decentralized federated learning with unreliable communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 487–500, 2022.

[116] O. Wang, S. Zhou, and G. Y. Li, "Few-shot learning for new environment adaptation," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 351–356.

[117] O. Wang, J. Gao, and G. Y. Li, "Learn to adapt to new environments from past experience and few pilot blocks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 2, pp. 373–385, 2022.

[118] Y. Liu, Z. Qin, and G. Y. Li, "Energy-efficient distributed spiking neural network for wireless edge intelligence," *IEEE Transactions on Wireless Communications*, 2024.

[119] O. Wang, S. Zhou, and G. Y. Li, "BADM: Batch ADMM for deep learning," *arXiv preprint arXiv:2407.01640*, 2024.

[120] H. Cai, C. Gan, L. Zhu, and S. Han, "TinyTL: Reduce memory, not parameters for efficient on-device learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11 285–11 297, 2020.

5.5G 时代的人工智能：场景、关键技术与未来趋势

林英沛，陈雁，秦熠，孙琰，徐瑞，杨玉雯，张征明，陈家璇，田洋，曹佑龙，柴晓萌，陈宏智，齐鸿，庞旭
无线网络研究部

摘要

人工智能（Artificial Intelligence，AI）作为当代技术革命的核心驱动力，其与 5.5G 的融合预示着通信领域的一次革命性飞跃。本文将探讨 AI 的演进趋势，分析 AI 在 5.5G 网络中的关键价值，讨论 AI 时代下新出现的应用场景，例如人工智能生成内容（Artificial Intelligence Generated Content，AIGC）和具身智能，并识别这些应用场景对 5.5G 网络的关键需求与挑战。针对这些关键挑战，本文进一步阐述了 AI 如何在 5.5G 时代提升网络性能，以及 5.5G 网络如何提供高质量的 AI 服务的关键技术。最后，文章对 AI 与 5.5G 深度融合的未来进行了展望，期待这一融合能够引领我们进入一个智能化和个性化的通信新时代。

关键词

人工智能，5.5G，网络性能，AI 服务，AIGC，具身智能

1 引言

在信息时代，通信技术的发展一直是推动社会变革的重要力量。5.5G 作为最新一代的通信技术，以其高速率、低延迟和高可靠性的特点，为各种新兴技术的应用提供了广阔的平台。在这一背景下，人工智能（Artificial Intelligence，AI）技术与 5.5G 的结合，预示着通信领域的一次革命性飞跃。

AI 技术，以其强大的数据处理能力和智能决策支持，已经成为现代技术革命的核心。从基础理论研究到广泛的行业应用，AI 正不断拓展其影响力，尤其在移动通信领域，AI 的应用已经开始重塑网络架构和服务模式。本文旨在探讨 5.5G 时代 AI 技术的关键作用和发展趋势，分析 AI 如何助力 5.5G 网络性能的提升，并展望 AI 技术在未来通信领域的应用潜力。

本文通过回顾 AI 模型、算力、数据和应用范式的发展进程，总结了 AI 演进的趋势（第 2 节），分析了 AI 在 5.5G 网络中的关键价值（第 3 节），讨论了 AI 时代下新出现的应用场景，例如人工智能生成内容（Artificial Intelligence Generated Content，AIGC）和具身智能，并识别了这些应用场景对 5.5G 网络的关键需求与挑战（第 4 节）。针对这些关键挑战，本文进一步阐述了 AI 如何在 5.5G 时代提升网络性能，以及 5.5G 网络如何提供高质量的 AI 服务的关键技术（第 5 节）。最后，本文对 AI 与 5.5G 深度融合的未来进行了展望，期待这一融合能够开启一个智能化和个性化的通信新时代（第 6 节）。

2 AI 的演进趋势

AI 不仅已成为当代技术革命的核心驱动力，而且通过深度挖掘数据中的高价值信息，在各行各业产生了无可争辩的深远影响。AI 作为一门多学科交叉的学科，不仅涉及基础理论和前沿研究，更是一种广泛应用于各个领域的技术，具有显著的实际影响和商业价值。AI 技术的突破性发展，特别是在移动互联网等行业的重塑，不仅重新定义了我们对可能性的认知，还推动了社会和经济的深刻变革。AI 的发展历程充满了令人难以置信的突破、质疑与不断演变。站在 AI 进一步发展的新起点，我们清晰地认识到 AI 的潜力是无限的。随着 5.5G 时代的到来，AI 的爆发式发展要求我们积极拥抱这一技术，以实现不断的突破和持续的进步。

2.1 AI 模型的发展

以神经网络模型为代表的 AI 模型于 1943 年，由 Warren Sturgis McCulloch 和 Walter Pitts 共同提出 [1]，在经历跌宕起伏的发展与质疑的岁月后，以 Frank Rosenblatt 提出的感知机 [2] 为基础的 AI 模型于 1993 年正式进入蓬勃发展阶段。在此之后，机器硬件快速迭代更新，运算能力及储存空间大幅提升，AI 模型渐渐地在各个领域应用，并再次受到企业各界的重视。研究者通过为感知机添加更多的层创建了可以在硬件实体上工作的多层感知机，并将其用于图像的识别。

在多层感知机显示出解决图像识别等复杂问题的潜力之后，研究者开始思考如何发明新型的 AI 模型以对文本等数据进行建模。最终于 1997 年长短期记忆（Long Short-Term Memory，LSTM）[3] 递归神经网络被提出并对后来的 AI 的研究产生了深远影响。得益于计算与存储设备的迭代升级，涉及由三个门（输入门、遗忘门和输出门）控制的记忆单元的 LSTM 架构可以通过逻辑门决定 AI 模型的记忆增减和输出。LSTM 作为循环神经网络的代表在处理长序列的问题上表现出强大的能力，并成为文本分类、情感分析、语音识别、图像标题生成和机器翻译等序列任务的经典神经网络架构。然而，为获得最佳的效果，研究者发现这类 AI 模型的参数规模较大并且计算成本较高，在计算能力受限的情况下仍不能满足需求。

2006 年，Geoffrey Hinton 发明出了受限玻尔兹曼机模型与深度信念网络可以训练多层神经网络，并将多层神经网络正式命名为“深度学习”（Deep Learning，DL）[4]。从此，AI 模型进入深度神经网络时代。以深度神经网络为基座堆叠更多的层和设计更巧妙的结构为代表的 AI 模型不断涌现出来，并开始取得让人振奋的结果。例如，由 5 个卷积层、1 个最大池化层、3 个全连接层和一个 softmax 层组成的 AlexNet [5] 赢得了 ImageNet 大规模视觉识别挑战赛。业内迅速认识到深度卷积神经网络可以很好地处理视觉识别任务，并由此开启了卷积神经网络大规模开发与应用的时代。最具代表性的 VGG [6] 和 ResNet [7] 被先后提出，并成功将计算机视觉任务的性能提升到前所未有的高度。AI 模型从原来的简单两层神经网络发展到了具有 10 层以上卷积层的 VGG 模型和具有 50 层以上卷积层的 ResNet 模型。同时，AI 模型的能力从原来识别光学字符识别（Optical Character Recognition，OCR）等简单图像变成了实现语义分割，实例分割等高级任务。但与此同时，研究者也注意到这些卷积神经网络在自然语言处理任务上还存在一定的局限性，AI 模型在自然语言的理解方面仍效果欠佳。

随着以注意力机制为重要组成单元的 Transformer [8] 横空出世，AI 模型遇到的自然语言理解的挑战取得了重要突破。Transformer 是一类基于注意力机制的神经网络模型，这类模型不使用循环网络或卷积，而是由多头注意力、残差连接、层归一化、全连接层和位置编码组成，在保留序列顺序的同时挖掘序列信息的相关性。Transformer 彻底改变了自然语言处理，并迅速成为主导其他方向（例如，计算机视觉领域）的核心 AI 模型。在自然语言处理中，这类 AI 模型被用于机器翻译、文本摘要、语音识别、文本

补全、文档搜索等，并最终导致 ChatGPT 等大语言模型（Large Language Model，LLM）的问世。由 OpenAI 于 2022 年 11 月发布的 ChatGPT 成为生成式 AI 的现象级应用。该 AI 模型基于 GPT-3.5 架构并通过强化学习进行训练，完成了 AI 模型从图像识别到图像理解再到生成式人工智能（Generative Artificial Intelligence，GenAI）的华丽转变 [9]。而 AI 模型的参数规模也从原来的几千个变成了以 million（百万）为单位和以 billion（十亿）为单位。

2.2 AI 算力的发展

计算能力，也即算力，是 AI 发展与进步的关键驱动力。在过去的 10 年中，用于训练 AI 模型的计算量增加了 3.5 亿倍。AI 的许多进步源于用于训练和运行 AI 模型的算力的显著增加。在 LLM [9]、AlphaGo [10]、蛋白质折叠模型 [11] 和自动驾驶模型 [12] 中，与以往训练和部署小规模 AI 模型不同的是，开发人员成功地利用了巨大的算力，在庞大的数据集上训练模型，以使得 AI 模型学习如何解决问题。在许多 AI 领域，研究人员发现了缩放规律：训练目标（例如“预测下一个单词”）的性能会随着用于训练模型的计算量的增加而增加。

得益于硬件的改进，无论是终端设备还是基站和云端，算力都得到了前所未有的提升。随着 CPU、GPU 和 TPU 等先进计算设备的发展，计算不仅限于传统的数据中心，计算开始走向边缘，并逐渐走向全栈场景。通过部署具备计算能力的智能设备，云、边、端的计算能力都成为了 AI 的算力，有效提升了 AI 的计算效率和性能。其中，云将收集到的数据存储到数据中心，在中心点进行计算和处理；基站设施组为边缘计算设备，也开始具有并发挥出强大的数据计算和处理能力；并且，随着终端设备智能化程度加深和 AI 模型的升级迭代，智能终端处理器的性能大幅提升，终端算力得到了增强。

2.3 AI 数据的发展

训练 AI 模型需要高质量和大规模的数据，这些数据是 AI 的燃料，是 AI 取得成功的决定性因素之一。在 AI 发展初期，研究者以人力采集的方式构建极其有限量的数据集，这些数据被用于 AI 模型的训练和测评。然而，随着 AI 模型规模和能力的发展，研究者很快注意到训练数据成为了获得更高性能 AI 模型的瓶颈。对高质量训练数据的渴求成为了当前 AI 发展的重大挑战。例如，当前大规模的 AI 语言模型是利用从互联网上获取的文本建立的，包括科学研究、新闻报道、维基百科条目等，这些数据被分解为词元。据研究人员估计，GPT4 模型所使用的训练数据已经包含了高达 12 万亿的词元。AI 模型在未来继续遵循当前的增长轨迹则需要 60 万亿到 100 万亿的词元用于训练。为应对训练数据枯竭，AI 领域需要更强的数据获取能力。

2.4 AI 应用范式的变化

在 AI 发展的早期，AI 应用主要集中在功能拟合上，即通过神经网络模型去模拟完成特定功能或任务，如分类、预测、优化等。这些应用通常基于规则或统计模型，目标是提高效率和准确性。随着大数据技术的发展，AI 应用开始转向数据驱动的模型，依赖大量数据来训练算法，实现更准确的功能拟合。

近年来，AI 大模型的快速发展不仅使得 AI 能够更高效地处理数据，而且还能够创造新的数据和内容。生成式 AI [13] 通过深入到输入数据的底层分布，理解数据和标签的联合概率分布，生成类似于训练数据的内容，从而允许创造性的数据合成和增强。AIGC 这一新兴范式，不仅拓宽了 AI 应用的深度与宽度，而且利用 AI 取代了手动内容生成，成为未来最具潜力的发展方向。

AIGC 能够生成的内容包括文本、语音、视频等。自然语言生成技术的进步，例如 OpenAI 推出的 ChatGPT，展示了 AI 在创造性文本生成方面的巨大潜力。ChatGPT 不仅能基于预训练阶段学到的模式和统计规律生成回答，还能完成撰写论文、邮件、脚本、文案、翻译、代码等多样化任务。同样，在图像和视频领域，深度学习技术如生成对抗网络和扩展模型等，也被用于生成逼真的图像和视频内容，包括动画、模拟场景和特效等。OpenAI 的 Sora 展示了 AI 在视频生成方面的能力，能够根据文本指令生成具有多个角色、特定类型运动及精确主题和背景细节的 60 秒复杂视频场景 [14]。

总的来说，AI 应用范式从简单的功能模拟，演进到能够理解、学习和创造的复杂智能系统。AIGC 作为 AI 发展的新篇章，不仅为 AI 领域带来了新的可能性，也为各行各业提供了创新的工具和解决方案，预示着一个充满想象力和创造力的新时代的到来。

3 AI 在 5.5G 的关键价值

3.1 提升网络性能

AI 在语言、音频、图像处理等诸多领域的技术突破，体现了数据驱动的方法在多模态信息特征提取与问题求解上的性能优势。而随着 AI 技术的飞速发展，无线网络与 AI 的融合逐渐加深，利用 AI 技术提升无线网络性能也成为通信领域的研究热点。学界以及业界的研究表明，利用 AI 技术可以从空口、核心网智能化等多个维度对无线网络进行增强。特别是无线智能空口技术的设计，是提升网络性能与标准化推进中的关键方向。

- 网络性能提升：传统通信模块设计为基于模型驱动的设计方法，即依赖于高斯假设、线性假设等对通信系统进行简化建模，进而对调制解调、编解码、信道测量等模块进行设计。AI 技术为基于数据驱动的设计方法，基于训练数据学习不同通信模块的输入输出映射关系，而未依赖于系统建模中的非实际假设。因此，结合数据驱动的 AI 技术实现对不同通信模块的性能增强，例如信道测量精度的提升、信号解调性能的提升等。最终，达成网络整体通信性能的增强，包括用户体验速率的提升以及通信覆盖能力的提升。
- 确定性业务保证：通信环境的时变特征导致通信链路性能的时变性，因此如何在变化的通信环境下保证业务的确定性传输是无线网络一直以来的重要研究方向。基于数据驱动的 AI 技术对通信环境的变化特征进行学习，并适应性调整通信策略，是实现确定性业务保证的重要技术方向。

基于 AI 的智能空口设计的演进，也促使了无线接入网（Radio Access Network，RAN）AI 标准的讨论。基于 AI 的空口设计在 3GPP R18 中进行了研究项目的立项，研究内容包括基于 AI 的设计对无线网络整体框架的影响，以及一些典型用例的性能及标准化影响 [15]。在项目开展过程中定义了 AI 相关基础概念、基础仿真验证方法论、基站终端合作方式，并针对模型/功能注册、数据传输、模型传输、模型/功能选择、模型/功能激活去激活等生命周期管理过程展开研究。

3.2 提供优质 AI 服务

在无线网络中，基于 AI 的部署不仅能够显著提升网络性能，而且无线网络本身作为一个具有通信连接、分布式算力部署、分布式数据处理和 AI 算法能力的系统，为优质 AI 服务的广泛构建提供了广阔的可能性。这些服务包括 AI 图像增强、游戏、扩展现实（Extended Reality，XR）、沉浸式通信等强体验型业务，它们对网络性能有着极高的要求。

随着 AI 领域大模型的设计与应用成为发展趋势，大模型规模的成倍增长也带来了 AI 服务落地的重要挑战。为了应对这一挑战，分布式训练和推理的技术架构应运而生，被认为是下一代 AI 架构的基本特征。基站部署的边缘 AI 算力，结合端网融合，为无线网络中构建分布式训练和推理能力提供了可能。这种架构的优势在于，通过将 AI 计算需求从中心云服务器卸载到更接近用户的网络设备，可以有效优化 AI 服务的延时和能耗，从而提升用户体验。

进一步地，无线网络的数据处理能力随着服务的扩展变得更加丰富，能够高效支持端到端的数据采集、传输、存储和共享。这与无线网络中的分布式训练和推理能力深度融合，支持更大规模、更高智能的模型优化、训练和推理。然而，如何将数据高效、安全地提供给网络内部功能或网络外部功能，依然是一个需要深入研究的课题。

展望未来，网络无处不在的边缘计算能力将为 AI 服务提供强大的支持。借助网络集成的通信、感知和计算的平台优势，我们可以开辟网络参与 AI 服务的新型市场空间，为 AI 时代的繁荣提供强劲动力。具体而言，RAN 在传输、协同、感知方面具有天然优势，基于“通算融合、感算融合”的技术路线，实现 AI 服务的低时延、高智能、高覆盖、低功耗的业务需求，满足未来 AI 业务的爆炸性增长的需求。

4 5.5G AI 的应用案例

4.1 场景用例分析

4.1.1 基于通用人工智能终端的 AIGC 应用

AIGC 的多模态处理能力可以大幅提升生产效率并降低原本繁杂的人力劳工成本 [16]。AIGC 作为当今最火爆的通用人工智能（Artificial General Intelligence，AGI）类应用之一，正在逐步改变我们创造和消费内容的方式。本节从 5.5G 网络的视角，选取电商直播、云游戏和视频通话三个场景来具体说明 AIGC 的能力和作用。

- **电商直播**：AIGC 可以用于生成吸引人的直播内容，包括自动生成产品描述、回答常见问题，甚至创建虚拟主播进行 24/7 的不间断直播。通过分析观众的互动和反馈，AIGC 技术能够实时调整直播内容，提高用户参与度和购买转化率。自动化内容生成不仅节省了商家的人力资源，还提供了更加个性化和多样化的购物体验 [17]。
- **云游戏**：AIGC 在云游戏服务中提供个性化的游戏推荐和动态生成的游戏内容。它可以根据玩家的游戏历史和偏好，生成定制化的游戏关卡或任务，例如定制化多样化的 NPC 设计，为用户带来独特的游戏体验，增加游戏的多样性和趣味性。AIGC 进一步使能 AI 对战，为追求更好竞技性的玩家提供游戏性乐趣 [18]。
- **视频通话**：AIGC 可以用于改善通话质量，提供实时的背景替换、语音增强和情感分析等功能。通过分析通话内容，AIGC 能够自动生成会议摘要、关键词标签或情感反馈，帮助用户更好地理解和回顾通话内容。这使得视频通话变得更加智能和高效，特别是在远程工作和在线协作的场景中。

AIGC 的核心优势在于其高度的自动化和智能化。通过深度学习模型，AIGC 能够分析大量数据，学习人类的创作习惯和风格，然后独立生成高质量的内容。无论是撰写新闻报道、编写故事情节、设计视觉艺术作品，还是制作音乐和视频，AIGC 都能够以其独特的视角和创意，提供令人耳目一新的作品。

此外，AIGC 的可定制性为用户提供了极大的灵活性，可以根据需求设定不同的参数和条件，指导 AIGC 生成特定主题、风格或情感的内容。这种定制化服务在广告、市场营销和娱乐产业中尤为受欢迎。创新性是 AIGC 的另一大亮点，它不受限于传统的思维模式和创作边界，能够探索未知领域，创造出前所未有的内容形式，为艺术创作、科学研究和教育领域带来新的视野，激发无限的想象力和创造力。

随着技术的不断进步，我们可以期待 AIGC 在未来将带来更多令人激动的应用和成果。例如，苹果在 2024 年 WWDC 全球开发者大会展示的 Apple Intelligence [19]，接入 GPT-4o 赋能 iOS，将 Siri 完全改造成“终极虚拟助手”，并准备将其开发为“最强大的杀手级 AI 应用程序”。通过 on-device 处理 + 私有云计算的端云协同实现个性化交互、个人化智能，这将成为 iPhone 16 及后续机型的标配。而 AGI 终端的 on-device 推理和 on-device 边缘训练能力也将消耗流量的主体从人类扩展到机器，这将会对通信管道提出更快且更可靠的要求。运营商 ToC 的下一个超级应用，或许就是 AIGC。

4.1.2 基于机器人的具身智能应用

随着 AI 技术的快速发展，智能机器人已经从科幻小说中走进现实生活，它们在工业、医疗、服务业、家庭等多个领域展现出巨大的潜力和价值 [20]。具身智能，作为 AI 领域的一个新趋势，指的是将智能系统嵌入具有实体形态的机器中，使它们能够直接与环境进行互动。这种智能的实现依赖于多学科的融合，包括机械工程、电子学、计算机科学和认知科学等，通过这些学科的交叉，智能系统能够学习如何适应环境，优化行为，并在复杂场景中做出决策。

GPT-4o 等先进的 AI 模型实现了文本、音频和图像的多模态交互，并具备了一定的情感解读能力。但具身智能的关键在于，智能系统能够与物理世界产生互动，这标志着 AI 从依赖人工提供 Prompt 的阶段，走向了更加自主和智能化的形态。这些智能载体可以是机器人、无人机、无人车或其他形式的自动化设备，它们通过集成的传感器网络来感知外界，并将感知数据转化为对环境的理解和响应。

在实际应用中，具身智能展现出了广泛的潜力。无论是在工业自动化中提高生产效率，还是在服务行业中提供更加个性化的客户体验，或是在探索未知领域中执行高风险任务，具身智能都扮演着越来越重要的角色。随着 AGI 产业的蓬勃发展和大模型技术的百花齐放，计算机视觉、计算机图形学、自然语言处理、认知科学等技术的成熟，具身智能正快速从理论走向实践，从实验室走向日常生活。

面向 5.5G+ 时代，ToC 的个人/家庭助理型具身智能机器人将应用在如取快递、帮忙家庭采购、辅助看护等一系列场景，其中典型应用场景包括：

- **陪伴机器人**：具身智能在家庭环境中的一个典型应用就是陪伴机器人。它们不仅能与人类进行对话，还能通过面部表情、语音语调、身体语言等识别用户的情绪状态，理解和回应人类的情感和需求。对于老年人，陪伴机器人可以监测健康状态，提醒用药或在紧急情况下提供帮助。对于儿童，它们可以提供教育内容，辅助学习，同时提供游戏、音乐、故事讲述等娱乐内容，成为儿童成长的良师益友。
- **快递运输机器人**：在物流和快递行业中扮演着越来越重要的角色。这些机器人利用传感器信息和地图数据进行导航，避开障碍物，规划最佳路径，提高配送效率，降低物流成本，为用户提供便捷快速的服务体验。

5.5G 网络的高速度、低延迟和高可靠性将赋能具身智能终端进行更精确的环境感知，并为室内外机器人应用提供高可靠的端到端确定性时延服务。这将为人类社会带来深远的影响，开启一个更加智能化和个性化的新时代。

4.2 关键需求与挑战

随着 AI 技术的不断进步，AIGC 和智能机器人的应用领域不断拓展，对无线网络提出了新的诉求和挑战，以支持更丰富、更复杂的内容生成和交互体验。

4.2.1 AIGC 应用的需求与挑战

AIGC 技术对 5.5G 网络传输具有低时延的严格要求，尤其是那些需要实时交互的应用场景，如在线游戏、虚拟现实、远程控制等。低时延能够确保内容的实时生成和交互，从而显著提升用户体验。以虚拟直播购物为例，从用户发出请求到内容呈现的整体延迟需控制在 70~100 ms 范围内。进一步细分，空口的单向时延需达到 5~10 ms 的量级。

AIGC 应用可能涉及大量数据的传输，包括高清图像、视频流和复杂的模型参数。因此，高带宽网络对于支持这些数据的快速传输至关重要，以满足 AIGC 对数据处理的需求。例如，1080p 视频的典型上传码率在 5~8 Mbps，而 AIGC 生成内容的下行速率可能达到百兆级别 [20, 21]。此外，无线网络还需要支持多模态数据的高效传输，包括文本、图像、音频和视频，并为 AIGC 应用提供差异化的服务质量保证，确保关键任务应用获得必要的网络资源。

4.2.2 具身智能的需求与挑战

智能机器人，得益于 AI 技术尤其是 LLM 的快速迭代，日益变得聪明，能够快速、可靠地在非结构化环境中执行任务。然而，非结构化环境的不确定性和任务的多样性也带来了一系列挑战。

首先，如果所有计算都在机器人本体进行，将需要较高的计算能力和相应的功耗。但轻量化设计是智能机器人在现实环境中工作的重要要求，限制了它们无法配备大量的CPU/GPU 单元和大容量电池。其次，机器人本体的感知信息相对受限，对于视野之外的目标物体，仅依靠自身的感官信息可能导致任务成功率降低或任务完成时间延长。

为解决这些问题，计算卸载至网络成为关键解决方案。机器人本体传感器收集的多模态数据，连同任务指令一起传输到网络，网络执行推理生成最终输出，如目标检测、路径规划等，再发送回机器人执行。此外，网络的优越感知能力可以提供综合的环境信息，辅助机器人进行任务执行的推理，如路径规划。

在这些应用场景中，端到端推理时延应控制在 200 ms 以内，推理准确率应不低于 90% [22]。同时，网络需要保证能够满足这些 AI 服务时延和准确率要求的智能机器人数量，每个小区至少支持 30 个智能机器人的稳定运行。

5 适用于 5.5G AI 的关键技术

5.1 提升网络性能的关键技术

以香农信息论为基础，无线空口传输技术自 20 世纪五六十年代开始经历了长久的发展，衍生出调制解调技术、导频与信道估计技术、信道测量技术、波形技术等各个细分领域，并成功在 2G 到 5G 蜂窝网商用通信系统中获得了广泛应用。

AI/ 机器学习（Machine Learning，ML）的发展也同样经历了长时间的积累和沉淀。1954 年图灵提出了著名的图灵测试，随后在 1956 年的达特茅斯会议上“AI”的概念被首次提出，之后 AI 经历了两轮技术发展周期。2006 年之后，以深度学习算法和大型数据集为新的切入点，第三次 AI 发展浪潮快速席卷了各个领域。

将 AI 技术与 5.5G 结合，利用 AI 技术提升网络性能，是一种连接通信理论和 AI 方法的跨领域技术。通过两个领域在数学模型、系统架构、算法设计的有效结合和扩展，为无线通信网络打造智能化内核，赋予更优的传输性能、更高的运行和维护效率、量身定制式的用户体验。本节将从 AI 星座设计、AI 灵活导频、AI 高精度信道测量和 AI 覆盖增强波形等方面介绍 AI 提升网络性能的关键技术。

5.1.1 AI 星座设计

星座调制是一种数字调制技术，是将信息 bit 承载到载波信号上，调制后的信号可以通过星座图形象地表示在二维平面上。现有无线通信系统中通常使用正交幅度调制（Quadrature Amplitude Modulation，QAM），即在 I 路和 Q 路两个正交载波上进行幅度调制的调制方式。QAM 调制根据星座图中星座点的数量可具体细分为 *N*-QAM，*N* 为 QAM 调制的阶数，即每个调制符号能够承载 $\log_2 N$ 个 bit 的信息。通常情况下，I/Q 两路正交载波上幅度调制的候选集合相同，因此 *N* 一般为 2 的偶数次幂，即 16QAM、64QAM、256QAM、1024QAM 等，如图 1 所示。

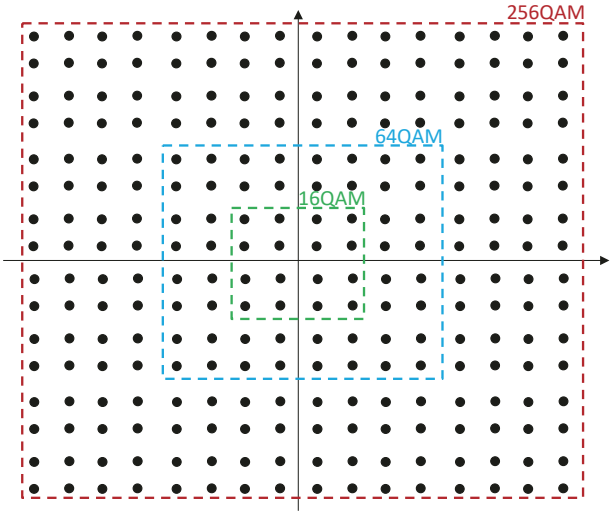


图 1 16QAM、64QAM、256QAM 星座图

QAM 是一种规则的调制技术，即每个星座点与它对应的信息 bit 之间的关系都可以通过同一个公式表示，例如 16QAM 的星座点与 4 bit 之间的对应关系可表示为：

$$s = \frac{1}{\sqrt{10}}((1-2b_1)(2-(1-2b_3)) + 1j*(1-2b_2)(2-(1-2b_4)))$$

这使得 QAM 调制和解调在实现上比较简单。但是，这种规则性也导致它在性能上并非最优的调制方式。在 AWGN 信道下，理论上星座图越接近高斯分布，性能越逼近香农信道容量。QAM 调制的星座图显然不是高斯分布，因此与香农信道容量存在一定差距，且随着阶数的增加，差距越来越大，如图 2 所示。

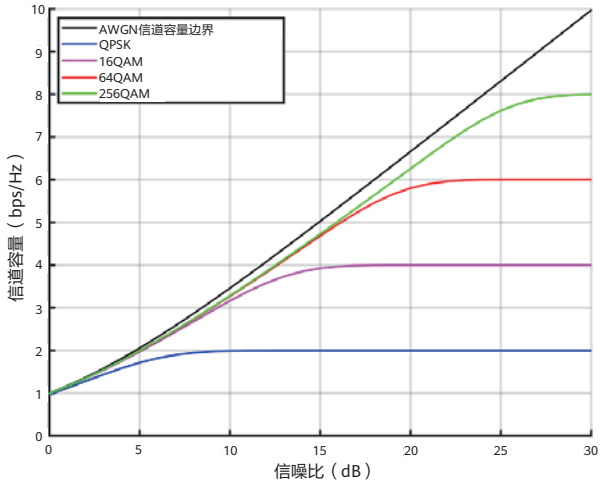


图 2 QAM 调制的信道容量

传统非规则星座调制可以分为几何成型和概率成型两种。几何成型是通过改变星座图中星座点的位置坐标，使其趋于服从高斯状分布；概率成型是不改变星座图几何形状（如仍然是 QAM 星座图），但改变发送信号星座点的出现概率，使其接近高斯分布。

几何成型和概率成型都是通过理论的方式改变星座图的分布，使其更接近高斯分布。但实际上，考虑到非理想的接收机，高斯分布也可能不是最优的星座图分布；考虑到不同的信道条件（如信噪比等），不同情况下最优的星座图分布可能也不相同，这就导致很难从理论上给出最优的星座设计。

AI 星座设计则可以根据具体的信道条件通过端到端训练的方式，找到最适配目标信道条件的最优星座设计。相比于 QAM 调制、几何成型和概率成型，AI 星座设计也更为灵活。如图 3 所示，在 AWGN 信道下，可以通过端到端的方式，联合训练星座图的几何成型、bit 映射和对应的解调器，使其性能逼近香农信道容量。在发送端，AI 模型输出星座图，包括该星座图中所有星座点（调制符号）的位置（I/Q 两路的取值），bit 映射通过星座点的顺序体现，例如对于 4 阶星座图，bit00 对应第 1 个星座点，bit01 对应第 2 个星座点，以此类推。在接收端，AI 模型输入叠加了噪声的调制符号，输出为每个 bit 的对数似然比（Log-Likelihood Ratio，LLR）。损失函数可以用二元交叉熵（Binary Cross Entropy，BCE）函数，使得 AI 解调器输出的 LLR 逼近发送的信息 bit。经过信道均衡模块后，信道和多用户干扰的影响已经被大部分移除，对于解调器来说，可以近似认为经历了 AWGN 信道，因此，在 fading 信道和多用户场景，都可以使用 AWGN 信道训练得到的星座。

图 4 所示为 AI 设计的两个具体的星座图示例，可以看出，AI 星座图是一种非规则的星座图。虽然从星座图分布上，AI 星座相比于 QAM 更接近高斯分布，但与传统几何成型相比，其星座图分布也更为灵活，更容易适应各种信道条件。

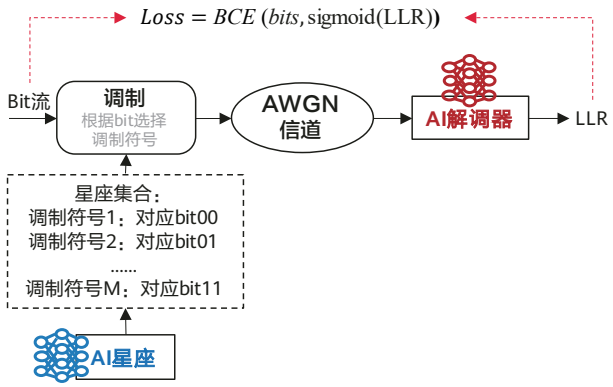


图 3 AI 星座设计训练示意图

5.1.2 AI 灵活导频

5G 系统中存在许多作用不同的参考信号，其中解调参考信号（Demodulation Reference Signal，DMRS）是用于数据解调的参考信号，即可以利用 DMRS 估计数据所占用的时频资源的信道响应。信道估计的精确度和 DMRS 的密度/开销之间存在一个折中关系。如果信道表现出较为严重的频率选择性（即信道在频域变化较大），则应增加频域内的 DMRS 密度。类似地，如果信道在时域变化较快，则需要在时域上占用更多资源来部署参考信号。在确定时频域的 DMRS 密度后，我们需要进一步考虑 DMRS 在时频资源块中的位置。例如，在信道平稳的条件下，为了减小插值误差和降低实现复杂度，可以采用在频率和时域上均匀地分配 DMRS 信号。由于 DMRS 本身不传输任何对用户有用的数据信号，因此需要以适当的密度分配 DMRS 以最大化吞吐量。

现有协议以及前述方案中 DMRS 与数据资源仍然是正交分配时频资源的，虽然保证了信道估计性能，但是需要预留较多的资源给 DMRS（尤其在移动性场景中），由此引发信道估计性能与数据可用资源之间的矛盾。进一步地，通过 AI 的手段还能实现更加灵活的 DMRS 设计，通过将 DMRS 与数据叠加的传输方式：

- 打开了更大的优化空间，释放了 DMRS 资源，使能极限资源利用率。
- 突破对正交 DMRS 的依赖，打破 DMRS 端口数限制。
- 引入 AI 来完成叠加传输模式下数据与 DMRS 间的功率分配以及接收机（利用 AI 融合多个收端模块，至少包括信道估计与均衡）的联合设计也能有效解决复杂度的问题。
- 在所有可用时频资源上覆盖 DMRS，通过功率分配形成非规则的 DMRS 图样也能提高系统对通信环境的适配性。

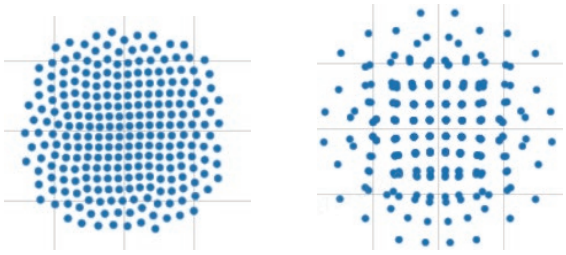


图 4 AI 星座设计示意图

具体地，每个资源粒子（Resource Element，RE）既承载调制数据符号又承载参考信号，并且仍然需要保证功率归一。图 5 给出水印导频传输模式下单个资源块中每个 RE 上导频与数据符号的形态，其中蓝色代表数据符号占用的功率比例，黄色代表导频符号占用的功率比例。收端通过相对应的 AI 接收机，可以对数据叠加导频进行信道估计和数据解调。

针对上行场景，AI 是部署在基站端；针对下行场景则是部署在 UE 端。功率分配因子是通过 AI 训练得到的，基站可以根据配对流数分配功率。承载的导频序列可以复用现有标准中的序列生成方式。

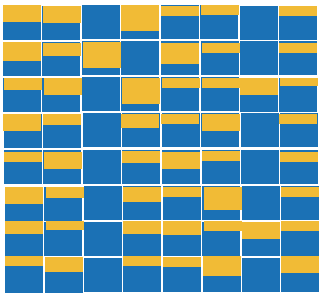


图 5 数据叠加导频模式下资源块中数据与导频符号的传输形式

5.1.3 面向时分双工的高精度测量

为实现无线网络的覆盖提升与性能提升，天线规模的扩大不可避免。但是，天线规模的增加导致高精度信道测量所需的空口资源占用显著提升。如何在有限的空口占用下，实现大规模 MIMO 系统的高精度信道测量成为 5.5G 时分双工（Time Division Duplex，TDD）系统的关键挑战。

实现有限空口占用下的高精度信道测量的关键点在于，如何有效利用基于不同参考信号获取的有限信道信息，恢复全维度信道信息。例如，UE 根据下行参考信号获取下行信道信息，基站根据上行参考信号获取上行信道信息，但是来源于不同参考信号的信道信息其实都源于相同的通信环境，因此可视为对相同通信环境的不同视角或不同模态的测量。

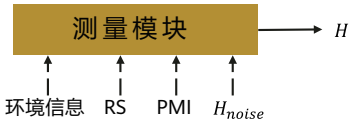


图 6 多模态信息融合的 AI 技术使能通信环境和通信信道的高精度恢复

利用多模态信息融合的 AI 技术，可实现对来源于不同参考信号获取的信道信息的有效融合，从而实现对通信环境和通信信道的高精度恢复。

5.1.4 AI 覆盖增强波形

目前 NR 系统上行有两种波形，分别是多载波正交频分复用（Orthogonal Frequency Division Multiplexing，OFDM）波形和单载波 OFDM 波形，其中单载波 OFDM 波形通过在传统的 OFDM 处理过程之前增加一个额外的离散傅立叶变换（Discrete Fourier Transform，DFT）处理，将其转变为单载波，因此也成为离散傅里叶–扩频–正交频分复用（DFT-Spread OFDM，DFT-s-OFDM）波形。由于多载波 OFDM 波形的发送信号是由多个子载波的信号叠加而成，有些时刻多个子载波的信号同向叠加形成很高的峰值，因此单载波 OFDM 波形的峰均功率比（Peak to Average Power Ratio，PAPR）较低，即有更好的覆盖性能。但对于极致深度覆盖的用户而言（比如小区边缘用户或经过多次建筑物穿损的用户），DFT-s-OFDM 波形的 PAPR 仍然较高，因此需要针对极致深度覆盖的用户设计比 DFT-s-OFDM 波形覆盖性能更好的新波形。

传统降低 OFDM 波形 PAPR 的技术主要包括滤波器技术和削波技术等。滤波器技术是通过设计适当的频域滤波器，改变时域波形，降低 PAPR；削波技术是通过将超出门限的信号按比例缩小，从而降低 PAPR。这些技术均能够对 OFDM 波形的 PAPR 进行一定程度的优化，但代价是会损失一些数据传输吞吐率。

AI 覆盖波形设计是通过 AI 训练的方式，优化波形的覆盖性能。AI 覆盖波形设计相比于传统波形降 PAPR 技术，主要的优势在于可以通过多目标联合优化，实现覆盖和吞吐的折中，在提升覆盖同时减少吞吐的损失。

例如，可以通过 AI 实现子载波预留技术。在实际传输带宽外，预留若干子载波，通过 AI 训练优化该子载波上的信号，改变波形形状，降低 PAPR，如图 7 所示。AI 模型的输入为原始数据符号的时域信号，输出为预留子载波上的信号。为了避免预留子载波上功率的浪费，在训练时可以约束子载波的功率。

图 7 AI 子载波预留设计

5.2 提供优质 AI 服务的关键技术

5.2.1 分布式推理

随着大模型的参数规模逐渐增大，训练和推理需求的硬件水平也逐渐提升，AI 模型不断地被推至边缘。基站作为最靠近终端的生态位，且能实时感知信道变化，可充分利用通算资源的深度耦合优化 AI 业务推理性能。

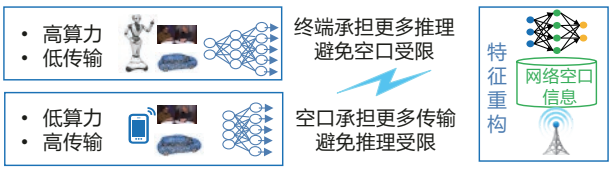


图 8 基于通算联合的空口资源动态调度

针对边缘 AI 分布式推理优化，目前主要有四类技术：

- **稀疏量化、结构冻结类**：主要通过减少计算和内存使用来实现推理优化。例如，量化将模型参数从高精度浮点数（如 32 bit）压缩到低精度整数（如 8 bit）或定点类型，从而减少模型大小和计算量。冻结即将模型中不需要更新的部分（例如预训练的层）冻结，使其不再参与训练和推理，从而减少模型大小和计算量。稀疏化是指通过去除模型中的冗余连接或参数，降低模型的复杂度，从而减少计算量和存储空间。常见的稀疏化方法包括权重剪枝（Weight Pruning）和结构化稀疏化（Structured Sparsification）。
- **流水线串行类**：主要将模型的不同层分配到不同的边缘设备上，通过串行执行的方式进行推理。例如，将图像预处理层分配到低功耗设备，将卷积层分配到更高性能的设备，最终将分类层分配到基站/云端。这种方法可以利用不同设备的优势，提高推理效率。
- **张量并行类**：即将模型的计算任务分配到多个边缘设备上，通过并行执行的方式进行推理。例如，将一个大规模的矩阵乘法操作分解成多个子操作，分别分配到不同的设备上执行，然后将结果进行整合。这种方法可以充分利用多核处理器的并行计算能力，加速模型推理。
- **批处理类**：将多个推理请求进行合并，一次性进行处理，从而提高推理效率。例如，将多个用户的图像识别请求合并成一个批次，一起进行推理，然后将结果分别返回给用户。这种方法可以有效地降低推理延迟，提高资源利用率。

从空口的角度分析，如何基于通信资源和计算资源的动态变化，从而实现动态稀疏化、动态流水线串行、动态张量并行、动态批处理，是当前的关键挑战。基于基站通算融合的独特优势，设计端网动态算力协同方案，开辟未来网络参与 AI 计算市场的新空间。

5.2.2 检索增强

在快速发展的无线通信世界中，终端设备和接入网络设备常常受限于它们的处理器、存储和计算能力。这些限制成为了大规模模型训练的绊脚石。然而，分布式训练技术的出现，为这一难题提供了优雅的解决方案。通过这种技术，我们能够跨多个设备共享计算任务，不仅减轻了单个设备的负担，还使得每个设备能够利用其独特的数据信息。

面向无线通信系统的分布式训练，涵盖了几项关键技术，它们共同构成了这一领域的技术支柱：

- **模型并行方案：**这是解决大型模型训练难题的一种有效手段。面对单一设备无法承载的庞大模型参数，模型并行通过将模型拆分为多个部分，并将这些部分分布到不同的计算设备上，从而降低了对单个设备内存的需求，并显著提升了计算效率。在这一过程中，模型分割是核心步骤，它可以根据模型的层级结构或者结合终端、接入网设备以及云端服务器的功能特性来进行。
- **分布式知识库的构建：**这是另一项关键技术。考虑到不同终端和接入网设备所能获取的信息存在差异，构建一个分布式知识库变得尤为重要。这一过程包括数据的收集与预处理，以及知识表示的设计。例如，基站具备全局感知能力，它需要从多样化的环境中收集数据，并对这些数据进行统一的预处理，以确保数据格式和质量的一致性。随后，我们需要确定如何在知识库中表示这些知识，可能涉及特征表示、模型参数或中间结果的表达方式。

如图 9 示例，基站负责收集环境的点云数据，并构建起一个点云知识库。这一知识库随后被共享至 LLM，以供进一步分析和使用。机器人在执行任务时，会实时收集环境信息，并将这些信息与导航请求一同发送至 LLM。LLM 结合基站提供的点云数据和机器人采集的视频信息，进行综合分析，以制定出精确的导航路径。最终，LLM 将制定好的导航规划结果发送回机器人，指导其完成导航任务。

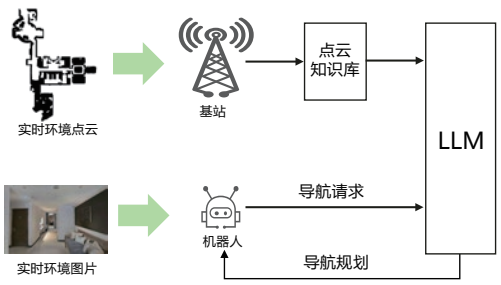


图 9 分布式知识库构建示意图

针对室内场景中的机器人导航规划任务，基站所提供的全局环境点云数据的比例是影响任务成功率的一个关键因素。如图 10 可以清晰地看到，随着基站所提供的点云数据量比例增加，机器人在室内执行导航规划任务的成功率也随之提升。这一现象表明，详尽的环境数据对于提高机器人导航的准确性和可靠性至关重要。具体而言，点云数据的丰富性直接影响大模型对环境的认知深度，进而优化其路径规划算法，使得机器人能够更加高效地完成导航任务。

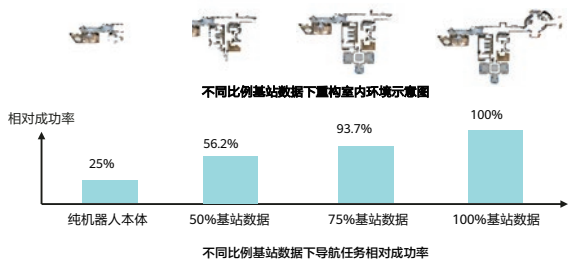


图 10 基站点云辅助机器人导航规划成功率

- **高精度小模型设计：**它是提升整体系统性能的重要技术。结合终端基站和云端服务器的特性，设计多个小模型，并通过高质量的标注数据进行训练。在训练过程中，可以采用迁移学习等先进技术，以提高模型的学习效率。除此之外，还有小模型猜测和大模型校验技术，小模型的猜测结果需要被转换成适合大模型输入的格式，例如向量或编码形式。最后，大模型将根据小模型的结果进行进一步的校验和调整，以确保最终输出的准确性。

通过这些技术的综合应用，无线通信系统中的分布式训练不仅能够克服设备能力的局限，还能够充分利用每个设备的优势，实现更加高效和精准的模型训练。

5.2.3 特征流通信

特征流通信是一种新的通信范式，这种技术的核心在于，它不仅仅关注数据的比特流，而是进一步分析和理解这些数据所代表的实际意义。特征流通信深入理解传输数据的语义内容，有助于接收端识别最有价值的信息，即对于获取发送端意图而言最重要的信息，以此来提高通信效率和准确性。

语义感知的空口通信技术将语义信息的理解和处理融入通信过程，根据语义特征来优化传输策略，具体的技术可以体现在两个方面：（1）特征流的空口差异化传输；（2）特征流容错的空口通信。下文分别介绍这两个技术方向。

- 特征流的空口差异化传输

  在传统的通信系统中，数据包的传输和接收通常是基于信号强度和错误率等技术指标。然而，特征流通信通过引入语义层面的理解，能够更加智能地处理数据。例如，若某个数据包包含较为重要的信息，在进行空口传输时，可以提高该数据包的传输优先级，确保其快速且可靠地到达目的地。

  此外，特征流通信还能够根据不同特征流对接收端语义恢复的贡献度来动态调整编码和调制策略。

  以图 11 所示的一种语义引导的特征流空口差异化传输为例，信源侧的数据通过语义转化切分成多个特征流，这些特征流在传输前根据其不同语义重要性的进行重要性标签映射。在信道编码过程中，根据特征流的重要性采取不同的信道编码策略。例如，针对能够对接收端语义恢复过程贡献度较大的特征流，分配更多的传输资源或采取更可靠的调制编码策略等，以此保证整个语义引导的特征流空口差异化传输系统可以根据语义的重要性来选择最合适的传输方式，从而在保证通信质量的同时，也优化了频谱资源的使用。

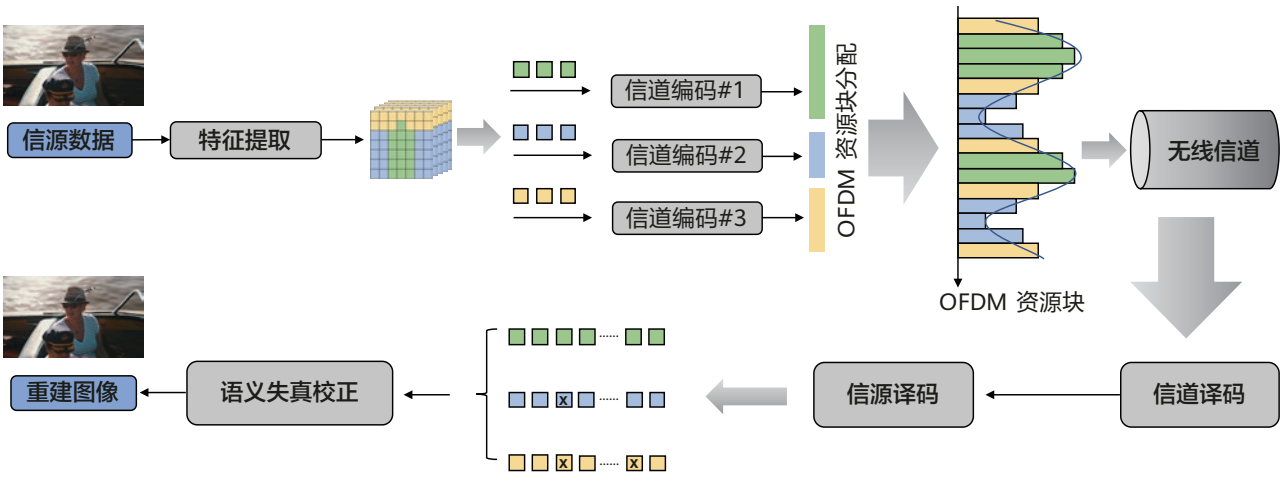


图 11 语义引导的特征流空口差异化传输示意图

传统通信中部分比特出错恢复画面　　特征流通信中部分比特出错恢复画面

图 12 特征流容错的空口通信示意图

- 特征流容错的空口通信

  特征流容错的空口通信可以抵抗无线信道不稳定性和可能发生的错误。这种通信方式可以在部分特征信息受损或丢失的情况下，恢复出原始信息的语义内容。

  如图 12 所示 [23]，对于传统通信方法主要关注于比特流的正确传输，若部分比特传输出错，在接收端进行画面恢复时会无法恢复发送端所传输的意图或信息，则在图 12 显示为对应出错的比特呈现马赛克的状态，整个画面则出现花屏。

  但对于利用了特征流容错的空口通信，即使在传输过程中比特传输出现错误，特征流通信在接收端仍然能够恢复出或推断出发送端的意图或信息，则在图 12 中显示为即便出现误码，画面出错区域仍可较好恢复 [24]。因此，利用特征流的容错性，可以提高空口通信的容忍度和恢复能力。

6 总结与展望

本文深入探讨了 5.5G 时代 AI 的演进趋势、关键技术及其在多个领域的应用前景。从 AI 模型的发展历史，到算力和数据的飞速进步，再到 AI 在 5.5G 时代的关键价值和应用案例，本文全面展示了 5.5G AI 技术面临的机遇与挑战。特别是 AI 在提升网络性能、提供优质 AI 服务以及具身智能和 AIGC 应用中的关键作用，突显了 AI 技术在未来通信和网络发展中的核心地位。

展望未来，5.5G 与 AI 的深度融合预示着一个智能化新时代的到来。随着技术的不断进步，AI 将在更多领域展现其潜力，特别是在自然语言处理、图像和视频生成以及多模态交互等方面。AI 技术的进一步发展将带来更加个性化和智能化的服务，改善用户体验，并为各行各业提供创新解决方案。同时，AI 的算力和数据需求将持续推动硬件和网络基础设施的升级，为实现更高效的分布式训练和推理提供支持。此外，AI 在提升无线网络性能方面的应用，如 AI 星座设计、灵活导频、高精度信道测量和覆盖增强波形等关键技术，将进一步优化网络资源的使用，提高通信效率。

尽管 AI 技术前景广阔，但同时也面临着不少挑战。例如：

- **网络能力的新突破**：5.5G 技术的持续演进将推动网络速度、连接密度和时延性能的进一步提升。
- **网络运维的新模式**：智能化和自动化工具将改变网络的运维方式，提高效率和准确性。
- **智能化服务的普及**：AI 技术将更广泛地应用于各行各业，极大提升生产效率和用户体验。
- **技术创新与融合**：边缘计算、网络切片等新技术将与 AI 深度融合，为特定应用场景提供定制化服务。
- **数据安全与隐私保护**：面对 AI 处理数据量的增加，确保数据安全和用户隐私的技术和法规将得到加强。
- **AI 的可解释性**：提高 AI 决策的透明度和可解释性，确保系统的公正性和安全性。

总之，5.5G 时代的 AI 正站在一个新的起点上，它的发展将深刻影响未来的社会结构和人类生活。

参考文献

[1] McCulloch W S and Pitts W, "A logical calculus of the ideas immanent in nervous activity[J]," The bulletin of mathematical biophysics, 1943, 5: 115–133.

[2] Rosenblatt F, "Principles of neurodynamics: Perceptrons and the theory of brain mechanisms[R]," Cornell Aeronautical Lab Inc Buffalo NY, 1961.

[3] Schmidhuber J and Hochreiter S, "Long short-term memory[J]," Neural Comput, 1997, 9(8): 1735–1780.

[4] LeCun Y, Bengio Y, and Hinton G, "Deep learning[J]," nature, 2015, 521(7553): 436–444.

[5] Krizhevsky A, Sutskever I, and Hinton G E, "ImageNet classification with deep convolutional neural networks[J]," Advances in neural information processing systems, 2012, 25.

[6] Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition[J]," arXiv preprint arXiv:1409.1556, 2014.

[7] He K, Zhang X, Ren S, *et al.*, "Deep residual learning for image recognition[C]," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.

[8] Vaswani A, Shazeer N, Parmar N, *et al.*, "Attention is all you need[J]," Advances in neural information processing systems, 2017, 30.

[9] Wu T, He S, Liu J, *et al.*, "A brief overview of ChatGPT: The history, status quo and potential future development[J]," IEEE/CAA Journal of Automatica Sinica, 2023, 10(5): 1122–1136.

[10] Silver D, Huang A, Maddison C J, *et al.*, "Mastering the game of Go with deep neural networks and tree search[J]," nature, 2016, 529(7587): 484–489.

[11] Jumper J, Evans R, Pritzel A, *et al.*, "Highly accurate protein structure prediction with AlphaFold[J]," nature, 2021, 596(7873): 583–589.

[12] Liu H X and Feng S, "Curse of rarity for autonomous vehicles[J]," nature communications, 2024, 15(1): 4808.

[13] Ho J, Jain A and Abbeel P, "Denoising diffusion probabilistic models[J]," Advances in neural information processing systems, 2020, 33: 6840–6851.

[14] https://openai.com/index/sora/

[15] TR 38.843 v18.0.0, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," (Release 18), December 2023.

[16] 中国 AIGC 应用全景报告 , 量子位智库，2024.03.

[17] 易观分析 . 2023 年中国直播电商发展洞察 [R]. 2023. https://www.analysys.cn/article/detail/20020949.

[18] 中国信息通信研究院 . 全球云游戏产业深度观察及趋势研判研究报告 [R]. 北京：中国信息通信研究院，2023.

[19] WWDC 2024 — June 10 Apple, https://www.youtube.com/watch?v=RXeOiIDNNek.

[20] CES2023 最新机器人【技术大盘点】今年的消费电子展有这些不一样 , https://www.youtube.com/watch?v=_UflhU8pIJU.

[21] Sim, D. You, Y. Lee, *et al.*, "User-selectable stereoscopic video streaming using video enhancement information: System design and implementation."

[22] Tanya Stivers, "Universals and cultural variation in turn-taking in conversation."

[23] http://toflow.csail.mit.edu/

[24] Dai J, Zhang P, Niu K, *et al.*, "Communication beyond transmitting bits: Semantics-guided source and channel coding[J]," IEEE Wireless Communications, 2022.

预见 6G，通信 + AI

唐雄燕，王友祥，隋腾飞
中国联通研究院

摘要

从移动互联网的发展历程来看，3G/4G 时代主要面向连接。2016 年 3 月，AlphaGo 的诞生开启了人工智能（Artificial Intelligence，AI）赋能网络（“AI for Net”）的新开端。5G 核心网引入了智能化网元 NWDAF（Network Data Analytics Function，网络数据分析功能），标志着网络向自动化、智能化的道路迈进。同时，5G 引入 SDN/NFV、边缘计算等技术，促进了网络使能 AI 的进程，即“Net for AI”，加速了 AI 产业的发展和应用落地。

面向 6G 时代，AI 和网络将深度融合，AI 赋能网络（“AI for Net”）将由外挂式 AI 转变为内生 AI，通过统一的内生 AI 架构，高效地支撑“Net for AI”。6G 将超越连接，向连接、计算和智能融合一体的方向发展，成为智能泛在的综合型数字信息基础设施，实现普惠智能。

关键词

6G，内生 AI，大模型

1 5G 与 AI 融合发展，带来乘法效应

5G 具有大带宽、低时延、大连接等特点，为智能化应用的发展提供坚实的底层基础。这也引发了 5G 与人工智能（Artificial Intelligence，AI）结合的广泛研究和探讨，开启了 5G 网络智能化的快速发展之路。

3GPP 在 R15 版本中引入了新的核心网网元 NWDAF（Network Data Analytics Function，网络数据分析功能），主要用于网络数据采集和分析，并提供网络切片负载信息。随着标准化进程不断推进，NWDAF 的功能不断扩充、增强，最终实现对集中式智能网络架构、服务化接口以及分层可信的智能化网络架构的全方位支持。R18 的研究涵盖了 AI 在网络中应用的两项重要工作。首先是引入了联邦学习，将模型训练和推理阶段的数据交由 NWDAF 分析和处理。其次，提出了 5G 网络支持基于AI/机器学习（Machine Learning，ML）的服务，包含无线接入网和核心网的跨域协同、模型识别/管理一致性、AI/ML 应用训练以及相关的 QoS 增强等。在 R19 阶段，5G 智能化相关标准工作将进一步推进，支撑网络持续优化，进一步提升资源利用率、网络能效和客户体验。

3GPP 在 R17 阶段开启了针对无线 AI/ML 的研究工作，主要涉及网络节能、负荷均衡、移动性优化等功能。与此同时，也定义了基本的运行结构，包括数据采集、模型训练、模型推理和执行。在 R18 阶段研究了信道状态信息（Channel State Information，CSI）反馈增强、毫米波波束管理以及定位精度增强，跨入系统和终端协作的新阶段。R19 将研究新的 AI/ML 用例，包括对分布式学习的支持等。

AI 可以提升网络的智能化程度，改善网络性能和用户体验，全面提升网络生产力。目前，AI 已经广泛应用于 5G 网络的规、建、维、优、营各个方面。例如在网络规划与建设方面，AI 技术可以实现网络的拓扑优化、资源/容量规划、智能验收以及新站点选择，提升网络服务质量与用户体验；在网络运维、优化与节能方面，AI 技术能够实现网络的自我优化，包括故障识别、预测、智能诊断和自愈等，从而提高网络的稳定性和可靠性；在网络运营和客户服务方面，AI 技术能够基于用户行为和偏好，提供个性化的业务推荐，提升客户满意度。

5G 作为重要的数字基础设施，凭借高速率、低时延等特性，以及强大的算力和数据归集能力，为 AI 应用提供了强大的支撑和赋能，促进 AI 技术与各行业深度融合，推动行业的全面数字化、智能化转型。5G 通过分布式的边缘计算，实现云边端算力协同，支持 AI 算法在云边端设备之间更高效地运行数据，将 AI 推理任务从云端下沉到了边缘和端侧。创新的各类 AI 应用和服务涵盖了各个领域，从智能家居到智能城市，再到智慧医疗和智能交通系统等，推动了智能经济的快速衍生和发展。

2 6G 与 AI 和谐共生，双向赋能

2023 年 6 月，国际电信联盟 ITU-R WP 5D 完成了《IMT 面向 2030 及未来发展的框架和总体目标建议书》（以下简称《建议书》），针对 6G 提出了六大典型使用场景，分别是：沉浸式通信、超大规模连接、超高可靠低时延通信、AI 和通信、通信感知一体化以及泛在连接。其中，前三个场景是对 5G 三大场景的演进与增强。AI 和通信则是新增的场景，AI 的引入使得网络更加智能化，同时也为模型、算法和数据分析提供更强的支撑。

《建议书》中同时提出了 6G 系统的设计原则，包括但不限于：可持续性、安全/弹性、连接未连接的用户、提高整体系统性能的泛在智能。未来，6G 网络将在传统连接能力的基础上，向连接、计算和智能融合一体的方向发展，成为新一代数智服务使能平台。

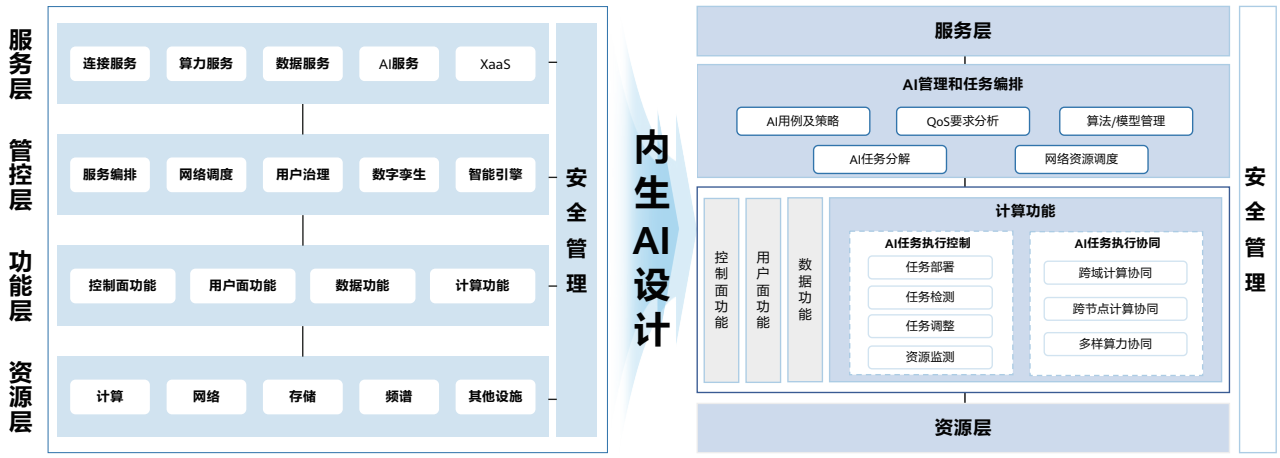


图 1 中国联通 6G 网络架构内生 AI 设计

2.1 内生 AI 的 6G 网络架构

基于面向“超越连接”、面向“多元用户”和面向“平台即服务”的设计理念，继承了 CUBENet3.0 的内涵理念，中国联通系统性提出了层次化的 6G 网络体系架构，目标是成为新一代数智服务使能平台。在架构设计中，将内生 AI 的理念自上而下贯穿于全要素中，由管控层进行 AI 管理和任务编排，控制面功能协同控制，用户面功能、数据功能和计算功能一起完成 AI 任务的部署、调度和执行，共同构成内生 AI 的网络架构。

图 1 展示了中国联通 6G 网络架构内生 AI 设计，自下而上分别为资源层、功能层、管控层、服务层，四层统一进行安全管理 [1]。其中：

- 资源层包括计算、存储、网络和频谱等各种基础设施，提供功能部署、运行和服务支撑，是整个网络运行的基础。
- 功能层根据管控层下发的指令，由控制面功能进行协同控制，联合用户面功能、数据功能和计算功能，共同完成 AI 任务的部署、调度和执行。其中，用户面功能负责执行接入锚点、业务数据转发、网络数据转发功能；数据功能包括数据的采集、预处理、存储、共享等；计算功能分为 AI 任务执行控制和 AI 任务执行协同两个模块，对计算节点等要素进行协同调度从而执行业务部署、调整等 AI 任务。
- 管控层负责对 AI 的管理和任务编排，支撑服务层提供 AI 用例及策略、QoS 要求分析、算法/模型管理、AI 任务分解以及网络资源调度等功能。
- 服务层直接面向用户，提供连接、算力、数据以及 AI 服务等。

内生 AI 的设计需要满足四个基本条件 [2]：（1）采用标准化的数据格式，促进数据的互操作性，简化 AI 模型的训练与集成；（2）定义 AI 服务质量评价体系 QoAIS（Quality of Artificial Intelligence Service，人工智能服务质量），在传统的网络服务质量基础上扩展如 AI 服务响应时间、可靠性和准确性等指标；（3）融合 AI 的计算能力与通信网络的传输能力，实现数据处理的最优化；（4）要求 AI 可信，确保 AI 系统决策透明、公正，提供用户隐私保护，过程符合伦理和合规性要求。

未来，AI 能力将会以人工智能即服务（AI as a Service，AIaaS）的形式进行网络内部功能调用和第三方能力开放，通过 6G 网络（包括核心网、无线接入网和终端等）内的连接、计算、数据、模型等资源和功能，构建分布式的高效、节能、安全的 AI 服务和开放生态 [3]。

2.2 大模型为通信网络带来新机遇

自 2017 年谷歌发布 Transformer 模型到 2024 年文本视频模型 Sora 的发布，大模型发展势头迅猛，经历了从感知理解世界到生成创造世界、从单模态到多模态的跃迁，完成了从专用到通用的发展历程，并扩展到不同的领域和应用场景。AI 正逐步进入通用智能时代，大模型的出现将为 6G 与 AI 的融合带来新的机遇。

得益于 6G 内生 AI 的设计，每个网元将原生集成通信、计算和存储等能力，形成一个“云边端”多级算力协同的融合架构。通过强大的网络连接，实现对更广泛的算力资源的管控和动态按需调度。面向未来大模型的部署和应用，采用云边端协同的方式，实现大模型从训练到推理的全过程。其中，核心云负责模型训练，满足模型训练对存储和算力的要求；网络负责数据和模型的传输；边和端侧则利用训练好的模型完成推理过程，从而实现更快的响应时间、更高的可靠性和更好的资源利用。

一方面，AI 大模型与 6G 融合将使得网络在数据处理、分析、优化等方面达到前所未有的高度，提升网络效能，实现网络智能运营。另一方面，通过 6G 技术的加持，在网络内开展 AI 大模型的训练和推理，进而提供适配不同应用场景的 AI 能力，为用户提供高性能的 AI 服务。相较于传统小模型，大模型通过内容生成训练实现模型能力泛化、任务多样化，具备极强的通用能力。但与之相对的是，大模型的专业能力较弱。因此，利用 AI Agent 逐步形成“大模型 + 小模型”的协同模式，也是未来 AI 产品的重要趋势之一。图 2 以网络优化场景为例，展示了大小模型协同模式。这种模式可以更好地满足不同应用场景的差异化需求，提高 AI 产品的性能和效率。

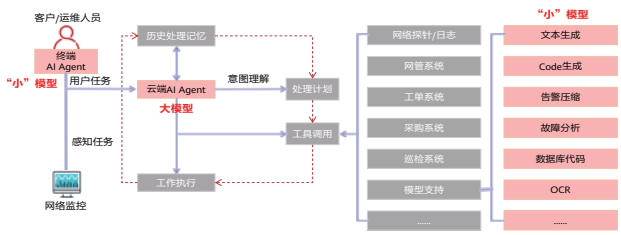


图 2 以网络优化场景为例的 AI Agent“大模型 + 小模型”协同模式

3 产业共同推动通信与 AI 融合发展

AI 为 5G 演进、6G 发展注入了新活力，逐渐成为通信技术迭代、系统和设备创新的关键驱动力。随着 6G 标准即将启动，中国联通将持续深入研究 6G 与 AI 融合的关键技术，加强与业界合作，共同推进 6G 网络智能化，助力数字经济发展。

参考文献

[1] 中国联通. 中国联通 6G 网络体系架构白皮书 [R]. 2023.06.

[2] 6GANA. 6G 网络原生 AI 技术需求白皮书 [R]. 2021.01.

[3] IMT-2030（6G）推进组. 6G 网络架构愿景与关键技术展望 [R]. 2021.09.

