

沉浸式体积视频表征、编解码与跨层优化综述

黄亚坤¹⁾ 朱原玮²⁾ 邱丽飏¹⁾ 李 岩^{3),4)} 张 昊⁵⁾
吕 婧⁵⁾ 乔秀全¹⁾

¹⁾(北京邮电大学网络与交换技术全国重点实验室 北京 100876)

²⁾(安徽大学大数据与统计学院 合肥 230601)

³⁾(浪潮计算机科技有限公司 济南 250101)

⁴⁾(山东省先进计算重点实验室 济南 250101)

⁵⁾(中电万维信息技术有限责任公司 兰州 730030)

摘 要 随着虚拟现实、增强现实和混合现实等沉浸式交互技术的迅速发展,具备六自由度(6-DoF)交互体验的体积视频已成为实现真实感与沉浸感融合的关键技术。体积视频通过重建真实场景的几何与外观信息,使用户能够在三维空间中自由移动与视角切换,广泛应用于通信、娱乐、教育及工业仿真等领域。然而,其数据量通常是二维视频的数十倍甚至数百倍,对存储、传输与实时渲染带来严峻挑战,促使高效编解码技术成为研究热点。本文在系统梳理现有研究的基础上,从处理速度与时延预算出发,提出新的综述视角,将体积视频编解码技术划分为离线与近实时、面向交互式点播的实时,以及面向低时延直播的实时三类;分别总结了基于点云、体素、网格等显式几何表示的传统压缩方法,以及神经辐射场与三维高斯溅射等隐式神经表示的最新进展。同时,概述了针对不同应用场景的系统优化策略,并归纳了现有技术的关键特点、适用场景及未来挑战,展望了面向沉浸式通信的体积视频编解码的发展趋势。

关键词 体积视频;沉浸式交互;编解码技术;神经表示;跨层优化

中图法分类号 TP393 **DOI号** 10.11897/SP.J.1016.2026.01733

A Survey on Representation, Coding and Cross-Layer Optimization for Immersive Volumetric Video

HUANG Ya-Kun¹⁾ ZHU Yuan-Wei²⁾ QIU Li-Rui¹⁾ LI Yan^{3),4)} ZHANG Hao⁵⁾
LV Jing⁵⁾ QIAO Xiu-Quan¹⁾

¹⁾(State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876)

²⁾(School of Big Data and Statistics, Anhui University, Hefei 230601)

³⁾(Inspur Computer Technology Co., Ltd., Jinan 250101)

⁴⁾(Shandong Key Laboratory of Advanced Computing, Jinan 250101)

⁵⁾(China Telecom Wanwei Information Technology Co., Ltd., Lanzhou 730030)

Abstract With the rapid advancement of virtual, augmented, and mixed reality technologies, volumetric video with six degrees of freedom (6-DoF) interaction has become a key enabler for realistic and immersive experiences. By reconstructing the geometry and appearance of real

收稿日期:2025-06-27;在线发布日期:2026-05-07。本课题得到国家自然科学基金重点项目(No. 62431015)、国家自然科学基金面上项目(No. 62571055)资助。黄亚坤,博士,副教授,主要研究领域为沉浸式通信、体积视频传输、分布式系统、智能AR/VR服务。E-mail: ykhuang@bupt.edu.cn。朱原玮,博士,讲师,主要研究领域为体积视频流、强化学习。邱丽飏,博士研究生,主要研究领域为体积视频编解码、三维高斯溅射。李 岩,工程师,主要研究领域为异构加速技术。张 昊,硕士,高级工程师,主要研究领域为空间计算、沉浸式通信。吕 婧,硕士,工程师,主要研究领域为空间计算、沉浸式通信。乔秀全(通信作者),博士,教授,长江学者,主要研究领域为沉浸式通信、6G全息通信、数字孪生与三维重建、智能AR/VR服务。E-mail: qiaoxq@bupt.edu.cn。

scenes, volumetric video allows users to freely move and change viewpoints in three-dimensional space, finding broad applications in communication, entertainment, education, and industrial simulation. However, its data scale—often tens to hundreds of times that of 2D video—poses severe challenges for storage, transmission, and real-time rendering, driving intensive research on efficient compression and coding. This paper presents a comprehensive survey of volumetric video coding from the perspective of processing speed and latency constraints, classifying existing approaches into three categories: offline and near real-time, real-time for interactive on-demand streaming, and real-time for low-latency live streaming. We summarize conventional methods based on explicit geometric representations such as point clouds, voxels, and meshes, and highlight recent advances in implicit neural representations, including Neural Radiance Fields (NeRF) and 3D Gaussian Splatting. In addition, we review system-level optimization strategies for various applications and analyze the key characteristics, applicable scenarios, and challenges of current techniques. Finally, future trends in volumetric video coding for immersive communication are discussed.

Keywords volumetric video; immersive interaction; coding and compression techniques; neural representations; cross-layer optimization

1 引 言

随着计算机图形处理能力的提升、网络通信能力的增强以及交互设备的不断迭代,数字内容的消费方式正呈现出由传统二维平面视听向更加逼真且具有交互性的三维沉浸式体验转变的趋势。虚拟现实(Virtual Reality, VR)、增强现实(Augmented Reality, AR)以及混合现实(Mixed Reality, MR)等沉浸式技术的迅猛发展,使用户能够从被动接受信息转变为主动参与和自由交互,从而成为未来人机交互的重要方向^[1-2]。然而,360度全景视频虽然增强了视觉包围感,但其交互自由度仍局限于固定视点,难以支持在三维空间中的自由移动与动态视角切换,限制了真正的沉浸式体验^[3-5]。

体积视频(Volumetric Video),又称自由视点视频(Free-Viewpoint Video, FVV)或六自由度(Six Degrees of Freedom, 6-DoF)视频,通过采集与重建真实场景的几何与外观信息,使用户能够在三维空间内自由改变位置与视角,实现高度真实感的沉浸式交互体验。体积视频在远程通信、文化遗产保护、虚拟旅游、互动娱乐、医学培训、工业仿真以及体育与演艺直播等多个应用领域展现出广阔前景^[6-10]。然而,体积视频的数据规模通常是二维视频的数十倍甚至数百倍,原始数据每分钟可达TB级别,这给存储、传输、实时解码和渲染带来了严峻挑战^[11-12]。如何在有限带宽与算力条件下,实现高保

真、低延迟的体积视频传输与呈现^[13-14],成为推动6G沉浸式通信落地的核心问题。

为应对上述挑战,高效的体积视频编解码技术近年来受到学术界和工业界的广泛关注。已有综述工作大多从数据表征形式或压缩方法类型出发,对体积视频编解码方法进行梳理^[3-4,11-12]。本文在此基础上以处理速度与时延预算为分类主线,将现有方法系统划分为三大类:离线高保真编解码、面向交互式点播的实时编解码以及面向低时延交互直播的实时编解码,对应不同应用场景对性能的核心需求(如延迟约束与交互复杂度)。此外,本文系统回顾了体积视频编解码从传统显式几何表示向新兴神经表示的演进路径,深入分析了其压缩原理、关键技术与最新进展,总结了相关优化策略与跨层系统实现,并对现有实验评价指标与标准数据集进行了整理与归纳。

本文主要贡献如下:(1)提出以处理速度与时延预算为核心的分类体系,构建符合实际应用需求的综述框架;(2)梳理体积视频表示的演进脉络,系统总结显式几何与神经表示在编解码中的特点与进展;(3)归纳现有优化方法与系统实现路径,覆盖编解码器、传输、客户端与硬件协同;(4)总结现有实验评估指标与数据集,讨论未来挑战与发展趋势。

本文内容组织结构如图1所示,其余章节安排如下:第2节介绍体积视频的定义与主流表征方法;第3节按照处理速度与时延预算,综述离线、点播实时与直播实时三类编解码方法;第4节总结跨层优

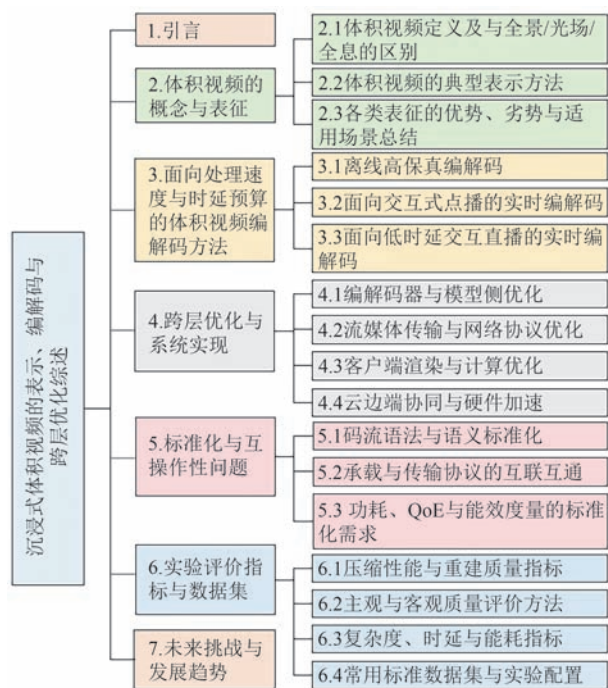


图1 本文内容组织结构

化与系统实现,包括编解码器、传输协议、客户端和硬件加速;第5节讨论标准化与互操作性问题;第6节介绍常用的实验评价指标与数据集;第7节分析未来研究挑战与趋势;第8节对全文进行总结。

2 体积视频的概念与表征

2.1 体积视频定义及与全景/光场/全息的区别

体积视频是对时变三维场景的数字表示^[15-16]。记体积视频为序列 $\{S(t)\}_{t=1}^T$,每一时刻 $S(t)=(\mathcal{G}_t, \mathcal{L}_t)$,其中, $\mathcal{G}_t \subset \mathbb{R}^3$ 表示几何与结构(如点集、网格、体素或显式基元集合等), $\mathcal{L}_t: (x, d) \mapsto (\sigma, c)$,为位置-方向辐射场,把空间点 x 与观察方向 d 映射到体密度/不透明度 σ 与颜色/辐照度 c 。体积视频的6-DoF播放要求解码后在有限空间内支持位置与朝向的自由变化,并在既定时延预算内完成渲染交互。

与体积视频密切相关的沉浸式媒体形式包括全景视频、光场视频和全息视频,但四者在自由度、信息量与应用场景上均存在显著差异(如图2所示)。全景视频通过单点采集获得全向影像,提供方向上的三自由度体验,但视点固定,缺乏空间移动自由度,无法实现真实的6-DoF交互。光场视频通过稠密采样并合成连续视图,在一定程度上降低了对三维重建的依赖,支持更自然的视角切换,但本质上仍为视点插值,未形成可编辑和可压缩的三维模

型。全息视频则直接编码并调制光波的振幅与相位,从物理层面再现波前传播特性,虽具备独特的真实感优势,却对光学硬件高度依赖,难以与现有通信和编解码体系兼容。总结来看,体积视频的核心特征在于构建可压缩、可传输与可渲染的三维表示,以满足交互场景下低时延与高保真度的需求。

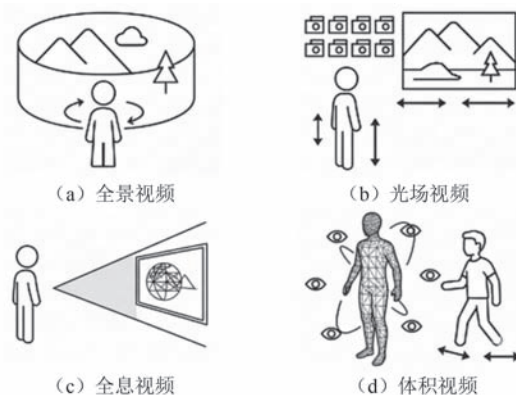


图2 不同沉浸式媒体对比示意图

2.2 体积视频的典型表示方法

体积视频的表征方式既决定其编码、传输与渲染方法,也是理解不同技术路线演化的核心。现有表征主要分为两类:显式几何表示与隐式表示。前者利用离散点、网格或体元直接存储几何与外观信息,具有结构直观、工具成熟和渲染高效的特点;后者则通过连续函数或神经网络间接刻画物体表面与体积属性,在表达能力和压缩性能上展现独特优势。隐式表示进一步可分为基于解析函数的隐式方法与基于学习的神经隐式方法,分别对应传统几何连续场建模与最近兴起的神经场景建模。下文对这两类方法进行介绍。

2.2.1 显式几何表示

显式几何表示利用离散数据结构直接刻画场景几何,具有直观、可编辑、易于硬件加速等优势。点云是最基础且常见的形式(如图3所示),以带有颜色、法线或纹理属性的离散点集合近似表面,采集简便且无需显式拓扑关系,但稀疏性、不均匀分布和噪声问题显著影响压缩与渲染效率^[3,17-25]。多边形网格通过顶点、边和面的拓扑关系构建连续表面,尤其三角网格特点与GPU渲染管线兼容性好、工具链成熟,因而得到了广泛应用(如图4所示)^[26-32]。然而,刻画光滑曲面或复杂几何时需大量面片,导致数据量增加,且难以表示烟雾、火焰等参与介质。体素网格通过规则划分将空间离散为体素单元,每个单元记录颜色、密度或占用状态,可自然表达内部结构与

体积效应,因此在医学成像等领域具有重要应用^[33]。但其存储成本随分辨率呈立方增长,限制了在大规模动态场景中的使用。为此,研究者提出了八叉树、稀疏体素八叉树(Sparse Voxel Octree, SVO)与体素动态B+树(Volumetric Dynamic B+ Tree, VDB)等结构,以缓解存储与计算压力^[34-39]。

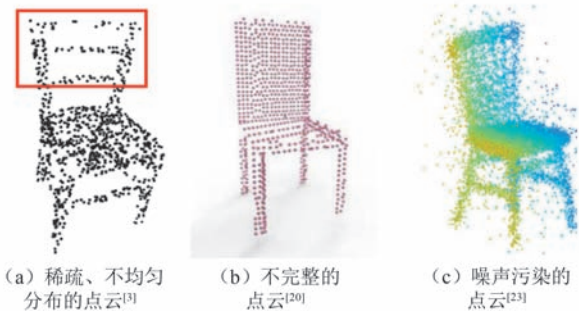


图3 不同类型点云示意图

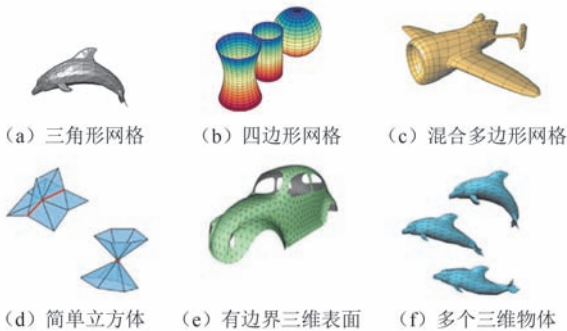


图4 多边形网格模型^[3]

2.2.2 隐式表示

2.2.2.1 基于解析函数的隐式表示

解析隐式表示通常利用连续标量函数定义物体

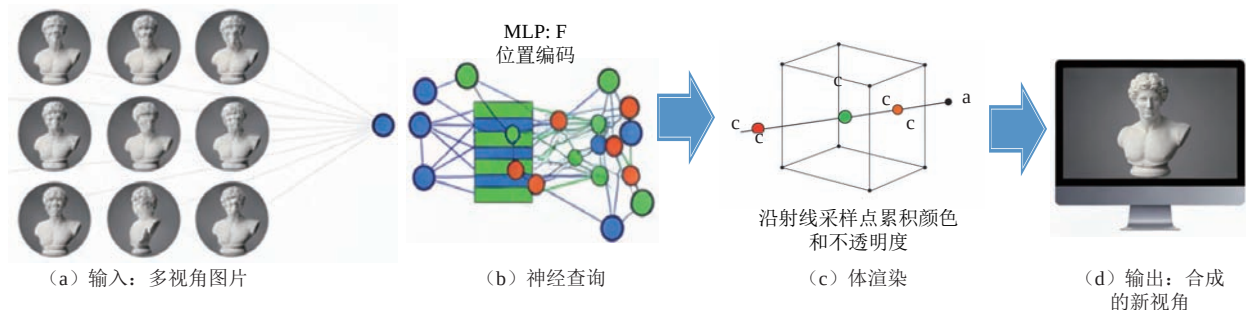


图5 神经辐射场技术路线图

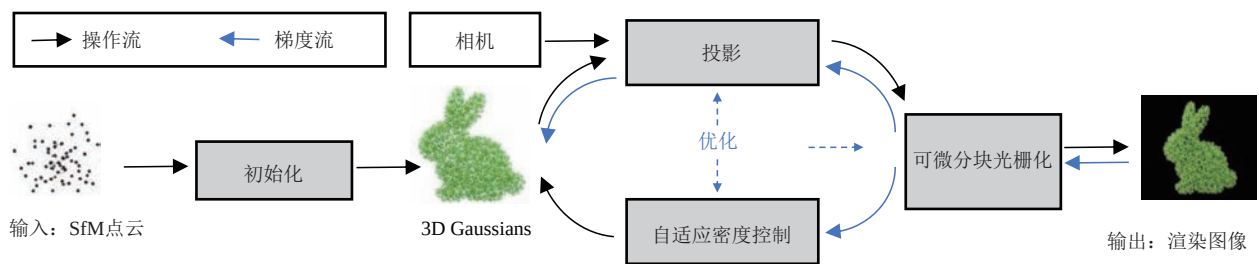
在 NeRF 之后,3D 高斯溅射(3D Gaussian Splatting, 3DGS)迅速成为神经隐式/混合表示的新兴方向^[63-64]。如图6所示,3DGS将场景建模为高斯基元集合,每个基元承载位置、协方差矩阵、不透

表面。符号距离场(Signed Distance Field, SDF)^[40-41]通过计算空间点到表面的有符号最短距离确定几何边界,其零等值面即为目标表面。其梯度满足Eikonal方程,因此在法向估计、光线追踪和几何运算中具有优势。常见扩展包括无符号距离场(Undsigned Distance Field, UDF)^[42]与截断符号距离场(Truncated Signed Distance Field, TSDF)^[43],后者已广泛应用于KinectFusion^[44]、Voxel Hashing^[45]等实时三维重建系统。尽管SDF/TSDF在连续曲面表达和几何运算支持方面表现优异,但在高分辨率下数据开销巨大,从稀疏观测恢复全局一致场亦较困难,且渲染依赖光线步进,计算代价较高^[46]。总体而言,解析隐式方法在数学严谨性与几何灵活性上具优势,但在效率和交互性方面仍受制约。

2.2.2.2 基于学习的神经隐式表示

近年来,基于深度网络的神经隐式表示已成为体积视频与自由视点渲染的重要方向。这类方法通过端到端学习,将空间坐标与观察方向映射为体密度与颜色,实现连续场景的高保真重建。神经辐射场(Neural Radiance Field, NeRF)^[47]及其变体奠定了基础:通过密集采样体积渲染模型实现新视角合成(如图5所示)。NeRF虽然在静态场景中达到了高质量渲染,但训练与推理效率低下,难以应用于实时或大规模动态场景。研究者提出引入稀疏体素网格(VoxelNeRF)、哈希编码(Instant-NGP)、混合显式/隐式结构(MipNeRF、Plenoxels等)以加速训练与渲染^[48-55]。在动态场景建模方面,D-NeRF、NR-NeRF、NeRFPlayer等通过引入时序变形场或残差模式实现帧间一致性^[56-62]。然而,这些技术依旧受限于端侧实时性和模型压缩开销。

明度以及颜色/方向关联属性,并通过GPU加速的光栅化溅射渲染实现实时显示(典型帧率可达60 FPS以上)。这一表示兼具显式结构与隐式表达能力,在效率与交互性方面都较NeRF有优势。

图6 3DGS技术路线图^[63]

针对3DGS存储庞大与视觉伪影问题,研究者提出了量化、熵编码、敏感性聚类、矢量聚类、剪枝等压缩技术:如Compressed 3DGS采用敏感性向量聚类 and 量化训练,压缩率可达 $31\times$,视觉误差极低^[65];LightGaussian通过剪枝与紧凑重参数化实现 $15\times$ 压缩率并将渲染速度提升至200 FPS以上^[66];TC-GS利用三平面特征压缩高斯参数,提高了可伸缩性与带宽适配性^[67];而文献[68]提出了提供多层结构化压缩以适配大场景流式传输。此外,FCGS^[69]采用前馈网络直接学习压缩映射,大幅降低训练时间;文献[70]则结合标准视频编解码器实现跨模态压缩。这些研究使3DGS在压缩率、实时性与可编辑性方面持续优化。

在动态场景方向,4D高斯溅射(4D Gaussian Splatting, 4DGS)已被应用于实时动态重建与渲染,通过在规范空间建立静态高斯集合并利用形变场映射至时变状态,有效平衡了压缩效率与时间一致性。后续研究如Instant Gaussian Stream^[71]提出了关键帧驱动与锚点引导的高斯运动网络以支持流式重建;MBGS^[72]利用运动图控制高斯参数插值以提高动态稳定性;EvolvingGS^[73]采用增量演化策略自适应更新高斯参数,实现帧间一致性与局部精度平衡;MEGA^[74]从架构层面优化显存与缓存管理,在保持质量的前提下显著降低内存占用;4DGS SLAM^[75]将4DGS扩展至动态SLAM任务,实现同步相机位姿估计与场景重建;而SaRO-GS^[76]则通过尺度感知残差建模进一步增强了运动一致性。

2.3 各类表征的优势、劣势与适用场景总结

不同体积视频表征方式在几何精确性、外观拟合、存储开销与渲染效率等方面各具优势与不足,因此在实际应用中呈现出差异化特征。显式方法具有结构直观、工具链成熟等优点,但在连续性和压缩潜力上存在局限;解析函数型隐式方法在数学严谨性和几何灵活性方面表现突出,却受限于效率与交互性;神经隐式方法则在渲染质量、压缩与动态建模上展现潜力,但实时性与资源开销仍是挑战。表1总

结呈现了不同方法在数学属性、计算复杂度与应用价值上的差异。

2.3.1 显式几何表征:直观性与成熟度

显式几何表征是三维建模最早发展且应用最广的形式。点云直接记录离散采样点及其属性(颜色、法向),具备采集便捷、拓扑灵活等优势,但因缺乏显式连接关系,稀疏不规则数据需经复杂预处理才能高效渲染与压缩。多边形网格通过顶点、边与面的拓扑关系构建连续表面,具有较高的可视化效率,与GPU光栅化管线高度兼容,并受益于成熟的建模与动画工具支持。然而,其在高精度曲面建模中依赖密集面片,易导致数据冗余,且动态场景下的拓扑一致性维护困难。体素网格将空间划分为规则单元,可自然表达内部结构与体积特性,常用于医学成像和参与介质建模。但其存储需求随分辨率呈立方增长,在稀疏区域尤显冗余。

