

# Ev-NeRF : Event Based Neural Radiance Field

Inwoo Hwang<sup>1</sup>, Junho Kim<sup>1</sup>, and Young Min Kim<sup>1,2,\*</sup>

<sup>1</sup>Department of Electrical and Computer Engineering, Seoul National University

<sup>2</sup>Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

## Abstract

We present *Ev-NeRF*, a Neural Radiance Field derived from event data. While event cameras can measure subtle brightness changes in high frame rates, the measurements in low lighting or extreme motion suffer from significant domain discrepancy with complex noise. As a result, the performance of event-based vision tasks does not transfer to challenging environments, where the event cameras are expected to thrive over normal cameras. We find that the multi-view consistency of NeRF provides a powerful self-supervision signal for eliminating the spurious measurements and extracting the consistent underlying structure despite highly noisy input. Instead of posed images of the original NeRF, the input to Ev-NeRF is the event measurements accompanied by the movements of the sensors. Using the loss function that reflects the measurement model of the sensor, Ev-NeRF creates an integrated neural volume that summarizes the unstructured and sparse data points captured for about 2-4 seconds. The generated neural volume can also produce intensity images from novel views with reasonable depth estimates, which can serve as a high-quality input to various vision-based tasks. Our results show that Ev-NeRF achieves competitive performance for intensity image reconstruction under extreme noise conditions and high-dynamic-range imaging.

## 1. Introduction

Event cameras are neuromorphic sensors, where individual pixels detect changes of brightness that exceed a threshold. The output of event cameras is a sequence of asynchronous events composed of the polarity, pixel location, and the time stamp, occurring only at a sparse set of locations where the brightness change is detected. They have many advantages over conventional cameras such as high temporal resolution, low energy consumption, and high dynamic range [15]. However, the measurements of the same object change significantly under different motion or light-

ing conditions causing domain discrepancy in real-world deployment [24, 12, 63]. While event cameras are expected to prosper under extreme environmental conditions, the performance of event-based vision tasks often deteriorates due to the significant domain shift with severe noise.

The output of event streams is very different from an ordinary image, which is a two-dimensional array with dense color values. Many existing approaches using event data compile them into a more structured form for denoising [10, 11, 30, 14, 2, 13], or directly perform downstream tasks such as motion estimation [33, 29, 36, 52] or pose estimation [34, 5]. Nonetheless, training data is often limited and the performance of event-based vision is often inferior to the performance of the same tasks with conventional images [24]. The complex noise characteristics and domain discrepancy further complicate developing practical algorithms for event cameras.

Inspired by the recent success of Neural Radiance Fields (NeRF) [32], we propose Ev-NeRF, a neural radiance field built directly from raw event data, as shown in Figure 1(a). Ev-NeRF builds a 3D volumetric representation that can concurrently explain events associated with the camera movement. Given the 5D input of location and viewing direction, NeRF outputs the volume density and emitted color, which can be aggregated to synthesize an image from an arbitrary viewpoint by the volume rendering. While NeRF is trained to minimize the color discrepancy between the synthesized image and the ground truth image, Ev-NeRF is trained with a new loss function that incorporates the sensor movement and the resulting events triggered by the difference of brightness.

Ev-NeRF properly handles the complex noise in event cameras without ground truth supervision, and at the same time, enjoys the technical advantages of the sensor over conventional cameras. The volumetric aggregation in the formulation effectively reduces the prevalent noise in event measurements [15], as the complex spatial and temporal noises lack multi-view consistency. Further, the associated intensity values in Ev-NeRF are in a high dynamic range (HDR), as the provided measurements of event cameras are sensitive to extreme lighting beyond the dynamic range of

\*Young Min Kim is the corresponding author.

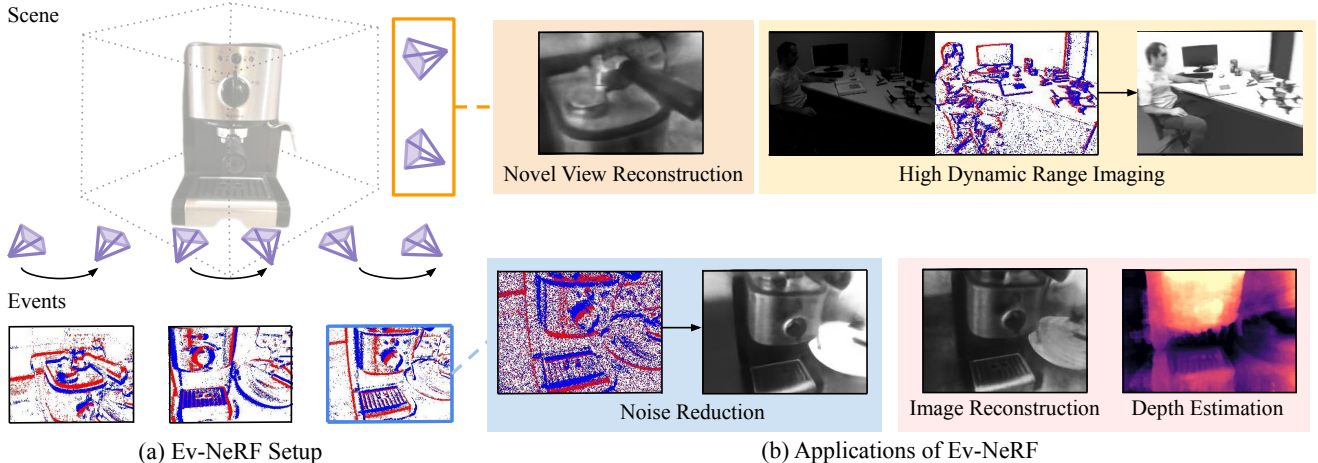


Figure 1. (a) Ev-NeRF operates with event data obtained from a moving event camera. (b) Ev-NeRF learns the implicit volume with the raw event output of the sensor and serves as a solution for various event-based applications, such as high dynamic range imaging, noise reduction, depth estimation, intensity image reconstruction, and novel-view intensity image reconstruction.

the conventional camera.

