Supplementary Material for
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

In this supplementary document, we first show how event accumulation results in an effective amplification of pixel-to-pixel contrast threshold variation (Sec. A) and discuss the optimality of normalization in the threshold-normalized difference loss (Sec. B). Next, we present implementation details of Robust *e*-NeRF and the baselines used in the experiments (Sec. C). We also provide a detailed justification on the qualitative nature of our real experiments (Sec. D). Lastly, we present additional quantitative and qualitative results on all experiments (Sec. E).

A. Amplification of Threshold Variation

As alluded in Sec. 3.3.3, the accumulation of successive events at each pixel over time intervals leads to the effective amplification of pixel-to-pixel contrast threshold variation. This can be shown by simply analyzing the distribution of the target log-radiance difference after event accumulation, at any given pixel.

The time-independent contrast threshold of polarity p can be modeled as a random variable $c_p \sim \mathcal{N}(C_p, {\sigma_{C_p}}^2)$ (Sec. 3.2). Assuming N_p number of polarity p events are accumulated at the pixel within the specified time interval, the target log-radiance difference $\Delta \log L_{acc}$ is then given by:

$$\Delta \log L_{acc} = \sum_p p N_p c_p \ , \tag{9}$$

which follows the Gaussian distribution below:

$$\mathcal{N}\left(\sum_p p N_p C_p, \ \sum_p {N_p}^2 {\sigma_{C_p}}^2 - 2 N_{+1} N_{-1} \sigma_{c_{+1}, c_{-1}}\right) \ , \tag{10}$$

where $\sigma_{c_{+1},c_{-1}} \in \left[-\sigma_{c_{+1}}\sigma_{c_{-1}}, \ \sigma_{c_{+1}}\sigma_{c_{-1}}\right]$ is the covariance between c_{+1} and c_{-1}.

Note that when N_{+1} and N_{-1} increases by a factor of K, the standard deviation of $\Delta \log L_{acc}$ will also increase by the same factor, which results in noisier targets. Moreover, assuming that c_{+1} and c_{-1} do not have a strong positive correlation (*i.e.* $\sigma_{c_{+1},c_{-1}} \ll \sigma_{c_{+1}}\sigma_{c_{-1}}$, with respect to the range of $\sigma_{c_{+1}c_{-1}}$), which is highly likely to be true, it can also be shown that standard deviation of $\Delta \log L_{acc} \gg |\sum_p p N_p \sigma_{C_p}| \geq 0$ under non-zero N_{+1} and N_{-1}. This suggests that when $N_{+1}C_{+1} \approx N_{-1}C_{-1}$, which often holds true in practice over sufficiently long accumulation intervals (relative to the speed of motion and amount of scene texture), the mean of $\Delta \log L_{acc} = \sum_p p N_p C_p \approx 0$ whereas the standard deviation remains very much larger than 0, especially for large N_{+1} and N_{-1}. Such a cancellation between positive and negative accumulated events further aggravates the target noise. All these observations suggest an effective amplification of threshold variation when event accumulation is involved.

B. Optimality of Normalization in ℓ_{diff}

As mentioned in Sec. 3.3.3, the threshold-normalized difference loss ℓ_{diff} (Eq. 6) is optimal in the sense that the magnitude of the normalized target $|{}^{pC_p}/_{\bar{C}}|$, which is essentially the normalized threshold ${}^{C_p}/_{\bar{C}}$, is always centered at 1 regardless of the threshold ratio ${}^{C_{+1}}/_{C_{-1}}$, as follows:

$$\left|\frac{pC_p}{\bar{C}}\right| = \frac{C_p}{\bar{C}} = 1 + p\frac{\tilde{C}}{\bar{C}} \tag{11}$$

where $\tilde{C} = \frac{1}{2}(C_{+1} - C_{-1})$ and the magnitude of the offset ${}^{\tilde{C}}/_{\bar{C}}$ can be interpreted as the normalized threshold difference. This facilitates the scale consistency of the loss, thus enabling the adoption of a single, global loss weight λ_{diff} for arbitrary contrast threshold values. Nevertheless, the variance of the normalized target increases as the thresholds become more asymmetric.