上述显式表征的共性在于几何结构直观、语义清晰、易于编辑并支持硬件加速,为压缩编码提供了明确的数据结构基础。例如,MPEG标准化组织提出了V-PCC和G-PCC点云与网格压缩方案^[77-79]。然而,此类方法在表达视角相关外观、复杂光照和超精细几何细节方面存在限制,且原始数据规模庞大,压缩前处理开销较高。

2.3.2 隐式函数表征:连续性与几何精确性

函数型隐式方法通过连续场间接定义物体边界,典型代表为符号距离场(SDF)及其TSDF/UDF变体。这类方法在几何连续性、拓扑灵活性和解析几何运算方面具有优势,并已在多视融合和实时建图中得到验证。其边界由零等值面刻画,梯度满足Eikonal性质,可用于法向估计,因而在表面重建与布尔运算中具备数值稳定性与可解释性。然而,在编解码与实时渲染中,通常需在体素网格上离散采样以支持存储与可视化,高分辨场景下数据规模和计算成本显著增加;同时,依赖光线步进的体绘制难以在端侧时延预算下实现交互式帧率。因此,函数型隐式方法更适合作为高精度重建和几何分析

表1 各类体积视频表征方法特性比较

特性维度	点云	多边形网格	体素网格	SDF/TSDF	NeRF	3DGS	4DGS
数据性质	显式,离散点	显式,表面元	显式,体元	隐式 (需离散采样)	隐式(MLP/混合 显式结构)	神经隐式 (高斯基元渲染)	神经隐式 (时空联合高斯基元)
连续性	离散	分段连续	离散	场连续(等值面)	场连续	基元集合 (近似场)	时空连续 (规范空间映射)
主要优势	采集直接、 结构简单	渲染高效、 编辑成熟	体积/内部 结构自然	几何平滑、 拓扑灵活	视角相关外观与 保真度高	实时高质量渲染、 训练较快	动态一致性强、 压缩高效、流式重建
主要劣势	缺拓扑、 渲染空洞	高精度时 面片冗余	数据量极高、 块效应	高分辨存储与 渲染代价高	训练/推理成本 高、编辑困难	基元数量大、 潜在伪影	形变场复杂、 内存开销较高
数据量 (原始)	中-高 (随密度)	中-高 (随复杂度)	极高 (随分辨率)	极高(随分辨率)	低(MLP)/ 中-高(混合)	中-高 (随基元数)	中(跨帧共享与 稀疏优化后)
视角相关	有限	依赖纹理/ 着色器	有限 (体素属性)	无直接支持	优异	优异	优异(时空一致 球谐特征)
外观	(点属性)						
体积效应	无直接支持	有限	优异(密度/ 透明度)	有限 (符号/等值面)	良好	良好	良好
渲染速度	快(点渲染)	极快 (GPU光栅化)	中-慢 (体渲染)	中-慢(光线步进)	极慢(快速变体 可实时)	极快(GPU溅射)	快-实时(时变 缓存与增量渲染)
训练/构 建时间	采集/预处理	建模/扫描 重建	采集/预处理	三维重建	慢(加速变体 显著提升)	中-快	中(动态形变场 需额外训练)
可编辑性	中(点级操 作)	高(顶点/ 面编辑)	低(体素级 编辑)	中(隐函数运算)	低(隐式 MLP)	中(基元操作)	中(规范空间编辑 与运动约束)
压缩潜力	G-PCC/ V-PCC等	Draco、MPEG FAMC等	八叉树/稀疏 结构	八叉树/稀疏结构	网络参数与 特征压缩	参数量化与熵编码	基元共享、跨帧稀 疏调度、形变场压缩

表征,在实时传播链路中需结合层次稀疏结构或学习型模块,以缓解存储与渲染效率的瓶颈^[40-46]。

2.3.3 基于学习的神经表征:高保真度与紧凑性新范式

神经隐式方法将连续场的表达能力与数据驱动的拟合机制结合,在高保真重建、视角相关外观建模与动态扩展方面推动了体积视频表征的演进。早期NeRF利用多层感知机将空间坐标与方向映射为颜色与体密度,并结合体积渲染实现新视角重建^[47],在外观与视角效应建模方面具有先天优势,但其训练与推理代价高昂,限制了实时交互能力。为应对这一瓶颈,后续研究^[48-55]融入结构化加速,例如稀疏体素和分块缓存、哈希网格编码、浅层网络结构等机制,这些方法以较低代价换取较高帧率或收敛速度。

在此框架之上,3DGS为神经隐式表征提供了新的折中方案:训练仍在神经网络优化范式下进行,而在推理阶段通过显式基元渲染实现高效实时合成。相比NeRF,3DGS在渲染效率与交互能力上具有显著提升,但其基元规模庞大与伪影问题成为新的技术瓶颈。为此,研究者提出了一系列紧凑化和流式化策略,如高斯参数的量化与熵编码、基元剪枝、上下文聚类、渐进加载和可编辑结构^[63-64,70,80-85]。

此外,面向动态体积视频的神经表征正在快

速演化。动态高斯方案通过对每帧参数变化进行建模以表达时变场景^[86],而4DGS则在规范空间中维护一组静态高斯基元,并辅以形变场将其映射到不同帧,实现更稳定一致的时间建模^[87]。近年来,为增强压缩与渲染效率,还出现了若干新工作:NeRFCoder^[88]将NeRF模型编码为可流式比特流,以支持端对端压缩与解码;Gaussianflow^[89]提出用光流场引导高斯参数演化从而提升时序一致性;HybridGS^[90]尝试融合网格结构与高斯表示以兼顾结构约束与表达能力。

总结来看,神经隐式表示,尤其是3DGS和4DGS,使体积视频有望同时兼顾高保真、紧凑表示和实时性。NeRF在参数紧凑性与外观拟合上仍具优势,但受限于推理效率;3DGS/4DGS在交互性和渲染能力上表现突出,却受存储压力与伪影控制挑战;显式几何虽然应用基础成熟,但尚难以完成复杂外观建模。因此,融合显式结构与神经隐式表达的混合表征方案,将是协调高保真、低延迟与系统可落地性的关键方向。

3 面向处理速度与时延预算的体积视频编解码方法

本章以处理速度与时延预算为主线,对体积视

频编解码的设计空间与方法演进进行系统梳理。在统一的显式几何与神经表示两条表征路径下,分别讨论离线编解码的高精度处理、面向交互式点播的实时处理以及面向直播交互的低时延处理,并对比三类模式下在率失真、复杂度与时延维度上的表现。其中,离线编解码主要面向数据归档与高精度后期制作,数据为本地预先准备,编解码追求极致的率失真性能,可容忍较高的解码复杂度。点播场景数据同样为远端预先准备,但其编码结构区别在于通常采用非对称编解码架构,编码端需预先构建空间索引、层次化结构或视点感知机制,以支撑客户端在毫秒级时延内完成视图的随机访问与实时解码。直播场景其数据为实时采集生成,必须在百毫秒级的端到端时延预算内完成从采集、在线编码、传输到本地交互渲染的全线同步。本章将基于上述架构与时延特性的差异,系统梳理现有技术。

3.1 离线高保真编解码

离线高保真编解码方法适用于端到端时延要求宽松的场景,如存储、后期制作和高精度分发。在此类应用中,编码端采用高复杂度模型以提升压缩率和重建保真度,解码端则依赖批处理或准实时渲

染。针对点云、网格、体素、隐式函数和神经表示等不同表征形式,已有方法在编解码策略上呈现出显著差异,以适配各自的数据特性与应用需求。

3.1.1 面向显式几何表示的离线编解码方法

显式几何表示的离线编解码方法主要涉及点云、网格与体素,已形成较为成熟的研究体系。在点云压缩方面,早期方法多依赖基于规则的空间划分与变换。典型如八叉树(Octree)^[91],可通过递归划分实现分辨率自适应压缩,但深层结构计算开销较大。为提升属性压缩效率,Zhang等提出基于图傅里叶变换(Graph Fourier Transform, GFT)的点云方法^[92],但成本较高;De Queiro等提出的区域自适应分层变换^[93]显著降低复杂度,并被采纳入MPEG G-PCC标准^[79]。V-PCC^[78]将点云投影为二维图像序列,并借助视频编码器(HEVC、VVC)实现压缩,在自由视点与沉浸式应用中广泛使用^[78,94]。总体上,V-PCC更适合稠密、纹理丰富的动态点云,而G-PCC在稀疏点云与CAD模型中表现更好。为更直观地呈现上述面向显式几何表示的离线编解码方法在性能与适用性上的差异,表2中对比了各典型算法的核心特点、代表性量化性能及其关键总结。

表2 面向显式几何表示的离线编解码方法典型性能对比

方法	具体算法/方法	核心特点与适用性	典型量化性能	关键总结
点云压缩	八叉树(Octree) ^[91]	几何编码,递归划分,分辨率自适应;计算开销随深度增加	CR:10:1-30:1(几何) 几何质量:中等PSNR	基础几何编码,结构简单,计算复杂度相对高
	基于图傅里叶变换(GFT) ^[92]	属性压缩,利用图结构;成本较高,通用性受限	属性CR:优于传统变换 质量:高属性PSNR	属性编码特化,计算成本与通用性挑战
	RAHT(MPEG G-PCC子集) ^[83]	几何编码核心,区域自适应,低复杂度;适用于稀疏点云	CR:G-PCC中几何核心 几何质量:高PSNR	G-PCC几何核心,低复杂度,适用于稀疏点云
	V-PCC ^[78]	投影2D图像,借用视频编码器;稠密、纹理丰富动态点云	CR:20:1-100:1 几何/纹理质量:高PSNR, D1 PSNR > 40 dB	适用于稠密动态、纹理点云,利用视频编码器,速度快
	G-PCC ^[79]	独立几何编码;稀疏点云与CAD模型,高精度几何	CR:20:1-80:1 几何质量:极高PSNR, D2 PSNR > 45 dB 速度:适中	适用于稀疏点云/CAD,几何保真度高,与V-PCC互补
体素压缩	SVO/VDB ^[34,95]	稀疏数据存储,非编码;静态稀疏体素,实时渲染	存储效率:相较密集体素显著降低(>10x),需结合编码	稀疏数据管理,非直接压缩,需结合编码
	3D-DCT/三维小波变换 ^[96-97]	能量聚集,高压缩比;高分辨率/复杂动态下效率受限	CR:20:1-50:1 质量:中高PSNR,易块效应	传统方法,高压比,但高分辨率/动态受限,易失真
	动态体素方法(差分/运动补偿等)	差分更新/运动补偿,利用时序相关性;适用于变化缓慢/可预测运动	CR:随帧间相似度,高质量:中高PSNR	针对动态体素,利用时序相关性,适用于预测性运动
学习方法	基于学习的几何/属性压缩 ^[98-105]	神经网络端到端学习,处理非刚性形变与高维属性;泛化性/计算开销仍是瓶颈	CR:可达传统方法相当 质量:可优化感知质量	潜力大,适用于非刚性/高维属性,泛化与计算待优化

体素压缩则更注重稀疏性与块结构利用。静态体素常基于SVO^[95]或VDB^[34]存储,并结合

3D-DCT^[106]或三维小波变换^[107]进行能量聚集式编码;动态体素则采用八叉树差分更新^[108-109]、运动补

偿残差编码^[110-111], 或(3D+t)-DCT等联合变换^[96-97]。这类方法在理论上已较为成熟,但在高分辨率场景和复杂动态条件下,效率仍然受限。

学习方法近年来逐渐应用于显式数据压缩。Xu等^[98]提出可学习的几何下采样/上采样机制, Sun等^[99]利用频率分解实现多尺度压缩; Akhtar等^[100]与Shen等^[101]探索动态点云预测与局部几何重建。PolyGen^[102]与MeshCNN Autoencoder^[103]展示了基于图神经网络的压缩潜力。3D CNN自编码器框架^[104-105]可建模复杂时空结构。尽管学习方法在非刚性动态与高维属性建模上具优势,但泛化性与计算开销仍是瓶颈。

总结来看,传统方法如八叉树与基于变换的体素压缩,在特定场景下表现稳健,但普遍面临高分辨率或复杂动态数据处理的效率瓶颈。MPEG标准中的V-PCC和G-PCC则针对稠密动态点云与稀疏/CAD模型形成了互补优势,通过成熟的工业级方案实现了高效编码。而学习型方法,尽管在非刚性形变与高维属性建模方面展现出巨大潜力,其泛化性、训练成本与计算开销仍是当前研究需克服的挑战。

3.1.2 面向隐式与神经表示的离线编解码方法

隐式函数与神经表示近年来成为体积视频离线编解码的重要研究方向。隐式函数(如SDF及其变体TSDF/UDF)可连续描述几何边界与拓扑,离散化后常存储为体素网格,因而部分压缩策略与体素类似。但其零等值面信息对重建精度尤为关键,需要在表面邻域设计高保真保护机制^[112-113]。常见方法包括稀疏八叉树与差分更新,以及基于3D CNN或自编码器的端到端学习压缩^[114]。神经隐式表示(DeepSDF^[41]、SIREN^[115]等)通过MLP参数化函数实现紧凑几何建模。其压缩主要借鉴网络压缩技术,如剪枝^[116]、量化^[117]、权重共享^[118]、低秩分解^[119]与知识蒸馏^[120-121]。在动态场景下,通常将时间编码输入空间、学习潜在序列,或结合规范空间与形变场建模^[122],相应需压缩网络参数与潜在驱动向量。

表3对各类典型方法进行了对比总结。这些方法普遍通过连续函数或可学习基元紧凑表示复杂场景,在几何与外观的高保真拟合及渲染真实感方面展现出巨大潜力。其压缩重点从传统的数据结构转向网络参数、辅助显式结构或高斯基元的优化。

表3 面向隐式与神经表示的离线编解码方法典型性能对比

方法类别	具体方法	核心特点与建模机制	主要压缩策略	优势	挑战
隐式函数	SDF/TSDF/UDF ^[112-113] , 3D CNNs ^[114]	连续描述 几何边界与拓扑离散化后为体素网格	稀疏八叉树; 3D CNN/端到端学习	几何连续性; 高保真零等值面重建潜力	对零等值面保护机制设计复杂; 高分辨率下效率受限
神经隐式表示	DeepSDF ^[41] , SIREN ^[115]	MLP参数化函数,紧凑几何建模; 动态场景下时间序列编码	剪枝、量化、权重共享、低秩分解、知识蒸馏	紧凑模型; 可学习几何; 动态场景建模潜力	参数规模大; 训练成本高; 网络压缩效果受限
NeRF	NeRF ^[123-124] , PlenOctrees ^[49] , Instant-NGP ^[125] , DVGO ^[126]	MLP学习场景辐射场; 结合显式辅助结构(网格、哈希表、稀疏体素)	量化、稀疏编码、预测、熵编码; 动态NeRF需压缩形变场/潜在编码	真实感渲染; 高保真外观 几何拟合; 多视角合成	训练成本高; 参数规模大; 泛化能力弱; 动态场景实时性挑战
3DGS	HAC ^[127] , HAC++ ^[128] , Scaffold-GS ^[129] , Compact3DGS ^[130] , 4DGS ^[131] , 3DGS ^[86] , LightGaussian ^[66] , MEGA ^[74] , 4DGS-1K ^[132]	显式高斯基元场景建模; 3DGS:静态高效渲染, 4DGS:引入规范空间形变场以建模时变场景	参数量化、向量聚类、基元剪枝、锚点表示、上下文熵编码、时空差分与跨帧共享;端到端率失真优化学习压缩	训练与推理快; 实时渲染; 支持动态一致性与流式传输; >30×压缩能力	存储与基元数量大; 伪影控制与形变场拟合复杂; 标准化与跨场景泛化能力不足

基于神经场的方法中,NeRF^[123-124]及其加速变体(PlenOctrees、Instant-NGP、DVGO等^[49, 125-126])最具代表性。其压缩对象包括网络参数与显式辅助结构(特征网格、哈希表、稀疏体素),常结合量化、稀疏编码、预测与熵编码等手段^[49, 133];动态NeRF^[134-135]还需处理形变场与帧级潜在编码。相比之下,3DGS与4DGS^[74, 86-87, 129-132, 136]通过数百万高斯基元直接建

模场景,压缩多依赖参数量化、基元剪枝、差分与向量量化。Lu等人提出的Scaffold-GS方法通过双层的体素-高斯表征显著提升了隐式神经场的编码效率^[129],该方法首先基于稀疏体素网格生成锚点构建场景的空间支架,每个锚点携带可学习的局部特征和尺度参数;然后在渲染时,仅对视锥内的锚点由小型MLP动态解码出若干局部3D高斯的属性(如不透

明度、颜色、方向和尺度等),实现了体素锚点与神经高斯表示相结合的分层结构。由于每个锚点仅在需要时生成有限的高斯,该模型实现了参数的共享与压缩,大幅减少了冗余存储和计算开销,并通过锚点的生长/裁剪策略自适应优化场景结构覆盖。在保证接近实时的渲染速度下,以更紧凑的模型达到了与原始3DGS方法相当的渲染质量,在大视角变化、透明、高光等复杂场景下表现出更强的鲁棒性^[129]。

此外,近期如 Compact3DGS^[130]、Light4GS^[131]、HAC^[127]与 HAC++^[128]等进一步探索端到端学习压缩,将率失真优化与表示学习结合,以兼顾效率与实时性。其中,HAC^[127]针对 3DGS 模型进行端到端压缩,通过学习上下文概率模型将每个高斯的空间位置、外观特征和透明度等属性联合编码,预测出熵编码所需的概率分布,从而充分挖掘高斯参数间的相关性以最大化压缩效率。HAC 引入了结构化的二值哈希网格对原本稀疏无序的 3D 高斯进行结构性建模,根据锚点位置查询哈希网格并获得插值哈希特征,用作上下文来预测锚点属性的值分布,从而显式捕获邻近高斯之间的空间相关性。同时,其设计了自适应掩码机制以剪除无效高斯和锚点,进一步消除冗余,对 3DGS 模型压缩平均达到 75× 以上。

此外,HAC 思路已在后续工作中得到拓展,例如 HAC++ 可将压缩率提升至近 100× 量级,同时保持高保真重建^[128]。总体来看,隐式与神经表示在几何/外观拟合、压缩潜力与渲染真实感方面优势显著,但仍面临训练成本高、参数规模大与泛化不足等挑战。

3.1.3 离线与近实时编解码特性分析

离线与近实时编解码通常在率失真、复杂度与时延之间权衡。显式几何方法根据规则化数据结构与标准化体系(如 V-PCC、G-PCC),在较低复杂度下可获得稳定的率失真性能,适合高精度存储与分发,但在动态场景与复杂外观建模上存在不足。隐式函数方法(如 SDF/TSDF)具有几何连续性与拓扑灵活性优势,但依赖密集采样和光线步进,效率与时延表现不及显式方法。神经表示(如 NeRF、3DGS/4DGS)则在外观建模与动态适应性上表现突出,能实现高保真重建与实时渲染,但面临参数规模庞大、压缩与低延迟优化困难等挑战。

从表 4 关键指标对比的结果来看,三类方法在不同性能维度上形成互补:显式几何强调稳定性与标准兼容性,隐式函数突出几何连续性,神经表示则引领视觉质量与拟合能力。

表 4 各类体积视频表征方法关键指标比较

方法	率失真	复杂度	时延	关键特征
显式几何	稳定、成熟,标准化支持好	依赖结构化数据	适合离线	高兼容性,几何精度适中
隐式函数	高几何保真度与拓扑灵活性	依赖稠密采样	实时性受限	表面连续光滑,压缩率低
神经表示	视觉保真度最高	训练/渲染开销大	离线	高质量渲染,参数冗余

3.2 面向交互式点播的实时编解码

与离线场景侧重压缩比和高保真不同,实时交互式点播更强调端到端低延迟与即时响应。在此类应用中,编解码方法需在压缩率与重建质量的平衡基础上,进一步优化解码复杂度与交互时延。本节将结合显式几何与神经表示两类主流表征,综述典型实时点播编解码方法,并从率失真、复杂度与时延等维度进行比较分析。

3.2.1 VoD 场景特点与性能需求

如图 7 所示,体积视频点播(Video on Demand, VoD)展现出明显的编解码不对称结构:服务端可在离线阶段部署高复杂度优化模型,以获取优异的压缩率与重建质量^[137-139];而客户端则必须在受限算力和存储预算下完成解码与渲染,以维持沉浸式交互体验,因此解码/渲染效率成为系统瓶颈。传统的

VoD 系统中常依赖提前生成视点质量图谱、几何简化代理或空间索引等辅助信息,以指导客户端快速加载与渲染^[140]。通常,客户端解码必须在几十到数百毫秒的时延预算内完成,尤其在移动终端或轻量头显设备上,对并行性、内存访问和渲染调度的效率要求极高^[141-142]。相比于传统二维视频,体积视频在数据规模、三维几何复杂性与视角自由度方面呈指数级增长,其对带宽与存储的要求也远超常规流媒体管线。为适配这种高负载环境,渐进式数据加载、视点预测驱动的自适应传输与轻量化代理渲染等方案被广泛采纳^[143]。

随着神经表示方法的兴起,VoD 场景下的研究方向正在向可流式压缩与自适应加载演进。文献^[144]通过残差式场的构建支持自由视点视频的可流式传输;NeRFFCodec^[88]提供可块化与熵模型化的参数压

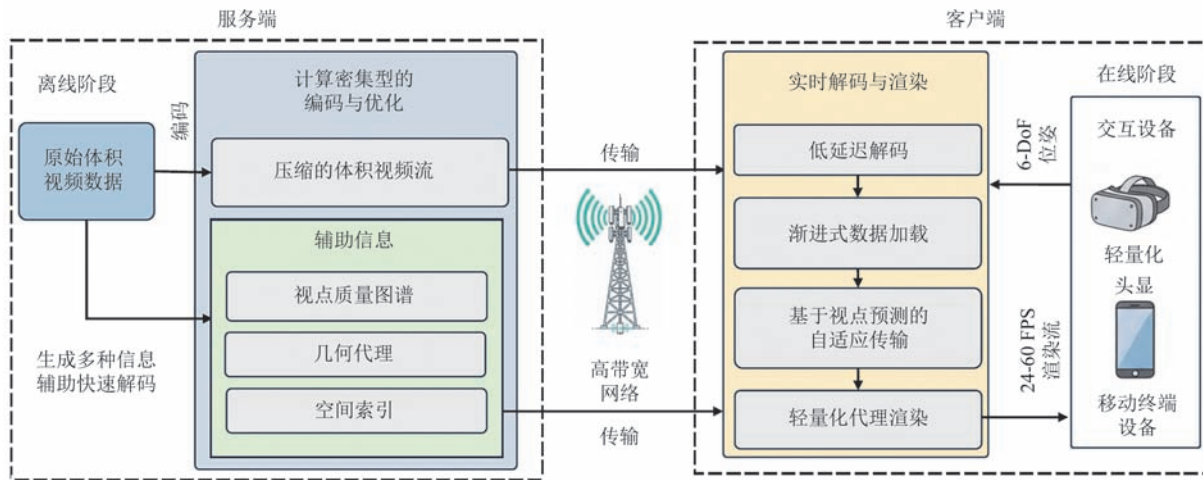


图7 VoD场景流程示意图

缩机制,使模型可被渐进传输;VRVVC^[145]引入可变码率训练机制,实现不同网络环境下的自适应比特流;而HPC^[146]则提出层级渐进编码结构,便于客户端根据带宽与视口动态分层请求加载。VoLUT^[147]提出将体积视频降采样后传输、客户端使用查询表驱动超分还原,显著减小带宽需求(最高可节省~70%带宽)且兼顾画质与渲染速度。

