计算机科学
COMPUTER SCIENCE

基于三维高斯溅射的低码率实时多视点视频流传输

王义总, 宁泓博, 王昊峰, 马思伟, 高文

引用本文

王义总, 宁泓博, 王昊峰, 马思伟, 高文. 基于三维高斯溅射的低码率实时多视点视频流传输[J]. 计算机科学, 2026, 53(3): 225-230.

WANG Yizong, NING Hongbo, WANG Haofeng, MA Siwei, GAO Wen. Low-bitrate and Real-time Multiview Video Streaming with 3D Gaussian Splatting [J]. Computer Science, 2026, 53(3): 225-230.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250700104

基于三维高斯溅射的低码率实时多视点视频流传输

王义总[1]　宁泓博[2]　王昊峰[2]　马思伟[1]　高　文[1]
1 北京大学计算机学院　北京 100871
2 北京大学信息工程学院　广东 深圳　518055
(wang@pku.edu.cn)

摘　要　多视点视频能够为用户提供沉浸式体验并支持多种应用,但其传输带宽需求远高于传统视频。现有多视点编码算法主要利用二维视点间的冗余信息,未考虑三维空间冗余。为此,提出一种多视点视频流传输方法,将多视点视频转换为稀疏视点紧凑表示来降低三维空间冗余,并基于该表示在接收端进行三维重建,合成剩余视点。具体包括:1)提出一种基于稀疏视点的多视点视频紧凑表示,利用三维高斯重建与溅射合成剩余视点;2)设计视点选择方法,以优化合成视点的视觉质量。实验表明,提出的系统相比基线方法可降低至少44.6%的码率,同时支持端到端30 FPS以上的实时传输。

关键词:视频流传输;多视点视频;三维高斯溅射;沉浸式视频;三维重建
中图分类号　TP391

Low-bitrate and Real-time Multiview Video Streaming with 3D Gaussian Splatting

WANG Yizong[1],NING Hongbo[2],WANG Haofeng[2],MA Siwei[1] and GAO Wen[1]
1 School of Computer Science,Peking University,Beijing 100871,China
2 School of Electronic and Computer Engineering,Peking University,Shenzhen,Guangdong 518055,China

Abstract　Multiview videos can offer viewers immersive experiences and enable a variety of applications,but they require times of transmission bandwidth compared to traditional videos. Current multiview coding algorithms mainly leverage redundancy between 2D views and do not consider 3D spatial redundancy. This paper presents a multiview video streaming approach that transforms multiview video content into a compact sparse-view representation to reduce redundancy in 3D space. At the receiver side,the remaining views are synthesized through 3D reconstruction based on this representation. Specifically,this paper proposes a compact multiview video representation based on sparse-views,where the remaining views are synthesized using 3D Gaussian reconstruction and splatting,and a view selection method that selects views to optimize visual quality of synthesized views. Experiments show that the proposed method achieves at least a 44.6% bitrate reduction compared with the baseline and supports end-to-end streaming at over 30 FPS.

Keywords　Video streaming,Multiview video,3D Gaussian splatting,Immersive video,3D reconstruction

1　引言

沉浸式应用的快速发展推动着多视点视频技术在虚拟现实(VR)[1]、增强现实(AR)[2]、三维场景理解[3]、自动驾驶[4]及医学成像[5]等领域的广泛应用。以VR/AR为代表的沉浸式应用依赖高质量多视点视觉内容构建用户体验,其产生的海量数据对存储与传输提出严峻挑战。与此同时,裸眼三维显示技术可支持高密度视点呈现(如32视点[6])。这使得高效编码与传输成为应对上述领域数据需求增长的关键技术。

当前主流编码标准通过扩展基础视频编码框架并利用多视点间冗余信息实现多视点视频压缩,典型代表包括基于H.264的MVC标准[7]、基于H.265的MV-HEVC标准[8],以及基于H.266的VVC-ML标准[9]。此外,MPEG组织开发的沉浸式视频标准MIV[10-11]利用深度信息和纹理信息,可支持六自由度沉浸式体验。在深度学习领域,基于神经网络的多视点图像编码方法已取得显著进展[12-15],但受限于计算复杂度,其在视频编码中处理速度受限,仅能达到约10帧每秒[16]。

