



研究与开发

边云协同驱动的沉浸式 Web3D

许娅男¹, 赵桂丹², 冯大权^{1,3,4}, 施文哲^{3,5}, 陆平^{3,5}

(1. 深圳大学电子与信息工程学院, 广东 深圳 518060;

2. 重庆信息通信研究院, 重庆 401336;

3. 移动网络和移动多媒体技术国家重点实验室, 广东 深圳 518055;

4. 广东省哲学社会科学重点实验室 (文化数字化与文化创新发展), 广东 深圳 518060;

5. 中兴通讯股份有限公司, 广东 深圳 518052)

摘要: 以浏览器为载体, Web3D 凭借其天然的跨平台、易访问优势, 在虚拟现实、3D 建模与可视化等数字内容领域展现广泛的应用前景。然而, 研究如何构建多元、共享的沉浸式 Web3D 服务平台尚处于起步阶段。在应对网络环境动态变化、浏览器端渲染性能的限制及未来海量交互需求方面, 如何确保无缝通信、计算和存储服务成为核心挑战。随着移动互联网、5G 通信等前沿技术的推进以及边缘计算的兴起, Web3D 服务模式正逐步从单一的终端或云端渲染方式过渡到边云协作方式, 通过有效整合云端、边缘节点和终端用户资源来满足快速响应、数据速传、实时交互需求。围绕边云协同框架在沉浸式 Web3D 服务中的应用, 探讨这一计算模式如何应对 Web3D 平台搭建中的复杂难题。首先, 分析了面向 Web3D 应用场景的边云协同框架, 以及边云协同在构建沉浸式 Web3D 服务平台时面临的关键挑战。其次, 分别从资源调度、传输与加载策略、预缓存机制, 以及模型轻量化的角度详细讨论了 Web3D 边云协同优化方案。最后, 探讨了开放性研究问题。

关键词: Web3D; 边云协同; 沉浸式交互

中图分类号: TP393

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024193

Edge-cloud collaboration empowered immersive Web3D

XU Yanan¹, ZHAO Guidan², FENG Daquan^{1,3,4}, SHI Wenzhe^{3,5}, LU Ping^{3,5}

1. College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China

2. Chongqing Information and Communication Research Institute, Chongqing 401336, China

3. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China

4. Guangdong Provincial Key Laboratory of Philosophy and Social Sciences, Shenzhen 518060, China

5. ZTE Co., Ltd., Shenzhen 518052, China

Abstract: Leveraging the natural cross-platform accessibility, Web3D has shown wide application prospects in digital

收稿日期: 2024-06-03; 修回日期: 2024-07-17

通信作者: 冯大权, fdquan@szu.edu.cn

基金项目: 国家自然科学基金优秀青年基金项目 (No. 62422112); 国家重点研发计划项目 (No. 2023YFF0612400); 深圳市科技计划项目 (No. KJZD20230923114707016)

Foundation Items: NSFC Excellent Young Scientists Fund (No.62422112), The National Key Research and Development Program of China (No. 2023YFF0612400), Shenzhen Science and Technology Plan Project (No. KJZD20230923114707016)



content fields with a browser as a carrier, including virtual reality, 3D modeling and visualization, to name a few. Nonetheless, constructing a diversified and shared immersive Web3D service platform is still in its infancy. How to ensure seamless communication, computing, and storage services has become a core challenge in dealing with the dynamic changes in network environments, limitations in browser-based rendering capabilities, and the impending surge in interactive demands. The progression of cutting-edge technologies such as mobile Internet and 5G networks, coupled with the emergence of edge computing, has propelled Web3D service modes from solitary terminal or cloud-based rendering toward a collaborative edge-cloud paradigm. Such a paradigm effectively integrates the resources from the remote cloud, edge nodes, and end users, thereby fulfilling the requirements of rapid response, high-speed data transmission, and real-time interaction. Focusing on the application of edge-cloud collaboration frameworks in immersive Web3D services, how this computing mode tackle the intricate issues encountered in establishing Web3D platforms was discussed. Firstly, a Web3D scenario-oriented edge-cloud collaboration framework and the key challenges faced in constructing immersive Web3D service platforms were discussed. Secondly, the optimization strategies for edge-cloud collaboration-driven Web3D were discussed from the perspectives of resource scheduling, transmission and loading, pre-caching schemes, and model lightweighting, respectively. Finally, open research questions were explored.

Key words: Web3D, edge-cloud collaboration, immersive interaction

0 引言

构建虚拟世界或将物理世界的人、物、场景转化为数字3D形态,并实现用户与3D场景的深度交互,已成为数字媒体、在线教育、文化旅游等多个领域发展的重要趋势。然而,传统3D应用通常要求用户在目标操作系统上预先下载并安装特定的应用程序,尤其是扩展现实(extended reality, XR)原生应用,还需考虑XR设备兼容性问题,这进一步提高了3D内容的开发和使用门槛。不同于传统3D服务对于设备或平台兼容性的要求,Web技术的跨平台、易访问、动态交互特性提供了更加便捷、灵活的内容服务方式。通过在Web平台上集成交互式3D技术——Web3D,可允许用户在无须额外安装特定软件或插件的情况下,随时随地访问3D内容并与之交互^[1],这进一步扩大了3D应用的规模。随着Web技术的持续演进与迭代,网页三维图形处理、实时通信与应用性能优化等方面的突破进一步为沉浸式Web3D应用提供了强大的技术支撑,例如支持硬件3D渲染加速的计算机图形库(Web graphics library, WebGL),实现网页端通

信、传输和协作的Web实时通信(Web real-time communication, WebRTC)技术,提升Web应用开发效率和执行速度的WebAssembly编程接口,以及支持更通用渲染计算的Web图形处理器标准(Web graphics processing unit, WebGPU)。这些关键技术的不断革新与Web平台天然优势相结合,极大丰富了用户交互方式,提供不受地理或设备限制的沉浸式视觉体验,已广泛应用于3D建模、虚拟展览、沉浸式社交等多元数字内容领域^[2]。

Web3D体验根据是否配备辅助设备(例如头戴式显示器、手柄、传感器等),可分为裸眼Web3D体验和网页虚拟现实(WebXR)两类。前者以工程建筑领域为例,将Web技术与建筑信息模型(building information modeling, BIM)结合,即WebBIM,在网页端呈现城市建筑与道路交通等模型,实现跨平台的建筑模型在线漫游与消防疏散规划演练等模拟功能,打破了BIM数据可视化的特定操作系统环境约束^[3]。另一方面,随着人们生活方式和社会需求的数字化转变,以XR为代表的沉浸式技术凭借深度虚拟交互体验得到了广泛关注。Web技术与沉浸式技术的融合