计算即服务：探索移动终端的无限可能

袁雁南，吴琦，康艳超，刘健康，孙晓文，姜大洁，秦飞
维沃移动通信有限公司

摘要

“AI 和通信”使用场景为 6G 拓展了计算这一新服务。6G 的计算即服务将有助于催生新形态的终端和优化用户体验，为移动终端发展提供更多可能性。6G 系统需基于典型用例（如虚拟人等）确定计算服务的性能指标，如算力密度、计算连接密度、用户峰值算力、用户计算时延等。为了支持“AI 和通信”场景，本文提出了 6G 潜在的移动算网融合技术方向包括基于演进的路线和基于变革的路线。基于演进的路线包括增强边缘计算（Edge Computing，EC）和增强 IP 多媒体子系统（IP Multimedia Subsystem，IMS），基于变革的路线则是在移动网络中新增计算控制功能、计算节点和计算数据传输通道。本文还介绍了 vivo 移动算网融合样机。通过样机测试验证当算力和传输带宽高于阈值时，6G 网络提供的 AI 和通信服务相比于用户设备（User Equipment, UE）本地计算具有更高的精度和更低的时延。

关键词

6G，AI 和通信，增强 EC，增强 IMS，计算控制功能，移动算网融合样机

1 引言

国际电信联盟无线电通信部门（International Telecommunication Union – Radio communication Sector，ITU-R）在《IMT 面向 2030 及未来发展的框架和总体目标建议书》[1] 提出了沉浸式通信、超大规模连接、超高可靠低时延通信、AI 和通信、通信感知一体化、泛在连接六大场景。“AI 和通信”场景在传统高用户体验速率和低时延等通信功能的基础上，还需要将人工智能 (Artificial Intelligence，AI) 和计算相关的功能集成到 6G 系统。从而使能 6G 系统支持不同节点间的模型共享和模型推理，以及多个数据源的数据采集和分布式 AI 模型训练等。因此，AI 和通信应用场景为 6G 拓展了计算这一新服务。

目前云计算是计算服务（包含 AI、渲染等）的主要形式，并且云计算市场规模持续增长 [2]。那么，未来 6G 的 AI 和通信场景有可能在哪些方面发挥优势和提供差异化服务呢？我们认为，面向自由连接的物理与数字融合世界的 6G 愿景，终端将成为连接物理世界与数字世界的桥梁。因此，6G 的 AI 和通信场景应该围绕与移动终端相关的应用场景，在现有移动通信基础上扩展计算。当移动终端相关的 AI 和通信场景具备如下特征时，6G 可能在性能、效率或安全方面提供相比目前云计算更优的差异化服务。

- AI 和通信场景所需的计算能力要求超过大多数移动终端的计算能力，并且具有低时延要求。对应的业务例如交互式高保真 3D 虚拟人等。
- AI 和通信场景所需的部分计算数据由 6G 网络提供。对应的业务例如数据集包含部分网络数据的 AI 模型训练等。
- AI 和通信场景由于终端移动性等导致信道状态和通信性能的波动，需要实时调度通信资源和计算资源才能满足需求。6G 系统基于实时信道状态信息和计算资源状态信息可以更好地实现计算和通信的协同优化调度。
- 基于 IP 多媒体子系统（IP Multimedia Subsystem，IMS）提供运营商原生的 AI 和通信服务。类似现有运营商提供的话音或短消息服务，移动终端可以在不需要安装应用的情况下获得 AI 和通信服务。
- 需要移动终端和网络节点之间通过相互协作完成复杂任务的 AI 和通信场景。例如分布式 AI 代理（agent）等。

面向 6G 的 AI 和通信场景，潜在的服务流包括移动终端的按需计算卸载和移动终端的按需在网计算。如图 1 所示，绿色虚线表示的按需计算卸载是指移动终端根据本地算力、计算任务需求、功耗等确定是否将计算卸载至网络来处理。计算卸载的计算数据流的发送节点和接收节点相同。如果 6G 终端需要计算卸载，那么发送待计算数据给网络，并获得网络的计算结果。蓝色虚线表示的按需在网计算是指在网络传输过程中按需对传输的数据提供计算服务。进一步，应用功能可以由 6G IMS 应用服务器提供也可以由 OTT（Over-the-Top）应用服务器提供。根据计算数据流的发送节点和接收节点，在网计算可以分为：

- 6G 终端发送数据，经网络传输和计算处理后，应用功能接收数据。
- 应用功能发送数据，经网络传输和计算处理后，6G 终端接收数据。
- 6G 终端发送数据，经网络传输和计算处理后，另一个 6G 终端接收数据。

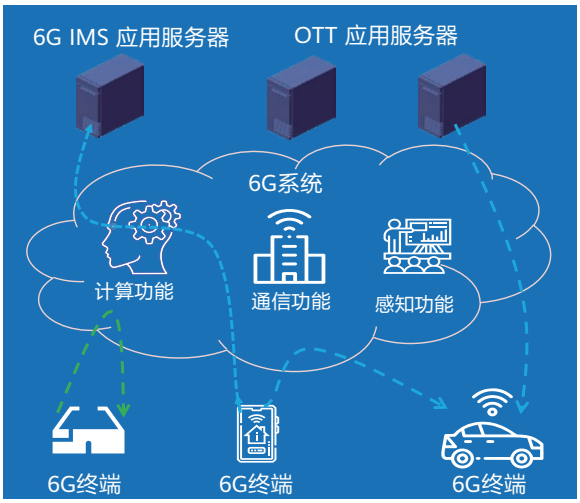


图 1 与移动终端相关的计算服务流

2 计算服务的性能指标

6G 系统提供的 A 计算服务包含 AI、渲染等，其中 AI 服务的性能指标包括可达性能（如归一化均方误差、余弦相似度等）、AI 模型复杂度、收敛速度（或训练时间）、泛化能力、数据依赖性、推理时间、训练的算力开销、模型的传输开销和模型的存储开销等 [3]。AI 服务的性能指标取决于 AI 算法和大数据技术等计算机领域相关技术在 2030 年及以后的发展水平。

本文聚焦在由 6G 系统中部署的计算及通信相关资源和性能综合决定的潜在性能指标 [4]。以交互式 3D 虚拟人为例，假设实时交互体验要求数字人对人的语言和动作的反馈的总时延不超过 200 ms，考虑传输时延等其它开销，预计 6G 系统中的计算时延要求在 10 ms 到 100 ms。考虑虚拟人的精细度和计算复杂度等，初步估算 50 万面以上高保真度智能交互型虚拟人所需的驱动和渲染的计算需求不小于 10 Tera FLOPS（Floating-Point Operations Per Second，每秒浮点运算次数）。考虑终端功耗和低能力终端等因素，移动终端通过按需计算卸载，由 6G 辅助支持交互式 3D 虚拟人。同时，假设三扇区覆盖情况下站间距为 100m，活跃用户密度为每 5 m^2 一个人，每人每天使用数字人的平均时间为 30 分钟，忙时集中率为 10%。那么，以上述交互式 3D 虚拟人作为典型应用所需的性能指标 [5] 如表 1 所示。

表 1 计算服务的性能指标

计算服务的性能指标		定义	以虚拟人为典型用例的需求
系统性能指标	算力密度	移动通信网络单位覆盖面积能提供的算力	~100000 Tera FLOPS/km^2
	计算连接密度	移动通信网络单位覆盖面积能提供的计算服务连接数量	~10000/km^2
用户性能指标	峰值算力	单用户可获得的峰值计算性能	~10 Tera FLOPS
	计算时延	从用户发起计算服务请求到接收到计算响应的整体时延	10 ms ~ 100 ms

3 移动算网融合的潜在技术方向

本节将结合现有标准探讨未来 6G 支持 AI 和通信服务的潜在技术方向。支持 AI 和通信服务的本质是移动算网融合 [6]。移动算网融合的关键在于如何打破各层之间的界限。拉通底层的计算资源和上层应用的计算和通信需求，这也是不同算网融合技术方向中计算控制功能的重要任务。针对 AI 这一典型业务类型，计算控制功能还可以进一步融合控制数据、模型等功能。从而面向 AI 业务按需进行算力、数据和算法相关的功能定制和优化，满足应用功能等的 AI 模型推理和训练等需求。

如图 2 所示，下面将 6G 潜在的移动算网融合技术方向分为基于演进的路线和基于变革的路线两类。基于演进的路线包括增强边缘计算（Edge Computing，EC）和增强 IMS，基于变革的路线则是在移动网络中新增计算控制功能和计算数据传输通道。对于 EC 增强和 IMS 增强，计算资源分别位于边缘应用服务器和 IMS 应用服务器，而不能位于 UPF 或基站等网络节点。基于演进的路线中，计算服务和通信协议仍然是上层应用和下层传输协议的关系。基于变革的路线则是在 3GPP 标准的控制面协议中实现了计算服务和通信的协同。因此，前述两类技术方向的差异在于计算服务与通信协议的关系。

基于 EC 增强的技术方向在用户设备（User Equipment, UE）侧边缘域名系统客户端（Edge DNS Client，EDC）功能 [7]，以及网络侧边缘应用服务器发现功能（Edge Application Server Discovery Function，EASDF）和边缘应用服务器（Edge Application Server，EAS）的基础上进行扩展。在边缘引入计算控制节点，计算控制节点通过用户面从 UE 侧接收应用的计算需求。计算控制节点还可以获得 EAS 计算节点的实时状态信息。6G 网络基于网络和计算的状态信息选择与 UE 计算需求匹配最优的 EAS。6G 网络还可以支持应用服务器重定向和基于计算状态的动态路由规划。从而，将 UE 需求实时调度到计算最优和访问时延最短的 EAS，从计算维度进一步提升用户体验。

基于 IMS 增强的技术方向既可以在已有 IMS 基础上扩展，也可以新增支持新业务的 IMS 标准。这一技术方向中

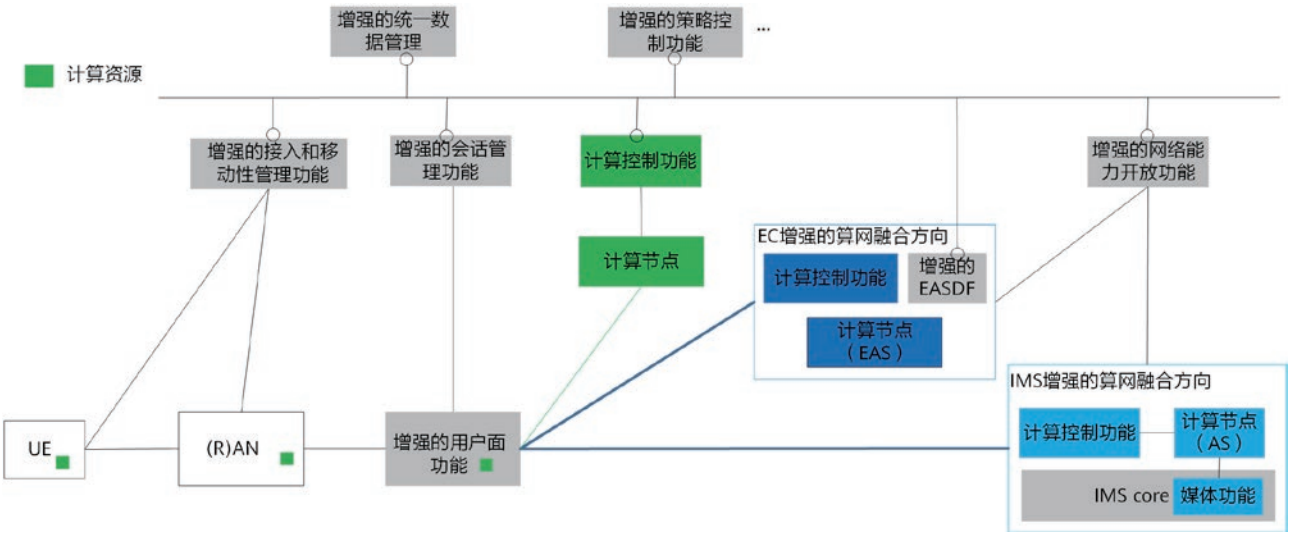


图 2 6G 潜在的移动算网融合技术方向

UE 和网络之间的计算控制信息和计算数据均通过用户面传输。下面以基于现有 IMS 扩展为例，现有 IMS 支持音频通话、视频通话和数据通道 [8]。面向 AR（Augmented Reality，AR）或者虚拟人等应用场景，数据通道应用服务器（Data Channel Application Server，DCAS）或媒体功能（Media Function，MF）具备图像和视频的计算处理能力（例如图像渲染等）。但是，目前媒体业务处理仅支持前述数据通道服务，且 IMS 不具备数据通道服务所需的计算节点的控制能力。因此，6G 支持基于数据通道的按需计算服务是潜在的演进方向。这将要求在数据通道的建立和使用过程中增强媒体应用管理，对于当前数据通道可以获取到的计算资源信息进行管理。应用服务器的计算能力和负载情况，甚至终端的计算能力将作为数据通道使用和建立的重要因素。同时，增加终端与网络的协商机制，最大程度的支持不同类型 UE 获取高算力需求的数据通道媒体业务。

基于变革的路线则需要在核心网控制面引入计算控制功能。计算控制功能还可以获得应用业务器的动态计算信息。计算控制功能可以通过有效的手段对网络中的异构计算资源进行有效度量和高效资源管理。计算控制功能采用服务化接口（Serivce-based Interface），UE 和网络之间的计算控制信息通过控制面传输。从通信协议层的角度来说，计算控制功能支持与 UE 对等的端到端的计算控制协议层，聚焦于计算控制信息交互，包括计算需求的交互，以及计算资源分配和更新的流程等。计算控制功能与通信控制功能基于策略进行协同，在图 2 具有计算资源的节点中选择合适的计算节点，以及管理计算数据传输所需的计算承载。从而实现通信和计算资源的联合动态调度，保证用户的多样性业务体验需求。

4 移动算网融合样机验证

基于变革的路线的 6G 移动算网融合方向，vivo 开发了移动算网融合样机。其中，4.1 节介绍移动算网融合样机系统架构和主要技术方案；4.2 节阐述计算资源评价指标和调度策略；4.3 节展示一个演示用例，并以时延作为指标对比不同计算资源或通信资源下的测试结果。

4.1 移动算网融合样机系统简介

如图 3 所示，移动算网融合样机主要由 vivo 自研终端验证平台、基站、通算融合核心网和管理编排系统四部分组成。vivo 自研终端验证平台作为计算服务的需求方，请求和使用网络计算服务。基站为终端提供无线数据传输，目前该移动算网融合样机暂不涉及基站功能的优化和测试。通算融合核心网响应终端的计算服务请求，并为终端提供计算服务。管理编排系统用于对通算融合核心网和计算服务应用功能所在的服务器集群通过 Kubernetes（K8s）统一管理。

vivo 自研终端验证平台采用了 Arm + FPGA 的架构。如图 4 所示，终端验证平台硬件主要由 FPGA 加速卡、全球用户识别卡（Universal Subscriber Identity Module，USIM）读卡器、软件无线电设备（Software Defined Radio，SDR）和上位机组成。在软件架构上，vivo 自研终端验证平台主要包括射频处理单元、物理层基带处理单元、高层协议栈、上位机程序四部分。高层协议栈和物理层基带处理模块分别运行在 Arm 和 FPGA 上。目前上述验证终端已经支持 NR（New Radio）Rel-15 的主要特性及 Rel-17 的部分特性，并支持面向 6G 的大带宽低延迟等新特性的原理验证。针对移动算网融合的验证需求，扩展开发了 vivo 自研终端验证平台的 NAS 模块，支持处理来自应用功能的计算需求和向网络发送计算请求。具体的，通过 AT Command（ATentionCommand，AT 命令）获取来自上位机（即终端应用处理器）的计算请求，将计算请求和响应消息分别通过 NAS 消息承载。最终，实现终端与通算融合核心网的计算和通信交互。

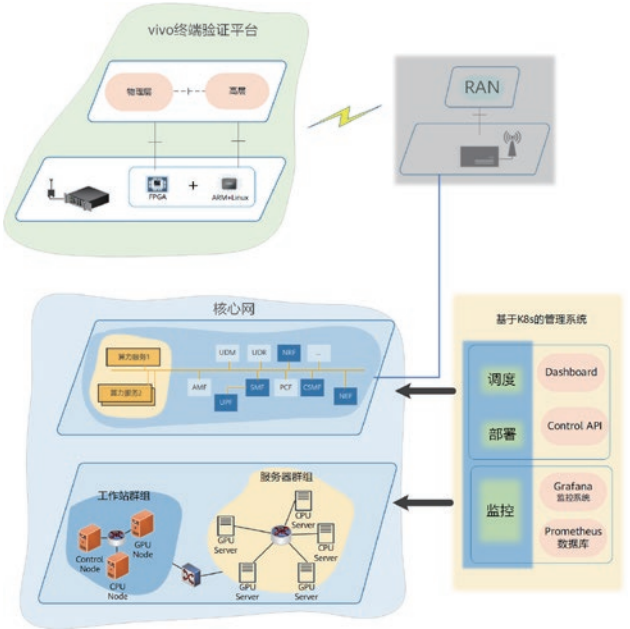


图 3 vivo 移动算网融合样机架构示意图

图 4 vivo 终端验证平台实物照片

通算融合核心网是以 5G 核心网为基础，新增研发了计算服务管理功能（Computing Service Management Function，CSMF）和计算服务器单元（Computing Server，CS）。CSMF 负责计算调度，以及计算资源状态的管理和维护。CS 负责监控当前计算节点状态并响应 CSMF 的管理和调度信息。增强会话管理功能（Session Management Function，SMF）和用户平面功能（User Plane Function，UPF）。SMF 负责统计各计算节点和 UPF 之间的网络状态信息，并响应 CSMF 对各种网络状态信息的订阅。

为实现计算节点间的负载均衡和可视化监控，管理编排系统通过通算融合核心网和现有云监控技术协同实现。对于计算节点的本地状态信息，使用基于监控数据库 Prometheus、节点监控器 Node-Exporter 和可视化工具 Grafana 的监控方案。CSMF 通过定时获取 Prometheus 数据库中的监控信息，从而基于所获取的信息实现计算和通信资源的动态调度。

4.2 计算资源评价指标与调度策略

本节将介绍移动算网融合样机采用的计算资源评价指标和调度策略。计算资源评价指标是计算与移动通信融合的重要问题之一，统一的算力模型是保障计算节点有效入网、管理和运营的基础。移动算网融合样机的计算资源的评价指标主要分为两大类，计算性能和网络传输性能。计算性能表征了节点所能承载的计算量、计算速度及占用情况，主要表示为 CPU、内存、GPU 等主要硬件的参数指标及各自的占用率。网络传输性能主要表示为计算节点和用户之间的网络带宽、延迟、抖动等参数及当前的网络负载情况等。网络传输性能参数与计算节点和用户之间的物理距离，当前网络路由和拥塞程度等相关。

前述两类评价指标中，网络传输性能和计算性能中的负载情况会随时间发生变化。因此需要建立一套实时监控的动态评价指标，为计算服务分配合适的计算节点。 此外，不同的应用对于计算资源的需求也是不同的。对于计算密集型应用，如 AI 模型的训练、推理等，倾向于使用拥有更高计算性能的节点。例如计算节点具有更高的 CPU 频率、更多 CPU 核心数、更多 CUDA 核等。对于实时交互等带宽密集型应用，如实时视频串流、云游戏等，更大的带宽、较低的时延、较小的时延抖动等网络传输性能更加重要。因此，需要为不同的计算业务设计不同的评价指标和调度策略。

最后，对于实际应用场景，计算业务的需求可能是刚性的也可能是弹性的。例如，对于 AI 模型的训练或推理这一类计算任务，对 GPU 的显存大小的需求通常是刚性的。如果不能满足需求，那么无法提供该计算服务。对于能满足显存需求的 GPU，更高的性能的 GPU 往往计算服务质量更高。但是，较差的性能的 GPU 也同样能支持该计算服务。因此，资源评价指标和调度策略的设计应该综合考虑刚性需求和弹性需求，使业务在能运行的基础上获得最好的表现。

综上所述，设计了一套能适配多样性的业务需求，能区分业务的刚性需求和弹性需求。基于实时监控的动态评价指标和调度策略如下所示：

$$S = \prod_{i=0}^{n} r_i \cdot (\omega^T \cdot f)$$

其中 f 和 w 都是 n 维向量，n 表示可量化的计算资源评价参数个数。f 为弹性需求对应的归一化的参数向量，如 CPU 主频，当前内存大小等；w 为弹性需求参数的加权向量，根据类型不同，不同业务可能具有不同的加权向量；r_i 为各刚性需求的满足情况，取值为 0 或 1。注意到，上述刚性需求和弹性需求所对应的评价参数是随负载、网络状况等因素动态变化的，即上述公式是时变的。

计算资源根据以上原则完成评价后由大至小排序，就可以获得该计算服务的最优计算节点列表。移动算网融合样机就采用了上述方法。需要注意的是，在实际场景中，除了用户的计算服务质量，网络的功耗控制也应该纳入调度策略的设计过程。例如，假设一种极端情况，网络中存在 n 个相同计算性能和相同网络传输性能的节点。当处理 n 个相同的计算服务请求时，如果仅以用户服务质量最优作为调度准则的话，那么会倾向于每个用户单独占用一个计算节点。这会导致网络中所有计算节点均无法进入休眠状态，功耗会显著增加，这一问题在较大规模的网络中会尤其明显。因此，调度策略的设计还需要权衡用户服务质量和网络功耗等开销，权衡服务质量与代价。

4.3 演示用例与测试结果

基于 4.1 节所述的移动算网融合样机，本节以基于 AI 的实时目标检测作为演示用例。将 UE 使用通算融合网络进行远程视频处理与 UE 本地视频处理进行对比，展示在该应用场景下移动算网融合的优势，该样机于 2023 年通过了 IMT-2030（6G）推进组的移动算力网络认证测试。

演示用例中终端对摄像头采集的视频做实时目标检测，标识出视频内被识别出的物体和人。基于 AI 的实时目标检测通常应用于实时人脸识别、障碍物检测、安防监控等场景。由于终端的硬件规模和功耗等限制，如果直接在 UE 本地进行处理，可能难以达到所需的效果。例如时延无法满足实时性需求，受硬件限制而无法运行更高精度的 AI 模型。此外，UE 本地处理还可能对终端的功耗和其它应用的计算资源造成影响。如果 UE 调用网络侧的计算资源，那么有望解决上述问题，更好地完成 AI 实时目标检测。因此，终端可以向网络请求计算卸载服务。网络利用服务器搭载的高性能 GPU 完成处理。本节所述的验证用例将详细演示这一过程的实现方法，具体步骤如下：

- 计算节点启动后自动注册至核心网的 CSMF 网元，并定时维护自身信息；
- 当 vivo 自研终端验证平台有实时目标检测需求时，终端选择使用本地算力或网络侧算力；
- 当本地算力不能满足要求或者本地不支持该类型的业务时，上位机通过 AT Command 控制终端向网络侧发起特殊的 PDU 会话建立请求，承载实时目标检测业务，并携带计算业务信息；
- 核心网 CSMF 网元依据终端的计算请求，选择合适的计算节点，并在被选中的计算节点上启动终端所需的计算服务。在本样机中，计算服务是实时图像检测和视频推拉流；
- 计算节点和终端建立计算服务连接，通过视频推流的方法，终端获得网络提供的实时目标检测结果，实现更高精度的快速实时目标检测。

测试中 6G 网络计算节点的计算能力强于 UE 本地计算的硬件的计算能力。网络提供的计算服务和 UE 本地计算处理的性能对比如图 5 所示。端到端时延为双向传输时延和画面帧处理时延之和。对于 UE 本地计算，受 CPU 和 GPU 性能限制，画面帧时延相比在网络侧计算更高。但是由于不涉及网络传输，其传输时延近似为 0。对于网络侧计算，端到端时延相比 UE 本地计算更低。网络对画面帧处理的计算时延增益大于双向传输时延开销，因此获得了比 UE 本地计算更低的时延。此外，由于 UE 本地计算受限，网络提供的计算服务可以支持更大参数量的 AI 模型，实现更高精度的目标检测。

在本例中，相比于本地计算，网络侧计算可以实现更低的平均时延和更小的时延抖动，为进一步说明节点计算性能和网络传输性能对端到端时延或业务效果的影响，固定视频流分辨率和目标检测模型，分别修改计算节点的 CPU 频率限制和网络带宽限制，统计平均时延如下图 6 和图 7 所示。

从图 6 和图 7 中可以看到，随着 CPU 频率增加，处理时延逐渐减小；而网络带宽在低于一定阈值时，性能严重受限，无法正常完成业务，在高于阈值时，随着带宽提升，性

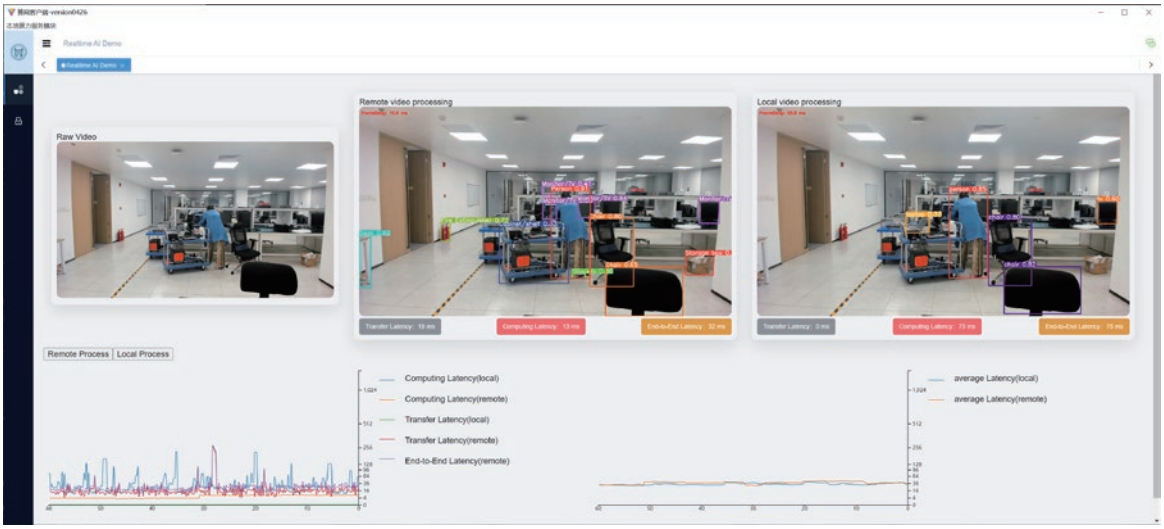


图 5 实时目标检测演示用例展示

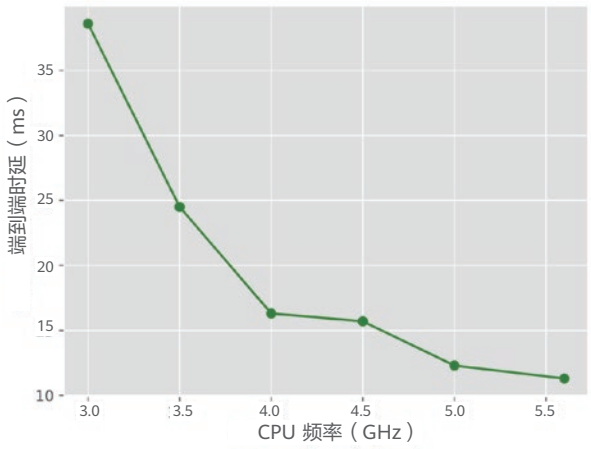


图 6 端到端时延与 CPU 频率关系曲线

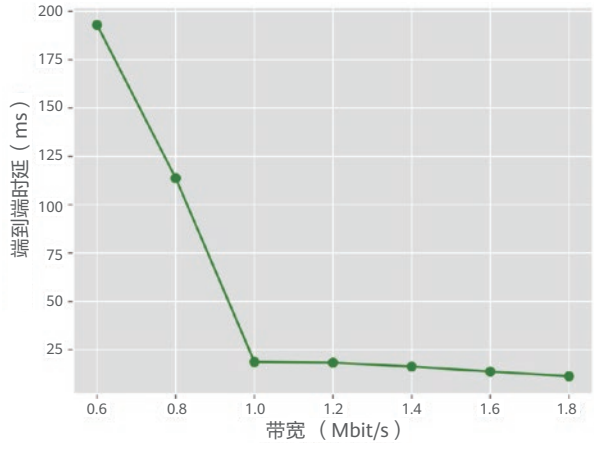


图 7 端到端时延与带宽关系曲线

能无明显增益。另外，本地时延并不总是高于 6G 计算服务的时延。当网络带宽较低，或者网络侧计算节点性能相比于 UE 本地计算性能差异不大时，6G 计算服务所引入的传输时延将高于计算处理时延的减小量。因此，在这种情况下该计算任务更适合于在 UE 本地处理。需要注意的是，在实际中，普通 UE 通常是基于 Arm 架构，而移动算网融合样机中 vivo 自研终端验证平台处理实时目标检测的算力远高于普通 UE，因此可以预期，相比本文的移动通算融合样机的结果，手机终端的计算时延会更高。因此，将计算业务卸载至网络侧可能带来更大的时延增益。

5 总结与展望

AI 和通信应用场景为 6G 拓展了新服务。我们建议了一种基于典型用例（如虚拟人等）确定 6G 系统 AI 和通信服务的性能指标，如算力密度、计算连接密度、用户峰值算力、用户计算时延等。为了支持 AI 和通信服务，6G 潜在的移动算网融合技术方向包括基于演进的路线和基于变革的路线两类。基于演进的路线包括增强 EC 和增强 IMS，基于变革的路线则是在移动网络中新增计算控制功能和计算数据传输通道。最后，通过移动算网融合样机对基于变革的路线进行了验证。基于 UE 请求的 AI 实时目标检测用例进行了对比测试，在测试配置下 6G 网络提供的 AI 和通信服务相比于 UE 本地计算具有更高的识别精度和更短的端到端时延。

移动通信作为经历了 1G 到 5G 长期商用的技术，其成熟度经过了实际应用的验证。而面向 6G 的 AI 和通信服务仍然存在一些挑战。目前，高端手机已经可以支持十万面级别虚拟人和 7B 参数量的大模型推理等。因此，移动终端上更大计算量的新应用，特别是具有超低时延和高计算量需求的应用将更能展现 6G AI 和通信服务的优势。同时，6G AI 和通信服务是提升新形态终端（如 XR 设备、机器人、无人驾驶车等）用户体验的方法之一。未来如果新形态的移动终端达到了更好的用户体验，那么这些新形态终端将有望具备和手机相似量级的用户普及率和使用率。新形态终端的普及也是 6G AI 和通信广泛应用的生态基础之一。计算即服务，将助力探索终端的无限可能。

参考文献

[1] ITU-R, "Framework and overall objectives of the future development of IMT for 2030 and beyond," November 2023.

[2] China Academy of Information and Communications Technology (CAICT), *Cloud Computing White Paper*, July 2023.

[3] Dinh C. Nguyen, P. Cheng, M. Ding, D. Lopez-Perez, P. N. Pathirana, J. Li, A. Seneviratne, Y. Li, and H. V. Poor, "Enabling AI in future wireless networks: A data life cycle perspective," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 553-595, September 2020.

[4] vivo Communications Research Institute, "6G Services, Capabilities and Enabling Technologies [R]," 2022.

[5] Jiang Dajie, Yuan Yannan, Zhou Tong, *et al.*, "6G-oriented services integrating communication, sensing and computing, system architecture, and key technologies [J]," *Mobile Communications*, 2023, 47(03): 2-13.

[6] vivo Communications Research Institute, "6G Network Architecture [R]," 2023.

[7] 3GPP TS 23.548 V18.5.0, "5G system enhancements for edge computing," March 2024.

[8] 3GPP TS 23.228 V18.5.0, "IP multimedia subsystem (IMS)," March 2024.

NetGPT 十大问题

童文[1]，彭程晖[1]，杨婷婷[2]，王飞[1]，邓娟[3]，李荣鹏[4]，杨璐[1]，张宏纲[4]，王栋[5]，艾明[6]，杨立[7]，刘光毅[3]，杨旸[8]，肖遥[1]，岳烈骥[3]，孙万飞[6]，李泽旭[5]，孙文文[7]

[1]华为无线技术实验室
[2]鹏城实验室
[3]中国移动研究院
[4]浙江大学
[5]中国电信研究院
[6]中信科移动通信
[7]中兴通讯
[8]香港科技大学（广州）

摘要

当前移动网络中应用的 AI 算法通常为“专用”模型，为特定任务进行定制化设计，在通用性、性能增益、管理、协作等方面存在明显不足。随着大模型（Foundation Model）的快速发展和应用，可以预见大模型将在未来移动通信中扮演重要角色。本文将应用在移动网络中的大模型定义为 NetGPT（全名为 Network Generative Pre-trained Transformer），并系统阐述了 NetGPT 在设计和应用过程中需要解决的十大问题。

关键词

NetGPT，大模型，无线，移动通信，十大问题

1 引言

1.1 背景

随着国际电信联盟无线电通信部门（International Telecommunication Union – Radiocommunication Sector，ITU–R）将“AI 和通信”列为 IMT–2030 的主要场景之一 [1]，近年来已有很多研究人员专注于这一领域，并取得了丰硕的成果，惠及移动网络的各个技术域，包括无线接入网、核心网、运维和用户设备。例如，有的专家利用深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）生成核心网策略，也有人将深度 Q 学习（Deep Q–Learning，DQN）用于网络运维。而不论哪种方式，都需要为各个具体用例定制一套 AI 算法，这种做法存在诸多不足，如通用性低、性能增益有限、管理复杂、协作困难等 [2–4]。

大模型是 AI 技术的一项突破性发展，其应用领域十分广泛，特别是大语言模型（Large Language Model，LLM），在对话、编程等任务中展现出了令人惊叹的能力。通过大模型，移动网络就有望能实现高通用性、极致性能、简化管理、方便协作、多任务处理的目标。为了区别于其他行业应用，我们将移动通信网络中使用的大模型定义为“NetGPT”（全名为 Network Generative Pre–trained Transformer）。目前，NetGPT 的设计和应用还处于起步阶段，需要从模型本身的设计与实现，以及网络架构如何进行更好的支撑等角度进行全面的分析和推进，为此本文总结了 NetGPT 的十大核心问题，为后续的研究梳理方向。

1.2 NetGPT 的定义

移动网络包含无线接入网、核心网、运维等不同的技术域，它们在功能特性、数据结构、性能需求上都有着明显的区别。比如，针对运维域，可以利用参数高效微调（Parameter–Efficient Fine–Tuning，PEFT）技术直接对 LLM 进行微调，从而得到适合运维的 NetGPT 模型。如果是在网络边缘，则可以对 LLM 进行蒸馏或剪枝，得到一个体量较小的 NetGPT 模型。除了这两种 NetGPT 模型，我们还可以从零开始训练一种全新的 NetGPT 模型，保留与 LLM 类似的神经网络架构。因此，NetGPT 并不是适配所有移动通信场景的单一模型，而是一系列模型的组合。

本文为 NetGPT 定义了三层体系架构，即 L0、L1 和 L2。其中，NetGPT–L0 代表全网通用大模型；NetGPT–L1 代表网络不同领域大模型（含无线接入网大模型、核心网大模型、运维大模型）；NetGPT–L2 代表各域特定场景下的网络模型——如图 1 所示，无线接入网有针对物理层（PHY）的 NetGPT–L2 大模型，运维域有网络调优专用的 NetGPT–L2 大模型，等等。

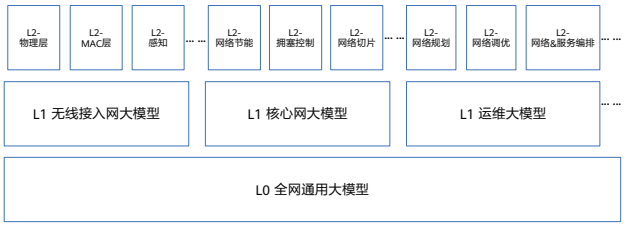


图 1 NetGPT 的三层体系架构

2 NetGPT 的十大基础问题

本节将介绍 NetGPT 在 6G 无线通信中所面临的问题，这些问题可分为两类：一类是 NetGPT 本身的设计问题，另一类是移动网络架构设计如何支撑 NetGPT 应用的问题。

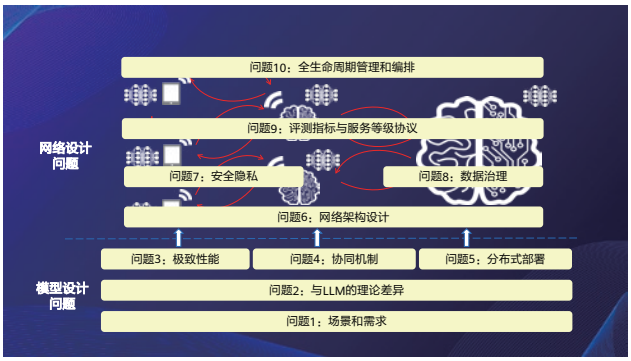


图 2 NetGPT 的十大问题

问题 1：场景和需求

NetGPT 需要基于庞大的数据集和数十亿的参数进行训练，对算力要求极高。然而，无线网络通常由多个边缘设备和移动终端组成，具有高异构、广分布的特点，缺乏像云端一样集中且强大的计算能力。因此，将云端 AI 模型和算法直接部署到无线网络的做法是行不通的，需要重新设计出能适配无线网络特点的算法以及能原生支持 NetGPT 的网络架构。值得注意的是，无线网络的层级越低，对实时性和准确性等服务质量（Quality of Service，QoS）的要求就越高。然而，LLM 的复杂性（如幻觉问题、体量超大等）可能会阻碍这些 QoS 的实现。这就引出了一个关键问题：NetGPT 在无线网络中的应用是否存在边界？例如，NetGPT 是否只适用于层级更高的空口层，而不适用于物理层？这些边界也会影响 NetGPT 在每个具体应用中可能发挥的作用，如 NetGPT 能在多大程度上支撑未来的运维系统——能让网络完全自智还是部分自智 [5]？因此，在研究 NetGPT 时，必须要明确 NetGPT 的场景和边界。

问题 2：与 LLM 的理论差异

LLM 是最具代表性的大模型，也是 NetGPT 的底座 [6–8]。然而，通信与自然语言处理（Natural Language Processing，NLP）是截然不同的两个领域，因此 NetGPT 与 LLM 之间也势必会存在一些理论差异，主要体现在以下几个方面：

- 数据特征：NetGPT 使用的数据集包含通信信息（如信道信息），这类信息用高维张量的形式表示，与 LLM 中基于令牌（Token）的数据形式有本质区别。
- 后端任务：无线网络需要处理各种不同类型的任务，因此 NetGPT 的输出是多样的，与 LLM 仅用令牌作为输入/输出的形式明显不同。
- 模型大小：NetGPT 定义了一种多层模型架构，每一层都部署了不同体量的模型。有些 NetGPT 模型，特别是在网络边缘（如基站）部署的 NetGPT–L2 模型，可能仅有 1~10 亿参数，对比集中式 LLM 50~2000 亿的参数规模有相当大的差异。

这就需要我们研究清楚以下问题：NetGPT 的神经网络架构能否与 LLM 保持一致？ NetGPT 能否激发 AI 理论和相关神经网络架构方面的重大创新?

问题 3：极致性能

未来无线网络（如 6G 网络）在实时推理、可靠性等方面都提出了极致性能的诉求，远超过当前大模型应用所能达到的水平。NetGPT 必须满足这些极致性能的诉求，才可以在无线网络中应用。

相较于 5G，未来无线网络不仅需要将通信性能提 10 倍，还需要支持各类对性能要求极高的新服务，如自动驾驶、工业机器人等。为保证未来无线网络能实现极致性能，NetGPT 将 AI 与通信融合，这样就可以在极短的时间内得出推理结果，进而实时适应高度动态的无线网络环境。然而，NetGPT 计算过程极为复杂、参数量十分庞大，要在未来无线网络中实现 0.1 ms 级的实时推理绝非易事，亟需更加高效的模型算法和推理加速方法。

除此之外，未来无线网络对可靠性的要求也极高 [1]。大模型可能出现的幻觉问题，容易导致网络决策错误，带来不可预测的风险。因此，目前的大模型大多仍局限在外围辅助，无法触及核心的网络服务功能。为满足未来无线网络的高可靠性要求，可以从提高数据质量、改善模型结构、引入知识图谱等方面入手，探索提高 NetGPT 可靠性的方法。

问题 4：协同机制

大模型对算力、通信资源和能源的消耗极高，因此像 NetGPT–L2 这样体量较小的模型更适合部署在无线网络边缘（如基站等位置）。NetGPT–L2 的优势是善用本地知识处理特定场景，但对其他场景缺乏通用性。如果能与集中部署的 NetGPT–L0/L1 实现协同，就可弥补 NetGPT–L2 在通用性上的不足。

因此，不同大小的模型之间如何协同（包括模型训练和推理方面的协同），是未来无线网络部署 NetGPT 必须要解决的问题。训练协同的具体场景包括：（1）NetGPT–L0/L1 整合本地数据，通过变换、蒸馏、剪枝等方式生成更适合本地的 NetGPT–L2 模型；（2）NetGPT–L0/L1 辅助 NetGPT–L2 进行训练；（3）NetGPT–L2 向 NetGPT–L0/L1 反馈成效信息，帮助 NetGPT–L0/L1 持续强化自身的能力，形成有机循环的智能体系。推理协同的具体场景包括：（1）若 NetGPT–L2 无法独立达到预期的置信度，可与 NetGPT–L0/L1 协同完成实时推理；（2）在每个端侧部署多个 NetGPT–L2 模型，协同完成实时推理。

上述场景涉及的一些关键算法还有待进一步研究，包括如何根据大模型 NetGPT–L0 参数的相关性进行参数修剪，如何高效微调快速适应新任务等。除了算法上的挑战，NetGPT 还需要解决跨厂商协同的问题，因此有必要定义一套标准化的协同机制，包括功能和程序在内的协同内容的标准化、协同集的生成方法、协同事故的系统控制等，都是需要继续深入研究的课题。

问题 5：分布式部署

在“端–边–云”协同方面，根据服务需求，可能需要在无线网络边缘或移动终端部署整个或部分 NetGPT–L0/L1 或第三方大模型。而无线网络中的“端–边–云”环境是动态变化的，为高效应对环境的动态性和异构性，需要研究大模型的拆分方式。由于每个节点的计算资源、存储能力和传输速率都有所不同，需要根据节点的实际情况合理拆分模型，在实现节点间负载均衡的同时，保证资源利用率最大化。

在增量训练过程中，如果网络中分布的子网出现参数更新延迟，那么大模型就可能会遇到参数不一致的问题。为了保证无线网络边缘之间的参数一致性，需要探索跨节点增量训练机制。

此外，分布式节点之间的高效通信也需要重点关注。一方面可以从算法入手，通过剪枝和量化等手段压缩模型，降低通信开销。另一方面，还应细化网络接口与协议，让网络节点间的数据流动更加高效、便捷。

问题 6：网络架构设计

NetGPT 不断融入未来无线网络，会对网络架构产生深远影响，包括以下几个方面：

- 网元智能化演变：现有的网络功能可以全部或部分交由 NetGPT 实现，但随着 NetGPT 广泛应用于端到端场景，某些经典的网络功能可能不再适用，需要深度重组或重新设计，包括当前 5G 核心网的网元分类和组织方式（如服务化架构）[9]。
- 网络接口/协议变化：传统的接口协议需要基于标准化的信号和元素串，在 NetGPT 中，这类协议会被模型间的协作接口（如 Token 接口）取代。
- 网络能力更新：未来无线网络需要原生支持 NetGPT 和第三方大模型，实现各种 NetGPT 模型的在线更新与演进。针对具体场景，大模型需要利用低秩适应（Low-Rank Adaptation，LoRA）等 PEFT 手段开展增量训练。相比之下，NetGPT 除了要具备大模型的这些通用能力外，还需增量学习新知识，即具备终身学习的能力。考虑到 NetGPT 的庞大参数量和数据量，已有的数据并行（如联邦学习）和模型并行（如拆分学习）方式可能存在一定的局限性，需要深入探索二者的涌现能力。
- 网络服务优化：基于大模型解释语义信息的能力，未来无线网络可以为每个应用程序单独生成一个专有网络，从业务逻辑、网络逻辑和网络资源等不同维度提供服务。

问题 7：安全隐私

在端到端无线网络中应用 NetGPT 时，需特别关注安全问题。海量的数据和参数增加了 NetGPT 受攻击的风险，特别是容易被攻击者恶意植入后门，导致模型被非法指令误导。此外，模型越变越大之后，会变得越来越有偏见，可信度越来越低。因此，我们需要对 NetGPT 进行有针对性的安全设计。

与其他 AI 算法一样，NetGPT 的算法也缺乏可解释性，因此在网络中应用 NetGPT 会存在一定的风险。这方面虽然已经有一些重要的理论研究成果，但还缺乏一个完善的理论框架，特别是还没有大模型定量分析相关的数学方法和分析方法。

此外，在使用 NetGPT 的过程中，用户的隐私数据，如用户账户信息、历史对话记录、互动中携带的各种隐私信息，可能会暴露给供应商、服务提供商及其关联公司。在世界各地发生多起数据泄露事件之后，全球监管机构也开始关注大模型带来的数据隐私风险。因此，亟待出台 NetGPT 相关的数据隐私原则和使用规定。

问题 8：数据治理

由于 NetGPT 的性能在很大程度上取决于数据的质量，我们需要为 NetGPT 重新设计数据治理服务 [10]，涉及的内容有：

- NetGPT 会收到各方、各技术域的数据，为了高效处理这些海量的异构数据，未来无线网络需要搭建一套完善的数据治理框架，覆盖数据所有权、数据格式、数据质量、数据隐私等方方面面。
- 为保障 NetGPT 的推理和在线更新能顺利进行，未来无线网络必须实现大规模数据的分布式存储和实时供给。对于要求极致性能的 NetGPT 而言，必须设计出能保障 QoS 的数据服务。
- 为了提供更加可靠的服务，减少大模型的幻觉现象，未来无线网络需要将 NetGPT 与网络知识图谱结合起来。

综上，未来无线网络需要针对不同的 NetGPT 模型设计出一套统一的数据治理框架。

问题 9：评测指标与服务等级协议

如何全面、客观地评测 NetGPT 的性能也是一个亟待解决的问题。一方面，通过对 NetGPT 性能的评测，可以为优化和改进 NetGPT 提供强有力的依据，取得更好的应用效果和更大的商业价值。另一方面，NetGPT 评测可以作为基准，用来判断不同厂商 NetGPT 的性能和适用性。

现有的准确率（Accuracy）、F1 评分（F1 Score）[11] 和 BLEU 值（Bilingual Evaluation Understudy）[12] 等指标远不能满足 NetGPT 的评测需求，因此需要根据网络特征制定更有针对性的指标，如功能正确性、通信任务成功率等指标。这些指标可以按需组合，对 NetGPT 在特定场景下的表现进行更精细化的评测。在处理网络专业领域任务时，还需关注模型对领域特定术语、概念和规则的理解和应用，以确保评测结果的可靠性。

与语言领域不同，网络相对比较封闭，能公开获取的标注数据有限，收集小样本数据比较困难，导致模型训练难以覆盖实践中可能遇到的所有场景。因此，评测 NetGPT 对不同网络场景的通用性也至关重要。

问题 10：全生命周期管理和编排

网络中部署的 NetGPT 模型可能来自于不同的厂商，这就需要建立统一的机制，在整个生命周期内对这些模型进行高效的编排和管理（包括模型的添加、更新、转移和移除等任务），并在此过程中妥善保护 NetGPT 模型的知识产权。然而，NetGPT 模型的所有者并不一定就是网络运营商，并且他们可能并不愿意将模型的控制权交给运营商。因此有必要建立一套平衡的协同管理机制，更好地保护双方利益。

此外，如何组织和调度这些模型是 NetGPT 面临的另一大挑战。首先，各厂商的模型语言必须标准化，才能建立统一的接口和交互方法；其次，NetGPT 的部署将涉及连接、计算、存储等多维资源，如何根据场景特征和需求来编排这些模型及网络资源，以提高系统性能和资源利用率，将是一项艰巨的任务。

3 总结与展望

大模型的出现有望给未来无线网络带来革命性的变化，但许多相关的方向和标准化工作还有待进一步研究。NetGPT 是无线网络和大模型深度双向融合的趋势，本文深入探讨了 NetGPT 的十大基础问题，包括基础理论、场景需求、网络架构、部署管控、数据治理等。面向未来，NetGPT 的进一步发展仍需要大家共同的努力，唯有持续投入和深入研究，才能将大模型真正融入无线网络。

参考文献

[1] ITU-R, "Framework and overall objectives of the future development of IMT for 2030 and beyond," 2023.

[2] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.

[3] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6G: AI empowered wireless networks," *IEEE Communications Magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[4] Y. Yang, M. Ma, and H. Wu, "6G network AI architecture for everyone-centric customized services," *IEEE Network*, pp. 1–10, 2022.

[5] T. McElligott, "Network automation using machine learning and AI," TMForum, 2020.

[6] Y. Chen, R. Li, Z. Zhao, C. Peng, J. Wu, E. Hossain, and H. Zhang, "NetGPT: A native-AI network architecture beyond provisioning personalized generative services," arXiv e-prints, p. arXiv:2307.06148, July 2023.

[7] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?," arXiv e-prints, p. arXiv:2306.10249, June 2023.

[8] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," arXiv e-prints, p. arXiv:2307.02779, July 2023.

[9] 6GANA, "Ten questions of 6G native AI network architecture," https://www.6g-ana.com/, 2022.

[10] 6GANA, "6G data service - concept and requirements," https://www.6g-ana.com/, 2022.

[11] C. J. Van Rijsbergen, "Information retrieval (2nd ed.)," Butterworth-Heinemann, 1979.

[12] K. Papineni, S. Roukos, and T. Ward, "BLEU: A method for automatic evaluation of machine translation[C]," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318, 2002.

6G 网络通信大模型关键问题与技术探索

杨婷婷，张平，郑孟帆，李楠，马帅
鹏城国家实验室

摘要

本文旨在阐述 6G 网络通信大模型的技术挑战和未来发展方向。首先，概述了 6G 网络通信大模型的研究背景与现状，探讨了网络通信大模型的关键问题与挑战。然后，探索了生成式人工智能赋能的通信感知一体化及语义通信技术，介绍了 6G 基础大模型的开发进展。最后，对当前 IEEE 通信学会“生成式人工智能新兴技术倡议委员会（Large Generative AI Models in Telecom Emerging Technology Initiative，简称 GenAINet ETI）”的建设情况进行了介绍。

关键词

6G，网络通信大模型，生成式人工智能

1 引言

国际电信联盟无线电通信部门（International Telecommunication Union – Radiocommunication Sector，ITU-R）在近期发布的 IMT–2030 6G 愿景中明确了“通信与 AI”这一新使用场景。未来无线网络的作用将发生根本性变革，从单纯提供连接服务的基础设施，逐步演变为原生支持计算、数据、AI 功能及通算一体的智能系统。图 1 展示了从 2020 年到 2030 年，“通信+AI”的发展历程和未来展望。在“通信+AI 1.0”时代（传统 AI 与通信结合阶段），研究主要集中在 R17 到 R18 5G–A 时期，标志着 6G 趋势及愿景的初步形成。2023 年起，进入“通信+AI 2.0”时代（大模型与通信融合阶段），从 R19 开始，6G 研究与提案准备阶段将持续至 2026 年，随后在 R20 阶段进行 6G 技术研究，并在 2028 年左右的 R21 阶段制定 6G 规范，最终在 2030 年实现 6G 的商用。这一过程中，候选技术的提交、技术规范与标准的制定也将同步进行，全面推动通信网络向智能化系统转型。

在“通信+AI 1.0”时代，AI 与无线通信融合的研究已持续超过五年 [1]。在此期间，传统“N 层 N 面”等 AI 与无线融合网络架构以及相关算法研究取得了显著进展，然而，随着研究的深入，这一阶段也逐渐达到了瓶颈期。近期，随着大模型技术的涌现，“通信+AI 2.0”时代正逐步开启。大模型技术发展迅速 [2–4]，并且有望与 5G/6G 技术一起成为信息和通信技术（Information and Communications Technology，ICT）的基础设施。面对“通信+AI 1.0”时代存在的模型泛化能力弱、定制样本代价高、常识推理能力低、复杂任务处理难等应用痛点，大模型技术具备解决这些难题挑战的潜力。

2 6G 网络通信大模型的关键问题

作为人工智能领域的前沿技术，生成式预训练大模型技术将在无线网络服务能力增强、资源配置优化以及泛在连接智能等方面发挥重要作用，有望重构下一代无线通信网络的设计范式。与此同时，未来无线网络也将在网络架构、无线空口、接入网与核心网等方面进行变革或增强，以赋能多种基于生成式预训练大模型的应用。

然而，当前大模型的研究也存在诸多挑战，在 6GANA 网络大模型专题研讨会及第三届通算一体网络基础理论和技术研讨会的基础上，来自通信、网络、计算、数学、人工智能等不同行业领域的专家进行了深入探讨与总结，并由 6GANA 发布了《网络大模型十大问题白皮书》[5]。白皮书系统阐述了网络通信大模型 NetGPT 的十大重点研究问题，包括 NetGPT 的场景和需求问题、基础理论问题、极致性能要求问题、协同问题、原生分布式部署问题、网络架构设计问题、安全隐私问题、数据服务问题、评判体系与方法问题，以及全生命周期管控与编排问题。白皮书不仅系统性地阐述了上述问题，还深入探讨了每个问题的具体内涵和挑战。例如，在场景和需求方面，白皮书指出网络通信大模型需要在各种复杂网络环境中进行适配，以满足不同应用场景的需求，包括提高网络性能、实现智能化管控优化等。在基础理论方面，白皮书强调了网络通信大模型与大语言模型（Large Language Model，LLM）的区别，提出需要开发针对通信领域的专用模型架构，以提升网络通信大模型的泛化能力和处理效率。

此外，白皮书还特别关注了性能要求、协同工作、分布式部署、网络架构设计、安全隐私、数据服务、评判体系和全生命周期管控等方面的问题。具体来说，在性能要求方面，强调了网络通信大模型在实时性、可靠性、高可用性和灵活性扩展性方面的严格要求，并提出了相应的优化策略，如通过高效硬件加速和模型压缩优化来提升推理效率。在协同方面，探讨了大模型和小模型之间的协同进化，提出了在边缘和云端之间进行模型能力的有效传递和协同工作机制。针对分布式部署问题，提出了模型拆分、分布式训练和高效节点间通信机制等具体方案。在网络架构设计方面，建议在架构层面实现通信、计算、数据和 AI 算法模型的深度融合，以提升数据处理和决策推理的效率。在安全隐私方面，提出了模型可靠性、可解释性和隐私保护方面的具体挑战和解决方案。在数据服务方面，指出需要为网络通信大模型提供高效的数据支持，包括处理大量异构数据、实现大规模数据的分布式部署和实时供给，以及构建网络知识图谱以提升大模

图 1 通信 AI 融合技术演进路线图

型的推理能力。评判体系方面，提出了针对网络领域的特定评测指标和方法，强调了对网络通信大模型性能和安全性的全面评估。最后，探讨了网络通信大模型全生命周期管控和编排的相关问题，提出了任务分解、任务编排和任务执行的具体框架，以实现网络通信大模型在不同场景中的高效应用。白皮书为网络通信大模型的研究和应用提供了系统性的指导，致力于推动网络通信大模型全面赋能 6G 通信网络。

3 6G 网络通信大模型技术探索

“鹏城云脑”是国家网络智能领域的重大科技基础设施，致力于提供高性能的计算平台（如图 2 所示）。从 2019 年开始推出“鹏城云脑 I”，具备 120P FLOPS 的算力，作为模型研发的高规格算力平台。2020 年推出的“鹏城云脑 II”基于国产芯片的 E 级智能计算平台，具备 1E FLOPS 的算力，支持千亿级参数 AI 模型训练。计划于 2025 年推出的“鹏城云脑 III”，将算力进一步提升至 16E FLOPS，面向新一代智能网络通信研究，提供高能效、全产业链可控、可生态化的计算能力。“鹏城 · 脑海”大模型平台是“鹏城云脑”的重要组成部分，涵盖了如基于 200B 参数的中文大模型和 33B 参数的通用大模型等多项应用，显著提升了 AI 大模型在不同领域的性能。该平台提供了跨领域的 AI 算力支持，并具备通用智能和定制化应用的能力，支撑复杂任务处理和高效能计算需求。

本节首先对 6G 网络通信大模型“鹏城 · 脑海”技术进行概述，并以生成式人工智能赋能的通信感知一体化以及生成式人工智能赋能的语义通信两个方面为例，对网络通信大模型方向的研究进展进行概述。

3.1 6G 网络通信大模型“鹏城 · 脑海”技术概述

基于“鹏城云脑”和“鹏城 · 脑海”大模型，笔者所在研究团队面向下一代 6G 网络进行网络通信大模型的技术开发，过程包括：首先，针对通信小模型存在的泛化能力弱、定制样本代价高、常识推理能力低、复杂任务处理难等应用痛点，提出通过运用“鹏城 · 脑海”大模型，结合运营商合作数据、开源获取数据和通信专家指令数据，进行领域应用的微调。其次，通过模型和数据的扩充（如从 10M 参数、10B 令牌的小规模通信仿真数据扩充到 100B 参数、10T 令牌的大规模仿真与实测通信数据），探索大模型的智能涌现能力，旨在实现从通用到专有、从小模型到大模型的转变，以推动 6G 网络通信大模型的设计开发。

3.2 生成式人工智能赋能的通信感知一体化技术探索

笔者所在团队针对 6G 网络通信数据及自然语言语料的多模态数据统一表征进行了相关的数学理论研究，并基于 NVIDIA Sionna 神经无线电框架 [6] 构建了多径信道生成、无线信道数据采集以及感知成像的散射点信息提取仿真平台。同时，基于 Transformer 架构，分别构建了 30M 和 3.5B 参数量的生成式预训练大模型，实现了感知成像任务的训练与推理。这些模型支持 Sensiverse [7] 等已开源的无线信道数据集，致力于为业界提供通信感知一体化大模型测试平台和评估基准。目前，模型已在鹏城启智社区开源，开源代码网址：https://openi.pcl.ac.cn/Foundation-Model-of-6G-Communication。

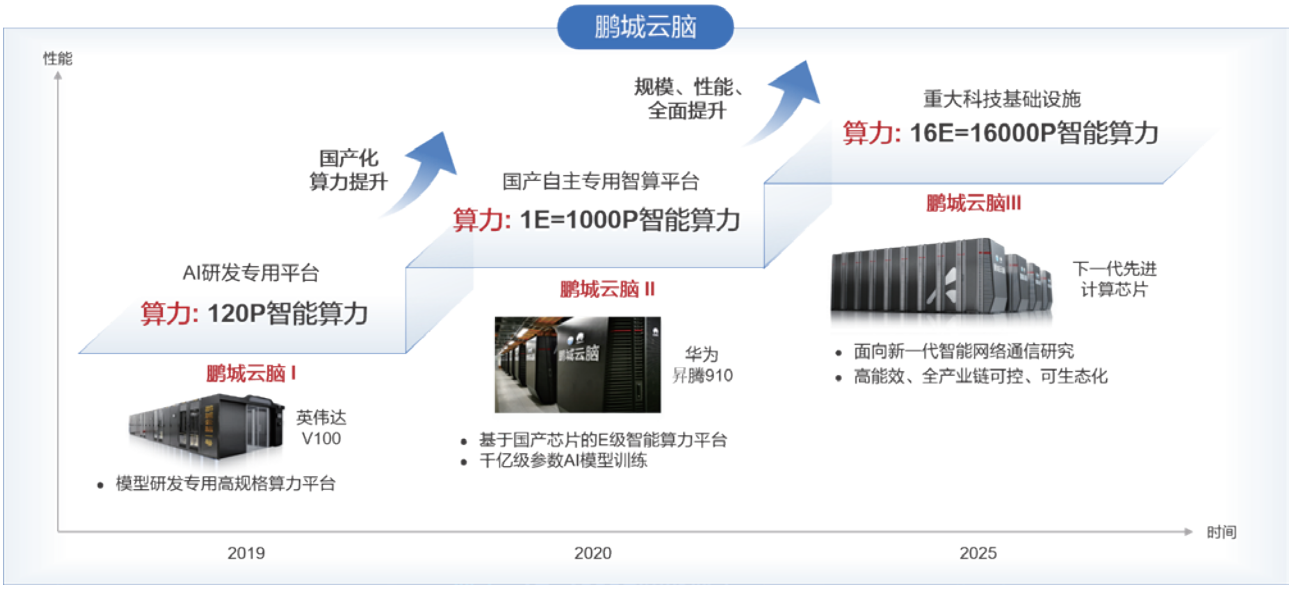


图 2 “鹏城云脑”网络智能重大科技基础设施

3.3 生成式人工智能赋能的语义通信术探索

生成式人工智能赋能的语义通信从语义维度对信源数据进行特征提取、编码、传输和译码，其本质是利用算力资源置换通信物理资源开销（带宽、功耗或时延等），侧重信息语义级的准确传输，优势在于先理解后传输，可大幅提升通信系统的传输效率。利用“鹏城·脑海”大模型的强大算力，可以设计并实现生成式人工智能赋能的语义通信。与传统语法通信相比，生成式人工智能赋能的语义通信具有以下优势：1）超高压缩比：传统数据压缩性能受香农熵极限制约，语义编码侧重提取并编码任务相关的语义信息，忽略与任务无关的信息，其压缩性能可超越香农熵极限。2）抗噪声能力强：语义通信采用信源信道联合编码，对比低密度奇偶校验（Low-Density Parity Check，LDPC）等现有的信道编码方式，其抗噪声能力更强。特别是在低信噪比条件下，可以避免传统信道编码方式所带来的“悬崖效应”问题。3）显著提升网络容量：与传统技术路线通过增加带宽、天线数目或能耗来提升传输容量的方式不同，语义通信主要通过超高压缩比来减少传输数据量，显著提升端到端数据传输量。生成式人工智能赋能的语义通信将传统的语法通信转变为面向内容的语义通信，将引发移动通信网络演进的范式变革，成为6G移动通信的基础理论之一。

笔者所在团队基于“鹏城·脑海”大模型对生成式人工智能赋能的语义通信进行了深入研究，在发射端挖掘数据的语义新维度，实现信息的高度抽象表征和智能简约传输；在接收端利用生成式人工智能强大的内容生成能力，恢复高质量信源数据内容，为突破语法通信瓶颈提供了一条全新的技术路线。

4 IEEE 生成式人工智能新兴技术倡议委员会介绍

为了进一步推动6G网络通信大模型技术领域的生态创新和发展，并构建网络通信技术国际合作联盟，笔者所在团队作为创始成员单位，联合华为技术有限公司、Khalifa University、中国移动及6GANA等单位，在IEEE通信学会发起成立“生成式人工智能新兴技术倡议委员会（Large Generative AI Models in Telecom Emerging Technology Initiative，简称GenAINet ETI）”[8]并担任学术主席。作为IEEE通信学会目前唯一的网络通信大模型技术交流与合作的学术平台，该委员会旨在通过数学、信息理论、无线通信、人工智能、信号处理、网络以及信息安全等多学科交叉碰撞，共同推进网络通信大模型的研究。目前，委员会已经吸引了来自全球近200家机构的参与，为业界学者、研究人员以及行业领导者提供了一个跨学科的前沿创新平台。未来还将召开多次国际研讨会，并发布研究报告和白皮书，共同促进6G网络通信大模型技术标准的制定和推广。

5 结语

本文系统阐述了6G网络通信大模型的关键问题和技术探索。首先介绍了其在未来无线网络中的重要地位和研究现状，然后结合6GANA发布的《网络大模型十大问题白皮书》，详细探讨了基础理论、场景需求、网络架构、部署管控、数据治理等方面的十大重点研究问题及其内涵和挑战。此外，本文还展示了笔者所在团队在6G网络通信大模型方面的探索和成果，并详细介绍了IEEE通信学会“生成式人工智能新兴技术倡议委员会GenAINet ETI”的建设情况。

参考文献

[1] Khaled B. Letaief, Yuanming Shi, Jianmin Lu, and Jianhua Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," IEEE Journal on Selected Areas in Communications 40, no. 1 (2021): 5–36.

[2] Hang Zou, Qiyang Zhao, Lina Bariah, Yu Tian, Mehdi Bennis, Samson Lasaulce, Merouane Debbah, and Faouzi Bader, "GenAINet: Enabling wireless collective intelligence via knowledge transfer and reasoning," arXiv preprint arXiv:2402.16631 (2024).

[3] Hongyang Du, Guangyuan Liu, Dusit Niyato, Jiayi Zhang, Jiawen Kang, Zehui Xiong, Bo Ai, and Dong In Kim, "Generative AI-aided joint training-free secure semantic communications via multi-modal prompts," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 12896–12900. IEEE, 2024.

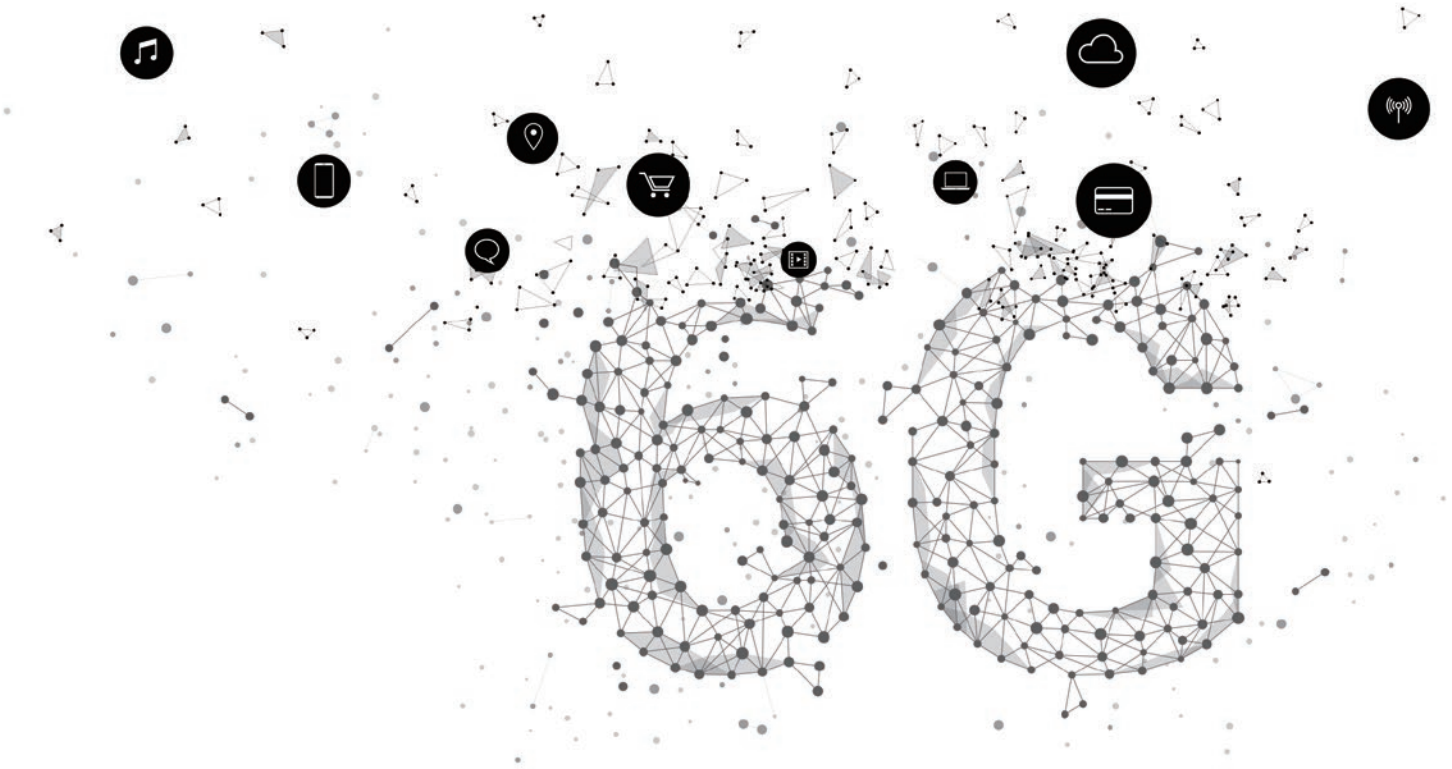
[4] Yifei Shen, Jiawei Shao, Xinjie Zhang, Zehong Lin, Hao Pan, Dongsheng Li, Jun Zhang, and Khaled B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," IEEE Communications Magazine (2024).

[5] Wen Tong, Chenghui Peng, Tingting Yang, Fei Wang, Juan Deng, Rongpeng Li, Lu Yang, *et al*., "Ten issues of NetGPT," arXiv preprint arXiv:2311.13106 (2023).

[6] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller, "Sionna: An open-source library for next-generation physical layer research," arXiv preprint arXiv:2203.11854 (2022).

[7] Jiajin Luo, Baojian Zhou, Yang Yu, Ping Zhang, Xiaohui Peng, Jianglei Ma, Peiying Zhu, Jianmin Lu, and Wen Tong, "Sensiverse: A dataset for ISAC study," arXiv preprint arXiv:2308.13789 (2023).

[8] https://www.comsoc.org/about/committees/emerging-technologies-initiatives/large-generative-ai-models-telecom-genainet



6G AI 和通信的性能要求和评估方法

张公正 [1]，王坚 [1]，李榕 [2]，陈雁 [3]，邵家枫 [3]，林辉 [3]，王俊 [1]，马江镭 [4]，朱佩英 [4]

[1] 无线技术实验室

[2] 法国先进无线技术实验室

[3] 无线网络研究部

[4] 渥太华先进无线技术实验室

摘要

国际电信联盟无线电通信部门（International Telecommunication Union – Radiocommunication Sector，ITU-R）已将“AI 和通信”定义为下一代移动通信系统的六大使用场景之一，新场景意味着新能力，亟需对其关键性能指标（Key Performance Indicator，KPI）和最低性能要求进行定义。本文先介绍什么是“AI 和通信”场景，以及针对该场景下一代移动通信系统会提供哪些典型的人工智能（Artificial Intelligence，AI）服务；接着阐述性能定义的一般性原则，并详细给出具体的性能指标与要求。针对这些性能指标，本文还提供了一套评估方法，并具体阐述了评估流程。

1 引言

随着人工智能（Artificial Intelligence，AI）技术，特别是深度学习和大型预训练模型的快速发展，AI 将走进千行百业，成为人们日常生活不可或缺的一部分。广泛部署的移动通信系统或许是 AI 和通信融合的最佳选择。作为统一的基础设施，移动通信系统能为所有联网的人和机器提供无处不在的 AI 服务，这也将推动移动通信系统的变革。

为促进下一代移动通信系统的开发，国际电信联盟无线电通信部门（International Telecommunication Union – Radiocommunication Sector，ITU-R）明确定义了 IMT-2030 的六大典型使用场景。除了增强 IMT-2020 已有的三个场景外，IMT-2030 还纳入了两项超越通信的服务——AI 和感知。这两项服务将由 6G 网络提供，对系统的新能力和性能指标提出了要求，因此需要对相关技术、性能要求和评估方法展开研究。然而，近年来大部分研究工作只聚焦于技术，在性能要求和评估方法层面鲜有涉及。

因此，本文重点研究 6G AI 和通信的性能要求和评估方法，旨在为下一代移动通信系统的设计提供指导，为用户提供有保障的 AI 服务。在接下来的章节中，我们会先介绍 IMT-2030 框架中定义的“AI 和通信”场景，重点关注这一新场景在 6G 中有哪些典型的 AI 服务和能力要求。随后梳理相关性能指标的现状、介绍设计原则和性能要求的定义（这些性能要求分为定性和定量两种）。最后，提供相应的评估方法和示例，并给出总结。

2 AI 和通信

作为下一代移动通信系统，6G 旨在通过提供人工智能即服务（Artificial Intelligence as a Service，AIaaS），实现普惠智能。对于无线网络中广泛分布的 AI 大模型而言，训练将会更简单、分发更迅速、推理更精准。利用分布式智能终端提供的数据和资源，6G 可提供 AI 模型训练服务，其逻辑是先在分布式终端上进行本地训练，然后终端之间通过网络进行模型交互，这种方式可以有效地保护用户数据隐私。此外，针对资源受限的终端，6G 可以联合调度通信资源和 AI 资源，为其提供高精度推理服务。因此，AIaaS 成为 6G 的一个典型应用场景。本节将介绍 ITU 中的标准化进展及典型服务。

2.1 IMT-2030 框架中的 AI 和通信

为促进 IMT-2030 的开发，ITU-R 5D 工作组（Working Party 5D，WP 5D）批准了一个新框架和总体目标 [1]，确定了下一代移动通信系统的动机、应用、技术趋势、频谱、使用场景和能力。泛在智能既是重要的应用趋势，也是关键的使能技术——AI 能增强无线接口的性能、实现无线网络自动化和网络服务智能化，IMT 系统设计的一个关键目标是高效支撑无线网络中的 AI 服务。

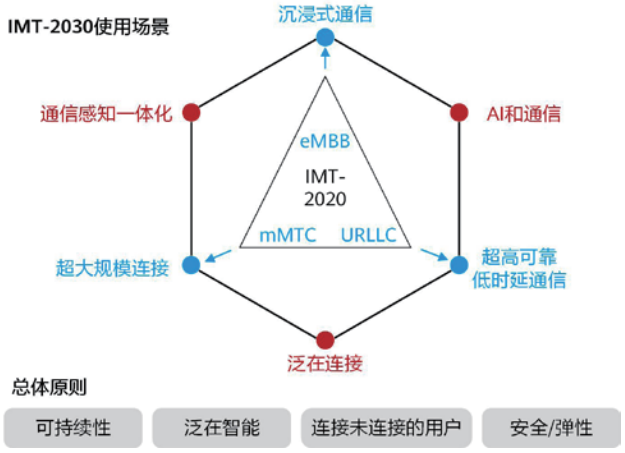


图 1 IMT-2030 使用场景 [1]

在 ITU-R 确定的 IMT-2030 六大使用场景（见图 1）中，“AI 和通信”作为超越通信的服务场景，支持分布式计算和 AI 应用，包括数据采集、本地/分布式计算卸载、分布式 AI 模型训练和推理等。典型用例包括：IMT-2030 辅助的自动驾驶、医疗辅助设备间自主协作、跨设备/网络计算密集型操作的卸载、数字孪生的创建并用于预测。

为了支持新使用场景，IMT-2030 除了传统的通信能力外，还需要新增 AI 能力与感知能力（详见表 1）。对

表 1 IMT-2030 的能力 [1]

增强能力	IMT-2020	IMT-2030
峰值速率（Gbps）	下行：20；上行：10	如 50、100、200
用户体验速率（Mbps）	下行：100；上行：50	如 300、500
频谱效率（bps/Hz）	（峰值）下行：30；上行：15	如 x1.5、x3
区域通信容量（$Mbps/m^2$）	10	如 30、50
连接密度（每平方公里设备数）	10^6	10^6 ~ 10^8
移动性（km/h）	500	500 ~ 1000
时延（ms）	1	0.1 ~ 1
可靠性	$1 - 10^{-5}$	$1 - 10^{-5}$ ~ $1 - 10^{-7}$
IMT-2030 中新增的能力		**取值**
覆盖		***待定***
感知		***待定***
AI		***待定***
可持续性		***待定***
定位精度（cm）		1 ~ 10

于“AI 和通信”这一场景，在通信方面，“AI 和通信”需要更高的区域通信容量和用户体验速率，同时要求低时延和高可靠性，具体要求视实际情况而定。在 AI 方面，IMT-2030 也需要具备一些新能力，包括从不同来源获取、准备并处理数据，进行分布式 AI 模型训练，在不同 IMT 系统之间共享模型并进行分布式推理，以及编排和链接计算资源等。下面将结合典型的 AI 服务来介绍相关的 AI 能力和性能要求。

2.2 “AI 和通信”场景中的典型服务

IMT-2030 将为 AI 应用提供端到端的高效支撑，通过在分布式智能之间建立连接，提供泛在 AI 服务（如 AI 模型训练、推理、部署等）。为实现这一目标，IMT-2030 可利用网络中的连接、数据和模型资源与能力，构建一个分布式高效 AI 服务平台。AI 应用范围非常广泛，既可以使用端到端 AI 算法为无线接口和无线网络进行定制化调优和自动化运维（即，提供智能化网络调优能力），也可以作为分布式学习的基础设施，利用网络原生的通信 AI 一体化能力，实现从云端集中式智能到边缘深度泛在智能的演进（即，提供智能化用户服务能力）。

2.2.1 IMT-2030 的 AI 应用示例

业界普遍认为，协作机器人将是未来 6G 的一个重要应用场景，但这种机器人需要依赖时延低、学习/推理精度高的 AI 服务才能工作。在该示例应用中，多个机器人共同协作完成工业环境中的复杂任务，每个机器人都配备了摄像头等传感器，借助 AI 实现部分自主。对于完全自主或复杂任务，协作机器人系统应通过感知、认知、规划和控制来实现任务的最终目标。例如，当人类用语音指示机器人去拿某个物品时，首先要能理解自然语言指令，然后规划每个机器人负责的子任务。无论是理解还是规划，都离不开高效训练的大（语言）模型，而这些模型会消耗大量的计算资源和内存资源。借助本地视觉或控制模型，机器人能从感知的图像中检测到物体，并规划子任务的路径轨迹和相应的控制决策。

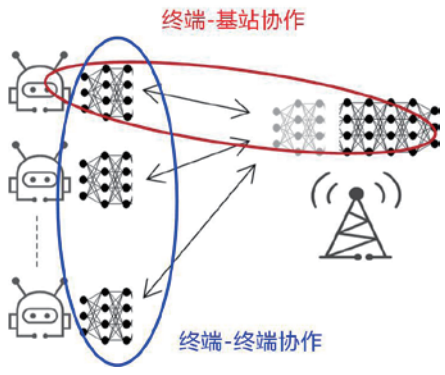


图 2 AI 应用之协作机器人

这样，AI 机器人就能与网络协作，利用网络提供的超级 AI 能力，实现复杂任务规划。机器人还可以基于网络相互合作，通过协作训练、分享和学习彼此的经验提升本地模型的性能。接下来，我们将详细介绍本例中涉及的两个典型 AI 服务——模型推理和模型训练。

2.2.2 模型推理服务

AI 模型推理是 AI 应用的一项基础功能。它会根据给定输入运行 AI 模型，并产生预期输出。通过泛在连接，6G 网络原生智能可以提供模型实时推理能力，以满足不同需求。在分布式 AI 模型推理服务中，6G 网络利用通信能力与 AI 能力，通过模型协作为用户提供实时高精度的模型推理服务，弥补用户的能力短板。在图 3 所示的典型 AI 模型推理服务中，我们可以将一个大模型拆成两部分，分别部署在网络侧和用户侧，分工合作。其中，对资源要求较高的那部分部署在网络侧，以发挥强大的网络 AI 能力，为用户提供模型联合推理服务。

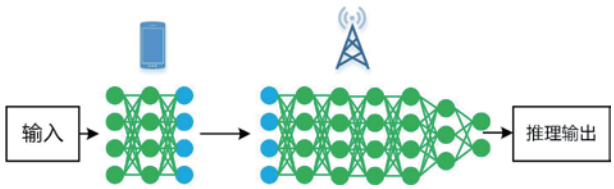


图 3 AI 模型推理服务

2.2.3 模型训练服务

AI 模型训练对获取高精度模型至关重要。6G 网络原生智能可以根据不同的用户特性和网络特性，提供合适的算法和资源进行模型训练编排，从而提高模型训练的速度和精度。在大规模分布式 AI 模型训练服务中，网络作为管理平台，提供高速数据通道和高效调度机制，方便分布式终端交换数据或模型参数。这样既能支持模型的快速聚合和分发，又能保护用户隐私。图 4 展示了一个典型的分布式模型训练服务，