综上,VoD场景的性能需求主要体现在三方面:(1)在维持高压缩率与重建质量的基础上,极大降低客户端解码与渲染负担;(2)在严格时延预算内(通常低于几百毫秒)实现高帧率(24-60FPS或更高)自由视角渲染;(3)在网络波动与设备异构条件下实现稳定鲁棒播放,通过渐进、分层、可变码率等机制维持连续性与体验连贯性。

3.2.2 面向显式表示的点播编解码方法

在点播场景中,点云与网格作为主流显式三维表征形式,凭借其几何直观性、拓扑表达能力和标准兼容性,能够较好地融入现有流媒体传输与渲染管线中。对于点云表示,主流方法通常构建层次化结构(如差分二叉树、Differential Octree)并结合增量更新机制(如GROOT^[141]),通过对连续帧之间的结构或属性差异进行编码,辅助客户端并行解码以实现高效重建。渐进式传输机制(例如Yuzu^[142])则在服务端首先传输低密度版本的点云,客户端再借由学习驱动的上采样模块(PU-GAN^[148]、MPU^[149])填充细节,从而在带宽受限环境下仍可维持较好视觉质量。相较之下,网格表示因具有显式的顶点-边-面结构,其压缩设计得以利用顶点预测、拓扑简化、参数化映射等机制降低比特率,并借助GPU加速达到实时渲染性能。Draco库正因其低解码延迟和成

熟生态在交互系统中广泛采用^[30],而MPEG的FAMC标准则针对动态网格提出运动补偿与残差编码策略,适合具备较高计算资源平台^[26,150]。

近几年,为进一步提升显式几何表示在VoD链路中的适应性和性能,学术界提出了一些具有代表性的改进方向。例如,PatchVVC^[151]提出基于点云Patch的压缩框架,通过将体积视频划分为多个局部Patch,并对每个Patch投影至二维进行编码,实现了在保证较高压缩比的同时具备实时解码能力,且适合视域(Field of View, FoV)自适应流式传输场景。又如文献[152]在V-PCC Patch生成阶段引入基于形状相似性和遮挡关系的细分策略,从而缓解投影误差、提升率失真(Rate-Distortion, R-D)性能并缩短Patch生成时间。在V-PCC编码器层面,uvgVPCCenc^[153]为快速V-PCC编解码提供了工程可落地的实现,其设计目标即是兼顾编码效率与实时性,被视作点播场景中显式压缩的工程支撑平台。此外,深度压缩框架patchDPCC^[154]通过将点云划分为一致大小的Patch、利用帧间特征迁移模块压缩冗余,并在保证重建精度的基础上显著提升压缩比(与V-PCC相比平均提升47%-85%),为显式压缩注入端到端学习的潜力方向。

总体来看,点播场景下的显式表示压缩在标准成熟度、客户端渲染效率和低解码延迟方面具备显著优势;但在面对带宽波动、复杂运动与几何细节极端变化时,其压缩效率与灵活性仍显不足。因此,未来研究趋势是将渐进式加载、Patch自适应划分、运动补偿机制与学习式预测和插值模型融合应用,从而在压缩率、解码复杂度、渲染实时性与交互时延之间寻求新的平衡。

3.2.3 面向神经表示的点播编解码方法

与显式几何相比,神经表示在点播场景中提供了新的路径,其优势在于利用紧凑的参数集合刻画复杂几何与外观,从而在带宽与存储受限条件下提升传输效率。以NeRF为代表的隐式方法实现了高保真重建,但原始版本因训练与渲染缓慢而难以应用于实时点播。后续研究通过引入结构化先验与混合表征(如Instant-NGP^[104]的多分辨率哈希网格)显著降低时延,使其逐渐具备部署可行性。在点播流程中,典型方案分为两类:其一,服务端完成训练并压缩权重或特征网格,客户端仅需快速推理,减少传输与解码负担;其二,采用渐进式或分块式传输(如NR-PCC^[155]),客户端可按带宽逐层恢复,保证播放流畅性与灵活性。此外,部分方法尝试在端侧进行轻量化再训练或微调,以增强对动态视角和环境变化的适应性^[65]。

3DGS/4DGS通过高斯基元直接建模场景,结合GPU加速实现高效渲染,具备天然的渐进式解码

特性。相关工作通过参数量化与熵编码在低比特率下保持良好视觉质量^[69],并探索基于区域划分的分块传输,使客户端可仅加载与兴趣区域相关的参数集,在带宽与延迟受限的条件下依然维持交互体验。总体而言,神经表示在点播编解码中带来了带宽、质量与交互灵活性的平衡,但其在模型压缩、传输稳定性和端侧低时延推理方面仍存挑战,在异构硬件上端到端优化是未来研究的重点。

为了直观对比点播场景下显式与神经表示的编解码方法,表5总结了它们在传输策略、关键性能目标及优劣势上的特点。显式表示凭借其标准化与渲染效率优势,通过层次化、渐进式传输及预测补偿策略,在确保可适配性的同时平衡了带宽与时延。而神经表示,特别是优化后的NeRF及高斯溅射方法,则以其紧凑的参数化能力,在带宽受限下提供了高保真重建和交互灵活性的新范式,但仍面临模型压缩与低时延端侧推理的优化挑战。

表5 点播场景下显式与神经表示编解码方法对比

类别	具体方法/策略	核心特点与传输策略	关键性能目标	优势	挑战
显式表示	点云: 层次化/差分编码 (GROOT ^[141]), 渐进式传输 (Yuzu ^[142]) + 学习型上采样 (PU-GAN ^[148] , MPU ^[149])	逐帧增量更新, 并行解码; 服务端低分辨率传输, 客户端上采样恢复	高效重建; 带宽受限下高视觉质量; 降低延迟	标准成熟; 几何直观; 适配现有流媒体管线; 渲染效率高	带宽敏感下, 纯几何表示效率受限; 动态场景需结合学习模型
	网格: Draco ^[30] , FAM ^{C[26,15]} , MeshCNN AE ^[103]	低延迟压缩/解码; 利用拓扑关系降低比特率, GPU实时渲染	低延迟传输/解码; 实时渲染; 高渲染帧率	业界常用工具 (Draco); 高效渲染; 拓扑关系利于压缩	复杂动态场景压缩效率挑战
神经表示	NeRF/NeRF-like: Instant-NGP ^[104] , NeRFCodec ^[88] , VRVVC ^[145] , NR-PCC ^[155]	紧凑参数传输; 多分辨率哈希/渐进分块结构; 端侧快速推理与分层加载	高压缩率; 可变码率; 带宽自适应	高保真重建; 可流式加载; 参数传输效率优	模型压缩与熵编码不统一; 端侧推理耗时与功耗高
	JointRF ^[156] , Neural Residual Radiance Fields ^[144]	动态辐射场残差编码与联合优化; 支持流式自由视点视频传输	流式自由视点重建; 多帧时域一致性	码率控制精准; 端到端可训练	训练代价高; 时序漂移对交互鲁棒性影响大
	3DGS/4DGS: FCGS ^[69] , Hierarchical Compression for 3DGS ^[68]	高斯基元直接渲染; 参数量化与层次化渐进加载; GPU实时解码	高保真渐进重建; 低码率传输; 高帧率渲染	GPU管线兼容性强; 视觉质量领先NeRF同类方法	模型体积控制与动态场景压缩仍具挑战

3.2.4 交互式点播的实时编解码特性分析

在交互式点播场景中,率失真、复杂度与时延的表现与离线编解码形成显著对比。显式几何方法依托成熟标准和硬件加速,能够在中低复杂度下提供稳定的压缩与渲染性能。例如V-PCC在低带宽条件下仍能维持可接受的几何与纹理质量,适合标准化与兼容性要求较高的应用。然而,随着场景复杂度增加,其码率与存储开销迅速攀升,导致端到端时延与资源消耗难以满足严格的交互需求。

神经表示在率失真特性上展现更大潜力。

NeRF及其快速变体通过紧凑参数高效拟合复杂几何与视角效应,使其具备在VoD系统中的可行性,但客户端解码仍依赖较强算力,限制了轻量终端的普适性。3DGS/4DGS进一步改善不同特性之间的平衡,通过GPU加速实现实时渲染,并通过量化与熵编码降低参数存储需求,在动态场景和交互灵活性上优于显式几何与传统NeRF。然而,其庞大的参数规模若缺乏高效压缩,仍会带来存储与传输负担。总体而言,显式几何方法强调稳定与兼容,神经方法突出高保真与灵活性,而高斯溅射在

实时性与视觉质量间提供折中。未来的发展趋势可能是混合式方案,结合显式结构的高效性与神经表示的强表现力,在不同特性维度间实现动态优化。

3.3 面向低时延交互直播的实时编解码

相比交互式点播,直播场景更强调端到端低时延、帧率稳定性与多模态同步,通常需要在百毫秒级内完成采集、压缩、传输与渲染,以保障自然交互与沉浸体验。同时,系统还需在动态场景和带宽波动条件下保持长期稳定运行,因此直播编解码更强调围绕率失真、复杂度与时延展开跨层协同优化,这也不同于点播场景中离线重压缩、在线低复杂度的非对称模式。

3.3.1 直播场景特点与性能需求

直播体积视频广泛应用于远程会议、演出、体育和沉浸式社交,其核心需求在于端到端低时延与视听一致性:采集—重建—编码—传输—解码—渲染必须压缩在百毫秒级内,避免交互卡顿与运动-视觉错位。如图8所示,给出了沉浸式体积视频直播技术路线图。同时,系统需在动态内容与网络波动下维持稳定帧率与可接受质量。相比点播,直播更强调在线可更新的表示与轻量化编码,要求具备渐进、差分 and 并行特性,并与低时延传输协议(如WebRTC)及端侧GPU渲染深度协同。在有限算力与能耗约束下,如何在同一时延预算内提升率失真性能,成为算法与系统设计的关键挑战。



图8 沉浸式体积视频直播技术路线图^[12]

3.3.2 面向显式表示的直播编解码方法

在直播系统中,点云仍是核心研究方向。基于Draco的轻量化几何与属性压缩因解码效率高、生态成熟,被多种实时原型验证有效^[30,157-160]。为降低延迟与带宽开销,研究者提出帧间差分与属性渐进更新等策略,使时间冗余能够快速编码传输。学习驱动方面,如AITransfer^[161]和Zhu等系统^[162],通过端到端语义特征提取、渐进式重建和自适应网络控制,在重建质量与鲁棒性上取得进展,并通过剪枝和量化适配移动终端^[6]。

网格表征在直播场景中更适合对几何与纹理精度要求较高的应用。几何与纹理分离的混合编码已

成为主流:几何通常采用Draco或轻量无损压缩,纹理通过UV展开后交由H.264/HEVC/AV1等视频编码器处理^[163-165]。相关系统如Jin等^[166]、Ma等^[167]Avatars与Orts-Escolano等^[168],均展示了几何压缩与GPU加速渲染结合的低时延链路,而文献^[169]强调了UV参数化稳定性与时间一致性对整体性能优化的重要性。

3.3.3 面向神经表示的直播编解码方法

直播场景下神经表示的核心难点,在于在线建模、压缩和渲染需要在严格时限内协同完成。多数快速神经方法在点播场景中往往依赖离线优化或较长时间的微调,这种条件在直播中通常并不具备。现有研究主要分为两条路径:一是服务器端渲染(Server-Side Rendering, SSR),由云端完成重建与渲染,再通过传统视频编码低时延传输,避免了终端侧的神经解码压力,但牺牲了6-DoF自由交互能力^[50,170];二是可流式更新的紧凑神经表示,即对NeRF/3DGS等模型进行增量训练,仅传输参数增量或潜在编码,并在端侧快速合并与渲染,从而在一定程度上平衡延迟与交互能力^[131,171-172]。

在可实时渲染的表示中,3DGS通过与GPU光栅化管线的高度适配和高帧率渲染能力,被认为是最具潜力的直播候选。其围绕参数量化、稀疏剪枝、差分与熵编码的压缩策略,已在离线场景显著缩减模型体积,为直播中的低码率传输与快速更新提供了基础^[49,62,134,168,173]。然而,无论是SSR还是流式更新方法,在高动态场景下要同时实现高保真、低码率和百毫秒级时延,仍面临模型稳态训练与带宽敏感性控制等难题,尚缺乏统一且可复现的方案。

表6总结对比了直播场景下显式与神经表示的编解码方法在传输机制、核心目标、主要优势及挑战。显式表示通过帧间差分、渐进更新或视频编码复用,实现了低时延传输,但处理复杂形变仍是瓶颈。神经表示则通过云端渲染或流式增量更新来应对在线建模挑战,其中3DGS因其渲染效率而展现出巨大潜力,有望在带宽、质量和交互之间取得新平衡,但其工程化落地仍需克服高动态场景下的稳态训练、带宽敏感性与端到端优化难题。

3.3.4 低时延直播的实时编解码特性分析

如表7所示,在严格时延约束下,显式与神经表示展现出互补的特征。基于Draco的点云与网格方案依托成熟标准与并行友好结构,在有限算力与抖动带宽下可维持稳定帧率和低端到端延迟^[30,157,159,166],但在非刚性形变、细粒度纹理与视角相

表 6 直播场景下显式与神经表示编解码方法对比

类别	具体方法/策略	核心特点与传输策略	关键性能目标	优势	挑战
显式表示	点云和网格两行排版对齐;Draco ^[30] , 帧间差分,属性渐进更新,学习 驱动 ^[161-162]	轻量化几何/属性压 缩;语义特征提取,自 适应网络	低延迟传输;带宽开 销降低;鲁棒性与重 建质量提升	解码效率高,生态成 熟;学习方法改善视 觉效果	学习方法依赖强算力 和数据;剪枝/量化适 配性
	网格;Draco ^[30] 轻量无损几何, H.264/HEVC/AV1纹理 ^[163-165] , GPU加速系统 ^[166-168]	几何与纹理分离混合 编码;UV展开纹理	高几何/纹理精度; 低时延链路;实时 渲染	复用视频编解码与渲 染生态;GPU加速渲 染效率高	非刚性形变/拓扑变化 下几何压缩复杂;UV 参数化稳定性
神经表示	服务器端渲染(SSR) ^[50,170]	云端完成重建与渲染, 传统视频编码低时延 传输	避免终端侧神经解 码压力;低时延视频 传输	终端侧负载轻;利用 成熟视频编码技术	牺牲6-DoF自由交互 能力;渲染分辨率受 传输带宽限制
	可流式更新的紧凑神经 表示 ^[131,171-172]	增量训练/更新模型或 潜在编码;端侧快速合 并与渲染	平衡延迟与交互能 力;高帧率渲染;低 码率传输	高帧率渲染,与GPU 管线适配;高保真;模 型体积潜力大	高动态场景稳态训练 困难;带宽敏感性;工 程化方案缺乏

关外观的表达存在上限。学习驱动的显式方法(如AITransfer及其系统化扩展)改善了同码率下的几何与属性重建,并通过非对称编解码降低端侧负担,但对训练数据与模型泛化依赖较强,导致跨场景直播中的稳健性与一致时延控制仍存瓶颈^[6,161-162]。神经表示在重建质量和动态适应性上优势显著,其中

3DGS凭借GPU渲染管线,在实时性与端侧复杂度之间达成较好平衡,相关的参数压缩与流式更新研究正推动其向可部署落地演进^[63,130,174-175]。相比之下,NeRF及其加速变体即便结合显式结构优化,仍需更高效的在线学习与参数更新机制,才能在不牺牲延迟的前提下实现直播级协同^[131,170-172]。

表 7 直播场景典型编解码方法的特性对比

方法	率失真	复杂度	时延	典型代表
显式:点云(规则型)	压缩效率受限,细节表达不足	算法简洁,硬件友好	低延迟,易实时解码	[30,157] [159,166]
显式:点云(学习型)	更优几何/属性重建质量	编码复杂度较高,解码端可轻量化	满足实时约束,但需算力支撑	[6,161] [162]
显式:网格(几何+纹理分离)	纹理保真度高,几何透度有损	利用视频编解码器,复杂度受控	端到端低延迟,依赖UV稳定性	[63,174] [175]
神经: NeRF系列	高保真建模,动态适应性强	训练/推理成本高	难以满足百毫秒级约束	[171,172] [170]
神经: 3DGS/4DGS	重建质量与实时渲染平衡良好	解码高效,依赖GPU渲染	接近实时,但更新仍具挑战	[63,130] [131]

低时延直播更适合采用显式表示与神经表示相结合的分层设计:主干链路利用显式结构保证较低复杂度和稳定吞吐,关键视域或语义重要区域则引入神经增强,并结合参数量化、差分更新和自适应带宽调度,实现性能与时延之间的平衡。

4 跨层优化与系统实现

4.1 编解码器与模型侧优化

在体积视频系统中,编解码器与模型设计直接决定了压缩效率、解码时延与重建质量。仅依靠单一压缩算法往往难以满足实际部署中带宽受限与终端异构约束,因此研究逐渐转向跨层优化:通过在编

码、模型与系统层之间协同设计,在有限资源下实现高效压缩、低延迟解码与高质量重建。现有工作可归纳为三类:空间与结构优化、感知驱动与内容自适应优化,以及终端侧增强与计算优化。

4.1.1 空间与结构优化

空间分割与分块传输是提升编解码效率的基础手段。通过将体积视频划分为可独立管理的子单元(如空间图块、面片或三维子块),服务端可并行编码,客户端则依据视点按需加载,从而在带宽与算力受限条件下实现资源高效利用,同时具备错误隔离与CDN部署优势^[176-184]。在不同表征下,这一机制有多样化实现:点云常用立方体划分或八叉树组织^[185,186],网格可拆分为子网格或面片,体素天然适

配三维子块^[177-178], MPEG PCC 标准也已将分块机制纳入以增强流媒体适配性^[182,184]。其主要难点是块边界的视觉不连续与元数据开销^[187-190]。

在此基础上,分层编码与可伸缩流媒体提供了更灵活的结构优化思路。通过基础层与增强层的层次化码流设计,系统可根据带宽、终端性能与用户需求灵活选择解码层级^[191]。具体而言,空间可伸缩性支持多分辨率自适应^[192],质量可伸缩性逐步叠加几何与纹理细节^[193],复杂度可伸缩性则让解码路径随算力动态调整。不同表征形式均可结合分层机制:点云可由稀疏到稠密逐步增强^[194],网格通过拓扑细化或位移贴图提升精度^[195],体素依赖多分辨率特性^[196],而神经表示则借助分层特征或渐进加载实现渐进重建^[197]。

4.1.2 感知驱动与内容自适应优化

在带宽与算力受限的条件下,感知驱动策略通过用户行为与场景语义实现自适应调度,成为提升效率的重要途径。视口自适应流媒体通过追踪头部姿态与视线方向,并结合预测模型优先加载视野相关区域,从而保证用户所见范围的高质量呈现,同时减少全局解码与渲染开销^[198-200]。在神经表示中,这可通过局部参数动态加载、特征裁剪或激活特定高斯子集实现^[172,201]。然而,预测误差可能导致未加载区域出现黑边或模糊,而突发视角切换也对预取与延迟控制提出更高要求。

进一步的优化方法利用场景语义的重要性将有限资源优先分配给用户关注的主要人物或交互对象,而对背景区域采用更高压缩率或较低优先级^[162,202-204]。具体实现上,语义权重可通过目标检测、显著性预测或语义分割等技术映射到三维数据结构(如图块、面片或高斯组件),再据此差异化压缩与传输。这类方法在资源受限下显著提升了主观体验,但在低延迟语义提取、跨表征统一映射以及个性化权重建模方面仍存在挑战^[137]。

4.1.3 终端侧增强与计算优化

在资源受限场景中,客户端增强成为缓解传输与解码瓶颈的重要手段。典型做法是通过轻量化的深度学习模型执行超分辨率与细节补偿:服务端仅传输低分辨率或高压压缩率数据,终端在本地重建几何与纹理细节,从而实现“以算力换带宽”的优化路径。已有研究表明,生成式AI模型在此方面具有巨大潜力,这类方法在点云、体素和神经表示等多种表征下均能有效提升视觉质量,例如PU-GAN^[148]能通过生成对抗网络从稀疏点云恢复完整表面。

3D-FM GAN^[205]将输入人脸及其三维编辑结果映射到StyleGAN隐空间,以实现对人脸体积视频的三维可控操纵与合成,通过这种方式,只需对潜在特征进行编辑,便可在接收端生成具有目标视角或表情的逼真人脸序列,在虚拟交互等应用中减少了传输完整体积视频数据的开销。Yuzu系统^[142]通过客户端上采样增强点播体验。而针对体素与隐式场的工作则利用特定网络实现高分辨率几何与属性重建^[147]。GaussianSR^[206]是一个代表性实例,它将从海量2D图像训练的扩散模型知识通过SDS蒸馏到3D高斯表示中,对低分辨率的体积场景进行超分辨率重建,从而在低码率下依然恢复出高清细节。S2Gaussian^[207]仅使用稀疏和低分辨率的视图超分辨率重建结构精确和更具细节的3D场景。总体而言,生成式AI方法与NeRF、3DGS等现有技术在终端侧增强与计算优化方面形成了良好的互补,弥补了传统方法在细节恢复和灵活编辑上的不足,通过在解码端引入智能生成来提升最终重建质量和沉浸感。

然而,终端增强仍面临关键挑战。其一,推理延迟必须控制在实时范围,尤其是在低时延直播中;其二,增强结果需兼顾真实性与泛化能力,否则可能引入伪影或失真,降低沉浸感^[208];其三,模型的存储与更新对移动和嵌入式设备提出额外负担。未来的研究方向在于设计更高效、鲁棒且适应性强的增强模型,并结合硬件加速与跨端协同机制,使终端侧在带宽受限和能耗受控条件下依然能够保障沉浸式体验。

如表8所总结,空间与结构优化、感知驱动与内容自适应优化,以及终端侧增强与计算优化这三类主要方法,分别从核心目标、具体实现机制、显著效益及其面临的局限性与挑战等方面进行了归纳。这些跨层优化手段通过协同设计,力求在资源受限环境中实现体积视频的高效压缩、低延迟传输和高质量重建,以支撑多样化沉浸式应用场景的需求。

