

# AE-NeRF: Augmenting Event-Based Neural Radiance Fields for Non-ideal Conditions and Larger Scenes

Chaoran Feng<sup>1</sup>, Wangbo Yu<sup>1,2</sup>, Xinhua Cheng<sup>1</sup>, Zhenyu Tang<sup>1</sup>, Junwu Zhang<sup>1</sup>,  
Li Yuan<sup>1,2\*</sup>, Yonghong Tian<sup>1,2\*</sup>

<sup>1</sup>School of Electronic and Computer Engineering, Peking University, China

<sup>2</sup>Peng Cheng Laboratory, China

{chaoran.feng, wbyu, chengxinhua, zhenyutang, junwuzhang}@stu.pku.edu.cn, { yuanli-ece, yhtian}@pku.edu.cn

## Abstract

Compared to frame-based methods, computational neuro-morphic imaging using event cameras offers significant advantages, such as minimal motion blur, enhanced temporal resolution, and high dynamic range. The multi-view consistency of Neural Radiance Fields combined with the unique benefits of event cameras, has spurred recent research into reconstructing NeRF from data captured by moving event cameras. While showing impressive performance, existing methods rely on ideal conditions with the availability of uniform and high-quality event sequences and accurate camera poses, and mainly focus on the object level reconstruction, thus limiting their practical applications. In this work, we propose AE-NeRF to address the challenges of learning event-based NeRF from non-ideal conditions, including non-uniform event sequences, noisy poses, and various scales of scenes. Our method exploits the density of event streams and jointly learn a pose correction module with an event-based NeRF (e-NeRF) framework for robust 3D reconstruction from inaccurate camera poses. To generalize to larger scenes, we propose hierarchical event distillation with a proposal e-NeRF network and a vanilla e-NeRF network to resample and refine the reconstruction process. We further propose an event reconstruction loss and a temporal loss to improve the view consistency of the reconstructed scene. We established a comprehensive benchmark that includes large-scale scenes to simulate practical non-ideal conditions, incorporating both synthetic and challenging real-world event datasets. The experimental results show that our method achieves a new state-of-the-art in event-based 3D reconstruction.

## Introduction

The rapid advancement of 3D reconstruction techniques has enabled the generation of high-fidelity novel views from camera captures of a scene, further fostering numerous downstream applications, including robotics (Zhang et al. 2023a; Zhou et al. 2023; Kerr et al. 2022; Adamkiewicz et al. 2022), 3D games (Xia et al. 2024; Condorelli and Luigini 2024), and scene understanding (Nguyen et al. 2024; Li et al. 2024). However, in environments with suboptimal lighting or rapid object motion, standard RGB cameras often struggle to capture enough scene information and may experience overexposure, underexposure, or motion blur, making

\*Corresponding author.

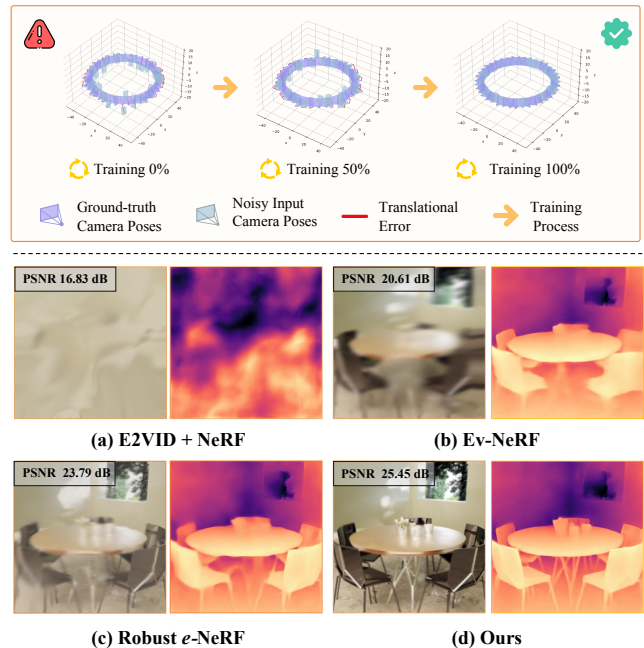


Figure 1: **Comparison of novel view synthesis (NVS) and pose correction using existing event-based methods.** The scene is captured by an event camera with 360-degree non-uniform motion and poses are estimated from COLMAP.

the captured images unsuitable for 3D scene reconstruction. In contrast, neuromorphic sensors like event cameras (Gallego et al. 2020), which detect individual changes in brightness through a sequence of asynchronous events based on polarity rather than absolute intensities, provide significant advantages in such challenging situations due to their high dynamic range and temporal resolution.

Unfortunately, integrating event cameras into 3D reconstruction techniques remains challenging because these cameras capture relative brightness changes, which cannot be directly used to reconstruct scenes in alignment with human visual perception. Some methods combine depth map (Li et al. 2023) or standard cameras with event cameras to reconstruct 3D scenes, sacrificing the advantages of high temporal resolution offered by event cameras. Other approaches use stereo visual odometry (VO)(Zhou, Gallego,

and Shen 2021) or SLAM(Gao et al. 2023) to address these issues, but they can only reconstruct sparse 3D models like point clouds. The sparsity limits their broader applicability. Additionally, another method (Nehvi et al. 2021) initially represents objects as rough templates and then updates their deformations to align with events. However, these methods rely on template initialization, cannot address the impact of inaccurate poses, limited to specific object category scenes.

Neural Radiance Fields (NeRFs) (Mildenhall et al. 2020) has revolutionized the field of 3D scene reconstruction by learning neural 3D scene representation from dense image captures. It also inspired Event-based NeRF reconstruction methods, such as Ev-NeRF (Hwang, Kim, and Kim 2023), PAEv3d (Wang et al. 2024), Event-NeRF (Rudnev et al. 2023), and Robust *e*-NeRF(Low and Lee 2023). These methods bridge NeRF with event stream (or additional RGB frame) captured by event cameras for 3D reconstruction and are desinged for handling scenes with extreme light condition or fast object motion. Nevertheless, these methods face fundamental challenges in non-ideal conditions that align with real-world scenarios. Firstly, these methods rely on ground truth poses (for synthetic data) or poses derived from complicated motion capture system (for real-world data) to train NeRF. This reliance is impractical in everyday settings, where the common approach for estimating poses from captured images is to use off-the-shelf Structure-from-Motion techniques, such as COLMAP (Schonberger and Frahm 2016). However, the accuracy of COLMAP rapidly degrades when handling low-quality RGB frames produced by event cameras, posing substantial challenges for the real applications of existing event-based NeRFs. Secondly, non-uniform camera movement is common in real-world scenarios, which can lead to inconsistencies in the density of the event stream captured by event camera, and further affect the reconstruction quality of event-based NeRFs. Additionally, existing methods mainly focus on reconstructing simple objects and suffer significant performance degradation when generalized to larger scenes. Figure 1 illustrates an example of using existing event-based NeRFs in a large scene with non-ideal conditions. It is evident that both E2VID+NeRF (Rebecq et al. 2019) and Ev-NeRF (Hwang, Kim, and Kim 2023) fail to reconstruct the 3D scene, while Robust-e-NeRF (Low and Lee 2023) also shows low fidelity.

