3D Gaussian splatting From Unposed Spike Streams

Yijia Guo [1] Tong Hu[2] Liwen Hu [1] Lei Ma [1,2*] Tiejun Huang [1]
[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2] College of Future Technology, Peking University

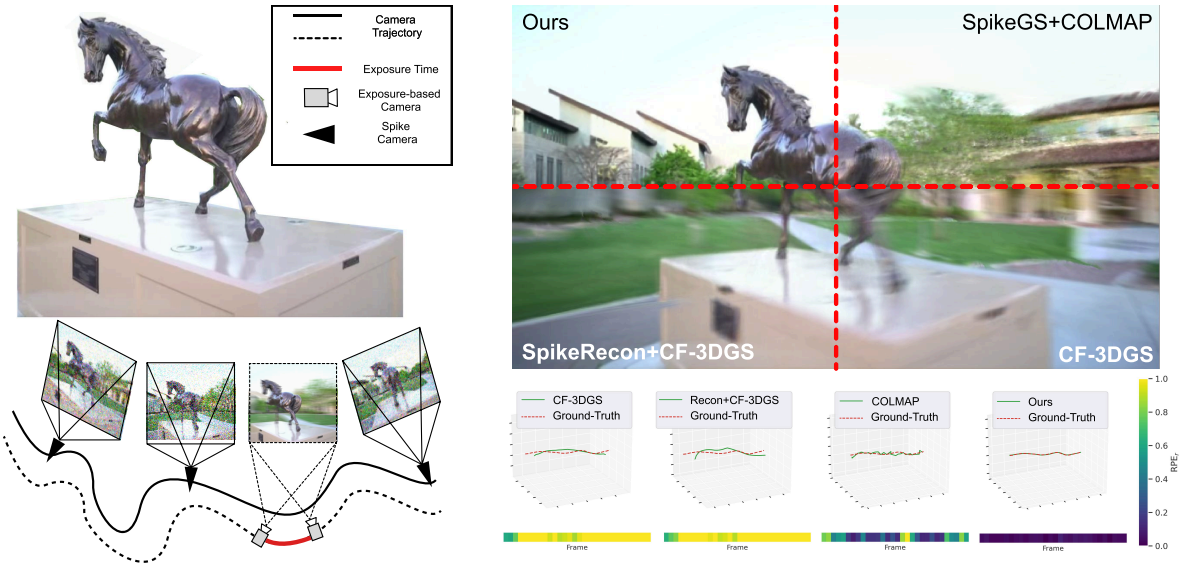

Figure 1. **Left:** An example of scene capturing under high-speed scenarios. Exposure-based camera captures discretely with the exposure window, which lead to severe motion blur. Spike camera continuously records the scene with high temporal resolution, but its output is unstable. **Right:** Comparisons of different methods. Our Nope-SGS surpasses current state-of-the-art pose-free and spike-based methods in terms of both novel view synthesis (top) and pose estimation (bottom).

Abstract

3D Gaussian Splatting (3DGS) has significantly advanced 3D reconstruction with its impressive performance. However, its reliance on sharp images and precise camera pose priors limits its effectiveness in high-speed scenarios. Recent advances have integrated spike camera, a bio-inspired sensor with a high temporal resolution, to enhance 3DGS in such conditions. Although spike-based methods reduce the need for sharp images, they still face challenges in achieving precise camera pose estimation due to unstable observations and visual texture deficiency. To address these challenges, we propose Nope-SGS, the first framework that reconstructs high-speed 3D scenes from ***unposed captures*** *of the bio-inspired high-temporal-resolution spike camera. To achieve robust 3D reconstruction and pose estimation, we first reformulate the spike model from a probabilistic perspective and extend its application to keyframing, effectively alleviating the instability caused by the spike stream. Building upon this foundation, we devise a progressive optimization framework to facilitate swift 3D reconstruction. The experimental results demonstrate that our method achieves up to 7.4dB higher PSNR and 40% lower Absolute Trajectory Error (ATE) compared to state-of-the-art methods under challenging high-speed scenarios while maintaining the fastest reconstruction speed among spike-based methods.*

1. Introduction

3D Gaussian Splatting (3DGS) [16] has revolutionized 3D reconstruction through its exceptional geometric fidelity and real-time rendering capabilities. Despite its success, methods based on 3DGS tend to fail in high-speed scenarios due to their reliance on sharp images and precise camera pose priors which are hard to obtain because of insufficient observations and motion blur. Extensive studies [3, 10, 32, 34] have introduced spike camera [36], a bio-inspired sensor with a high temporal resolution, to aid 3D reconstruction in high-speed scenarios with significant superiority. Although the spike-based methods eliminate the

*Corresponding authors

dependence on sharp images, they still struggle with precise camera pose preparation. Directly utilizing the current go-to pose estimation frameworks, such as COLMAP or VGGT [28], for spike cameras entails several inherent challenges: a) The texture details in spike camera outputs are generally less refined and exhibit significant instability compared to exposure-based cameras, as shown in Fig. 1. This can lead to unsatisfactory feature extraction, erroneous estimation results and complete failures. b) Spike cameras typically require a larger number of frames for 3D reconstruction compared to exposure-based cameras, introducing additional computational complexity.

