

基于深度先验引导的神经动态场景渲染重建

李璋林, 孙益辉, 李煜, 姚献艺, 兰世儒

(湖南三一工业职业技术学院 数智工程学院, 湖南 长沙 410129)

摘要: 目前要实现实时且高质量的动态场景渲染是一项具有挑战性的任务。为此, 提出一种基于深度先验引导的四维时空高斯场景表示方法。首先, 通过在三维高斯中引入时间相关的不透明度以及可参数化的运动与旋转, 使得模型能够同时刻画静态结构与瞬态动态内容。其次, 利用深度信息对渲染过程进行约束与指导, 从而进一步提升细节恢复能力与几何一致性。最后, 结合训练误差与粗略深度的联合引导策略, 实现对场景中稀疏区域的有效补充与建模。相比现有的其他先进方法, 该方法在渲染质量、渲染速度以及模型紧凑性方面均具有明显优势。在高分辨率场景下, 该方法能够在常规图形处理平台 NVIDIA RTX 4090 GPU 上实现每秒 160 帧的实时渲染。

关键词: 动态场景渲染; 3DGS; 深度引导; 场景重建

DOI: 10.11907/rjdk.241796

扫描二维码阅读全文: 

中图分类号: TP391.4

文献标识码: A

文章编号: 1672-7800(2026)001-0184-07

Neural Dynamic Scene Rendering and Reconstruction Based on Deep Prior Guidance

LI Zhanglin, SUN Yihui, LI Yu, YAO Xianyi, LAN Shiru

(School of Digital Intelligence Engineering, Hunan SANY Polytechnic College, Changsha 410129, China)

Abstract: Achieving real-time and high-quality rendering of dynamic scenes remains a highly challenging task. To address this problem, this paper proposes a four-dimensional spatiotemporal Gaussian scene representation guided by depth priors. First, temporal opacity together with parameterized motion and rotation is introduced into three-dimensional Gaussians, enabling the model to capture both static structures and transient dynamic content. Second, depth information is used to constrain and guide the rendering process, thereby improving detail recovery and geometric consistency. Third, a joint guidance strategy combining training error and coarse depth is adopted to effectively supplement and model sparsely covered regions in the scene. Compared with other advanced methods, the proposed approach exhibits clear advantages in rendering quality, rendering speed, and model compactness. Under high-resolution settings, the method is capable of achieving real-time rendering at one hundred and sixty frames per second on a standard NVIDIA RTX 4090 graphics platform.

Key Words: dynamic scene rendering; 3DGS; deep guidance; scene reconstruction

0 引言

从图像重建三维场景一直是计算机视觉领域的重要研究方向。近年来, 新视角合成技术迅速发展, 尤其是神经辐射场 (Neural Radiance Fields, NeRF) 在多视角场景建模方面取得了显著突破^[1]。围绕提升渲染质量与加快渲

染速度的目标, 研究者开展了大量工作, 从不同角度对场景表示与渲染流程进行了改进^[2-7]。文献[8]提出使用各向异性的径向基函数来构建三维场景表示, 不仅能更好地适应场景的结构特性, 而且显著提升了细节建模能力。在可微分体渲染方向上, 近期提出的三维高斯喷溅 (3D Gaussians Splatting, 3DGS) 方法采用光栅化渲染范式, 将三维高斯直接投影至图像平面, 实现了更高的渲染效率,

收稿日期: 2024-09-13

基金项目: 2022年湖南省教育信息技术研究省级重点课题 (HNETR22042)

作者简介: 李璋林 (1981-), 男, 湖南三一工业职业技术学院数智工程学院讲师, 研究方向为智能无人系统、计算机视觉; 孙益辉 (1975-), 男, 湖南三一工业职业技术学院数智工程学院副教授, 研究方向为计算机视觉与图像处理; 李煜 (1995-), 男, 湖南三一工业职业技术学院数智工程学院讲师, 研究方向为计算机视觉; 姚献艺 (1995-), 男, 湖南三一工业职业技术学院数智工程学院讲师, 研究方向为计算机视觉; 兰世儒 (1990-), 男, 湖南三一工业职业技术学院数智工程学院讲师, 研究方向为计算机视觉。本文通讯作者: 兰世儒。