C. Implementation Details

C.1. Robust e-NeRF

Architecture. Robust e-NeRF adopts Instant-NGP [5] as the NeRF backbone, as it allows for high-quality reconstructions given relatively low training time and memory cost. More precisely, we employ the implementation provided by the NerfAcc toolbox [4], due to its simple and flexible Python APIs, but with some slight modifications.

In particular, parameters of the embedded *Multi-Layer Perceptron* (MLP) are initialized using the PyTorch-default method, instead of *Xavier* initialization [2]. Furthermore, we replace all *Rectified Linear Unit* (ReLU) hidden layer activations with *SoftPlus* ($\beta = 100$) as it is infinitely differentiable everywhere, thereby facilitating the optimization of ℓ_{grad}.

Since the predicted *log*-radiance is at most accurate up to an offset per color channel (Sec. 3.3.2), or equivalently the predicted *linear* radiance (modeled by NeRF) is at most accurate up to a scale per color channel, we also replace the bounded sigmoid radiance output activation with the lower-bounded SoftPlus (default $\beta = 1$). In addition, we add a small $\epsilon = 0.001$ to the positive raw radiance output from the NeRF model (*i.e.* $\hat{\boldsymbol{L}} = \hat{\boldsymbol{L}}_{raw} + \epsilon$) to improve the numerical stability of the predicted log-radiance $\log \hat{\boldsymbol{L}}$. This augmentation imposes a lower bound of ϵ on the radiance our method can *model*, as $\hat{\boldsymbol{L}} > \epsilon$. Nevertheless, this is not a cause for concern given the minimum per-channel scale ambiguity of $\hat{\boldsymbol{L}}$, non-upper bounded range of $\hat{\boldsymbol{L}}_{raw}$ and non-zero scene radiance (*i.e.* absolute darkness is virtually impossible in practice).

For synthetic scenes, we also alpha composite $\hat{\boldsymbol{L}}_{raw}$ with a learnable background radiance, which is parameterized via SoftPlus to ensure that it is always positive, prior to ϵ-augmentation. In contrast, common NeRF backbones and EventNeRF [6] adopt a fixed background, which is inappropriate given the scale ambiguity.

As only the threshold ratio can be recovered during the joint optimization of contrast threshold (Sec. 3.3.3), we keep the negative threshold C_{-1} fixed at an arbitrary value and only optimize the learnable positive-to-negative contrast threshold ratio ${}^{C_{+1}}/_{C_{-1}}$, which is parameterized via SoftPlus to ensure that it is always positive. Moreover, since the refractory period is lower bounded at 0 and upper bounded by the minumum time interval between successive events at any pixel (Sec. 4.1), we parameterize the refractory period via a scaled sigmoid that preserves the gradient profile of the default, unscaled sigmoid function. We additionally clamp the parameterized refractory period between ε and $(1 - \varepsilon)\times$ its range to limit the minimum gradient of the scaled sigmoid to approximately $\varepsilon\times$ the range. This prevents vanishing gradients at the extremes, which implicates the optimization of the refractory period.

For real scenes, we appropriately predefine the *Axis-Aligned Bounding Box* (AABB), as well as the near and far bounds of the back-projected rays used for volume rendering, for each scene. Furthermore, we employ the spherical space contraction proposed in mip-NeRF 360 [1] to better model unbounded scenes. We also increase the occupancy grid resolution to 256^3 and set the cone angle (*i.e.* ray marching step size increment scale) to 0.004, which is approximately ${}^{1}/_{256}$ as suggested by Instant-NGP.