While numerous methods [4, 7, 8] have successfully reduced the reliance of traditional 3DGS on camera poses, these approaches are specifically designed for sharp images captured by exposure-based cameras under slow camera motion. Directly applying these methods to high-speed scenarios or trivially combining them with spike cameras leads to suboptimal results, as shown in Fig. 1. In this paper, we propose Nope-SGS, a framework that eliminates the dependency of spike-based 3D reconstruction on precise camera pose priors. To achieve robust 3D reconstruction and pose estimation, we first reformulate the spike model from a probabilistic perspective, enabling the extraction of stable information from inherently unstable single-frame spike data for effective supervision. Building upon this foundation, we meticulously design optimization and keyframing strategies tailored specifically for spike cameras. These strategies are seamlessly integrated into a progressive optimization framework, facilitating rapid 3D reconstruction. Our approach not only addresses the unique challenges posed by spike stream but also significantly enhances the accuracy and efficiency of 3D reconstruction in high-speed scenarios. Our main contributions are as follows:

- We propose the first framework that reconstructs high-speed 3D scenes from unposed spike streams. Our Nope-SGS is the fastest spike-based 3D reconstruction method to the best of our knowledge.
- We first reformulate the spike camera model from a probabilistic perspective, providing a new insight for all future work based on spike cameras.
- To further evaluate the robustness of our approach, we additionally introduce a dataset captured with a new-generation spike camera.

2. Related Works

2.1. Novel View Synthesis from Unposed Sequence

Recent advances in neural scene representation have driven significant progress in eliminating SfM dependency. Neural Radiance Fields (NeRFs) pioneered this direction through joint pose-scene optimization [29], with subsequent works addressing optimization stability via frequency-domain regularization [19] and implicit spectral networks [30]. Nope-NeRF [1] further incorporates geometric priors for dynamic motion handling. However, these implicit representations face fundamental limitations in rendering speed and scene editability.. The emergence of 3D Gaussian Splatting (3DGS) [16] shifted paradigms through explicit volumetric representations, sparking new pose-free optimization strategies. CF-3DGS [8] introduces progressive Gaussian primitives growth, though struggles with motion discontinuities. While Instantsplat [7] enables rapid large-scale reconstruction through stereo priors, and ZeroGS [4] handles unordered image collections, both remain constrained to low-speed scenarios. The latest frontier explores bio-inspired sensors for high-speed reconstruction. Event-based approaches like EF-3DGS [18] and IncEventGS [14] leverage microsecond temporal resolution from event cameras, demonstrating feasibility for free-viewpoint high-speed rendering. However, their dependence on threshold-triggered events introduces inherent information loss, high-fidelity reconstruction remains an open challenge.

2.2. Novel View Synthesis on Bio-inspired Cameras

Bio-inspired vision sensors have emerged as pivotal tools for high-speed 3D reconstruction. Event-based approaches like EventNeRF [26] pioneer neural radiance field optimization under event supervision, with extensions addressing dynamic deformations [22] and motion blur [25]. Recent advancements integrate physical event formation models [20] and explicit 3D Gaussians [27, 31], achieving real-time performance. However, the requirement of SfM pose remains, which seriously affects the efficiency and accuracy of bio-inspired 3D reconstruction. EF-3DGS [18] and IncEventGS [14] integrate event data into the scene optimization process to reconstruct 3D scenes from freely moving high-speed videos, eliminating SfM poses from the training pipeline. However, the inherent lack of texture details in event data continues to limit the efficacy of these approaches. In comparison, spike cameras [6, 12, 13] are capable of capturing richer texture information. NeRF and 3DGS techniques that utilize spike cameras [9, 10, 32, 34, 38] have been shown to provide superior 3D scene reconstruction quality relative to event-based methods. All these methods require precise camera poses, which are difficult to obtain in real-world scenes. USP-Gaussian[3] mitigates erroneous reconstruction caused by imprecise pose estimation. However, USP-Gaussian still relies on relatively accurate camera pose priors and cannot handle RGB spike streams, which limits its practical applications.

3. Method

In this section, we introduce our method for reconstructing stable 3D scenes and restore camera trajectories from

unposed spike streams. We first initialize a rough 3DGS scene and imprecise camera poses using spike intervals (Sec. 3.3). Then, we leverage the normalized binomial distribution spike (Sec. 3.2) to filter out keyframes and sequentially optimize camera pose (Sec. 3.4) and scene (Sec. 3.5) using these keyframes.

3.1. Preliminary: 3D Gaussian Splatting

3DGS models the scene as a set of 3D Gaussian primitives. Each Gaussian primitive G is defined by:

$$G(x) = e^{-1/2(x)^T \Sigma^{-1}(x)} \tag{1}$$

where Σ is a full 3D covariance matrix and $x \in \mathbb{R}^3$ is the center (mean) of Gaussian primitive. Gaussian primitive can also be represented utilizing a rotation matrix denoted as R and a scale matrix denoted as S, permitting independent optimization of both components.

$$\Sigma = RSS^T R^T \tag{2}$$

In rendering, the splatting technique is used to position Gaussian primitives on the camera planes. The final color rendered can be obtained by alpha-blending N ordered overlapping points:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{3}$$

Where c_i is the color of each point and α_i is given by evaluating a 2D Gaussian.

3.2. Rethinking Spike Camera Model

In the spike camera model, each pixel functions by continuously capturing incident light signals, converting them into corresponding current signals, and integrating these input currents over time. This mechanism allows the sensor to efficiently capture dynamic scenes with exceptional temporal resolution and sensitivity. As illustrated in Fig. 2, for pixel $\mathbf{x} = (x, y)$, if the accumulation of input current reaches a fixed threshold ϕ, a spike is fired and then the accumulation can be reset as:

$$A(\mathbf{x}, t) = \hat{A}(\mathbf{x}, t) \bmod \phi = \int_0^t L_C(\mathbf{x}, \tau) d\tau \bmod \phi, \tag{4}$$

where $\hat{A}(\mathbf{x}, t)$ is the accumulation at time t, $A_{\mathbf{x}}(t)$ is the accumulation without reset before time t, $L_C(\mathbf{x}, \tau)$ is the input current of a certain color at time τ (proportional to light intensity). The spike camera generates a spatiotemporal binary stream as its output, where each binary value represents the occurrence or absence of a spike at a specific pixel location and time instance. This output can be formally represented as: $S \in \{0, 1\}^{H \times W \times N}$. The H and W

