Robust e-NeRF: NeRF from Sparse & Noisy Events under Non-Uniform Motion

Weng Fei Low
Gim Hee Lee
The NUS Graduate School’s Integrative Sciences and Engineering Programme (ISEP)
Institute of Data Science (IDS), National University of Singapore
Department of Computer Science, National University of Singapore
\{wengfei.low, gimhee.lee\}@comp.nus.edu.sg
https://wengflow.github.io/robust-e-nerf

Abstract

Event cameras offer many advantages over standard cameras due to their distinctive principle of operation: low power, low latency, high temporal resolution and high dynamic range. Nonetheless, the success of many downstream visual applications also hinges on an efficient and effective scene representation, where Neural Radiance Field (NeRF) is seen as the leading candidate. Such promise and potential of event cameras and NeRF inspired recent works to investigate on the reconstruction of NeRF from moving event cameras. However, these works are mainly limited in terms of the dependence on dense and low-noise event streams, as well as generalization to arbitrary contrast threshold values and camera speed profiles. In this work, we propose Robust e-NeRF, a novel method to directly and robustly reconstruct NeRFs from moving event cameras under various real-world conditions, especially from sparse and noisy events generated under non-uniform motion. It consists of two key components: a realistic event generation model that accounts for various intrinsic parameters (e.g. time-independent, asymmetric threshold and refractory period) and non-idealities (e.g. pixel-to-pixel threshold variation), as well as a complementary pair of normalized reconstruction losses that can effectively generalize to arbitrary speed profiles and intrinsic parameter values without such prior knowledge. Experiments on real and novel realistically simulated sequences verify our effectiveness. Our code, synthetic dataset and improved event simulator are public.

1. Introduction

Event cameras are bio-inspired sensors that represent a paradigm shift in visual acquisition and processing. This is attributed to its fundamentally distinctive principle of operation, where its pixels independently respond to log-intensity changes in an asynchronous manner, yielding a stream of events, rather than measuring absolute linear intensity synchronously at a constant rate, as done in standard cameras. Such unique properties contribute to their multitude of advantages over standard cameras [9]: low power, low lat-
tency, high temporal resolution and high dynamic range, thereby the recent success of event-based [15, 37, 10, 7, 38] or event-image hybrid [48, 47, 11] applications.

The success of many downstream visual applications in robotics, computer vision, graphics and virtual/augmented reality also hinges on an efficient and effective representation to encode various information of the scene being interacted with. Neural scene representations [24, 28, 49, 54, 20], especially neural fields [52], have recently emerged as promising candidates for future applications, owing to their continuous nature and memory efficiency. This trend is further driven by the exceptional capabilities and photorealism of Neural Radiance Field (NeRF) [24]-based works [49, 23, 45, 30, 31, 32].

Motivated by such promise and potential of event cameras and NeRF, we are interested in studying the following research question: How to robustly reconstruct a NeRF from a moving event camera under general real-world conditions? One simple way is to retrofit an events-to-video reconstruction method [38, 44, 29] to NeRF. However, such a naïve approach is inherently limited by the low photometric accuracy and consistency of the reconstructed video frames, since they are assumed to be true observations of the scene.

On the contrary, recent works, such as EventNeRF [41], Ev-NeRF [12] and E-NeRF [18], have proposed to reconstruct NeRFs directly using events via alternative reconstruction losses inspired or derived from an event generation model. Nonetheless, these works heavily rely on a temporally dense and low-noise event stream, which is generally inaccessible in practice due to the presence of refractory period (i.e., pixel dead-time after generating an event) and pixel-to-pixel variation in the contrast threshold (i.e., minimum log-intensity change for event generation). Such a limitation can be partly attributed to the accumulation of successive events at each pixel over time intervals, as performed in these works. Moreover, the effective reduction in contrast sensitivity and amplification of threshold variation resulting from the event accumulation also leads to a less detailed and robust reconstruction, respectively.

In addition, these methods do not directly and effectively generalize to arbitrary contrast threshold values and camera speed profiles, as their optimal hyper-parameter configuration greatly depends on the contrast threshold and speed of motion. EventNeRF and E-NeRF assumes symmetric positive and negative thresholds, which generally does not hold true in practice. While joint optimization of the contrast threshold is supported in Ev-NeRF, an additional regularization is necessary to prevent degeneracy. Furthermore, the assumption of time-varying thresholds made in Ev-NeRF and E-NeRF, which is not well supported by the literature, also leads to a reduction in reconstruction accuracy as shown in our experiments.