现有多视点处理方法主要建立视点间的平面级关联(见图1左侧),而视点与拍摄对象间存在空间级关联,即多视点

到稿日期:2025-07-16　返修日期:2025-10-18
基金项目:北京市自然科学基金(L242014);鹏城实验室科教基金会—中国移动科创专项(2024ZY1C0040);"国家资助博士后研究人员计划"和"中国博士后科学基金"(BX20250382)
This work was supported by the Beijing Natural Science Foundation (L242014), PCL-CMCC Foundation for Science and Innovation (2024ZY1C0040) and Postdoctoral Fellowship Program and China Postdoctoral Science Foundation(BX20250382).
通信作者:马思伟(swma@pku.edu.cn)

像素间的空间对应关系。典型代表如三维重建算法正是通过挖掘这种空间级关联构建体素模型或动态三维序列。因此，本研究着力于利用空间级关联实视点数量精简（见图 1 右侧）。具体而言，通过对部分视点进行紧凑表征编码以显著降低码率，解码端则基于传输视点重建三维场景模型，进而实时渲染输出全部视点内容。

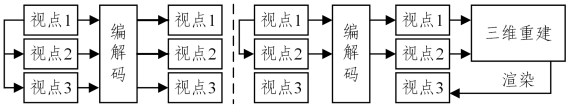


图 1 多视点视频传输路线对比图

Fig. 1 Comparison system frameworks of multiview video streaming

本文引入三维高斯溅射（3DGS）[17]技术，该技术通过多视点图像生成三维场景表征并实现新视点合成。其核心在于采用彩色高斯元混合表征场景，通过概率化高斯分布精确建模几何细节与连续深度纹理变化。该方法具有两大显著优势：1）高斯函数的概率特性可实现对复杂曲面的平滑精确建模，构建具有丰富细节的三维环境；2）原生的可微分渲染器支持高效实时推理，满足实时流媒体需求。

本文提出基于三维高斯溅射（3DGS）的多视点视频流系统 GaussianViews，该系统相比现有方案显著降低了码率，并且能够以 30 FPS 运行。如图 1 所示，GaussianViews 在编码端采用视点选择策略对部分视点进行压缩编码，只需传输两个视点，显著降低了带宽消耗；解码端则基于传输的稀疏视点，通过 3DGS 渲染输出所有视点。

然而，GaussianViews 的系统设计面临两个主要技术挑战：

1）三维高斯模型实时重建与多视角合成。尽管 GaussianViews 显著降低了传输视点的数量，但三维重建过程仍存在较大的计算开销，导致处理时延显著增加，难以满足实时性需求。3DGS 不仅需要多视角输入及较长的优化周期，且需逐场景独立优化，未能有效挖掘视频间的共性特征信息，严重制约了视频流媒体场景下的实时三维重建应用。

本研究提出基于三维高斯泼溅的多视点视频编码增强算法。该算法通过立体视差估计提取选定视角的深度特征，构建了无需迭代优化的三维高斯模型参数回归机制，有效保障系统实时性。同时充分利用高斯泼溅的快速渲染特性，实现剩余视点的实时合成。通过在大规模人体数据集上进行预训练，确保参数回归器具备跨场景泛化能力，最终在稀疏视点条件下达成免微调的三维重建与实时渲染。

2）合成视点视觉质量最优化的稀疏视点选择。选择过近的视点会导致场景覆盖不全，进而引发重建完整性缺失；而选择过远的视点则会造成视差估计精度下降，导致重建质量劣化。视点选择不仅受视点索引影响，更与视点-目标物间距呈现强相关性，这使得视点选择十分困难。

针对三维高斯重建需求，创新性地设计基于双视角夹角及视物距离的视点优选机制。通过公开数据集的大规模实验确定最优视角间距阈值及视物距离比，建立视点选择参数优化模型，显著提升合成视角的视觉保真度。

在 8i 和 Microsoft upper body 公开数据集上的实验表明：相比传统 MV-HEVC 流媒体方案，本方案展现出显著优势，其带宽消耗降低 44.6%～73.2%，峰值信噪比（PSNR）提升 1.03%～3.72%，结构相似性（SSIM）提升 6.59%～15.66%。同时系统支持实时编码解码，帧率可达 30 FPS 以上，满足流媒体实时性要求。

本文的主要贡献如下：

1）研究了基于稀疏视点的多视点视频流传输问题，即如何通过编码部分视点，在接收端进行三维重建并合成剩余视点。

2）提出了三维高斯泼溅辅助的多视点视频紧凑表示，实现了基于稀疏视点的多视点视频流传输，显著降低了带宽消耗；提出视点选择方法，最大化合成视点视觉质量。

3）在公开数据集上进行了实验，结果表明，GaussianViews 相比基线方法具有更低的码率、更高的视觉质量，并且可以实时运行。

2 相关工作

传统多视点视频编解码器（如 MVC 标准[6]、MV-HEVC[8]及 VVC-ML[9]）通过扩展 H.264，H.265 和 H.266 框架，引入视间相关性建模，以消除多视角间的信息冗余。为支持六自由度沉浸式媒体传输，MPEG 组织提出的沉浸式视频编码标准 MIV[10-11]在多视点视频压缩方面取得重要进展。然而，上述方法依赖人工设计模块，难以充分挖掘跨视点信息的潜在关联。

