SCIGaussian-D: Dynamic Scene Reconstruction from a Single Snapshot Compressive Image

Yunhao Li[1,2*] Yuze Yang[2*] Yanan Hu[1,2] Xiaoyue Li[1] Yong Tang[2]
Xin Yuan[2†] Peidong Liu[2†]

[1]Zhejiang University [2]Westlake University

liyunhao@westlake.edu.cn

Input SCI Measurement Rendered Multi-view Images

Figure 1. Given a single snapshot compressed image, our method is able to recover the underlying dynamic 3D scene representation. Leveraging the fast deformable radiance field representation of 3D Gaussian Splatting, we can render high-quality images of a dynamic scene from a **single measurement** in **real-time**.

Abstract

In this paper, we explore the potential of snapshot compressive imaging (SCI) for dynamic 3D scene reconstruction from a single temporal compressed image. SCI is a low-cost imaging technique that captures high-dimensional information-such as temporal data—using 2D sensors and coded masks, significantly reducing data bandwidth while offering inherent privacy advantages. While recent advances in Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) have enabled 3D reconstruction from SCI measurements, these methods are fundamentally limited to static scenes and fail to generalize to dynamic content. To address this, we propose SCIGaussian-D, a novel framework that enables dynamic 3D reconstruction from a single SCI image. Our method represents the scene with 3D Gaussians defined in a canonical space and models motion using learnable deformation fields. By incorporating the SCI imaging model into the training loop, SCIGaussian-D directly reconstructs the dynamic 3D scene and recovers the corresponding camera motion from a single SCI. We evaluate our method on both synthetic and real SCI datasets, demonstrating significant improvements in reconstruction quality over existing baselines. Our results establish a new state of the art for dynamic scene reconstruction within the SCI framework, paving the way for practical applications in high-speed imaging and real-time scene rendering.

1. Introduction

Reconstructing dynamic 3D scenes from a single image remains a fundamental and unresolved challenge in computational imaging. To tackle this problem, we propose a practical method for recovering dynamic 3D scene geometry from a single snapshot compressive image. Our approach builds upon the video Snapshot Compressive Imaging (SCI) framework [53], a computational imaging paradigm originally developed for high-speed video acquisition [24]. Traditional high-speed imaging systems often entail substantial hardware costs and storage demands. Motivated by the theory of compressive sensing (CS) [3, 9], video SCI aims to alleviate these limitations. A typical SCI system comprises a hardware encoder and a software decoder. During image acquisition, the encoder employs a sequence of random binary 2D masks to modulate the incoming light over time, resulting in a single compressed measurement. This design enables low-cost 2D sensors such as CCD and CMOS cameras to capture high-speed events efficiently, reducing both acquisition costs and data volume. Subsequently, the decoder reconstructs high frame-rate video frames by leverag-

*Equal contribution.
†Corresponding authors.

ing the encoded measurement along with the corresponding binary masks.

In recent years, significant progress has been made on SCI reconstruction algorithms, spanning model-based techniques [21, 23, 51] and deep learning-based approaches [5, 7, 8, 26, 36, 43, 44]. While these methods demonstrate impressive image and video reconstruction quality, they typically disregard the 3D structure underlying the scene, which is critical for achieving multi-view consistency. To address these limitations, Li et al.[20] proposed SCINeRF, which integrates Neural Radiance Fields (NeRF)[30] into the SCI framework, enabling 3D scene representation recovery within the exposure window. Building on this, SCISplat [19] employs 3D Gaussian Splatting (3DGS) [16] to improve rendering fidelity and computational efficiency. However, both SCINeRF and SCISplat are restricted by the static scene assumptions inherent to NeRF and 3DGS, and thus encounter significant limitations when applied to dynamic 3D scenes captured via SCI.

To overcome these challenges, we propose **SCIGaussian-D**, a novel dynamic 3D scene representation framework for SCI. SCIGaussian-D employs a deformable 3DGS framework, where 3D Gaussians are learned in a canonical space and subsequently conditioned on time through implicit deformation fields. Training SCIGaussian-D on a single SCI measurement involves performing differentiable Gaussian rasterization, followed by simulating the physical SCI imaging model to generate a synthesized SCI measurement. To stabilize the training of SCIGaussian-D, we employ a two-stage optimization strategy. In the first warm-up stage, the camera poses and canonical Gaussians are jointly optimized while deliberately excluding the deformation fields, allowing the model to obtain a stable scene representation and poses. In the second stage, the camera poses are fixed, and both Gaussians and deformation fields are optimized to capture scene motions across different time stamps. With the help of SCIGaussian-D, we can recover dynamic 3D scene structures captured by a fast-moving camera within extremely short exposure durations, e.g., less than 20ms or even less than 10ms.

To comprehensively evaluate the performance of our method, we build a **real hardware platform** to capture SCI images. Additionally, we conduct quantitative evaluations using synthetic datasets generated via Blender. Experimental results on both synthetic and real-world datasets demonstrate that SCIGaussian-D effectively reconstructs dynamic 3D scenes from a single SCI image. In terms of image restoration and rendering quality, our method outperforms previous state-of-the-art (SOTA) methods.

In summary, our key contributions are listed as follows:

- We present a novel method to restore dynamic 3D scenes, with 3D aware multi-view images of the captured scenes, from a single snapshot compressive image.
- We experimentally validate that our approach is able to synthesize high-quality multi-view images from represented dynamic 3D scenes, surpassing existing state-of-the-art SCI image/video reconstruction methods and static NeRF and 3DGS-based methods.
- Our method also presents alternative approach for efficient 3D scene storage/transmission, privacy protection, and enables the practical deployment of SCI systems in real-world scenarios involving the capture of dynamic, high-speed 3D scenes.