Interestingly, the created volumetric representation is a solution to many of the vision problems tackled in previous works using event data, as shown in Figure 1(b). While the event data only contain the relative changes in the brightness instead of the absolute term, the trained volume can directly synthesize the intensity image for ordinary computer vision, which is one of frequently tackled problems in the community [9, 22, 3, 45, 44, 43, 37, 53, 62, 46, 57, 6, 56]. Further, the reconstructed density volume can represent the approximate 3D structure of the scene. This is inherent from the original NeRF formulation enforcing the multi-view consistency, and the quality of 3D reconstruction is superior to the 3D structure built from previous approaches [23, 42, 65].

Our contributions can be summarized as follows:

- We suggest Ev-NeRF, which combines the popular NeRF formulation with the raw event output of a neuromorphic camera for the first time.
- Ev-NeRF is highly robust to event noise and builds a coherent 3D structure that can provide high-quality observations.
- The created neural volume serves as solutions for various event-based applications, namely intensity image reconstruction, novel-view image synthesis, 3D reconstruction, and HDR imaging.
- Ev-NeRF demonstrates performance comparable to many of existing event-vision algorithms that are dedicated to a specific task in the experimental result.

Given the strong experimental result, we expect Ev-NeRF to expand the possible application area of event-based vision

that fully leverages the potential of the sensor.

## 2. Related works

In this section, we review the key tasks in event-based vision, along with existing work on neural implicit 3D representations.

**Processing Event Data** Although event cameras can acquire visual information in challenging conditions such as low-lighting or extreme motion, a significant domain gap occurs due to the large amount of noise which further leads to performance degradation [24, 49, 40, 12, 25]. Wu et al. [63] first demonstrated that event-based vision can deteriorate due to increased noise levels, although the assessments were mainly conducted in synthetic events. Kim et al. [24] further introduced a large-scale dataset enabling systematic assessment of object recognition tasks, and demonstrated that large camera motion or illumination change leads to greater amounts of noise that ultimately deteriorates performance. Existing approaches denoise the raw data to cope with such adversaries [63, 58, 59], or suggest stacking events to overcome domain gaps under extreme lighting condition [50]. On the other hand, Ev-NeRF can compensate for the spurious noises by enforcing multi-view consistency for the scene geometry without extra noise processing.

Instead of handling complex data characteristics from the raw data, many approaches aggregate the sequential measurements into an ordinary image or 3D geometry. Early attempts for intensity image reconstruction are inspired from statistical methods [9, 22, 3]. Several subsequent approaches suggest various network architecture designs to improve the image quality or computational

cost [46, 57, 6, 56, 62]. Because the sensor has a high dynamic range, the intensity image restoration can be explicitly designed for HDR images [66, 57] or by applying domain adaptation to day-light condition [51]. For estimating 3D geometry, recent event-based SLAM methods utilize classical techniques [23, 22, 16, 21, 65, 17], minimizing the energy function formulated over the image-like event representations. On the other hand, for event-based depth estimation both classical method [19, 60] and learning-based approaches [20, 55] coexist.

However, for any of the aforementioned tasks, it is challenging to obtain a large-scale dataset with the ground-truth label or formulate the correct measurement model that represents the wide range of possible sensor characteristics. Rebecq et al. [44, 45] suggested generating the training data using simulator [43]. Stoffregen et al. [53] examined the statistical aspect to reduce the gap. Paredes et al. [37] proposed a self-supervised learning framework with the aid of optical flow and does not require ground truth, but their reconstructed images are characterized by several artifacts. On the contrary, Ev-NeRF works without the ground truth or synthetic data and shows stable results comparable to the state-of-the-art in intensity image reconstruction or depth estimation.

**Neural Implicit 3D Representation** Neural Implicit 3D Representation is gaining popularity due to its strong advantage of memory requirements, no restrictions on spatial resolution, and representation capability. Several works [38, 31, 8] showed the advantage of neural implicit representations with 3D supervision. NeRF (Neural Radiance Fields) [32] proposes an implicit representation of 3D coordinates and viewing direction which can synthesize images with volume rendering techniques. The resulting neural volume contains information about 3D volume density and emitted radiance for rendering images. Motivated by the photo-realistic quality of the produced images, a large number of subsequent works spurred to overcome the limitation of the original NeRF including: enabling fast convergence and rendering [35, 54, 1]; handling the input images with unknown or noisy camera poses [61, 28]; or processing dynamic scenes [41, 27]. Our method learns the NeRF volume with event data. By enforcing the multi-view consistency of collected measurements, Ev-NeRF produces a high-quality image or depth in a novel view and effectively removes spurious noises of an event camera.

### 3. Background

For the completeness of discussion, we include the event generation model of the sensor followed by the mathematical formulations of the neural radiance fields (NeRF), which serves as the two main components for deriving Ev-NeRF.

**Event Generation Model** Instead of recording the absolute color values of the image pixels, an event camera records asynchronous changes of the brightness as a sequence of events  $E_k = (u_k, v_k, t_k, p_k)$ , indicating that the brightness change at the pixel coordinate  $(u_k, v_k)$  reaches a specific threshold  $B$  at time  $t_k$ ,

$$|L(u_k, v_k, t_k) - L(u_k, v_k, t_k - \delta t)| \geq |B|, \quad (1)$$

where  $L = \log(I)$  is the logarithm of brightness  $I$  and  $\delta t$  is time that has passed since the last event.  $p_k \in \{+, -\}$  is the polarity denoting whether the brightness change is positive or negative. It is known that the threshold for triggering positive events is different from the one for the negative events [15], which we denote  $B^+$  and  $B^-$ , respectively. If we accumulate the events occurring for a given period of time  $\Delta t$ , the brightness change in a specific pixel can be approximated by [15]

$$\Delta L(u, v, \Delta t) = \sum_{\substack{t_k \in \Delta t, \\ (u_k, v_k) = (u, v)}} p_k |B^{p_k}|. \quad (2)$$

The threshold  $B^{p_i}$  can be different under various physical conditions, which further challenges the event-based vision, in addition to complex noise characteristics of the sensor.