In this work, to tackle the challenges of event-based NeRF reconstruction from non-ideal conditions and large scene, we make the following contributions:

- We propose **AE-NeRF**, a joint pose-NeRF training framework, which facilitates event-based NeRF reconstruction under various non-ideal conditions, particularly with inaccurate poses and uneven event density. As presented in Figure 1, it can effectively correct the inaccurate poses for better 3D reconstruction.
- We introduce a proposal network-based sampling strategy to address local minima optimization and large-scale generalization issues.
- We propose an event-based 3D reconstruction dataset with different complex scenes based on an improved version of ESIM(Rebecq et al. 2018), setting a benchmark

for event-based NeRF reconstruction in large scenes.

Comprehensive experiments validates that our method significantly outperforms prior state-of-the-art on both synthetic and real-world datasets.

## Related Work

### 3D Scene Representations

Prior research has investigated various methods for representing 3D scenes. Traditional approaches using explicit representations, such as point clouds (Qi et al. 2017; Achlioptas et al. 2018), meshes (Wang et al. 2018; Liu et al. 2020), and voxels (Byravan et al. 2023; Sitzmann et al. 2019), often struggle with fixed topology and limited quality in novel view synthesis. To address these issues, 3D Gaussian Splatting (3D-GS) (Kerbl et al. 2023) has been proposed, offering advantages in fast rendering speed and high-quality novel view synthesis. However, it requires point cloud initialization and does not effectively utilize the single-pixel characteristics of event stream, limiting the development of event data streams in 3D Gaussian Splatting.

NeRF (Mildenhall et al. 2020) has made significant strides by encoding continuous volumetric representations of shape and color within a multi-layer perceptron (MLP). This success has driven extensive research across computer vision, including large-scale scene reconstruction (Zhang et al. 2020; Barron et al. 2021; Tancik et al. 2022; Byravan et al. 2023), scene editing (Cheng et al. 2024), scene understanding (Zhi et al. 2021), SLAM (Zhu et al. 2024; Rosinol, Leonard, and Carlone 2023), and 3D generation (Zhang et al. 2023b; Yu et al. 2023). For unbounded or large-scale scenes, methods like Mip-NeRF360 (Barron et al. 2022) have enabled scene-level reconstruction, while Block-NeRF (Tancik et al. 2022) and BungeeNeRF (Xiangli et al. 2022) have extended this to city-scale reconstruction. The integration of single-pixel event data with NeRF through pixel ray marching leverages the high dynamic range and temporal resolution of event data, ensuring high geometric consistency and texture fidelity in the reconstructed results.

### NeRF-based Reconstruction with Event Stream

In contrast to traditional 3D reconstruction methods, the application of event cameras in NeRF-based 3D reconstruction remains underexplored. Current NeRF techniques mainly focus on dense or sparse multi-view image reconstruction (Lu et al. 2024; Zou et al. 2024), often incorporating optical flow (Wang et al. 2023a), depth maps (Deng et al. 2022), or point clouds (Truong et al. 2023). Early attempts to reconstruct NeRFs from event data include Ev-NeRF (Hwang, Kim, and Kim 2023), E-NeRF(Klenk et al. 2023), EventNeRF (Rudnev et al. 2023), and Robust *e*-NeRF (Low and Lee 2023). However, these approaches face limitations, such as reliance on precise camera trajectories in EventNeRF and the need for high-quality event streams in Ev-NeRF. In cases of inaccurate pose estimation or complex scenes, these methodologies struggle to maintain geometric consistency and texture fidelity.