2. Related Works

2.1. NeRF and 3DGS

Neural Radiance Fields (NeRF) [30] have significantly advanced novel-view synthesis by enabling high-fidelity 3D scene representation. However, NeRF requires accurate camera poses as input, which are often unavailable or unreliable in real-world datasets. Most NeRF-based pipelines depend on Structure-from-Motion (SfM) tools such as COLMAP [38] to estimate poses, but these can produce noisy results or fail altogether. To address this, several pose-free or jointly optimized variants have emerged. NeRF– [46] estimates camera intrinsics and extrinsics during training; Jeong [15] introduces a self-calibration approach; iMAP [40] integrates NeRF with SLAM to estimate camera poses; and GNeRF [28] leverages GANs [13] for more robust pose and NeRF optimization. Other methods, such as BARF [22] and BAD-NeRF [45], refine poses via coarse-to-fine strategies and simulate motion blur to enhance reconstruction quality. Another key limitation of NeRF is its low computational efficiency. The implicit representation based on multi-layer perceptron (MLP) and volumetric rendering incur high training and inference costs, often requiring hours or days. To improve efficiency, grid-based models like TensoRF [6], Plenoxels [11], and HexPlane [4], as well as hash-based methods such as Instant-NGP [31], have been proposed. Despite these advances, real-time rendering with high quality remains challenging. 3D Gaussian Splatting (3DGS) [16] offers a compelling alternative by replacing implicit fields with an explicit, point-based scene representation. Using a tile-based rasterization of 3D Gaussians, 3DGS achieves real-time rendering with high visual fidelity. However, like NeRF, it requires accurate camera poses and sparse geometry—typically from COLMAP—for initialization, limiting its applicability in pose-unknown settings.

2.2. Snapshot Compressive Imaging

Early SCI image reconstruction methods primarily rely on regularized optimization frameworks [21, 23, 51, 52], where compressed measurements are recovered by solving

inverse problems with handcrafted priors. Common regularizers include sparsity constraints [49] and total variation (TV) [51], often solved using the alternating direction method of multipliers (ADMM) [2] for improved stability and flexibility across imaging setups. Representative methods such as DeSCI [8] and GAP-TV [51] have achieved notable performance improvements, but their iterative nature leads to high computational costs, limiting scalability to high-resolution or real-time applications.

With the rise of deep learning, recent SCI reconstruction methods increasingly leverage neural networks. Architectures such as U-Net [37] and GANs [13] have been employed to learn end-to-end mappings from compressed measurements to full images. Due to the scarcity of large-scale real SCI datasets, models are typically trained on synthetic data with simulated measurements and masks. For example, Qiao et al. [36] proposed an end-to-end CNN (E2E-CNN), while Cheng et al. [7] introduced BIRNAT, a bi-directional RNN-based model for video SCI. To address scalability, RevSCI designed a multi-group reversible 3D CNN to reduce training memory and computation. Hybrid approaches like PnP-FFDNet [52] and FastDVDNet [54] combined learned denoisers with traditional solvers for improved speed and flexibility. More recently, Wang et al. [5, 44] introduced spatial-temporal Transformers (STFormer) and EfficientSCI to better capture correlations across time and space, which lead to SOTA results.

While deep models achieve high-quality reconstructions, they are typically constrained to 2D outputs aligned with specific masks, and often generalize poorly to real-world data due to reliance on synthetic pretraining. To address this, Li et al. [20] proposed SCINeRF, which jointly optimizes NeRF and camera poses to recover the 3D scene from a single SCI image. Building on this, SCISplat [19] employs 3D Gaussian Splatting (3DGS) for improved rendering quality and efficiency. Both methods leverage test-time optimization (TTO) to mitigate generalization issues and enable novel-view synthesis with high temporal resolution. However, they assume static scenes and cannot handle dynamic motion from a single SCI snapshot, limiting their applicability in real-world dynamic settings.

2.3. Dynamic 3D Scene Representation

Conventional NeRF and 3DGS methods assume static scenes during multi-view image capture, which limits their applicability to dynamic environments such as human motion capture and autonomous driving. A central challenge in modeling dynamic 3D scenes lies in incorporating temporal information while maintaining spatiotemporal consistency. Early works extend NeRF by conditioning the radiance field on time. However, naively treating time as an additional input often fails to disentangle spatial deformation from appearance, leading to degraded rendering quality. To address this, later approaches decouple geometry and appearance by learning radiance fields in a static canonical space, paired with an implicit deformation field to model scene motion. D-NeRF [35] introduces a deformation MLP to model temporal displacements of points along rays. HyperNeRF [33] constructs a 5D radiance field, treating each image as a slice through this space. Nerfies [32] improves robustness via elastic regularization of the deformation field. Other methods incorporate motion segmentation [39, 41], depth priors [1], or 2D hex-plane encodings [12] to better capture scene dynamics.

With the rise of 3D Gaussian Splatting (3DGS), attention has shifted toward explicit representations for dynamic scenes. Compared to NeRF's implicit formulation, 3DGS-based methods achieve higher rendering quality and efficiency. Wu et al. [47] combine 4D point clouds with K-Planes [12] for dynamic modeling. Grid4D [48] enhances plane-based representations with hash encoding and attention modules. Duan et al. [10] introduce 4D rotors to capture Gaussian motion, while Luiten et al. [25] model trajectories via per-frame transformations. Although effective for tracking, the latter suffers from rendering artifacts. Inspired by deformation-based NeRF methods, Yang et al. [50] propose a deformable 3DGS framework using an MLP-based deformation field to modulate canonical Gaussians. Building on this implicit-explicit paradigm, our method also leverages deformation fields to estimate motion, while optimizing the Gaussians in canonical space to ensure consistent scene representation across time.

