SCIGS: 3D Gaussians Splatting from a Snapshot Compressive Image

Zixu Wang[1,2,3,*]   Hao Yang[1,2,4,*]   Yu Guo[1,2,4,†]   Fei Wang[1,2,4]

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
[2]National Engineering Research Center for Visual Information and Applications
[3]School of Software Engineering, Xi'an Jiaotong University
[4]Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

Input SCI Image   Rendered Novel View Images

Figure 1. Given a single compressed image of a dynamic scene as input, the proposed SCIGS can reconstruct a high-quality dynamic 3D scene and recover multi-view consistent images.

Abstract

Snapshot Compressive Imaging (SCI) offers a possibility for capturing information in high-speed dynamic scenes, requiring efficient reconstruction method to recover scene information. Despite promising results, current deep learning-based and NeRF-based reconstruction methods face challenges: 1) deep learning-based reconstruction methods struggle to maintain 3D structural consistency within scenes, and 2) NeRF-based reconstruction methods still face limitations in handling dynamic scenes. To address these challenges, we propose SCIGS, a variant of 3DGS, and develop a primitive-level transformation network that utilizes camera pose stamps and Gaussian primitive coordinates as embedding vectors. This approach resolves the necessity of camera pose in vanilla 3DGS and enhances multi-view 3D structural consistency in dynamic scenes by utilizing transformed primitives. Additionally, a high-frequency filter is introduced to eliminate the artifacts generated during the transformation. The proposed SCIGS is the first to reconstruct a 3D explicit scene from a single compressed image, extending its application to dynamic 3D scenes. Experiments on both static and dynamic scenes demonstrate that SCIGS not only enhances SCI decoding but also outperforms current state-of-the-art methods in reconstructing dynamic 3D scenes from a single compressed image. The code will be made available upon publication.

* Equal contribution   † Corresponding author

1. Introduction

High-speed imaging techniques are widely used in science research, sports, aerospace, etc. However, conventional high-speed imaging techniques typically entail significant expenditure on hardware and substantial storage requirements. Facing these challenges, Compressed Sensing (CS)[3, 8] and video Snapshot Compressive Imaging (SCI)[42] technology has been developed. A SCI system usually has two components: a hardware encoder and a software decoder. During an exposure time, the hardware encoder uses multiple designed masks to divide an exposure process into multiple frames and modulate them into a compressed image. The software decoder can then use the masks to decode the compressed image into high frame rate images. This makes it possible for capturing high-speed video with ordinary cameras, which can reduce the hardware costs and storage costs.

For the hardware encoder of SCI systems, several mature approaches [17, 18] are proposed, though the decoding part still faces challenges. Existing decoding methods are categorized into model-based methods [15, 36, 40] and deep learning-based methods [5, 6, 21, 25, 30, 31]. Model-based methods use iterative optimization based on natural image priors, offering flexibility across resolutions and compression rates, but they suffer from the long processing time cost and low output quality. In contrast, deep learning-based methods leverage network architectures for end-to-end

decoding of compressed images, achieving better real-time performance and higher image quality. However, both methods neglect the structure of the 3D scene, leading to inconsistencies across views. To address this, SCINeRF [14] recovers 3D NeRF representations from a single compressed image by jointly optimizing NeRF and camera poses, yielding promising results. Yet, it performs suboptimally in dynamic scenes, common in high-speed photography. Additionally, NeRF-based reconstruction, with its large number of parameters in MLP, CNN, and Transformer architectures, demands significant training time and memory.

In view of the fact that existing NeRF-based 3D reconstruction methods are incompetent for dynamic scene reconstruction and deep learning-based reconstruction methods struggle to maintain 3D structural consistency within scenes, inspired by the impressive representation capabilities and flexibility of 3DGS, to adapt dynamic scene, we propose SCIGS, the first method to construct a 3D explicit scene from a single compressed image and further extend this to dynamic 3D scenes. Due to the inability to extract the initialization of the camera pose and Gaussians from the compressed image, and constrained by the discrete nature of 3D Gaussians, it is challenging to optimize the camera poses and 3D gaussians simultaneously. to address this issue, a transformation network is proposed, which not only can decouple the transformation field from compressed image for adapting to the dynamic scenes, but also provide a solution for the oscillations of camera poses during optimization. A high-frequency filter is subsequently applied to suppress artifacts generated during transformation. Extensive experiments both on static and dynamic scenes demonstrate that the proposed method achieves superior image quality in SCI decoding tasks and outperforms other methods for 3D scene reconstruction from compressed images, particularly in dynamic scenarios.

Our main contributions can be summarized as follows:

- The proposed SCIGS is the first to recover explicit 3D representations from a single snapshot compressed image within the 3D Gaussian Splatting (3DGS) framework.
- Introducing camera pose stamps and a Gaussian primitive-level transformation network, we substitute the optimization of camera poses with a transformation of Gaussians, tackling the oscillations of camera poses during optimization equivalently.
- Extensive experiments demonstrate that SCIGS synthesizes high-quality novel-view images in both static and dynamic scenes, surpassing existing SCI image reconstruction methods.

2. Related Work

In this paper, two main fields of the prior works are reviewed: the SCI image decoding methods and the 3D reconstruction methods.

SCI image decoding. Decoding a SCI image back to the original images is an ill-posed problem, to solve which, various types of priori knowledge are utilized in traditional methods, including total variation (TV)[40], low-rank prior[7, 17, 20], over-complete dictionary[11, 16], Gaussian mixture model (GMM)[36, 37], etc. These traditional methods are flexible in different scenarios, but has high computational cost and poor reconstruction quality.

