

基于神经面元融合溅射学习的实时真实 场景重建算法

邬鹏坤¹, 吴星辰²

(1. 沈阳航空航天大学设计艺术学院, 沈阳 110136; 2. 季华实验室, 广东 佛山 528200)

摘要: 基于一组二维图像对三维场景进行重建的过程中, 现有算法难以建模不同时间图片间的关联关系, 导致产生较差的大型环境渲染结果, 针对该问题, 提出一种用于实时真实场景重建的神经面元融合溅射算法。该算法从流式输入数据中逐步构建和更新场景模型, 包括基于 2D 高斯基元的神经面元表示和基于门控循环单元 (GRU) 的融合机制。首先, 通过将高斯面元溅射集成到渲染管道中并采用显式光线溅射交点, 实现了透视校正渲染, 从而增强了跨多个视点的深度一致性和几何保真度; 然后, 为了进一步提高重建表面的质量, 该方法结合了深度拉取和法线一致性正则化, 从而促进了更平滑的表面重建并有助于提取高质量的网格; 最后, 在 DeepBlending 和 Tanks&Temples 两个大型室内/室外场景重建数据集对算法性能进行测试。实验结果表明: 该算法在定量和定性评估方面均优于现有算法。

关键词: 姿态估计; 特征对齐; 序列信息建模; 卷积神经网络

中图分类号: TP391.4 **文献标志码:** A **文章编号:** 1671-5497(2026)07-2020-06

DOI: 10.13229/j.cnki.jdxbgxb.20241303

Real-time scene reconstruction algorithm based on neural voxel fusion splatting learning

WU Peng-kun¹, WU Xing-chen²

(1. College of Design and Art, Shenyang Aerospace University Shenyang, Shenyang 110136, China; 2. JiHua Laboratory, Foshan 528200, China)

Abstract: Three-dimensional scene reconstruction from a set of two-dimensional images remains a significant and challenging research task in computer vision. Existing algorithms struggle to effectively model the relationships between images captured at different times, resulting in poor rendering quality for large-scale environments. To address this challenge, this paper proposes a neural surfels fusion splatting algorithm for real-time scene reconstruction. The algorithm progressively builds and updates scene models from streaming input data, incorporating neural surfel representations based on 2D Gaussian primitives and a fusion mechanism based on Gated Recurrent Units (GRU). First, the algorithm achieves perspective-correct rendering by integrating Gaussian surfel splatting into the rendering pipeline and employing explicit

收稿日期: 2024-12-04.

基金项目: 辽宁省教育厅一般科研项目 (LJKMR20220546).

作者简介: 邬鹏坤 (1980-), 男, 副教授. 研究方向: 数字媒体艺术. E-mail: 2677675362@qq.com

通信作者: 吴星辰 (1990-), 男, 高级工程师, 博士. 研究方向: 计算机视觉. E-mail: wuxingchen3687@sina.com

ray-splat intersections, thereby enhancing depth consistency and geometric fidelity across multiple viewpoints. Then, to further improve the quality of reconstructed surfaces, the method combines depth pulling and normal consistency regularization, promoting smoother surface reconstruction and facilitating high-quality mesh extraction. The algorithm's performance was evaluated on two large-scale indoor/outdoor scene reconstruction datasets: DeepBlending and Tanks&Temples. Extensive experimental results demonstrate that the algorithm outperforms existing methods in both quantitative and qualitative assessments.

Key words: pose estimation; feature alignment; sequence information modeling; CNN

0 引言

大型室内场景的照片级真实感重建和渲染是计算机视觉和图形学领域的基本挑战,在虚拟现实、机器人技术和沉浸式模拟中有着广泛的应用。传统方法,例如视觉同步定位和地图构建(SLAM)系统^[1-3],可以实现实时3D重建,但通常使用具有顶点颜色或纹理贴图的实体表面来表示场景。这些表示缺乏高质量照片级真实感渲染所需的细粒度外观细节,限制了它们在需要逼真视觉体验的场景中的有效性。