为浏览器提供了增强现实（augmented reality, AR）、虚拟现实（virtual reality, VR）等功能接口^[4]。在网页虚拟现实应用中，用户通过接入沉浸式设备并访问网页的统一资源定位符（uniform resource locator, URL）即可随时随地接入虚拟环境，极大降低了沉浸式交互的门槛。万维网联盟的沉浸式 Web 工作组也正在积极推进与 WebXR 相关的技术规范^[5]。

3D 内容交付的核心目标在于实现逼真的视觉效果与实时交互同步，尤其是在基于浏览器的运行环境中，渲染性能与网络传输性能是优化的关键。尽管交互设备的硬件能力不断加强，Web 应用相较于运行在操作系统上、可不受限制访问硬件资源的 3D 应用，受浏览器自身的限制（包括内存约束、JavaScript 性能以及对于底层硬件资源的访问限制等），仍存在一定的性能损耗。此外，Web3D 应用对网络条件提出了更高的要求，当网络不稳定或断开时，网页端需花费更多时间重新加载页面内容。另一方面，移动终端和外接设备的能源限制进一步增加了处理计算密集型任务（例如目标识别、碰撞检测和光照计算）的难度。传统将任务交由远程数据中心处理的解决方案虽然能有效提升计算性能，但是较高的网络时延成为其主要瓶颈^[6]。边缘计算的兴起则为实时计算提供了可行方案，其通过将云资源下沉到更靠近终端用户或数据源的网络边缘，为关键应用提供分布式计算环境和类云服务，从而在减轻核心网压力的同时提升响应速度^[7-8]。在该范式下，终端可以将复杂渲染任务卸载至边缘服务器，同时在边缘节点就近预存热点内容以快速响应其服务范围内的访问请求^[9-10]。然而，边缘资源由于异构性、分布不均及规模限制，其在面对复杂场景精细处理、高质渲染及高并发访问需求时，难以保证多地多用户的快速接入、实时访问^[11]。此外，大规模部署边缘基础设施来补充算力支撑，也意味着高昂的部署成本和较长的部署周期。

利用边缘与远程云的协作计算成为了应对 3D 服务需求快速增长的解决方案。边云协同通过有效整合网络后端和边缘的计算、存储和通信能力，在丰富的计算资源和高响应性之间合理折中，为 Web3D 应用的灵活交付提供了具有前景的解决方案。然而，边云协同下的沉浸式 Web3D 应用在传输、加载和资源协同管理等方面仍存在亟待解决的关键问题。对此，本文围绕边云协同机制如何赋能 Web3D 的高效传输和协作渲染等方面，介绍一种面向 Web3D 应用的边云协同框架，分析该框架在支撑沉浸式服务中所面临的关键挑战，从分布式资源调度、传输与加载策略、动态预缓存机制以及模型轻量化处理角度回顾现有优化方案，同时展望未来的研究方向。

1 Web3D 服务框架

随着 Web3D 应用对沉浸式体验需求的日益增长，传统单一渲染方案已难以满足用户对沉浸式、流畅体验的需求。在边云协同计算环境下，算力和存储资源丰富的远程数据中心提供大规模计算和数据处理能力，而分布在网络边缘的边缘节点分担骨干网的负载压力、就近提供与渲染相关的服务。终端产生的计算密集型或高能耗型任务可以根据系统状态、任务属性、网络条件等因素选择性地卸载到远程云或边缘节点上远程处理，以充分利用网络后端和边缘的计算资源，从而提供灵活的分布式渲染解决方案。对此，本节将讨论驱动 Web3D 的边云协同框架，以及沉浸式服务提供过程中所面临的主要挑战。

1.1 边云协同框架

边云协同利用地理分布的边缘节点就近提供其覆盖范围内的低时延服务，同时由远程云提供海量算力支持。在网络边缘侧，基于终端间协同的通信技术，例如终端直通（device-to-device, D2D）通信技术和对等（peer-to-peer, P2P）网络技术，具备一定硬件性能的终端设备同样可充



当“算力提供者”参与复杂任务的协作处理。例如,在D2D通信技术辅助下,相邻设备之间可以不经基站直接通信,从而提高网络的总体频谱效率并降低后端计算负载压力^[12]。因此,通过终端协同,相邻设备间可以共享计算、通信等资源,实现轻量级任务的高效协作。总的来说,边云协同的核心在于构建终端间、终端与网络间、边缘间以及云与边缘间的多域协同机制,动态聚合分布式通信、计算和存储资源,以提供实时交互响应和高效数据传输能力。边云协同框架可大致分为云层和边缘层,边云协同基本架构如图1所示。其具体描述如下。

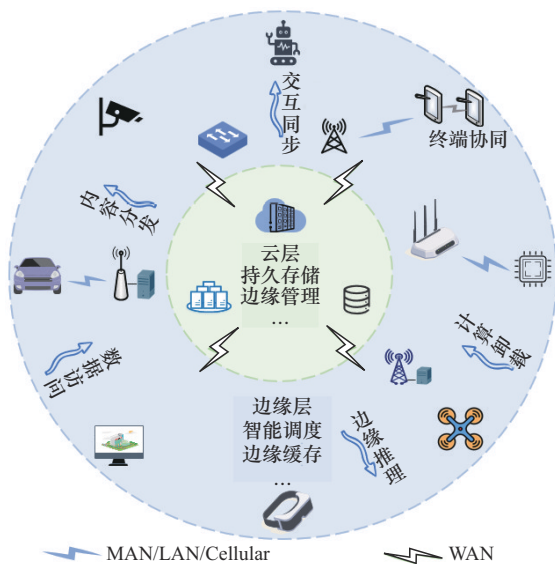


图1 边云协同基本架构

(1) 云层: 云数据中心、服务器集群作为核心计算资源池,承担非实时性的高强度计算与海量数据处理任务,负责数据持久存储与管理,执行全局性能监控以掌握边缘侧的负载和网络条件,并统筹管理边缘节点(包括服务放置、任务调度、资源分配与内容分发等)。