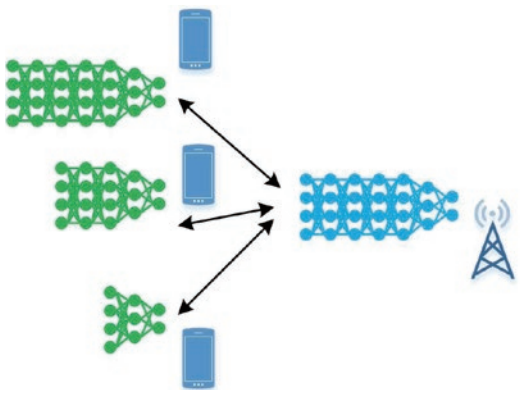


图 4 分布式 AI 模型训练服务

在每轮训练中，分布式终端先基于本地数据在本地训练模型，然后将更新后的模型上传到网络；网络将终端更新的本地模型聚合成一个全新的全局模型，再分发给各个终端。这一“聚合 + 分发”的过程不断迭代，在实现联合学习的同时，也保护了用户的原始数据。

3 “AI 和通信”场景的性能要求

性能要求是系统设计的主要驱动力，也是每一代移动通信系统发展乃至革新的根本所在。现有的移动通信系统主要为面向连接的数据传输而设计，因此其关键性能指标（Key Performance Indicator，KPI）主要是连接的传输速率和时延。然而，AI 服务不仅仅是传输，还涉及 AI 相关的资源，这就衍生出新的 KPI，即 AI 模型学习/推理的精度和时延。从通信角度来看，6G 网络应提供很高的通信容量（尤其是上行方向），以满足模型训练和模型推理过程中数据/模型交互的需求。而从 AI 角度来看，6G 网络应支持大规模分布式学习和实时推理。因此，6G 网络设计从一开始就应该综合考虑 AI 和通信两个方面。以下小节将分析现状，并为“AI 和通信”场景定义具体的性能要求，详述其原则和架构。

3.1 现状

从 2G 到 5G，前几代移动通信系统均聚焦于提供通信服务，数据传输几乎是其唯一的任务。5G R18 开始研究系统对 AI/机器学习（Machine Learning，ML）操作的支持。3GPP TR 22.874 [2] 识别并报告了三种典型的 AI/ML 操作，分别是分割推理、模型/数据分发和分布式/联邦学习。此外，还定义了各种应用，如图像识别、实时媒体编辑、机器人之间的分割推理与控制、多智能体之间的协作学习等。所有 AI/ML 操作预计都需要在云服务器中执行，而 5G 系统仍然只提供用户和云服务器之间的数据传输（即通信服务），这对数据速率提出了更高的要求。

下一代移动通信系统将引入超越通信的新能力（如 AI 相关的能力），因此 6G 相关的研究小组通常会考虑支持 AI 服务。譬如，中国的 IMT-2030 推进组和欧洲的 Hexa-X 不约而同地在其白皮书 [3, 4] 中指出，6G 提供的 AI 服务将是下一代网络设计的关键因素，同时还建议纳入新能力，以确定 AI 服务的性能要求。针对 AI 空口和 AI 服务，这两大 6G 研究组织在传统通信性能指标之外，还提出了 AI 相关的性能指标，包括 AI 模型推理精度和时延。然而，他们并未清晰地论证些性能指标，也没有详细定义 6G 的性能要求和评估方法。

为评估 AI 软硬件系统的能力，计算机科学界定义了一些训练和推理方面的 KPI。例如，MLCommons 定义了 MLPerf 基准 [5]，通过参考应用、模型和数据集来衡量模型训练和推理的 AI 性能。但是，这些指标只能用来度量集中部署的软硬件能力，而 6G 网络中的 AI 服务是分布式部署的，且涉及通信，需要新的指标来度量。

3.2 6G AI 和通信的性能定义原则

利用 6G 网络中的连接、模型、数据等资源，6G AIaaS 能提供适配不同应用场景的 AI 能力。与传统移动网络不同，6G 网络不仅需要连接，还需要其他资源，以便为用户提供高性能的 AI 服务。因此，6G AIaaS 需要实现通信与 AI 能力的融合，为 AI 服务构建全面的性能指标和评估方法，这对于 6G 网络设计和网络资源配置具有指导意义。

AI 相关能力的性能定义主要遵循以下原则：

- **端到端 AI 能力：**为保障用户体验到的服务质量，AI 服务需要端到端的性能指标。AI 服务质量不仅依赖于通信能力，也依赖于 AI 能力。然而，现有的性能指标和评估方法只关注通信能力，无法保证 AI 服务质量。因此，IMT-2030 系统需要考虑如何实现通信与 AI 能力的融合。
- **典型服务：**IMT-2030 系统是实现泛在智能的关键。需要利用网络内部的 AI 能力并通过协作，为不同用户提供统一平台，支撑大规模的分布式模型训练和统一的高精度模型推理。因此，AI 相关能力的指标需要结合训练、推理等典型服务来定义。
- **核心性能：**通信 AI 一体化的宗旨是高效使能 AI 服务，包括模型训练和实时高精度模型推理。同时，AIaaS 面向数十亿用户，需要特别关注影响用户体验的关键因素。围绕 AI 服务有不少性能指标，而要实现“高效使能 AI 服务”的宗旨，IMT-2030 系统必须优先考虑最核心的几个指标。

3.3 6G AI 和通信的性能要求

AI 和通信的 KPI 是根据 6G 网络提供的服务（包括 AI 模型训练和推理）来定义的，这些服务的性能取决于 AI 模型能力（受系统 AI 资源限制）以及用户与网络间的通信能力。本文提出的 AI 服务性能要求包括一组功能要求和三个量化要求。其中，功能要求可通过检查来评估，而量化要求则需要通过仿真来评估，具体如下：

- **AI 服务功能要求**

  AI 相关能力的功能要求是：无线接口技术（Radio Interface Technology，RIT）或成套无线接口技术（Set of Radio Interface Technologies，SRIT）需要为功能提供相应的机制和/或信令，包括作为能力开放给外部应用的功能（如分布式数据处理、分布式学习、AI 计算、AI 模型执行、AI 模型推理等），或是在候选 RIT/SRIT 提出者看来能更好地支撑 AI 应用的功能。

- **AI 服务精度（或 AI 服务质量）**

  AI 服务精度即 AI 推理/学习服务的精度，是指在给定的时延要求内，AI 服务的输出与给定输入对应的真值一致的程度（或与参考情况的相对程度）。对于给定的 AI 任务，AI 服务精度主要取决于任务特性、AI 模型部署方式以及 AI 相关数据的传输。不同应用可能对 AI 服务精度有不同的要求，譬如，自动驾驶对物体识别的精度要求远高于普通消费者拍照识花的要求。因此，可以为 6G 网络中有一定精度要求的具体应用定义最低性能要求。给定一个 AI 推理/学习任务，最低精度要求可以定义为部署环境中该任务在特定时间内能达到的精度需要高于的值。若能满足最低要求，则说明 6G 网络可以支持所有精度要求更低的应用。

- **AI 服务时延**

  AI 服务时延是指 AI 推理/学习服务从开始到结束所需的时间，包括 AI 相关数据的传输时间与 AI 模型处理时间，其中处理时间的长短取决于实际设备和实现方式。与 AI 服务精度类似，不同应用对 AI 服务的时延要求可能也有所不同。因此，我们可以为 6G 网络中有一定时延要求的具体应用定义最低性能要求。给定一个 AI 推理/学习任务，最低时延要求可以定义为部署环境中该任务达成特定精度要求所花的时间需要小于的值。若能满足最低要求，则说明 6G 网络可以支持所有时延要求更低的应用。

- **AI 服务密度**

  AI 服务密度是指在单位面积内，网络中有多少 AI 服务能同时满足特定的精度要求和时延要求。IMT-2030 系统将 AI 服务密度作为系统容量指标，在不同的应用要求（即精度或时延）下，系统可以支持不同的 AI 服务密度。这意味着，6G 网络的最低性能要求既可以针对具有特定精度和时延要求的具体应用或应用的组合来定义。因此，对于给定的 AI 推理/学习任务，可以将 6G 网络 AI 服务密度的最低要求定义为部署环境中每平方公里的服务数量。

4 评估方法与示例

上一节定义了典型分布式 AI 模型训练和推理服务的量化性能要求，由于服务性能由通信和 AI 资源共同决定，评估时需要对通信和 AI 做一些假设。本节将介绍一套性能评估方法，并在示例中给出详细的假设和结果。

4.1 评估方法

性能要求可以从两大基本 KPI 得出，即 AI 服务精度和 AI 服务时延。AI 服务精度反映的是 AI 服务的输出与给定输入对应的真值一致的程度，具体取决于 AI 模型以及 AI 相关的数据/模型传输。AI 服务时延是 AI 模型处理时间与数据传输时间之和，同样也取决于 AI 模型以及 AI 相关的数据/模型传输。这两个指标的定义对于模型训练和模型推理服务同等适用，二者使用的无线资源类似，只是前者涉及模型交互，后者涉及数据交互。

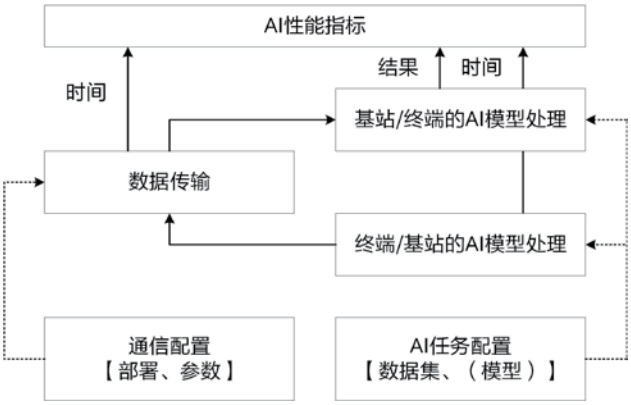


图 5 AI 服务性能评估体系

性能评估可按服务流程进行，图 5 所示的 AI 服务性能评估体系包括以下几个关键部分：

- **资源假设：**通信性能评估中已经定义了一套测试环境，AI 的评估也应在类似环境中进行 [6]。该环境需要配置无线资源，包括带宽、终端（UE）天线数量和基站天线数量等。同时，AI 任务应使用 AI 相关配置来定义，包括由输入和对应目标输出组成的数据集以及精度计算方法。
- **AI 服务流程：**整个流程从终端侧的 AI 模型处理开始，在终端侧生成中间数据（模型输出或模型权重）。然后，在假定的无线配置下，数据从终端传往基站。基站收到数据后，再用基站侧 AI 模型进行处理，得到服务结果，并计算性能指标。
- **AI 服务性能计算：**根据服务结果、AI 模型处理时间以及传输时间，可以计算出 AI 服务精度和 AI 服务时延这两大指标。如前所述，AI 服务精度是指 AI 模型处理后的输出与数据集中各输入对应的目标值一致的程度，这个程度需根据具体 AI 任务来定义。AI 服务时延则是终端和基站的 AI 模型处理时间与中间数据的传输时间之和。

AI 服务密度是指满足特定 AI 服务精度和 AI 服务时延要求的 AI 服务数量。因此，评估 AI 服务密度需要仿真 AI 服务精度和 AI 服务时延。例如，先将终端数量 N 设为最小值并由终端发起服务请求，利用测试环境的评估参数进行系统仿真，在服务时延范围内统计 AI 服务精度。然后逐步增大 N 并重复仿真，直至 AI 服务精度不再达标为止，此时的 N 值为 Nmax。然后，就能根据公式 $C = Nmax/覆盖面积$ 计算出 AI 服务密度。

4.2 评估示例

本节以分布式 AI 推理服务为例来论证上述性能评估方法。根据相应的服务流程修改评估流程后（这是未来需要研究的一个方面），该方法也可用于协作训练和推理服务。在典型的未来智能工厂中，AI 机器人需要通过摄像头感知环境（如实时检测物体），基于收集的图像实时进行高精度的 AI 模型推理。

图 6 分布式 AI 推理服务示例

图 6 展示了用户的 AI 推理服务流程，该流程不仅涉及用户侧和网络侧的 AI 模型处理，还包括用户和网络之间的传输过程。具体分为三个步骤：1）终端使用终端侧 AI 模型处理输入的数据，生成中间数据；2）终端将中间数据发往基站；3）基站使用基站侧 AI 模型处理接收到的中间数据，得到推理结果。需要注意的是，本示例流程假设服务是从终端侧开始的，由终端处理输入数据并上传中间数据到基站，从而算出结果。下行方向也可采用类似流程：先由基站处理输入数据，再将中间数据发给终端，从而算出结果。以下评估方法同样适用于这种下行场景：

- **评估配置：**评估配置定义如下，中括号【 】内为示例。
  - 测试环境：【密集城区】
  - 无线配置：【同沉浸式通信（用户体验速率：500 Mbps）】
  - AI 任务：【图像识别】
  - AI 数据集：【ImageNet-1k 验证数据集 [7]】
  - AI 模型：【AlexNet [8]，左边由终端处理，右边由基站处理，如图 7 所示】

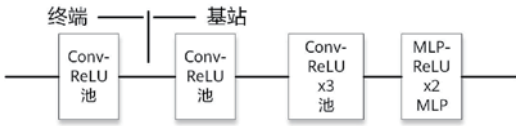


图 7 AlexNet 模型部署示例

  - AI 模型处理时间：【终端侧模型：0.75 ms；基站侧模型：0.45 ms】

- **评估流程**
  - AI 服务精度：可以通过仿真来评估 AI 服务精度。首先，终端利用终端侧 AI 模型对数据集中每个样本 $S_i, i = 1, \ldots, n$ 的输入进行处理，得到中间数据 Z_i。根据测试环境和传输配置，终端发送中间数据到基站处理。以经典传输方案为例，中间数据先量化并转化为比特，再编码、调制成符号进行无线传输。然后，基站再利用基站侧 AI 模型处理接收到的中间数据 $\tilde{Z}_i$，得到每个样本对应的推理结果 $\tilde{Y}_i$。接着，将推理结果与每个样本的目标输出或标签 Y_i 进行比较或计算，从而得到输出与真值一致的程度 $\text{acc} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\tilde{Y}_i == Y_i\}}$，即为 AI 服务精度。

    对于参考情况的精度，可以用整个 AI 模型处理数据集中的每个样本 S_i，得出推理结果 $\tilde{Y}_i'$，即参考情况的输出。然后将推理结果与每个样本的标签 Y_i 进行比较，即参考情况下该输出与真值一致的程度表示为 $\text{acc}_{\text{ref}} = \frac{1}{n}\sum_{i=1}^{n} 1_{\{\tilde{Y}_i' == Y_i\}}$。最后，可以用 $\text{acc}/\text{acc}_{\text{ref}}$ 计算出 AI 服务的相对精度。
  - AI 服务时延：AI 服务时延是中间数据的传输时间 t_{comm} 与终端和基站侧的 AI 模型处理时间 $t_{\text{proc,UE}}, t_{\text{proc,BS}}$ 之和。因此，AI 服务时延可以用 $t_{\text{service}} = t_{\text{comm}} + t_{\text{proc,UE}} + t_{\text{proc,BS}}$ 表示。在本例中，我们通过有效载荷比特数除以数据速率来计算数据传输时间，其中有效载荷比特数由中间数据的元素数量及每元素量化比特数决定。如果有新技术可用，也可使用其他方案。

- **评估结果**

表 2 总结了不同的传输配置（即每元素量化比特数）对 AI 服务性能的影响。从 AI 服务精度和 AI 服务时延的数值变化情况可以看出，由于中间数据传输的限制，AI 服务时延和 AI 服务精度难以同时兼顾——精度越高，时延往往也越大。但我们可以通过优化传输配置或提升传输技术来满足性能要求，以表 2 为例，将每元素比特数设为 8，可保证 AI 服务精度大于 56%、AI 服务时延小于 2 ms。

表 2 AI 服务性能评估结果

每元素比特数	2	4	6	8	10	12	16	32
AI 服务精度（%）	0.14	10.35	52.94	56.47	56.53	56.55	56.55	56.56
AI 服务相对精度（%）	0.24	18.30	93.61	99.84	99.95	99.99	99.99	100
AI 服务时延（ms）	1.4	1.6	1.8	1.9	2.1	2.3	2.7	4.2

5 结语

“AI 和通信”是 IMT-2030 中定义的 6G 新场景之一，本文重点分析了该场景的动机、典型 AI 服务和性能要求。笔者从用户体验和网络容量两个角度出发，针对通信与 AI 能力及资源的一体化定义了全新的性能指标，从而为系统设计提供指导，并能更好地支撑 AI 服务。与此同时，本文还提供了相应的性能评估方法，并结合详细示例展开讲解。以上研究，只是 6G 从愿景走向技术设计的第一步。

参考文献

[1] ITU-R, Recommendation ITU-R M.2160-0, "Framework and overall objectives of the future development of IMT for 2030 and beyond," Nov. 2023.

[2] 3GPP TR 22.874, "Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," Release 18, 2021.

[3] IMT-2030 (6G) Promotion Group, "White paper on typical usage scenarios and key capabilities in 6G," July 2022.

[4] Hexa-X Deliverable D1.3, "Targets and requirements for 6G – Initial E2E architecture," Feb. 2022.

[5] https://mlcommons.org/en/, accessed on Aug. 10, 2022.

[6] ITU-R M.2412-0, "Guidelines for evaluation of radio interface technologies for IMT-2020," 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," IEEE CVPR, 2009.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS, 2012.

AI 原生 6G 网络的数据面设计

严学强[1]，张馨然[2]，王俊凡[1]，张翼[3]
[1] 华为无线技术实验室
[2] 北京邮电大学
[3] 浙江大学

摘要

相较以往移动网络，6G 网络将不仅提供增强的连接服务，还将利用人工智能（Artificial Intelligence，AI）和无线感知能力推动智能应用取得前所未有的发展。通过集成射频感知能力，6G 网络将从物理世界采集大量数据，并将这些数据用于 AI 应用。伴随通信感知一体化（Integrated Sensing and Communication，ISAC）和网络原生 AI 应用而产生的数据需要进行有效管理。为此，我们提出了数据即服务（Data as a Service，DaaS）的概念，并提出在 6G 网络架构中包含一个专用数据面的设计，旨在高效处理分布式感知和 AI 应用中的数据流。该数据面主要由两类组件组成：数据编排网元（Data Orchestration，DO）和数据代理（Data Agent，DA）。数据编排网元负责解读业务需求并将其转化为网络配置，而数据代理负责数据的采集、处理、存储和共享。我们还提出了一个无状态、分布式数据通信代理（Data Communication Proxy，DCP）的设计。该代理利用发布/订阅（Publish/Subscribe，Pub/Sub）这种异步通信模式，实现数据生产者和数据消费者的相互解耦。DCP 为数据面带来了多项关键优势，包括超低时延的数据转发、数据随路处理、以及灵活的数据处理拓扑结构等。

关键词

数据面，DO，数据服务，DCP，AI，ISAC

1 引言

数据在 6G 网络的人工智能（Artificial Intelligence，AI）部署中发挥至关重要的作用。AI 部署需要大量的高质量训练数据和测试数据。基于不完整或不正确的低质量数据训练出来的模型可靠性较差，而此类模型生成的结果质量也较差。然而，海量数据的采集过程往往耗时且费力，这严重阻碍了 AI 应用大规模获取高质量数据。

得益于其固有的射频感知能力，6G 网络将催生大量的数据驱动型应用。这些应用生成、传输、处理和存储的数据量将达到前所未有的规模。充分利用射频感知能力，6G 网络可以将自身转化为一个“传感器”，从物理世界读取数据。因此，6G 感知数据可以成为 AI 模型训练的一个重要数据源，推动生成式和交互式 AI 服务的发展。若将 6G 网络 1% 的能力用于感知，端侧 AI 模型的 6G 感知数据量将达到每天 10^{30} 字节量级，基站上云侧 AI 模型的 6G 感知数据量将达到每天 10^{21} 字节量级。若能有效将这些数据用于 AI 模型训练，模型可以更好地与现实世界对齐，提供定制化的 AI 服务。

网络原生 AI 是 6G 的基本特性之一，其将采用深度边缘架构，助力实现分布式、协作式、广泛的机器学习（Machine Learning，ML）。网络原生 AI 的首要目标是将大量的分布式智能代理进行智能连接，以实现 AI 的规模化部署。智能代理是数据即服务（Data as a Service，DaaS）生态系统中的重要组成，在智能代理之间，模型参数等海量数据需要以高效率、高吞吐量且低时延的方式传输。从 DaaS 的角度，网络必须为机器学习运维（Machine Learning Operations，MLOps）提供灵活支持，具体而言，网络需要自动化并简化 ML 的工作流以及相关部署。

1.1 当前的连接服务网络架构

当前典型的移动网络架构广泛采用控制面和用户面设计为移动用户提供无线接入服务。控制面处理连接建立和用户面数据转发控制等任务。在 5G 核心网领域，此类任务统称为会话管理。用户面的主要作用是将网络数据包高效转发至对应的目标节点。

1.1.1 控制面：基于服务的接口

图 1 展示了 3GPP 基于服务的架构（Service-based Architecture，SBA）。在此架构中，5G 网络的控制面功能和公共数据存储库分布在相互连接的网络功能（Network Function，NF）中间。每个 NF 都有权限访问其他 NF 提供的服务 [1]。在 SBA 架构中，5G 核心网控制面的 NF 之间

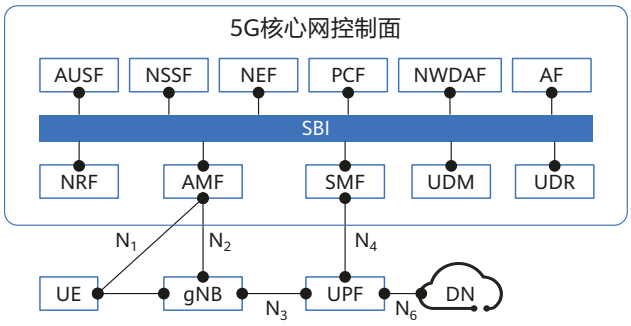


图 1 5G 核心网的 SBA 架构

有一个基于服务的接口（Service-based Interface，SBI）。该接口作为通信总线，使用 HTTP/2 协议在控制面所有 NF 之间进行通信。

5G 网络中的 HTTP/2 连接是点对点的，即消息在 NF 之间双向流动。一个 NF 向另一 NF 发送请求消息前，会先与其建立连接，而对端 NF 也通过该连接返回响应消息。若对端 NF 需要向源端 NF 发送消息，其需要另行建立一条反向的连接。这些信令消息由预先设定规模的小数据包组成，因此，整体数据流量与活动用户设备（User Equipment，UE）数量成正比。

1.1.2 用户面：GTP-U 隧道

如图 2 所示，用户面上的数据流量由横跨 UE、无线接入网（Radio Access Network，RAN）和核心网用户面功能（User Plane Function，UPF）的协议数据单元（Protocol Data Unit，PDU）会话承载。为保证数据传输质量，数据包首先会匹配对应的服务质量（Quality of Service，QoS）业务流，然后通过 UE 和 RAN 节点之间的数据无线承载（Data Radio Bearer，DRB）通道传输，再通过 RAN 节点和核心网 UPF 之间的 N3 GTP-U 隧道传输。此范式中，数据在专用的隧道中传输，因此，此范式是基于会话的。

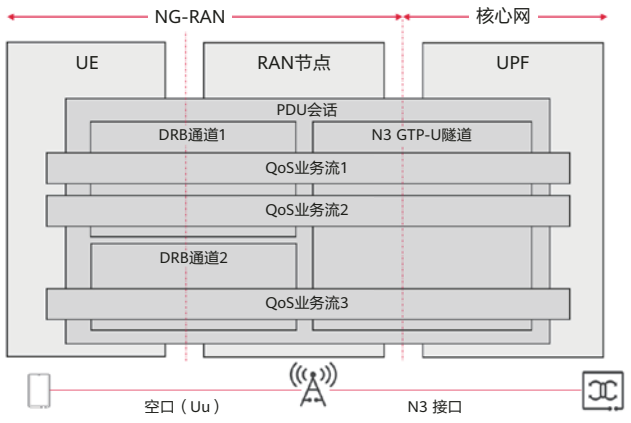


图 2 5G 网络中的 PDU 会话

总而言之，当前的控制面和用户面均运行在基于通信会话的架构之中。信令流程和用户数据包与对应的 UE 之间存在内在关联。控制面为每个会话建立用户面数据通道，实现 UE、基站、核心网功能等网络节点之间的信令消息交互，以保持状态同步。

1.2 挑战

6G 网络的主要功能是为 UE 提供连接服务。此外，6G 网络还将扩展提供 AI 和感知类服务，统称为“超越连接的服务”。“超越连接的服务”与传统的连接服务有明显差异，这也迫切需要网络架构从“以连接为中心”向“以数据为中心”转变。而控制面和用户面的架构无法满足新的服务需求，具体原因如下：

- **海量数据：**6G 网络处理的数据量将达到前所未有的规模，其中的很大部分数据量将由“超越连接的服务”产生。6G 网络的实际数据量将远超 ITU-R 当前的预测，因为该预测仅考虑了移动用户的数据流量 [2]。6G 网络的泛在感知能力将催生海量数据，这些数据也将被用于构建物理世界数字孪生的算法。同时，具备分布式智能能力的 UE、基站、核心网和云等将产生 AI 数据，如梯度、参数、模型、词元（Token）等，此类数据是另一类大规模的非连接数据。如此大规模的数据量是无法由控制面（如 SBI 接口）处理的，因为控制面的特点是小数据包的可靠传输和短时连接。
- **复杂的数据拓扑：**在新兴的 6G 服务中，数据生产者和数据消费者之间的数据拓扑可能相当复杂，其中的 RAN 节点、UE 和 NF 之间并非以线性的、点对点方式连接。与面向 UE 提供的连接服务不同，“超越连接的服务”可以触达任何通过身份验证和授权的网络接入用户。而当前的用户面（如 PDU 会话）仅在 UE 和 UPF 之间建立一对一通信通道，这种面向连接的范式无法处理 6G“超越连接的服务”中的复杂数据拓扑。
- **数据编排需求：**与无线接入模式不同，AI/感知服务是按需运行的，即不同的 AI/感知服务需要不同的 UE、基站和 NF 组合协作才能实现。因此，具备感知能力/算力的各类 UE、基站和 NF 需要进行统筹编排以提供服务，比如分布式学习服务。这就需要网络能够自动解读服务需求并将其转化为网络配置。而这是当前网络架构无法满足的。

自 5G 时代起，智能连接设备生态系统不断扩大，需要实时处理的数据量激增。而传统的、以会话为中心的模式高度依赖集中式代理，无法满足大规模的分布式 AI 系统和协作感知网络中的通信流量管理需求。此外，传统模式在满足扩展性要求、以及端边云统一计算中的低时延和隐私性要求方面也面临巨大挑战。因此，迫切需要定制化设计一个创新的网络架构，以满足 AI 原生 6G 网络中的数据管理和处理需求。

2 DaaS 数据面

受核心网控制面解耦和用户面设计的启发，我们提出在 6G 网络架构中设计一个专用的数据面，以提高网络的灵活性和运行效率。数据面能够助力各类拓扑中的数据流动，从而灵活提供数据服务，实现高效的随路数据处理。具体而言，数据面将服务需求分解为数据处理功能，如数据收集、预处理和分析，并统筹相关网络组件协作以满足服务需求 [3]。数据面传输非连接数据，其主要目标是简化管理和控制。非连接数据将被注入 AI 算法中，助力算法迭代优化。数据面是为内外部应用和服务提供数据服务的基础。数据面主要由两类组件组成：数据编排网元（Data Orchestration，DO）和数据代理（Data Agent，DA）。

2.1 DO

在连接服务中，网络只需在 UE 和 UPF 之间建立通道，而非连接服务的发放和交付需要多个网元协同。这些网元必须进行恰当的编排。如图 3 所示，DO 是数据面中的“指挥官”，监管网元间的数据流动。DO 解读服务需求并将其转换为网络配置，以此确保数据在数据面的精确流动——在正确的时间到达正确的处理单元/应用。

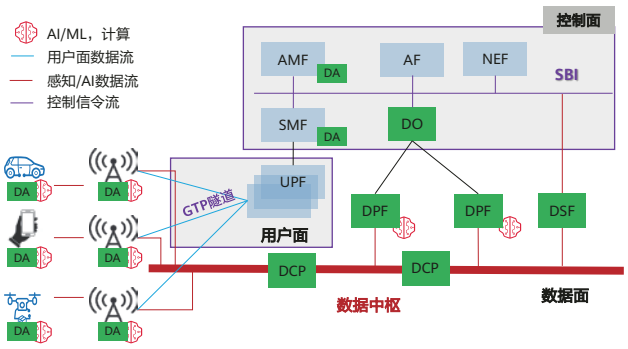


图 3 6G 数据面架构图

DO 包括四个基本组件：数据代理注册、数据流引擎、数据流管理和策略控制。数据代理注册组件负责记录 DO 管理的各 DA 具备的能力，基于此信息，数据代理注册组件可以保障资源的有效利用率。数据流引擎组件负责基于服务需求确定需要执行的操作和数据转换（如过滤、聚合和转发）的顺序，确保数据在系统中有效流动。数据流管理组件负责监管数据包在编排好的数据拓扑中的实际路由情况，其会综合考虑网络拥塞、时延要求和资源可用性等因素。策略控制组件负责数据安全、访问控制和隐私相关的策略决策，确保敏感数据的妥善处理。

2.2 DA

DA 在整个数据服务生命周期中充当“智能中介”的角色，可以提供数据预处理、数据转换、本地决策、数据缓存和安全实施等多种功能。具体而言，DA 可以对原始数据 —— 如通信感知一体化（Integrated Sensing and Communication，ISAC）数据的同相正交（In-Phase and Quadrature，I/Q）信号 —— 进行过滤、聚合和压缩，以降低网络流量和数据处理开销。DA 可以将感知数据转换为适合网络原生 AI 算法/应用的格式，方便数据传输。在分布式数据处理场景中，DA 可以根据本地收集的数据进行基本分析和决策，无需与中心实体持续通信。此外，DA 可以对频繁的数据访问和过程处理结果实施数据缓存，从而加快检索速度，提高处理效率。DA 还可以执行访问控制机制和加密协议，强化数据安全。

DA 可以部署在多类网络节点上，如 UE、基站、网络边缘节点、核心网节点等。部署在 UE 或基站时，DA 可以在数据传输前执行初步的数据获取和本地处理，如数据过滤、压缩等。部署在边缘计算节点时，DA 可以对数据进行进一步处理和分析，并可能触发实时操作，或将处理/分析结果返回支持 AI 算法的下游 DA 等。此外，部署在核心网的集中式 DA 还可能处理更复杂的任务，如聚合多个数据源的数据或将数据馈送至集中式 AI 算法等。

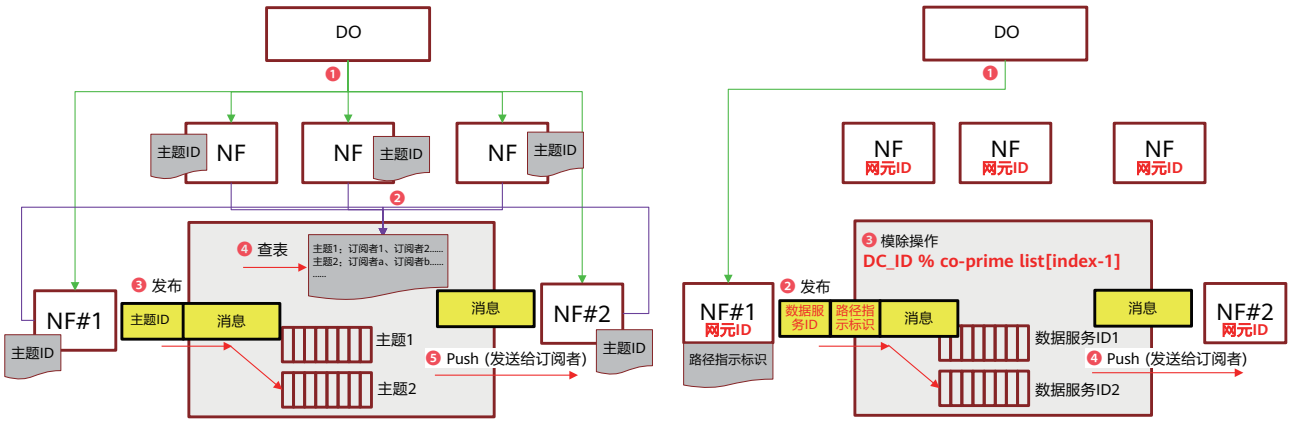
基于不同的实现方式，DA 可以分为两类：嵌入式部署 DA 和独立部署 DA。嵌入式部署 DA 作为网络实体的一部分，一般负责数据采集和预处理。独立部署 DA 依照 DO 的指令承担相应角色。数据处理功能（Data Processing Function，DPF）和数据存储功能（Data Storage Function，DSF）是可以部署在 RAN 或核心网中的两种独立部署 DA。DPF 负责 ISAC 和 AI 等数据的处理，而 DSF 是数据长期存储的仓库。

2.3 DCP

DO 和 DA 的设计带来了另一个挑战，即，传统的数据流模式不再能满足新型数据拓扑的需求。一般情况下，在具备网络原生 AI 或无线感知能力的“超越连接的服务”中，数据拓扑会涉及多个数据源、数据处理功能和数据宿。也就是说，由于 AI 和感知能力，移动网络将变成一个数据密集的计算和大数据分析平台。以会话为中心的数据传输模式无法满足该平台“多进多出”的数据传输需求。鉴于此，我们提出一个面向 6G 网络的全新逻辑功能网元，称为数据通信代理（Data Communication Proxy，DCP）。该网元处理 6G 网络中拓扑复杂、且需随路处理的海量感知和 AI 数据。如图 4 所示，DCP 克服了传统的“先连接再通信”模式的缺点，可以高效进行数据分发。

在当前网络架构中，数据经由 UPF 流至对应的 NF。而在我们提出的架构中，由 DCP 担当数据收集中介。DCP 接收数据，并将数据高效分发至对应的 NF。通过数据生产者和数据消费者之间的解耦，DCP 实现了异步数据交换，即数据生产者可以独立进行数据处理，而不必依赖消费者响应。DCP 可以部署为一个分布式系统，由一个消息代理和多个消息队列组成。消息代理作为数据发送者和接收者之间的中介，基于消息主题，实施从数据发布者到订阅者的消息路由。消息队列是一种数据结构（或容器），通过参与消息发送、接收和存储，协助应用间的通信。同时，采用发布/订阅（Publish/Subscribe，Pub/Sub）通信模式作为异步消息传输架构。在此架构中，消息发送者和接收者无需了解彼此的身份便可进行消息传输。数据生产者向消息代理发布特定主题（即数据类别）的消息，而订阅者向消息代理订阅自己感兴趣的消息主题。

DCP 有两种实现方式：有状态 DCP 和无状态 DCP。有状态 DCP 将订阅信息和主题信息存储在一张表格中，消息



（a）有状态 DCP　　（b）无状态 DCP

图 4 DCP 实现方案

路由基于查表进行。无状态 DCP 在本地不存储订阅信息和主题信息，消息路由基于消息内携带的信息进行。该信息是由 DO 编排和配置的数据拓扑的一种数字表征。在我们的方案中（详见第 3 节），数据转发通过模除操作实现，其中数据拓扑的数字表征是模除操作中的输入。与有状态 DCP 相比，无状态 DCP 具有超低时延的数据转发、高扩展性等优势。超低时延是通过基于模除的数据转发实现的，此数据转发方式的效率远超有状态 DCP 中的查表方式。高扩展性是通过免除了数据业务发放过程中 DO 和 NF 之间大量的配置交互而实现的。在无状态 DCP 中，DO 在配置消息内携带的数据拓扑的数字表征时，只需与源数据节点通信。

DCP 可以部署为独立的 NF，也可以部署在已有 NF 上。图 5 为一个计算机视觉任务的示例，该任务涉及基于人脸检测的年龄和性别分类 AI 模型，也涉及基于犬类检测的品种预测 AI 模型。在此例中，DO 接收服务请求，并根据注册 DA/DPF 的能力构建数据处理拓扑。与 DPF 一样，数据源也作为 DA，被分配一个唯一的主题 ID。在此例中，摄像机即为数据源 DA，其将主题 ID 为“Video”的视频帧数据发布至 DCP。其后，DCP 将视频帧数据推送至消息队列，再推送至订阅了“Video”主题的数据消费者。在此例中，数据消费者是两个 DA/DPF，其一配备了人脸检测模型，另一个配备了犬类检测模型。两个 DA/DPF 进行推理后，向 DCP 发布主题 ID 分别为“Human face”和“Dog”的新主题。DCP 再将年龄和性别检测结果推送至 DA/DPF。在此过程中，数据收集与模型推理是异步进行的，这样，模型推理期间便可以收集到更多的感知数据。

这种基于 DCP 的 ML 模型编排的优势在于 DCP 允许任意数量的消息队列与消息代理绑定。数据可以跨 ML 应用重用，一个模型/训练周期的输出也可以作为下一模型/训练周期的输入。具体而言，DCP 方案具备以下优势：

- DCP 在数据生产者和数据消费者之间引入了缓冲，这可以确保无缝、高效的数据流动，从而避免潜在的数据传输瓶颈。这一优势在处理海量数据时尤为重要，因为数据流的任何延迟或中断都会极大影响 ML 模型的性能。
- DCP 可以适应不同的处理速度，在允许所有 DPF 以各自速度运行的同时，又能确保数据不丢失。
- 与 DO 配合，DCP 可以使任务执行更有序，这也是 ML 模型成功完成推理的一个基本要素。

3 基于无状态 Pub/Sub 范式的分布式消息代理

3.1 Pub/Sub 范式

随着 AI 应用（尤其是采用高级神经网络架构的 AI 应用）的出现，迫切需要一种在 6G 网络内计算中实现 AI 流程编排和调度的新方法。因此我们提出了一个无状态 Pub/Sub 范式，旨在有效管理大规模分布式 AI 系统上从数据源到订阅者的信息流动。

此 Pub/Sub 范式将通信端点相互解耦，并支持分布式学习、推理和协同无线感知等场景中的多对多数据分发。此范式的独特性在于其高效的数据流管理，这使其支持涉及多个数据生产者和消费者的应用，甚至可能可以满足 ML 工作流的管理需求，从而大幅增强网络原生 AI 和 MLOps 能力。此外，Pub/Sub 范式可以使 ML 工作流配置更为灵活，尤其是在响应实时数据或动态变化方面。在传统的消息控制流中，业务流是确定性的、预先定义的。而 Pub/Sub 范式是

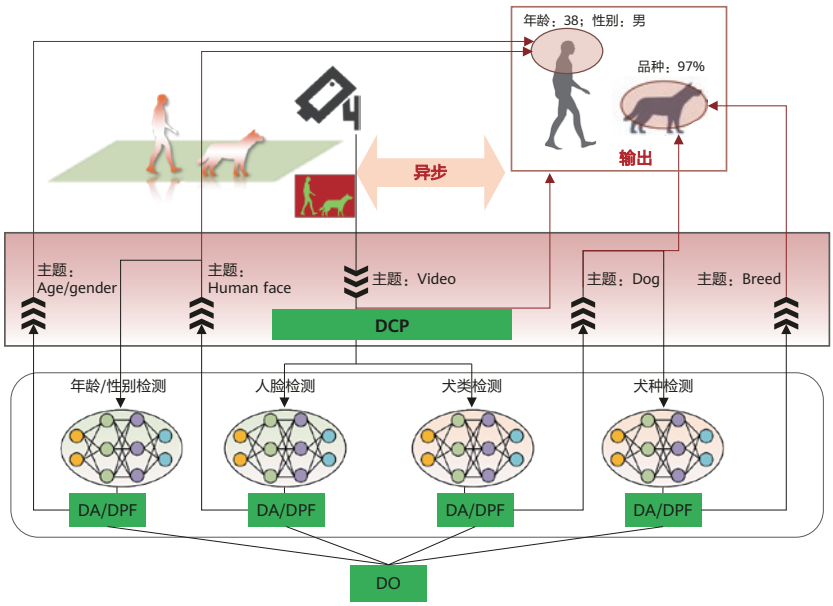


图 5 基于 DCP 的 ML 模型编排

一种更加积极的事件驱动过程，其中数据管道上的各个节点均可作为订阅者，接收其他节点或数据发布者发布的特定主题消息。Pub/Sub 范式中数据生产者和消费者的解耦提升了系统的扩展性和容错性，因为某个组件的故障不会直接影响其他组件。

3.2 基于 CRT 的无状态消息代理

“超越连接的服务”以数据驱动，其中数据的隐藏价值需要采用基于统计或 ML 的方法进行挖掘。鉴于数据拓扑中的所有节点均需处理数据并向下游节点发送数据，我们提出了一种用于非连接数据的随路数据处理方法。此方法有助于减少网络带宽，并保障数据的私密性。

传统的有状态消息代理会维护一个数据转发表，其中包含每个主题所有订阅者的目标地址。此类消息代理在每个消息到达时都需要查找数据转发表，这会额外增加数据传输的时延。另外，此模式的可扩展性非常有限，因为消息代理的所有节点都必须维护数据转发表。

为克服上述缺点，我们提出将消息代理“去状态化”，即采用无状态消息代理的模式。其中一种方案是在消息数据中携带一个编码的“状态”作为带内信息。这样，无状态消息代理的设计目标则是找到一个能将数据拓扑打包或转换为适合封装于数据包中的数字表征的函数，同时，其逆函数的计算必须足够简单。综合以上考虑，我们基于余数系统（Residue Number System，RNS）和中国剩余定理（Chinese Remainder Theorem，CRT）设计了一个函数 f 及其逆函数 f^{-1}。根据 RNS 文献的惯例，我们将 f 称为反向转换器，将 f^{-1} 称为正向转换器。

- **RNS**

RNS 是一种表征数值以执行快速运算的方法。它将一个整数表征为使用模除而得到的余数。RNS 运算以数字为单元进行，其中无进位操作，这样，一个较大数字则可以分解为多个较小数字做并行运算 [4]，从而提高运算效率。如图 6 右侧所示，正向转换器基于一个模数组 $M = (m_1, m_2, m_3)$ 将一个较大整数转换为一组较小的整数余数 (x_1, x_2, x_3)。其中，这个模数组中的数字两两互质。将数字 X 转换为 RNS 表征的过程非常简单——只需用每个模数对 X 进行模除，得到的余数集就是 X 的 RNS 表征。公式如下：

$$x_i = X \ mod \ m_i \tag{1}$$

例如，若 $X = 11$，模数组为 $(3, 4, 5)$，则 11 的 RNS 表征为 $(2, 3, 1)$，公式如下：

$$11 \equiv (2, 3, 1)_{RNS(3,4,5)} \tag{2}$$

- **CRT**

CRT 可以用作反向转换器。利用 CRT，如果我们知道一个整数 n 被若干互质整数除后的余数，那么我们就可以唯一确定 n（如图 6 左侧所示）。公式如下：

$$P = \prod_{i=1}^{3} m_i = m_1 m_2 m_3 = 3 * 4 * 5 = 60, \tag{3}$$

$$M_i = \frac{P}{m_i}, i = 1, 2, 3, \tag{4}$$

$$CRT\ X = \left| \sum_{i=1}^{3} x_i M_i \left| M_i^{-1} \right|_{m_i} \right|_P \tag{5}$$

在以上公式中，$\left| M_i^{-1} \right|_{m_i}$ 是 M_i 的逆元。

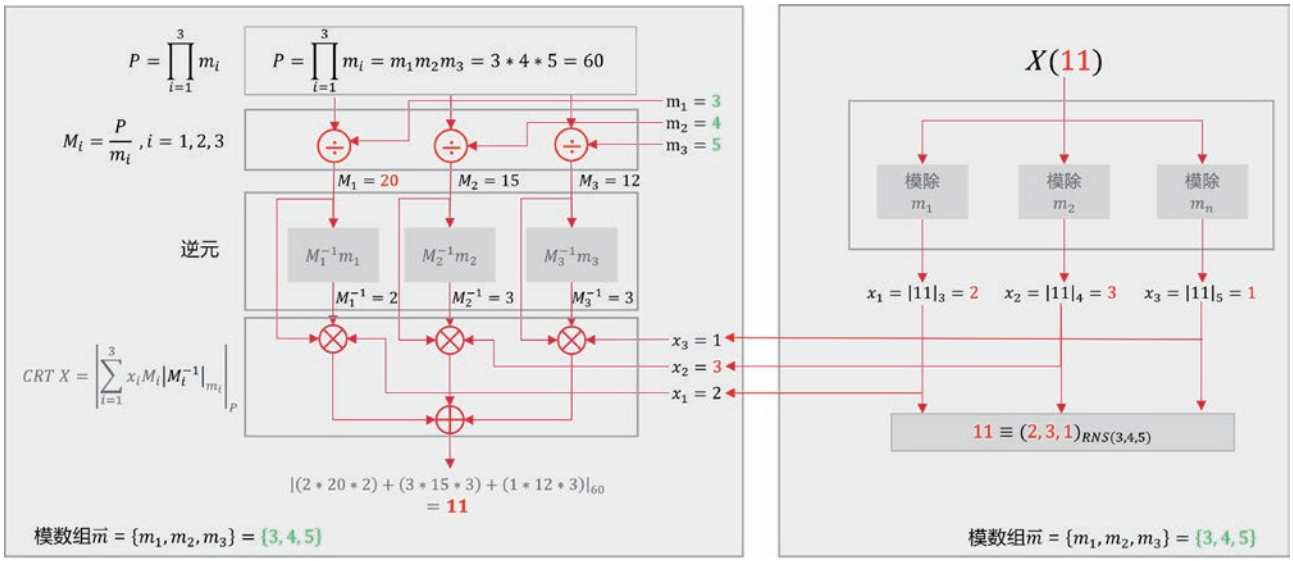


图 6 RNS 和 CRT [4]

在 DCP 中，数据拓扑由包含模数组和余数组的数组（$\vec{m}, \vec{n}$）表示，而数据拓扑的数字表征 X 由消息数据包携带。正向转换耗时较多，且可以提前进行，因此，可将其作为数据编排过程的一部分，用于数据转发，其中，余数组表示数据拓扑中的下游数据处理节点。

3.3 性能评估

基于无状态 Pub/Sub 范式的分布式消息代理的主要设计目标是最小化消息转发时延，这对于协作感知和基于 AI/ML 的自动驾驶等“超越连接的服务”至关重要。这里，我们将消息转发时延定义为输入消息的第一个比特进入队列到其最后一个比特退出队列之间的时间间隔。端到端时延受多种因素影响，包括消息大小、消息到达率、DCP 的处理能力等。

表 1 DCP 的实验环境设置

项目	详细说明
操作系统	Ubuntu 22.04.4 LTS（Jammy Jellyfish）
内核版本	5.15.0-113-generic
CPU	双插槽架构，每个插槽包含 40 个物理核心，每个核心支持双线程，总共 160 个逻辑处理器，以 2.30 GHz 的基本频率运行。
内存	503 GiB 内存
存储	894.3 GB 固态硬盘 1.8 TB 硬盘
RocketMQ	5.1.4
Java	1.8.0_392
Docker	24.07

我们将 DCP 方案与 RocketMQ [5] 进行了性能对比实验。RocketMQ 是一种广泛用于 IT 领域的分布式消息队列系统，具有大容量、实时性和零错交互的特点。RocketMQ 以单副本模式运行，即系统中只有一个消息代理，没有副本节点。我们对 DCP 也进行了相同的设置，并在相同的硬件和软件环境下（详见表 1）进行了对比实验。

图 7 展示了消息大小为 $10^1 \sim 10^5$ 个字节区间时 DCP 和 RocketMQ 系统的消息转发时延。在测试中，每个系统转发 5 条大小不等的消息，每条消息重复转发 5 次，并计算这 5 次的平均转发时延。如图所示，随着消息大小的增加，RocketMQ 系统的消息转发时延也增加了。当消息大小达到 10^5 字节时，消息转发时延大幅增加。而 DCP 系统的消息转发时延一直稳定在 2 ms 左右。总体上，综合所有消息大小，DCP 系统的平均转发时延较 RocketMQ 系统低 65% 以上，而对于较大消息（如 10^5 字节），DCP 系统的转发时延较 RocketMQ 系统的优势更加明显。

DCP较RocketMQ系统的优势主要源于以下两个因素：

- **无状态设计：**DCP 采用无状态设计，无需为每个消息转发操作维护和管理复杂的状态信息。这种设计减少了处理每个消息的计算开销，可以显著降低消息转发时延。
- **无需消息持久存储：**DCP 应用场景不要求支持消息持久存储，因此我们从系统中移除了消息持久存储相关的组件。这进一步简化了 I/O 操作，降低了数据写入磁盘的开销，提高数据转发效率。相比之下，RocketMQ 系统要求支持消息持久存储，这也导致其处理较大消息的时延显著增加。

我们还测试了 DCP 系统在不同消息发送间隔（2 ms、3 ms、4 ms、5 ms、10 ms、50 ms、100 ms、300 ms、500 ms 和 1,000 ms）情况下不同规模消息的发送成功率。如图 8 所示，当消息大小在 $10^3 \sim 10^5$ 字节之间时，在不同消息发送间隔的情况下，消息发送成功率都接近 100%。这说明 DCP 系统能够有效处理并成功发送较小规模的消息。但是，当消息规模增加到 10^5 字节及以上时，消息发送成功率在所有消息发送间隔的情况下都有所下降。

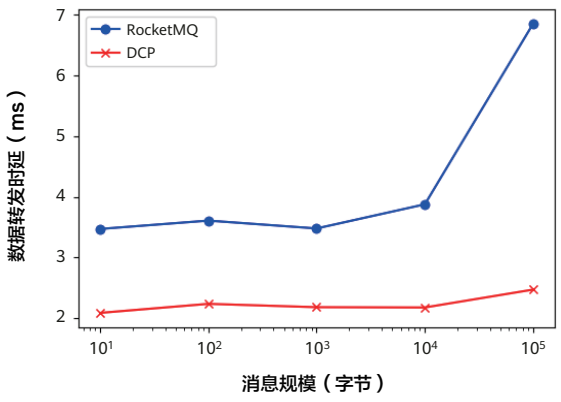


图 7 DCP 和 RocketMQ 系统的数据转发时延

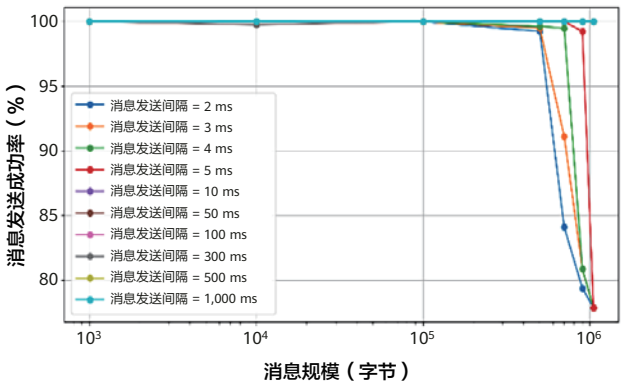


图 8 DCP 系统的消息发送成功率

以上结果表明，对于较小规模的消息，消息发送间隔对发送成功率影响不大；而对于较大规模的消息，消息发送间隔越短，发送成功率越低。此结果主要是由于网络拥塞或系统处理能力不足导致的。因此，在设计消息传输策略时，平衡消息大小和发送间隔是至关重要的，而这对于重载场景下的发送成功率保障尤为重要。

此外，我们测试了分别配置 1 个、2 个、5 个和 10 个消息代理情况下 DCP 系统处理不同规模消息（1 KB、10 KB、64 KB 和 128 KB）的吞吐量。如图 9 所示，DCP 系统的吞吐量随着消息代理数量的增加而提升。当消息规模为 1 KB 时，配置 1 个消息代理的情况下，DCP 系统的吞吐量为 228.0 MB/s，而配置 10 个消息代理的情况下，吞吐量达到了 1,061.3 MB/s。另外，当消息规模为 128 KB 时，配置 1 个消息代理的情况下，DCP 系统的吞吐量为 714.6 MB/s，而配置 10 个消息代理的情况下，吞吐量为 2,394.6 MB/s。

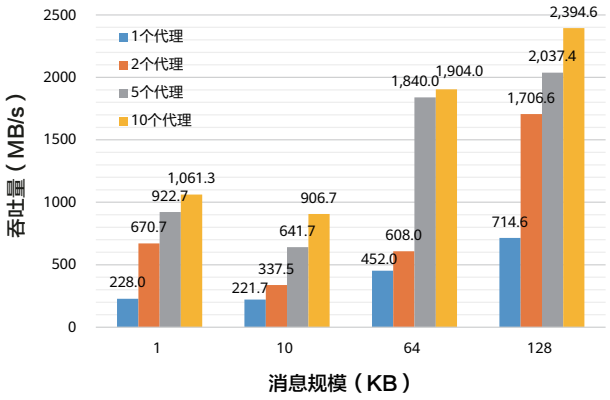


图 9 DCP 系统的吞吐量

以上结果显示，DCP 系统的吞吐量随着消息规模的增加而提升。这说明，只要妥善管理消息发送间隔以避免消息发送成功率下降，DCP 系统处理较大消息时更为高效。此外，增加消息代理的数量对较大消息的吞吐量提升效果更为明显。

4 结语

网络原生 AI、无线感知等“超越连接的服务”产生和消费的数据量巨大，因此，需要一个高效的数据传输机制，以支持复杂的数据拓扑以及数据的随路处理。然而，当前网络架构采用的基于会话控制和管理 UE 通信数据的模式与非连接业务采用的基于业务的模式有很大不同。为克服这一限制，本文提出一个专用的数据面设计方案，用来传输非连接数据。此方案引入了 DO 和 DA 组件，用于管理各类数据拓扑中的数据链，通过随路处理分布式数据并利用泛在计算能力，实现 DaaS，并助力网络原生 AI 和无线感知业务的部署。本文还提出了一种异步数据交换机制 DCP，以实现数据生产者和消费者之间的高效数据交换。该 DCP 是一个无状态、分布式消息代理，其效果得到了实验结果证明，尤其在需要多个 AI 模型协作或利用 ISAC 进行环境重建的场景下，其优势尤为明显。

参考文献

[1] 3GPP, "3GPP TS 23.501 - System architecture for the 5G system (Rel 15)," 2021.

[2] International Telecommunication Union (ITU), "IMT traffic estimates for the years 2020 to 2030," [Online]. Available: https://www.itu.int/pub/r-rep-m.2370 (accessed on Oct. 25, 2023)

[3] Z. Qin *et al*., "6G data plane: A novel architecture enabling data collaboration with arbitrary topology," *Mobile Networks and Applications*, vol. 28, no. 1, pp. 394–405, 2023.

[4] H. L. Garner, "The residue number system," *IRE Transactions on Electronic Computers*, pp. 140–147, 1959.

[5] RocketMQ, [Online]. Available: https://rocketmq.apache.org/docs/

无线网络在网学习
基于潜在结构蒸馏的分布式 LLM

Abdellatif Zaidi [1]，Romain Chor [1]，Piotr Krasnowski [1]，Milad Sefidgaran [1]，Rong Li [1]，Fei Wang [2]，
Chenghui Peng [2]，Shaoyun Wu [2]，Jean-Claude Belfiore [1]
[1] 法国先进无线技术实验室
[2] 无线技术实验室

摘要

本文面向无线接入网（Radio Access Network，RAN）推理场景提出一种无需传输原始数据的分布式学习算法：在网学习（In-Network Learning，INL）。而对于**多模态**数据和**异构**数据的学习（一般涉及分布式提取并融合特征以支撑可靠决策），在网学习算法的表现尤为出色。相比水平/垂直联邦学习（Federated Learning，FL）和水平/垂直拆分学习（Split Learning，SL）等方案，在网学习的精度和带宽收益优势尤为显著。在网学习算法还支持扩展，可以在无线网络上支持大语言模型（Large Language Model，LLM）的部署和知识蒸馏。

1 无线网络分布式推理

传统工作普遍把机器学习算法视作黑盒而直接取代通信模块。实际上，无线接入网（Radio Access Network，RAN）具有独特的内在属性，如果利用得当，将有望实现机器学习和网络通信的共存与协作。不同于计算机视觉和神经科学等数据集中的领域，无线网络的数据往往分散于各个站点。无线网络通常基于信道状态信息（Channel State Information，CSI）或无线信号强度指示（Radio-Signal Strength Indicator，RSSI）来评估信道质量，这些信息可以进一步用于定位、预编码及波束对齐等操作 [1]。

在无线网络上实现机器学习的一个常规做法是，通过某一站点（如云服务器或宏基站）采集所有相关数据，然后基于全部可用数据和算力训练出一个完整的机器学习模型。但由于数据量庞大、网络资源（包括算力和带宽）稀缺，这套做法在很多情况并不适用。有些应用（如自动驾驶）对时延的要求极为严苛，数据共享程度受限。对用户隐私有要求的场景，也不适合共享原始数据。而且，小型基站、车载传感器和智能手机等边缘设备的内存和算力也非常有限。再加上无线网络环境瞬息万变，连接波动时有发生，随时有设备接入或退出网络。另外，从数据形态来说，无线网络数据往往是多模态的异构数据，可能同时跨设备和跨用户。RAN 推理的主要特点总结如下表。

1.1 无线网络边缘 AI

以上种种挑战需要一种全新的范式应对：“边缘学习”，或称“分布式学习”，将 AI 从网络中心推广到网络边缘。该范式的通信设计至关重要，因为在无线网络上，作为机器学习算法核心组件的数据和算力都是高度分散的。RAN 分布式推理，就是利用数据的分散分布，在一个或多个站点上做出一个或多个任务决策。整体框架见图 1，基站（Base Station，BS）和用户设备（User Equipment，UE）等都有各自的神经网络，有的设备负责对通信/感知采集得到的数据进行处理，还有的设备则负责提供算力。

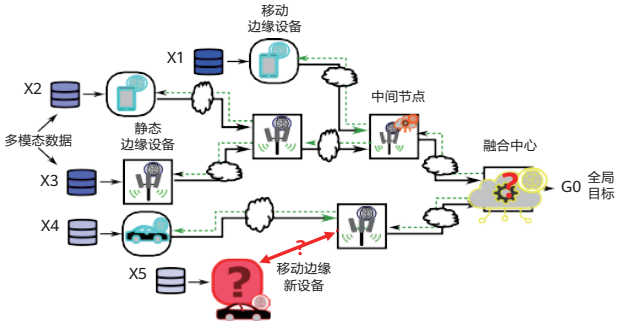


图 1 无线网络分布式推理

1.2 业界算法综述

按分布式阶段对无线网络 AI 方案分类，则有仅支持分布式训练的水平联邦学习（Horizontal Federated Learning，HFL）和水平拆分学习（Horizontal Split Learning，HSL），以及既支持分布式训练也支持分布式推理的垂直联邦学习（Vertical Federated Learning，VFL）。

- **HFL 算法：** [2] 提出的 HFL 当属目前最流行的分布式学习架构。这种架构最适用于分布式训练 + 集中式推理的场景。在训练阶段，神经网络（Neural Network，NN）的多个副本在不同的客户端上基于本地数据集进行训练。训练完成后，云服务器或参数服务器（Parameter Server，PS）对各客户端最新学习到的权重参数进行聚合（如求平均）。这个过程可以多次迭代，每次都应用最新的聚合后模型重新初始化，直到收敛。这种架构的核心就是对模型做渐进调整，不仅关注本地数据集，同时兼顾数据的全局变化。
- **VFL 算法：** VFL [3] 是联邦学习架构的一个变体。VFL 将数据进行垂直切分，既支持分布式训练也支持分布式推理。HFL 和 VFL 架构区别见图 2。该架构下，每个客户端有不同特性的数据。对于跨客户端或多模态的异构数据，各客户端采用不同的 NN 模型以适配不同的模态。模型完成联合训练后，提取共同特征，支撑融合中心完成最终决策。VFL 的架构示例见图 3a。有关 VFL 及在无线网络的应用，见 [4, 5] 及其参考文献。

表 1 无线网络的推理特点

数据	推理	网络连接/拓扑	隐私	计算资源
• 训练阶段分散分布 • 推理阶段分散分布 • 多模态 • 异构	• 分布式 • 存在通信瓶颈 • 有融合需求 • 要求极低延迟（≈ 0.1ms）	• 瞬息万变： – 用户入网/退网 – 链路故障 – 信道质量下降	• 重要 – 原始数据携带用户隐私	• 有限 • 跨站点分布

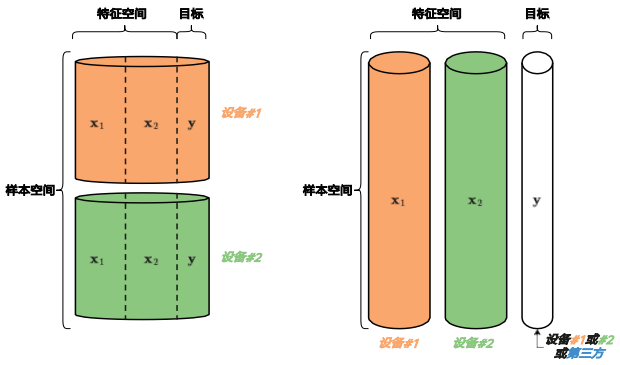


图 2 HFL（左）和 VFL（右）

- **拆分学习算法**：拆分学习（Split Learning，SL）最早出自 [6]，也有两大变体：水平拆分学习（HSL）和垂直拆分学习（VSL）。尽管 VSL 的出现早于 VFL，但目前 VSL 被视为 VFL 的一种特例，本文不再赘述。HSL 算法将 NN 模型拆分为编码器和解码器两大部分。每个边缘设备都拥有一个编码器副本，编码器和解码器按顺序学习。解码器没有自己的数据，而每一轮训练会给 NN 编码器输入一个设备的数据，并基于上一轮的学习结构进行参数初始化。最后，将学习到两部分模型应用于集中推理。

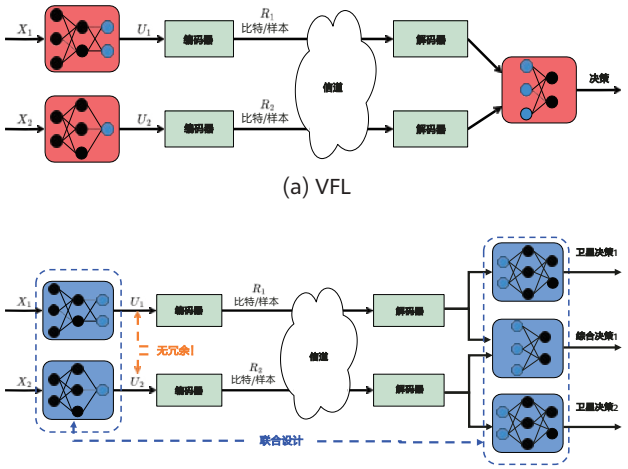
2 在网学习算法

在网学习（In-Network Learning，INL）的起源可以追溯到 [7, 8]，后来由 [9–11] 进一步发扬光大。

2.1 概述

INL 算法擅长对异构和多模态数据进行分布式推理。该架构下，每个设备都拥有一个完整的 NN 模型。推理过程中，每个设备独立地从输入数据中基于推理任务提取特征。特征经网络传输至融合中心进行融合，最终用于推理决策。也就是说，每个设备都相当于一台**编码器**，对各自拥有的数据进行特征提取，互不干涉。训练过程中，INL 确保编码器只提取**互补**特征，设备间的冗余特征则被丢弃，从而大幅节省带宽。因此，INL 的关键组件包括：

- **网络特征融合**：基于分布式方式提取出的特征通过网络传输到融合中心进行融合，支撑融合中心输出可靠的决策。
- **冗余功能去除**：INL 的一大特点在于，编码器经过训练后，能够在推理过程中**只提取非冗余特征**。具体来说，在推理阶段，每个编码器仅从输入数据中提取对当前推理任务有用的特征，不重复提取由其他编码器提取的特征。



(a) VFL

(b) INL

图 3 INL 去除冗余特征

- **网络信道质量决定特征提取**：编码器的特征提取受融合中心的信道质量影响。特征提取的前提是，该特征能够被可靠传输给决策中心。
- **卫星解码器**：融合中心配有一个主解码器和多个卫星解码器，经过训练可以根据编码器传输的具体特征输出软决策。系统架构见图 3b。

2.2 形式描述

给定一个由N个节点组成的网络，表示为一个有向无环图$\mathcal{G}=(\mathcal{N},\mathcal{E},C)$，其中$\mathcal{N}=[1:N]$为节点集合，$\mathcal{E}\subset\mathcal{N}\times\mathcal{N}$为边集合，而$C=\{C_{jk}:(j,k)\in\mathcal{E}\}$为各边权重的集合。每个节点代表一个设备，每条边代表一个容量为C_{jk}的信道。集合$\mathcal{J}$的每个节点给每条边$(j,l)\in\mathcal{E}$分配一条索引或消息$m_{jl}\in[1,M_{jl}]$（$x_j\in\mathcal{X}_j$），接收的索引元组为$(m_{ij}:(i,j)\in\mathcal{E})$。对于$j\in\mathcal{J}$和$l$，满足$(j,l)\in\mathcal{E}$，则$\mathcal{M}_{jl}=[1:M_{jl}]$。节点$j$处的编码函数为：

$$\phi_j:\mathcal{X}_j\times\left\{\bigtimes_{i:(i,j)\in\mathcal{E}}\mathcal{M}_{ij}\right\}\longrightarrow\bigtimes_{l:(j,l)\in\mathcal{E}}\mathcal{M}_{jl}, \tag{1}$$

其中$\times$表示集合的笛卡尔积。同理，对于$k\in[1:N-1]/\mathcal{J}$，节点k给每条边$(k,l)\in\mathcal{E}$分配一条索引$m_{kl}\in[1,M_{kl}]$，索引元组为$(m_{ik}:(i,k)\in\mathcal{E})$。即：

$$\phi_k:\bigtimes_{i:(i,k)\in\mathcal{E}}\mathcal{M}_{ik}\longrightarrow\bigtimes_{l:(k,l)\in\mathcal{E}}\mathcal{M}_{kl}. \tag{2}$$

编码函数$\{\phi_i\}$的范围有如下大小限制：

$$\log|\mathcal{M}_{ij}|\le C_{ij}\quad\forall i\in[1,N-1]\quad\text{and}\quad\forall j:(i,j)\in\mathcal{E}. \tag{3}$$

节点N基于传入消息，对随机变量$Y\in\mathcal{Y}$进行推理：

$$\psi:\bigtimes_{i:(i,N)\in\mathcal{E}}\mathcal{M}_{iN}\longrightarrow\hat{\mathcal{Y}}. \tag{4}$$

基于分布集$\mathcal{Y}$构造重建集$\hat{\mathcal{Y}}$，且$\hat{\mathcal{Y}}=\mathcal{P}(\mathcal{Y})$。然后，

计算$Y \in \mathcal{Y}$的真值与其关于平均对数损失的拟合差异，满足$(y, \hat{P}) \in \mathcal{Y} \times \mathcal{P}(\mathcal{Y})$：

$$d(y, \hat{P}) = \log \frac{1}{\hat{P}(y)}. \tag{5}$$

给定网络拓扑和带宽预算$\{C_{ij}\}$，INL 算法的性能优劣由推理结果的**相关性**给出：

$$\Delta = H(Y) - \mathbb{E}\left[d(Y, \hat{Y})\right]. \tag{6}$$

分类的相关性 (6) 取决于分类误差。

实际上，在有监督学习场景下，(1)、(2) 和 (4) 这三个映射通过学习训练样本$\{(x_{1,i}, \ldots, x_{J,i}, y_i)\}_{i=1}^{n}$得到。训练样本的分布满足：节点$j$（$j \in \mathcal{J}$）有样本$\mathbf{x}_j := (x_{j,1}, \ldots, x_{j,n})$，而末端决策节点$N$有预测结果$\mathbf{y} := (y_1, \ldots, y_n)$。接下来，利用 NN 模型对映射 (1)、(2) 和 (4) 进行参数化。用到的 NN 模型可以任取且互无联系，因为 INL 架构不要求使用完全一样的 NN 模型，这点与 FL 架构相反。为便于反向传播，须满足以下条件：对于$j \in \mathcal{J}$和$x_j \in \mathcal{X}_j$[1]，

$$\begin{aligned}&\text{NN } (j)\text{首层的大小} = \\ &\text{维度 } (x_j) + \sum_{i:(i,j)\in\mathcal{E}} (\text{NN } (i)\text{末层的大小})\end{aligned} \tag{7}$$

同理，对于$k \in [1:N]/\mathcal{J}$：

$$\begin{aligned}&\text{NN } (k)\text{首层的大小} = \\ &\sum_{i:(i,k)\in\mathcal{E}} (\text{NN } (i)\text{末层的大小})\end{aligned} \tag{8}$$

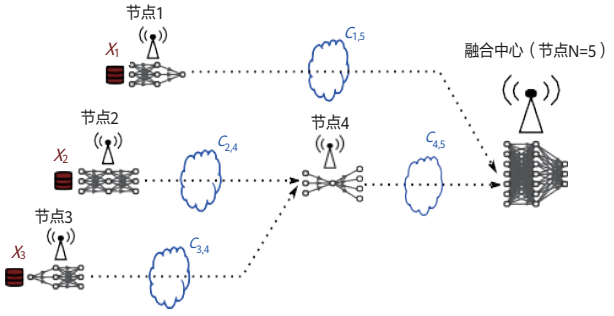


图 4 无线网络 INL 框架示例

图 4 所示网络的训练和推理过程概述如下。

1. **训练阶段**：前传过程中，各节点$j \in \{1, 2, 3\}$对训练数据集$\mathbf{x}_j$进行小批次处理，小批次大小为b_j。节点 2 和 3 分别将各自 NN 末层激活值向量发送给节点 4，在节点 4 的 NN 输入层完成垂直拼接，见 (7)。激活值向量在节点 4 的 NN 上继续前传，抵达 NN 末层。然后，节点 1 和 4 分别把各自 NN 的末层激活值发送给节点 5。节点 1 和 4 的 NN 末层大小满足 (8)，因此同样在节点 5 的 NN 输入层进行垂直拼接。同理，前传继续，直至节点 5 的 NN 末层。各节点基于标准反传算法更新参数：对于节点$t \in \mathcal{N}$，令L_t表示节点t的 NN 末层索引，设$\mathbf{w}_t^{[l]}$、$\mathbf{b}_t^{[l]}$和$\mathbf{a}_t^{[l]}$分别表示权重、偏置以及节点t的 NN 第$l \in [2 : L_t]$层的激活值，σ是激活函数，$\mathbf{a}_t^{[1]}$表示 NN 的输入，则节点t按下式计算误差向量：

$$\boldsymbol{\delta}_t^{[L_t]} = \nabla_{\mathbf{a}_t^{[L_t]}} \mathcal{L}_s^{NN}(b) \odot \sigma'(\mathbf{w}_t^{[L_t]} \mathbf{a}_t^{[L_t-1]} + \mathbf{b}_t^{[L_t]}) \tag{9a}$$

$$\boldsymbol{\delta}_t^{[l]} = [(\mathbf{w}_t^{[l+1]})^T \boldsymbol{\delta}_t^{[l+1]}] \odot \sigma'(\mathbf{w}_t^{[l]} \mathbf{a}_t^{[l-1]} + \mathbf{b}_t^{[l]}) \ \forall\ l \in [2, L_t - 1], \tag{9b}$$

$$\boldsymbol{\delta}_t^{[1]} = [(\mathbf{w}_t^{[2]})^T \boldsymbol{\delta}_t^{[2]}] \tag{9c}$$

并按下式更新其权重和偏置参数：

$$\mathbf{w}_t^{[l]} \to \mathbf{w}_t^{[l]} - \eta \boldsymbol{\delta}_t^{[l]} (\mathbf{a}_t^{[l-1]})^T, \tag{10a}$$

$$\mathbf{b}_t^{[l]} \to \mathbf{b}_t^{[l]} - \eta \boldsymbol{\delta}_t^{[l]}, \tag{10b}$$

其中η为学习率[2]。

在反传阶段，各 NN 按式 (9) 和 (10) 进行参数更新。反传过程从节点 5 开始更新该节点 NN 的参数，即从该节点的 NN 末层开始依次应用式 (9) 和 (10)。

误差最终反传至节点 5 的 NN 首层。然后，节点 5 将其输入层的误差向量(9c)水平切分为两个子误差向量，上方的子误差向量的大小即节点 1 的 NN 末层大小，而下方的子误差向量大小即节点 4 的 NN 末层大小，详见图 5。然后，节点 1 和节点 4 同时继续类似的反传操作。节点 4 将其输入层的误差向量 (9c) 水平切分为两个子误差向量。上方的子误差向量的大小即节点 2 的 NN 末层大小，而下方的子误差向量大小即节点 3 的 NN 末层大小。最后，节点 2 和节点 3 的 NN 反传继续，整个过程重复迭代，直至收敛。

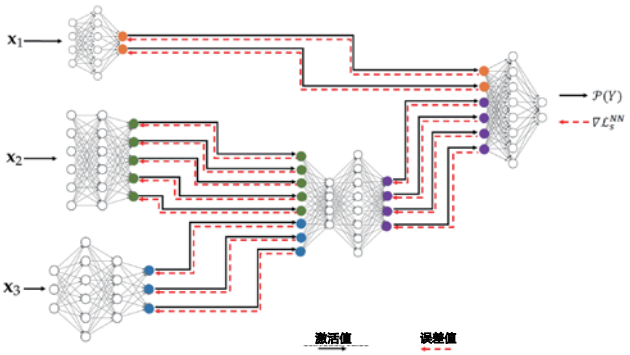


图 5 网络拓扑 4 的正传和反传过程

2. **推理阶段**：推理过程中，节点 1、2 和 3 负责观测不同的数据，设为$\mathbf{x}_1$、$\mathbf{x}_2$和$\mathbf{x}_3$。节点 1 用其 NN 处理数据$\mathbf{x}_1$，并将编码值$\mathbf{u}_1$发送给节点 5；同理，节点 2 和 3 将其编码值发送给节点 4。节点 4 接收到来自节点 2 和 3 的$\mathbf{u}_2$和$\mathbf{u}_3$后，垂直拼接为一个向量并用其 NN 进行处理，最后输出$\mathbf{u}_4$给节点 5。节点 5 对激活值$\mathbf{u}_1$和$\mathbf{u}_4$执行类似处理，然后，该节点以软输出的形式给出标签$\mathbf{y}$的预测结果$Q_{\phi_5}(\mathbf{y}|\mathbf{u}_1, \mathbf{u}_4)$。