**Contributions.** We propose Robust e-NeRF, a novel method to directly and robustly reconstruct NeRFs from moving event cameras under various real-world conditions, especially from temporally sparse and noisy event streams given by event cameras in non-uniform motion.

In particular, we incorporate a more realistic event generation model that accounts for various intrinsic parameters (e.g., time-independent, asymmetric contrast threshold and refractory period) and non-idealities (e.g., pixel-to-pixel threshold variation). Furthermore, we introduce two complementary normalized reconstruction losses that are not only effectively invariant to the camera motion speed and threshold scale, but also minimally influenced by asymmetric thresholds. This allows for their effective generalization, as well as the regularization-free joint optimization of unknown contrast threshold and refractory period from poor initializations. The first loss serves as the primary loss for high-fidelity reconstruction, while the second acts as a smoothness constraint for better regularization of texture-less regions. As both loss functions do not involve event accumulation, detailed and robust reconstruction from sparse and noisy events can be achieved. Our experiments on novel sequences, simulated using an improved variant of ESIM [36], and real sequences from TUM-VIE [17] verify the effectiveness of Robust e-NeRF. We publicly release our code, synthetic event dataset and improved ESIM.

## 2. Related Work

**Event-based Scene Reconstruction.** Successful reconstruction of geometry, and possibly appearance, from events has been demonstrated in notable works, such as [35, 10, 56, 37, 15, 57]. Nonetheless, reconstruction of scene appearance is largely limited to diffuse surfaces due to the adoption of Lambertian surface assumption via brightness constancy [37, 15]. Moreover, existing methods can mainly recover semi-dense geometry, in the form of discrete depth maps or point clouds, corresponding to edges (more precisely, locations with high perceived spatial intensity gradient) in the scene. This is due to the fact that events are primarily generated along edges under relative motion [9].

While [51] achieved dense diffuse reconstruction by applying the classic Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipelines on video frames reconstructed from events [38], its performance is intrinsically limited by the accuracy of the recovered frames, as similarly discussed in Sec. 1. Although limited to simple, object-level mesh or specialized parametric models, [42, 25, 53] also demonstrated dense diffuse reconstruction via Analysis-by-Synthesis or equivalently Vision-as-Inverse-Graphics [13]. In contrast, we aim to reconstruct dense, continuous scene geometry and view-dependent appearance, in the form of a NeRF, directly from the raw event stream.
Reconstructing Neural Radiance Fields. In general, visual reconstruction of neural scene representations, including NeRFs, is achieved via Analysis-by-Synthesis. Nonetheless, NeRF derivatives mainly focus on the reconstruction from dense [21, 4, 5] or sparse [26, 16] multi-view images, possibly with depth maps [2, 3] or point clouds [40, 6, 39].

The reconstruction of NeRFs from events was first proposed and investigated in EvNeRF [12], E-NeRF [18] and EventNeRF [41]. E-NeRF also explored on using a combination of events and images for NeRF reconstruction. However, these works suffer from various limitations, as outlined in Sec. 1. Moreover, EventNeRF also relies on the access to the analytic camera trajectory, in contrast to E-NeRF and our work which only require constant-rate camera poses. Furthermore, inconsistent settings of loss functions and hyper-parameters were also adopted across different scenes in E-NeRF. In addition, E-NeRF with normalized and no-event losses also requires the contrast threshold to be known as a priori, which is hard to achieve in practice.

3. Our Method

We first briefly introduce the Neural Radiance Field (NeRF) scene representation (Sec. 3.1). Next, we detail the event generation model (Sec. 3.2) and normalized training losses (Sec. 3.3) proposed to robustly reconstruct NeRFs from event cameras. Lastly, we describe a Gamma Correction-based approach to align the radiance levels of the synthesized views to a set of given reference views (Sec. 3.4).

3.1. Preliminaries: Neural Radiance Fields

Neural Radiance Field (NeRF) [24] represents a scene using a Multi-Layer Perceptron (MLP) \( F_\theta : (x, d) \mapsto (c, \sigma) \) that maps 3D position \( x = (x, y, z) \) and 2D viewing direction \( d = (\theta, \phi) \) to its corresponding directional emitted radiance, or simply color, \( c = (r, g, b) \) and volume density \( \sigma \). From this representation, the estimated incident radiance \( L \) at a given pixel \( u \) can be computed using the volume rendering equation with quadrature [46], as follows:

\[
L(u) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) c_i ,
\]

where \( \sigma_i \) and \( c_i \) are the volume density and emitted radiance, respectively, of a sample \( x_i \) along the back-projected ray through the pixel, which has direction \( d \) from the camera center \( o \). The sample \( x_i = o + s_i d \) has a distance \( s_i \) from the camera center and a distance of \( \delta_i = s_{i+1} - s_i \) between its adjacent sample \( x_{i+1} \).