(2) 边缘层: 主要包括具备计算、存储、虚拟化、感知等能力的边缘节点(边缘服务器、边缘网关、路由器、交换机等),以及终端设备(智能手机、计算机、摄像头、智能穿戴设备

等)。边缘服务器作为云的延伸部署,属于中量级基础资源,负责实时监测终端与网络状态,承担与用户地理位置紧密相关的Web3D业务,例如内容分发、光线追踪、渲染等。另一方面,终端设备可根据实时性能状态提供额外的计算能力来处理部分轻量级任务。

在边云协同架构下,边云协同驱动Web3D服务提供方式示例如图2所示。如前所述,Web3D用户无须在终端上预先下载特定应用程序,通过浏览器访问预定义的URL即可启动3D服务请求。首先,用户发出访问请求,全局负载均衡器或者中心调度节点确定用户附近可用的边缘服务器;其次,全局负载均衡器或者中心调度节点分配最优边缘服务器与终端建立连接并响应终端请求。边缘节点间可通过WebRTC技术进行交互与同步,终端用户之间可通过D2D技术建立通信链路以协作计算和消息转发。建立连接后,远程云根据用户请求处理相应3D内容(包括3D模型简化、纹理压缩等轻量化操作),并将所有数据发送到指定边缘服务器来响应后续访问请求。此外,终端设备一般具有一定的计算和存储能力,因此在发送内容访问或任务处理请求时可以在本地负责中量/轻量级计算任务以分担服务器负载、加速计算。对于浏览、编辑类3D内容呈现业务(例如Web-BIM),当视点变化或进行其他交互操作时,终端可在本地计算更新视点后的可视场景列表后再对边缘服务器请求相应内容。需要注意的是,对于WebAR类增强型业务,一般首先由终端自身负责灰度化等图像帧预处理,再将需进行目标识别的帧传至边缘节点进行位姿计算、特征提取等操作,最后接收识别结果并在页面进行渲染和呈现。对于WebVR类业务,终端主要负责实时动作的捕捉、运动和视点等相关数据的上传等操作,再由指定的服务器做后续处理,然后返回相应3D内容或全景视频到终端页面上渲染并呈现。

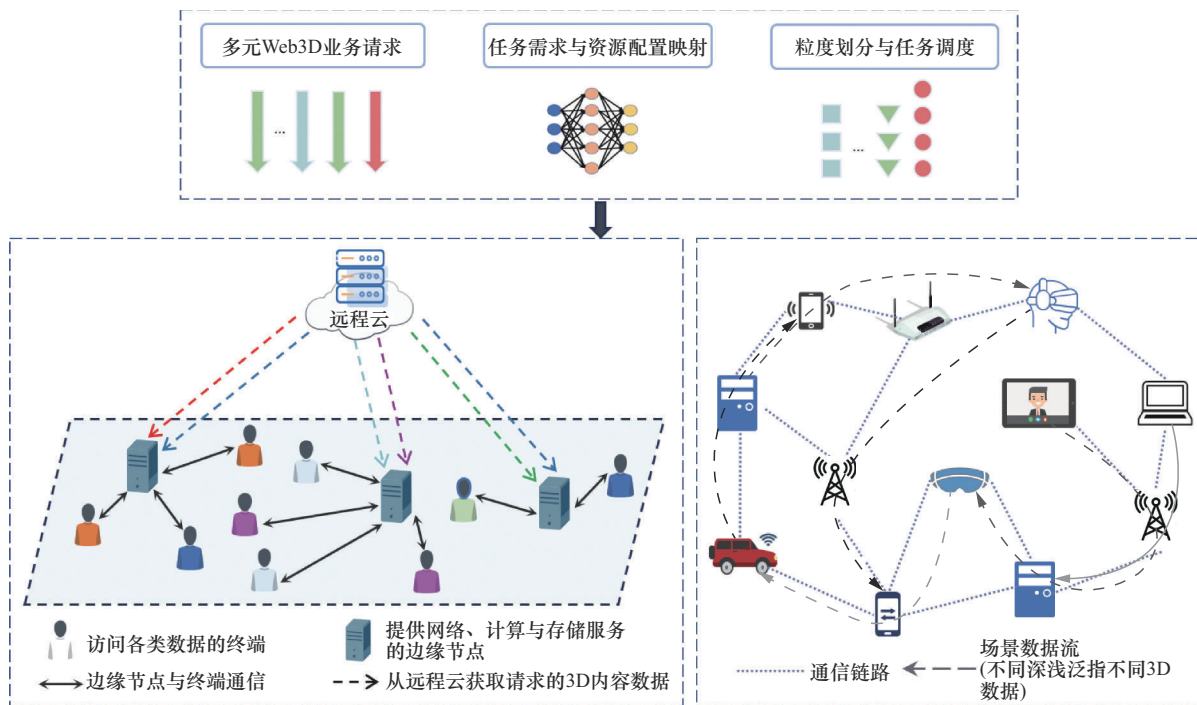


图2 边云协同驱动Web3D服务提供方式示例

1.2 关键挑战

边云协同范式有望为Web3D提供灵活可扩展、按需调度的服务方案。然而,要实现云层与边缘层的高效协作、Web3D数据高效传输与浏览器端快速加载以及Web3D内容的高质量渲染,边云协同驱动的沉浸式Web3D服务平台构建仍面临许多技术难题。具体如下。

(1) Web3D业务的多样性与异构性导致了计算资源需求的极大差异,从简单的3D模型展示到复杂的虚拟现实体验,不同场景下所需的计算能力、存储空间和网络带宽各不相同。特别是大量并发连接和局部高密度交互引起的边缘节点算力需求,给系统的负载均衡能力和响应速度带来了严峻考验。此外,强交互Web3D业务要求至少毫秒级的端到端响应时延才能保障基本的低用户体验;而移动性较强的Web3D应用进一步对服务连续性、流畅体验提出了更高的要求。当用户快速移动时,如何在边云协作环境下各节点间保持无缝的服务切换,避免断连或重连时延,也变得

尤为困难。

(2) Web3D业务涉及多模态数据(例如3D模型、几何与纹理信息、多维感知信号),其传输不仅包含静态的视觉元素,还有动态变化的交互性信息,例如用户操作的实时反馈、物理环境的实时变化等。因此,传统的多媒体(例如音视频)传输方案与自适应码率传输机制在面对信息维度更丰富、结构更复杂的Web3D内容请求与传输时适用性受到限制。特别是当采用D2D辅助通信时,需要在多个设备之间保持一致的状态更新,同时还要考虑终端能耗限制、网络拓扑动态性和网络资源优化,这进一步增加了交互同步与协同传输规划的难度。

