

S2D-LFE: Sparse-to-Dense Light Field Event Generation

Supplementary Material

Yutong Liu Wenming Weng Yueyi Zhang Zhiwei Xiong*

University of Science and Technology of China

{ustcclt, wmweng}@mail.ustc.edu.cn {zhyuey, zwxiong}@ustc.edu.cn

Overview

This supplementary material is organized as follows:

Section 1 provides a more detailed explanation and visualization of the proposed synthetic and real-world dataset.

Section 2 provides comparisons of computational efficiency and more extensive ablation studies.

Section 3 provides additional qualitative experimental results.

1. Additional Dataset Details

1.1. Real-world LFE Dataset

We collect a real LFE dataset comprising 25 dynamic scenes using our LFE capture system, which is termed “LFE-real”. The dataset consists of two main categories: 15 indoor sequences captured by moving our camera array through various indoor environments and 10 autonomous driving sequences. Each sequence spans between 3 to 5 minutes. The detailed system specifications adopted to collect this dataset, including resolution, baseline measurements, and exposure settings, are provided in Table 1. Moreover, to provide a more intuitive visualization of our LFE-real dataset, we randomly select 9 sequences, convert them to RGB format [12], and display one frame from each sequence. The results are presented in Fig. 1. It can be observed that, whether in autonomous driving scenarios or indoor scenes, our recorded dataset achieves clear textures and rich content.

Table 1. Parameters of the LFE capture system.

Parameter	Specification
Cameras	DAVIS346 \times 4
Resolution	346 \times 260 pixels
Baseline	6 cm
Exposure time	100 μ s
Focal Length	7 mm, 14 mm

1.2. Synthetic LFE Dataset

To generate the synthetic LFE dataset, referred to as LFE-syn, we leveraged the Carla simulator [4] to create diverse sequences under controlled conditions. The dataset consists of 162 sequences, each spanning one minute, and is generated using 18 different maps provided by the simulator. Each map was divided into 9 regions, with a designated starting point in each region where virtual vehicles equipped with a 5 \times 5 event camera array were deployed. The baseline lengths between adjacent cameras were randomly set between 3 cm and 9 cm to simulate varying spatial configurations, while each camera was configured with a resolution of 346 \times 260 pixels and a field of view of 90°. Detailed parameter configurations are summarized in Table 2. The scenes in LFE-syn primarily consist of autonomous driving road scenarios, featuring pedestrians, vehicles, vegetation, various traffic signs, and buildings from the simulation maps. To highlight the texture details of LFE-syn, we randomly selected 9 sequences, converted them to RGB format [12], and extracted one representative frame from each. The results are presented in Fig. 2. The dataset exhibits clear textures and rich content across diverse environments.