4.2 流媒体传输与网络协议优化

为系统性地阐明体积视频在流媒体传输与网络协议层面的优化,表9中对比了传输协议与网络机制优化以及内容感知编码优化这两大类策略。该表详细分析了各类优化的核心目标、具体实现、所带来的关键效益以及当前面临的挑战。这些优化手段共同致力于在带宽受限且对时延高度敏感的网络环境中,实现体积视频的低延迟、高吞吐量与鲁棒传输,进而保障高质量的沉浸式体验。

表 8 编解码器与模型侧优化策略对比					
类别	具体优化策略	核心目标与思想	主要实现机制/应用场景	主要优势/效益	挑战/限制
空间与结构优化	空间分割与分块传输	划分为可独立子单元,实现并行编码与按需加载	点云;立方体/八叉树 ^[185-186] ,网格;子网格/面片,体素三维子块 ^[187-188] ,MPEG PCC 标准 ^[182,184]	带宽/算力高效利用;CDN 友好部署;3DGS 层级加载与渐进渲染	块边界视觉不连续;元数据开销;高斯块索引与层级细节 (Level of Detail, LoD)切换导致跳变
	分层编码与可伸缩流媒体	根据带宽、终端算力与用户需求灵活选择解码层级	空间可伸缩性 ^[192] ,质量可伸缩性 ^[193] ,复杂度可伸缩性;应用于点云 ^[194] 、网格 ^[195] 、体素 ^[196] ;神经表示与渐进加载 ^[197]	多分辨率自适应;兼容 3DGS 渐进式高斯加载与多层结构	多维联合优化复杂;层级元数据冗余;层间平滑过渡难以控制
感知驱动与内容自适应	视口自适应流媒体	追踪用户行为,优先加载视野区域,减少全局解码渲染开销	头部姿态/视线预测;Light4GS ^[131] ;神经表示中局部参数加载、特征裁剪或高斯子集激活 ^[172,201]	所见区域高质量呈现;显著减少带宽/算力消耗;支持动态 4DGS	预测误差导致黑边/模糊;突发视角切换对预取与延迟控制要求高
	场景语义优先级调度	根据场景语义重要性,将有限资源优先分配给关注区域	语义权重映射至图块、面片或高斯组件;差异化压缩与传输 ^[162,202-204]	在有限带宽下优先保障主要对象质量;与语义解码/调度联动优化主观体验	低延迟语义提取难;语义与几何层映射不统一;个性化建模复杂
终端侧增强与计算优化	客户端超分辨率与细节补偿	服务端传输低分辨率数据,终端本地重建细节,实现以算力换带宽	轻量 DL 模型,如 YuYu ^[142] ;PU-GAN ^[148] ;端侧重建体素/隐式场 ^[147]	缓解传输与解码瓶颈;提升视觉细节;端侧加速结合 GPU 管线实现高帧率渲染	实时推理延迟受限;增强真实性与泛化能力不足;模型更新/设备负担较高

表 9 流媒体传输与网络协议优化策略对比					
类别	具体优化策略	核心目标与思想	主要实现机制/应用场景	主要优势/效益	挑战/限制
传输协议与网络机制	基于 UDP 的实时传输协议	实现低延迟、高吞吐量,规避 TCP 重传延迟	WebRTC ^[209] 、SRT ^[210] 、RIST ^[211]	端到端低时延;异构网络鲁棒性;支持多路径自由视点交互	协议复杂性高;安全与拥塞控制机制仍待标准化
	自适应比特率 (Adaptive Bitrate, ABR) 控制	根据网络状态动态调整数据层级,避免卡顿,提升流畅性	可伸缩码流;Neural-ABR ^[212] ;结合强化学习预测带宽变化	保持播放连续性;避免卡顿;动态质量一致性	带宽预测精度有限;层级切换平滑性与实时性平衡
	内容分发网络 (Content Delivery Network, CDN) 部署	降低访问延迟与服务器负载,支持大规模分发	缓存边缘节点 ^[213] ;结合语义分块调度	低延迟大规模分发;支持视口驱动内容传输	缓存一致性与多用户协调复杂;边缘部署成本高
内容感知编码优化	显著性引导编码	利用视觉注意偏好,优先保障用户感知敏感区域	运动主体、面部、纹理丰富区域高比特分配;背景强压缩 ^[214-217]	显著提升主观视觉质量;有损码率下资源优化	显著性预测准确性与实时性;跨场景泛化能力
	语义驱动编码	检测与分割关键实体,优先保障其高保真还原	人脸、手部、交互对象高比特分配;次要区域开销降低 ^[218-222]	保证关键实体清晰度;优化远程会议/虚拟交互体验	低延迟语义提取;跨表征统一映射;个性化权重建模
	运动特性自适应编码	根据运动速度/类型分配比特,保证动态场景真实感	快速运动/非刚性变形区域高比特;静态/缓慢变化区域长时预测/低频更新 ^[223-226]	保证动态场景连续性与真实感;有效节约带宽	非刚性形变建模复杂;运动预测精度;刚性与非刚性差异化处理

4.2.1 传输协议与网络机制

网络传输协议的选择与优化是实现低延迟、高吞吐量、可交互体积视频流的关键。与传统二维视频相比,体积视频在网络侧的核心矛盾不仅是数据量更大,更在于其 6-DoF 交互带来的请求与渲染闭环,用户视角与位姿变化会驱动客户端在短时间内

对特定空间区域发起突发请求。同时,体积内容往往由可独立解码和渲染的空间单元构成(例如点云、网格的分块或 patch、体素子块以及神经表示中的局部特征块或高斯子集等)。这些需要网络侧支持按需到达甚至可渐进细化的传输模式。因此网络传输协议除了追求低时延,还需与体积视频的分块或分

层结构、重要性标注与端侧渲染时延预算相互协同,对基础可用的关键数据(如基础层、关键几何纹理、块索引与时序元数据)优先保障可达性,对增强细节则允许一定丢失以避免交互卡顿。

基于上述特点,系统往往优先采用基于UDP的轻量级实时传输协议,以避免TCP重传引入的排队与头阻塞延迟,更契合体积视频交互对时延抖动的敏感性,如WebRTC^[209]、SRT^[210]和RIST^[211]。这些协议在保障数据传输时效性的基础上,进一步集成了丢包恢复、带宽估计与拥塞控制机制,从而在异构网络环境中展现出良好的鲁棒性与适应性。在体积视频场景中,这类协议还便于实现分级可靠性与优先级调度,可针对不同重要性的体积数据单元采用差异化策略。例如对基础层或关键块使用更强的丢包恢复(选择性重传等)与更严格的时延上界控制,而对增强层或远离当前视锥的细节块采取更激进的丢弃与快速替换,从而在丢包或带宽波动时优先维持可交互的连续体验,再逐步补齐细节质量。该策略本质上将体积视频可渐进、局部更新的内容组织方式映射到传输层的可靠性与调度策略上,是典型的跨层协同。

在应用层,传输机制通常与自适应比特率(Adaptive Bitrate, ABR)控制结合,以根据实时网络状态动态调整传输负载并维持体验稳定性^[212]。在体积视频场景下,ABR决策对象不应仅是码率档位,而应面向其三维内容结构进行调度。例如,在可伸缩编码或分块结构下,客户端可根据带宽与时延预算选择不同LoD、不同空间块、不同层级进行请求与解码,确保视锥内区域优先达到可用质量和服务的连续性,并在资源允许时再请求增强层以提升细节一致性。与此同时,在大规模分发场景中,内容分发网络(Content Delivery Network, CDN)仍是降低访问延迟与服务器负载的基础设施^[213],但体积视频的CDN使用方式同样具有内容结构特征。缓存对象不再只是线性时间分片,而更适合以空间、patch层级为粒度进行组织,并结合视口驱动或语义感知策略进行边缘侧的预取与调度,以应对视点切换导致的突发请求与热点区域聚集。综上,这些面向体积视频的传输协议与网络机制的协同应用,使得体积视频能够在受限网络环境中实现可扩展、低延迟且高鲁棒性的实时传输。

4.2.2 内容感知编码优化

除传输机制外,近年来相关研究还引入了内容感知的优化思路,以弥补传统无差别压缩策略的不

足,并改进比特分配与流媒体调度方式。这类方法通过刻画三维场景的语义结构、视觉显著性和运动特征,使系统能够在有限码率条件下优先保障对用户感知更为敏感的内容,从而提升主观质量。

显著性引导方法利用人类视觉系统的注意偏好,在运动主体、面部或纹理丰富区域投入更多比特,而对背景采用更强压缩,从而显著提升主观体验^[214-217]。语义驱动方法则通过检测与分割关键实体(如人脸、手部或交互对象),在远程会议或虚拟交互中保证高保真还原,同时降低次要区域开销^[218-222]。此外,运动特性也是重要维度:快速运动或非刚性变形区域需分配更多比特以保证连续性与真实感,而静态或缓慢变化区域则通过长时预测和低频更新节约带宽^[137, 223-226]。除了显著性与语义,运动特性同样是编码优化的重要维度。对于快速运动或非刚性形变区域,传统预测方法往往难以建模,容易造成模糊或失真,因此需要投入更高比特率以保证连续性与真实感。相比之下,静态背景或缓慢变化区域则可通过长时预测与低频更新减少数据量,从而有效节约带宽。刚性与非刚性运动的差异化处理方式被提出,例如通过刚体的6-DoF变换参数实现简化编码,而对复杂非刚性变形采用更精细的时空建模,从而在有限资源下取得更优的平衡。

总结来看,内容感知优化提供了自上而下的智能调控机制,使系统不再仅依赖网络协议与分层结构的被动适应,而是主动结合用户注意力与场景特征实现资源分配。

4.3 客户端渲染与计算优化

在体积视频系统中,客户端既是解码结果的呈现端,也是承担大量计算任务的关键环节。相比传统二维视频,其在三维重建、自由视点选择与交互式渲染上的需求显著增加了终端算力压力,而不同设备(移动端、头显、桌面平台)的计算能力差异又加剧了异构性。因而,如何在客户端侧平衡计算负荷、降低延迟并保障渲染质量,成为系统优化的重要方向。本节讨论了基于学习的超分辨率与细节增强、渲染管线的优化与轻量化设计,以及软件层面的协同机制三类优化策略。表10对客户端计算和渲染中的主要优化策略进行了对比。

4.3.1 客户端超分辨率与细节增强

受带宽和传输效率限制,服务端通常需要以较低分辨率或较高压缩率传输体积视频。为补偿由此带来的质量损失,客户端超分辨率和细节增强方法被用于在终端侧改善重建结果。借助轻量化深度模型,

表 10 客户端渲染与计算优化策略对比					
优化策略类别	典型策略	核心目标与思想	优势	挑战/限制	主要应用场景
传输协议与网络机制	基于学习的超分辨率 ^[142,147-148,172]	通过 AI 模型在客户端恢复高分辨率细节	提升视觉质量，以算力换带宽	推理延迟、伪影、模型部署(存储/能耗)	低带宽传输、高沉浸需求
	渐进加载与局部高分辨率重建 ^[172]	分阶段传输与渲染，优先显示关键区域	快速初始化、缓解卡顿、提升交互体验	数据分层管理、无缝过渡	大型场景漫游、VR/AR 交互
内容感知编码优化	GPU 光栅化与子网格分片 ^[181]	利用 GPU 并行能力高效渲染网格	高效处理显式几何、渲染吞吐量高	仅限于网格/几何体，数据量大时仍有瓶颈	几何模型为主的场景、游戏
	稀疏数据结构与八叉树索引 ^[187-188]	减少计算冗余，优化点云/体素处理	降低计算复杂度、节省内存	数据结构维护开销、不规则数据处理	大规模点云/体素渲染
	神经表示加速 ^[172,197,201]	提升 NeRF/3DGS 等神经表示的渲染效率	实时渲染、高质量三维重建	模型训练复杂、内存占用、泛化性	自由视点视频、神经渲染应用
	异步调度与任务流水线 ^[227]	并行处理数据加载、解码、渲染等任务	掩盖延迟、提高整体响应速度	任务依赖管理、同步开销	实时交互、多任务并发
	SIMD 指令、高性能库(MKL, cuDNN 等) ^[228-229]	利用底层硬件特性加速数值计算	高效利用 CPU/GPU、显著降低计算时间	依赖特定硬件架构、库的集成与维护	任何计算密集型任务、AI 推理
软件层面优化与软硬件协同	多线程并行与性能分析工具 ^[227,230,231]	并发执行任务、识别性能瓶颈	充分利用多核 CPU、精确指导优化	线程同步、死锁风险、工具的准确性	复杂系统优化、性能调试

终端可将稀疏或低质量数据恢复为更高密度的点云、更精细的几何网格或纹理更丰富的体积场，从而提升用户感知质量。典型工作包括：PU-GAN^[148]从稀疏点云生成结构完整表面，Yuzu 系统^[142]在客户端上采样改善视觉质量；针对体素与隐式函数，已有方法利用神经网络实现低分辨率场到高分辨率场的补全^[147]；而在 NeRF、3DGS 等神经表示中，则通过渐进加载或局部高分辨率重建提升视觉体验^[172]。

尽管该策略有效平衡了资源利用与质量提升，其应用仍受制于三方面挑战：其一，推理延迟需满足实时交互与直播的严格预算；其二，增强结果必须保持真实性与泛化能力，避免伪影破坏沉浸感；其三，模型部署需兼顾移动终端的存储与能耗限制^[208]。因此，未来研究方向在于优化轻量化网络结构与高效推理框架，并结合端侧 AI 加速硬件，实现低功耗条件下的高质量增强。

4.3.2 渲染管线与轻量化优化

渲染作为客户端的核心任务之一，直接决定了用户在自由视点条件下的沉浸体验。体积视频的渲染往往涉及点云投影、网格光栅化、体素体绘制，以及基于 NeRF 或 3DGS 的光线追踪与高斯溅射，这些过程均为计算密集型任务。为了提升实时性能，研究者提出多种渲染管线优化与轻量化策略。

在显式几何表示下，利用 GPU 光栅化管线可高效完成网格渲染^[181]，同时通过子网格分片与并行绘制提高渲染吞吐量。在点云和体素渲染中，采用稀

疏数据结构、八叉树索引与并行数据流处理可有效降低计算冗余^[187-188]。对于隐式与神经表示，渲染效率优化的需求更为迫切。NeRF 原始模型渲染效率较低，但通过稀疏体素加速、哈希编码和浅层网络融合等方法显著缩短了推理时间^[197]。3DGS 则通过 GPU 加速实现实时渲染，其在客户端推理速度与可交互性上展现出优越性能^[172,201]。

除了算法层面的轻量化优化，异步调度和任务划分同样在客户端渲染管线中发挥重要作用。通过将数据加载、解码、光栅化和光线步进等任务分解为独立流水线并并发执行，可以显著掩盖单个环节的延迟，提高整体响应速度^[227]。此外，轻量级的缓存与预测机制能够进一步降低交互时的卡顿风险，提升系统鲁棒性。

4.3.3 软件层面优化与软硬件协同

为了充分释放终端硬件潜能，软件层面的优化与软硬件协同设计不可或缺。与二维视频以连续像素块处理为主不同，体积视频在端侧往往需要在毫秒级时延预算内完成几何和属性(或神经参数)的解码与解包、空间索引、可见性裁剪(视锥裁剪、遮挡剔除、LoD 选择)以及面向显式几何或神经表示的实时渲染(点云/网格光栅化、体素体绘制、NeRF 推理或 3DGS)。因此，数据结构与算法实现必须充分考虑底层架构特性和体积数据的组织方式。例如，通过内存连续布局提升缓存命中率，利用单指令多数据流(Single Instruction Multiple Data, SIMD)指令

加速底层数值计算,减少冗余数据拷贝与格式转换等方法,均可有效降低处理延迟^[228]。同时,体积视频系统中常见的矩阵、向量计算与神经网络算子(例如视点预测、轻量增强、NeRF 或 3DGS 相关推理模块)可直接复用主流硬件厂商提供的高性能库(如 Intel MKL、Arm Compute Library、NVIDIA cuDNN/cuBLAS/Thrust),为端侧优化提供了便利^[229]。

在复杂任务场景中,体积视频端侧性能更依赖异步流水线与并行调度来稳定帧率与交互延迟。例如,在头显设备中,可通过并行线程同时处理姿态追踪、数据解码和渲染,从而避免阻塞并保持低延迟交互^[229-230]。此外,体积视频常见的热点瓶颈并非单纯算术运算,而是 GPU 显存带宽、CPU 与 GPU 数据传输以及渲染阶段的排序/裁剪/批处理(例如高斯的可见性剔除与按深度组织)导致的同步与内存压力。因此,引入性能分析工具有助于识别关键性能瓶颈,从而进行针对性优化^[231]。

未来的发展趋势是实现端侧软硬件的深度协同。

例如,可在算法设计阶段面向特定体积视频 AI 加速器(如移动 NPU)调整网络结构,也可在硬件开发阶段原生支持体积视频中的典型操作,如稀疏点云处理、光线追踪和几何变换,从而实现算法与硬件之间的协同适配^[229,232]。这种协同路径不仅有助于在低功耗设备上实现高质量渲染与实时交互,也为支持新兴的神经表示与混合表征提供了可持续的技术基础。

4.4 云边端协同与硬件加速

随着体积视频规模的不断扩大以及实时交互需求的增强,单纯依赖客户端计算或服务器集中处理都难以在延迟、能耗和带宽之间取得理想平衡。云边端协同架构通过充分利用不同层级计算与存储资源的互补性,结合专用硬件加速技术,成为推动沉浸式体积视频系统落地的重要技术路径。本节从并行处理与多线程优化、专用硬件加速与新兴架构,以及云边端协同计算三个角度讨论上述优化策略的性能特点,表 11 中对云边端协同与硬件加速的主要策略进行了对比。

表 11 云边端协同与硬件加速策略对比

优化策略类别	典型策略	核心目标与思想	优势	挑战/限制	主要应用场景
并行处理与多线程优化	GPU 并行计算 (CUDA, OpenCL) ^[230]	利用 GPU 大规模并行能力加速渲染/AI 推理	实时渲染、神经表示加速、高吞吐	数据传输瓶颈、内存管理复杂	实时渲染、NeRF/3DGS、AI 推理
	CPU 多线程 (OpenMP, TBB) ^[233]	利用多核 CPU 并行处理可分解任务	提升预处理/解码吞吐量,通用性强	线程同步开销、SIMD 利用效率	离线处理、高吞吐量需求
	硬件编解码器 ^[234-235]	利用专用芯片高效编解码视频序列	极低功耗、高效率	仅限于视频流格式,非通用三维数据	属性信息视频化、低功耗设备
专用硬件加速	FPGA 加速 ^[236]	可重构逻辑实现特定算法加速	高度定制化、实时性强、能效比高	开发难度大、通用性相对较低	多视角数据处理、几何变换
	AI 加速器 (NPU, TPU, DPU) ^[237]	针对 AI 运算优化,加速神经网络推理	高效推理、低能耗、提升 AI 任务性能	依赖特定模型、部署复杂性	视口预测、内容感知编码、超分辨率
	新兴三维数据专用硬件 ^[238]	针对点云/体素等三维数据定制硬件	突破传统瓶颈、高吞吐、低延迟	研发成本高、生态不成熟	大规模点云处理、体素级计算
	硬件编解码器 (ASIC for H. 26x/AV1) ^[234-235]	利用专用芯片高效编解码视频序列	极低功耗、高效率	仅限于视频流格式,非通用三维数据	属性信息视频化、低功耗设备
	云端卸载与高保真渲染 ^[239-240]	云端承担重计算、渲染结果流传输	减轻终端负担、低门槛高保真体验	网络延迟敏感、数据传输量大	离线重建、云 XR、高品质内容制作
云-边-端协同计算	边缘计算(多视角融合,局部推理) ^[177,240-241]	边缘节点处理时延敏感任务	降低端到端延迟、提高鲁棒性	边缘资源有限、任务调度复杂	远程协作、云 XR 游戏、交互式会议
	任务动态调度与服务质量 (Quality of Service, QoS) 保障 ^[232]	根据网络/设备状态智能分配任务	灵活部署、资源最优利用、服务质量	调度算法复杂、系统开销	动态交互场景、异构网络环境

4.4.1 并行处理与多线程优化

体积视频编解码与渲染流程包含大量可并行化的子任务,例如点云子集处理、网格分片渲染、神经表示查询、几何变换编码与量化等。然而,其并行优化的关键不是简单加线程,而是并行粒度需要与体

积表示的空间组织单元一致。典型地,显式表示(点云/网格/体素)往往以空间块、patch、八叉树节点或子网格作为最小可调度单元;神经表示(NeRF、3DGS/4DGS)则常以特征网格块、哈希表分辨率层、或高斯簇/锚点集合等作为可增量更新与渲染的

基本对象。基于这些空间可分割单元组织任务队列与工作调度,有助于并行计算同时适应体积视频中的两类要求:一方面,视点切换带来的突发访问需要低时延响应;另一方面,可见区域的质量需要具备渐进细化能力。

现代多核CPU和GPU的并行计算能力在此过程中发挥了关键作用。在CPU层面,多线程框架(如OpenMP、Intel TBB等)适合用于吞吐与延迟敏感的准备阶段任务:例如块/patch级的码流解析与熵解码、八叉树/层次结构遍历、几何与属性的去量化与坐标变换、以及视锥/遮挡剔除所需的空间索引更新等。此类任务通常呈现不规则控制流与分支较多的特点,若以“空间块/节点”为单位进行任务切分,可降低线程间共享状态与锁竞争;同时配合SIMD向量化用于批量点属性处理、矩阵/四元数变换等内核,可在不牺牲交互时延稳定性的前提下提升解码与更新速度^[233]。

在GPU层面,由于具备数千核心的大规模并行架构,在体积视频处理中的作用尤为突出。其不仅支撑点云与网格的实时渲染,还能够加速神经网络推理、光线步进和高斯溅射等操作,是当前NeRF与3DGS系统不可或缺的算力支撑^[230]。然而,并行处理也带来新的挑战,例如高斯基元数量巨大导致显存带宽压力、线程间同步开销、CPU与GPU间数据传输瓶颈(如PCIe带宽限制)以及内存访问冲突等问题。为此,研究者提出在算法划分、内存管理与任务调度中进行针对性设计,以最大化并行硬件潜力并维持低延迟^[227]。

4.4.2 专用硬件加速与新兴架构

在能效比和任务定制化方面,通用CPU和GPU仍存在一定局限。因此,专用硬件加速逐渐成为体积视频系统的重要补充。主流移动SoC已集成专用ASIC单元,用于高效执行H.264/AVC、H.265/HEVC、H.266/VVC和AV1等视频标准^[234]。当体积视频的属性信息(如纹理图、神经特征网格或高斯参数集)被组织为图像序列时,这些现有的硬件编解码器能够以极低功耗完成高质量压缩与传输^[235]。

