DOI 10.15302/J-SSCAE-2024.01.011

端云协同智能计算的关键问题、方法和应用

张圣宇[1,2]，况琨[2]，吕承飞[2,3]，李纪为[2]，肖俊[2]，吴帆[4]，吴飞[2,5*]

（1. 浙江大学软件学院，杭州 310013；2. 浙江大学计算机科学与技术学院，杭州 310013；3. 淘宝（中国）软件有限公司，杭州 310012；4. 上海交通大学计算机科学与工程系，上海 200240；5. 浙江大学上海高等研究院，上海 201203）

摘要：端云协同智能计算是大数据、云计算、边缘计算发展的产物，可在保护用户隐私的前提下显著提升数据利用率，实现智能计算实时响应能力与服务鲁棒性的优势互补，而相应技术研发和实践应用具有复杂性。本文剖析了端云协同智能计算的应用价值，凝练了端学习效率优化、端少样本过拟合、端模型定制化、分布差异下虚假关联学习、通信开销与计算效率平衡等方面的技术难题；系统梳理了端云协同智能计算中主流方法研究进展，涉及作为应用基石的高效计算硬件、以端为中心的协同计算、以云为中心的协同计算、端云双向协同计算、可信端云协同智能计算等主要方向；总结了推荐系统、自动驾驶、安防系统、教育模式等端云协同智能计算的垂直领域应用情况。着眼端云协同智能计算的未来发展，需重点研究云资源在端模型个性化中的应用策略、端云协同多目标优化算法、端–端与云协同计算的优化策略。

关键词：端云协同；大小模型协同计算；端计算；可信协同；机器学习

中图分类号：TP18；TP393　　**文献标识码：**A

Device-Cloud Collaborative Intelligent Computing: Key Problems, Methods, and Applications

Zhang Shengyu[1,2], Kuang Kun[2], Lyu Chengfei[2,3], Li Jiwei[2], Xiao Jun[2], Wu Fan[4], Wu Fei[2,5*]

(1. School of Software Technology, Zhejiang University, Hangzhou 310013, China; 2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310013, China; 3. Taobao (China) Software Co., Ltd., Hangzhou 310012, China; 4. Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; 5. Shanghai Institute for Advanced Study of Zhejiang University, Shanghai 201203, China)

Abstract: Device-cloud collaborative intelligent computing, an emergent result of the development in big data, cloud computing, and edge computing, offers significant improvements in data utilization while protecting user privacy. This approach synergizes the real-time response capabilities of intelligent computing with service robustness. The study explores the application value of this computing paradigm, highlighting technical challenges such as optimizing on-device learning efficiency, mitigating overfitting with limited samples at the device, customizing on-device models, learning false associations under distributional discrepancies, and balancing communication overhead with computational efficiency. We systematically review the progress in mainstream methods within device-cloud collaborative intelligent computing, encompassing efficient computation hardware as the application cornerstone, device-centric collaborative computing, cloud-centric collaborative computing, bidirectional device-cloud collaborative computing, and trustworthy device-cloud collaborative computing. The study also summarizes applications in vertical domains such as recommendation systems,

收稿日期：2023-11-10；**修回日期：**2023-12-20

通讯作者：*吴飞，浙江大学计算机科学与技术学院教授，研究方向为人工智能、多媒体分析与检索、跨媒体计算；E-mail: wufei@zju.edu.cn

资助项目：科技创新2030—“新一代人工智能”重大项目“大小模型端云协同进化与系统”(2022ZD0119100)；中国工程院咨询项目“新一代人工智能及产业集群发展战略研究”(2022-PP-07)

本刊网址：www.engineering.org.cn/ch/journal/sscae

autonomous driving, security systems, and educational models. Looking toward the future of device-cloud collaborative intelligent computing, it underscores the need for focused research on cloud resource application strategies in device model personalization, multi-objective optimization algorithms for device-cloud collaboration, and optimized collaborative strategies between devices and the cloud.

Keywords: device-cloud collaboration; large and small model collaboration computing; on-device computing; trustworthy collaboration; machine learning

一、前言

当前，人工智能（AI）系统运行大多依赖中心化的云机器学习架构，使收集、处理、计算中心化数据面临着数据治理、隐私保护、带宽负载、网络时延方面的成本压力。针对于此，端云协同计算范式提供云上服务、端侧推理能力，推动云上模型、端侧模型的协同进化，从云计算、端智能向着端云协同进化计算模式演进，能够充分发挥云上、端侧、端云链中各类计算资源的应用效能。相应地，协同发展云服务与边缘计算服务、聚焦AI关键技术等内容纳入国家发展规划，成为加快数字化发展、建设数字中国的重要内容[1]。

端云协同智能计算的技术研发及落地应用具有挑战性和复杂性，也属新兴技术方向，因而梳理和总结相应研究进展具有价值。关于边缘计算方法及其产业应用已有一些研究和探讨：评述了传统边缘计算中的端至云卸载方案[2]，而未分析智能计算领域的特殊性；梳理了边缘计算在视频监控[3]、智慧交通[4]、元宇宙[5]等关键方向上的应用情况，而未讨论端云协同智能计算的新应用进展。近期，有研究对以隐私保护为核心的端云协同学习技术[6]、端云协同驱动的大规模预训练模型技术[7]等进行了详细分析，但仅将端云协同技术作为实现其他目标（如隐私保护）、支持其他技术（如预训练大模型）的辅助手段，而非针对端云协同智能计算这一独立方向开展的深入讨论。

本文侧重开展端云协同智能计算方向的多维度、系统性综述，在理论分析和技术对比的基础上，辨识核心问题、梳理关键方法、明确关键应用，进而提出技术与应用发展建议，以期为端云协同智能计算的技术探索、应用突破、产业升级等提供参考。需要指出的是，为了增强可读性并保持简洁性，文中的端云协同计算、端云协同、端云计算、端计算、云计算等，均指代基于AI机器学习方法的智能计算背景。

二、端云协同智能计算的应用价值

在传统的中心化云机器学习架构中，中心化的云服务器负责收集来自端设备的数据并据此训练和部署AI模型。端设备在服务过程中，将实时数据上传至云服务器以获得推理结果，据此进行端侧的决策制定或内容呈现。中心化的云机器学习架构为AI服务提供了灵活且具扩展性的平台，端设备能够访问并共享云侧的计算能力及存储资源，支持AI模型的构建与部署。中心化的云机器学习架构也有不足之处：① 集中式的数据收集与计算面临着数据治理和隐私保护方面的突出问题，而在国内外，相关政策法律法规在不断强调数据治理的重要性；② 随着数据量的爆发式增长，中心化的云服务器在广泛收集、实时处理海量数据的过程中，不可避免地会遭遇带宽负载、网络时延、应用成本等压力，这在相当程度上限制了服务的实时性、影响了用户体验；③ 大量的端侧设备在离线状态下功能受到限制，也易受中心服务器故障的影响，这在很大程度上降低了系统的稳定性和可靠性。

