# Shape of Motion: 4D Reconstruction from a Single Video

**Fig. 1: Shape of Motion.** Our approach enables joint 3D long-range tracking and novel view synthesis from a monocular video of a complex dynamic scene. Here, we demonstrate our ability to render moving scene elements at fixed viewpoint with different moments in time. Additionally, we visualize the estimated long-range 3D motion as colorful trajectories. These trajectories reveal distinct geometric patterns that encapsulate the movement of each point through 3D space and time, which leads us to the term "Shape of Motion".

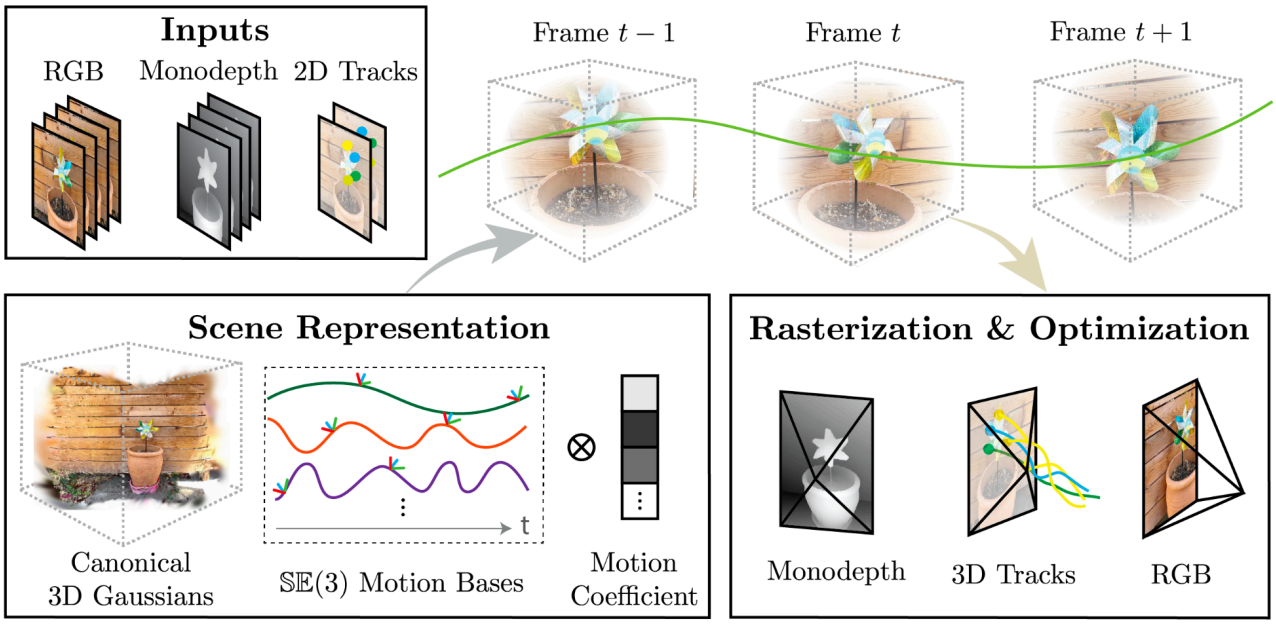


**Fig. 2: System Overview.** Given a single RGB video sequence with known camera poses, along with monocular depth maps and 2D tracks computed from off-the-shelf models [14,106] as input, we optimize a dynamic scene representation as a set of persistent 3D Gaussians that translate and rotate over time. To capture the low-dimensional nature of scene motion, we model the motion with a set of compact $\mathbb{SE}(3)$ motion bases shared across all scene elements. Each 3D Gaussian's motion is represented as a linear combination of these global $\mathbb{SE}(3)$ motion bases, weighted by motion coefficients specific to each Gaussian. We supervise our scene representation (canonical 3D Gaussian parameters, per-Gaussian motion coefficients, and global motion bases) by comparing the rendered outputs (RGB, depths and 2D tracks) with the corresponding input signals. This results in a dynamic 3D representation of the scene with explicit long-range 3D scene motion.