3. Method

In this paper, we consider the input SCI image as the compressed multi-view monocular images of a dynamic scene. Our proposed method takes a single SCI image and encoding masks as input and recovers the underlying dynamic 3D scene structure, as well as camera poses. Then, we can render compressed multi-view images from the reconstructed 3D scene. To achieve this, we first employ an initialization protocol to estimate point clouds and poses from a single SCI measurement to start the training procedure. Then we decouple the scene structure and motion by learning the scene represented by 3D Gaussians in the canonical space, and applying implicit deformation fields to model the scene motion. We follow the image formation model of video SCI to synthesize snapshot compressive images from Gaussians. By maximizing the photometric consistency between the synthesized image and the actual SCI measurement, we optimize Gaussian-based 3D scene representation, camera poses and deformation fields. An overview of our method is shown in Figure 2.

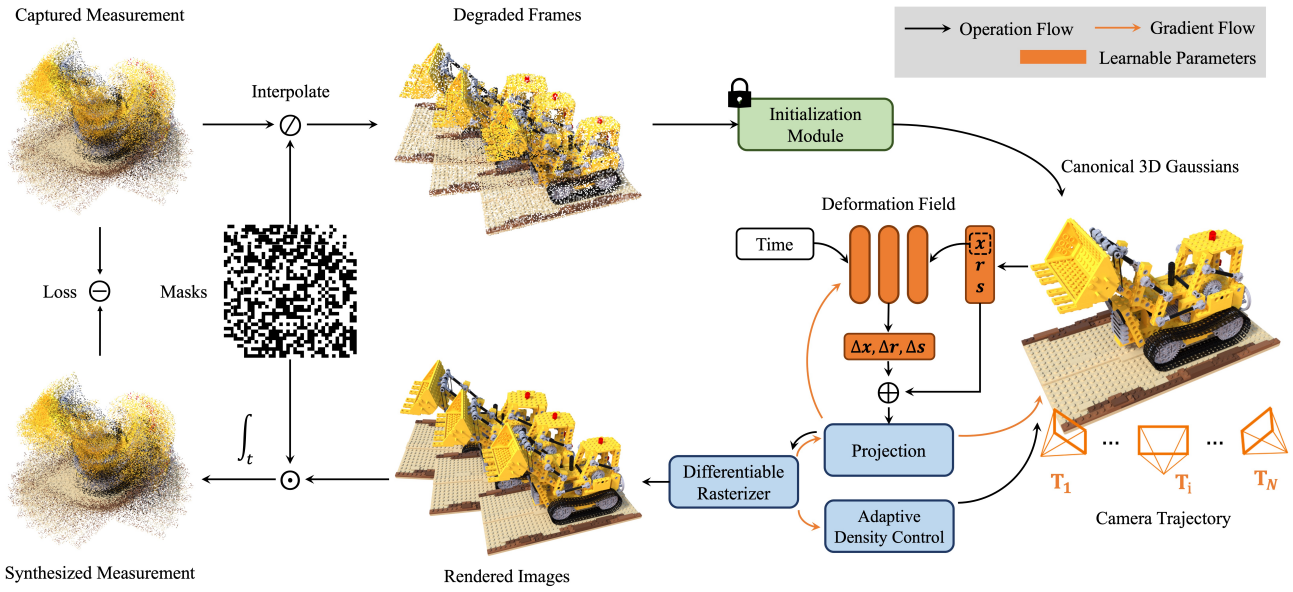

Figure 2. **Overview of the proposed SCIGaussian-D.** Our method takes a single snapshot compressive image and corresponding encoding masks as input, and recovers the underlying dynamic 3D scene representation as well as the camera motion trajectory within a single exposure time. A set of degraded frames are first reconstructed from the input real SCI measurement and modulation masks using pixel interpolation. These frames are then fed into a learning-based SfM module to initialize Gaussian parameters. To decouple the dynamic scene representation, we apply an implicit deformation field apart from canonical 3D Gaussians. When training SCIGaussian-D, we first take a warm-up phase to jointly optimize canonical 3D Gaussians and camera poses without deformation fields. Subsequently, we lock on camera poses and jointly optimize deformation fields and and canonical Gaussians. The canonical Gaussians, deformation fields and camera poses are optimized by minimizing the photometric loss between the synthesized measurement (from the rendered multi-view images via differential Gaussian rasterization) and the captured SCI measurement.

3.1. Image Formation Model of Video SCI

The formation process of a video SCI system is similar to that of a blurry image. The difference is that the captured images $\mathcal{X} = \{\mathbf{X}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N_I}$ of a video SCI system are modulated by N_I binary masks $\mathcal{M} = \{\mathbf{M}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N_I}$ throughout the exposure time, where both H and W are image height and width, respectively. For practical hardware implementation, those masks are achieved by displaying different 2D patterns on the digital micro-mirror device (DMD) or a spatial light modulator, e.g., liquid crystal on silicon. The image sensor then accumulates the modulated photons across exposure time to a compressed/coded image. The number of masks or different patterns on the DMD within the exposure time determines the number of coded frames, i.e. the temporal compression ratio (CR). Due to mask modulation, the N_I virtual images can be recovered from a single compressed image alone by solving an ill-posed inverse problem.

Mathematically, the imaging process can be described as follows:

$$\mathbf{Y} = \sum_{i=1}^{N_I} \mathbf{X_i} \odot \mathbf{M_i} + \mathbf{Z}, \tag{1}$$

where $\mathbf{Y}, \mathbf{X}_i \in \mathbb{R}^{H\times W}$ are the compressed captured image and the i^{th} virtual image within exposure time, respectively, N_I is the temporal CR, $\odot$ denotes element-wise multiplication, and $\mathbf{Z} \in \mathbb{R}^{H\times W}$ is the measurement noise. The individual pixel value in the binary mask is generated randomly. For masks N_I throughout the exposure time, the probability of assigning 1 to the same pixel location is fixed and defined as overlapping ratio.

