Supplementary Material for Ev-NeRF: Event Based Neural Radiance Field

1. Additional Implementation Details

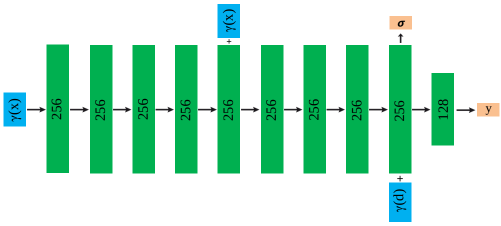


Figure 1. Network Architecture of Ev-NeRF. The network consists of fully-connected layers. The numbers in the colored blocks indicate the dimension of corresponding layers.

We implement Ev-NeRF using PyTorch using the NeRF formulation. For positional encoding, we use 10 frequencies for $\mathbf{x}$ and 4 for $\mathbf{d}$. The weight λ 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 [3] with a learning rate of 5×10^{-4}. Ev-NeRF takes about an hour in RTX 3090 GPU to train per scene and 1.5 seconds to render a single image.

Similar to NeRF, the neural network takes the 3D coordinate and the ray direction as input and outputs the volume density and emitted radiance. Sinusoidal positional encoding, $\gamma(\cdot)$, is applied to input variables. Figure 1 shows the detailed architecture of our network. The architecture mostly follows that of NeRF [4], however it predicts the emitted luminance value instead of color values. Following NeRF [4], we add zero-mean, unit-variance Gaussian random noise to the density σ for slightly improved performance.

2. Dataset Description

For the real-world data, we use a sub-sequence of the event sequences from IJRR [5], HQF [10] and Stereo DAVIS [12] for training. The data includes intensity images at regular time intervals (about 24 Hz) and asynchronous event data. These datasets are generated with a DAVIS240C [1] event camera and both intensity images and events have a resolution of 240×180. In our setup, we assume camera poses are given, which are calculated from running SfM [8, 9] with the intensity frames. Except for this process, the intensity frames are not available during training and are used only for evaluation. We use sub-sequences with a length of 50 to 100 intensity frames for training and the exact frame index corresponding to the original dataset [5, 10] is described in Table 1. For comparison with event-based SLAM, we use two sequences (simulation_3planes, reader) from [12].

For the synthetic data, we examine the extracted scene structure using models widely used for NeRF [4], namely lego, hotdog, mic, drums, and chair. Figure 2 shows examples of models and generated events.

Dataset	Scene	start frame index	end frame index
IJRR	office_zigzag	0	100
	office_spiral	0	100
	boxes	230	330
	dynamic_6dof	30	130
	hdr_boxes	70	120
HQF	reflective_materials	60	150
	high_texture_plants	930	1000
	still_life	300	350
Stereo DAVIS	monitor	0	100
	reader	0	100

Table 1. The start and end frame index for intensity frames used for our experiments. We use sub-sequences of original datasets, namely IJRR [5], HQF [10] and Stereo DAVIS [12].

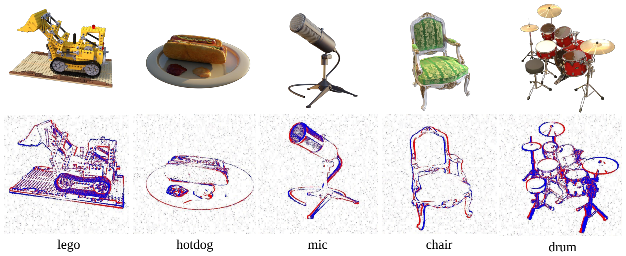


Figure 2. Visualization of generated events for each scene.

3. Description on Convergence

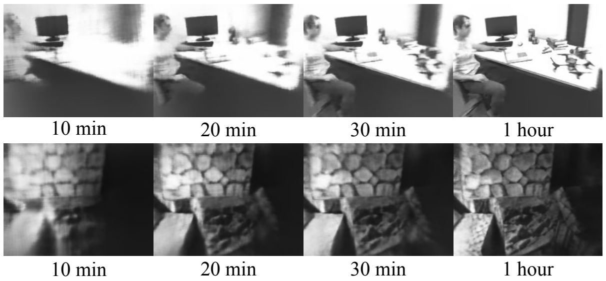
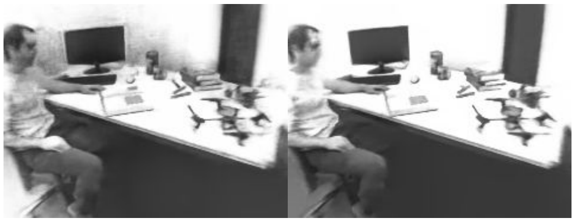


Figure 3. Intermediate results over the training time for each scene.