神经渲染领域的最新进展,尤其是神经辐射场(NeRF),通过学习场景的隐式表示,在新型视图合成方面取得了显著成功。Wei等^[4]提出一种结合SfM和先验知识约束的NeRF算法,通过建模多视角图像信息实现对三维场景的重建;Chen等^[5]提出一种针对稀疏输入图像的重建算法,通过在输入图像中嵌入结构提示来解决NeRF算法难以约束几何构型的问题;Liu等^[6]使用受限的NeRF对三维场景进行建模,通过将特征和场景坐标回归到同一框架中实现场景重建。然而,这些方法通常需要对每个场景进行大量的优化,并且渲染速度较慢,尤其是在扩展到大型环境时。此外,它们通常需要在优化之前对所有输入图像进行离线处理,这阻碍了它们在实时或交互式应用中的适用性。

3D高斯飞溅(3DGS)提出了一种前景广阔的三维场景建模技术,该方法不仅能实现快速重建与实时渲染,还能提升渲染质量。其核心在于将复杂场景分解为一系列具有明确位置、不透明度、各向异性协方差及球谐系数的3D高斯分布^[7]。SuGaR在此基础上进一步发展,通过增加正则化项来促使这些高斯分布与场景表面更好地对齐,从而高效且精确地提取网格^[8]。尽管如此,实现

各向异性3D高斯与实际表面之间的紧密拟合仍然是一项具有挑战性的任务。

为解决这些限制,本文提出了一种用于在线照片级真实感重建和渲染大型室内场景的新方法,利用基于面元的表示的效率和神经网络的强大功能来实时建模复杂的场景几何和外观。具体来说,本文使用一种增强的2D高斯基元的神经面元表示对表面信息进行编码,从而辅助模型更为准确高效地建模场景表面。然后,采用一种基于门控循环单元(GRU)的融合机制,逐步将传入的图像数据集成到神经面元表示中。这种在线融合过程允许在获取新观测值时实时更新场景模型。GRU有效地从多个视图聚合信息,增强了重建场景的稳健性和完整性,同时保持了计算效率。通过使用GRU融合不断更新神经面元,本文方法支持即时反馈,这对于交互式应用程序至关重要。

为了实现高质量的渲染和准确的几何表示,将2D高斯溅射技术融入渲染管道中。与可能导致深度感知不一致和计算效率低下的传统体积表示不同,本文方法利用明确的射线溅射与2D高斯基元的交集,这会产生透视校正溅射,确保不同视点之间的深度一致性,并显著提高渲染图像的质量。2D高斯基元本身模拟了表面法线,从而可以实现更精确的阴影和照明效果。

本文使用两个大型室内/室外场景重建基准数据集(DeepBlending和Tanks&Temples)对模型性能进行评估。实验结果表明:与现有模型相比,本文模型生成的结果表现出更高的渲染质量。

1 本文算法

本文提出的基于神经面元融合溅射学习的实时真实场景重建算法,整体框架如图1所示。首先,从视频中获取有序的RGB图像以及相机位姿,对这些图像首先进行预处理,得到每一帧的点

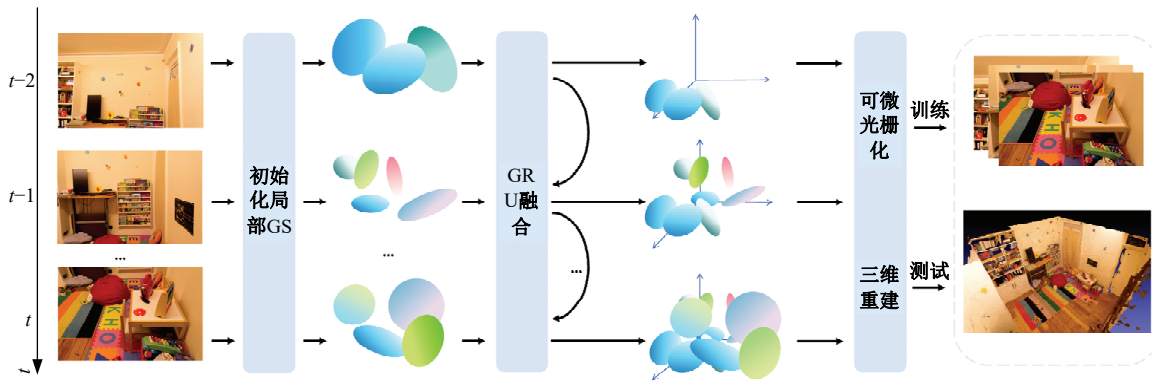


图1 基于神经面元融合溅射学习的实时真实场景重建算法整体流程图

Fig. 1 Overall pipeline of real-time real-world scene reconstruction via neural facet fusion splatting

