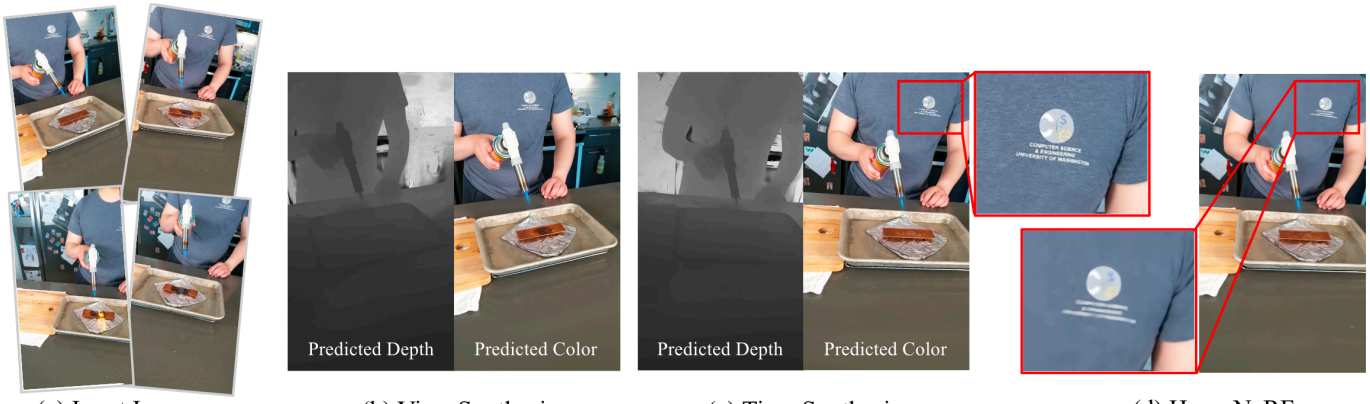
Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction



(a) Input Images (b) View Synthesis (c) Time Synthesis (d) HyperNeRF

Figure 1. Given a set of monocular multi-view images and camera poses (a), our proposed method can reconstruct accurate dynamic scene geometry and render high-quality images in both the novel-view synthesis (b) and time interpolation (c) tasks. In real-world datasets with intricate details, our method outperforms *HyperNeRF* [31] (d) in terms of rendering quality and time performance.

Introduction

背景

高质量的动态场景重建和真实感渲染对于许多应用至关重要，例如增强现实/虚拟现实（AR/VR）、3D 内容制作和娱乐。这些任务的目标是通过一组输入图像，生成高保真几何结构和逼真的视图效果。

传统方法的不足

- **网格表示方法**：基于网格的方法（如文献 [9, 14, 18, 40]）被广泛使用，但存在以下问题：
  - **细节缺失**：无法捕捉复杂的几何细节。
  - **语义信息不足**：缺乏对场景中语义元素的表达能力。
  - **拓扑变化支持不足**：难以处理动态场景中的拓扑变化。

神经渲染的引入

神经渲染技术（例如 NeRF [28]）引入后，为以下任务带来了显著改进：

- 新视角合成
- 场景重建
- 光照分解

然而，NeRF 等隐式表示方法在动态场景中表现出以下缺点：

- **效率低下**：训练时间长，推理速度慢。
- **过拟合问题**：对复杂动态场景表现不足。

现有加速方法

在静态场景中，已有加速方案，例如：

- **基于网格的结构**：如 [7, 46] 提出的网格结构。
- **预计算策略**：如 [44, 52] 中的优化方法。
- **哈希编码**：Instant-NGP [29] 使用哈希编码大幅提升了训练速度。

动态场景的挑战

- **高维表示的计算复杂度**：动态场景的表示通常需要考虑时间维度，导致显著的计算开销。
- **实时渲染难度**：现有方法难以同时实现高质量与实时性。

本文的贡献

为了解决上述挑战，本文提出了一种基于 **可变形的 3D 高斯分布 (Deformable 3D Gaussians)** 的框架：

1. **高效动态场景建模**：通过可变形的 3D 高斯和时间相关的变形场，描述动态场景。
2. **自适应平滑训练机制**：通过退火平滑训练（AST）机制，解决姿态误差引起的时间插值任务抖动问题。
3. **实时渲染性能**：借助差分高斯光栅化管道，同时实现高质量渲染和实时性。

Related Work

本文的研究围绕两个主要方向：**动态场景的神经渲染** 和 **神经渲染的加速**。

Neural Rendering for Dynamic Scenes

背景

神经渲染技术凭借其生成高保真图像的能力，近年来备受关注。NeRF [28] 是该领域的一个里程碑，通过多层感知机（MLPs）建模场景的辐射场，为新视角合成任务提供了显著改进。此后，该方法被扩展至多种任务，例如：

