

# SpikeGS: Learning 3D Gaussian Fields from Continuous Spike Stream

Jinze Yu<sup>1</sup>, Xin Peng<sup>2</sup>, Zhengda Lu<sup>3</sup>, Laurent Kneip<sup>4</sup>, and Yiqun Wang<sup>1\*</sup>

<sup>1</sup> College of Computer Science, Chongqing University, Chongqing, China

<sup>2</sup> Motovis Co., Ltd., Shanghai, China

<sup>3</sup> University of Chinese Academy of Sciences, Beijing, China

<sup>4</sup> Mobile Perception Lab, ShanghaiTech University, Shanghai, China

**Abstract.** A spike camera is a specialized high-speed visual sensor that offers advantages such as high temporal resolution and high dynamic range compared to conventional frame cameras. These features provide the camera with significant advantages in many computer vision tasks. However, the tasks of novel view synthesis based on spike cameras remain underdeveloped. Although there are existing methods for learning neural radiance fields from spike stream, they either lack robustness in extremely noisy, low-quality lighting conditions or suffer from high computational complexity due to the deep fully connected neural networks and ray marching rendering strategies used in neural radiance fields, making it difficult to recover fine texture details. In contrast, the latest advancements in 3DGs have achieved high-quality real-time rendering by optimizing the point cloud representation into Gaussian ellipsoids. Building on this, we introduce SpikeGS, the method to learn 3D Gaussian fields solely from spike stream. We designed a differentiable spike stream rendering framework based on 3DGs, incorporating noise embedding and spiking neurons. By leveraging the multi-view consistency of 3DGs and the tile-based multi-threaded parallel rendering mechanism, we achieved high-quality real-time rendering results. Additionally, we introduced a spike rendering loss function that generalizes under varying illumination conditions. Our method can reconstruct view synthesis results with fine texture details from a continuous spike stream captured by a moving spike camera, while demonstrating high robustness in extremely noisy low-light scenarios. Experimental results on both real and synthetic datasets demonstrate that our method surpasses existing approaches in terms of rendering quality and speed.

**Keywords:** Spike camera · 3D Gaussian splatting · Novel View Synthesis · 3D reconstruction

## 1 Introduction

In recent years, neuromorphic cameras have made significant advancements, most notably in the development of spike cameras [8, 20] and event cameras

---

\* Corresponding author

[10, 43]. These types of cameras excel in capturing intensity changes in high-speed scenes due to their high temporal resolution, and high dynamic range, offering significant potential for applications in the field of computer vision [5, 9, 16, 17, 27, 30, 62]. The output of a spike camera is a spike stream, fundamentally different from the data produced by traditional frame cameras. When the accumulated photons exceed the threshold, each pixel independently responds to the accumulation of photons by generating asynchronous spikes and then resets the accumulation. At each timestamp, the spike camera outputs a binary matrix, known as an event frame, indicating the presence of spikes at all pixels. This distinctive feature enables the spike camera to have complete texture sensing and high-speed sensing capabilities, and to record full visual details with an ultra-high temporal resolution of up to 40 kHz. Thanks to these capabilities, spike cameras excel in various fields of computer vision tasks [7, 16, 18, 50, 61].

Meanwhile, the field of computer vision is increasingly inclined to explore neural radiance fields [33] and 3DGS [23] as solutions for 3D scene reconstruction and novel view synthesis. However, due to the distinctive data modality of spike camera, current algorithms for 3D reconstruction and novel view synthesis [28, 32, 38, 41, 46, 48] primarily rely on high-quality images obtained by traditional frame-based cameras under optimal lighting conditions. This raises the question of whether we can reconstruct dense and realistic 3D scene representations solely from the spike stream captured by a moving spike camera, and whether such reconstructions can maintain robustness in extremely noisy and low-light scenarios (where, in real-world situations, the spike camera inevitably generates a large amount of noise due to its internal circuit structure and external ambient light).

One of the most representative algorithms in the field of novel view synthesis today is NeRF [33], which implicitly represents a scene as a neural radiance field. By combining implicit function representation of MLPs with differentiable rendering, NeRF has garnered widespread attention for its ability to recover high-quality 3D scene representations from 2D images. However, due to the use of deep fully connected neural networks and the need for per-pixel ray sampling during the rendering process, NeRF suffers from high sampling costs, potential noise generation, and considerable computational complexity. A recent advancement, 3D Gaussian Splatting (3DGS) [23], explicitly represents scenes by optimizing Gaussian ellipsoids. Thanks to the tile-based multithreaded parallel rendering mechanism of 3D Gaussian, which achieves real-time rendering by splatting three-dimensional ellipsoids onto a two-dimensional plane, surpassing NeRF in both rendering quality and speed.

Although some studies have explored the application of 3DGS and NeRF in a unique type of neuromorphic camera known as an event camera, characterized by differential sampling. Representative examples include Ev-NeRF [21], EventNeRF [44], and EvGGS [49], all of which introduce neural radiance fields or Gaussian fields derived specifically from event streams. However, the inherent lack of texture detail in event data limits the effectiveness of these methods, resulting in suboptimal outcomes. The concurrent work [66] [67] explored spike

stream reconstruction under high-speed motion cameras. SpikeNeRF [63] and Spike-NeRF [14] have both explored methods for learning neural radiance fields from spike stream. However, the deep fully connected neural networks and ray marching-based rendering strategies used in neural radiance fields make it difficult to achieve high-quality real-time rendering. Additionally, Spike-NeRF [14] does not demonstrate robustness in extremely noisy, low-light scenarios.

Based on the above, we leverage the multi-view consistency of 3D Gaussian and the tile-based multi-threaded parallel rendering mechanism in conjunction with spiking neurons to establish robust self-supervision, mitigating the impact of erroneous measurements under high noise levels and diverse illumination conditions. At the same time, we achieve high-quality real-time rendering. The main contributions of this paper are:

- 1) We proposed a novel differentiable rendering framework that learns 3D Gaussian fields solely from spike stream (Fig. 3). SpikeGS (Fig. 1) exhibits high robustness in extremely noisy, low-quality lighting scenarios.
- 2) We proposed a novel spike stream rendering loss function based on 3D Gaussian splatting (3DGS) and spiking neurons capable of generalizing across varying illumination conditions (Fig. 5).
- 3) Experiments on synthetic and real datasets demonstrate that our method outperforms prior state-of-the-art implicit neural rendering methods in terms of rendering quality and speed.

## 2 Related work

### 2.1 Neural Radiance Fields and 3DGS

NeRF [33] employs MLPs to represent neural implicit fields and has garnered widespread attention for its excellent performance in synthesizing high-quality novel views and accurately representing 3D scenes. Improved works based on NeRF have also emerged subsequently. In terms of fast rendering, instant-ngp [36] replaces NeRF’s fully connected neural networks with a smaller MLP and introduces multi-resolution hash encoding to enhance NeRF’s rendering speed. In the realm of sparse view reconstruction, methods like PixelNeRF [56] and RegNeRF [39] achieve high-quality novel view synthesis using minimal input images. In the domain of deblurring, approaches like Deblur-NeRF [29] and DP-NeRF [25] aim to reconstruct clear scene representations from blurred input views by modeling the physical processes of motion blur. Enhancing rendering quality, Mip-NeRF [1] introduces a sampling strategy based on view frustum for NeRF-based anti-aliasing and addressing aliasing artifacts. Works such as Block-NeRF [47] and BungeeNeRF [53] extend NeRF’s rendering scale from small to city-scale large scenes. A recent revolutionary 3D reconstruction method, 3D Gaussian Splatting (3DGS) [23], represents the scene using optimizable Gaussian ellipsoids, which is fundamentally different from NeRF’s MLP-based implicit representation. The novel representation of 3DGS makes it possible to render images in real-time, furthermore improving the training time. A plenty of 3DGS-based

techniques [3, 6, 22, 24, 31, 58] have been proposed recently. Deformable-GS [55] and 4DGS [52] proposed dynamic scene reconstruction based on 3D Gaussians. VastGS [26] extended the reconstruction capabilities of 3D Gaussians to large-scale scenes. SuGaR [12] and 2DGS [19] improve surface fitting by flattening Gaussian ellipsoids, allowing them to better conform to the scene’s surface.