云。然后,初始化神经面元融合溅射系统,将每一帧的点云构建局部的高斯面元组。接着,使用GRU网络融合所有顺序输入帧的局部高斯面元,得到整个场景的全局一致高斯面元组。最后,使用光栅化,将世界坐标系下的全局高斯面元渲染至图像平面,使用泊松重建生成3D mesh。

1.1 基于2D高斯基元的神经面元表示

现有大多三维重建算法需要学习场景隐式表示,从而对场景表面信息进行编码。这类算法。需要对每个场景进行大量优化建模,不仅难以对细节信息进行处理,并且渲染速度较慢。因此,本文受基于面元表示的启发,利用一种增强的2D高斯基元的神经面元表示,从而能准确高效地建模场景表面。

首先,对数据集进行预处理,从视频中获取有序的RGB图像以及相机位姿,对这些图像首先进行预处理,获得每个帧在相机坐标系下的稀疏点云,然后计算相机参数与世界坐标系对齐。接着,初始化神经面元融合溅射系统。通过获取的稀疏点云进行初始化操作获得初始化高斯面元,初始时设置随机旋转矩阵和缩放向量,用于表示颜色的球谐编码与不透明度也被随机初始化,为上述的这些属性创建优化器,并初始化学率。因此,椭圆的局部坐标系内的点的高斯分布可以表示为:

$$\zeta(u) = \exp\left(-\frac{u^2 + v^2}{2}\right) \quad (1)$$

式中: $\zeta(u)$ 为 uv 空间中点 u 的高斯分布; u 和 v 分别为椭圆坐标系内的 x 轴和 y 轴坐标。

为解决椭圆面元在特定视角观察下的形状退化问题,将式(1)扩展为:

$$\hat{\zeta}(x) = \max\left\{\zeta(u(x)), \zeta\left(\frac{x-c}{\sigma}\right)\right\} \quad (2)$$

式中: $u(x)$ 为像素 x 对应的高斯面元的 uv 局部坐标系中的坐标; c 为高斯面元中心点的投影。

直观地讲, $\hat{\zeta}(x)$ 的下界由一个固定的屏幕空间高斯低通滤波器确定,该滤波器的中心为 c ,半径为 σ ,实验中,设置 $\sigma = \sqrt{2}/2$ 以确保在渲染过程中使用足够的像素。

1.2 基于门控循环单元的融合机制

现有大多数算法独立学习每一帧图像进行三维重建,然而相邻帧中含有互补的细节特征,忽略相邻帧中的信息会导致重建性能的降低。门控循环单元(GRU)可以动态调节信息流动,使用更新门控制对所捕捉的历史信息进行选择性保留,而重置门可以控制新输入与历史信息的融合,从而高效捕捉长期依赖。在三维场景重建时,每帧图像间具有互补的信息,从不同视角图像中挖掘相关特征,可以辅助模型生成更为准确的重建结果。因此,本文提出一种基于门控循环单元的融合机制,首先将每一帧的特征编码为独立的局部高斯面元,然后再由门控循环单元从第一帧开始逐帧进行特征读取,将前一帧的全局特征与当前帧的局部特征进行融合,生成当前帧的全局特征。以第 t 帧的全局高斯面元获取过程为例,该计算过程可以描述为:

$$\hat{\zeta}_t^{\text{corr}} = \text{GRU}(\hat{\zeta}_t^{\text{merge}}, \hat{\zeta}_{t-1}^{\text{corr}}) \quad (3)$$