(3) 在边云协作的渲染架构中,为了减轻终端设备的计算负担,复杂的3D物理模拟和高分辨率渲染任务通常被分配到边缘节点或云端进行处理,以一定的网络时延来抵消计算开销。对于需要即时反馈的场景(例如VR游戏或远程培训模拟),用户动作与屏幕反馈之间的时间差需控



制在毫秒级内（例如小于 20 ms）。然而在网络环境不稳定时，远程渲染和结果传输过程中可能出现数据包丢失、重传等情况，导致更多的网络时延，并影响渲染效果的一致性和流畅度。

综上所述，边云协同在赋能 Web3D 服务的过程中，不仅要考虑业务多样性和异构性带来的资源需求波动，还需克服多模态数据传输的复杂性，并应对渲染与响应速度的严格要求。为了满足多元业务的动态需求，边云协同要充分考量访问内容、边缘资源使用情况、复杂网络环境进行合理的资源调度决策。决策优化目标包括内容初始加载时间最小化、终端能耗最小化、网络/计算/存储资源效用最大化等。为了克服云中心化模式的局限、减轻边缘节点交互同步压力，需设计动态、高效的协同传输规划方案。同时，综合传输能耗、网络传输时延、频谱资源、丢包率等多维度指标来设计高效的自适应传输机制能进一步确保 Web3D 服务在各种环境下实现稳定、流畅的体验。而在自适应传输基础上，浏览器端还需考虑渐进式加载等内容加载方式以进一步降低用户等待请求内容的时间。此外，使用边缘缓存技术来提升内容加载速度是一种可行方案，包括缓存放置和内容预测（例如估计 VR 视频内容未来流行度并预先缓存内容）。预测模型需处理更高维度的数据、考虑用户交互的非线性特性要求，并解决同步问题。

2 Web3D 边云协同优化方案

Web3D 边云协同优化研究近年来已取得了显著进展，本文针对边云协同驱动的 Web3D 平台搭建面临的关键挑战，对现有工作进行了整体回顾，重点聚焦以下几方面的优化方法：

（1）分布式资源调度优化，以提高 Web3D 任务的处理速度和异构资源的利用效率；

（2）传输与加载策略优化，以应对网络波动下稳定响应的需求、加速数据传输和接收加载流程；

（3）动态预缓存机制优化，就近预先缓存用户感兴趣的内容以提高用户响应速度。

此外，本节还回顾了可以提升数据分发传输效率和内容加载速度的 3D 模型轻量化处理优化研究。

2.1 分布式资源调度优化

在边云协同框架下，网络边缘侧具有高度动态性、异构性和不确定性，有效的分布式资源调度是满足 Web3D 业务在多层级间的动态需求的核心，这要求合理分配多节点异构资源以及高效进行计算卸载决策，确保在提升系统性能（例如端到端时延）的同时实现执行成本（例如能耗）最小化。

一方面，借助计算卸载技术，性能受限终端可以将计算密集型或高能耗型任务发送到资源充足的边缘服务器或云上远程处理。Polys 等^[13]研究了面向视觉障碍者的 Web3D 情境感知技术，并提出基于计算机视觉的端到端服务架构。其中，云基础设施平台上部署了场景解析、字幕生成等机器学习算法模块，用户访问三维场景时可截取屏幕并将截图传输至云上解析图像、生成描述及语音旁白。Li 等^[14]提出了一种包含云、边缘节点和终端的分布式协同渲染服务框架，通过卸载渲染任务到边缘节点就近处理以缓解远程云的负载和传输压力。此外，该框架采用了 JSON（JavaScript 对象表示法）作为交换数据格式，并使用粒度计算任务调度和冗余重用机制，构建轻量渲染环境。对于 AR 应用，其计算任务涵盖了从图像帧的预处理和目标跟踪等轻量运算，以及特征提取和目标识别等计算密集型任务。考虑 AR 设备的能耗预算与硬件资源有限，终端同样可以将复杂计算卸载至云或边缘云进行处理。例如，Zhang 等^[15]提出了边缘计算辅助的移动 AR 框架——EdgeXAR，移动客户端主要负责目标跟踪和内容渲染操作，目标识别任务则卸载到边缘服务器上的多目标图像识别管线处理。终端与边缘服务器

之间基于非阻塞式无连接传输协议通信以应对不可靠网络环境的阻塞问题,并设计了识别与跟踪分离的时延补偿机制以缓解计算卸载带来的额外传输开销与对齐问题。XR类应用常涉及人工智能(artificial intelligence, AI)模型训练与推理,而Web平台的存储与计算能力的约束导致难以在其上部署模型并提供实时推理服务。考虑深度神经网络(deep neural network, DNN)在AR目标识别任务上的出色效果以及Web平台的性能限制,Huang等^[16]提出了面向移动Web平台的DNN分布式协同计算框架,针对多分支DNN结构建立了推理时延与AR设备能耗预测模型,并基于深度强化学习方法设计了满足时延与能耗约束的在线DNN卸载策略。此外,为了提升无线VR渲染性能,Guo等^[17]研究了基于移动边缘计算的自适应VR实时渲染框架,以用户体验(由时延和帧率定义)为目标并同时考虑用户能耗与边缘节点缓存资源的限制,联合优化了VR渲染卸载决策与缓存策略。

另一方面,边云资源的合理分配与灵活服务迁移同样是优化系统性能和成本效益的重要手段。Zhao等^[18]在为用户动态分配云服务器时,综合考量了参与用户的地理分布、网络时延、体验质量、计算资源占比等因素,引入用户活跃度参数来对网络时延进行缩放,并基于遗传算法迭代搜索,提出了能够在降低服务成本的同时提升体验质量的服务器分配方案。为了实现高效的边缘辅助移动WebAR应用,Ren等^[19]提出了分布式边缘系统编排方案,该方案包括基于位置的任务调度机制以应对多服务器协作下网络动态变化和负载不均衡的问题,以及合理的服务迁移方案以处理边缘服务器过载的情况并提高服务提供效率。此外,对于引入D2D通信辅助边缘计算的场景,实现异构终端与边缘服务器各方资源的动态协作是至关重要的。对此,Ren等^[20]构建了5G环境下支持多用户协同的移动WebAR框架——Edge AR X5,优