端云协同智能计算体系（见图1）可有效应对上述挑战，融合了端侧的实时感知响应能力、云侧的大规模计算存储能力，为多方（尤其是多终端）

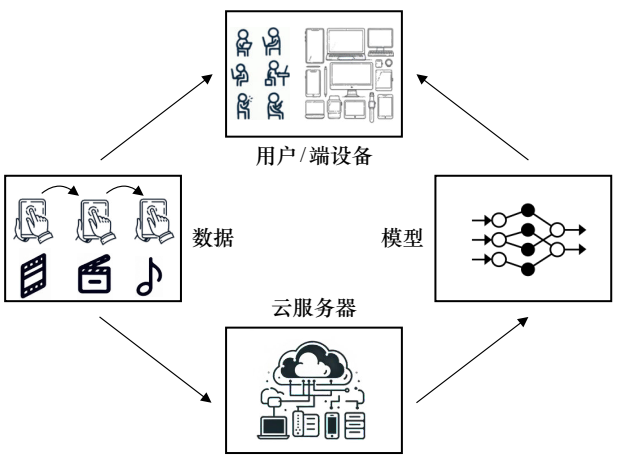

图1　端云协同智能计算的基本要素

数据共享与融合应用提供了良好的解决方案。端云协同架构的主要特征体现为：将端侧、云侧的计算资源进行优势互补，在云端训练大模型并将模型能力输出至端上的小模型，进而由端模型负责实际的推理并向云模型反馈算法与执行的成效，形成“有机循环”的能力增强体系，实现数据资源的高效利用并提升服务的鲁棒性。

端云协同智能计算体系的应用价值表现在：① 能够在保护用户隐私的前提下，显著提高数据资源的利用率，支持实现多方数据共享与融合应用，推动诸多领域的技术研究和应用创新；② 端云协同能够显著提高系统的实时响应能力和服务的鲁棒性，为具有高的实时性和可靠性要求的应用场景提供关键解决方案；③ 端云协同架构提供灵活且可扩展的平台，使云端、端侧的资源得到充分利用和优化配置，推动云计算、边缘计算技术的融合与创新；④ 端云协同应用驱动 AI、物联网（IoT）等领域的技术创新和业务模式探索，为相应领域的中长期演进提升筑牢基础。

三、端云协同智能计算的关键问题

（一）计算资源限制条件下端侧学习效率难优化问题

在端云协同架构中，受计算资源和存储空间的限制，优化端设备的学习效率成为核心问题，需要针对在端环境下执行的特定模型训练和推断任务，利用有限的计算资源来显著提升任务的实时性、效率和可行性。

学术界和工业界普遍采用模型压缩技术来提高边缘设备上的学习效率。例如，模型裁剪[8]、参数量化[9]技术广泛用于减小模型规模，能够提高计算速度并降低边缘设备的计算负担；模型裁剪可移除不必要的模型参数或层以简化模型，参数量化则减少参数所需的位数来降低模型的存储和计算需求。这些方法有效缓解了边缘设备上模型部署的约束条件（如计算能力、存储空间的限制）。

边缘设备硬件性能的持续进步也为边缘学习效率优化提供了直接支持，在一定程度上缓解了部署于边缘设备上的模型规模和结构方面的限制。通过软硬件协同，提高边缘设备上模型的训练推断速度和整体运算效率，使 AI 模型在端计算环境下更高效地执行。

（二）端侧少样本学习条件下的过拟合问题

在端云协同系统中，边缘设备通常只能收集有限的数据量且数据的标签注释质量往往不高，因而端侧的少样本学习问题及其引发的过拟合现象成为挑战。考虑到用户隐私保护的需求，数据直接共享受到限制，而不同边缘设备根据其类型、地理位置、传感器特性收集的数据存在显著差异性，将加剧数据的稀疏性问题。解决这一问题有两种典型策略：① 联邦学习[10]允许在不共享原始用户数据的前提下合并不同设备上的模型更新，据此实现模型优化，能够降低数据传输风险，在保护用户隐私的同时提高全局模型的性能；② 隐私保护[11]条件下的数据共享，应用差分隐私等技术维护用户数据隐私并引入特定的噪声干扰，使数据中提取的统计信息不会泄露个人敏感信息，在安全的基础上实现一定程度的数据共享。

（三）差异化端侧需求和场景下的模型定制化

端模型个性化旨在为各类边缘设备以及用户提供定制化的模型，满足特定的需求和偏好（见图 2）。在端侧，不同的用户往往展示出独特的行为模式，导致数据分布在用户之间存在显著差异。对于个别用户，人的数据分布与全局数据通常差异较大[12]。因此，若使用基于全局数据分布训练的模型来处理个别用户的数据，推理的准确性可能会显著下降。以推荐系统为例，每位用户都有自己独特的习惯和兴趣，如果用不同的数据分布训练出的模型来服务所有用户，则每个人的体验难言良好[13]。

可行的策略之一是先在云端训练出通用模型，随后在边缘设备上对其进行适用性调整，从而形成匹配特定设备或应用环境的个性化模型[14]。对模型参数进行精细调整，可以提高模型性能，确保在边

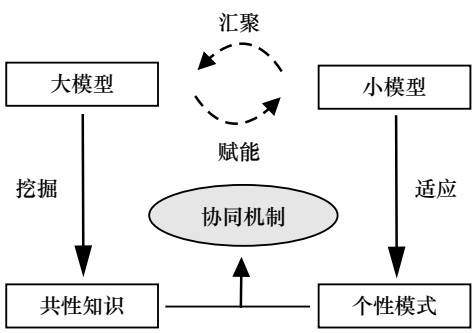

图 2　一种典型的端云协同智能计算模式

缘设备上的高效运行。此外，元学习技术能够较好地初始化模型参数，使模型更加灵活地适应新任务或新设备。通过一系列任务的训练，模型学习了更为通用的特征表达，在面对新任务时可表现出更好的适应性。在端云协同模式下，模型先在云端接受元学习训练，之后可更迅速地适应各种边缘设备[15]。

（四）端云数据分布异构条件下的虚假关联挑战

在云端与边缘设备（端）之间以及不同的边缘设备（端）之间，普遍存在数据异构性，对许多关键问题产生了影响，如让云端训练的模型有效泛化到数据分布各异的边缘设备上存在的泛化性问题。以联邦学习为例，端云协作算法在面对数据异构情况时易出现“客户漂移”现象，即各个独立更新的客户端模型，其参数逐渐偏离全局模型参数，导致系统收敛不稳定或收敛速度下降[16]。

异构数据上训练的模型容易吸收不存在因果关系的数据以及特征与标签之间的虚假关联信息。虚假关联对端云协同智能计算构成的负面影响有：① 干扰云模型在广泛端环境中的泛化能力不强，使云模型难以适应不同的应用场景和数据分布；② 抑制模型汇聚过程中偏见知识与稳定知识的有效分离，降低模型的可靠性和公平性；③ 劣化端云协同推断中信息知识传递的鲁棒性，可能导致模型推断结果不稳定并累积误差。