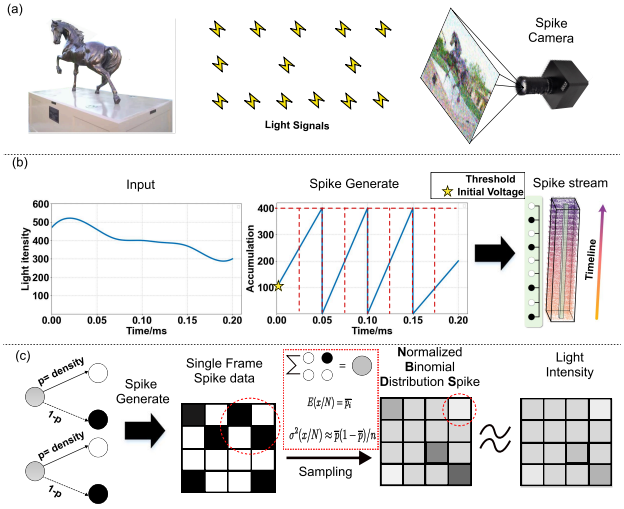

Figure 2. **(a)** Working principle of spike camera. **(b)** Coding mechanism of spike camera. Note the existence of random initial voltage. **(c)** Spike streams from probabilistic perspective. Due to the presence of initial voltage, single frame spike stream can be modeled as Bernoulli distribution and can be further sampled as an approximate binomial distribution. Its expectation is approximately equal to the light intensity with relatively low variance.

are the height and width of the sensor, and N is the temporal window size of the spike stream. Since a single frame spike stream is binary in nature, consisting solely of values 0 and 1, it lacks rich and stable visual texture information. This limitation restricts its direct applicability for 3D reconstruction and camera trajectory estimation. In this paper, we reformulate the spike camera model from probability theory and introduce a novel approach for recovering coarse visual textures from single-frame spike data, which enables robust camera trajectory and scene optimization. As illustrated in Fig. 2, in real-world scenarios, each pixel in a spike camera exhibits a random initial voltage V_X following a uniform distribution $V_x \sim U[0, \phi]$. Eq. 4 can be rewritten as:

$$A(\mathbf{x}, t) = (\int_0^t L_C(\mathbf{x}, \tau) d\tau \ + V_x) \bmod \phi, \tag{5}$$

Therefore, the output of any pixel in a given frame of the spike stream $s(x, y, k)$ can be modeled as a Bernoulli distribution $s(x, y, k) \sim B(1, p(x, y, k))$. Here:

$$p(x, y, k) = (\int_t^{t+\mathbb{T}} L_C(\mathbf{x}, \tau) d\tau \ )/\phi, \tag{6}$$

A proof is provided in Appendix. C. Where $\mathbb{T}$ is designated time interval. The existence of the V_x introduces temporal independence between different pixels. Consequently, the temporal summation of M adjacent pixels yields an approximate binomial distribution $B[n, p]$ if we assume $p(x, y, k) \approx p(x + \delta, y + \delta, k) \approx \overline{p}$. The expectation for

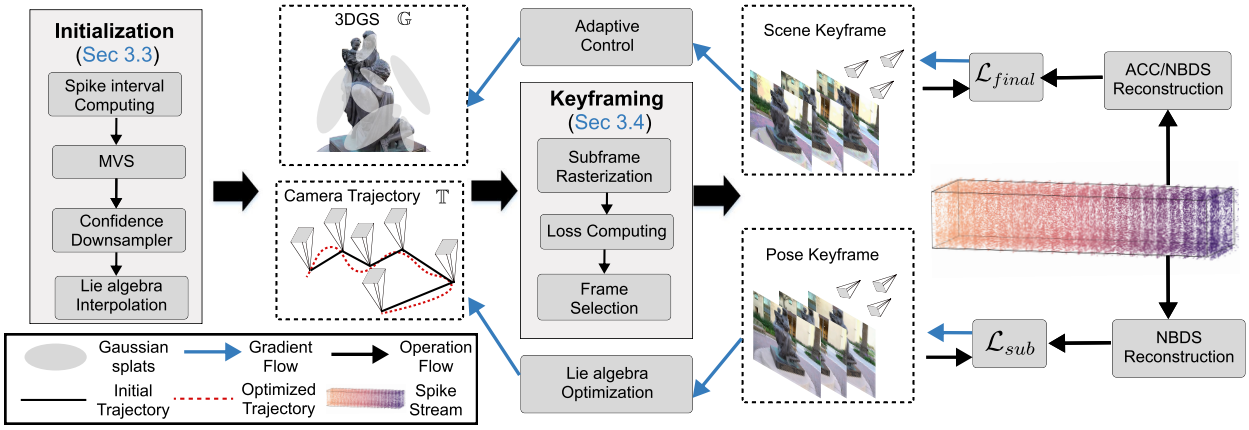

Figure 3. Overview of our Nope-SGS. We first initialize a rough 3DGS scene and imprecise camera trajectory (solid line) from sparse spike intervals. Then, we apply the normalized binomial distribution spike (NBDS, Sec. 3.2) to filter out keyframes and progressively optimize camera trajectory (Sec. 3.4) and scene (Sec. 3.5) based on keyframes. By leveraging the dense spike stream, we refine the trajectory into a smooth motion path (dashed line) and reconstruct a high-quality 3DGS scene.