化了时延、设备传输能耗和公共频谱资源占用,同时还引入了AR设备协作的特征提取方案以合理利用邻近AR设备的空闲资源、缓解服务初始化问题,但未充分考虑资源配置方案。

2.2 传输与加载策略优化

实现终端协同通信动态规划和3D数据的网络自适应传输机制,同时考虑渐进加载和协同渲染3D内容,可以有效保证各种交互环境下对Web3D请求数据的稳定分发、降低浏览器端处理压力并提升用户体验。

面对云中心化瓶颈与边缘云环境下的协同难题,特别是面对多人在线社交应用的实时交互和同步需求,能够根据时变网络拓扑、当下可用资源进行动态通信路径规划是提高通信效率、实现带宽与数据高效共享的关键途径。为保障D2D通信协助移动边缘计算环境下的多用户交互同步与计算效能,Edge AR X5框架设计了基于启发式算法的动态分组通信规划方案、基于自回归差分移动平均模型的运动感知关键帧选取机制。然而,该框架虽然采用分组形式来优化启发式通信规划方案决策时间,算法执行效率还需进一步提升。许昊天^[21]针对中心服务器负载过重、网络传输性能差的情况,设计了一种层次化的P2P分发方案,通过将大型场景细分为多个子场景,设置超级节点负责管理各自子场景的模型信息,用户漫游时利用WebRTC通信技术实现从访问场景对应的超级节点请求所需数据。然而,在网络请求稀疏/负载较轻时,此分发方案相较于传统集中式模式在传输性能上的优势不突出。而当网络不稳定、网络资源动态变化时,自适应传输是确保传输流畅的有效方法。通过动态调整三维模型数据的质量或数据包大小来应对网络波动,可以保障Web3D访问数据在各种网络环境下的传输流畅性。例如,Wu等^[22]建立了交互时延、码率与网络条件之间的关系模型,并基于该模型设计了一种综合考虑游戏内容、网络 and 用户操作等因素的



自适应比特率选择算法来降低交互时延。

在传输数据的同时,如何高效处理并加载这些数据成为提升用户体验的又一关键步骤。浏览器端处理复杂场景加载任务(包括光照贴图、可见性计算等)时采用的传统一次性加载方式往往会导致较长的用户等待时间,甚至出现明显卡顿、浏览器崩溃等情况。对此,渐进加载方式,即划分大规模三维模型并在网页侧加载时逐步显示当前视点下的关键构件,可以有效避免一次性渲染全部3D内容的长时间等待现象,使交互体验更为流畅。李柯等^[23]提出了轻量级云-边-页协同框架,定义了兴趣度来考量WebBIM场景各模型构件对视觉效果的重要性,在此基础上提出了由兴趣度驱动的自适应传输机制来优化WebBIM场景的数据传输与分布式渲染效率,并设计了初始视点选取算法以进一步缩减初始加载时间。为了实现glTF 2.0格式的3D网格模型渐进加载,Lemoine等^[24]设计了基于视场角感知的调度方案以确定场景构件的加载优先级,并提出了一种允许客户端在文件下载过程中逐步渲染场景的网络流式传输机制,从而降低启动等待时间。在综合利用云与边缘节点资源来加载3D内容时还需确保渲染结果的一致性。在云-浏览器协同渲染框架下,Liu等^[25]提出了动态云烘焙方案来权衡全局光照带来的逼真视觉效果与算力开销之间的矛盾。其中,云服务器存储三维场景轻量化预处理后的渐进网格以便后续流式传输,并负责全局照明计算;客户端接收数据后计算环境照明,与云端的全局照明结果进行帧融合(即对应像素值加权平均),同时采用3D Warping技术和基于采样的孔洞填充方法来应对视点变化时协同渲染的帧融合伪影、孔洞问题。然而,对于用户在线编辑的场景,该烘焙方案的全局照明预测性能较差,特别是面对遮挡、光源改变和快速移动等情景。

现有方案在终端协同通信规划、带宽/码率自适应传输、细粒度化的三维场景渐进加载和协同

渲染方面的进展,能够显著改善Web3D应用在复杂网络环境下的性能表现,减少用户等待时间,提升整体体验流畅度。然而,综合考虑实时性要求和高视觉保真度目标时,动态内容处理和适应性策略等方面的深入优化还需进一步研究。

2.3 动态预缓存机制优化

在用户漫游或与Web3D场景进行交互的过程中,对用户未来可能访问的内容或做出的动作进行精准预测并提前缓存至边缘节点上,能进一步降低交互时延、提升应用响应能力。

针对边云协同驱动的WebBIM应用场景,Li等^[3]采用基于密度的聚类算法预测用户移动路径,通过聚类得出热点区域并定义能够反映用户对每个热点区域关注度的区域影响因子(regional impact factor, RIF),当所有区域的RIF值超出给定阈值时,激活路径预测模块并按需传输加载路径上每个视点相较于当前视点的可视增量模型。然而,该方法主要用于用户在3D场景中的漫游,未深入考虑多用户交互和动作预测。对于应用玩家的运动预测,Chen等^[26]研究了分布式虚拟环境下的实体轨迹预测方法。其中,考虑航位推算(dead reckoning, DR)方法在面对预测不准确的频繁回滚操作时,产生过多通信开销的情况,提出了一种基于无监督聚类的DR优化方法来解决虚拟环境中参与者之间的不一致性并缓解通信瓶颈。该方法包括离线训练热点区域图和在线预测算法两个环节:首先,对热点区域聚类并创建热点区域图辅助在线预测;其次,客户端执行的在线预测算法根据热点区域是否可用灵活切换,当识别的热点区域可用时则利用其信息进行精准预测,否则依靠实体运动状态执行常规DR预测。然而,该算法仅构建一个时间步长内的热点区域图且难以适用于动态时变场景,回滚能力需进行优化,对快速变化场景的适应性及预测回滚机制尚有优化空间。