（五）端云通信开销与计算效率的平衡问题

在将模型部署到真实场景时，通信预算通常有限，因而通信成本成为需要考虑的重要因素之一。在实际场景中，数据从边缘设备传输到云端需要消耗昂贵的带宽费用；其他的通信成本包括云端数据存储和处理服务的费用，在大规模部署时相应占比更为明显。在联邦学习中，大量设备都将本地更新发送到中央服务器，此时通信带宽成为主要瓶颈。针对端云通信开销，可行的应对策略有：① 采用数据压缩和筛选技术来降低需要传输的数据量，着重传输应用中的关键数据以减小通信成本和延迟；② 将部分计算任务卸载至本地，运用端侧计算能力以减少对云的依赖，从而降低数据传输量、减少网络操作延迟、提高传输效率，海量数据在本地进行处理、分析和执行能够提高数据交互的安全性及效率[17]。

四、端云协同智能计算的主流方法

（一）端智能计算方法的基石：高效计算硬件

经过一段时期的发展和积累，云计算市场规模显现[18~20]，也涌现出了一批知名云计算厂商。与此同时，IoT快速发展，带来边缘到数据中心的数据传输量爆发式增长[21]。端计算可以减少数据上传，从而降低对网络条件的依赖，兼顾保护用户隐私[20]。因此，IoT的发展推动了具有分散式计算范式特征的端计算的出现，也使端计算成为减少传输延迟的通用解决方案[21]。端计算的主要挑战在于计算、存储、能源资源条件是有限的，因而在边缘设备上执行复杂的AI模型变得困难[20]。例如，算法在模型构建过程中经常占用大量的内存，才能存储模型参数和其他辅助变量；查询模型参数并进行处理和推断，耗时又耗能[22]，导致功耗、散热性能等成为应用瓶颈。

目前，用来执行特定智能任务（如图像处理、机器学习模型推断）的智能计算硬件[20]主要有3类，目标都是尽量满足端侧AI应用的需求。① 视频处理单元（VPU）是一种新兴的微处理器类别，专为边缘AI设计，能够高效执行边缘视觉工作负载，同时在电源效率、计算性能之间达到良好平衡。② 端侧张量处理器（Edge TPU）由美国谷歌公司率先研发，用于加速边缘设备上的机器学习推断任务，可以运行MobileNets、MobileNets SSD、Inception、TensorFlow Lite等卷积神经网络模型，已在端侧检测、分割任务中获得实际应用。③ 移动图形处理器（GPU）可用于并行优化边缘计算，通常要求高度的能效比、较小的物理尺寸、优化的图形管线、机器学习AI加速能力等。

在移动平台上，端智能计算技术、高效计算硬件取得了显著进展，促成了以虚拟助手、个性化推荐系统为代表的新兴智能应用。国内外的主要信息技术企业都推出了相应试点项目，以展示端侧AI服务的实用性和效果，涉及的应用方向有实时视频分析[23]、农业认知辅助[24]、智能家居[25]、工业IoT[24]等。

（二）以端为中心的协同计算

以端为中心的协同计算作为新兴的计算模式，侧重云辅助下的端侧设备（如移动设备、IoT设备

等）计算和数据处理；网络边缘产生的数据量急剧增长，促进去中心化计算模式的发展[21]。例如，2021 年边缘网络产生的数据量约为 850 ZB，而全球数据中心的流量仅为 20.6 ZB[20]。

以端为中心的协同计算通常设置辅助云模型，结合边缘端的分散优势进行个性化，其中的模型拆分将部分推理从云端卸载到边缘或从边缘卸载到云端。这一过程通常涉及将模型分割为部署在云端、部署在边缘端两部分，相应策略可称为“分割部署”[26~29]。在执行推理任务时，云侧模块通常负责复杂任务的计算，过程中的执行结果可作为端侧模型开展预测的必要输入。由于端侧模型负责最终的预测，因而“分割部署”多演变为以端侧为中心的端云协同智能计算架构，以充分发挥端侧计算的隐私保护能力，同时减少数据传输到云端所需的带宽。自动分割解决方案[30]可自动将深度神经网络模型分为用于边缘计算、云端计算的两部分，基于拍卖机制的边缘计算方法[31,32]顾及边缘容量以更有效地分配资源。

针对端侧服务个性化的需求，在云侧计算辅助的端侧模型个性化学习方面进行了较多探索。基于协作学习方法[33]的用户位置预测，为每个设备建立个性化模型，允许云和边缘共同学习；云模型作为全局聚合器，从多个边缘模型中提取知识。在协作框架[34]下，云侧卷积神经网络（CNN）为每个本地的边缘 CNN 提供软监督以便开展边缘训练，再由边缘 CNN 执行与视觉输入交互的实时推理。在端云快速个性化框架[35]中，引入超网络技术建立数据分布到模型参数的映射关系，根据单个设备的实时点击序列得到与设备当前数据分布相关的模型权重；端模型在无需任何反向传播的条件下实现大规模的实时个性化，有效解决了端模型实时泛化问题。在设备与云协同的模型个性化框架[36]中，每个用户从云端检索出与自身本地数据分布相似的外部样本并将之与本地数据结合，再进行设备上的增量训练，实现模型个性化并降低过拟合的风险。

在以端为中心的协同计算中，端侧数据和计算资源有限，通常需要请求云侧资源支持训练及推理。端 – 云通信收益计算模型[37]能够解决频繁请求带来的通信开销问题，部署在设备上并以较低的资源消耗来计算端 – 云通信收益，使端模型失效后无需重新泛化。云边协同的边缘工作负载预测框架[38]分为两个阶段：在云端进行粗粒度的空间 – 时间预测，捕捉不同边缘节点之间的相关性；在每个边缘节点进行细粒度的空间 – 时间预测，捕捉同一节点内不同虚拟机之间的相关性。相应框架基于多视图和协同预测，提高了预测准确率并降低了带宽消耗。

目前，以端为中心的协同计算在工业界得到广泛应用，如无人机在云端执行复杂的图像识别任务，而在端上进行实时导航和避障[39]。游戏系统则面临着在云计算和边缘计算之间划分工作的挑战[40~42]；在训练和推理过程中，传输潜在表示或留下高强度但对延迟不敏感的训练任务，维持边缘和云之间的交互反馈，可为用户带来更多的创意和自由度，也为元宇宙[43,44]等先进游戏系统创造了可能性[45]。

（三）以云为中心的协同计算

联邦学习属于分布式机器学习方法，允许用户端保留原始数据，而只将本地模型参数发送到模型管理器，然后在不共享原始数据的情况下进行模型聚合以及进一步的训练[46]。在端云协同架构下，联邦学习被视为未来网络服务的关键基础设施，能够保护用户数据隐私，同时充分发挥边缘数据的应用潜能[46]，尽管应用前景广阔，但面临着异质性、攻击防御、个性化方面的挑战。