并进一步提升了新视角图像的重建质量^[9]。尽管以上方法在静态场景取得了成功,但是由于模型大小和训练时间的开销,将其应用于动态场景仍然具有挑战性。动态视角合成方法将多个帧在单个模型中进行表示,例如 Hyper-Reel^[10]和 NeRFPlayer^[11]利用自然视频中相邻帧通常表现出高度相似性的特性,将静态空间表示时间特征与共享插值相结合,以提高模型的紧凑性。HexPlane^[12]和 K-Planes^[13]将 4D 时空域分解为多个 2D 平面,以实现紧凑地表示场景。类似地, MixVoxels 使用时变函数,并通过内积将其与空间特征连接起来^[14]。但这些方法的一个局限性是其网格状表示不能完全适应动态场景下的几何结构,从而阻碍了场景精细细节的建模。与此同时,以上方法致力于在不牺牲质量的情况下实时进行高分辨率的渲染。

本文方法可以在实现高质量渲染效果的同时保持较小的模型规模。本文在三维高斯的基础上将其扩展至四维时空表示,并为其引入时间相关的不透明度以及多项式形式的运动与旋转参数化,从而能够同时刻画场景中的静态结构、持续动态变化以及瞬态内容。同时,为了增强模型的紧凑性并考虑到时空外观,本文提出拼接式的特征聚合网络结构。具体而言,对于每个时空高斯,不用球形谐波系数,而是存储编码基本颜色。此外,还存储与视图相关的信息和计时信息,并将这些函数进行栅格化以得到高维空间,之后通过微小的多层感知机网络来生成最终颜色。尽管这些特征的规模小于球谐系数,但依然具有较强的表达能力。在此基础上,本文进一步引入基于深度先验的高斯引导采样策略,以提升复杂场景下的渲染质量。此外还可以观察到,对于高斯分布较为稀疏的远距离区域,直接初始化往往会导致渲染质量下降。为解决这一问题,本文在训练误差与粗略深度的联合引导下,对三维动态场景中高斯不足的区域进行增量采样,从而实现更完整、稳定的场景建模。

本文主要贡献如下:

- (1) 提出一种引入时间变量的时空高斯核表示方法,用于高效建模动态场景,并提升新视角合成效果。
- (2) 构建一种新的喷射式高斯特征渲染机制,增强了特征表达能力,有助于捕获时空变化信息。
- (3) 引入基于深度先验的高斯引导采样策略,可有效改善远距离稀疏区域的渲染质量。

1 渲染方法

NeRF 是近年来用于新视角合成的重要技术之一,其本质上是一种连续辐射场函数 f_θ ,用于将三维空间点 $x \in \mathbb{R}^3$ 及其视角方向 d 映射为对应的体密度 σ 和颜色 c 。具体公式如下:

$$(c, \sigma) = f_\theta(x, d) \quad (1)$$

之后,进一步采用体积渲染的方法将体密度 σ 和颜色

c 累加为射线 r 的预测颜色 $C(r)$ 。具体公式如下:

$$\hat{C}(r) = \sum_{i=1}^N T_i \cdot \alpha_i \cdot c_i \quad (2)$$

其中, $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$, $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$, N 是采样点数量, δ_i 是邻近采样点两点之间的距离 $\delta_j = \|x_{i+1} - x_j\|_2$ 。

最后,以渲染损失为优化目标,将射线颜色合并为图像。

与 NeRF 类方法不同,三维高斯喷射技术(3DGS)通过可微光栅化对一组各向异性的三维高斯进行优化,属于一种显式的点云式场景表示方法,可用于描述静态三维场景。由于该方法的光栅化过程十分高效,优化后的模型能够实时生成具有高保真度的新视角图像。3DGS 的基本元素包括三维高斯的位置 h_i 、协方差矩阵 h_i 、不透明度 h_i 以及球面谐波系数(Spherical Harmonic Coefficient, SHC) h_i 等。最终,三维高斯球在任意空间点 x 的不透明度公式如下:

$$\alpha_i = \sigma_i \cdot \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i)\right) \quad (3)$$