3.2. 3D Gaussian Splatting

3DGS leverages 3D Gaussian as a more efficient scene representation. A set of 3D Gaussians $G = \{\mathbf{g}_i\}_{i=1}^{M}$, parameterized by their mean position $\mathbf{x}_i \in \mathbb{R}^3$, 3D covariance $\mathbf{\Sigma}_i \in \mathbb{R}^{3\times3}$, opacity $o_i \in \mathbb{R}$, and color $\mathbf{c}_i \in \mathbb{R}^3$, is introduced to faithfully represent the 3D scene. The distribution of each Gaussian $\mathbf{g}_i$ is defined as:

$$\mathbf{g}_i(\mathbf{p}) = \exp\left(-\frac{1}{2}(\mathbf{p}-\mathbf{x}_i)^\top \mathbf{\Sigma}_i^{-1}(\mathbf{p}-\mathbf{x}_i)\right), \tag{2}$$

To ensure that the 3D covariance $\mathbf{\Sigma}_i$ remains positive semi-definite, which is physically meaningful, and to reduce optimization difficulty, 3DGS represents covariance matrix using a rotation quaternion $\mathbf{r}$ and a 3D scaling vector $\mathbf{s}$, which can be transformed into the corresponding scale matrix $\mathbf{S}_i \in \mathbb{R}^{3\times3}$ (diagonal matrix) and rotation matrix

$\mathbf{R}_i \in \mathbb{R}^{3\times3}$:

$$\mathbf{\Sigma}_i = \mathbf{R}_i \mathbf{S}_i \mathbf{S}_i^\top \mathbf{R}_i^\top. \tag{3}$$

The input to 3DGS also consists of a series of multi-view images $\mathcal{X} = \{\mathbf{X}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N}$ of the target 3D scene, along with their corresponding intrinsic and camera poses. Additionally, at the beginning of 3DGS training, a sparse point cloud $\mathcal{Q} = \{\mathbf{Q}_i \in \mathbb{R}^3\}_{i=1}^{M}$ of the target 3D scene, typically obtained through Structure-from-Motion (SfM) software COLMAP, is used to initialize the Gaussians g.

To render multi-view images, 3DGS employs a differentiable Gaussian rasterization process. In this process, 3D Gaussians are projected onto the 2D image plane based on a given camera pose $\mathbf{T}_i = \{\mathbf{R}_{c,i}, \mathbf{t}_{c,i}\}$, where $\mathbf{R}_{c,i} \in \mathbb{R}^{3\times3}$ is the camera rotation matrix and $\mathbf{t}_{c,i} \in \mathbb{R}^3$ is the translation vector. The projection is described by the following equation:

$$\mathbf{\Sigma}_i' = \mathbf{J}_i \mathbf{R}_{c,i} \mathbf{\Sigma}_i \mathbf{R}_{c,i}^\top \mathbf{J}_i^\top,$$

where $\mathbf{\Sigma}_i' \in \mathbb{R}^{2\times2}$ is the 2D covariance matrix, $\mathbf{J}_i \in \mathbb{R}^{2\times3}$ is the Jacobian matrix of the affine approximation of the projective transformation.

Then, the image pixels are rendered by rasterizing these sorted 2D Gaussians based on their depths:

$$C = \textstyle\sum_{i=1}^{M} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{4}$$

where C represents the pixel in the rendered multi-view images, and α_i denotes the alpha value computed by evaluating the 2D covariance $\mathbf{\Sigma}_i^{'}$ multiplied with the learnable Gaussian opacity o_i:

$$\alpha_i = o_i \exp(-\sigma_i), \quad \sigma_i = \frac{1}{2}\Delta_i^\top \mathbf{\Sigma}_i'^{-1} \Delta_i, \tag{5}$$

where $\Delta_i \in \mathbb{R}^2$ is the offset between the pixel center and the projected 2D Gaussian center. Finally, the 3D Gaussians g are updated by minimizing the photometric loss computed between rendered images and real captured images.

3.3. Proposed Framework

To initialize the training of Gaussians, we need both camera poses and sparse point clouds as a coarse estimation of 3D scenes. Since the SCI measurement contains compressed frames in the form of a single 2D image, it is unrealistic to estimate camera poses and point clouds using COLMAP, as in most of the 3DGS-based methods.

To deal with the initialization problem, in this paper, we employ the pre-trained SfM models to initialize the Gaussians. Specifically, inspired by Wang et al. [44] and SCISplat [19], we first normalize the real SCI measurement $\mathbf{Y}$ using the sum of modulation masks $\mathbf{M}_i$.

$$\overline{\mathbf{Y}} = \mathbf{Y} \oslash \sum_{i=1}^{N} \mathbf{M_i}, \tag{6}$$

where $\overline{\mathbf{Y}}$ is the normalized measurement, and $\oslash$ denotes element-wise division. Then, degraded frames $\widetilde{\mathbf{I}} = \{\widetilde{\mathbf{I}}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N_I}$ can be obtained by interpolating the normalized measurement after modulated by a filtered version of each mask $\mathbf{C}_i \odot \mathbf{B}_i$,

$$\widetilde{\mathbf{I}}_i = \text{Interp}\left(\overline{\mathbf{Y}} \odot (\mathbf{M}_i \odot \mathbf{B}_i)\right), \tag{7}$$

$$(\mathbf{B}_i)_{j,k} = \begin{cases} 1, & \text{if } (\mathbf{M}_i)_{j,k} \geq \eta \\ 0, & \text{otherwise} \end{cases}, \tag{8}$$