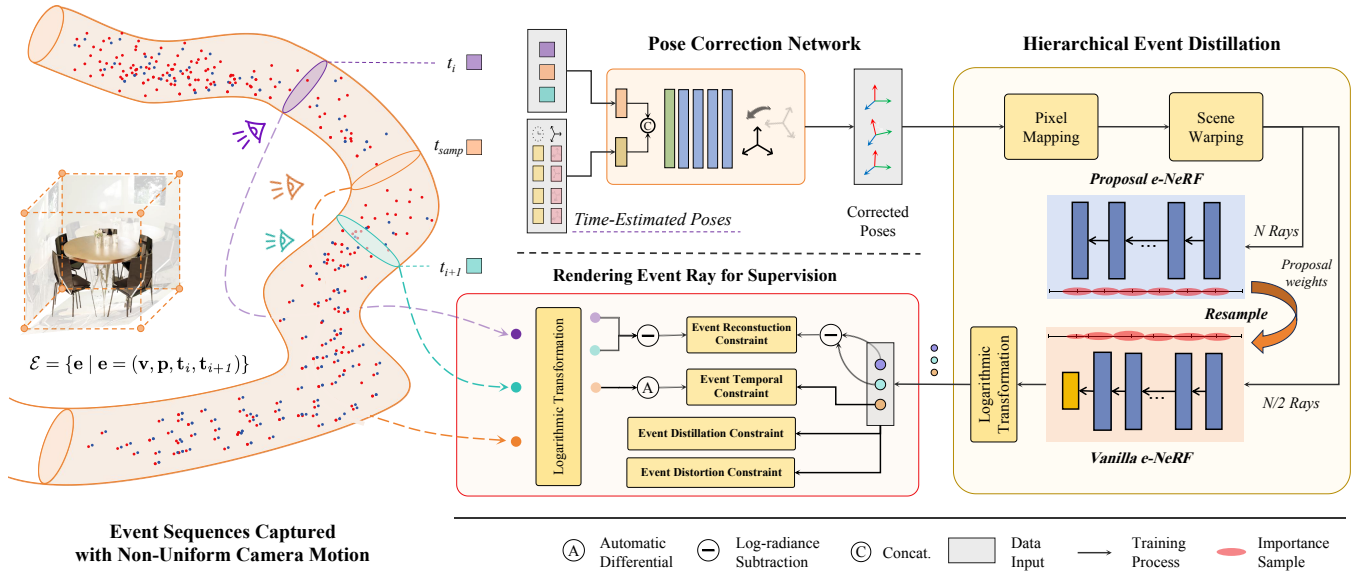


Figure 2: **Overview of AE-NeRF.** For each event  $e$  in the batch  $\mathcal{E}$ , randomly sampled from the raw sequences, we sample the timestamp  $t_{\text{samp}}$  between the previous timestamp  $t_i$  and the current timestamp  $t_{i+1}$ . We then use a pose correction network with timestamps-poses pairs to interpolate discrete poses with dense timestamps, yielding corrected poses at  $t_i$ ,  $t_{i+1}$ , and  $t_{\text{samp}}$ . With these corrected poses, we process the event ray through scene warping and apply a two-stage e-NeRF to resample weights and distances, which infers the predicted log-radiance of pixel  $\mathbf{v}$ . The predicted event reconstruction difference and temporal gradient are then computed against the ground truth, utilizing distillation loss and distortion loss for regularization. Finally, a learning-based approach is employed for color correction to refine tone mapping.

## Preliminary

### Neural Radiance Fields

Our model draws its inspiration from the NeRF approach and the neural network  $F_\theta(\cdot)$  processes a 3D coordinate  $\mathbf{x}_i \in \mathbb{R}^3$  and a ray direction  $\mathbf{d}_i \in \mathbb{S}^2$ , outputting the density  $\sigma_i \in \mathbb{R}$  and the emitted radiance  $\mathbf{c}_i \in \mathbb{R}^3$ :

$$F_\theta : (\gamma_x(\mathbf{x}_i), \gamma_d(\mathbf{d}_i)) \rightarrow (\sigma_i, \mathbf{c}_i), \quad (1)$$

Here,  $\gamma(\cdot)$  is a sinusoidal positional encoding function capturing high-frequency spatial information. Rendering each pixel involves sampling  $N$  points along a ray  $r(\mathbf{x}_0, \mathbf{d})$ , where  $\mathbf{x}_0$  is the ray's origin at the camera's focal point. The pixel color  $\hat{\mathbf{L}}(r)$  is computed as:

$$\hat{\mathbf{L}}(r) = \sum_{i=1}^N w_i \mathbf{c}_i, \delta_i = \|\mathbf{x}_{i+1} - \mathbf{x}_i\|, \quad (2)$$

$$w_i = \exp\left(-\sum_{l=1}^{i-1} \sigma_l \delta_l\right) (1 - \exp(-\sigma_i \delta_i)).$$

The ray is rendered by sampling coarse distances  $t^c$  from a uniform distribution and sorted, followed by generating coarse weights  $w^c$  with an MLP. Fine distances  $t^f$  are then sampled from the histogram with  $t^c$  and  $w^c$ , and sorted:

$$\begin{aligned} t^c &\sim \mathcal{U}[t_n, t_f], & t^c &= \text{sort}(\{t^c\}), \\ t^f &\sim \text{hist}(t^c, w^c), & t^f &= \text{sort}(\{t^f\}). \end{aligned} \quad (3)$$

What's more, we adopt the normalization trick proposed in (Barron et al. 2022) to compute the ray distance  $s$ .

### Event Generation Model

An event  $e_k = (\mathbf{v}_k, p_k, t_k)$  represents a brightness change detected by an event camera at time  $t_k$ , with pixel location  $\mathbf{v}_k = (x_k, y_k)$  and polarity  $p_k \in \{-1, +1\}$ . The polarity indicates positive or negative changes in logarithmic illumination, based on thresholds  $C^{+1}$  and  $C^{-1}$ . We adopt the event generation model from Robust e-NeRF (Low and Lee 2023), where an event camera captures log-radiance changes, producing an event stream  $\mathcal{E}$ :

$$\mathcal{E} = \{e \mid e = (\mathbf{v}, \mathbf{p}, t_i, t_{i+1})\}, \quad (4)$$

where each event records the current timestamp  $t_{i+1}$  and the previous timestamp  $t_i$  from the same pixel  $\mathbf{v}$ .

An event with polarity  $p$  is triggered when the log-radiance difference at a pixel reaches the contrast threshold  $C^p$  and the condition is expressed as:

$$\Delta \log \mathbf{L} := \log \mathbf{L}(\mathbf{v}, t_{i+1}) - \log \mathbf{L}(\mathbf{v}, t_i) = p C^p. \quad (5)$$

For color event cameras,  $\mathbf{L}$  represents the radiance of light after passing through the color filter in front of the pixel.

## Methodology

This work addresses the challenge of novel view synthesis based on event neural implicit representations, in the non-uniform motion and unbounded-scene regime. Especially, we assume access to only discontinuous input views with noisy camera pose estimates in real-world scenarios.

To correct pose estimation and obtain continuous poses, we introduce a **pose correction network**  $\psi(\cdot)$  using dense

timestamps as the main driving signal for the joint pose-NeRF training, thereby solving the challenge of imperfect poses. Moreover, to enhance eNeRF’s ability to represent unbounded scenes, we draw inspiration from (Barron et al. 2022) and utilize **hierarchical event distillation**. This approach trains two MLPs, with one resampling and predicting volumetric density and the other handling color estimation and image rendering, and thus encourages the learned scene geometry to be consistent across all viewpoints performance. Next, we propose and improve several normalization loss functions to **render event rays for supervision**, based on the event generation model. These functions generalize effectively to various real-world conditions, allowing joint optimization on randomly sampled event batches  $\mathcal{E}_{batch}$  to boost novel view rendering quality and further tackle the overfitting problem. Finally, instead of using gamma correction, we employ a learning-based approach for **color correction** to refine tone mapping. It restores photorealistic colors from relative light intensities, enhancing overall model performance. The overall pipeline is shown in Figure 2.