其中, Σ_i 为半正定矩阵,可以分解为缩放矩阵 S_i 和旋转矩阵 R_i 。具体公式如下:

$$\Sigma_i = R_i S_i S_i^T R_i^T \quad (4)$$

其中, S_i 是对角缩放矩阵,其 3 个对角元素由三维向量 s_i 给定; R_i 为旋转矩阵,由四元数 q_i 计算得到。

渲染图像之前,首先通过近似透视变换将三维高斯投影到二维图像空间^[15]。具体来说,三维高斯的投影被近似为一个二维高斯,中心为 μ_i^{2D} ,协方差为 Σ_i^{2D} ,使 W 、 K 分别成为视角矩阵和投影矩阵。 μ_i^{2D} 和协方差 Σ_i^{2D} 计算公式如下:

$$\mu_i^{2D} = (K((W\mu_i)/(W\mu_i)_z))_{1:2} \quad (5)$$

$$\Sigma_i^{2D} = (JW\Sigma_i W^T J^T)_{1:2, 1:2} \quad (6)$$

其中, J 是射影变换的雅可比行列式。将高斯按深度顺序排列后,通过体积渲染获得像素的颜色:

$$I = \sum_{i \in \mathcal{N}} c_i \alpha_i^{2D} \prod_{j=1}^{i-1} (1 - \alpha_j^{2D}) \quad (7)$$

其中, α_i^{2D} 是公式(3)在二维投影空间中的形式,计算中使用 μ_i^{2D} 、 Σ_i^{2D} 、 x^{2D} (像素坐标)分别替代三维空间中的 μ_i 、 Σ_i 、 x ; c_i 是根据视图方向和系数 h_i 对球谐系数评估的 RGB 颜色。

2 本文方法

本文提出一种基于深度信息引导时空高斯的新型场景表示法,用于动态三维场景建模。本文方法将多视角视频作为输入,允许在新视角进行渲染。

2.1 时空高斯

为了表示四维动态场景,本文构建了一种基于时间变

化特性的时空高斯表示方式,将三维高斯与时间维度相结合,从而能够模拟场景中内容的出现、消失以及运动变形。方法的整体框架如图1所示。

具体来说,引入时间径向基函数来编码时间不透明度,可以有效模拟视频持续时间内出现或消失的场景内容。同时,利用三维高斯的位置和旋转的时间条件参数函

数模拟场景中的运动及变形。对于一个时空点 (x, t) ,其不透明度表示如下:

$$(x, t) = \sigma_i \cdot \exp\left(-\frac{1}{2}(x - \mu_i(t))^T \Sigma_i^{-1}(t)(x - \mu_i(t))\right) \quad (8)$$

其中, $\sigma_i(t)$ 是时间透明度, $\mu_i(t), \Sigma_i(t)$ 是与时间相关的位置和协方差, i 是第 i 个高斯核。下文将详细介绍每个组成部分。

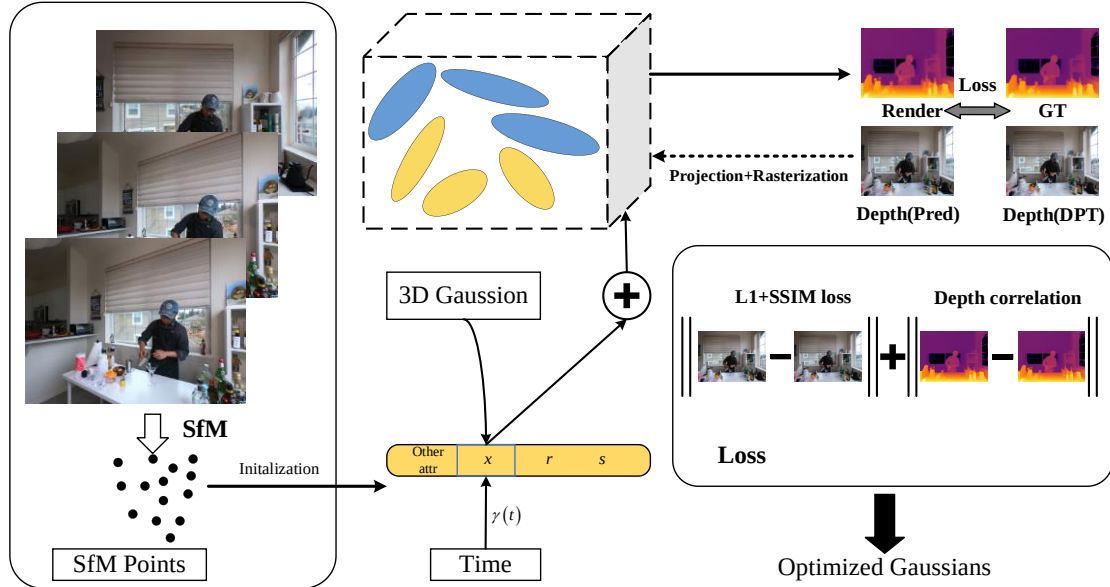


Fig. 1 Network framework of the proposed algorithm

图1 本文算法网络框架

本文采用一维高斯函数来建模高斯核在任意时间 t 的时间不透明度,从而能够有效描述其随时间的变化特性。时间不透明度 $\sigma_i(t)$ 公式如下:

$$\sigma_i(t) = \sigma_i^s \exp(-s_i^T |t - \mu_i^T|^2) \quad (9)$$

其中, μ_i^T 是时空中心, S_i^T 是时间缩放因子, σ_i^s 是时间独立的空间透明度, μ_i^T 是时空高斯最明显的时间戳,而 S_i^T 决定其有效持续时间(即其时间不透明度较高的持续时间)。将 σ_i^s 包括在内,以允许不同时空高斯核的空间不透明度存在差异。

2.2 多项式运动轨迹

对于每个时空高斯核,都采用一个时间条件函数来模拟其运动。受文献[16]、[17]的启发,选择多项式函数如下:

$$\mu_i(t) = \sum_{k=0}^{n_p} b_{i,k}(t - \mu_i^T)^k \quad (10)$$

其中, $\mu_i(t)$ 是时空高斯在某个时间的空间位置, $\{b_{i,k}\}_{k=0}^{n_p}, b_{i,k} \in \mathbb{R}$ 是相应的多项式系数,并在训练过程中进行优化。通过结合公式(9)和公式(10),复杂的长轨迹运动可用多个运动较简单的片段来表示。在实验过程中设 $n_p = 3$,旨在平衡模型的代表能力与参数规模。

2.3 多项式旋转

根据文献[18-20],使用四元数方式对公式(4)中的旋

转矩阵 R_i 进行参数化。与运动轨迹类似,采用多项式函数表示四元数,具体公式如下:

$$q_i(t) = \sum_{k=0}^{n_q} c_{i,k}(t - \mu_i^T)^k \quad (11)$$

式中, $q_i(t)$ 表示时空高斯在 t 时刻的旋转(以四元数形式表示), $\{c_{i,k}\}_{k=0}^{n_q}$ 和 $c_{i,k} \in \mathbb{R}$ 是多项式系数,并在训练过程中进行优化。四元数 $q_i(t)$ 可进一步转换为旋转矩阵 $R_i(t)$,其协方差矩阵 $\Sigma_i(t)$ 可由公式(4)在时间 t 下计算得到。在实验过程中,设置 $n_q = 1$ 。

其中,公式(4)中的缩放矩阵 s_i 与时间无关,因为在实验中没有观察到对该参数应用时间条件函数时渲染质量的改善。

2.4 融合深度先验高斯

本文采用预训练的密集预测网络(Dense Prediction Transformer, DPT)生成单目深度图,该网络基于约140对深度图像进行训练,因而具备较多的深度先验知识。

为了降低真实场景尺度和估计深度之间的尺度模糊性,在估计深度图和渲染深度图之间引入一种相对损失皮尔逊相关性的变量,可测量二维深度图之间的分布差异性,并遵循以下公式:

$$\text{Corr}(\hat{D}_{\text{ras}}, \hat{D}_{\text{est}}) = \frac{\text{Cov}(\hat{D}_{\text{ras}}, \hat{D}_{\text{est}})}{\sqrt{\text{Var}(\hat{D}_{\text{ras}}) \text{Var}(\hat{D}_{\text{est}})}} \quad (12)$$

其中, $Corr$ 表示相关性变量, D_{ras} 、 D_{est} 分别表示渲染光栅化得到的深度图和单目深度估计生成的深度图。设置该软约束的目的在于强调深度的相对结构, 从而降低绝对深度值不一致所带来的干扰, 避免产生异常与模糊现象, 并提升最终的渲染质量。

2.5 可微分深度光栅化

为使深度先验能够反向指导高斯参数的优化, 本文构建了一个可微分的深度光栅化模块, 用于接收渲染深度 D_{ras} 与估计深度 D_{est} 之间的误差信号。在具体实现中, 采用 3DGS 中的 alpha-blend 混合渲染方式进行深度光栅化。对像素产生贡献的高斯按照深度顺序存储于顺序高斯缓冲区 (Sequential Gaussian Buffer, SGB), 并依次累积以生成最终的深度值:

$$d = \sum_{i=1}^n d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (13)$$

其中, d_i 表示第 i 个高斯的 z-buffer, 参数 α 的计算方式与公式 (2) 中的表达一致。完全可微分的实现使深度相关损失成为可能, 进一步提高了渲染深度与估计深度之间的相似性。

