Deformable-3D-Gaussians

为进一步提高渲染质量，研究团队提出了一种基于光栅化 （rasterization）的单目动态场景建模 pipeline，首次将变形场（Deformation Field）与 3D 高斯（3D Gaussian Splatting）结合，实现了高质量的重建与新视角渲染。

单目动态场景（Monocular Dynamic Scene）是指使用单眼摄像头观察并分析的动态环境，其中场景中的物体可以自由移动。单目动态场景重建对于理解环境中的动态变化、预测物体运动轨迹以及动态数字资产生成等任务至关重要。

随着以神经辐射场（Neural Radiance Field，NeRF ）为代表的神经渲染的兴起，越来越多的工作开始使用隐式表征（implicit representation）进行动态场景的三维重建。尽管基于 NeRF 的一些代表工作，如 D-NeRF，Nerfies，K-planes 等已经取得了令人满意的渲染质量，他们仍然距离真正的照片级真实渲染（photo-realistic rendering）存在一定的距离。

来自浙江大学、字节跳动的研究团队认为，上述问题的根本原因在于基于光线投射（ray casting）的 NeRF pipeline 通过逆向映射（backward-flow）将观测空间（observation space）映射到规范空间（canonical space）无法实现准确且干净的映射。逆向映射并不利于可学习结构的收敛，使得目前的方法在 D-NeRF 数据集上只能取得 30 + 级别的 PSNR 渲染指标。

为了解决这一问题，该研究团队提出了一种基于光栅化（rasterization）的单目动态场景建模 pipeline，首次将变形场（Deformation Field）与 3D 高斯 （3D Gaussian Splatting）结合，实现了高质量的重建与新视角渲染。研究论文《Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction》已被计算机视觉 顶级国际学术会议 CVPR 2024 接收。值得一提的是，这是首个使用变形场将 3D 高斯拓展到单目动态场景的工作。

- 项目主页：https://ingra14m.github.io/Deformable-Gaussians/
- 论文链接：https://arxiv.org/abs/2309.13101
- 代码：https://github.com/ingra14m/Deformable-3D-Gaussians

实验结果表明，变形场可以准确地将规范空间下的 3D 高斯前向映射（forward-flow）到观测空间，不仅在 D-NeRF 数据集上实现了 10 + 的 PSNR 提高，而且在相机位姿不准确的真实场景也取得了渲染细节上的增加：

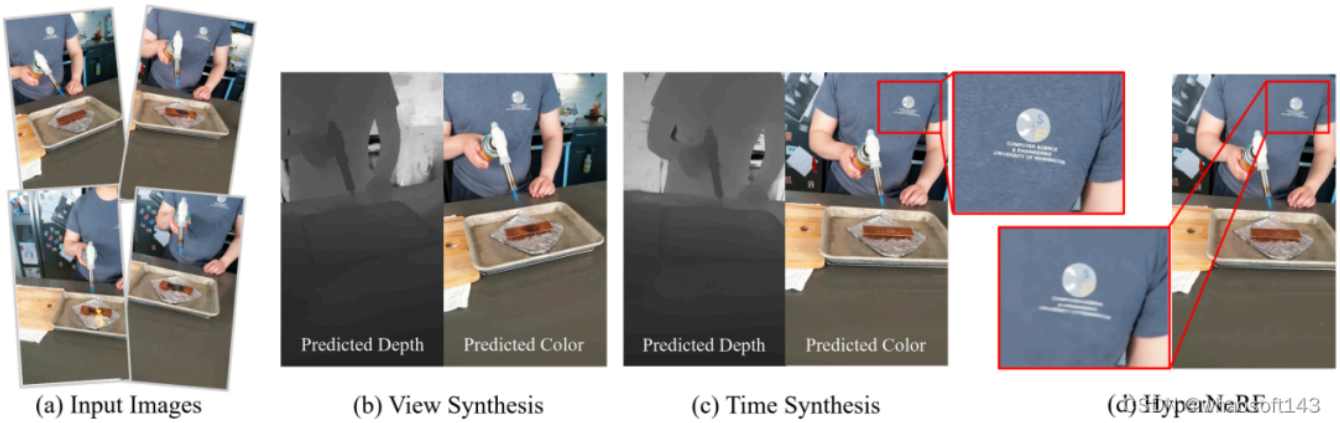

图 *1 HyperNeRF* 真实场景的实验结果。

相关工作

动态场景重建一直以来是三维重建的热点问题。随着以 NeRF 为代表的神经渲染实现了高质量的渲染，动态重建领域涌现出了一系列以隐式表征作为基础的工作。D-NeRF 和 Nerfies 在 NeRF 光线投射 pipeline 的基础上引入了变形场，实现了稳健的动态场景重建。TiNeuVox，K-Planes 和 Hexplanes 在此基础上引入了网格结构，大大加速了模型 的训练过程，渲染速度有一定的提高。然而这些方法都基于逆向映射，无法真正实现高质量的规范空间和变形场的解耦。