# Introduction

动态场景的几何和三维运动的重建是理解和交互物理世界的关键任务。然而，从单一视频中恢复复杂动态场景的几何和运动信息是一项非常具有挑战性的问题。尽管近年来在静态三维场景建模方面取得了显著进展，动态三维场景的建模仍然具有诸多未解难题。

## 背景与挑战

动态场景重建通常面临以下主要挑战：

1. **不充分的约束条件**：动态场景的重建问题高度欠约束，尤其是从单视点视频中重建时。
2. **依赖额外设备或条件**：许多现有方法需要同步多视点视频、激光雷达(LIDAR)或深度传感器等额外设备。
3. **难以捕捉全局一致的三维运动轨迹**：许多方法只关注短程运动，例如基于逐帧的场景流(scene flow)，而未能捕获跨时间段的持久运动。

## 方法与贡献

为解决上述问题，本文提出了一种新的方法，名为“Shape of Motion”，可以从随意拍摄的单目视频中重建复杂动态场景的几何和完整的三维运动。核心创新点包括：

1. **低维运动表示**：将场景的运动表示为紧凑的 SE(3) 运动基（motion bases），每个点的运动是这些基的线性组合。这种表示方法可以实现对场景的柔性分解，将场景划分为多个刚性运动的组。
2. **数据驱动先验的融合**：利用单目深度图和长时间段的二维轨迹等数据先验，设计了一种融合噪声监督信号的方法，从而获得全局一致的场景几何和运动表示。

## 核心流程

- **表示方法**：使用持续的三维高斯分布（3D Gaussians）描述动态场景。每个高斯分布的运动由一组共享的 SE(3) 运动基进行建模。
- **监督信号**：通过比较渲染输出（包括 RGB 图像、深度图和二维轨迹）与对应输入信号进行监督，从而优化动态场景表示。

## 实验结果

实验表明，本文方法在长程三维/二维运动估计和动态场景新视点合成任务上达到了最先进的性能。此外，该方法能够在保持时间一致性的前提下，高效地处理真实动态场景。

## 数学公式

动态场景的运动表示为：

$$T_{0\to t} = \sum_{b=0}^{B} w^{(b)} T_{0\to t}^{(b)} \tag{1}$$

其中：

- $T_{0\to t}$ 表示从初始时间 $t=0$ 到时间 $t$ 的运动变换。
- $T_{0\to t}^{(b)}$ 是全局共享的第 $b$ 个运动基。
- $w^{(b)}$ 是第 $b$ 个基的权重。

结合以上思路，本文首次实现了在单目视频上同时进行动态场景重建和长期三维运动追踪。

# Related Work

## 1. Correspondences and Tracking

单目长距离3D跟踪在文献中尚未被广泛研究，但许多方法已经在二维图像空间中进行了跟踪工作。
典型的二维点对应方式依赖于**光流**方法，这涉及估计图像对之间的稠密运动场：