3.2. Event Generation Model

Fig. 2 shows the illustration of the event generation model. An event camera responds to log-radiance changes in the scene and outputs an Event Stream \( \mathcal{E} \), given by:

\[
\mathcal{E} = \{ e \mid e = (u, p, t_{\text{prev}}, t_{\text{curr}}) \},
\]

where \( e \) is an Event generated by pixel \( u \), with polarity \( p \in \{-1, +1\} \), at timestamp \( t_{\text{curr}} \). For convenience of discussion, we augment each event with the timestamp of the previous event that was generated by the same pixel, \( t_{\text{prev}} \).

An event of polarity \( p \) is generated when the difference in log-radiance at a pixel, measured with respect to a reference log-radiance at timestamp \( t_{\text{ref}} \), has the same sign as \( p \) and a magnitude that equals to the Contrast Threshold associated to polarity \( p, C_p \) [9]. In short, the condition is given by:

\[
\Delta \log L := \log L(u, t_{\text{curr}}) - \log L(u, t_{\text{ref}}) = p C_p ,
\]

where \( L \) denotes the incident radiance at the given pixel and timestamp. For color event cameras, \( L \) corresponds to the radiance of the incident light after passing through the specific color filter in front of the pixel.

After an event is generated, the pixel will be deactivated for an amount of time specified by the Refractory Period \( \tau \). During this period of time, the pixel is invariant to any change in log-radiance and thus will not generate any new events. At the end of the refractory period, the pixel will be reactivated and the current log-radiance value at the pixel will be set as the new reference value, enabling the next event to be generated at this pixel [9, 19]. In essence:

\[
t_{\text{ref}} = t_{\text{prev}} + \tau .
\]
The refractory period gives rise to temporal sparsity in the event stream. While this leads to partially observable log-radiance changes, it also limits the event generation rate of the camera [9, 19]. This is crucial to prevent the corruption of event timestamps due to Address Event Representation (AER) bus saturation and readout congestion, especially when the resolution and/or the speed of the camera are high [9, 1, 34, 8]. The bounded event rate also facilitates longer periods of recordings given a fixed memory budget. Furthermore, long refractory periods, as well as asymmetric contrast thresholds, contribute to a lower Signal-to-Noise Ratio (SNR).

Event sensors also suffer from non-uniformity of pixel response, similar to standard image sensors. This is characterized by the pixel-to-pixel variation in the contrast threshold, which can be viewed as the event sensor analogue of Fixed-Pattern Noise (FPN). Studies suggest that the threshold variation can be modeled as a Gaussian distribution with mean $C_p$ [9, 19, 33].

The contrast threshold is also effectively constant over time, as it is independent of temperature. The key assumption for temperature-independence is the absence of junction leakage [27], which holds true for modern event cameras [43]. Regardless of leakage, temperature-independence remains observed over a large temperature range [27].

3.3. Training

3.3.1 Assumptions

To reconstruct a NeRF from an event camera in motion, we assume that we are given the intrinsic camera matrix, lens distortion parameters and constant-rate camera poses of sufficiently high sampling rate for accurate interpolation at arbitrary time instants. Similar to [18], we perform Linear Interpolation (LERP) on camera positions and Spherical Linear Interpolation (SLERP) on camera orientations. While our method does not rely on event camera-specific intrinsic parameters such as contrast threshold and refractory period to be known as a priori, this information generally facilitates a more accurate reconstruction. As we will see in Sec. 3.3.2, it is also acceptable to provide only the (positive-to-negative) contrast threshold ratio $C_{+1}/C_{-1}$, which can be inferred from event camera bias settings.

3.3.2 Fundamental Limitations

As event cameras only provide (partial) observations of log-radiance changes, not absolute log-radiance, the predicted log-radiance $\log L$, given by volume renderings from the reconstructed NeRF (Eq. 1), is only accurate up to an offset per color channel. Furthermore, there will be an additional scale ambiguity that is consistent across all channels, when only the contrast threshold ratio is known. This is similar to an image with unknown black levels and ISO. Nonetheless, these ambiguities can be easily dealt with or corrected for post-reconstruction, e.g. using a set of reference images of the same scene (Sec. 3.4), thereby not a cause for concern. We are also restricted to the reconstruction of a NeRF with single-channel directional emitted radiance, when a monochrome event camera is employed.