this new distribution is:

$$E = \sum_{i \in N} p_i = \sum_{i \in N} \left(\int_{t}^{t+\mathbb{T}} L_C^i(\mathbf{x}, \tau) d\tau \right) / \phi \tag{7}$$

Here L_C^i is the light intensity of pixel i. And if we average them:

$$E(x/N) = \overline{p_i} = \overline{\left(\int_{t}^{t+\mathbb{T}} L_C^i(\mathbf{x}, \tau) d\tau \right) / \phi} \tag{8}$$

and

$$\sigma^2(x/N) \approx \overline{p}(1-\overline{p})/n \tag{9}$$

We define the result of the above operation as **N**ormalized **B**inomial **D**istribution **S**pike (NBDS), $\hat{S} \in [0,1]^{H \times W}$. NBDS provides a more reliable supervisory signal than raw single-frame spikes, as it effectively mitigates the inherent noise and instability associated with spike streams. We provide additional implementation details of NBDS in subsequent sections and demonstrate its effectiveness.

3.3. Initialization

Similar to Instantsplat [7], we integrate Multi-View Stereo (MVS) with 3DGS to initialize 3D Gaussian primitives from sparse spike intervals. Following [10, 37], the spike interval is defined as:

$$I_{in} = \frac{\phi}{t_1 - t_2} \tag{10}$$

The entire initialization process only takes about 1 minute. More details are provided in Appendix. F.2. Through our initialization process, we generate an approximate sparse camera trajectory $\mathbb{T}_{sparse}$ and an initial set of Gaussian primitives $\mathbb{G}$. We utilize linear interpolation in the Lie algebra to estimate the dense but inaccurate camera trajectory $\mathbb{T}$, as shown in Fig. 3. These initial estimates undergo iterative optimization to achieve progressively higher accuracy, ultimately converging to precise reconstruction results through our refinement pipeline, mathematically formulated as

$$\{\mathbb{G}, \mathbb{T}\} \rightarrow \{\mathbb{G}^*, \mathbb{T}^*\} \tag{11}$$

Our optimization goal is

$$\{\mathbb{G}^*, \mathbb{T}^*\} = \arg\min_{\mathbb{G},\mathbb{T}} \mathcal{L}(\mathbb{G}, \mathbb{T}) \tag{12}$$

We will provide a detailed explanation on how to choose a loss function $\mathcal{L}$ and accelerate the optimization process.

3.4. Camera Pose Optimization

Optimization. In the monocular case, we aim to identify the optimal camera pose $P_\theta \in \mathbb{T}$ that minimizes the following photometric errors:

$$\mathcal{L}_{pho}^k = |C(P_\theta^k) - I_{gt}^k| \tag{13}$$

where $C(P_\theta)$ is the rendering result of the Gaussian primitives $\mathbb{G}$ from $P_\theta \in SE(3)$ and k represents the frame number. The primary challenge we face lies in the inherent instability of the available spike streams, which precludes their direct utilization for the computation of photometric errors. To address this challenge, we successfully recovered a stable supervisory signal (NBDS) from the unstable spike stream avoiding introducing any bias, as elaborated in detail in Sec. 3.2. Our photometric errors can be rewritten as:

$$\mathcal{L}_{nbds}^k = |\hat{C}(P_\theta^k)) - (\hat{S}_{gt}^k)| \tag{14}$$

Where $\hat{S}_{gt}$ represents the NBDS frame and $\hat{C}(P_\theta)$ is the average pooling results of $C(P_\theta)$ to align with NBDS. To further enhance the stability of the camera trajectory, we use the optimized camera trajectory to calibrate the current trajectory to be optimized. Photometric errors can be further rewritten as:

$$\mathcal{L}_{sub}^{k} = |(\hat{C}(P_\theta^k) - \hat{C}(P^{k-n})) - (\hat{S}_{gt}^k - \hat{S}_{gt}^{k-n})| \quad (15)$$

where $\hat{C}(P^{k-n})$ is the rendering result from camera pose $P^{k-n} \in SE(3)$. n is a hyperparameter. This method not only facilitates the stabilization of the supervisory signal but also amplifies error multiplication at critical positions, such as edges, while simultaneously reducing the optimizer's emphasis on rendering errors. Furthermore, considering the inherent noise in spike cameras, as demonstrated by [11], the independence of $\hat{S}_{gt}^k$ and $\hat{S}_{gt}^{k-n}$ is guaranteed when the temporal interval n is sufficiently large. Same as [23], we use Lie algebra to derive minimal Jacobians, aligning their dimensionality with the system's degrees of freedom to eliminate redundancy and improve computational efficiency of gradient. Appendix. D shows more details.

Keyframing. Since using all the frames from spike stream to optimize the camera poses is infeasible and unnecessary, we carefully select frames that deviate significantly from the correct trajectory and optimize them. Specifically, we first compute the $\mathcal{L}_{nbds}$ for all frames based on Eq. 14, this process only takes a few seconds. Then we select frames with significantly higher than average photometric errors $\mathbb{F}_{key} = \{f^k, ...|\mathcal{L}_{nbds}^k > \delta * mean(\mathcal{L}_{nbds})\}$. δ is a hyperparameter and $f^k = P^k, S^k$. The photometric error is caused by two parts: incorrect Gaussian set and incorrect camera pose. We further select frames with significantly higher than average photometric errors caused by camera pose error $\mathbb{F}_{pose} = \{f^k, ...|\mathcal{L}_{sub}^k > \delta * mean(\mathcal{L}_{sub})$ based on Eq. 15. $\mathbb{F}_{pose}$ will be used for camera pose optimization, and $\mathbb{F}_{scene} = \overline{\mathbb{F}_{pose}} \cap \mathbb{F}_{key}$ will be used for scene optimization.