同时,FPGA因其可重构性和高度并行性,被广泛应用于实时性要求严格的环节,如多视角传感器数据预处理、空间划分和几何变换等^[236]。随着NeRF和3DGS等神经表示在体积视频中的普及,AI加速器(如NPU、TPU和DPU)也逐渐成为核心组件。这类专用处理器针对矩阵乘法、卷积运算和激活函数进行了深度优化,在视口预测、内容感知编

码和超分辨率重建等任务中显著提升推理效率和能效表现^[237]。

近年来,研究者还探索了面向三维数据的新兴专用硬件架构。例如,针对点云处理的数据流加速器、支持高效KD-Tree构建的定制芯片,以及具备体素级并行计算能力的协处理器^[238]。这些新兴硬件不仅能突破传统平台的性能瓶颈,也为未来构建高性能、低功耗的沉浸式体积视频系统奠定基础。

4.4.3 云边端协同计算

尽管硬件加速在客户端和边缘侧显著提升了处理效率,但在面对复杂任务和动态交互时,单一层级仍难以满足所有需求。云边端协同架构通过分层分工与任务卸载实现了计算与存储资源的最优利用,其核心在于围绕体积视频链路将任务按时延敏感度与数据形态分层。在云端,强大的计算集群适合承担离线/近实时的全局任务,如离线三维重建、高精度编码和大规模神经表示训练,从而减轻终端负担并提升全局效率^[239-240]。在交互式场景中,云端还能进行高保真渲染并将结果以二维流或轻量化数据流传输至终端,实现低门槛的沉浸体验,但该方案高度依赖网络时延控制。

边缘计算则是体积视频低时延与可交互需求下更常用的折中,通过将多视角数据融合、视口预测、局部解码或轻量化推理部署在接近用户的边缘节点(如5G基站或本地服务器),系统能够显著降低端到端延迟,提高鲁棒性与响应速度^[177,239,241]。

然而,云边端协同仍面临诸多挑战,包括跨层数据分发与同步机制、任务调度的动态优化、服务质量(Quality of Service, QoS)保障以及隐私与安全问题。未来的发展趋势是通过软硬件协同和智能调度机制,在保证低延迟和高可扩展性的同时实现灵活部署。这不仅有助于推动体积视频走向大规模实用化,也为支持新兴神经表示和混合表征提供了基础支撑^[232]。

5 标准化与互操作性问题

体积视频的标准化覆盖整个端到端链路,包括表征、码流、承载、传输、呈现与评测等多个层面。其中,在编解码与码流语法方面,已有V3C、V-PCC、G-PCC与MIV等标准体系;在封装与分发上,ISO/BMFF、OMAF与CMAF提供了容器支持;在传输承载上,则依赖DASH/HTTP等渐进式方案与WebRTC/RTP等低时延协议。另一方面,体验

质量(Quality of Experience, QoE)与能效度量仍在演进中。MPEG与ITU-T已提出主观测试与客观指标框架,但针对点云、体素与神经表示的统一评价方法和互操作机制尚未成熟。同时,中国AVS标准工作组也在体积视频方向推进相关国家标准制定,包括点云压缩(AVS-PCC)、虚拟现实视频内容、高效多媒体编码传输等标准。

5.1 码流语法与语义标准化

MPEG-I为体积视频建立了由语法、工具和承载构成的分层体系。其中,V3C(ISO/IEC 23090-5)是这一体系的重要组成部分。该标准采用体素化和视频化的处理思路,将三维内容映射到二维视频编码工具链中,以兼容现有基础设施。在此框架下,V-PCC通过投影方式利用多路视频序列压缩点云几何与属性,并与V3C在语法上配套;G-PCC则直接在三维空间编码几何与属性,适合稀疏或高精度场景。MPEG-I还推出了MPEG沉浸式视频(MPEG Immersive Video, MIV)通过多视点与深度输入构建自由视点合成框架,主要面向6-DoF/3-DoF+体验。总体来看,V3C、V-PCC、G-PCC和MIV已成为当前点云与多视体积媒体领域的主要标准体系。

在国内标准方面,AVS工作组于2024年发布了首个点云压缩标准AVS-PCC,侧重于三维几何的直接编码路径(类似于MPEG G-PCC)并改进了八叉树预测和上下文熵编码等技术细节,在测试数据上相比MPEG参考模型TMC13在几何与属性压缩性能上均有所提升,并可满足实时低时延解码需求。其时空图形数据编码系列标准包括系统、点云、点云主观质量评价、符合性测试等,涵盖时空图形数据编码从传输,编码到质量评价等方面。

在系统实现层面,MPEG定义了V3C/V-PCC在ISO/BMFF(ISO/IEC 14496-12)中的承载规范(23090-10),确保轨道/采样、时序与同步与DASH、CMAF、OMAF等流媒体生态对接。ISO/BMFF作为通用容器,支持元数据、索引与扩展语义的封装;OMAF(23090-2)则为全景与沉浸式媒体提供了分片、视口自适应与信令机制,逐步与V3C的分块与视口调度对齐。整体来看,V3C定义了语法抽象,V-PCC与G-PCC分别代表投影与几何路径,MIV覆盖多视/深度场景,而ISO/BMFF与OMAF则保证系统互通。当前挑战在于:如何在点云、视图和神经表示之间对齐语义层次(如块/切片与视口映射),以及在分层、增量更新下保持码流一致性,仍是后续标准演进的重点。

5.2 承载与传输协议的互联互通

在点播与近实时分发场景中,体积内容常依赖MPEG-DASH(ISO/IEC 23009-1)与CMAF在HTTP/2或HTTP/3(QUIC)之上的自适应传输机制。对于V3C/V-PCC,标准23090-10已定义在ISO/BMFF内的轨道化承载与时间同步,结合DASH/OMAF的MPD描述与视口/质量选择机制,可驱动分块与层级的自适应请求。这一设计使V-PCC、MIV等采用视频化思路的体积媒体能够较好地复用现有CDN和边缘网络,从而具备与传统视频相近的部署条件。与此同时,AVS高效多媒体编码系列中新制定的智能媒体传输标准(GB/T 33475.6-2024)提出了面向异构分组网络的自适应流媒体传输方案,涵盖数据模型、信令、媒体呈现以及HTTP/2低时延直播和自适应前向纠错等技术。该方案为体积视频高可靠传输提供了标准化支持,在设计理念上与MPEG DASH/OMAF的自适应传输机制相辅相成。

随着低时延与交互需求的提升,系统更依赖IETF生态的WebRTC/RTP/QUIC(RFC 9000)等协议栈。对于直接基于RTP的传输,可利用私有负载与扩展头承载块/视口级信令,并结合OMAF元数据完成终端拼接与渲染。尽管V3C/V-PCC的RTP载荷标准化工作仍在推进,但基于23090-10、ISO/BMFF与RTP/FU片段之间的映射方案已在试验系统中得到验证。总体上,跨层互通的核心在于语义对齐:如服务器端V3C分块/切片与客户端视口预测之间的空间索引映射,以及DASH/OMAF的可伸缩层与终端解码能力描述的配合。当前较为稳妥的做法是以ISO/BMFF统一封装,并在MPD或会话层显式暴露可请求的体积块与质量层,从而兼顾CDN友好性与低耦合实现。

5.3 功耗、QoE与能效度量的标准化需求

在QoE与客观指标方面,MPEG已在V-PCC/G-PCC的标准化过程中提出基于几何与属性误差的测试条件(Common Test Conditions, CTC),采用点到点/点到面误差、色彩误差与重采样一致性等度量。在主观评测方面,ITU-T P. 913和P. 919为互联网视频与沉浸式媒体建立了较为通用的评价方法,可用于体积视频中的显示、交互与评分流程。然而,点云、体素与神经表示仍缺乏统一的端到端主观评测标准,当前多依赖MPEG CTC与ITU-T建议的组合,并辅以任务驱动或交互式的实验流程以保证可比性与可重复性。

在能效与功耗方面,尚无统一国际标准,但业界已开始借鉴传统流媒体的能耗分析框架,将终端解码—重建—渲染—显示的功耗剖面与网络传输(分块请求、拥塞控制、重传/前向纠错)的能耗统计结合。随着HTTP/3/QUIC与WebRTC逐渐成为主流方案,协议栈层面的遥测信息也开始被纳入端到端能效计量。相关研究普遍关注功耗与主观质量之间的关系,并据此指导可伸缩层级、视口自适应和语义驱动的比特分配。较为可行的做法是在现有QoE与能耗度量体系中补充体积块级统计信息以及端侧GPU和AI加速器占用情况,从而提高不同设备和实现之间结果的可比性与可审计性。

综上,标准化进展与体积视频表征、编解码和跨层优化技术路线紧密相连。在表征与压缩算法层面,MPEG提出的V3C架构及其衍生的V-PCC/G-PCC/MIV等标准实际对应并验证了前文所述的显式几何表示压缩方案。体系演进方面,随着NeRF、3DGS等新兴神经表示在压缩效率和渲染质量上的优势日益凸显,标准化正积极探索将这些隐式表征纳入统一语法体系的途径(例如通过扩展码流语义以承载体素特征网格或NeRF参数块),从而使前沿的神经编解码器成果能够在标准框架下互操作,推动神经网络驱动的高保真压缩策略实现标准落地。跨层系统优化方面,标准中的承载与传输规范(如MPEG DASH/OMAF自适应流和IETF的WebRTC/RTP/QUIC低时延传输)为跨层协同策略提供了接口与保障。在质量评价和能效方面引入统一的主客观指标和功耗度量标准,也体现了跨层优化中平衡主观体验与资源开销的理念,使感知质量优化、能耗约束等策略有据可依。综上,体积视频标准化工作正在将多种数据表征形式、压缩技术手段与跨层优化方法有机融合,搭建起从内容表示、编码压缩、网络传输到终端呈现各环节协同的统一框架。这不仅巩固了前期研究成果向实际可部署系统演进的路径,也为构建可互操作、可扩展的6-DoF沉浸式媒体服务生态奠定了重要基础。

目前,MPEG-I已形成覆盖语法、承载和系统层的标准体系。其中,V3C、V-PCC、G-PCC和MIV负责语法层定义,23090-10将其映射到ISO/BMFF,OMAF、DASH、HTTP/3以及WebRTC、RTP则分别支持点播和低时延交互。下一阶段标准化可重点围绕三个方向推进:完善分块、视口和可伸缩机制之间的跨层语义对齐与增量更新;推动主观相关客观指标与功耗度量的统一;以及扩展神经表示的承载

和互操作支持范围,包括体素特征网格以及3DGS、NeRF等参数块表示,从而提升体积媒体服务的可测性、可比性和可部署性。

6 实验评价指标与数据集

对沉浸式体积视频编解码技术进行客观、可复现且可比的评估,是推动该领域发展的关键。完善的评价体系需同时覆盖压缩效率、重建质量、处理速度与资源开销等核心维度,并依托标准化数据集确保公平比较,避免因内容差异引入偏差。本章将系统梳理体积视频编解码研究中常用的主客观评价指标与标准测试集,以构建一套支持实验设计、性能分析与方法对比的统一评估框架。

6.1 压缩性能与重建质量指标

体积视频编解码的性能评价首先集中在压缩效率与重建质量两个核心维度,这也是衡量算法优劣和实用性的基本依据。压缩效率直接决定了数据在存储与传输中的带宽和存储开销,而重建质量则反映了压缩后数据在几何、纹理及感知层面的保真度。一个完整的评估体系通常需要在率失真(Rate-Distortion, R-D)性能与计算复杂度之间建立平衡,以确保技术既能在资源受限条件下高效运行,又能满足沉浸式体验对高保真内容的需求。

在压缩效率方面,最常用的指标是比特率(Bitrate),即压缩后数据流的平均传输速率,通常以kbps或Mbps表示^[242]。对于静态对象或单帧分析,也会引入更细粒度的度量方式,如每点比特数(bits per point, bpp)、每体素比特数(bits per voxel, bpv)或每帧比特数(bits per frame, bpf),以反映不同数据表示下的编码开销。另一类常见指标是压缩比(Compression Ratio, CR),其定义为原始数据大小与压缩后数据大小的比值^[243]。在R-D性能对比中,Bjontegaard Delta Rate(BD-Rate)与BD-PSNR是两种标准化指标^[244]。其中,BD-Rate衡量在相同质量下编码器所需比特率的差异,而BD-PSNR则评估在相同比特率下的质量提升幅度。这类指标已广泛应用于标准测试与跨方法比较,为不同编解码方案提供公平对照。

在重建质量评估中,由于体积视频可采用点云、网格、体素或隐式场等多种表征,相关指标也呈现出差异化。对于显式几何表征,几何失真度量尤为重要。倒角距离(Chamfer Distance, CD)^[245]是点云质量评估的主流指标,衡量两点集之间的平均最小

距离。此外,豪斯多夫距离(Hausdorff Distance, HD)^[246]更关注两点集在最坏情况下的最大偏差,反映了一个集合相对于另一个集合的最远误差。由于HD对异常点高度敏感,实际应用中往往采用其改进版本,如平均豪斯多夫距离(Average Hausdorff Distance, AHD)或修正豪斯多夫距离(Modified Hausdorff Distance, MHD)。MPEG PCC标准推荐的D1 PSNR与D2 PSNR也常用于几何质量评估,分别基于点到面的距离与点到点的距离,前者更关注表面拟合精度,后者则与CD类似。对于体素化表示,体积交并比(Volumetric IoU)被广泛使用^[247]。

在属性质量度量方面,峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)是最常见的指标^[247]。对于基于神经表示的方法(如NeRF和3DGS),由于其评估更偏向图像级别的合成质量,因此更常用图像感知指标,如结构相似性指数(Structural Similarity Index Measure, SSIM)^[249]多尺度结构相似性指数(Multi-Scale Structural Similarity Index Measure, MS-SSIM)^[256]、视觉信息保真度(Visual Information Fidelity, VIF)^[250],以及基于深度特征的感知相似性度量LPIPS(Learned Perceptual Image Patch Similarity, LPIPS)^[251]。这些指标与人类主观感知高度相关,因而在新视角合成的质量评估中应用广泛。

综上所述,压缩性能与重建质量指标不仅在度量方法上因表征形式而异,还需要结合主客观评价体系进行交叉验证。当前趋势是构建跨表征、跨任务的统一指标体系,以同时兼顾几何精度、外观保真与感知体验,从而为未来体积视频编解码标准化与跨系统对比提供坚实的评价基础。

6.2 主观与客观质量评价方法

在体积视频编解码的研究与应用中,如何有效评估内容的真实感与体验质量,是贯穿整个体系的核心问题。尽管客观指标(如比特率、PSNR、SSIM、CD等)能够提供量化、可复现的性能比较,但它们往往难以全面反映人类用户在沉浸式场景中的实际感知。尤其是在自由视点观看、交互式渲染或动态内容播放的情境下,几何误差、纹理失真、光照伪影以及时延抖动等因素均可能对用户体产生显著影响,而这些细节往往超出了传统客观指标的捕捉能力。因此,主观与客观质量评价需要协同使用,以在实验评估中建立更为可靠的多维度框架。

主观评价方法仍是衡量体积视频质量最常采用的依据。其中,最常见的评价方式是平均意见分(Mean Opinion Score, MOS),其评估流程通常遵循

ITU-R BT. 500-13建议^[252]。在受控实验环境下,组织一定数量的观察者观看不同编解码器输出的体积视频,并依据五级评分体系(1表示“差”,5表示“优秀”)对整体感知质量进行打分。MOS的统计结果能够直观地反映人类感知体验,但实施成本较高,且结果容易受到个体差异、实验场地、显示设备和观看角度的影响。为提高评价结果的稳定性和区分度,研究中常采用成对比较(Pairwise Comparison)^[253]或排序法(Ranking Method)^[254]等方法,以减小主观评分偏差。此外,在交互式场景下,A/B测试^[255]也常用于比较不同方案的用户偏好,从而更接近实际使用过程中的体验差异。

相比之下,客观评价方法因其自动化与低成本特性,在体积视频研究中应用更为广泛。对于显式几何表征,常用指标包括点云的几何距离(CD、HD、D1/D2 PSNR)、体素IoU等,这些指标能够较为准确地量化几何保真度。然而,它们在反映视觉感知层面的效果时存在局限性。例如,几何误差较小的点云仍可能出现纹理模糊或运动伪影,从而导致较差的主观体验。针对这一问题,研究者提出将图像级别的指标引入体积视频评估中,如SSIM^[248]、MS-SSIM^[249]、VIF^[250]和LPIPS^[251]等,它们能够更好地刻画局部结构与感知特征的差异,尤其适用于基于神经表示的体积视频重建与渲染任务。

近年来,随着沉浸式应用需求的增长,融合主观与客观评价的混合型方法逐渐受到关注。这类方法结合两类评价方式的特点,例如利用少量主观实验结果对客观指标进行回归建模,从而提高其与人类感知之间的一致性^[252-253]。此外,也有研究探索基于深度学习的感知质量预测模型,通过端到端的训练直接学习从体积视频数据到主观评分的映射^[254]。这种方法不仅能够自动化预测大规模数据的感知质量,还能够为未来的自适应编码与QoE驱动优化提供重要支撑。

总体而言,主观与客观质量评价在体积视频编解码评估体系中各具优势:主观方法能够反映用户体验的真实性,客观方法则保证了量化分析的可复现性。两者的结合不仅能够提升评估的科学性与实践性,还为后续跨方法比较、标准化指标建立及用户体验驱动的系统优化奠定了坚实基础。

6.3 复杂度、时延与能耗指标

在体积视频编解码的实验评价中,除了压缩效率与重建质量外,复杂度、时延与能耗指标同样至关重要。这些指标直接决定了算法在不同硬件环境、应用场景和交互模式下的可用性与实用性,尤其是

在实时点播和低时延直播等对延迟敏感的场景中,其重要性甚至超过单纯率失真表现^[256-257]。

计算复杂度通常通过理论运算量与实际运行时间来刻画。理论运算量可采用浮点运算次数(Floating Point Operations, FLOPs)或乘加运算次数(Multiply-Accumulate operations, MACs)进行估算,能够反映算法在硬件无关条件下的计算需求。然而,FLOPs并不能直接等价于真实运行时间,因此实际测试往往同时测量编码时间与解码时间。编码时间对于离线存储应用相对宽松,但在直播与交互式点播中则需要显著压缩;解码时间则直接决定播放流畅性与用户交互延迟,是评价算法能否进入实际应用的关键指标。处理帧率(Frames Per Second, FPS)常被用作直观度量,反映系统在实时条件下能够支持的最大输出速度。

时延指标涵盖端到端的多个环节,包括采集延迟、编码延迟、传输延迟、解码延迟与渲染延迟。在体积视频应用中,端到端时延(End-to-End Latency)通常要求控制在数百毫秒以内,以保证交互的自然性。对于VR/AR等高沉浸式场景,更为严格的“运动到光子”延迟(Motion-to-Photon Latency)需低于20毫秒,以避免用户产生眩晕或不适^[258]。此外,不同表征方式在延迟敏感度上差异明显:显式几何编码通常依赖块级或帧级处理,延迟较可控;而神经表示则涉及大量并行矩阵运算与推理步骤,其推理延迟常成为制约实时性的主要瓶颈^[11]。因此,如何在不同编码策略中实现延迟优化,已成为近年来的重要研究方向。

能耗指标在移动端与边缘计算环境中尤其关键。随着体积视频应用逐渐从实验室走向消费级终端,编解码过程的能效比成为不可忽视的问题。典型的度量方法包括:单位帧能耗(Energy per Frame)、单位比特能耗(Energy per Bit)以及单位运算能耗(Energy per FLOP)。这些指标不仅反映了算法的功耗需求,还决定了电池供电设备在长时间运行中的可持续性。此外,一些研究还将能效与用户体验相结合,提出能耗感知的QoE(Energy-aware QoE)模型^[259],在保证感知质量的同时最大限度降低能量消耗。

近年来,已有研究尝试采用综合性指标,对编解码方法在率失真、复杂度、时延与能耗上的权衡进行统一描述。例如,R-D-C曲线^[256]在传统率失真曲线基础上引入复杂度维度,可同时反映压缩效率、重建质量与计算代价之间的关系;R-D-C-L四维分析框

架^[260]则进一步纳入时延因素,为离线、点播和直播场景下的性能评估提供了更完整的分析框架。这类多维指标有助于比较不同方法的综合性能,为后续跨层优化与系统设计提供参考。

6.4 常用标准数据集与实验配置

在沉浸式人体建模与体积视频压缩研究中,标准化测试数据集对于算法评估、性能比较和技术迭代具有重要意义。这些数据集通常由多视角相机阵列或深度传感器采集,并通过三维重建技术生成高质量点云、网格或神经表示,广泛用于评估压缩效率、重建质量、实时传输策略以及基于学习的编码方法。表12中介绍了当前主流的人体相关体积视频数据集及其核心特征。

(1) MPEG V-PCC 动态人体数据集(ISO/IEC 23090-5)^[261]: 该数据集由国际标准化组织 MPEG 在其体积视频压缩标准制定过程中发布,是目前最具权威性和广泛应用性的数据集之一。V-PCC 数据集包含多个动态人体序列,如 longdress_vox10、redandblack_dancer_vox10 和 soldier_mithai_vox10 等,均由多视角同步相机捕获并重建,每帧点数可达百万级,适用于评测动态点云压缩算法在复杂运动场景下的性能表现。该数据集格式统一、时间一致性良好,已成为行业内的重要基准测试集。

(2) 8i Voxelized Full Bodies (8iVFB)^[262]: 8iVFB数据集是早期广泛采用的公开动态人体点云基准,包含多名演员执行舞蹈、行走、对话等自然动作的三维序列,提供体素化后的表面点云,每帧约含100万-300万个点,并附带颜色信息。该数据集因其动作丰富、采集质量高,已成为自由视点视频、NeRF与3DGS表征压缩研究的重要评测资源。

(3) Microsoft Voxelized Upper Bodies (MVUB)^[263]: MVUB数据集更侧重于上半身交互场景,如虚拟会议和远程通信。该数据集提供人物上半身的动作序列(如演讲、头部转动、手势交互),适合评估低带宽条件下的压缩与重建平衡。尽管其点云密度略低于8iVFB,但其对面部表情与手部细节的捕捉较为精细,对语义感知编码与内容优先级优化研究具有重要价值。

(4) INTERACT Point Cloud Dataset^[264]: INTERACT数据集聚焦多人协作与交互行为,涵盖双人互动、物体传递、手势识别等场景。该数据集支持时间一致性分析与局部压缩策略研究,尤其适用于直播体积视频中的视口预测与ROI编码优化。