While events provide only changes of brightness, we empirically found it converges to reliable absolute brightness as the learning progresses. The supplementary video shows how Ev-NeRF converges on absolute brightness as training progresses. Figure 3 contains a few representative images. Ev-NeRF obtains a rough 3D structure in about 10 min, and the subsequent timesteps focus on capturing further details.

	MSE ↓			SSIM ↑			LPIPS ↓		
Scene	w/o joint	w/o noise inj.	full	w/o joint	w/o noise inj.	full	w/o joint	w/o noise inj.	full
office_zigzag	0.03	0.04	**0.03**	0.40	0.41	**0.42**	0.29	0.27	**0.27**
office_spiral	0.04	0.04	**0.03**	0.41	0.41	**0.41**	0.29	0.28	**0.27**
boxes	0.04	0.06	**0.04**	0.47	0.45	**0.48**	0.32	0.33	**0.31**
dynamic_6dof	0.21	0.24	**0.19**	0.25	0.24	**0.26**	0.42	0.43	**0.41**
reflective_materials	0.06	0.07	**0.05**	0.39	0.38	**0.40**	0.35	0.36	**0.35**
high_texture_plants	0.04	0.03	**0.03**	0.43	0.43	**0.44**	0.34	0.35	**0.34**
still_life	0.04	0.05	**0.03**	0.52	0.51	**0.53**	0.21	0.19	**0.18**
monitor	0.05	0.08	**0.03**	0.30	0.26	**0.32**	0.38	0.39	**0.37**
reader	0.11	0.12	**0.09**	0.44	0.43	**0.45**	0.38	0.36	**0.35**

Table 2. Ablation study on the effect of joint training of the sensor threshold values and noise injection. The columns display results without joint training, without noise injection, and the full model, respectively. Our full training method shows the optimal reconstructed image quality.



Without noise injection With noise injection

Figure 4. Visual analysis on noise injection, which helps solve ambiguity in places such as walls where events do not occur.

4. Ablation Studies

Joint Training As shown in Equation (8), we propose joint training that concurrently optimizes ΔI and the threshold values B_j^+, B_j^- of all timestamps with $\lambda = 1000$. We verify the advantage of our joint training on intensity reconstruction in Table 2. We compare the proposed joint training against the ablated version with pre-fixed threshold values ± 0.3 and $\lambda = 0$. Results show that the joint training scheme is beneficial to the reconstructed image quality.

Noise Injection We verify that the additional random noise slightly improves the quality of Ev-NeRF. When we compile the training data of events occurred during time slice $[T_j, T_{j+1})$, we add random events whose amount is 5% of the number of events that occurred at time. Table 2 numerically compares the quality of reconstructed images with random noise injection against the original event slice $\mathcal{E}_j$ and verifies that the noise injection is beneficial to the image quality. Figure 4 visualizes the effect of noise injection. The additional noise helps the neural volume to resolve ambiguity in the areas where events do not occur, such as the solid-colored wall behind the monitor.

5. Additional Results

Here we show more examples of quantitative and qualitative results. The results are presented in various scenes with three baselines: E2VID [7], E2VID+ [10] and ssl-E2VID [6], which are designed to reconstruct intensity images.

5.1. Intensity Image Reconstruction

Additional Quantitative Results Table 3 displays the quantitative comparison against baseline methods with real-world datasets, namely IJRR [5], HQF [10], and Stereo DAVIS dataset [12]. We compared using three metrics: mean squared error (MSE), structural similarity (SSIM) and perceptual similarity (LPIPS) [11]. For real-world data, Ev-NeRF outperforms ssl-E2VID and is on par with E2VID and E2VID+ without observing the ground truth intensity frames.