$$\mathbf{O} = f(\mathbf{I}_1, \mathbf{I}_2) \tag{2}$$

其中，$\mathbf{O}$ 表示光流场，$f$ 为估计模型。这种方法在连续帧中有效，但在视频中实现精确的长时间跟踪仍然是一个挑战。

稀疏关键点匹配方法（如SURF）可以生成长轨迹，但它们主要用于稀疏3D重建：

$$\mathbf{p}_t \approx g(\mathbf{p}_{t-1}, \mathbf{K}) \tag{3}$$

其中，$\mathbf{p}_t$ 为时间 $t$ 的关键点位置，$g$ 是匹配函数，$\mathbf{K}$ 是内参矩阵。

一些方法通过手工设计的先验生成运动轨迹，而最近的方法采用**数据驱动策略**：

- **优化方法**：整合短距离运动估计，生成长时间对应。
- **神经网络方法**：基于合成数据学习长时间对应。

然而，这些方法缺乏对3D场景几何和运动的理解。

---

# 2. Scene Flow 和 3D Motion Trajectory

**场景流（Scene Flow）** 是建模3D场景运动和点对应的常见表示形式。
以激光雷达点云或RGBD图像为输入的场景流估计方法占主流：

$$\mathbf{S}_t = h(\mathbf{X}_t, \mathbf{X}_{t+1}) \tag{4}$$

其中，$\mathbf{S}_t$ 表示时间 $t$ 的场景流，$\mathbf{X}_t$ 和 $\mathbf{X}_{t+1}$ 是点云。

单目场景流估计方法依赖于**自监督学习**或**测试时优化策略**。
然而，这些方法要么聚焦于单一物体，要么依赖模板先验，或只能生成短距离的运动对应。

---

# 3. Dynamic Reconstruction and View Synthesis

动态3D场景重建和新视图合成（Novel View Synthesis）是相关领域的重点问题：

- 早期方法依赖于 **RGBD 传感器** 或强先验。
- 最近方法（如NeRF）在动态场景中展示了显著进展，但通常需要同步多视角视频。

**无模板的单目方法** 使用不同类型的表示形式建模动态场景，如：

1. 视频深度图
2. 时间依赖的NeRF
3. 动态3D高斯分布

然而，这些方法通常无法处理真实世界中的动态单目视频。

# 4. 本研究的改进

本研究的创新点：

1. 提出了一种新的动态场景表示方法，结合全局一致的3D跟踪和实时的新视图合成。
2. 利用物理运动先验和数据驱动先验设计框架，优化单目视频的表示。

# Method

本研究的方法以单目视频为输入，旨在恢复整个动态场景的几何形状以及场景中每个点的完整3D运动轨迹。
与传统的动态NeRF方法不同，我们采用显式的点云表示形式，通过一组共享的SE(3)运动基追踪动态场景元素的全时段运动。

---

## 3.1 Preliminaries: 3D Gaussian Splatting

我们使用全局一致的3D高斯分布表示动态场景的外观和几何属性。这种显式且可微的场景表示形式适用于高效的优化和渲染。

**3D高斯参数化**

每个3D高斯在标准参考帧 $t_0$ 中的参数定义如下：

$$g_0 \equiv (\mu_0, R_0, s, o, c) \tag{5}$$

- $\mu_0 \in \mathbb{R}^3$：高斯分布的中心位置。
- $R_0 \in SO(3)$：高斯的方向（旋转矩阵）。
- $s \in \mathbb{R}^3$：尺度。
- $o \in \mathbb{R}$：不透明度。
- $c \in \mathbb{R}^3$：颜色。

**投影和光栅化**

将3D高斯从世界坐标投影到图像平面，可以近似为2D高斯分布，其参数为：

$$\mu_0'(K,E) = \Pi(KE\mu_0), \quad \Sigma_0'(K,E) = J_{KE}\Sigma_0 J_{KE}^\top \tag{6}$$

其中：

- $K$ 是相机内参矩阵。
- $E$ 是世界到相机的外参矩阵。
- $\Pi$ 是透视投影函数。
- $J_{KE}$ 是透视投影的雅可比矩阵。

然后通过以下公式完成RGB图像和深度图的光栅化：

$$\hat{I}(p) = \sum_{i\in H(p)} T_i\alpha_i c_i, \quad \hat{D}(p) = \sum_{i\in H(p)} T_i\alpha_i d_i \tag{7}$$

其中：

- $\alpha_i = o_i \cdot \exp\left(-\frac{1}{2}(p-\mu_i')^\top \Sigma_i'(p-\mu_i')\right)$：每个高斯的权重。
- $T_i$：累积透明度。