[1] 设$\mathcal{X}_j$各元素的维度相同。

[2] 为简化运算，我们对所有 NN 模型采用同一套η和σ。

3 性能收益

本节基于数据分类任务，在精度和带宽维度上对比 INL 算法与联邦学习和拆分学习算法。

3.1 INL vs. HFL 和 HSL

实验 1：基于 CIFAR-10 数据集 [12] 构造五组加噪图像。具体过程是，先对图像做归一化，然后进行加性高斯噪声处理，标准差分别设置为 0.4, 1, 2, 3, 4。INL 框架对五个输入 NN 分别在同一图像的不同加噪版本上训练。NN 网络均采用 VGG 架构 [13]，以分类交叉熵作为损失函数，网络架构如图 6。CIFAR-10 图像的五个加噪版本由不同节点的不同 NN 同时处理，经过拼接后在节点$(J+1)$上再经过一系列全连接（Fully Connected，FC）层处理。而在 HFL 方案下，每个客户端节都拥有一张完整的网络（图 6），数据集等分成五份，同一张 CIFAR-10 图像的五个加噪版本会送入同一个客户端的 NN 网络，不同的客户端观测不同的图像。

HSL 方案中，除$(J+1)$外的其他节点都配备的是卷积层，而节点$(J+1)$用到的是全连接层，整体架构如图 6 所示。训练过程中，各 NN 网络对卷积层的输出做垂直拼接，然后传给节点$(J+1)$，最后该节点将误差向量回传。每个 NN 完成一个轮训（Epoch）后，将更新后的权重向下一客户端传递，下一客户端再对本地数据集执行相同操作。

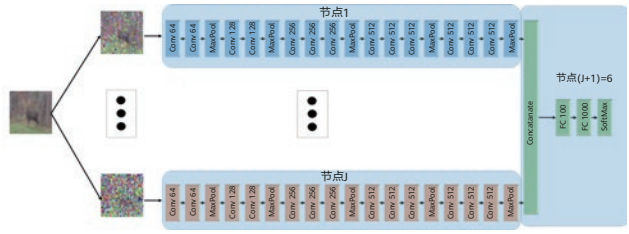


图 6 网络架构（Conv 代表卷积层，FC 代表全连接层）

对于不同的分类精度要求，各节点数据交换所需带宽资源见图 7。由图可知，相比 HFL 和 HSL 方案，INL 能以更低的传输带宽成本满足同等的分类精度要求。

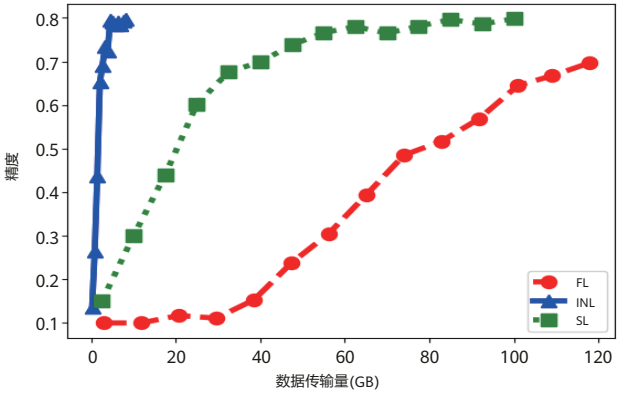


图 7 实验 1 的精度 vs. 带宽成本关系

实验 2：上一个实验因考虑到 INL 和 HFL 的架构差异对训练数据集做了差异化的切分。而接下来这个实验是在相同的数据上完成的。各客户端的 NN 在完整的 CIFAR-10 图像样本上进行训练。对同一个客户端来说，其本地数据集与其他客户端数据集唯一的区别在于添加的高斯噪声多少（标准差分别为 0.4, 1, 2, 3, 4）。而且各节点在不同方案下采用的 NN 模型没有差异，以确保 INL 和 HFL、HSL 三种方案比较的公平性，NN 架构见图 8。

本实验中三种方案的推理表现见图 9。HFL 的推理输入图像取的是 INL 五个加噪图像的平均质量图像。结果显示，相较 HFL 和 HSL 方案，INL 方案的分类精度和带宽要求都更具优势。

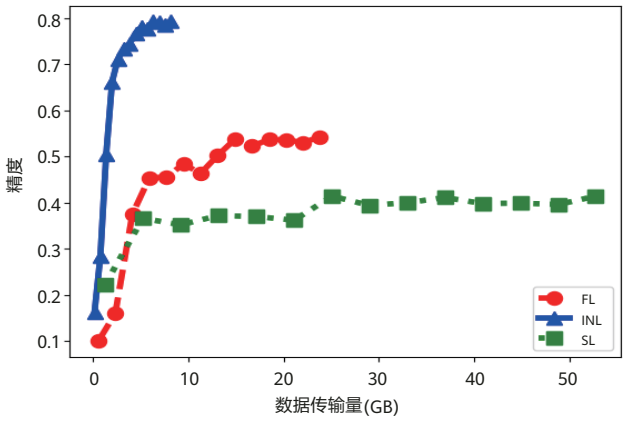


图 9 实验 2 的精度 vs. 带宽成本关系

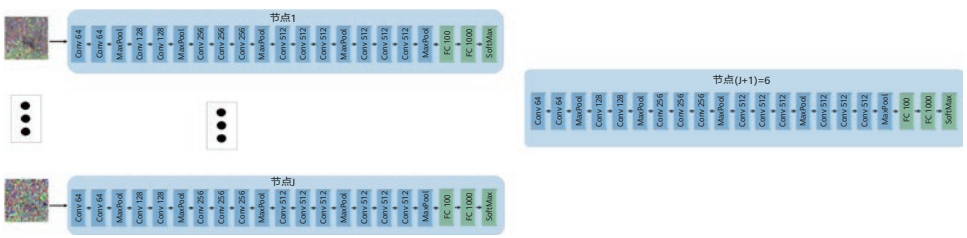


图 8 实验 2 采用的 NN 架构

3.2 INL vs. VFL

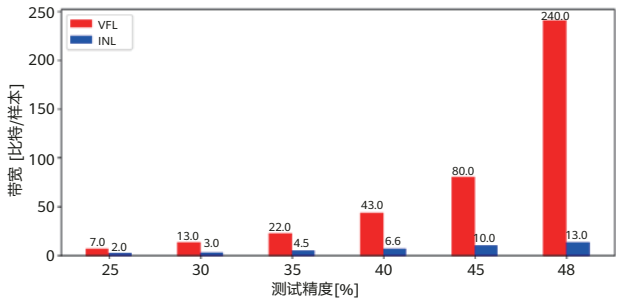
实验 3：本实验基于 CIFAR−10 数据集的变体比较 INL 和 VFL 两个方案（因 VSL 是 VFL 的一种特例，不再赘述）。实验用到两个编码器和一个解码器，三者通过无线网络通信。编码器 A 到解码器的单信道容量为 R_1 比特，编码器 B 到解码器的单信道容量为 R_2 比特，如图 3b。设本征向量 U_1 和 U_2 的维数为 64。在推理阶段的每个时刻，编码器 A 采用 CNN（含少量卷积层）或 ResNet−18 网络 [14]，输入给它的是给定 CIFAR−10 图像的半遮挡副本，而编码器 B 由全连接层组成，输入给它的是同一给定 CIFAR−10 图像的加噪版本。

如图 10（CNN 编码器）和图 11（ResNet−18 编码器）所示，在同样的精度要求下，VFL 所需带宽比 INL 多将近 3 倍。某些特定精度下，INL 的通信成本优势更为突出。同样对于每样本 $R_1 = R_2 = 20$ 比特的带宽要求，INL 方案可实现 45% 的精度，而 VFL 的精度仅为 28%。

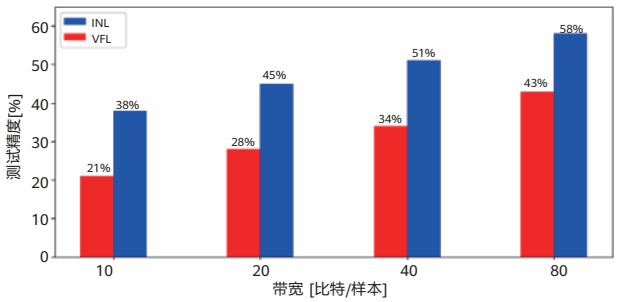
4 大语言模型

大语言模型（Large Language Model，LLM）功能强大，正在颠覆 AI 开发方式，重塑我们的未来。然而，LLM 的多模态数据处理要求也给云端部署带来了诸多挑战，包括（1）响应时间长，（2）通信带宽成本高，（3）数据隐私风险。实际上，可以利用移动边缘计算（Mobile Edge Computing，MEC）技术，在数据源上或靠近数据源的地方微调和部署 LLM，也能有效保障最终用户的数据隐私。面向 6G 时代，无线网络正在向“网络使能 AI”（Network for AI，NET4AI）演进 [15]。基于本文提出的 INL 架构，可以利用 6G MEC 技术将 LLM 部署在网络边缘。我们的方案具有以下关键优势：

- **目标分解**：移动网络系统的各网络层协同执行推理。融合中心将推理目标分解为一个个子目标，并根据下层 BS 的能力特点进行子目标分配。BS 收到子目标后为下层设备做进一步分解，直到最终边缘设备都接收到目标的一小部分。完整过程示意见图 12a。
- **跨视图注意力**：Transformer 的自注意力机制只计算本地传感器数据。如果在同一个推理任务上采用多个传感器获取多视图数据，就需要相应计算跨视图注意力。跨视图注意力定义某一传感器采集的数据的令牌需要在多大程度上关注其他传感器采集或观测的数据的令牌。我们可以把特征投影到特征空间中的超平面，然后在融合中心处计算跨视图注意力。参见图 12b、12c 和 12d。
- **潜在结构知识蒸馏**：迈向 6G 时代，移动网络有望在边缘侧实现在网分布式 AI 计算 [15]。但是，如何能在 6G 边缘设备上运行 LLM 这样对内存和算力要求极高的大模型？此外，网络带宽是否足够支持不同的代理/设备进行 LLM 交换以实现模型聚合和协作？针对这些挑战，[16] 给出了一个可行方案：设备之间利用 INL 算法仅互相交换各自提取的特征结构，而不交换特征本身。这样，各设备可以利用彼此交换的特征结构对本地提取的特征做进一步微调。

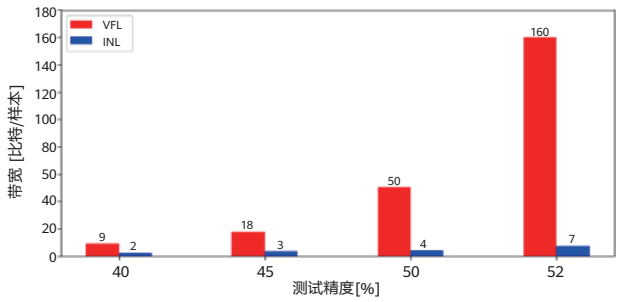


(a) 不同精度要求下的所需带宽

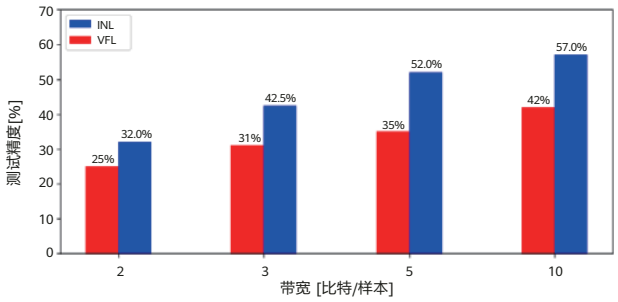


(b) 测试精度 vs. 可用带宽（$R_1 = R_2 = R$ 比特/样本）

图 10 实验 3：INL 和 VFL 性能比较（基于 CNN 编码器）

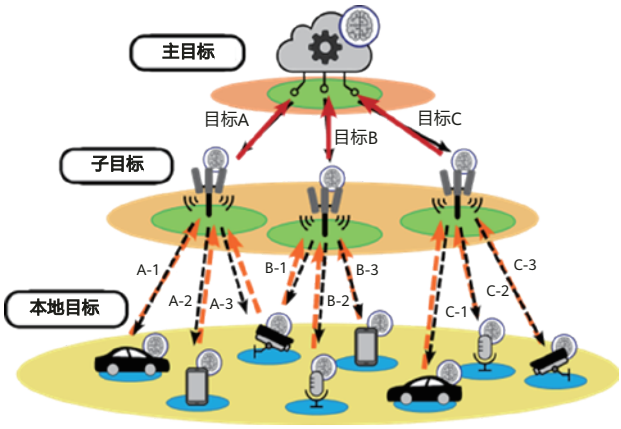


(a) 不同精度要求下的所需带宽

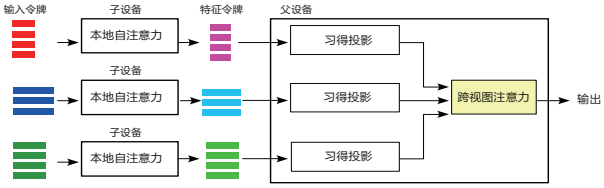


(b) 测试精度 vs. 可用带宽（$R_1 = R_2 = R$ 比特/样本）

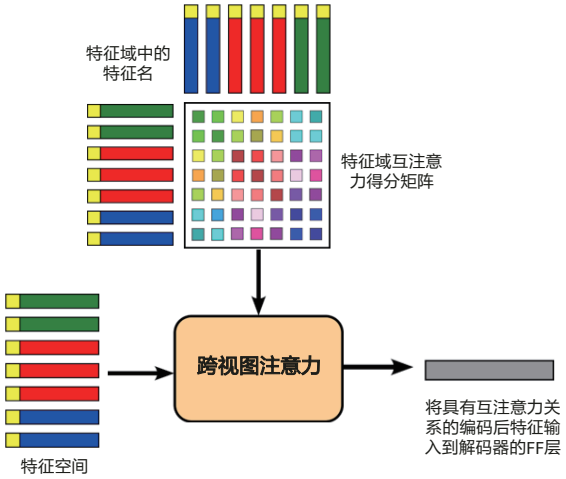
图 11 实验 3：INL 和 VFL 性能比较（基于 ResNet−18 编码器）



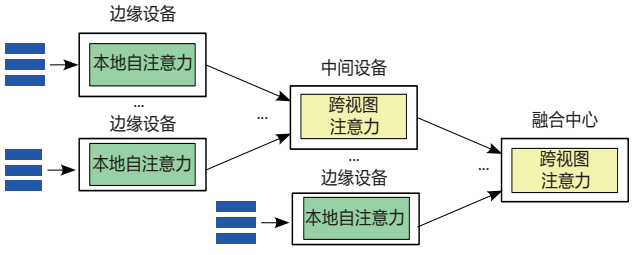
(a) 无线网络的决策目标分解



(b) 基于特征投影计算跨视图注意力



(c) 在特征空间中计算跨视图注意力



(d) 跨视图注意力的分层计算

图 12 LLM INL 架构的关键组件

参考文献

[1] X. Wang, L. Gao, S. Mao, and S. Pandey, "CSI-based fingerprinting for indoor localization: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, 2017.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[3] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, and T. Ranbaduge, "Vertical federated learning: Challenges, methodologies and experiments," *arXiv preprint arXiv*: 2202.04309, 2022.

[4] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al*., "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[6] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.

[7] I. E. Aguerri and A. Zaidi, "Distributed Variational Representation Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 120–138, 2021.

[8] I. Estella Aguerri and A. Zaidi, "Distributed information bottleneck method for discrete and Gaussian sources," in *International Zurich Seminar on Information and Communication (IZS 2018). Proceedings*. ETH Zurich, 2018, pp. 35–39.

[9] M. Moldoveanu and A. Zaidi, "In-network Learning for Distributed Training and Inference in Networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[10] ——, "On in-network learning. A comparative study with federated and split learning," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2021, pp. 221–225.

[11] ——, "In-network learning: Distributed training and inference in networks," *Entropy, vol*. 25, no. 6, p. 920, 2023.

[12] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical report, University of Toronto, Toronto, Ontario, Tech. Rep. 0, 2009.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv*: 1409.1556, 2014.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] L. Huawei *et al*., "6G: The next horizon from connected people and things to connected intelligence," *Huawei, White Paper*, 2022.

[16] M. Sefidgaran, A. Zaidi, and P. Krasnowski, "Minimum description length and generalization guarantees for representation learning," *Advances in Neural Information Processing Systems*, vol. 36, 2023.

基于联邦学习架构的 6G 空口数据分布式生成新方式

徐明枫，李阳，周伟，刘慧 *，江甲沫
中国信息通信研究院 移动通信创新中心

摘要

人工智能技术与通信系统的深度融合将为通信系统开辟新的智能维度，推动其向普惠智能方向演进，实现“万物智联”的 6G 愿景。然而，高质量人工智能模型的诞生依赖于海量数据。为减少海量化、多样化空口数据采集带来的高昂成本，本文探索了应用生成对抗网络模型合成人工空口信道数据的可行性。为了充分利用边缘节点算力并降低网络传输负载，本文研究了基于联邦学习架构的分布式生成对抗网络训练方法。仿真结果表明，在联邦式训练下，模型生成的信道数据质量可以逼近集中式训练水平，为未来无线空口数据采集提供了一种新思路。

关键词

6G，生成对抗网络，联邦学习

* 通信作者

1 引言

自 21 世纪 20 年代以来，业界涌现出一批现象级的人工智能应用，如 AlphaFold、ChatGPT、DALL-E3 和 Sora 等，这些新型应用不仅将改变人们的生活方式，还将推动生产效率提升 [1]。通信系统与人工智能技术的深度融合将助力 6G 无线网络迈向智能化革新阶段，支撑普惠智能的社会建设，实现万物智联的 6G 愿景 [2]。然而，人工智能模型的训练和优化需要海量高质量数据作为支撑，而采集多样化的大规模无线空口数据具有挑战性 [3]。由于无线空口信道与周围传播环境密切相关，环境微小波动也会导致空口信道发生变化，为确保训练出人工智能模型具备足够泛化能力，需要尽可能充分采集各种物理环境下的空口数据，这将带来巨额的人力物力开销。为解决这一问题，业界提出了应用生成式模型扩充数据集的方法 [4]，通过将高质量合成数据混入原始数据集中，可进一步提升模型收敛性能和泛化性能。

然而，训练高效的生成式模型仍需基于充足的的数据集 [5]。考虑到网络中的数据通常由众多边缘节点采集并本地存储，将这些海量数据回传至集中节点处理会导致巨大的链路负载和通信资源开销 [6]。为降低系统开销，一种方法是应用联邦学习 (Federated Learning) 的分布式训练框架，实现生成式模型的分布式训练 [7]。这种方法不仅能有效减少通信开销，还能充分利用各边缘节点的闲置算力，加快模型收敛速度 [8]。此外，由于在训练过程中数据不离开本地，联邦学习能保护数据隐私 [9]。

本文研究了基于联邦学习架构的分布式生成对抗网络 (Generative Adversarial Network) [10] 在无线空口信道数据上的训练方法，并比较分析其与传统集中式训练的生成对抗网络模型在数据生成效果和性能上的差异。

2 基于联邦学习的生成对抗网络模型训练方法

2.1 生成对抗网络模型基本概念

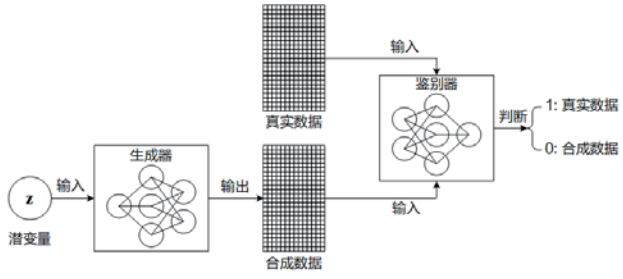


图 1 生成对抗网络模型的基本框架 [11]

图 1 展示了生成对抗网络的基本架构，整体由两个独立的神经网络组成，分别为生成器模型 G 和鉴别器模型 D。生成对抗网络对这两个模型的具体结构没有限制，可以配置为多层感知机、卷积神经网络或 Transformer 等。其核心思想是在传统生成模型基础上引入了一个监督模型，用于判断生成样本是否符合训练样本分布，从而使两个模型在相互对抗的训练中逐渐达到纳什均衡。

生成器模型旨在学习从低维空间的潜变量 z 到高维空间真实信道数据的映射关系。通过随机采样得到的 z 输入到生成器模型，可以输出新的合成数据，记为 $G(z;\theta_g)$，其中 θ_g 表示生成器模型参数。而鉴别器模型的核心目标是区分输入数据是真实还是合成数据。通常，生成器模型的输出层会连接一个 Sigmoid 函数，使其输出值 $D(x;\theta_d)$ 处于 [0, 1] 之间，并作为判定输入是否为真实数据的概率，其中 θ_d 表示鉴别器模型参数，x 表示输入数据。由于生成器训练目标是混淆鉴别器使其误判，而鉴别器训练目标是识别真伪，两者构成极大极小博弈问题，因此生成对抗网络整体优化目标可以建模为：

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}(x)}\{\log D(x;\theta_d)\} + \mathbb{E}_{z\sim P_g(z)}\{\log(1-D(G(z)))\}, \quad (1)$$

当两个模型达到最优时，生成器模型生成的数据分布 P_g 接近训练数据分布 P_{data}，而鉴别器的判别概率约为 0.5。

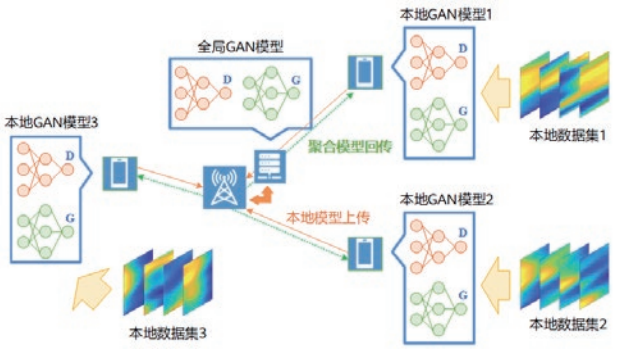


图 2 基于联邦学习的生成对抗网络训练框架

2.2 基于联邦学习的生成对抗网络模型架构

如图 2 所示，考虑有 N 个用户参与联邦训练全局生成对抗网络模型的场景，且各用户所存储的信道数据集不同，记为 $D_n(n=1,\ldots,N)$，每个数据集包含 $|D_n|$ 个信道样本，其分布记为 $P_n(x)$。本文只考虑所有用户的数据集服从独立同分布的情况。为了完成联邦训练，需要进行多轮模型聚合。以第 t 轮模型参数聚合为例，每个用户先基于本地数据集进行 M 轮本地生成器模型和鉴别器模型训练，两个模型参数分别记为 $\theta_n^g(tM)$ 和 $\theta_n^d(tM)$，其中，tM 表示训练次数。然后将这两个模型的参数上传至集中节点，由集中节点对多个用户上传的参数进行加权平均后将其回传，加权平均规则

按照各用户数据量占比决定，数据量越大的用户其模型参数的重要性越高 [12]，最终聚合结果如下：

$$\theta_{FL}^{d}(t)=\frac{\sum_{n=1}^{N}|D_n|\,\theta_n^{d}(tM)}{\sum_{n=1}^{N}|D_n|},\qquad \theta_{FL}^{g}(t)=\frac{\sum_{n=1}^{N}|D_n|\,\theta_n^{g}(tM)}{\sum_{n=1}^{N}|D_n|},\tag{2}$$

其中，$\theta_{FL}^{d}(t)$ 和 $\theta_{FL}^{g}(t)$ 分别表示第 t 轮的生成器和鉴别器的聚合结果。在模型参数回传阶段，集中节点将更新后的全局模型参数 $\theta_{FL}^{d}(t)$ 和 $\theta_{FL}^{g}(t)$ 分发给所有参与用户，然后用户将本地模型参数 $\theta_n^{d}(t)$ 和 $\theta_n^{g}(t)$ 更新为全局模型参数，完成一轮联邦聚合训练。重复上述过程直到模型收敛。

2.3 训练方法

经过 T 轮聚合后，模型趋近于收敛，模型的完整训练过程详见算法 1。

算法 1 基于联邦学习的生成对抗网络训练过程

初始化： 联邦训练轮数 T，每轮各用户本地训练次数 M，各用户初始化的生成器模型参数 $\theta_n^{g}(0)$ 和鉴别器模型参数 $\theta_n^{d}(0)$，各用户数据量 $|D_n|$，收敛阈值 δ。

循环： 对于第 t 轮联邦训练，$1 \leqslant t \leqslant T$，

1. 各用户根据文献 [11] 中的算法 1 进行 M 轮本地训练，得到生成器和鉴别器模型参数 $\theta_n^{d}(tM)$ 和 $\theta_n^{g}(tM)$；
2. 各用户将 $\theta_n^{d}(tM)$ 和 $\theta_n^{g}(tM)$ 上传至集中节点，集中节点根据式 2 更新全局鉴别器和生成器模型参数为 $\theta_{FL}^{d}(t)$ 和 $\theta_{FL}^{g}(t)$，然后回传给各用户；
3. 各用户将本地鉴别器和生成器模型参数更新为全局模型参数。

终 止： 当 $t>T$ 时 或 $\left\|\theta_{FL}^{d}(t)-\theta_{FL}^{d}(t-1)\right\| \leq \delta$ 且 $\|\theta_{FL}^{g}(t)-\theta_{FL}^{g}(t-1)\| \leq \delta$ 时。

输出： 训练好的全局生成器模型 $\theta_{FL}^{g,*}$ 和鉴别器模型 $\theta_{FL}^{d,*}$。

3 仿真与分析

3.1 无线空口信道参数设置

在本实验中，无线空口信道数据集基于 3GPP TS 38.901 标准中的 TDL-C 信道模型 [13] 构建，该模型可在非视距传输条件下对城市宏基站场景的瑞利信道进行建模。特别是，本实验针对 300 km/h 高速移动场景进行仿真评估。其余仿真参数如下：每帧信道数据维度为 48 个子载波 ×14 个符号，载波频率为 3.5 GHz，子载波间隔为 30 kHz。此外，考虑 4 个用户参与联邦训练过程，每个用户的本地数据集包含 2500 条独立同分布的信道数据。同时，我们选择集中式训练作为对比，其中集中节点的数据集汇总了各用户的本地数据，共计 10000 条信道数据。

3.2 生成对抗网络模型参数设置

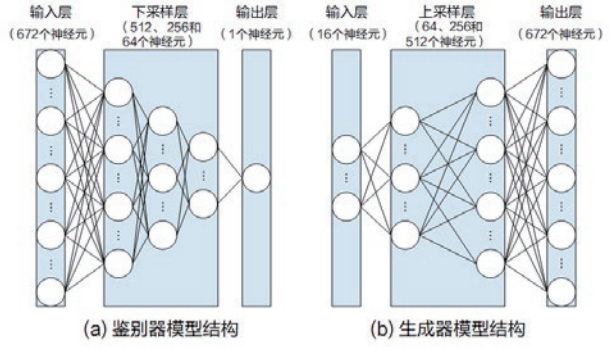


图 3 生成对抗网络模型结构 [11]

本实验采用的生成对抗网络模型结构参考文献 [11]，其中生成器和鉴别器均为 5 个线性层的全连接网络，具体如图 3 所示。

对于鉴别器模型，首先将输入的信道数据展开为向量形式，考虑到信道值是复数形式，因此将信道的实部和虚部作为两条数据样本进行处理。因此，输入层设计为 672 个神经元。隐藏层包含三个下采样线性层，分别具有 512、256 和 64 个神经元。前四层输出连接 LeakyReLU 激活函数和 LayerNorm 层归一化函数。最后，输出层设 1 个神经元，并接 Sigmoid 函数，用于判别输入样本属于真实数据的概率。

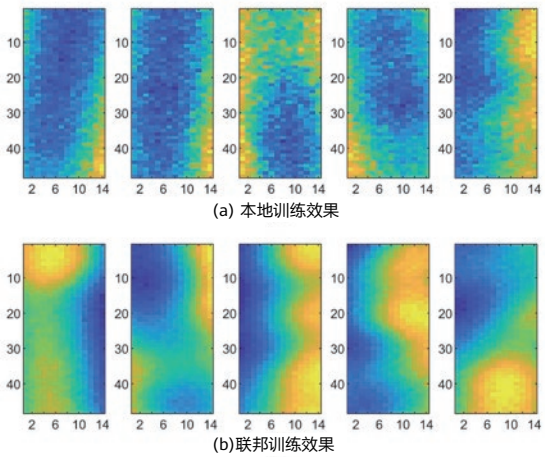
对于生成器模型，输入的潜变量 z 为维度 16 的标准高斯分布随机向量。隐藏层包含三个上采样线性层，各层神经元结构与鉴别器模型的下采样层镜像对应。最后，输出层为 672 个神经元，以还原完整信道矩阵大小，并接 Tanh 激活函数。

3.3 仿真评估

图 4 展示了在联邦训练轮数为 20 次、每轮本地训练次数为 10 次时无线空口信道强度仿真结果，横坐标表示符号，纵坐标表示子载波。图 4a 显示仅本地训练的用户信道生成效果，图 4b 则是联邦训练全局模型的生成效果。从图中可以看出，在有限训练次数下，仅靠本地训练的模型生成性能较差，生成的信道图像高度模糊，而联邦训练模型显著提高了生成性能，噪点明显减少。这表明联邦架构能够充分利用分散在多节点的数据，提高模型训练效果。

图 5 展示了联邦训练轮数为 50 次、每轮本地训练次数为 20 次的仿真结果。从图 5a 可以看出，由于数据量不足，

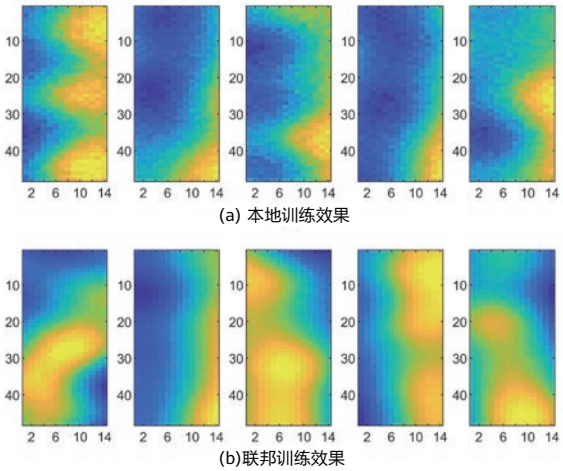
本地模型生成的信道样本质量仍然较低，尽管相比图 4a 有明显改善，但样本中依然存在明显噪声。图 5b 显示了联邦训练模型的生成效果，所生成的信道样本几乎没有噪声，已经接近真实信道样本，与纯粹本地训练模型相比，其性能有了质的提升，接近同等数据量下集中式训练模型的水平。



(a) 本地训练效果

(b)联邦训练效果

图 4 训练次数为 200 次时，联邦训练与本地训练的生成对抗网络模型生成效果。



(a) 本地训练效果

(b)联邦训练效果

图 5 训练次数为 1000 次时，联邦训练与本地训练的生成对抗网络模型生成效果。

为了进一步评估基于联邦学习的生成对抗网络模型性能，本实验在信道估计任务上进行了测试，指标为估计值与真实值间的归一化均方误差。图 6 展示了应用联邦训练和集中式训练所得生成对抗网络模型进行信道估计的性能结果，其中信道估计方法详见参考文献 [11] 中的算法 2。由图可知，在不同导频数情况下，联邦训练模型在信道估计性能上接近集中式训练模型。考虑到集中式训练是串行计算方式，而联邦训练是并行计算方式，当各节点的算力相当时，由于实验中模型参数量较少且传输时延远小于模型训练时延，可忽略参数上传和回传的时延 [9]，此时联邦训练的时延开销与参与用户数成反比，即参与用户越多，模型训练耗时越短。

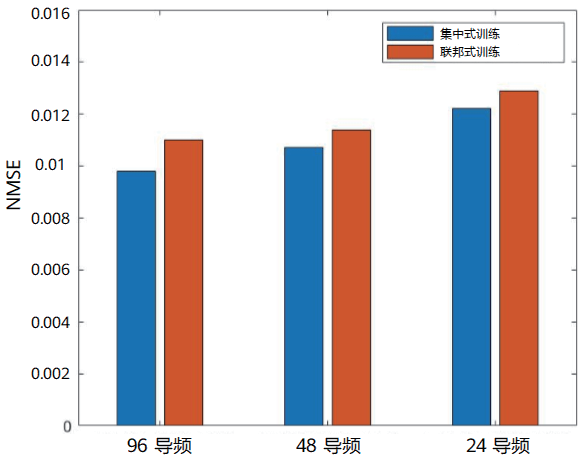


图 6 信噪比为 15 dB 时，集中式训练生成对抗网络模型和联邦训练生成对抗网络模型的信道估计性能比较

4 结语

本文研究了一种基于联邦学习的无线空口信道数据分布式生成对抗网络训练方法。仿真结果表明，采用联邦学习架构训练的生成对抗网络模型，其数据生成质量与传统集中训练模型几乎一致。这为未来无线空口数据采集提供了新指引。

目前本文仅讨论了 Sub 6G 频段下的空口信道数据生成问题，毫米波以及更高频段等复杂场景的空口信道数据生成效果尚待验证。此外，实采信道数据和仿真信道数据之间的差异对数据生成的影响也需深入研究。对于联邦学习等分布式模型训练架构，还存在用户数据非独立非同分布、模型异构和模型聚合异步等潜在问题，需要进一步探索。另外，生成模型能提供的有效数据量也需进一步实验验证。

参考文献

[1] D. Zhang, Z. P. Bhat, K.-H. Lai, *et al.*, "Data-centric artificial intelligence: A survey," arXiv preprint, arXiv: 2303.10158, 2023.

[2] IMT-2030(6G) 推进组 . 6G 总体愿景与潜在关键技术 , 2021 年 6 月 .

[3] M. Xu, Y. Li, M. Li, *et al.*, "A denoising diffusion probabilistic model based data augmentation method for wireless channel," in Proc. International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, Nov. 2023.

[4] X. Li, K. Wang, X. Gu, *et al.*, "Paralleleye pipeline: An effective method to synthesize images for improving the visual intelligence of intelligent vehicles," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 53, no. 9, pp. 5545–5556, May. 2023.

[5] B. Shaker, G. P. R. Papini, M. Saveriano, and K.-Y. Liang, "Generating synthetic vehicle data using decentralized generative adversarial networks," IEEE Access, Jul. 2024.

[6] Z. Zhao, S. Bu, T. Zhao, *et al.*, "On the design of computation offloading in fog radio access networks," IEEE Transactions on Vehicular Technology, vol. 68, no. 7, pp. 7136–7149, Jun. 2019.

[7] C. Hardy, E. L. Merrer, and B. Sericola, "MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets," in Proc. IEEE International Parallel and Distributed Processing Symposium (IPDPS), Rio de Janeiro, Brazil, May. 2019.

[8] 王志勤 , 江甲沫 , 刘沛西等 . 6G 联邦边缘学习新范式：基于任务导向的资源管理策略 [J]. 通信学报 , 2022.

[9] C. Feng, Z. Zhao, Y. Wang, *et al.*, "On the design of federated learning in the mobile edge computing systems," IEEE Transactions on Communications, vol. 69, no. 9, pp. 5902–5916, Jun. 2021.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.

[11] Y. Du, Y. Li, M. Xu, *et al.*, "A joint channel estimation and compression method based on GAN in 6G communication systems," Applied Sciences, vol. 13, no.4, 2023.

[12] H. B. McMahan, E. Moore, D. Ramage, *et al.*, "Communication-efficient learning of deep networks from decentralized data," in Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, USA, Apr. 2017.

[13] 3GPP, "3GPP TR 38.901 v16.1.0 3rd generation partnership study on channel model for frequencies from 0.5 to 100 GHz (Rel 14)," 2020.

面向 6G 网络内生 AI 的服务质量保障

刘光毅[1]，王凯悦[1]，邓娟[1]，吴佳骏[1]，胡焕然[2]，林冠臣[2]

[1] 中国移动有限公司研究院（北京）

[2] 北京邮电大学

摘要

ITU 框架建议书指出，AI 与通信的融合是 6G 六大场景之一，智慧内生成为 6G 研究的重要方向。6G 智慧内生网络的设计目标是有效支撑 6G 空口性能提升、高水平网络自治以及高性能 AI 服务，如何以统一的网络 AI 服务质量（Quality of Artificial Intelligence Service，QoAIS）体系来满足多样化 AI 服务的需求，并以统一的 QoAIS 机制来保障差异化场景的人工智能（Artificial Intelligence，AI）服务质量是业界研究的重点。本文研究 6G 网络 AI 服务质量保障的架构设计、指标体系、协议流程和保障技术，提出分层管控的 6G 网络 QoAIS 架构设计、AI 服务质量指标体系、通算融合的协议设计、面向 AI 大模型的指标体系、基于 MADDPG-Adv 算法的 QoAIS 保障技术等，在网络内部实现 AI 全生命周期管理和 AI 四要素的按需调度。最后，本文围绕准确度和时延两个指标仿真验证了 AI 推理应用场景的服务质量保障效果，所提强化学习算法相比 MADDPG 算法可以带来 6% 的性能增益，满足 QoAIS 要求的用户任务数增加 23%。

关键词

6G，内生 AI，AI 服务质量保障

1 引言

人工智能（Artificial Intelligence，AI）在各行各业的应用对未来 6G 网络提出了新要求，6G 网络需要内生 AI 或者网络 AI 的设计来打造 AI 即服务（AI as a Service，AIaaS）的能力 [1]，为网络自身和第三方用户提供泛在普惠的智能服务。内生 AI 的 6G 网络将支持 AI 赋能网络和网络使能 AI 两类场景。在 AI 赋能网络方面，AI 与空口融合是重要方向，6G 系统的带宽、速率和容量等 KPI 的提升需要通过内生 AI 技术来支撑。在网络使能 AI 方面，移动通信网络是实现 AI 泛在普惠的基础平台，可以利用 6G 网络特色优势来使能 AI 的高价值场景、用例和业务。在 AI 赋能网络方面，ITU、3GPP 等标准化组织积极投入空口 AI 的典型用例及关键技术的研究和标准化工作。目前，AI 技术在复杂未知环境的建模与估计、智能信号调制与编码、网络智能调度、网络智能优化部署等方面具有重要的应用潜力，对 6G 技术的研究和演进具有重要价值 [2]。在网络使能 AI 方面，中国移动联合华为等公司提出面向任务的智慧内生无线接入网（Radio Access Network，RAN）架构，为 6G 内生 AI 架构提供思路 [3]。[4] 在分布式 AI 基础理论模型研究方面，通过设计关键算法并开展仿真实验，为内生 AI 算法研究提供有效验证，并进一步聚焦网络大模型的基础理论算法研究。[5] 提出了内生 AI 无线网络中 AI 训练业务的任务调度和资源分配方案，以实现灵活的计算服务器选择、数据质量调整和资源分配，保障 AI 任务的服务质量（Quality of Service，QoS）需求。[6] 提出了面向 6G 网络的内生 AI 网络切片架构，能够实现 AI 与网络切片的协同，促进网络智能管理，支持新兴 AI 业务。针对 6G 内生 AI 应用，[7] 提出了一种新型联邦迁移学习框架，该框架在分布式边缘多智能体上训练本地模型，并在边缘基站上进行全局聚合，能够显著提高模型精度并降低能耗。6G 与 AI 融合是新机遇，AI 和通信一体化助力实现“智慧泛在，数字孪生”的 6G 总体发展愿景。

6G 网络 AI 为网络高水平自治和各类用户提供智能服务。其中，网络自治的 AI 应用场景可分为运维智能、网络智能和网元智能三类。考虑实时性、隐私性、移动性、端边协同的优势，6G 网络 AI 外部价值场景包括移动机器人、车联网和 XR。典型用例有外出家用机器人、工厂运输机器人、AI 辅助自动驾驶、AR 导航等。根据 GGII 统计数据，全球移动机器人在 2027 年市场规模将达到 1800 亿元以上，机器人的种类多样，可以满足不同行业、不同任务的需求 [8]。一些远程机器人控制需要更严格的 KPI，例如，直升机、类人机器人的时延小于 5 ms，需要端边协同的设计，6G 网络 AI 可以为其提供更低的时延与端边协同优势 [9]。根据 IDC 预测，2026 年全球自动驾驶车辆规模为 8930 万辆，车联网成为未来经济发展重中之重 [10]。车联网的远程驾驶单向时延 10 ms，图像识别小于 5 ms，故障检测、安全警告、视频识别时延需求 500 ms ~ 1s，需要 6G 网络来满足实时性的技术需求。XR 技术在工业应用、医疗领域、教育、军事训练、电子商务等多个场景中广泛应用。相比云 AI，网络 AI 更靠近用户，从而减少非接入网的信令交互，提高 XR 智能服务请求的响应速度。相比于端侧 AI，6G 网络可以提供更强大的算力来支撑 XR 所需的 AI 推理、训练服务，减小 AI 计算的时延。另外，针对端侧用户信息、工业场景的敏感数据等容易受到窃听、污染等问题，6G 网络 AI 相比云 AI 的数据传输范围更小，有利于数据隐私保护。对于用户在移动过程中发出的请求，网络可以更灵活地响应 AI 服务，例如模型的推理及实时下载等。

由于场景和用例的多样化、AI 服务类型和统一 QoS 需求存在差异，6G 网络 AI 需按需提供 AI 服务类型，并保障 AI 服务质量（Quality of Artificial Intelligence Service，QoAIS）。6G 网络 AI 可以按需向被服务方提供的 AI 服务类型主要包括 AI 推理、AI 训练、模型优化、数据处理与预处理四大类。AI 推理适用于网络 AI 场景下，对时延要求高的用例，AI 模型接收输入数据并立即生成输出，例如语音识别、实时翻译等；同时适用非实时的推理，例如大规模图像分类、视频分析等。AI 训练包括监督学习训练、无监督学习训练、半监督学习训练、强化学习训练，模型优化包括模型剪枝、量化、知识蒸馏、结构优化等。数据处理与预处理包括去除或修正数据中的噪声和错误、确保数据质量的数据清洗服务和通过生成新的训练样本来扩展数据集、常用于图像、文本等领域的数据增强服务等 [11]。

6G 对 QoAIS 的保障包括架构设计、指标体系、协议流程和保障技术四个方面。在架构设计方面，传统通信架构设计以会话为中心进行 QoS 保障，无闭环保障，面向 AI 服务，架构设计需要内生支持连接、计算、数据和算法等多要素资源的深度融合，必须构建新的 QoS 架构；在指标体系方面，传统的通信 QoS 仅包括时延、可靠性、速率、优先级等指标，而 AI 服务涉及连接、算力、数据、模型等多维要素资源，无法用 5G QoS 指标体系进行准确刻画 [12]。在协议流程方面，AI 服务的申请、AI 任务的控制和 AI 资源的控制需要端到端协议和流程设计，新型 AI 承载需要相应的协议设计，QoS 指标的传递和映射也需要流程支持。现有 QoS 机制的业务区分颗粒度较粗、优化调整周期较长，难以适配以任务为中心的 AI 服务模式。在保障技术方面，5G QoS 保障以会话和连接为基础，核心网网元根据 IP 五元组映射获得每个数据包对应的 QoS Flow 及 QoS 指标，然后传递给 RAN，并由 RAN 进行相应的数据无线承载映射和空口资源调度，进而保证不同业务和数据包的通信 QoS [13]，而 AI 服务质量的保障需要连接、计算、数据和模型四维度资源均达到 QoS 要求、且三层 QoS 间均有合理的映射关系，保障技术将更加复杂。综上分析，现有方案无法满足 6G 网络 QoAIS 保障需求。面对 AI 服务场景业务和 QoS 需求差异大、通算数智协同调配资源要素多等挑战，6G 需要设计新的 QoAIS 保障的方案体系，包括网络架构、指标体系、协议流程和保障技术等。

2 QoAIS 架构设计

6G 网络 QoAIS 架构采用分层管控设计，以任务为中心进行任务动态编排，在网络内部实现 AI 的全生命周期管理和 AI 的四要素按需调度。

由于 AI 引入了计算、数据、模型要素，6G 网络 QoAIS 架构需要在增强已有通信网络功能的基础上新增计算、数据和模型相关功能。在图 1 所示的 6G 内生 AI 的逻辑功能架构图中，最上层是管理编排功能，完成 AI 服务需求解析、映射和 AI 服务功能链编排，满足服务层的质量需求（QoS）。管理编排功能无法直接管理 UE，因此无法实现四要素的协同来管控和保障任务 QoS，且管理编排层信令时延较大，任务管控不及时，难以满足严格的任务 QoS 保障要求。在控制面引入 AI 任务控制功能，负责 AI 任务的生命周期管理和任务控制器的选择。通信控制功能在现有功能基础上，新增数据承载建立、模型传输承载等控制功能；计算控制功能进行计算资源感知、计算节点选择、计算资源分配等计算控制管理；数据控制功能进行数据采集节点选择、数据采集配置等数据管理功能；AI 模型控制功能进行模型注册、模型选择等模型控制管理。AI 任务控制功能可以基于 AI 任务的 QoS 需求完成任务部署、启动、删除、修改、监控等，包括调控通信、计算、数据、模型四要素资源来进行任务部署阶段的 QoS 保障。在用户面引入任务执行功能，负责 AI 训练、推理、验证等 AI 任务的具体执行。

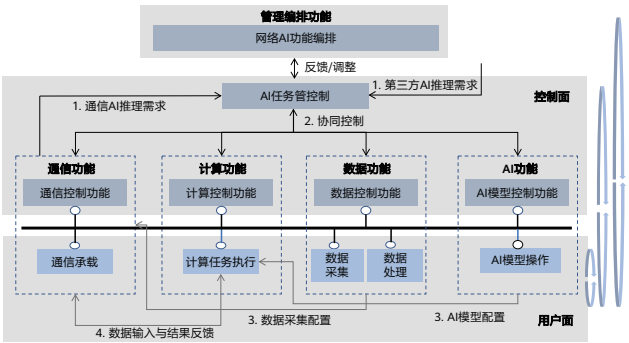


图 1 6G 内生 AI 的逻辑功能架构

为保障 6G 网络中的 AI 服务质量，需要形成闭环反馈的 QoS 保障机制。任务执行层通过实时监测和优化四要素资源，保证资源 QoS 的达成；任务控制层保障任务 QoS 的达成，当资源 QoS 无法达成时，任务控制层对资源配置和任务 QoS 到资源 QoS 的映射进行调整；编排管理层保障服务 QoS 的达成，当任务控制层无法提供任务 QoS 保障时，编排管理层优化 AI 服务到 AI 任务的映射。三层闭环的 QoAIS 机制，完成 AI 服务需求到底层多维网络资源的按需协同映射，实现差异化的 AI 服务质量保障。

图 2 所示为 6G 网络内生 AI 逻辑功能架构中各功能可能的部署位置。AI 服务编排功能可部署于核心网，接收用户发起的 AI 服务请求，完成需求解析以及 AI 服务到 AI 任务的分解。AI 服务编排功能按需编排 AI 任务管理，并将 AI 任务部署到 AI 任务管理功能中。AI 任务管理根据 AI 任务 QoS 要求，部署四要素相关功能，并进行资源配置，完成任务 QoS 到资源 QoS 的映射，考虑 AI 服务的实时性要求，四要素功能可部署在 RAN 内；再由四要素的控制面和用户面功能完成 AI 任务执行。

3 QoAIS 指标体系

QoAIS 指标体系是网络对 AI 服务的质量和效果进行保障所使用的一套指标体系，能够体现 6G 网络 AI 优势，解决用户智能需求、公平性与有限资源的矛盾。QoAIS 指标

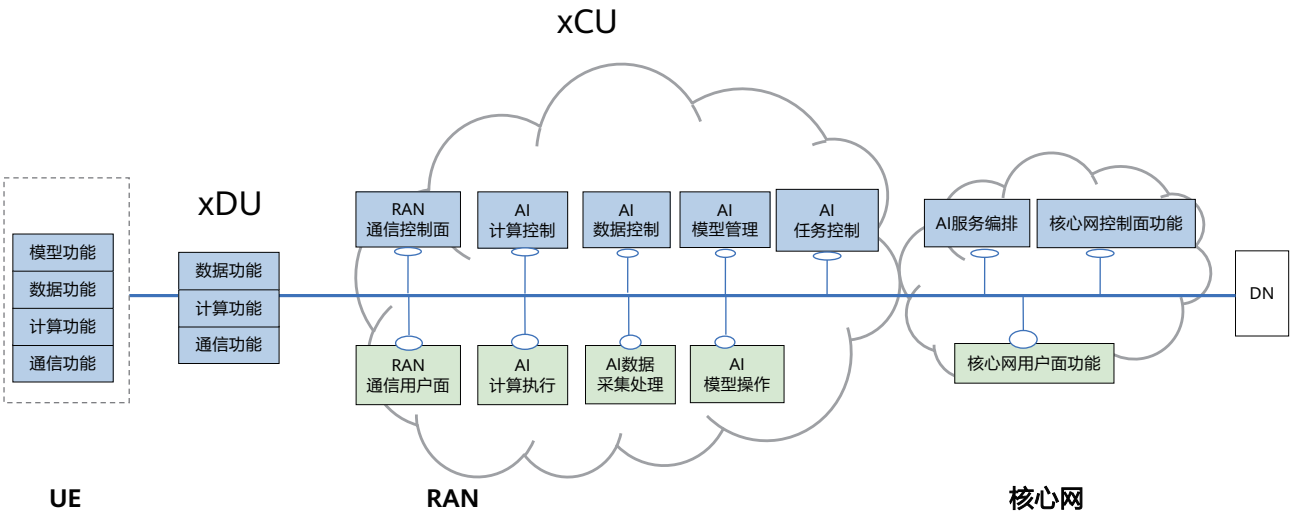


图 2 6G 网络内生 AI 的逻辑功能部署

体系包含 AI 服务 QoS、AI 任务 QoS 和 AI 资源 QoS，三层指标之间具有映射关系。AI 服务是指按需向被服务方提供 AI 技术、业务或 AI 三要素（计算、数据和模型）等。AI 服务 QoS 是指 AI 服务提供的性能、可用性、安全性等方面的质量指标。AI 服务可以分解为一个或多个 AI 工作流，AI 工作流可进一步分解为一个或多个任务。AI 任务是指协同通信、计算、数据、模型来完成 AI 服务的某个特定目标。AI 任务 QoS 包括部署阶段的任务编排优先级和执行阶段的四要素保障优先级。AI 资源是指保障 AI 任务实现涉及的通信、计算、数据、模型四要素资源，包含 CPU、GPU 等计算资源或空口时频域资源等物理资源。AI 资源 QoS 是指网络衡量四要素资源质量的一系列指标。比如计算 QoS 包括计算时延、计算可靠性、计算并行度等；算法 QoS 包括训练收敛速度、泛化性、可解释性等；数据 QoS 包括共线性度、安全隐私等级等；连接 QoS 包括优先级、时延、丢包率等。针对 AI 推理任务，表 1 给出 QoAIS 分层的映射指标。

表 1 QoAIS 分层映射指标

类型		指标
AI 服务		时延、能耗、传输速率、抖动等
AI 推理任务		推理速度、推理精度
AI 资源	数据资源	特征冗余度、完整度、数据准确度、数据准备耗时
	模型资源	性能指标界、训练耗时、收敛性、优化目标匹配度
	计算资源	计算精度、时长、效率
	通信资源	带宽及抖动、时延及抖动、误码率及抖动、可靠性等

第三方大模型发展迅速，在网络使能大模型方面，需要结合大模型特点聚焦 AI 服务和 QoAIS 指标设计。在 AI 服务方面，网络使能大模型服务包括四个阶段：在模型预训练阶段，需要更新全部参数、训练数据海量，对算力需求大；在模型微调阶段，可更新部分参数、训练数据较少，算力需求相对预训练较小；在模型部署阶段，无需参数更新和训练数据、仅执行前向传播、算力需求相对最小；在模型优化阶段，算力需求相对预训练也较小。考虑聚焦在模型推理服务、微调和优化服务，表 2 针对网络大模型的推理、微调和优化服务，从不同的评估维度简化 QoAIS 指标设计，给出网络使能大模型的指标体系。

表 2 网络使能大模型 QoAIS 指标体系

服务类型	评估维度	QoAIS 指标
微调或优化	性能	性能指标界、微调耗时、可解释性、损失函数与优化目标的一致性、公平性
	开销	存储开销、计算开销、传输开销、能耗
	安全	存储安全、计算安全、传输安全
	隐私	数据隐私等级、算法隐私等级
	自治	完全自治、部分人工可控、全部人工可控
推理	性能	推理速度、并发度、推理误差、推理精度等
	自治	模型可控性，如推理任务自动调度、推理资源自动分配等

4 QoAIS 协议流程

6G 网络 QoAIS 保障需要设计端到端协议流程。QoAIS 是管理编排功能、任务控制功能的重要输入，管理编排功能需要将用户需求转化为网络可理解的对 AI 服务能力的要求，即 AI 服务 QoS，再进一步将服务 QoS 分解为任务 QoS，以作为控制面的输入；控制面再将任务 QoS 映射到对连接、

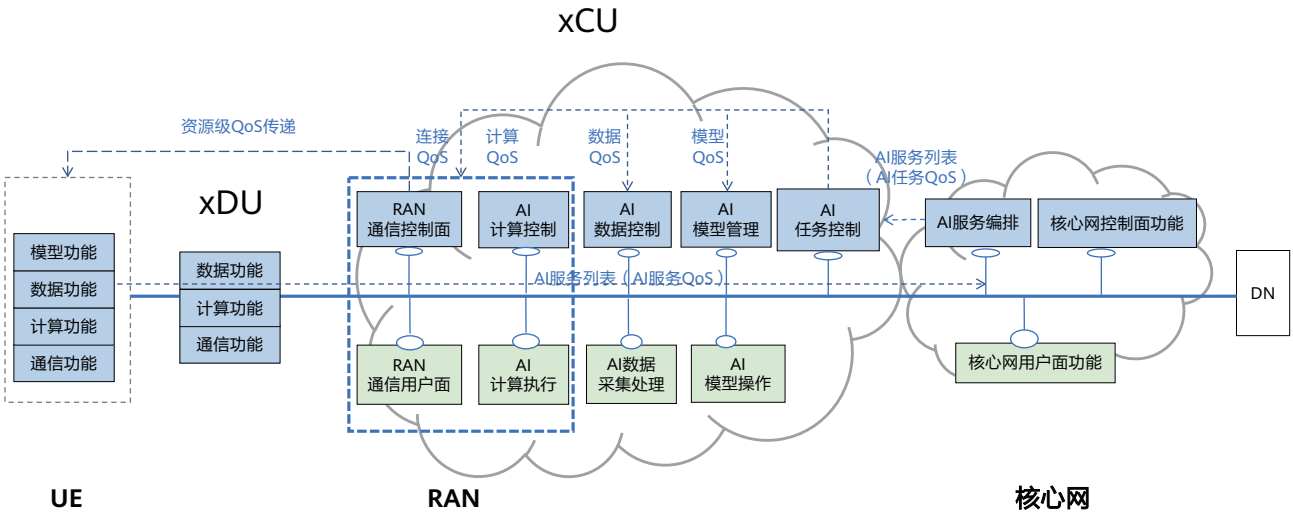


图 3 QoAIS 端到端传递流程示意图

计算、数据和算法等方面的 QoS 要求上，通过实时监测和优化四要素资源，进行 AI 异构多维资源融合控制和承载协议的设计，保障 AI 任务传输和执行过程中的高 QoS 需求。图 3 所示为 QoAIS 的端到端传递流程示意图。

基于 6G 网络的内生 AI 功能部署架构，可进一步进行体系化协议设计。在通算融合的协议设计方面（图 3 蓝色虚框所示），通过新增计算执行层进行 AI 任务的计算处理，新增计算连接承载来传输计算数据。如图 4 所示，在现有 5G 空口用户面协议栈基础上，引入计算相关协议层 CTAP，完成 AI 计算任务的 QoS 映射与更新。网络侧 CTAP 实体根据收到的计算任务的通信 QoS 和计算 QoS 要求，为计算任务数据包添加包头索引，相同通信 QoS 的计算任务数据包过滤映射为一个 QoS Flow，并发送给下层 SDAP 实体，数据适配协议层 SDAP 根据 QoS 指示信息将 QoS Flow 映射到相应无线承载上，并添加 SDAP 包头索引，通过 PDCP、RLC、MAC、PHY 等协议的处理传输，实现不同类型的计算数据到计算连接承载的映射。计算执行层主要负责网络侧或者用户侧计算任务的执行，从上层接收计算数据包以及计算 QoS 要求，将具有类似计算 QoS 要求的计算任务映射到同一个计算执行功能中，再将计算结果递交给下一层。通过新增 CTAP 层，实现 AI 计算执行 QoS 和 AI 连接 QoS 向计算资源和连接承载的映射，保障计算传输与计算执行的 QoS 要求。

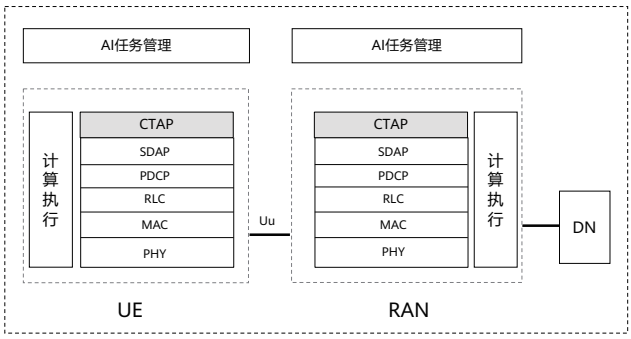


图 4 QoAIS 协议设计

5 QoAIS 保障技术

QoAIS 保障技术通过对用户的 AI 推理、AI 训练等 AI 服务进行调度并进行通信、计算和 AI 模型资源的控制分配，尽可能满足用户对 AI 服务质量的要求，提高系统的整体性能。本文聚焦 AI 服务的准确度和时延指标，在 QoAIS 架构的任务控制层对通信、计算、模型进行融合控制，以保障 AI 服务质量。

本文针对 AI 推理服务提出基于云边端协同的 AI 服务质量保障技术。云端能够提供 AI 功能和计算资源，并存储全量 AI 模型，AI 模型的版本越高，模型结构越复杂，精度越高。云端算力资源强大，计算时延最低。基站提供 AI 任务控制功能、通信与计算功能以及 AI 模型管理。相比于云端，基站缓存空间有限，只能缓存部分 AI 服务模型。端侧具有一定的计算资源，用户可以处理神经网络部分层，与基站或者云端进行联合拆分推理。

如图 5 所示，基站作为一个智能体，从云端缓存多个 AI 服务，可向用户分配带宽和算力，基站间可协作。用户申请 AI 服务，上报 QoAIS（模型准确性、时延）需求给网络，基站的 AI 任务控制功能输出 AI 服务质量保障策略（即给出空口通信带宽和计算资源的分配决策、AI 模型的缓存决策），由云边端协同执行 AI 推理任务。

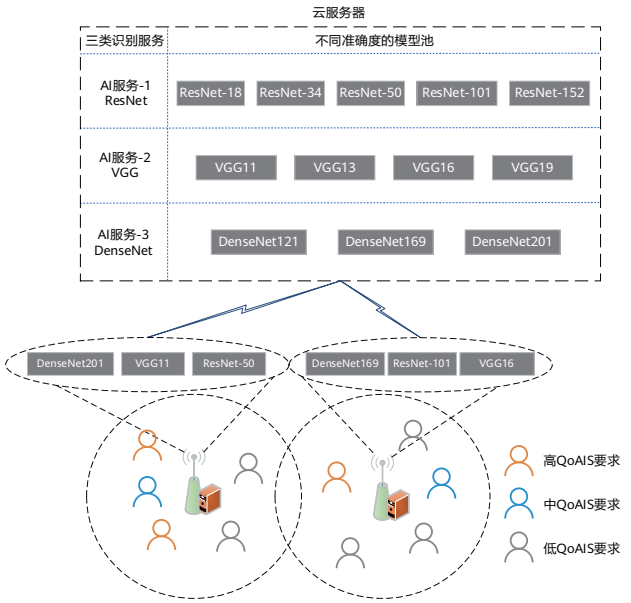


图 5 云边端协同的 AI 推理示意图

评估 AI 推理服务质量的指标可以包含准确度和时延等。AI 服务准确度得分 $Q_p^{u_i}$ 计算公式如下：

$$Q_p^{u_i} = \begin{cases} v_b \\ v_b - \Delta \times v_c \end{cases} \tag{1}$$

当分配的 AI 模型满足用户需求或者大于用户需求时，分数为基准分数 v_b；当分配 AI 模型未满足用户需求时，在基础分数上减去惩罚分数，惩罚分数为惩罚因子 v_c 与 Δ 的乘积，表示用户请求的 AI 模型准确度和分配的 AI 模型准确度之间的差异。

AI 服务时延得分 $Q_d^{u_i}$ 计算公式如下：

$$Q_d^{u_i} = \begin{cases} v_d \\ 0 \end{cases} \tag{2}$$

当时延在要求范围内时，时延得分为 v_d；当用户的时延大于要求范围时，时延得分为 0。

为在动态复杂的无线网络状态下，持续保障 AI 服务质量，保障技术方案采用强化学习算法优化 AI 任务的服务准确度和 AI 服务时延，马尔可夫过程的算法状态、动作、奖

励设置如下（智能体 1 和智能体 2 分别表示图 5 中的基站 1 和基站 2）：

- 状态空间：智能体 1 状态：[用户接入基站，用户申请服务索引] * 用户数；智能体 2 状态：[用户接入基站，用户信道状态，实际承载用户计算的基站，用户实际计算的服务索引] * 用户数。
- 动作空间：智能体 1 动作：[基站存储的 3 个 AI 服务] * 基站数；智能体 2 动作：[分配给用户的计算资源、带宽资源] * 用户数。
- 奖励函数：奖励函数与优化目标强相关。设置如下：

$$R_1 = \left(\sum_{u_i \in U} (Q_p^{u_i} - q_1 - q_2) - \sum_{j=0}^{1} q_3 \right) \times \rho_1 \tag{3}$$

$$R_2 = \sum_{u_i \in U} Q_d^{u_i} \times \rho_2 \tag{4}$$

式中，ρ_1、ρ_2表示缩放因子。考虑到缓存策略在一定程度上影响用户的计算位置，q_1、q_2分别为两个基站之间、基站与云端之间性能损失引入的惩罚因子。q_3为基站重复缓存引入的惩罚因子。

考虑到资源分配策略之间的相关性会影响资源的分配算法，本文对 MADDPG 算法进行优化，提出 MADDPG-Adv 算法。该算法将通信、计算和模型三种资源的共同分配问题分解为两个子问题：获取缓存的全局策略和获取通信和计算资源的全局策略。此外，MADDPG-Adv 算法还增强了 MADDPG 中不同智能体之间的信息传递。如图 6 所示，不同的代理可以解决对应基站上的不同问题。此外，考虑到 AI 模型的缓存位置（云端或基站）对通信和计算资源的分配策略有直接影响，而在状态观察期间无法直接获取 AI 模型缓存信息，需等待动作执行完成后；因此，本技术方案引入了一种额外的信息传输机制：当动作完成后，将观察到的网络状态信息与传输的缓存信息进行合成，共同完成策略输出，以实现算法操作的异步执行。

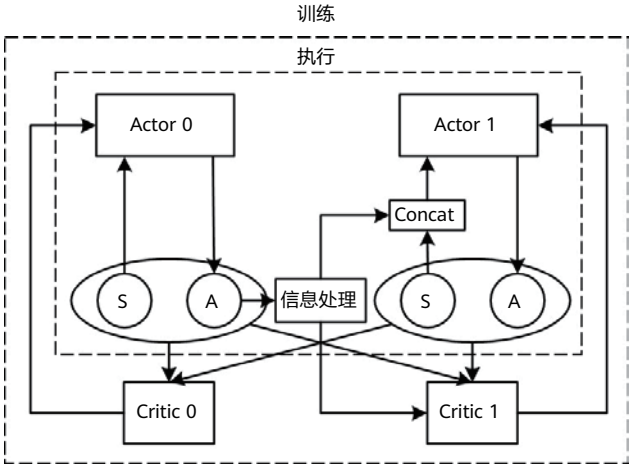


图 6 MADDPG-Adv 架构图

6 仿真方案与结果

在 AI 推理应用场景中，仿真设置包括一组基站和 8 个用户。云端存储 12 种识别类 AI 模型，每种 AI 模型有 3 ~ 5 个模型版本，在仿真验证中，从 AI 服务准确度和 AI 服务时延两个维度判断 AI 服务是否满足 QoAIS 需求，优化目标是最大化 AI 服务准确度得分和 AI 服务时延得分。

本文参考 3GPP TR.38.901，采用正交频分多址（Orthogonal Frequency Division Multiple Access，OFDMA）技术进行系统建模 [14]。用户的上行链路传输速率 e_{u_i} 建模为：

$$e_{u_i} = B_{u_i} \log_2 \left(1 + \frac{\beta_{u_i} h_{u_i} p_{u_i}}{\sigma B_{u_i}}\right) \tag{5}$$

式中，β_{u_i}为用户与基站之间的信道损耗，h_{u_i}为天线增益，p_{u_i}为用户的上行传输功率，σ为噪声功率谱密度。

用户的传输时延T_{t-u_i}可以表示为：

$$T_{t-u_i} = \begin{cases} \dfrac{D_{air}}{r_{u_i}} \\ \dfrac{D_{air}}{r_{u_i}} + T_c \\ \dfrac{D_{air}}{r_{u_i}} + T_b \end{cases} \tag{6}$$

式中，D_{air}为空口传输负载，T_c 表示到云端的往返时延（用户涉及云端计算），T_b 表示两个基站之间的有线传输时延（用户涉及协作基站计算）。

用户设备和基站的计算资源是有限的，分别用C_l 和C_b表示（单位为 FLOPS）。假设云上的计算资源是无限的，可以忽略云上的计算时延。D_l为本地计算负载，D_b为基站计算负载，在一组基站中，由于采用深度神经网络分割计算方法，计算分为本地计算和剩余计算两部分。本地计算或者协同计算时，用户的计算时延可以分别表示为式 7 的上下两部分：

$$T_{c-u_i} = \begin{cases} \dfrac{D_l}{C_l} \\ \dfrac{D_l}{C_l} + \dfrac{D_b}{C_{b-u_i}} \end{cases} \tag{7}$$

为了验证本文提出的算法的性能，将 MADDPG-Adv 算法与 MADDPG 和通信神经网络（CommNet）两种算法进行比较。MADDPG 算法架构允许不同的代理拥有独立的动作和评判网络，使得不同智能体之间的动作更加灵活多样，更容易扩展到高度动态的网络场景。CommNet 算

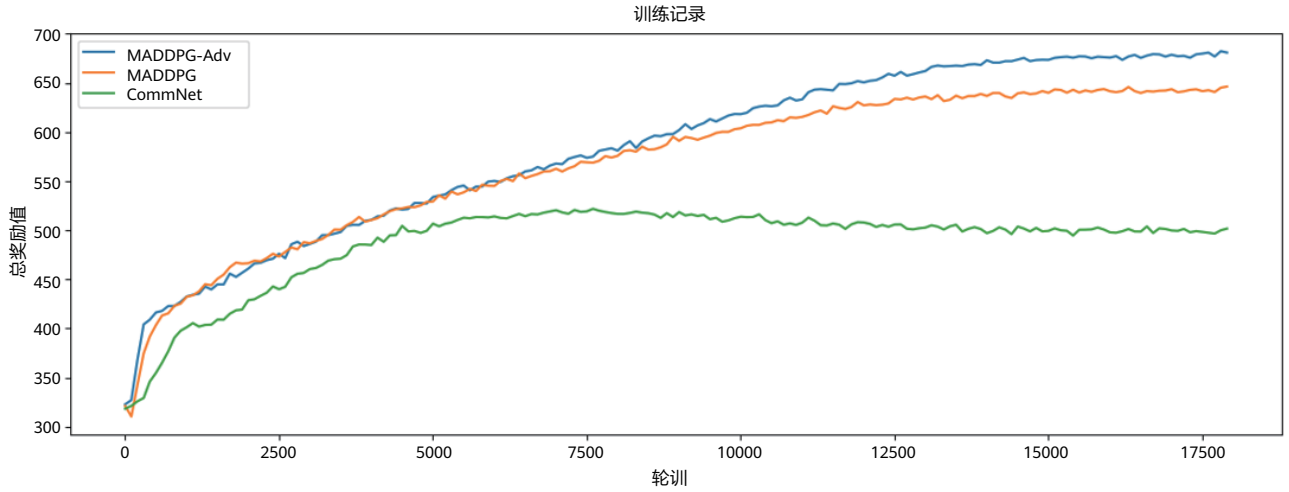


图 7 算法性能比较

法需要更复杂的通信机制来协调智能体之间的学习，算法灵活性受限。如图 7 所示，对比三种算法的训练曲线和收敛结果，MADDPG-Adv 算法奖励值要高于 MADDPG 和 CommNet。MADDPG-Adv 算法的性能比 MADDPG 算法高 6%，比 CommNet 算法高 32%，因此 MADDPG-Adv 更适合实现 AI 任务控制功能。

为了验证系统对用户 QoAIS 性能的保障，可测量同时满足准确度和时延要求的用户任务数量。准确度满足的条件定义为分配给用户的 AI 服务版本大于或等于用户请求的版本，时延满足的条件为用户的时延小于或者等于时延要求范围。如图 8 所示，对比三种算法满足 QoAIS 需求的用户任务数，可以发现 MADDPG-Adv 算法满足 QoAIS 需求的用户任务数最多，相比 MADDPG 算法，满足 QoAIS 需求的用户任务数增加 23%。

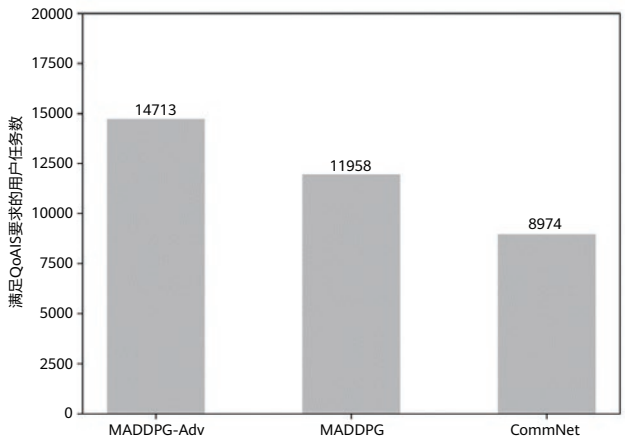


图 8 满足 QoAIS 需求的用户任务数比较

7 总结与展望

本文针对 6G 网络内生 AI 服务质量保障需求，研究 6G 网络 AI 服务质量保障的架构、指标体系、协议设计和保障技术，提出三层闭环管控的 6G 网络 QoAIS 机制及网络功能架构设计，提出 AI 服务质量指标体系、通算 QoS 到承载映射的协议设计；针对 AI 推理服务提出基于云边端协同和多智体强化学习算法的 QoAIS 保障技术，围绕准确度和时延两个指标仿真验证了 AI 推理应用场景的服务质量保障效果，所提强化学习算法相比 MADDPG 算法可以带来 6% 的性能增益，满足 QoAIS 需求的用户任务数增加 23%，验证了 QoAIS 保障技术的性能增益。

由于 QoAIS 的保障需要从架构、协议流程和技术等方面综合实现，后续还需进一步优化通信、计算、模型等资源的融合控制机制与承载，并将所提算法与架构和协议流程的完整设计相结合，针对不同智能服务场景，为资源 QoS、任务 QoS 和服务 QoS 提供更全面的保障效果。

参考文献

[1] Guangyi Liu, Juan Deng, Qingbi Zheng, Gang Li, Xin Sun, and Yuhong Huang, "Native intelligence for 6G mobile network: Technical challenges, architecture and key features[J]," The Journal of China Universities of Posts and Telecommunications, 2022, 29(1): 27–40.

[2] IMT-2030（6G）推进组：无线人工智能技术研究报告 [R]，2023.

[3] Y. Yang *et al.*, "Task-oriented 6G native-AI network architecture," in IEEE Network, vol. 38, no. 1, pp. 219–227, Jan. 2024, doi: 10.1109/MNET.2023.3321464.

[4] 6GANA. 6G 网络分布式学习白皮书 [R]，2023.

[5] T. Chen, Q. Tang, and G. Liu, "Efficient task scheduling and resource allocation for AI training services in native AI wireless networks," in 2023 IEEE International Conference on Communications Workshops (ICC Workshops), 2023, pp. 637–642.

[6] W. Wu, C. Zhou, M. Li, H. Wu, H. Zhou, N. Zhang, *et al.*, "AI-native network slicing for 6G networks" IEEE Wireless Communications, vol. 29, no. 1, pp. 96–103, 2022.

[7] M. Hua, T. Chen, N. Li, and H. Zhang, "Energy-efficient federated transfer learning in 6G native AI networks," 2023 IEEE Globecom Workshops (GC Wkshps), Kuala Lumpur, Malaysia, 2023, pp. 1746–1751.

[8] GGII. 移动机器人行业深度研究报告.

[9] 3rd Generation Partnership Project (3GPP), "TR 22.874: Technical specification group services and system aspects; Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," Release 18, December 2021.

[10] IDC(2023). 中国车联网安全解决方案市场洞察.

[11] 6GANA TG1. 6G 网络 AI 场景用例业务应用需求详解白皮书 [R]，2022.

[12] 3rd Generation Partnership Project (3GPP), "TS 23.503: Policy and charging control framework for the 5G system (5GS)," 2020.

[13] 3rd Generation Partnership Project (3GPP), "TS 23.501: System architecture for the 5G system," 2020.

[14] 3rd Generation Partnership Project (3GPP), "TR.38.901: Study on channel model for frequencies from 0.5 to 100 GHz," 2022.