### Pose Correction for Continuity

Since we are accustomed to capturing RGB images and event data streams with event cameras like DAVIS346 (Taverni 2020), it is inaccurate to directly estimate poses  $\hat{P}_{\mathcal{E}}$  from RGB or gray images  $I_{\mathcal{E}}$  using COLMAP with fixed sampled time  $T_{\mathcal{E}}$ . The timestamp of each event  $t_i \in T_{\mathcal{E}}$  and image  $I_i \in I_{\mathcal{E}}$  can form a time-image pair, serving as prior information. We achieve continuous time-pose mapping through the correction network  $\psi(\cdot)$ . The time-image pairs are embedded via a sparse head, and each timestamp  $t_i$  is also embedded. As shown in Figure 3 (a), this maps time to a helical axis representation  $\psi := (t; T_{\mathcal{E}}, \hat{P}_{\mathcal{E}}) \rightarrow SE(3)$ .

This pose correction network generates a continuous 6-DoF pose as a function of time, making it highly suitable for handling asynchronous events. Unlike other event-based NeRF approaches (Low and Lee 2023) that use trajectory interpolation or turntable poses, we address the joint problem of learning neural 3D representations and optimizing imperfect event poses, similar to BARF (Lin et al. 2021). The process can be formulated as follows:

$$\hat{P}_{corr}(t_i) = \psi(t_i, T_{\mathcal{E}}, \hat{P}_{\mathcal{E}}) \quad (6)$$

where  $\hat{P}_{corr}(t_i)$  is the corrected pose at time  $t_i$ ,  $\psi(\cdot)$  is the correction module. This mapping transforms the time-image pairs into a continuous  $SE(3)$  pose representation, ensuring accurate pose estimation for asynchronous event streams.

### Hierarchical Event Distillation

The pose correction network favors a global and continuous solution that remains consistent across all training event sequences. However, the reconstructed scene often exhibits inconsistencies when viewed from novel viewpoints, due to the lack of fine sampling strategies for unbounded scenes during training. We propose a two-phase jointly optimized e-NeRF, designed to ensure that the learned geometry is consistent from any viewing direction.

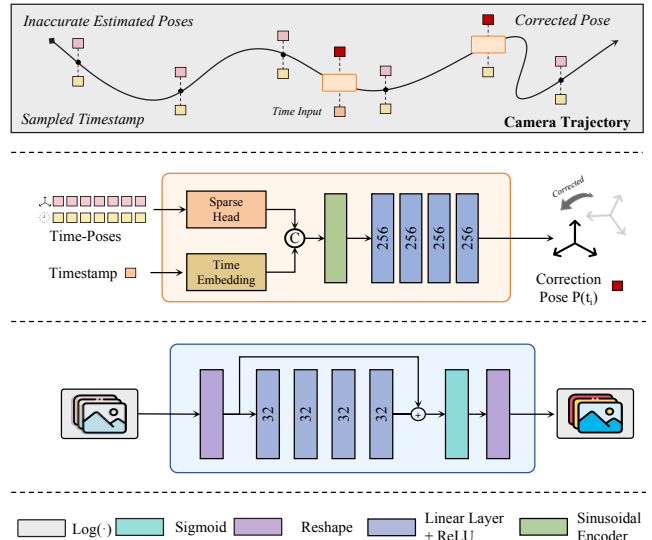


Figure 3: Framework of Pose Correction Network and Color Correction Network.

**Scene warping** It has been demonstrated that space warping functions, such as NDC warping (Mildenhall et al. 2020) and inverse sphere warping (Barron et al. 2023; Wang et al. 2023b), are effective for rendering unbounded scenes. We primarily use NDC warping for object-level scenes and employ uniform space warping  $\mathcal{C}(\cdot)$  for unbounded scenes, as defined below:

$$\mathcal{C}(\mathbf{x}) = \begin{cases} \mathbf{x} & \|\mathbf{x}\| \leq 1 \\ \left(2 - \frac{1}{\|\mathbf{x}\|}\right) \left(\frac{\mathbf{x}}{\|\mathbf{x}\|}\right) & \|\mathbf{x}\| > 1 \end{cases} \quad (7)$$

**Two-phase Jointly Optimized e-NeRF** We use a larger vanilla e-NeRF alongside a smaller proposal e-NeRF, repeatedly evaluating and resampling numerous samples from the proposal e-NeRF. This approach allows our model to exhibit higher capacity than existing e-NeRF methods, with only slightly increased training costs. Utilizing a small MLP to model the proposal distribution does not compromise accuracy, indicating that distilling the NeRF MLP is simpler and more effective than view synthesis, as shown below:

$$F_{\varphi}^p(t, \hat{w}) \circ F_{\theta}^v(t, w) : (\gamma_x(\mathcal{C}(\mathbf{x})), \gamma_d(\mathbf{d})) \rightarrow (\sigma, \mathbf{c}) \quad (8)$$

In joint optimization process, a supervised method is needed to ensure consistency between the histograms produced by the proposal e-NeRF  $F_{\varphi}^p(t, \hat{w})$  and the vanilla e-NeRF  $F_{\theta}^v(t, w)$ , as detailed in the following sections. Inspired by Mip-NeRF 360 (Barron et al. 2021), we adopt its histogram boundary function  $\mathcal{B}(\cdot)$ , which calculates the sum of proposal weights that overlap with interval  $T$ :

$$\mathcal{B}(\hat{t}, \hat{w}, T) = \sum_{j: T \cap \hat{T}_j \neq \emptyset} \hat{w}_j. \quad (9)$$

To maintain consistency between the two histograms, for all intervals  $T_i$ ,  $w_i$  within  $(t, w)$ , the condition  $w_i \leq \mathcal{B}(\hat{t}, \hat{w}, T_i)$  must be satisfied. This requirement resembles the additivity property of outer measures in measure theory.

This consistency ensures that the distributions between two-phase e-NeRFs remain relatively stable during resampling processes. With the boundary function, we can effectively constrain weights distribution of the proposal e-NeRF to match that of the vanilla e-NeRF, thus achieving a more efficient sampling process. It boosts the precision of sampling and the rendering quality during the training process.

### Rendering Event Ray for Supervision

Overfitting directly to the training event streams leads to a compromised event neural radiance field that collapses towards the provided views, even when assuming perfect camera poses. With noisy and imperfect input poses, this issue is further exacerbated, making the L2 reconstruction loss unsuitable as the primary signal for joint pose-eNeRFs training. We apply several event regularization losses, to enforce learning a globally consistent 3D solution across the optimized scene geometry and camera poses.