*Corresponding author

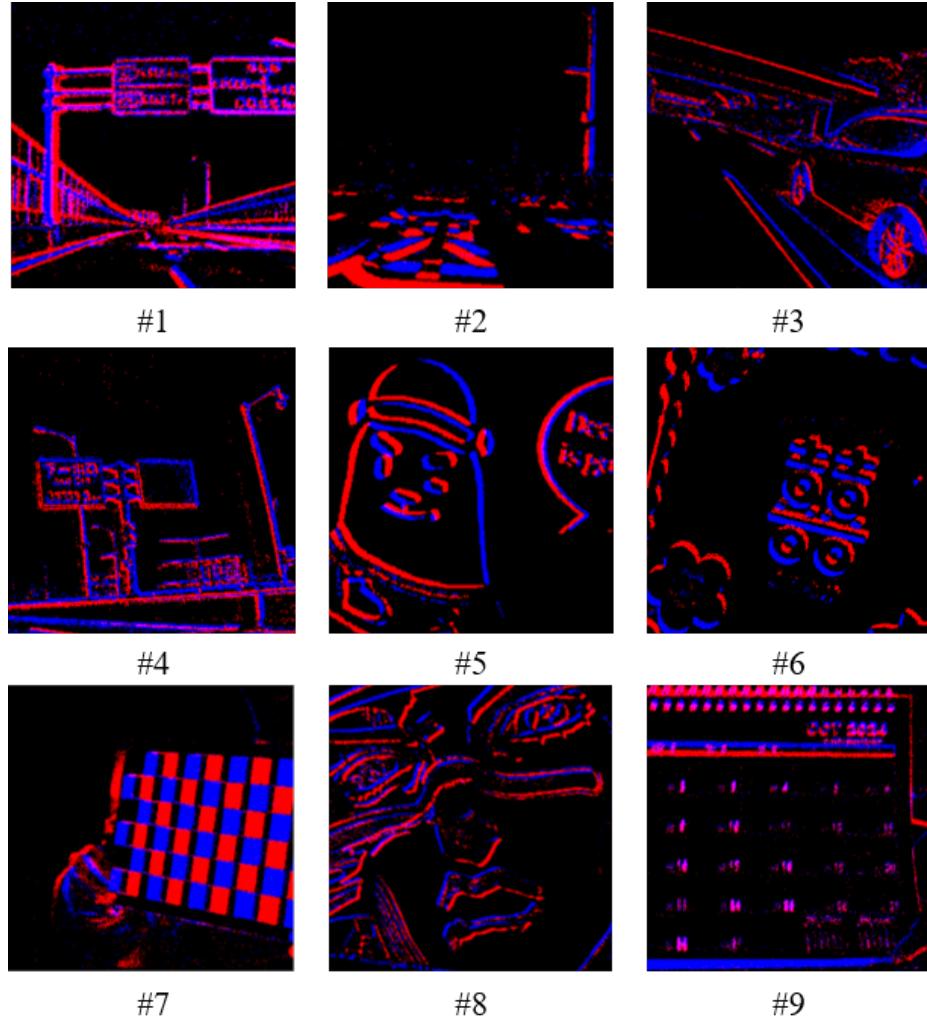


Figure 1. Visualization of sample frames from our real LFE dataset, “LFE-real.” Frames #1–#4 represent autonomous driving scenes, while frames #5–#9 correspond to indoor scenes. The dataset exhibits clear textures and rich content across diverse environments.

Table 2. Parameters of the synthetic LFE dataset LFE-syn.

Parameter	Specification
Simulator	Carla [4]
Number of maps	18
Sequences per map	9
Camera array	5×5 event cameras
Resolution	346×260 pixels $\times 25$
Baseline length	Randomly set between 3 cm and 9 cm
Event trigger threshold (positive)	0.1 to 0.15
Event trigger threshold (negative)	-0.15 to -0.1
Field of view	90°
Sequence length	1 minute per sequence
Total synthetic sequences	162