6G 智能内生网络多维资源联合编排管控技术研究

王栋[1]，郭建章[2]

[1] 中国电信股份有限公司研究院

[2] 中电信数智科技有限公司

摘要

5G 移动通信网络主要围绕核心网推动云化和智能化发展。面向 6G，在端到端网络持续云化和智能化发展的趋势下，智能内生将成为 6G 网络的典型特征，进而需要通信、数据、算力、人工智能模型等多维资源的联合编排管控。本文聚焦 6G 智能内生网络中多维资源联合编排管控的技术需求，系统阐述了 6G 网络架构的展望和智能内生网络方案，并提出了面向连接、智能、感知、安全融合的一体化编排管控方案。同时，进一步介绍了后续技术研究、标准化和原型样机研发的建议。

关键词

6G，智能内生网络，多维资源，联合编排管控

1 引言

随着新型工业化战略的深入实施，我国信息与通信技 术（Information and Communication Technology，ICT） 产业迎来了前所未有的发展机遇 [1]。以通信技术（Communication Technology，CT）的基础上融合信息技术（Information Technology，IT）的 ICT 融合趋势在业务需求驱动的领域中不断发展壮大 [2]。6G 技术是 5G ICT 融合的延续。IT 与 CT 技术不断融合，促使服务化、云原生、AI 等创新技术应用于移动网络。ICT 技术的融合也使得网络基础更加强大，不仅具备工业领域所需的低时延、低抖动、高可靠性等确定性特征，同时还能满足高清视频、车联网和工业互联网等领域对大带宽网络基础的需求。

与此同时，伴随着 5G 网络的规模商用，5G 时代的 ICT 融合围绕核心网云化，控制面、用户面分离的分布式网络架构已逐步推进并趋于完善。面向未来，业界已经系统性开展了 6G 技术研究，而 6G 网络架构是其中重要的研究领域，也是 6G 愿景落地的重要环节和技术手段。以沉浸式 XR、元宇宙、数字孪生等为代表的新业务对网络提出了新的需求，为满足此类业务在高带宽的同时也要满足低时延、网络确定性以及边缘高算力的需求，端到端云网融合是 6G 网络的必然发展趋势 [3]。有鉴于此，6G 网络架构需要融合最新技术进一步助力。

因此，当前 ICT 技术发展呈现出多元、融合、智能化的趋势，数据技术（Data Technology，DT）和运营技术（Operational Technology，OT）融合引入正不断驱动下一代移动网络变革和能力升级。在一方面，引入 DT 技术将在网络智能化增强和用户体验优化方面起到有力推动。人工智能（Artificial Intelligence，AI）大模型能力的持续跃升，智能算力需求的异军突起，以及智能算力设施多元化、普遍化的发展趋势，都标志着智能化技术将成为未来产业发展的新引擎。在另一方面， OT 的引入也将加快网络化自动化和智能化改造，提升网络安全、数据安全、业务安全等方面能力。《工业互联网创新发展行动计划（2021–2023 年）》中指出，要加速工业设备的网络化改造，推进企业内部网络的升级，促进 IT 网络与 OT 网络的融合。

此外，为了应对未来个人用户对 6G 网络的服务需求，6G 时代的移动通信网络应更智能、更弹性、更安全。6G 网络需要具备自主学习的智能使能和智能内生的能力，可以根据用户实际的情境感知信息、业务体验和个性化需求，进行智能化决策和自适应组网。这同样要求需要具备从感知资源层到功能控制层，再到服务应用层的无处不在的分布式计算和智能内生能力，在此之中智能内生在优化和增强网络组织和智能服务能力方面将发挥关键作用。综上所述，6G 网络需要具备对包括连接、数据、算力、模型、安全等多维资源的联合编排管控能力，从而满足用户日趋多样化的个性需求。

2 现状与挑战

IMT-2030（6G）推进组在 2021 年就发起了“面向 DOICT 融合的 6G 网络架构技术发展”倡议。推进组在 6G 网络架构领域的阶段性成果中，明确提出了面向数据、运营、信息、通信技术（Data, Operation, Information, and Communication Technology，DOICT）融合的网络发展方向 [4]。面向 2030 年及未来，5G 已无法满足沉浸式云 XR、全息通信、智慧交互、数字孪生等新兴业务场景的需求，业界已经认识到 6G 时代需要一个高度智能的自动化网络，进一步提升网络智能化水平。作为 CT 技术的重要呈现，移动网络已经充分引入了 IT 技术，将 NFV、容器、SDN、基于 API 的能力开放等技术在系统中获得充分应用 [5, 6]。面向未来将有更多的来自生产运营的需求，通过 OT 技术为移动网络带来新的基因，DT 技术也将为网络演进注入新的活力，数字经济的发展基础是海量连接、数据采集以及建模和分析。然而，目前 DOICT 融合仍处于初步阶段，如何有效编排和管控 6G 智能内生网络中 DOICT 融合，是一个需要重点研究和突破的关键问题。

此外，在智能使能和智能内生方面，当前网络架构仍需要进一步改进以实现 6G 网络的愿景。目前的 5G 网络主要采用集中式的独立 AI（例如 NWDAF），各网元（如 AMF、SMF 等）本身缺乏 AI 能力，需要依赖于独立 AI 的集中分析。然而，这种集中式的处理方式可能带来一些问题。首先，集中式 AI 需要大量数据，这会严重消耗通信网络资源。其次，在规模化组网下，集中式处理、分析和反馈方式难以满足时效性等要求 [7]。此外，AI 分析通常需要算力资源，而集中的 AI 分析则需要高度聚合的算力资源，导致难以有效利用多点和边缘的算力资源。如何高效利用网络 AI 能力，包括多智能体协同等 AI 资源的协同应用，实现内生智能和智能使能的 6G 网络的愿景目标，各界针对此课题纷纷开展研究。

行业内，ETSI 设立了体验式网络智能（Experiential Networked Intelligence，ENI）、零接触网络和服务管理（Zero-touch network and Service Management，ZSM）等标准化工作组，开始了与自智网络和网络使能相关的标准制定和研究项目，并已发布了一些与网络智能化相关的框架、应用、接口等方面的技术报告和标准。3GPP TS 28.535 [8] 中定义的闭环控制（Closed-loop Service Level Specification Assurance，COSLA）是用于通信服务保证的管理控制环，由监控、分析、决策和执行组成。用于通信服务的资源的调整是通过管理控制环中的步骤的不断迭代来完成的。通信服务的闭环控制部署在准备阶段，在生命周期管理的准备阶段和运行阶段生效。后续，仍需继续深化面向多维智能要素的闭环控制研究。

在学术界，6G 网络智能使能和智能内生的研究也正如火如荼地开展。北京邮电大学的张平院士团队 [9] 提出了一项名为 Ubiquitous-X 的 6G 信息交换中枢网络，基于对智

慧和意识的理解，将“人–机–物–灵”全面融合于现有网络中，为未来的 6G 网络构建了新的通信对象。文献 [10] 提出一种 AI 使能的智简 6G 无线接入网体系架构，该架构面向空天地海一体化组网需求，基于雾无线接入网，构建了通信、计算、缓存和控制协同的柔性可重构体系。文献 [11] 提出 AI 驱动的 6G 智能内生网络架构，在用户层、边缘层、中心层和核心网层引入 AI 模块，使网络具备智能感知和预测能力，然后进行集中式和分布式的卸载、调度等决策。文献 [12] 进一步提出**智简无线接入网**，提出采用单一的一体化无线接入网体系结构，实现广覆盖、巨容量、巨连接、超低时延和高自组织等性能目标。构建 4C 协同和 AI 使能的“外简内繁”柔性可重构无线接入网体系架构。

智能内生已经成为 6G 网络的核心要素，如何实现 6G 智能内生网络的高效服务，围绕网络资源的感知管控和用户差异化需求，需要深入开展 6G 多维资源联合编排管控技术研究。

3 6G 网络架构展望

新技术的跨域融合，将促进 6G 网络架构的快速发展以及行业的数字化变革。6G 是通信技术、算力网络、AI 技术等深度融合的新一代移动通信系统，呈现出极强的跨学科、跨领域发展特征。DOICT 深度融合是新一代移动通信系统发展的趋势。6G 将在 5G 的基础上全面支持整个世界的数字化变革，实现新一代信息通信技术与各产业深度融合，以数字化变革催生发展新动能。

面向未来，6G 网络架构的演进必须以用户/客户需求为中心，以云网融合为指导，全面提升移动网络服务能力，并全面升级用户的业务体验。因此 6G 网络需开展端到端一体化架构设计，以具备更强的灵活性和能力开放，从而实现多维资源的端到端编排管控。

新型 6G 网络架构需满足以下需求：

首先，针对新业务新场景的需求，如沉浸式 XR、全息通信、人机交互、机机交互等，必须展开面向“连接+”的 6G 网络架构研究。这些业务不仅需要基础连接服务，还需要边缘算力服务、高精度环境及物体感知服务、网络 AI 服务等。因此，6G 网络将综合提供连接、感知、智能、算力等一体化服务。

其次，需要吸收技术创新与融合的精髓，同时兼顾技术先进性和可落地能力。因此，6G 网络架构应结合移动通信自身演进趋势和 DOICT 融合发展趋势进行设计，评估新兴技术对 6G 网络的影响，如 AI、云计算、大数据、区块链等。此外，需要注重端到端体系化设计，避免出现割裂和断层。在关键技术方案和协议的设计方面，需要终端、无线网络、核心网络的端到端横向协同，同时也需要跨层跨域的纵向配合，基于多维资源联合编排管控实现对包括运营管理域、承载网络域、云网资源域在内的联合编排管控。

再者，需要加强 6G 网络的灵活性设计，以构建更加开放的产业生态。6G 网络应支持按需编排和部署、业务按需加载、流量按需调度，从而增强网络对业务的适配能力。

最后，应充分考虑 2B 领域的需求，进行统筹设计。在数字经济发展和产业数字化转型的过程中，5G 网络发挥了重要的作用。面向未来，6G 网络服务 2B 领域前景广阔，在数字化转型中将发挥越来越重要的作用。实践证明，2B 和 2C 的网络在业务特征、部署需求、管理方案等方面存在明显差异，因此 6G 网络架构设计应加强对 2B、2C 等的统筹考虑。有鉴于此，如图 1 所示，6G 网络总体框架将是“三层四面”的分层架构。

其中，基于网络云化、服务化的思路，“三层”自下而上分别为云网资源层、网络功能层和应用使能层。

- 云网资源层：作为 6G 网络部署的基础设施和资源依托，云网资源层主要用于承载网络功能。资源类型涵盖各种计算资源（如 CPU、GPU、FPGA 等）、存储资源、网络资源以及异构资源。网络资源方面主要指 IPv6/SRv6 承载网络，实现云化后的网络功能之间的连接。
- 网络功能层：6G 网络应支持的网络功能层，旨在提供“连接 +”的网络服务。这一层面可以进一步分为控制面、用户面、数据面、智能面，以满足不同的服务需求和应用场景。

图 1 6G 网络分层架构

- 应用使能层：该层聚合网络服务能力和通用的应用服务组件，通过能力开放、应用使能框架等方式为应用或周边生态提供服务。这样的设计实现了统一的应用使能管理，为应用的开发和部署提供了便利。

云网运营管理贯穿和覆盖各个层面的运营管理，包括继承当前的网络管理、业务受理、计费结算等功能。随着云网融合的演进趋势，6G 网络的运营管理还将引入新的功能，如云网智能编排调度功能、数字孪生网络管理功能、基于区块链的共建共享网络管理等关键功能。这些新功能将提升网络的灵活性和智能化水平，促进网络资源的高效利用和业务的优化部署。

“四面”是在 3GPP 对移动核心网描述基础上的延伸。面向“连接+”的功能需求，网络功能层的“四面”逻辑功能如下：

- 控制面：作为网络控制的核心，控制面实现了对网络连接服务、智能服务、算力服务、感知服务等功能的统一管理。它与其他层级密切协作，完成多种接入方式的整合控制、身份验证、移动性管理、会话管理、策略控制、算力资源分配和管理等一体化管控。
- 用户面：用户面支持网络的可编程性，允许灵活定义数据处理策略。其主要功能包括隧道管理、数据流标识、业务感知、确定性通信保障、数据封装、数据转发以及流量导向等。作为用户与数据层之间的策略、动作和路由节点，以实现用户数据、环境物体感知数据、AI 任务数据等各类数据的处理和传输。
- 数据面：为了进一步分离数据和业务逻辑，网络引入了独立的数据面，负责统筹管理各类数据，并通过标准接口向控制面、用户面、智能面提供数据服务，包括静态数据和动态实时数据，如用户注册数据、网络状态数据、连接数据等。
- 智能面：智能层支持核心网和接入网的全面智能化，是 6G 网络的智能核心。智能层不仅能够支撑和满足网络自身的智能化需求，还应对用户和业务应用的智能化需求提供支持。它提供网络 AI 相关的功能，包括数据建模、模型训练、推理决策、知识图谱、反馈与评估等，以实现网络的智能化管理和服务优化。

4 6G 智能内生网络

随着 AI 技术逐步成熟，智能化技术作为 6G 网络的核心能力已成为业界共识。为满足用户的 AI 服务体验以及网络的智能化运营需求，6G 网络的智能化需要处理分散在各网络节点中的海量、多态、异构、隐私性高的数据和知识，需要支持服务所需的各种模型训练，需要实现推理、决策、评估、优化的全自动闭环等要求。目前 5G 网络采用集中外挂式的 AI 模式，主要聚焦数据分析和预测功能，存在分布式学习能力弱、数据处理不够灵活、缺少决策反馈机制、未考虑数据和知识的积累等问题，因此 6G 网络智能内生需要具备以下几点典型特征：

- 6G 分布式网络需要支持群智协同： 6G 内生智能网络因海量数据的多态异构性、多级节点的能力多元化、高性能机器学习（Machine Learning，ML）的高资源消耗等趋势，呈现广连接、超分布、多点跨域协同的特点。
- 6G 极致服务体验需要满足各种 AI 学习方法：6G 网络服务更加丰富多样，为满足用户体验，6G 网络应支持多种具备不同特点的 AI 技术。除监督学习、强化学习、深度学习等基本的 ML 方法外，应重点聚焦分布式 AI 架构多点协同的学习方法，例如联邦学习、群体学习、多智体强化学习等。此外，还应充分利用知识储备提升效率的学习方法，例如知识驱动的迁移学习、终身学习等 [12, 13]。
- 6G 内生 AI 需要增强数据的感知和处理：引入独立的数据处理和分析网元，网元新增专用的感知处理模块，将数据与转发/控制进一步分离，整体形成分布协同的数据面。通过对数据感知和处理规则的灵活配置、对网络数据实时处理和 AI 分析，实现数据内生。
- 6G 复杂异构网络需要引入自主决策：6G 网络新增 AI 智能决策能力或网元，与各网元后端的推理决策模块交互，形成多点闭环的决策系统。根据网络的实时状态和 AI 的分析推理结果自适应选择或动态执行满足需求和当前网络状态的智能决策，并辅以决策评估和决策验证能力，提供实时高效的服务保障。
- 6G 高效运营需要注重知识的积累和储备：6G 网络应具备知识构建和知识提取能力，能够自主构建网络知识图谱。通过知识仓库，存储和积累网络图谱、专家经验和先验规则等知识；支持网络各节点智能检索并实时获取 AI 服务相关的知识，提升学习效率和智能化水平，最终实现数据和知识双驱动。

5 多维资源编排管控

为实现多维资源联合编排管控，需关注 6G 网络 DOICT 融合的资源维度。一般认为，在网络连接基础之上，借助网络云化，首先应实现通信和智能计算的融合，包括数据存储、模型和算力资源的统一调度；其次，随着无线感知技术的不断成熟，通感一体化（Integrated Sensing and Communication，ISAC）应用也将进一步融合。最后，信息安全问题一直是网络通信要解决的基础问题。面向 6G，也要一体化考虑内生安全需求，将安全性评估作为端到端一体化资源编排管理的重要考虑因素。

图 2 展示的是一种多维资源联合编排管控需求的系统分析方案。6G 智能内生网络的多维资源编排管控需要从广度和深度同时进行融合。在广度方面，以连接为基础，按照技术成熟度逐步融合云存储能力，AI 模型的基本应用。同时，伴随算力需求的提升，对于不同算力资源需求的业务实例，综合考虑算力时延和传输时延，统筹选择最优的算力资源池完成在本地难以及时完成的运算需求。在深度方面，目前通信技术和云计算技术分属于不同的专业，相互依存又相互独立，标准体系也差距明显。二者的深度融合将是一个长期的过程，也是一个非常有挑战性的课题。

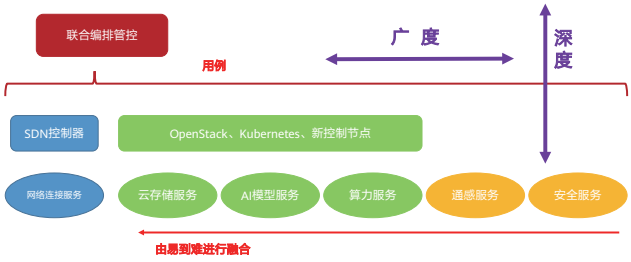


图 2 多维资源联合编排管控需求分析

目前，通感一体化仍局限在接入网，利用无线空口电磁波开展感知应用。感知业务基于独立的控制器和数据处理，暂未实现与通信系统的融合。未来随着通感融合的应用成熟，将逐步基于与移动通信相同的核心网和业务平台开展业务应用，共用 6G 核心网的管控能力，并基于通用的智算数据资源进行感知信号和业务的处理。

如图 3 所示，为了给用户提供差异化与个性化的智能服务，灵活调度全局云网算力资源，全业务场景随愿服务针对各种智能化用例，实现对 6G 接入网、承载网、核心网、空天网络域、云算力资源域等的端到端一体化动态编排管理。在闭环自智角度，方案设计了双闭环自智架构，其中外层闭环又包括两类数据流，分别实现意图的创建和下发功能，意图的修改和满足功能。其中，意图的创建和下发的数据流为用户界面（Use-case User Interface, UUI）–> 服务编排器（Service Orchestrator, SO）–> SDN 控制器（Software-Defined Networking Controller, SDN-C）–> 数据收集、分析和事件触发模块（Data Collection Analytics and Events, DCAE）–> UUI；意图的修改和满足功能数据流为 UUI –> AAI –> DCAE –> 策略（Policy）–> SO –> SDN-C –> DCAE –> UUI。内层闭环结构实现用户的意图保证，包括四个阶段，分别为监控到分析（Monitoring to Analysis, M2A）、分析到决策（Analysis to Decision, A2D）、决策到执行（Decision to Execution, D2E）和执行到监控（Execution to Monitoring，E2M）阶段。每个阶段的相应任务由方案架构中模块定义并实现（如意图实例模块、策略模块、服务编排模块等），并在开放的网络自动化平台（Open Network Automation Platform，ONAP）中具体部署实现。本方案采用轻量级解耦方式实现模块解耦合，采用规范的网络意图验证流程和接口，具有较高实施可行性。

意图网络中的意图输入和转译模块基于 UUI 实现。这一模块负责解析用户输入的自然语言信息，获取对应的网络意图需求信息。用户输入的自然语言信息包括输入语音或网

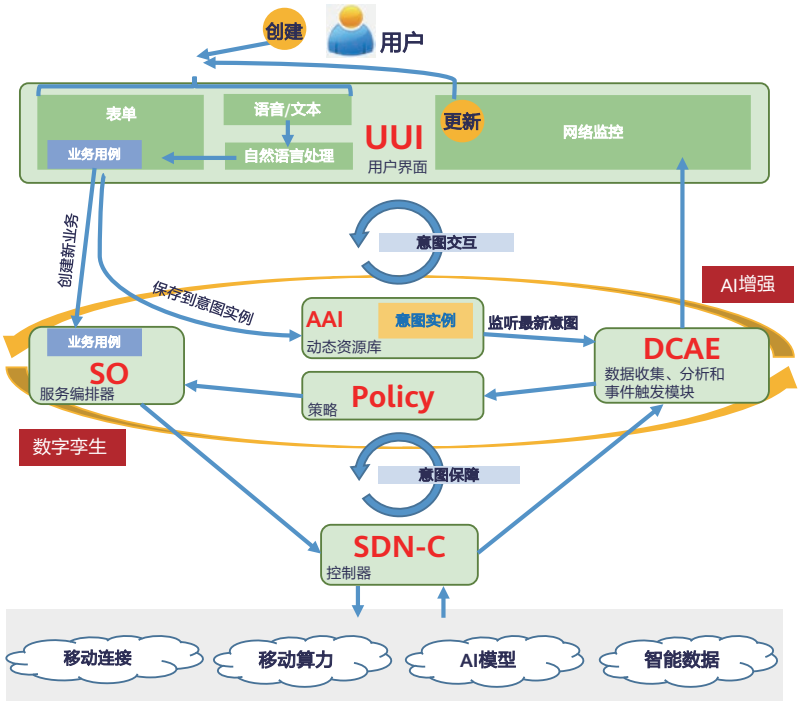


图 3 多维资源联合编排管控架构

络意图需求文本信息等；网络意图需求信息包括网络服务质量需求、网络基本配置信息以及网络状态信息等；网络服务质量需求，即网络配置参数，包括基于用户意图感知获取的网络服务质量需求；网络基本配置信息包括用户位置区域信息等；网络状态信息包括实时流量、时延、抖动等，用于判断当前网络状态是否满足网络意图需求，以对网络配置策略进行调整。用户输入的语音或文本网络意图需求信息，可基于自然语言处理（Natural Language Processing）方法（例如 BERT 算法）进行实体抽取、实体识别等。经过自然语言处理流程，用户使用自然语言表达的需求被转译成相应的网络参数。SDN-C 和 SO 负责进行 SDN 的控制，进而履行意图再向 DCAE 反馈，通过意图验证模块对这些策略进行验证，判断网络状态是否符合服务水平协议（Service Level Agreement，SLA），最终 DCAE 再向 UUI 发送判定结果供用户决策。

意图网络中的意图实例管理技术可实现用户意图闭环管理和网络状态监控。当 DCAE 发现网络状态发生改变即向 UUI 发送通知，请求用户干预。用户通过 UUI 修改存储在 AAI 中的意图实例参数，进而 AAI 向 DCAE 发送修改通知，更新用户意图存储、意图转译结果参数存储、读取和更新网络状态反馈信息存储。

UUI 中的自然语言处理功能将用户意图分类到匹配的用例或服务中，从文本中提取参数，并将其填充到自动生成的服务请求表单中，然后显示给用户以进行确认或修改。服务请求表单是用户意图的标准格式化表达，它指定了服务类型、需求和 SLA。

SO 将客户服务模型转译为服务交付模型，其定义了如何在网络中设计服务。SDN-C 将服务交付模型转化为相应的网络配置模型，并应用于物理网络。

这样的模型驱动设计的先进性体现在三个方面：（1）转译步骤中使用的模型可以通过 SO 进行标准化，这有助于意图驱动网络解决方案的标准化。（2）每个步骤中的模型实例都保存在 AAI 数据库中，并且可以通过 RESTful 接口检索。（3）模型驱动的设计将数据模型与访问 (读/写) 数据的代码逻辑解耦，使代码逻辑能够独立于数据模型改进和发展，从而将对其他组件的影响降到最低。

意图保障的内闭环由 4 个阶段组成：监测、分析、决策和执行。监控阶段在 SDN-C 上实现。它从网络控制器收集监测和性能数据，并转发给 DCAE。SDN-C 根据在意图转译过程中从用户意图中获得的 SLA 参数来确定收集哪些数据。

分析阶段是指 DCAE 对从 SDN-C 接收的监测数据进行分析，并向 Policy 提供反馈。其意义在于检测网络异常，并通知 Policy 采取纠正措施。

决策阶段在 Policy 上实现。它根据从 DCAE 接收到的数据做出闭环决策，并向 SO 发出执行服务更改的适当建议。

执行阶段是典型的 SDN 编排和控制工作流。SO 和 SDN-C 将新的网络配置应用到物理网络中。

所有的数据，包括监测数据、业务模型、网络资源和配置，都保存在 AAI 中，以便在各个阶段检索和共享。

6 意图实例：多维资源编排的需求实体

意图实例是一种面向 6G 智能内生网络的重要实体，包含用户输入的原始意图、意图转译结果和网络状态信息等内容，是为了实现用户意图的有效存储、管理和交互更新所创建的数据实体，用于规范和关联用户意图存储和交互更新、意图转译结果参数存储、读取和更新、网络状态反馈信息存储等。意图实例管理技术可实现用户意图闭环管理和网络状态监控，对用户意图的闭环感知进行反馈。同时，意图实例管理提供了用户原始意图交互信息、意图转译结果信息、网络状态反馈信息等用户信息脱敏数据的汇总导出功能，并建立标准数据规范，可为后续意图网络相关智能算法模型的训练和应用提供数据基础。

闭环编排管理应根据用户意图需求，利用编排器抽象资源状态，并以网络切片、虚拟网络等形式完成不同业务意图的部署配置，实现端到端的服务自动化部署，负载均衡和服务保障。端到端的服务自动化部署技术依托 ONAP，实现复杂计算机系统、中间件和服务的自动化安排、协调和管理。完成全场景下的网络资源自动化、智能化的按需管理和编排。考虑到各类服务需求的多样化和海量化，闭环管理系统可采用容器化部署方式，提供满足高可靠、可扩展的闭环服务，以避免需求激增导致的服务暂停问题。同时，可在边缘侧使用云边协同边缘平台，通过云边自适应协同的方式解决无线通信云计算平台数据量大带来的高负载、无线通信边缘计算平台数据分析的不完整性等问题。

针对意图实例管理模型，需设计可用资源库完成对意图实例和相关信息配置等的存储，承接来自 UUI 的意图转译结果；需设计数据收集分析模块，查询可用资源库中的用户意图处理，处理可能的意图更新；需设计服务编排模块实现意图用例、场景、资源等的管理。

意图实例管理技术应实现意图实例的增删改查。在意图实例创建过程中，UUI 读取用例服务 ID，将用户意图绑定新创建的意图实例 ID 和读取的用户服务 ID 存入可用资源库，同时可用资源库提供用户意图监控接口供数据收集分析模块监听可能的用户意图更新；在意图修改更新过程中，UUI 读取新的用例服务 ID，将新用户意图用例服务 ID 存入可用资源库并同现有意图实例进行关联；意图实例管理技术同时提供意图查询和意图删除功能。

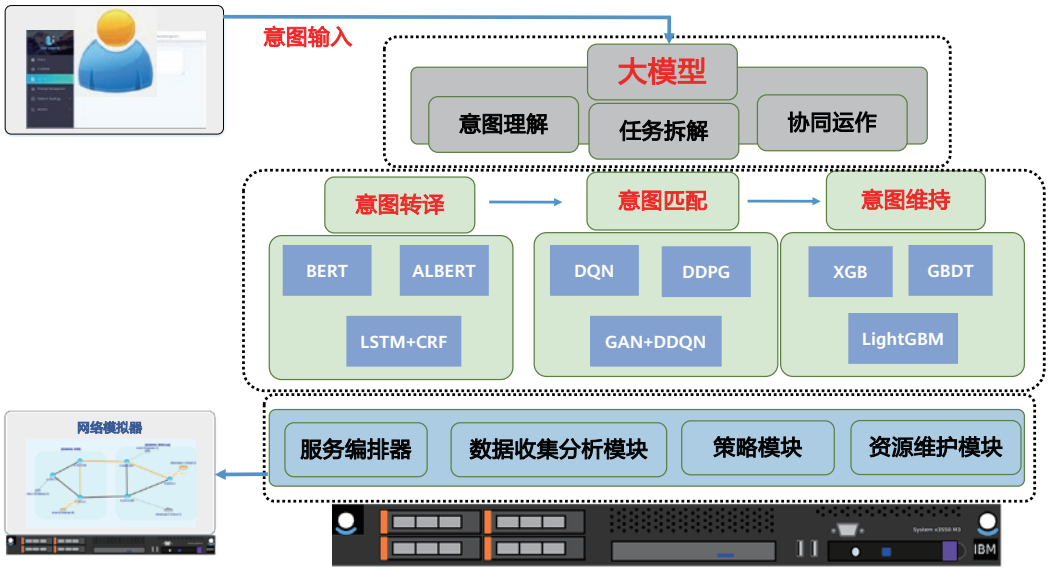


图 4 智能编排管控模型应用方案

采用意图实例管理技术，通过意图实例和数据表间的映射（如以意图实例 ID 作为 Key ID），能够关联意图相关信息，如（1）考虑不同场景用例服务，可将不同场景用例服务 ID 同意图实例 ID 关联，满足用户多场景需求；（2）用户不断变换的意图和意图转译信息，与统一意图实例进行关联；（3）网络变化信息和意图保障信息，与统一意图实例进行关联，能够评估和优化网络对用户意图的支持能力。

意图实例管理技术提供脱敏数据汇总导出功能，将网络意图相关信息导出为后续意图网络相关智能算法模型的训练和应用提供数据基础。

如图 4 所示，意图驱动网络可以将客户的服务意图，经过意图输入、意图转译、冲突检测、意图执行、意图保障等步骤，翻译成一系列的可执行的网络任务。能够根据用户和运营商的意图自动进行转换、验证、部署、配置和优化，以达到目标网络状态，并能自动解决异常事件，保证网络的可靠性。然而，面向全业务场景的差异化需求与复杂的网络环境，当前意图驱动网络仍然面临意图转译不准确、意图策略不灵活、意图管理复杂等问题。

相比于传统 AI 模型，大模型在意图理解、推理和判断决策等方面具备明显的优势。将大模型与意图驱动网络相结合，能够有效提升意图转译准确度，简化意图管理与网络运维流程。将大模型应用于意图识别与转译，能够提升对用户意图和需求识别的准确度；应用于策略匹配与执行，能够智能匹配用户需求与网络配置方案，实现面向用户差异化需求的网络自动化定制；应用于网络监控与维护，能够识别监控网络环境的变化，触发网络自优化，更新网络部署或配置。

网络大模型部署于运营域，打造“大模型为中心，小模型协同”的总体架构。大模型对用户意图任务拆解分类，小模型完成特定任务（意图转译、意图匹配、意图维持）。通过大小模型协同，以意图驱动为抓手，提高端到端编排管控智能化水平。

更进一步，下一代网络架构提出了对于端到端多要素联合调度编排的需求，大模型通过赋能端到端意图驱动，有助于全面优化端到端编排，提高网络性能。

7 未来工作建议

本研究聚焦于 6G 智能内生网络多维资源联合编排管控技术，探讨了以智能内生为核心能力的 DOICT 融合架构方案及相应的多维资源联合编排管控技术。然而，当前 6G 网络架构研究仍处于初级阶段，还有大片空白需要补充。

首先，在 6G 网络的演进过程中，必须更加注重支持多模态异构、全场景接入，并深入探讨如何满足各行各业用户对网络的个性化需求。同时，也需要思考如何在网络架构设计中解决当前网络存在的难以彻底解决的问题。

其次，未来的研究应当致力于进一步优化智能内生网络架构，实现多维资源的一体化管控和调度。提升区域 AI、特定服务 AI 或边缘 AI 的定制化能力，以及完善网元内嵌 AI 的感知处理和决策能力。通过不断优化网络架构，提高网络智能化水平，进一步满足不断增长的用户需求和新兴应用场景的挑战。

最后，未来的工作应满足沉浸式 XR、元宇宙、数字孪生等为代表的新业务的极高的网络性能需求，并进一步拓展智能内生网络的应用场景，包括智慧城市、智能交通、工业互联网等领域。通过将智能化技术与各行

业深度融合，实现智能化服务的全面覆盖，为社会经济发展和人民生活带来更多便利和改变。

综上，围绕 6G 智能内生网络多维资源联合编排管控，仍有较多的研究工作，需要业界一起努力，协同开展原型样机研发和统一标准制定，加速推动实现 6G 全场景按需服务目标。

8 结语

智能内生将是 6G 网络的典型特征，这一观点已是业界共识。围绕 6G 智能内生，需要加强多维资源编排管控能力，以支持智能内生网络的 DOICT 融合。一方面，需要以云网融合为导向，全面提升网络服务能力；另一方面，为满足网络智能化需求，以智能内生为核心，需要强化连接、数据、算力、算法等要素的联合调度。本文还对 6G 智能内生网络多维资源编排管理技术的未来研究进行了探讨和展望。

参考文献

[1] 2024 年信息通信业（ICT）十大趋势 [J]. 互联网天地，2023(12): 12.

[2] 王卫斌，周建锋，黄兵. ODICT 融合的网络 2030 [J]. 中兴通讯技术，2022，28(01): 47-56.

[3] 张平，李文璟，牛凯. 6G 需求与愿景 [M]. 人民邮电出版社，2021.

[4] IMT-2030(6G) 推进组发布 6G 白皮书和技术报告 描绘 6G 未来 [J]. 信息技术与标准化，2021(10): 6.

[5] IMT-2030. 6G 网络架构愿景与关键技术展望 [R]. 2021.

[6] 龙振兴. SDN 与 DFV 技术在电信核心网演进中的应用 [J]. 中国新通信，2020，22(21)：86-87.

[7] 雷波，陈运清. 边缘计算与算力网络——5G+AI 时代的新型算力平台与网络连接 [J]. 中国信息化，2020(12)：113.

[8] 3GPP.TR 28.535, "Management and orchestration; Management services for communication service assurance; Requirements [R]," 2024.

[9] Zhang P, Xu W, Gao H, *et al*. "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks[J]," Engineering, 8: 60-73, 2022.

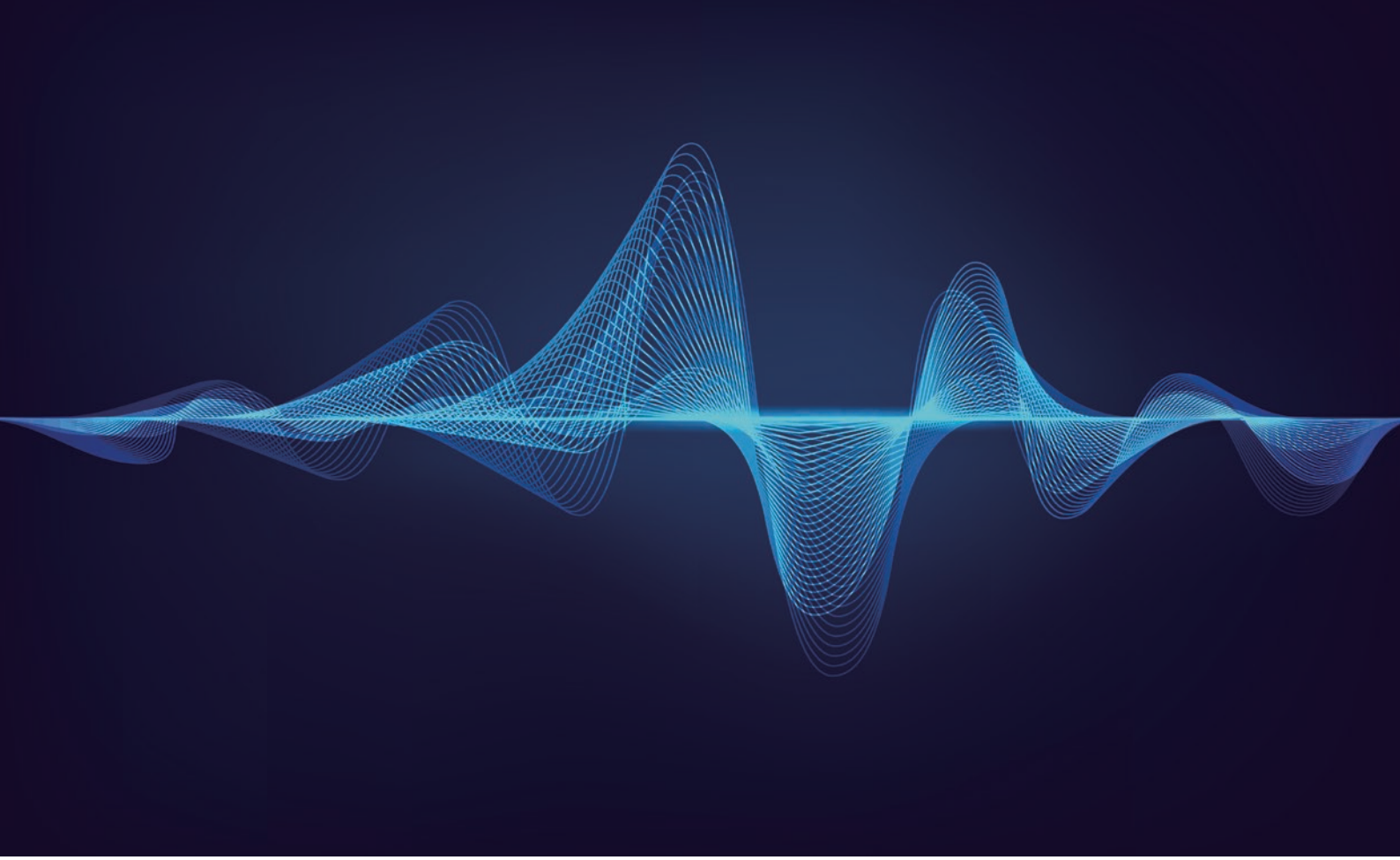
[10] 姜宁，章川扬之，亢晨宇，等. 基于通信感知计算融合的低轨卫星网络体系架构与关键技术 [J]. 无线电通信技术，2023，49(05)：842-852.

[11] 张彤，任奕璟，闫实，等. 人工智能驱动的 6G 网络：智慧内生 [J]. 电信科学，2020，36(09)：14-22.

[12] 彭木根，孙耀华，王文博. 智简 6G 无线接入网：架构、技术和展望 [J]. 北京邮电大学学报，2020，43(03)：1-10. DOI: 10.13190/j.jbupt.2020-079.

[13] Yu M, Li P, Xing Y, *et al*. "A method to improve the performance of network data analytics function based on transfer learning [C]," 2023 IEEE international Symposium on Broadband Multimedia Systems and Broadcasting (BMSB). IEEE, 2023: 1-5.

[14] 刘玉芹，邢燕霞，陈鹏. 6G 网络架构展望 [J]. 中兴通讯技术，2023，29(05)：16-20.



面向 6G 的非正交叠加导频传输与接收方案研究

肖寒，田文强，郑旭飞，刘文东，沈嘉
OPPO 研究院

摘要

导频设计是无线通信系统设计的基本问题。当前系统中的导频传输与数据传输存在对无线传输资源的竞争关系，而随着 AI 接收机的引入，强大的非线性处理能力将使得发送端导频设计中一直以来遵循的正交约束变得松弛，这也为 6G 系统带来了重新探索导频与数据间资源分配关系的可能。本文介绍了针对非正交叠加导频及 AI 接收机的技术研究。首先，针对该技术方案的总体框架进行阐述，并通过基础场景下的仿真展示了性能优势；其次，在复杂场景下概述了多种潜在的实现方案，多方验证了该技术的可行性；进一步地，考虑到方案在多流传输等复杂场景的有效性、实际部署时的可扩展性等问题，通过设计一种基于干扰消除的 AI 接收机进一步适应实际部署。上述内容为未来 6G 中有关导频与数据资源分配关系相关课题的技术研究、推进和标准化提供有益参考。

关键词

叠加导频，AI 接收机，干扰消除，多流传输

1 引言

精确的信道估计是无线通信系统的关键问题之一，对于确保通信链路的可靠性和有效性至关重要，而导频的设计对估计性能的影响可谓举足轻重。在现有5G系统中引入了一系列的具有预定义图样及序列的导频[1]，例如解调参考信号、信道状态信息参考信号以及探测参考信号，以支持包括信道估计、资源调度、链路自适应、波束管理等其他多种功能。同时，基于AI的导频设计也在性能提升方面展现巨大潜力，例如基于深度学习的导频序列及图样设计[2–7]。然而，在上述方案中，数据符号与导频符号在时间与频率资源上是正交分配的，即导频与数据相互竞争有限的传输资源。这无疑带来了显著的导频开销，进而降低了数据传输的频谱效率，限制了系统的吞吐性能。此外，对于大量潜在的6G候选技术来说，例如更大规模的多输入多输出（Multiple-Input Multiple-Output，MIMO）、更精细的波束管理、超高速移动、场景感知、高精度定位等[8, 9]，均存在对于导频传输的深度依赖，产生围绕导频的更多样化的设计需求和资源开销问题，进而加剧数据与导频传输在有限传输资源利用上的竞争。综上，亟需考虑设计新的导频和数据的传输策略以面对上述问题和挑战。

本文介绍了针对非正交叠加导频及 AI 接收机的技术研究，其中发送端将导频与数据在时频资源上非正交叠加；在接收端辅以 AI 接收机以进行有效的信道估计及符号检测。该方案实现了导频与数据的时频资源共享，大大提高了频谱效率。具体来说，本文首先给出叠加导频的总体实现框架，以及基础传输场景下的仿真验证；同时，针对更多复杂场景，介绍多样性的实现方式及对应的多方验证结果；为了进一步解决叠加导频在多流、高速等复杂场景的有效性、实际部署时的可扩展性等问题，提出一种引入干扰消除机制的 AI 接收机方案。上述研究内容及仿真验证结果可以较好地展示叠加导频及 AI 接收机的技术优势以及应用潜力，为后续 6G 的相关研究及标准化提供相应的参考和支撑。

2 叠加导频及 AI 接收机

2.1 基本框架

叠加导频方案 [10] 的核心实现框架是将导频符号与数据符号在发送端进行叠加发送，以突破传统的正交传输方式中数据与导频资源竞争的桎梏。通过在相同的时间和频率资源上同时传输导频和数据，该方案可以实现频谱资源的共享从而显著提高系统的频谱效率。

具体来说，如图 1 所示，发送端可以考虑将时频资源内的导频符号矩阵与数据符号矩阵进行加权叠加，生成用于传输的叠加符号矩阵。在频域 S 子载波，时域 T 个符号的

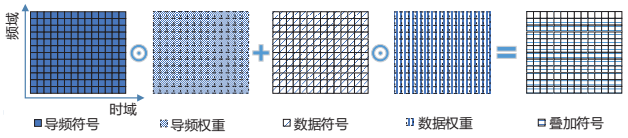


图 1 非正交叠加导频构建方式

传输资源配置下，进行 N_t 发送天线、N_r 接收天线、L 流的下行传输。其中发射机将导频与数据进行非正交叠加获得叠加符号，即：

$$\mathbf{S} = sqrt(\mathbf{W}) \odot \mathbf{D} + sqrt(\mathbf{V}) \odot \mathbf{P}$$

式中，$\mathbf{D} \in C^{L\times T\times S}$ 表示数据符号张量，C 表示复数集合，$\mathbf{P} \in C^{L\times T\times S}$ 表示导频符号张量，$\mathbf{S} \in C^{L\times T\times S}$ 为叠加符号张量，$\mathbf{W} \in R^{L\times T\times S}$ 及 $\mathbf{V} \in R^{L\times T\times S}$，$R$ 表示实数集合，分别表示数据及导频权重张量，$sqrt(\cdot)$ 表示平方根计算，$\odot$ 表示哈德玛乘积。叠加符号经过信道传输后发送至接收端，获得接收信号，即：

$$\mathbf{Y}_r = \sum_{l=1}^{L} \mathbf{H}_{rl} \odot \mathbf{S}_l + \mathbf{N}_r$$

式中，$\mathbf{Y}_r \in C^{T\times S}$ 表示第 r 根接收天线的接收信号，$1 \le r \le N_r$ 及 $1 \le l \le L$ 分别表示接收天线及传输流索引，$\mathbf{H}_{r,l} \in C^{T\times S}$ 表示第 r 根接收天线第 l 条传输流的等效信道，$\mathbf{N}_r \in C^{T\times S}$ 表示加性高斯白噪声。最后，将 N_r 根接收天线获得的接收信号 $\mathbf{Y}_r$ 进行拼接，获得最终的接收信号 $\mathbf{Y} \in C^{T\times S\times N_r}$。

显然，上述框架中每个传输资源单元都同时包含了导频信息和数据信息，这使得 AI 接收机能够利用叠加符号中的导频信息进行信道估计，同时也实现了数据传输的高频谱效率。值得注意的是，在叠加过程中需合理配置导频与数据符号的功率分配比例，以在数据传输的等效信噪比（Signal-to-Noise Ratio，SNR）与信道估计性能之间取得较好的平衡，确保接收端能够高效率地恢复信号。在接收端采用先进的 AI 接收机，对叠加信号有效地处理和接收。AI 接收机的输入是通过调制、导频非正交叠加及信道传输后的接收信号，输出为对数似然比或信息比特流。AI 接收机的实现不仅可以是同时进行隐式信道估计及数据恢复的一体化接收机，也可以是分别进行显式信道估计和数据恢复的模块化接收机。

2.2 典型场景下的仿真分析

在上述实现框架下，本节首先给出针对叠加导频（SIP+AI）在基础配置下（载波频率4 GHz、子载波间隔30 kHz、16QAM调制、LDPC信道编码、SVD预编码）的仿真结果，其中基线（4P+LMMSE）为4符号正交导频设计及LMMSE接收机，用于LMMSE信道估计的协方差矩阵来自10^5 信道样本的统计。AI接收机基于ResNet实现[11]。

图 2 首先给出了在城区宏小区（Urban Macrocell，UMa）信道模型下的块误码率（Block Error Rate，BLER）链路级仿真结果。从图中可以看出，叠加导频方案在 BLER 性能上获得了与基线正交导频方案可比较的性能，这表明叠加导频的引入没有带来额外的 BLER 的性能损失。显然，接收端的 AI 接收机可以有效地进行叠加导频方案下的数据接收。

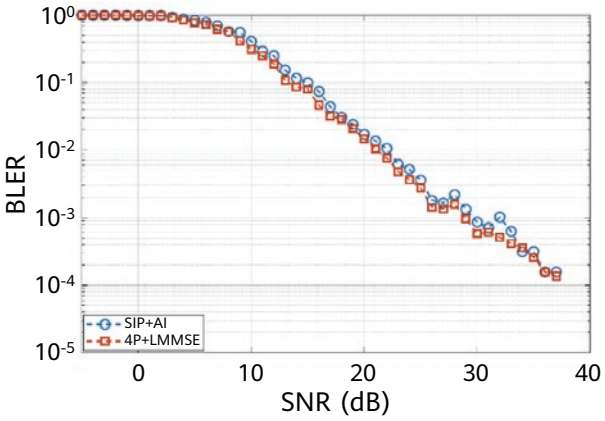


图 2 典型场景 BLER 性能对比（UMa, $N_t = 1$, $N_r = 1$, $L = 1$, $T = 12$, $S = 624$, 300 km/h, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$）

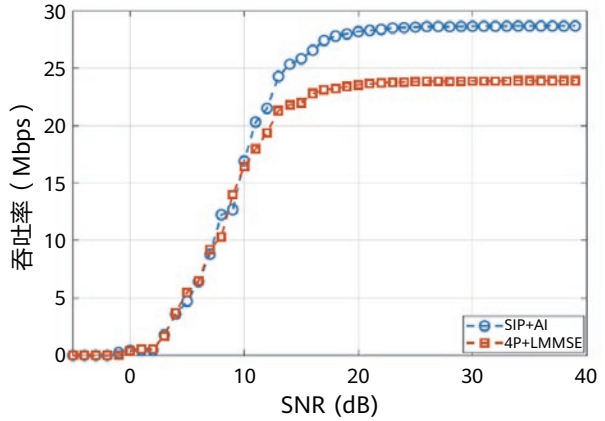


图 3 典型场景吞吐率性能对比（UMa, $N_t = 1$, $N_r = 1$, $L = 1$, $T = 12$, $S = 624$, 300 km/h, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$）

相应地，图 3 给出了对应的吞吐性能对比结果，即在可比较的 BLER 性能下，叠加导频方案由于节省了原来正交导频带来的资源开销，可以在相同码率下传输更多的数据信息比特，从而具有更高的频谱效率、吞吐率的性能优势。

值得一提的是，在针对数据与导频叠加传输问题现有研究的基础上 [11–15]，对于复杂信道环境下，尤其是考虑到多流以及实际终端部署的场景下，叠加导频方案性能以及适配度还有待进一步提升，这对叠加导频及 AI 接收机的多样化改进设计提出了需求。

3 复杂场景及多方验证

3.1 复杂场景下的叠加导频

多流传输利用多个发射和接收天线，允许同时发送和接收多个独立的数据流，可以显著增强通信系统的数据传输能力，提高频谱效率。在现有系统的正交导频方案中，不同流之间的导频是正交配置的，使得不同流的信道可以较好地估计，进一步通过预编码带来的等效信道的低相关性降低不同流间数据的干扰，实现有效的接收。然而，在多流传输条件下，叠加导频较比传统的正交导频方案会引入了更多的流间、流内干扰。具体来说，在多流传输条件下 AI 接收机在接收每一流时，不仅要处理流内的数据、导频的干扰，还要处理其他流的数据、导频对当前流的数据、导频的干扰。因此本文进一步探索叠加导频及 AI 接收机更多的改进设计。在不同干扰程度下，可以考虑使用不同的叠加导频图样。

例如，如图 4 所示，在两流传输下，流间干扰相对较小，仍可以考虑将导频在时频资源上叠加铺满。如图 5 所示，在具有更多干扰的四流传输下，可考虑将不同流的叠加导频在时频资源上正交配置，进一步降低流间的导频干扰。

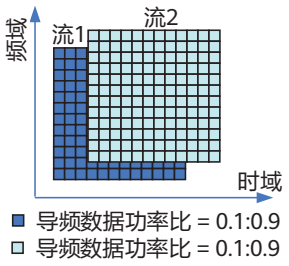


图 4 两流配置下叠加方式示意图

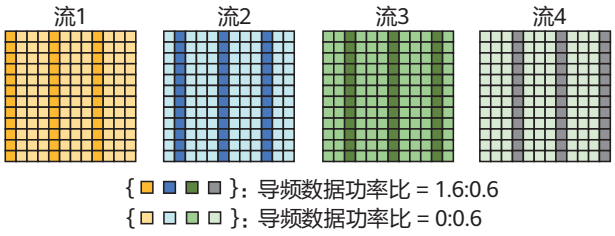


图 5 四流配置下叠加方式示意图

另外，针对处理叠加导频的 AI 接收机的实际部署也应考虑到终端侧模型的存储、计算复杂度等实际挑战。因此，在多流传输模式的叠加导频下，设计终端部署友好的轻量化 AI 接收机，同时实现有效的接收，获得非正交导频的系统增益，是值得研究和攻关的课题。

3.2 潜在解决方案

3.2.1 数据处理及增强

数据在 AI 接收机的训练和构建过程中至关重要，可以说数据本身、数据处理和增强方法直接决定了 AI 接收机最终的性能。可以考虑使用多种数据增强技术来增强模型的接收性能和泛化能力。其中一类是借鉴现有计算机领域的普适性增强方法，另一类可以是借助无线通信领域的知识以衍生出的适应性增强方法。

首先，可以通过计算不同接收信号的线性求和生成新的接收信号样本，有效地增加了模型输入的多样性。此外，可以在天线、流、子载波、时域符号维度上进行数据翻转，可以进一步使得模型学习到复杂且多样的接收信号特征。

在引入通信知识后，一类方法考虑时域和频域的混叠，即将接收信号中某一资源单位的符号与临近资源单位的符号进行加权求和，作为数据增强后的资源单位符号，以模拟信号在实际传输过程中可能发生的符号间及载波间干扰现象。另一类考虑在频域进行剪切使得模型学习到不同带宽下的特征，不仅可以对整个接收信号以不同子带宽度进行接收后进行模型集成以提升性能，从实际部署的角度上更提高了接收的灵活性。同时，还可以针对接收信号增加随机旋转和高斯噪声，以模拟真实链路中的噪声和信号相位的变化，大大提高了模型在环境噪声和信道扭曲影响下的鲁棒性。

3.2.2 模型设计

在潜在的模型设计方面，不仅可以借鉴现有 AI 领域的高效模型架构，也可以基于此进一步进行一系列针对通信信号本身的优化。首先，一类是采用基于 Transformer 的架构，如 Swin Transformer 和 Vision Transformer，这些模型通过其自注意力机制大大提高了对时域和频域维度特征提取的能力。上述实现中的窗口划分的策略设计主要需考虑到不同的频域和时域尺度，以适配具有高度时空关联性的通信数据。另一类考虑卷积神经网络，在现有的卷积网络下，进一步针对通信场景进行优化改进。通过将传统的单一卷积层分解为分组卷积和线性层，显著优化了计算资源的使用。为了进一步提升特征提取效率，可以引入通道分割机制，并应用不同类型的卷积处理，如方形、横向和纵向卷积。这种设计不仅丰富了特征的维度，还通过直连映射保留了原始输入的关键信息。进一步地，潜在方案中可包括对多类基础模型进行联合设计，例如考虑将卷积层内置于 Transformer 的结构、或将 MLP-Mixer 内置于 U-Net 结构。

3.2.3 轻量化

在面对非正交导频与数据叠加传输问题时，考虑到终端的部署落地，模型的存储和计算复杂度成为关键考量。轻量化设计成为必要的策略。首先，可以考虑使用半精度存储，即参数量化。通过将模型参数从浮点数转换为低位宽的整数形式，进一步减小模型的内存占用。其次，网络结构的简化也是实现模型轻量化的一种有效手段。例如使用深度可分离卷积代替传统的卷积层或者优化模型的连接方式，不仅保持了模型处理高维信号的能力，还显著降低了其计算和存储复杂度。最后，知识蒸馏作为一种模型压缩技术，允许将大型模型的知识转移到小型模型中。通过上述方法，小型模型能够在保持较低资源消耗的同时，达到与复杂模型相似的性能，从而应对终端侧部署时计算资源有限等挑战。

3.3 多方案验证

针对 3.1 节所述叠加导频传输假设，由 IMT-2030（6G）推进组、IMT-2020（5G）推进组 5G 与 AI 融合研究任务组共同主办的 2024 6G 无线通信 AI 大赛围绕“非正交导频与数据叠加传输接收方案设计”开展了一系列的方案给出与验证工作。从最终结果来看，通过不同的数据处理方案、多样化的模型架构和创新优化、大量独创方案均可实现针对叠加导频传输的有效接收，展示了叠加导频的接收机设计方案的解法多样性以及其广泛应用的潜力。

表 1 首先给出了 3.1 中两类叠加导频配置下对应两个场景的系统配置参数，充分考虑了不同复杂场景下的评估。其中场景 1a 考虑到了多流下的流间/流内干扰、高速带来的复杂信道、不同速度下的泛化性问题；场景 1b 考虑到终端部署的复杂度问题；场景 2 考虑到多流下更高的流间、流内干扰以及较高阶调制的低抗干扰问题。

多方队伍针对潜在增强点的多种不同组合进行了多样化的总体方案设计。表 2 给出了这些队伍在 SNR 为 5～35 dB 下的平均数据恢复精确度（1-BER）。可以看出，在紧张的赛程下，较多参与者能够在短时间内针对叠加导频问题给出性能尚可的解决方案。

图 6 给出了叠加导频方案（SIP+AI）与 4 导频符号的传统 LMMSE 方案（4P+LMMSE）的 BLER 性能对比。可以看出在复杂场景下，叠加导频方案可以获得与传统正交方案可比较的性能。图 7 进一步给出的吞吐率性能对比中可以看出，显然由于节省了正交导频的资源开销，叠加导频方案可以在有效工作的 SNR 区间（即 BLER 为 10^{-2} 时约 15 dB）

表 1 场景参数

参数	场景 1	场景 2
S	624	96
T	12	12
N_t	2	32
N_r	2	4
L	2	4
调制策略	16QAM	64QAM
移动速度	3 ~ 120 km/h	3 km/h
存储复杂度	1a: ≤ 100 MB 1b: ≤ 20 MB	≤ 100 MB

表 2 数据恢复精确度

方案	场景 1	场景 2
1	0.9393	0.9771
2	0.9390	0.9774
3	0.9391	0.9773
4	0.9386	0.9771
5	0.9386	0.9770
6	0.9386	0.9769
7	0.9384	0.9769
8	0.9384	0.9768
9	0.9382	0.9768
10	0.9390	0.9757

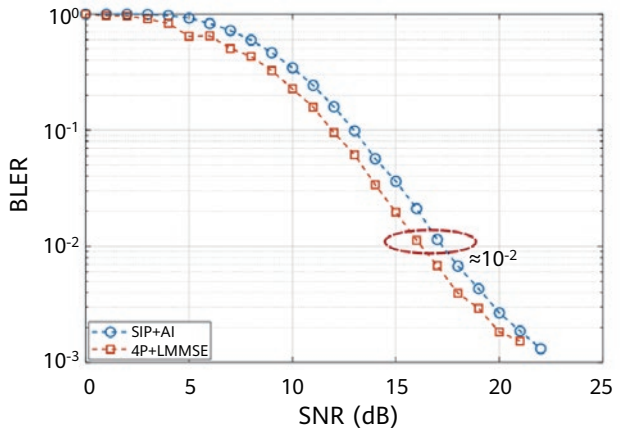


图 6 场景 1 轻量化模型 BLER 性能对比

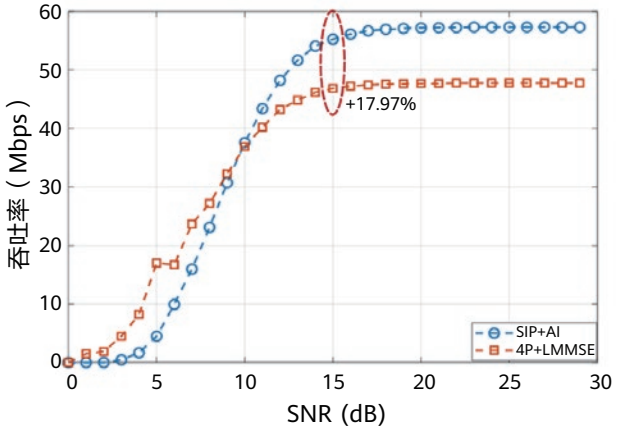


图 7 场景 1 轻量化模型吞吐率性能对比

下获得 17.97% 的吞吐率性能增益。值得注意的是，上述为在存储复杂度小于 20 MB 的轻量化模型下获得的性能。在多方验证中，在保证 90% 以上数据恢复精确度的有效性能下，最小的存储复杂度仅仅小于 2 MB，这可以较好地面对终端实际部署时复杂度的需求。

4 引入干扰消除机制的叠加导频 AI 接收机

4.1 设计思路

在 3.3 的场景 2 中，为了面对叠加导频在多流情况下的流间干扰挑战，引入了图 5 所示的图样。这种发端的特殊配置虽然可以缓解一定的干扰，但实际应用中图样的配置等流程增加了模型生命周期管理、信令配置等流程的复杂度 [16]。同时，接收端单一的 AI 接收机往往不能泛化适配发端不同的调制方式、流数等配置。在发端统一图样配置的情况下，设计能进一步解决流间干扰、具有面向调制方式和流数的可扩展性的 AI 接收机具有较高的研究意义。因此，[10] 中进一步提出了引入干扰消除机制的叠加导频 AI 接收机。特别地，方案在发送端的不同流的导频利用正交叠加码进行正交化，实现在低流间导频干扰的情况下统一图样，简化发端设计，并实现在多流下导频的铺满配置，更好地应对高速场景；在接收端的 AI 接收机设计中引入干扰消除结构，在进行每流的信道估计及符号检测时提前消除流内、流间的干扰；在接收端的 AI 接收机中设计相同模型并行处理不同流的机制以实现不同流数配置下同一 AI 接收机的可扩展性和低存储复杂度；在接收端的 AI 接收机中引入输出的裁剪机制以实现不同调制编码策略（Modulation and Coding Scheme，MCS）下同一 AI 接收机的泛化性。

4.2 仿真性能

在上述设计思路下，给出针对叠加导频（SIP+AI）在载波频率 4 GHz、子载波间隔 30 kHz、LDPC 信道编码的性能仿真结果，其中基线方案（BaselineI）为 4 符号正交导频设计及 LMMSE 接收机。

图 8 给出了 300 km/h 及 900 km/h 高速场景下的性能对比，可以看出提出的方案（Proposed）在导频简化配置下的多流场景同样获得了更高的吞吐率增益。图 9 及图 10 给出了方案在 MCS 泛化性及流数扩展性的对比结果。其中提出的方案（Mixed）同一模型可以在不同流数 L 及 MCS 配置（$m = \{3, 7, 14\}$ 对应 {QPSK, 16QAM, 64QAM} 调制及 {449/1024, 490/1024, 719/1024} 目标码率）下获得与在对应配置下分别单独训练的针对性模型（Specific）持平的性能，印证了较好的泛化性和可扩展性，很好地应对了实际部署时的一系列挑战。

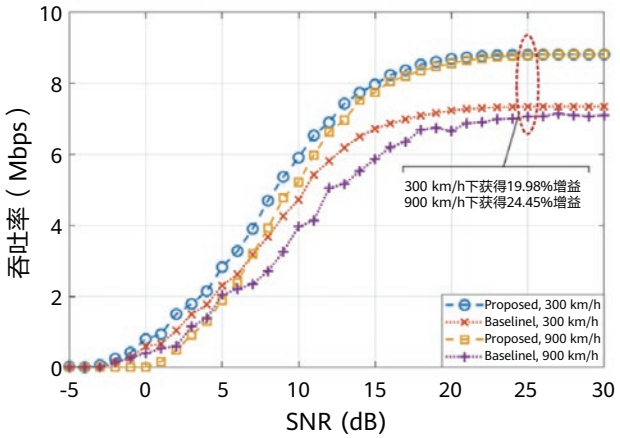


图 8 吞吐率对比（CDL-D 扩展，$N_t = 4$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$）

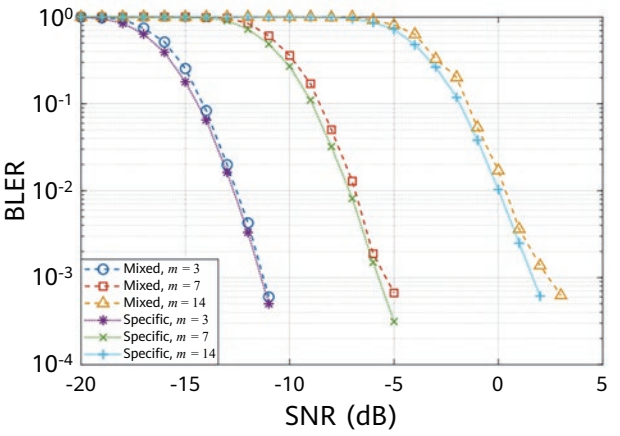


图 9 MCS 泛化性对比（3 km/h, CDL-C, $N_t = 32$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$）

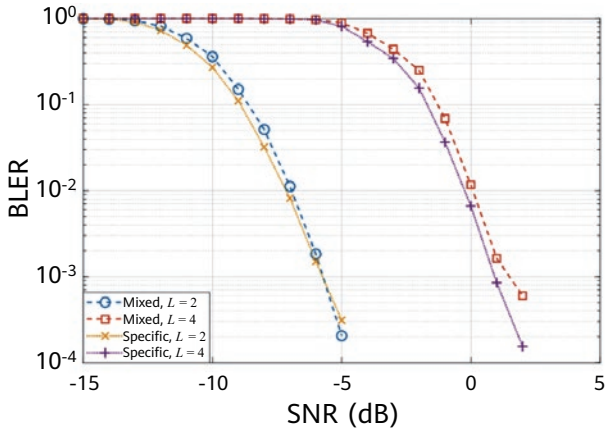


图 10 流数扩展性对比（3 km/h, CDL-C, $N_t = 32$, $N_r = 4$, $L = 2$, $T = 12$, $S = 96$, $\mathbf{V} \in \{0.05\}^{L \times T \times S}$）

5 结语

本文针对非正交叠加导频及 AI 接收机的技术研究进行了介绍。首先，阐述了技术方案的总体框架，并通过基础场景下的仿真展示了性能优势；其次，在复杂场景下概述了潜在的多种实现方式，多方验证了该技术的应用潜力；同时，通过设计一种基于干扰消除的 AI 接收机进一步使该技术适应实际部署；基于上述内容，全面地阐释了叠加导频及 AI 接收机方案可以为导频与数据资源分配关系相关课题的技术研究、推进和标准化提供有益参考，有望为 6G 系统设计带来新的变化和突破。

参考文献

[1] Tang H, Yang N, Zhang Z, Du Z, and Shen J, "5G NR and Enhancements: From R15 to R16[M]," Elsevier, 2021.

[2] Ma X and Gao Z, "Data-driven deep learning to design pilot and channel estimator for massive MIMO[J]," IEEE Transactions on Vehicular Technology, 2020, 69(5): 5677–5682.

[3] Sohrabi F, Attiah K M, and Yu W, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO[J]," IEEE Transactions on Wireless Communications, 2021, 20(7): 4044–4057.

[4] Chun C J, Kang J M, and Kim I M, "Deep learning-based joint pilot design and channel estimation for multiuser MIMO channels[J]," IEEE Communications Letters, 2019, 23(11): 1999–2003.

[5] Xu J, Zhu P, Li J, *et al*., "Deep learning-based pilot design for multi-user distributed massive MIMO systems[J]," IEEE Wireless Communications Letters, 2019, 8(4): 1016–1019.

[6] Soltani M, Pourahmadi V, and Sheikhzadeh H, "Pilot pattern design for deep learning-based channel estimation in OFDM systems[J]," IEEE Wireless Communications Letters, 2020, 9(12): 2173–2176.

[7] Mashhadi M B and Gündüz D, "Pruning the pilots: Deep learning-based pilot design and channel estimation for MIMO-OFDM systems[J]," IEEE Transactions on Wireless Communications, 2021, 20(10): 6315–6328.

[8] Wang C X, You X, Gao X, *et al.*, "On the road to 6G: Visions, requirements, key technologies, and testbeds[J]," IEEE Communications Surveys & Tutorials, 2023, 25(2): 905–974.

[9] Quy V K, Chehri A, Quy N M, *et al*., "Innovative trends in the 6G era: A comprehensive survey of architecture, applications, technologies, and challenges[J]," IEEE Access, 2023, 11: 39824–39844.

[10] Xiao H, Tian W, Jin S, *et al*., "Interference Cancellation Based Neural Receiver for Superimposed Pilot in Multi-Layer Transmission[J]," arXiv preprint arXiv:2406.18993, 2024.

[11] Aoudia F A and Hoydis J, "End-to-end learning for OFDM: From neural receivers to pilotless communication[J]," IEEE Transactions on Wireless Communications, 2021, 21(2): 1049–1063.

[12] Jing X, Li M, Liu H, *et al*., "Superimposed pilot optimization design and channel estimation for multiuser massive MIMO systems[J]," IEEE Transactions on Vehicular Technology, 2018, 67(12): 11818–11832.

[13] Ma J, Liang C, Xu C, *et al*., "On orthogonal and superimposed pilot schemes in massive MIMO NOMA systems[J]," IEEE Journal on Selected Areas in Communications, 2017, 35(12): 2696–2707.

[14] Hoeher P, Tufvesson F. Channel estimation with superimposed pilot sequence[C]//Seamless Interconnection for Universal Services. Global Telecommunications Conference. GLOBECOM'99.(Cat. No. 99CH37042). IEEE, 1999, 4: 2162–2166.

[15] Ye H, Li G Y, and Juang B H, "Deep Learning Based End-to-End Wireless Communication Systems Without Pilots[J]," IEEE Trans. Cogn. Commun. Netw., 2021, 7(3): 702–714.

[16] Chen W, Lin X, Lee J, *et al*., "5G-advanced toward 6G: Past, present, and future[J]," IEEE Journal on Selected Areas in Communications, 2023, 41(6): 1592–1619.