2.6 Splatted 特性渲染

为了既准确又紧凑地编码随视角和时间变化的辐射度, 在每个时空高斯中存储特征而非球谐系数。具体来说, 每个时空高斯的特征由 3 部分组成:

$$f_i(t) = [f_i^{base}, f_i^{dir}, (t - \mu_i^t) f_i^{time}]^T \quad (14)$$

其中, $f_i^{base} \in \mathbb{R}^3$ 包含一个基础 RGB 颜色, $f_i^{dir}, f_i^{time} \in \mathbb{R}^3$ 编码与视线方向和时间相关的信息。

特征拼接过程与高斯拼接方法类似, 只是公式 (7) 中的 RGB 颜色项 C_i 被相应的特征映射函数 $f_i(t)$ 所替代。在喷溅到图像空间后, 将每个像素处的喷溅特征分成 F^{base} 、 F^{dir} 、 F^{time} , 其通道与公式 (10) 中的 3 部分相对应。每个像素的最终 RGB 颜色在经过 2 层多层感知机 (Multilayer Perceptron, MLP) 网络 Φ 变换后得到以下公式:

$$I = F^{base} + \Phi(F^{dir}, F^{time}, r) \quad (15)$$

其中, r 是像素的视线方向, 并与输入的特征进行了连接。与传统的球谐函数 (Spherical Harmonics, SH) 编码方式相比, 基于特征的表示在每个时空高斯上所需的参数量更少, 仅需 9 个参数, 而三阶 SH 编码则需要 48 个参数。同时, 由于所采用 MLP 网络结构 Φ 的深度较浅且宽度较窄, 因此整体渲染过程仍能保持较高速度。为进一步提升渲染效率, 在训练和推理阶段还可以省略方向相关特征, 仅保留基础颜色特征进行计算。

2.7 损失函数

本文中损失函数主要包含 3 部分: 颜色重建损失 $\mathcal{L}_{color}$ 和 $\mathcal{L}_{ssim}$, 以及深度正则化先验损失 $\mathcal{L}_{prior}$ 。3DGS 的颜色重建损失由重建误差项 $\mathcal{L}_1$ 和 D-SSIM 项共同构成, 用于度量渲染图像 $\hat{I}$ 与真实图像 I 之间的差异。其渲染公式如下:

$$\mathcal{L}_{color} = \mathcal{L}_1(\hat{I}, I) + \lambda_1 \mathcal{L}_{ssim}(\hat{I}, I) \quad (16)$$

深度正则化先验损失 $\mathcal{L}_{prior}$ 公式如下:

$$\mathcal{L}_{prior} = \|Corr(D_{ras}, D_{est})\|_1 \quad (17)$$

通过可微分拼接和基于梯度的反向传播来优化以上参数, 并使用高斯密度控制进行交错。使用渲染损失来比较渲染图像和地面实况图像。渲染损失由 $\mathcal{L}_1$ 项和 D-SSIM 项组成。

训练总损失的构成如下所示:

$$\mathcal{L} = \lambda_0 \mathcal{L}_{color} + \lambda_2 \mathcal{L}_{prior} \quad (18)$$

其中, λ_1 , λ_2 , λ_3 分别设置为 0.8、0.1 和 0.05。

2.8 深度先验高斯引导取样

采用传统 3DGS 方法初始化高斯点云时, 对于高斯分布较为稀疏的区域, 渲染过程往往难以收敛到较高质量, 尤其是在这些区域距离训练摄像机较远的情况下。因此, 本文引入一种新策略, 在训练误差和粗深度的指导下采样新的高斯。沿着训练过程中误差较大的像素射线对新高斯进行采样, 为了保证采样效果, 在训练损失稳定后再进行采样。

由于误差图在训练过程中可能产生噪声, 因此逐块聚合训练误差, 以优先考虑具有重大误差的区域, 而不是异常像素。之后, 从每个具有较大误差的选定补丁的中心像素采样光线。为了避免在过大的深度范围内采样, 利用高斯中心的粗略深度图来确定更具体的深度范围, 再在光线的深度范围内对新高斯进行均匀采样, 并向新采样的高斯函数中心添加了小噪声。