With the development of deep learning, deep denoising networks are recognized as an effective image prior, based on which various plug-and-play (PnP) methods[26, 29, 34, 41] are proposed, which are able to achieve excellent performance while maintaining the flexibility of traditional methods. In addition, many end-to-end methods based on deep learning are also proposed, and these SCI decoders employ various network architectures[4, 6, 31] to extract the information in compressed image, including CNN[25], RNN[5], Tranformer[9, 30], etc. To address the computational cost of the current mainstream methods based on Transformer, EfficientSCI[4, 31] introduces hierarchical dense connections in residual block. Moreover, existing SCI image decoding methods do not consider the structure of the underlying 3D scene during reconstruction, resulting in the lack of multi-view consistency.

3D Reconstruction. Current mainstream works on 3D reconstruction mainly include NeRF-based reconstruction and 3DGS-based reconstruction.

Mildenhall et al. proposed Neural Radiance Fields (NeRFs) [23], which employs a Multi-Layer Perceptron (MLP) to implicitly learn 3D scene and renders images by Ray-Marching, has shown excellent performance compared to the previous methods, but its use of MLP leads to high train cost. Many NeRF-based variants have since emerged, such as Mip-NeRF360 [1], which balances image quality and rendering speed. Other works focused on camera-free methods, like NeRF-- [32], which jointly estimates scene and camera poses, and SCINeRF [14], which optimizes NeRF parameters and camera poses in static scenes. However, SCINeRF struggles with dynamic scene reconstruction due to NeRF's limitations.

On the other hand, 3DGS [13] uses 3D Gaussians for scene reconstruction, offering faster rendering and high image quality. Variants like 2DGS [12] and Mip-Splatting [39] improve multi-view consistency and reduce rendering artifacts. Methods for handling scenarios without camera poses or in dynamic scenes have

also been proposed, such as COLMAP-Free 3DGS [10] and GS-SLAM [35], which optimize camera poses and Gaussians iteratively. However, these methods struggle with large discrepancies between initial and target camera poses. iComMa [28] addresses this with a matching loss for optimization guidance, and 6DGS [2] estimates camera poses by reversing the 3DGS rendering process. For dynamic scenes, methods like Deformable 3DGS [38] introduce deformation fields, while 4D Gaussian Splatting [33] extends this idea with neural voxel encoding to handle dynamic scene changes.

3. Method

The pipeline of the proposed method is shown in Fig. 2. The input to the proposed method is a single compressed image and a set of masks. From a random initial point cloud, a set of initial 3D Gaussians $\mathcal{G}\left(\mu, r, s, \sigma\right)$ are created, which are defined by position μ, opacity σ, and a 3D covariance matrix Σ derived from quaternions r and scaling vectors s. Then a fixed viewpoint camera is defined by the random external parameters and the given internal parameters. The appearance of the Gaussians at each viewpoint is represented by spherical harmonics (SH). In order to substitute camera pose transformation by the camera-pose-aware transformation of the 3D Gaussians and to adapt to the dynamic scene, a transformation network $\mathcal{F}$ is introduced, which takes the positions of each 3D Gaussians and a camera pose stamp as inputs, outputs transformation of Gaussians. To eliminate the high-frequency artifacts generated during the transformation, a high-frequency filter follows, before the differential Gaussian rasterization pipeline that outputs intermediate frame images. Afterwards, simulating the modulation process of SCI system, the intermediate frame images are modulated into compressed images. Along with the adaptive density control of the Gaussians, the 3D Gaussians and the transformation network are simultaneously optimized by fast back-propagation.

3.1. 3D Gaussian Splatting

In this paper, the efficient differentiable rasterization pipeline proposed in [13] is employed to render images from 3D Gaussians. As the rendering primitive for 3DGS, a 3D Gaussian is defined as:

$$\mathcal{G}\left(\boldsymbol{x}\right) = \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right), \quad (1)$$

where Σ is parameterized as a combination of quaternion r and a 3D scaling vector s, defined as:

$$\Sigma = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^{\top}\boldsymbol{R}^{\top}. \quad (2)$$

In order to render 3D Gaussians to 2D images, it is necessary to project the 3D Gaussians onto a 2D imaging plane. The covariance of the resulting 2D Gaussians after projection can be approximated as:

$$\Sigma' = \boldsymbol{J}\boldsymbol{W}\Sigma\boldsymbol{W}^{\top}\boldsymbol{J}^{\top}, \quad (3)$$

where $\boldsymbol{J}$ denotes the Jacobian matrix for the affine approximation of the projection transformation, $\boldsymbol{W}$ stands for the view matrix transitioning from world coordinates to camera coordinates.

Subsequently, during the rendering process, the color of given pixel p can be calculated through the alpha blending of N ordered 2D Gaussians:

$$
\begin{aligned}
C(p) &= \sum_{i \in N} T_i \alpha_i c_i, \\
\alpha_i &= \sigma_i e^{-\frac{1}{2}(p-\mu_i)^{\top}\Sigma'(p-\mu_i)}, \\
T_i &= \prod_{j=1}^{i-1}\left(1 - \alpha_j\right),
\end{aligned}
\quad (4)
$$

where c_i represents the Gaussian color derived from the spherical harmonic coefficients.

3.2. Snapshot Compressive Imaging Model

In the process of capturing compressed images in the SCI system, an exposure time is divided into B time intervals by the corresponding B encoding masks. Within each time interval, each pixel's exposure is determined by the value at the corresponding position on the respective mask, and the image sensor accumulates the exposure data from each pixel in that time interval onto the compressed image, resulting in a final compressed image. Additionally, in the process of randomly generating binary masks, The probability of selecting a position on the mask for exposure is fixed, denoted as the Overlap Ratio (OR), which is selected through ablation experiments. The entire imaging process can be formulated as follow:

$$Y = \sum_{i=1}^{B} X_i \odot M_i + Z_i, \quad (5)$$

where Y, $X_i \in \mathbb{R}^{H \times W}$ are the modulated compressed image and the ith virtual image within exposure time respectively, B denotes the temporal Compression Ratio(CR), $\odot$ denotes element-wise multiplication, and $Z \in \mathbb{R}^{H \times W}$ is the measurement noise. Additionally, this process is entirely differentiable.