3.5. Scene Optimization

The inherent instability of visual textures in single-frame spike data renders it unsuitable for direct scene optimization. To mitigate this issue, we adopt NBDS (Sec. 3.4) to derive stable supervision signals. However, relying solely on NBDS incurs texture distortion. Inspired by SpikeGS [10], we use a spike stream accumulation mechanism to preserve high-frequency structural information:

$$I_{acc}(t_1, t_N) = \phi / N \sum_{t_i} S(P_i) \quad (16)$$

where I_{acc} denotes the accumulated results over N frames. To align with the temporal characteristics of spike accumulation, we render multiple synthetic frames C_{acc} through differentiable rendering. More details are provided in Appendix. F.1. Our final composite loss combines $\mathcal{L}_{acc}$ (long-term temporal modeling) and $\mathcal{L}_{nbds}$ (short-term detail preservation), balancing complementary aspects of the optimization:

$$\mathcal{L}_{final} = \lambda_{acc}\mathcal{L}_{acc} + \lambda_{nbds}\mathcal{L}_{nbds} \quad (17)$$

Each loss component integrates photometric and structural constraints:

$$\mathcal{L}_* = (1 - \lambda_1)\|C_* - I_*\|_2 + \lambda_1 SSIM(C_*, I_*) \quad (18)$$

Here C_* is the rendering results,I_*represents supervision signals (accumulated spikes I_{acc} and $*$ represents for acc and sample) and $* \in \{acc, nbds\}$. Finally, we employ keyframes $\mathbb{F}_{pose}$ $\mathbb{F}_{scene}$ to compute the pose estimation loss $\mathcal{L}_{sub}$ and the scene loss $\mathcal{L}_{final}$, progressively optimizing the scene and camera trajectories respectively to obtain stable motion estimates and precise geometric details. This operation is not applicable to pose-dependent methods, such as Spikegs and USP-Gaussian. As a result, in scenes with severely inaccurate poses, these methods produce rendered images with significant parallax.

4. Experiment

4.1. Experimental Settings

Datasets. We conduct extensive experiments on different synthetic and real-world datasets, including Tanks and Temples [17], Deblur-NeRF dataset[21], USP-Gaussian [3] dataset and our Nope-SGS dataset. **Tanks and Temples**: We evaluate novel view synthesis quality and pose estimation accuracy on 7 scenes covering both indoor and outdoor scenes. Similar to [3, 33, 35], we employ frame interpolation techniques to simulate high-speed camera motion within scene-specific video sequences while utilizing SOTA spike simulation frameworks [11] to generate corresponding spike streams. **Deblur-NeRF dataset:** Following SpikeGS [10], we generate spike streams using three synthetic scenes from the Deblur-NeRF dataset [21]. We configure the camera frame rate in Blender to match the spike camera's specifications while rendering corresponding images. **USP-Gaussian dataset:** To evaluate our method's performance in real-world high-speed scenarios, we employ the USP-Gaussian dataset, which captures spike data through rapid camera shaking. **Nope-SGS dataset:** To evaluate the real-world robustness of our method, we collected an additional dataset consisting of 8 indoor and outdoor scenes using the high-resolution Spike M1K40-H4-Gen3 camera (1000 × 1000). We adhere to the USP-Gaussian acquisition procedure to ensure consistency and comparability. All scenes are captured in under one second with fast camera motion throughout the sequence. All the data will be released to the public.

Figure 4. Qualitative comparison for novel view synthesis. Our approach produces more realistic rendering results with fine-grained details. **Better viewed when zoomed in.**

Metrics. For novel view synthesis, we employ quantitative metrics including PSNR, SSIM, and LPIPS for synthetic data. NIQE and IL-NIQE are applied for real-world data since no ground truth is provided. Regarding camera pose estimation, we quantify performance using both absolute and relative error metrics, specifically reporting ATE and RPE. For depth estimation, we quantify performance using δ and AbsRel.

Baselines. We first compare our method with pose-free methods: CF-3DGS [8], ht3DGS[15] and Instantsplat [7]. Then, we compare our method against spike-based 3D reconstruction methods, such as SpikeGS [10], Spike-nerf [9] and USP-Gaussian [3] with camera poses generated by COLMAP and VGGT [28]. Finally, we benchmark our approach against the combination of spike image reconstruction and pose-free methods: Spikerecon [37]+CF-3DGS/Instantsplat. Note that USP-Gaussian outputs only grayscale results. Explanations are listed in Appendix. B.

Implementation Details. All experiments were conducted on NVIDIA A800 GPU. For hyperparameters, we set $n = 32$, $M = 16$ and $\delta = 1.0$, other hyperparameters are same as 3DGS. For progressive optimization, we perform 500 iterations of camera pose optimization and 1000 iterations of scene optimization per epoch.

4.2. Experimental Results

Novel View synthesis Results: We report the comparison results of NVS tasks on Tanks and Temples datasets and Deblur-NeRF datasets in Tab. 1. Our method consistently outperforms the others across all metrics. On average, our method achieves **a +7.4dB higher PSNR, 30% higher SSIM and a 64% lower LPIPS** compared to the best performance among all baselines on Tanks and Temples datasets. Qualitative results in Fig. 4 demonstrate that our approach not only resolves the significant motion blur and geometric inconsistencies inherent in traditional pose-free methods under high-speed conditions but also mitigates the loss of geometric details caused by the inability of spike-based methods. More visualization results are provided in the appendix, we further conduct the visual comparison on real-world data in Fig. 5 and Tab. 2 to highlight the necessity of eliminating the dependency of spike-based 3D reconstruction on camera pose. Vanilla spike-based 3D reconstruction approaches, including SpikeGS, demonstrate significant limitations in generating accurate 3D scene reconstructions when processing erroneous and unstable camera trajectories. Our method demonstrates superior reconstruction quality across diverse spike stream patterns.