**Neural Radiance Fields** Ev-NeRF takes inspiration from NeRF [32] which is trained to accumulate the volumetric information with 2D supervision. The supervision signal for NeRF is the total squared error between the rendered and true pixel colors. Basically the neural network  $F_\theta(\cdot)$  receives the input of the 3D coordinate  $\mathbf{x}_i \in \mathbb{R}^3$  and the ray direction  $\mathbf{d}_i \in \mathbb{S}^2$  and outputs the density  $\sigma_i \in \mathbb{R}$  and emitted radiance  $\mathbf{c}_i \in \mathbb{R}^3$

$$F_\theta : (\gamma_x(\mathbf{x}_i), \gamma_d(\mathbf{d}_i)) \rightarrow (\sigma_i, \mathbf{c}_i). \quad (3)$$

Here  $\gamma(\cdot)$  is sinusoidal positional encoding function which successfully captures the high-frequency information along the spatial direction. With positional encoding and coarse-to-fine sampling techniques, a neural network is trained to synthesize high-quality novel view images.

Following the classical volume rendering technique, each pixel is rendered by sampling  $N$  points  $\mathbf{x}_1, \dots, \mathbf{x}_N$  of the volume density along the ray  $r(\mathbf{x}_0, \mathbf{d})$ .  $\mathbf{x}_0$  is the initial point of the ray located at the focal point of the camera using the pinhole camera model. The final rendered color of the pixel is aggregated along the ray as

$$\hat{C}(r) = \sum_{i=1}^N A_i \alpha_i \mathbf{c}_i. \quad (4)$$

$A_i = \exp\left(-\sum_{l=1}^{i-1} \sigma_l \delta_l\right)$  denotes the accumulated transmittance along the ray, and  $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$  denotes

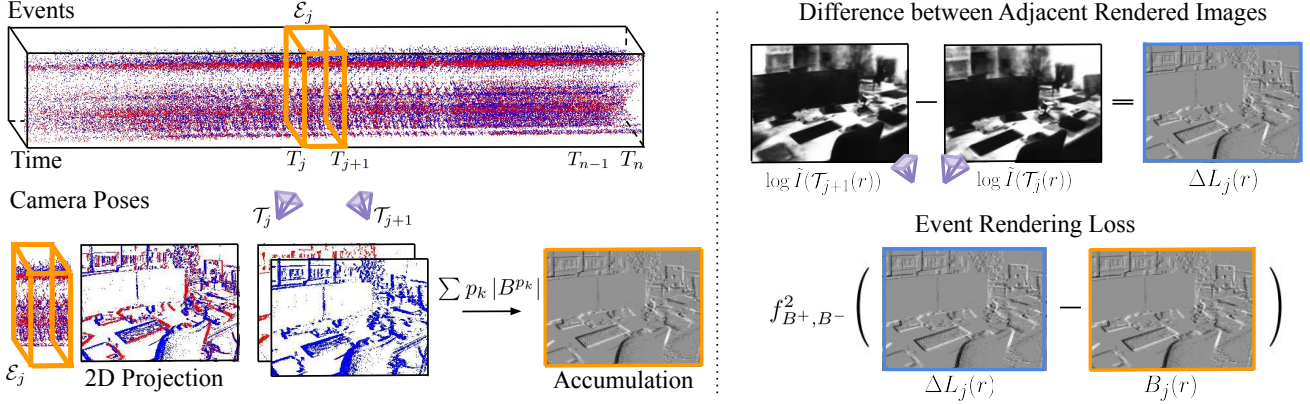


Figure 2. Overview of our method. According to the measurement model of the sensor, the events  $\mathcal{E}_j$  accumulated during a short time interval  $[T_j, T_{j+1})$  should reflect the difference in brightness. Using the implicit volume, we render intensity frames from the view points of two adjacent event camera poses,  $\mathcal{T}_j$  and  $\mathcal{T}_{j+1}$ . Event rendering loss is the discrepancy between the accumulated event  $B_j(r)$  and difference in the intensity of adjacent rendered frames  $\Delta L_j(r)$ .

the alpha value, where  $\delta_i = \|\mathbf{x}_{i+1} - \mathbf{x}_i\|$  is the distance between adjacent samples. Additionally, the depth along the ray direction can be approximated with a similar formulation:

$$\hat{D}(r) = \sum_{i=1}^N A_i \alpha_i s_i, \quad (5)$$

where  $s_i$  denotes the distance between  $\mathbf{x}_0$  and  $\mathbf{x}_i$ .

## 4. Method

Ev-NeRF creates neural implicit representation  $F_\theta$  of a static scene as NeRF. Since an event is triggered when the brightness changes, we use a slice of the event sequence  $\mathcal{E}_j = \{E_k = (u_k, v_k, p_k, t_k) | T_j \leq t_k < T_{j+1}\}$  during a small duration of time  $[T_j, T_{j+1})$ . The motion of the camera is provided by the starting and ending poses of the period,  $\mathcal{T}_j$  and  $\mathcal{T}_{j+1}$ . The neural network for Ev-NeRF regresses for one dimensional emitted luminance value  $y_i \in \mathbb{R}$  instead of RGB color values,

$$F_\theta : (\gamma_x(\mathbf{x}_i), \gamma_d(\mathbf{d}_i)) \rightarrow (\sigma_i, y_i). \quad (6)$$

This is a natural choice considering that the event only records the brightness change in a single channel.

After the neural network is trained, we can render the intensity image from an arbitrary viewpoint adapting Equation 4. To elaborate, if we define the camera ray  $r$  that passes through the pixel location  $(u, v)$  from a camera pose  $\mathcal{T}$ , we can sample  $N$  points along the ray and apply volume rendering technique to find the intensity of the pixel

$$\hat{I}(\mathcal{T}(r)) = \sum_{i=1}^N A_i \alpha_i y_i. \quad (7)$$

The depth measurement can also be approximated using Equation 5.

When we combine the formulation with the event generation model, we also jointly optimize for the unknown thresholds,  $B_j^+$  and  $B_j^-$ , in addition to the implicit neural volume  $F_\theta(\cdot)$ . We assume that the threshold is a function of time and polarity, but is spatially the same for all pixels. More specifically, we assume that the threshold is constant in each time interval  $[T_j, T_{j+1})$ , but changes when the time interval changes.

The total loss used to train Ev-NeRF is given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{event}} + \lambda \mathcal{L}_{\text{thres}}. \quad (8)$$

$\mathcal{L}_{\text{event}}$  is the event rendering loss which replaces the image rendering loss in conventional NeRF. Here we combine the event generation model in Section 3 with the volume rendering formulation of the original NeRF.  $\mathcal{L}_{\text{thres}}$  is the threshold bound loss, designed to avoid degenerate cases.