利用语义数字孪生增强无线通信与大语言模型推理的性能

陈佩瑶，葛屹群，张其蕾，史无限，魏哲远
渥太华无线先进系统能力中心

摘要

本文提出一种新的技术，结合通信感知一体化与大语言模型两方面的能力，从传感器数据中提取特征来创建语义数字孪生（Semantic Digital Twin，SDT）。这种 SDT 是包含语义令牌（Token）的一个动态集合，集合中的语义令牌通过特征聚类的一个变种方法（该方法使用同一聚类表示相同的事件或对象）进行更新。与传统的数字孪生不同，SDT 吸纳了历史数据，基于历史数据来整合时空方面的信息，从而增强无线通信系统和大语言模型推理的性能。由于结合了通信感知一体化原理和高级语言处理能力，这种新技术不仅能够实时复制物理环境，还可以全面理解系统行为。无线通信方面，SDT 可以实现精确的波束赋形和个性化的用户定位，从而提升系统灵活性与用户体验。AI 推理方面，通过精确视觉裁剪、上下文感知预测、有效提示工程等手段，SDT 能够增强大语言模型的推理结果，更好地支撑智能医疗诊断和交通等领域的应用。总的来说，这种两相结合的技术为无线通信和语言理解领域的性能提升都带来重要契机，将促进系统向更加智能化的方向发展。

关键词

通信感知一体化（ISAC），大语言模型（LLM），语义数字孪生（SDT）

1 引言

以大语言模型（Large Language Model，LLM）为代表的人工智能（Artificial Intelligence，AI）不断发展，在制造、医疗保健、交通等诸多行业展现令人振奋的前景。从生产线的智能管理，到交通系统的自动化控制，再到医疗诊断的精准预测，大语言模型的应用正逐步重塑我们的工作与生活，为社会创造一个更高效、更安全、更健康的未来。大语言模型能够理解和处理上下文信息，产生连贯、有意义的响应，更好地模仿人类沟通。

随着无线通信技术的进步，6G 无线系统在向更高的频率（包括毫米波乃至太赫兹）、更大的带宽和更大规模的天线阵列等方向演进 [1]。一方面，通信系统已逐步具备类似于感知系统的能力，通过广泛覆盖的移动通信网络，以及对直接信号、反射信号和散射信号的分析 [2]，可以从无线电波中提取距离、角度和材质等信息，实现对目标对象或环境属性与状态的感知。另一方面，感知技术借助高精度定位、环境重构等手段，可以实时复刻物理世界，构建一个平行的数字世界，即“数字孪生”。数字孪生提供精确的波束赋形和高效的信道状态信息检测等功能，有助于增强通信性能。这意味着通信感知一体化（Integrated Sensing and Communication, ISAC）将成为大势所趋 [3, 4]。同时，大语言模型与感知技术的结合，也为传感器和摄像头等设备赋予了智能感知能力。具备了智能感知能力以后，这些设备不仅可以识别、检测和采集多样化的海量数据，还能够分析并优化数据，进而感知和理解外部环境。未来通信感知一体化与大语言模型将不断融合，共同推进 6G“人工智能物联网（Artificial Intelligence of Things，AIoT）”时代的到来 [5]。

基于这样的愿景，在未来的通信系统尤其是 6G 系统中，传感器和机器人等各类智能设备之间的通信会占据很大比重 [6]。大语言模型的出现，让人与机器间、机器与机器间的通信变得更为直观和高效，将大大促进语义通信在 6G 研究中的发展。语义通信减少了数据传输量，不仅传输原始数据，还传递信息内涵，因而可以提高通信效率。利用大语言模型，可以从图像、音频和点云等各种模态中提取信息，并将信息转换为常见的令牌化表示。这些从大语言模型词表中提取的离散令牌，封装了底层数据的语义，无需考虑数据的原始模态。这一机制创造了振奋人心的可能性，让不同设备和系统间得以进行无缝通信与信息交换。借助这种基于令牌的语义通信方法，还可以更方便地把信息集成到知识图谱和其他语义表示框架中，从而促进对环境的全面理解及相关决策。通过上下文感知通信，设备可以根据周边环境和系统整体目标来动态调整行为。

高效通信及人工智能物联网要走进现实，就需要把大语言模型扩展到人类用户之外，去覆盖一张巨大的物联网设备网络。然而，由于这类设备的计算能力有限，直接在设备上执行大语言模型的推理通常不具备现实的可操作性。如果采用传统的云端解决方案，用户与远程数据中心之间来回通信会引入显著时延，难以满足实时应用的即时响应要求。这个问题在自动驾驶或工业控制等时间敏感型任务中尤为突出，此类任务中的时间以毫秒计，微小的时延都会产生重大影响。为了应对这样的挑战，需要有先进无线系统来支持大语言模型的在线推理，特别是在行将到来的 6G 时代，基站层面的支持至关重要。

如此一来，未来的无线通信系统不仅要管理控制无线通信、向用户设备提供连接与通信服务，它的基站还要充当连接 AI 模型的中心枢纽，而当中的每个模型都是为特定的功能和应用而设计。这些模型经过预训练和验证后，策略性地部署到核心网络的各个基站，让 AI 能力更贴近终端用户。图 1 展示了这样一个未来的 6G 系统。

本文提出通信感知一体化和大语言模型相结合的思路，通过建立语义数字孪生（Semantic Digital Twin，SDT）来提升无线通信与 AI 推理的效率。后续章节的内容如下。第 2 节介绍无线通信系统中 SDT 的具体框架，重点说明通信感知一体化和大语言模型的融合。第 3 节探讨 SDT 在无线通信和 AI 推理增强方面的应用。第 4 节给出本文小结。

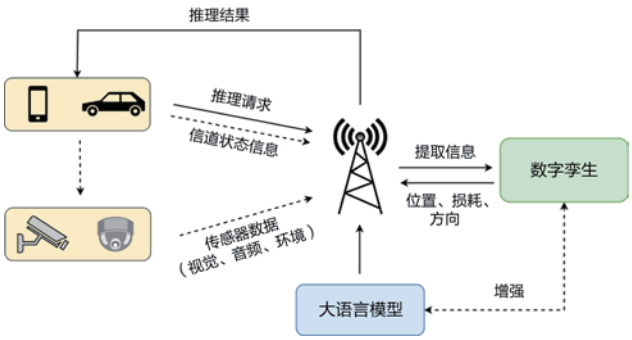


图 1 通信感知一体化和大语言模型在 6G 系统中的融合

2 语义数字孪生

数字孪生的概念正在颠覆我们对复杂系统的认知与管理。通过数字孪生，网络运营商能够识别未覆盖区域、降低信号干扰、高效分配资源，以达到网络性能优化的目的。当网络向 6G 及更远的未来演进，数字孪生将愈发变得不可或缺，因为它作为创建虚拟副本的关键所在，复刻了物理无线环境中从基站、用户设备到周边地形等林林总总的一切对象。这些数字化的表示还会随实时数据不断更新，实现对系统行为的持续监控、分析与预测。

在迈向高效通信愿景的进程中，语义通信和数字孪生技术的结合为未来 6G 智能系统的发展开辟了广阔前景。具体而言，两种技术的结合可以实现实时物理蜂窝网络的令牌化表示。传感器的运用，以及令牌化无线相关特征——比如信道状态信息（Channel State Information，CSI）和信道质量指示（Channel Quality Indicator，CQI）——的信息辅助，对构建准确有效的 SDT 将发挥关键作用。

2.1 语义传感器数据

大语言模型的应用使传感器在处理原始数据时能够理解特定的任务或目标，提高处理效率。换言之，传感器可以将注意力与处理能力聚焦到具体场景和任务的相关方面，针对性地提取更有意义的信息。提取的信息包含两方面，一是与任务相关的语义概念，二是目标的附加属性，这两项信息都通过语义令牌 T^s 来表示。譬如，相机的功能远不止显示基本的图像像素。它还可以检测特定场景中执行具体动作的人，并对人的位置和运动等附加属性进行编码。与此类似，环境传感器除了上报温度值，还可以根据预定义的阈值和环境模型传达“舒适”、“潮湿”或“污染”等语义概念，并在环境异常时发出警报。

2.2 令牌化的无线信道测量

由于整个通信网络充当了一个巨型传感器，利用无线相关特征可以显著增强对物理世界的感知和理解。具体而言，从无线信号中提取的距离、速度和角度等信息可以增强感知。例如，分析信道状态信息可以揭示障碍物的存在，识别不同类型的干扰（如同频干扰或外部干扰源），并检测目标对象在覆盖区域内的运动。这些感知与理解的结果会编码为令牌 T^c，令牌值包括“障碍物”、“干扰”或“运动”等。同时，位置、损耗或方向等相关参数也会一并编码。此外，如果基站拥有相关环境的综合射频地图和充足的算力，那么基于令牌化的表示来重建粗略但完整的信道状态信息，也具备潜在的可行性。这种重建能力使信道状态信息的表示变得更加高效、紧凑，从而在保留无线环境基本信息的同时，减少了需要传输的数据量。

2.3 语义数字孪生表示

如图 2 所示范式，语义数字孪生是包含语义令牌的一个动态集合，令牌会基于传感器和无线信道测量所获得的信息持续更新。每个令牌序列表示环境的特定方面或环境中发生的特定事件，不仅承载了自身固有的语义内涵，还包含了时间和空间维度的上下文信息。这意味着，数字孪生中的每条信息都被打上时间戳和位置戳，进而构建出时间、空间和语义（用于事件描述）环境的三维表示。这样构建出来的数字孪生提供了更为丰富的信息，它的角色不再是被动地收集数据，而是主动地参与到理解和解释环境的活动中来。

上述语义数字孪生是由基站建立的。在建立数字孪生的过程中，每个时间戳的令牌融合是我们面临的主要挑战。设令牌 T^s 和 T^c 的长度相等，或使用额外的神经网络投影来对齐两者的长度。受 [7] 启发，我们根据令牌特征将令牌 $T = \{T^s, T^c\}$ 分成若干聚类，再融合同一特征聚类中的令牌，如图 3 所示。需要注意，同一特征聚类中的令牌对应多个相同的事件或对象，各个特征聚类中融合的令牌数量也不尽相同。本文采用的聚类方法是一种使用了语义令牌的混合特征聚类法。该方法主要包括两部分：一是令牌的最近邻聚类（KNN），也就是先基于特征的空间相似度将令牌聚类；二是令牌融合，即利用大模型再来考察令牌的语义相似度。在模型训练过程中，会使用语义图将属于同一目标或事件的多个语义令牌属性聚合到同一个聚类中。

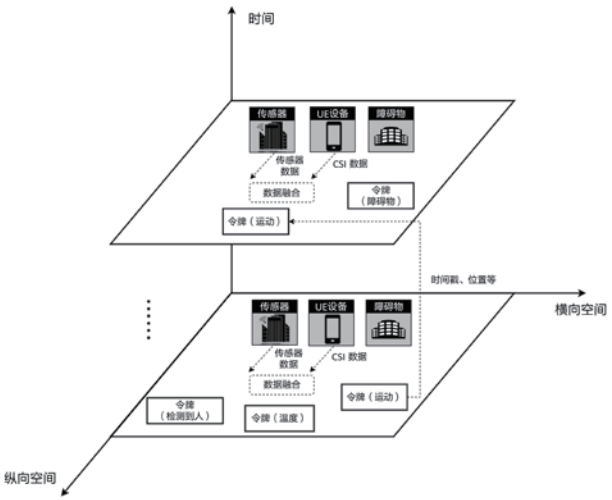


图 2 语义数字孪生

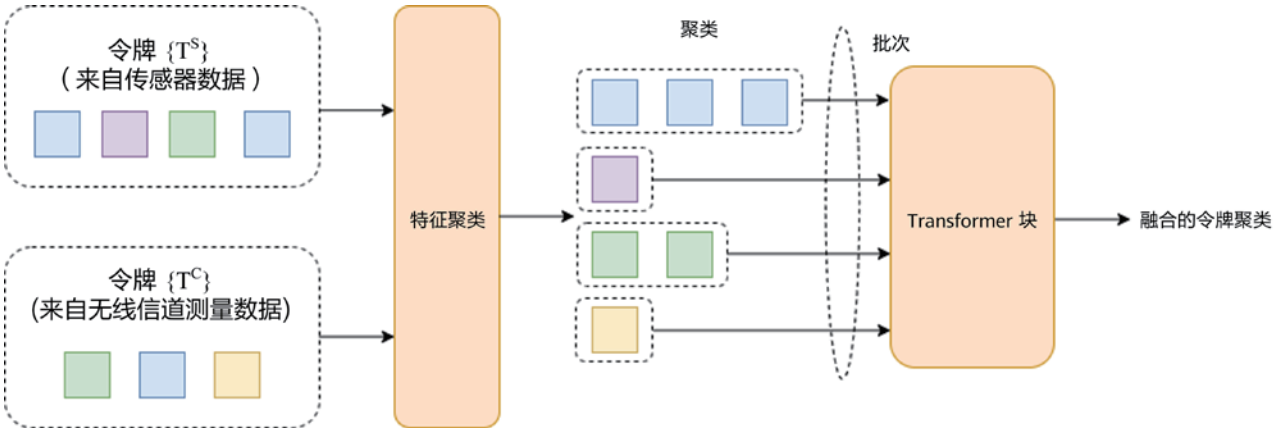


图 3 令牌融合过程

- **特征聚类：**创建特征聚类采用的是 DPC-KNN（Density Peaks Clustering based on k-Nearest Neighbors）[8] 的一个变种算法。由于聚类中心的密度比邻居令牌更大，与较高密度令牌之间的距离也相对更远，因此应同时考虑密度 ρ 和相对距离 δ。给定一组令牌 T，设 $\mathrm{NN}_k(t_i)$ 为在语义相似度上与 t_i 最邻近的第 k 个令牌。那么，t_i 的 k-最近邻 $\mathrm{KNN}(t_i)$ 定义如下：

$$\mathrm{KNN}(t_i) = \left\{ j \in T \middle| \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|} \leq \frac{t_i \cdot \mathrm{NN}_k(t_i)}{\|t_i\| \, \|\mathrm{NN}_k(t_i)\|} \right\}. \quad (1)$$

然后，通过计算 t_i 到 k 个最近邻的平均距离，可以得到令牌 t_i 的局部密度 ρ_i：

$$\rho_i = \exp\left(-\frac{1}{k} \sum_{t_j \in \mathrm{KNN}(t_i)} \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|} \right). \quad (2)$$

相对距离的计算公式如下：

$$\delta_i = \begin{cases} \min\limits_{j:\rho_j > \rho_i} \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|}, & \text{if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max\limits_{j} \frac{t_i \cdot t_j}{\|t_i\| \, \|t_j\|}, & \text{otherwise} \end{cases}, \quad (3)$$

式中，ρ_i 是令牌 t_i 的局部密度。

令 $s_i = \rho_i \times \delta_i, i \in \{1, \ldots, |T|\}$ 表示每个令牌 t_i 的得分。通过选取具有最高得分 s_i 的令牌，可以确定一系列聚类中心，然后根据语义距离将其他令牌分配到最近的聚类中心。

- **令牌融合：**每个特征聚类使用一个 Transformer 块，用于捕捉同一特征聚类中不同令牌间的语义关系和信息交互，以生成融合的令牌聚类 $\tilde{T}_n$。

对于不同时间戳的特征聚类，会基于相似度距离进行配对，也就是说，只有当各聚类的中心之间相似度距离小于给定的阈值 d_c 时，这些聚类才能互相匹配。在执行决策或类似任务时，会考虑跨越不同时间和空间的所有对应特征聚类，这一做法能提高准确度，为物理世界与数字世界之间更和谐的交互开辟了新的可能性。

3 语义数字孪生的应用

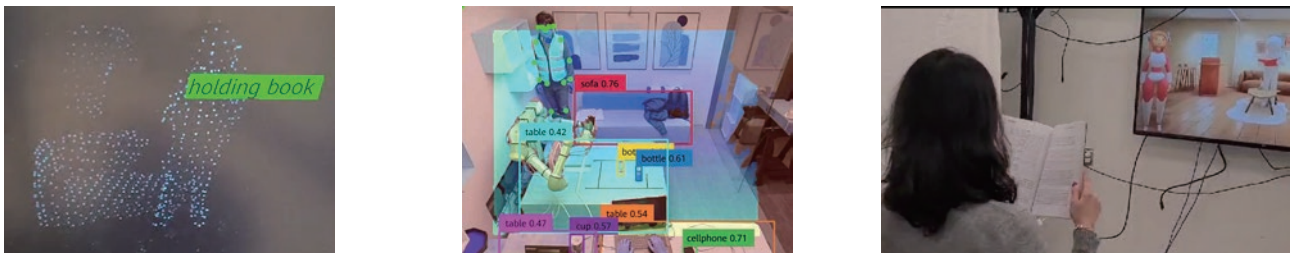
时空 SDT 在无线通信和大语言模型推理中起着至关重要的作用。

3.1 无线通信中的作用

通过对历史与实时数据的分析整合，SDT 可以优化资源分配和信号处理。具体来说，在波束赋形等技术中，SDT 可以精确定位信号传输方向，从而最大化信号接收效率。

在传统的波束赋形方法中，定向传输通常依赖设备或特定信号源的地理位置。而借助 SDT，系统可以识别和理解特定的用户活动或状态，比如识别用户读书时的姿态或行为。这种个性化定位超越了严格的地理界限和信号源限制，关注点转移到了用户行为与需求。基于识别到的信息，系统可以调整天线阵列的波束赋形方向，精确定位到特定的用户设备。此外，系统还可以快速响应用户姿态或环境条件的变化，动态调整波束方向以维持通信的连续性和效率。这些能力增强了通信系统的灵活性与适应性，显著提升了用户体验和业务质量。

图 4 是 SDT 检测到人手持书本的实时演示。演示使用了摄像头和激光器等多种类型的传感设备。为了对齐各类传感设备所采集的数据，我们利用第 2.3 节提出的令牌融合方法来提取特征并匹配多设备间的目标与对象。检测分为环境检测和语义检测。前者检测静态对象，这可以由当前的大语言模型来轻松处理。后者理解并检测人的动作，需要分析并整合目标个人和周围物体的相对位置与状态。演示中维护两个队列：一个语义状态队列 $S(p)$；一个相对位置队列 $L(p, o)$，指示对应语义状态的人和物体所处的相对位置，其中 p 表示被检测人的索引，o 表示被检测物体的索引。之后，利用语义状态和相对位置间的对比学习来提高人体姿态检测与理解的精度。整个过程如图 5 所示。



(a) 语义检测　(b) 环境检测与定位　(c) 可视化的 SDT

图 4 人手持书本的 SDT 演示

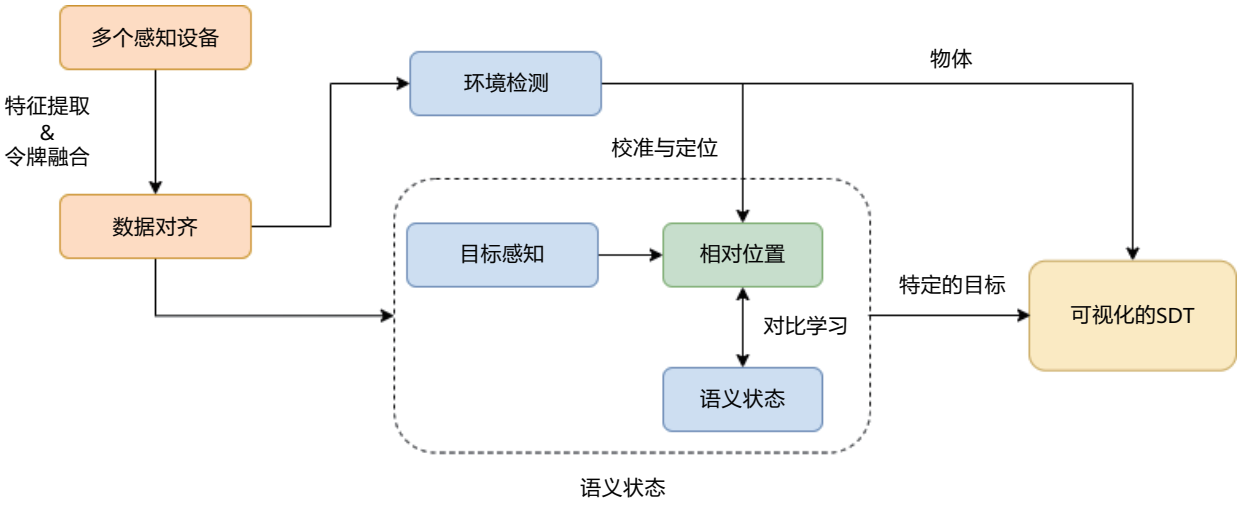


图 5 SDT 构建过程

3.2 对 AI 推理的增强

在以下几个关键方面，SDT 可以显著增强 AI 推理能力。

- **精确视觉裁剪：**使用多模态大语言模型处理视觉问答任务时，有效的处理性能对医疗诊断和智能交通应用极为重要。如 [9] 介绍，问题中视觉主体的大小显著影响模型灵敏度。较大的视觉主体往往会提高相关问题回答的准确度。相反，较小或模糊的细节常常增加模型处理难度，削弱模型有效处理细微视觉线索的能力。因此，精确的图像裁剪可以让模型聚焦关键视觉区域，从而显著提高视觉问答任务的准确度和效率。传统方法（如 [9]）专注于单图像裁剪，而 SDT 与之不同，通过令牌表示来提供环境的全局视图，实现更准确的裁剪。
- **上下文感知预测：**当前的视觉大语言模型是面向单图像任务做的优化，缺乏时间记忆。由于视频数据的海量性质，如果直接训练视频大语言模型，资源消耗会非常大。涉及“拿起”和“放下”等动作的任务，需要上下文信息才能做到准确判读，而仅做单帧分析可能无法提供这类信息。要增强推理任务——尤其是对规律性动作或场景进行预测的任务，可以在大语言模型中有效利用 SDT 的时空知识。例如，在机器人疏导拥挤区域的场景中，SDT 可以洞察障碍物和行人移动，并提供社交距离指引，显著提高机器人动作的准确度和有效性。
- **有效提示工程：**SDT 可以分析理解过往的语言数据，借此来改进大语言模型并优化提示工程。经过精调，推理引擎做出的决策能够更充分地考虑上下文与相关信息。以机器人取食物的任务场景为例。如果机器人仅依赖自身的机载传感器，它的能力就受到天然的限制，只能基于当前所处的周边环境开展任务，无法获取历史环境信息。这种情况下，如果附近没有可见的食物，机器人可能无法完成任务。而一旦将 SDT 的时空感知能力融合到推理过程中，就可以将机器人的感知范围扩展到它所处的直接环境之外。SDT 的集体记忆功能可以洞察过往事件和历史环境，填补机器人的知识空白。比如，即使机器人无法直接观察到食物，SDT 也可以提示具体位置信息，告知机器人食物在某个特定抽屉中。基于这种背景知识，推理引擎可以有效地引导机器人成功取到食物。这个例子展示了 SDT 技术与机器人推理相结合所带来的变革性影响，表明 SDT 能够增强机器人在复杂环境中的智能程度和适应性。

4 结语

本文介绍了一种新方法，将通信感知一体化与大语言模型相结合来建立 SDT。在 SDT 中，传感器数据由传感设备和无线信道测量功能来采集，并使用语义令牌来表示。语义令牌又按照特征聚类进行融合。由于整合了历史数据，SDT 可以增强无线通信性能，尤其是能够提供精确的波束赋形和个性化的用户定位。同时，通过精确视觉裁剪、上下文感知预测和有效提示工程，SDT 还能提升 AI 推理任务的精度和效率。这种融合方法为通信感知一体化和大语言模型两个领域的智能系统发展都带来广阔前景。未来的研究还可以进一步探索 SDT 在自动驾驶、智能制造和环境监测等多种应用中的潜力，从而实现物联网技术的全面部署与推广。

参考文献

[1] Wen Tong and Peiying Zhu, "6G: The next horizon – From connected people and things to connected intelligence," Cambridge University Press, May 2021.

[2] Zhi Zhou, Xianjin Li, Jia He, *et al.*, "6G integrated sensing and communication - sensing assisted environmental reconstruction and communication," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 1–5.

[3] Danny Kai Pin Tan, Jia He, Yanchun Li, *et al.*, "Integrated sensing and communication in 6G: Motivations, use cases, requirements, challenges and future directions," in Proceedings of 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S). 2021: 1–6.

[4] Alireza Bayesteh, Jia He, Yan Chen, *et al.*, "Integrated sensing and communication (ISAC) — From concept to practice," Communications of Huawei Research, 2022: 4–25.

[5] Jing Zhang and Dacheng Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," IEEE Internet of Things Journal, 2020, 8: 7789–7817.

[6] one6G, whitepaper, "6G & robotics, use cases and potential service requirements," June 2023. Available online: https://one6g.org/resources/publications/

[7] Wang Zeng, Sheng Jin, Wentao Liu, *et al.*, "Not all tokens are equal: Human-centric visual analysis via token clustering transformer," in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11101–11111.

[8] Mingjing Du, Shifei Ding, and Hongjie Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," Knowledge-Based Systems, 2016, 99: 135–145.

[9] Jiarui Zhang, Mahyar Khayatkhoe, Prateek Chhikara, and Filip Ilievski, "Visual cropping improves zero-shot question answering of multimodal large language models," arXiv preprint arXiv:2310.16033, 2023.

数字孪生在线信道建模：愿景、进展与挑战

李俊伶 [1,2]，张惟天 [1]，王承祥 [1,2]，黄晨 [2,1]

[1] 东南大学信息科学与工程学院移动通信全国重点实验室

[2] 紫金山实验室普适通信研究中心

摘要

不同于传统的离线信道建模，数字孪生在线信道建模能够实时感知和准确表征动态无线信道，实现高效的 6G 网络优化。本文提出了一种新型的数字孪生在线信道模型（Digital Twin Online Channel Model，DTOCM），通过实现对动态信道变化的持续可视化和准确预测，DTOCM 能够同步数字网络与物理网络之间的性能。我们首先探讨了数字孪生在线信道模型的发展，强调其愿景和相关挑战。然后，阐述了 DTOCM 的原理、构建机制以及在典型 6G 场景中的应用。接着，基于 DTOCM 平台，说明了其实时信道信息提供和可视化功能。最后，讨论了未来的研究方向和待解决的挑战。

关键词

数字孪生，6G 在线信道建模，信道地图，环境感知，机器学习

1 引言

第六代（Sixth Generation，6G）无线通信网络预计将具备“全覆盖、全应用、全数字、全频谱、全感官、强安全”的特点 [1]。传统网络优化依赖于信道测量、反复试错和工程经验等成本高昂、耗时极长、风险极大的方法。且实际无线网络可调参数多，网络性能和用户流量的非线性预测复杂度极大。因此，离线网络优化效果不佳，大部分网络只能发挥约 60% 的性能，仍有显著的提升空间 [2]。能够准确反映真实物理环境的在线信道模型，辅助实现在线网络优化对于 6G 网络的部署和性能提升至关重要。然而，现网通信系统测试性能与实验室仿真性能不一致，主要原因有两点：首先，现网测试的场景未经信道测量，实验室仿真用的信道模型与现网实测环境（信道）不匹配；其次，即便有符合现网实测场景的信道模型，其信道模型参数仍面临固化局限，难以适应时变的现网实测场景，导致信道仿真模型不能实时匹配现网测试的场景 [3, 4]。因此，亟需能够实时反映现网通信环境信道特性的数字孪生在线信道模型（Digital Twin Online Channel Model，DTOCM）[16]。

DTOCM 允许对信道状况进行持续监测和实时分析，提供动态信道变化的可视化和预测 [5]。随着 6G 网络逐步支持混合服务、沉浸式通信和动态环境，DTOCM 能够实现低成本、低复杂度和高精度的信道信息获取，从而解决大规模导频开销的问题，减少估计误差，提高整体网络性能 [6]。DTOCM 的可行性依赖于两个关键因素。首先，物理环境与信道特征之间的紧密耦合关系可以通过物理定位和环境特征部分捕捉到。这种内在关联表明，相似的物理位置和环境特征往往会产生相似的信道特征，使基站生成的大量数据可重复利用。这种可重复利用性显著有助于构建准确的信道数字孪生模型。其次，随着通信网络和频段的扩展，定位和感知精度的提升增强了精确定位和环境认知的能力。这些增强的能力使得对无线信道的数字化更为详细和准确，与 DTOCM 的目标完美契合。

本文提出了一种新型的数字孪生在线信道建模方法，通过利用物理环境传感器和多种感知方法，实现对动态环境的实时感知，建立了实际通信环境参数、信道参数和信道特性之间的映射关系。此外，通过将环境感知与信道建模相结合，DTOCM 提高了对无线信号传播环境变化的预测精度，使得对现实世界无线传播场景的模拟更加精准，为通信网络优化和决策提供了强有力的支持。

本文的组织结构如下。在第 2 节中，我们将探讨 DTOCM 的优势与愿景。在第 3 节中，我们将详细介绍所提出的数字孪生在线信道建模框架，包括其整体工作原理、三步构建机制以及在 6G 通信场景中的典型应用。在第 4 节中，我们将说明所提出的数字孪生在线信道建模框架的实时提供和可视化信道信息的特性。随后，在第 5 节中，我们将讨论与 DTOCM 相关的一些开放性研究课题，最后在第 6 节中对本文进行总结。

2 DTOCM 的优势及愿景

2.1 DTOCM 的优势

本文提出的 DTOCM 具有四个关键优势：

- **减少导频开销：**DTOCM 能够预加载基本信道状态信息（Channel State Information，CSI），使设备在通过连续 3D 空间时获取信道参数。这可以通过访问数据库或基于位置、方向和天线参数进行计算实现。DTOCM 通过直接在信道研究中解决通信网络设计挑战，显著减少导频开销，提高 6G 网络性能；
- **实时 CSI 提供与预测：**DTOCM 具有较强的实时信道信息提取和预测能力，适应动态网络条件。这一能力确保了 6G 通信网络对环境变化保持响应，有助于校准信道模型仿真与实际环境条件之间的性能，从而提升网络性能的可靠性和稳健性；
- **信道信息可视化：**DTOCM 提供信道信息变化的可视化，预测并突出通信环境的变化。它允许显示节点位置和运动状态，提供信道特性和其他相关信息的实时监测。这种可视化增强了未来 6G 网络中的环境感知和决策能力；
- **优化网络性能：**DTOCM 能够在虚拟环境中模拟场景，提供优化现实通信网络部署的感知和反馈。这能够在受控的虚拟环境中进行全面测试，从而实现快速迭代和改进。通过利用这些仿真获得的信息，未来的 6G 网络可以有效优化，实现性能与效率的平衡。

2.2 DTOCM 的愿景

如图 1(a) 所示，DTOCM 捕捉了实际动态环境中的 CSI 变化，显示物体和节点的位置及运动情况，同时实时输出信道大小尺度衰落信息、信道特性等。DTOCM 可以促进实时信道信息预测，并辅助通信网络优化决策。利用物理环境传感器实现动态环境的实时感知，将实际通信环境参数映射为数字孪生信道的信道参数和特征。结合 AI 算法，它可以预测无线传播环境在空间、时间和频率域内的变化。

上述愿景可以在 6G 网络的三个典型应用场景中得到阐明，如图 1(b) 所示。第一种场景中，发射机和接收机是静止的，而散射环境动态变化。DTOCM 可以显示实时的 CSI，如幅度、相位、延迟、多普勒和角度。第二种情况下，发射机和接收机在动态环境中移动。DTOCM 提供实时的 CSI 以及各个节点的位置和运动情况。第三种情况下，一个移动实体如无人机在运动中，DTOCM 可以根据信道信息的变化推断出无人机的位置。

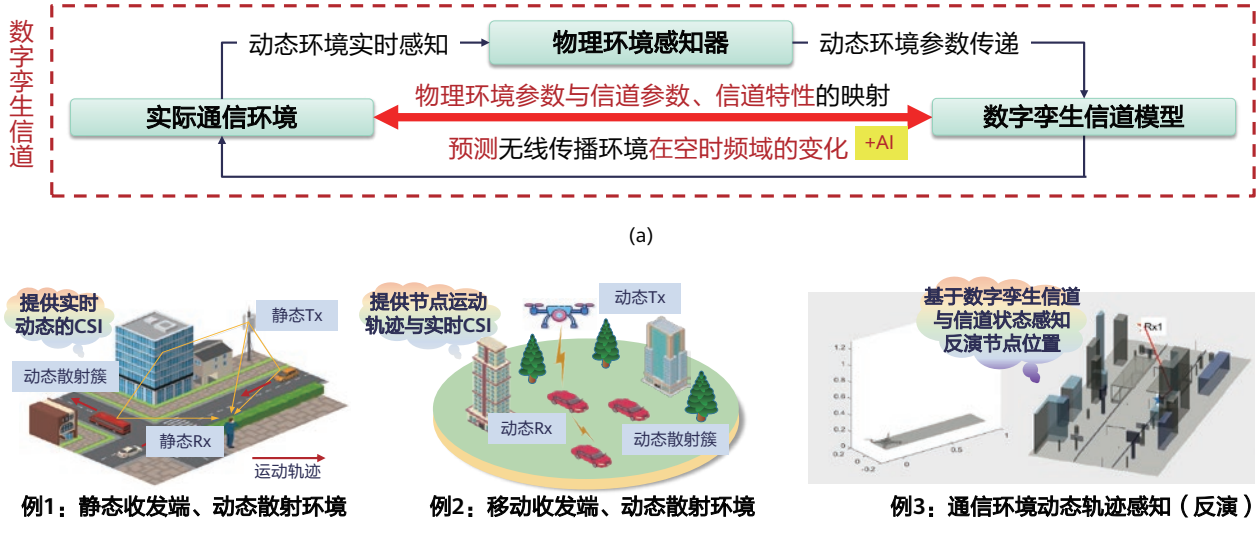


图 1 DTOCM 的 (a) 愿景和 (b) 典型应用场景

3 数字孪生在线信道建模框架

3.1 总体框架介绍

DTOCM 首先在对 6G 场景进行全面、排他的分类的基础上，通过环境信息实时感知识别具体的通信场景，然后将其与 6G 普适信道模型（6G Pervasive Channel Model，6GPCM）[7, 8] 的模型参数配对，以实现数据驱动的通信场景识别和模型驱动的全场景信道仿真。首先，在物理网络层收集环境和信道测量数据，用于环境重构和数据孪生，创建数字孪生层。数字孪生层利用射线追踪（Ray Tracing，RT）和基于几何的随机建模（Geometric-Based Stochastic Modeling，GBSM）生成无线信道地图。这张地图作为 AI 预测的粗模型，其中基于 AI 的模型利用已知时间信道数据库来实现有效且准确的空时频预测 [9]。这些预测填补了无线信道地图中的空白，如未知场景、未来时间段和不熟悉的频段。通过这一过程，可以实现一个相对完整的数字孪生信道地图。该地图可用于物理网络层的网络参数优化和环境信息反演。此外，物理网络层可以为数字孪生信道地图提供网络性能校准所需的参数或数据。

3.2 基于机器学习的场景识别

当前的 5G 标准化信道模型对通信场景的划分相对粗略，无法充分描述未来 6G 网络在全频段、全覆盖和全应用愿景下的所有通信场景。为了构建适用于所有频段和场景的实时且精确的数字孪生在线信道模型，需要对 6G 无线通信场景进行详细分类，并探索不同通信场景的信道特性。

如图 2 所示，构建 DTOCM 的第一步是创建全面、详细且分层的 6G 通信场景分类，然后匹配适当的信道模型参数。具体而言，基于环境感知数据，采用数据驱动的机器学习方法来识别和建模通信场景，步骤如下：

1. **实时环境感知：**为了实现全面的通信场景识别，使用各种环境数据源，如电子地图、遥感图像和点云数据，提供对环境的实时感知 [10, 11]。这些数据源有助于提取环境参数，这对于准确的信道建模至关重要；
2. **参数提取和场景分类：**通过机器学习模型（如神经网络）提取和处理环境参数，将通信场景分类为不同类别，如航空通信、航天通信、陆地通信和海洋通信。这些类别进一步细分为更具体的环境，如卫星、太空站、飞机、无人机、室内、城市、海面和岛屿。基于神经网络的分类网络是这一过程的关键，它可以对各种环境进行更精细的理解；
3. **信道模型匹配：**最后，识别不同的通信场景并匹配 6GPCM [7, 8] 的不同信道模型参数。该步骤将适当的信道模型参数映射到各个场景，使得可以进行准确的多场景信道测量，并确保与 6G 信道模型的兼容性。

3.3 离线信道地图的初始化与环境重构

基于场景识别获得相应结果后，下一步是重构数字孪生环境以进行离线信道图初始化。可以通过首先使用多模态环境感知，如图像、视频、3D 电子地图和点云，进行 3D 场景重构来实现。重构后的虚拟场景将被导入 RT 软件以模拟信道特性。通过比较 RT 模拟结果与实际测量数据之间的差异，可以校准模拟中使用的电磁系数，以提高 RT 信道重

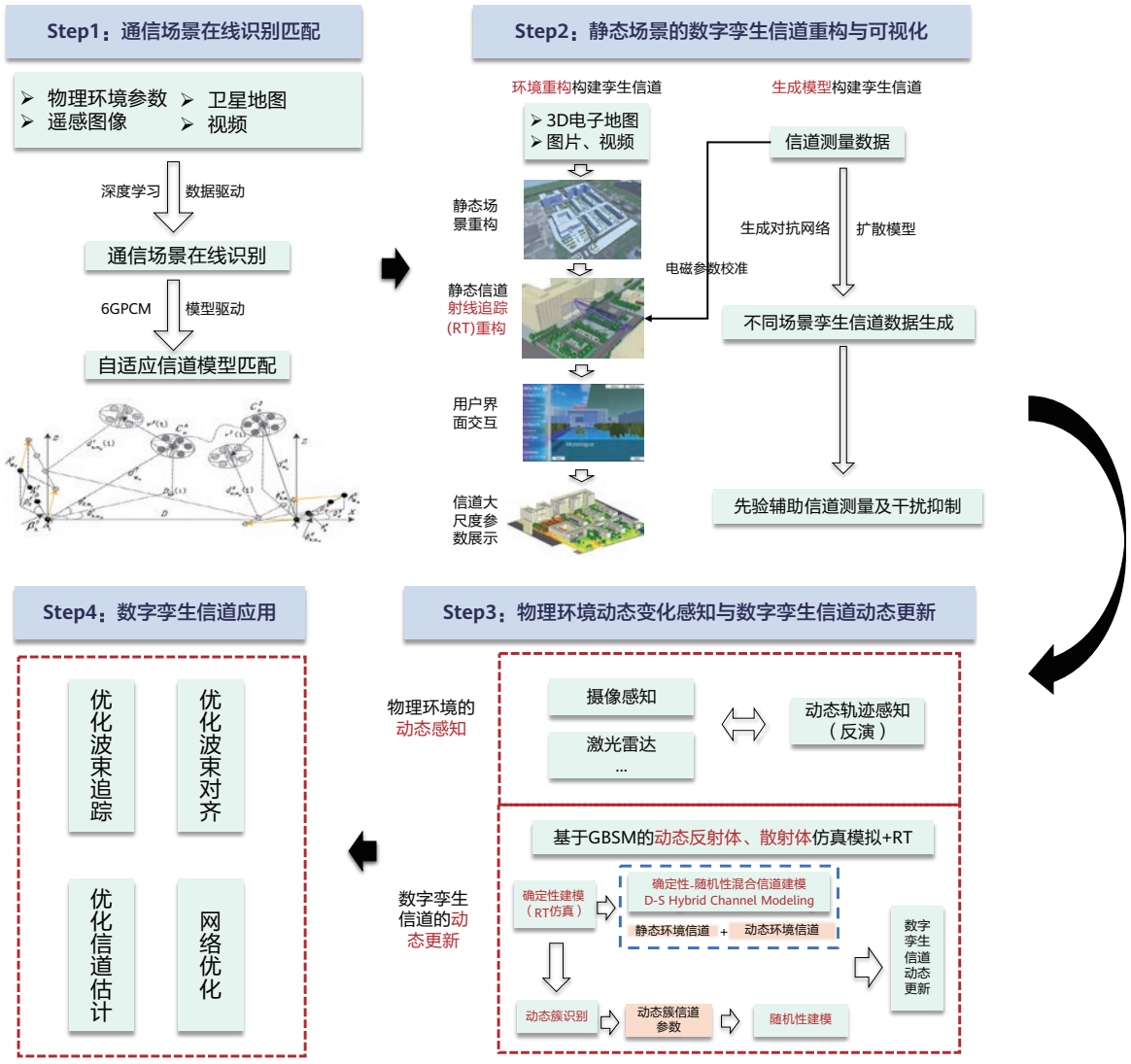


图 2 DTOCM 构建流程

构的准确性。除了孪生环境外，数字孪生层还包含孪生数据。可以使用生成模型从测量数据中创建不同场景的孪生环境，为信道测量和干扰抑制提供先验支持。

3.4 环境感知与数字孪生信道动态更新

最后一步是实现动态环境信道特性的实时更新。我们使用物理环境传感器，如摄像机传感、激光雷达等相关感知设备，实时监测环境动态变化，如移动的车辆和行人。然后，这些变化将反映在孪生环境中，并重新模拟相应的信道特性。DTOCM 采用一种新颖的静态–动态混合信道建模算法来进行动态信道信息更新 [12]。环境中的静态物体使用 RT 建模进行模拟（详见 3.3 节）。对于动态物体，我们使用从测量数据中提取的动态簇参数，并采用 GBSM 来模拟动态散射信道。

3.5 数字孪生在线信道建模的应用

基于上述步骤构建的数字孪生信道可以用于各种应用场景，增强 6G 通信网络如波束对齐、波束追踪、信道估计和网络优化等能力 [13]。

- **优化波束对齐：**利用 DTOCM 得出的信道信息可以优化波束对准。它支持选择最佳性能的波束，减少对准波束所需的努力并提高精度，从而实现最佳信号质量。DTOCM 基于功率谱密度和其他相关指标预先选择这些波束，减少了导频开销和信道估计的复杂性。
- **优化波束追踪：**DTOCM 可以通过生成优化检测波束范围的信道统计数据来增强波束跟踪，从而提高波束跟踪的速度和精度。通过稀疏贝叶斯学习利用信道稀疏性，选择具有最大预期信噪比的波束，促进快速波束对准和高效跟踪。这种优化在快速响应动态环境变化时尤为关键。

- **优化信道估计：**优化的信道估计可以利用 DTOCM 指导导频设计并前置信道信息获取步骤，从而显著减少信道估计的开销。通过信道地图的初步数据，简化了信道估计过程，提高了 6G 通信网络的整体效率。
- **网络优化：**与传统信道建模不同，DTOCM 不仅旨在表征特定场景、历史时间数据和已知频段内的信道特性，还致力于预测未知环境、未来时间和未知频段内的信道特性。这使得能够基于无线传播信道的感知数据预测未来的信道孪生数据，然后将这些数据反馈到网络层，从而基于 CSI 进行网络参数优化。

4 DTOCM 的可视化平台展示

本文所提出的数字孪生在线信道建模框架在 6G 网络中具备实时信道信息的提取与可视化功能，这使得网络性能分析和智能网络资源调度成为可能。演示平台从场景环境的重构开始，其中静态场景使用 Blender 建模“无线谷”的物理特征，而动态场景如无人机或行人等移动对象则使用 Unity 开发。它实现了交互式仿真，用户可以通过鼠标点击选择和修改发射机（Transmitter，Tx）和接收机（Receiver，Rx）的位置，从而增强模型的真实性和适用性。系统输出详细的信道特性，包括 CSI、路径损耗和其他相关数据，并根据用户交互和场景变化动态更新。同时利用 RT 和 GBSM 准确预测和仿真无线信道变化，实现实时更新 [14]。

如图 3 所示，3(a) 展示了实时感知到的环境变化，3(b) 展示了对应的数字孪生环境，该环境实时更新以反映行人的移动。3(c) 通过深度图显示深度分析，用不同颜色表示距离和深度。3(d) 展示了对应的信道特性，包括方位角和仰角的功率分布。当行人移动遮挡或改变信号反射面时（如行人经过大型金属结构或建筑物的反射面前），方位角的功率分布会显示出新的峰值或现有峰值的减少。这些变化在功率分布图中可以观察到新的峰出现或原有峰的位置、高度变化。如果行人在不同的楼层移动，或者他们的活动改变了地面与周围建筑的相对位置关系，则仰角的功率分布会显示信号的垂直传播路径发生变化。例如，行人在楼宇间小道上行走，某些仰角下的功率可能增加，表明从那些角度反射的信号增强。这些变化反映出数字孪生信道对实际信道的准确模拟。

图 4 展示了基于 GAN-GRU（Generative Adversarial Network-Gated Recurrent Unit）的空时域预测信道模型的仿真结果，其中 GAN 网络能够对测量数据进行特征学习并实现数据量翻倍，GRU 网络可以捕捉物理环境变化与空时域信道特性变化之间的映射。由图 4 可知，GAN-GRU 空时域预测信道模型在室内走廊的视距（Line of Sight, LoS）和非视距（Non-LoS, NLoS）场景下均取得了良好的信道预测效果。预测信道的接收功率与实测信道的接收功率能够准确地匹配，且信道的大多数径都能成功预测，在信道冲激响应（Channel Impulse Response, CIR）对比图中得到了良好的表征，证明了数字孪生信道建模的可行性 [15]。

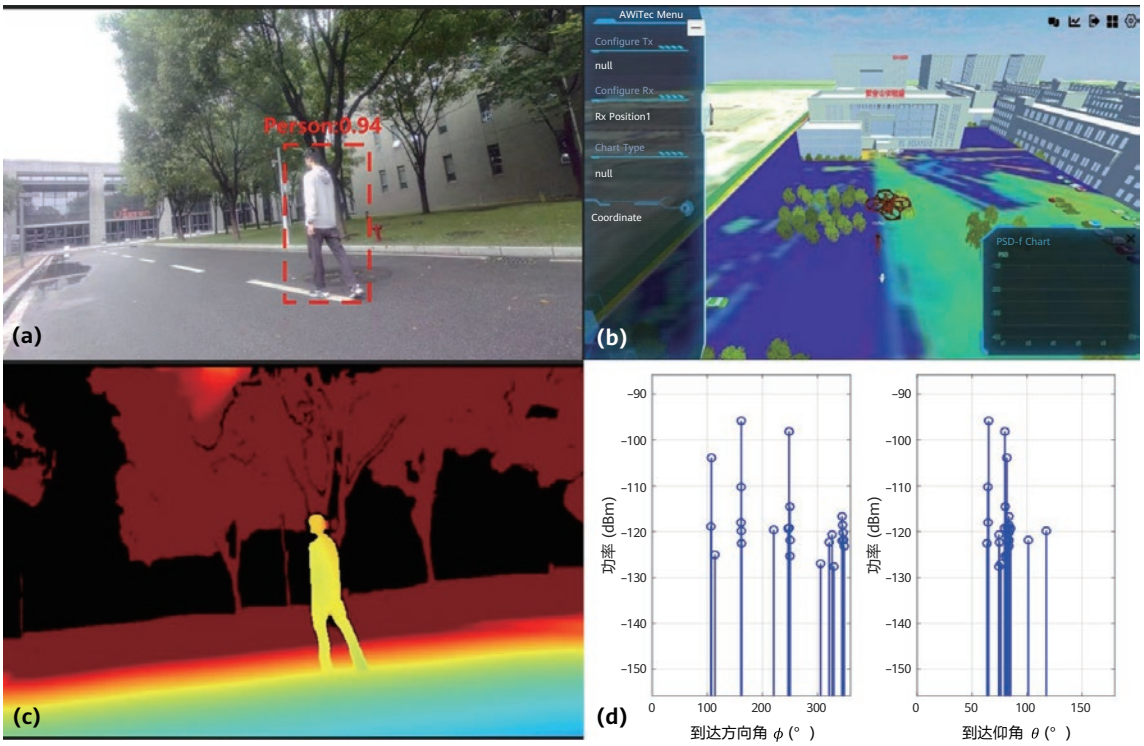


图 3 DTOCM 的实时信道信息可视化
(a) 感知信息；(b) 数字孪生环境；(c) 深度图；(d) 信道特性 [14]

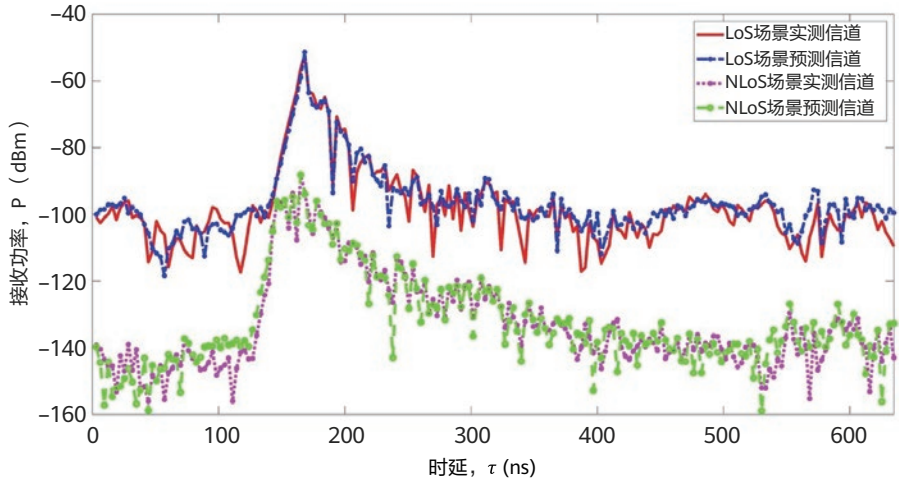


图 4 在 6 GHz 频段上 LoS 和 NLoS 场景中实测信道和预测信道的信道冲激响应 [15]

5 研究展望

尽管 DTOCM 在促进 6G 网络管理方面有多种应用，但仍然存在一些关键挑战。在本节中，我们将讨论未来 6G 网络中与 DTOCM 相关的一些开放性研究课题。

5.1 高精度射线追踪

RT 仿真是 DTOCM 的关键组成部分，其精确度受三个主要因素的影响。首先，RT 仿真的精确度受到所使用 3D 模型的复杂性和细节的显著影响。然而，构建高度详细的 3D 模型需要大量的计算资源和强大的数据输入，这在频繁发生变化的动态环境中是一个挑战。其次，模型中所代表材料的电磁特性也对 RT 的精确度起着至关重要的作用。正确分配材料特性如反射率、吸收率和介电常数是关键，因为这些特性决定了波与表面相互作用的方式。第三，RT 仿真中使用的基础算法会影响结果的精确度。提高算法处理诸如衍射、散射和多次反射等复杂交互的能力是增加仿真精度的必要条件。这些算法的计算效率也影响实时仿真的可行性，这是 DTOCM 所必需的。

5.2 多模态数据融合

如何有效感知和融合不同来源、不同类型的多模态数据是推进数字孪生在线信道建模的关键挑战。多模态数据不仅包括从传感器和地图收集静态数据，还涉及从无人机和车辆等移动实体收集动态的实时数据。为融合利用这些多模态数据，需要更高效、凝练的 AI 算法，促进数字孪生在线信道建模的发展与成熟。

5.3 实时处理与降低延迟

感知数据实时处理以及降低延迟对于确保数字孪生在线信道建模的准确性和响应性至关重要。有效实时地整合来自不同来源（如传感器、无人机和用户设备）的动态数据，根据物理世界中收发端位置变化及通信场景反射体、散射体变化同步计算相应的空时频域信道特性变化有助于感知数据实时处理。此外，环境感知和数据处理中的延迟最小化对于信道地图的及时更新至关重要。这需要开发低延迟处理算法，并可能需要采用边缘计算解决方案来处理更接近源的数据。解决这些挑战将增强模型提供准确、实时信道状态信息的能力。

5.4 辅助连续立体空间无线电信道建模

对于未来的网络，信道从离散局部空间无线传播信道演变为 3D 连续立体空间无线电信道，即多用户、网络级、基站和用户在连续移动（连续立体空间）、考虑双端天线所有参数的无线电信道。连续立体空间传播环境复杂，地理信息缺失，不同传输技术对信道表征需求差异大，给在 3D 连续立体空间无线电环境中应用 DTOCM 带来了额外的困难。因此，迫切需要构建基于静态环境重建和动态环境感知的数字孪生模型，获取实时准确的信道信息。

6 结语

本文中提出了 DTOCM 框架，能够准确表征动态信道，并大大促进 6G 网络优化。我们概述了 DTOCM 的愿景、相关挑战、构建流程和操作原理，并强调了其实时信道信息提供与可视化的能力。希望本文的研究成果能够促进业界对数字孪生在线信道建模的进一步研究与开发，最终提高未来网络的效率、适应性和性能。

参考文献

[1] C.-X. Wang *et al*., "On the road to 6G: Visions, requirements, key technologies, and testbeds," *IEEE Commun. Surv. Tutor*., vol. 25, no. 2, pp. 905–974, 2023.

[2] Z.-Q. Luo *et al*., "SRCON: A data-driven network performance simulator for real-world wireless networks," *IEEE Commun. Mag*., vol. 61, no. 6, pp. 96–102, 2023.

[3] C.-X. Wang, J. Huang, H. Wang, X. Gao, X. You, and Y. Hao, "6G wireless channel measurements and models: Trends and challenges," *IEEE Veh. Technol. Mag*., vol. 15, no. 4, pp. 22–32, 2020.

[4] R. He *et al*., "A kernel-power-density-based algorithm for channel multipath components clustering," *IEEE Trans. Wirel. Commun*., vol. 16, no. 11, pp. 7138–7151, 2017.

[5] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, "Holistic network virtualization and pervasive network intelligence for 6G," *IEEE Commun. Surv. Tutor*., vol. 24, no. 1, pp. 1–30, 2022.

[6] X. Lin, L. Kundu, C. Dick, E. Obiodu, T. Mostak, and M. Flaxman, "6G digital twin networks: From theory to practice," *IEEE Commun. Mag*., vol. 61, no. 11, pp. 72–78, 2023.

[7] C.-X. Wang, Z. Lv, Y. Chen, and H. Haas, "A complete study of space-time-frequency statistical properties of the 6G pervasive channel model," *IEEE Trans. Commun*., vol. 71, no. 12, pp. 7273–7287, 2023.

[8] C.-X. Wang, Z. Lv, X. Gao, X. You, Y. Hao, and H. Haas, "Pervasive wireless channel modeling theory and applications to 6G GBSMs for all frequency bands and all scenarios," *IEEE Trans. Veh. Technol*., vol. 71, no. 9, pp. 9159–9173, 2022.

[9] C. Huang, C.-X. Wang, Z. Li, Z. Qian, J. Li, and Y. Miao, "A frequency domain predictive channel model for 6G wireless MIMO communications based on deep learning," *IEEE Trans. Commun*., vol. 72, no. 8, pp. 4887–4902, 2024.

[10] F. Zhang *et al*., "A radio wave propagation modeling method based on high-precision 3-D mapping in urban scenarios," *IEEE Trans. Antennas Propag*., vol. 72, no. 3, pp. 2712–2722, 2024.

[11] P. Koivumaki, G. Steinbock, and K. Haneda, "Impacts of point cloud modeling on the accuracy of ray-based multipath propagation simulations," *IEEE Trans. Antennas Propag*., vol. 69, no. 8, pp. 4737–4747, 2021.

[12] T. Qi, C. Huang, J. Shi, J. Li, S. Chen, and C.-X. Wang, "A Novel Dynamic Channel Map for 6G MIMO Communications," in *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 2024, pp. 809–814.

[13] R. Levie, Ç. Yapar, G. Kutyniok, and G. Caire, "RadioUNet: Fast radio map estimation with convolutional neural networks," *IEEE Trans. Wirel. Commun*., vol. 20, no. 6, pp. 4001–4015, 2021.

[14] S. Xiao, H. Zhang, M. Yao, C. Cui, J. Li, C. Huang, and C.-X. Wang, "Demo: A novel 3D environment-aware digital twin online channel modeling platform," in *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 2024, pp. 297–298.

[15] Z. Li *et al*., "A GAN-GRU based space-time predictive channel model for 6G wireless communications," *IEEE Trans. Veh. Technol*., vol. 73, no. 7, pp. 9370–9386, 2024.

[16] J. Li, C.-X. Wang, C. Huang, T. Qi, and T. Wu, "Digital Twin Online Channel Modeling: Challenges, Principles, and Applications," *IEEE. Veh. Technol. Mag*.

基于 AI 的射频及天线设计

王光健，阳棣均，Jimmy Jian，Chandan Roy，潘立，黄国龙，蔡华，童文

摘要

随着人工智能（Artificial Intelligence，AI）技术的飞速发展，AI 在射频及天线设计领域的应用日益广泛，展现出了显著的潜力和发展前景。本文深入探讨了 AI 技术在射频及天线设计中的多元化应用，包括射频电路设计、天线形状优化、阵列综合、和电磁高效仿真等关键环节。在关键设计环节中，AI 技术通过自动化设计流程、快速仿真和优化、智能设计工具的开发，显著提高了设计效率。同时，AI 在多目标优化、精细化设计、高维度问题求解、非线性问题分析、多物理场耦合等方面，展现出其独特且优越的复杂问题处理的能力。另一方面，我们也讨论了目前 AI 在应用过程中的挑战、潜在解决方案以及未来射频及天线技术的发展趋势。本文为射频及天线领域的发展提供了有益的参考，有助于推动通信技术的进步。

关键词

人工智能，射频设计，天线设计，电磁仿真，机器学习，滤波器

1 引言

在当今的信息时代，射频及天线技术宛如通信领域的魔法钥匙，开启了无线通信的无限可能。它们不仅是现代通信系统的核心基石，更是连接人与人、人与世界的重要纽带。想象一下，没有射频及天线技术，我们将无法畅享便捷的移动通信，无法与远方的亲朋好友随时保持联系；卫星通信将陷入瘫痪，无法实现全球范围内的信息传递；物联网的宏伟愿景也将成为泡影，智能设备之间的“对话”将无从谈起。射频及天线设计是无线通信系统中的关键环节，其性能对整个系统的质量和效率具有重要影响，射频及天线技术的重要性不言而喻 [1]。

然而，随着科技的飞速发展，我们对通信的需求日益增长，这也对射频及天线技术提出了更高的要求。如何提高射频信号的传输效率？如何设计更小巧、更高效的天线？如何应对日益紧张的频谱资源？这些都是摆在我们面前的严峻挑战。面对新出现的挑战，传统的设计方法往往依赖于经验和试错，需要大量的时间和精力，尤其是当器件参数众多且相互依赖时计算效率将显著下降。而另一方面，随着人工智能技术的迅速发展，其在射频及天线设计领域的应用正逐渐引起人们的关注 [2]。AI 具有强大的数据处理和学习能力，有望为射频及天线设计带来新的思路和方法 [3]。

本文旨在探索射频及天线技术的奥秘，寻求提升其性能的创新方法。我们将深入研究射频电路设计、天线设计和电磁仿真等关键领域，挖掘人工智能在其中的巨大潜力。通过不断的实验和分析，我们期待揭示 AI 技术如何模拟复杂的电磁现象，以及通过强化学习来自动调整仿真参数，实现更高效的仿真过程。这篇文章的组织结构如图 1 所示，引言之后主要包含下内容。第二节讨论了 AI 的分类和一些用于射频和天线设计的具体 AI 模型。第三节举例说明了 AI 辅助下所实现的射频电路，天线和电磁仿真。第四节专门讨论了 AI 在射频电路和天线设计中的具体优势。第五节给出了基于 AI 设计电路的具体实施示例。第六节讨论了 AI 在射频电路设计中的挑战和未来的展望。最后在第七节中给出结论。

2 用于射频和天线设计的 AI 模型

在射频和天线设计领域， AI 正以其独特的优势重塑传统的设计方法。AI 技术，尤其是机器学习（Machine Learning， ML）和深度学习（Deep Learning，DL），通过模仿人类的认知过程，使得机器能够通过识别数据中的模式来学习并解决复杂问题。这些技术在射频及天线设计中的应用，不仅提高了设计效率，还优化了设计结果，特别是在处理高维度、非线性和多物理场耦合问题时表现出色。

在射频设计中，前馈神经网络（Feedforward Neural Network，FFNN）如多层感知机（Multilayer Perceptron，MLP）和径向基函数（Radial Basis Function，RBF）网络，已被证明在解决非动态建模问题方面非常有效 [4]。小波神经网络（Wavelet Neural Network， WNN）由于其隐藏神经元的局部性质，适合于具有高度非线性或急剧变化的问题，有助于更容易的训练并取得更高的模型精度。极限学习机（Extreme Learning Machine，ELM）是一种单隐藏层 FFNN，其优点是快速学习速度和在复杂电磁参数建模中（特别是在训练数据集较小的情况下）仍然保持较好性能。它们能够精确地模拟射频电路的行为，从而优化滤波器和放大器的设计。动态神经网络 [5]、递归神经网络（Recurrent Neural Network， RNN）和时延神经网络（Time-Delay Neural Network， TDNN）则在表征非线性设备或电路的时间域动态行为方面发挥着关键作用 [6]。深度神经网络（Deep Neural Network，DNN）通过其多层结构，为复杂建模问题提供了更深层次的解决方案 [7]，而生成式对抗网络（Generative Adversarial Network，GAN）则在新颖设计生成方面展现出显著优势 [8, 9]。此外，基于知识的神经网络（Knowledge-based Neural Network，KBNN） [10] 利用现有的等效电路和经验模型，辅助微波组件的 CAD，减少了对大量训练数据的依赖，并提高了模型的外推能力。这些技术的综合应用，不仅加速了设计过程，还提升了设计的整体性能和可靠性。

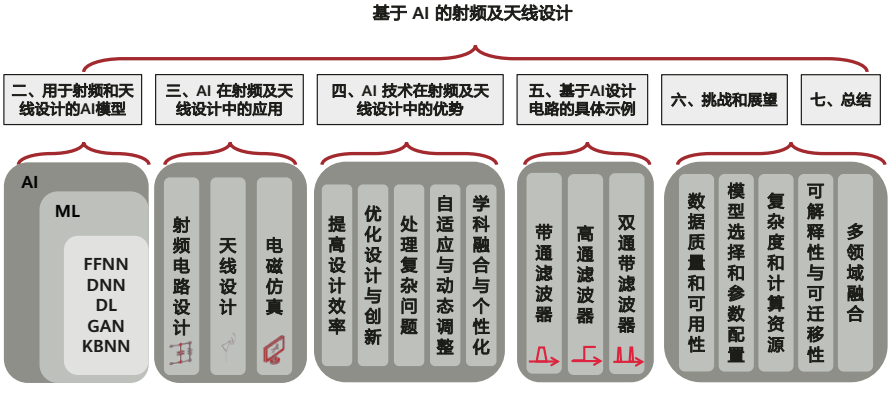


图 1 文章结构组成

3 AI 在射频及天线设计中的应用

3.1 射频电路设计

在射频电路设计领域， AI 正以其卓越的能力重塑传统设计方法。AI 技术的应用不仅加速了设计过程，还显著提升了电路性能。例如在滤波器、耦合器设计方面，AI 能够精确计算并优化频率响应、耦合衰减特性和群延迟，满足多样化的频率和功率选择需求 [13, 14]。对于放大器设计，AI 技术通过智能算法优化增益、噪声系数等关键参数，确保放大器在多变的工作条件下保持卓越性能 [15]。在振荡器设计中，AI 助力实现高稳定性和低相位噪声的振荡器，即使在复杂电磁环境中也能保持高频谱纯度。此外，AI 在阻抗匹配方面同样展现出巨大潜力，通过精确计算匹配网络，有效提升射频电路的传输效率和整体性能。

具体在滤波器设计方面，有多种技术工艺可用于实现所需的滤波器要求，具体技术工艺的选择取决于未来的应用。例如，微带线和带状线技术适合与平面电路更好地集成。另一方面，传统波导技术适用于低传输损耗的高频应用。基片集成波导（Substrate Integrated Waveguide，SIW）是一种新兴的技术，可用于高频应用，具有低传输损耗条件，以更好地与平面系统结构集成。所有这些技术都有自己的滤波器结构的设计变量。例如，微带线的长度和宽度是微带滤波器的主要几何变量。过孔的直径和后续过孔之间的距离被认为是 SIW 滤波器的主要变量。光阑长度和窗口长度是波导滤波器的主要几何变量。ML 技术通常用于表征或评估滤波器性能如何随其几何参数的变化而变化。我们期望一个经过良好训练的机器学习模型能够预测几何参数对滤波器响应的影响。利用这一模型，我们可以快速改进设计，避免了使用计算成本高昂且速度慢的电磁模型 [16]。

3.2 天线设计

在天线设计领域，AI 技术正推动着天线设计向更高效、更精准的方向发展。一方面 AI 能够作为代替模型快速生成天线的等效仿真结果，另一方面 AI 技术通过对天线形状的精细调整，如改变尺寸、形状和结构，或添加特定的结构元素，实现了天线性能的优化，包括提升增益、带宽和方向性等关键指标。此外，AI 在极化方式选择上的应用，通过分析不同极化对信号传输和接收的影响，使工程师能够根据具体场景需求，选择最合适的极化方式，如线极化、圆极化或椭圆极化，进一步提升信号传输效率和接收质量。在天线阵列设计方面，AI 算法的辅助使天线布局和元素数量的确定更加科学，优化了波束形成和控制，显著增强了天线的方向性和增益。同时，AI 技术在多频段天线设计中的应用，考虑了不同频段的天线参数和性能要求，成功实现了天线在多个频段的兼容性和高效工作，满足了现代通信对频谱资源的多样化需求。总体而言，AI 技术的集成和应用为天线设计带来了创新的解决方案，提高了设计效率和系统性能，为无线通信的进步提供了强大的技术支撑。

具体在辅助天线设计方面，完整的设计流程如图 2 所示。在天线设计流程中，首先根据工程师的经验选择一个基本的几何形状以满足所需的性能，然后进行参数优化以找到最优设计参数。在参数优化过程中，更新模型参数的全波电磁仿真计算最为耗时。而基于机器学习算法的代理模型可以用来替代耗时的全波仿真过程，节省计算资源并加快天线设计。该过程首先需要运行全波仿真来获取数据集，即通过改变天线的几何形状在工作频段内来生成 S 参数和增益等输出参数。之后构建并通过生成的数据训练代理模型。在训练过程中，通过代理模型生成的结果会逐步逼近真实全波仿真结果，以在优化过程中替代耗时的全波仿真。迄今为止，大多数代理模型辅助的天线优化方法采用高斯过程（Gaussian Process, GP）代理模型 [17]，其优点在于易于实现、可解析和不确定性可量化等。在所有数据集中，大约 80% 的数据集用于训练代理模型，其余 20% 用于测试其准确性。如果测试误差不满意，则可以生成更多数据或进一步改进代理模型重复上述过程。通过代理模型得到期望的器件参数后，需要用全波仿对该器件进行验证和微调，以得到最佳设计性能。在天线分析中，机器学习能够从仿真或测量数据中推测出辐射模式和共振频率等特性。具有多个输出的复杂代理模型可以与多目标演化算法结合使用，以在多种需求指标下找到最优的天线设计 [2]。另外多种不同神经网

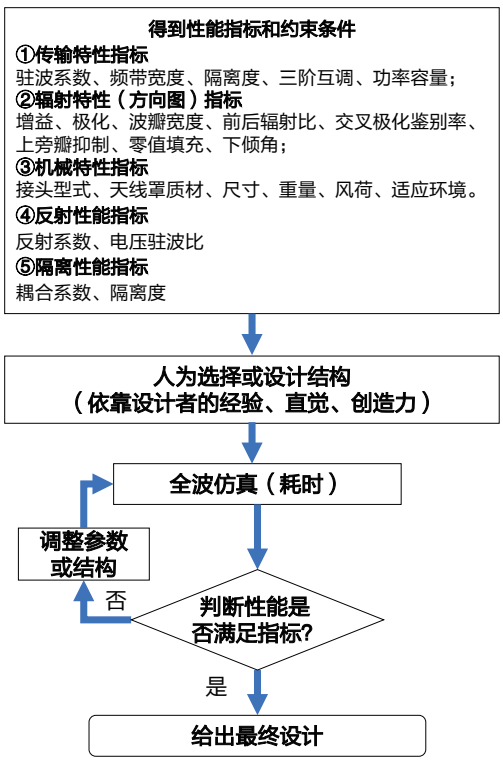


图 2 天线设计过程

络组合完成天线设计同一样是值得探索的方面，如图 2 所示的天线设计流程，第一步是选择合适的天线类型。基于支持向量机（Support Vector Machine，SVM）的推荐系统 [18] 可以协助工程师进行这一选择。一旦确定了天线的基本几何形状，就可以使用人工神经网络（Artificial Neural Network，ANN）和堆叠集成学习模型等 AI 方法计算模型参数的值。

3.3 电磁仿真

在电磁仿真领域，通过结合 AI 进行多物理场耦合仿真，如电磁、热和结构等，可以更全面地评估系统在实际工作环境中的性能。这种跨学科的仿真方法为射频及天线设计提供了深入的洞察，确保了设计的鲁棒性和适应性。一方面，AI 算法在电磁辐射和散射计算中快速得到计算结果。另一方面，AI 算法在电磁仿真的反演分析中也显示出巨大潜力，能够从测量数据中提取关键信息，进而优化设计参数。这一过程不仅提高了设计的精确度，还加快了从概念到产品的过程。实时仿真是 AI 技术的另一应用领域，它使得电磁仿真能够即时反馈设计结果，显著加快了设计迭代的速度。这种快速迭代能力对于快速原型设计和测试至关重要，极大地缩短了产品开发周期。最后，AI 在不确定性量化方面也发挥着重要作用。在电磁仿真中，不确定性因素可能会影响设计的可靠性。AI 技术能够评估这些不确定性因素，为设计决策提供了更加可靠的依据，从而提高了设计的成功率。

具体在源与散射场之间的正反演计算方面，传统的电磁散射和辐射全波仿真中，需要做费时的矩阵求逆来计算感应电流如下：

$$\bar{J} = \left(\bar{\bar{I}} - \bar{\bar{\chi}} \cdot \bar{\bar{G}}_D\right)^{-1} \cdot \left(\bar{\bar{\chi}} \cdot \bar{E}_{inc}\right) \tag{1}$$

为了节省这方面的时间开销，人们提出了一些基于深度学习的非迭代方法。特别是在涉及复杂散射体的情况下，研究者发现基于 GAN 的人工智能方法优于其它类型的神经网络（如 U-Net）。一个代表性的例子是前向感应电流学习方法（Forward-Induced Current Learning Method，FICLM）[19]，它通过神经网络映射来计算感应电流。然后，通过格林函数与求得的感应电流相乘来计算散射场。而从结果发现如果能使用多种输入方案作为输入，其计算得到的散射场更准确。如图 3 所示，输入方案涵盖了入射场和背景相对介电常数的多样化组合，满足了波动物理学中求解电磁散射问题的要求。通过 AI 模型和现有物理知识的相互结合，为快速求解电磁散射问题开拓了一条崭新的道路。

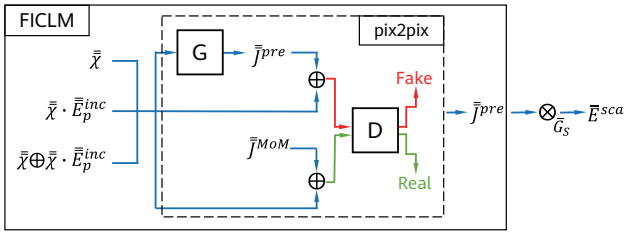


图 3 FICLM 的流程图

在反演与优化方面，研究者可以利用 AI 模型算法对电磁现象进行电磁反演分析，即从测量电磁数据中提取背景信息和优化器件的设计参数。已经开发了三种类型的神经网络求解器用于求解电磁反演问题。第一种模型的设计流程图如图 4 所示。直接从测量电磁结果反演散射体的物理参数，其学习过程可以用下面的模型表示：

$$R_l = \min_{R_\theta,\theta} \sum_{m=1}^{M} f\left(R_\theta\left(\bar{E}_m^s\right), \bar{\bar{\chi}}_m\right) + g(\theta) \tag{2}$$

图 4 第一种模型的流程图

通过拟合训练数据中每一对物理参数 $\bar{\bar{\chi}}_m$ 和散射场 $\bar{E}_m^s$，我们得到了从散射场到物理参数的直接映射的神经网络 R_l，注意，为了避免过拟合，我们还引入了正则化项 $g(\theta)$。由于该模型学习了太多的冗余信息（已知的波动物理学信息），导致其反演能力非常有限 [20]。

第二种模型仍然使用目标函数方法的传统框架，其中神经网络被训练为学习迭代求解器的一些组件 [21]。第三种将近似求解器（如反向传播算法）与 DNN 结合，通过将 DNN 的输入从测量的电磁场 $\bar{E}_m^s$ 转化成了物理参数的近似解 $\bar{\bar{\chi}}_m^a$（见下面图 5 的模型流程图），从而减轻了神经网络学习的负担，也简化了学习过程 [22, 23]。对于第三种方法，逆散射问题（Inverse Scattering Problem，ISP）中的迭代方法和 DNN 的体系结构之间的相似性启发研究人员改进了 DNN 的拓扑，例如由三个 CNN 模块组成的级联 DNN，其中每个模块单独训练。一开始，研究人员专注于第三种方法的对比度信息，其中对比度信息的近似结果和真实结果被用作 CNN 的输入和输出。此后受到波动物理学的启发，开发了增强型的神经网络 [24]，将

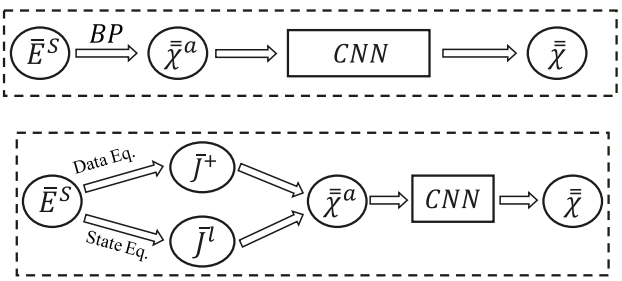


图 5 第三种模型的流程图

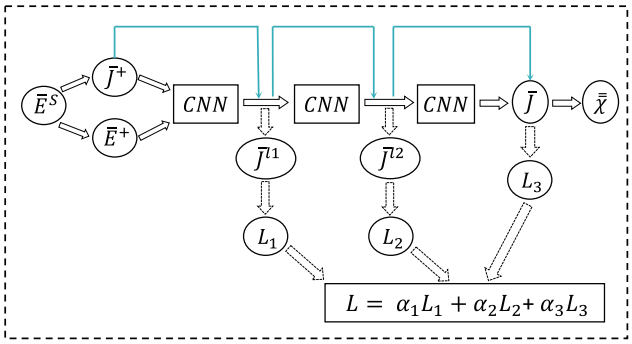


图 6 神经网络电磁反演的流程图

感应电流和电场也包括在输入和输出中，如图 6 所示，使得反演性能得到显著提高。这种结合了电磁物理的神经网络的学习过程可以用下面的模型表示：

$$R_l = \min_{R_\theta,\theta} \sum_{m=1}^{M} f(R_\theta(\bar{J}^+, \bar{E}^+), \bar{J}^{l1}, \bar{J}^{l2}, \ldots, \bar{J}) + g(\theta) \quad (3)$$

其中 $\bar{J}^{l1}, \bar{J}^{l2}, \ldots, \bar{J}$ 分别代表各级 CNN 输出的感应电流密度。

4 AI 技术在射频及天线设计中的优势

4.1 提高设计效率

AI 技术在射频及天线设计领域带来了显著的效率提升和创新。AI 的自动化设计流程通过自动执行参数计算和模型生成等重复性任务，显著节省了设计师的时间和精力。此外，AI 算法的快速仿真和优化能力，使得设计师能在极短的时间内获得高质量的设计方案。智能设计工具的开发进一步增强了这一效率，提供了更直观、高效的设计手段。

其中 AI 代理模型在提高效率方面起到了显著效果。AI 代理模型是通过机器学习的算法模型对复杂的射频 / 微波组件进行等效代替，既保持了电磁模型的准确性，又提升了电路模型的计算速度。但是，这要求 AI/ML 模型必须经过充分的训练，并拥有足够的数据支持，以确保其能够具有足够高的准确度，替代耗时的全波电磁仿真。这种技术的应用不仅提高了设计效率，而且降低了研发成本，加速了射频及天线设计从概念到实现的整个过程 [25]。

4.2 优化设计与创新

AI 在射频及天线设计中展现出其卓越的优化能力，多目标 AI 模型能够同时考虑性能、成本和尺寸等多个设计目标，寻找到最佳的设计平衡点。另一方面，AI 强大的数据分析能力，能够精细化把控设计细节，使全局优化成为可能，进一步提升了设计质量。此外，AI 的创新设计能力，能够激发了设计师的创造力，推动了新颖设计思路和方法的产生。

其中在复杂射频器件优化设计方面，经过数值数据和实验数据训练的 AI 模型，具有作为优化工具的巨大潜力，能够设计出传统方法难以实现的复杂天线、阵列和射频器件。一方面这些射频器件设计通常涉及时域、频域和谱域，且受到多重约束，其实质是高维和非凸的优化问题。这些问题具有非线性和多尺度特性，以及强的相互耦合，使得它们的优化设计特别具有挑战性。另一方面，优化算法通常需要大量的仿真结果来达到期望的性能，而迭代次数取决于所选优化算法和处理问题的复杂性。由于全波电磁仿真在能源和时间成本上的消耗极大，尤其是对大尺寸电磁模型的快速仿真优化，其可行性更是显得不切实际。然而，机器学习的发展为解决这些难题提供了有效的手段。它无需漫长的仿真时间就能提供与电磁模型相当的精度，极大地提高了优化的效率和可行性 [26]。

4.3 处理复杂问题

AI 技术在射频及天线设计领域提供了强大的问题解决能力，特别是在处理高维度和具有复杂约束的设计问题上，AI 能够准确找到最优解。AI 算法在捕捉非线性问题的特征和行为方面表现出色，提供了更为精确的分析结果，这在传统方法中往往难以实现。AI 的另一个显著优势在于多物理场耦合的分析能力，它能够综合考虑电磁、热、结构等不同物理场的信息，从而全面评估射频及天线系统的性能。这种跨学科的分析方法为设计提供了更深入的洞察，确保了系统在多方面性能上的优化。此外，AI 技术在管理设计中的不确定性因素方面也显得尤为重要。面对材料参数的变化、工作环境的波动等不确定性因素，AI 能够进行鲁棒性设计，确保射频及天线系统在各种条件下都能保持稳定性和可靠性。这些能力使得 AI 成为射频及天线设计中不可或缺的工具，推动了设计方法的革新和设计质量的提升。

在面对复杂射频/微波结构的模型分析时，通常涉及电磁模型（精细模型）和集总参数等效电路模型（粗略模型）两种表示方法。精细模型虽然准确，但计算成本高昂；而粗略模型虽然计算速度快，但准确性较低。传统的空间映射技术能够在这两种模型之间转换参数，以优化几何参数并达到预期的性能 [27]。AI/ML 模型提供了另一种更为高效的解决方案，它们能够执行类似的参数转换，但所需的计算步骤更少。AI/ML 模型通过学习大量的数据，能够洞察精细模型和粗略模型之间的复杂关系，并快速准确地调整参数，从而辅助优化过程。这种方法不仅减少了对计算资源的需求，而且加快了设计迭代速度，使得设计者能够以更低的成本和更快的速度实现高性能的射频/微波结构设计 [28]。

4.4 自适应动态调整

在射频与天线系统中具有自适应与动态调整能力将极大地提高系统的适用性和鲁棒性。通过多种不同功能的 AI 系统实时监测并调整射频与天线系统，将显著提高系统多方面性能如图 7 所示。

图 7 AI 在射频电路与天线设计中的自适应调整

基于 AI 的设计能够实时监测系统参数状态，并通过智能算法自动调整射频电路与天线参数，以适应不断变化的工作条件，从而确保通信系统的性能和稳定性。这种实时监测与调整能力显著提升了射频电路与天线系统的灵活性和响应速度。

AI 技术在智能故障诊断方面同样发挥着重要作用。通过分析系统运行数据，AI 能够及时发现故障并预测潜在问题，从而采取预防或纠正措施，提高系统的可用性和可靠性。这不仅减少了系统维护的人力和时间成本，也避免了因故障导致的服务中断。