Pose Estimation: The quantitative results for camera pose accuracy are presented in Tab. 1. The learned camera poses of all methods are post-processed and aligned with ground truth using a consistent method following [1, 8, 24]. The results show that the camera pose accuracy of our method consistently not only outperforms all previous pose-free methods but also is superior to Colmap and VGGT results. These findings provide compelling evidence supporting our hypothesis that accurate camera pose estimation remains challenging when relying solely on blurry images, spike-based reconstruction results, or single-frame spike streams. Our method represents a significant breakthrough in addressing this fundamental limitation, demonstrating superior performance across all evaluated scenarios. Qualitative results in Fig. 7 demonstrate that our approach not only produces camera pose estimates with unprecedented proximity to ground truth values but also effectively mitigates motion fluctuations caused by inherent spike stream instability.

Depth Estimation: Following MVSplat [5] and PixelSplats [2], we evaluate the quality of the reconstructed geometry by assessing the predicted depth maps. As ground-truth depth maps are unavailable, we employ DepthAnythingV2 to generate pseudo depth maps. Prior to evaluation, all predicted depth maps are normalized and aligned to ensure fair comparison. The results in Tab. 3 and Fig. 6 demonstrate that our method significantly outperforms previous

Datasets	Tanks						Deblur-NeRF					
Methods|Metrics	PSNR ↑	SSIM ↑	LPIPS ↓	RPE_t↓	RPE_r↓	ATE↓	PSNR ↑	SSIM ↑	LPIPS ↓	RPE_t↓	RPE_r↓	ATE↓
CF-3DGS	20.245	0.650	0.411	3.803	0.766	0.045	22.775	0.662	0.472	11.173	0.386	0.038
Instantsplat	21.197	0.657	0.432	1.783	0.653	0.038	22.025	0.634	0.507	8.528	2.286	0.129
ht3DGS	20.157	0.634	0.416	1.543	0.585	0.034	18.771	0.599	0.576	12.091	1.361	0.137
Spikerecon+CF-3DGS	22.375	0.695	0.380	4.027	0.817	0.039	20.513	0.562	0.517	9.739	0.345	0.079
Spikerecon+Instant	22.295	0.703	0.387	0.701	0.250	0.012	22.319	0.617	0.513	5.903	1.395	0.094
USP-Gaussian	19.09	0.555	0.474	3.423	0.357	0.035	18.699	0.523	0.567	12.712	0.474	0.400
Spikenerf	18.29	0.623	0.422	1.346	0.482	0.018	17.641	0.613	0.453	4.147	0.383	0.055
SpikeGS+VGGT	20.192	0.688	0.343	1.185	0.274	0.012	24.143	0.693	0.353	3.481	0.352	0.073
SpikeGS+Colmap	20.801	0.691	0.339	1.346	0.482	0.018	22.255	0.682	0.412	4.147	0.383	0.055
Ours	30.184	0.911	0.122	0.229	0.069	0.003	28.058	0.765	0.295	1.817	0.287	0.030

Table 1. Quantitative comparisons of novel view synthesis and pose estimation. **Best** results are marked in red and the second best results are marked in yellow .

Figure 5. Visualizations on real-world datasets. Rows correspond to recordings from different generations of spike cameras (Vidar1, Spike M1K40-H4-Gen3, Spike M1K40-H4-Gen3-Color). **Better viewed when zoomed in.**

Datasets	Nope-SGS		USP	
Metrics	NIQE ↓	IL-NIQE ↓	NIQE ↓	IL-NIQE ↓
SpikeGS	9.93	87.92	10.92	88.77
USP-Gaussian	7.86	67.42	7.48	56.14
Ours	5.72	44.97	4.84	46.20

Table 2. NR-IQA metrics on Real-world dataset and USP-Gaussian dataset.

Method / Metrics	δ_1 ↑	δ_2 ↑	δ_3 ↑	AbsRel ↓
CF-3dgs	26.30	48.72	64.27	1.78
SpikeRecon+CF-3DGS	28.87	53.33	66.94	1.58
Instantsplat	52.15	75.56	85.21	0.79
SpikeRecon+Instant	59.46	79.77	87.23	0.61
USP-Gaussian	17.19	30.55	42.35	3.16
SpikeGS	18.92	34.95	48.48	2.48
Ours	67.11	85.38	91.88	0.31

Table 3. Evaluation of depth estimation.

approaches in reconstructed scene geometry quality.

Efficiency To provide a comprehensive performance evaluation, we conducted comparative timing analyses of all spike-based 3D reconstruction methods, measuring the total computational time required for complete scene reconstruction from spike streams under identical experimental conditions. Results in Fig. 8 indicate that our method achieves a 3× faster reconstruction speed compared to previous SOTA approaches, demonstrating significant computational efficiency improvements.