## 2.2 Scene reconstruction based on event cameras and spike cameras

Event cameras generate events asynchronously based on changes in scene brightness. Each event records the pixel position, timestamp of occurrence, and polarity change. Spike cameras capture the absolute brightness of each pixel and output spike stream. Specialized methods like Event-NeRF [44] and Ev-NeRF [21] have been proposed to derive neural radiance fields directly from event streams. E2NeRF [42] integrates event data and blurred frames to guide the reconstruction of clear radiance fields from blurry inputs. Evdeblurnerf [2] combines a series of previous works [25, 29, 40, 42], integrating blur kernels, adaptive weighting networks, and the EDI [40] model with an event camera, achieving state-of-the-art (SOTA) results in deblurring reconstruction on NeRF. [57] and [54] introduce event cameras into the framework of 3D Gaussian based on previous NeRF methods to assist in deblurring reconstruction. EvGGS [49] introduces a generalized event-based Gaussian splatting learning framework. SpikeNVS [7] combines spike stream and RGB images synergistically to recover clear neural radiance fields from blurry inputs. SpikeNeRF [63] and Spike-NeRF [14] both propose novel view synthesis methods based on neural radiance fields. However, constrained by the limitations of neural radiance fields, they struggle to recover fine texture details of scenes and have extremely slow training speed, whereas [14] focuses only on simple synthetic datasets without significant noise.

In general, although spike cameras offer advantages that traditional cameras lack, current NeRF-based methods with spike cameras cannot achieve high-quality real-time rendering. Therefore, we propose SpikeGS.

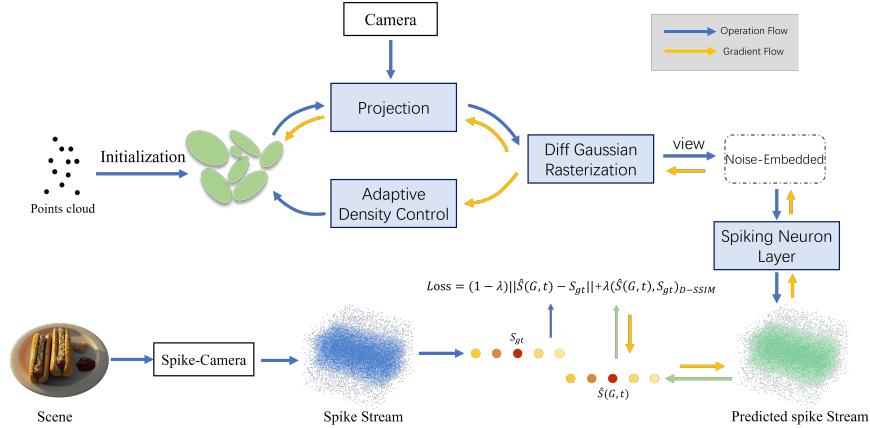
## 3 Method

### 3.1 3D Gaussian Splatting

The 3D Gaussian approach does not rely on neural radiance fields. instead, it represents the scene as a series of 3D Gaussian distributions. Based on the initialized sparse point cloud or randomly generated point cloud, a set of 3D Gaussians, defined as  $\mathbf{G}$ , is parameterized by its mean position  $\boldsymbol{\mu} \in \mathbb{R}^3$ , 3D covariance  $\boldsymbol{\Sigma} \in \mathbb{R}^{3 \times 3}$ , opacity  $\alpha \in \mathbb{R}$  and color  $c \in \mathbb{R}^3$ .  $c$  is represented by spherical harmonics for view-dependent appearance. The distribution of each Gaussian is defined as:

$$\mathbf{G}(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})} \quad (1)$$

It is essential to note that directly optimizing the covariance matrix  $\boldsymbol{\Sigma}$  can result in a non-positive semi-definite matrix, which would not adhere to the physical



**Fig. 1:** The architecture of SpikeGS. we establish the connection between the 3D Gaussian fields and the real-world spike stream  $S$ . Firstly, the input is randomly initialized point clouds or sparse point clouds generated by structure-from-motion (SfM) [45]. The point clouds are then converted into Gaussian ellipsoids, and rasterization is used to render the pixel values from corresponding viewpoints. Next, the rendered pixels are converted into spike stream through a spike stream generation pipeline (Noise Embedded and Spiking Neurons), and a rendering loss is established by comparing them with real-world spike stream. Finally, the loss is backpropagated to update the learnable parameters (mean position  $\mu$ , 3D covariance  $\Sigma$ , opacity  $\alpha$  and color  $c$ ) of the Gaussian ellipsoids for optimization. In Section 3, we will model the event stream generation process based on 3D Gaussian distributions and derive the gradient backpropagation process for our model.

interpretation typically associated with covariance matrices. To circumvent this issue, 3D-GS chooses to optimize a quaternion  $\mathbf{q}$  and a 3D vector  $\mathbf{s}$  represent rotation and scale, respectively. This approach allows the covariance matrix  $\Sigma$  to be reconstructed as follows:

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T \quad (2)$$

where  $\mathbf{R}$  and  $\mathbf{S}$  denote the rotation and scaling matrix derived from  $\mathbf{q}$  and  $\mathbf{s}$ , respectively. There is a complex computational graph to obtain the opacity  $\alpha$ :  $\mathbf{q}$  and  $\mathbf{s} \rightarrow \Sigma$ ,  $\Sigma \rightarrow \Sigma'$ ,  $\Sigma' \rightarrow \alpha$  [4].

To enable differentiable Gaussian rasterization, 3D gaussians are projected into the 2D image space from a given camera pose  $\mathbf{T}_c = \{\mathbf{R}_c \in \mathbb{R}^{3 \times 3}, \mathbf{t}_c \in \mathbb{R}^3\}$  for rasterizing and rendering using the following equation, as described in [65]. Given the viewing transformation  $\mathbf{W}$  and 3D covariance matrix  $\Sigma$ , the projected 2D covariance matrix  $\Sigma'$  is computed using:

$$\Sigma' = \mathbf{J} \mathbf{W} \Sigma \mathbf{W}^T \mathbf{J}^T \quad (3)$$

where  $\mathbf{J}$  is the Jacobian of the affine approximation of the projective transformation.

Subsequently, the transformed Gaussian ellipsoids are sorted based on their depth and the sorted Gaussian ellipsoids are rasterized to render pixel values using the following volume rendering equation:

$$C = \sum_{i \in N} c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j) \quad (4)$$

where  $c_i$  is the learned color and the final opacity  $\alpha'_i$  is the multiplication result of the learned opacity  $\alpha_i$  and the Gaussian:

$$\alpha'_i = \alpha_i \times \exp\left(-\frac{1}{2} (\mathbf{x}' - \boldsymbol{\mu}'_i)^T \boldsymbol{\Sigma}'^{-1} (\mathbf{x}' - \boldsymbol{\mu}'_i)\right) \quad (5)$$

where  $\mathbf{x}'$  and  $\boldsymbol{\mu}'_i$  are coordinates in the projected space.