此外，AI 技术使射频及天线系统在动态环境中具有更好的自适应能力。在移动通信等应用场景中，信号干扰和其他环境因素可能会影响通信质量。AI 能够实时分析这些变化，并自动调整天线阵列的波束状态，以优化信号传输和接收，确保通信的连续性和稳定性。

4.5 跨学科融合与个性化

AI 技术在电路与天线设计领域的一项显著优势是其促进了跨学科融合与个性化设计的能力。AI 不仅能够整合来自不同领域的先进技术和理念，如材料科学、电子工程和计算机科学，实现创新的融合，还能够根据特定需求和应用场景提供定制化的射频及天线设计方案。这种跨学科的融合为射频电路和天线的设计带来了新的视角和解决方案，推动了技术的创新和发展。

个性化设计方面，AI 技术的应用使得设计过程更加灵活和客户化。AI 可以分析特定应用的需求，如通信范围、频率带宽和环境因素，从而生成满足这些需求的定制化设计方案。这种能力不仅满足了市场的多样化需求，还为特定应用提供了优化的解决方案，增强了射频系统的性能和适用性。

5 AI 模型设计实例

射频电路及天线设计是无线通信系统中最为核心且充满挑战的环节之一。随着 AI 技术的不断进步，其在射频电路设计和天线设计中的应用已经展现出显著的优势。随着我们深入探讨 AI 在射频及天线设计中的应用，我们将转向更具体的案例——滤波器设计。

5.1 带通、高通滤波器设计

通过对器件拓扑结构的像素化，使其在变量上具有了较高的自由度，从而可提高了达到预期性能的可能性。因此，该像素方案受到了微波研究界的广泛关注。各种微波电路都可以通过这种基于像素的结构实现，包括功率放大器和天线 [15]。我们也根据像素化结构的思想，设计了不同类型的微波滤波器。 设计目标是开发一种适用于 Ka 频段（26.5 GHz 至 40 GHz）的滤波器，该滤波器需满足特定的频率选择性要求，并已被广泛应用于自动驾驶车辆领域。首先，以设计带通和高通滤波器为例。采用厚度为 0.127 mm 的氧化铝作为衬底，采用 4 μm 厚的金作为传输线设计滤波器。对滤波器的设计可以遵循以下步骤： 第一步：在 CST 和 HFSS 等商业软件上建立具有两个端口 50 Ω 微带线的实心贴片的电磁模型。 第二步：根据谐振器的数量，将完整的 Patch 切分成不同大小的网格。 第三步：每个网格元素可以被指定为金属或非金属，在神经网络模型训练阶段将被对应为数值 1 或 0。 第四步：设置 Python 环境，实现自动执行电磁软件。 第五步：从电磁软件完成数据生成过程。 第六步：用二进制流作为输入（金属 = 1；非金属 = 0），S_{11} 和 S_{21} 参数作为输出，来训练 CNN 模型。 第七步：用训练好的 CNN 模型替换电磁模型，该模型在计算速度上比电磁模型更快。第八步：在所开发的 CNN 模型上建立一个合适的优化算法，该算法以通带回波损耗和阻带插入损耗为输入，能够输出所需的滤波器形状。

图 8 显示了在 CST 环境中设置的初级过滤器结构。为了减少数据生成时间，我们在目标结构中施加了对称平面，其中右半结构是左半结构的镜像。我们将左半平面分成 4 x

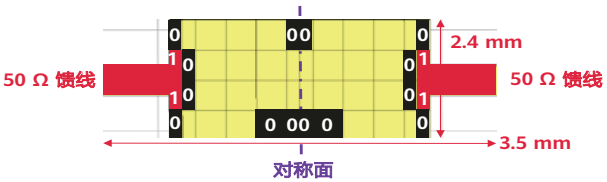


图 8 像素化滤波器的仿真示意图

5 = 20 个矩形块。如图 8 所示，根据之前联轴器的设计经验，将部分区域在优化之前进行了预设置，即图中红黑色方块的 1 或 0。另一方面，我们对于 I/O（输入输出）、I/R（谐振器间）耦合区域和谐振器区域设计了不同的网格尺寸，I/O 区域和耦合区域每个矩形块网格的尺寸为 0.12 mm x 0.3 mm，而谐振器区域每个矩形块网格的尺寸为 0.27 mm x 0.3 mm。选择不同尺寸带来的优势是尽可能保证精度的前提下，减少全波仿真所用的时间。通常，与谐振器尺寸相比，耦合区域太小（耦合间隙）。因此，我们在可能的耦合区域中保持较小的网格尺寸，而谐振器网格尺寸设置较大。通过这种方式，可以减小设计目标空间中的变量总数。在这种情况下，总共 15 个网格被视为变量。通过引入金属和非金属件，我们总共找到了 32768 个结构。因此，从 34 GHz 到 46 GHz 进行了 32768 次全波电磁仿真，共 51 个样本。我们将 15 个二进制变量和频率变量作为输入，而 S_{11} 和 S_{21} 的实部和虚部作为模型输出来训练 CNN 模型。最终训练出了能等效代替全波仿真的 EM-CNN 模型。

在成功开发了准确的 EM-CNN 模型后，我们结合了遗传算法（Genetic Algorithm，GA）优化方案，通过改变金属和非金属的排列组合（分布情况），以实现具有不同滤波性能的滤波器。首先，我们的目标是设计一个通带滤波器，其阻带定义为：

- 阻带 1（SB-1）≤ 40 GHz
- 阻带 2（SB-2）≥ 44 GHz
- 通带范围 41 GHz ≤ Passband (PB) ≤ 43 GHz
- 通带插入损耗 max[S_{21}(PB)] > −1 dB
- 通带最大回波损耗 max[S_{11}(PB)] < −10 dB
- 阻带抑制 max[S_{21}(SB)] < −25 dB

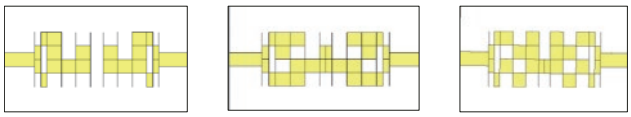
其次，我们设计了另一个高通滤波器，其中阻带定义为：

- 阻带 1（SB-1）≤ 40 GHz
- 通带范围 ≥ 42 GHz
- 带内最小插入损耗 max[S_{21}(PB)] > −1 dB
- 通带最大回波损耗 max[S_{11}(PB)] < −10 dB
- 阻带抑制 max[S_{21}(SB)] < −25 dB

由上述设计目标，我们以下列方式建立滤波器的损失函数：

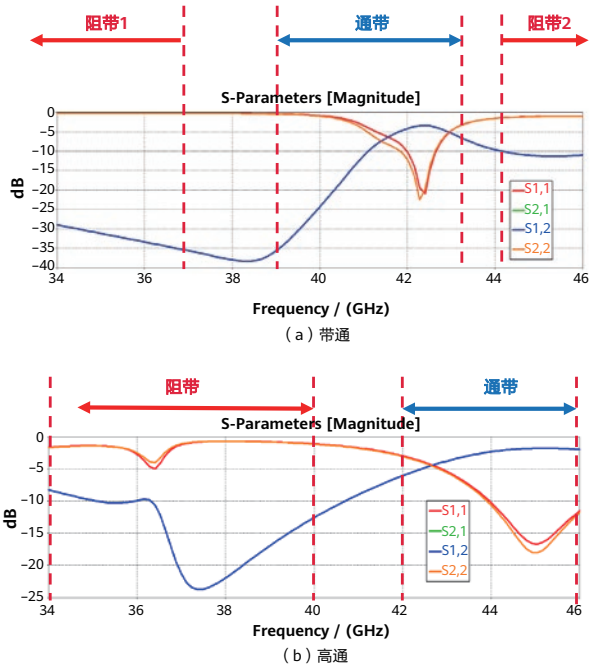
$$K = \max[(S_{11})_{PB}, -RL] + w * \max[(S_{21})_{SB}, -IL] \quad (4)$$

通带回损（RL）和阻带插入损耗（IL）之间的加权因子用 (4) 中的 w 表示，加权因子 w 是在通带回波损耗和阻带插入损耗之间起调节作用的关键因素。该系数是根据优化参数的灵敏度及其对通带和阻带中回波损耗和插入损耗特性的综合影响来确定的。对于带通滤波器，将对应的损失函数带入 EM-CNN 模型优化，可生成相应的二进制序列为 [1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1]。将这个序列转换成 CST 中的几何形状如图 9a 所示。对于高通滤波器，优化生成的二进制序列是 [0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1]，它被转换为 CST 中的几何形状如图 9b 所示。这些滤波器的全波仿真性能分别如图 10a 和图 10b 所示。滤波器带内带外性能基本满足预期要求，验证了我们的设计方法。



（a）带通　（b）高通　（c）双通带

图 9 像素化滤波器的优化结果



（a）带通

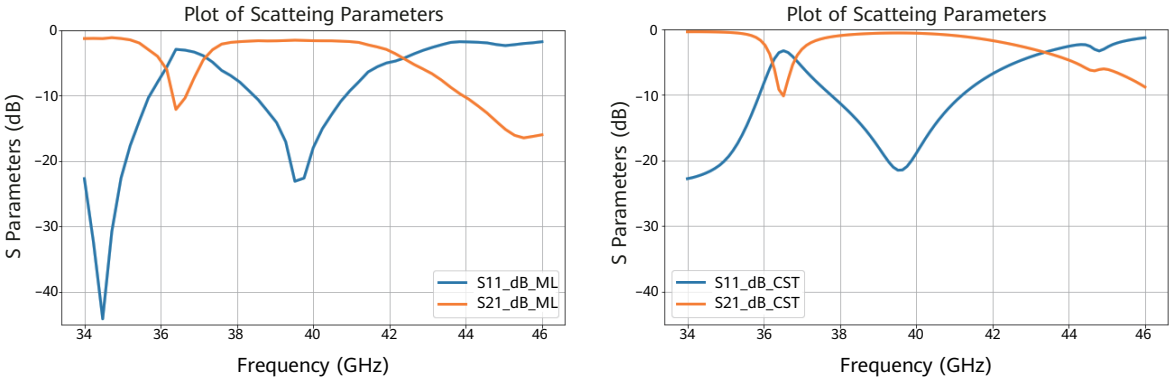
（b）高通

图 10 优化后滤波器的仿真结果

5.2 双通带滤波器设计

在设计了带通和高通两个滤波器之后，我们的目标是根据开发的 CNN 模型设计一个双通带滤波器。双通带滤波器较之前的单一通带需要设计更复杂的目标函数。

- 阻带 1: 36 GHz ≤ Stopband（SB-1）≤ 38 GHz
- 阻带 2: 44GHz ≤ Stopband（SB-2）≤ 46 GHz
- 通带 1: 34 GHz ≤ Passband (PB1) ≤ 36 GHz
- 通带 2: 38 GHz ≤ Passband (PB2) ≤ 42 GHz
- 通带插入损耗 max[S_{21}(PB)] > –2 dB
- 带内最大回波损耗 max[S_{11}(PB)] < –15 dB
- 带外抑制 max[S_{21}(SB)] < –25 dB



（a）CNN预测　　（b）全波电磁仿真（CST）

图 11 优化后滤波器的响应比较

我们选择 w = 0.8 来平衡通带回波损耗和阻带插入损耗，以实现损失函数最小化。与前两个示例类似，我们使用 GA 优化算法来获得二进制序列，该二进制序列代表滤波器的几何形状与对应的目标响应。GA 提供的优化二进制序列是 [1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1]。优化后的二进制序列对应的滤波器的几何形状如图 9c 所示。这里我们进一步对比了 CNN 预测结果与全波仿真结果，CNN 预测的滤波器 S 参数响应如图 11a 所示，CST 软件计算的 S 参数响应如图 11b 所示。我们可以看出 CST 计算的响应与 CNN 预测的响应具有极高的相似，同时也基本达到了所期望设计的滤波器性能。在上述的例子中仅是为了展示可行性，为了减少设计周期，其性能指标并没有严格满足。显然，通过进一步扩大滤波器自由度和加密网格能够实现更优的滤波器性能。

6 挑战与展望

6.1 数据质量和可用性

在天线设计领域， AI 模型的成功应用高度依赖于数据的产生、质量和可用性。大规模、高质量的数据是训练有效 AI 模型的关键因素，但在射频及天线设计领域，面临数据有限、专业性强和复杂性高的挑战。这些数据通常来源于电磁模型的仿真，其计算成本高昂，进一步增加了数据获取的难度。

一个准确的 AI/ML 模型需要充分的训练数据、验证数据和测试数据，以全面代表目标问题。在射频/微波结构的背景下，这些数据集通常由电磁模型仿真生成，这不仅成本高昂，而且耗时。此外，为了提高模型的准确性和泛化能力，必须努力收集、整理和标注相关数据，这涉及到对数据的细致选择和采样，确保数据集的质量和可用性能够代表设计空间的关键区域。

为应对射频及天线设计中 AI 模型训练数据的挑战，可以采取多种策略来优化数据生成和模型准确性之间的平衡：（1）主动学习（Active Learning）：通过主动学习策略，我们能够有选择性地生成最具信息量的数据点，这种方式可以在减少所需数据总量的同时，仍然保持模型的高准确性。（2）迁移学习（Transfer Learning）：利用在相关任务上预训练的模型，可以显著减少所需的新数据量，因为模型可以基于之前学习到的特征和模式进行构建。（3）数据增强（Data Augmentation）：应用数据增强技术可以帮助人为增加数据集的规模，减少对额外电磁仿真的需求，同时保持训练数据的多样性。（4）降维（Dimensionality Reduction）：使用降维技术可以集中关注最关键的参数，简化数据生成过程并降低计算成本。平衡最小化数据生成时间和最大化模型准确性的需求涉及战略性的权衡。

除了传统的数据收集和整理方法，还可以利用 AI 技术自身来生成所需的数据。如扩散模型（Diffusion Model）和其他类型的生成模型，提供了一种创新的数据生成方式。这些模型能够学习现有数据的分布特征，并生成新的、与真实数据相似的数据样本。在射频及天线设计中，这意味着可以从有限的真实测量或仿真数据出发，生成更多的数据以扩充训练集。这种方法可以捕捉到数据的复杂结构，使得生成的数据既具有多样性又保持了与原始数据集的一致性。此外，GAN 也是一种强大的数据生成工具，它们通过对抗训练的方式生成逼真的数据样本。在射频及天线设计中，GAN 可以用于生成电磁仿真数据，帮助设计者在早期设计阶段评估不同设计方案的性能。采用这些策略可以帮助优化这种平衡，并为射频/微波结构开发有效的 AI/ML 模型。这些方法的运用不仅加速了 AI 模型的开发过程，也为射频及天线设计领域带来了创新和高效的解决方案。

6.2 模型选择和参数配置

在射频/微波结构设计中，开发 AI/ML 模型时面临的一个主要问题是模型选择和超参数配置。由于不同问题的特性差异，没有单一模型能够适用于所有场景，因此必须根据问

题陈述选择最合适的模型架构。此外，超参数的选择，如层数、神经元数量、数据分割比例、激活函数等，对模型性能有显著影响。目前，超参数的选择很大程度上依赖于专家经验和试错方法，这不仅耗时耗力，而且可能引入模型开发的不确定性。因此，如何高效、准确地选择模型和配置超参数，是实现高性能 AI/ML 模型的关键挑战。

为了解决模型选择和超参数配置的问题，可以采取以下策略：（1）开发自动化的模型选择工具，这些工具能够基于问题特性和数据特征推荐或选择最合适的模型架构。（2）使用超参数优化技术，如网格搜索、随机搜索或贝叶斯优化，来系统地探索超参数空间，找到最优的超参数组合。（3）开发自动化流程，集成数据预处理、模型训练、超参数优化和模型评估，以减少人工干预并提高开发效率。

6.3 算法复杂度和计算资源

在射频及天线设计中应用 AI/ML 算法时，算法复杂度和计算资源成为了显著的瓶颈问题。高级的 AI 算法，尤其是深度学习模型，往往需要大量的计算资源来进行训练和推理。这些资源包括但不限于高性能的 GPU、大量的存储空间以及快速的数据处理能力。然而，这些资源可能昂贵且难以获得，特别是对于研究和小型开发团队。此外，高度复杂的算法可能导致训练时间过长，从而减缓了设计迭代速度和创新进程。因此，如何平衡算法的复杂度与可用计算资源，成为实现高效 AI 辅助设计的关键问题。

为了解决算法复杂度和计算资源的问题，可以考虑以下解决方案：（1）算法优化：研究和开发更高效的算法，减少计算步骤和资源消耗，同时保持或提高模型性能。（2）模型简化：通过模型剪枝、知识蒸馏等技术，简化模型结构，减少模型参数，以降低计算负担。（3）硬件加速：利用专用硬件如张量处理器（Tensor Processing Unit，TPU）、现场可编程门阵列（Field Programmable Gate Array，FPGA）等进行 AI 算法的加速，这些硬件为深度学习提供了优化的计算能力。（4）云计算资源：利用云计算服务提供的强大计算资源，按需分配，以降低本地硬件投资成本。（5）并行计算：采用并行计算技术，将训练任务分配到多个处理器或设备上，以缩短训练时间。（6）资源调度和管理：开发智能的资源调度系统，优化计算资源的分配和使用，提高资源利用率。（7）轻量级模型：开发轻量级的 AI 模型，如 MobileNet、ShuffleNet 等，这些模型专为资源受限的环境设计。（8）模型量化：通过模型量化技术减少模型的精度要求，降低计算复杂度，同时减少模型对内存和存储的需求。（9）异构计算资源利用：结合使用不同类型的计算资源，如 CPU、GPU、专用集成电路（Application-specific Integrated Circuit，ASIC）等，以实现计算任务的最优分配。通过这些策略的实施，可以在有限的计算资源下，有效降低 AI 算法的复杂度，加快设计流程，提高射频及天线设计的效率和可行性。

6.4 可解释性与可迁移性

在射频及天线设计时，AI 模型的可解释性是一个重要但常被忽视的问题。尽管一些多目标、多功能 AI 模型，特别是深度学习模型，在解决特定复杂问题时表现出色，但它们通常被认为是“黑箱”，因为很难理解模型是如何做出特定决策的。在射频及天线设计中，设计者需要理解模型的决策过程，以确保设计满足物理原理和实际应用的要求。缺乏可解释性不仅增加了设计风险，也限制了 AI 模型在关键应用中的采纳。

为了解决射频及天线设计中 AI 模型的可解释性问题，可以采取以下策略：（1）开发和应用可解释的 AI 技术，如 LIME 和 SHAP，这些技术可以提供模型预测的解释。（2）利用可视化工具来展示模型内部的工作原理，包括特征的重要性和决策边界。（3）使用更简单、更易于理解的模型，如决策树或线性模型，尽管它们可能在某些情况下不如复杂模型性能优越。（4）通过模型蒸馏技术，将复杂模型的知识迁移到一个更简单的模型中，以提高可解释性。（5）采用后处理技术，如规则提取，从黑盒模型中提取可解释的规则。（6）在模型设计阶段就考虑可解释性，选择那些天然具有透明性的模型架构。通过这些方法，可以提高 AI 模型在射频及天线设计中的可解释性，增强设计者对模型的信任，并促进 AI 技术在这一领域的应用。

但是，AI 模型的可解释性和可迁移性之间的关系一般是矛盾的。例如，高度复杂的模型可能在新环境（可迁移性）中表现良好，但可能难以解释（可解释性差），而简单模型可能容易解释，但可能在多样化的环境或数据上（鲁棒性和可迁移性）表现不佳。缺乏可迁移性，不仅在设计初期需要多个模型来适配不同场景，而且还会在模型应用后期缺乏应变特殊情况的能力，即提高了设计成本也加大了维护难度。因此，正如物理学界追求的大一统理论一样，提出契合射频电路网络的新 AI 模型，在模型具备可解释的同时，使其充分考虑多种情况即具有较为广阔的泛化能力（可迁移性），是实现 AI 技术在射频及天线设计中广泛应用的关键挑战之一。

6.5 多领域融合

射频及天线设计是一个多学科交叉的领域，涉及电磁学、材料科学、电子工程等多个学科。AI 模型在这一领域的应用需要能够处理和融合来自不同学科的复杂数据和知识。然而，多领域融合在 AI 模型开发中面临诸多挑战。不同学科的数据可能具有不同的特性和格式，难以直接整合。此外，不同领域的专业知识和理论需要有效地结合，以确保 AI 模型能够全面地理解和解决问题。当前，AI 模型往往专注于单一领域，缺乏跨学科的集成和协同，限制了其在射频及天线设计中的性能和应用范围。

为了解决射频及天线设计中 AI 模型的多领域融合问题，可以考虑以下策略：（1）组建包含不同领域专家的团队，

共同参与 AI 模型的开发，确保模型能够综合考虑多学科的知识和数据。（2）开发数据标准化流程，将不同来源和格式的数据转化为统一格式，便于 AI 模型处理和分析。（3）应用多任务学习技术，使 AI 模型能够同时学习多个相关任务，促进不同领域知识的融合。（4）利用迁移学习技术，将在一个领域学到的知识迁移到另一个领域，提高模型在新领域的适应性和性能。（5）设计领域特定的模型架构，能够更好地表示和处理特定学科的数据和知识。（6）构建跨学科的知识图谱，整合不同领域的专业知识，为 AI 模型提供丰富的背景知识。

7 结语

本文详细探讨了人工智能在射频及天线设计中的应用、优势、挑战以及未来的发展方向。AI 技术在天线优化、射频电路设计和电磁仿真等方面展现出了巨大的潜力，能够提高设计效率、优化设计结果并处理复杂问题。然而，数据质量、模型选择、算法复杂度、模型可解释性与可迁移性等挑战也需要得到关注和解决。尽管面临一些挑战，AI 在射频及天线设计中的应用前景依然广阔。通过不断的研究和创新，我们有理由相信 AI 将为射频及天线设计带来更大的突破和发展。AI 为射频及天线设计带来了新的机遇，通过提高设计效率和优化性能，有望推动无线通信领域的进一步发展。然而，要实现其全部潜力，还需要解决一些技术和数据方面的挑战。未来的研究将集中在提高数据质量、开发更高效的算法、提高模型的可解释性上以及促进多领域融合等方面。深入研究 AI 技术在射频及天线设计中的应用，推动无线通信系统的发展，为人们的生活带来更多便利。

参考文献

[1] W. Tong and P. Zhu, Eds., "6G: The next horizon: From connected people and things to connected intelligence," Cambridge: Cambridge University Press, 2021. doi: 10.1017/9781108989817.

[2] S. K. Goudos, P. D. Diamantoulakis, M. A. Matin, P. Sarigiannidis, S. Wan, and G. K. Karagiannidis, "Design of antennas through artificial intelligence: State of the art and challenges," *IEEE Commun. Mag.*, vol. 60, no. 12, pp. 96–102, Dec. 2022, doi: 10.1109/MCOM.006.2200124.

[3] N. Sarker, P. Podder, M. R. H. Mondal, S. S. Shafin, and J. Kamruzzaman, "Applications of machine learning and deep learning in antenna design, optimization, and selection: A review," *IEEE Access*, vol. 11, pp. 103890–103915, 2023, doi: 10.1109/ACCESS.2023.3317371.

[4] H. J. KELLEY, "Gradient theory of optimal flight paths," *ARS J.*, Jun. 2012, doi: 10.2514/8.5282.

[5] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022, doi: 10.1109/TPAMI.2021.3117837.

[6] "Time-delay neural networks for control," *IFAC Proc. Vol.*, vol. 27, no. 14, pp. 967–972, Sep. 1994, doi: 10.1016/S1474-6670(17)47423-4.

[7] "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, Apr. 2017, doi: 10.1016/j.neucom.2016.12.038.

[8] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[9] GoodfellowIan *et al.*, "Generative adversarial networks," *Commun. ACM*, Oct. 2020, doi: 10.1145/3422622.

[10] "Knowledge-based artificial neural networks," *Artif. Intell.*, vol. 70, no. 1–2, pp. 119–165, Oct. 1994, doi: 10.1016/0004-3702(94)90105-8.

[11] F. Feng, W. Na, J. Jin, J. Zhang, W. Zhang, and Q.-J. Zhang, "Artificial neural networks for microwave computer-aided design: The state of the art," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4597–4619, Nov. 2022, doi: 10.1109/TMTT.2022.3197751.

[12] S. Haykin, "Neural networks: A comprehensive foundation," Subsequent. Upper Saddle River, N.J: Prentice Hall, 1998.

[13] C. Roy and K. Wu, "Homotopy optimization and ANN modeling of millimeter-wave SIW cruciform coupler," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4751–4764, Nov. 2022, doi: 10.1109/TMTT.2022.3200040.

[14] A. Pietrenko-Dabrowska and S. Koziel, "Low-cost design optimization of microwave passives using multifidelity EM simulations and selective broyden updates," *IEEE Trans. Microw. Theory Tech.*, vol. 70, no. 11, pp. 4765–4771, Nov. 2022, doi: 10.1109/TMTT.2022.3207482.

[15] E. A. Karahan, Z. Liu, and K. Sengupta, "Deep-learning-based inverse-designed millimeter-wave passives and power amplifiers," *IEEE J. Solid-State Circuits*, vol. 58, no. 11, pp. 3074–3088, Nov. 2023, doi: 10.1109/JSSC.2023.3276315.

[16] C. Roy and K. Wu, "A review of electromagnetics-based microwave circuit design optimization," *IEEE Microw. Mag.*, vol. 25, no. 7, pp. 16–40, Jul. 2024, doi: 10.1109/MMM.2024.3387036.

[17] B. Liu *et al.*, "An efficient method for complex antenna design based on a self adaptive surrogate model-assisted optimization technique," *IEEE Trans. Antennas Propag.*, vol. 69, no. 4, pp. 2302–2315, Apr. 2021, doi: 10.1109/TAP.2021.3051034.

[18] D. Shi, C. Lian, K. Cui, Y. Chen, and X. Liu, "An intelligent antenna synthesis method based on machine learning," *IEEE Trans. Antennas Propag.*, vol. 70, no. 7, pp. 4965–4976, Jul. 2022, doi: 10.1109/TAP.2022.3182693.

[19] Z. Ma, K. Xu, R. Song, C.-F. Wang, and X. Chen, "Learning-based fast electromagnetic scattering solver through generative adversarial network," *IEEE Trans. Antennas Propag.*, vol. 69, no. 4, pp. 2194–2208, Apr. 2021, doi: 10.1109/TAP.2020.3026447.

[20] Z. Wei and X. Chen, "Deep-learning schemes for full-wave nonlinear inverse scattering problems," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 1849–1860, Apr. 2019, doi: 10.1109/TGRS.2018.2869221.

[21] R. Guo, M. Li, F. Yang, S. Xu, G. Fang, and A. Abubakar, "Application of gradient learning scheme to pixel-based inversion for transient EM data," in *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*, Mar. 2018, pp. 1–3. doi: 10.1109/COMPEM.2018.8496518.

[22] L. Li, L. G. Wang, F. L. Teixeira, C. Liu, A. Nehorai, and T. J. Cui, "DeepNIS: Deep neural network for nonlinear electromagnetic inverse scattering," *IEEE Trans. Antennas Propag.*, vol. 67, no. 3, pp. 1819–1825, Mar. 2019, doi: 10.1109/TAP.2018.2885437.

[23] Y. Sun, Z. Xia, and U. S. Kamilov, "Efficient and accurate inversion of multiple scattering with deep learning," *Opt. Express*, vol. 26, no. 11, pp. 14678–14688, May 2018, doi: 10.1364/OE.26.014678.

[24] "Physics-inspired convolutional neural network for solving full-wave inverse scattering problems | IEEE Journals & Magazine | IEEE Xplore," Accessed: Jul. 10, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8741152

[25] Q.-J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural networks for RF and microwave design - from theory to practice," *IEEE Trans. Microw. Theory Tech.*, vol. 51, no. 4, pp. 1339–1350, Apr. 2003, doi: 10.1109/TMTT.2003.809179.

[26] C. Roy, W. Lin, and K. Wu, "Swarm intelligence-homotopy hybrid optimization-based ANN model for tunable bandpass filter," *IEEE Trans. Microw. Theory Tech.*, vol. 71, no. 6, pp. 2567–2581, Jun. 2023, doi: 10.1109/TMTT.2023.3236676.

[27] J. W. Bandler, R. M. Biernacki, S. H. Chen, P. A. Grobelny, and R. H. Hemmers, "Space mapping technique for electromagnetic optimization," *IEEE Trans. Microw. Theory Tech.*, vol. 42, no. 12, pp. 2536–2544, Dec. 1994, doi: 10.1109/22.339794.

[28] C. Roy and K. Wu, "Surrogate model-based filter optimization by a field-circuit model mapping," *IEEE Trans. Microw. Theory Tech.*, vol. 72, no. 5, pp. 3144–3157, May 2024, doi: 10.1109/TMTT.2023.3318692.

物理启发的智能通信：机遇、进展和趋势

邢子青，黎日东，陈子瑞，杨照辉，张朝阳
浙江大学信息与电子工程学院

摘要

近年来，人工智能（Artificial Intelligence，AI）正在逐渐成为推动无线网络演进的关键使能技术。将深度神经网络用于解决无线问题，已经被证实可以改善许多任务的性能，甚至催生崭新的无线用例。但不可忽视的是，现有的无线 AI 主要依赖于在常规的深度学习模型的基础上以无线数据驱动产生无线智能，这种智能的泛化性、可靠性和可拓展性往往不够理想。而与数据驱动相对的，物理启发的智能通信（Physics-inspired Intelligent Communication，PIC）通过在深度学习的过程中融合无线问题中的先验物理规律，能够显著提高无线智能的质量和可用性。在本文中，我们对 PIC 对于无线 AI 发展的意义进行介绍。我们首先提炼了 PIC 的核心思想，并以 PIC 在信道状态信息（Channel State Information，CSI）压缩反馈、信道预测和用户定位等任务中的应用为例展示了它的典型技术方案。在此基础上，我们以移动信道预测问题为例阐述了如何综合运用多种 PIC 技术来解决复杂无线问题。最后，我们探讨了 PIC 技术在无线系统中的前景和挑战。

关键词

无线 AI，PIC

1 引言

以深度神经网络（Deep Neural Network，DNN）为代表的人工智能（Artificial Intelligence，AI）技术，凭借其在隐式特征提取、高维数据表征、复杂决策等方面的出色能力，正在赋能大量的科学和工业技术的发展。因此，在 6G 网络的演进中，AI 和通信的融合既被认为是令人期待的重要愿景，也被视为解决诸多关键性技术难题的必要手段 [1]。近年来，深度学习技术已经被用于包括信道状态信息（Channel State Information，CSI）压缩反馈 [2]、信道预测、高精度无线定位 [3] 等多种典型无线任务，并表现出了相比于传统信号处理显著的性能增益。

然而，现有无线 AI 模型的设计大多是根据无线信道在数据结构和数值特性上的某些表层特征，将其类比为传统的图像或者序列，并借鉴计算机视觉和自然语言处理领域的模型进行相应的处理。文献 [2] 根据角度时延域信道的邻域特性，将 CSI 反馈任务视为图像压缩任务并用卷积神经网络（Convolutional Neural Network，CNN）进行处理。文献 [4] 将时变信道视为时间序列，并采用长短时记忆（Long Short-Term Memory，LSTM）网络来处理信道预测问题。事实上，无线信道本质上是由基站位置、用户移动、以及信号传播机制决定的，具有明确的内在物理特性，与经典的图像和序列模型的先验条件之间存在显著的差异。首先，信道矩阵的索引对应了明确的物理量，比如天线和子载波，因此其不具有图像中的平移不变性。另一方面，由于电磁波的周期和波长通常远小于对信道进行时空采样的尺度，信道序列具有波动的相位变化特性，是高度非平滑的，而传统序列模型的预测精度往往依赖于序列数据的数值相关性和平滑性。这种无线信道内在特征和 AI 模型先验之间的不匹配从根源上限制了其在无线任务中的表现，致使现有的无线 AI 技术在训练代价、泛化性、可靠性等方面表现出令人担忧的一面，这些局限性使得 AI 技术仍无法在无线系统中得到全面的应用。

为了抑制深度学习对数据和训练过程的过分依赖，进一步提高无线 AI 技术的性能和泛化能力，物理启发的方法正在得到研究者们的关注。这类方法的核心思想在于通过将学习过程与具有普适性的物理规律进行融合，使学习到的智能兼具 DNN 的数据驱动与物理规律的模型驱动的优点。具体来说，有别于主要根据任务中的数据结构和数据形式选择或组合当下流行的 DNN 结构的思路，物理启发的智能通信（Physics-inspired Intelligent Communication，PIC）强调挖掘任务与数据在无线系统中所代表的物理含义，并通过先验的设计向神经网络传递这些可利用的归纳偏置，以引导神经网络学习到更贴近物理本质的无线智能，从而根本性地满足无线系统的多样化需求 [5]。当前，已有一些工作 [6–11] 在某些具体无线任务上对 PIC 的设计准则和效用进行了初步的探索和分析，反映出了物理启发的无线 AI 技术在性能、泛化性以及应用灵活性等方面的巨大潜力。

因此，在现有技术性工作的基础上，本文对 PIC 技术进行一个阶段性的全面分析、评估和展望，着重分析一些有代表性的设计思路和技术方案，并评估这类方法在无线系统中的具体意义，同时对物理启发的无线 AI 的发展趋势和未来研究方向进行展望。

本文后续的行文结构如下：

- **第 2 节介绍 PIC 的技术框架**。我们将以 CSI 压缩反馈、静态信道预测、用户定位等具体无线 AI 任务为例，介绍如何从神经网络的连接结构、激活函数、损失函数以及非完全黑盒模型这四个层面设计 PIC 方案。
- **第 3 节介绍 PIC 的综合应用**。我们以移动信道预测这一富有挑战性的任务为例，介绍如何根据无线环境的物理特性，综合运用多种 PIC 技术以应对更为复杂的无线任务，并通过实验结果证明 PIC 相比于传统无线 AI 方案的性能优势。
- **第 4 节介绍 PIC 当前面临的挑战以及未来的研究方向**。

2 物理启发的智能通信技术框架

鉴于 PIC 技术在提升无线通信性能和效率方面的巨大潜力，我们需要发掘无线问题的深层物理机制，并设计与之相适配的神经网络架构。

在本节中，我们将介绍 PIC 的技术框架，如图 1 所示。具体而言，我们可以根据无线通信场景中的内在物理机制，从如下四个层面考虑物理启发的设计：

- **连接结构的设计**。比如利用空频域信道的二维序列性质，设计相应的神经网络结构。
- **激活函数的设计**。比如利用周期性激活函数来捕获信道在时空中的相位变化。
- **损失函数的设计**。比如设计累积损失函数来融合多个频率下 CSI 的定位信息。

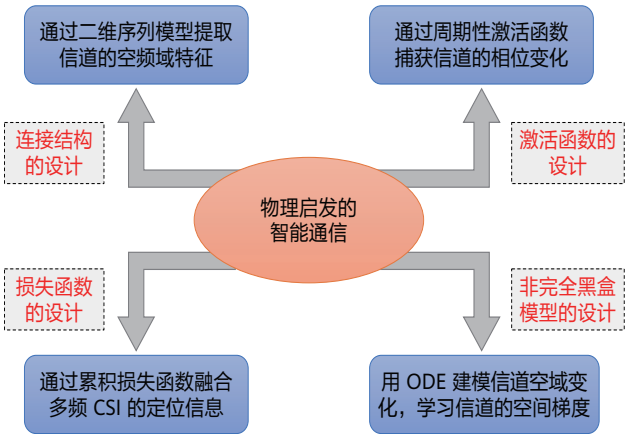


图 1 PIC 的技术框架和设计案例

- **非完全黑盒模型的设计。**比如采用常微分方程（Ordinary Differential Equation，ODE）建模信道空域变化，并通过 DNN 学习信道的空间梯度。

下面，我们将分析 CSI 信道反馈、信道预测、用户定位等重要无线任务的内在物理机制，详细介绍针对神经网络连接结构、激活函数、损失函数以及非完全黑盒模型的设计方案。

2.1 连接结构的设计

在 MIMO-OFDM 通信系统中，基站需要获取下行信道的 CSI 来进行波束赋形、载波分配和功率控制。在频分双工（Frequency Division Duplexing，FDD）模式下，需要在用户设备（User Equipment，UE）侧估计下行 CSI，并将其反馈给基站侧。随着无线通信系统的演进，MIMO-OFDM 系统的天线和子载波数量将不断增加，导致 CSI 反馈的成本过高。因此，如何有效压缩 CSI 以降低反馈开销已成为重要的课题。

文献 [2] 提出了 CsiNet，将空频域信道转化到角度时延域，并利用 CNN 对 CSI 进行特征提取以对其进行压缩表征，这一设计利用了角度时延域信道的稀疏性和邻域相关性。然而，不同于自然图像，无论是空频域还是角度时延域的 CSI，其元素的索引都具有明确的物理含义，这意味着特征的索引位置对于 CSI 而言是至关重要的，而 CNN 具有平移不变性，当 CSI 通过 CNN 时将会损失索引位置的信息。此外，角度时延域信道虽然具有邻域相关性，但并不具备和图像类似的平滑特性，这也给基于 CNN 的特征提取带来了挑战。

根据 MIMO-OFDM 多径信道的表达形式，空频域 CSI 矩阵 $\mathbf{H} \in \mathbb{C}^{N_t \times N_c}$ 的元素为

$$\mathbf{H}_{n,m} = \sum_{p=1}^{P} \alpha_p e^{-j2\pi(f_0+(m-1)\Delta f)\tau_p + j\varphi_p} e^{-j\chi(n-1)\cos\theta_p} \quad (1)$$

其中，$\alpha_p e^{j\varphi_p}$，τ_p，θ_p 分别为第 p 条径的复系数、时延和到达角，$\chi = 2\pi d f_c/c$，d 为天线间距，Δf 为子载波间隔。观测式 (1)，考虑 $\mathbf{H}$ 的第 m 列（即第 m 个子载波上的空域 CSI），任何两个相邻天线之间的 CSI 差异都是由 P 条多径的空域相位变化 $\chi\cos\theta_p$ 引起的。类似的，对于 $\mathbf{H}$ 的第 n 列（即第 n 根天线上的频域 CSI），任何两个相邻子载波之间的 CSI 差异都是由 P 条多径的频域相位变化 $2\pi\Delta f\tau_p$ 引起的。因此，空频域 CSI 矩阵 $\mathbf{H}$ 具有独特的二维序列性质。基于此，我们将 $\mathbf{H}$ 划分为 $L_{\text{ver}} \times L_{\text{hor}}$ 个子块，并设计了如图 2 所示的 2D LSTM 对其进行特征提取，每一个 2D LSTM 块能够同时沿着空间和频率两个物理维度传递信息 [6]。

不同于 CsiNet 中 CNN 的邻域连接方式，该模型根据多径信道的物理模型，设计了全新的 2D Seq2Seq 连接结构，以实现空域和频域两个维度上的序列特征提取。我们将其作为 CSI 反馈任务中编码器和解码器的主体结构，实现了更高效的 CSI 反馈模型。

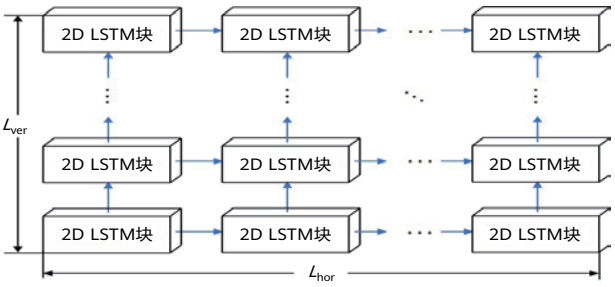


图 2 基于信道二维序列性质的 CSI 特征提取

2.2 激活函数的设计

由于无线散射环境的复杂性，无线信道通常包含大量未知参数，若对 CSI 进行实时的精确估计将会产生难以承受的信令开销。然而，在实际通信场景中，基站总是服务某一固定区域，其中的主要散射体（例如建筑物）并不会发生明显变化。因此，我们可以利用历史通信数据来训练一个静态信道预测模型，采用神经网络来隐式表示散射环境和电磁波传播特性，根据用户的位置坐标预测其对应的信道冲激响应（Channel Impulse Response，CIR），从而大幅降低获取 CSI 的信令开销。

信道预测任务是富有挑战性的，因为它需要从低维的坐标信息恢复高维的 CIR 向量。除此之外，微小的空间偏移量就能导致 CIR 相位的巨大变化，而采用 ReLU 等传统激活函数的神经网络难以拟合 CIR 在时空中的高频变化。

在静态信道下，任意 TX-RX 对之间的 CIR 可以表述为

$$h(f,\tau) = \sum_{p=1}^{P} \alpha_p \delta\left(\tau - \frac{d_p}{c}\right) e^{-j2\pi f\frac{d_p}{c}} \quad (2)$$

其中传播衰减 $\alpha_p \propto 1/d_p$，d_p 为第 p 条路径的长度。可以发现，CIR 呈现出振幅上的反比例特性和相位上的周期特性。在远场条件下，反比例函数很适合用一系列径向基函数（Radius Basis Function，RBF）进行拟合。同时，文献 [12] 表明采用周期性激活函数的多层感知机（Multilayer Perceptron，MLP）更适合对具有高频细节的场景进行隐神经表征。根据上述性质，我们在 [7] 中提出了创新的静态信道预测模型——余弦-高斯径向基函数网络（Cosine-Gaussian Radial Basis Function Network，C-GRBFNet），通过 MLP 提取散射环境中的像点坐标，并采用具有如下形式的余弦-高斯径向基函数（Cosine-Gaussian Radial Basis Function，C-GRBF）作为激活函数

$$\phi(\boldsymbol{x}) = \cos(\omega|\boldsymbol{x}-\boldsymbol{a}|+b)\cdot\exp(\beta|\boldsymbol{x}-\boldsymbol{c}|^2) \quad (3)$$

以模拟不同传播路径带来的信道幅值和相位响应。为了处理复值 CIR，激活函数分别输出同相和正交分量，如图 3 所示。

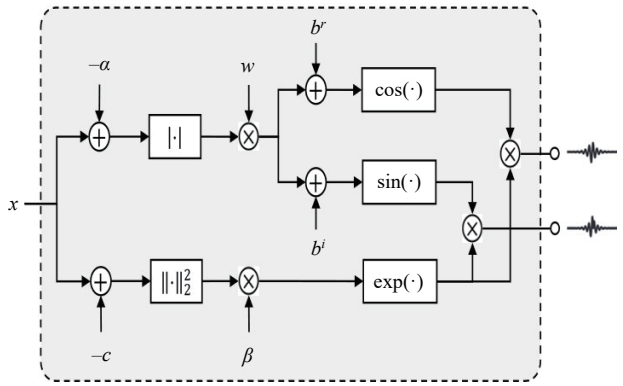


图 3 基于信道衰减和波动特征的 C-GRBF 激活函数

由于在激活函数的设计上引入了信道响应的先验结构，C-GRBFNet 只用一层 C-GRBF 函数便能够捕获信道的衰减和相位特性。在低采样密度情况下，C-GRBFNet 展现出比传统的基于自编码器的信道预测方案和 [12] 中的正弦表示网络（Sinusoidal Representation Network，SIREN）更高的预测准确度。

2.3 损失函数的设计

高精度无线定位将成为下一代无线系统的重要服务，从 CSI 指纹推断位置信息是一种模式识别任务，非常适合采用深度学习技术进行处理。在 Massive MIMO 系统下，可以有效地对单个子载波的 CSI 进行多径分离，即

$$\varphi_1: \mathbf{h}(f) \to \theta_p, \alpha_p \tag{4}$$

根据理想的反射模型，路径衰减因子为

$$\alpha_p = \frac{c \cdot \prod_{i=1}^{k} \gamma_i}{4\pi f d_p} \tag{5}$$

其中 γ_i 为第 i 次反射的衰减。对于固定场景和某一到达角 θ_p，随着路径总长度 d_p 的增加，反射次数 k 将会增加，因此 α_p 相对于 d_p 单调递减，即

$$\varphi_2: \theta_p, \alpha_p \to \theta_p, d_p \tag{6}$$

同时，在固定场景下，根据已知的基站位置、到达角 θ_p 和传播路径长度 d_p，用户位置可以唯一确定，即

$$\varphi_3: \theta_p, d_p \to \boldsymbol{x} \tag{7}$$

因此，我们发现此时用户位置和单载波 CSI 之间存在一一映射关系

$$\varphi: \mathbf{h}(f) \to \theta_p, \alpha_p \to \theta_p, d_p \to x \tag{8}$$

在实际应用中，由于噪声、散射等非理想因素的存在，上述映射关系虽然难以准确建立，但是采用单载波 CSI 进行用户定位的想法依然是有价值的。

在 [13] 中，我们提出不直接将整个 MIMO-OFDM CSI 矩阵 $\mathbf{H}$ 作为指纹，而是将多个子载波的指纹定位结果进行融合。根据这一设计思路，我们提出了多载波累积学习神经网络（Multi-Channel Correction Network，MCCNet），通过 LSTM 网络传递各个子载波提取出的指纹特征，并输出多载波融合后的定位结果。在训练阶段，我们希望累积过程中的定位结果 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{N_c}$ 逐渐接近真实坐标 $\boldsymbol{x}$，因此设计了如下的累积学习损失函数

$$Loss(\Theta_{\mathrm{MCC}}) = \sum_{n=1}^{num} \sum_{i=1}^{N_c} \| w_i \cdot (\boldsymbol{x}_i - \boldsymbol{x})_n \|_2^2 \tag{9}$$

其中加权系数为

$$w_i = \frac{2i}{N_c + 1}, i = 1,2, \ldots, N_c \tag{10}$$

在推理阶段，我们采用子载波指纹信息积累最充分的 $\boldsymbol{x}_{N_c}$ 作为最终定位结果。整体训练和推理过程如图 4 所示。

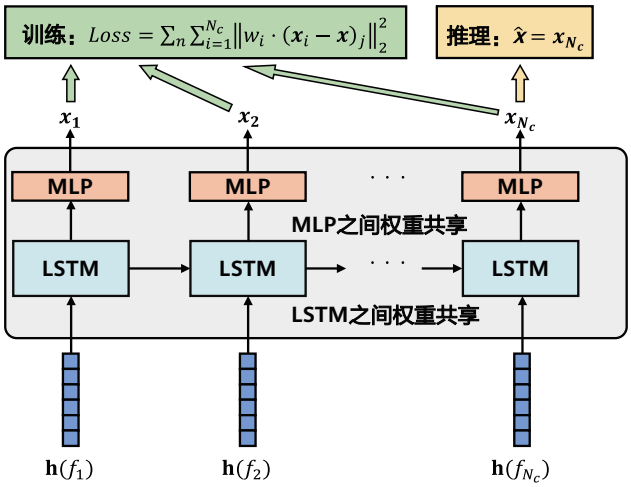


图 4 基于单载波指纹提取和多载波融合的定位方案

得益于 Massive MIMO 信道和用户位置之间的潜在映射关系，这种单载波特征提取和多载波累积学习的方案引入了更充足的物理先验，使得 MCCNet 在定位任务上相比于采用 CNN 对角度时延域信道进行特征提取的方案 [3] 展现出显著的性能优势。

2.4 非完全黑盒模型的设计

我们已经介绍了针对无线 AI 模型的连接结构、激活函数、损失函数的 PIC 设计方案，这些方案使得模型先验更符合无线问题的物理特性，让 PIC 展现出在性能、泛化能力等方面的优势。然而，由于神经网络黑盒化的训练方式，仅对神经网络的某些模块进行物理启发的设计并不能保证神经网络的运行方式符合物理过程，这也影响了无线 AI 模型的

可解释性。为了将物理信息进一步融入无线 AI 模型的设计之中，我们可以将解析或部分解析的物理方程直接嵌入无线 AI 模型的架构之中，实现非完全黑盒的 PIC 方案。

在准静态散射环境中，用户坐标和信道响应之间存在一一映射关系，这使得信道矩阵 $\mathbf{H}$ 对空间位移 $\mathbf{m}$ 的导数 $\frac{\partial \mathbf{H}}{\partial \mathbf{m}}$ 仅与当前信道 $\mathbf{H}$ 和用户移动方向 θ_m 有关，这驱使我们采用 ODE 来建模信道的空间变化率：

$$\frac{\partial \mathbf{H}}{\partial \mathbf{m}} = f(\mathbf{H}, \mathbf{m}) \tag{11}$$

然而，由于散射环境存在大量未知参数，空间梯度函数 $f(\cdot)$ 无法显式表述。在 [9] 中，我们采用 Neural ODE 来隐式表征信道 $\mathbf{H}$ 的空间梯度 $\frac{\partial \mathbf{H}}{\partial \mathbf{m}}$，将其称为空间信道梯度网络（Spatial Channel Gradient Network，SCGNet）。完成训练之后，可以通过前向积分的方式，实现从历史信道预测当前位置的信道，如图 5 所示，

$$\mathbf{H}_{\text{pred}} = \mathbf{H}_0 + \int_0^S f_\theta(\mathbf{H}, \mathbf{m}) d\mathbf{m} \tag{12}$$

这一设计使得信道预测模型不再是直接学习空间位置 $\mathbf{x}$ 到信道矩阵 $\mathbf{H}$ 的高维映射函数，而是学习信道在空间中的变化率，然后根据历史 CSI 和用户位置，通过解析的积分运算完成信道预测任务，大幅降低了模型的学习难度。

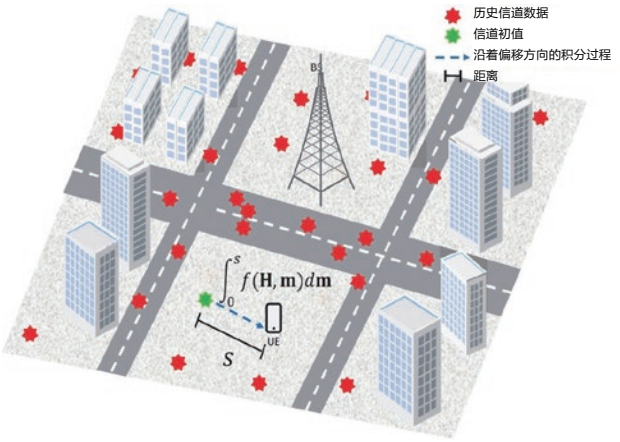


图 5 基于空间梯度学习的信道预测方案

3 物理启发的智能通信技术的综合应用

上一节中的四种 PIC 技术方案并非是分立的，而是互为补充的。在应对实际的无线 AI 问题时，需要针对具体的物理模型和应用场景，综合运用各种 PIC 技术方案。

在特定的 MIMO-OFDM 移动通信场景中，主要散射体的数量和位置固定。然而，用户移动会给信道带来多普勒频移的影响，并且不同移动速度、不同子载波频率的多普勒频移各不相同。因此，移动信道比静态信道的预测更为复杂，需要充分挖掘 MIMO 信道的物理特性以及用户的运动信息，综合运用多种 PIC 技术来提升信道预测性能。

3.1 物理启发的移动信道预测方案

我们针对信道的序列特性 [6] 对连接结构进行物理启发的设计，并且基于 Neural ODE 将信道空域变化的物理方程直接嵌入无线 AI 模型的架构之中 [9]，实现非完全黑盒的 PIC 方案，其整体流程如图 6 所示，主要分为迭代定位与运动信息提取、基于 Neural ODE 的静态信道预测、角度时延域多普勒补偿这三个步骤。

迭代定位与运动信息提取网络如图 7 所示。利用 [13] 中基于 LSTM 的静态信道 CSI 指纹定位网络，我们可以根据不同时隙的 CSI，建立一个关于时间的位置序列。将用户在该时间段内的运动建模为匀速直线运行，则可以通过最小二乘回归拟合这个位置序列，

$$\boldsymbol{x} = \boldsymbol{v}^i t + \boldsymbol{\sigma} \tag{13}$$

其中 $\boldsymbol{v}^i$ 为第 i 次迭代计算的速度矢量，$\boldsymbol{\sigma}$ 为一常矢量。然后，根据估计的速度矢量，对信道响应执行多普勒消除。将执行多普勒消除后的 CSI 序列再次输入到定位网络中，我们将获得新的位置序列，从而实现对用户定位和速度矢量的迭代更新。该 CSI 指纹定位网络引入了单载波特征提取和多载波累积学习的思想，而迭代算法的设计利用了信道数据的时间序列特性，结合了用户的运动模式先验知识，因此提高了定位和速度估计的准确度和鲁棒性。

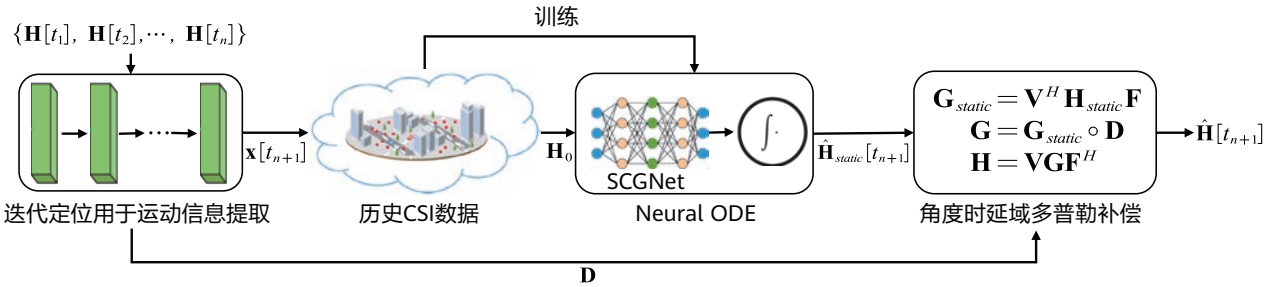


图 6 基于 Neural ODE 的移动信道预测流程

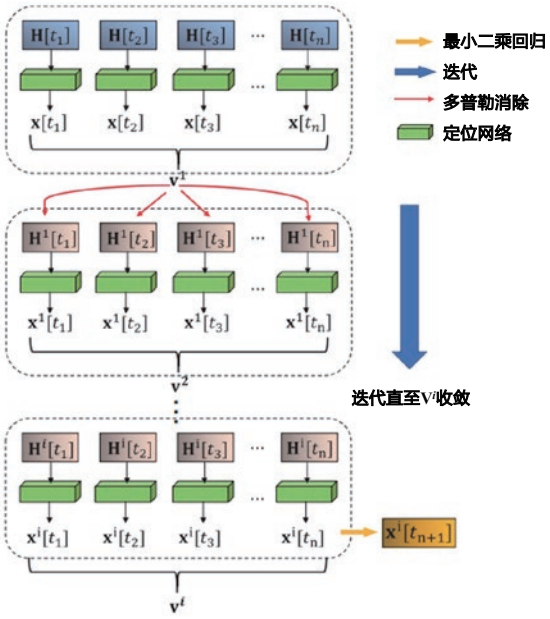


图 7 用户定位和速度矢量的迭代更新

SCGNet 的设计遵循角度-时延域信道变化的物理微分方程，

$$\frac{\partial g_{\tau,\theta}^{real}}{\partial \mathbf{m}} = \left(-\frac{1}{d_p} g_{\tau,\theta}^{real} + \rho g_{\tau,\theta}^{imag}\right)\frac{\partial d_p}{\partial \mathbf{m}}, \tag{14}$$

$$\frac{\partial g_{\tau,\theta}^{imag}}{\partial \mathbf{m}} = \left(-\frac{1}{d_p} g_{\tau,\theta}^{real} - \rho g_{\tau,\theta}^{imag}\right)\frac{\partial d_p}{\partial \mathbf{m}}. \tag{15}$$

其中 $g_{\tau,\theta}^{real}$ 和 $g_{\tau,\theta}^{imag}$ 分别代表角度-时延域信道矩阵中的任意一个表示到达角为 θ、到达时延为 τ 的矩阵元素的实部和虚部。具体而言，SCGNet 由散射环境学习网络和方向嵌入网络两部分组成，如图 8 所示。散射环境学习网络采用全连接结构学习 CSI 矩阵到 $-\frac{1}{d_p}$ 的映射，方向嵌入网络采用方向矢量作为输入以获得方向导数 $\frac{\partial d_p}{\partial \mathbf{m}}$ 。散射环境学习网络和方向嵌入网络的输出通过矩阵运算结合，得到信道矩阵对空间的梯度。因此，SCGNet 是一种在连接结构上进行了物理启发设计的 Neural ODE。完成训练后，可以通过对 SCGNet 的输出进行前向积分，实现从历史 CSI 预测当前用户位置的静态信道。

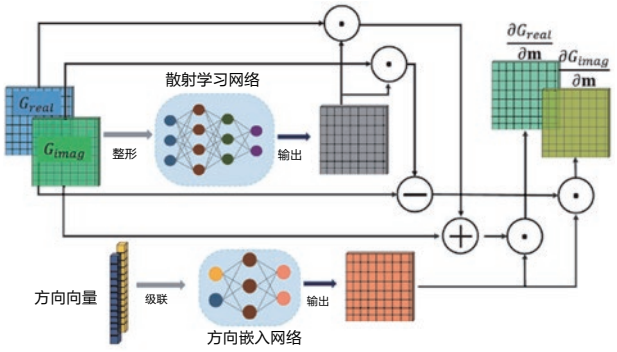


图 8 信道空间梯度网络 SCGNet 的结构示意图

最后，根据第一步估计的用户速度矢量，在角度时延域对静态信道预测结果进行多普勒补偿，就能得到用户移动信道的预测结果。

3.2 仿真结果

我们对浙江大学信电楼周围环境进行 3D 建模并采用 Remcom 公司的 Wireless InSite 软件做射线追踪生成信道数据集进行仿真验证，用于仿真的 3D 模型场景如图 9 所示，其中 1 号楼高 25 m，2 号楼高 35 m，3 号楼和 4 号楼高 8 m，建筑材料设定为水泥。5 号区域为一片树林。用户分布在一个 120 m × 60 m 的区域。电磁波中心频率设定为 3.5 GHz。基站位于 2 号建筑上方 10 m 高处，配备了一个 ULA 天线阵列。OFDM 系统带宽为 100 MHz，最大计算路径数为 25。考虑到实际系统中的信道获取时间间隔较小，可以假设用户在序列信道获取时间内以匀速直线运动方式向任意随机方向移动。

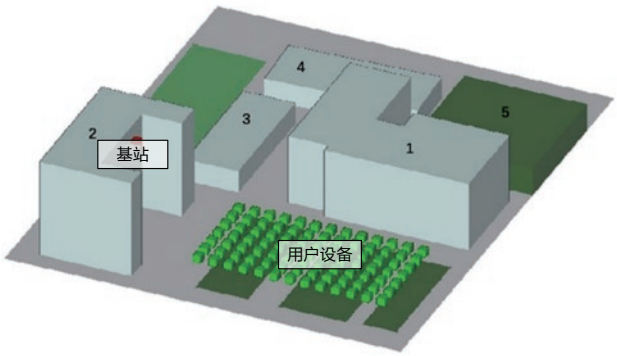


图 9 用于仿真的 3D 模型场景图

为了分别验证物理启发的 Neural ODE 和 SCGNet 网络的作用，我们将所提方案与静态数据库方法以及 Neural ODE 结合 MLP 的方法在不同历史数据采样密度下进行比较，结果如图 10 所示。与传统方法相比，采用基于 Neural ODE 的学

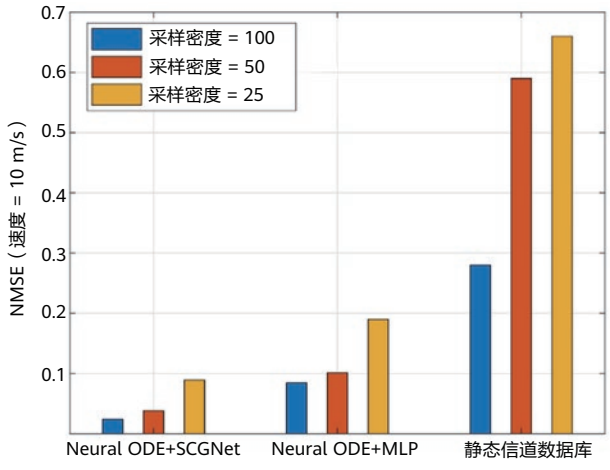


图 10 不同采样密度下三种方法的预测 NMSE

习结构可以带来出色的性能提升。在对比方案中，当信道采样密度从 50 转换到 25 时，预测精度会有大幅下降。然而，在基于 Neural ODE 的方法中，即使在较低的采样密度下，预测准确度也不会有明显下降。另一方面，将网络从普通的 MLP 切换为 SCGNet，在所有采样密度设定下，预测精度均获得明显提高。这是因为所提出的网络不是直接学习 CSI 与其空间梯度的映射，而是将信道物理变化的先验信息融合入网络设计中，极大降低了网络的学习难度。

为了更全面地比较所提方案和 LSTM 信道预测网络的差异，对比方案中 LSTM 网络采用了三种不同的长度。LSTM 的训练数据集的序列间隔为 1 ms，所有网络均输出角度-时延域信道矩阵。采样密度为 50 时所提方案和对比方案在不同的 UE 速度下的预测归一化均方差（Normalized Mean Square Error，NMSE）比较如图 11 所示。得益于综合运用多种 PIC 技术，所提方案在任何情况下都能达到远超对比方案的预测准确度。此外，从实验结果中可以看出，基于序列的网络结构的预测性能对 UE 的移动速度十分敏感。当 UE 速度增加时，这些网络的预测性能会大大降低。这背后的主要原因是，这类网络的性能高度依赖于序列数据之间的相关性。随着 UE 移动速度的增加，序列采样点之间的空间间隔也相应增加，序列之间的空间相关性降低。与之相反的是，所提方案即使在 UE 快速移动的情况下也能保持较高的预测精度。

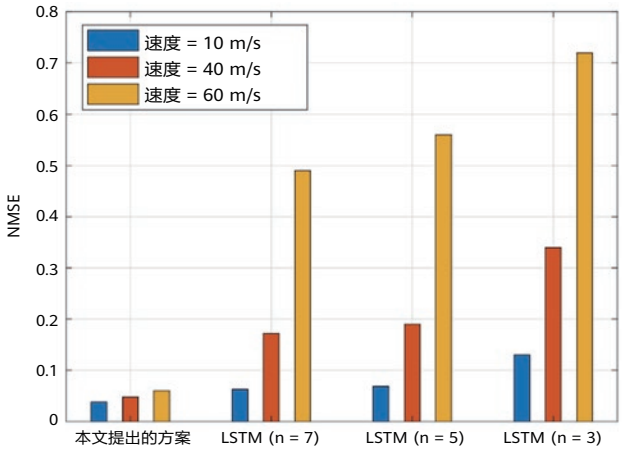


图 11 采样密度为 50 时不同 UE 速度下各方案的预测 NMSE

4 挑战和未来研究方向

现有的物理启发的方法显著改善了无线 AI 技术的性能和可用性，但这一技术仍有许多进一步发展的空间和方向。

首先，尽管物理启发的方法与主要依赖数据驱动的方法在学习方式有一些差异，但他们都仍然是典型的深度神经网络，在总体性的部署范式和工作模式上仍基本相同。因此，一些已经被证实可以改善深度神经网络在无线系统中部署质量的技术，比如量化神经网络、自适应推理等方法，也同样可以用于和现有的 PIC 方案进行融合，以减少 PIC 技术的推理开销，降低 AI 的使用成本。同时，一些更先进的 AI 架构，如 Transformer、MLP-Mixer 等方法，已经被证明在经过合理的建模和变式后可以很好的刻画无线信号的物理属性 [10, 11]，进一步改善网络对物理先验的利用效率。

其次，物理启发可以不局限于改善已有的典型无线子任务模块。借助深度学习出色的信息融合能力，下一步的研究中也可以思考直接从无线系统的基本需求和总体性功能出发，不再局限于替换已有的功能模块，而是追求直接围绕端到端的通信催生新的功能体系。物理启发将是这一演进思路的指导原则之一。因为如果要构建更加简约且更加高效的无线系统，就必须更贴近电磁传输的物理本质，更充分地利用无线信息的物理特性。某种意义上说，这使得物理启发从具体的结构设计上升到整体的架构设计，也将进一步释放 AI 方法对无线网络演进的潜力。

另外，由于现有的 PIC 技术的研究重点集中在如何恰当的引入物理机制的启发，在总体框架上沿用了无线 AI 技术中常见的任务且场景特定的智能范式，这种窄泛化使得智能的可用性、可拓展性以及在困难用例上的性能仍然面临挑战。为了解决现有无线 AI 在框架设计上带来的泛化性限制，研究者们提出了以集成多任务、统一多场景以及一体化调度为目标的无线大模型 [14] 愿景。PIC 技术势必会与无线大模型产生美妙的反应。一方面，由于物理法则自身具有的普适性，PIC 能够显著促进跨任务、跨场景的无线智能的建立；另一方面，大模型更大的学习容量和更复杂的推理机制也能够进一步改善 PIC 技术的性能和可用性。

同时，PIC 技术在元宇宙和数字孪生等新兴技术领域也有显著的应用潜力。通过在深度学习过程中融合无线问题的先验物理规律，PIC 技术能够提高通信效率、预测准确性、定位精度和系统融合度，为创建更加丰富和真实的数字体验提供坚实的技术基础。

5 结语

在本文中，我们介绍了 DNN 在 6G 网络中的重要角色，同时我们也指出了 DNN 依赖于大量训练数据而带来的泛化性和可靠性问题。通过引入 PIC 技术，我们能够将物理规律和数据驱动的 AI 模型相结合，从而提升无线 AI 的性能。我们详细探讨了 PIC 的技术框架，包括优化神经网络的连接结构、激活函数、损失函数，以及设计非完全黑盒模型。同时，我们展示了 PIC 技术在 CSI 压缩反馈、信道预测和用户定位等无线任务上的应用案例，其展现出比传统无线 AI 方案更优越的性能，这证明了 PIC 在实际通信系统中的应用潜力。尽管 PIC 技术为无线系统带来了显著的改进，但其在模型优化、系统复杂性等方面仍存在挑战。未来的研究应集中在对 PIC 的实际部署进行优化，并基于 PIC 技术对当前无线系统的功能模块进行重新设计，以进一步释放 PIC 的潜力。

参考文献

[1] I. Recommendation, "Framework and overall objectives of the future development of IMT for 2030 and beyond," *Int. Telecommun. Union ITU Recomm. ITU-R*, 2023.

[2] C.-K. Wen, W.-T. Shih, and S. Jin, "Deep learning for massive MIMO CSI feedback," *IEEE Wirel. Commun. Lett.*, vol. 7, no. 5, pp. 748–751, Oct. 2018, doi: 10.1109/LWC.2018.2818160.

[3] J. Vieira, E. Leitinger, M. Sarajlic, X. Li, and F. Tufvesson, "Deep convolutional neural networks for massive MIMO fingerprint-based positioning," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct. 2017, pp. 1–6. doi: 10.1109/PIMRC.2017.8292280.

[4] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 320–332, 2020, doi: 10.1109/OJCOMS.2020.2982513.

[5] A. Zappone, M. Di Renzo, and M. Debbah, "Wireless networks design in the era of deep learning: Model-based, AI-based, or both?," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019, doi: 10.1109/TCOMM.2019.2924010.

[6] Z. Chen, Z. Zhang, Z. Xiao, Z. Yang, and K.-K. Wong, "Viewing channel as sequence rather than image: A 2-D Seq2Seq approach for efficient MIMO-OFDM CSI feedback," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 11, pp. 7393–7407, Nov. 2023, doi: 10.1109/TWC.2023.3250422.

[7] Z. Xiao, Z. Zhang, C. Huang, X. Chen, C. Zhong, and M. Debbah, "C-GRBFnet: A physics-inspired generative deep neural network for channel representation and prediction," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2282–2299, Aug. 2022, doi: 10.1109/JSAC.2022.3180800.

[8] Z. Chen, Z. Zhang, Z. Xiao, Z. Yang, and R. Jin, "Deep learning based multi-user positioning in wireless FDMA cellular networks," *IEEE J. Sel. Areas Commun.*, p. 1, 2023, doi: 10.1109/JSAC.2023.3322799.

[9] Z. Xiao, Z. Zhang, Z. Chen, Z. Yang, C. Huang, and X. Chen, "From data-driven learning to physics-inspired inferring: A novel mobile MIMO channel prediction scheme based on neural ODE," *IEEE Trans. Wirel. Commun.*, p. 1, 2023, doi: 10.1109/TWC.2023.3338419.

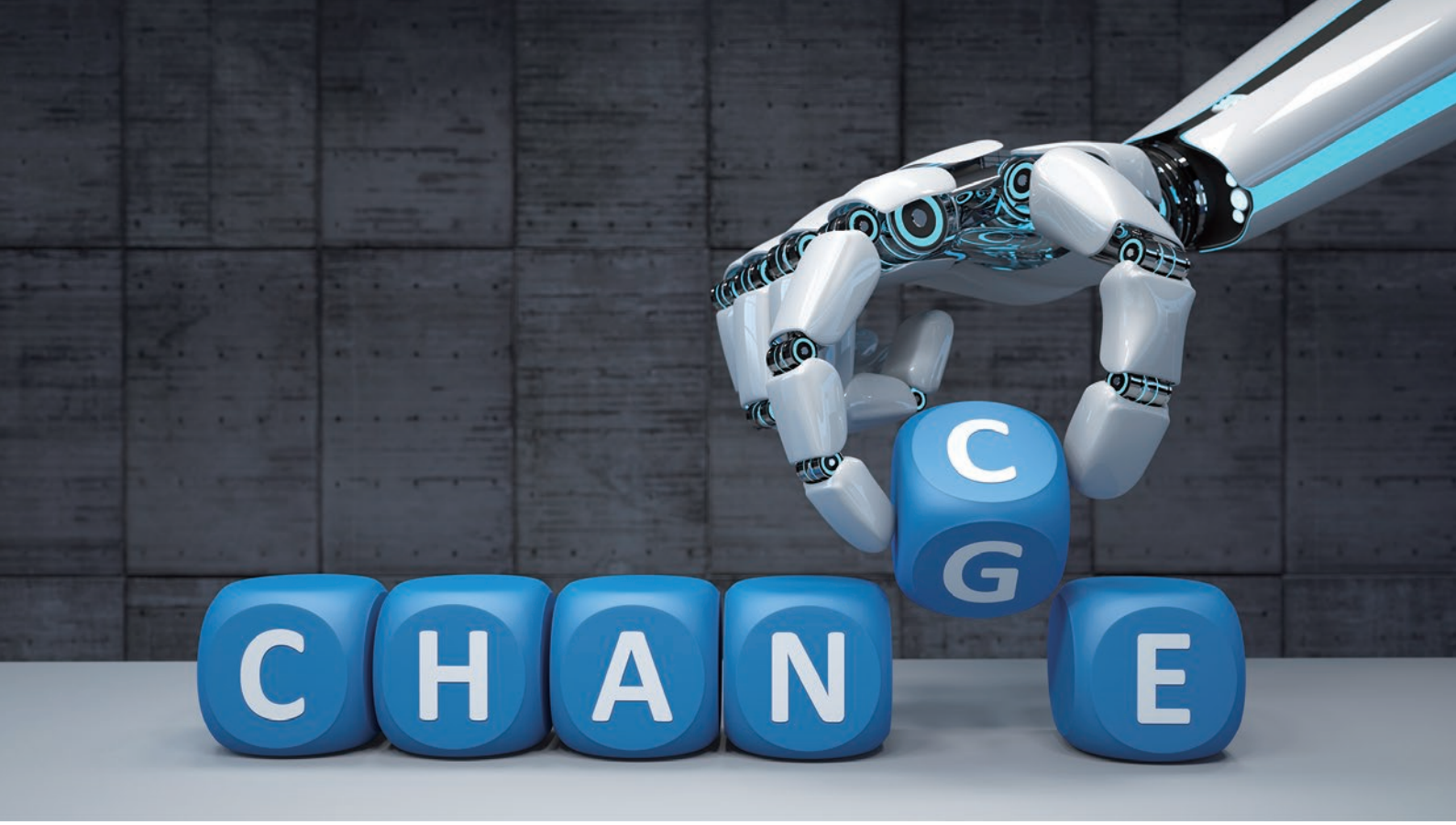
[10] Z. Chen, Z. Zhang, Z. Yang, and L. Liu, "Channel mapping based on interleaved learning with complex-domain MLP-mixer," *IEEE Wirel. Commun. Lett.*, p. 1, 2024, doi: 10.1109/LWC.2024.3370303.

[11] Z. Chen, Z. Zhang, Z. Yang, C. Huang, and M. Debbah, "Channel deduction: A new learning framework to acquire channel from outdated samples and coarse estimate," Mar. 28, 2024, *arXiv*: arXiv:2403.19409. doi: 10.48550/arXiv.2403.19409.

[12] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 7462–7473. Accessed: Jul. 03, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/53c04118df112c13a8c34b38343b9c10-Abstract.html

[13] Z. Chen, Z. Zhang, Z. Xiao, C. Zhang, and Z. Yang, "CSI of each subcarrier is a fingerprint: Multi-carrier cumulative learning based positioning in massive MIMO systems," in *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2023, pp. 1–7. doi: 10.1109/PIMRC56721.2023.10294059.

[14] Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE Wirel. Commun.*, pp. 1–9, 2024, doi: 10.1109/MWC.015.2300404.