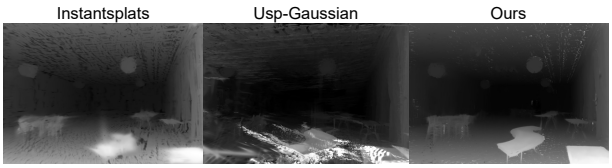

Figure 6. Visualizations of depth map

4.3. Ablation Study

Pose Optimization. We first validate the effectiveness of keyframing (ID: VI) by removing it from the optimization of camera pose. The quantitative results of pose estimation and training time in Tab. 4 demonstrate that we achieved a 10 × faster pose optimization process with almost no quality degradation in the presence of keyframing. This indicates that our keyframing scheme correctly identified the incorrect camera poses and optimized them correctly. Tab. 4 also demonstrates that the incorporation of keyframes helps

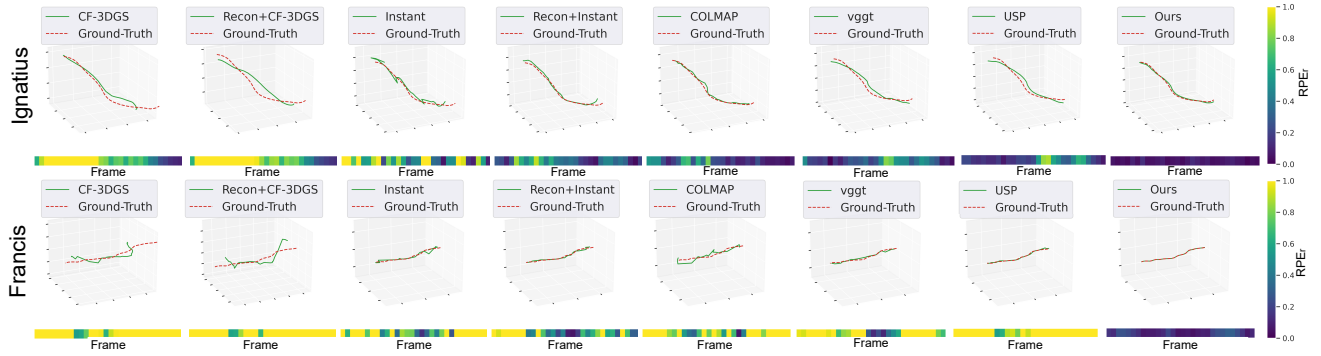

Figure 7. Pose Estimation Comparison on Tanks and Temples. We visualize the trajectory (3D plot) and RPEr (color bar) of each method.

ID	Pose Optim.				Scene Optim.		NVS			Pose Est.			Time/min
	$\mathcal{L}_{pho}$	$\mathcal{L}_{nbds}$	$\mathcal{L}_{sub}$	Key.	$\mathcal{L}_{nbds}$	$\mathcal{L}_{acc}$	PSNR↑	SSIM↑	LPIPS↓	RPE_t↓	RPE_r↓	ATE↓	
I	✗	✗	✔	✔	✗	✗	22.42	0.645	0.377	-	-	-	-
II	✗	✗	✔	✔	✔	✗	26.44	0.875	0.189	-	-	-	-
III	✗	✗	✔	✔	✗	✔	29.16	0.892	0.152	-	-	-	-
IV	✔	✗	✗	✔	✔	✔	28.59	0.872	0.147	0.387	0.129	0.007	11.5
V	✗	✔	✗	✔	✔	✔	29.69	0.902	0.126	0.285	0.075	0.004	9.0
VI	✗	✗	✔	✗	✔	✔	29.62	0.901	0.125	0.258	0.062	0.003	45.3
Ours	✗	✗	✔	✔	✔	✔	30.184	0.911	0.122	0.229	0.069	0.003	3.0

Table 4. Quantitative ablation on the Tanks and Temples dataset. Key. represent Keyframing and Time represents average optimization time per epoch.

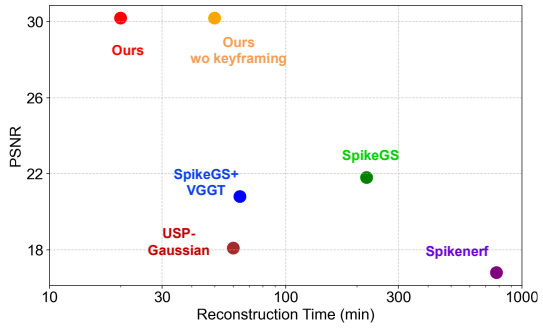

Figure 8. Comparison of rendering quality and efficiency with prior spike-based 3D reconstruction methods.

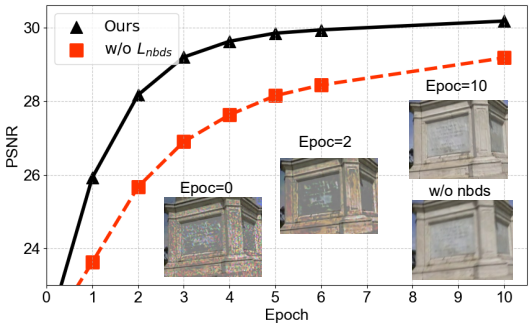

Figure 9. Progressive optimization curve and visualizations across epochs.

to minimize photometric loss and enhances the rendering of structural details, aligning closely with our design objectives. We also validate the effectiveness of $\mathcal{L}_{pho}$ (ID: IV), $\mathcal{L}_{nbds}$ (ID: V) and $\mathcal{L}_{sub}$ (ours). The quantitative results in Tab. 4 depict that $\mathcal{L}_{sub}$ yields significant improvement over $\mathcal{L}_{pho}$ and $\mathcal{L}_{nbds}$ with 33% lower ATE and 3 $\times$ faster optimization process. This indicates that our subframe scheme successfully recovered a stable supervisory signal from the unstable spike stream while accelerating the convergence of the optimization process. **Comparative videos in the supplementary materials** demonstrate that the camera trajectory is more stable with $\mathcal{L}_{sub}$.

Scene optimization and progressive optimization. We further validate the effectiveness of different losses in the scene optimization process. We remove $\mathcal{L}_{nbds}$ (ID: II), $\mathcal{L}_{acc}$ (ID: III) and both of them (ID: I) during scene optimization. by removing them respectively. Tab. 4 demonstrates that both $\mathcal{L}_{nbds}$ and $\mathcal{L}_{acc}$ accumulation loss are important to the scene optimization. Visualization details and optimization curve in Fig. 9 indicate that $\mathcal{L}_{nbds}$ facilitates accelerated convergence and enhances the recovery of geometric details. The curve depicted in Fig. 9 also illustrates how the progressive optimization process systematically transforms unstable scenarios into stable configurations.