### 3.2 Spike Signal Model

A spike camera reflects the light intensity of a scene through discharge events that occur when the voltage of a photodiode is released to the reference voltage (the received incoming photons will be transferred to voltage). The accumulator at each pixel accumulates the light intensity. For a pixel at position  $(x, y)$ , if the accumulated intensity reaches the scheduling threshold  $\Omega$ , a spike is emitted. Simultaneously, the corresponding accumulator is reset by subtracting the scheduling threshold from its own intensity. As shown in Equation (6) below,  $A_{t_i}$  and  $A_{t_{i-1}}$  represent the values of the accumulator at time  $t_i$  and  $t_{i-1}$ , respectively.  $I_{t_i}$  represents the input value at time  $t_i$ .

$$A_{t_i} = (A_{t_{i-1}} + I_{t_i}) \bmod \Omega \quad (6)$$

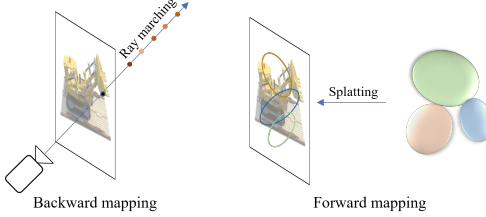
The integral form of the accumulator voltage can be expressed as:

$$A(x, y, T) = \int_0^T \eta \cdot I(x, y, t) dt \bmod \Omega \quad (7)$$

where  $I(x, y, t)$  represents the light intensity at pixel  $(x, y)$  at time  $t$ , and  $\eta$  is the photoelectric conversion rate. We will directly use  $I(t)$  to represent the luminance intensity to simplify our presentation. Due to the limitations of circuits, each spike is read out at discrete time  $t$ ,  $t \in T$  ( $T = Nt$ , where  $t$  represents the unit time step and  $N$  is the size of the time window). Thus, the output of the spike camera is a spatial-temporal binary stream  $S$  with  $H \times W \times N$  size. Here,  $H$  and  $W$  are the height and width resolution of the sensor, respectively, and  $N$  is the temporal window size of the spike stream. In our experiments, we set  $N$  to 256. The process of spike emission can be represented by the following equation:

$$S_{t_i} = \begin{cases} 1, & \text{if } A_{t_{i-1}} + I_{t_i} \geq \Omega \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

where, 0 indicates no spike, while 1 indicates a spike is sent.



**Fig. 2:** Rendering Diagrams for 3D Gaussians and NeRF.

### 3.3 Spike Noise Model

Inspired by [64], we understand that due to significant circuit differences, the noise characteristics of spike cameras differ greatly from traditional cameras. The photodetectors in spike cameras receive photons from the scene, but even under uniform illumination, the photons striking the diodes are not constant. The difference between the number of photons at a given moment and the ideal number of photons is referred to as shot noise  $N_p$ . The dark current noise generated by the thermal diffusion of charge carriers and defects within or on the surface of a PN junction is denoted as  $N_d$ . Additionally, differences in photodiode characteristics and capacitance contribute to variations in pixel sensitivity to incident light intensity, resulting in photo-response non-uniformity noise  $N_{rnu}$ . Moreover, the temporal delay between the generation of the reset signal and the subsequent release of the spike signal introduces quantization noise  $N_q$ . If the temporal length of the spike stream is not long enough, truncation noise will appear in the process from the spike to the image. we define truncation noise as  $N_c$ . Therefore, based on the above, we can define the equation for pulse noise as:

$$I + N = \frac{1}{\frac{Q_r}{L+N_p+N_d} + N_{rnu} + N_q} + N_c \quad (9)$$

Where  $I$  is the ideal image without noise,  $N$  is the total noise,  $L$  represents the scene light intensity,  $Q_r$  is the relative quantity matrix of electric charge.  $N_p$ ,  $N_d$ ,  $N_{rnu}$ ,  $N_q$ , and  $N_c$  represent shot noise, dark current noise, response nonuniformity noise, quantization noise, and truncation noise, respectively.

### 3.4 SpikeGS

The first spike-camera-based 3D Gaussian Splatting framework, SpikeGS, is introduced in this section. An integrate-and-fire mechanism is utilized to simulate the generation of spikes from intensity, which is estimated by splatting 3D Gaussians into an image plane. The noise generation mechanism is considered to imitate how a spike camera works under real scenarios, especially low-illumination

scenarios. Simulated spike stream are compared with real spike stream to indicate 3D Gaussians representing 3D scenes correctly. Details are demonstrated as follows.

**Establish the relationship between rendered pixel values and the real-world light intensity values  $I$ .** Before establishing the relationship between rendering spike stream and real spike stream, we first establish their relationship in terms of light intensity. The goal is to estimate the intensity values corresponding to the real light of the scene. Unlike NeRF, which is a backward mapping process that operates on a per-pixel basis, emitting rays from 2D pixels and integrating along sampled points on the ray via volume rendering to synthesize the corresponding pixel values shown as Fig. 2. On the other hand, 3D Gaussians utilize forward mapping. In this forward mapping, 3D Gaussian ellipsoids are splatted onto the 2D image plane through a rasterization pipeline (Each ellipsoid corresponds to multiple two-dimensional pixels). Therefore, in models based on 3D Gaussians, we cannot directly establish the relationship between the pixel ray  $r$  and the real-world light intensity values  $I$ . Therefore, we chose to establish the relationship between the view(pose) and the real-world light intensity values  $I$ . We denote the estimated intensity value corresponding to the real light of the scene as:

$$\hat{I}(R_G(P), t), \quad R_G(P) = \{C(x) | x \in X\} \quad (10)$$

where  $R_G(P)$  represents the rendering image  $R_G$  of Gaussian Splatting  $G$  from the pose  $P$ .  $C(x)$  represents the pixel value at the rendered pixel position  $x$ .  $X$  denotes the set of all pixels in the image space.

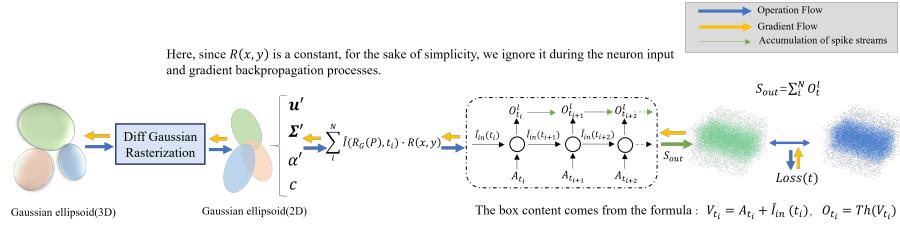
**Noise embedding.** To supervise  $\hat{I}(R_G(P), t)$  using real noise spike stream, we need to consider noise in multiple scenarios. As stated in Section 3.3, we can establish the following relationship:

$$\hat{I}(R_G(P), t) + I(N, t) = \frac{1}{\frac{Q_r}{L+N_p+N_d} + N_{rnu} + N_q} + N_c \quad (11)$$

The above equation can be written as:  $\hat{I}(R_G(P), t) + I(N, t) = I(S, t)$ ,  $N$  represents the total noise,  $I(N, t)$  represents the intensity variation caused by noise, and  $I(S, t)$  represents the intensity variation capturing the real spike stream. In fact, the deviation matrix  $R(x, y)$  [63] corresponding to the response nonuniformity noise can be obtained by capturing a uniform light scene and recording the intensity. So we use the matrix  $R(x, y)$  to simulate noise embedding. Choosing the pixel  $(x_m, y_m)$  which is closest to the average response value as the reference pixel.  $R(x, y)$  is obtained by calculating the ratio of the response value of a reference pixel to the response values of other pixels. Finally, we can rewrite Equation (11) as  $\hat{I}(R_G(P), t) \cdot R(x, y) = I(S, t)$ .