**Event Rendering Loss** Our loss for training  $F_\theta$  compares the recorded events and the difference in the rendered intensity, as shown in Figure 2. Let us denote the intensity values of the pixel ray  $r$  at time  $T_j$  and  $T_{j+1}$  as  $\hat{I}(\mathcal{T}_j(r))$  and  $\hat{I}(\mathcal{T}_{j+1}(r))$ , respectively, where the intensity images are obtained using Equation 7. Then we can calculate  $\Delta L_j$  at the pixel ray  $r$  as

$$\Delta L_j(r) = \log \hat{I}(\mathcal{T}_{j+1}(r)) - \log \hat{I}(\mathcal{T}_j(r)). \quad (9)$$

Using the event generation model in Equation 2,  $\Delta L_j(r)$  in Equation 9 should be measured by the accumulated sum of the events  $B_j(r) = \sum_{E_k \in \mathcal{E}_j(r)} p_k B_j^{p_k}$  within the time interval  $[T_j, T_{j+1})$ .

Our event-rendering loss  $\mathcal{L}_{\text{event}}$  is the total sum of discrepancy for all time intervals  $j$  and rays  $r$  available in the batch,

$$\mathcal{L}_{\text{event}} = \sum_j \sum_r f_{B_j^+, B_j^-}^2(\Delta L_j(r) - B_j(r)), \quad (10)$$



where the function  $f$  penalizes the discrepancy above the sensor threshold by incorporating a dead zone  $[B^-, B^+]$  as described in [3]:

$$f_{B^+, B^-}(x) = \begin{cases} x - B^+, & \text{if } x > B^+, \\ 0, & \text{if } B^- \leq x \leq B^+, \\ -x + B^-, & \text{if } x < B^-. \end{cases} \quad (11)$$

Therefore we only focus on where the measurements provide enough evidence for the brightness changes.

**Threshold Bound Loss** While the joint optimization over the unknown threshold values  $B_j^+, B_j^-$  improves the performance of Ev-NeRF, the additional parameters further challenge the optimization process which already is highly under-constrained with the unknown brightness values  $I$ . Without additional constraints, we empirically observed that the network often converges to the trivial solution with  $\Delta I = 0$  and the threshold value 0. The threshold bound loss is a simple prior to keep the threshold values within the reasonable bound:  $\mathcal{L}_{\text{thres}}$

$$\mathcal{L}_{\text{thres}} = u(B_0^+ - B_j^+) + u(B_j^- - B_0^-), \quad (12)$$

where  $u(\cdot)$  is a unit step function. In our experiments, we set  $B_0^+ = 0.3$  and  $B_0^- = -0.3$ , based on our prior knowledge about threshold scale [15].

## 5. Results

Once we train Ev-NeRF, we can generate high-quality images given corrupted input data, whose quality can be compared in various approaches developed for different purposes. We first compare the output of Ev-NeRF against previous works dedicated to noise reduction or HDR imaging under challenging environments. The Ev-NeRF volume also can generate images or depth estimates in a novel view, which are compared against previous works on image reconstruction or depth estimates.

**Implementation Detail** The input to Ev-NeRF is the stream of event data obtained from an event sensor moving around a static scene. The stream data is accompanied by a sequence of the sensor’s intermediate positions which are time-stamped. The sensor positions can be acquired from an additional sensor or structure from motion (SfM) using intensity images. In our implementation, we calculate the poses by running SfM [47, 48] with the intensity frames, which are provided in the datasets recorded at about 24 Hz. Except for this process, the intensity frames are not available to Ev-NeRF during training and are used only for evaluation. We use about 50 to 100 consecutive event slices  $\mathcal{E}_j$  to train a neural volume, where the length of each time slice  $[T_j, T_{j+1})$  is chosen to be the frame rate of intensity frames,

about 1/24s. The duration of the total event sequence used for Ev-NeRF is roughly 2-4 seconds. For each event slice  $\mathcal{E}_j$  corresponding to  $[T_j, T_{j+1})$ , we add random events equivalent to 5% of the number of events that occurred at time slice  $[T_j, T_{j+1})$  during training. We find that the additional random noise slightly improves the quality of neural representation in ambiguous regions, which is further described in the supplementary material.

We implement Ev-NeRF using PyTorch [39] using the NeRF formulation [32]. For positional encoding, we use 10 frequencies for  $\mathbf{x}$  and 4 for  $\mathbf{d}$ . The weight  $\lambda$  of  $\mathcal{L}_{\text{thres}}$  in Equation 8 is set large to 1000 to avoid thresholds from continuing to decrease. For all experiments, we use the Adam optimizer [26] with a learning rate of  $5 \times 10^{-4}$ . Ev-NeRF takes about an hour in RTX 2080 GPU to train per scene and 1.5 seconds to render a single image. Note that the speed can be accelerated by incorporating recent methods that enable fast learning and rendering of NeRF [35, 54, 1].

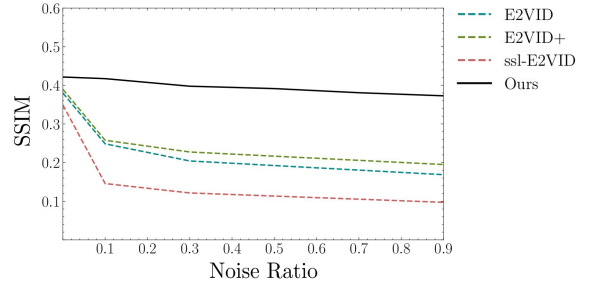


Figure 3. Effect of noise on various image reconstruction methods. Even in serious noise conditions, Ev-NeRF can robustly reconstruct images whose quality is comparable to that of other approaches in normal conditions.

**Noise Reduction** We compare the noise reduction of Ev-NeRF against three baselines: E2VID [44], E2VID+ [53] and ssl-E2VID [37]. E2VID [44] is trained in a supervised fashion with a large amount of synthetic data. E2VID+ [53] adjusts the synthetic training data to better fit the distribution of the real data. ssl-E2VID [37] tries to overcome the domain gap and suggests a self-supervised approach achieving results comparable to E2VID+ [53] without ground truth data. Unlike Ev-NeRF, which is trained for each scene, previous works are trained in a supervised fashion with a synthetic dataset composed of a pair of ground truth images and event measurements.