在联邦学习中，异质性属于广泛认知层面的挑战，又可细分为 3 类。① 统计异质性。不同的客户端可能拥有非独立同分布、不平衡分布等形式的数据，这种数据异质性会产生客户端模型漂移问题，进而影响模型的性能[46]。② 模型异质性。每个客户端可能有不同的任务和特定的要求，都希望独立设计本地模型，这就在异质参与者之间构成知识转移障碍，导致无法应用通用模型进行聚合或梯度操作[47]。③ 设备异质性。不同参与者的设备，其存储和计算能力可能存在差异，将导致一些参与节点出现错误和失活，已有一些在不同阶段解决此类问题的方法[48]。

在联邦学习中，系统可能面临各种安全和隐私威胁。根据机器学习的训练和预测阶段，可将攻击方法分为投毒攻击、推断攻击两类：前者指攻击者可能提供错误的数据、恶意修改模型参数，试图干扰系统学习过程[49]；后者指攻击者可能从模型更新

中推断出其他参与者的敏感信息，据此进行不合法应用[50]。相应地，目前已有多种可用的防御机制[51]。设计可抵抗不同攻击、更加安全且稳定的联邦学习系统，是未来发展方向[52]。

个性化旨在解决传统联邦学习模型的局限性，即通常仅为所有客户端提供统一的共享模型，而无法考虑到客户端的个体差异。当客户端的本地数据分布不均匀时，距离最近的个性化联邦学习可训练出顾及个体差异的模型[53,54]。此外，可以只训练出部分模型，共享参数和个本地参数可以在设备上同时或交替更新[55]。联邦个性化不仅提高了全局模型的性能，还可改善客户端模型的性能，将激励更多的用户参与联邦训练。

（四）端云双向协同计算

在端云双向协同计算中，端和云的模型各自有独立的优化目标，但在训练或推理中会进行双向反馈；应优化资源分配和计算能力，在数据隐私、实时性、系统负载等方面进行均衡。相较以端或云为中心的协同方法，双向协同的计算架构设计通常更显复杂，有待深入研究。

国内“产学研”联合构建的“洛犀”端云协同平台，应用了端云协同推荐的“慢–快”学习机制[56]：慢速组件（云模型）传递辅助的潜在表示，支持快速组件（边缘模型）进行预测；快速组件也将从实时暴露项目中获得的反馈传递给慢速组件，以更好捕捉用户兴趣信息。这种机制类似于人类认知中的系统Ⅰ和系统Ⅱ的作用[57,58]，即系统Ⅱ进行缓慢但全面的推理，而系统Ⅰ快速进行准确识别，两者交互则及时交换先验/领域信息以适应不同的环境需求。

个性–共性发现与协同学习算法[59]在端侧和云侧进行独立的模型构建及训练，也在训练与推理阶段进行交互和反馈。云模型结合因果发现方法，可挖掘海量端设备的共性模式，实现知识的抽象归纳；以共性知识为基础，端侧进一步解耦个体特性知识，实现因果共性模式、个性化模式的自适应互补增强与协同决策推断。该算法考虑了海量端设备的共性特点，支持针对个体特性的高度定制化，在减少计算负担、降低通信成本、提高协同推断精度等方面表现出优势。

云端和设备协同的连续域自适应方法[60]采用具有不确定性的采样策略，从设备端筛选出最重要的样本并发送到云端；在云端采用视觉提示进行知识迁移，将云模型的泛化能力传递给设备模型。

效率优化是端云双向协同的又一重要目标，在保障数据隐私、减轻系统负载的同时，提高计算任务的实时性和准确性。优化数据的流动和处理过程，在保证响应速度和准确度的基础上，合理降低通信成本和计算负担。大–小模型因果协同推断算法[61]用于实时监测海量异构端设备环境的动态变化，据此开展端云协同推断的动态规划。该算法结合因果增益模型，可准确评估端云模块化网络的双向调度效果，减少端云模块中的无效调度，从而降低通信和计算开销，最大化计算资源受限条件下的协同推断收益，为解决端云协同智能计算中的动态资源分配及优化问题提供了一种有效且可量化的方法。

在设备与云协作学习框架（DC-CCL）[62]中，位于云端的大型视觉模型可利用移动设备端的样本进行学习，大型云模型、轻量端模型共享底层特征编码器，云侧训练共享编码器和云端子模型；通过知识蒸馏技术训练控制模型并模拟云端子模型，在移动设备和云端上协作训练共享的底层子模型，促进端云异构模型的知识迁移和共享。

端云双向协同计算已在工业应用中取得重要进展。率先研制了产业级端云协同学习通用系统[63]，覆盖研发期、部署期、运行期，支持300多种学习任务，日均调用超过1×10^{11}次；构建了以端云协同智能计算架构为核心，包含个性化信息流推荐、多模态内容理解、三维重建及实时渲染在内的新技术体系，开源了高性能、轻量级的深度学习端侧推理引擎。

（五）可信端云协同智能计算

在端云协同智能计算的复杂生态系统中，可信端云协同是重要的研究课题。端设备和云服务器在数据、模型、算力上存在明显的异构性，为协同计算带来了应用复杂性，突出表现在模型迁移、去偏泛化、多端公平性、异构协同等。针对端到端、端到云之间的数据分布和特征空间异构问题，迁移学习方法可提高模型在不同环境中的适应性，成为端云之间知识迁移的有力工具。在去偏泛化、多端公平性方面，因果理论提供了分析和解决方案框架，

如用于消除模型偏见和混杂效应的因果推断方法。可信端云协同是多维度、多层次、高复杂度、综合性的问题，需要运用多种计算模型、算法、理论来进行深入研究。

在去偏泛化、多端公平性等问题研究中，因果理论因其应用价值逐步明确而成为解决端云协同相关复杂问题的关键工具。基于因果后门调整的多任务大规模预训练算法[64]、基于混淆因子解耦的多场景泛化算法[65]，能够解决端任务异构、端场景异构等对云模型向端迁移泛化的不利影响。因果泛化算法的核心理念在于，深入分析端云协同训练闭环中数据与特征的关联性及依赖性，揭示虚假关联的生成机制及其结构，解耦混淆因子；使用因果后门调整技术消除混淆因子对变量间关联效应预估的影响，捕捉数据、特征、预测之间的可泛化真实因果效应。以因果效应作为模型学习的指导因素，构建云模型向端迁移中的因果抗偏差学习框架，缓解云–端数据非独立同分布条件下的模型迁移问题，支持模型在未知和差异化的海量端设备环境下稳定开展预测。

端向云因果蒸馏算法[66]用于解决不同端数据偏差、训练偏差等条件下个性化端模型向云汇聚学习的公平性问题。引入因果推断的理论框架，蒸馏提取异构端模型中稳定不变的因果知识，据此判定约束模型汇聚学习对特征、标签关联的重要性，设计正则项缓解云侧汇聚学习偏向特定端群体的不足，增强面向长尾端用户的服务性能和整体公平性。