**Spike neuron layer.** Based on Section 3.2, we construct the process of converting scene light intensity into spike stream using an integrate-and-fire mechanism [11] in the spike camera as a spike neuron layer (Fig. 3). We denote it by  $SNL$  and represent its discrete form as:

$$A_t = A_{t-1} + I_{in}(t), \quad S_t = Thr(A_t) \quad (12)$$



**Fig. 3:** Diagram of the backpropagation process in spiking neurons based on 3D Gaussians. After passing through the rasterization pipeline, three-dimensional Gaussian ellipsoids are projected onto a two-dimensional plane, resulting in two-dimensional Gaussian ellipsoids (rendered pixels) with corresponding learnable parameters  $\theta$  ( $\mathbf{u}'$ ,  $\Sigma'$ ,  $c$ ,  $\alpha$ ). Using  $N = T$  as the step size (window size) and discrete time steps as units, the predicted light intensity values are input into the spiking neurons. Each spiking neuron receives the accumulated value  $A_t$  from the previous moment and the current input  $\hat{I}(R_G(P), t)$  (summed as  $V_t$ ). The output consists of a spike stream  $S$  and the residual voltage in the accumulator after reset, which is used as the input for the next neuron.

Where  $A_t$  represents the value of the accumulator at time  $t$ ,  $A_{t-1}$  represents the value of the accumulator at the previous time step, and  $I_{in}(t)$  represents the input value at time  $t$ . Function  $Thr$  is represented by formula (8). When  $A_t$  is greater than or equal to the threshold, it outputs 1; otherwise, it outputs 0. Note that if the accumulator's value exceeds the threshold, it releases a potential and then resets itself by subtracting the threshold from its own value.

**Spike stream rendering loss.** To measure the difference between the generated spike stream and the input spike stream, we propose a spike stream rendering loss function based on 3D Gaussians:

$$L = (1 - \lambda) \|\hat{S}(G, t) - S_{gt}\|_1 + \lambda \left( \hat{S}(G, t), S_{gt} \right)_{D-SSIM} \quad (13)$$

Where  $\hat{S}(G, t) = SNL(\hat{I}(R_G(P), t) \cdot R(x, y))$ . Note that for the Structural Similarity assessment, we transposed the shape of the spike stream to  $N \times H \times W$ , where  $N$  is the spike stream window size.

**Gradient Derivation.** In this section, we will derive the backpropagation process of our SpikeGS model. Based on Section 3.1, we will uniformly represent the learnable parameters of the 2D Gaussian ellipsoid using  $\theta$ . where  $\theta = [\mathbf{u}' \Sigma' c \alpha]$ . We set the time step of the spiking camera to  $N$  (in our experiments,  $N$  is set to 256), corresponding to the total time  $T$ . As shown in Fig. 3, after rasterizing the Gaussian ellipsoid, we use  $\sum_t^T \hat{I}(R_G(P), t)$  to represent the total light intensity over the entire time step  $T$ . The light intensity at each step is processed by the spiking neurons to generate the corresponding spike flow at that moment. The accumulated spike flows from all time steps result in the final output spike stream  $S_{out}$ .

In Fig. 3 (To distinguish the input and output values of the spiking neurons, for convenient we denote  $V_{t_i} = A_{t_{i-1}} + I_{t_i}$ . ) we can see that each spiking neuron has two output values: the accumulated value of the accumulator at the next time step  $A_t$  (including the reset case) and the corresponding transmit signal (emit spike)  $O_t^l$ . The accumulated value  $A_t$  of the accumulator continues to be combined with the light intensity  $\hat{I}_{in}(t)$  at the next step as input for the next spiking neuron.

The network is unrolled for all discrete steps, and the total gradient for all discrete steps is calculated. The gradient transmission formula is shown below:

$$\frac{\partial L_{total}}{\partial \hat{I}_{in}} = \sum_i^N \frac{\partial L_{total}}{\partial O_{t_i}^l} \frac{\partial O_{t_i}^l}{\partial V_{t_i}} \frac{\partial V_{t_i}}{\partial \hat{I}_{in}(t_i)}, \quad \frac{\partial O_{t_i}^l}{\partial V_{t_i}} = Thr'(V_{t_i}) \quad (14)$$

Since  $\frac{\partial O_{t_i}^l}{\partial V_{t_i}}$  itself is non-differentiable, we refer to [37] and use the surrogate gradient method to compute it. The calculation of the learnable parameters for a 2D Gaussian ellipsoid is as follows:

$$\frac{\partial L_{total}}{\partial \theta} = \begin{bmatrix} \frac{\partial L_{total}}{\partial \hat{I}_{in}} \frac{\partial \hat{I}_{in}}{\partial \mathbf{u}'} & \frac{\partial L_{total}}{\partial \hat{I}_{in}} \frac{\partial \hat{I}_{in}}{\partial \Sigma'} & \frac{\partial L_{total}}{\partial \hat{I}_{in}} \frac{\partial \hat{I}_{in}}{\partial \alpha'} & \frac{\partial L_{total}}{\partial \hat{I}_{in}} \frac{\partial \hat{I}_{in}}{\partial c} \end{bmatrix} \quad (15)$$

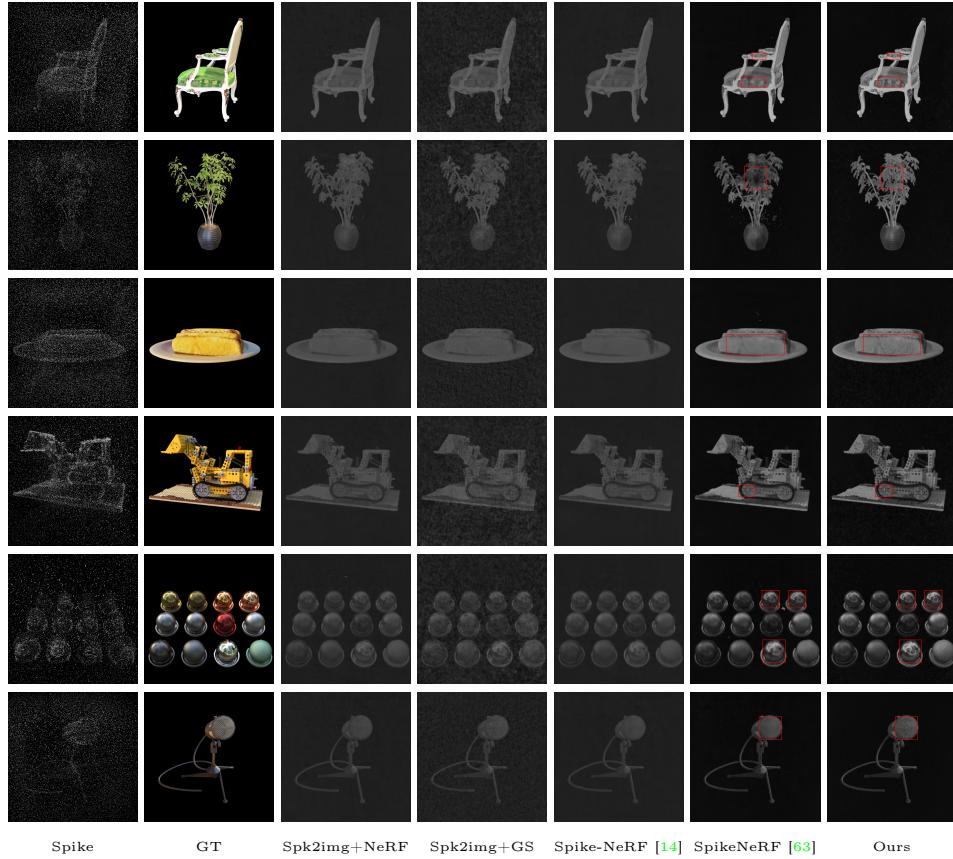
Subsequently, we can derive the gradients of the learnable parameters for the 3D Gaussian ellipsoids. For a detailed explanation of the gradient propagation process from 2D to 3D, please refer to the original 3D Gaussian splatting paper [23].

In this section, we propose the first learnable spike stream generation pipeline based on 3D Gaussians. Next, we will present the advanced results of our model on both synthetic and real-world datasets in the experimental section.