Training. The training loss weights used in all experiments are given by $\lambda_{diff} = 1$ and $\lambda_{grad} = 0.001$. As suggested by Instant-NGP, we also impose a weight decay of 10^{-6} on the MLP to prevent overfitting. The model is trained for 40 000 iterations with a learning rate decay of 0.33 at 20 000, 30 000 and 36 000 iterations (*i.e.* 50%, 75% and 90% progress, as done in NerfAcc), using the Adam optimizer [3] with a learning rate of 0.01 and PyTorch-default hyper-parameters. During joint optimization of contrast threshold, its parameter is assigned a higher learning rate of 0.1 to facilitate to its early convergence. Moreover, since the scaled sigmoid function preserves its gradient profile, but the range of the refractory period may vary greatly, the learning rate assigned to the (unscaled logit) parameter of refractory period is set to $50\times$ the range. The event batch size is determined dynamically based on the average number of ray samples used to render a single pixel, similar to Instant-NGP, to maximize the utilization of the GPU memory. Specifically, we ensure that every batch of events involves approximately $2^{20} = 1\ 048\ 576$ samples in total, for either the rays at t_{ref}, t_{curr} (relevant to ℓ_{diff}) *or* t_{sam} (relevant to ℓ_{grad}). As a side note, the poses of the target novel views in the real experiments are interpolated from the given unsynchronized constant-rate camera poses using LERP and SLERP.

C.2. Baselines

As alluded in Sec. 4, both baselines have been carefully reimplemented on the same NerfAcc backbone and trained with the same hyper-parameters (including the weight decay), when applicable, to facilitate a fair comparison. However, we only train the naïve baseline of E2VID + NeRF for 20 000 iterations with a learning rate decay of 0.33 at 10 000, 15 000 and 18 000 iterations (*i.e.* 50%, 75% and 90% progress) due to its comparably faster convergence, as a result of the direct absolute radiance supervision. Similar to the target novel views, the poses of the E2VID-reconstructed training views are also interpolated from the given unsynchronized constant rate camera poses using LERP and SLERP. Furthermore, we extend the implementation of E2VID to support the RGGB *Bayer* pattern adopted in ESIM.

D. Justification of Qualitative Real Exps.

As mentioned in Sec. 4.2, we mainly perform qualitative evaluation for the real experiments. This is done because the target novel views, given by a separate standard camera, suffer from saturation due to the comparably smaller dynamic range of the standard camera, and are not raw images that have not been processed by the lossy in-camera image processing pipeline. Moreover, the spectral sensitivity curve of the event camera adopted is also not documented, hence gamma correction may not accurately align the synthesized views.

Furthermore, the comparably narrower *field-of-view* of the event camera and the limited camera motion also leads to a relatively smaller coverage of the scene, thereby causing artifacts in the synthesized novel views near the borders, as observed in the qualitative results. This further complicates the quantitative evaluation as it is non-trivial to delineate the valid synthesis regions. Other event camera datasets also suffer from similar issues, as all are not specifically suited for novel view synthesis.

E. Additional Experiment Results

E.1. Per-Scene Breakdown

Tab. 6 and Fig. 5, 6 show the quantitative and qualitative results of all methods, respectively, for each of the seven synthetic scene sequences simulated with the default settings, which is optimal for all methods. The per-scene quantitative results is generally consistent with the aggregate metrics, which is also presented in Sec. 4.1, as our method outperforms the baselines in most scenes and has comparable performance in others. The per-scene qualitative results reveal our superior performance in reconstructing fine details and maintaining high color accuracy, especially at the background, as previously observed in Sec. 4.1.

E.2. Qualitative Analysis of ℓ_{grad}

Fig. 7 illustrates the effect of target-normalized gradient loss ℓ_{grad} on the `hotdog` and `chair` scene sequences simulated with the easy and hard settings, respectively, as similarly done in Sec. 4.3. It can be observed that with ℓ_{grad}, the plate of the hotdog and the back of the chair exhibit less noise, especially the latter. This is achieved while preserving high-frequency details on the hotdog and the cushion of the chair. This further validates the effectiveness of ℓ_{grad} in regularizing textureless regions, particularly under challenging conditions.

E.3. Qualitative Results on `office-maze`

Apart from `mocap-1d-trans` and `mocap-desk2`, we also benchmark all methods on the `office-maze` sequence from the TUM-VIE dataset. We only employ the subsequence before the 395$^{\text{th}}$ target novel view, as it captures a bounded space of an office (in approximately 2 loops around the office). The qualitative results reported in Fig. 8 clearly shows our effectiveness in recovering details and resolving the scene structure without suffering from severe fogs in free space.