可信端云协同智能计算的代表性应用场景是推荐系统。推荐系统有助于用户快速查找和探索信息，在电子商务、微视频门户、社交媒体网站等在线应用中获得了越来越多的应用。尽管如此，推荐系统仍然存在用户导向的偏见/公平性问题，也受隐私泄露、响应高延迟等的困扰。

五、端云协同智能计算的垂直领域应用

端云协同智能计算凭借“云侧计算资源丰富、端侧贴近用户和数据”的独特优势，在增强端侧推理能力的同时，提升了云侧模型的泛化能力，为用户提供了定制化解决方案以满足特定业务需求。目前，端云协同智能计算在推荐系统、自动驾驶、安防系统、教育等实际生活场景中得到了广泛应用。

（一）推荐系统

推荐系统在电子商务、社交媒体、音乐、电影、新闻等在线平台中得到广泛应用，基于用户的历史数据和关联信息，预测用户的兴趣和喜好并向其推荐可能感兴趣的商品。推荐系统仍面临一些挑战，如用户不公平性、端设备存储、网络传输延迟、隐私泄露等。以云计算为主的推荐系统倾向于关注交互和消费比较多的用户，而忽视交互较少的大多数用户群体[45,46,67]，导致大多数用户无法得到有效推荐。每个端设备都依赖云端进行模型推理，带来频繁的网络传输需求。

端云协同有助于解决这些问题。云侧将预先训练好的模型通过网络传输到每个端设备，用户可以利用端设备有限的计算资源对模型进行本地微调，从而得到个性化模型，减小网络传输导致的时延[11]。云端也可利用用户的个性特征训练个性化模型[35]，将训练好的模型传输给每个用户，用户在本地直接进行推理。对于特定的用户端设备，云侧可根据设备的存储空间选择合适的模型，采用模型压缩等方法进一步减小网络传输的延迟，以改善用户体验。随着联邦学习的发展和应用，用户可在不上传自身数据的情况下更新云端模型，防止隐私泄露和潜在的安全问题[68,69]。

（二）自动驾驶

作为正在发展中的先进车辆技术，自动驾驶集成传感器、计算机视觉、AI、控制系统，使车辆能够在道路上自动行驶并执行驾驶任务，相关过程完全由机器控制而无需人类驾驶员的直接干预。自动驾驶也是边缘计算发展阶段的重要应用，对室外覆盖和移动性提出了严苛要求，因而快速的端云信息传输成为必要条件。

第五代移动通信（5G）网络的高带宽、低时延、广连接特性，高度契合自动驾驶需求，成为自动驾驶技术的重要支持。自动驾驶车辆配备边缘计算系统，集成定位、感知、规划、控制等实时功能模块。边缘计算系统使车辆可在本地高效进行实时数据处理和决策[70]，显著提升行驶安全性和运行效能。端云协同学习技术将进一步提升自动驾驶安全性、功能性、隐私性方面的能力。自动驾驶车辆的位置和行为信息中包含敏感信息，不应被共享[71]，为此较多采用数据匿名化、加密数据、增强集中式

训练等方法来保护用户的隐私。

（三）安防系统

安防系统用于监测和保护特定区域的安全，传统上主要有监控摄像头、入侵检测设备、门禁系统等。智能手机、平板电脑、智能家居设备等端设备具备计算和通信能力，云平台可提供大规模数据存储、分析、处理能力；将端设备和云平台结合起来，执行安防监控、数据处理、管理协同等工作，将形成功能强大、更为智能的安防解决方案，进一步提高安全性和响应能力。

为了更好利用设备端的样本数据、充分发挥移动设备贴近用户和数据源的优势，将安防设备上的数据上传到云端并用于辅助云端模型的训练成为常见做法，但存在隐私泄露风险。为此，在安防系统中引入跨设备联邦学习[7]，以保护协同训练过程中的隐私不被泄露，但在实施中需要在设备端部署整个模型。随着深度学习方法的进步，VGG19[72]、ResNet152[73]等视觉模型的参数越来越多，对设备端的存储空间提出了更高要求，加大了协同训练难度。为此，DC-CCL[62]引入一种轻量级模型，基于知识蒸馏方式模拟大型云端子模型；相应轻量级模型可部署到移动设备上，用于控制小型子模型的优化方向。DC-CCL 支持在移动设备上部署模型，规避了存储空间不足的问题，仍可控制子模型的优化方向。

（四）教育

近年来，智能校园成为应用热点，教育模式朝着数字化、智能化方向转变，有望革新校园活动样式[74]。端云协同技术为教育领域数字化、智能化发展创造了机遇，丰富了典型的教育应用，如测验生成、导学问答、非事实内容检测等。学生和教师可以随时随地获取教材、课件、学习工具[75]，知识追踪、多样性控制等技术可用于个性化教学，自动评分、反馈迭代算法可支持错误自纠正和学习效果的精细化管理。端侧个性化小模型、高性能推理训练加速也是关键技术，软硬件协同加速及量化、低精度推理相结合可实现高效计算（见图3）。

智能校园发展，重在以大模型基础建设（含分布式计算、并行训练）为依托。通过人类反馈强化学习，模型将更好适应教育场景。迁移学习、领域适应、跨学科知识整合与交叉应用等是云侧大模型的侧重点，轻量化微调、多源数据集成、教育数据挖掘等保障了相应模型的适用性[76]。相应的挑战主要涉及数据隐私、模型的解释性及可靠性、高效率教育垂直大模型微调与学科大图谱构建。确保端云协同智能计算在教育中的深化应用，切实提升教育质量并保障平等取向（而非加剧教育不平等现象），是有待深入研究的关键问题。

六、端云协同智能计算的未来研究建议

（一）云资源在端模型个性化中的应用策略

探索云端数据资源与端侧模型协同预训练机制，展示云端丰富的数据资源及强大的并行计算能力在端侧模型初始性能提升上的积极作用，增强模型对特定端侧环境或用户偏好的适应性。

设计端侧模型的云端实时更新与优化策略，分析云为端侧模型提供即时更新及优化服务的方式、方法、时机。针对端侧设备遭遇新数据分布或用户行为变化工况，利用云端计算资源进行模型微调和优化，降低端侧设备的计算及存储负担。

开展云端数据资源在端侧模型个性化学习中的应用研究。利用云端数据资源进行分析和挖掘，探寻端侧模型个性化学习能力提升的新途径，如通过全局数据集聚类分析，识别用户或设备的不同类型并针对性定制优化模型。

研究端云协同个性化学习模型的数据安全、模型同步、资源调度等关键问题。针对端侧模型个性化学习的高效性、时效性需求，提出数据安全性、模型同步、云端资源调度等的解决方案，提高端云协同个性化学习模型的实用性。