(5) HoloLens2/Azure Kinect 实时采集数据

表 12 常用标准体积视频数据集

数据集名称	类型	场景类别	数据规模 (点数/帧或重建方式)	包含属性	数据格式/重建方法	应用重点
MPEG V-PCC ^[261]	动态点云	人体表演	100 万-300 万点	几何、颜色、时间	多视角重建点云	视频点播、直播、动态 PCC 压缩
8iVFB ^[262]	动态点云	全身动作	100 万-300 万点	几何、颜色、时间	体素化表面点云	自由视点视频、NeRF/3DGS 建模
MVUB ^[263]	动态点云	上半身交互	50 万-100 万点	几何、颜色、时间	体素化表面点云	虚拟会议、远程呈现、低比特率传输
INTERACT ^[264]	动态点云	多人协作交互	50 万-200 万点	几何、颜色、时间	实时重建点云	实时交互、手势识别、视口预测
HoloLens2/ Azure Kinect ^[265]	动态点云、网格	实时采集交互场景	数十万-百万点/帧	几何、颜色、网格	深度相机实时重建点云/网格	移动设备压缩、边缘计算、实时 VR/AR
EPIC-KITCH-ENS 扩展版 ^[266]	动态点云	日常厨房活动	实时捕获点云	几何、颜色、时间	RGB-D 视频扩展为点云	动作驱动压缩、语义增强编码
Human3. 6M + NeRF 扩展 ^[267]	网格/NeRF 训练数据	室内动作捕捉	网格或用于 NeRF 训练的图像序列	几何、纹理、动作标签	骨架/图像数据扩展为网格/NeRF 输入	动作建模、参数化压缩、SMPL 变形场
THuman 系列 ^[268]	高质量网格/ NeRF 训练数据	人体姿态模型	高精度网格模型和多视角图像	几何、纹理、姿态	扫描重建网格，多视角图像	神经表示训练、参数化编码、人体建模
Neural 3D Video ^[269]	多视角 RGB 视频/动态 NeRF 训练数据	真实动态交互场景	6 个场景；每场景由约 18 台同步相机拍摄 10 秒 30FPS 视频	几何、颜色、时间、动作	多视角同步 RGB 视频/相机位姿	动态 NeRF 模型训练与高质量 3D 视频合成、新视角渲染和运动插值
D-NeRF ^[56]	合成动态 RGB 图像数据/NeRF 训练数据	合成动态场景	8 个合成场景；单目相机不同视角序列 (800×800 分辨率)	几何、颜色、时间	稀疏单目 RGB 图像序列/精确相机轨迹	动态 NeRF 研究基准、单目视频的动态场景重建和新视角合成评测
NeRF-DS ^[134]	动态镜面物体；RGB 训练数据	镜面物体动态场景	视频序列(左训练/右测试，分辨率约 480×270)	几何、颜色、时间、镜面反射	双相机 RGB 序列，含前景物体 mask 和相机标定参数	动态 NeRF 镜面场景、运动高光物体新视角渲染准确性
HyperNeRF ^[270]	非刚性动态 RGB 视频/真实动态 NeRF 训练	真实动态场景	17 个场景；单目或双目相机一致运动采集，提供每帧深度图和相机位姿	几何、颜色、时间、拓扑结构变化	单/双相机 RGB 图像序列/每帧深度和相机姿态	拓扑变化动态 NeRF、非刚性拓扑变化场景的重建与新视角合成
Google Immersive Dataset ^[271]	多视角沉浸式 光场视频数据集	室内/外 大范围运动场景	15 个场景；46 台同步相机半球阵列拍摄 (部分场景相机数较少)	几何、颜色、时间、复杂运动/遮挡模式	多视角 RGB 视频/完整相机参数标定	动态场景自由视点视频合成与压缩评测基准、复杂动态场景重建与渲染

集^[265]：微软开源的一系列 HoloLens2 或 Azure Kinect 捕获的点云与网格数据虽未形成统一的大规模官方基准，但因其贴近实际 XR 设备的采集特性，已被多个研究工作用于构建轻量级人体压缩与边缘端处理的实验平台，适用于移动端或嵌入式设备上的体积视频应用。

(6)EPIC-KITCHENS 扩展点云版本^[266]：虽然 EPIC-KITCHENS 最初为 RGB-D 视频数据集，但部分后续研究将其扩展为点云形式，用于评估日常动作捕捉与自由视点重建效果，适合结合动作理解、语义分割与基于学习的压缩方法进行研究。

(7)MPI-INF-3DHP/Human3. 6M 的隐式场扩

展^[267]：Human3. 6M 是经典的人体姿态与动作捕捉数据集，近年来一些研究尝试将其拓展为点云、网格乃至神经表示形式，用于评估动态人体压缩与重建能力，尤其是在结合 SMPL^[272]或 SMPL-X^[273]模型进行参数化压缩的研究中具有代表性意义。

(8)Neural 3D Video^[269]：专为动态 NeRF 等神经体积建模任务设计的数据集，涵盖多个真实室内交互场景。该数据集由 18-21 台同步相机组成的多视角阵列捕获，记录了如烹饪、物体摆放等人物与器具互动过程，兼具刚性和非刚性运动特征。每个序列时长约 10 秒，约含 300 帧图像，提供完整的 RGB 多视角图像与相机参数，适用于建模视角一致性强、

结构变化复杂的动态体积表示,是当前研究动态隐式场建模任务的重要基准之一。

(9)D-NeRF^[56]:D-NeRF是一个广泛用于隐式动态重建的合成数据集,包含8个不同场景,每个序列均由单个动态物体构成,如弹跳球、跳跃人物与刚性变形的乐高积木等。该数据集在800×800分辨率下以不同相机姿态渲染生成,场景中普遍存在大幅位移与刚性形变,适用于评估基于NeRF的时间建模与空间一致性约束方法的表达能力。

(10)NeRF-DS^[134]:NeRF-DS是面向镜面物体的新视角合成评估的标准数据集,包含8个具有高反射属性的动态场景,涵盖移动或刚性变形的镜面对象。数据由两台固定相机同步采集,分别用于训练与测试,提供了面向动态NeRF系列模型在视角依赖性强的镜面材料建模方面的评测基准。

(11)HyperNeRF^[270]:HyperNeRF旨在支持非刚性动态场景中拓扑结构变化的建模与合成,数据集中共包含17个序列,采用单目或双目相机以一致路径采集,避免帧间过度运动干扰。每帧均提供高质量相机姿态与真实深度图,其中最常用的为包含拓扑变化的四个vrig场景,如人物张口、肢体遮挡与断连等结构变化,适用于评估动态神经表示在不连续形变场景下的建模能力。

(12)Google Immersive Dataset^[271]:面向自由视角视频的高质量动态数据集,覆盖15个动态场景,最多由46台相机构成半球形采集阵列。该数据集提供结构光重建的精确相机标定信息,支持从多视角动态图像中重建高保真的3D运动场景,适用于多视角压缩、体积视频渲染与视点迁移研究。

7 未来挑战与发展趋势

(1)数据的高效表示与压缩

尽管NeRF、3DGS/4DGS等神经表示显著提升了体积视频重建质量,高效表示与压缩仍是关键问题。这类方法常依赖特征网格、哈希表和高斯基元参数等辅助结构,存储开销较大,在高分辨率、长时序动态场景中更为突出。现有量化、剪枝和特征压缩方法尚难兼顾保真度与效率,对时序相关性的利用也不充分,多数方法仍以逐帧压缩为主。同时,基于语义的高层压缩开始显现潜力,例如在体积视频会议中,仅传输人体骨架运动等参数即可重建动作。尽管已有NeRFFlow^[88]、GaussianFlow^[89]、HybridGS^[90]等探索,神经表示的存储成本仍然偏

高。点云、网格和体素等显式表示虽有较成熟的标准与工具,但在复杂动态场景和高保真需求下同样存在瓶颈。未来需发展更紧凑的神经表示与面向6-DoF沉浸通信的可学习编解码方法,充分利用时间相关性,并支持多维可伸缩与渐进编码,以适应不同带宽和终端条件。

(2)端到端实时神经编解码

动态体积视频的端到端实时编解码需要在采集、重建、编码、传输、解码和渲染全流程中将时延控制在数百毫秒内,这仍是当前的主要难点。现有NeRF和3DGS方法虽可在特定场景下实现接近实时的渲染,但通常不包含实时采集、表示更新和压缩过程。当前研究多集中于预扫描场景的快速重建与渲染,或受限动态内容的局部更新,尚难支持复杂未知场景下的即时自由交互。近期有方法将动态辐射场序列转换为二维特征视频流,借助现有视频编解码器和GPU着色器实现移动端实时解码与绘制^[170],在一定程度上降低了端到端时延,但距离任意新场景下的实时沉浸式传输仍有差距。未来需发展超轻量、快速收敛的神经表示,以及面向流式输入的增量更新与高效压缩机制,结合低时延神经编解码与软硬件协同,进一步提升实时处理能力。

(3)沉浸式体验的质量评估与优化

当前体积视频质量评估常用的PSNR、SSIM、LPIPS、CD和HD等客观指标多源自传统二维图像、视频或静态三维模型,难以全面刻画6-DoF沉浸式体验。6-DoF质量不仅与几何精度、纹理清晰度、运动流畅性和视角相关效应有关,还受到交互时延、视点切换平滑性、视野范围以及视觉辐辏—调节冲突等因素影响。现有客观指标在主观相关性和可部署性方面仍显不足,这也限制了体积视频编解码与优化的发展。未来需结合视觉感知与认知心理学,构建面向6-DoF体验的客观评价模型,并重点推进体积视频质量数据库与基准建设、感知驱动的编解码优化,以及晕动不适的量化与缓解。

(4)标准化与互操作性

体积视频涵盖采集、建模、编码、传输与渲染的全链路,但在数据格式、编解码、元数据与流媒体协议上尚无统一标准,已成为跨平台互操作与生态建设的主要障碍。MPEG在点云压缩(V-PCC、G-PCC)和沉浸式媒体表示(MPEG IMR)上已有进展,Khronos Group也推动glTF等3D资产格式标准化,但针对NeRF、3DGS等快速演化的神经表示,标

准化仍处于起步阶段。未来需建立开放、可扩展的标准体系,兼容点云、网格、体素及神经表示等多种格式,并具备前后向兼容性。重点在于规范神经表示的核心组件(如网络拓扑、特征网格、高斯参数)的表示与压缩,以支持模型共享与高效解码。同时,应统一描述场景空间、对象属性、光照参数及交互能力的元数据,确保跨平台体验一致性与内容可发现性。

(5) 高效内容创建与编辑工具

高质量体积视频的生成与编辑仍受成本高、流程复杂和工具不足等因素限制,内容生态发展因而受阻。专业采集通常依赖多相机阵列与精密同步,重建也需要较高算力;现有工具对体积表示,尤其神经表示的建模与编辑支持仍较有限,3DGS中密集分布的高斯基元也增加了精细交互的难度。未来需要面向多种表示发展更高效的创作与编辑工具,包括利用消费级设备实现低成本采集与重建,引入扩散模型、GAN等生成式方法支持文本、图像和草图驱动的内容生成与增强,并设计更直观的神经表示编辑界面以支持对象级变形和材质调整。近年来,DreamTime^[274]、GVGEN^[275]和TADA^[276]等方法已表明,生成式模型能够基于有限输入高效生成细节丰富的三维场景或可驱动数字人,从而减轻体积内容创作对高成本逐帧采集的依赖。与此同时,云原生协作平台以及标准化交换格式和API也有助于提升互操作性与系统集成度,推动沉浸式内容生态发展。

(6) 硬件适应性与普适计算

沉浸式体积视频的目标是实现跨终端普适可用,从高性能PC/VR头显到移动终端均能流畅体验。然而,设备在计算能力、存储、显示性能、网络质量与能耗等方面差异显著,高质量渲染方案多依赖高端GPU,难以在资源受限设备上运行,成为规模化落地的瓶颈。为此,需要在算法和系统架构上提高对异构硬件环境的适应性,应对策略包括:构建可伸缩的编解码与渲染框架,依据终端能力自适应调整复杂度、分辨率与质量;推动跨平台标准(如扩展的glTF、统一的WebGPU API),提升异构硬件利用效率;利用移动端AI协处理器发展轻量解码与能效优化的渲染算法,并叠加超分辨率、运动插帧等增强功能;同时,建立云边端协同架构,将高复杂度任务卸载至云/边缘,终端侧聚焦轻量渲染与交互,从而在性能与视觉质量之间取得平衡。

(7) 安全性与隐私保护

体积视频具有高保真、高维度特性,能够细致

呈现人物形态、表情、姿态及其与环境的交互,但也带来突出的隐私与安全风险。高分辨率人像数据可能被用于面部识别、行为建模和三维深度伪造,进而威胁个人隐私与肖像权;仅凭体积视频模型即可提取面部轮廓、步态等生物特征,从而在未经授权的情况下识别个人身份。若扫描数据涉及家庭、医疗等敏感场景,一旦泄露,后果尤为严重,商业内容也面临盗版和非法传播风险。与此同时,用户在6-DoF环境中的头部和视线运动轨迹同样属于敏感信息。已有研究表明,头部运动数据与用户心理和健康状态相关,可反映孤独症、创伤后应激障碍和早期痴呆等特征。当前相关防护机制仍不成熟,有必要从采集、传输、存储到内容分发全过程进行系统设计,包括在采集与重建阶段引入三维匿名化和差分隐私,在传输与存储阶段结合端到端加密、访问控制与溯源认证,并在点云、网格、体素及神经表示等三维内容中嵌入数字水印,以支持完整性保护和责任追踪。同时,还需发展抗伪造身份验证机制,并结合联邦学习、安全多方计算、同态加密和区块链等技术,构建用户可控的数据管理框架;在此基础上,进一步完善法律法规与伦理规范,明确数据权属与使用边界,为体积视频的可信发展提供支撑。

(8) 多模态体积视频编码

体积视频与音频等模态分开处理,缺乏统一的跨模态压缩与同步机制,因而容易造成音画不同步和资源浪费。例如,音视频独立编码未能利用模态间相关性提升整体效率,空间音频若不能随用户视角实时调整,还会引起声源定位偏差。已有研究表明,在沉浸式场景中,空间音频保真度对感知质量的影响可与视频分辨率相当,准确的声源定位和环境音效对于增强沉浸感十分重要。然而,现有体积视频标准仍主要面向视觉内容,对音频、触觉等模态的支持相对有限,多模态体积视频编码因而成为值得进一步研究的方向。未来需要发展高效的多模态融合压缩方法,实现视觉、音频及其他传感信息的联合编码与传输;一方面挖掘视觉与音频之间的冗余和关联,另一方面保证各模态在解码与呈现阶段的同步与一致。对于空间音频,还需依据用户6-DoF视听位置实时调整渲染,并结合低时延音频处理或视点元数据辅助解码。对于触觉反馈等其他模态,也需研究相应的同步传输与压缩机制。总体来看,多模态体积视频编码需要面向多感官感知进行统一建模与协同优化,并推动跨模态数据格式与信令设计的发展。

8 总 结

沉浸式体积视频作为VR/AR与未来互联网的重要支撑,凭借六自由度交互与真实三维重建能力,正在重塑数字媒体形态。但其高维、大规模数据特性对存储、传输与实时处理提出严峻挑战,使高效表征与编解码成为核心问题。本文系统梳理了体积视频表征的演进脉络,从显式几何与函数式隐式方法,到近年神经表示(NeRF、3DGS等),并提出以离线压缩与实时处理为主线的编解码框架,综述不同表征下的压缩机制、编码策略与最新进展。同时,总结了跨表征的通用优化路径,包括模型轻量化、内容感知编码、视点预测、硬件加速与流媒体调度,并结合现有数据集与指标体系,提供了统一的研究与实践视角。总结来看,体积视频编解码正处于由规则驱动向学习驱动、由静态向动态智能压缩演进的关键阶段。尽管已有显著成果,但距离普适、实时与高质量的沉浸式体验仍存技术瓶颈,未来需在算法、系统与标准层面进一步协同创新。

参 考 文 献

- [1] Qiao X, Ren P, Dustdar S, et al. Web AR: A promising future for mobile augmented reality—State of the art, challenges, and insights. *Proceedings of the IEEE*, 2019, 107(4): 651-666
- [2] Xu Ya-Nan, Zhao Gui-Dan, Feng Da-Quan. Edge-cloud collaboration empowered immersive Web3D. *Telecommunications Science*, 2024, 40(9): 123-135 (in Chinese)
(许娅男, 赵桂丹, 冯大权, 等. 边云协同驱动的沉浸式Web3D. *电信科学*, 2024, 40(9): 123-135)
- [3] Zeng Huan-Qiang, Kong Qing-Wei, Chen Jing. Overview of immersive video coding. *Journal of Electronics & Information Technology*, 2024, 46(2): 602-614 (in Chinese)
(曾焕强, 孔庆伟, 陈婧, 等. 沉浸式视频编码技术综述. *电子与信息学报*, 2024, 46(2): 602-614)
- [4] Bentaleb A, Lim M, Hammoudi S, et al. Solutions, challenges, and opportunities in volumetric video streaming: an architectural perspective. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025, 21(7): 1-35
- [5] Yaqoob A, Bi T, Muntean G M. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 2020, 22(4): 2801-2838
- [6] Huang Y, Bai B, Zhu Y, et al. ISCom: Interest-aware semantic communication scheme for point cloud video streaming on metaverse XR devices. *IEEE Journal on Selected Areas in Communications*, 2023, 42(4): 1003-1021
- [7] Zhu Y, Huang Y, Qiao X, et al. HiVAT: Improving QoE for hybrid video streaming service with adaptive transcoding. *IEEE Transactions on Mobile Computing*, 2024, 23(12): 11209-11226
- [8] Zhang Y, Li X, Dai M. Virtual tourism immersive experience system//*Proceedings of the 2020 International Conference on Virtual Reality and Visualization (ICVRV)*. Recife/Porto de Galinhas, Brazil. 2020: 312-313
- [9] Fasogbon P, Bisht S, Kernen J, et al. Volumetric video use cases for xr immersive streaming//*Proceedings of the 2024 8th International Conference on Education and Multimedia Technology*. Tokyo, Japan. 2024: 1-8
- [10] Xu Z, Xie T, Peng S, et al. Easyvolcap: Accelerating neural volumetric video research//*SIGGRAPH Asia 2023 Technical Communications*. New York, USA: ACM, 2023: 1-4
- [11] Viola I, Cesar P. Volumetric video streaming: Current approaches and implementations. *Immersive Video Technologies*, 2023: 425-443
- [12] Huang Y, Zhu Y, Qiao X, et al. Toward holographic video communications: A promising AI-driven solution. *IEEE Communications Magazine*, 2022, 60(11): 82-88
- [13] Qian F, Han B, Pair J, et al. Toward practical volumetric video streaming on commodity smartphones//*Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications*. Santa Cruz, USA. 2019: 135-140
- [14] Hu K, Chen Y, Han K, et al. LiveVV: Human-centered live volumetric video streaming system. *IEEE Internet of Things Journal*, 2025, 12(11): 17233-17243
- [15] Collet A, Chuang M, Sweeney P, et al. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 2015, 34(4): 1-13
- [16] Liu K, Cheng R, Wu N, et al. Toward next-generation volumetric video streaming with neural-based content representations//*Proceedings of the 1st ACM Workshop on Mobile Immersive Computing, Networking, and Systems*. Madrid, Spain. 2023: 199-207
- [17] Alexiou E, Nehmé Y, Zerman E, et al. Subjective and objective quality assessment for volumetric video//*Immersive video technologies*. Cambridge, Massachusetts, USA: Academic Press, 2023: 501-552
- [18] Marvie J E, Nehmé Y, Graziosi D, et al. Crafting the MPEG metrics for objective and perceptual quality assessment of volumetric videos. *Quality and User Experience*, 2023, 8(1): 4
- [19] Liu Y, Yang Q, Xu Y, et al. Point cloud quality assessment: Dataset construction and learning-based no-reference metric. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023, 19(2s): 1-26
- [20] Tious A, Vigier T, Ricordel V. New challenges in point cloud visual quality assessment: a systematic review. *Frontiers in Signal Processing*, 2024, 4: 1420060
- [21] Yang M, Luo Z, Hu M, et al. A comparative measurement study of point cloud-based volumetric video codecs. *IEEE Transactions on Broadcasting*, 2023, 69(3): 715-726
- [22] Wu C H, Hsu C F, Hung T K, et al. Quantitative comparison of point cloud compression algorithms with PCC arena. *IEEE Transactions on Multimedia*, 2022, 25: 3073-3088

- [23] Zeng Huan-Qiang, Ding Rui, Huang Hai-Liang. Immersive video quality assessment: An overview. *Journal of Signal Processing*, 2022, 38(9): 1853-1862 (in Chinese)
(曾焕强, 丁瑞, 黄海靓, 等. 沉浸式视频质量评价综述. *信号处理*, 2022, 38(9): 1853-1862)
- [24] Wang Xu, Liu Qiong, Peng Zong-Ju. Research progress of six degree of freedom(6DoF)video technology. *Journal of Image and Graphics*, 2023, 28(06): 1863-1890 (in Chinese)
(王旭, 刘琼, 彭宗举, 等. 6DoF 视频技术研究进展. *中国图象图形学报*, 2023, 28(06): 1863-1890)
- [25] Jin Y, Liu J, Hu K, et al. A networking perspective of volumetric video service: Architecture, opportunities and case study. *IEEE Network*, 2024, 38(6): 138-145
- [26] Mamou K, Zaharia T, Prêteux F. FAMC: The MPEG-4 standard for animated mesh compression//*Proceedings of the 2008 15th IEEE International Conference on Image Processing*. San Diego, USA. 2008: 2676-2679
- [27] Cavagna R, Bouville C, Gioia P, et al. A MPEG-4 AFX compliant platform for 3D contents distribution in peer-to-peer//*Proceedings of the 2008 15th IEEE International Conference on Image Processing*. San Diego, USA. 2008: 2688-2691
- [28] Botsch M, Kobbelt L, Pauly M, et al. *Polygon mesh processing*. Boca Raton, USA: CRC Press, 2010
- [29] Campen M, Attene M, Kobbelt L. A practical guide to polygon mesh repairing. *Eurographics (Tutorials)*. 2012: 32
- [30] Google. Draco: 3D Data Compression. 2017. GitHub, <https://github.com/google/draco>. Accessed 28 May 2025
- [31] Graziosi D B. Video-based dynamic mesh coding//*Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*. Anchorage, USA. 2021: 3133-3137
- [32] Zhou X, Wu R, Zeng Y, et al. Choice of Voxel-based Morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. *Communications Biology*, 2022, 5(1): 913
- [33] Bunyamin J, Sinclair B, Law M, et al. Voxel-based and surface-based cortical morphometric MRI applications for identifying the epileptogenic zone: A narrative review. *Epilepsia Open*, 2025, 10(2): 380-397
- [34] Meagher D. Geometric modeling using octree encoding. *Computer Graphics And Image Processing*, 1982, 19(2): 129-147.
- [35] Crassin C, Neyret F, Sainz M, et al. Interactive indirect illumination using voxel cone tracing: An insight//*ACM Siggraph-Talk*. Vancouver, Canada. 2011:1
- [36] Museth K. VDB: High-resolution sparse volumes with dynamic topology. *ACM Transactions on Graphics (TOG)*, 2013, 32(3): 1-22
- [37] Espe A E, Gjermundnes Ø, Hendseth S. Efficient animation of sparse voxel octrees for real-time ray tracing. *arXiv preprint arXiv:1911.06001*, 2019
- [38] Kim D, Lee M, Museth K. Neuralvdb: High-resolution sparse volume representation using hierarchical neural networks. *ACM Transactions on Graphics*, 2024, 43(2): 1-21
- [39] Ruijters, D. Common artifacts in volume rendering. *arXiv preprint arXiv:2109.13704*, 2021
- [40] Frisken S F, Perry R N, Rockwood A P, et al. Adaptively sampled distance fields: A general representation of shape for computer graphics//*Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*. New Orleans, USA. 2000: 249-254
- [41] Park J J, Florence P, Straub J, et al. Deepsdf: Learning continuous signed distance functions for shape representation//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, USA. 2019: 165-174
- [42] Long X, Lin C, Liu L, et al. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada. 2023: 20834-20843
- [43] Oleynikova H, Millane A, Taylor Z, et al. Signed distance fields: A natural representation for both mapping and planning//*Proceedings of the RSS 2016 workshop: Geometry and Beyond-Representations, Physics, and Scene Understanding for Robotics*. Ann Arbor, USA. 2016:1-6
- [44] Izadi S, Kim D, Hilliges O, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera//*Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. Santa Barbara, USA. 2011: 559-568
- [45] Nießner M, Zollhöfer M, Izadi S, et al. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 2013, 32(6): 1-11
- [46] Keller M, Lefloch D, Lambers M, et al. Real-time 3d reconstruction in dynamic scenes using point-based fusion//*Proceedings of the International Conference on 3D Vision-3DV 2013*. Seattle, USA. 2013: 1-8
- [47] Mildenhall B, Srinivasan P P, Tancik M, et al. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021, 65(1): 99-106
- [48] Garbin S J, Kowalski M, Johnson M, et al. Fastnerf: High-fidelity neural rendering at 200fps//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada. 2021: 14346-14355
- [49] Yu A, Li R, Tancik M, et al. Plenooctrees for real-time rendering of neural radiance fields//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada. 2021: 5752-5761
- [50] Hedman P, Srinivasan P P, Mildenhall B, et al. Baking neural radiance fields for real-time view synthesis//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada. 2021: 5875-5884
- [51] Duckworth D, Hedman P, Reiser C, et al. Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration. *ACM Transactions on Graphics (TOG)*, 2024, 43(4): 1-13
- [52] Reiser C, Peng S, Liao Y, et al. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps//*Proceedings of the*