E.4. Robustness to Temporal Event Sparsity

To evaluate the robustness of our method to temporal sparsity of the event stream (*i.e.* data efficiency), we benchmark it on a set of nine sequences simulated on the synthetic `chair` scene with different refractory periods. Apart from the standard image similarity performance metrics, we also report some statistics such as the percentage of τ relative to the duration of the event sequence, as well as the degree of sparsity of the event stream, as defined in Sec. 4.1. Moreover, we also report the number of images that occupy an equivalent amount of memory as the event sequence disregarding compression, assuming 8 bits per image pixel channel and 47 bits per event (*i.e.* 2×11 bits for position, 1 bit for polarity and 24 bits for timestamp), as implied after decompression of the Prophesee EVT 3.0 [7] event encoding format.

The quantitative and qualitative results given in Tab. 7, Fig. 9 and Fig. 10 demonstrate our astonishing robustness under severely sparse event streams, which suggests that our method is highly data efficient. It is worth noting that our method can still reconstruct the scene with reasonable accuracy with $\tau = 1000ms$, where only 3 equivalent views are used and each pixel can only generate at most 4 events throughout the sequence. The event stream is also around $200\times$ sparser than the default with $\tau = 0ms$.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010. 2

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[4] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. NerfAcc: A General NeRF Acceleration Toolbox, 2022. arXiv:2210.04847 [cs]. 2

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics*, 2022. 2

Metric	Method	Synthetic Scene							Mean
		`chair`	`drums`	`ficus`	`hotdog`	`lego`	`materials`	`mic`	
PSNR ↑	E2VID + NeRF	19.62	19.52	22.44	17.33	17.41	18.13	18.02	18.92
	Ev-NeRF	28.93	**23.89**	28.37	25.22	**29.10**	**26.50**	32.03	27.72
	Robust *e*-NeRF	**30.24**	23.15	**30.71**	**28.07**	27.34	24.98	**32.87**	**28.19**
SSIM ↑	E2VID + NeRF	0.869	0.842	0.863	0.859	0.710	0.835	0.844	0.832
	Ev-NeRF	0.932	0.889	0.948	0.940	0.930	**0.926**	0.979	0.935
	Robust *e*-NeRF	**0.958**	**0.897**	**0.971**	**0.953**	**0.934**	0.923	**0.981**	**0.945**
LPIPS ↓	E2VID + NeRF	0.277	0.277	0.289	0.341	0.406	0.282	0.337	0.316
	Ev-NeRF	0.085	0.203	0.085	0.103	**0.058**	0.054	**0.024**	0.087
	Robust *e*-NeRF	**0.040**	**0.091**	**0.022**	**0.095**	0.074	**0.052**	0.029	**0.057**

Table 6. Per-synthetic scene breakdown under the default setting.