式中: $\hat{\zeta}_t^{\text{corr}}$ 和 $\hat{\zeta}_t^{\text{merge}}$ 分别为第 t 帧的全局高斯面元和局部高斯面元; $\hat{\zeta}_{t-1}^{\text{corr}}$ 为第 $t-1$ 帧的全局面元全局高斯面元;GRU($\cdot$)是门控循环单元。

通过上述方法,可以建模不同视角图像中的相关特征,从而辅助模型有效理解场景全局信息,进而生成令人满意的三维场景重建结果。

1.3 颜色渲染过程

通过重复执行上述基于门控循环单元的融合机制获取每一帧图像的全局高斯面元信息,然后进行颜色渲染。首先,将待渲染的图像分为均等大小的块,将各自块所观察到的高斯面元进行深度排序,并沿相机光线的方向对每个相交的面元的颜色进行 alpha 混合,得到像素颜色信息。以像素 x 为例,颜色渲染过程描述为:

$$\hat{c}(x) = \sum_{i=1} c_i \alpha_i \hat{\zeta}_i(u(x)) \prod_{j=1}^{i-1} (1 - \alpha_j \hat{\zeta}_j(u(x))) \quad (4)$$

式中: c_i 为图像块沿深度递增方向上的第 i 个高斯椭球的颜色; α_i 为其对应的不透明度; $\hat{\zeta}_i(u(x))$ 为像素 x 在第 i 个高斯局部坐标系中的全局高斯分布。

1.4 损失函数

根据现有工作^[9-11],本文使用颜色损失函数、深度损失函数和法线一致性损失函数训练模型。颜色损失函数是减小渲染颜色和真实颜色间的 L2 距离。该损失函数可表示为:

$$\ell_c = \frac{1}{|X|} \sum_{x \in X} |\hat{c}(x) - c(x)|_2 \quad (5)$$

式中: $\hat{c}(x)$ 和 $c(x)$ 分别为渲染像素颜色和真实像素颜色; X 为一系列像素点的集合。

深度拉取损失函数通过降低像素点的渲染深度和真实深度的 L2 距离来优化模型。该损失函数可表示为:

$$\ell_d = \frac{1}{|X_{vd}|} \sum_{x \in X_{vd}} |\hat{D}(x) - D(x)|_2 \quad (6)$$

式中: X_{vd} 为像素集合 X 中具有有效深度信息的相机光线; $\hat{D}(x)$ 和 $D(x)$ 分别为像素点 x 的渲染深度和真实深度。

此外,由于本文的表示基于高斯面元,因此本文必须确保所有的高斯面元与实际表面局部对齐。本文将渲染的法线图与法线先验图对齐,如下所示:

$$\ell_n = \sum_i \omega_i (1 - n_i^T / N) \quad (7)$$

式中: i 为沿射线相交的面元的索引; ω_i 为交点的混合权重; n_i^T 为朝向相机条带的法线; N 为通过深度图的梯度估计的法线。

2 实验结果与分析

本文使用两个室内/室外场景基准数据集 DeepBlending^[12] 和 Tanks&Temples^[12] 对模型性

能进行评估。

2.1 实验设置

2.1.1 数据集

DeepBlending 数据集:该数据集共包含 19 个不同场景,共计 2 630 张图片,每个场景包含 12~418 个输入图像,分辨率从 $1\,228 \times 816$ 到 $2\,592 \times 1\,944$ 不等。其中,包括 5 个室内场景和 5 个含有大量植被的场景等具有挑战性的重建数据。数据集中 90% 的数据为训练集,剩余 10% 的图片作为测试集。

Tanks&Temples 数据集:该数据集是一个基于图像的三维重建数据集,包括火车、汽车、礼堂和博物馆等 14 个室内和室外场景。Ground Truth 由高质量工业激光扫描仪采集,并提供高分辨率视频用于训练三维重建算法。

2.1.2 基线和评价指标

本文将 3 种先进的算法作为对比来证明模型的性能,包括 NeuS^[14]、3DG^[7] 和 SuGaR^[8]。根据现有工作^[14,15],使用标准评价指标 PSNR、SSIM 和 LPIPS 对重建质量进行评估。