## 4 Experiments

### 4.1 Experimental Settings

**Benchmark Datasets.** We used the spike dataset provided by SpikeNeRF [63] to evaluate our model. The synthetic dataset includes six scenes (chair, ficus, hotdog, lego, materials and mic), with each scene comprising 100 images from different viewpoints. This synthetic dataset is generated based on NeRF’s synthetic dataset using the spike generator provided in [64]. The original size of the synthetic dataset is 800x800. The real-world dataset includes four scenes (dolls, box, toys, grid), each with 35 images from different viewpoints. This dataset is recorded using a handheld spike camera, capable of capturing spike stream at a spatial resolution of 250x400 and a temporal resolution of 20 KHz. **Baselines.** Due to the relative lack of methods for novel view synthesis based on spike camera, we compared our approach with four spike cameras image reconstruction methods: Spk2img+NeRF, Spk2img+GS, SpikeNeRF [63] and Spike-NeRF [14] (This paper is not open-source, therefore, we manually reproduced



**Fig. 4:** Qualitative comparison of our model with other methods on the synthetic dataset. The scenes depicted above, from top to bottom, are "chair", "ficus", "hotdog", "lego", "materials" and "mic".

the method. Since the [14] paper does not include experiments on real-world datasets, we do not test [14] on real datasets for the sake of fairness). Spk2img (Spk2imgNet) [60] is a neural network-based learning method capable of recovering images from event streams. We combined Spk2imgNet with NeRF and 3DGS model. Spk2imgNet first recovers multi-view images from the event stream, and then these images are input into the NeRF model and the 3DGS model for novel view synthesis and comparison.

**Training Details.** For both synthetic and real datasets, the inputs consist of the corresponding spike stream and poses. Since 3D Gaussian inputs require an initial point cloud, we randomly generate the initial point cloud for the synthetic dataset. For the real dataset, we use the sparse point cloud (points3D) provided by SpikeNeRF and exported by COLMAP. Our experiments were conducted on a single NVIDIA 4090, with 20K iterations for Spike2img+GS and our model,



**Fig. 5:** Qualitative comparison of our model with other methods on the real datasets under different lighting conditions. The scenes depicted above, from top to bottom, are "dolls"(high light intensity)", "box"(lowest illumination), "toys"(moderate illumination), and "grid"(moderate illumination) .

and 200K iterations for Spk2img+NeRF, SpikeNeRF and Spike-NeRF. Notably, our framework is approximately 15 times faster than SpikeNeRF, and its memory consumption is only half that of NeRF-based methods.

**Experimental Evaluation Metrics.** In the synthetic dataset, we use PSNR, SSIM [51], and LPIPS [59] as our experimental evaluation metrics to quantify the distance between the synthesized novel views and the ground truth RGB images. For the real dataset, which lacks corresponding ground truth images, we employ NIQE [15] [35] and BRISQUE [13] [34] as no-reference image quality evaluation metrics. Below, we present the quantitative and qualitative comparisons of our model on both synthetic and real datasets. In the table, each color shading indicates the best and second-best result.

## 4.2 Quantitative and Qualitative Results

Table 1 and 2 present the quantitative comparison results for the synthetic and real-world data, respectively (Note that since Spike-NeRF [14] is not suitable for extremely noisy spike data and does not incorporate noise embedding or spiking neurons, its final performance is only comparable to Spike2img+NeRF). We tested our model’s performance with different window sizes (128, 256, 512). Our model outperforms existing methods across almost all metrics in each scene,

**Table 1:** Quantitative comparison for novel view synthesis on the real dataset.

Method	Brisque $\downarrow$	NIQE $\downarrow$	Time $\downarrow$
Spk2img+NeRF(200K)	47.45	25.08	>3 hours
Spk2img+GS(30K)	19.59	16.91	$\sim$ 3 mins
SpikeNeRF [63](200K)	39.91	27.22	>8 hours
Ours N = 128(30K)	23.39	16.15	$\sim$ 20 mins
Ours N = 512(30K)	20.40	20.41	>1 hours
Ours N = 256(30K)	13.11	19.04	$\sim$ 30mins

**Table 2:** Quantitative comparison for novel view synthesis on the synthetic dataset.

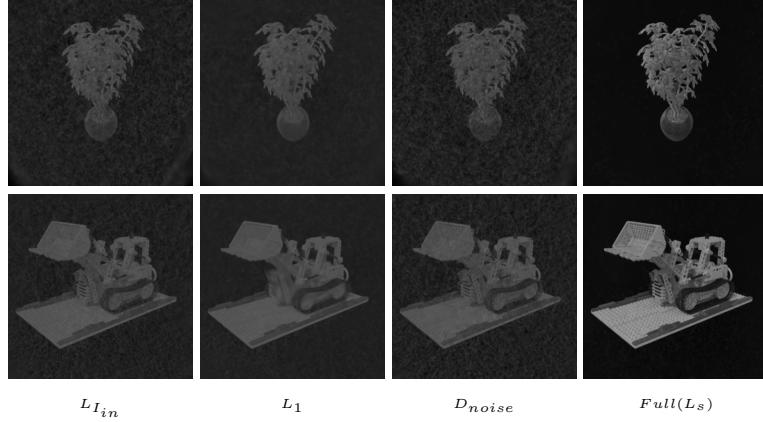
Method	Chair	Ficus	Hotdog	Lego	Materials	Mic	Time $\downarrow$
	PSNR $\downarrow$ SSIM $\uparrow$ LPIPS $\downarrow$						
Spk2img+NeRF(200K)	14.47 .0876 .2242	17.21 .0487 .1964	15.48 .1702 .2351	14.73 .1188 .3057	16.93 .1386 .2314	17.78 .0682 .1826	>3 hours
Spk2img+GS(30K)	14.30 .0870 .4953	16.85 .0498 .5345	15.28 .1636 .5015	14.56 .1190 .5319	16.51 .1281 .5260	17.40 .0644 .5031	$\sim$ 5 mins
Spike-NeRF [11](200K)	14.44 .0865 .2240	17.21 .0486 .1975	15.47 .1701 .2346	14.72 .1186 .3043	16.92 .1386 .2247	17.79 .0677 .1780	>3 hours
Spike-NeRF [63](200K)	20.06 .1871 .1271	21.65 .1081 .1649	19.94 .2530 .1393	18.62 .2247 .1987	21.84 .2319 .1396	23.62 .1299 .1235	>10 hours
Ours N = 128(30K)	14.13 .1526 .3427	14.22 .0656 .4605	15.39 .1962 .3996	14.78 .1964 .4003	15.73 .1576 .4087	16.56 .0778 .4157	$\sim$ 30 mins
Ours N = 512(30K)	14.70 .5317 .1692	21.03 .5388 .1430	15.71 .5403 .1917	15.08 .5134 .2566	17.77 .5258 .1786	21.88 .5576 .1350	>2 hours
Ours N = 256(30K)	20.24 .1984 .1213	21.86 .1201 .1820	20.17 .2567 .1612	18.63 .2335 .2470	22.21 .2493 .1335	24.38 .1406 .1397	$\sim$ 40mins

meanwhile also being faster compared to NeRF-based methods. As illustrated in the last column of the two tables, our method requires only around 40 minutes(synthetic dataset) or 30 minutes(real dataset), whereas NeRF-based methods typically take over ten hours(synthetic dataset) or eight hours(real dataset). Furthermore, our method consumes only half the memory of NeRF-based methods, and our rendering speed can reach 100 FPS, whereas NeRF achieves less than 10 FPS. It is important to note that the computation time for Spk2img+NeRF and Spk2img+GS does not include the time spent on image reconstruction by the Spk2imgNet network, only the training time for NeRF and GS is considered.

Qualitative results are demonstrated both on synthetic and real-world data. Fig. 4 presents the rendered images with different methods based on synthetic data (The visualized spike images are simulated by point clouds). Our method shows significant results with noisy spike inputs. The rendered images with real data are shown in Fig. 5. In the figure are four real scenes under different lighting conditions. SpikeGS produces clearer images with fine textures than other methods.