3.3. Transformation Network

The proposed method takes a compressed image as input, making it impractical to extract camera poses and
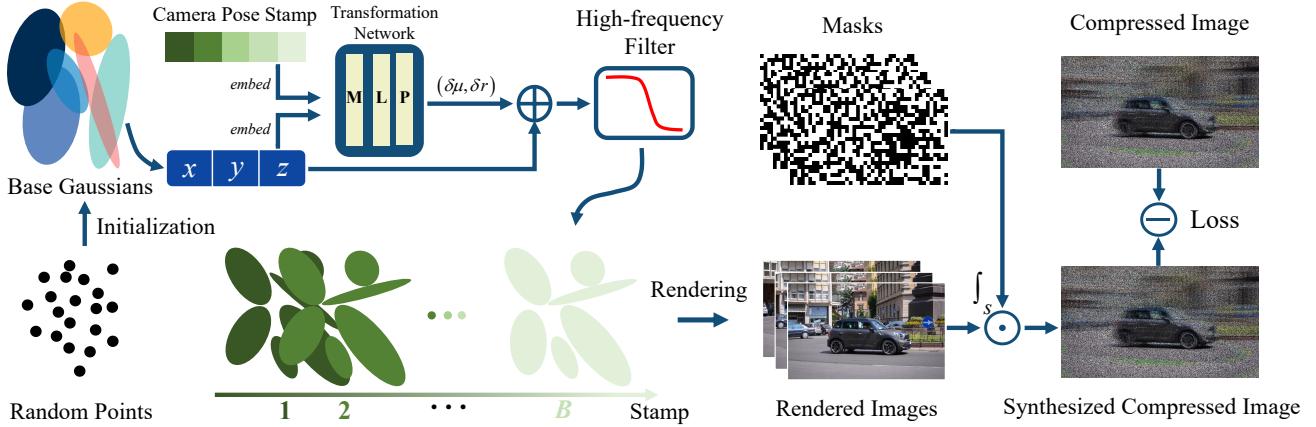
Figure 2. **The pipeline of the proposed SCIGS.** Given a set of randomly initialized 3D Gaussians and a camera pose, and introducing the same number of camera pose stamps as the compression ratio, our transformation network takes the Gaussian primitives and the camera pose stamps as inputs, followed by a high-frequency filter, outputs 3D Gaussians under different camera pose stamps. These camera-pose-aware transformed 3D Gaussians are then rendered to images under the given camera viewpoint, and are modulated by a given set of masks to generate compressed images.

point clouds by COLMAP[27], which necessitates optimizing camera poses and 3D Gaussians jointly.

When directly optimizing the camera by gradient descent, The gradient of the loss $\mathcal{L}_c$ on the camera extrinsic matrix E_c is shown as follow:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_c}{\partial E_c} &= \sum_{i=1}^{H \times W} \frac{\partial \mathcal{L}_c}{\partial C} \frac{\partial C}{\partial E_c} \\
&= \sum_{i=1}^{H \times W} \frac{\partial \mathcal{L}_c}{\partial C} \sum_{j=1}^{M} \left(\frac{\partial C}{\partial c_j} \frac{\partial c_j}{\partial E_c} + \frac{\partial C}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial E_c} \right),
\end{aligned}
\tag{6}
$$

where C and c_j denote the color of the rendered pixel and the color of the Gaussian $\mathcal{G}_j$ respectively. M indicates the number of visible 3D Gaussians.

As shown in Fig. 3(a) and Fig. 3(b), the possible conditions of 3D Gaussians can be roughly divided into two types: the projection of the Gaussian overlapping with the correct region and the projection of the Gaussian not overlapping with the correct region, hereinafter referred to as the effective Gaussians and the ineffective Gaussians respectively.

When the Gaussian $\mathcal{G}_i$ is an ineffective Gaussian, it is worth noting that the gradients $\frac{\partial C}{\partial c_j}$ and $\frac{\partial C}{\partial \alpha_j}$ provide no guidance, which lead the camera optimized in chaotic direction according to Eq. 6. In cases where there is a significant difference between the target camera pose and the initial camera pose, the number of ineffective Gaussian far exceeds the effective ones, hindering the oscillations of camera pose.

To avoid the above issue in camera pose optimization, this paper approaches the problem from the per-

spective of transforming Gaussian primitives and introduces a camera-pose-aware transformation network, with a multi-layer perceptron (MLP) as its core component. As formulated in Eq. 7, The input of this network consists of the positions of Gaussians and a camera pose stamp, while the output is the increments of positions and quaternions, i.e., $\delta\mu$, δr . Since moving the initial camera to the correct pose is equivalent to moving the 3D Gaussians to the correct position in front of the camera. Following the prior of local smoothness in natural images, contiguous Gaussians often have similar colors. With the presence of both effective and ineffective Gaussians in a neighborhood, as shown in Fig. 3(c), the correct gradients from effective Gaussians guide the transformation network optimized towards correct direction. Due to the continuity of the MLP, as shown in Fig. 3(d), nearby ineffective Gaussian points are also moved towards the correct direction with the transformation field, and gradually transformed into effective Gaussians. This process ultimately leads the 3D Gaussians to move towards the correct pose.