基于深度学习的多视点图像编码研究主要聚焦于两类范式：针对小基线立体图像对的编码方法[12-13,15,18-20]，以及利用解码端统计相关性的分布式图像编码框架[14-22]。现有方法通过显式视点间像素坐标匹配或基于注意力机制的隐式关联建模来捕捉视间相关性，但其建模维度仍局限于平面层级（Plane-level），未能有效表征视点与目标物间的空间层级（Spatial-level）关联特性。

三维高斯泼溅（3DGS[1,23]）技术创新性地提出基于点的可微分渲染技术，将三维空间点表征为包含均值、方差、不透明度及色彩特征的高斯函数，通过投影变换生成目标视角图像。该技术的可微特性支持通过反向传播更新三维高斯属性，确保其几何与纹理属性与原始三维场景高度一致。值得注意的是，3DGS 已被应用到人物重建任务中，取得了良好的效果[24-25]。这一特性启发了本研究利用 3DGS 获取原始人物场景几何先验，进而辅助解决多视点视频压缩任务。

3 方法

本文方法系统架构如图 2 所示。首先设计了一种三维高斯溅射辅助的多视点视频编码算法，该算法通过传输指定稀疏视点并在客户端完成三维重建，进而实时合成剩余视点（见 3.1 节）。具体实现流程包含以下关键步骤。1）稀疏视点编解码：采用 MV-HEVC 标准对选定视点进行编解码处理。2）特征与深度联合提取：从解码视点中提取多尺度图像特征，并通过立体匹配算法估计深度图。3）三维高斯参数建模：融合图像特征与深度信息，构建像素级三维高斯参数估计模型。4）多视点实时合成：基于三维高斯泼溅（3DGS）技术，实现剩余视点的快速渲染。

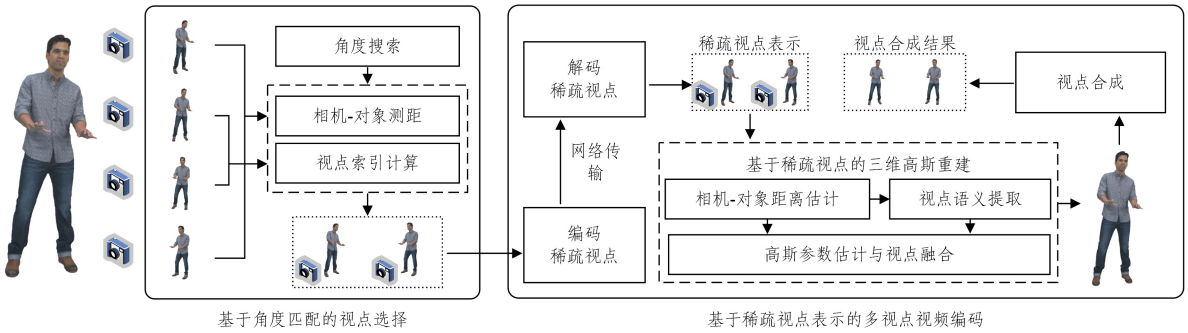


图 2 方法架构图

Fig. 2 Schematic diagram of the overall system

为达成合成视角视觉质量最优化目标，本文设计视点优选方法（见 3.2 节），具体实施包含 3 个技术环节。1）目标物-相机空间距离建模：构建基于深度感知的目标物表面特征点三维坐标估计模型，实现目标物与相机阵列间空间距离的精准度量。2）对称式视点对选择：建立以中心视点为基准的对称式相机对选择准则，确保所选视点具备最优视场覆盖特性。3）最佳夹角参数优化：通过大规模数据集测量确定相机对-目标物最佳夹角参数，建立角度选择与视觉质量的量化关系模型。