3.3.3 Loss Functions

An overview of the training pipeline is illustrated in Fig. 3. In line with the event generation model (Sec. 3.2), we propose two complementary normalized loss functions: threshold-normalized difference loss $\ell_{\text{diff}}$ and target-
normalized gradient loss $\ell_{\text{grad}}$ that directly and effectively generalize to various real-world conditions. The weighted sum of the two losses form the total training loss, which is optimized on a batch of events $E_{\text{batch}}$, sampled randomly from the raw, asynchronous event stream. Formally, the total training loss is given by:

$$L = \frac{1}{|E_{\text{batch}}|} \sum_{e \in E_{\text{batch}}} \lambda_{\text{diff}} \ell_{\text{diff}}(e) + \lambda_{\text{grad}} \ell_{\text{grad}}(e), \quad (5)$$

where $\lambda_{\text{diff}}$ and $\lambda_{\text{grad}}$ are the respective loss weights.

We specifically refrained from optimizing on a reduced event stream, obtained by accumulating successive events at each pixel over time intervals, as done in \cite{41, 12, 18}. This enables us to account for the refractory period and prevent unnecessary effective reduction in contrast sensitivity, which leads to lower reconstruction fidelity. Moreover, it can also be shown that event accumulation results in the effective amplification of threshold variation (proof in supplementary materials), thereby reduction in noise robustness.

**Threshold-Normalized Difference Loss.** This loss enforces the mean contrast threshold $C = \frac{1}{2}(C_{-1} + C_{+1})$ normalized squared consistency between the observed log-radiance difference $\Delta \log L = pC_p$ from an event (Eq. 3) and the predicted log-radiance difference $\Delta \log \hat{L} := \log \hat{L}(u, t_{\text{curr}}) - \log \hat{L}(u, t_{\text{ref}})$, given by renders from the NeRF model (Eq. 1), as follows:

$$\ell_{\text{diff}}(e) = \left( \frac{\Delta \log \hat{L} - pC_p}{C} \right)^2. \quad (6)$$

Note that when a color event camera is employed, $\hat{L}$ refers to the single-channel rendered radiance, where the color channel is governed by the pixel color filter.

The loss serves as the primary reconstruction loss and can be effectively interpreted as a squared percentage error, especially under symmetric contrast thresholds. The normalization entails that the loss is invariant to the common scale of the positive and negative thresholds, as well as the predicted log-radiance, and only dependent on their ratio. Moreover, the normalization is optimal in the sense that the magnitude of the normalized target $|pC_p/C| = C_p/C$ is always centered at 1 regardless of the threshold ratio (proof in supplementary materials). The loss can therefore effectively generalize to arbitrary threshold values.

Unlike Ev-NeRF \cite{12}, these properties also enable the joint optimization of the unknown contrast threshold without additional regularization, as it does not suffer from any degeneracy. However, only the contrast threshold ratio can be recovered, thereby an additional scale ambiguity in the predicted log radiance, as mentioned in Sec. 3.3.2. In addition, our experiments also demonstrate the viability of jointly optimizing the refractory period $\tau$ via $t_{\text{ref}}$ (Eq. 4).

**Target-Normalized Gradient Loss.** This loss is simply the Absolute Percentage Error (APE) of the predicted log-radiance temporal gradient derived using auto-differentiation $\frac{\partial}{\partial t} \log \hat{L}(u, t)$, with respect to the finite difference approximation of the target log-radiance gradient $\frac{\partial}{\partial t} \log L(u, t) \approx \frac{pC_p}{t_{sam} - t_{ref}}$, at a timestamp $t_{sam}$ sampled between $t_{\text{ref}}$ and $t_{\text{curr}}$, as follows:

$$\ell_{\text{grad}}(e) = \text{APE} \left( \frac{\partial}{\partial t} \log \hat{L}(u, t_{sam}) \right) \frac{pC_p}{t_{sam} - t_{\text{ref}}}) \quad (7)$$

where $\text{APE}(\hat{y}, y) = \left| \frac{\hat{y} - y}{y} \right|$. As the finite difference approximation error is minimum at the midpoint and maximum at the endpoints, we sample $t_{sam}$ from a truncated normal distribution that is centered at the midpoint and has a standard deviation of $1/4$ the interval.

The loss acts as a smoothness constraint for log-radiance changes between $t_{\text{ref}}$ and $t_{\text{curr}}$, which lack explicit regularization from $\ell_{\text{diff}}$. This is particularly important for the effective reconstruction of textureless regions in the scene, where events associated have comparatively longer intervals. In contrast to analogous regularization losses adopted in recent works \cite{41, 12, 18}, $\ell_{\text{grad}}$ is specifically invariant to the speed of motion, hence generalizable to arbitrary speed profiles. An unnormalized gradient loss would over-emphasize events generated under high-speed motion, as they have relatively larger target gradients. Furthermore, the loss is invariant to the common scale of the threshold and predicted log-radiance, similar to $\ell_{\text{diff}}$. It is therefore also able to effectively generalize to arbitrary threshold values and facilitate joint optimization of unknown threshold.