Figure 3 evaluates the SSIM of the reconstructed intensity image compared against the ground truth intensity image with the office\_zigzag scene in the IJRR dataset [34]. We use the event camera simulator v2e [18] to synthetically add realistic sensor noise due to photon fluctuations or invalid threshold values. The simulator allows us to control the amount of noise, which we indicate with the ratio of

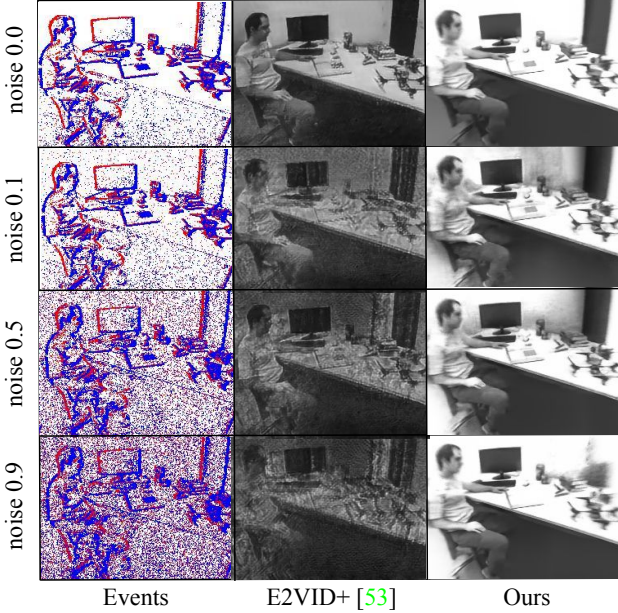


Figure 4. Qualitative comparison on image reconstruction given input with various noise levels. Ev-NeRF shows little performance degradation even with extremely noisy inputs.

the number of noisy events added to the number of existing events. The effectiveness of Ev-NeRF is prominent for noisy events, where it compensates the complex noise despite over 70% of noise and achieves comparable quality as the state-of-the-art method. In contrast, other methods rapidly deteriorate under noisy data, as they suffer from the domain shift caused by severe noise, which is not observed during the training neural network under the supervised setup. Additional results are provided in the supplementary material.

Figure 4 shows the visual comparison of reconstructed images of NeRF for inputs with different noise levels against E2VID+ [53], which is the state-of-the-art method for intensity image reconstruction. While E2VID+ [53] is vulnerable even to small noise, the intrinsic multi-view consistency of Ev-NeRF results in stable performance regardless of noise level and alleviates the effect of noise.

**HDR Image Reconstruction** In addition to noise reduction, the reconstructed images from Ev-NeRF naturally contain high dynamic range information without further processing. Figure 5(a) shows the qualitative results on HDR imaging. Compared to the intensity images concurrently captured in low-light set-up ((b), left), which are unknown to the algorithm, the sensor measurements detect the subtle details within the scene ((a), left), which are compiled to produce the HDR intensity images ((a), middle). This is because Ev-NeRF generates the intensity values to reflect the fine-grained changes in illumination without saturation and

is agnostic to any prior on the absolute values of the intensity that might be clipped to a smaller range. While we use intensity images to find the poses, the neural volume trained with event data contains variations and details that could not be captured with low-quality intensity images, especially in challenging lighting. In Figure 5(b), we also present the rendering using the neural volume of ordinary NeRF for comparison, which is trained with the intensity images of the same sequence obtained from a DAVIS camera [4]. The intensity-based NeRF is trained with the MSE loss between the rendered and measured intensity, following [32]. The reconstructed images and depths are superior when using Ev-NeRF, therefore fully exploiting the dynamic range of the sensor.

**Intensity Image Reconstruction** Since Ev-NeRF creates the NeRF volume that can generate novel-view images, we compare the quality of the synthesized image against previous works on intensity image reconstruction. We use three available real-world datasets that are widely used for event-based image reconstruction, namely IJRR [34], HQF [53], and Stereo DAVIS dataset [65] for quantitative comparison. For the stereo DAVIS dataset, we only use the measurements from a single event camera. Unless otherwise noted, we use four sequences (dynamic\_6dof, office\_spiral, office\_zigzag, hdr\_boxes) from the IJRR dataset, three sequences (reflective\_materials, high\_texture\_plants, still\_life) from HQF dataset and two sequences (monitor, reader) from Stereo DAVIS dataset.

Table 1 displays the comparison against baseline methods using three metrics: mean squared error (MSE), structural similarity (SSIM) and perceptual similarity (LPIPS) [64]. Ev-NeRF outperforms E2VID and ssl-E2VID and is on par with E2VID+ without observing the ground truth intensity frames. Figure 6 shows exemplar reconstructed images, and different approaches exhibit different kinds of artifacts. The results for all of the tested scenes are available in the supplementary material. Figure 6(f) also contains the intensity images used as ground truth for evaluation. Not only the ground truth images are in low resolution ( $240 \times 180$ ), but also the quality is suboptimal with limited dynamic range. While this might indicate that the performance metric should be only used to comprehend the approximate performance, we can still conclude that Ev-NeRF has the capacity to correctly capture the scene structure. Also note that previous approaches are trained in a fixed resolution dedicated to a specific measurement condition, whereas Ev-NeRF is not susceptible to such domain gaps. We further validate intensity image reconstruction results on the CED Dataset [7] with a different resolution ( $346 \times 240$ ) in the supplementary material.