3.1 基于稀疏视点的多视点视频紧凑表示

给定 N 个输入视点 $\{V_i\}_{i=1}^{N}$，以及它们固定的相机参数 $\{P_i\}_{i=1}^{N}$，其中所有相机具有相同的焦距和成像分辨率。输入视点的数量 N 取决于接收端显示设备支持的视点数量，通常为 30 到 50。需要编码和传输的输入视点子集表示为 $\{V_s\}_{s\in S}$，其中 S 表示编码视点的索引。该子集包含两个视点，以满足深度估计的要求的同时最小化视点数量。$\{V_s\}_{s\in S}$ 中的视点使用 MV-HEVC 进行编码，并通过互联网传输给接收端。接收端使用 MV-HEVC 解码比特流以获得选定的视点。其余视点利用选定视点实时解码，如下所示。

1）视点语义提取。首先通过一个共享的二维 UNet 特征提取器（由多个残差块和下采样层构成），从选定的视点图像集合 $\{\boldsymbol{I}_s\}_{s\in S}$ 中提取多尺度图像特征。提取的多尺度特征图表示为 $\{\boldsymbol{F}_{s,k}\in\mathbb{R}^{H/2^k\times W/2^k\times C_k}\}_{k=1}^{K}$，其分辨率由高到低逐级递减，其中 K 表示特征提取器的层级总数。最后一个阶段的特征图 $\boldsymbol{F}_{s,k}$ 将被用于在深度估计模块中构建三维相关体。

2）相机-对象距离估计。为了建立 2D 图像平面和 3D 高斯表示之间的联系，需要估计选定视点的深度图。本文设计了一个类似于 RAFT-stereo 的立体视差估计模块，来估计所有选定视点 $\{\boldsymbol{I}_s\}_{s\in S}$ 的深度，以实现更高的深度估计精度和计算效率，考虑到立体视差估计只需要两个光轴平行且成像平面共面的相机，这里 $S=\{l,r\}$。此外，通过以下公式从视点对的视差估计中获得深度估计：

$$\boldsymbol{D}(u,v)=\frac{\boldsymbol{B}\cdot f}{\boldsymbol{d}(u,v)},(u,v)\in[0,H-1]\times[0,W-1]$$

其中，(u,v) 是像素坐标，$\boldsymbol{D}$ 是深度图，$\boldsymbol{B}$ 是两个立体相机之间的基线，f 是焦距，$\cdot$ 是实数乘法，$\boldsymbol{d}$ 是视差图。这意味着在矫正后的立体图像中，一个视点对中的像素 (u,v) 对应于另一个视点中的坐标 $(u+\boldsymbol{d}(u,v),v)$。

为了量化选定的视点对的视觉相似性，计算特征图之间的点积，由于相关性计算仅限于具有相同 y 坐标的像素，而不是像在 RAFT 中那样计算所有像素对的视觉相似性来构建 4D 相关体，因此该点积被称为 3D 相关体。

然后，通过多级 GRU 使用迭代更新机制，在 3D 相关体中通过查找来预测一系列视差估计 $\{\boldsymbol{d}_{l,t}\}_{t=1}^{T}$ 和 $\{\boldsymbol{d}_{r,t}\}_{t=1}^{T}$。关于视差序列更新的更多细节，请参考 GPS-Gaussian[26]。通过迭代得到的最终预测视差场 $\{\boldsymbol{d}_{l,T},\boldsymbol{d}_{r,T}\}$ 的分辨率是输入视点分辨率的 $1/2^K$，通过类似于 RAFT 中的凸上采样方法，将其提升到全图像分辨率。

3）高斯参数估计与视点融合。3D 空间中的每个高斯点由属性 $G=\{x,c,r,s,\alpha\}$ 来表征，这些属性表示其 3D 位置、颜色、旋转、缩放和不透明度。为每个选定视点定义一组高斯参数映射，如下所示：

$$\mathcal{G}(p)=\{\boldsymbol{M}_x(p),\boldsymbol{M}_c(p),\boldsymbol{M}_r(p),\boldsymbol{M}_s(p),\boldsymbol{M}_\alpha(p)\}$$

其中，p 是相机坐标系中图像平面上的一个像素；$\boldsymbol{M}_x\in\mathbb{R}^{H\times W\times 3}$，$\boldsymbol{M}_c\in\mathbb{R}^{H\times W\times 3}$，$\boldsymbol{M}_r\in\mathbb{R}^{H\times W\times 4}$，$\boldsymbol{M}_s\in\mathbb{R}^{H\times W\times 3}$，$\boldsymbol{M}_\alpha\in\mathbb{R}^{H\times W\times 1}$ 分别表示 3D 位置、颜色、旋转、缩放和不透明度的高斯参数图。这些参数图是基于提取的特征图 $\{\boldsymbol{F}_{s,K}\}_{s\in S}$ 和深度图 $\{\boldsymbol{D}_s\}_{s\in S}$ 推断出来的。关于参数图推断的更多细节，请参考 GPS-Gaussian[26]。最后，融合从两个源视点 $\{\boldsymbol{I}_l,\boldsymbol{I}_r\}$ 构建的高斯图 $\{\mathcal{G}_r,\mathcal{G}_l\}$，并通过高斯溅射渲染其余视点。