考虑深度学习方法在处理强时空特性数据方

面展现的显著优势,近年研究也着重于采用基于深度学习的内容/交互预测来提高预缓存的效率与准确度。Yang等^[27]设计了基于图神经网络(graph neural network, GNN)的协作缓存策略来应对VR背景内容缓存效率低的问题,通过构建GNN内容请求预测模型来学习VR背景内容的时空变化规律、用户请求与空间位置的关联性,同时还支持邻近节点间共享缓存内容以提高缓存命中率。Kumar等^[28]针对移动边缘计算环境下360°视频流传输面临的回程拥塞和时延问题,将360°视频文件划分为可独立编码的瓦片数据,利用长短期记忆(long short-term memory, LSTM)网络模型预测视频的未来流行度,再训练卷积神经网络识别视频中最适合缓存的关键瓦片,该方法可以减少至少35%的回程数据使用。然而,基于深度图像渲染的预测方法对动态对象的预测性能较差(如伪影)。对此,ZGaming框架^[22]通过划分图像帧中的动态前景和静态背景,并采用LSTM模型从历史帧的三维几何信息来预测前景对象的运动和变形趋势。此外,由于缓存完整内容带来的存储开销过大,还需考量用户偏好、模型构件间的关联性等诸多因素来合理替换缓存内容。ZGaming框架提供了质量驱动的3D块缓存机制,通过分析灯光效果、用户行为和修复算法等因素,为每个游戏帧的分块定义效用函数以量化该块对未来长期预测性能增益的影响,替换低效用值的块,以确保缓存始终存储对预测质量提升最有利的信息。然而,当突发交互请求或时延较高时,LSTM的预测性能会降低。分离前景与背景内容,虽然可以更精准地预测动态对象,但对输入帧的要求较为严格,当连续两帧色差明显或缺乏足够纹理细节时,静态对象可能会被错分到动态前景中。

动态预缓存机制通过前瞻性的内容准备,可以减轻终端存储负担并提高数据访问速度,以便在面对复杂且数据密集的场景中,用户与Web3D

环境的交互也能保持体验的连贯性。然而,面对动态更新的场景内容、用户行为模式以及网络条件的不确定性,当前预缓存方案仍需进一步优化,包括适应用户交互频率和场景时变的在线更新机制、提升预测准确度和响应速度的用户动作、动态内容预测机制。

2.4 模型轻量化处理方法

在边云协同计算环境下,优化资源调度、传输与加载策略,以及动态预缓存,可以有效应对由计算密集型操作和网络不确定性带来的挑战,在多样化运行环境下保持服务的高质量交付。然而,随着3D模型复杂性和规模的增长,沉重计算负担和网络传输压力仍旧是沉浸式Web3D体验的瓶颈。对此,简化模型和数据量实现快速网络传输与渲染的模型轻量化操作,是平衡模型复杂度与快速渲染需求的关键步骤^[29]。

现有轻量化方案涵盖了从模型简化(例如网格)、存储方式优化到数据压缩等多个层面。为了减少首屏时间(即3D模型首次加载到浏览器页面渲染的耗时)和场景漫游带来的计算和存储负荷,许昊天^[21]从复杂模型存储方式角度进行轻量化预处理,通过参数化表示、三角带化来代替复杂网格信息的存储以降低文件大小。为进一步降低传输数据大小,该工作还使用了由谷歌发布的Draco库对轻量化后的网格进行压缩。张韵等^[30]从减少大体量模型三角面片数角度出发,提出曲线近似曲率的求解公式并改进顶点法向量计算从而丰富几何信息,改进边折叠算法存在的局部几何特征损失问题,以提高渲染速率、减少内存占用。一些研究工作还考虑了数字孪生模型所面临的处理和传输效率低问题。Zhou等^[31]侧重网格的局部特性,设计了二次误差度量算法改进传统边折叠方法,提高简化操作后的模型精度。而Leng等^[32]聚焦机械结构的数字孪生模型,在边折叠基础上引入顶点显著性因子提高简化后尖锐边缘等特征的保真度。



除了考虑针对工业生产等情境下对模型精度的严苛需求外,加快3D模型建模流程同样是提升Web3D应用性能的关键。随着渲染算法和硬件性能的提升,3D模型的构建方式日益多样化,不再局限于对技能和人力成本要求高的专业建模软件。如今,3D扫描智能建模、基于深度传感器的点云三维重建以及深度学习驱动的3D形状生成方法,极大放宽了对输入的严格限制并降低了对使用者掌握复杂建模流程、光照渲染等专业知识的硬性要求^[33]。以物体级别三维重建为例,O²-Recon框架^[34]引入了预训练的2D扩散模型来应对物体重建的遮挡问题,能够在较少人工干预下对输入的深度图像(red green blue-depth, RGB-D)视频序列生成高质量的掩模,同时还设计了两阶段训练法来应对符号距离场(signed distance field, SDF)在稀疏视图下的过拟合问题,以准确学习物体的几何结构。BundleSDF^[35]从单目RGB-D视频序列中实现在遮挡、运动等干扰下对未知物体进行三维重建,该框架通过设计关键帧内存池来改进位姿估计存在的跟踪漂移问题,并同步使用神经对象场(neural object field)来学习三维物体形状和外观。另一方面,3D Gaussian Splatting以其突出的实时渲染性能、逼真的场景重建效果以及可编辑特性,成为近年来被广泛关注的三维重建技术。例如,Octree-GS模型^[36]借助细节层次(level of detail, LOD)结构来锚定场景中的3D高斯分布从而对场景内容进行多分辨率分配,并根据视点和场景细节的丰富程度从一组锚点中自适应选择合适的LOD,以保证不同细节层次的渲染质量。然而,对于在线强交互应用而言,现有轻量化和重建方法还需对渲染效率和质量进行优化。

3 开放性研究问题

面对边云协同驱动Web3D平台构建面临的

关键挑战,可以从资源调度、数据传输、内容加载与预缓存以及模型轻量化等方面优化来提升Web3D应用的性能和用户体验。然而,为了构建一个更加智能、高效、适应性强的Web3D生态系统,仍有许多研究问题可以深入探讨。例如,如何更精细地调度资源以适应不断变化的用户需求,或在不牺牲质量的前提下,有效处理多样化的3D数据格式。对此,本节对边云协同驱动沉浸式Web3D的未来研究方向进行讨论。