在采样的高斯中, 不需要的高斯经过步骤训练后会具有较低的不透明度并被剪枝。对于实验中的场景, 上述采样过程只需要进行不超过 3 次。深度引导采样策略与高斯方法中的密度控制技术相辅相成。密度控制是通过分裂的方式在现有高斯核附近逐渐增加高斯核数量, 而本文提出的策略能够在高斯稀疏甚至完全缺失的区域生成新高斯, 从而补全点云分布, 并对场景进行更均匀的溅射覆盖, 最终实现更高质量的场景渲染。

3 实验结果与分析

3.1 实验环境

本文使用来自所有可用时间戳的运动结构稀疏点云来初始化高斯球。在密度控制方面, 本文使用了比原始 3DGS 方法更有效的剪枝策略, 以减少高斯球数量, 并使模型具有相对较小的体积。使用 Adam 优化器, 50 帧序列的训练时间为 40~60 min^[21]。同时, 对拼接过程进行了调整, 以支持真实世界数据集中的不同相机模型, 并利用预训练的密集预测变换器 (DPT) 模型进行零镜头单目深度估计。所有实验均在 NVIDIA RTX 4090 GPU 上完成。

3.2 数据集设置

(1) Neural 3D Video Datasets 数据集^[12]。在本文实验中, DyNeRF 数据集包含由 18~21 个摄像机以 2 704×2 028 分辨率捕获的 6 个室内多视图视频序列。按照常规设置, 训练与评估均在半分辨率下进行, 即将原始图像的长和宽缩小为 50%, 并使用第一台相机的视角作为测试视图。每个场景的帧数为 300 帧。

(2) Mip-NeRF360 数据集^[3]。由 9 个不同场景组成, 每个场景都有一个复杂的中心物体或区域与详细背景。本文方法在已公开的 7 个场景上进行评估, 使用 24 个训练视图, 并将原始图像分别下采样为 1/4 和 1/8 分辨率进行实验。测试图像选择方式与 LLFF 数据集的设置保持一致^[1]。本文旨在构建一种简单而高效的模型, 以应对复杂大规模场景中少量训练视图背景下新视角合成的挑战。

3.3 实验结果

在定量评价中, 将本文方法与多种代表性方法进行全面的比较, 实验结果如表 1、表 2 所示。其中, 图像结构一致性指标包括: 学习式感知图像块相似度 (Learned Perceptual Image Patch Similarity, LPIPS), 数值越小表明感知质量越好; 结构相似性指数 (Structural Similarity Index Measure, SSIM), 数值越大表明结构保持越好。

为便于更直观地评价各方法的表现, 图 2、图 3 展示了相应的定性比较结果。

Table 1 Comparison results of different methods on the DyNeRF dataset

表 1 不同方法在 DyNeRF 数据集上的比较结果

Method	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	FPS ↑	Size/MB ↓
Hexplane ^[11]	31.71	—	0.075	47.600	200
DyNeRF ^[12]	29.58	0.020	0.083	0.015	28
HyperReel ^[14]	31.30	0.036	0.096	2.000	360
K-Planes ^[15]	31.63	0.018	—	0.300	311
MixVoxels ^[16]	31.34	0.017	0.096	37.700	500
4DGS ^[24]	38.30	0.098	0.020	82.000	18
SpaceTimeGS ^[25]	34.05	0.026	0.044	160.000	200
Ours	38.72	0.078	0.020	160.000	110

注: Ours 代表本文方法, ↑ 代表对应指标数值越大性能越好, ↓ 代表对应指标数值越小性能越好。

Table 2 Comparison results of different methods on the Mip-NeRF360 dataset

表 2 不同方法在 Mip-NeRF360 数据集上的比较结果

Method	PSNR(dB) ↑	SSIM ↑	LPIPS ↓	FPS ↑	Size/MB ↓
Hexplane ^[11]	31.71	—	0.075	47.600	200
DyNeRF ^[12]	29.58	0.020	0.083	0.015	28
HyperReel ^[14]	31.30	0.036	0.096	2.000	360
K-Planes ^[15]	31.63	0.018	—	0.300	311
MixVoxels ^[16]	31.34	0.017	0.096	37.700	500
4DGS ^[24]	38.30	0.098	0.020	82.000	18
SpaceTimeGS ^[25]	34.05	0.026	0.044	140.000	200
Ours	38.72	0.078	0.020	160.000	110