### 3.4. Gamma Correction of Synthesized Views

As alluded in Sec. 3.3.2, the channel-consistent scale and per-channel offset ambiguity in the predicted log-radiance can be corrected for post-reconstruction, given a set of reference images of the same scene. To better account for the likely mismatch of spectral sensitivity, thereby color balance, between the event camera and the standard camera used to capture the reference images. Akin to \cite{41}, we further relax the channel-consistent scale to per-channel scales. This entails an affine correction of the predicted log-radiance, or equivalently a gamma correction of the predicted linear radiance, for each color channel as follows:

$$\log \hat{L}_{\text{corr}} = a \odot \log \hat{L} + b, \quad (8)$$

where $a$ and $b$ are the correction parameters, is necessary and sufficient for the reference alignment. Consequently, given the target log-radiance $\log L$ from the reference images, the optimal correction parameters can be simply derived via ordinary least squares.
4. Experiments

We adopt Novel View Synthesis (NVS) as the standard task to verify that our method, Robust e-NeRF, can indeed directly and robustly reconstruct NeRFs from moving event cameras under various real-world conditions, particularly from sparse and noisy events generated under non-uniform motion. The NVS benchmark experiments are conducted on both synthetic (Sec. 4.1) and real sequences (Sec. 4.2). In addition, we perform ablation studies (Sec. 4.3) to investigate the significance of various components in our method.

Metrics. We adopt a consistent set of metrics to assess the performance of a method in all experiments. In particular, we employ the three following standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [50] and AlexNet-based Learned Perceptual Image Patch Similarity (LPIPS) [55] to quantify the similarity between the gamma-corrected (Sec. 3.4) synthesized novel views and given target novel views.

Baselines. We benchmark our method, Robust e-NeRF against a recent work, Ev-NeRF [12] and a naive baseline, E2VID + NeRF, given by cascading the seminal events-to-video reconstruction method, E2VID [38] to NeRF. For Ev-NeRF, events are accumulated over non-overlapping time intervals of $\frac{1}{24}$ $s \approx 41.67$ ms, as suggested in their paper. To facilitate a fair comparison, all methods, including ours, have been (re-)implemented to adopt a common NeRF backbone. While we do not explicitly compare against E-NeRF [18] and EventNeRF [41], our experiment results should still provide an accurate indication on their performance relative to ours since they share a lot of similarities with Ev-NeRF as detailed in Sec. 1.

Datasets. The synthetic experiments are performed on a novel set of sequences simulated on the “Realistic Synthetic 360°” scenes, which were proposed in the seminal NeRF paper [24]. These scenes contain a wide variety of photo-realistic objects with complicated structure and non-Lambertian effects, thereby effective for NVS evaluation. The new synthetic event dataset allows for a retrospective comparison between event-based and image-based NeRF reconstruction methods, as the sequences were simulated under highly similar conditions, unlike in [41, 12].

The events are generated from a virtual event camera moving in a hemi-spherical spiral motion about the object at the origin, as shown in Fig. 1(b, d), using an improved variant of ESIM [36], which is an efficient and realistic event simulator. Specifically, we added support to time-independent pixel-to-pixel threshold variation and Blender, circumvented singularities in the trajectory orientation interpolation, as well as improved event simulation accuracy, especially with non-zero refractory periods. On the contrary, the real experiments are performed on the mocap-1d-trans and mocap-desk2 sequences of the TUM-VIE dataset [17], which are mainly forward-facing captures of a desk with various objects on top, under linear and spiral camera motion, respectively. These real sequences were chosen for their relative suitability for NVS and their adoption of a modern, high-resolution event sensor — Prophese Gen 4.

4.1. Synthetic Experiments

The synthetic experiments form the core of the benchmark, as they allow for realistic controlled experiments under various real-world conditions with absolute ground truth, which are otherwise infeasible using real sequences.

All sequences are simulated with a symmetric contrast threshold of 0.25 ($i.e., C_{1} = C_{-1} = 0.25$), which is the approximate nominal threshold for event sensors such as the Prophese Gen 3.1, Gen 4.1 and Sony IMX636 sensor, hence providing a known threshold ratio of 1 ($i.e., C_{1}/C_{-1} = 1$). The event camera trajectory is also sampled at a high rate of 1 kHz to minimize the influence of pose errors on the performance assessment. Furthermore, the virtual event camera revolves the object 4 times, with uniform 1 revolution per second speed about the object vertical axis by default, similar to [41]. Unless otherwise stated, a sequence is also simulated with zero pixel-to-pixel threshold standard deviation and refractory period ($i.e., \sigma_{\tau} = 0, \tau = 0$). By default, Ev-NeRF and our method are trained with a constant symmetric contrast threshold, given by the prior knowledge of time-independent threshold ratio.