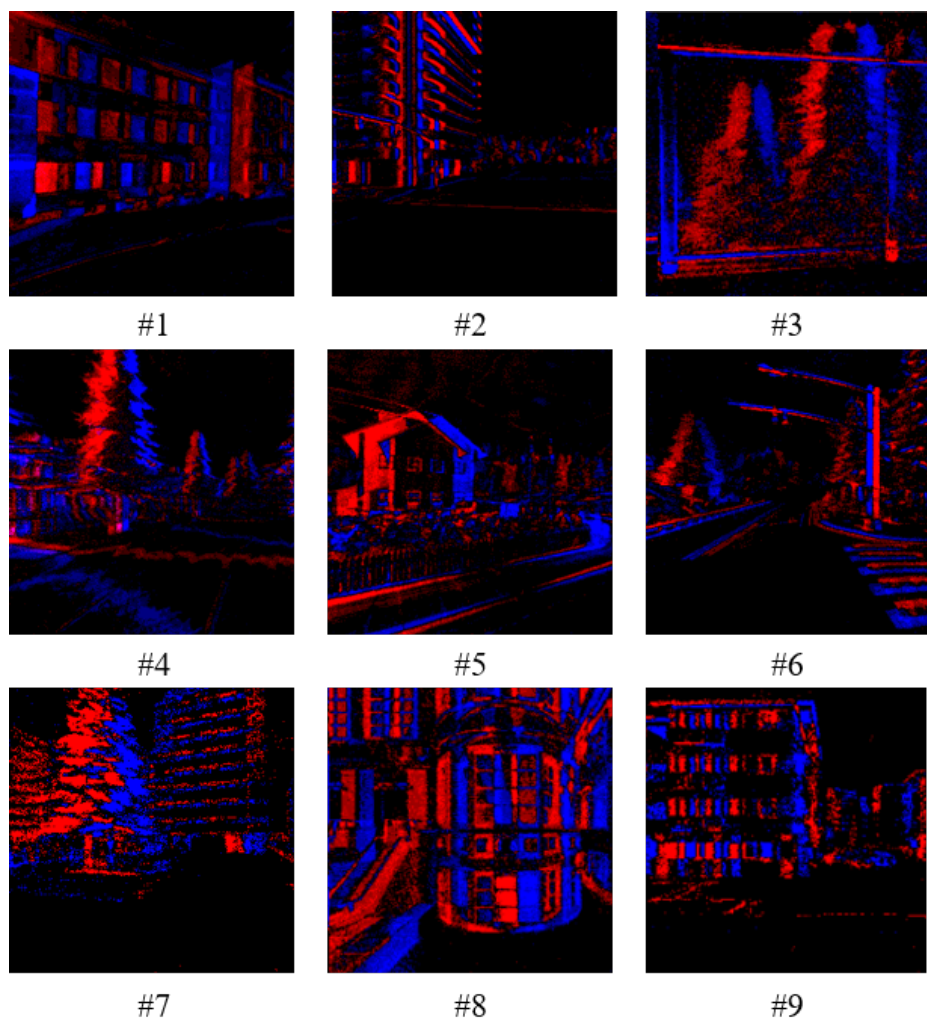


Figure 2. Visualization of sample frames from our LFE synthetic dataset. The dataset exhibits clear textures and rich content across diverse environments.

2. Additional Ablations

2.1. Efficiency Comparisons

Below we provide the model size (M) and execution time (sec) for generating $5 \times 5 \times 256 \times 256$ LFEs with one NVIDIA A100 GPU. As can be seen, although S2D-LFE has a larger model size due to its VAE+diffusion architecture, it shows superior efficiency compared to other baselines (beyond distinct performance advantage), which validates its feasibility in real-world scenarios.

Table 3. Efficiency comparisons of Guo *et al.* [6], R2L [9], and our S2D-LFE. Below we provide the model size (M) and execution time (sec) for generating $5 \times 5 \times 256 \times 256$ LFEs.

Method	Parameters	Time
Guo <i>et al.</i> [6]	7.45 M	3.622 sec
R2L [9]	23.77 M	3.171 sec
S2D-LFE	130.39 M	1.484 sec

2.2. Ablation of ContBlock and GeoBlock

To further validate the effectiveness of our LFE-adapter, we construct two variants by individually removing the ContBlock and GeoBlock, replacing each with ResBlocks. As shown in Table 4, removing either block leads to perceptible performance drops across PSNR, SSIM, and LPIPS. This finding indicates that both ContBlock and GeoBlock play a pivotal role in retaining high reconstruction accuracy, structural fidelity, and perceptual quality. The comparisons confirm that each component is crucial for sustaining the overall performance of S2D-LFE.