**Event Distillation Constraint.** As mentioned earlier, we jointly optimize the proposal e-NeRF and vanilla e-NeRF, penalizing only the proposal weights that underestimate the distribution implied by the vanilla e-NeRF. Overestimation is expected since proposal weights are generally coarser and form an upper envelope over NeRF weights. This loss is similar to a half-quadratic version of the chi-squared histogram distance used in statistics and computer vision. We introduce the event threshold  $\bar{C} = \frac{1}{2}(C^{-1} + C^{+1})$  to normalize this loss, further constraining the sampling distribution. Formally, the proposal loss is defined as:

$$\ell_p(t, w, \hat{t}, \hat{w}) = \sum_i \frac{1}{w_i} \max(0, \frac{w_i - B(\hat{t}, \hat{w}, T_i)}{\bar{C}})^2 \quad (10)$$

**Event Reconstruction Constraint.** The main idea is aimed to use the log-radiance map  $\Delta \log \hat{L} := \log \hat{L}(\mathbf{v}, t_{i+1}) - \log \hat{L}(\mathbf{v}, t_i)$  rendered from the training viewpoints to match ground truth relative light intensity  $\Delta \log L$ . Instead of L2 Reconstruction loss of NeRF, this difference is normalized using the event threshold for supervision, as follows:

$$\ell_r(e) = \frac{\text{MSE}(\Delta \log \hat{L}, \Delta \log L)}{\bar{C}^2} \quad (11)$$

Note that when using a color event camera,  $\hat{L}$  refers to the single-channel rendered radiance, where the color channel is determined by the pixel’s color filter (Low and Lee 2023).

**Event Temporal Constraint.** To accurately capture the density of event streams under non-uniform motion, we compute the error between the predicted log-radiance gradient and the target log-radiance gradient, leveraging the high sampling rate of event cameras (Gallego et al. 2020). The predicted time gradient, obtained via auto-differentiation, is represented as  $\mathcal{G}_{\text{pred}} = \frac{\partial}{\partial t} \log L(\mathbf{v}, t)$ , and is defined by:

$$\ell_g(e) = \left| \frac{\mathcal{G}_{\text{gt}} - \mathcal{G}_{\text{pred}}}{\mathcal{G}_{\text{gt}}} \right| \quad (12)$$

Here, the target gradient  $\mathcal{G}_{\text{gt}} \approx \frac{pC^p}{t_{i+1} - t_i}$  is computed as a finite difference approximation, with sampling at  $t_{\text{sam}}$ .

**Event Distortion Constraint.** However, we notice that the depth consistency of the rendered results was not satisfactory, exhibiting some pathological depth issues. Inspired by total variance regularization (Michel et al. 2011), we adapted the smoothing from neighboring pixels to an integral over the distances between all points along the normalized ray. This integral is scaled by the weights assigned to each point by the vanilla e-NeRF, enforcing self-supervised smoothness and consistency of the ray weights:

$$\ell_d(s, w) = \sum_{i,j} w_i w_j \left| \frac{s_i + s_{i+1}}{2} - \frac{s_j + s_{j+1}}{2} \right| \quad (13)$$

In short, we use four normalized loss functions to supervise the model training with sampling a batch of events  $\mathcal{E}_{\text{batch}}$  and formally, the total training loss  $\mathcal{L}$  is defined as:

$$\mathcal{L} = \frac{1}{|\mathcal{E}_{\text{batch}}|} \sum_{e \in \mathcal{E}_{\text{batch}}} (\alpha \ell_r + \beta \ell_g + \gamma \ell_p + \eta \ell_d) \quad (14)$$

### Color Correction of Synthesized Views

Event cameras capture changes in log-radiance, leading to an offset in the predicted log-radiance  $\log \hat{L}$  from the NeRF reconstruction. Additionally, consistent color channel ambiguities, akin to unknown black levels and ISO in images, arise from spectral sensitivity differences between event and standard cameras. These can be corrected reconstruction with affine adjustments based on reference images, with parameters optimized via ordinary least squares.

However, this method tends to perform poorly in scenes with multiple objects or complex textures environments. To further improve this process, we adopt a learning-based approach. As shown in Figure 3(b), we employ a color correction network  $\mathcal{F}(\cdot)$  to learn the correction process for RGB images of validation views in each scene and predict the color  $\hat{\mathbf{c}}$  of test viewpoints from log-radiance intensity:

$$\hat{\mathbf{c}} = \mathcal{F}(\log \hat{L}) \quad (15)$$

This approach allows for adaptive and precise tone mapping, capable of handling complex variations in the data.

## Experiments

We utilize novel view synthesis (NVS) as a benchmark to demonstrate that our method can effectively reconstruct NeRF from event cameras, particularly in scenes with inaccurate pose estimation and sparse, noisy data caused by non-uniform motion. The NVS benchmark tests are conducted on both synthetic and real sequences. In addition, we perform ablation studies on pose optimization and losses to assess the impact of each component in our method.

**Event Datasets.** For synthetic scenes, similar to Ev-NeRF and Event NeRF, we utilize the synthetic dataset from NeRF (Mildenhall et al. 2020) and design additional synthetic event sequences, including event data streams, estimated poses from COLMAP (Schonberger and Frahm 2016), ground truth poses, and RGB images, using Blender (Community 2018). Our dataset includes four scenes inspired by Deblur-NeRF (Ma et al. 2022) and four additional

Methods	Easy Settings			Hard Settings			Easy Settings			Hard Settings		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
E2VID+NeRF	18.92	.8328	.3167	18.92	.8328	.3167	15.60	.6212	.3298	15.60	.6212	.3298
Ev-NeRF	27.72	.9356	.0891	25.44	.8903	.1275	18.26	.7256	.2718	17.82	.6959	.2943
Event NeRF	26.81	.9183	.1007	24.92	.8783	.1377	17.99	.7048	.2881	17.68	.6933	.2996
Robust e-NeRF	<u>28.38</u>	<u>.9462</u>	.0578	<u>28.37</u>	<u>.9464</u>	.0578	21.48	.9033	.1204	21.41	.9028	.1207
<b>Ours(w.o. <math>\psi</math>)</b>	28.17	.9441	<u>.0541</u>	28.13	.9436	<u>.0550</u>	<u>22.83</u>	<u>.9105</u>	<u>.1169</u>	<u>22.79</u>	<u>.9098</u>	<u>.1172</u>
<b>Ours(w. <math>\psi</math>)</b>	<b>28.96</b>	<b>.9512</b>	<b>.0478</b>	<b>28.90</b>	<b>.9503</b>	<b>.0485</b>	<b>24.28</b>	<b>.9324</b>	<b>.0975</b>	<b>24.23</b>	<b>.9324</b>	<b>.0981</b>

Table 1: **Comparison of NVS for synthetic scenes.** Average performance is shown for object-level scenes in blue and for scenes from the proposed dataset in orange. The best and second-best results are highlighted in bold and underline.