$$
(\delta\mu, \delta r) = \mathcal{F}\left(embed(\mu), embead(stamp)\right), \tag{7}
$$

$$
embed(\mu) = \left(sin(2^k \pi \mu), cos(2^k \pi \mu)\right)_{k=0}^{L-1}. \tag{8}
$$

Benefiting from the transformation network acting directly on Gaussian primitives instead of camera poses, the Gaussian primitives can undergo different transformations under varying camera poses. This approach offers the transformation network a potential means to learn the movement of objects within the
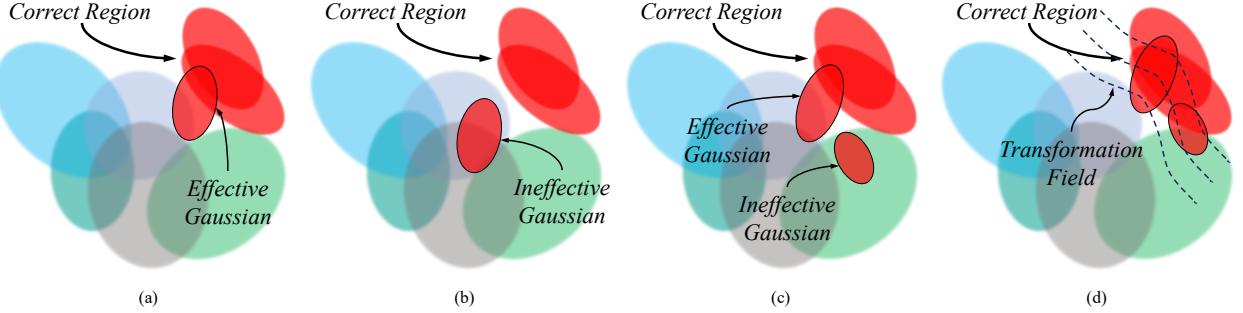
Figure 3. (a) illustrates an effective Gaussian, (b) illustrates an ineffective Gaussian. (c) and (d) show how the transformation network converts ineffective Gaussians to effective Gaussians.
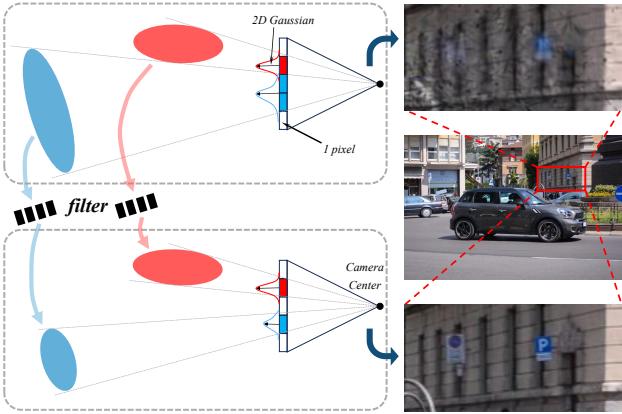

Figure 4. **The illustration of the principle of the proposed high-frequency filter.** The Gaussians that cause high-frequency artifacts are filtered to eliminate the artifacts.

scene, enabling SCIGS to reconstruct dynamic scenes from a single SCI image.

3.4. High-frequency Filter

When projecting 3D gaussians on imaging plane, a fixed 2D dilation factor is employed to ensure the projected 2D Gaussians are larger than one pixel, which leads to a systematically underestimation of scale. Then, in the case of moving 2D Gaussians closer, the rendered Gaussian is thinner than they actually appear, which exhibits high-frequency artifacts on rendered image. As shown in Fig. 4, in the proposed frame, the transformation network moves the base Gaussians to the appropriate position, and similar to the above phenomenon, when optimizing based on the positions under a certain stamp, Gaussians under another stamp may be rendered as high-frequency artifacts. To address this issue, inspired by Mip-Splatting[39], a high-frequency filter is introduced to eliminate the high-frequency artifacts.

Consider the rendering process as a sampling of 3D Gaussians, according to the Nyquist-Shannon sampling theorem: to recover the original signal from the sampled signal without distortion, the sampling frequency should be greater than two times the highest frequency of the signal. For a camera with focal f, its sampling interval is one pixel in screen space, the sampling interval $\hat{T}$ and sampling frequency $\hat{v}$ for an object in depth d in the camera coordinate are shown in Eq. 9.

$$\hat{T} = \frac{1}{\hat{v}} = \frac{d}{f}. \tag{9}$$

According to the sampling theorem, a primitive smaller than $2\hat{T}$ may result in artifacts during the splatting process, since the sampling frequency is below twice the signal frequency. Then the maximal sampling frequency of Gaussian $\mathcal{G}_i$ can be calculated by

$$\hat{v}_i = max\left(\left\{\frac{f_s}{d_s}\right\}_{s=1}^{S}\right), \tag{10}$$

where S denotes the total number of camera pose stamps. Given the maximal sampling frequency, as shown in Eq. 11, the high-frequency filter is achieved by convolving a low-pass filter Gaussian with 3D Gaussian before the rasterization pipeline.

$$\mathcal{G}_i'(x) = (\mathcal{G}_i * \mathcal{G}_{low})(x)$$
$$= \sqrt{\frac{|\Sigma_i|}{|\Sigma_i + \frac{\gamma}{\hat{v}_i} \cdot \mathbf{I}|}} e^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^\top (\Sigma_i + \frac{\gamma}{\hat{v}_i} \cdot \mathbf{I})^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})\right)}, \tag{11}$$

where γ is a hyperparameter controlling the filter size. the scale of the low-pass filtered Gaussian is $\frac{\gamma}{\hat{v}_i}$, which ensures that the scale of the filtered 3D Gaussian is not larger than the sampling interval after convolution. In the additional experiments, an ablation experiment demonstrated the validity of the high-frequency filter.