5. Conclusion

This paper presents Nope-SGS, the first framework that enables spike-based 3D reconstruction without camera pose priors. Our progressive optimization framework and specially designed strategies for spike streams achieve end-to-end reconstruction, outperforming existing spike-based and pose-free methods on multiple datasets.

6. Acknowledgments

This work was supported by National Science and Technology Major Project (Grant No. 2022ZD0116305) and the Beijing Natural Science Foundation (Grant Nos. F251020 and JQ24023).

References

[1] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 2, 6

[2] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 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*, pages 19457–19467, 2024. 6

[3] Kang Chen, Jiyuan Zhang, Zecheng Hao, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Usp-gaussian: Unifying spike-based image reconstruction, pose correction and gaussian splatting. *arXiv preprint arXiv:2411.10504*, 2024. 1, 2, 5, 6

[4] Yu Chen, Rolandos Alexandros Potamias, Evangelos Ververas, Jifei Song, Jiankang Deng, and Gim Hee Lee. Zerogs: Training 3d gaussian splatting from unposed images. *arXiv preprint arXiv:2411.15779*, 2024. 2

[5] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 6

[6] Siwei Dong, Tiejun Huang, and Yonghong Tian. Spike camera and its coding methods. *arXiv preprint arXiv:2104.04669*, 2021. 2

[7] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2, 4, 6

[8] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 2, 6

[9] Yijia Guo, Yuanxi Bai, Liwen Hu, Mianzhi Liu, Ziyi Guo, Lei Ma, and Tiejun Huang. Spike-nerf: Neural radiance field based on spike camera. *arXiv preprint arXiv:2403.16410*, 2024. 2, 6

[10] Yijia Guo, Liwen Hu, Lei Ma, and Tiejun Huang. Spikegs: Reconstruct 3d scene via fast-moving bio-inspired sensors. 2024. 1, 2, 4, 5, 6

[11] Liwen Hu, Lei Ma, Yijia Guo, and Tiejun Huang. Scsim: A realistic spike cameras simulator. 2024. 5

[12] Liwen Hu, Ziluo Ding, Mianzhi Liu, Lei Ma, and Tiejun Huang. Learning to robustly reconstruct dynamic scenes from low-light spike streams. In *European Conference on Computer Vision*, pages 88–105. Springer, 2025. 2

[13] Liwen Hu, Yijia Guo, Mianzhi Liu, Yiming Fan, Rui Ma, Shengbo Chen, Lei Ma, and Tiejun Huang. Learn to enhance sparse spike streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2026. 2

[14] Jian Huang, Chengrui Dong, and Peidong Liu. Inceventgs: Pose-free gaussian splatting from a single event camera. *arXiv preprint arXiv:2410.08107*, 2024. 2

[15] B Ji and A Yao. Sfm-free 3d gaussian splatting via hierarchical training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21654–21663, 2025. 6

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 1, 2

[17] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36 (4):1–13, 2017. 5

[18] Bohao Liao, Wei Zhai, Zengyu Wan, Tianzhu Zhang, Yang Cao, and Zheng-Jun Zha. Ef-3dgs: Event-aided free-trajectory 3d gaussian splatting. *arXiv preprint arXiv:2410.15392*, 2024. 2

[19] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021. 2

[20] Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18335–18346, 2023. 2

[21] 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. 5

[22] Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Deformable neural radiance fields using rgb and event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3590–3600, 2023. 2

[23] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. 5

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 6

[25] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. E2nerf: Event enhanced neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13254–13264, 2023. 2

[26] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single colour event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4992–5002, 2023. 2

[27] Jiaxu Wang, Junhao He, Ziyi Zhang, Mingyuan Sun, SUN Jingkai, and Renjing Xu. Evggs: A collaborative learning framework for event-based generalizable gaussian splatting. In *Forty-first International Conference on Machine Learning*. 2

[28] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 6

[29] 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

[30] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. *arXiv preprint arXiv:2210.04553*, 2022. 2

[31] Tianyi Xiong, Jiayi Wu, Botao He, Cornelia Fermuller, Yiannis Aloimonos, Heng Huang, and Christopher A Metzler. Event3dgs: Event-based 3d gaussian splatting for fast egomotion. *arXiv preprint arXiv:2406.02972*, 2024. 2

[32] Jinze Yu, Xin Peng, Zhengda Lu, Laurent Kneip, and Yiqun Wang. Spikegs: Learning 3d gaussian fields from continuous spike stream. In *Proceedings of the Asian Conference on Computer Vision*, pages 4280–4298, 2024. 1, 2

[33] Jiyuan Zhang, Shanshan Jia, Zhaofei Yu, and Tiejun Huang. Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 137–147, 2023. 5

[34] Jiyuan Zhang, Kang Chen, Shiyan Chen, Yajing Zheng, Tiejun Huang, and Zhaofei Yu. Spikegs: 3d gaussian splatting from spike streams with high-speed camera motion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9194–9203, 2024. 1, 2

[35] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11996–12005, 2021. 5

[36] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1432–1437, 2019. 1

[37] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1432–1437. IEEE, 2019. 4, 6

[38] Lin Zhu, Kangmin Jia, Yifan Zhao, Yunshan Qi, Lizhi Wang, and Hua Huang. Spikenerf: Learning neural radiance fields from continuous spike stream. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6285–6295, 2024. 2