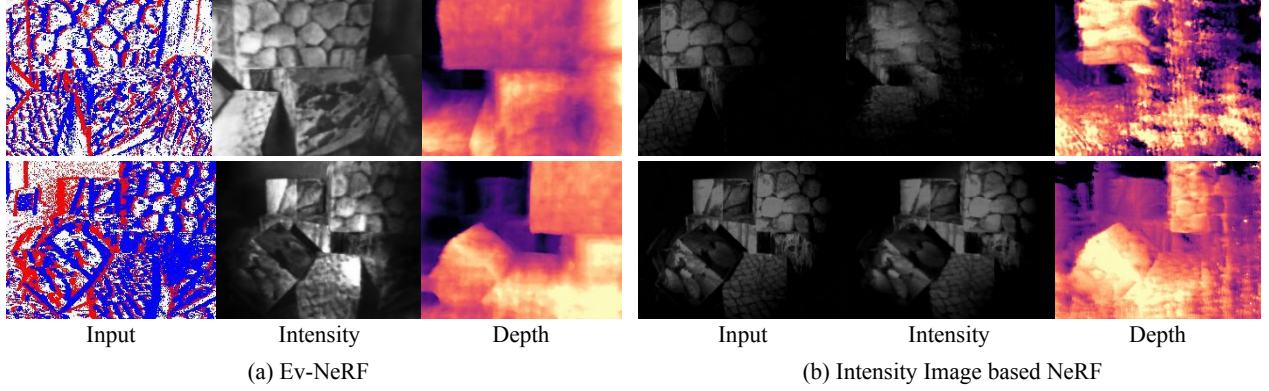


Figure 5. Qualitative comparison on intensity image reconstruction and depth estimation of the NeRF volume trained with (a) event data and (b) intensity images captured in low-light conditions.

Scene	MSE ↓				SSIM ↑				LPIPS ↓			
	E2VID	E2VID+	ssl-E2VID	Ours	E2VID	E2VID+	ssl-E2VID	Ours	E2VID	E2VID+	ssl-E2VID	Ours
office_zigzag	0.07	<u>0.05</u>	0.08	<b>0.03</b>	0.38	<u>0.39</u>	0.34	<b>0.42</b>	0.34	<b>0.25</b>	0.40	<u>0.27</u>
office_spiral	0.06	<u>0.05</u>	0.07	<b>0.03</b>	0.38	<u>0.39</u>	0.37	<b>0.41</b>	0.35	<u>0.27</u>	0.39	<b>0.27</b>
boxes	0.06	<b>0.03</b>	0.08	<u>0.04</u>	0.47	<b>0.60</b>	0.45	<u>0.48</u>	0.31	<b>0.20</b>	0.37	<u>0.31</u>
dynamic_6dof	<u>0.12</u>	<b>0.06</b>	0.15	0.19	<u>0.29</u>	<b>0.34</b>	0.28	0.26	<u>0.40</u>	<b>0.33</b>	0.54	0.41
reflective_materials	0.07	<u>0.05</u>	0.08	<b>0.05</b>	0.39	<b>0.45</b>	0.30	<u>0.40</u>	<u>0.31</u>	<b>0.24</b>	0.38	0.35
high_texture_plants	0.04	<b>0.02</b>	0.05	<u>0.03</u>	0.42	<b>0.55</b>	0.42	<u>0.44</u>	<u>0.21</u>	<b>0.12</b>	0.23	0.34
still_life	0.05	<b>0.02</b>	0.08	<u>0.03</u>	0.50	<b>0.61</b>	0.40	<u>0.53</u>	0.23	<b>0.13</b>	0.29	<u>0.18</u>
monitor	0.04	<u>0.04</u>	0.09	<b>0.03</b>	0.31	<b>0.36</b>	<u>0.33</u>	0.32	0.39	<b>0.20</b>	<u>0.34</u>	0.37
reader	<u>0.07</u>	<b>0.04</b>	0.09	0.09	0.42	<u>0.43</u>	0.38	<b>0.45</b>	<u>0.31</u>	<b>0.25</b>	0.40	0.35

Table 1. Quantitative comparison of image reconstruction on scenes from the IJRR [34], HQF [53] and Stereo DAVIS [65] dataset. The results with the best performance are in bold. We additionally underline the runner-up metric.

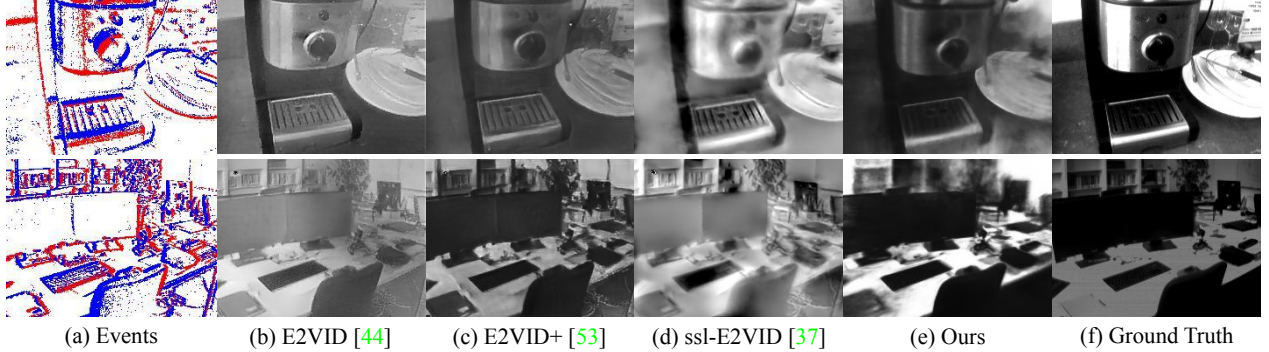


Figure 6. Qualitative comparison on intensity image reconstruction.

Scene	MSE ↓		SSIM ↑		LPIPS ↓	
	Given	Novel	Given	Novel	Given	Novel
office_zigzag	0.03	0.04	0.42	0.41	0.27	0.28
office_spiral	0.03	0.03	0.41	0.40	0.27	0.27
boxes	0.04	0.04	0.48	0.46	0.31	0.33
dynamic_6dof	0.19	0.20	0.26	0.26	0.41	0.43
reflective_materials	0.05	0.06	0.40	0.40	0.35	0.38
high_texture_plants	0.03	0.03	0.44	0.42	0.34	0.36
still_life	0.03	0.04	0.53	0.52	0.18	0.18

Table 2. Quantitative results of novel view reconstruction. The given views are provided during training and only a small performance gap is observed for reconstructing images of unknown novel views.