本文使用两阶段训练策略来训练特征提取、深度估计和高斯参数回归模块。在第一阶段，通过最小化视差预测序列与真实视差之间的 L1 距离来优化特征提取和深度估计模块，权重呈指数增长，如 RAFT-stereo 中那样。

$$\mathcal{L}_{\text{depth}}=\sum_{t=1}^{T}\gamma^{T-t}\parallel\boldsymbol{d}_t-\boldsymbol{d}_{\text{gt}}\parallel_1$$

其中，γ 是权重参数，实验中设置为 0.9；$\boldsymbol{d}_{\text{gt}}$ 是视差的真值。在第二阶段，通过额外使用渲染图像与真实图像之间的 L1 距离和 SSIM 损失，来联合优化特征提取、深度估计和高斯参数回归模块，如 GPS-Gaussian[26] 中那样。

$$\mathcal{L}_{\text{img}}=\lambda_1\mathcal{L}_1+\lambda_2\mathcal{L}_{\text{ssim}}$$

第二阶段的损失函数是深度损失 $\mathcal{L}_{\text{depth}}$ 和渲染图像损失 $\mathcal{L}_{\text{img}}$ 的总和。

3.2 基于角度匹配的视点选择

如 3.1 节所述，3D 高斯溅射辅助多视点视频编码算法需要视点对作为深度估计的输入，因此，选择视点对以最小化计算和传输开销。

选择一对关于相机阵列中心轴对称的视点，直观地说，选择距离较远的一对视点可以覆盖更多的空间信息，这可以减少多视点视频解码中的信息损失。然而，视点之间的距离越大，用于高斯重建的视点输入就越稀疏，视点间相关性的降低可能会导致 3D 重建中出现意外错误。因此，需要在空间信息和视点间相关性之间取得平衡，选择在 3D 高斯溅射辅助解码后能产生最佳视觉质量的视点对。

本文在两个公共数据集上进行了测量，并分析所选视点对之间的距离，以及所有视点解码结果的平均 PSNR 和 SSIM 之间的关系。将视点总数设置为 20，相机与拍摄对象的距离分别为 1.5m，2m 和 3m，同时保持相机阵列固定。本文发现：1）峰值出现在特定的视点对，这验证了某些视点对可以实现合成视点的最佳视觉质量；2）不同相机-对象距离下的最佳视点对不同，这表明最佳视点对与距离有关；3）最佳视点对中两个相机光心与拍摄对象连线之间的夹角都比较接近，这一观察结果可以指导在不同距离下选择最佳视点对。

为此，本文提出一种视点选择方法：1）深度估计，在编码端进行深度估计以获得前景对象到相机的平均距离；2）角度匹配，计算对称位置的每对相机光心与拍摄的前景对象连线之间的夹角，选择夹角最接近 21°的视点对作为结果。基于经验测量，这对视点能最大化合成的新视点的视觉质量。

4 实验

4.1 实验设置

1）数据集。通过从两个公开数据集中渲染 4 个高质量的体视频来构建一个多视角视频数据集，即来自 8i 数据集[27]的 Long dress（Long）视频和 Red and black（Red）视频，以及来自微软上半身数据集[28]的 Ricardo 视频和 Andrew 视频。这些都是空间分辨率为 1024×1024×1024 的高密度动态点云视频，时长在 7～10 秒。从排列在 60 度圆弧上的 30 个虚拟视点渲染这些视频，圆弧半径为 2 米，分辨率为 1024×1024。

2）基线方法。本文构建了一种多视点视频流传输方法作为基线方法，名为 MV-Streaming。使用 MV-HEVC 编解码器（HEVC 的扩展），它通过消除视角间的冗余编码多视角视频。具体来说，使用 x265 作为编码器，FFmpeg 作为解码器，它们是 MV-HEVC 的高效实现方式。

3）评估指标。本文使用比特率来衡量带宽消耗，使用 RGB 通道的平均峰值信噪比（PSNR）和 RGB 通道的平均结构相似性指数（SSIM）[29]来衡量视觉质量，使用帧率来衡量时间效率。