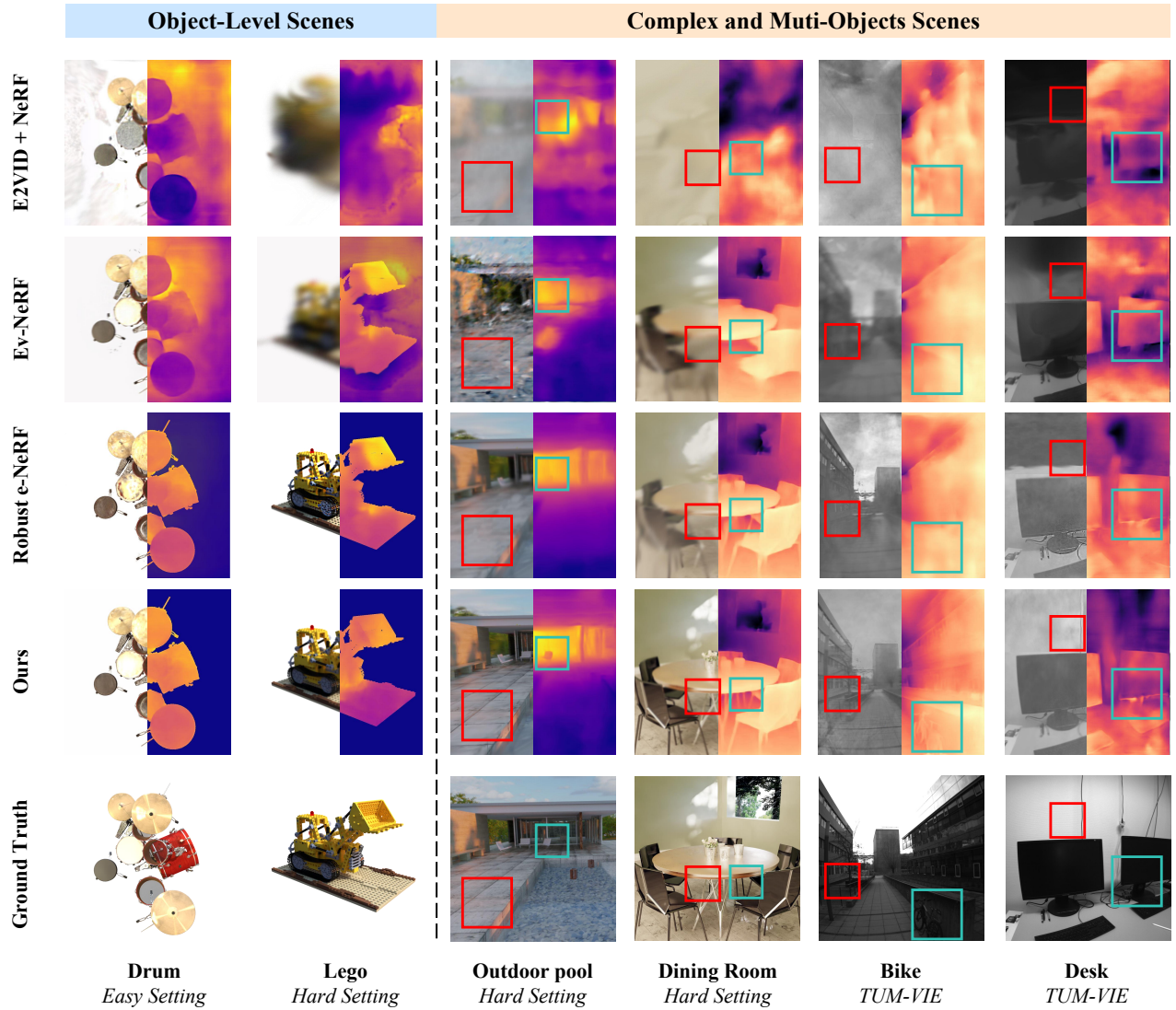


Figure 4: Qualitative Comparison of Novel View Synthesis with AE-NeRF.

custom scenes created with Blender and ESIM. These sequences are simulated in “Realistic Synthetic 360-degree”

environments, which are rich in complexity and texture, making them ideal for NVS evaluation. Following the ap-

Methods	E2VID NeRF	Ev- NeRF	Event NeRF	Robust <i>e</i> -NeRF	Ours (w. $\psi$ )
PSNR $\uparrow$	15.81	17.79	17.61	18.97	<b>19.69</b>
SSIM $\uparrow$	.6338	.7856	.7672	.8223	<b>.8470</b>
LPIPS $\downarrow$	.3072	.2138	.2203	.1964	<b>.1882</b>

Table 2: NVS Comparison the TUM-VIE dataset.

proach of Robust *e*-NeRF (Low and Lee 2023), we classify the camera trajectories into simple and challenging categories, detailed in the appendix. For real-world experiments, we use event sequences from the TUM-VIE dataset (Klenk et al. 2021), which primarily capture forward motion with a Prophesee Gen 4 event camera under linear and spiral movements, providing event data, ground truth poses, and grayscale images. To simulate real-world conditions, we also estimate images using COLMAP to obtain noisy poses.

**Baseline Methods.** We compare our method with a simple baseline method E2VID+NeRF which combines the well-known event-to-video reconstruction method E2VID with NeRF, and recent excellent works Ev-NeRF, Event NeRF and Robust *e*-NeRF.

**Evaluation Metrics.** The experimental results on both synthetic and real-world datasets are evaluated through the novel view synthesis task using quantitative metrics and qualitative comparisons of rendered images. Akin to previous studies, we use widely adopted evaluation metrics to compare synthesized images with corresponding ground truth images: PSNR, SSIM, and LPIPS to quantify the similarity between color-corrected synthesized novel views and target novel views. Moreover, we use rotation error (RE) and translation error (TE) to measure the accuracy of pose learning, further validating the model’s performance.