- IEEE/CVF International Conference on Computer Vision. Montreal, Canada. 2021: 14335-14345
- [53] Fridovich-Keil S, Yu A, Tancik M, et al. Plenoxels: Radiance fields without neural networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA. 2022: 5501-5510
- [54] Müller T, Evans A, Schied C, et al. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 2022, 41(4): 1-15
- [55] Sun C, Sun M, Chen H T. Improved direct voxel grid optimization for radiance fields reconstruction. *arXiv preprint arXiv:2206.05085*, 2022
- [56] Pumarola A, Corona E, Pons-Moll G, et al. D-nerf: Neural radiance fields for dynamic scenes//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA. 2021: 10318-10327
- [57] Tretschk E, Tewari A, Golyanik V, et al. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada. 2021: 12959-12970
- [58] Wang L, Zhang J, Liu X, et al. Fourier plenotrees for dynamic radiance field rendering in real-time//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA. 2022: 13524-13534
- [59] Wang F, Tan S, Li X, et al. Mixed neural voxels for fast multi-view video synthesis//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France. 2023: 19706-19716
- [60] Liu J W, Cao Y P, Mao W, et al. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *Advances in Neural Information Processing Systems*, 2022, 35: 36762-36775
- [61] Gu J, Jiang M, Li H, et al. Ue4-nerf: Neural radiance field for real-time rendering of large-scale scene. *Advances in Neural Information Processing Systems*, 2023, 36: 59124-59136
- [62] Reiser C, Szeliski R, Verbin D, et al. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 2023, 42(4): 1-12
- [63] Kerbl B, Kopanas G, Leimkühler T, et al. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023, 42(4): 139:1-139:14
- [64] Fei B, Xu J, Zhang R, et al. 3d gaussian splatting as new era: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 2024: 1-20
- [65] Niedermayr S, Stumpfegger J, Westermann R. Compressed 3d gaussian splatting for accelerated novel view synthesis//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 10349-10358
- [66] Fan Z, Wang K, Wen K, et al. Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. *Advances in Neural Information Processing Systems*, 2024, 37: 140138-140158
- [67] Wang T, Yu Z, Xu Y. TC-GS: Tri-plane based compression for 3D Gaussian Splatting. *arXiv preprint arXiv: 2503.20221*, 2025
- [68] Huang H, Huang W, Yang Q, et al. A hierarchical compression technique for 3d gaussian splatting compression//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India. 2025: 1-5
- [69] Chen Y, Wu Q, Li M, et al. Fast feedforward 3d gaussian splatting compression. *arXiv preprint arXiv:2410.08017*, 2024
- [70] Lee S, Shu F, Sanchez Y, et al. Compression of 3d gaussian splatting with optimized feature planes and standard video codecs. *arXiv preprint arXiv:2501.03399*, 2025
- [71] Yan J, Peng R, Wang Z, et al. Instant gaussian stream: Fast and generalizable streaming of dynamic scene reconstruction via gaussian splatting//Proceedings of the Computer Vision and Pattern Recognition Conference. Nashville, USA. 2025: 16520-16531
- [72] Zhang X, Chang H, Liu Y, et al. Motion Blender Gaussian Splatting for Dynamic Scene Reconstruction. *arXiv preprint arXiv:2503.09040*, 2025
- [73] Zhang C, Zhou Y, Wang S, et al. Evolvinggs: High-fidelity streamable volumetric video via evolving 3d gaussian representation. *arXiv preprint arXiv:2503.05162*, 2025
- [74] Zhang X, Liu Z, Zhang Y, et al. Mega: Memory-efficient namic scenes. *arXiv preprint arXiv:2410.13613*, 2024
- [75] Li Y, Fang Y, Zhu Z, et al. 4D Gaussian Splatting SLAM. *arXiv preprint arXiv:2503.16710*, 2025
- [76] Yan J, Peng R, Tang L, et al. 4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia. 2024: 7871-7880
- [77] Schwarz S, Preda M, Baroncini V, et al. Emerging MPEG standards for point cloud compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2018, 9(1): 133-148
- [78] Li G, Gao W, Gao W. MPEG video-based point cloud compression (V-PCC) standard//Point Cloud Compression: Technologies and Standardization. Singapore: Springer Nature Singapore, 2024: 199-218
- [79] Li G, Gao W, Gao W. MPEG geometry-based point cloud compression (G-PCC) standard//Point Cloud Compression: Technologies and Standardization. Singapore: Springer Nature Singapore, 2024: 135-165
- [80] Radl L, Steiner M, Parger M, et al. Stopthepop: Sorted gaussian splatting for view-consistent real-time rendering. *ACM Transactions on Graphics (TOG)*, 2024, 43(4): 1-17
- [81] Schonberger J L, Frahm J M. Structure-from-motion revisited//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA. 2016: 4104-4113
- [82] Grubert G, Barthel F, Hilsmann A, et al. Improving adaptive density control for 3D gaussian splatting. *arXiv preprint arXiv: 2503.14274*, 2025
- [83] Sun X, Lee J C, Rho D, et al. F-3dgs: Factorized coordinates

- and representations for 3d gaussian splatting//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia. 2024: 7957-7965
- [84] Bagdasarian M T, Knoll P, Li Y H, et al. 3dgs. zip: A survey on 3d gaussian splatting compression methods. arXiv preprint arXiv:2407.09510, 2024
- [85] Chen Y, Chen Z, Zhang C, et al. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 21476-21485
- [86] Luiten J, Kopanas G, Leibe B, et al. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis//2024 International Conference on 3D Vision (3DV). Davos, Switzerland. 2024: 800-809
- [87] Wu G, Yi T, Fang J, et al. 4d gaussian splatting for real-time dynamic scene rendering//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 20310-20320
- [88] Li S, Li H, Liao Y, et al. NeRFCoder: Neural feature compression meets neural radiance fields for memory-efficient scene representation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 21274-21283
- [89] Gao Q, Xu Q, Cao Z, et al. Gaussianflow: Splatting gaussian dynamics for 4d content creation. arXiv preprint arXiv: 2403.12365, 2024
- [90] Yang Q, Yang L, Van Der Auwera G, et al. HybridGS: High-efficiency Gaussian splatting data compression using dual-channel sparse representation and point cloud encoder. arXiv preprint arXiv:2505.01938, 2025
- [91] Schnabel R, Klein R. Octree-based point-cloud compression. PBG@ SIGGRAPH, 2006, 3: 111-121
- [92] Zhang C, Florencio D, Loop C. Point cloud attribute compression with graph transform//2014 IEEE International Conference on Image Processing (ICIP). Paris, France. 2014: 2066-2070
- [93] De Queiroz R L, Chou P A. Compression of 3D point clouds using a region-adaptive hierarchical transform. IEEE Transactions on Image Processing, 2016, 25(8): 3947-3956
- [94] Jia W, Li L, Li Z, et al. Deep learning geometry compression artifacts removal for video-based point cloud compression. International Journal of Computer Vision, 2021, 129(11): 2947-2964
- [95] Laine S, Karras T. Efficient sparse voxel octrees-analysis, extensions, and implementation. NVIDIA Corporation, 2010, 2(6):1-30
- [96] Xue D, Ma H, Li L, et al. AiWave: Volumetric image compression with 3-D trained affine wavelet-like transform. IEEE Transactions on Medical Imaging, 2022, 42(3): 606-618
- [97] Wang C, Shen H W. A framework for rendering large time-varying data using wavelet-based time-space partitioning (wtsp) tree. Technical Report OSUCISRC-1/04-TR05, Department of Computer and Information Science, the Ohio State University, 2004
- [98] Xu H, Zhang X, Wu X. Fast Point Cloud Geometry Compression with Context-based Residual Coding and INR-based Refinement. arXiv preprint arXiv:2408.02966, 2024
- [99] Sun X, Wang J, Chen J, et al. MC-PCGC: A Space-channel mixed contextual coding for point cloud geometry compression//Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME). Niagara Falls, Canada. 2024: 1-6
- [100] Akhtar A, Li Z, Van der Auwera G. Inter-frame compression for dynamic point cloud geometry coding. IEEE Transactions on Image Processing, 2024, 33: 584-594
- [101] Shen W, Zhang B Y, Xu H, et al. Multi-space point geometry compression with progressive relation-aware transformer. IEEE Transactions on Multimedia, 2024, 26: 8969-8980
- [102] Nash C, Ganin Y, Eslami S M A, et al. Polygen: An autoregressive generative model of 3d meshes//Proceedings of the International conference on machine learning. Vienna, Austria. 2020: 7220-7229
- [103] Hahner S, Garcke J. Mesh convolutional autoencoder for semi-regular meshes of different sizes//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA. 2022: 885-894
- [104] Choy C B, Xu D, Gwak J Y, et al. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction//Proceedings of the Computer vision-ECCV 2016: Amsterdam, the Netherlands, 2016: 628-644
- [105] Gregor K, Danihelka I, Graves A, et al. Draw: A recurrent neural network for image generation//Proceedings of the International Conference on Machine Learning. Lille, France. 2015: 1462-1471
- [106] Li X, Dick A, Shen C, et al. Incremental learning of 3D-DCT compact representations for robust visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(4): 863-881
- [107] Rodler F F. Wavelet based 3d compression for very large volume data supporting fast random access. BRICS Report Series, 1999, 6(34)
- [108] Que Z, Lu G, Xu D. Voxelcontext-net: An octree based framework for point cloud compression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA. 2021: 6042-6051
- [109] Fu C, Li G, Song R, et al. Octattention: Octree-based large-scale contexts model for point cloud compression//Proceedings of the AAAI Conference on Artificial Intelligence. Online. 2022, 36(1): 625-633
- [110] Fan T, Gao L, Xu Y, et al. D-dpcc: Deep dynamic point cloud compression via 3d motion prediction. arXiv preprint arXiv: 2205.01135, 2022
- [111] De Queiroz R L, Chou P A. Motion-compensated compression of dynamic voxelized point clouds. IEEE Transactions on Image Processing, 2017, 26(8): 3886-3895
- [112] Tang D, Singh S, Chou P A, et al. Deep implicit volume compression//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA. 2020: 1293-1303

- [113] Wei J, Leutenegger S. Gsfusion: Online rgb-d mapping where gaussian splatting meets tsdf fusion. *IEEE Robotics and Automation Letters*, 2024
- [114] Liang R, Sun H, Vijaykumar N. Coordx: Accelerating implicit neural representation with a split mlp architecture. *arXiv preprint arXiv:2201.12425*, 2022
- [115] Sitzmann V, Martel J, Bergman A, et al. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 2020, 33: 7462-7473
- [116] Cheng H, Zhang M, Shi J Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 10558-10578
- [117] Lang J, Guo Z, Huang S. A comprehensive study on quantization techniques for large language models//*Proceedings of the 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*. Xiamen, China. 2024: 224-231
- [118] Shakerinava M, Sohrabi M M S, Ravanbakhsh S, et al. Weight-sharing regularization//*Proceedings of the International Conference on Artificial Intelligence and Statistics*. Valencia, Spain. 2024: 4204-4212
- [119] Moar C, Tahmasebi F, Pellauer M, et al. Characterizing the accuracy-efficiency trade-off of low-rank decomposition in language models//*Proceedings of the 2024 IEEE International Symposium on Workload Characterization (IISWC)*. Vancouver, Canada. 2024: 194-209
- [120] Mansourian A M, Ahmadi R, Ghafouri M, et al. A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*, 2025
- [121] Gou J, Yu B, Maybank S J, et al. Knowledge distillation: A survey. *International Journal of Computer Vision*, 2021, 129(6): 1789-1819
- [122] Wang Y, Han Q, Habermann M, et al. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Paris, France. 2023: 3295-3306
- [123] Pham T, Mandt S. Neural NeRF Compression. *arXiv preprint arXiv:2406.08943*, 2024
- [124] Chen Y, Wu Q, Harandi M, et al. How far can we compress instant-ngp-based nerf?//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA. 2024: 20321-20330
- [125] Wang Y, Gong Y, Zeng Y. Hyb-NeRF: a multiresolution hybrid encoding for neural radiance fields//*Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, USA. 2024: 3677-3686
- [126] Lee Y, Yang L, Fan D. MFNeRF: Memory efficient NeRF with mixed-feature Hash table//*Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Tucson, USA. 2025: 2686-2695
- [127] Chen Y, Wu Q, Lin W, et al. Hac: Hash-grid assisted context for 3d gaussian splatting compression//*Proceedings of the European Conference on Computer Vision*. Milan, Italy. Cham: Springer Nature Switzerland, 2024: 422-438
- [128] Chen Y, Wu Q, Lin W, et al. Hac++ : Towards 100x compression of 3d gaussian splatting. *arXiv preprint arXiv:2501.12255*, 2025
- [129] Lu T, Yu M, Xu L, et al. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA. 2024: 20654-20664
- [130] Lee J C, Rho D, Sun X, et al. Compact 3d gaussian splatting for static and dynamic radiance fields. *arXiv preprint arXiv:2408.03822*, 2024
- [131] Liu M, Yang Q, Huang H, et al. Light4gs: Lightweight compact 4d gaussian splatting generation via context model. *arXiv preprint arXiv:2503.13948*, 2025
- [132] Yuan Y, Shen Q, Yang X, et al. 1000+ fps 4d gaussian splatting for dynamic scene rendering//*Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems*. San Diego, USA. 2025
- [133] Tang L, Zhai Y, Wang R. Compression-aware tuning for compressing volumetric radiance fields// *Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP)*. Abu Dhabi, United Arab Emirates. 2024: 2250-2256
- [134] Yan Z, Li C, Lee G H. Nerf-ds: Neural radiance fields for dynamic specular objects//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vancouver, Canada. 2023: 8285-8295
- [135] Zhan Y, Li Z, Niu M, et al. KFD-NeRF: Rethinking dynamic NeRF with Kalman filter//*Proceedings of the European Conference on Computer Vision*. Milan, Italy. Cham: Springer Nature Switzerland, 2024: 1-18
- [136] Wang H, Zhu H, He T, et al. End-to-end rate-distortion optimized 3d gaussian representation//*Proceedings of the European Conference on Computer Vision*. Milan, Italy. 2024: 76-92
- [137] Li B, Chen Y, Zhang D, et al. DeformStream: Deformation-based adaptive volumetric video streaming. *arXiv preprint arXiv:2409.16615*, 2024
- [138] Li X, Salehi M A, Bayoumi M, et al. Cost-efficient and robust on-demand video transcoding using heterogeneous cloud services. *IEEE Transactions on Parallel and Distributed Systems*, 2017, 29(3): 556-571
- [139] Bukhari S M A H, Baccour E, Bilal K, et al. To transcode or not? A machine learning based edge video caching and transcoding strategy. *Computers and Electrical Engineering*, 2023, 109: 108741
- [140] Pakdaman F, Adelimanesh M A, Gabbouj M, et al. Complexity analysis of next-generation VVC encoding and decoding//*Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP)*. Abu Dhabi, United Arab Emirates. 2020: 3134-3138
- [141] Lee K, Yi J, Lee Y, et al. GROOT: A real-time streaming system of high-fidelity volumetric videos//*Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. London, UK. 2020: 1-14

- [142] Zhang A, Wang C, Han B, et al. YuZu: Neural-enhanced volumetric video streaming//Proceedings of the 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22). Renton, USA. 2022: 137-154
- [143] Viola I, Cesar P. Volumetric video streaming. *Immersive Video Technologies* (2022), 2022, 425
- [144] Wang L, Hu Q, He Q, et al. Neural residual radiance fields for streamably free-viewpoint videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada. 2023: 76-87
- [145] Hu Q, Zhong H, Zheng Z, et al. VRVVC: Variable-rate NeRF-based volumetric video compression//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA. 2025, 39(4): 3563-3571
- [146] Zheng Z, Zhong H, Hu Q, et al. Hpc: Hierarchical progressive coding framework for volumetric video//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia. 2024: 7937-7946
- [147] Wang C, Zhang A, Yang Y, et al. VoLUT: Efficient Volumetric streaming enhanced by LUT-based super-resolution. *arXiv preprint arXiv:2502.12151*, 2025
- [148] Li R, Li X, Fu C W, et al. Pu-gan: a point cloud upsampling adversarial network//Proceedings of the IEEE/CVF international conference on computer vision. Seoul, Republic of Korea. 2019: 7203-7212
- [149] Yifan W, Wu S, Huang H, et al. Patch-based progressive 3d point set upsampling//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Long Beach, USA. 2019: 5958-5967
- [150] Preda M. 3D graphics compression core experiments description. ISO/IEC JTC 1/SC 29/WG 11 N8499, 2006
- [151] Chen R, Xiao M, Yu D, et al. patchVVC: A real-time compression framework for streaming volumetric videos//Proceedings of the 14th Conference on ACM Multimedia Systems. Vancouver, Canada. 2023: 119-129
- [152] Tohidi F, Paul M, Ulhaq A, et al. Improved video-based point cloud compression via segmentation. *Sensors*, 2024, 24(13): 4285
- [153] Fréneau L, Gautier G, Mercat A, et al. uvgVPCcenc: Practical open-source encoder for fast V-PCC compression//Proceedings of the 16th ACM Multimedia Systems Conference. Stellenbosch, South Africa. 2025: 235-241
- [154] Pan Z, Xiao M, Han X, et al. patchdpcc: A patchwise deep compression framework for dynamic point clouds//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada. 2024, 38(5): 4406-4414
- [155] Yin Z X, Qiu J, Cheng M M, et al. Multi-space neural radiance fields//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada. 2023: 12407-12416
- [156] Zheng Z, Zhong H, Hu Q, et al. JointRF: end-to-end joint optimization for dynamic neural radiance field representation and compression//Proceedings of the 2024 IEEE International Conference on Image Processing (ICIP). Abu Dhabi, United Arab Emirates. 2024: 3292-3298
- [157] Hu J, Shaikh A, Bahremand A, et al. Characterizing real-time dense point cloud capture and streaming on mobile devices//Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges. New Orleans, USA. 2021: 1-6
- [158] Guan Y, Hou X, Wu N, et al. Metastream: Live volumetric content capture, creation, delivery, and rendering in real time//Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. Madrid, Spain. 2023: 1-15
- [159] De Fré M, van der Hooft J, Wauters T, et al. Scalable MDC-based volumetric video delivery for real-time one-to-many WebRTC conferencing//Proceedings of the 15th ACM Multimedia Systems Conference. Bari, Italy. 2024: 121-131
- [160] De Fré M, van der Hooft J, Wauters T, et al. Demonstrating adaptive many-to-many immersive teleconferencing for volumetric video//Proceedings of the 15th ACM Multimedia Systems Conference. Bari Italy. 2024: 453-458
- [161] Huang Y, Zhu Y, Qiao X, et al. AITransfer: Progressive AI-powered transmission for real-time point cloud video streaming//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China. 2021: 3989-3997
- [162] Zhu Y, Huang Y, Qiao X, et al. A semantic-aware transmission with adaptive control scheme for volumetric video service. *IEEE Transactions on Multimedia*, 2022, 25: 7160-7172
- [163] Collet, Y. LZ4: Extremely fast compression algorithm. <https://lz4.github.io/lz4/>, 2023
- [164] NVIDIA. (2022). Improving Video Quality and Performance with AV1 and NVIDIA Ada Lovelace Architecture. Retrieved from <https://developer.nvidia.com/blog/improving-video-quality-and-performance-with-av1-and-nvidia-ada-lovelace-architecture/>
- [165] Foxglove. (2023). Announcing: Support for H.265, VP9, and AV1 Video Codecs. Retrieved from <https://foxglove.dev/blog/announcing-support-for-h265-vp9-and-av1-video-codecs>
- [166] Jin T, Dasa M, Smith C, et al. Meshreduce: Scalable and bandwidth efficient 3d scene capture//Proceedings of the 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). Orlando, USA. 2024: 20-30
- [167] Ma S, Simon T, Saragih J, et al. Pixel codec avatars//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA. 2021: 64-73
- [168] Orts-Escolano S, Rhemann C, Fanello S, et al. Holoportation: Virtual 3d teleportation in real-time//Proceedings of the 29th annual symposium on user interface software and technology. Tokyo, Japan. 2016: 741-754
- [169] Doumanoglou A, Drakoulis P, Zioulis N, et al. Benchmarking open-source static 3D mesh codecs for immersive media interactive live streaming. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019, 9(1): 190-203
- [170] Wang L, Yao K, Guo C, et al. Videorf: Rendering dynamic radiance fields as 2d feature video streams//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 470-481
- [171] Bao Z, Li Q, Liao G, et al. MGStream: Motion-aware 3D Gaussian for Streamable Dynamic Scene Reconstruction. *arXiv*