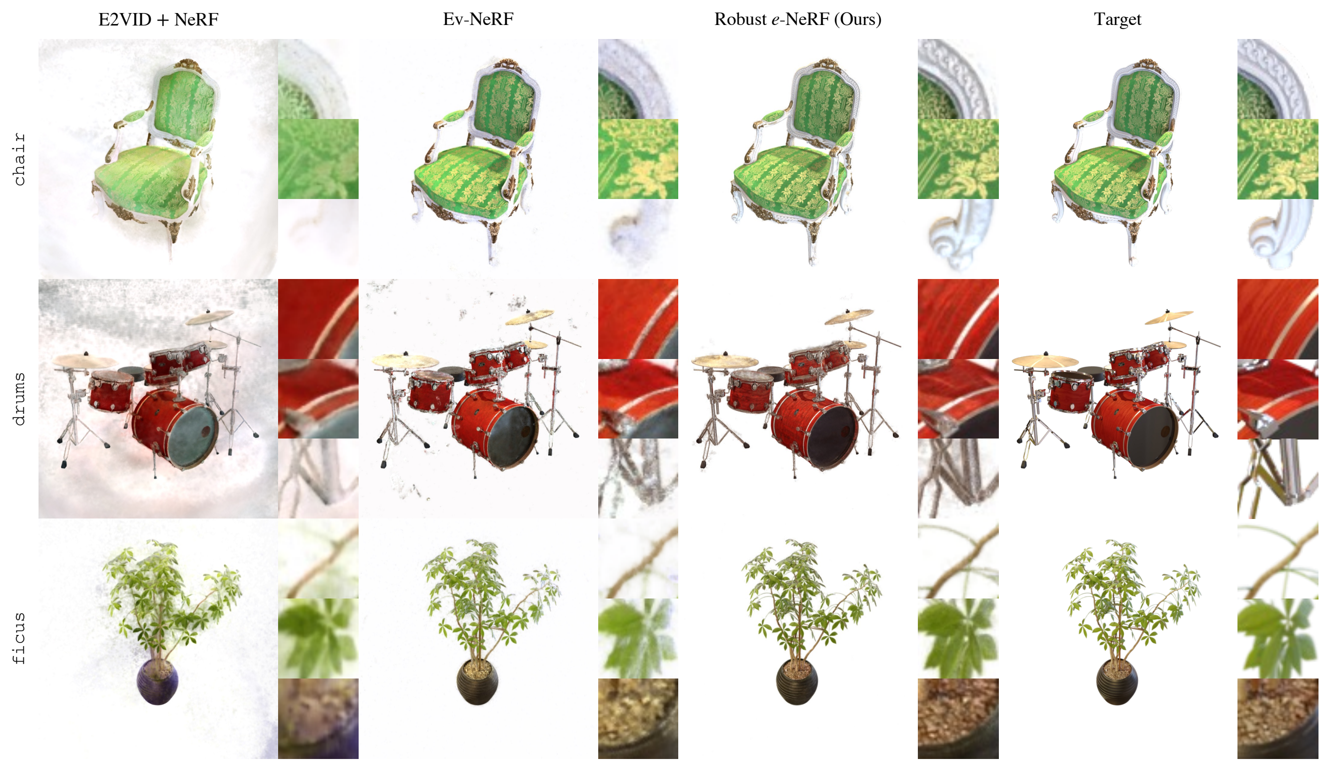

Figure 5. Synthesized novel views on `chair`, `drums` and `ficus` under the default setting.

[6] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single colour event camera, 2022. arXiv:2206.11896 [cs]. 2

[7] Prophesee S.A. Evt 3.0 format. https://docs.prophesee.ai/stable/data/encoding_formats/evt3.html. 3