**Novel View Synthesis** Ev-NeRF inherits the capability of NeRF and can render an image from an arbitrary viewpoint. To the best of our knowledge, Ev-NeRF is the first work to reconstruct a novel-view image from an event sensor. We use the same dataset for the intensity image reconstruction to evaluate the performance of novel view synthesis. We divide each sub-sequence in the dataset into a training set and test set, and synthesize the images at the viewpoints included in the test set with Ev-NeRF trained with the training set. Table 2 compares the reconstructed images against ground truth intensity images using the same three metrics: MSE, SSIM, and LPIPS. The reported metrics for novel-



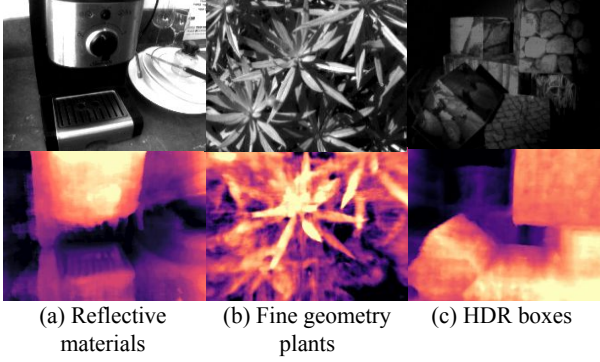


Figure 7. Qualitative results on depth estimation on real dataset. The intensity images (top) are not observed by the algorithm.

view images are compatible with the results for the given view. Thus, the ability of NeRF is nicely transferred to Ev-NeRF.

**3D Structure Estimation** The density values of NeRF volume can provide estimates for 3D structure as presented in Equation 5. Figure 5 contains the depth estimation from low lighting conditions. Additionally, Figure 7 shows that Ev-NeRF creates reasonable depth estimates for challenging real-world scenes with reflective materials, fine geometry details, or HDR measurements where existing approaches might fail.

Note that there is no prior work that is bound to the exact set-up as Ev-NeRF. As an alternative, we compare the acquired geometry against the 3D information extracted from an event-based SLAM approach by Zhou et al. [65]. Event-based SLAM computes the camera pose in addition to the geometric structure, often with the help of additional hardware to augment the input data. Specifically, Zhou et al. [65] use a pair of temporally synchronized event cameras to recover pose and reconstruct the semi-dense 3D structure. On the other hand, Ev-NeRF uses a single event camera with known poses. Even though the detailed setup is different, both estimate 3D structure from the measurements of a moving sensor observing a static scene, whose results are provided in Figure 8. Zhou et al. [65] reconstruct the semi-dense 3D structure and only show depth in the area where the events have occurred (second column in Figure 8). In contrast, Ev-NeRF reconstructs implicit 3d volume and can produce denser depth with intensity images (third and last column in Figure 8).

**Limitation and Discussion** Our approach is the first to suggest incorporating NeRF with event data, and uniquely uses the neural representation of the scene. There are no previous approaches that handle the exact same input and output format, and admittedly comparisons are made under different set-ups. Most of the previous works are trained

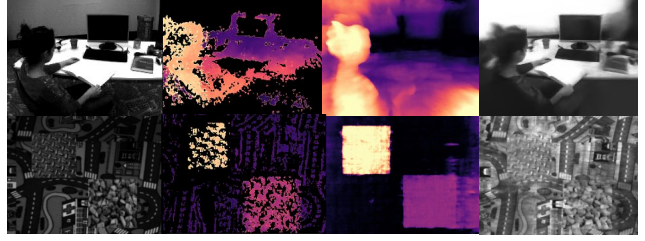


Figure 8. Qualitative comparison on 3D structure estimation. Ground-truth intensity frame, semi-dense depth by [65], dense depth, and intensity image by Ev-NeRF in order.

with the input-output pairs for specific tasks, for example, intensity images, depth reconstruction, or noise handling. After being trained in a supervised setting, the base-lines quickly process new input and can handle the dynamic scenes without observation from multiple views. Nonetheless, the pipelines can be susceptible to performance degradation with complex noise characteristics, different sensor configurations, or large domain shifts in extreme conditions.

On the other hand, Ev-NeRF does not require any labels or domain-specific priors and trains a neural representation from the input sequence in a completely self-supervised way. It is trained to overfit for a static scene. After training, the acquired volume compensates for the diverse unknown adversaries and discovers a consistent yet comprehensive structure of an arbitrary scene. It is a desirable characteristic considering that it is impossible to collect the data that exhaustively represents the various sources of noise in the sensor which are known to deteriorate the performance of event-based vision [24, 25]. Therefore Ev-NeRF can prosper in various measurement settings where ground-truth data is not available.

## 6. Conclusions and Future Work

We present Ev-NeRF, which learns the implicit volume of the neural radiance field from the raw stream of events generated by a neuromorphic camera. The inherent multi-view consistency creates a representation remarkably robust to noisy inputs, which is a critical challenge for using a neuromorphic sensor, yet exploits the subtle brightness changes detected from the sensor. Further, the created NeRF volume can generate intensity images or estimate depth, whose quality is comparable to many existing supervised methods exclusively designed to solve a specific task. To the best of our knowledge, Ev-NeRF is the first attempt to incorporate the NeRF formulation with raw event data and can advance with the abundant subsequent works that overcome the limitations of NeRF, such as handling dynamic scenes to better incorporate the fast temporal resolution of the sensor [41, 27], reducing training and rendering time [35, 54, 1], and alleviating the requirements of known camera poses [61, 28].