### Evaluation on Synthetic Event Stream

We evaluate our method on the NeRF synthetic dataset and proposed dataset comprising eight scenes. As shown in Table 1, our method demonstrates notable improvement in large-scale scenes, though its impact is less pronounced in object-level scenes. It effectively preserves structural integrity, particularly at geometric discontinuities, such as the wheel in the “*lego*” scene and the ground in the “*outdoor pool*” scene. In contrast, the baseline method introduces significant background noise in the depth maps, whereas ours achieves clearer depth representations. Furthermore, our approach produces sharper renderings, while the benchmark results appear blurrier. Quantitative and qualitative comparisons are detailed in Table 1 and Figure 4.

### Evaluation on Real Event Stream

We randomly select four scenes (*desk1*, *desk2*, *office*, *bike*) with different objects and materials for establishment. As stated in Figure 4, our approach faithfully reconstructs the main structures of objects even if there are some fog noises around them. In the scene of “*bike*”, the benchmark infers wrong geometries of the bike and leads to large variances in depth predictions. Moreover, the texture and depth map of

Scenes (w. $\psi$ )	Synthetic Easy $\checkmark$		Synthetic Hard $\times$		TUM-VIE $\checkmark$ $\times$	
RE $\downarrow$	<b>0.085</b>	0.086	<b>0.127</b>	0.881	<b>0.431</b>	1.283
TE $\downarrow$	<b>0.867</b>	0.872	<b>1.472</b>	3.751	<b>1.676</b>	4.487

Table 3: Ablation study of pose correction.

$\ell_p$	$\ell_r$	$\ell_g$	$\ell_d$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
$\checkmark$	$\checkmark$			19.80	.8356	.1811
$\checkmark$	$\checkmark$	$\checkmark$		23.19	.8852	.1494
$\checkmark$	$\checkmark$		$\checkmark$	20.37	.8413	.1775
$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$	<b>24.67</b>	<b>.8971</b>	<b>.1337</b>

Table 4: Ablation study of constrained loss.

wall in the scene of “*desk*” is wrongly rendered. In contrast, ours maintains relatively clean shapes and sharper boundaries. And we additionally compute the metrics for the four scenes and the results are listed in Table 2.

### Ablations Study

**Pose Correction.** We evaluate proposed model between the ground truth poses and corrected estimated poses in outdoor pool and diningroom with hard settings, as well as office and bike in real scenes, shown in Figure 1 and Table 3. It is evident that the pose correction network is significant in rectifying camera poses, particularly in complex environments where the camera motion is non-linear.

**Loss Functions.** We conclude the evaluation in the scenes same with pose correction ablation settings, in Table 4, the contribution of all the losses introduced above. Adding event time derivative for supervision from Eq.(12) improves PSNR by +3.39dB which is further increased when the event camera’s trajectory is irregular. Similarly, we evaluate on adding the event distortion loss in Eq.(13) and performance increases in this case, with a +0.57dB increase when adding this loss. The best performance is achieved when both are combined with an overall increase of +4.87dB in PSNR.

### Conclusion

This paper presents **AE-NeRF**, a novel approach designed to robustly reconstruct objects and scenes directly from moving event cameras under diverse real-world conditions. By leveraging the multi-view consistency of Neural Radiance Fields and the unique advantages of event cameras, our method addresses the challenges of inaccurate camera pose estimation and unbounded scene modeling through nonlinear scene parametrization, hierarchical distillation, and innovative regularizers. Comprehensive experiments on both synthetic and real-world datasets demonstrate that our method significantly outperforms state-of-the-art approaches and baseline benchmarks in terms of rendering quality and robustness. Our results highlight the effectiveness in enhancing novel view synthesis, even in complex environments. We will release our code and dataset to support further research and development in this field.

## Acknowledgments

This work is supported in part by the Natural Science Foundation of China (No. 62332002, 62202014, 62425101, 62027804, 62088102).

## References

- Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, 40–49. PMLR.
- Adamkiewicz, M.; Chen, T.; Caccavale, A.; Gardner, R.; Culbertson, P.; Bohg, J.; and Schwager, M. 2022. Vision-only robot navigation in a neural radiance world. *IEEE Robotics and Automation Letters*.
- Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *International Conference on Computer Vision (ICCV)*.
- Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Computer Vision and Pattern Recognition (CVPR)*.
- Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2023. Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields. In *International Conference on Computer Vision (ICCV)*.
- Byravan, A.; Humpalik, J.; Hasenclever, L.; Brussee, A.; Nori, F.; Haarnoja, T.; Moran, B.; Bohez, S.; Sadeghi, F.; Vujatovic, B.; et al. 2023. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9362–9369. IEEE.
- Cheng, X.; Yang, T.; Wang, J.; Li, Y.; Zhang, L.; Zhang, J.; and Yuan, L. 2024. Progressive3D: Progressively Local Editing for Text-to-3D Content Creation with Complex Semantic Prompts. arXiv:2310.11784.
- Community, B. O. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam.
- Condorelli, F.; and Luigini, A. 2024. Rapid and Low-Cost 3D Model Creation Using Nerf for Heritage Videogames Environments. In *Advances in Representation: New AI-and XR-Driven Transdisciplinarity*.
- Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised NeRF: Fewer Views and Faster Training for Free. In *Computer Vision and Pattern Recognition (CVPR)*.
- Gallego, G.; Delbrück, T.; Orchard, G.; Bartolozzi, C.; Taba, B.; Censi, A.; Leutenegger, S.; Davison, A. J.; Conradt, J.; Daniilidis, K.; et al. 2020. Event-based vision: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*.
- Gao, L.; Su, H.; Gehrig, D.; Cannici, M.; Scaramuzza, D.; and Kneip, L. 2023. A 5-point minimal solver for event camera relative motion estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8049–8059.
- Hwang, I.; Kim, J.; and Kim, Y. M. 2023. Ev-nerf: Event based neural radiance field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 837–847.
- Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*.
- Kerr, J.; Fu, L.; Huang, H.; Avigal, Y.; Tancik, M.; Ichnowski, J.; Kanazawa, A.; and Goldberg, K. 2022. Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects. In *6th annual conference on robot learning*.
- Klenk, S.; Chui, J.; Demmel, N.; and Cremers, D. 2021. Tum-vie: The tum stereo visual-inertial event dataset. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8601–8608. IEEE.
- Klenk, S.; Koestler, L.; Scaramuzza, D.; and Cremers, D. 2023. E-nerf: Neural radiance fields from a moving event camera. *IEEE Robotics and Automation Letters*, 8(3): 1587–1594.
- Li, H.; Huang, J.; Jin, P.; Song, G.; Wu, Q.; and Chen, J. 2023. Weakly-supervised 3d spatial reasoning for text-based visual question answering. *IEEE Transactions on Image Processing*, 32: 3367–3382.
- Li, H.; Zhang, D.; Dai, Y.; Liu, N.; Cheng, L.; Li, J.; Wang, J.; and Han, J. 2024. GP-NeRF: Generalized Perception NeRF for Context-Aware 3D Scene Understanding. In *Computer Vision and Pattern Recognition (CVPR)*.
- Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5741–5751.
- Liu, S.; Li, T.; Chen, W.; and Li, H. 2020. A general differentiable mesh renderer for image-based 3D reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 50–62.
- Low, W. F.; and Lee, G. H. 2023. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18335–18346.
- Lu, Z.; Zheng, Q.; Shi, B.; and Jiang, X. 2024. Pano-NeRF: Synthesizing High Dynamic Range Novel Views with Geometry from Sparse Low Dynamic Range Panoramic Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3927–3935.
- Ma, L.; Li, X.; Liao, J.; Zhang, Q.; Wang, X.; Wang, J.; and Sander, P. V. 2022. Deblur-nerf: Neural radiance fields from blurry images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12861–12870.
- Michel, V.; Gramfort, A.; Varoquaux, G.; Eger, E.; and Thirion, B. 2011. Total Variation Regularization for fMRI-Based Prediction of Behavior. *IEEE Transactions on Medical Imaging*, 30(7): 1328–1340.
- Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision (ECCV)*.