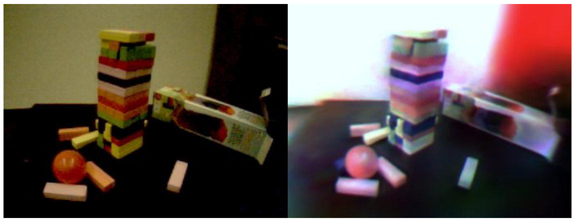
However, we used sub-sequences of real datasets and event camera trajectory from sub-sequences is not optimal for NeRF, which is typically trained with cameras located on the hemisphere around the object from roughly constant distances. This causes Ev-NeRF to show lower performance compared to E2VID+ for some sequences. As mentioned in Table 1 from the main paper, with typical trajectories for NeRF, the quality of intensity images of Ev-NeRF is then consistently superior to the baseline, especially in the presence of noise.

Additional Qualitative Results for Intensity Image and Novel View Synthesis Figure 6 contains the results of intensity image reconstruction for all of the sequences we use dwith a qualitative comparison with various baselines and Figure 7 shows image reconstruction results observed from various camera poses. Also, the supplementary video contains the event sequences used for training, paired with corresponding reconstructed intensity images and depth results. Figure 8 shows novel view synthesis observed from the viewpoints that are not available in the input dataset. Ev-NeRF maintains comparable performance for all scenes. Additional results on novel view synthesis are shown in the supplementary video.

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

Table 3. Quantitative comparison of image reconstruction on scenes from the IJRR [5], HQF [10] and Stereo DAVIS [12] dataset. The results with the best performance are in bold. We additionally underline the runner-up metric.

Intensity Image Reconstruction on Different Sensor Resolution We further validate the domain-invariance of Ev-NeRF with intensity image reconstruction results on the Color Event Camera Dataset (CED) [2]. The data is composed of three channels of color events in a different resolutions (346×240). All of the datasets presented in other sections are processed in the resolution of 240×180 bound to the sensor resolution. Ev-NeRF does not assume any fixed resolution of the scene and can be applied in multiple color channels without fine-tuning. We extend our approach to three color channels following the method suggested in [2]; we first reconstruct the intensity images via red, green, and blue channels respectively, and upsample the individual channels of the image to the original resolution to produce a single color image. Figure 5 shows an exemplar color DAVIS frame and reconstructed color image for simple_jenga scene from CED [2]. Ev-NeRF can find the scene structure within reasonable ranges. More importantly, there is no additional training to account for the domain shift in sensor characteristics or resolution.



(a) Color DAVIS frame (b) Our Reconstruction

Figure 5. Color image reconstruction through Ev-NeRF.

5.2. Noise Resistant Image Reconstruction

Ev-NeRF is extremely robust under noise and maintains the quality of the reconstruction under data of various noise levels. In addition to the visual example shown in Figure 4 of the main text, we show results given data corrupted with different levels of noise in Figure 9. Figure 10 contains more results of image reconstruction with severe noise of ratio 0.9. Ev-NeRF exhibits little degradation in performance and therefore can be useful in an extreme environments subject to unknown noise characteristics.

References

[1] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[2] 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.

[3] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] 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.

[10] 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.

[11] 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.