- preprint arXiv:2505.13839, 2025
- [172] Sun J, Jiao H, Li G, et al. 3dstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA. 2024: 20675-20685
- [173] Liang Z, Liu J, Dasari M, et al. Fumos: Neural compression and progressive refinement for continuous point cloud video streaming. *IEEE Transactions on Visualization and Computer Graphics*, 2024, 30(5): 2849-2859
- [174] Shi Y, Morin G, Gasparini S, et al. Lapisgs: Layered progressive 3d gaussian splatting for adaptive streaming. arXiv preprint arXiv:2408.14823, 2024
- [175] Katsumata K, Vo D M, Nakayama H. A compact dynamic 3d gaussian representation for real-time dynamic view synthesis//Proceedings of the European Conference on Computer Vision. Milan, Italy. 2024: 394-412
- [176] Zhang A, Wang C, Han B, et al. Efficient volumetric video streaming through super resolution//Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications. Online. s2021: 106-111
- [177] Liu Y, Han B, Qian F, et al. Vues: practical mobile volumetric video streaming through multiview transcoding//Proceedings of the 28th Annual International Conference on Mobile Computing And Networking. Sydney, Australia. 2022: 514-527
- [178] He Z, Yang W, Li L, et al. Iterative removal of G-PCC attribute compression artifacts based on a graph neural network. *Electronics*, 2024, 13(18): 3768
- [179] Cesium. (n.d.). Tiler data types and formats. Retrieved from <https://cesium.com/learn/3d-tiling/tiler-data-formats/>
- [180] Janea Systems. (2025). How 3D tiles and cesiumJS transform massive 3D data. Retrieved from <https://www.janeasystems.com/blog/streaming-large-3d-scenes-with-3d-tiles>
- [181] Cesium. Draco compressed meshes with glTF and 3D tiles. Retrieved from <https://cesium.com/blog/2018/04/09/draco-compression/>, 2018
- [182] MPEG Group. [MPEG127-m48892][G-PCC] CE13.2 Report on Point Cloud Tile and Slice Based Coding. Retrieved from <https://mpeg-pcc.org/index.php/public-contributions/mpeg127-m48892g-pcc-ce13-2-report-on-point-cloud-tile-and-slice-based-coding/>
- [183] InterDigital. V-PCC: The first MPEG codec for point cloud compression. Retrieved from <https://www.interdigital.com/videos/v-pcc-the-first-mpeg-codec-for-point-cloud-compression>
- [184] Chen A, Mao S, Li Z, et al. An introduction to point cloud compression standards. *GetMobile: Mobile Computing and Communications*, 2023, 27(1): 11-17
- [185] Rossignac, Jarek. 3D Mesh Compression. (2005): 359-379.
- [186] AWS. (2023). Convert glTF to 3D tiles for streaming of large models in AWS IoT twinMaker. Retrieved from <https://aws.amazon.com/blogs/iot/convert-gltf-to-3d-tiles-for-streaming-of-large-models-in-aws-iot-twinmaker/>
- [187] Lam, K. Y., ResearchGate. VVGLTF: Efficient Streaming of Volumetric Video with glTF. Retrieved from https://www.researchgate.net/publication/385243485_VVGLTF_EFFICIENT_STREAMING_OF_VOLUMETRIC_VIDEO_WITH_GLTF
- [188] AGI.. Stream Massive 3D Datasets with 3D Tiles. Retrieved from <https://www.agi.com/blog/2019/04/stream-massive-3d-datasets-with-3d-tiles>
- [189] Zhou C, Xiao M, Liu Y. Clustile: Toward minimizing bandwidth in 360-degree video streaming//Proceedings of the IEEE INFOCOM 2018-IEEE Conference on Computer Communications. Honolulu, USA. 2018: 962-970
- [190] Kattadige C, Thilakarathna K. VAD360: Viewport aware dynamic 360-degree video frame tiling. arXiv preprint arXiv: 2105.11563, 2021
- [191] Hadizadeh H, Bajić I V. Learned scalable video coding for humans and machines. *EURASIP Journal on Image and Video Processing*, 2024, 2024(1): 41
- [192] Bonneineau C, Hamidouche W, Travers J F, et al. Caesar: Conditional autoencoder and super-resolution for learned spatial scalability//Proceeding of the 2021 International Conference on Visual Communications and Image Processing (VCIP). Munich, Germany. 2021: 1-5
- [193] Van der Auwera G, David P T, Reisslein M, et al. Traffic and quality characterization of the H. 264/AVC scalable video coding extension. *Advances in Multimedia*, 2008, 2008(1): 164027
- [194] Zhang J, Wang J, Ding D, et al. Scalable point cloud attribute compression. *IEEE Transactions on Multimedia*, 2023, 27: 889-899
- [195] Stefanoski N, Ostermann J, Klie P. Scalable compression of time-consistent 3D mesh sequences: U. S. Patent 8, 502, 815. 2013-8-6
- [196] Mao X, Yuan H, Guo T, et al. SPAC: Sampling-based progressive attribute compression for dense point clouds. *IEEE Transactions on Image Processing*, 2025, 34: 2939-2953
- [197] Yang L, Deng B, Zhang J. Scalable and high-quality neural implicit representation for 3D reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2025: 1-17
- [198] Li J, Li Z, Liu Z, et al. Viewport prediction for volumetric video streaming by exploring video saliency and trajectory information. arXiv preprint arXiv:2311.16462, 2023
- [199] Hu K, Yang H, Jin Y, et al. Understanding user behavior in volumetric video watching: Dataset, analysis and prediction//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada. 2023: 1108-1116
- [200] Fang X, Zhang Y, Tan H, et al. Performance evaluation and optimization of 3D gaussian splatting in indoor scene generation and rendering. *ISPRS International Journal of Geo-Information*, 2025, 14(1): 21
- [201] Wang S. Low-latency adaptive media streaming: From 2D to 6-DoF. Binghamton, USA: State University of New York at Binghamton, 2022
- [202] Yan Z, Pei J, Wu H, et al. Semantic-aware adaptive video streaming using latent diffusion models for wireless networks. arXiv preprint arXiv:2502.05695, 2025
- [203] Mari D, Camuffo E, Milani S. CACTUS: Content-Aware

- Compression and Transmission Using Semantics for Automotive LiDAR Data. *Sensors*, 2023, 23(12): 5611
- [204] Wang Y, Chan P H, Donzella V. Semantic-aware video compression for automotive cameras. *IEEE Transactions on Intelligent Vehicles*, 2023, 8(6): 3712-3722
- [205] Liu Y, Shu Z, Li Y, et al. 3d-fm gan: Towards 3d-controllable face manipulation// *Proceedings of the European Conference on Computer Vision*. Tel Aviv, Israel. 2022: 107-125
- [206] Yu X, Zhu H, He T, et al. Gaussiansr: 3d gaussian super-resolution with 2d diffusion priors. *arXiv preprint arXiv: 2406.10111*, 2024
- [207] Wan Y, Shao M, Cheng Y, et al. S2Gaussian: Sparse-view super-resolution 3D Gaussian splatting//*Proceedings of the Computer Vision and Pattern Recognition Conference*. Nashville, USA. 2025: 711-721
- [208] Liu Y, Chen R, Li Y, et al. SPU-PMD: Self-supervised point cloud upsampling via progressive mesh deformation//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, USA. 2024: 5188-5197
- [209] Zhang Z, Chen H, Cao X, et al. Anableps: Adapting bitrate for real-time communication using VBR-encoded video//*2023 IEEE International Conference on Multimedia and Expo (ICME)*. Brisbane, Australia. 2023: 1685-1690
- [210] SRT Alliance. (2023). Secure Reliable Transport (SRT). <https://www.srtalliance.org/>
- [211] Video Services Forum. (2023). Reliable Internet Stream Transport (RIST). <https://www.videoservicesforum.org/activities/rist/>
- [212] Maura F, Casasnovas M, Bellalta B. Experimenting with adaptive bitrate algorithms for virtual reality streaming over Wi-Fi//*Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. Washington, USA. 2024: 1930-1937
- [213] Cloudflare. What is Adaptive Bitrate Streaming? <https://www.cloudflare.com/learning/video/what-is-adaptive-bitrate-streaming/>
- [214] Dai Y, Xue C, Zhou L. Visual saliency guided perceptual adaptive quantization based on HEVC intra-coding for planetary images. *Plos One*, 2022, 17(2): e0263729
- [215] Xie L, Gao W, Zheng H, et al. Roi-guided point cloud geometry compression towards human and machine vision// *Proceedings of the 32nd ACM International Conference on Multimedia*. Melbourne, Australia. 2024: 3741-3750
- [216] Moskalenko A, Bryncev A, Vatolin D, et al. AIM 2024 challenge on video saliency prediction: Methods and results// *Proceedings of the European Conference on Computer Vision*. Milan, Italy. 2025: 178-194
- [217] Perugachi-Diaz Y, Sautière G, Abati D, et al. Region-of-interest based neural video compression. *arXiv preprint arXiv: 2203.01978*, 2022
- [218] Schult J, Engelmann F, Hermans A, et al. Mask3d: Mask transformer for 3d semantic instance segmentation//*Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*. London, UK. 2023: 8216-8223
- [219] Wu Y, Yan H, Ding K. Transformer based 3D tooth segmentation via point cloud region partition. *Scientific Reports*, 2024, 14(1): 28513
- [220] Qian G, Li Y, Peng H, et al. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 2022, 35: 23192-23204
- [221] Sun Y, Zhang X, Miao Y. A review of point cloud segmentation for understanding 3D indoor scenes. *Visual Intelligence*, 2024, 2(1): 14
- [222] Zeng Z, Hu Q, Xie Z, et al. Small but mighty: Enhancing 3d point clouds semantic segmentation with u-next framework. *International Journal of Applied Earth Observation and Geoinformation*, 2025, 136: 104309
- [223] Yang B, Zhang L, Li F W B, et al. Motion-aware compression and transmission of mesh animation sequences. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2019, 10(3): 1-21
- [224] Zhang Z, Lu G, Liang H, et al. Efficient dynamic-nerf based volumetric video coding with rate distortion optimization// *Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME)*. Niagara Falls, Canada. 2024: 1-6
- [225] Fan T, Hu Y, Wang Y. U-Motion: Learned point cloud video compression with U-structured motion estimation. *arXiv preprint arXiv:2411.14501*, 2024
- [226] Jiang Y, Shen Z, Guo C, et al. RePerformer: Immersive human-centric volumetric videos from playback to photoreal reperformance. *arXiv preprint arXiv:2503.12242*, 2025
- [227] Hashimoto N, Takamaeda-Yamazaki S. Fadec: FPGA-based acceleration of video depth estimation by hw/sw co-design// *2022 International Conference on Field-Programmable Technology (ICFPT)*. Hong Kong, China. 2022: 1-9
- [228] Intel Corporation. (2025). Optimizing Explicit SIMD Kernels. Retrieved from <https://www.intel.com/content/www/us/en/docs/oneapi/optimization-guide-gpu/2025-0/optimizing-explicit-simd-kernels.html>
- [229] Zhang Y, Liang J, Peng J, et al. SpNeRF: Memory efficient sparse volumetric neural rendering accelerator for edge devices// *Proceedings of the 2025 Design, Automation & Test in Europe Conference (DATE)*. Verona, Italy. 2025: 1-7
- [230] GitHub. (n.d.). A curated list of awesome GPGPU (CUDA/ OpenCL/Vulkan) resources. Retrieved from <https://github.com/jslee02/awesome-gpgpu>
- [231] Xiao Q, Zheng S, Wu B, et al. Hasco: Towards agile hardware and software co-design for tensor computation// *Proceedings of the 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. Online. 2021: 1055-1068
- [232] Zhang Y, Kamath A K, Wu Q, et al. Data-model-circuit tri-design for ultra-light video intelligence on edge devices// *Proceedings of the 28th Asia and South Pacific Design Automation Conference*. Tokyo, Japan. 2023: 745-750
- [233] Aydin S, Samet R, Bay O F. Real-time parallel image processing applications on multicore CPUs with OpenMP and GPGPU with CUDA. *The Journal of Supercomputing*, 2018,

- 74: 2255-2275
- [234] Medialooks. Mastering Video Codecs: H.264, H.265, and AV1 Compared. Retrieved from <https://medialooks.com/articles/mastering-video-codecs-h-264-h-265-and-av1-compared/>
- [235] AMD. AMD Alveo™ MA35D Media Accelerator. Retrieved from <https://www.amd.com/en/products/accelerators/alveo/ma35d.html>
- [236] Callaba. How does FPGA make real-time video processing possible?. Retrieved from <https://callaba.io/how-does-fpga-make-real-time-video-processing-possible>
- [237] Gateworks. Choosing the Right AI Accelerator | NPU or TPU for Edge and Cloud Applications. Retrieved from <https://www.gateworks.com/choosing-the-right-ai-accelerator-npu-or-tpu-for-edge-and-cloud-applications/>
- [238] Lin Y, Zhang Z, Tang H, et al. Pointacc: Efficient point cloud accelerator//Proceedings of the MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture. Online. 2021: 449-461
- [239] Gül S, Podborski D, Buchholz T, et al. Low-latency cloud-based volumetric video streaming using head motion prediction//Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. Istanbul, Türkiye. 2020: 27-33
- [240] Chen W, Chen S, Leng J, et al. A review of cloud-edge SLAM: Toward asynchronous collaboration and implicit representation transmission. IEEE Transactions on Intelligent Transportation Systems, 2024; 25(11): 15437-15453
- [241] Grulich P M, Nawab F. Collaborative edge and cloud neural networks for real-time video processing. Proceedings of the VLDB Endowment, 2018, 11(12): 2046-2049
- [242] Zerman E, Gao P, Ozcinar C, et al. Subjective and objective quality assessment for volumetric video compression. Electronic Imaging, 2019, 31: 1-7
- [243] Mekuria R, Blom K, Cesar P. Design, implementation, and evaluation of a point cloud codec for tele-immersive video. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 27(4): 828-842
- [244] Barman N, Martini M G, Reznik Y. Bjøntegaard delta (bd): A tutorial overview of the metric, evolution, challenges, and recommendations. arXiv preprint arXiv:2401.04039, 2024
- [245] Wu T, Pan L, Zhang J, et al. Density-aware chamfer distance as a comprehensive metric for point cloud completion. arXiv preprint arXiv:2111.12702, 2021
- [246] Javaheri A, Brites C, Pereira F, et al. A generalized Hausdorff distance based quality metric for point cloud geometry//Proceedings of the 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). Athlone, Ireland. 2020: 1-6
- [247] Hore A, Ziou D. Image quality metrics: PSNR vs. SSIM//Proceedings of the 2010 20th International Conference on Pattern Recognition. Istanbul, Türkiye. 2010: 2366-2369
- [248] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600-612
- [249] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment//Proceedings of the Thirtieth Asilomar Conference on Signals, Systems & Computers, 2003. Pacific Grove, USA. 2003, 2: 1398-1402
- [250] Sheikh H R, Bovik A C. Image information and visual quality. IEEE Transactions on Image Processing, 2006, 15(2): 430-444
- [251] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA. 2018: 586-595
- [252] Series B. Methodology for the subjective assessment of the quality of television pictures. Recommendation ITU-R BT, 2012, 500(13)
- [253] Mao Dong-Xing, Yu Wu-Zhou, Wang Zuo-Min. Statistical validation and criterion for paired comparison data in sound quality evaluation. Acta Acustica, 2005, 30(5): 468-472 (in Chinese)
(毛东兴, 俞悟周, 王佐民. 声品质成对比较主观评价的数据检验及判据. 声学学报, 2005, 30(5): 468-472)
- [254] Union I T. Methods for the subjective assessment of video quality audio quality and audiovisual quality of internet video and distribution quality television in any environment. SERIES P: Terminals And Subjective And Objective Assessment Methods, 2016
- [255] Seufert M, Egger S, Slanina M, et al. A survey on quality of experience of HTTP adaptive streaming. IEEE Communications Surveys & Tutorials, 2014, 17(1): 469-492
- [256] Chen Y H, Ho K W, Benjak M, et al. On the rate-distortion-complexity trade-offs of neural video coding//Proceedings of the 2024 IEEE 26th International Workshop on Multimedia Signal Processing. West Lafayette, USA. 2024: 1-6
- [257] Zhang C, Gao W. AdaDPCC: Adaptive Rate Control and Rate-Distortion-Complexity Optimization for Dynamic Point Cloud Compression//Proceedings of the AAAI Conference on Artificial Intelligence. Philadelphia, USA. 2025, 39(12): 13188-13196
- [258] Warburton M, Mon-Williams M, Mushtaq F, et al. Measuring motion-to-photon latency for sensorimotor experiments with virtual reality systems. Behavior Research Methods, 2023, 55(7): 3658-3678
- [259] Herglotz C, Robitza W, Kränzler M, et al. Modeling of energy consumption and streaming video QoE using a crowdsourcing dataset//Proceedings of the 2022 14th International Conference on Quality of Multimedia Experience (QoMEX). Lippstadt, Germany. 2022: 1-6
- [260] Gao Y, Feng R, Guo Z, et al. Exploring the rate-distortion-complexity optimization in neural image compression. Journal of Visual Communication and Image Representation, 2024, 105: 104294
- [261] ISO/IEC. Information technology — Coded representation of immersive media — Part 5: Video-based point cloud compression: ISO/IEC 23090-5:2021. Geneva: International Organization for Standardization, 2021
- [262] d'Eon E, Harrison J, Myers K. 8i Voxelized Full Bodies - A V-PCC Dataset. ISO/IEC JTC1/SC29/WG11, MPEG2017/

- M40059, 2017 [2024-05-30], <https://github.com/MPEGGroup/mpeg-pcc-tmc2>
- [263] Valenzise G, Tagliasacchi M, Tubaro S, et al. The Microsoft voxelized upper bodies dataset//Proceedings of the 10th ACM Multimedia Systems Conference. Amherst, USA. 2019: 223-228
- [264] INTERACT Project. INTERACT: Intelligent multimodal interfaces for interaction with active spaces. EU FP7 Project 600578, 2012-2015 [2024-05-30]
- [265] Grassal A, Keil J, Wölfel M. Real-time point cloud data acquisition using Azure Kinect and HoloLens 2. Proceedings of the ACM on Human-Computer Interaction, 2021, 5(EICS): 1-22
- [266] Damen D, Doughty H, Farinella G M, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany. 2018: 720-736
- [267] Ionescu C, Papava D, Olaru V, et al. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325-1339
- [268] Xiang X, Ma L, Wang S, et al. THuman: A real-time multi-view human capture system with skeleton-aware reconstruction. arXiv:1912.11216, 2019. <https://arxiv.org/abs/1912.11216>
- [269] Li T, Slavcheva M, Zollhoefer M, et al. Neural 3d video synthesis from multi-view video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA. 2022: 5521-5531
- [270] Park K, Sinha U, Hedman P, et al. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228, 2021
- [271] Broxton M, Flynn J, Overbeck R, et al. Immersive light field video with a layered mesh representation. ACM Transactions on Graphics (TOG), 2020, 39(4): 86: 1-86: 15
- [272] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (TOG), 2015, 34(6): 248:1-248:16
- [273] Pavlakos G, Choutas V, Ghorbani N, et al. Expressive body capture: 3D hands, face, and body from a single image//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA. 2019: 10975-10985
- [274] Huang Y, Wang J, Shi Y, et al. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. arXiv preprint arXiv:2306.12422, 2023
- [275] He X, Chen J, Peng S, et al. Gvgen: Text-to-3d generation with volumetric representation// Proceedings of the European Conference on Computer Vision. Milan, Italy. 2024: 463-479
- [276] Liao T, Yi H, Xiu Y, et al. Tada! text to animatable digital avatars// Proceedings of the 2024 International Conference on 3D Vision (3DV). Davos, Switzerland. 2024: 1508-1519



HUANG Ya-Kun, Ph. D., associate professor. His current research interests include immersive communication, volumetric video streaming, distributed systems, and intelligent AR/VR services.

ZHU Yuan-Wei, Ph. D., lecturer. His current research interests include volumetric video streaming and reinforcement learning.

QIU Li-Rui, Ph. D. candidate. Her current research interests include volumetric video codec and 3D Gaussian splatting.

LI Yan, engineer. His current research interests include heterogeneous acceleration technology.

ZHANG Hao, M. S., senior engineer. His current research interests include spatial computing, immersive communication, point cloud streaming and AR/VR intelligent services.

LV Jing, M. S., engineer. Her current research interests include spatial computing, immersive communication, point cloud streaming and AR/VR intelligent services.

QIAO Xiu-Quan, Ph. D., professor. His current research interests include immersive communication, digital twin and metaverse, 6G hologram communication, and intelligent AR/VR service.

Background

Immersive volumetric video, a cornerstone of next-generation virtual reality (VR), augmented reality (AR), and mixed reality (MR), has gained unprecedented attention due to its ability to enable six degrees of freedom (6DoF) experiences. Unlike traditional 2D or panoramic video, volumetric content captures the full 3D geometry and appearance of scenes, but its massive data volume poses critical challenges in storage,

transmission, and real-time rendering. Efficient compression and coding techniques are thus imperative to realize scalable immersive communication. This survey, supported by the National Natural Science Foundation of China (NSFC) (No. 62431015) and (No. 62571055), systematically reviews advancements in volumetric video coding, bridging classical methods and cutting-edge neural representations to address these

challenges.

Volumetric video compression lies at the intersection of computer graphics, computer vision, and multimedia communication. Internationally, research has evolved from standardized geometric approaches—such as MPEG’s Video-based Point Cloud Compression (V-PCC) and Geometry-based Point Cloud Compression (G-PCC)—to data-driven neural representations like Neural Radiance Fields (NeRF) and 3D Gaussian Splatting. While traditional methods excel in offline scenarios (e. g. , movie production), they struggle with dynamic content and real-time constraints. Neural representations, though promising for fidelity and adaptability, face hurdles in computational efficiency and standardization. Our research group has previously contributed to this field through efficient point cloud compression algorithms and

latency-aware neural coding frameworks, laying the groundwork for this comprehensive analysis.

This survey advances the NSFC project’s overarching goal of “intelligent multimedia communication for immersive systems” by focusing on the coding efficiency and real-time adaptability sub-objectives. We categorize techniques into offline (e. g. , MPEG standards, NeRF compression) and real-time paradigms (e. g. , sensor fusion, edge-assisted decoding), critically evaluating their trade-offs in rate-distortion-accuracy (RDA) performance. By synthesizing optimization strategies (e. g. , hybrid geometric-neural pipelines) and identifying open challenges (e. g. , cross-modal synchronization), this work not only consolidates fragmented knowledge but also informs future standards, such as MPEG Immersive Video (MIV).