- **网格重建**（如 [20, 45]）

- **反演渲染**（如 [5, 25, 54]）
- **相机参数优化**（如 [21, 47, 48]）
- **少样本学习**（如 [10, 51]）

动态场景的挑战

动态场景中的时间维度引入了额外的复杂性，尤其是对于单目（monocular）动态场景的重建。此类任务通常涉及从稀疏视点进行重建。

方法分类

1. **时间变量耦合方法**
   - 在 NeRF 的输入中添加时间变量 t，使辐射场依赖时间 t。
   - **问题**：这种耦合策略会将时间变化与辐射场混杂，缺乏对时间影响场景几何的几何先验，需强正则化来保证渲染结果的时间一致性。
2. **时间变量解耦方法**
   - 使用变形场（deformation field）将点坐标映射到时间 t 对应的规范空间（canonical space）。
   - 适用于描述显著的刚体运动以及拓扑变化的场景。

改进方向

为提升动态神经渲染的质量，研究者从多个方面进行优化：

- **场景分割**：如 [39, 42]，将静态与动态对象进行分离。
- **引入深度信息**：如 [1]，通过几何先验提升渲染质量。
- **场景先验编码**：如 [22, 33]，通过 2D CNN 提取场景特征。
- **多视图冗余信息**：如 [19]，通过关键帧压缩加速渲染。

本文的改进

本文基于可区分的点渲染框架，进一步分离变形场和辐射场，解决现有基于 MLP 表示的动态场景渲染质量不佳的问题，同时提高动态场景中中间状态的可编辑性。

Acceleration of Neural Rendering

静态场景的加速方法

为实现实时渲染，研究者们探索了多种加速方案，包括：

1. **预计算方法**

- 通过球谐函数系数或特征向量的预计算优化推理速度（如 [12, 35]）。
- **优点**：显著提升推理速度。
- **缺点**：训练时间长、存储需求高。

2. **混合方法**
   - 将显式网格结构嵌入神经网络，兼顾训练效率与渲染质量（如 [4, 7, 24]）。
   - 已扩展至时间条件下的 4D 特征建模（如 [6, 36]）。
3. **点基渲染**
   - 通过显式点渲染实现实时加速（如 [15]），并通过 CUDA 管道优化高斯点的光栅化。

动态场景的加速方法

在动态场景中，渲染质量与实时性之间的权衡更加复杂。本文受到 3D-GS [15] 启发，使用可区分的高斯点渲染框架扩展至动态场景，并结合优化的变形场实现更高效的训练和渲染。

结论

现有方法虽然在静态场景的加速与动态场景的神经渲染方面取得了进展，但在动态场景中，实时高质量渲染仍然面临挑战。本文结合上述两类技术，提出了可变形的 3D 高斯框架，为动态场景建模提供了一种新思路。