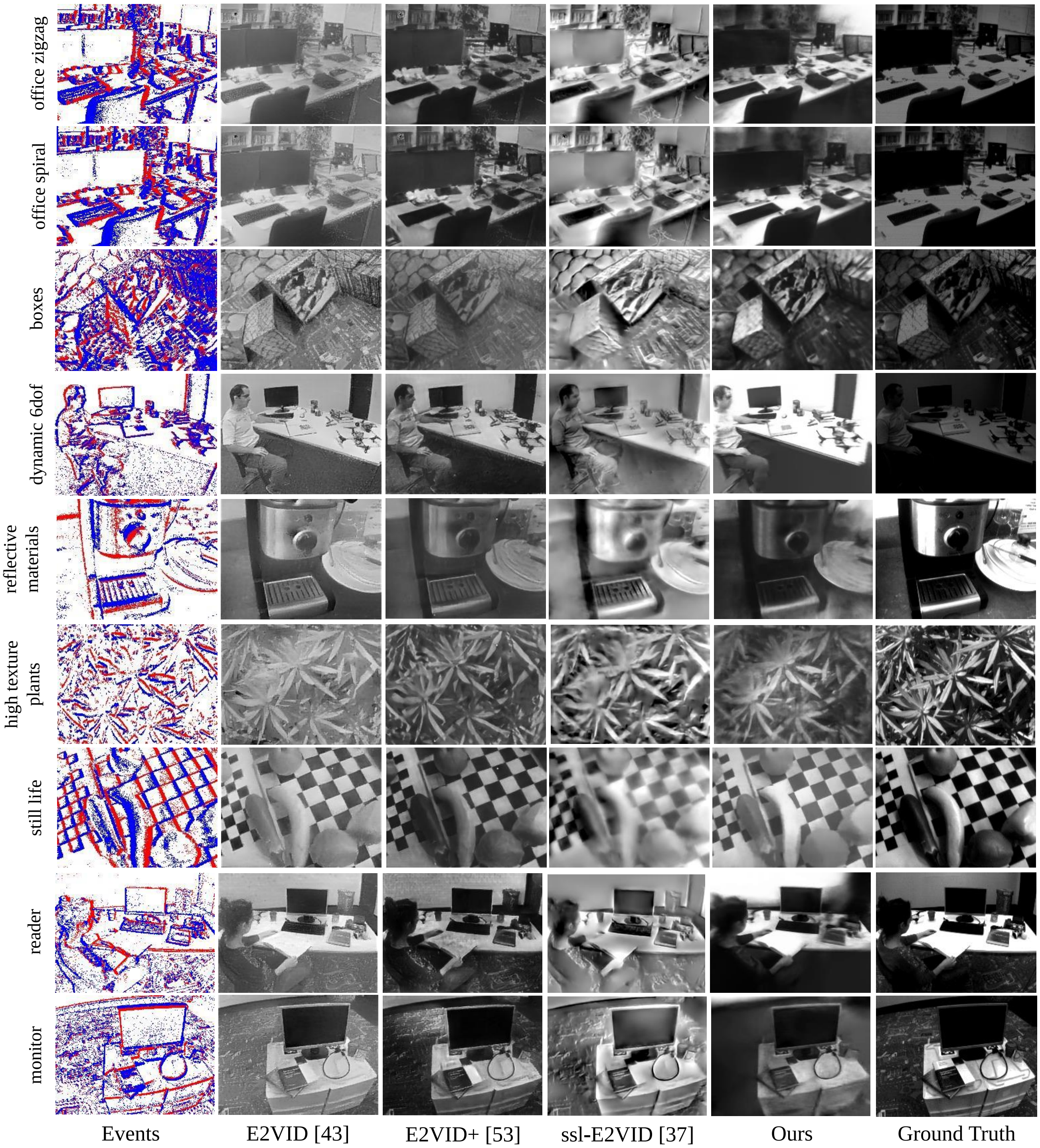


Figure 6. Qualitative comparison on intensity image reconstruction under various scenes.

[12] 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.