where $\mathbf{B}_i$ is a selection matrix that only preserve the value of $\mathbf{C}_i$ positioned at (j,k) if its value exceeds η. For synthetic data $\eta = 1$, since modulation masks are binary, with only 0 and 1 values. For real data, we follow the SCISplat [19] to set η as 0.8. These degraded frames contain significant amount of noise, thus cannot be processed by COLMAP. However, the recent progress on deep learning-based MVS methods provide decent guesses from these noise frames. Specifically, we use VGGSfM [42], one of the current state-of-the-art deep learning-based 3D scene reconstruction frameworks, to estimate camera poses $\mathbf{T}$ and point cloud $\mathcal{Q}$, i.e.,

$$\mathbf{T}, \mathcal{Q} = f_\theta(\widetilde{\mathbf{I}}), \tag{9}$$

After initialization, we conduct a two-stage training strategy to train our SCIGaussian-D framework. In the first stage, we bypass the deformation fields and perform the joint optimization strategy applied by Li et al. [20], where the canonical 3D Gaussians and camera poses are jointly optimized. This warm-up stage is done to optimize the camera poses using the static parts (usually backgrounds) while also provide a good initial value in the next stage of training.

In second stage, we keep the optimized poses locked and let the model focus on optimizing the deformation field. The deformation field contains an MLP, which takes the center positions of the 3D Gaussians $\mathbf{x}$ and current time t as input and transforms the canonical Gaussians in the deformed space:

$$(\delta\mathbf{x}, \delta\mathbf{r}, \delta\mathbf{s}) = \text{MLP}_{\text{deform}}(\gamma(sg(\mathbf{x})), \gamma(t)), \tag{10}$$

where $\text{MLP}_{\text{deform}}$ denotes deformation MLP, $sg(\cdot)$ indicates the stop-gradient operation, and γ represents positional encoding. Then the deformed Gaussian of current time t can be obtained by adding deformation to the canonical Gaussian:

$$G(t) = G(\mathbf{x} + \delta\mathbf{x}, \mathbf{r} + \delta\mathbf{r}, \mathbf{s} + \delta\mathbf{s}), \tag{11}$$

where $G(t)$ denotes deformed Gaussians at time stamp t. With deformed Gaussians at different time stamps, we can render multi-view dynamic images $\widehat{\mathcal{X}} = \{\widehat{\mathbf{X}}_i \in \mathbb{R}^{H\times W}\}_{i=1}^{N_I}$ by differential Gaussian rasterization procedure described in Eq. 4.

Conventional 3DGS-based methods compute loss on multi-view images, while we only have one SCI image as input. To train our framework on a single SCI image, we follow the physical image formation model of SCI as described in Eq. 1 to transfer the rendered multi-view images from differentiable Gaussian rasterization process into a synthesized SCI measurement:

$$\widehat{\mathbf{Y}} = \sum_{i=1}^{N_I} \widehat{\mathbf{X}}_i \odot \mathbf{M}_i, \tag{12}$$

where $\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W}$ represents the synthesized SCI measurement, and $\widehat{\mathbf{X}}_i$ is the i-th rendered image. Here, we omit the measurement noise term $\mathbf{Z}$ in Eq. 1 to facilitate the recovery of the originally captured image. Finally, we compute and back-propagate the photometric loss between the synthesized SCI measurement and real SCI measurement as in original 3DGS framework:

$$\mathcal{L}_{\text{photo}} = (1-\lambda) \cdot \mathcal{L}_1(\widehat{\mathbf{Y}}, \mathbf{Y}) + \lambda \cdot \mathcal{L}_{D-SSIM}(\widehat{\mathbf{Y}}, \mathbf{Y}), \tag{13}$$

Additionally, we introduce two regularization terms, including a scale term $\mathcal{L}_{\text{s}}$ and an opacity term $\mathcal{L}_{\text{o}}$. These two terms minimize the scale and opacity of the current Gaussians to encourage a lower number of effective Gaussians. Therefore, the complete loss function $\mathcal{L}$ includes the photometric loss in Eq. 13 and two regularization terms:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \lambda_{\text{o}} \cdot \mathcal{L}_{\text{o}}(G) + \lambda_{\text{s}} \cdot \mathcal{L}_{\text{s}}(G). \tag{14}$$

During 3DGS training, Gaussian primitives are progressively refined to capture scene details. The original adaptive density control (ADC) mechanism [16] introduces new Gaussians in regions with high reconstruction error. However, ADC often increases opacity after cloning or splitting Gaussians, leading to brightness inconsistencies and unstable gradients from the photometric loss. This instability can cause pose drift and reconstruction collapse, particularly under the ambiguous pixel mappings in SCI measurements. To address this, we adopt an MCMC-based strategy [17], which recalculates opacity after densification to avoid abrupt changes. This results in more stable optimization dynamics and mitigates noise artifacts caused by local optima in the ill-posed SCI imaging process.

4. Experiment

4.1. Datasets

We generate synthetic datasets using the software Blender. Each scene in synthetic datasets contains dynamic objects with a static background. These scenes come from DeblurNeRF [27], HDR-NeRF [14] and D-NeRF [35] datasets. There are five virtual scenes in the synthetic datasets, including *Clock, Lego, Jump, Punch* and *Tank*. For these Blender-generated datasets, we use 600×400 resolution. The compression ratio of the dataset is 8. In this paper, we refer to the ablation study conducted by Li et al. [19] and set the mask overlapping rate to be 0.25. Furthermore, we also tested our proposed method on static SCINeRF synthetic datasets, which include 6 scenes *Airplants, Hotdog, Cozy2room, Tanbata, Factory, Vender* derived from LLFF [29], NeRF Synthetic 360 [30], and DeblurNeRF [27] datasets. For real dataset, we collect the SCI measurement from a real video SCI setup. The setup consists of an iRAYPLE A5402MU90 camera and a FLDISCOVERY F4110 DMD. The compression ratio of the real datasets is 8, with the resolution of SCI measurement as 1024×750.