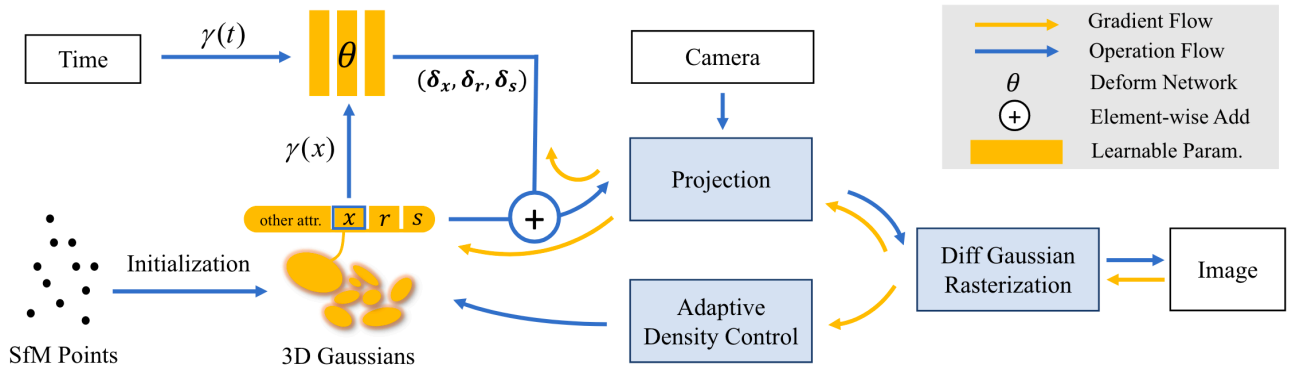


Figure 2. **Overview of our pipeline.** The optimization process begins with Structure from Motion (SfM) points derived from COLMAP or generated randomly, which serve as the initial state for the 3D Gaussians. We use the position (detached) of 3D Gaussians $\gamma(\mathrm{sg}(\boldsymbol{x}))$ and time $\gamma(t)$ with positional encoding as input to a deformation MLP network to obtain the offset $(\delta\boldsymbol{x}, \delta\boldsymbol{r}, \delta\boldsymbol{s})$ of dynamic 3D Gaussians in canonical space. We use a warm-up phase for the 3D Gaussians during the first 3k iterations without optimizing the deformation field. Following that, we use the fast differential Gaussian rasterization pipeline to perform joint optimization of the deformation field and the 3D Gaussians, as well as to adaptively control the density of the set of Gaussians.

Method

本文提出了一种基于可变形 3D 高斯分布的动态场景建模框架，通过结合 3D 高斯与变形场，实现对单目动态场景的高效、高质量建模。本节从框架概述、可区分渲染、高斯变形及退火平滑训练机制等方面进行详细说明。

1. Framework Overview

输入数据包括：

1. 一组单目动态场景的图像
2. 时间标签 t
3. 相机位姿 θ，通过 SfM [37] 估计，生成稀疏点云。

从稀疏点云中初始化 3D 高斯分布集合，表示为：

$$G(x, r, s, \sigma) \tag{1}$$

其中：

- x：中心位置
- r：由四元数表示的旋转
- s：缩放因子
- σ：不透明度

为了表示动态场景，本文解耦 3D 高斯和变形场，变形场以当前时间 t 为输入，输出偏移量 $(\delta x, \delta r, \delta s)$，最终得到变形的 3D 高斯：