3D 高斯泼溅是一种基于光栅化的点云渲染 pipeline。其 CUDA 定制的可微高斯光栅化 pipeline 和创新的致密化使得 3D 高斯不仅实现了 SOTA 的渲染质量，还实现了实时渲染。Dynamic 3D 高斯首先将静态的 3D 高斯拓展到了动态领域。然而，其只能处理多目场景非常严重地制约了其应用于更通用的情况，如手机拍摄等单目场景。

研究思想

Deformable-GS 的核心在于将静态的 3D 高斯拓展到单目动态场景。每一个 3D 高斯携带位置，旋转，缩放，不透明度和 SH 系数用于图像层级的渲染。根据 3D 高斯 alpha-blend 的公式，不难发现，随时间变化的位置，以及控制高斯形状的旋转和缩放是决定动态 3D 高斯的决定性参数。然而，不同于传统的基于点云的渲染方法，3D 高斯在初始化之后，位置，透明度等参数会随着优化不断更新。这给动态高斯的学习增加了难度。

该研究创新性地提出了变形场与 3D 高斯联合优化的动态场景渲染框架。具体来说，该研究将 COLMAP 或随机点云初始化的 3D 高斯视作规范空间，随后通过变形场，以规范空间中 3D 高斯的坐标信息作为输入，预测每一个 3D 高斯随时间变化的位置和形状参数。利用变形场，该研究可以将规范空间的 3D 高斯变换到观测空间用于光栅化渲染。这一策略并不会影响 3D 高斯的可微光栅化 pipeline，经过其计算得到的梯度可以用于更新规范空间 3D 高斯的参数。

此外，引入变形场有利于动作幅度较大部分的高斯致密化。这是因为动作幅度较大的区域变形场的梯度也会相对较高，从而指导相应区域在致密化的过程中得到更精细的调控。即使规范空间 3D 高斯的数量和位置参数在初期也在不断更新，但实验结果表明，这种联合优化的策略可以最终得到稳健的收敛结果。大约经过 20000 轮迭代，规范空间的 3D 高斯的位置参数几乎不再变化。

研究团队发现真实场景的相机位姿往往不够准确，而动态场景更加剧了这一问题。这对于基于神经辐射场的结构来说并不会产生较大的影响，因为神经辐射场基于多层感知机（Multilayer Perceptron，MLP），是一个非常平滑的结构。但是 3D 高斯是基于点云的显式结构，略微不准确的相机位姿很难通过高斯泼溅得到较为稳健地矫正。

为了缓解这个问题，该研究创新地引入了退火平滑训练（Annealing Smooth Training，AST）。该训练机制旨在初期平滑 3D 高斯的学习，在后期增加渲染的细节。这一机制的引入不仅提高了渲染的质量，而且大幅度提高了时间插值任务的稳定性与平滑性。

图 2 展示了该研究的 pipeline，详情请参见论文原文。

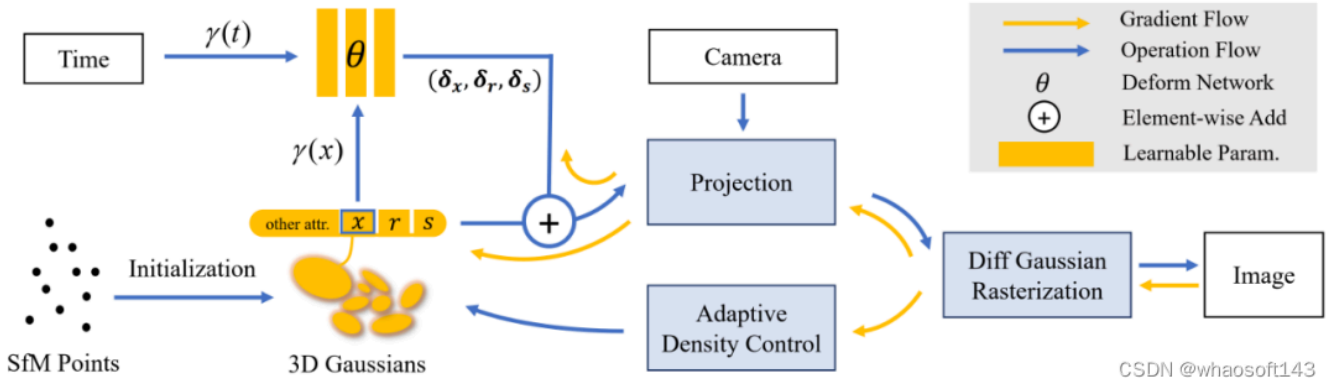

图 *2* 该研究的 *pipeline*。

结果展示

该研究首先在动态重建领域被广泛使用的 D-NeRF 数据集上进行了合成数据集的实验。从图 3 的可视化结果中不难看出，Deformable-GS 相比于之前的方法有着非常巨大的渲染质量提升。