注: Ours 代表本文方法, ↑ 代表对应指标数值越大性能越好, ↓ 代表对应指标数值越小性能越好。

如图 2 和表 1 所示, 在 DyNeRF 数据集上的实验结果表明, 本文方法在渲染质量方面具有最优表现, 并且在速度与模型紧凑性上也具备明显优势。具体而言, 本文模型在 PSNR、SSIM 和 LPIPS 等指标上均获得最佳结果, 同时在视觉质量上显著优于动态 3DGS 和 4DGS。在运行效率方面, 本文模型与 4DGS 均可在 DyNeRF 数据集上实现超过 160 FPS 的渲染速度。但与 4DGS 相比, 本文模型的大小仅为约 6%, 并在 PSNR 上进一步提升了 0.42 dB。此外, 动态 4DGS 的结果中存在明显的时间伪影, 而本文方法由于采用序列级训练方式, 能更好地保持帧间一致性, 从而避免了此类问题。

如图 3 和表 2 所示, 在 Mip-NeRF360 数据集上的实验结果表明, 本文方法在渲染速度和多项定量指标上均优于其他方法。与基于 NeRF 的多种基线方法相比, 本文方法的峰值信噪比 PSNR 提升了 4.67 dB, 渲染速度 FPS 由 140 提升至 160 FPS。图 3 的可视化结果进一步表明, 本文方法在结构细节保留、视觉一致性和整体图像质量方面均具有明显优势。在不同场景下的图像对比结果显示, 4DGS 和 HexPlane 在细节区域普遍存在模糊、伪影和信息缺失的问题, 而本文方法能够持续生成更清晰、真实且细节丰富的图像, 充分验证了其在 Mip-NeRF360 数据集上的有效性。



Fig. 2 Visualization results of different methods on the DyNeRF dataset

图 2 不同方法在 DyNeRF 数据集上的可视化结果

4 消融实验

为进一步验证深度引导策略在本文方法中的有效性,

最后进行了消融实验。定量实验结果如表 3 所示, 对应的可视化结果如图 4 所示。由表 3 可见, 引入深度先验正则可以显著提升实验效果。由图 4 可见, 深度正则化能够有效消除在渲染过程中的伪影现象^[24-25]。深度可视化显示



Fig. 3 Visualization results of different methods on the Mip-NeRF360 dataset

图3 不同方法在Mip-NeRF360数据集上的可视化结果

渲染的深度图可以有效地对齐场景的几何结构,而深度正则化可以显著改善高频细节。

从实验结果可以看出,仅采用全局归一化即可在一定程度上提升模型的拟合能力,这主要得益于局部补丁损失在优化过程中提供的稳定约束。进一步引入局部归一化后,渲染质量得到了明显改善,尤其是在细节表达方面表现更加突出。与不适用于该任务的全参数深度正则化相

比,本文提出的深度先验在全局与局部两个尺度上均展现出更好的适应性和优化效果,从而验证了标准化策略的有效性。

Table 3 Depth-guided Gaussian method ablation experiment

表3 深度引导高斯方法消融实验

Depth-guided	PSNR (db) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
$\times$	36.71	0.071	0.041
$\checkmark$	38.72	0.078	0.022

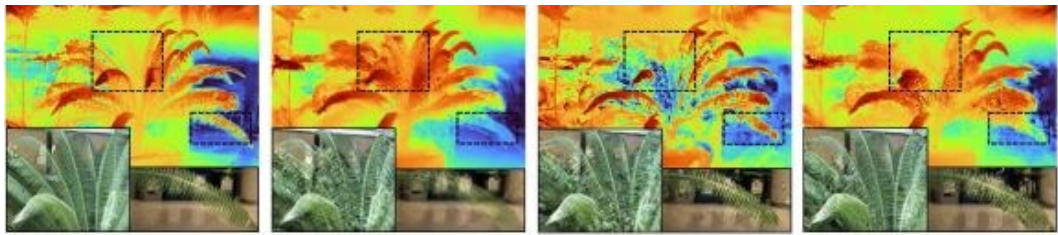


Fig. 4 Depth visualization results

图4 深度可视化结果

5 结语

本文提出一种结合深度引导的4D高斯速率神经渲染方法,通过引入基于神经特征的高斯溅射渲染与深度引导采样策略,有效提升了模型的紧凑性、动态建模能力和复杂场景下的渲染质量。实验结果表明,该方法在高分辨率与高帧率条件下实现了很好的渲染性能,同时保持了很小的模型规模。未来,可进一步向更大规模动态场景、更高效的时空特征建模以及多模态深度先验融合方向拓展,并

在实时渲染、增强现实和机器人视觉等应用中展现更广阔的潜力。

参考文献:

[1] MILDENHALL B, SRINIVASAN P P, TANCIK M, et al. NeRF: representing scenes as neural radiance fields for view synthesis[J]. Communications of the ACM, 2021, 65(1): 99-106.