（二）端云协同多目标优化算法

研究端云协同多目标优化策略，设计相应的多目标优化方法（如帕累托最优解[77]或其他权衡策略），实现模型压缩率、压缩后模型效果、模型泛化性之间的均衡优化。

建立统一框架，结合模型压缩与跨场景泛化双重目标，在通过模型压缩减小模型尺寸及计算量的同时，探讨模型在异构端侧环境下泛化性能的量化及优化策略。

研究适配端侧计算资源的模型压缩方法，在执

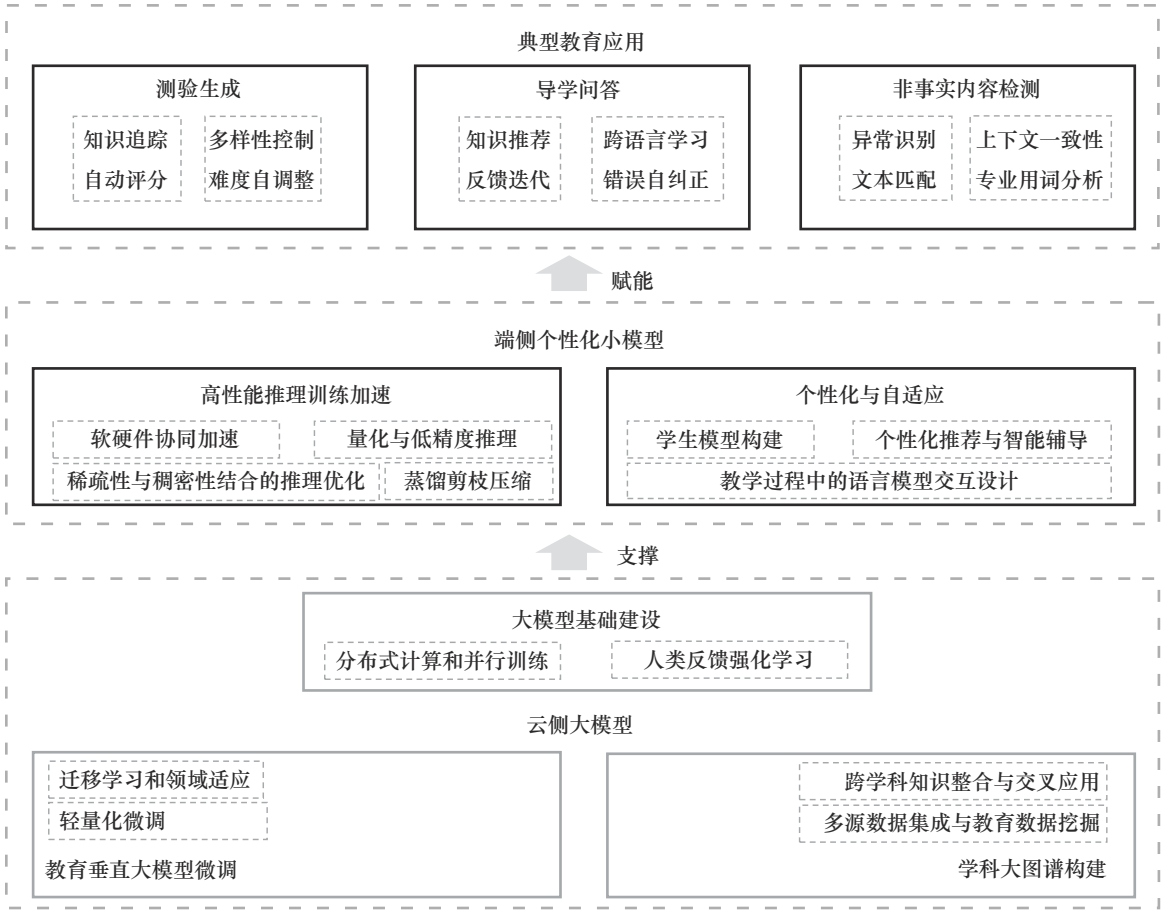

图3　教育垂直领域端云协同技术应用

行模型压缩时，根据端侧设备的计算能力和存储限制自动调整资源配置，定制化压缩后的模型结构。

（三）端–端与云协同计算的优化策略

探索多方智能计算在协同策略中的应用，发挥云服务器作为计算中心和高级调度器的作用，引导端–端设备的高效协作。利用多代理强化学习[78]、图神经网络[79]模拟并优化端到端之间的协作关系。

设计全局与局部调度优化方法，在通信和计算资源受限的条件下，探索全局优化与局部效率之间的均衡策略。基于数据驱动的优先级、延迟敏感度的任务调度算法，实现具有动态自适应特征的资源分配机制。

研发端–端协同计算中的资源管理和分配策略，建立分布式计算资源管理的高效方法，实现相应资源利用、任务执行效率的优化。研究端–端协同计算的数据同步、一致性机制，设计保障数据准确性及一致性的策略和协议。

利益冲突声明

本文作者在此声明彼此之间不存在任何利益冲突或财务冲突。

Received date: November 11, 2023; **Revised date:** December 20, 2023

Corresponding author: Wu Fei is a professor from the College of Computer Science and Technology, Zhejiang University. His major research fields include artificial intelligence, multimedia analysis and retrieval, cross-media computing. E-mail: wufei@zju.edu.cn

Funding project: Scientific and Technological Innovation 2030—"New-Generation Artificial Intelligence" Major Project "Co-evolution and System of Small-Large Device-Cloud Model Collaboration" (2022ZD0119100); Chinese Academy of Engineering project "Strategic Research on New Generation of Artificial Intelligence and Industrial Clusters" (2022-PP-07)

参考文献

[1] 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要 [EB/OL]. (2021-03-13)[2023-11-15]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.
The outline of the Fourteenth Five-year Plan for national economic and social development of the People's Republic of China

and the vision 2035 [EB/OL]. (2021-03-13)[2023-11-15]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[2] 张依琳, 梁玉珠, 尹沐君, 等. 移动边缘计算中计算卸载方案研究综述 [J]. 计算机学报, 2021, 44(12): 2406–2430.
Zhang Y L, Liang Y Z, Yin M J, et al. Survey on the methods of computation offloading in mobile edge computing [J]. Chinese Journal of Computers, 2021, 44(12): 2406–2430.

[3] 贾晓千, 陈刚, 李白冰. 边缘计算在视频侦查中的应用 [J]. 计算机工程与应用, 2020, 56(17): 86–92.
Jia X Q, Chen G, Li B B. Application of edge computing in video investigation [J]. Computer Engineering and Applications, 2020, 56(17): 86–92.

[4] 曹行健, 张志涛, 孙彦赞, 等. 面向智慧交通的图像处理与边缘计算 [J]. 中国图象图形学报, 2022, 27(6): 1743–1767.
Cao X J, Zhang Z T, Sun Y Z, et al. The review of image processing and edge computing for intelligent transportation system [J]. Journal of Image and Graphics, 2022, 27(6): 1743–1767.

[5] 王智, 夏树涛, 毛睿. 基于边缘智能的沉浸式元宇宙关键技术与展望 [J/OL]. 大数据, [2023-11-13]. https://link.cnki.net/urlid/10.1321.G2.20231110.1017.004.
Wang Z, Xia S T, Mao R. Edge-intelligence-based immersive metaverse: Key technologies and prospects [J/OL]. Big Data Research, [2023-11-13]. https://link.cnki.net/urlid/10.1321.G2.20231110.1017.004.