2.2 实验结果

2.2.1 DeepBlending 数据集实验结果

将本文算法与现有 3 种先进的姿态估计模型在 DeepBlending 数据集上进行比较。量化评价结果如表 1 所示,在 Playroom 场景中本文所提出模型的 PSNR 为 30.87,优于其他 3 种现有模型。与用于进行三维场景重建的经典算法 NeuS^[14] 相比,本文算法的 PSNR 提升了 11.5%。在 Dr. Johnson 场景中与最先进的 SuGaR^[8] 相比,本文算法的 LPIPS 降低了 2.3%。相较于现有算法,在所有场景中本文算法表现出更优的 SSIM,这说明本文方法可以生成整体更令人满意的结果。

可视化结果如图 2 所示,在背景信息复杂的

表 1 本文算法与现有方法在 DeepBlending 数据集的定量结果比较

Table 1 Quantitative results comparison between our algorithm and existing methods on the deepblending dataset

方法	Playroom			Dr. Johnson		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeuS ^[14]	27.69	0.883	0.284	26.30	0.864	0.314
3DG ^[7]	30.01	0.898	0.260	28.45	0.888	0.277
SuGaR ^[8]	30.12	0.897	0.257	28.71	0.888	0.268
本文算法	30.87	0.898	0.253	29.15	0.889	0.262



图 2 本文算法在 DeepBlending 数据集上室内场景的可视化结果

Fig. 2 Quantitative results comparison between this algorithm and existing methods on the deepblending dataset

室内场景,无论是明亮环境,还是较暗环境中本文模型均可实现较为精准的三维场景重建,进一步证明了该模型的有效性。与 SuGaR 的生成结果相比,本文算法所构建的三维场景中细节渲染更为准确,且没有产生扭曲的物体。

2.2.2 Tanks&Temples 数据集实验结果

本文在公开基准数据集 Tanks&Temples 对模型性能进行检测。量化评估结果如表 2 所示,与现有的 3 种常见模型相比,本文模型展示出最优的 PSNR、SSIM 和 LPIPS。本文模型的 PSNR 为 21.36,相较于最先进的 SuGaR 可提升 4.2%。可视化结果如图 3 所示,与其他算法所生成结果

表 2 本文算法与现有方法在 Tanks&Temples 数据集的定量结果比较

Table 2 Quantitative results comparison between this method and existing approaches on the tanks & temples dataset

方法	Train			Truck		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeuS ^[14]	18.13	0.694	0.335	20.08	0.787	0.235
3DG ^[7]	19.86	0.749	0.274	22.24	0.822	0.187
SuGaR ^[8]	20.50	0.763	0.258	22.67	0.827	0.174
本文算法	21.36	0.765	0.256	23.51	0.831	0.169

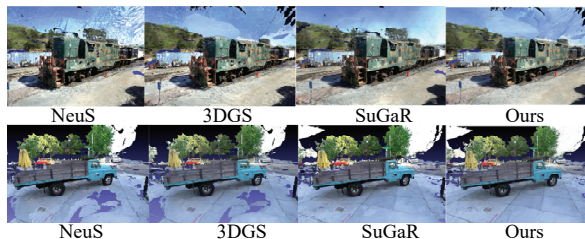


图 3 本文算法在 Tanks&Temples 数据集上室外场景的可视化结果

Fig. 3 Visualization results of the algorithm on outdoor scenes in the Tanks & Temples dataset

对比发现,本文算法对物体的细节、地面和光影的渲染更为准确,进一步证明所提出方法的有效性。

2.3 消融实验分析

为进一步了解本文所提出的模型,通过设计如下消融实验对各个模块的性能进行检测。该实验在 DeepBlending 数据集的 Playroom 场景中进行,实验结果如表 3 所示。

表 3 消融实验

Table 3 Ablation experiment

方 法	PSNR	SSIM	LPIPS
使用神经辐射场表示	29.83	0.892	0.257
删除基于门控循环单元的融合机制	30.12	0.885	0.261
完整模型	30.87	0.898	0.253