[2] BARRON J T, MILDENHALL B, TANCIK M, et al. Mip-NeRF: a multi-scale representation for anti-aliasing neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 5855-5864.

- [3] BARRON J T, MILDENHALL B, VERBIN D, et al. Mip-NeRF 360: unbounded anti-aliased neural radiance fields[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5470-5479.
- [4] BARRON J T, MILDENHALL B, VERBIN D, et al. Zip-NeRF: anti-aliased grid-based neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 19697-19705.
- [5] CHEN A, XU Z, GEIGER A, et al. TensorRF: tensorial radiance fields[C]//Munich: European Conference on Computer Vision, 2022: 333-350.
- [6] CHEN Z, LI Z, SONG L, et al. NeuRBF: a neural fields representation with adaptive radial basis functions[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 4182-4194.
- [7] FRIDOVICH-KEIL S, YU A, TANCIK M, et al. Plenoxels: radiance fields without neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5501-5510.
- [8] SHAO R, ZHENG Z, TU H, et al. Tensor4D: efficient neural 4D decomposition for high-fidelity dynamic reconstruction and rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16632-16642.
- [9] KERBL B, KOPANAS G, LEIMKÜHLER T, et al. 3D Gaussian splatting for real-time radiance field rendering[J]. ACM Transactions on Graphics, 2023, 42(4): 1-14.
- [10] CAO A, JOHNSON J. Hexplane: a fast representation for dynamic scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 130-141.
- [11] FRIDOVICH-KEIL S, MEANTI G, WARBURG F R, et al. K-planes: explicit radiance fields in space, time, and appearance[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 12479-12488.
- [12] SUN C, SUN M, CHEN H T. Direct voxel grid optimization: super-fast convergence for radiance fields reconstruction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5459-5469.
- [13] ATTAL B, HUANG J B, RICHARDT C, et al. HyperReel: high-fidelity 6-DoF video with ray-conditioned sampling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16610-16620.
- [14] LI T, SLAVCHEVA M, ZOLLHOEFER M, et al. Neural 3D video synthesis from multi-view video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 5521-5531.
- [15] SONG L, CHEN A, LI Z, et al. NeRFPlayer: a streamable dynamic scene representation with decomposed neural radiance fields[J]. IEEE Transactions on Visualization and Computer Graphics, 2023, 29(5): 2732-2742.
- [16] WANG F, TAN S, LI X, et al. Mixed neural voxels for fast multi-view video synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 19706-19716.
- [17] ZWICKER M, PFISTER H, BAAR J V, et al. EWA volume splatting[C]//Proceedings of the Conference on Visualization, 2001: 29-538.
- [18] FANG L, JIANG Q, SHI J, et al. TPNet: trajectory proposal network for motion prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6797-6806.
- [19] HOUENOU A, BONNIFAIT P, CHERFAOUI V, et al. Vehicle trajectory prediction based on motion model and maneuver recognition[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013: 4363-4369.
- [20] BROXTON M, FLYNN J, OVERBECK R, et al. Immersive light field video with a layered mesh representation[J]. ACM Transactions on Graphics, 2020, 39(4): 86: 1-86: 15.
- [21] KINGMA D P, BA J. A method for stochastic optimization [DB/OL]. <https://arxiv.org/abs/1412.6980>.
- [22] HU W, WANG Y, MA L, et al. Tri-MipRF: Tri-Mip representation for efficient anti-aliasing neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023: 19774-19783.
- [23] LUITEN J, KOPANAS G, LEIBE B, et al. Dynamic 3D Gaussians: tracking by persistent dynamic view synthesis [DB/OL]. <https://arxiv.org/html/2308.09713>.
- [24] WU G, YI T, FANG J, et al. 4D Gaussian splatting for real-time dynamic scene rendering[DB/OL]. <https://arxiv.org/abs/2310.08528>.
- [25] LI Z, CHEN Z, LI Z, et al. Spacetime Gaussian feature splatting for real-time dynamic view synthesis [DB/OL]. <https://arxiv.org/abs/2312.16812>.

(责任编辑:黄健)