Table 4. Ablation study on S2D-LFE by individually removing the ContBlock and GeoBlock. The table compares the performance under a setting of generating 25 views in the synthetic dataset. The best results are highlighted in bold. ‘↑’: the higher the better performance, ‘↓’: the opposite.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
w/o ContBlock	23.55	0.691	0.254
w/o GeoBlock	23.61	0.690	0.262
LFE-adapter	24.06	0.701	0.239

3. Additional Experimental Results

3.1. In-training-scale Comparisons

In Fig. 3 and Fig. 4, we present an extended qualitative evaluation on four additional scenes (two synthetic and two real-world) to further demonstrate the advantages of our proposed S2D-LFE (labeled “Ours”) over existing approaches. In the synthetic scenes, it can be observed that our method surpasses competing approaches in terms of detail preservation and texture fidelity. Notably, both DistgASR [10] and SAV [3] suffer from misalignment and artifacts, while ET-Net [11]+Guo *et al.* [6]+Vid2E [5] exhibits localized aliasing. Moreover, R2L [9] occasionally loses fine-grained details, causing a pronounced degradation in structural information. In contrast, our method more faithfully recovers the target views, effectively preserving scene geometry and yielding visually superior results. Similarly, for real-world scenes, our S2D-LFE also delivers higher-quality texture details and more accurate view alignment compared to the baselines.

In our manuscript, we only employ one learnable 3D representation method (R2L [9], a NeRF-based approach specialized for LFs) as baseline. Recently, with the rise of learnable 3D representations, several methods [1, 2, 7, 8] have demonstrated the ability to synthesize target views from only a small set of input view. However, these methods encounter several issues when extended to the LFE modality. Specifically, (1) the inherent sparsity of LFEs increases the difficulty of modeling coherent spatial relationships; (2) the extremely limited number of views makes it challenging for the model to learn a continuous 3D representation; and (3) methods that do not target LFs cannot exploit the inherent prior knowledge of view positions in a LF setup. To substantiate these conclusions, we additionally compare our approach with MVSplat [2], a state-of-the-art novel view synthesis method based on 3D Gaussian Splatting. The comparison results are listed in Table 5 below. It can be observed that our approach outperforms MVSplat on multiple qualitative metrics, thereby providing evidence in support of the above conclusion.