2.3.1 基于 2D 高斯基元的神经面元表示

将本文模型所使用的基于 2D 高斯基元的神经面元表示替换为神经辐射场表示。从表 3 的第 1 和 3 行的对比可以发现,PSNR 有明显降低,这说明重建图像的细节受到严重的负面影响。其原因在于本文方法可以对场景表面进行更为有效地编码,避免视觉信息的衰减。

2.3.2 基于门控循环单元的融合机制

本文通过将基于门控循环单元的融合机制直接删除的方式探究其作用。当该模块缺失时,算法难以从全局角度理解所渲染三维场景的特征,导致 SSIM 和 LPIPS 受到重大影响。此外,PSNR 也降低,进一步证明邻近帧中含有关键的三维重建相关特征。

3 结束语

现有的三维场景重建算法大多基于神经辐射场编码场景信息,需要学习大量隐式表示,导致渲染速度较慢,限制了其在实时交互式应用中的使用。此外,视频数据中相邻帧中含有潜在的关联特征,忽略这些信息会为重建带来负面影响。为解决上述问题,本文提出一种基于神经面元融合喷射学习的实时真实场景重建算法。首先,提出一种增强的 2D 高斯基元的神经面元表示,能够准确高效地建模场景表面。然后,本文设计了一种基于门控循环单元的融合机制,通过在视频信息的观测过程中,根据新观测值来实时更新场景特征信息,辅助模型从多个视图聚合信息,增强了重建场景的稳健性和完整性。实验结果表明,本文所提出的模型在 DeepBlending 和 Tanks&Tem-

ples室内/室外场景重建基准数据集上与现有方法相比,在定量和定性方面均可实现较高的重建效果。未来的工作可关注于对所重建场景中物品边缘的特征进行优化,提高物品间的区分度。

参考文献:

- [1] Cheng J, Zhang L Y, Chen Q H, et al. A review of visual SLAM methods for autonomous driving vehicles[J]. Engineering Applications of Artificial Intelligence, 2022, 114: 104992.
- [2] Kazerooni I A, Fitzgerald L, Dooly G, et al. A survey of state-of-the-art on visual SLAM[J]. Expert Systems with Applications, 2022, 205: 117734.
- [3] Lee T J, Kim C H, Cho D D. A monocular vision sensor-based efficient SLAM method for indoor service robots[J]. IEEE Transactions on Industrial Electronics, 2018, 66(1): 318-328.
- [4] Wei Y, Liu S H, Rao Y M, et al. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 5610-5619.
- [5] Chen Z, Wang C, Guo Y C, et al. Structnerf: Neural radiance fields for indoor scenes with structural hints[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15694-15705.
- [6] Liu J L, Nie Q, Liu Y, et al. Nerf-loc: Visual localization with conditional neural radiance field[C] //IEEE International Conference on Robotics and Automation(ICRA), London, United Kingdom, 2023: 9385-9392.
- [7] Kerbl B, Kopanas G, Leimkühler T, et al. 3D gaussian splatting for real-time radiance field rendering[J]. ACM Trans. Graph., 2023, 42(4): 139:1-14.
- [8] Stier N, Ranjan A, Colburn A, et al. Finerecon: Depth-aware feed-forward network for detailed 3d reconstruction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 18377-18386.
- [9] Wei W F, Wang J, Xie X, et al. Real-time dense visual slam with neural factor representation[J]. Electronics, 2024, 13(16): 3332.
- [11] Tretschk E, Kairanda N, Br M, et al. State of the art in dense monocular non-rigid 3D reconstruction[J]. Computer Graphics Forum, 2023, 42(2): 485-520.
- [12] Hedman P, Philip J, Price T, et al. Deep blending for free-viewpoint image-based rendering[J]. ACM Transactions on Graphics, 2018, 37(6): 1-15.
- [13] Knapitsch A, Park J, Zhou Q Y, et al. Tanks and temples: benchmarking large-scale scene reconstruction[J]. ACM Transactions on Graphics, 2017, 36(4): 1-13.
- [8] Guédon A, Lepetit V. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2024: 5354-5363.
- [14] Wang P, Liu L, Liu Y, et al. Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction[J]. arXiv preprint arXiv:2106.10689, 2021.
- [15] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 2018: 586-595.