4.2. Baseline methods and evaluation metrics.

Since our SCIGaussian-D can render high-quality images from the represented dynamic 3D scene, we compare our method with prior SOTA SCI image reconstruction methods, including GAP-TV [51], PnP-FFDNet [52], PnP-FastDVDNet [54], and EfficientSCI [44]. We also compared our methods with existing NeRF and 3DGS-based SCI 3D reconstruction methods, including SCINeRF [20] and SCISplat [19]. For fair comparisons, we fine-tuned EfficientSCI using our datasets. For evaluation metrics, we adopt widely-used structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and learned perceptual image patch similarity (LPIPS) [55].

4.3. Implementation details.

We implement our framework using PyTorch [34] on a single NVIDIA RTX4090 GPU. The training process, including first stage (i.e., warm-up stage) and the second stage, takes around 20K iterations. For optimization, we apply individual Adam [18] optimizers for pose optimization, canonical Gaussian and deformation field optimization. The pose learning rate decays from 1×10^{-3} to 1×10^{-5}, and the learning rate of deformation MLP decays from 1×10^{-3} to 1.6×10^{-5}. The learning rate of canonical Gaussians are the same as original 3DGS paper [16]. Adam's β value range is set to (0.9, 0.999). When initializing the Gaussians, the number of points are downsampled to 10,000. Both of the additional regularization terms λ_o and λ_s are set to be 0.01.

4.4. Results.

The experimental results on the synthetic dataset provides empirical evidence on the efficacy of our SCIGaussian-D in estimating and representing high-quality dynamic 3D scenes from a single SCI measurement, as shown in Figure 3 and Table 1. We notice that compared to static 3D scene representation algorithms such as SCINeRF and SCISplat, our method exhibits superior performance by better recovering dynamic part of the scene. When comparing with con-

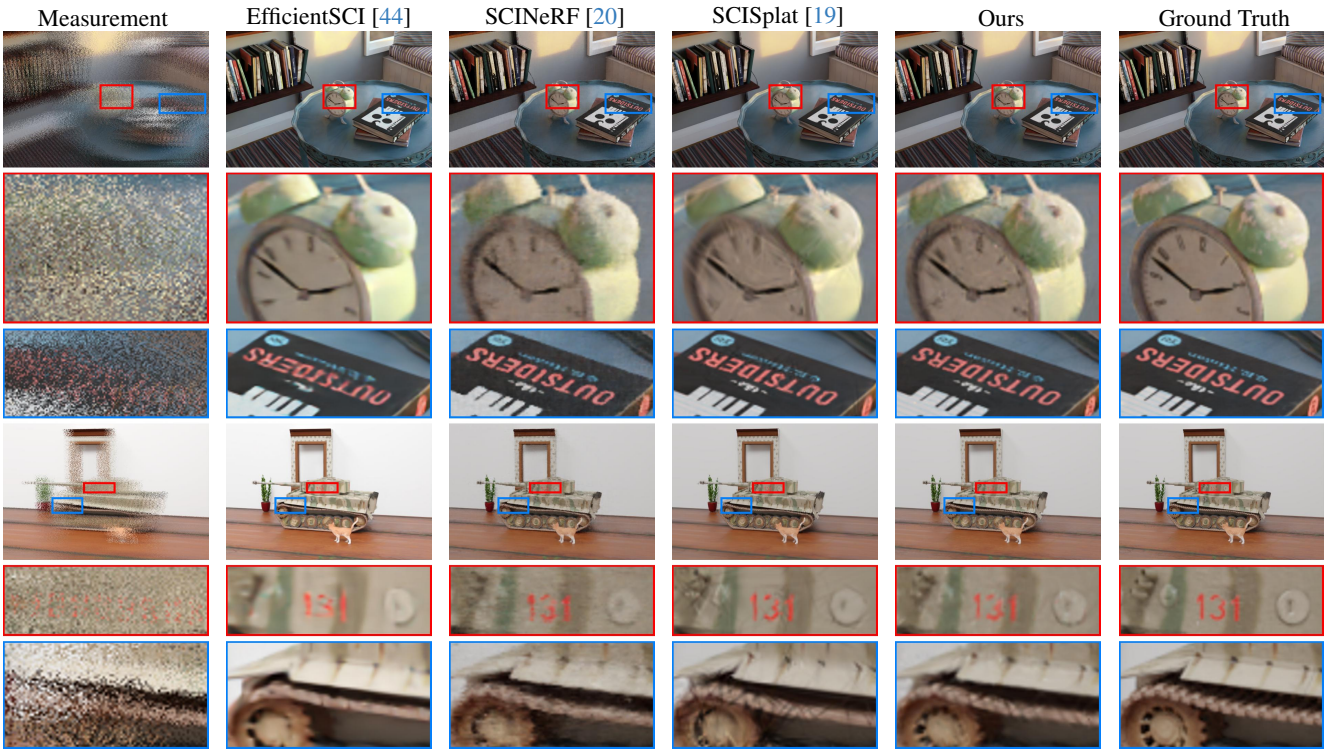

Figure 3. **Qualitative evaluations of our method against SOTA SCI image restoration methods on synthetic datasets** Top to bottom shows the results for different scenes, including *Clock* and *Tank*. The experimental results demonstrate that our method achieves superior performance on image restoration from a single compressed image (the far-left column). Our method effectively restores the dynamic part of the scenes while maintaining high rendering quality in static regions of the scenes.