### 4.3 Ablation Study

In this subsection, we perform ablation studies on each component of our framework. Specifically, we analyze the impact of our proposed 3D Gaussian-based spike generation pipeline and the spike rendering loss function. We denote the complete spike rendering loss function as  $L_s$ .  $L_{I_{in}}$  represents directly using estimated light intensity and reconstructed images as supervision.  $L_1$  denotes the supervision using only  $L_1$  loss. We also investigate the effect of removing the noise embedding, represented by  $D_{noise}$ . *Full* represents our complete model.



**Fig. 6:** In the qualitative comparison of the ablation study (the full results are provided in the supplementary materials), it can be observed that the images rendered using  $L_{I_{in}}$  and  $D_{noise}$  contain significant noise, while the images rendered using  $L_1$  are overly smooth with blurred details.

**Table 3:** Ablation Study for novel view synthesis on the synthetic dataset.

Method	Synthetic dataset										Real dataset	
	Chair		Ficus		Hotdog		Lego		Materials		Mic	
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
$L_{I_{in}}$	14.34	.0887	.4304	16.90	.0512	.4779	15.30	.1668	.4502	14.58	.1197	.4850
$L_1$	14.37	.0873	.2657	16.96	.0498	.2350	15.36	.1691	.2874	14.62	.1101	.3636
$D_{noise}$	14.31	.0877	.4903	16.86	.0504	.5312	15.28	.1656	.4894	14.56	.1198	.5288
$Full(L_s)$	20.24	.1984	.1213	21.86	.1201	.1820	20.17	.2567	.1612	18.63	.2335	.2470
										22.21	.2493	.1335
										24.38	.1406	.1397
										13.11		16.15

## 5 Conclusion

This paper introduces SpikeGS, the work to learn 3D Gaussian fields solely from spike stream. we propose a novel rendering framework based on spike stream. We model the generation process of spike stream using 3DGS and embed the spike generation pipeline into the differentiable rasterization process of 3DGS, deriving the backpropagation accordingly. Additionally, we introduce a novel loss function for spike stream. Our model can recover clear novel views with fine details from extremely noisy spike stream under low-quality lighting conditions, using only the spike stream as supervision. We demonstrate the effectiveness of our approach on both synthetic and real datasets.

## References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
2. Cannici, M., Scaramuzza, D.: Mitigating motion blur in neural radiance fields with events and frames. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9286–9296 (2024)
3. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19457–19467 (2024)
4. Chen, G., Wang, W.: A survey on 3d gaussian splatting. arXiv preprint [arXiv:2401.03890](https://arxiv.org/abs/2401.03890) (2024)
5. Chen, K., Chen, S., Zhang, J., Zhang, B., Zheng, Y., Huang, T., Yu, Z.: Spikereveal: Unlocking temporal sequences from real blurry inputs with spike streams. arXiv preprint [arXiv:2403.09486](https://arxiv.org/abs/2403.09486) (2024)
6. Chen, Z., Wang, F., Wang, Y., Liu, H.: Text-to-3d using gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21401–21412 (2024)
7. Dai, G., Wang, Z., Xu, Q., Cheng, W., Lu, M., Shi, B., Zhang, S., Huang, T.: Spikenvs: Enhancing novel view synthesis from blurry images via spike camera. arXiv preprint [arXiv:2404.06710](https://arxiv.org/abs/2404.06710) (2024)
8. Dong, S., Zhu, L., Xu, D., Tian, Y., Huang, T.: An efficient coding method for spike camera using inter-spike intervals. arXiv preprint [arXiv:1912.09669](https://arxiv.org/abs/1912.09669) (2019)
9. Duwek, H.C., Shalumov, A., Tsur, E.E.: Image reconstruction from neuromorphic event cameras using laplacian-prediction and poisson integration with spiking and artificial neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1333–1341 (2021)
10. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
11. Gerstner, W., Kistler, W.M., Naud, R., Paninski, L.: Neuronal dynamics: From single neurons to networks and models of cognition. Cambridge University Press (2014)
12. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5354–5363 (2024)
13. Guha, R.: Blind/referenceless image spatial quality evaluator (brisque). San Francisco(CA): github; <https://github.com/rehanguha/Brisque> (2021)
14. Guo, Y., Bai, Y., Hu, L., Liu, M., Guo, Z., Ma, L., Huang, T.: Spike-nerf: Neural radiance field based on spike camera. arXiv preprint [arXiv:2403.16410](https://arxiv.org/abs/2403.16410) (2024)
15. Gupta, P.: Niqe for iqa in python. San Francisco(CA): github; <https://github.com/guptapraful/niqe> (2019)
16. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1730–1739 (2020)

17. Hidalgo-Carrió, J., Gallego, G., Scaramuzza, D.: Event-aided direct sparse odometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5781–5790 (2022)
18. Hu, L., Zhao, R., Ding, Z., Ma, L., Shi, B., Xiong, R., Huang, T.: Optical flow estimation for spiking camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17844–17853 (2022)
19. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. arXiv preprint [arXiv:2403.17888](https://arxiv.org/abs/2403.17888) (2024)
20. Huang, T., Zheng, Y., Yu, Z., Chen, R., Li, Y., Xiong, R., Ma, L., Zhao, J., Dong, S., Zhu, L., et al.: 1000× faster camera and machine vision with ordinary devices. Engineering **25**, 110–119 (2023)
21. Hwang, I., Kim, J., Kim, Y.M.: Ev-nerf: Event based neural radiance field. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 837–847 (2023)
22. Jiang, Y., Tu, J., Liu, Y., Gao, X., Long, X., Wang, W., Ma, Y.: Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5322–5332 (2024)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023)
24. Lee, B., Lee, H., Sun, X., Ali, U., Park, E.: Deblurring 3d gaussian splatting. arXiv preprint [arXiv:2401.00834](https://arxiv.org/abs/2401.00834) (2024)
25. Lee, D., Lee, M., Shin, C., Lee, S.: Dp-nerf: Deblurred neural radiance field with physical scene priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12386–12396 (2023)
26. Lin, J., Li, Z., Tang, X., Liu, J., Liu, S., Liu, J., Lu, Y., Wu, X., Xu, S., Yan, Y., et al.: Vastgaussian: Vast 3d gaussians for large scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5166–5175 (2024)
27. Lin, S., Zhang, Y., Huang, D., Zhou, B., Luo, X., Pan, J.: Fast event-based double integral for real-time robotics. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 796–803. IEEE (2023)
28. Liu, T., Wang, G., Hu, S., Shen, L., Ye, X., Zang, Y., Cao, Z., Li, W., Liu, Z.: Fast generalizable gaussian splatting reconstruction from multi-view stereo. arXiv preprint [arXiv:2405.12218](https://arxiv.org/abs/2405.12218) (2024)
29. Ma, L., Li, X., Liao, J., Zhang, Q., Wang, X., Wang, J., Sander, P.V.: Deblurnerf: Neural radiance fields from blurry images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12870 (2022)
30. Massa, R., Marchisio, A., Martina, M., Shafique, M.: An efficient spiking neural network for recognizing gestures with a dvs camera on the loihi neuromorphic processor. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2020)
31. Matsuki, H., Murai, R., Kelly, P.H., Davison, A.J.: Gaussian splatting slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18039–18048 (2024)
32. Mildenhall, B., Hedman, P., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T.: Nerf in the dark: High dynamic range view synthesis from noisy raw images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16190–16199 (2022)

33. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **65**(1), 99–106 (2021)
34. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **21**(12), 4695–4708 (2012)
35. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal processing letters* **20**(3), 209–212 (2012)
36. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)* **41**(4), 1–15 (2022)
37. Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine* **36**(6), 51–63 (2019)
38. Niedermayr, S., Stumpfegger, J., Westermann, R.: Compressed 3d gaussian splatting for accelerated novel view synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10349–10358 (2024)
39. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5480–5490 (2022)
40. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6820–6829 (2019)
41. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9054–9063 (2021)
42. Qi, Y., Zhu, L., Zhang, Y., Li, J.: E2nerf: Event enhanced neural radiance fields from blurry images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 13254–13264 (2023)
43. Rebecq, H., Gehrig, D., Scaramuzza, D.: Esim: an open event camera simulator. In: *Conference on robot learning*. pp. 969–982. PMLR (2018)
44. Rudnev, V., Elgharib, M., Theobalt, C., Golyanik, V.: Eventnerf: Neural radiance fields from a single colour event camera. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4992–5002 (2023)
45. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4104–4113 (2016)
46. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7495–7504 (2021)
47. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P.P., Barron, J.T., Kretzschmar, H.: Block-nerf: Scalable large scene neural view synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8248–8258 (2022)
48. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dy-

- namic scene from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
49. Wang, J., He, J., Zhang, Z., Sun, M., Sun, J., Xu, R.: Evggs: A collaborative learning framework for event-based generalizable gaussian splatting. arXiv preprint [arXiv:2405.14959](https://arxiv.org/abs/2405.14959) (2024)
  50. Wang, Y., Li, J., Zhu, L., Xiang, X., Huang, T., Tian, Y.: Learning stereo depth estimation with bio-inspired spike cameras. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022)
  51. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
  52. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
  53. Xiangli, Y., Xu, L., Pan, X., Zhao, N., Rao, A., Theobalt, C., Dai, B., Lin, D.: Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In: European conference on computer vision. pp. 106–122. Springer (2022)
  54. Xiong, T., Wu, J., He, B., Fermuller, C., Aloimonos, Y., Huang, H., Metzler, C.A.: Event3dgs: Event-based 3d gaussian splatting for fast egomotion. arXiv preprint [arXiv:2406.02972](https://arxiv.org/abs/2406.02972) (2024)
  55. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20331–20341 (2024)
  56. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
  57. Yu, W., Feng, C., Tang, J., Jia, X., Yuan, L., Tian, Y.: Evagaussians: Event stream assisted gaussian splatting from blurry images. arXiv preprint [arXiv:2405.20224](https://arxiv.org/abs/2405.20224) (2024)
  58. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19447–19456 (2024)
  59. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
  60. Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T.: Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11996–12005 (2021)
  61. Zhao, J., Xiong, R., Xie, J., Shi, B., Yu, Z., Gao, W., Huang, T.: Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Transactions on Computational Imaging* **8**, 12–27 (2021)
  62. Zhao, J., Xiong, R., Zhang, J., Zhao, R., Liu, H., Huang, T.: Learning to super-resolve dynamic scenes for neuromorphic spike camera. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3579–3587 (2023)
  63. Zhu, L., Jia, K., Zhao, Y., Qi, Y., Wang, L., Huang, H.: Spikenerf: Learning neural radiance fields from continuous spike stream. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6285–6295 (2024)

64. Zhu, L., Zheng, Y., Geng, M., Wang, L., Huang, H.: Recurrent spike-based image restoration under general illumination. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8251–8260 (2023)
65. Zwicker, M., Pfister, H., Van Baar, J., Gross, M.: Ewa volume splatting. In: Proceedings Visualization, 2001. VIS'01. pp. 29–538. IEEE (2001)
66. Guo, Y., Hu, L., Ma, L., Huang, T.: Spikegs: Reconstruct 3d scene via fast-moving bio-inspired sensors. arXiv preprint [arXiv:2407.03771](https://arxiv.org/abs/2407.03771) (2024)
67. Zhang, J., Chen, K., Chen, S., Zheng, Y., Huang, T., Yu, Z.: Spikegs: 3d gaussian splatting from spike streams with high-speed camera motion. arXiv preprint [arXiv:2407.10062](https://arxiv.org/abs/2407.10062) (2024)

# SpikeGS: Learning 3D Gaussian Fields from Continuous Spike Stream

## Supplementary Material

### 1 Different lighting intensity experiments

We conducted experiments on three synthetic datasets (ficus, lego and materials) with varying lighting intensities to further demonstrate the generalization of our method under different illumination conditions (We set our model N to 256 for the experiment). For each scene, we performed experiments under three lighting intensity conditions: extremely low lighting, moderately low lighting, and original lighting intensity. The qualitative results are shown in Figure. 1 and Figure. 2. As can be observed, our method outperforms existing methods under all lighting conditions. Even in extremely low-light scenarios, our method is able to reconstruct complete scene structures and render fine texture details. In contrast, other methods either fail to reconstruct the complete scene structure or produce very blurred texture details, and even generate significant noise at relatively higher lighting intensities.

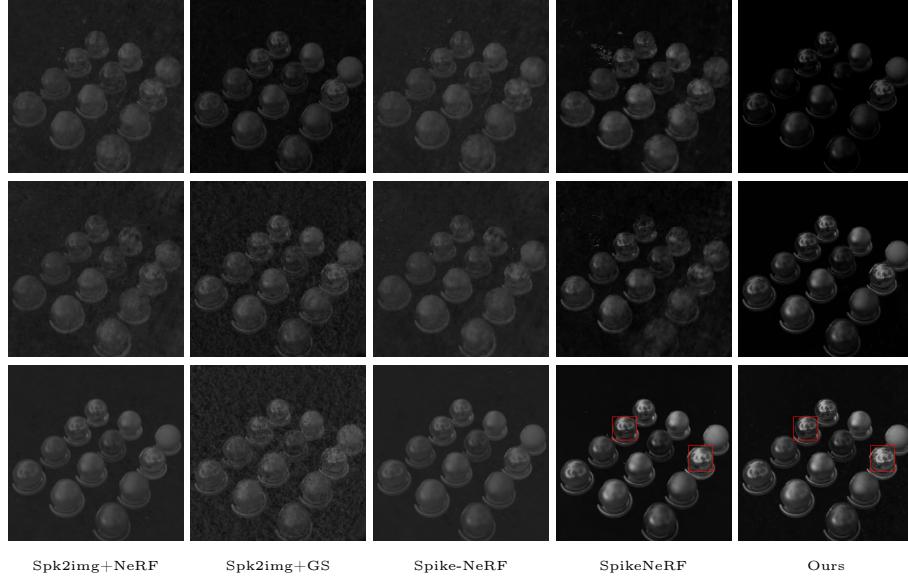
The quantitative comparison results are shown in Table 1. For each lighting condition in each scene, we calculated three metrics: PSNR, SSIM, and LPIPS (using the ground truth RGB images as reference, consistent with the full paper). As shown, our method outperforms existing methods on nearly all metrics.

### 2 Complete visualization results of the ablation experiments

According to the experimental setup described in the paper, we have provided the complete visual results for both the synthetic and real datasets in the supplementary materials. We conducted ablation experiments on 6 scenes from the synthetic dataset and 4 scenes from the real dataset. As shown in Fig. 3 and Fig. 4, when only using  $L_1$  as the loss function, the texture details of the scenes appear overly smooth. Additionally, when the noise embedding pipeline is removed or when using estimated light intensity directly (bypassing the spike generation pipeline) as input, the synthesized views exhibit a significant amount of noise. In contrast, the complete framework demonstrates high robustness to noise and is capable of recovering fine texture details(The quantitative comparison of ablation experiments is presented in Table 3 of the full paper. here, we only showcase the complete visual results(qualitative comparison)).