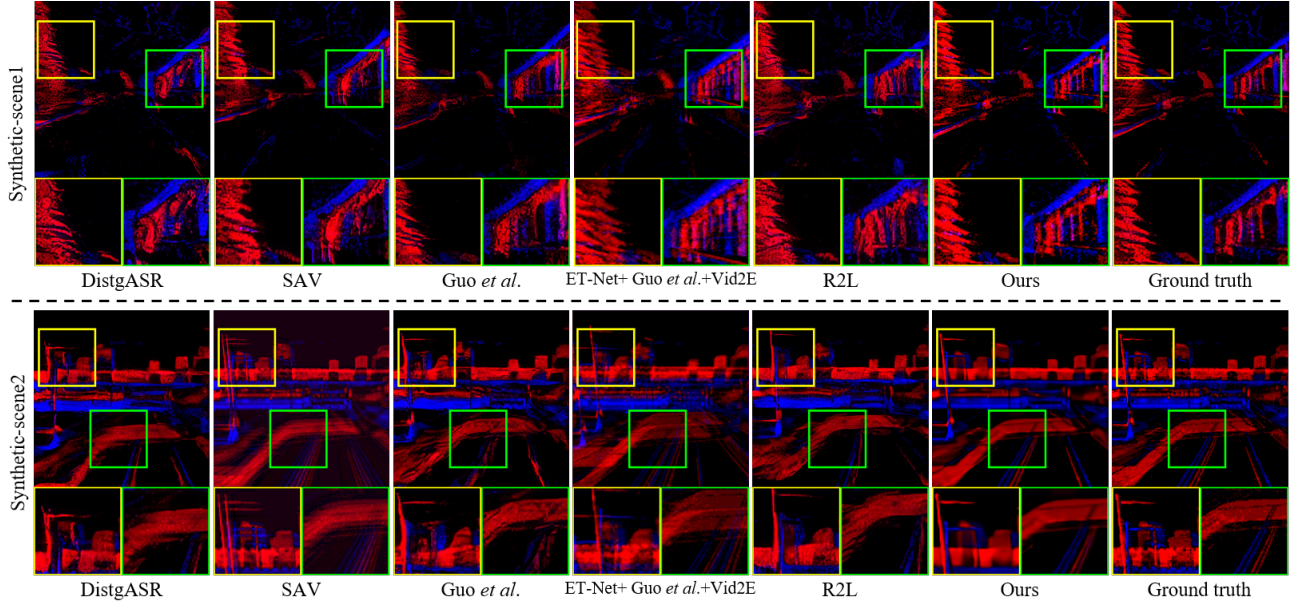


Figure 3. The figure presents a qualitative comparison on the central view of the generated LFE, using synthetic testset. The proposed S2D-LFE method (labeled as "Ours") is compared against other existing techniques, including SAV [3], Guo et al. [6], ET-Net [11] + Guo et al. + Vid2E [5], R2L [9], DistgASR [10], and the ground-truth. The highlighted regions (yellow and green boxes) magnify specific areas to emphasize the differences in event generation quality.

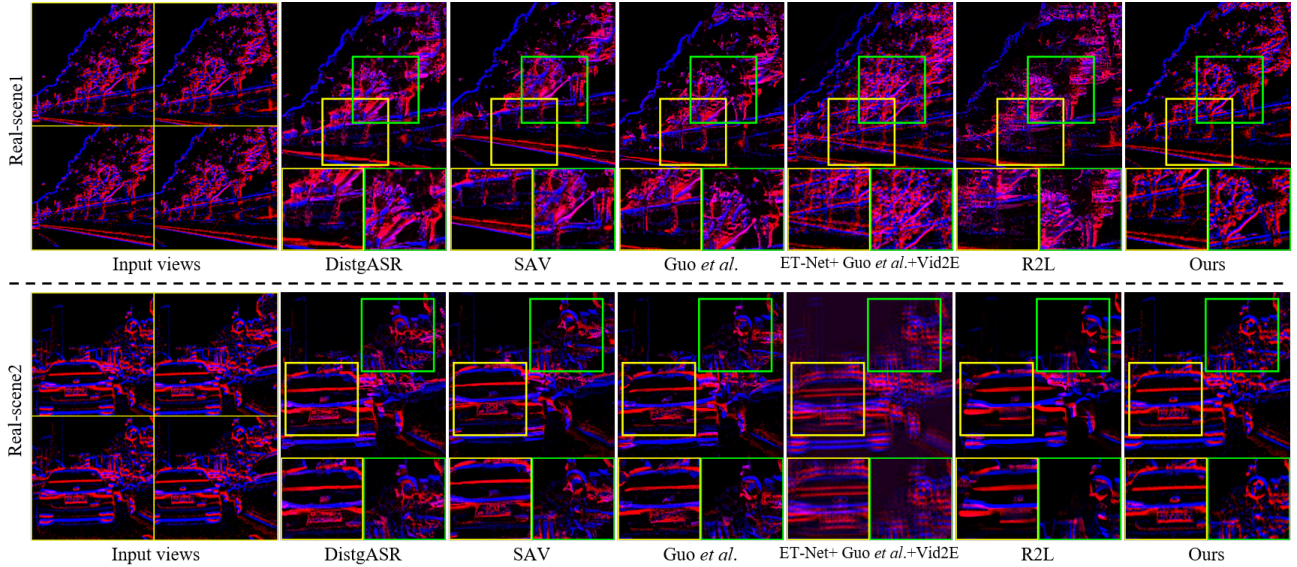


Figure 4. The figure presents a qualitative comparison on the central view of the generated LFE, using real-world testset. The proposed S2D-LFE method (labeled as "Ours") is compared against other existing techniques, including SAV [3], Guo et al. [6], ET-Net [11] + Guo et al. + Vid2E [5], R2L [9], DistgASR [10], and the ground-truth. The highlighted regions (yellow and green boxes) magnify specific areas to emphasize the differences in event generation quality.