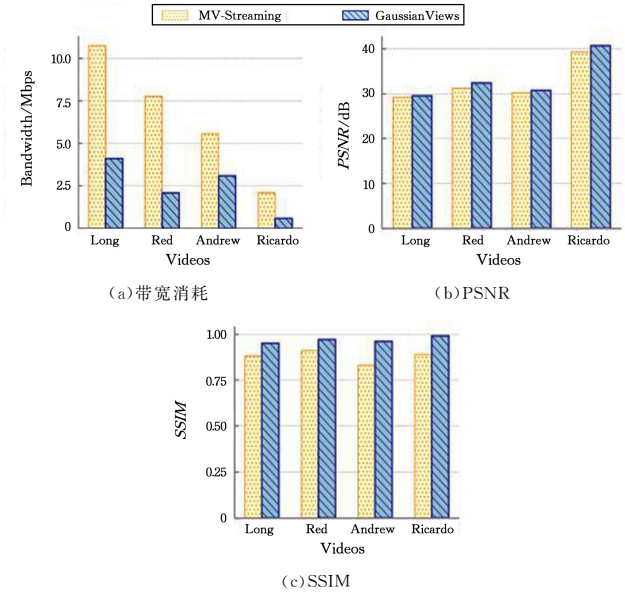
4）实验环境。实验在一个由发送端和接收端组成的系统上进行，发送端和接收端均配备英特尔酷睿 i7-13700K CPU、64GB 内存和 NVIDIA RTX 4090 GPU。发送端和接收端通过 1 Gbps 的以太网连接。使用 x265 对选定的视角进行压缩，并使用 FFmpeg 对其进行解码。使用 AdamW 优化器[30]，以 2×10^{-3} 的初始学习率训练 3D 高斯 splatting 辅助解码模型。损失函数权重超参数设置为 $\gamma=0.9$，$\lambda_1=0.8$，$\lambda_2=0.2$。两个阶段分别训练 4 万次和 10 万次迭代。使用 TensorRT 来加速神经网络推理。

4.2 整体评估

为了了解 GaussianViews 的带宽消耗和视觉质量，本文在所有视频上对 GaussianViews 和 MV-Streaming 进行了评估。

1）带宽消耗。如图 3（a）所示，与 MV-Streaming 相比，GaussianViews 在 Long，Red，Andrew 和 Ricardo 视频上的带宽消耗分别显著降低了 61.82%，73.2%，44.6%和 72.38%。有了如此显著的提升，GaussianViews 在这 4 个视频上仅需 0.58 ～4.10 Mbps 的带宽，这在广泛的网络条件下都能得到支持。这些结果验证了 GaussianViews 通过去除冗余视角，有效降低了比特率。

2）视觉质量。如图 3（b）和图 3（c）所示，与 MV-HEVC 相比，GaussianViews 在 4 个视频上分别使 PSNR 提高了 1.03%，3.72%，1.69%和 3.41%，使 SSIM 提高了 7.95%，6.59%、15.66%和 11.24%。尽管 GaussianViews 降低了带宽消耗，但它仍然比 MV-Streaming 具有更好的视觉质量。因为 GaussianViews 只编码两个视角，它使用较高的目标比特率来减少失真，并且所提出的三维高斯溅射辅助解码器也能合成高质量的视角。



（a）带宽消耗 （b）PSNR

（c）SSIM

图 3 GaussianViews 的整体评估

Fig. 3 Overall evaluation of GaussianViews

4.3 消融实验

本文将 GaussianViews 与其两个变体“基于深度图像的渲染（DIBR）”和“随机选择”进行比较，分别验证了 3.1 节和 3.2 节中设计的有效性。原始的 GaussianViews 利用高斯 splatting 进行 3D 重建和视角合成，而 DIBR 则根据估计的深度图将选定的视角变换到其余视角。原始的 GaussianViews 选择提供最佳视觉质量的两个视角，而 RandomSelection 则随机选择两个视角。本文省略了带宽消耗指标，因为 GaussianViews 和这些变体都对两个视角进行编码，并且在实验中它们的带宽消耗表现相似。

图 4 展示了视觉质量的结果，其中 GaussianViews 显著提高了视觉质量。具体来说，与 DIBR 和 RandomSelection 相比，GaussianViews 使 PSNR 平均分别提高了 58.05%和