- Nehvi, J.; Golyanik, V.; Mueller, F.; Seidel, H.-P.; Elgharib, M.; and Theobalt, C. 2021. Differentiable event stream simulator for non-rigid 3d tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1302–1311.
- Nguyen, T.-A.-Q.; Bourki, A.; Macudzinski, M.; Brunel, A.; and Bennamoun, M. 2024. Semantically-aware Neural Radiance Fields for Visual Scene Understanding: A Comprehensive Review. *arXiv preprint arXiv:2402.11141*.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.
- Rebecq, H.; Gehrig, D.; Scaramuzza, D.; and Feng, C. 2018. ESIM: an open event camera simulator. In *Conference on robot learning*, 969–982. PMLR.
- Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High Speed and High Dynamic Range Video with an Event Camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*.
- Rosinol, A.; Leonard, J. J.; and Carlone, L. 2023. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *International Conference on Intelligent Robots and Systems (IROS)*.
- Rudnev, V.; Elgharib, M.; Theobalt, C.; and Golyanik, V. 2023. EventNeRF: Neural Radiance Fields from a Single Colour Event Camera. In *Computer Vision and Pattern Recognition (CVPR)*.
- Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion Revisited. In *Computer Vision and Pattern Recognition (CVPR)*.
- Sitzmann, V.; Thies, J.; Heide, F.; Nießner, M.; Wetzstein, G.; and Zollhofer, M. 2019. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2437–2446.
- Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.
- Taverni, G. 2020. *Applications of Silicon Retinas: From Neuroscience to Computer Vision*. Ph.D. thesis, Universität Zürich.
- Truong, P.; Rakotosaona, M.-J.; Manhardt, F.; and Tombari, F. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4190–4200.
- Wang, C.; MacDonald, L. E.; Jeni, L. A.; and Lucey, S. 2023a. Flow supervision for deformable nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21128–21137.
- Wang, J.; He, J.; Zhang, Z.; and Xu, R. 2024. Physical Priors Augmented Event-Based 3D Reconstruction. *arXiv preprint arXiv:2401.17121*.
- Wang, N.; Zhang, Y.; Li, Z.; Fu, Y.; Liu, W.; and Jiang, Y.-G. 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, 52–67.
- Wang, P.; Liu, Y.; Chen, Z.; Liu, L.; Liu, Z.; Komura, T.; Theobalt, C.; and Wang, W. 2023b. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4150–4159.
- Xia, H.; Lin, Z.-H.; Ma, W.-C.; and Wang, S. 2024. Video2Game: Real-time Interactive Realistic and Browser-Compatible Environment from a Single Video. In *Computer Vision and Pattern Recognition (CVPR)*.
- Xiangli, Y.; Xu, L.; Pan, X.; Zhao, N.; Rao, A.; Theobalt, C.; Dai, B.; and Lin, D. 2022. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, 106–122. Springer.
- Yu, W.; Yuan, L.; Cao, Y.-P.; Gao, X.; Li, X.; Quan, L.; Shan, Y.; and Tian, Y. 2023. Hifi-123: Towards high-fidelity one image to 3d content generation. *arXiv preprint arXiv:2310.06744*.
- Zhang, J.; Dai, L.; Meng, F.; Fan, Q.; Chen, X.; Xu, K.; and Wang, H. 2023a. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6672–6682.
- Zhang, J.; Tang, Z.; Pang, Y.; Cheng, X.; Jin, P.; Wei, Y.; Yu, W.; Ning, M.; and Yuan, L. 2023b. Repaint123: Fast and high-quality one image to 3d generation with progressive controllable 2d repainting. *arXiv preprint arXiv:2312.13271*.
- Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.
- Zhi, S.; Laidlow, T.; Leutenegger, S.; and Davison, A. J. 2021. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15838–15847.
- Zhou, A.; Kim, M. J.; Wang, L.; Florence, P.; and Finn, C. 2023. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Computer Vision and Pattern Recognition (CVPR)*.
- Zhou, Y.; Gallego, G.; and Shen, S. 2021. Event-Based Stereo Visual Odometry. *IEEE Transactions on Robotics*, 37(5): 1433–1450.
- Zhu, Z.; Peng, S.; Larsson, V.; Cui, Z.; Oswald, M. R.; Geiger, A.; and Pollefeys, M. 2024. Nicer-slam: Neural implicit scene encoding for rgb slam. In *2024 International Conference on 3D Vision (3DV)*, 42–52. IEEE.
- Zou, Y.; Li, X.; Jiang, Z.; and Liu, J. 2024. Enhancing Neural Radiance Fields with Adaptive Multi-Exposure Fusion: A Bilevel Optimization Approach for Novel View Synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7882–7890.