	Clock			Lego			Jump			Punch			Tank		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAP-TV [51]	18.74	0.435	0.568	17.71	0.518	0.581	19.59	0.436	0.611	20.55	0.437	0.490	22.62	0.650	0.397
PnP-FFDNet [52]	26.65	0.887	0.083	26.02	0.942	0.120	30.38	0.962	0.146	28.65	0.885	0.143	29.51	0.913	0.085
PnP-FastDVDNet [54]	26.76	0.884	0.082	26.21	0.956	0.111	31.72	0.962	0.117	30.72	0.916	0.099	31.26	0.929	0.073
EfficientSCI [44]	32.35	0.943	0.036	30.96	0.972	0.047	36.43	0.986	0.046	32.20	0.942	0.065	33.16	0.958	0.051
SCINeRF [20]	31.99	0.936	0.051	29.56	0.965	0.069	33.46	0.956	0.096	31.55	0.940	0.083	32.69	0.940	0.071
SCISplat [19]	32.40	0.948	0.023	31.11	0.978	0.025	36.23	0.955	0.023	32.25	0.942	0.082	32.45	0.965	0.045
ours	**34.76**	**0.970**	**0.014**	**33.73**	**0.987**	**0.015**	**37.24**	**0.991**	**0.013**	**33.74**	**0.962**	**0.024**	**34.66**	**0.980**	**0.022**

Table 1. **Quantitative SCI image reconstruction comparisons on the synthetic datasets** The results are computed from the rendered images from estimated scenes via our proposed method, recovered images from SOTA SCI image restoration methods, and rendered images from prior SOTA static 3D scene representation methods SCINeRF and SCISplat. The experimental results demonstrate that our proposed method can render images with higher quality than those from existing methods.

ventional 2D SCI image recovery approaches, our method performs higher in scenes with rich textures and characters, where these existing methods fail to retrieve these details as shown in Figure 3. For static SCINeRF datasets, our method realizes comparable performance with SCISplat (please see supplementary materials for more details). The comparative results on both static and dynamic synthetic datasets prove that our method realizes superior performance against prior SCI image recovery methods such as SCINeRF and SCISplat on dynamic regions of the scenes, and preserves high reconstruction quality for static regions of the scenes.

Furthermore, we evaluate the computational efficiency of various SCI reconstruction algorithms by comparing their training times (if applicable) and inference speeds. For SOTA methods that output 2D frames directly from SCI images, we compare their inference speed with the rendering speed (in FPS) of our approach, which first reconstructs a 3D scene and then renders 2D images. Our method achieves less than 2.5 hours for training with 56 FPS on rendering speed, which remains real-time rendering capabilities. More details are available in supplementary materials.

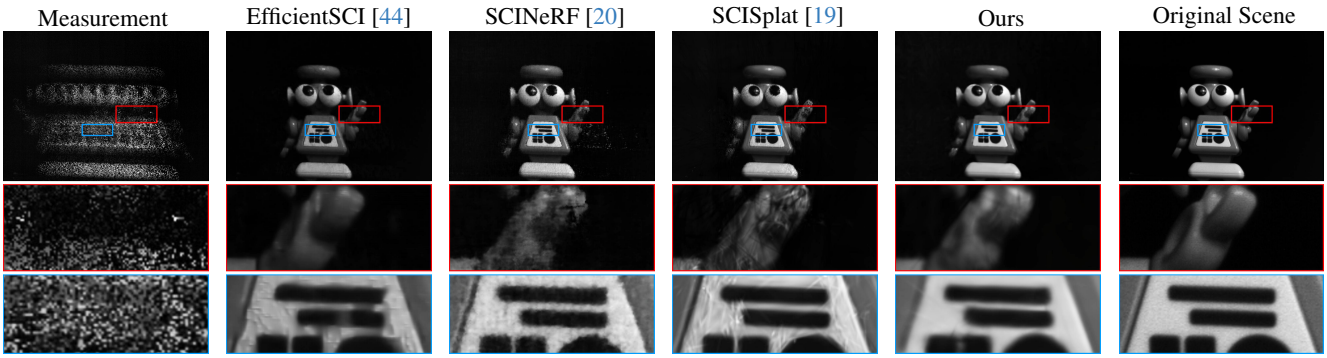

Figure 4. **Qualitative evaluations of our method against SOTA SCI image restoration methods on the real dataset captured by our system.** After capturing the snapshot compressed measurement, we capture separate scene images used for reference since the compressed ground truth images in real datasets are unavailable. The results demonstrate that our proposed method surpasses existing image restoration methods on real datasets by effectively retrieving dynamic regions of the scene and avoiding introducing noise defects which appear in existing methods.

Model	Synthetic Dataset		
	PSNR↑	SSIM↑	LPIPS↓
Ours w/o PO	33.57	0.951	0.033
Ours w/o MCMC	32.16	0.950	0.035
Ours w/o $\mathcal{L}_o, \mathcal{L}_s$	32.85	0.940	0.049
Ours (full)	**34.83**	**0.978**	**0.017**

Table 2. **Quantitative ablation studies results on the synthetic dataset.** The results demonstrate that the effectiveness of pose optimization (PO), MCMC strategy and additional loss regularization terms in our framework.

To evaluate the performance of SCIGaussian-D on real datasets, we also conduct qualitative comparisons against SOTA methods. Figure 4 illustrates the experimental results, depicting the outcomes for real datasets. Notably, existing SOTA techniques either fail to model the dynamic parts of the scenes, or introduce additional noisy artifacts in returned images. In contrast, our proposed method surpasses these methods on real datasets by effectively recovering scenes with fine details, thereby achieving superior performance.

4.5. Ablation Study

To prove the effectiveness of the components of our SCIGaussian-D, including pose optimization (PO), MCMC strategy, and additional loss regularization terms $\mathcal{L}_o$ and $\mathcal{L}_s$ presented in Eq. 14, we conducted an ablation study of our model on the synthetic datasets, as is shown in Table 2. It has been verified that PO, MCMC and loss regularization terms are effective and all of them contribute to improving the quality of reconstructed dynamic 3D scenes.