该过程是完全可微的，允许对3D高斯参数进行直接优化。

---

## 3.2 Dynamic Scene Representation

为了表述动态3D场景，我们在时间序列中追踪3D高斯的位置和方向变化。

**运动参数化**

每个3D高斯在时间 $t$ 的位姿通过刚体变换从参考帧 $t_0$ 映射得到：

$$\mu_t = R_{0\to t}\mu_0 + t_{0\to t}, \quad R_t = R_{0\to t}R_0 \tag{8}$$

其中：

- $R_{0\to t}$ 和 $t_{0\to t}$ 是参考帧到时间 $t$ 的旋转矩阵和平移向量。

为了简化计算，我们使用全局共享的 $B \ll N$ 个SE(3)运动基 $\{T_{0\to t}^{(b)}\}_{b=1}^{B}$ 表示整个场景的运动：

$$T_{0\to t} = \sum_{b=1}^{B} w^{(b)} T_{0\to t}^{(b)}, \quad \|w^{(b)}\| = 1 \tag{9}$$

其中，$w^{(b)}$ 是每个高斯的运动系数。

**轨迹光栅化**

对于时间 $t$ 的查询帧，其像素级3D运动轨迹通过以下公式计算：

$$\hat{\mathbf{X}}_{t\to t'}(p) = \sum_{i\in H(p)} T_i \alpha_i \mu_{i,t'} \tag{10}$$

并且对应的二维位置和深度为：

$$\hat{\mathbf{U}}_{t\to t'}(p) = \Pi(K_{t'}\hat{\mathbf{X}}_{t\to t'}(p)), \quad \hat{D}_{t\to t'}(p) = (\hat{\mathbf{X}}_{t\to t'}(p))_z \tag{11}$$

---

## 3.3 Optimization

**输入处理**

在优化过程中，我们使用以下信息：

1. 动态物体的掩码 $M_t$。
2. 单目深度图 $D_t$。
3. 长时间2D轨迹 $U_{t\to t'}$。

**初始化**

1. 选择参考帧 $t_0$。
2. 从初始轨迹中随机采样3D点作为高斯的中心位置。
3. 使用加权Procrustes方法初始化运动基。

### 损失函数

优化过程中，我们定义了两类损失函数：

1. **重建损失**：匹配逐帧的颜色、深度和掩码输入：

$$\mathcal{L}_{\text{recon}} = \|\hat{I} - I\|_1 + \lambda_{\text{depth}}\|\hat{D} - D\|_1 + \lambda_{\text{mask}}\|\hat{M} - M\|_1 \quad (12)$$

2. **轨迹一致性损失**：约束不同时间帧之间的对应关系：

$$\mathcal{L}_{\text{track-2d}} = \|U_{t\to t'} - \hat{U}_{t\to t'}\|_1, \quad \mathcal{L}_{\text{track-depth}} = \|D_{t\to t'} - \hat{D}_{t\to t'}\|_1 \quad ($$

3. **刚性约束损失**：保持动态高斯及其最近邻的距离不变：

$$\mathcal{L}_{\text{rigidity}} = \|\text{dist}(X_t, \mathcal{C}_k(X_t)) - \text{dist}(X_{t'}, \mathcal{C}_k(X_{t'}))\|_2^2 \quad (14)$$

### 最终优化目标

总损失为以上各部分的加权和：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{track}}(\mathcal{L}_{\text{track-2d}} + \mathcal{L}_{\text{track-depth}}) + \lambda_{\text{rigidity}}\mathcal{L}_{\text{rigidity}} \quad (15$$

**Table 1: Evaluation on iPhone dataset**. Our method achieves SOTA performance all tasks of 3D point tracking, 2D point tracking, and novel view synthesis. The baselines that perform best on 2D and 3D tracking (TAPIR [14]+DA [106] and CoTracker [38]+DA [106]) are unable to synthesize novel viewpoints of the scene, while the methods that perform best in novel view synthesis (T-NeRF [21] and HyperNeRF [65]) struggle with or fail to produce 2D and 3D tracks. Our method achieves a significant boost in all three tasks above baselines.