[6] 高祥云, 孟丹, 罗明凯, 等. 支持隐私保护的端云协同训练 [J]. 华东师范大学学报 (自然科学版), 2023 (5): 77–89.
Gao X Y, Meng D, Luo M K, et al. Privacy-preserving cloud-end collaborative training [J]. Journal of East China Normal University (Natural Science), 2023 (5): 77–89.

[7] 杨洋, 况琨, 陈政聿, 等. 基于端云协同体系的预训练大模型及其服务化 [J]. 人工智能, 2022, 9(6): 103–120.
Yang Y, Kuang K, Chen Z Y, et al. Pre-training model based on end-cloud collaboration system and its service [J]. AI-View, 2022, 9(6): 103–120.

[8] Lecun Y, Denker J S, Solla S A, et al. Optimal brain damage [C]. Denver: Advances in Neural Information Processing Systems 2, 1989.

[9] Han S, Mao H Z, Dally W J. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding [EB/OL]. (2015-10-01)[2023-11-15]. https://arxiv.org/abs/1510.00149.

[10] McMahan H B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data [EB/OL]. (2016-02-17)[2023-11-15]. https://arxiv.org/abs/1602.05629.

[11] Li Y G, Yu R, Shahabi C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting [EB/OL]. (2017-07-06)[2023-11-15]. https://arxiv.org/abs/1707.01926.

[12] Li T, Sahu A K, Talwalkar A, et al. Federated learning: Challenges, methods, and future directions [J]. IEEE Signal Processing Magazine, 2020, 37(3): 50–60.

[13] Yao J C, Wang F, Jia K Y, et al. Device-cloud collaborative learning for recommendation [EB/OL]. (2021-04-14)[2023-11-15]. https://arxiv.org/abs/2104.06624.

[14] Sim K C, Zadrazil P, Beaufays F. An investigation into on-device personalization of end-to-end automatic speech recognition models [EB/OL]. (2019-09-14)[2023-11-15]. https://arxiv.org/abs/1909.06678.

[15] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [EB/OL]. (2017-03-09)[2023-11-15]. https://arxiv.org/abs/1703.03400.

[16] Tan A Z, Yu H, Cui L Z, et al. Towards personalized federated learning [J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(12): 9587–9603.

[17] Deng S G, Zhao H L, Fang W J, et al. Edge intelligence: The confluence of edge computing and artificial intelligence [J]. IEEE Internet of Things Journal, 2020, 7(8): 7457–7469.

[18] Pimenidis E, Polatidis N, Mouratidis H. Mobile recommender systems: Identifying the major concepts [J]. Journal of Information Science, 2019, 45(3): 387–397.

[19] Shuja J, Bilal K, Alasmary W, et al. Applying machine learning techniques for caching in next-generation edge networks: A comprehensive survey [J]. Journal of Network and Computer Applications, 2021, 181: 103005.

[20] Cisco global cloud index: Forecast and methodology, 2016—2021 [EB/OL]. (2018-01-15)[2023-11-15]. https://virtualization.network/Resources/Whitepapers/0b75cf2e-0c53-4891-918e-b542a5d364c5_white-paper-c11-738085.pdf.

[21] Le Duc T, Leiva R G, Casari P, et al. Machine learning methods for reliable resource provisioning in edge-cloud computing: A survey [J]. ACM Computing Surveys, 52(5): 94.

[22] Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: A tutorial and survey [J]. Proceedings of the IEEE, 2017, 105(12): 2295–2329.

[23] Ananthanarayanan G, Bahl P, Bodik P, et al. Real-time video analytics: The killer app for edge computing [J]. Computer, 2017, 50(10): 58–67.

[24] Ha K Y, Chen Z, Hu W L, et al. Towards wearable cognitive assistance [C]. Bretton Woods: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, 2014.

[25] Cao J, Xu L Y, Abdallah R, et al. EdgeOS_H: A home operating system for Internet of everything [C]. Atlanta: 2017 IEEE 37th International Conference on Distributed Computing Systems, 2017.

[26] Dey S, Mondal J, Mukherjee A. Offloaded execution of deep learning inference at edge: Challenges and insights [C]. Kyoto: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops, 2019.

[27] Xu Z C, Zhao L Q, Liang W F, et al. Energy-aware inference offloading for DNN-driven applications in mobile edge clouds [J]. IEEE Transactions on Parallel and Distributed Systems, 2021, 32(4): 799–814.

[28] Pacheco R G, Couto R S, Simeone O. Calibration-aided edge inference offloading via adaptive model partitioning of deep neural networks [C]. Montreal: ICC 2021-IEEE International Conference on Communications, 2021.

[29] Pacheco R G, Oliveira F D V R, Couto R S. Early-exit deep neural networks for distorted images: Providing an efficient edge offloading [C]. Madrid: 2021 IEEE Global Communications Conference, 2021.

[30] Banitalebi-Dehkordi A, Vedula N, Pei J, et al. Auto-split: A general

framework of collaborative edge-cloud AI [C]. Singapore: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.

[31] Asheralieva A, Niyato D, Xiong Z H. Auction-and-learning based Lagrange coded computing model for privacy-preserving, secure, and resilient mobile edge computing [J]. IEEE Transactions on Mobile Computing, 2023, 22(2): 744–764.

[32] Shyuan J, Lim W, Xiong Z H, et al. A double auction mechanism for resource allocation in coded vehicular edge computing [J]. IEEE Transactions on Vehicular Technology, 2022, 71(2): 1832–1845.

[33] Lu Y, Shu Y C, Tan X, et al. Collaborative learning between cloud and end devices: An empirical study on location prediction [C]. Washington DC: Proceedings of the 4th ACM/IEEE Symposium on Edge Computing, 2019.

[34] Ding C T, Zhou A, Liu Y X, et al. A cloud-edge collaboration framework for cognitive service [J]. IEEE Transactions on Cloud Computing, 2022, 10(3): 1489–1499.

[35] Lyu Z Q, Zhang W Q, Zhang S Y, et al. DUET: A tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization [C]. New York: Proceedings of the ACM Web Conference, 2023.

[36] Yan Y K, Niu C Y, Gu R J, et al. On-device learning for model personalization with large-scale cloud-coordinated domain adaptation [C]. Washington DC: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022.

[37] Luy Z Q, Chen Z Y, Zhang S Y, et al. IDEAL: Toward high-efficiency device-cloud collaborative and dynamic recommendation system [EB/OL]. (2023-02-14)[2023-11-15]. https://arxiv.org/abs/2302.07335.

[38] Li Y N, Yuan H T, Fu Z, et al. ELASTIC: Edge Workload Forecasting based on Collaborative Cloud-Edge Deep Learning [C]. New York: Proceedings of the ACM Web Conference 2023, 2023.

[39] Zhou Z, Chen X, Li E, et al. Edge intelligence: Paving the last Mile of artificial intelligence with edge computing [J]. Proceedings of the IEEE, 2019, 107(8): 1738–1762.