(1) 多元性能指标与成本开销折中关系

边云平台的异构性与Web3D业务的多元性给服务提供商和用户带来了定制化性能指标与成本控制的难题。例如,AI支撑业务主要关注模型精确度、推理时延及服务收益;3D内容访问则侧重码率、交互响应时间、网络带宽、可靠性;而服务商需平衡服务质量、网络开销、成本控制。面对这些复杂的性能指标和成本控制需求,可以设计性能与成本的综合评估模型,分析不同类型任务的处理性能、执行成本与资源配给之间的动态关系,指导设计具备自主学习和适应能力的任务调度与高效资源调配决策尤为关键。

(2) 边云资源与任务需求精准适配

面对复杂的网络环境和多样化3D内容需求,需要进一步精细化任务调度策略和动态资源调整方案。一方面,需以初始访问时延最小化为目标,根据系统状态、网络条件和业务的时空动态特性,制定访问数据初始放置策略,包括虚拟计算单元在内的分布式资源配置;另一方面,对于由多个具有依赖关系的任务组成的应用而言,可进一步细化任务的卸载粒度以提升并行处理效率。例如,对应用构建有向无环图并分析子任务执行需求,为其制定有效的卸载策略与资源配置方案。

(3) 用户移动性管理与迁移机制

尽管边缘节点在靠近用户的位置提供即时

数据处理能力来加速交互响应,但对于计算密集且具有较强移动性和频繁交互的应用(例如基于位置服务的WebAR应用),维持无缝3D体验仍是一大挑战。对此,一方面需结合时空特征、历史行为数据,实现对用户轨迹的高精度预测以及动态任务迁移策略优化;另一方面,在业务分配初期,同样需要前瞻性考虑未来可能的跨云迁移,综合考虑业务特性、地理位置、网络状况、边缘负载等多方因素,并借助性能与成本的评估模型来制定合理的迁移机制与协同传输方案来减少服务中断和资源重新配置带来的成本。

(4) 多元3D数据格式高效处理

在Web3D应用中无缝集成多种格式数据是提升Web3D通用性的有效途径。以点云为例,密集三维点云数据相比网格模型可以实现更逼真的3D内容表示,但其数据量较大、处理复杂,对于阴影、纹理等属性的提取效果不够理想。且此类格式通常需转化为结构化格式才能在Web平台上有效解析和渲染,这进一步增加了数据处理的复杂性。因此,需深入研究在不稳定的网络条件下实现复杂数据的高效传输,同时简化其在Web端的解析和渲染流程。另一方面,数据传输涉及的3D内容、声触觉等多维度感知信号以及音视频等不同模态数据,在表征形式和执行需求上存在显著差异,但模态内在信息可能存在一定的语义关联性和潜在的冗余信息。在设计传输与资源配置方案时忽略这些属性容易造成不合理的信道资源占用、传输效率低下等问题。因此,需挖掘待传数据内在关联性以优化数据压缩及协同传输方案,以提升Web3D数据传输效率。

4 结束语

Web3D技术推动了沉浸式应用的普及,但受限于浏览器端与集中式计算的瓶颈,难以保证沉浸式视觉体验与超低时延服务。而将边缘分布式

资源与集中云强大算力相结合的边云协同架构为沉浸式Web3D服务提供了新途径,在弥补浏览器端渲染瓶颈的同时缓解核心网的繁重负载与网络传输压力。对此,本文聚焦边云协同框架下Web3D服务优化方案,介绍了面向Web3D的边云协同框架,分析了其在实施过程中面临的关键挑战,并从资源调度、传输与加载、预缓存机制和模型轻量化角度阐述了相应优化方案。最后,探讨了边云协同进一步提升Web3D体验的开放性研究问题。

参考文献:

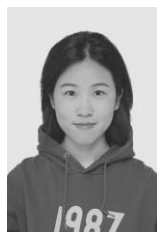
- [1] SEHNAL D, BITTRICH S, DESHPANDE M, et al. Mol* Viewer: modern Web App for 3D visualization and analysis of large biomolecular structures[J]. Nucleic Acids Research, 2021, 49(W1): W431-W437.
- [2] YU G, LIU C, FANG T, et al. A survey of real-time rendering on Web3D application[J]. Virtual Reality & Intelligent Hardware, 2023, 5(5): 379-394.
- [3] LI K, ZHANG Q, ZHAO H T, et al. User interests driven collaborative cloud-edge-browser architecture for WebBIM visualization[C]//Proceedings of the 25th International Conference on 3D Web Technology. New York: ACM Press, 2020: Article 8.
- [4] FLECK P, SCHMALSTIEG D, ARTH C. Creating IoT-ready XR-WebApps with Unity3D[C]//Proceedings of the 25th International Conference on 3D Web Technology. New York: ACM Press, 2020: Article 1.
- [5] MACLINTYRE B, SMITH T F. Thoughts on the future of WebXR and the immersive Web[C]//Proceedings of 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct). Piscataway: IEEE Press, 2018: 338-342.
- [6] LIU Q, HUANG S Q, OPADERE J, et al. An edge network orchestrator for mobile augmented reality[C]//Proceedings of IEEE INFOCOM 2018 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2018: 756-764.
- [7] ZHOU Y, LIU L, WANG L, et al. Service-aware 6G: an intelligent and open network based on the convergence of communication, computing and caching[J]. Digital Communications and Networks, 2020, 6(3): 253-260.
- [8] CAI Q, ZHOU Y Q, LIU L, et al. Collaboration of heterogeneous edge computing paradigms: how to fill the gap between