	Cozy2room			Tanabata			Factory			Vender			Airplants			Hotdog		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAP-TV[40]	21.77	.4321	.6031	20.42	.4264	.6250	24.05	.5666	.5149	20.00	.3678	.6882	22.85	.4057	.4986	22.35	.7663	.3179
PnP-FFDNet[41]	28.98	.8916	.0984	29.17	.9032	.1197	31.75	.8977	.1142	28.70	.9235	.1315	27.79	.9117	.1817	29.00	.9765	.0511
PnP-FastDVDNet[43]	30.19	.9132	.0793	29.73	.9333	.0980	32.53	.9165	.1055	29.68	.9395	.1043	28.18	.9092	.1757	29.93	.9728	.0522
EfficientSCI[31]	31.47	.9327	.0476	32.30	.9587	.0600	32.87	.9259	.0709	33.17	.9401	.0456	30.13	.9425	.1129	30.75	.9568	.0461
SCINerf[14]	33.23	.9492	.0445	33.61	.9638	.0374	36.60	.9638	.0221	36.40	.9840	.0298	30.69	.9335	.0728	31.35	.9878	.0310
SCIGS(ours)	33.78	.9191	.0423	35.12	.9580	.0271	37.75	.9646	.0291	36.00	.9641	.0192	27.18	.7267	.3003	29.31	.9369	.0809

Table 1. **Quantitative SCI image reconstruction comparisons on the static datasets.** The results demonstrate that our method outperforms or approaches the existing image reconstruction methods and 3D reconstruction methods for SCI image on most datasets from static scenes. The best results are shown in bold and the second-best results are underlined.

	Bear			Roundabout			Turn			Flamingo			Dance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
EfficientSCI[31]	26.81	.8759	.1040	22.08	.7854	.2934	22.30	.7763	.3613	25.97	.8795	.1386	24.83	.9055	.6268
SCINerf[14]	26.57	.7974	.1192	26.02	.8394	.1265	25.68	.6596	.2330	26.78	.7954	.1207	22.78	.6960	.2737
SCIGS(ours)	30.44	.9137	.0548	31.07	.9222	.0729	31.78	.8951	.0953	31.33	.9022	.0533	27.89	.9096	.0580

Table 2. **Quantitative SCI image reconstruction comparisons on the dynamic datasets.** The results demonstrate that our method surpasses the current image reconstruction methods and 3D reconstruction methods for SCI image on all of datasets from dynamic scenes. The best results are shown in bold.

4. Experiments

To evaluate the effectiveness of this method, extensive experiments were conducted using compressed images from both static and dynamic scenes. The results demonstrate that the proposed method delivers higher performance compared with existing works in SCI image decoding task and enhance the image quality in 3D reconstruction from compressed image.

4.1. Experiment Setup

Datasets. For fair evaluation of the proposed SCIGS, we follow SCINeRF and use six static scenes, including *Airplants* in LLFF[22] with solution 512×512, *Hotdog* in NeRF Synthetic360[23] with solution 400×400 and four datasets generated from the scenes in DeblurNeRF[19] (*Cozy2room*, *Tanabata*, *Factory*, and *Vender*) with solution 400×600. For dynamic scenes, we use five dataset from DAVIS2017[24] with resolution $480 \times 894(480p)$.

Baseline methods and evaluation metrics. In the comparison experiments, the SOTA SCI image decoding methods and SOTA reconstruction method for compressed image are used for comparison, including GAP-TV[40], PnPFFDNet[41], PnP-FastDVDNet[43], EfficientSCI[31], and SCINeRF[14]. Widely used metrics are employed for quantitative evaluations, including the structural similarity index(SSIM), peak signal-to-noise-ratio(PSNR), and learned perceptual image patch similarity(LPIPS)[44].

Implementation details. We use PyTorch framework with NVIDIA RTX A6000 GPUs for training. We used two independent Adam optimizers for the 3D Gaussians and the transformation network , and set the number of layers $L = 6$ in position coding function of the transformation network, the depth of the MLP $D = 8$ and the dimension of the hidden layer $W = 512$.

4.2. Result and Analysis

Static scenes. The existing SOTA SCI image decoding methods and SCINeRF are compared with the proposed method on static datasets. As shown in Table 1 and Fig. 5, in the majority of datasets, our method surpasses the SOTA SCI image decoding methods and exceeds or approaches the image quality of SCINeRF. This empirical result demonstrates the efficacy of our SCIGS in recovering 3D scenes from compressed images. It is noteworthy that our approach did not outperform SCINeRF and EfficientSCI in *airplants* and *hotdog*, which can be attributed to the fact that our method indirectly decodes the compressed image by recovering 3D representation, and the detail information lost due to the lack of texture information in *airplants* and the large-scale camera movements in *hotdog*.

Dynamic scenes. We also compared our method against SOTA SCI image restoration methods (EfficientSCI) and the SOTA 3D reconstruction method from compressed images (SCINeRF) in dynamic scenes. As shown in Table 2 and Fig. 5, in all scenes, our SCIGS outperforms SCINeRF on all metrics. This result demonstrates the efficacy of SCIGS in recovering dynamic scenes from compressed images. Qualita-

Figure 5. **Qualitative evaluations on the synthetic dataset compare our proposed method (SCIGS) with the SOTA SCI image method (SCINeRF).** From top to bottom are two static scenes (factory and tanabata) and two dynamic scenes (roundabout and flamingo). The experiments show that our method achieves comparable image recovery performance from a single compressed image in static scenes, while demonstrating superior performance in dynamic scenes.

tively, as shown in Fig. 5, we note that SCIGS demonstrates excellent performance in recovering the details of the moving objects in the scene, where existing methods fails to restore.