$$G(x + \delta x, r + \delta r, s + \delta s, \sigma) \tag{2}$$

整个优化流程基于一个高效的差分高斯光栅化管道，支持动态高斯的联合优化以及高斯密度的自适应控制。

---

2. Differentiable Rendering Through 3D Gaussians Splatting

2.1 投影与渲染

为了优化规范空间中的 3D 高斯参数，需要通过差分渲染生成 2D 图像。每个 3D 高斯投影到 2D 平面时，其协方差矩阵表示为：

$$\Sigma' = JV\Sigma V^T J^T \tag{3}$$

其中：

- J：射影变换的仿射近似的雅可比矩阵
- V：从世界坐标到相机坐标的视图变换矩阵
- Σ：3D 高斯的协方差矩阵

为简化学习，协方差矩阵 Σ 被分解为：

$$\Sigma = RSS^T R^T \tag{4}$$

其中：

- R：由旋转四元数 r 生成的旋转矩阵
- S：由缩放向量 s 生成的缩放矩阵

投影后，图像平面上像素 p 的颜色 $C(p)$ 使用基于点的体渲染技术计算：

$$C(p) = \sum_{i \in N} T_i \alpha_i c_i \tag{5}$$

其中：

- T_i：透射率，定义为 $\prod_{j=1}^{i-1}(1-\alpha_j)$
- α_i：第 i 个高斯的不透明度，计算公式为：

$$\alpha_i = \sigma_i e^{-\frac{1}{2}(p-\mu_i)^T \Sigma'(p-\mu_i)} \tag{6}$$

  其中 μ_i 为高斯投影到 2D 图像平面的坐标。

- c_i：第 i 个高斯的颜色

2.2 高斯密度的自适应控制

在优化过程中，适应性密度控制是渲染的关键组件，主要有两方面功能：

1. **稀疏区域填充**：对几何细节不足的区域进行填充。
2. **重叠区域细分**：对过大或重叠的高斯进行拆分。

具体操作：

- 若高斯过大且重叠显著，将其拆分并缩小比例，缩放因子为超参数 $\xi = 1.6$。
- 对于不足以捕捉细节的小高斯，复制并沿梯度方向移动。

---

3. Deformable 3D Gaussians

3.1 解耦动态与几何

为有效建模动态场景，本文将动态场景的运动和几何结构解耦：

1. 通过 3D 高斯建模时间独立的几何结构。
2. 使用变形网络建模时间相关的动态变化。

变形网络为一个 MLP，输入为 3D 高斯的中心位置 x 和时间 t，输出偏移量：

$$(\delta x, \delta r, \delta s) = F_\theta(\gamma(sg(x)), \gamma(t)) \tag{7}$$

其中：

- $\gamma(\cdot)$：位置编码函数，定义为：

$$\gamma(p) = (\sin(2^k \pi p), \cos(2^k \pi p))_{k=0}^{L-1} \tag{8}$$

  - $L = 10$：对于位置 x 和时间 t 的编码长度

- $sg(\cdot)$：停止梯度操作，防止偏移影响输入高斯。

---

4. Annealing Smooth Training (AST)

真实数据集中，姿态估计误差会导致时间插值任务中的抖动问题。为此，本文提出了一种退火平滑训练机制，其关键公式为：

$$\Delta = F_\theta(\gamma(sg(x)), \gamma(t) + X(i)) \tag{9}$$

其中：

- $X(i) = N(0,1) \cdot \beta \cdot \Delta t \cdot (1 - i/\tau)$
- $N(0,1)$：标准正态分布
- $\beta = 0.1$：缩放因子
- Δt：平均时间间隔
- $\tau = 20k$：退火阈值

退火平滑训练的优势：

1. 早期阶段提高模型的时间广义性。
2. 后期避免过度平滑，保留动态细节。
3. 显著降低时间插值任务中的抖动。

1. 背景与动机

在动态场景建模中，由于相机位姿估计误差（如使用 COLMAP 生成的位姿可能不精确），在时间插值任务中常出现如下问题：

1. **帧间不连续性**：即场景几何结构在不同时间点发生抖动。
2. **过拟合风险**：模型可能对训练帧过度拟合，导致时间插值时无法生成平滑的动态效果。

传统的隐式表示（如基于 MLP 的方法）通过其固有的平滑特性可以缓解这一问题，但显式点渲染方法（如 3D-GS）会放大位姿误差对结果的影响。

为此，本文提出 **退火平滑训练机制 (AST)**，在训练初期通过添加动态噪声提升模型的时间泛化能力，在后期逐步减小噪声以保留动态细节。

---

2. 方法描述

AST 的核心是为变形网络的时间输入 t 添加随训练迭代动态衰减的高斯噪声，从而优化变形场 F_θ 的时间平滑性。其公式为：

$$\Delta = F_\theta(\gamma(sg(x)), \gamma(t) + X(i)) \tag{10}$$

其中：

- $X(i)$：在第 i 次训练迭代中加入的动态噪声，定义为：

$$X(i) = N(0,1) \cdot \beta \cdot \Delta t \cdot (1 - i/\tau) \tag{11}$$

- $\gamma(\cdot)$：位置编码函数，将时间 t 转换为高维表示。
- $sg(x)$：停止梯度操作，防止偏移量 Δ 的优化反作用于输入高斯。

参数解析

1. $N(0,1)$：标准高斯分布，用于生成随机噪声。
2. β：噪声缩放因子，控制噪声的初始强度，本文中取值为 0.1。
3. Δt：时间间隔的平均值，用于标准化噪声幅度。
4. i：当前训练迭代次数。
5. τ：退火阈值，定义噪声衰减的结束点（例如 $\tau = 20k$ 次迭代）。

---

3. 训练过程

退火过程分为三个阶段：

1. **初期阶段（$i \ll \tau$）**：

- 动态噪声 $X(i)$ 较大，帮助模型探索时间维度的广义性。
- 加强变形场 F_θ 对姿态误差的鲁棒性，避免过拟合训练帧。

2. **中期阶段（$i \approx \tau/2$）**：
   - 噪声逐步减小，模型逐渐学习动态细节。
   - 时间插值效果开始平滑，减少帧间抖动。
3. **后期阶段（$i \to \tau$）**：
   - 噪声衰减至接近零，保留动态细节而不影响时间一致性。

---

4. 优势与效果

AST 机制的主要优势：

1. **增强时间泛化能力**：
   - 在训练初期，添加噪声能够帮助模型跳出局部最优，避免对训练帧的过拟合。
2. **减少帧间抖动**：
   - 在时间插值任务中，帧间几何和纹理的过渡更加自然。
3. **无额外计算开销**：
   - 与传统的平滑损失（如 L_{smooth}）相比，AST 仅通过简单的噪声添加实现平滑效果，不增加额外的计算复杂度。

5. 实验验证

定量分析

实验结果表明，在 NeRF-DS 和 HyperNeRF 数据集中，使用 AST 的模型在以下指标上均优于无 AST 的模型：

- **PSNR**（峰值信噪比）提高约 1-2 分贝。
- **SSIM**（结构相似性）提高 0.01-0.02。
- **LPIPS**（感知损失）降低 0.02-0.03。