| Method | 3D Tracking | | | 2D Tracking | | | View Synthesis | | |
|---|---|---|---|---|---|---|---|---|---|
| | EPE ↓ | $\delta_{3D}^{.05}$ ↑ | $\delta_{3D}^{.10}$ ↑ | AJ ↑ | $<\delta_{\text{avg}}$ ↑ | OA ↑ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| T-NeRF [21] | - | - | - | - | - | - | 15.60 | 0.55 | 0.55 |
| HyperNeRF [65] | 0.182 | 28.4 | 45.8 | 10.1 | 19.3 | 52.0 | 15.99 | 0.59 | 0.51 |
| DynIBaR [52] | 0.252 | 11.4 | 24.6 | 5.4 | 8.7 | 37.7 | 13.41 | 0.48 | 0.55 |
| Deformable-3D-GS [108] | 0.151 | 33.4 | 55.3 | 14.0 | 20.9 | 63.9 | 11.92 | 0.49 | 0.66 |
| CoTracker [38]+DA [106] | 0.202 | 34.3 | 57.9 | 24.1 | 33.9 | 73.0 | - | - | - |
| TAPIR [14]+DA [106] | 0.114 | 38.1 | 63.2 | 27.8 | 41.5 | 67.4 | - | - | - |
| Ours | **0.082** | **43.0** | **73.3** | **34.4** | **47.0** | **86.6** | **16.72** | **0.63** | **0.45** |

# Experiment

我们对方法进行了定量和定性评估，涉及长时间3D点跟踪、2D点跟踪以及动态场景的新视图合成任务。主要评估数据集包括真实世界的 iPhone 数据集和合成的 Kubric 数据集。

---

## 4.1 Task Specification

**长时间 3D 点跟踪 (Long-range 3D Tracking)**

主要任务是估计场景中任意像素在长时间内的3D运动轨迹。我们扩展了 RAFT-3D 提出的场景流评估指标：

1. **3D 终点误差 (End-Point Error, EPE)**：

$$\mathrm{EPE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{X}_i^{\mathrm{pred}} - \mathbf{X}_i^{\mathrm{gt}}\|_2 \tag{16}$$

   其中，$\mathbf{X}_i^{\mathrm{pred}}$ 是预测的3D位置，$\mathbf{X}_i^{\mathrm{gt}}$ 是真实的3D位置。

2. **准确点百分比 (Percentage of Accurate Points)**：

$$\delta_x^{3D} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\|\mathbf{X}_i^{\mathrm{pred}} - \mathbf{X}_i^{\mathrm{gt}}\|_2 < x) \tag{17}$$

   常用阈值 $x$ 为 5cm 和 10cm，即 $\delta_{0.05}^{3D}$ 和 $\delta_{0.10}^{3D}$。

---

**长时间 2D 点跟踪 (Long-range 2D Tracking)**

2D点跟踪的性能通过以下指标评估：

1. **平均 Jaccard 指数 (Average Jaccard, AJ)**：计算预测轨迹与真实轨迹的重叠程度。
2. **平均位置准确率 (<δavg)**：衡量位置偏差。
3. **遮挡准确率 (Occlusion Accuracy, OA)**：检测被遮挡点的预测能力。

---

**动态场景新视图合成 (Dynamic Novel View Synthesis)**

新视图合成的质量通过以下指标综合评估：

1. **峰值信噪比 (PSNR)**：衡量预测视图与真实视图的像素级差异。
2. **结构相似性指数 (SSIM)**：评估图像质量的一致性。
3. **感知损失 (LPIPS)**：衡量感知层次的图像差异。