Effect of Speed Profile. To investigate the influence of camera speed, we evaluate all works on 4 sets of sequences simulated with different speed profiles, where each set contains a sequence for each of the 7 scenes. Particularly, the 1st set is simulated with the default settings, thus providing a reference (uniform azimuth) speed of motion ($i.e., v = 1 \times$). We oscillate the speed of motion between $\frac{1}{8} \times$ and $8 \times$ the original speed at a frequency of 1 Hz ($i.e., v = v_{b} \sin 2\pi ft$, where $v_{b} = 8 \times$ and $f = 1 Hz$) for the 2nd set, and scale the speed of motion to $\frac{1}{8} \times$ and $8 \times$ of the original speed ($i.e., v = \frac{1}{8} \times, 8 \times$) for the 3rd and 4th sets, respectively. Fig. 1(b) illustrates the uniform motion in the 1st, 3rd and 4th sets, whereas Fig. 1(d) illustrates the non-uniform, oscillating motion in the 2nd set.

The quantitative results reported in Tab. 1 underscores the significance of speed invariance in enabling the effective generalization to arbitrary speed profiles, as it can be observed that the performance of Ev-NeRF deteriorates as the speed deviates from the optimal at $v = 1 \times$. Our method also outperforms all baselines, including Ev-NeRF, at the default setting ($i.e., v = 1 \times$), which is optimal for all. Qualitative results shown in Fig. 4 suggests an overall improvement in synthesis quality, including high-frequency details.
Table 1. Effect of speed profile. $v$ denotes the speed of motion relative to the default hemi-/spherical spiral motion with uniform azimuth speed, whereas $v_b$ denotes the oscillation factor of the relative speed of motion (i.e. $v = v_b \sin(2\pi ft)$, $f = 1$ Hz).

<table>
<thead>
<tr>
<th>Method</th>
<th>$v = 1 \times$</th>
<th>$v_b = 8 \times$</th>
<th>$v = \frac{1}{3} \times$</th>
<th>$v = 8 \times$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
</tr>
<tr>
<td>E2VID + NeRF</td>
<td>18.92 0.832 0.316</td>
<td>18.92 0.832 0.316</td>
<td>18.92 0.832 0.316</td>
<td>18.92 0.832 0.316</td>
</tr>
<tr>
<td>Ev-NeRF</td>
<td>27.72 0.935 0.087</td>
<td>26.25 0.926 0.102</td>
<td>19.79 0.792 0.326</td>
<td>20.83 0.862 0.198</td>
</tr>
<tr>
<td>Robust e-NeRF</td>
<td>28.19 0.945 0.057</td>
<td>28.19 0.945 0.057</td>
<td>28.19 0.945 0.057</td>
<td>28.19 0.945 0.057</td>
</tr>
</tbody>
</table>

Table 2. Effect of pixel-to-pixel threshold variation $\sigma_{C_p}$. “Opt. $C_p$” refers to jointly optimizing thresholds $C_p$ with NeRF parameters $\Theta$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Opt. $C_p$</th>
<th>Opt. $\tau$</th>
<th>$\sigma_{C_p} = 0.00$</th>
<th>$\sigma_{C_p} = 0.03$</th>
<th>$\sigma_{C_p} = 0.06$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>E2VID + NeRF</td>
<td>–</td>
<td>–</td>
<td>18.92 0.832 0.316</td>
<td>18.68 0.827 0.330</td>
<td>18.03 0.808 0.363</td>
</tr>
<tr>
<td>Ev-NeRF</td>
<td>✓</td>
<td>✓</td>
<td>27.43 0.911 0.123</td>
<td>23.66 0.826 0.261</td>
<td>15.43 0.708 0.441</td>
</tr>
<tr>
<td>Robust e-NeRF</td>
<td>×</td>
<td>✓</td>
<td>28.19 0.945 0.057</td>
<td>28.14 0.946 0.058</td>
<td>28.23 0.947 0.057</td>
</tr>
</tbody>
</table>