10.78%，使 SSIM 平均分别提高了 9.68%和 2.41%。结果验证了高斯 splatting 提高了合成视角的视觉质量。

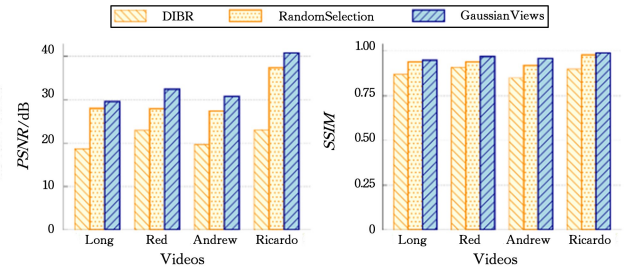


图 4 消融实验结果

Fig. 4 Results of ablation study

4.4 关键参数评估

为了了解所提出方法的可扩展性，评估了视角的数量。根据经验，尝试了 10 到 40 个视角的情况，发现：带宽消耗与不同数量的视角基本一致。这是因为无论视点密度高或低，本文都只选择两个视角，这意味着 GaussianViews 可以轻松扩展到具有更多视角的系统中。另一方面，所提出方法在不同数量的视角下也保持了良好的视觉质量，这得益于场景的 3D 高斯表示。

4.5 时间效率

本文测量并展示了 MV-Streaming 和 GaussianViews 的编解码时间。在 4.1 节所述的实验环境中，测试编解码计算所需的时间。记录采集的多视点图像将要输入编码器的时间 t_1，记录输出编码结果的时间 t_2，使用 t_2-t_1 作为编码时间；记录码流将要输入解码器的时间 t_3，解码器完成所有视点合成的时间 t_4，使用 t_4-t_3 作为解码时间。测试时间可能会受到操作系统其他计算任务的干扰，为了提高计算的可靠性和准确性，本文提供了在测试数据集上的每帧编码时间和解码时间的均值。

1)编码时间。在有 30 个视角的情况下，MV-Streaming 的编码需要 244.5 毫秒，而 GaussianViews 需要 32.65 毫秒。结果验证了 GaussianViews 显著提高了编码速度，达到了 7.5 倍，这支持其应用于超过 30 帧每秒(如直播)的实时应用场景。

2)解码时间。测量了 GaussianViews 和 MV-Streaming 的解码时间。对于分别有 10 个、20 个和 30 个视角的情况，MV-Streaming 分别需要 17.36 毫秒、33.28 毫秒和 50.43 毫秒；而 GaussianViews 分别需要 23.84 毫秒、31.42 毫秒和 38.99 毫秒。具体来说，使用 TensorRT 优化的神经网络大约需要 16 毫秒，并且每个视角的 3DGS 光栅化时间不到 1 毫秒。结果表明，GaussianViews 在 20 个视角以下可以实现实时执行。

结束语 本文介绍了一种用于多视角视频流的新颖框架 GaussianViews，它利用高斯 splatting 来实现低比特率和实时性能。本文提出了一种 3D 高斯 splatting 辅助的多视角视频编码算法，该算法直接从 2D 图像回归 3D 高斯参数，无需耗时的微调，实现了实时 3D 重建；还提出了一种视点选择算法，通过考虑角度关系和距离来优化稀疏视角的选择，从而最大化合成视角的质量。实验结果表明，GaussianViews 降低了 44.6%～73.2%的带宽消耗，同时提高了视觉质量，并且它可以以超过 30 帧每秒的速度运行，证明了其在直播应用中的实用性。

参 考 文 献

[1] ANTHES C, GARCÍA-HERNÁNDEZ R J, WIEDEMANN M, et al. State of the Art of Virtual Reality Technology[C]// Proceedings of IEEE Aerospace Conference. 2016: 1-19.

[2] SCHMALSTIEG D, HÖLLERER T. Augmented Reality: Principles and Practice[C]// Proceedings of IEEE Virtual Reality. 2017: 425-426.

[3] DAI A, CHANG A X, SAVVA M, et al. Scannet: Richly-annotated 3D Reconstructions of Indoor Scenes[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2432-2443.

[4] CHEN X, MA H, WAN J, et al. Multi-view 3D Object Detection Network for Autonomous Driving[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6526-6534.

[5] HOSSEINIAN S, AREFI H. 3D Reconstruction from Multi-view Medical X-ray Images Review and Evaluation of Existing Methods[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015, XL-1/W5: 319-326.

[6] YU X, SANG X, CHEN D, et al. AutostereoscopicThree-dimensional Display with High Dense Views and the Narrow Structure Pitch[J]. Chinese Optics Letters, 2014, 12(6): 060008.

[7] VETRO A, WIEGAND T, SULLIVAN G J. Overview of the Stereo and Multiview Video Coding Extensions of the H. 264/MPEG-4 AVC Standard[J]. Proceedings of IEEE, 2011, 99(4): 626-642.