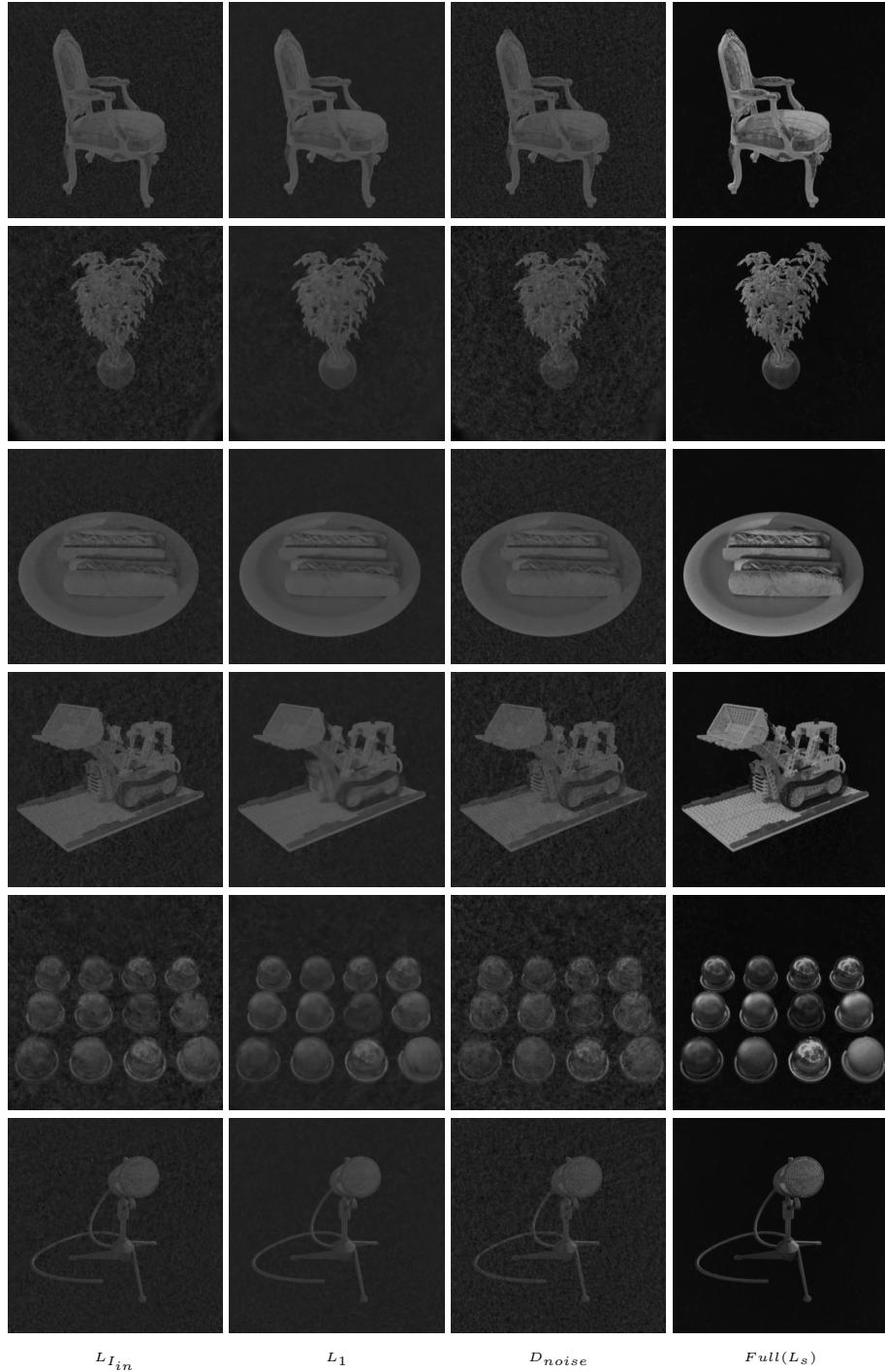
**Fig. 1:** Qualitative results on different light intensities. In the figure, every three rows represent one scene (The names of the two scenes are "ficus" and "lego"), with the first, second, and third rows corresponding to extreme low light intensity, medium low light intensity, and original light intensity, respectively. It is evident from the figure that our model consistently reconstructs the complete scene structure and fine details under all lighting conditions. In contrast, other methods often fail to reconstruct accurately and struggle to recover fine scene details under low lighting conditions, and they also produce significant noise at relatively higher lighting intensities.



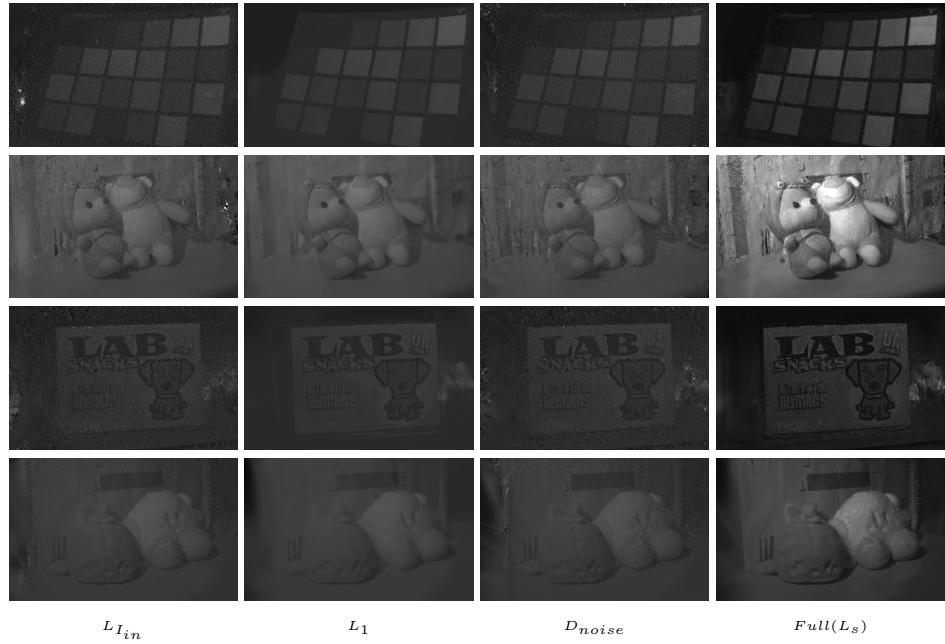
**Fig. 2:** Qualitative results on different light intensities. This figure is a continuation of Fig. 1, with the tested scene named "materials." The light intensity settings are consistent with Fig. 1. It can be observed that our model consistently recovers fine reflective details under different lighting intensities. In contrast, other methods struggle to recover fine details under lower lighting intensities and produce significant noise under relatively higher lighting intensities.

**Table 1:** Different Lighting Intensity Experiments. The terms "Light intensity (Low)," "Light intensity (Med)," and "Light intensity (Orig)" in the table correspond to extremely low lighting, moderately low lighting, and original lighting intensity, respectively. We calculated the average metrics for the three scenes (ficus, lego, and materials) under each lighting condition, and the results are shown below. Each color shading indicates the best and second-best result.

Method	Light intensity(Low)			Light intensity(Med)			Light intensity(Orig)			Time↓
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	
Spk2img+NeRF(200K)	15.73	.0786	.4162	15.83	.0788	.4582	16.92	.1020	.2445	>3 hours
Spk2img+GS(30K)	17.72	.1342	.2921	16.93	.1038	.4920	15.97	.0989	.5308	~5 mins
Spike-NeRF(200K)	15.76	.0791	.4011	15.83	.0792	.4433	16.28	.1019	.2422	>3 hours
SpikeNeRF(200K)	17.33	.1126	.3578	18.11	.1352	.3782	20.7	.1882	.1677	>10 hours
Ours(30K)	18.25	.8021	.1977	19.89	.7035	.1574	21.0	.2010	.1875	~40 mins



**Fig. 3:** Qualitative comparison of ablation experiments on the synthetic dataset. As shown in the figure, images rendered with  $L_{I_{in}}$  and  $D_{noise}$  contain noticeable noise, while images rendered with  $L_1$  are overly smooth and lack detail.



**Fig. 4:** Qualitative comparison of ablation experiments on the real dataset. As shown in the figure, images rendered with  $L_{I_{in}}$  and  $D_{noise}$  contain noticeable noise, while images rendered with  $L_1$  are overly smooth and lack detail.