Table 3. Effect of refractory period $\tau$. “Opt. $\tau$” refers to jointly optimizing refractory period $\tau$ with NeRF parameters $\Theta$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Opt. $C_p$</th>
<th>Opt. $\tau$</th>
<th>$\tau = 0\text{ms}$</th>
<th>$\tau = 8\text{ms}$</th>
<th>$\tau = 25\text{ms}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
</tr>
<tr>
<td>E2VID + NeRF</td>
<td>–</td>
<td>–</td>
<td>18.92 0.832 0.316</td>
<td>14.87 0.797 0.427</td>
<td>14.15 0.791 0.467</td>
</tr>
<tr>
<td>Ev-NeRF</td>
<td>✓</td>
<td>✓</td>
<td>27.43 0.911 0.123</td>
<td>13.17 0.707 0.559</td>
<td>12.75 0.759 0.528</td>
</tr>
<tr>
<td>Robust e-NeRF</td>
<td>×</td>
<td>✓</td>
<td>28.19 0.945 0.057</td>
<td>23.60 0.934 0.066</td>
<td>25.51 0.929 0.072</td>
</tr>
</tbody>
</table>

Table 4. Collective effect of speed profile, threshold variation and refractory period.

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$\ell_{\text{grad}}$</th>
<th>$v_b = 1 \times$, $\sigma_{C_p} = 0.00$, $\tau = 0\text{ms}$</th>
<th>$v_b = 4 \times$, $\sigma_{C_p} = 0.03$, $\tau = 8\text{ms}$</th>
<th>$v_b = 8 \times$, $\sigma_{C_p} = 0.06$, $\tau = 25\text{ms}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR ↑</td>
<td>SSIM ↑</td>
<td>LPIPS ↓</td>
<td>PSNR ↑</td>
</tr>
<tr>
<td>×</td>
<td>✓</td>
<td>28.19 0.945 0.057</td>
<td>12.77 0.799 0.372</td>
<td>12.41 0.798 0.412</td>
</tr>
<tr>
<td>✓</td>
<td>×</td>
<td>27.96 0.943 0.063</td>
<td>23.15 0.899 0.113</td>
<td>22.21 0.879 0.153</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>28.19 0.945 0.057</td>
<td>24.10 0.913 0.086</td>
<td>23.51 0.900 0.110</td>
</tr>
</tbody>
</table>

Table 5. Ablation studies on refractory period $\tau$ modeling and target-normalized gradient loss $\ell_{\text{grad}}$.

Effect of Pixel-to-Pixel Threshold Variation. To evaluate the robustness of our method to noise, in the form of threshold variation, we benchmark all works on three sets of sequences. Furthermore, we also benchmark Ev-NeRF and our method with jointly optimized contrast thresholds, which are poorly initialized with $C_{+1}/C_{-1} = 10$ (more
Figure 4. Synthesized novel views in the synthetic and real experiments. Results shown correspond to that with jointly optimized contrast thresholds and possibly refractory period, if applicable for the method in the particular experiment.

The quantitative results given in Tab. 2 clearly demonstrate the robustness of our method to threshold variation, as our performance is essentially unaffected by the degree of noise. In contrast, the baselines are visibly susceptible to threshold variation, as their performance declines with its severity, especially Ev-NeRF. It is interesting to note that even the naïve baseline outperforms Ev-NeRF at \( \sigma_{C_p} = 0.06 \). These conclusions are also supported qualitatively in Fig. 4. The quantitative results also illustrate the effectiveness and robustness of our threshold joint optimization, as the performance of our method with and without jointly optimized thresholds are virtually the same. In contrast, it can be observed that the joint optimization of time-varying contrast thresholds in Ev-NeRF leads to a slight decrease in performance on the default setting (i.e. \( \sigma_{C_p} = 0.00 \)).

**Effect of Refractory Period.** To study the effect of temporal sparsity in the event stream due to the refractory period, we benchmark all methods on three sets of sequences simulated with refractory periods of 0, 8 and 25ms (i.e. \( \tau = 0, 8, 25\text{ms} \)). Similar to the previous experiment, we additionally evaluate our method with jointly optimized refractory period, which is poorly initialized with half the maximum possible value, given by the minimum time interval between successive events at a pixel in the sequence.
Furthermore, we also benchmark Ev-NeRF with jointly optimized thresholds to validate the importance of accounting for the refractory period. $\tau = 8\text{ms}$ is just slightly less than mean event interval in the set of sequences with $\tau = 0\text{ms}$. Let $E_\tau$ be the event stream with refractory period $\tau$, the degree of sparsity due to $\tau$ can be quantified by the mean $|E_{\tau=0\text{ms}}|/|E_0|$ across all sequences in the set, which translates to $5.84\times$ and $11.37\times$ for $\tau = 8$ and $25\text{ms}$, respectively.