[8] HANNUKSELA M M, YAN Y, HUANG X, et al. Overview of the Multiview High Efficiency Video Coding (MV-HEVC) Standard[C]// Proceedings of IEEE International Conference on Image Processing(ICIP). 2015: 2154-2158.

[9] WANG Y K, SKUPIN R, HANNUKSELA M M, et al. The High-level Syntax of the Versatile Video Coding(VVC) Standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3779-3800.

[10] BOYCE J M, DORÉ R, DZIEMBOWSKI A, et al. MPEG Immersive Video Coding Standard[J]. Proceedings of the IEEE, 2021, 109(9): 1521-1536.

[11] VADAKITAL V K M, DZIEMBOWSKI A, LAFRUIT G, et al. TheMPEG Immersive Video Standard—Current Status and Future Outlook[J]. IEEE Multimedia, 2022, 29(3): 101-111.

[12] DENG X, YANG W, YANG R, et al. DeepHomography for Efficient Stereo Image compression [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2021: 1492-1501.

[13] LEI J, LIU X, PENG B, et al. DeepStereo Image Compression via Bi-directional coding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 19669-19678.

[14] ZHANG X, SHAO J, ZHANG J. Ldmic: Learning-based Distributed Multi-view Image Coding[C]// Proceedings of International Conference on Learning Representations. 2023.

[15] LIU Z, ZHANG X, SHAO J, et al. Bidirectional Stereo Image Compression with Cross-dimensional Entropy Model[C]// Proceedings of European Conference on Computer Vision. 2024.

[16] HUANG Y, CHEN B, LIAN N, et al. 3D-GP-LMVIC: Learning-based Multi-view Image Coding with 3D Gaussian Geometric Priors[J]. arXiv: 2409.04013, 2024.

[17] KERBL B, KOPANAS G, LEIMKUEHLER T, et al. 3D Gaussian Splatting for Real-time Radiance Field Rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 1-14.

[18] WÖDLINGER M, KOTERA J, XU J, et al. Sasic: Stereo Image Compression with Latent Shifts and Stereo Attention[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 661-670.

[19] ZHAI Y, TANG L, MA Y, et al. Disparity-based Stereo Image Compression with Aligned Cross-view Priors[C]// Proceedings of ACM International Conference on Multimedia. 2022: 2351-2360.

[20] DENG X, DENG Y, YANG R, et al. Masic: DeepMask Stereo Image Compression[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(10): 6062-6040.

[21] AYZIK S, AVIDAN S. Deep Image Compression Using Decoder Side Information[C]// Proceedings European Conference on Computer Vision. 2020: 699-714.

[22] HUANG Y, CHEN B, QIN S, et al. Learned Distributed Image Compression with Multi-scale Patch Matching in Feature Domain[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 4322-4329.

[23] HAMDI A, MELAS-KYRIAZI L, MAI J, et al. Ges: Generalized Exponential Splatting for Efficient Radiance Field Rendering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2024: 19812-19822.

[24] JIANG Y H, SHEN Z H, HONG Y, et al. Robust Dual Gaussian Splatting for Immersive Human-centric Volumetric Videos[J]. ACM Transactions on Graphics, 2024. 43(6): 1-15.

[25] LIU W, BAO Q, SUN Y, et al. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective [J]. ACM Computing Surveys, 2023, 55(4): 1-41.

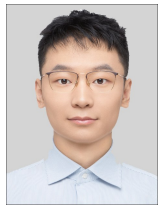
[26] ZHENG S, ZHOU B, SHAO R, et al. GPS-Gaussian: Generalizable Pixel-wise 3d Gaussian Splatting for Real-time Human Novel View Synthesis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2024: 19680-19690.

[27] D'EON E, HARRISON B, MYERS T, et al. 8iVoxelized Full Bodies-a Voxelized Point Cloud Dataset[EB/OL]. http://plenodb.jpeg.org/pc/8ilabs/.

[28] LOOP C, CAI Q, ESCOLANO S O, et al. Microsoft Voxelized Upper Bodies-a Voxelized Point Cloud Dataset[EB/OL]. http://plenodb.jpeg.org/pc/microsoft/.

[29] WANG Z, BOVIK A, SHEIKH H, et al. ImageQuality Assessment: from Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[30] LOSHCHILOV I, HUTTER F. DecoupledWeight Decay Regularization[C]// Proceedigns of International Conference on Learning Representations. 2019.



WANG Yizong, born in 1997, Ph.D, is a member of CCF (No. B7846M). His main research interest is immersive video streaming.

MA Siwei, born in 1979, professor, Ph.D supervisor. His main research interest is video coding.

(责任编辑:李亚辉)