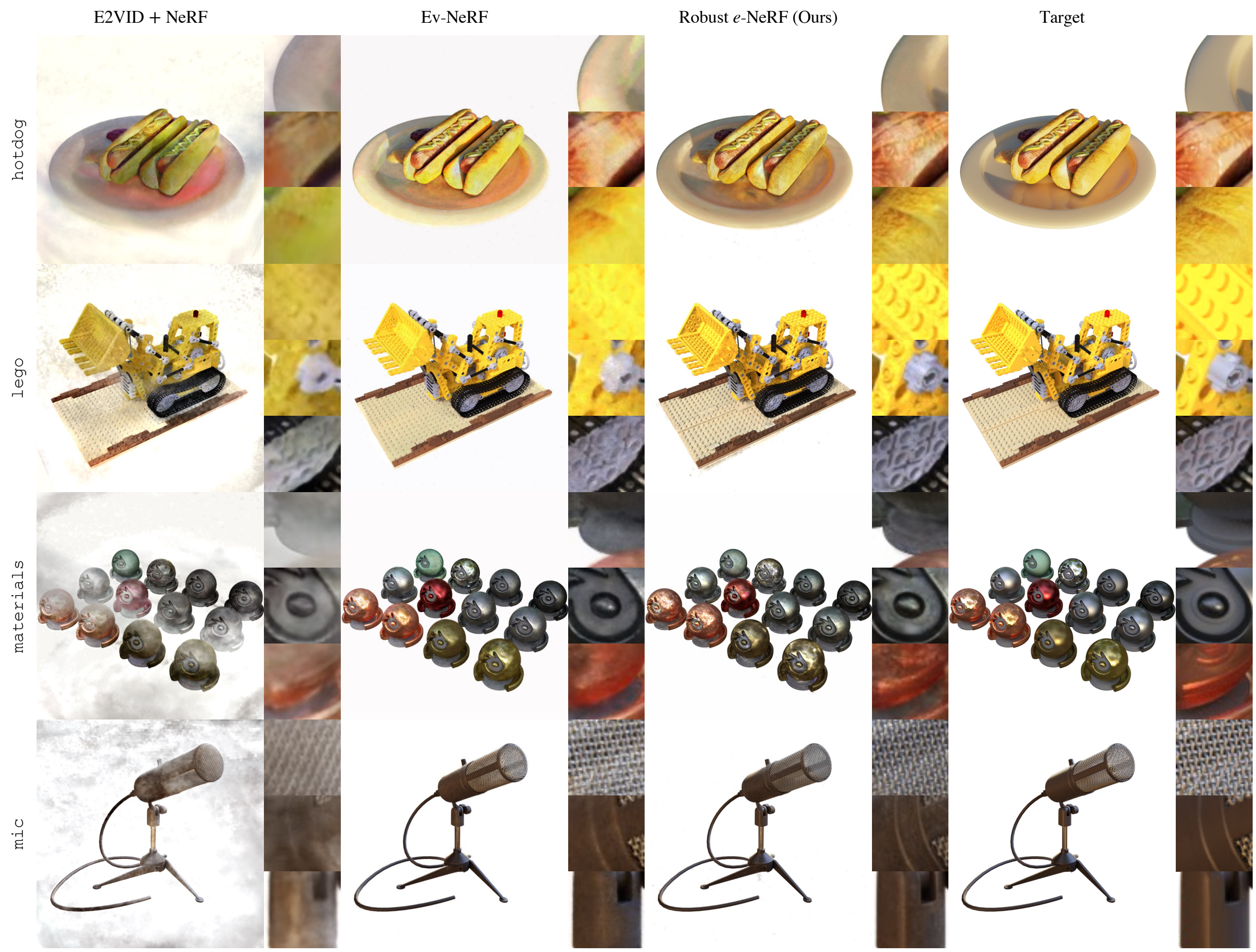

Figure 6. Synthesized novel views on `hotdog`, `lego`, `materials` and `mic` under the default setting.

Figure 7. Synthesized novel views with and without the target-normalized gradient loss ℓ_{grad}.

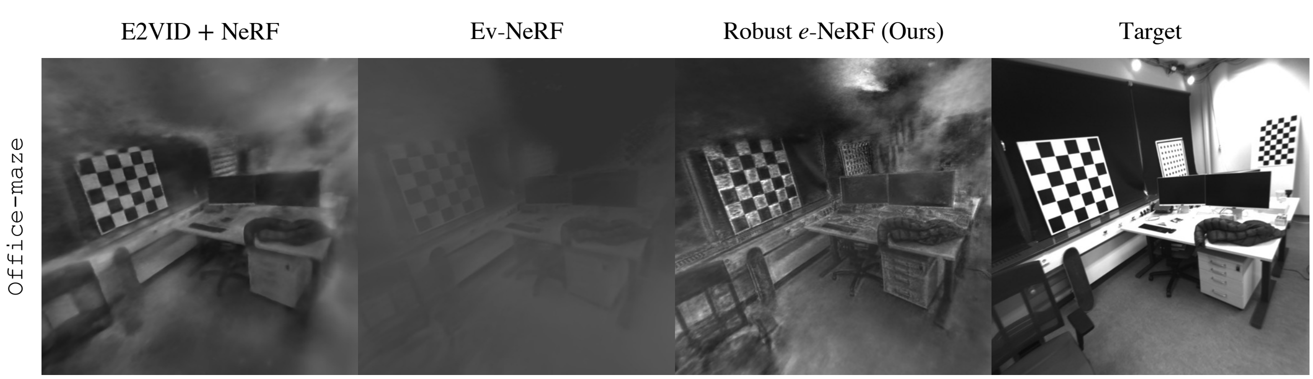

Figure 8. Synthesized novel views on the `office-maze` scene.