The quantitative results presented in Table 3 undoubtedly verifies the importance of refractory period modeling in enabling the robust reconstruction from temporally sparse event streams, as our method significantly outperforms all baselines across all values of $\tau$. Moreover, the attempt by Ev-NeRF to compensate $\tau$ with jointly optimized thresholds has also clearly failed, albeit contributing to slightly improved performance. It is also worth noting that the naive baseline achieves better reconstructions than Ev-NeRF under non-zero $\tau$. Our perceivable drop in performance, as $\tau$ increases, may be attributed to its associated decrease in observability. Our results also demonstrate the feasibility of jointly optimizing $\tau$, although less effective than threshold joint optimization, which reflects the complexities involved.

**Collective Effect.** To assess the performance of all methods under the collective effect of threshold variation, refractory period and camera speed, which resembles real-world operating conditions, we benchmark all works on three sets of sequences with different levels of difficulty: easy ($\sigma_{C_p} = 0.00$, $\tau = 0\text{ms}, v_b = 1\times$), medium ($\sigma_{C_p} = 0.03$, $\tau = 8\text{ms}, v_b = 4\times$) and hard ($\sigma_{C_p} = 0.06$, $\tau = 25\text{ms}, v_b = 8\times$). In addition, we also benchmark Ev-NeRF and our method with jointly optimized intrinsic parameters, similar to previous experiments. The medium and hard sets are $5.31\times$ and $8.27\times$ sparser than the easy set, respectively. This experiment is particularly challenging as a $K \times$ speed-up effectively scales $\tau$ by $K \times$ in the context of the original sequence. Moreover, methods that involve event accumulation will also experience a $K \times$ effective scaling in $\sigma_{C_p}$.

The quantitative results provided in Table 4 generally reflect that of the previous experiment, thus the same conclusions can be drawn. This should not be surprising, given the proven importance of accounting for $\tau$. Our decline in performance on the medium and hard settings, when compared to $\tau = 8$ and $25\text{ms}$, may be attributed to the complications arising from the interaction between $\sigma_{C_p}$, $\tau$ and $v_b$. Qualitative results given in Fig. 4 illustrate the strength and robustness of our method under challenging conditions.

### 4.2. Real Experiments

The real experiments mainly serve as a qualitative benchmark, as the dataset (and other similar ones) is not specifically suited for the task of NVS, thereby unable to provide an accurate quantification of performance (refer to supplementary materials for a detailed justification). For the real experiments, Ev-NeRF and our method are trained with jointly optimized intrinsic parameters, where the contrast threshold is initialized with $C_{+1}/C_{-1} = 1$ (more precisely, we set $C_{-1} = 1$ and $C_{+1} = 0.25$) this time. While the two sequences mainly involve an insignificant refractory period relative to their primarily non-uniform speed of motion, qualitative results shown in Fig. 4 still demonstrate our superior performance in recovering fine details while appropriately smoothing uniform regions, in line with the synthetic experiments. Exposure of the scene is also visibly more consistent and more accurately reconstructed. Visible artifacts near the borders of all synthesized views are due to the comparably narrower field-of-view of the event camera.

#### 4.3. Ablation Studies

The ablation studies are conducted in the same setting as the synthetic experiment on the collective effect, without jointly optimized intrinsic parameters. The quantitative results reported in Table 5 further validates the immense importance of accounting for $\tau$, even when event accumulation is not involved. Although the synthetic sequences mainly involves scenes with abundant texture, the results also verify the effectiveness of the target-normalized gradient loss in regularizing textureless regions, particularly under more challenging conditions (i.e. medium and hard settings).

### 5. Conclusion

In this paper, we introduce Robust $e$-NeRF, a novel approach to directly and robustly reconstruct NeRFs from moving event cameras under various real-world conditions, particularly from sparse and noisy events generated under non-uniform motion. Our method consists of two key components: a realistic event generation model that accounts for various event camera-specific intrinsic parameters and non-idealities, as well as a complementary pair of normalized reconstruction losses that can effectively generalize to arbitrary speed profiles and intrinsic parameter values without such prior knowledge. The proposed Analysis-by-Synthesis approach for event-based reconstruction naturally extends to other scene representations, such as 3D Gaussians [14]. In spite of its achievements, Robust $e$-NeRF still relies on the assumption of a static scene and given constant-rate camera poses, which we leave for future work. We also see the modeling of other spatial/temporal noise and artifacts of an event sensor as a promising future direction for more accurate reconstructions.

### Acknowledgements
This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-024), and the Tier 2 grant MOE-T2EP20120-0011 from the Singapore Ministry of Education.
References