5. Conclusion

In this work, we proposed SCIGaussian-D, a novel framework for dynamic 3D scene reconstruction from a single snapshot compressive image. By decoupling the scene into canonical 3D Gaussians and an implicit deformation field, our method effectively models both scene structure and motion. The physical image formation process of SCI is integrated into the training objective, and a two-stage optimization strategy is adopted—initially optimizing both Gaussians and camera poses, followed by deformation field learning with fixed poses—to stabilize training. Extensive experiments on synthetic and real-world datasets demonstrate that SCIGaussian-D achieves superior reconstruction quality compared to existing methods, enabling high-fidelity recovery of dynamic 3D scenes from a single SCI image. This opens the door for practical deployment of SCI systems in real-world scenarios such as autonomous driving, VR/AR, and high-speed imaging. Future work includes extending our framework to capture complex dynamic phenomena, including physical and chemical reactions.

Acknowledgements. This work was supported in part by NSFC under Grants 62202389 and 62271414, in part by a grant from the Westlake University-Muyuan Joint Research Institute, and in part by the Westlake Education Foundation, Science Fund for Distinguished Young Scholars of Zhejiang Province (LR23F010001), Research Center for Industries of the Future (RCIF) at Westlake University and and the Key Project of Westlake Institute for Optoelectronics (Grant No. 2023GD007).

References

[1] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew

O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in neural information processing systems*, 34:26289–26301, 2021. 3

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3 (1):1–122, 2011. 3

[3] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006. 1

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2

[5] Miao Cao, Lishun Wang, Mingyu Zhu, and Xin Yuan. Hybrid cnn-transformer architecture for efficient large-scale video snapshot compressive imaging. *International Journal of Computer Vision*, pages 1–20, 2024. 2, 3

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial Radiance Fields. In *ECCV*, pages 333–350. Springer, 2022. 2

[7] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan. Birnat: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In *European Conference on Computer Vision*, pages 258–275. Springer, 2020. 2, 3

[8] Ziheng Cheng, Bo Chen, Guanliang Liu, Hao Zhang, Ruiying Lu, Zhengjue Wang, and Xin Yuan. Memory-efficient network for large-scale video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16246–16255, 2021. 2, 3

[9] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006. 1

[10] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. In *CVPR*, pages 5501–5510, 2022. 2

[12] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 3

[14] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022. 6

[15] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 6

[17] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Jeff Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *arXiv preprint arXiv:2404.09591*, 2024. 6

[18] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Yunhao Li, Xiang Liu, Xiaodong Wang, Xin Yuan, and Peidong Liu. Learning radiance fields from a single snapshot compressive image. *arXiv preprint arXiv:2412.19483*, 2024. 2, 3, 5, 6, 7, 8, 1, 4

[20] Yunhao Li, Xiaodong Wang, Ping Wang, Xin Yuan, and Peidong Liu. Scinerf: Neural radiance fields from a snapshot compressive image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10542–10552, 2024. 2, 3, 5, 6, 7, 8, 1, 4

[21] Xuejun Liao, Hui Li, and Lawrence Carin. Generalized alternating projection for weighted-2,1 minimization with applications to model-based compressive sensing. *SIAM Journal on Imaging Sciences*, 7(2):797–823, 2014. 2

[22] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 2

[23] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2990–3006, 2018. 2

[24] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. *Optics express*, 21(9):10526–10545, 2013. 1

[25] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024. 3

[26] Jiawei Ma, Xiao-Yang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10223–10232, 2019. 2

[27] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12861–12870, 2022. 6

[28] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 6351–6361, 2021. 2

[29] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 6

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 6

[31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. 41(4):102:1–102:15, 2022. 2

[32] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5865–5874, 2021. 3

[33] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[35] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 3, 6

[36] Mu Qiao, Ziyi Meng, Jiawei Ma, and Xin Yuan. Deep learning for video compressive sensing. *Apl Photonics*, 5(3), 2020. 2, 3

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[38] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2

[39] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[40] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 2

[41] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 3

[42] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21686–21697, 2024. 5

[43] Lishun Wang, Miao Cao, Yong Zhong, and Xin Yuan. Spatial-temporal transformer for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[44] Lishun Wang, Miao Cao, and Xin Yuan. Efficientsci: Densely connected network with space-time factorization for large-scale video snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18477–18486, 2023. 2, 3, 5, 6, 7, 8, 1, 4

[45] Peng Wang, Lingzhe Zhao, Ruijie Ma, and Peidong Liu. Bad-nerf: Bundle adjusted deblur neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4170–4179, 2023. 2

[46] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[47] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 3

[48] Jiawei Xu, Zexin Fan, Jian Yang, and Jin Xie. Grid4d: 4d decomposed hash encoding for high-fidelity dynamic gaussian splatting. *arXiv preprint arXiv:2410.20815*, 2024. 3

[49] Peihao Yang, Linghe Kong, Xiao-Yang Liu, Xin Yuan, and Guihai Chen. Shearlet enhanced snapshot compressive imaging. *IEEE Transactions on Image Processing*, 29:6466–6481, 2020. 3

[50] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. 3, 1

[51] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International conference on image processing (ICIP)*, pages 2539–2543. IEEE, 2016. 2, 3, 6, 7, 1

[52] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447–1457, 2020. 2, 3, 6, 7, 1

[53] Xin Yuan, David J Brady, and Aggelos K Katsaggelos. Snapshot compressive imaging: Theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 38(2):65–88, 2021. 1

[54] Xin Yuan, Yang Liu, Jinli Suo, Fredo Durand, and Qionghai Dai. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7093–7111, 2021. 3, 6, 7, 1

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 586–595, 2018. 6