Moreover, in Fig. 5, we present two additional experimental comparisons (one synthetic scene and one real-world scene) to validate the superiority of our method in maintaining angular consistency under the in-training-scale setting. To provide a clearer view of alignment across different views, we highlight the vertical epipolar with yellow dashed lines. Under ideal conditions, the region delineated by the vertical epipolar in each viewpoint should exhibit only vertical disparity. It

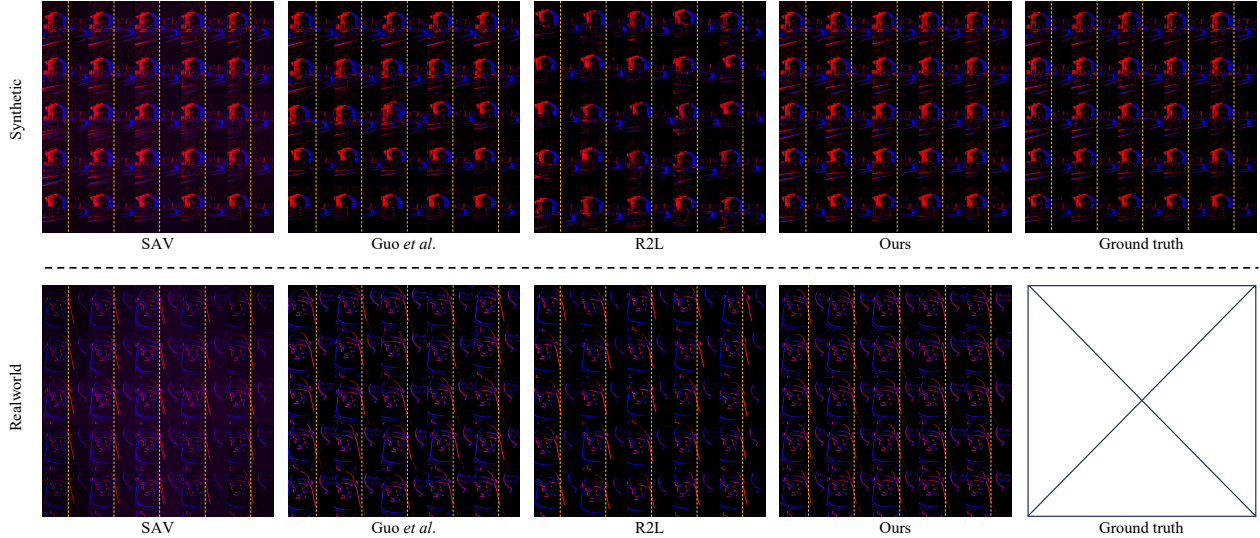


Figure 5. Qualitative comparison of angular consistency under the in-training-scale setting for both synthetic (top row) and real-world (bottom row) test datasets. To provide a clearer view of alignment across different views, we highlight the vertical epipolar with yellow dashed lines.

Table 5. Quantitative evaluation of MVSplat and our S2D-LFE. The table compares the performance under a setting of generating 25 views in the synthetic dataset. The best results are highlighted in bold. ‘↑’: the higher the better performance, ‘↓’: the opposite.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
MVSplat	19.54	0.558	0.442
S2D-LFE (Ours)	24.06	0.701	0.239