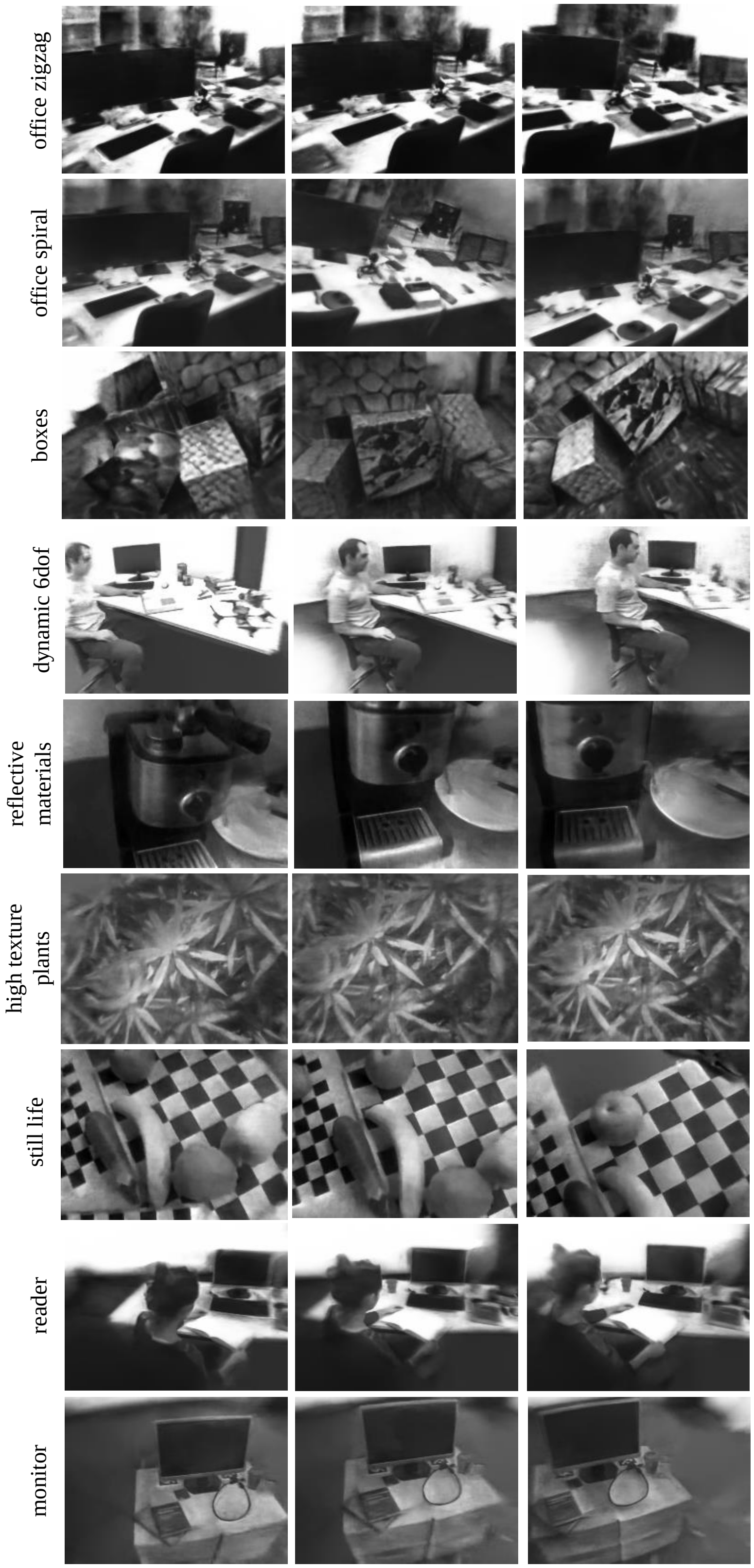


Figure 7. Qualitative intensity image reconstruction results at various time steps.

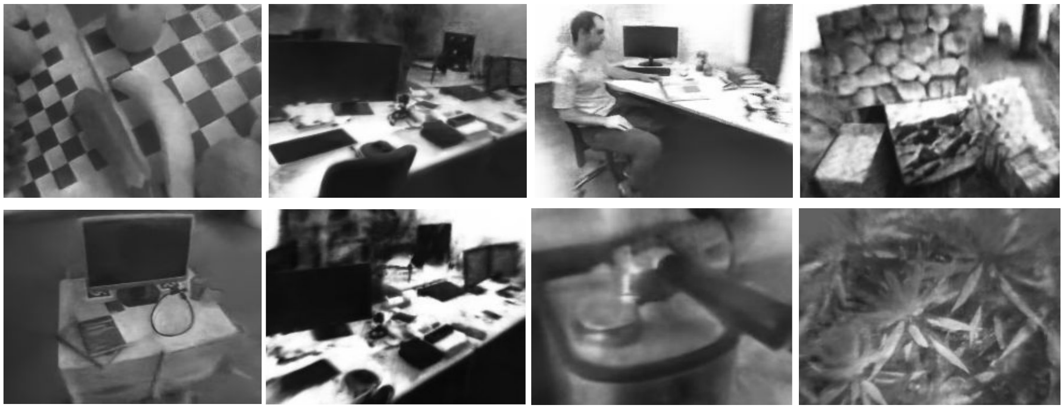


Figure 8. Qualitative results for novel view image reconstruction.