定性分析

对比有无 AST 的渲染结果：

- **无 AST**：时间插值任务中出现明显的几何抖动，尤其是在复杂动态场景中。
- **有 AST**：帧间过渡平滑，动态细节得到保留。

示例：对比 **Jumping Jacks** 场景中的插值结果，可以清晰看到 AST 在减少抖动和保留动态细节方面的显著效果。

---

6. 小结

AST 是一种简单高效的平滑训练机制，通过动态噪声退火，不仅提高了模型的时间一致性，还能显著改善动态场景中的渲染质量和细节表现。

总结

本文通过差分渲染、高斯变形和退火平滑机制，联合优化动态场景中的几何和运动，为高效、高质量的动态场景建模提供了一个新框架。

Experiment

本文实验旨在验证所提方法在动态场景建模中的有效性和高效性，实验分为以下几个部分：

1. 实验设置与实现细节。
2. 与现有方法的对比分析。
3. 消融实验验证各模块的贡献。

---

1. Implementation Details

1.1 数据集

- **合成数据集**：使用 D-NeRF 数据集 [34]，场景包含高质量几何结构与动态变化。
- **真实数据集**：
  - HyperNeRF 数据集 [31]
  - NeRF-DS 数据集 [50]

训练与测试分割：与原始论文设置保持一致，所有合成数据以 800x800 分辨率进行测试。

1.2 模型训练

- **优化策略**：使用 Adam 优化器，超参数设置为：
  - 学习率：3D 高斯为 $8e-4$，变形网络采用指数衰减，从 $8e-4$ 至 $1.6e-6$。
  - 动量系数 $\beta = (0.9, 0.999)$。
- **迭代设置**：
  - 前 3000 次迭代，仅优化 3D 高斯以稳定几何结构。
  - 后续联合优化 3D 高斯和变形场，确保动态结构的准确性。

1.3 渲染效率

- 在 NVIDIA RTX 3090 GPU 上测试，当 3D 高斯数量小于 250k 时，可以实现超过 30 FPS 的实时渲染。

---

2. Results and Comparisons

2.1 定量评估

本文方法在合成数据集和真实数据集上，与以下方法进行了对比：

1. **3D-GS** [15]：静态场景点渲染。
2. **D-NeRF** [34]：动态辐射场建模。
3. **TiNeuVox** [11]：时间感知的体素方法。
4. **Tensor4D** [38]：高效的 4D 神经网络。
5. **K-Planes** [36]：时间条件下的显式平面建模。

以下是关键指标：

- **PSNR**：峰值信噪比，用于评估图像质量。
- **SSIM**：结构相似性，用于评估图像的结构保真度。
- **LPIPS**：感知相似性，越小越好。

表格展示了本文方法在多种场景下的性能优势：

$$\text{PSNR} \uparrow, \text{SSIM} \uparrow, \text{LPIPS} \downarrow \tag{12}$$

2.2 定性评估

通过视觉化对比，本文方法在以下方面优于现有方法：

1. 更精细的动态几何细节恢复，例如手部动作或骨骼形状。
2. 时间插值任务中帧间的平滑性和一致性。

示例场景包括 **Lego**, **Jumping Jacks**, **Bouncing Balls** 等，展示了本文方法在渲染质量和动态一致性方面的显著优势。

---

3. Ablation Study

消融实验旨在分析各模块对模型性能的贡献，包括：

1. **无变形场（w/o deformation field）**：仅优化 3D 高斯。
2. **无退火平滑（w/o AST）**：不使用退火平滑机制。

3.1 变形场的贡献

移除变形场后，模型对动态几何结构的描述能力显著下降，尤其是在复杂运动场景中，如 **Mutant** 和 **T-Rex**。

3.2 退火平滑机制的贡献

实验显示，退火平滑机制能够：

- 显著提升时间插值任务中的帧间平滑性。
- 降低真实数据集中的渲染抖动。

退火平滑引入的高斯噪声能够在训练早期增强模型的时间泛化能力，同时在后期保持动态细节。

---

4. Limitations

实验也揭示了本文方法的局限性：

1. **相机位姿误差**：当位姿估计不准确时（如 HyperNeRF 数据集），可能会导致模型收敛困难。
2. **视点稀疏性**：在仅有少量训练视角的情况下，模型可能出现过拟合问题。

总结

实验结果表明，本文方法在动态场景建模中具有显著优势，包括：

1. **渲染质量**：在 PSNR 和 SSIM 等指标上优于现有方法。
2. **实时性**：在 3D 高斯数量适中的情况下，支持实时渲染。

3. **动态一致性**：通过变形场和退火平滑机制，提升了动态场景的时间一致性。