4.3. Additional Study

Mask overlapping rate. We assess the impact of various mask overlapping rates during the SCI image modulated. The mask overlapping rate is defined by the probability that a mask selects a specific pixel for exposure, which is formulated as follow:

$$OR(x, y) = \frac{\sum_{i=1}^{N} M_i(x, y)}{N} \quad (12)$$

where OR denotes mask overlapping rate, M_i indicates i-th mask and N is the number of Intermediate frame. From Eq. 12, lower the mask crossing rate means the sparser sampling, which result in the less image information retained, and conversely, the denser the sampling leads to the more information retained. However,

OR	PSNR↑	SSIM↑	LPIPS↓
0.125	30.32	.9066	.0634
0.25	30.41	.8954	.0814
0.5	29.04	.8569	.1145
0.75	27.50	.8294	.1336

Table 3. **The average metrics of image quility in the additional study on mask overlapping rate.** the quality of reconstruction increases first and then decreases with the overlapping rate ranging from 0.125 to 0.75.

too high a sampling rate will increase the ambiguity of the compressed image and may result in a blurred decoded image. As shown in Table 3, we tested different overlapping rates on multiple datasets, and the results showed that the image quality first increased and then decreased when the overlapping rate increased from 0.125 to 0.25, and decreased after 0.25. Empirically, we selected overlap rate of all experiments within 0.25.

Figure 6. **Ablation study on the high-frequency filter of SCIGS in both static and dynamic scenes.** The first two columns represent static scenes (factory and vender), while the last two columns show dynamic scenes (flamingo and roundabout). As shown, the addition of the high-frequency filter significantly enhances image quality in both static and dynamic scenes, with reduced artifacts and clearer details, further validating the effectiveness of the high-frequency filter across various scene types.

	PSNR↑		SSIM↑	
Filter	w/	w/o	w/	w/o
Factory	37.75	33.08	.9646	.9230
Vender	36.00	31.60	.9641	.9334
Flamingo	31.33	26.38	.9022	.7836
Roundabout	31.07	22.38	.9222	.7331

Table 4. **The results of ablation experiment validate the high-frequency filter.** The results show that the introduction of high-frequency filters significantly improves the image quality.

4.4. Ablation Experiment

To validate the efficacy of the proposed high-frequency filter, we conducted ablation experiments to compare the performance of our method before and after incorporating the filter, evaluated across both static and dynamic scenes. The results, presented in Table 4, demonstrate that the inclusion of the high-frequency filter significantly improves the quality of the recovered images, particularly in terms of visual fidelity and artifact reduction.

Moreover, the effectiveness of the high-frequency filter is evident in the qualitative experiments. As shown in Fig. 6, the filter almost completely eliminates artifacts, enhancing image sharpness and clarity, and leading to a more accurate and realistic reconstruction, particularly in high-frequency regions.

5. Conclusion

This paper proposes a novel method for recovering dynamic 3D scene representations from a single snapshot conpressive image, which is the first to introduce an dynamic explicit representation in this task, extending its application to dynamic scenes. We propose a transformation network to substitute for directly optimizing the camera poses and a high-frequency filter to eliminate the artifacts generated during the transformation. Different from previous works, our method can adequately reconstructing dynamic scenes from compressed image, and provides a new idea for optimizing camera poses with the absence of camera poses and pre-trained 3DGS representations. To access the effectiveness of SCIGS, extensive comparative experiment are conducted against the existing state-of-the-art SCI image recovery methods and SCI image-based reconstruction methods both in static scenes and dynamic scenes. And the result demonstrate the superior performance of our method, especially in dynamic scenes.

Due to the advantages of SCI technology in terms of storage cost, and the superiority and scalability of our 3DGS-based framework demonstrated in dynamic scenes, our method has great potential for low-cost and fast incremental reconstruction in high-speed dynamic scenes, such as autonomous driving.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 2

[2] Matteo Bortolon, Theodore Tsesmelis, Stuart James, Fabio Poiesi, and Alessio Del Bue. 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. *arXiv preprint arXiv:2407.15484*, 2024. 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] 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

[5] 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. 1, 2

[6] 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. 1, 2

[7] Weisheng Dong, Guangming Shi, Xin Li, Yi Ma, and Feng Huang. Compressive sensing via nonlocal low-rank regularization. *IEEE transactions on image processing*, 23(8):3618–3632, 2014. 2

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

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting, 2024. 3

[11] Yasunobu Hitomi, Jinwei Gu, Mohit Gupta, Tomoo Mitsunaga, and Shree K Nayar. Video from a single coded exposure photograph using a learned overcomplete dictionary. In *2011 International Conference on Computer Vision*, pages 287–294. IEEE, 2011. 2

[12] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2

[13] 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, 3

[14] 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, 6, 1, 3

[15] 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. 1

[16] Dengyu Liu, Jinwei Gu, Yasunobu Hitomi, Mohit Gupta, Tomoo Mitsunaga, and Shree K Nayar. Efficient space-time sampling with pixel-wise coded exposure for high-speed imaging. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):248–260, 2013. 2

[17] 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. 1, 2

[18] 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

[19] 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

[20] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising, deblocking, and enhancement through separable 4-d nonlocal spatiotemporal transforms. *IEEE Transactions on image processing*, 21(9):3952–3966, 2012. 2

[21] Ziyi Meng, Shirin Jalali, and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv preprint arXiv:2012.08364*, 2020. 1

[22] 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

[23] 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

[24] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object

segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[25] Mu Qiao, Ziyi Meng, Jiawei Ma, and Xin Yuan. Deep learning for video compressive sensing. *Apl Photonics*, 5(3), 2020. 1, 2