[40] Nguyen D V, Tran H T T, Thang T C. A delay-aware adaptation framework for cloud gaming under the computation constraint of user devices [C]. Thessaloniki: International Conference on Multimedia Modeling, 2019.

[41] Basiri M, Rasoolzadegan A. Delay-aware resource provisioning for cost-efficient cloud gaming [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(4): 972–983.

[42] Kassir S, de Veciana G, Wang N N, et al. Joint update rate adaptation in multiplayer cloud-edge gaming services: Spatial geometry and performance tradeoffs [C]. New York: Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing, 2021.

[43] Dionisio J D N, Gilbert R. 3D Virtual worlds and the metaverse: Current status and future possibilities [J]. ACM Computing Surveys, 45(3): 34.

[44] Lee L K, Braud T, Zhou P Y, et al. All one needs to know about Metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda [EB/OL]. (2021-10-06)[2023-11-15]. https://arxiv.org/abs/2110.05352.

[45] Jakob N. Usability Engineering [M]. San Francisco: Morgan Kaufmann Publishers Inc., 1994.

[46] Bao G M, Guo P. Federated learning in cloud-edge collaborative architecture: Key technologies, applications and challenges [J]. Journal of Cloud Computing, 2022, 11(1): 94.

[47] Zhou X, Lei X Y, Yang C, et al. Handling data heterogeneity in federated learning via knowledge fusion [EB/OL]. (2022-07-23)[2023-11-15]. https://arxiv.org/abs/2207.11447.

[48] Li T, Sahu AK, Zaheer M, et al. Federated optimization in heterogeneous networks [C]. Austin: Conference on Machine Learning and Systems, 2020.

[49] Ye M, Fang X W, Du B, et al. Heterogeneous federated learning: State-of-the-art and research challenges [J]. ACM Computing Surveys, 56(3): 79.

[50] Sun Z T, Kairous P, Suresh AT, et al. Can you really backdoor federated learning? [EB/OL] (2019-11-18)[2023-11-15]. https://arxiv.org/abs/1911.07963.

[51] Lyu L J, Yu H, Ma X J, et al. Privacy and robustness in federated learning: Attacks and defenses [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1–21.

[52] Chen Y, Gui Y J, Lin H, et al. Federated learning attacks and defenses: A survey [C]. Osaka: 2022 IEEE International Conference on Big Data, 2022.

[53] Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning [C]. London: 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), 2020.

[54] Marfoq O, Neglia G, Vidal R, et al. Personalized federated learning through local memorization [C]. London: Proceedings of the 39th International Conference on Machine Learning, 2022.

[55] Pillutla K, Malik K, Mohamed A, et al. Federated learning with partial model personalization [EB/OL]. (2022-04-18)[2023-11-15]. https://arxiv.org/abs/2204.03809.

[56] Chen Z Y, Yao J C, Wang F, et al. Mc^2-sf: Slow-fast learning for mobile-cloud collaborative recommendation [EB/OL]. (2021-09-25)[2023-11-15]. https://arxiv.org/abs/2109.12314.

[57] Kahneman D. Thinking, fast and slow [M]. London: Penguin Books Ltd., 2011.

[58] Madan K, Ke R N, Goyal A, et al. Fast and slow learning of recurrent independent mechanisms [EB/OL]. (2021-05-18)[2023-11-15]. https://arxiv.org/abs/2105.08710.

[59] Zhang S Y, Feng F L, Kuang K, et al. Personalized latent structure learning for recommendation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10285–10299.

[60] Gan Y L, Pan M J, Zhang R Y, et al. Cloud-device collaborative adaptation to continual changing environments in the real-world [C]. Vancouver: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[61] Qian X F, Xu Y, Lv F Y, et al. Intelligent request strategy design in recommender system [C]. Washington DC: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022.

[62] Ding Y C, Niu C Y, Wu F, et al. DC-CCL: Device-cloud collabora-

tive controlled learning for large vision models [EB/OL]. (2023-03-18)[2023-11-15]. https://arxiv.org/abs/2303.10361.

[63] Lyu C F, Niu C Y, Gu R J, et al. Walle: An end-to-end, general-purpose, and large-scale production system for device-cloud collaborative machine learning [EB/OL]. (2022-05-30)[2023-11-15]. https://arxiv.org/abs/2205.14833.

[64] Zhang S Y, Jiang T, Wang T, et al. DeVLBert: Learning deconfounded visio-linguistic representations [C]. Seattle: Proceedings of the 28th ACM International Conference on Multimedia, 2020.

[65] Zhang S Y, Feng X S, Fan W J, et al. Video audio domain generalization via confounder disentanglement [C]. Washington DC: Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[66] Zhang S Y, Jiang Z Q, Yao J C, et al. Causal distillation for alleviating performance heterogeneity in recommender systems [J]. IEEE Transactions on Knowledge and Data Engineering, 2023: 1–16.

[67] Naghiaei M, Rahmani H A, Deldjoo Y. CPFair: Personalized consumer and producer fairness re-ranking for recommender systems [C]. Madrid: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022.

[68] Huang H L, Wei S, Feng Y H, et al. Active client selection for clustered federated learning [J/OL]. IEEE Transactions on Neural Networks and Learning Systems, [2023-07-28]. https://doi.org/10.1109/TNNLS.2023.3294295.

[69] Yang Q. Federated recommendation systems [C]. Los Angeles: 2019 IEEE International Conference on Big Data, 2019.

[70] Liu S S, Liu L K, Tang J, et al. Edge computing for autonomous driving: Opportunities and challenges [J]. Proceedings of the IEEE, 2019, 107(8): 1697–1716.

[71] Deng Y, Zhang T H, Lou G N, et al. Deep learning-based autonomous driving systems: A survey of attacks and defenses [J]. IEEE Transactions on Industrial Informatics, 2021, 17(12): 7897–7912.

[72] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04)[2023-11-15]. https://arxiv.org/abs/1409.1556.

[73] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C]. Las Vegas: 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[74] Chen Q Y, Wang Z, Su Y, et al. Educational 5G edge computing: Framework and experimental study [J]. Electronics, 2022, 11(17): 2727.

[75] Wang C, Wang D. Managing the integration of teaching resources for college physical education using intelligent edge-cloud computing [J]. Journal of Cloud Computing, 2023, 12(1): 82.

[76] Caines A, Benedetto L, Taslimipoor S, et al. On the application of large language models for language teaching and assessment technology [EB/OL]. (2023-07-17)[2023-11-15]. https://arxiv.org/abs/2307.08393.

[77] Qian C, Yu Y, Zhou Z H. Subset selection by Pareto optimization [C]. Montréal: Advances in neural information processing systems, 2015.

[78] Gronauer S, Diepold K. Multi-agent deep reinforcement learning: A survey [J]. Artificial Intelligence Review, 2022, 55(2): 895–943.

[79] Wu F, Souza A, Zhang T Y, et al. Simplifying graph convolutional networks [C]. Long Beach: Proceedings of the 36th International Conference on Machine Learning, 2019.