---

## 4.2 Evaluation on iPhone Dataset

iPhone 数据集包含 14 个真实场景序列（200-500 帧），记录了真实动态场景，部分数据提供同步静态相机的新视图合成验证。

**定量结果**

我们对所有任务进行了定量评估，结果显示我们的方法显著优于基线方法。在 3D 和 2D 跟踪任务中，我们的方法在以下指标上表现最优：

1. **3D 终点误差 (EPE)**：我们的方法误差最小。
2. $\delta^{3D}_{0.05}$ **和** $\delta^{3D}_{0.10}$：准确点比例显著提升。

定量结果如表所示：

| Method | 3D Tracking (EPE ↓) | $\delta^{3D}_{0.05}$ ↑ | $\delta^{3D}_{0.10}$ ↑ | 2D Tracking (AJ ↑) | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|---|---|---|---|
| TAPIR + DA | 0.114 | 38.1 | 63.2 | 27.8 | - | - | - |
| HyperNeRF | 0.182 | 28.4 | 45.8 | 10.1 | 15.99 | 0.59 | 0.51 |
| Deformable-3D-GS | 0.151 | 33.4 | 55.3 | 14.0 | 11.92 | 0.49 | 0.66 |
| **Ours** | **0.082** | **43.0** | **73.3** | **34.4** | **16.72** | **0.63** | **0.45** |

**定性结果**

我们的方法不仅在3D点跟踪中表现优异，还生成了高质量的新视图合成。以下是对部分场景的可视化比较：

1. **3D轨迹可视化**：显示给定查询点的预测轨迹，并将其叠加在新视图中。

2. **动态场景新视图**：我们的方法生成的视图清晰且结构一致，而基线方法往往会出现模糊或失真。

---

## 4.3 Evaluation on Kubric Dataset

Kubric 数据集为合成数据集，包含 24 帧短视频场景（每个场景包含 10-20 个刚体物体）。这些场景提供了密集的真实标注，包括深度图、相机参数和跨时间的点对应。

**定量结果**

在 Kubric 数据集上的3D跟踪评估结果如下：

| **Method** | **3D Tracking (EPE ↓)** | $\delta^{3D}_{0.05}$ ↑ | $\delta^{3D}_{0.10}$ ↑ |
|---|---|---|---|
| TAPIR + DA | 0.20 | 34.0 | 56.2 |
| CoTracker + DA | 0.19 | 34.4 | 56.5 |
| **Ours** | **0.16** | **39.8** | **62.2** |

**定性结果**

优化后的运动系数表现出与场景中物体的运动一致的分组特性，进一步验证了我们方法的有效性。

---

## 4.4 Ablation Studies

为了验证方法中各组件的贡献，我们在 iPhone 数据集上进行了消融实验。

**消融设置**

1. **不同运动表示**：
   - **Per-Gaussian Transl.**：使用逐高斯的独立平移运动表示。
   - **Transl. Bases**：仅使用平移基，不考虑SE(3)。

2. **初始化影响**：
    - **No SE(3) Init.**：不使用SE(3)拟合初始化。
3. **监督信号**：
    - **No 2D Tracks**：移除2D轨迹监督信号。

### 定量结果

| Method | SE(3) Motion Basis | 2D Tracks | Initialization | EPE ↓ | $\delta^{3D}_{0.05}$ ↑ | $\delta^{3D}_{0.10}$ ↑ |
|---|---|---|---|---|---|---|
| **Ours (Full)** | ✓ | ✓ | ✓ | **0.082** | **43.0** | **73.3** |
| Transl. Bases | ✓ | ✓ | ✓ | 0.093 | 42.3 | 69.9 |
| Per-Gaussian Transl. | ✗ | ✓ | ✓ | 0.087 | 41.2 | 69.2 |
| No SE(3) Init. | ✓ | ✓ | ✗ | 0.111 | 39.3 | 65.7 |
| No 2D Tracks | ✓ | ✗ | ✓ | 0.141 | 30.4 | 57.8 |

消融结果显示：

- SE(3) 运动基对性能提升至关重要。
- 2D轨迹监督信号显著提高了跟踪精度。