AI 大模型赋能机器人及其为 6G 带来的机遇

Massimiliano Maule，Anh Vu Vu，曹瀚文，付廷中，Mohamed Gharba，
Daniel Gordon，Joseph Eichinger，张申飞，吴艺群，安雪莉，卢磊
先进无线技术实验室

摘要

全球范围内，AI 在机器人研究中的重要性日益凸显。尽管深度学习提高了训练的有效性，但 AI 的泛化能力仍有待提升。大模型（Foundation Model）具有海量参数并用大量数据进行了预训练，为这一问题提供了解决之道。机器人应用了大模型之后，可以自主理解自然语言指令并执行相关任务、动态分解复杂任务，并基于实时反馈调整动作，将人为干预程度降到最低。本文就大模型在机器人领域的应用，分析当前的业界研究，同时探讨 6G 技术在机器人应用中的潜力，并介绍一款采用 AI 和大模型技术的 6G 机器人系统原型。

关键词

机器人，大模型，6G，ISAC，AIaaS，3GPP

1 引言

当今世界致力于发展机器人技术，将 AI 融入机器人的重要性日益凸显。美国国家机器人计划（National Robotics Initiative）于 2024 年发布了一份技术规划报告——《美国机器人技术路线图：机器人技术创造美好未来》（A Roadmap for US Robotics: Robotics for a Better Tomorrow）[1]，阐述了 AI 的重要地位，并介绍了机器学习、通用人工智能（Artificial General Intelligence）、普遍自动化（Pervasive Automation）以及 AI 与机器人技术融合等课题的研究进展。此外，报告还着重强调了定制化 AI、AI 伦理、AI 辅助科学发现，以及它们对经济、劳动力和国家安全的影响。

欧盟发布了《人工智能、数据和机器人合作伙伴关系的联合战略研究、创新和部署议程》（Strategic Research Innovation and Deployment Agenda for AI, Data, and Robotics Partnership）[2]，强调 AI 和机器人技术的发展应以人为中心并且安全可信。《议程》的核心在于促进产业、学术和政策制定者之间的合作，推动相关技术的研究、开发与部署。《议程》还鼓励投资并积极应对关键挑战，以确立欧洲在 AI 和机器人领域的全球领导地位，推动经济、社会和环境朝着有利于发扬欧洲价值主张、维护欧洲权利的方向发展。

中国在《“十四五”机器人产业发展规划》强调，要推动人工智能、5G、大数据、云计算等技术的融合，提升机器人智能化、网络化水平。《规划》同时指出，要强化机器人系统的功能安全、网络安全和数据安全，不断提升国家的技术能力，推动行业应用。

在传统的 AI 机器人系统中，感知能力基于部署在受控环境中的深度学习方法实现。这样的 AI 机器人虽然可以有效学习各种技能，但训练耗时长，需要对每个不同的任务进行大量的工程处理，缺乏分布偏移和泛化能力。对于单一任务，这种方法看起来是不错的选择。然而，在实际的多任务并行实验中，该方法的学习成本和工作量呈指数增长，成为机器人研究领域的又一大难题。

构建可泛化的机器人系统面临诸多困难。但与此同时，一个全新研究领域的出现，有望指明问题的解决之道，帮助改进机器人系统，这个领域即大模型（Foundation Model）。大模型是一种大型 AI 模型，通过适配特定应用，可以为多类型下游任务提供通用框架。大模型的训练基于互联网海量数据，因此大模型拥有卓越的泛化能力，并拓展了迁移学习（Transfer Learning）和模型扩展（Scaling）的概念 [4]。

借助大模型，机器人可以自主理解高级自然语言指令并执行相关任务，动态分解复杂任务，基于实时反馈调整动作，从而最大程度减少人为干预。此外，利用摄像头、激光雷达（LiDAR）、麦克风等常见传感器所采集的多模态数据，机器人可以理解环境语义，提升态势感知能力。

传统的机器人往往只能僵化地执行预定义操作，受限于特定用途的模型。大模型让机器人既能理解环境，又能动态、智能执行各类任务，机器人因此更自主、更灵活、更高效。

本文就大模型在机器人领域应用的，分析学术界和产业界当前研究，以及未来的发展方向。本文还就 6G 技术对机器人的影响，重点介绍未来应用、与机器人与 AI 大模型的融合，以及相关组网要求。本文结构如下：第 2 节提供大模型在机器人应用的前沿分析；第 3 节简述主要机构的相关标准化工作；第 4 节概述 6G 和 AI 应用于机器人所带来的市场机会和研究前景；第 5 节介绍我们提出的 6G 机器人原型；最后第 6 节进行总结评述，并探讨未来的研究方向。

2 大模型在机器人领域应用的前沿成果

本节介绍机器人领域所应用的大模型的种类、作用和功能。相关术语参考以机器人和机器人设备 ISO 8373:2021 标准为参照——该标准确保了跨行业、跨学术领域以及跨区域交流机器人技术时术语表述清晰、一致。

2.1 大模型赋能机器人

大模型可赋予机器人以下能力：

- **综合知识库**：大模型提供广泛的多领域知识，让机器人可以理解并执行各类任务。具备这些知识，机器人可以轻松处理不同领域的复杂操作，无需针对特定任务进行繁复的重新编程。
- **自然语言理解**：大模型具备强大的自然语言处理能力，让机器人能够理解人类语言并基于自然语言交互。用户可以使用自然语言给机器人下达指令、接收反馈，任务指令和交流更简单。
- **多模态势感知**：大模型赋能的机器人具备多模态势感知能力，可借助 RGB 摄像头、LiDAR、麦克风等各类传感器理解周围环境的语义，还可以理解物体间逻辑与几何关系、评估态势、解释事件、预测未来事件。
- **零样本与小样本学习**：大模型在零样本和小样本学习方面表现出色，使机器人无需大量训练，就可以执行特定任务。机器人灵活度更高、适应性更强，在面对新任务、新环境时，也无需重新训练。

2.2 机器人领域的大模型宏观分类

大模型为机器人领域带来更多可能。具体来说，大模型中的一些预训练模型可提升机器人的感知、预测、规划和控制等方面性能 [6]。

- **大语言模型（Large Language Model，LLM）**[7]：LLM 让机器人可以理解自然语言指令，乃至用自然语言做出响应。
- **视觉 Transformer（Vision Transformer，ViT）或多模态 Transformer** [8]：机器人要理解摄像头、LiDAR 等传感器从环境中采集的视觉数据，这类模型发挥着至关重要的作用。
- **具身多模态语言模型** [9]：这是一类功能更广泛的模型，可同时具备 LLM、ViT 的能力，使机器人不仅可以理解自然语言，还可以理解指令的视觉上下文。
- **视觉生成模型（Visual Generative Model，VGM）**：从扩散模型背后的演进原理来讲 [10]，VGM 利用海量数据集进行训练，可以帮助构建真实场景，供机器人开展任务仿真演练。VGM 提供了丰富多样的训练数据，可提升机器人的感知能力、微调动作 [11]。

从以上几点可以看出，大模型的应用并非将现有的视觉和语言模型简单组合，而是可以帮助机器人领域开发更具针对性的模型。

2.3 机器人大模型：意图识别和视觉推理

近年来，基于 Transformer 架构的机器人 AI 因其强大的意图识别和视觉推理能力 [12] 而备受关注。这种架构以语言嵌入和观察为输入，输出预测的动作。在基于语言条件的机器人操控（Language-conditioned Robotic Manupulation，LcRM）中，应用视觉 – 语言 – 动作模型（Vision-Language-Action Model），可以获得长时程鲁棒性和可泛化的策略，缩小了机器人物理和 AI 之间的差距，具体体现在两方面。

- **高阶规划**：将复杂的语言指令转换并分解为基本动作原语序列，交由低阶控制器执行。规划与推理的高阶策略采用 PaLM-E 模型 [9]，该模型结合了 PaLM [13] 和 ViT [14]，参数规模达到 562B。
- **端到端学习**：通过训练得到 LLM，可以基于指令和观察直接生成动作。RT-1 [15] 和 RT-2 [16] 等多任务模型可以将机器人的输入分词，并输出动作，从而实现运行时的高效推理，有望让机器人实时控制成为现实。类似地，Octo [17] 采用基于 Transformer 的扩散方法训练和微调通用机器人策略（Generalist Robot Policy）。Octo 无需额外设置即可支持多 RGB 摄像头输入和多臂机器人，并且可以接受语言或目标图像形式的指令。此外，Octo 还在其 Transformer 主干架构中使用了模块化的注意力结构，即便目标领域数据集和可用计算资源很少，也可针对新的感知输入、动作空间和形态进行有效微调。

2.4 机器人与 AI 大模型联合仿真平台

目前已开发的一些框架，有的可用于带 AI 规划能力的机器人仿真，有的可用于算法控制，有的则兼具两种用途。我们认为有两类框架具有应用潜力，并将重点分析。英伟达 Isaac 实验室 [18] 的平台也具备应用潜力，但由于该平台需要专门的商业许可，因此未予考虑。

- **RoboCasa** [19] 是一种训练机器人执行日常任务的仿真框架，该框架提供了一套方法，可利用本体感知机器人数据（如关节编码读数）和图像（采集自机器人自带的或环境中的摄像头）来训练基于 Transformer 的模型。
- **MuJoCo（Multi-Joint dynamics with Contact，接触型多关节动力学）**[20] 是一款用于物理系统仿真的物理引擎，尤其适用于机器人仿真。MuJoCo 仿真效果逼真，可用于训练面向各种机器人任务的大模型。训练出的大模型可通过与虚拟环境交互、操控虚拟对象以及接收动作反馈来进行学习。训练数据可以传输给真实的机器人，使机器人能够在真实世界中执行类似任务。
- **HABITAT** [21] 是一款专用于机器人助手等具身 AI 智能体训练的高性能 3D 仿真环境。HABITAT 可用于各类常用机器人传感器（如 RGB-D 摄像头）的仿真，为大模型提供感知和决策所需的多样化感知信息。

3 6G 和 AI 赋能机器人的应用场景

在电信行业，科研和标准化组织一直在探索在探索如何将移动通信网络应用于机器人。3GPP SA1 基于对服务机器人的研究 [22]，确定了八种应用场景，包括实时协同安全防护、基于多机器人机载多模态传感器的智能通信数据采集与融合、面向矿山作业和交付的自主机器人与遥控机器人等。此外，这份研究报告中还讨论了触觉和多模态通信、通感一体化（Integrated Sensing and Communication，ISAC）、元宇宙、高级通信等技术。

one6G 协会的任务是发展、测试和推广下一代蜂窝和无线通信解决方案。该协会认为，机器人应用将融入多个行业领域和社会部门。此外，协会还公开发布了一系列 6G 和机器人相关的白皮书[1]，从通信、AI/机器学习、ISAC 等方面深入讨论了 6G 如何赋能机器人。同时，该协会还提出了由 6G 赋能的机器人的若干场景，如协作机器人、救灾、运动规划、工业机器人和医疗救助。

[1] https://one6g.org/resources/publications/

Hexa-X 和 Hexa-X-II [23] 是欧盟资助的旗舰级 6G 研究项目。这两个项目讨论、分析了各种 6G 应用场景和需求，重点关注了自主机器人。这类机器人可以互相通信，也可以与其他机器和周边人类通信，以执行单一任务或合作达成共同目标。协同移动机器人（Cooperating Mobile Robot，CMR）就是其中一种。

4 6G 机会点

机器人控制通常分为四级：**任务控制**、**动作控制**、**原语控制**和**伺服控制** [24, 25]。机器人融合 6G 的 AI 与传感能力后，有望获得超越传统任务控制的智能水平，我们把这种更高的能力称为“**元控制**”。元控制机器人可以完全自主地识别问题、定义任务，并根据角色、任务和规则的元定义适应动态环境，还拥有实时态势感知能力。图 1 中展示了不同级别、ISAC 功能和原生 AI 基础设施间的互操作关系。

根据我们的设想，未来智能机器人的控制可以划分为以下等级：

- **元控制**：该级别的机器人可基于角色、任务和规则的元定义自主识别问题、定义任务、适应动态环境，并具备实时态势感知能力。
- **任务控制**：该级别定义了机器人的总体目标和任务，包括高阶规划、决策和任务分解，例如清洁厨房地板或者来一杯低热量的气泡饮料。
- **动作控制**：该级别的机器人可将任务指令转换为具体的动作序列，包括轨迹规划、路径生成，例如规划一条避开孩子玩具、从客厅到厨房的路线。
- **原语控制**：该级别可以直接控制机器人致动器使机器人遵循规划轨迹、生成关节姿态、速度和力量的相关指令，例如控制机械臂沿路径精确移动并拾取物体。
- **伺服控制**：该级别为最低级别，重点是基于反馈回路精准控制致动器，确保指令可以高度准确、稳定地执行。

6G 愿景及相关初期研究和标准化工作中提及 ISAC、Network for AI（Net4AI）特性。这两大特性有望为未来 AI 大模型赋能的机器人注入重要的能力。

4.1 原生 AIaaS 提供 AI 模型和计算设施

6G 基于 Net4AI 提供 AI 即服务（AI as a Service，AIaaS）能力，将大模型和其他特定 AI 模型直接整合到网络基础设施。AIaaS 具备以下关键优势：

- **时延低**：AI 模型可显著降低 6G 网络时延。在无线接入网和核心网内就近处理数据，最大程度避免将数据传输到外部服务器处理，缩短了响应时间。
- **数据源丰富**：6G 框架内的 AI 模型可以访问来自无线接入网、核心网以及 ISAC 的丰富数据。海量的数据源有利于实现精准且基于情景感知的 AI 决策，从而增强 AI 应用性能。
- **数据集成增强**：6G 网络中感知和通信的无缝集成，允许 AI 模型利用多样的数据源，实现更稳健、全面的分析，支持实时环境监测、自适应机器人控制和动态资源管理等高级应用。

与传统多接入边缘计算（Multi-Edge Computing）相比，6G AIaaS 时延更低、带宽效率更高。这是因为 6G AIaaS 将 AI 能力整合到了网络基础设施中，边缘服务器和蜂窝系统间不再需要额外的数据路由。此外，6G 原生 AI 模型可以在全网范围内访问更广泛的数据（包括 ISAC 数据），AI 可以处理更全面的信息，业务交付更优质。6G 框架还支持在不同无线接入网和核心网实体中动态分配 AI 资源，AI 业务部署扩展性、灵活度更高。与机器人本地的 AI 系统相比，6G 原生 AI 优势显著。具体来说，网络中的 AIaaS 计算性能通常优于本地 AI，因此系统响应更快。将密集的 AI 计算迁移到网络，避免了本地处理所引入的功耗和散热问题，可延长机器人的寿命，也可以降低成本。此外，由于大量数据来自于网络，6G 原生 AI 模型可以更准确地理解上下文并作

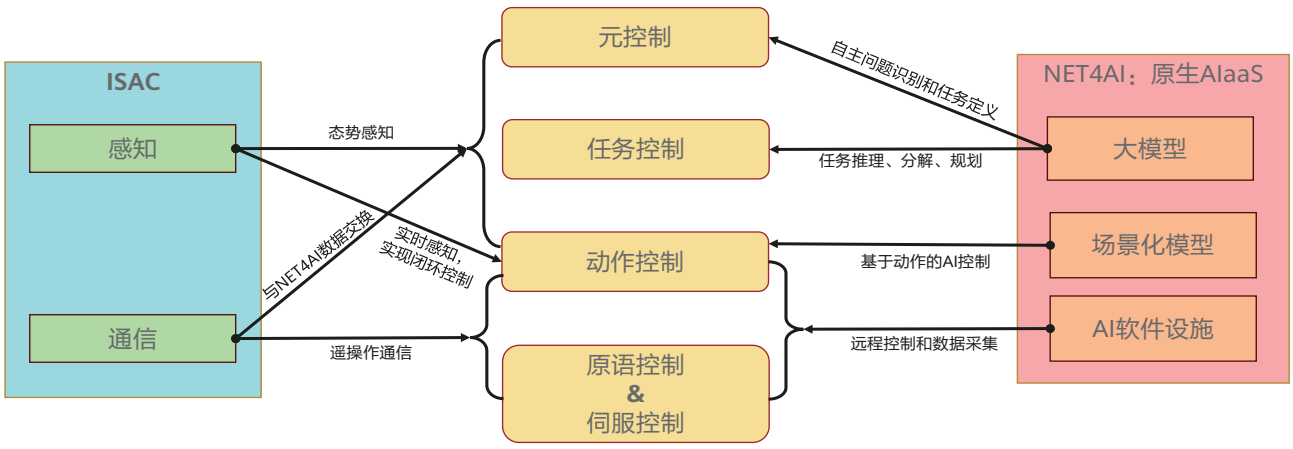


图 1 6G 能力与机器人控制等级的适用关系

出决策。综上，AIaaS 有望使网络的“大脑”灵活分布于本地和网络节点，以满足安全性等要求 [27]。

4.2 ISAC 用于机器人综合态势感知

3GPP 已启动 ISAC 的相关研究，并认为 ISAC 有望为机器人等诸多应用带来全新的面貌。SA1 基于 ISAC 的研究成果（FS_Sensing），在 TR22.837 [28] 中详述了 32 项 ISAC 应用场景，并在 TR22.137 [29] 定义了相关服务要求。这些研究提出的 ISAC 方案在 3GPP 无线网络感知的基础上，综合采用了非 3GPP 感知——如摄像头、LiDAR 等传感器。

未来移动网络的 ISAC 能力将从以下几个方面提升机器人应用：

- **将感知、通信、AI 整合至统一的标准化网络架构中**：整合了感知、通信、AI 大模型的统一 6G 网络架构，将为未来智能机器人带来革命性的变化，让机器人可以快速访问全面、实时的数据，实时决策能力、态势感知能力因此大大增强。
- **基于感知网络实现全面态势感知**：整合了感知、通信、AI 大模型的统一 6G 网络架构，为机器人提供网络化的感知能力，从而实现全面态势感知。除了机器人自身的传感器，ISAC 支持机器人访问网络中各感知节点（如其他机器人和环境传感器）上的丰富数据。
- **感知定位一体化**：移动机器人寻找物体、进行导航都需要定位能力。ISAC 可以融合移动网络的被动感知和主动定位，提高定位精度。
- **感知数字孪生（Digital Twin）构建**：为机器人构建数字孪生需要实时且准确的感知数据。未来，ISAC 或将创建精确、动态的虚拟副本，实现有效的数字孪生，提升多机器人协作水平。

4.3 6G 通信赋能未来机器人

6G 的脚步越来越近，其超高的可靠性、超低的时延、强大的业务质量（Quality of Service，QoS）保障，以及与机器人软件和协议的互联，将让未来的机器人变得更强大。

- **超高可靠、超低时延通信（Hyper Reliable and Low-Latency Communication，HRLLC）**：在机器人集中控制场景下，HRLLC 对工业应用来说必不可少。6G 提供了超高可靠、高稳定、低抖动的通信信道，确保机器人系统可以更流畅地运行与同步，这对精度和可靠性要求高的任务至关重要。
- **先进的 QoS 框架**：6G 引入了先进的 QoS 框架，可为 AI 大模型和专业机器人应用按需动态分配网络资源。基于强大的数据吞吐能力，6G 可高效传输 AI 训练数据、传感器数据和实时分析数据，支撑复杂的决策过程和算法学习。
- **互联新协议**：6G 支持与机器人软件和通信协议无缝互联，如数据分布式服务（Data Distribution Service，DDS）[30]、开放平台通信统一架构（Open Platform Communication Unified Architecture，OPC UA）[31]、消息队列遥测传输协议（Message Queuing Telemetry Transport，MQTT）[32]、Zenoh [33]。这样，现有系统无需重新设计，机器人就可以使用 6G 强大的能力。
- **实时闭环遥操作与训练**：6G 可以实现人类或 AI 对机器人的实时闭环遥操作，这对于解决未知的复杂任务、训练 AI 模型通过模仿学习获得新技能至关重要。利用 6G 强大的通信基础设施，运营商可以实时远程控制机器人，通过实际训练，缩短 AI 学习和适配所需的时间。
- **新商机**：AI 能力加上机器人和网络的 6G 感知，将为网络所有者和机器人服务提供商开拓新的商机。实时机器人操作需要集成感知、AI 和控制功能，还要求低延迟、高数据吞吐，以确保操作顺畅、性能高效。同时，还需要基于 AIaaS 智能体的部署情况，及时整合不同数据源的感知数据。此外，机器人运营商和供应商可能还需要融合移动网络所提供的资源密集型服务，以确保合同顺利履行、网络环境可信、运营连续可靠。

5 概念验证中的 6G 大模型机器人 MELISAC

本节介绍我们研发的 MELISAC（Machine Learning Integrated Sensing and Communication，机器学习通感一体化）复合型机器人，该机器人目前处于概念验证阶段。MELISAC 集成了多项先进技术，包括机器人智能控制、机器人在线训练，以及 ISAC。

5.1 硬件配置

MELISAC 是一款双臂复合机器人，由两个工业铰接式协作机器人（Collaborative Robot，Cobot）UR5e[2] 和自动导引车（Automated Guided Vehicle，AGV）组成。UR5e 安装在 AGV 顶部的铝框架上，可实现机器人自主导航和精确物体操控。MELISAC 配备了一对 MiaHand[3] 人型机械手作为末端执行器，能够以类似人类双手的方式执行任务，并且特别适合通过基于人类任务执行演示来训练 AI 模型控制机器人。

[2] https://www.universal-robots.com/products/ur5-robot
[3] https://www.mia-hand.com

此外，MELISAC 还配备了支持 ISAC 的亚太赫兹无线系统，无线系统的天线安装在机体框架上，也可用作末端执行器。它采用一台机载计算机来完成动作控制和信号处理相关的本地计算。

5.2 软件架构

如图 2 所示，在我们的方案中，传感器数据处理和运动规划由本地计算机负责，而计算密集型任务（例如 AI 推理）则由边缘服务器执行。

- **协作机械臂和 AMR 控制器**：机器人制造商提供的原生控制器。这些控制器提供一系列 API，用于执行低阶机器人功能，如紧急停止、障碍检测，以及正向与反向运动学。
- **适配 API**：作为适配层，将低阶控制指令抽象出来，用于高阶控制器，将基于大模型控制功能与硬件解耦。
- **人机接口（Human-Machine Interface，HMI）**：允许人类与机器人通过语音、动作等形式交互。
- **射频感知**：用于支撑无线 ISAC 实现的射频系统。在 RGB-D 摄像头和麦克风的基础上，无线 ISAC 提供了一个额外的感知层。

为满足计算和内存需求，目前最先进的大模型需要部署在位于边缘云的高性能服务器。在服务器上，每个大模型被加载到一个 AI 智能体，这一智能体会融合该大模型所需的软件栈。不同 AI 智能体之间通过部署在边缘云的文本多代理系统进行交互，机器人的本地计算机则通过 ROS2 协议与其他组件以及位于边缘云的 AI 智能体进行通信。

- **对话智能体**：基于 LLM 的 AI 智能体，拥有海量词汇和通用知识，支持与人类进行各种主题的对话。
- **视觉智能体**：视觉语言大模型的智能体，专门用于提取视频和图像输入中的语义，还可以对相关的对象进行分类和定位。
- **机器人智能体**：机器人大模型的智能体，基于对话智能体（用户请求）和视觉智能体（环境上下文）提供的输入，进行机器人动作的高阶规划。
- **语音智能体**：实时对语音和文本相互转换。

非结构化环境中执行陌生任务时，机器人大模型可能经常遇到困难。此时人类操作员可以介入，为机器人进行任务演示。遥操作员可以通过网络控制 MELISAC，并基于遥操作数据来训练模型。这种有人干预的在线训练，为预训练大模型增加了一个适配层，让训练可以在云端持续进行。

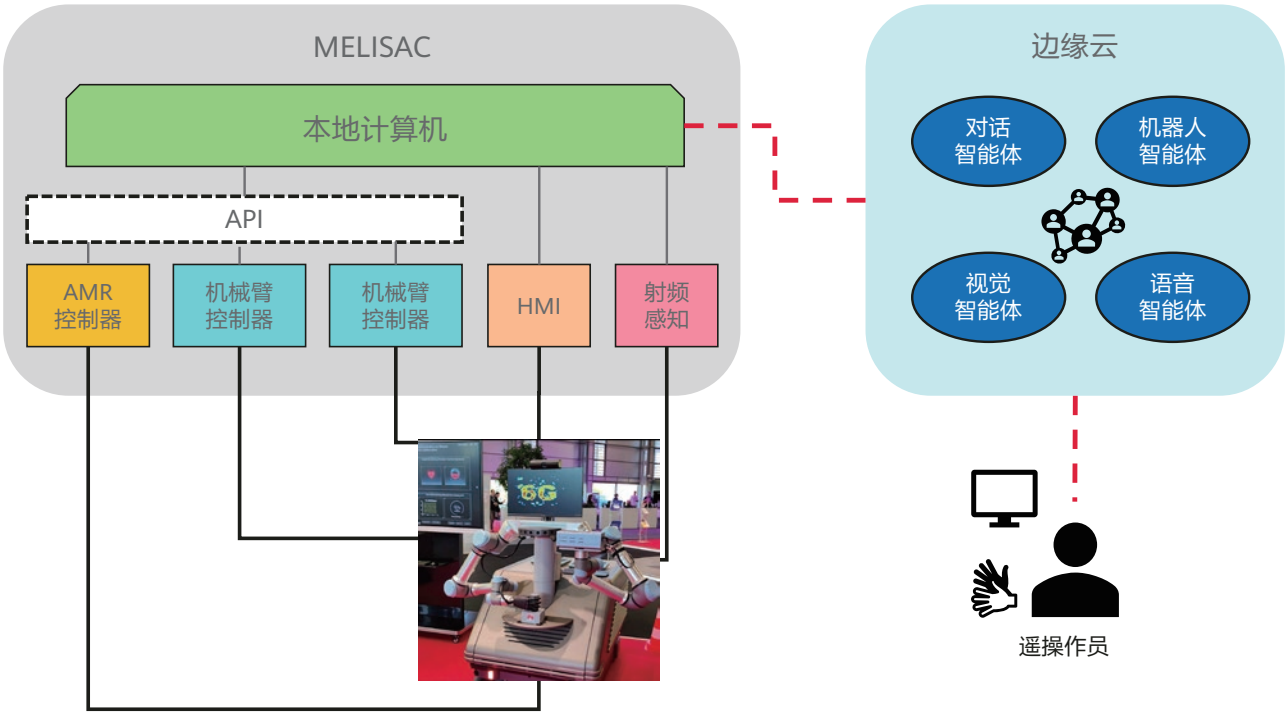


图 2 2023 年汉诺威工业博览会展出的 MELISAC 及其软件架构

5.3 技术探讨

端到端模型与模型链：对于大模型控制的机器人来说，使用单一的端到端模型还是多个模型组成的流水线来处理多模态输入，是一个重要问题。多模态数据一起训练，通过一次推理就可实现实时控制，因此 RT-1、RT-2 和 Octo 所采用的单一模型路线通常具有更好的泛化性能。模型流水线路线虽更灵活、透明度更高，也支持定制，但不可避免地会带来推理时间长、集成复杂的问题。不过，Tipsflow[4]、DSPy[5] 等现有框架可以帮助应对这些挑战。因此，具体方式的选择，还是要根据数据可用性和硬件适配性来决定。对于涉及机密数据的特定领域任务，采用由语言模型和动作模型组成的流水线视觉模型可能更合适。对于一般任务，使用由大规模互联网数据集训练的端到端模型则更为合适。

与机器人制造商 API 的集成：目前，制造商为同步定位与制图（Simultaneous Localization and Mapping，SLAM）和运动等功能提供了高阶 API 控制栈，但低阶运动控制仍受制于安全合规问题。这是因为，将 AI 集成至机器人，需要允许 AI 访问传感器和致动器。鉴于由大模型完全负责低阶控制并不可行，因此需要探索如何将大模型功能集成至现有系统。这时，检索增强生成（Retrieval-Augmented Generation，RAG）能力可以帮助大模型通过标准的低阶 API 文档学习如何控制机器人。一种较为合理的方法是，在高阶功能（可能是基于大模型实现的）和低阶 API 之间定义通用接口，这样可以在实现功能的同时满足安全性要求。实现这一方法则需要机器人制造商和大模型开发者展开合作。其中，对相关接口的标准化是一种有益的途径（虽然这种标准化并非必需）。

6 结语和评述

机器人大模型尽管在掌握基本物体和动作方面表现亮眼，但面对复杂任务时仍存在困难。机器人大模型无法细致入微地理解现实世界的物理原理，限制了它们执行高精操控动作的能力。目前机器人大模型在精度和灵活度方面也存在明显不足。除了物理上的限制，大模型还需要更多基本指令来完成复杂任务，而且也无法仅通过观察来学习复杂技能。较低的控制频率还限制了机器人在实时和高速环境中的操控能力。即使是执行包含流畅、精准动作的任务，大模型也不太适合。除此之外，如果不提供已有示例，训练机器人大模型学习全新动作仍面临巨大挑战，加之缺少可靠安全的机器人控制系统，机器人大模型还需进一步发展来弥补这些不足。

为适应未来的发展，大模型需要从专用 AI 模型、数字孪生技术、高性能计算资源等方面进行提升。集成专业 AI 有望提升精度和灵巧度，而数据处理技术则可以增强物理仿真和 AI 训练方面的能力，可以让机器人更深入地理解和预测物理相互作用。同时，还需开发智能混合控制系统，来整合大模型的高阶控制功能、面向专业领域的专用 AI 和低阶执行的传统方法，使操作更顺畅、更高效。还可以利用先进的计算和编程工具提高控制频率、提升响应实时性，让机器人处理动态任务更高效。如此全方位的方法将使机器人更自主、更灵活、更高效，让机器人拥有更强的能力驾驭现实世界中的复杂场景。6G 所具备的 AI 和感知能力，有望让机器人控制等级超越传统的任务控制，达到全新的“元控制”，让机器人有能力自主识别问题、定义任务、适应动态变化的环境。借助 6G 的 ISAC 和 AIaaS 能力，基于实时态势感知信息以及角色、任务、规则的元定义，机器人将可以更自主、更高效地识别任务、解决问题。

[4] https://github.com/microsoft/promptflow
[5] https://github.com/stanfordnlp/dspy

参考文献

[1] "A roadmap for US robotics: Robotics for a better tomorrow," National Robotics Initiative (NRI), 2024.

[2] "Joint strategic research innovation and deployment agenda (SRIDA) for the AI, data, and robotics partnership," European Union, 2020.

[3] "14th five-year plan for robot industry development," Chinese Government, 2021.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al*., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

[5] I. O. for Standardization, "ISO 8373: 2021 robotics–vocabulary," 2021.

[6] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, *et al*., "Foundation models in robotics: Applications, challenges, and the future," arXiv preprint arXiv:2312.07843, 2023.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, *et al*., "A survey on evaluation of large language models," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, 2024.

[8] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, *et al*., "A survey on vision transformer," IEEE transactions on pattern analysis and machine intelligence, vol. 45, no. 1, pp. 87–110, 2022.

[9] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Xu, *et al*., "PaLM-E: An embodied multimodal language model," arXiv preprint arXiv:2303.03378, 2023.

[10] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[11] N. Gothoskar, M. Lázaro-Gredilla, A. Agarwal, Y. Bekiroglu, and D. George, "Learning a generative model for robot control using visual feedback," arXiv preprint arXiv:2003.04474, 2020.

[12] K. Kawaharazuka, T. Matsushima, A. Gambardella, J. Guo, C. Paxton, and A. Zeng, "Real-world robot applications of foundation models: A review," arXiv preprint arXiv:2402.05741, 2024.

[13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al*., "PaLM: Scaling language modeling with pathways," Journal of Machine Learning Research, vol. 24, no. 240, pp. 1–113, 2023.

[14] A. Ulhaq, N. Akhtar, G. Pogrebna, and A. Mian, "Vision transformers for action recognition: A survey," arXiv preprint arXiv:2209.05700, 2022.

[15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al*., "RT-1: Robotics transformer for real-world control at scale," arXiv preprint arXiv:2212.06817, 2022.

[16] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al*., "RT-2: Vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.

[17] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al*., "Octo: An open-source generalist robot policy," arXiv preprint arXiv:2405.12213, 2024.

[18] F. F. Monteiro, A. L. B. Vieira, J. M. X. N. Teixeira, V. Teichrieb, *et al*., "Simulating real robots in virtual environments using NVIDIA's Isaac SDK," in Anais Estendidos do XXI Simpósio de Realidade Virtual e Aumentada. SBC, 2019, pp. 47–48.

[19] S. Nasiriany, A. Maddukuri, L. Zhang, A. Parikh, A. Lo, A. Joshi, A. Mandlekar, and Y. Zhu, "RoboCasa: Large-scale simulation of everyday tasks for generalist robots," arXiv preprint arXiv:2406.02523, 2024.

[20] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012, pp. 5026–5033.

[21] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al*., "HABITAT: A platform for embodied AI research," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9339–9347.

[22] 3GPP, "3rd generation partnership project; technical specification group TSG SA; study on network of service robots with ambient intelligence (Release 19)," 3GPP, Tech. Rep. TR TR 22.916 V1.0.0 (2023-12), 2023.

[23] S. Kerboeuf, P. Porambage, A. Jain, P. Rugeland, G. Wikström, M. Ericson, D. T. Bui, A. Outtagarts, H. Karvonen, P. Alemany, *et al*., "Design methodology for 6G end-to-end system: Hexa-X-II perspective," IEEE Open Journal of the Communications Society, 2024.

[24] D. Kortenkamp, R. Simmons, and D. Brugali, "Robotic systems architectures and programming," Springer handbook of robotics, pp. 283–306, 2016.

[25] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," Human factors, vol. 37, no. 1, pp. 32–64, 1995.

[26] X. Li, "Net4AI: Supporting AI as a service in 6G," HuaweiTech, 2022. [Online]. Available: https://www.huawei.com/en/huaweitech/future-technologies/net4ai-supporting-ai-as-a-sevice-6g

[27] C. E. DE NORMALISATION and E. K. F. NORMUNG, "Guidelines for the development and use of safety testing procedures in human-robot collaboration," 2022.

[28] 3GPP, "3rd generation partnership project; technical specification group TSG SA; feasibility study on integrated sensing and communication (Release 19)," 3GPP, Tech. Rep. TS 22.837 V19.3.0 (2024-03), 2024.

[29] ——, "3rd generation partnership project; specification group TSG SA; service requirements for integrated sensing and communication (Release 19)," 3GPP, Tech. Rep. TS 22.137 V19.1.0 (2024-03), 2023.

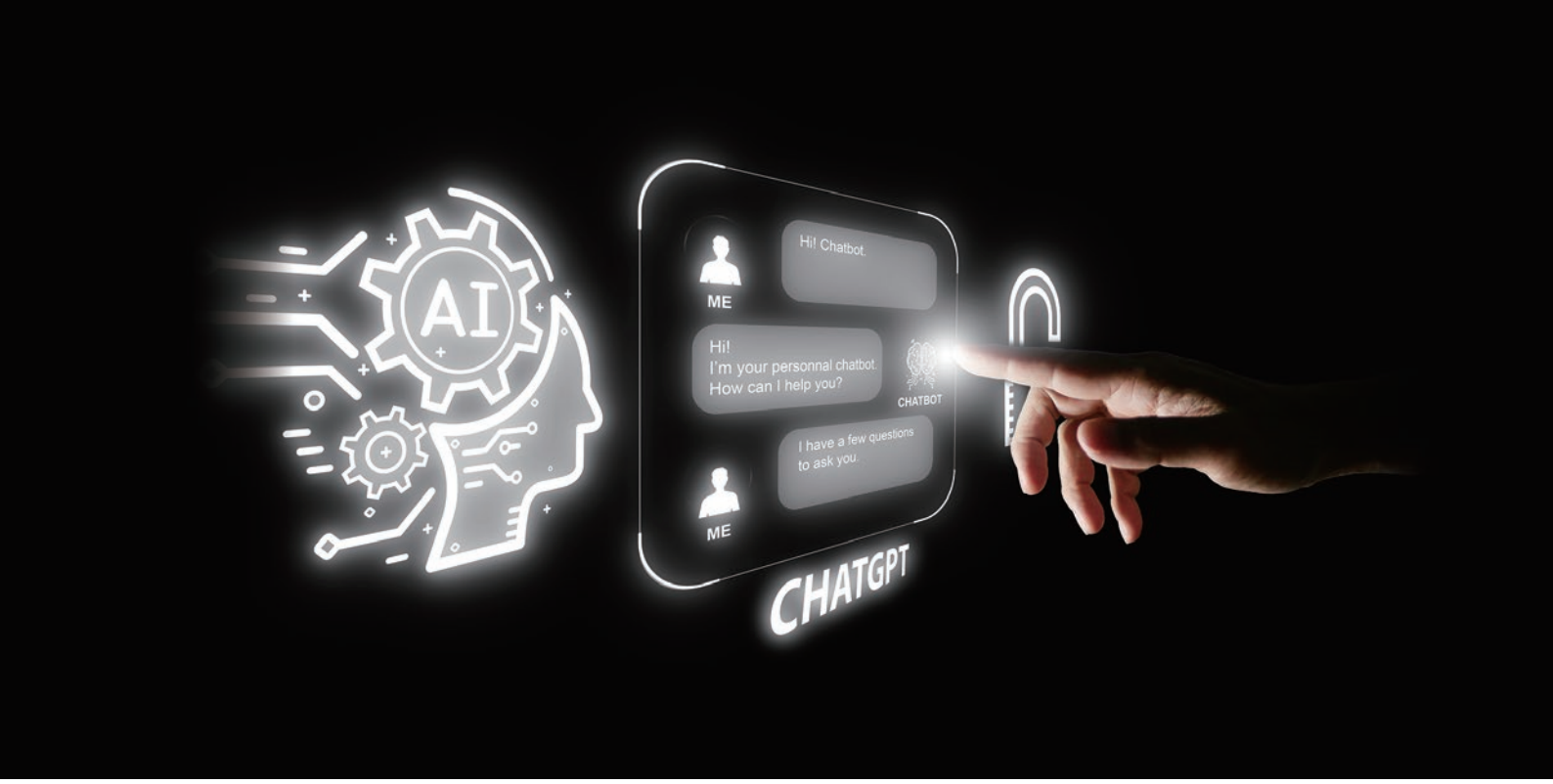
[30] Object Management Group (OMG), "Data Distribution Service (DDS)," 2015. [Online]. Available: https://www.omg.org/spec/DDS/1.4/PDF

[31] Y. S. Bareedu, T. Frühwirth, C. Niedermeier, M. Sabou, G. Steindl, A. S. Thuluva, S. Tsaneva, and N. Tufek Ozkaya, "Deriving semantic validation rules from industrial standards: An OPC UA study," Semantic Web, vol. 15, no. 2, pp. 517–554, 2024.

[32] Organization for the Advancement of Structured Information Standards (OASIS), "MQTT version 3.1.1," October 2014. [Online]. Available: https://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html

[33] Eclipse Zenoh Project, "Zenoh: A protocol for data-centric, resource-efficient, and location-transparent data sharing," May 2020. [Online]. Available: https://zenoh.io/

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474.



大语言模型在无线通信知识管理领域的实践探索

侯宏伟，马驰翔，杜笠弘，李俊辉

摘要

近年来，大语言模型取得了突破性的进展，引发了新一轮技术变革。在大语言模型卓越能力的基础上，知识管理领域展现出广阔的应用前景，但也同样面临一些严峻挑战：第一，为了最大化可访问性和适用性，大语言模型采用了互联网上的通用数据进行训练，对广泛且通用数据的侧重导致在它专业领域中的表现不尽如人意。第二，大语言模型容易生成虽具说服力但不准确的回应，这种现象被称为“幻觉”，可能会误导用户。为了解决大语言模型在知识管理领域遇到的挑战，业界主要有两种解决方案：微调和检索增强生成。本次实践聚焦于检索生成增强技术，完成无线通信知识库问答方案的总体设计，并设计测评方案，评估出大语言模型、嵌入模型和重排序模型的最佳组合。通过对大语言模型、嵌入模型和重排序模型的组合应用，并采用多种开源工具，我们成功实现了无线通信知识库问答的落地实践。检索增强生成技术通过结合大语言模型和数据库，有效减少了生成过程中的“幻觉”现象。然而，在实践过程中也面临一些挑战和局限，包括多模态数据处理、超参数选择、与现有企业知识库和搜索引擎的融合，以及如何处理数据的时间属性等。未来，进一步完善和优化检索增强生成技术，有望提升其在无线通信知识管理领域的应用效果和价值。

关键词

大语言模型，幻觉，微调，检索生成增强，嵌入模型，重排序模型，数据库

1 大语言模型的发展

大语言模型（Large Language Model，LLM）的发展历史可以追溯到 2005 年，当时主要是使用大型 n-gram 模型进行机器翻译 [1]。随后，2017 年，Transformer 架构的提出彻底改变了自然语言处理（Natural Language Processing，NLP）领域，这种基于注意力机制的新型网络架构，使得模型在多个任务上取得了显著的性能提升 [2]。到了 2018 年和 2019 年，BERT 模型的出现进一步推动了预训练语言模型（Pre-trained Language Model，PLM）的发展。BERT 通过双向编码器的方式，有效地利用了左右两侧的上下文信息，从而在多个 NLP 任务上达到了新的最佳性能 [3]。紧接着，RoBERTa 模型的提出，对 BERT 进行了优化，通过调整超参数和训练数据的大小，实现了更高的性能 [4]。

2020 年，GPT-3 的发布标志着大语言模型的一个重要里程碑，它展示了通过增加模型规模来提高语言模型的泛化能力和少样本学习能力的可能性。此外，GPT-3 还能生成难以与人类作品区分的新闻文章样本，显示了其在文本生成方面的强大能力 [5]。

近年来，大语言模型不仅在传统的文本处理任务上表现出色，还开始被应用于多模态任务，如结合图像和文本的任务 [6]。此外，随着技术的不断进步，大语言模型在保持知识更新、适应不断变化的世界知识方面也面临新的挑战和研究方向 [7, 8]。

总体来看，大语言模型的发展经历了从简单的统计模型到复杂的神经网络模型，再到现在的大规模预训练模型的演变过程。这一过程不仅涉及算法和架构的创新，还包括对模型训练、评估和应用等方面的深入研究 [9–11]。未来，大语言模型的发展可能会更加注重模型的可解释性、效率以及如何更好地融合和处理多种类型的数据 [6, 12]。

2 大语言模型在知识管理领域的典型技术

大语言模型的出色能力使得其在知识管理领域可以有很大发挥空间，但是也同样面临一些严峻挑战。第一，为了最大化可访问性和适用性，大语言模型采用了互联网上的通用数据进行训练，对广泛且通用数据的侧重导致在它专业领域中的表现不尽如人意。第二，大语言模型容易生成虽具说服力但不准确的回应，这种现象被称为“幻觉”，可能会误导用户。

为了解决大语言模型在知识管理领域遇到的挑战，业界主要有两种解决方案：微调（Fine-Tuning）和检索增强生成（Retrieval Augmented Generation，RAG）。

2.1 微调

大模型的微调是一种机器学习技术，用于在预训练的大型模型基础上，通过少量的特定任务数据进行再训练，以适应新的或特定的应用场景。这种方法通常涉及在预训练模型的基础上添加一个或多个输出层，并使用特定任务的数据集进行进一步训练，从而使模型能够更好地理解和执行该任务。微调的目的是利用预训练模型已经学到的通用知识，避免从头开始训练模型所需的大量计算资源和时间。例如，BERT 模型就是通过在大量文本数据上预训练，然后在特定任务上进行微调，实现了在多个自然语言处理任务上的突破性性能提升 [3]。

2.2 检索增强生成

检索增强生成是一种结合了预训练的参数化记忆（如大型语言模型）和非参数化记忆（如 Wikipedia 的密集向量索引）的方法，用于提升语言生成任务的性能。这种方法通过在生成过程中动态地从外部知识资源中检索信息，以提高生成内容的准确性、多样性和事实性。RAG 模型通常包括两个部分：一个大语言模型作为参数化记忆，以及一个访问非参数化记忆（如密集向量索引）的检索器 [13]。

2.3 RAG 和微调的优缺点对比

正如 [14] 所说，两种技术方案可以从 6 大维度进行对比：动态数据，外部知识，模型定制化，幻觉减少，透明度和技术实现难度。

我们希望利用大语言模型的能力赋能无线通信领域的知识管理，RAG 在动态数据、外部知识、幻觉减少、透明度、技术实现难度方面更有优势，因此选择采用 RAG 这条技术路线。

表 1 RAG 和微调优缺点对比

维度	RAG	微调
动态数据	Win	Lose
外部知识	Win	Lose
模型定制化	Lose	Win
幻觉减少	Win	Lose
透明度	Win	Lose
技术实现难度	Win	Lose

3 实践方案

3.1 无线通信知识库问答方案总体设计

整个方案包含两个部分：第一部分是离线构建无线通信知识库环节；第二部分是用户在线问答环节。

3.1.1 离线构建无线通信知识库

离线构建无线通信知识库环节的整体过程如图 1 的离线部分所示。首先，用户可以上传各类文档，包括但不限于代码文档、3GPP 协议等。接着，对用户上传的文档进行解析，清洗，切片。之后可以利用大语言模型对每个切片生成问答对。接下来，对问答对和原始切片数据建立向量索引和关键词索引，其中构建向量索引的过程用到了嵌入模型（Embedding Model）。然后将索引建立完成的切片和问答对都放入数据库。向量数据放入向量数据库，原始数据放入普通数据库。注意，生成问答对的环节是可选的，不是必须的。

3.1.2 在线问答

在线问答环节的整体过程如图 1 的在线部分所示。首先用户在输入框中输入问题进行提问，接着大语言模型可以对用户的提问进行意图识别。不同的意图可以调用不同的无线通信知识库或者进行不同的流程处理。这个意图环节部分是可选的，不是必须的。之后利用混合检索从数据库中召回和用户问题最相关的前 *K* 个知识片段。接着，利用重排序模型（Rerank Model）对召回的 *K* 个知识片段，结合用户的提问进行进一步排序，得到最相关的 *N* 个知识片段。然后，将 *N* 个知识片段和用户的提问，按照提示词模板的格式进行重新整理，放入 LLM 中。最后大语言模型根据输入给出相应的回答和回答所依赖的无线通信知识库片段。

3.1.2.1 混合检索

在线回答环节提到的混合检索包括语义检索和关键词检索。其中语义检索利用嵌入模型将用户的提问进行向量化，然后将问题向量和向量数据库里面的向量进行匹配，召回语义相近的 *K* 个知识片段。关键词搜索指的是利用关键词从数据库中进行搜索。

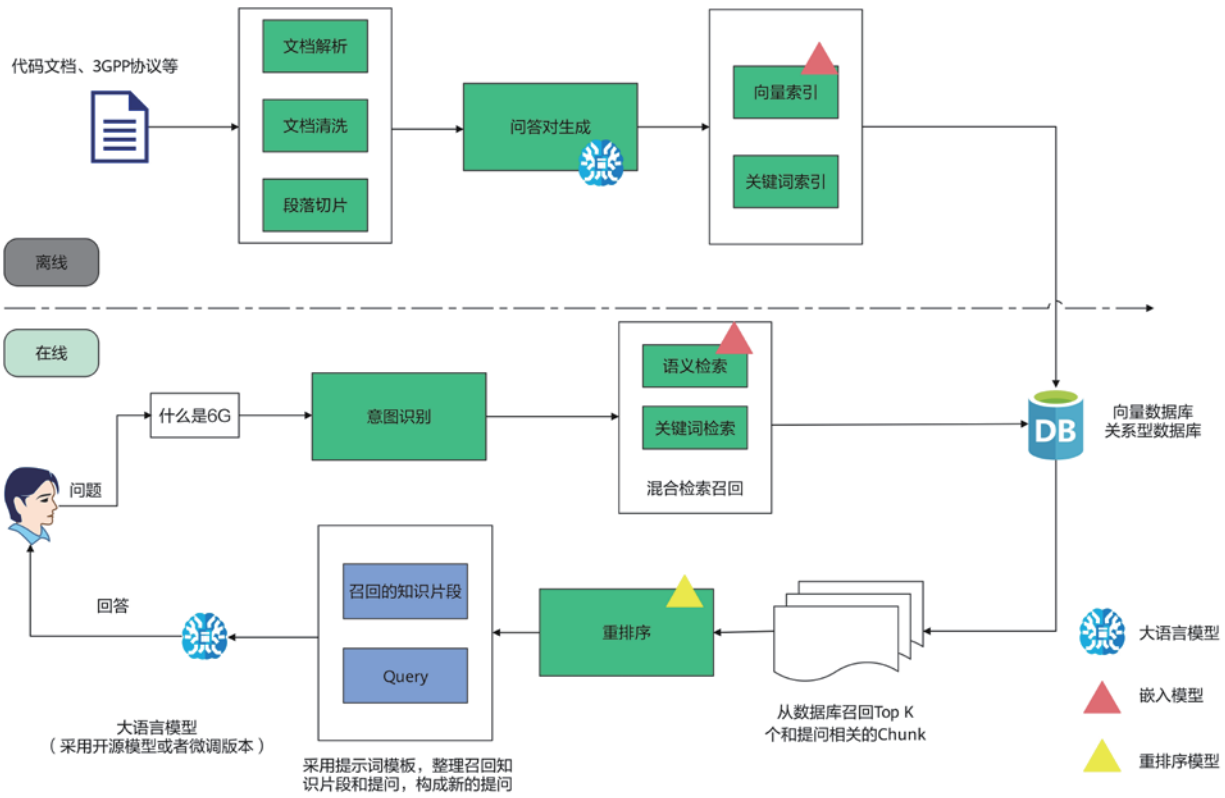


图 1 大语言模型在无线通信知识管理领域的实践探索设计图

语义检索能够实现复杂语义的文本检索，并且具有以下优势：

- 多语言理解能力：实现跨语言的理解，例如可以将中文输入匹配到英文内容。
- 多模态理解能力：支持文本、图像以及音视频等不同格式内容的相似匹配。
- 强大的容错性：能够处理拼写错误以及模糊不清的描述。

尽管语义检索在上述应用场景中展现出显著的优势，但在特定情况下可能效果不尽人意。例如：

- 当搜索特定的人名或物品名称时（如查询“华为 Mate 60”，使用语义检索可能会返回“Mate 50”等相关词条）；
- 搜索缩写词或简短短语（如“LLM”）；

这些局限性恰恰是传统关键词搜索所擅长的领域，传统关键词搜索优势在于：

- 精确匹配：能够准确匹配产品名称、人名等。
- 简短字符的快速检索：当用户仅输入几个关键字时，也能高效检索（尽管这在向量检索中效果不佳）。
- 低频词汇的强匹配能力：低频词汇在语言中往往承载更重要的含义，比如在“你想跟我去喝咖啡吗？”这句话中，“喝”、“咖啡”比“你”、“想”、“吗”具有更显著的意义。

在文本搜索的多种场景中，最关键的是确保能够呈现出与查询最相关的结果。向量检索与关键词检索各自在检索技术中占据独特的优势。混合检索技术正是结合了两者的优点，并有效弥补了各自的不足。

3.1.2.2 重排序

混合检索通过结合多种检索技术的优势，致力于提升搜索结果的召回效率。为了使来自不同检索系统的结果可以进行有效整合，采用了数据归一化策略：即将数据统一转换至一个标准范畴或分布模式，以便于后续的比对、分析与处理，并统一交予大型模型处理。此过程中，至关重要的一环是引入评分系统，即重排序模型。

重排序模型的作用在于通过衡量候选文档列表与用户查询语义的相符度，并据此进行重排，从而优化语义排序结果。该模型基于计算用户问题与每项候选文档之间的相关性分数，并按此相关性由高至低排列文档列表。

重排序不仅在汇集多个检索系统结果时发挥作用，即便在单一检索模式下，通过在关键词检索后加入语义重排序的步骤同样能显著提升文档的召回率。此外，为了提高计算效率，多数向量数据库通常会在某种程度上牺牲准确性。这会导致其检索结果具备一定的不确定性，召回的知识片段不一定是按照和用户的提问相关性从高到低排序的。因此，原始返回的最相关 Top *K* 结果可能并非最为精确，这就需要借助重排序模型对结果进行进一步的优化。

重排序模型并非检索技术的替代，而是一种强化现有检索系统的辅助手段。它的优势在于提供了一种简便且复杂度低的方式，不仅允许将语义相关性整合至现行的搜索系统，还免去了对基础设施的重大修改。

3.2 模型组合测评

在图 1 的设计图中用到了三类模型：大语言模型，嵌入模型和重排序模型。我们选择采用开源模型进行本地部署，利用本地文档构建无线通信知识库。

3.2.1 模型选择

3.2.1.1 大语言模型的选择

参考大语言模型的榜单“LMSYS Chatbot Arena”[15]，我们挑选了三款大语言模型：Llama-3-70b-Instruct、Command R+ 和 Qwen1.5-110B-Chat。这三款大语言模型都支持中文和英文。

3.2.1.2 嵌入模型的选择

参考嵌入模型的榜单“Massive Text Embedding Benchmark (MTEB) Leaderboard”[16]，主要关注“Retrieval”指标。

针对中文语言，我们挑选 4 款嵌入模型：360Zhinao-search、stella-mrl-large-zh-v3.5-1792d、PEG 和 bce-embedding-base_v1。

针对英文语言，我们同样挑选了 4 款嵌入模型：SFR-Embedding-Mistral、gte-large-en-v1.5、GritLM-7B 和 bce-embedding-base_v1。

3.2.1.3 重排序模型的选择

参考 [17, 18]，我们选择了两款重排序模型：bge-reranker-v2-gemma 和 bce-reranker-base_v1。这两个模型都支持中英文。

3.2.1.4 模型组合

我们对中英文分别进行测评，挑选出最佳中文模型组合和最佳英文模型组合。对于中文，LLM 有 3 种候选，嵌入模型有 4 种候选，重排序模型有 2 种候选，因此总计 24 种模型组合；对于英文，LLM 有 3 种候选，嵌入模型有 4 种候选，重排序模型有 2 种候选，因此总计 24 种模型组合。采用 vLLM [19] 运行 LLM，采用 Xinference [20] 运行嵌入模型和重排序模型。

3.2.2 测评方式

我们分别构建了中文和英文数据集。数据集来源为开源项目 RGB [21]。中文数据集采用 zh_refine.json 文件；英文数据集采用 en_refine.json 文件。每条数据的格式如图 2 所示。

```
"id": xxx,
"query": "xxx",
"answer": "xxx",
"positive": "xxx",
"negative": "xxx"
```

图 2 原始数据集格式

“id”表示数据编号，“query”表示本条数据的提问，“answer”表示本条数据的回答，“positive”表示和提问相关的文本，“negative”表示和提问不相关的文本，属于干扰文本。将 300 条中文数据的“positive”文本和“negative”文本统一放到一个文件中，利用前面章节提到的嵌入模型放入向量数据库中，构成中文知识库；采用同样的方式构建英文知识库。向量数据库采用开源的向量数据库 Chroma [22]。

测评框架采用 Ragas [23]。Ragas 的测评数据需要满足一定的格式，所需格式如图 3 所示。

```
data = {
    "question : questions,
    "answer": answers,
    "contexts": contexts,
    "ground_truths": ground_truths
}
```

图 3 Ragas 测评数据格式

其中“question”表示本次的提问，“ground_truths”表示本次提问对应的正确答案，这两个字段的内容都可以从 zh_refine.json 或者 en_refine.json 中获取。“answer”和“contexts”来自于大语言模型给出的回答。为了得到这两个字段，我们提供给大语言模型的提示词模板格式如图 4 所示。

对以上数据集和中英文模型组合分别进行测评。

3.2.3 测评结果

我们关注 faithfulness、answer_relevancy、context_precision 和 context_recall 四个测评指标，每个指标的详细含义可以参考 [24]。

3.2.3.1 中文测评结果

中文一共有 24 种模型组合。将 faithfulness、answer_relevancy、context_precision 和 context_recall 四个测评指标评分进行累加，得到每个组合的总分。

```
#Create prompt model
template = """You are an assistant for question-answering tasks.
Use the following pieces of retrieved context to answer the question.
If you don't know the answer, just say that you don't know.
Use two sentences maximum and keep the answer concise.
Question: {query}
Context: {context}
Answer:
"""
```

图 4 提示词模板格式

图 5 横坐标为各个模型组合的名字，其格式为“zh_x_y_z”。“x”代表 LLM 的名字，序号 0 表示 Command R+，序号 1 表示 Llama-3-70b-Instruct，序号 2 表示 Qwen1.5-110B-Chat；“y”代表嵌入模型的名字，序号“0”表示 stella-mrl-large-zh-v3.5-1792d，序号“1”表示 bce-embedding-base_v1，序号“2”表示 360Zhinao-search，序号“3”表示 PEG；“z”代表重排序模型的名字，序号 0 表示 bce-reranker-base_v1，序号 1 表 bge-reranker-v2-gemma。

从图中可以看出，中文最佳的模型组合为：Command R+、stella-mrl-large-zh-v3.5-1792d 和 bge-reranker-v2-gemma。采用该模型组合进行落地实践，用于中文应用场景。

3.2.3.2 英文测评结果

英文一共有 24 种模型组合。将 faithfulness、answer_relevancy、context_precision 和 context_recall 四个测评指标评分进行累加，得到每个组合的总分。

图 6 横坐标为各个模型组合的名字，其格式为“en_x_y_z”。“x”代表 LLM 的名字，序号 0 表示 Command R+，序号 1 表示 Llama-3-70b-Instruct，序号 2 表示 Qwen1.5-110B-Chat；“y”代表嵌入模型的名字，序号 0 表示 SFR-Embedding-Mistral，序号 1 表示 bce-embedding-base_v1，序号 2 表示 gte-large-en-v1.5，序号 3 表示 GritLM-7B；“z”代表重排序模型的名字，序号 0 表示 bce-reranker-base_v1，序号 1 表 bge-reranker-v2-gemma。

从图中可以看出，英文最佳的模型组合为：Llama-3-70b-Instruct、SFR-Embedding-Mistral 和 bce-reranker-base_v1。采用该模型组合进行落地实践，用于英文应用场景。

3.3 方案落地效果

我们采用 Dify [25] 作为落地无线通信知识库问答方案的底层框架，采用上一章节挑选出的最佳模型组合，利用华为内部文档构建无线通信知识库，采用工作流将 RAG 各个环节串联起来，最终构建出一个大模型应用软件。

在整个环节中，“问题分类器”承担意图识别的职责：如果用户的提问涉及到无线通信领域的问题，大模型应用软件就依次进行 RAG 的各个环节，直到给出用户问题的答案；如果用户的提问是其他领域的问题，大模型应用软件就拒绝回答，并且给出礼貌回复。

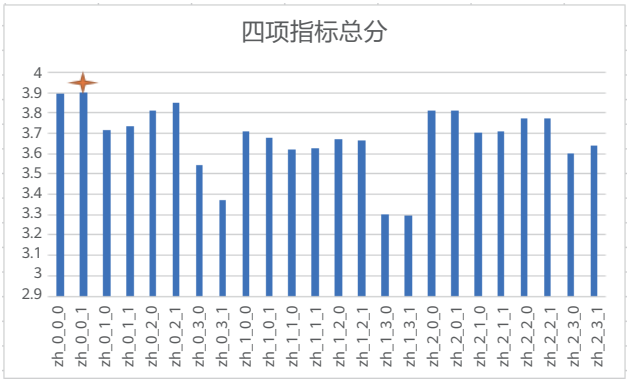


图 5 各个模型组合在中文测评的四项指标总分

图 6 各个模型组合在英文测评的四项指标总分

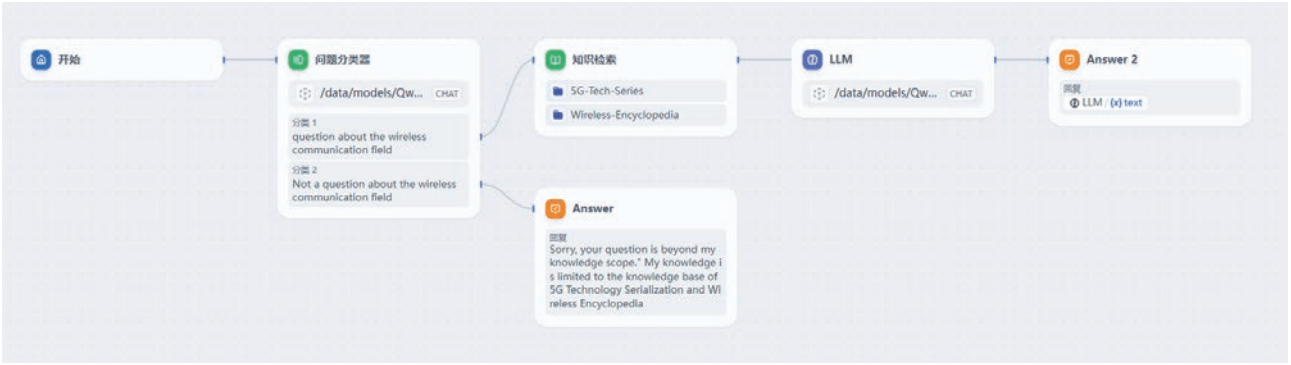


图 7 无线通信知识库问答方案落地

3.3.1 无线通信领域相关问题

从图 8 的提问可以看出，用户的提问涉及到 5G 通信领域的问题，大模型应用软件依次进行了 RAG 的各个环节，最终给出了用户问题的答案，并且给出相关的文献材料。

3.3.2 其他领域的问题

图 9 的提问不属于 5G 通信领域，大模型应用软件拒绝回答，并且给出了礼貌回复。

4 未来展望

RAG 通过结合大语言模型和数据库，大幅减少了生成过程中的“幻觉”现象，但是仍然有一些挑战需要克服。本章主要介绍 RAG 当前的挑战和未来 RAG 面临的研究方向。

- 多模态数据

  在企业的数字资产中，PPT 和 PDF 文件的比重较高，其中包含大量的图片、表格等非结构化数据，纯文字类型的占比极少。目前，大型语言模型最稳定的信息获取方式还是以纯文字为主。在应用模型时，如何确保其能够准确、稳定地从非结构化数据（如图片和表格）中提取关键信息，是提升模型性能的重要方向。

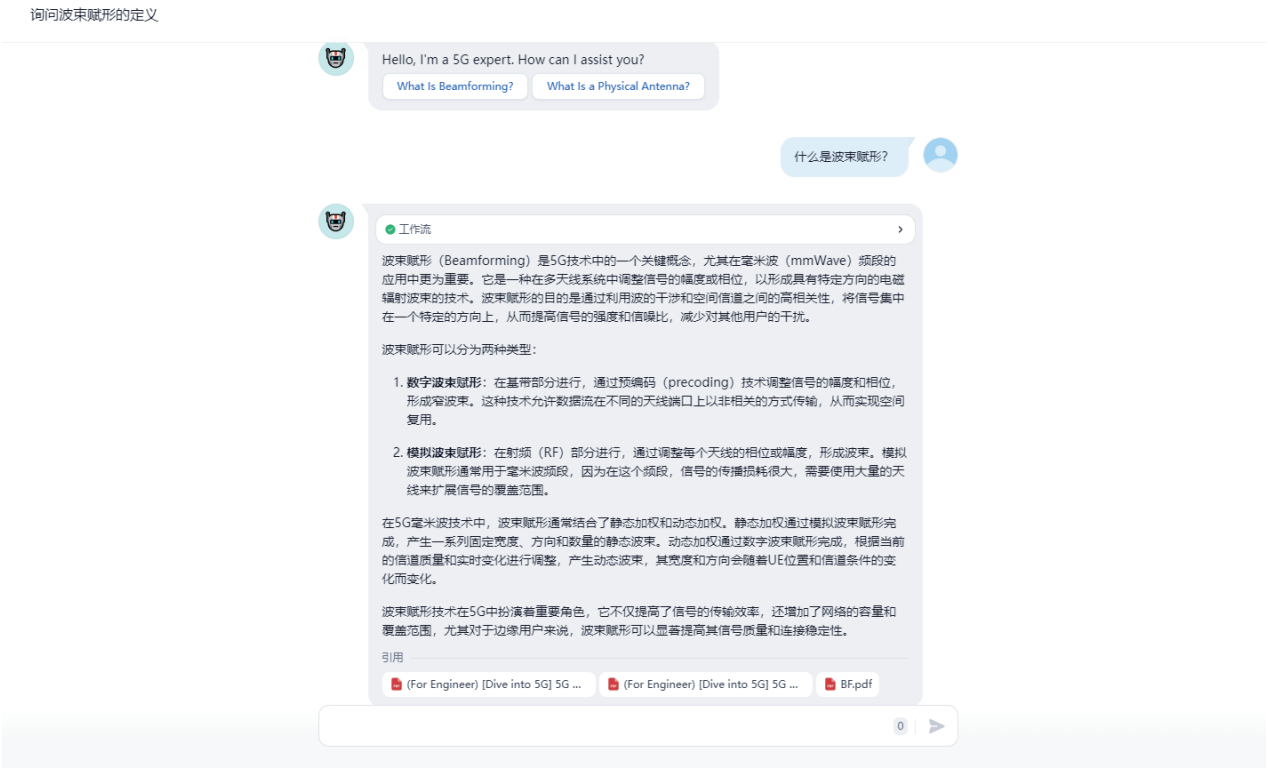


图 8 大模型应用回答 5G 通信领域相关问题

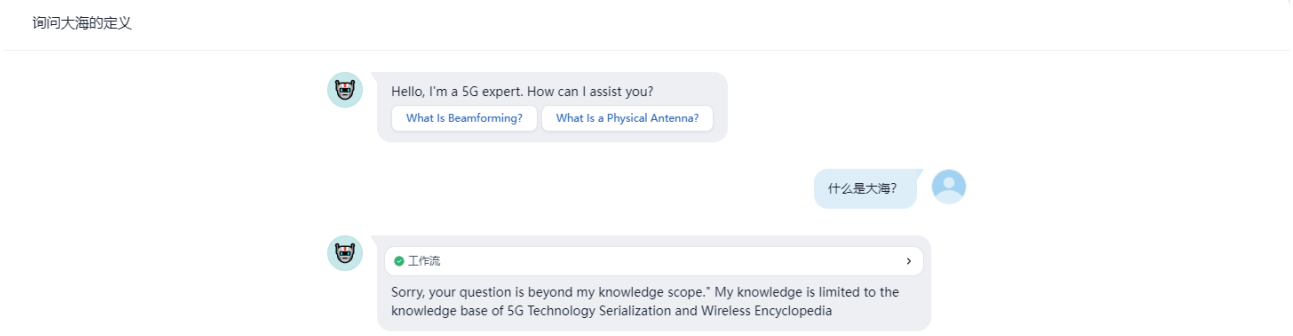


图 9 大模型应用回答其他领域相关问题

- RAG 链路组件多，超参数多

  RAG 在应用时通常包含多种组件，LlamaIndex 的 RAG 组件细分就有 80 多个，针对不同场景有多种解决方案。这导致 RAG 应用开发容易变成一项“调参工程”：切片方式、召回方式、前处理、后处理、路由等众多参数都需要调整和优化。如此庞大的参数空间给开发者和研究人员带来了挑战，如何在众多组件和参数中找到最佳组合，需要大量的实验和试错。这同时也凸显了简化 RAG 应用开发流程、提高开发效率的重要性。

- 与现有企业知识库和搜索引擎结合

  RAG 并不需要从零开始打造知识库，它可以很好地融合现有的企业搜索引擎（如 Elasticsearch 搜索）和关键词搜索引擎。但是，为了更好地适应大型语言模型的需要，需要对接口进行定制化设计。其中，需要关注的重点是解决上下文窗口限制、简化语言表达、减少接口参数数量并确保易于理解。通过量身定制接口，可以提升大型语言模型和搜索引擎之间的兼容性，从而充分发挥 RAG 模型的潜力，更有效地利用企业的知识库和数据资源。

- 缺乏时间属性

  现有的 RAG 策略可能面临一个问题：数据库中存在相互矛盾的数据集。例如，企业政策文档的更新可能会导致旧规则和新规则的冲突。当前的 RAG 召回策略没有考虑时间属性，在进行向量召回时，可能会同时召回矛盾的知识片段，进而导致大型语言模型给出错误答案。为解决此问题，未来的 RAG 策略可以引入数据的时间属性。在向量召回时，对更新的、更具相关性的数据给予更高的召回分数，这有助于减少矛盾知识片段被同时召回的可能性。此外，在模型训练和应用时，增加对时间属性的关注，也有助于提升模型的准确性和适应性。

- 针对领域数据微调嵌入模型和重排序模型

  开源的嵌入模型和重排序模型大多数是基于通用领域的语料进行训练得到的，在专用领域，比如无线通信领域，可能就无法发挥出最佳性能，RAG 的召回过程可能无法召回最相关的语料。为此，针对专用领域的数据进行进一步微调，进一步提高专用领域 RAG 的性能。

5 结语

近年来，大语言模型的突破性进展开启了一轮技术创新热潮，也为知识管理领域带来了广阔的发展前景。本文首先阐述了大语言模型的发展历程，之后又介绍了在知识管理领域的挑战和常用技术方案，包括微调和检索增强生成。随后，本文将重点转向检索生成增强技术，并阐述了实践方案的总体设计。通过精心设计评估方案，根据评估结果，选取了大语言模型、嵌入模型和重排序模型的最佳组合。最终，我们成功利用多种开源工具将无线通信知识库问答的大模型应用落地实现。检索增强生成技术通过整合大语言模型和数据库，显著减少了生成回答时的“幻觉”现象。然而，正如本文所讨论的，检索生成增强技术在实际应用中也面临一些挑战，如多模态数据处理、超参数选择的复杂性，以及如何顺利融合现有企业知识库和搜索引擎等。

总的来说，本文展示了检索增强生成技术在无线通信知识管理领域的巨大潜力。尽管未来仍有进一步完善的空间，但当前成果已经为提升大语言模型在无线通信知识管理中的应用效果和价值奠定了坚实基础。

参考文献

[1] Naomi Saphra, Eve Fleisig, *et al*., "First tragedy, then parse: History repeats itself in the new era of large language models," arXiv.org (2023).

[2] Ashish Vaswani, Noam M. Shazeer, *et al*., "Attention is all you need," *Neural Information Processing Systems* (2017).

[3] Jacob Devlin, Ming-Wei Chang, *et al*., "BERT: Pre-training of deep bidirectional Transformers for language understanding," *North American Chapter of the Association for Computational Linguistics* (2019).

[4] Yinhan Liu, Myle Ott, *et al*., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv.org (2019).

[5] Tom B. Brown, Benjamin Mann, *et al*., "Language models are few-shot learners," *Neural Information Processing Systems* (2020).

[6] Shukang Yin, Chaoyou Fu, *et al*., "A survey on multimodal large language models," arXiv.org (2023).

[7] Zihan Zhang, Meng Fang. *et al*., "How do large language models capture the ever-changing world knowledge? A review of recent advances," *Conference on Empirical Methods in Natural Language Processing* (2023).

[8] Boxi Cao, Hongyu Lin, *et al*., "The life cycle of knowledge in big language models: A survey," Machine Intelligence Research (2023).

[9] Lizhou Fan, Lingyao Li, *et al*., "A bibliometric review of large language models research from 2017 to 2023," arXiv.org (2023).

[10] Michael R Douglas, "Large language models," *Communications of the ACM* (2023). 7–7.

[11] 舒文韬，李睿潇，孙天祥等 . 大型语言模型：原理、实现与发展 [J]. 计算机研究与发展，2024，61(02)：351–361.

[12] Christopher Akiki, Giada Pistilli, *et al*., "BigScience: A case study in the social construction of a multilingual large language model," arXiv.org (2022).

[13] Patrick Lewis, Ethan Perez, *et al*., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Neural Information Processing Systems (2020).

[14] https://www.rungalileo.io/blog/optimizing-llm-performance-rag-vs-finetune-vs-both

[15] https://chat.lmsys.org/?leaderboard

[16] https://huggingface.co/spaces/mteb/leaderboard

[17] https://github.com/FlagOpen/FlagEmbedding/tree/master/FlagEmbedding/llm_reranker

[18] https://huggingface.co/maidalun1020/bce-reranker-base_v1

[19] https://github.com/vllm-project/vllm

[20] https://github.com/xorbitsai/inference

[21] https://github.com/chen700564/RGB?tab=readme-ov-file

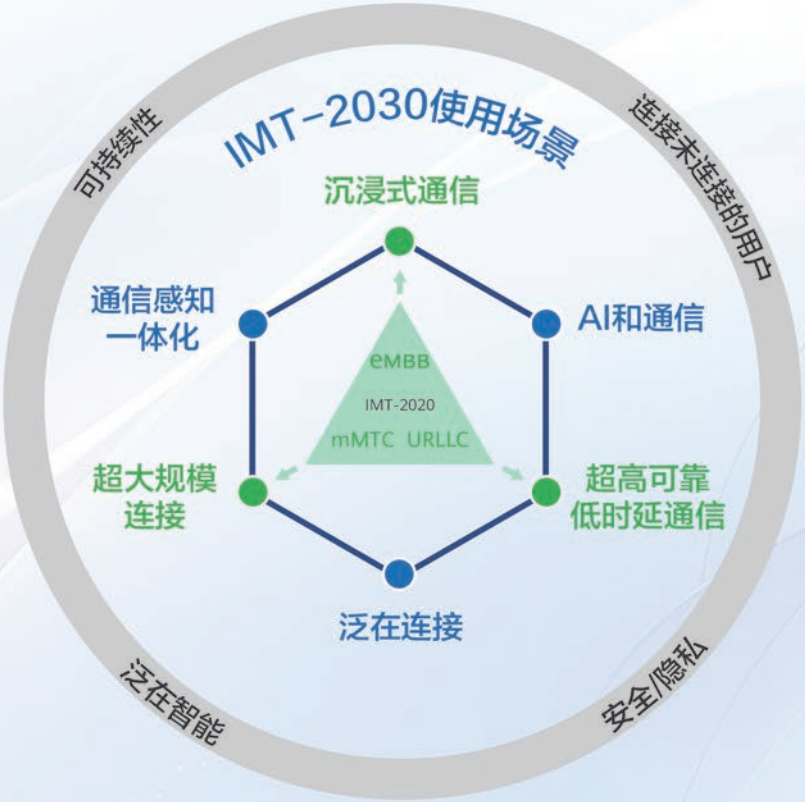
[22] https://github.com/chroma-core/chroma

[23] https://github.com/explodinggradients/ragas

[24] https://docs.ragas.io/en/stable/concepts/metrics/index.html

[25] https://github.com/langgenius/dify

跨越人联、物联，迈向万物智联



《华为研究》6G系列专刊下载

第1期
综合

第2期
6G

第5期
ISAC

第7期（本期）
AI和通信

华为技术有限公司
深圳市龙岗区坂田华为总部办公楼
电话：+86 755 28780808
邮编：518129

商标声明

免责声明