[26] Mu Qiao, Xuan Liu, and Xin Yuan. Snapshot temporal compressive microscopy using an iterative algorithm with untrained neural networks. *Optics Letters*, 46(8): 1888–1891, 2021. 2

[27] 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. 4

[28] Yuan Sun, Xuan Wang, Yunfan Zhang, Jie Zhang, Caigui Jiang, Yu Guo, and Fei Wang. icomma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching. *arXiv preprint arXiv:2312.09031*, 2023. 3

[29] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE global conference on signal and information processing*, pages 945–948. IEEE, 2013. 2

[30] 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*, 45(7):9072–9089, 2022. 1, 2

[31] 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. 1, 2, 6, 3

[32] 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

[33] 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

[34] Zongliang Wu, Chengshuai Yang, Xiongfei Su, and Xin Yuan. Adaptive deep pnp algorithm for video snapshot compressive imaging. *International Journal of Computer Vision*, 131(7):1662–1679, 2023. 2

[35] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 3

[36] Jianbo Yang, Xuejun Liao, Xin Yuan, Patrick Llull, David J Brady, Guillermo Sapiro, and Lawrence Carin. Compressive sensing by learning a gaussian mixture model from measurements. *IEEE Transactions on Image Processing*, 24(1):106–119, 2014. 1, 2

[37] Jianbo Yang, Xin Yuan, Xuejun Liao, Patrick Llull, David J Brady, Guillermo Sapiro, and Lawrence Carin. Video compressive sensing using gaussian mixture models. *IEEE Transactions on Image Processing*, 23 (11):4863–4878, 2014. 2

[38] 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

[39] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19447–19456, 2024. 2, 5

[40] 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. 1, 2, 6, 3

[41] 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, 6, 1, 3

[42] 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

[43] 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. 6, 1, 2, 3

[44] 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*, pages 586–595, 2018. 6

SCIGS: 3D Gaussians Splatting from a Snapshot Compressive Image

Supplementary Material

In this supplementary material, additional experiments are conducted on datasets from dynamic scenes and static scenes, comparing our SCIGS against current state-of-the-art SCI decoding methods (GAP-TV[40], PnP-FFDNet[41], PnP-FastDVDNet[43] and EfficientSCI[31]) and state-of-the-art SCI image-based reconstruction method(SCINeRF[14]).

A. Additional Experiments

A.1. Experiment Setup

To further validate the effectiveness of our method in dynamic scenes, additional qualitative and quantitative experiments were conducted on five datasets from dynamic scene (*Bear*, *Roundabout*, *Flamingo*, *Turn* and *Dance*). For fair comparisons, we fine-tuned EfficientSCI [31] with the masks used in our datasets. Additionally, the results of qualitative experiments conducted under all static scene datasets (*Factory*, *Tanabata*, *Vender*, *Cozy2room*, *hotdog* and *airplants*) are also presented in this supplementary material. For a better quantitative comparison, we also present the results of the experiments with static scenes, which are shown in Table B.

A.2. Result and Analysis

The results of the qualitative and quantitative experiments in dynamic scenes are shown in Fig. A and Table A, respectively. These results provide empirical evidence for the effectiveness of our SCIGS in reconstructing dynamic scenes from single compressed images. It is also worth noting that the metrics of our method do not exceed EfficientSCI in *Dance*. The observation can be attributed to the fact that our method recovers images by reconstructing the underlying scene. The images in the *Dance* dataset have dynamic blur, which leads to the loss of structural information, so SCIGS cannot accurately reconstruct this part of the scene, leading to a degradation in image quality. In contrast, as a traditional SCI image decoding methods, EfficientSCI uses only 2D image information without considering the structural consistency, and thus outperforms our method in this scene.

As shown in Fig. B and Table B, the proposed SCIGS shows comparable image recovery performance on static scene. In addition, we note that SCIGS outperforms existing methods in the reconstruction of the parts with rich textures and characters, which cannot be directly observed from metrics.

	Bear			Roundabout			Turn			Flamingo			Dance		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAP-TV[40]	22.63	.5698	.3734	22.26	.6976	.3823	25.28	.6774	.3437	23.68	.6986	.3404	22.20	.6981	.3953
PnP-FFDNet[41]	21.91	.6569	.3822	25.80	.8727	.1314	26.93	.8598	.2661	25.50	.8206	.2000	22.29	.8284	.1987
PnP-FastDVDNet[43]	26.77	.8561	.1413	27.01	.8938	.1006	27.58	.8723	.2090	29.27	.8978	.0994	28.10	.9465	.0569
EfficientSCI[31]	29.26	.9099	.0710	28.45	.9110	.0876	29.03	.8934	.1617	31.03	.9247	.0668	31.55	.9677	.0412
SCINerf[14]	26.57	.7974	.1192	26.02	.8394	.1265	25.68	.6596	.2330	26.78	.7954	.1207	22.78	.6960	.2737
SCIGS(ours)	30.44	.9137	.0548	31.07	.9222	.0729	31.78	.8951	.0953	31.33	.9022	.0533	27.89	.9096	.0580