τ, ms	Statistics			Metrics		
	% Seq. Duration	Sparsity, $\times$	Equiv. # Views	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
0	0	1.000	336.8	30.24	0.958	0.040
8	0.2	4.176	80.66	30.41	0.959	0.042
25	0.625	8.440	39.90	29.84	0.958	0.041
50	1.25	13.50	24.95	29.20	0.953	0.046
100	2.5	21.27	15.83	27.40	0.938	0.060
250	6.25	40.80	8.255	25.95	0.916	0.081
500	12.5	67.77	4.970	24.08	0.900	0.102
1000	25	110.5	3.048	22.10	0.854	0.204
2000	50	209.6	1.607	17.05	0.762	0.398

Table 7. Robustness of our method to temporal event sparsity on the `chair` scene.

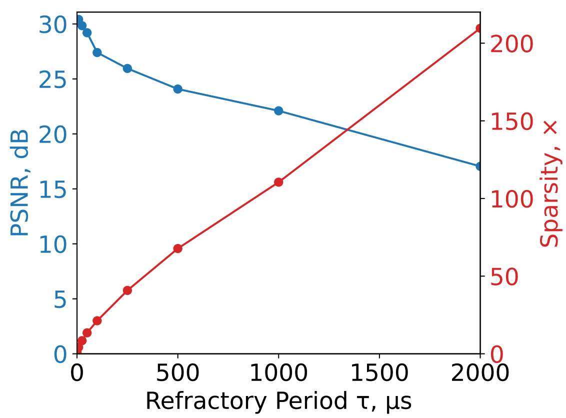

Figure 9. Plot of novel view synthesis PSNR and degree of event sparsity on the `chair` scene against refractory period τ.

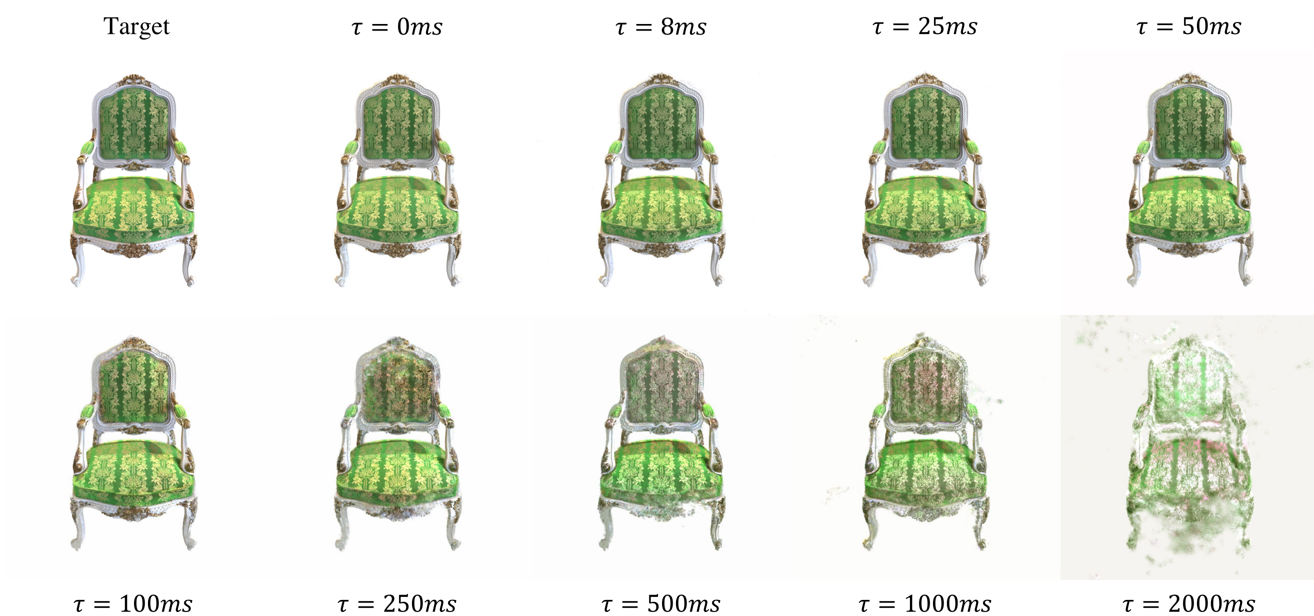

Figure 10. Synthesized novel views on the `chair` scene under numerous refractory periods τ.