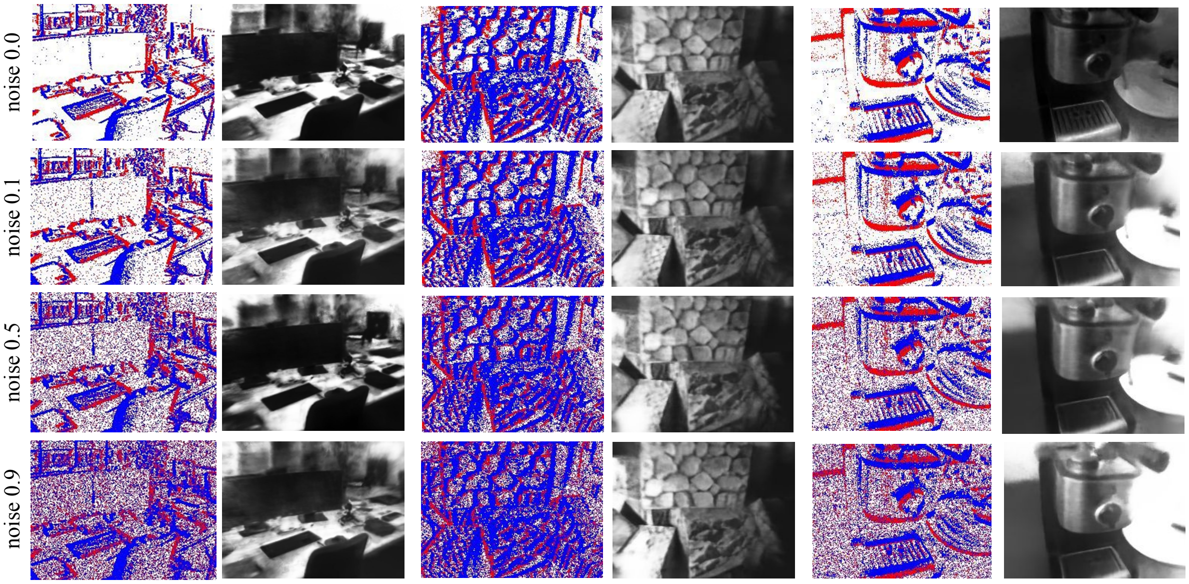


Figure 9. Qualitative comparison under different noise levels.

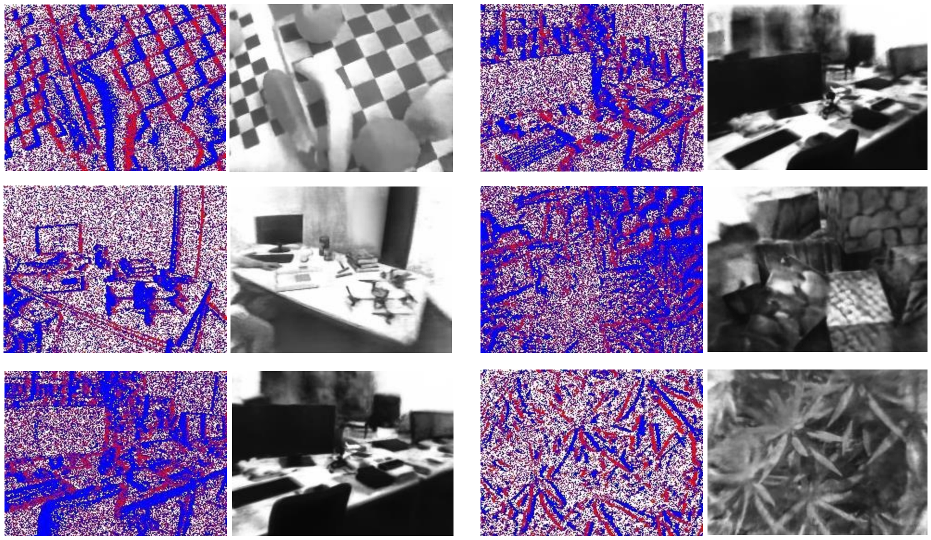


Figure 10. Noise reduction of events over various scenes with extreme noise ratio of 0.9.