- theory and practice[J]. IEEE Wireless Communications, 2024, 31(1): 110-117.
- [9] CAO J, LAM K Y, LEE L-H. Demo: real-time WebXR edge-based object detection for AR[C]//Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services. New York: ACM Press, 2023: 590 - 591.
- [10] XU Y N, FENG D Q, ZHAO M X, et al. Edge intelligence empowered metaverse: architecture, technologies, and open issues[J]. IEEE Network, 2023, 37(6): 92-100.
- [11] ZHOU Y Q, TIAN L, LIU L, et al. Fog computing enabled future mobile communication networks: a convergence of communication and computing[J]. IEEE Communications Magazine, 2019, 57(5): 20-27.
- [12] FENG D Q, LU L, WU Y Y, et al. Device-to-device communications underlying cellular networks[J]. IEEE Transactions on Communications, 2013, 61(8): 3541-3551.
- [13] POLYS N, WASI S M. Increasing Web3D accessibility with audio captioning[C]//Proceedings of the 28th International ACM Conference on 3D Web Technology. New York: ACM Press, 2023: Article 6.
- [14] LI L, HUANG Y K, QIAO X Q, et al. Towards distributed collaborative rendering service for immersive mobile Web[J]. IEEE Network, 2023, 38(3): 137-145.
- [15] ZHANG W X, LIN S K, BIJARBOONEH F H, et al. EdgeXAR: a 6-DoF camera multi-target interaction framework for MAR with user-friendly latency compensation[J]. The Proceedings of the ACM on Human Computer Interaction, 2022, 6 (EICS): Article 152.
- [16] HUANG Y K, QIAO X Q, DUSTDAR S, et al. AoDNN: an auto-offloading approach to optimize deep inference for fostering mobile web[C]//Proceedings of IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2022: 2198-2207.
- [17] GUO F X, YU F R, ZHANG H L, et al. An adaptive wireless virtual reality framework in future wireless networks: a distributed learning approach[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8514-8528.
- [18] ZHAO M Q, ZHENG J M, LIU E S. Server allocation for massively multiplayer online cloud games using evolutionary optimization[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2021, 17(2): Article 51.
- [19] REN P, LIU L, QIAO X Q, et al. Distributed edge system orchestration for web-based mobile augmented reality services[J]. IEEE Transactions on Services Computing, 2023, 16(3): 1778-1792.
- [20] REN P, QIAO X Q, HUANG Y K, et al. Edge AR X5: an edge-assisted multi-user collaborative framework for mobile web augmented reality in 5G and beyond[J]. IEEE Transactions on Cloud Computing, 2022, 10(4): 2521-2537.
- [21] 许昊天. WebVR大规模三维虚拟场景轻量对等传输关键技术[J]. 电脑知识与技术, 2019, 15(3): 215-217.
- XU H T. Key technologies of lightweight peer-to-peer transmission of large-scale 3D virtual scene in WebVR[J]. Computer Knowledge and Technology, 2019, 15(3): 215-217.
- [22] WU J K, GUAN Y, MAO Q, et al. ZGaming: zero-latency 3D cloud gaming by image prediction[C]//Proceedings of the ACM SIGCOMM 2023 Conference. New York: ACM Press, 2023: 710 - 723.
- [23] 李柯, 张乾, 贾金原. 云边页协同的WebBIM大场景多粒度兴趣加载调度算法[J]. 计算机辅助设计与图形学学报, 2021, 33(9): 1388-1397.
- LI K, ZHANG Q, JIA J Y. CEB-collaboratively multi-granularity interest scheduling algorithm for loading large Web-BIM scene[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(9): 1388-1397.
- [24] LEMOINE W, WIJNANTS M. Progressive network streaming of textured meshes in the binary glTF 2.0 format[C]//Proceedings of the 28th International ACM Conference on 3D Web Technology. New York: ACM Press, 2023: Article 7.
- [25] LIU C, OOI W T, JIA J Y, et al. Cloud baking: collaborative scene illumination for dynamic Web3D scenes[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(3): Article 59.
- [26] CHEN Y F, CAI W T, LIU E S. Hot area targeting dead reckoning for distributed virtual environments[C]//Proceedings of the 2021 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation. New York: ACM Press, 2021: 129 - 137.
- [27] YANG J C, GUO Z W, LUO J T, et al. Cloud-edge-end collaborative caching based on graph learning for cyber-physical virtual reality[J]. IEEE Systems Journal, 2023, 17(4): 5097-5108.
- [28] KUMAR S, BHAGAT L, A A F, et al. Multi-neural network based tiled 360° video caching with mobile edge computing[J]. Journal of Network and Computer Applications, 2022(201): 103342.
- [29] LIU X J, JIA J Y, LIU C H. Survey of lightweighting methods of huge 3D models for online Web3D visualization[J]. Virtual Reality & Intelligent Hardware, 2023, 5(5): 395-406.
- [30] 张韵, 王淑营, 郑庆, 等. 保持细节几何特征的三维网格模型轻量化算法[J]. 计算机应用, 2023, 43(4): 1226-1232.
- ZHANG Y, WANG S Y, ZHENG Q, et al. Lightweight algorithm of 3D mesh model for preserving detailed geometric features[J]. Journal of Computer Applications, 2023, 43(4): 1226-1232.
- [31] ZHOU C H, LIANG L. Research and implementation of lightweight technology for digital twin models of complex equip-

- ment[C]//Proceedings of 2023 IEEE 7th Information Technology and Mechatronics Engineering Conference (ITOEC). Piscataway: IEEE Press, 2023: 857-862.
- [32] LENG J W, LIN Z S, HUANG Z Q, et al. Rapid simplification of 3D geometry model of mechanisms in the digital twins-driven manufacturing system design[J]. Journal of Intelligent Manufacturing, 2023, 35: 2765 - 2786.
- [33] CHEN Z Q. A review of deep learning-powered mesh reconstruction methods[EB]. 2023.
- [34] HU Y B, YE S, ZHAO W, et al. O²-Recon: completing 3D reconstruction of occluded objects in the scene with a pre-trained 2D diffusion model[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(3): 2285-2293.
- [35] WEN B W, TREMBLAY J, BLUKIS V, et al. BundleSDF: neural 6-DoF tracking and 3D reconstruction of unknown objects[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 606-617.
- [36] REN K R, JIANG L H, LU T, et al. Octree-GS: towards consistent real-time rendering with LOD-structured 3D Gaussians[EB]. 2024.

[作者简介]



许娅男 (1996-), 女, 深圳大学电子与信息工程学院博士生, 主要研究方向为沉浸式通信、边缘计算。



赵桂丹 (1993-), 女, 重庆信息通信研究院无线部技术应用研究室主任, 主要研究方向为无线电频谱技术。



冯大权 (1986-), 男, 博士, 深圳大学电子与信息工程学院副教授、博士生导师, 广东省哲学社会科学重点实验室(文化数字化与文化创新发展)副主任, 主要研究方向为沉浸式通信、VA/AR、频谱共享技术。



施文哲 (1989-), 男, 移动网络和移动多媒体技术国家重点实验室高级工程师, 中兴通讯股份有限公司 XRExplore 产品规划经理, 主要研究方向为室内视觉AR导航、三维重建、实时云渲染。



陆平 (1971-), 男, 博士, 移动网络和移动多媒体技术国家重点实验室副主任, 中兴通讯股份有限公司副总裁、正高级工程师, 主要研究方向为云计算、增强现实、媒体大数据。