## References

- [1] Alex Yu and Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks, 2021. [3](#), [5](#), [8](#)
- [2] R Baldwin, Mohammed Almatrafi, Vijayan Asari, and Keigo Hirakawa. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [1](#)
- [3] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 884–892, 2016. [2](#), [5](#)
- [4] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A  $240 \times 180$  130 db 3  $\mu$ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. [6](#)
- [5] Samuel Bryner, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Event-based, direct camera tracking from a photometric 3D map using nonlinear optimization. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019. [1](#)
- [6] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing*, 30:2488–2500, 2021. [2](#), [3](#)
- [7] Timo Stoffregen Cedric Scheerlinck, Henri Rebecq and Davide Scaramuzza Nick Barnes, Robert Mahon. Ced: Color event camera dataset. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. [6](#)
- [8] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [3](#)
- [9] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776, 2011. [2](#)
- [10] Daniel Czech and Garrick Orchard. Evaluating noise filtering for event-based asynchronous change detection image sensors. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 19–24, 2016. [1](#)
- [11] Tobi Delbruck. Frame-free dynamic digital vision. In *Proceedings of Intl. Symposium on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 6–7, 2008. [1](#)
- [12] Yongjian Deng, Youfu Li, and Hao Chen. Amae: Adaptive motion-agnostic encoder for event-based object classification. *IEEE Robotics and Automation Letters*, 5(3):4596–4603, 2020. [1](#), [2](#)
- [13] Peiqi Duan, Zihao W. Wang, Xinyu Zhou, Yi Ma, and Boxin Shi. Eventzoom: Learning to denoise and super resolve neuromorphic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12824–12833, June 2021. [1](#)
- [14] Yang Feng, Hengyi Lv, Hailong Liu, Yisa Zhang, Yuyao Xiao, and Chengshan Han. Event density based denoising method for dynamic vision sensor. *Applied Sciences*, 10:2024, 2020. [1](#)
- [15] Guillermo Gallego, Tobi Delbruck, Garrick Michael Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jorg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. [1](#), [3](#), [5](#)
- [16] Cheng Gu, Erik Learned-Miller, Daniel Sheldon, Guillermo Gallego, and Pia Bideau. The spatio-temporal poisson point process: A simple model for the alignment of event camera data. In *International Conference on Computer Vision (ICCV)*, 2021. [3](#)
- [17] Javier Hidalgo-Carri , Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. [3](#)
- [18] Y Hu, S C Liu, and T Delbruck. v2e: From video frames to realistic DVS event camera streams. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2021. [5](#)
- [19] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. High-speed structured light based 3d scanning using an event camera. *Opt. Express*, 29(22):35864–35876, Oct 2021. [3](#)
- [20] Daniel Gehrig Javier Hidalgo-Carrio and Davide Scaramuzza. Learning monocular dense depth from events. *IEEE International Conference on 3D Vision.(3DV)*, 2020. [3](#)
- [21] Jianhao Jiao, Huaiyang Huang, Liang Li, Zhijian He, Yilong Zhu, and Ming Liu. Comparing representations in tracking for event camera-based SLAM. *CoRR*, abs/2104.09887, 2021. [3](#)
- [22] Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew Davison. Simultaneous mosaicing and tracking with an event camera. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. [2](#), [3](#)
- [23] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *ECCV*, 2016. [2](#), [3](#)
- [24] Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, and Young Min Kim. N-imagenet: Towards robust, fine-grained object recognition with event cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2146–2156, October 2021. [1](#), [2](#), [8](#)
- [25] Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. [2](#), [8](#)
- [26] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [5](#)
- [27] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. [3](#), [8](#)
- [28] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In

- IEEE International Conference on Computer Vision (ICCV)*, 2021. [3](#), [8](#)
- [29] Daqi Liu, Álvaro Parra, and Tat-Jun Chin. Globally optimal contrast maximisation for event-based motion estimation. In *European Conference on Computer Vision (ECCV)*, 2020. [1](#)
- [30] Hongjie Liu, Christian Brandli, Chenghan Li, Shih-Chii Liu, and Tobi Delbruck. Design of a spatiotemporal correlation filter for event-based sensors. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 722–725, 2015. [1](#)
- [31] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. [3](#)
- [32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. [1](#), [3](#), [5](#), [6](#)
- [33] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. [1](#)
- [34] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, Feb 2017. [1](#), [5](#), [6](#), [7](#)
- [35] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv:2201.05989*, Jan. 2022. [3](#), [5](#), [8](#)
- [36] Urbano Miguel Nunes and Yiannis Demiris. Entropy minimisation framework for event-based vision model estimation. In *European Conference on Computer Vision (ECCV)*, 2020. [1](#)
- [37] Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. *CVPR*, 2021. [2](#), [3](#), [5](#)
- [38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. [3](#)
- [39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. [5](#)
- [40] Etienne Perot, Pierre de Tournemire, Davide Nitti, Jonathan Masci, and Amos Sironi. Learning to detect objects with a 1 megapixel event camera. *arXiv preprint arXiv:2009.13436*, 2020. [2](#)
- [41] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. [3](#), [8](#)
- [42] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time. *Int. J. Comput. Vis.*, 126:1394–1414, Dec. 2018. [2](#)
- [43] Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. *Conf. on Robotics Learning (CoRL)*, Oct. 2018. [2](#), [3](#)
- [44] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019. [2](#), [3](#), [5](#)
- [45] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)*, 2019. [2](#), [3](#)
- [46] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 156–163, 2020. [2](#), [3](#)
- [47] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [5](#)
- [48] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. [5](#)
- [49] Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification. *arXiv preprint arXiv:2018.00186*, June 2018. [2](#)
- [50] I. S.MohammadMostafavi, Lin Wang, and Kuk-Jin Yoon. Learning to reconstruct hdr images from events, with applications to depth and flow prediction. *International Journal of Computer Vision*, pages 1–21, 2021. [2](#)
- [51] Yu Zhang Song Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *ECCV*, 2020. [3](#)
- [52] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. [1](#)
- [53] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahoney. Reducing the sim-to-real gap for event cameras. In *ECCV*, 2020. [2](#), [3](#), [5](#), [6](#), [7](#)
- [54] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, 2021. [3](#), [5](#), [8](#)

- [55] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3
- [56] Bishan Wang, Jingwei He, Lei Yu, Gui-Song Xia, and Wen Yang. Event enhanced high-quality image recovery. In *European Conference on Computer Vision*. Springer, 2020. 2, 3
- [57] Lin Wang, S. Mohammad Mostafavi I. , Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 3
- [58] Y. Wang, B. Du, Y. Shen, K. Wu, G. Zhao, J. Sun, and H. Wen. Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6351–6360, 2019. 2
- [59] Zihao Wang, Peiqi Duan, Oliver Cossairt, Aggelos Kat-saggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2
- [60] Ziwei Wang, Liyuan Pan, Yonhon Ng, Zheyu Zhuang, and Robert Mahony. Stereo hybrid event-frame (shef) cameras for 3d perception. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021. 3
- [61] 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. 3, 8
- [62] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2563–2572, October 2021. 2, 3
- [63] J. Wu, C. Ma, X. Yu, and G. Shi. Denoising of event-based sensors with spatial-temporal correlation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4437–4441, 2020. 1, 2
- [64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6
- [65] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 6, 7, 8
- [66] Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2024–2033, June 2021. 3