图 *3* 该研究在 *D-NeRF* 数据集上的定性实验对比结果。

该研究提出的方法不仅在视觉效果上取得了大幅度的提升，在渲染的定量指标上也有着相应的改进。值得注意的是，研究团队发现 D-NeRF 数据集的 Lego 场景存在错误，即训练集和测试集的场景具有微小的差别。这体现在 Lego 模型铲子的翻转角度不一致。这也是为什么之前方法在 Lego 场景的指标无法提高的根本原因。为了实现有意义的比较，该研究使用了 Lego 的验证集作为指标测量的基准。

	Hell Warrior			Mutant			Hook			Bouncing Balls		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM ↑	LPIPS↓
3D-GS	29.89	0.9155	0.1056	24.53	0.9336	0.0580	21.71	0.8876	0.1034	23.20	0.9591	0.0600
D-NeRF	24.06	0.9440	0.0707	30.31	0.9672	0.0392	29.02	0.9595	0.0546	38.17	0.9891	0.0323
TiNeuVox	27.10	0.9638	0.0768	31.87	0.9607	0.0474	30.61	0.9599	0.0592	40.23	0.9926	0.0416
Tensor4D	31.26	0.9254	0.0735	29.11	0.9451	0.0601	28.63	0.9433	0.0636	24.47	0.9622	0.0437
K-Planes	24.58	0.9520	0.0824	32.50	0.9713	0.0362	28.12	0.9489	0.0662	40.05	0.9934	0.0322
Ours	41.54	0.9873	0.0234	42.63	0.9951	0.0052	37.42	0.9867	0.0144	41.01	0.9953	0.0093
	Lego			T-Rex			Stand Up			Jumping Jacks		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM ↑	LPIPS↓
3D-GS	22.10	0.9384	0.0607	21.93	0.9539	0.0487	21.91	0.9301	0.0785	20.64	0.9297	0.0828
D-NeRF	25.56	0.9363	0.0821	30.61	0.9671	0.0535	33.13	0.9781	0.0355	32.70	0.9779	0.0388
TiNeuVox	26.64	0.9258	0.0877	31.25	0.9666	0.0478	34.61	0.9797	0.0326	33.49	0.9771	0.0408
Tensor4D	23.24	0.9183	0.0721	23.86	0.9351	0.0544	30.56	0.9581	0.0363	24.20	0.9253	0.0667
K-Planes	28.91	0.9695	0.0331	30.43	0.9737	0.0343	33.10	0.9793	0.0310	31.11	0.9708	0.0468
Ours	33.07	0.9794	0.0183	38.10	0.9933	0.0098	44.62	0.9951	0.0063	37.72	0.9897	0.0126

CSDN @whaosoft143

图 *4* 在合成数据集上的定量比较。

如图 4 所示，该研究在全分辨率（800x800）下对比了 SOTA 方法，其中包括了 CVPR 2020 的 D-NeRF，Sig Asia 2022 的 TiNeuVox 和 CVPR2023 的 Tensor4D，K-planes。该研究提出的方法在各个渲染指标（PSNR、SSIM、LPIPS），各个场景下都取得了大幅度的提高。

该研究提出的方法不仅能够适用于合成场景，在相机位姿不够准确的真实场景也取得了 SOTA 结果。如图 5 所示，该研究在 NeRF-DS 数据集上与 SOTA 方法进行了对比。实验结果表明，即使没有对高光反射表面进行特殊处理，该研究提出的方法依旧能够超过专为高光反射场景设计的 NeRF-DS，取得了最佳的渲染效果。

whaosoft aiot http://143ai.com

图 *5* 真实场景方法对比。

虽然 MLP 的引入增加了渲染开销，但是得益于 3D 高斯极其高效的 CUDA 实现与我们紧凑的 MLP 结构，我们依旧能够做到实时渲染。在 3090 上 D-NeRF 数据集的平均 FPS 可以达到 85（400x400），68（800x800）。

此外，该研究还首次应用了带有前向与反向深度传播的可微高斯光栅化管线。如图 6 所示，该深度也证明了 Deformable-GS 也可以得到鲁棒的几何表示。深度的反向传播可以推动日后很多需要使用深度监督的任务，例如逆向渲染（Inverse Rendering），SLAM 与自动驾驶等。

图6 深度可视化。