Table A. **Quantitative SCI image reconstruction comparisons on the dynamic datasets.** The results demonstrate that our method surpasses the current SCI decoding methods and 3D reconstruction methods for SCI image on datasets from dynamic scenes. The best results are shown in bold and the second-best results are underlined.
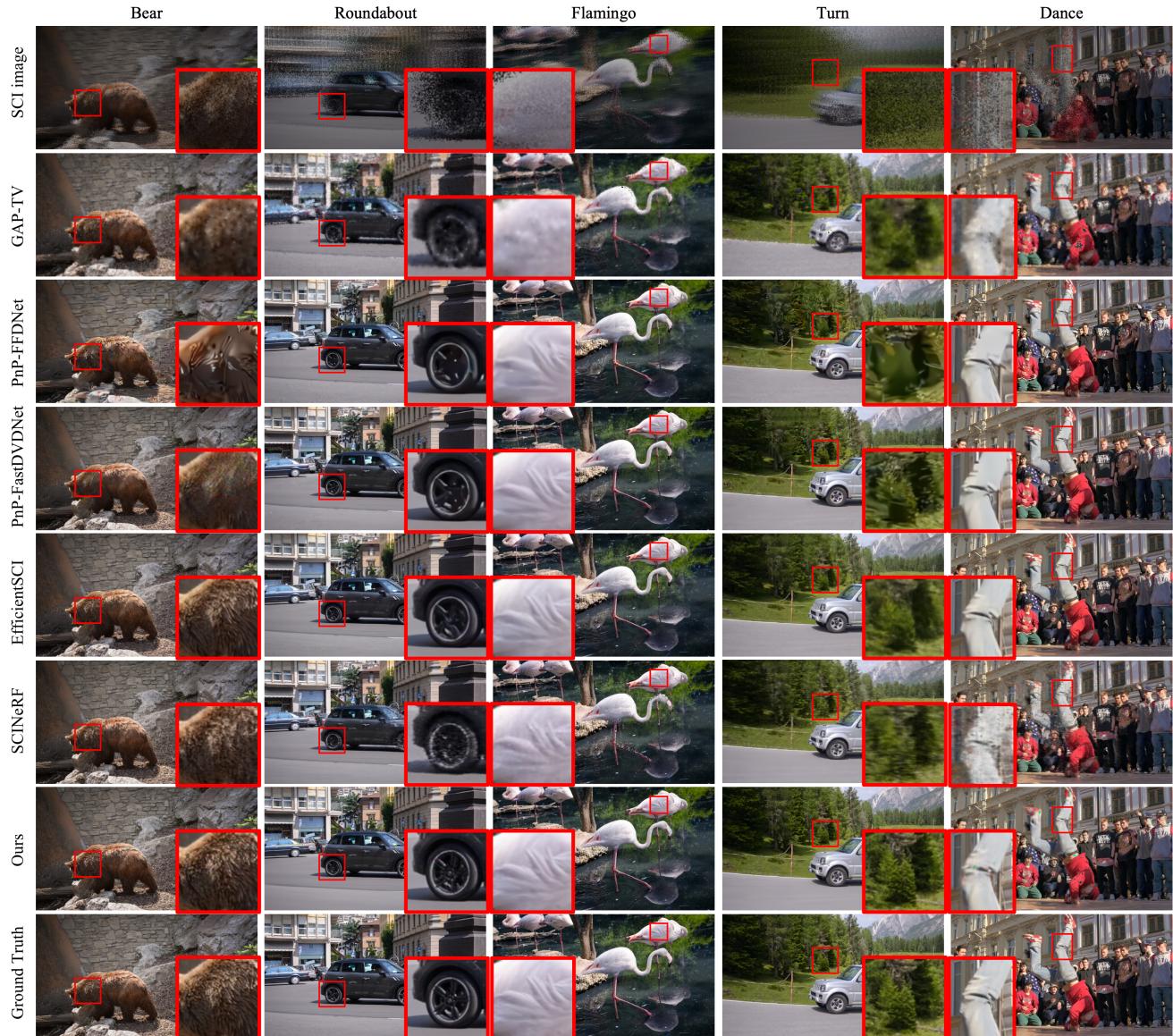

Figure A. **Qualitative evaluations on the datasets from dynamic scenes.** From left to right shows the results for five dynamic scenes including *Bear*, *Roundabout*, *Flamingo*, *Turn* and *Dance*. The experiments show that our method achieves superior performance in dynamic scenes.

	Cozy2room			Tanabata			Factory			Vender			Airplants			Hotdog		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAP-TV[40]	21.77	.4321	.6031	20.42	.4264	.6250	24.05	.5666	.5149	20.00	.3678	.6882	22.85	.4057	.4986	22.35	.7663	.3179
PnP-FFDNet[41]	28.98	.8916	.0984	29.17	.9032	.1197	31.75	.8977	.1142	28.70	.9235	.1315	27.79	.9117	.1817	29.00	.9765	.0511
PnP-FastDVDNet[43]	30.19	.9132	.0793	29.73	.9333	.0980	32.53	.9165	.1055	29.68	.9395	.1043	28.18	.9092	.1757	29.93	.9728	.0522
EfficientSCI[31]	31.47	.9327	.0476	32.30	.9587	.0600	32.87	.9259	.0709	33.17	.9401	.0456	30.13	.9425	.1129	30.75	.9568	.0461
SCINerf[14]	33.23	.9492	.0445	33.61	.9638	.0374	36.60	.9638	.0221	36.40	.9840	.0298	30.69	.9335	.0728	31.35	.9878	.0310
SCIGS(ours)	33.78	.9191	.0423	35.12	.9580	.0271	37.75	.9646	.0291	36.00	.9641	.0192	27.18	.7267	.3003	29.31	.9369	.0809

Table B. **Quantitative SCI image reconstruction comparisons on the static datasets.** The results demonstrate that our method outperforms or approaches the existing image reconstruction methods and 3D reconstruction methods for SCI image on most datasets from static scenes. The best results are shown in bold and the second-best results are underlined.

Figure B. **Qualitative evaluations on the datasets from static scenes.** From left to right shows the results for five static scenes including *Factory*, *Tanabata*, *Vender*, *Cozy2room*, *Hotdog* and *Airplants*. The experiments show that our method achieves comparable image recovery performance from a single compressed image in static scenes.