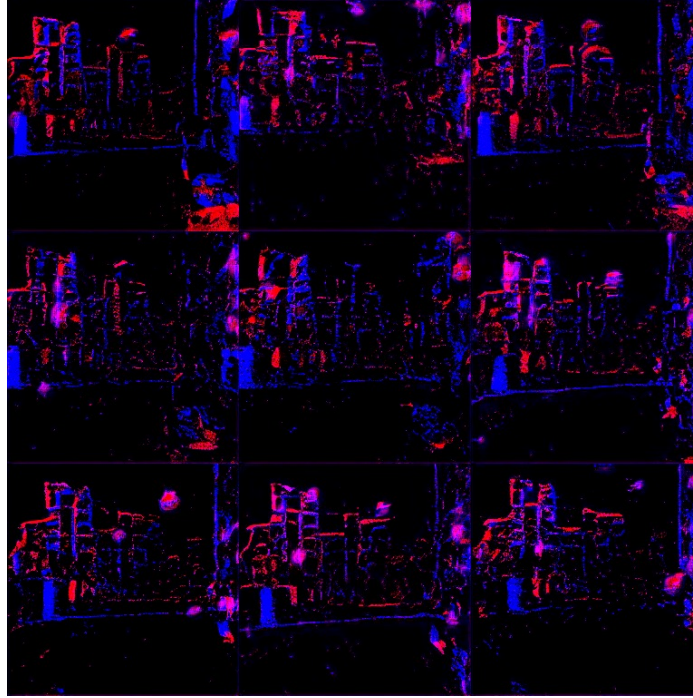
can be observed from Fig. 5 that, in the synthetic scenes, Guo et al. [6] and R2L [9] exhibit noticeable misalignments in reconstructing houses across different views, where horizontal viewpoints produce vertical disparities. Meanwhile, SAV [3] suffers from view aliasing. In contrast, our method provides consistent reconstructions with accurate inter-view alignment, closely matching the ground truth. Similarly, in the real-world scenes, S2D-LFE effectively preserves angular consistency while recovering the faithful details. These results demonstrate that S2D-LFE exhibits an advantage in maintaining angular coherency in reconstruction results.

3.2. Out-of-training-scale Comparisons

We conducted an additional experiment to evaluate the performance of our method under varying numbers of generated views, demonstrating its advantages in out-of-training-scale settings. Specifically, we selected a scene and compared our method with Guo et al. [6], using 2×2 input views to generate 3×3 , 5×5 , 7×7 , and 9×9 views. It can be observed from Fig. 6, Fig. 7, Fig. 8 and Fig. 9 that our method consistently maintains better angular consistency and reconstruction detail across all four settings. In contrast, Guo et al. exhibits degradation in both detail preservation and view alignment as the number of generated views increases. These findings demonstrate the superior generalization capability of our method, particularly in scenarios requiring the generation of dense views beyond the training configuration.

Generating
9 views

Guo et al.



Ours

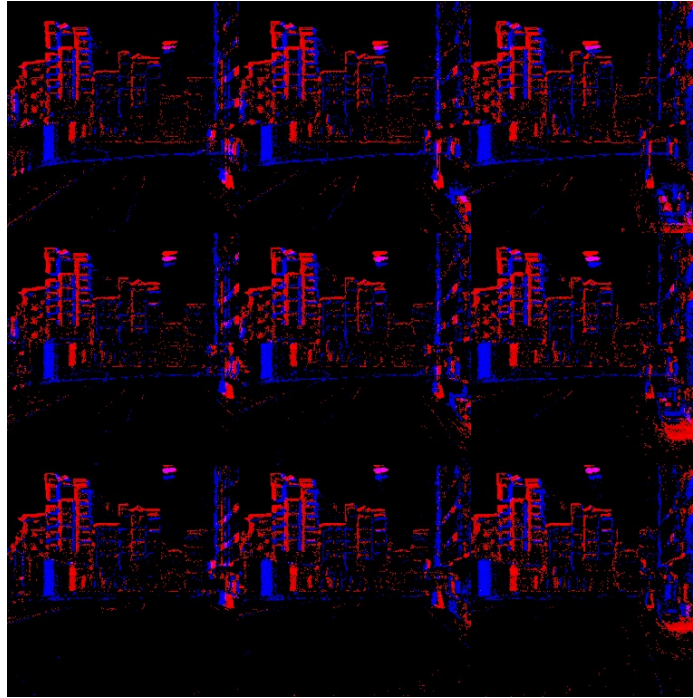
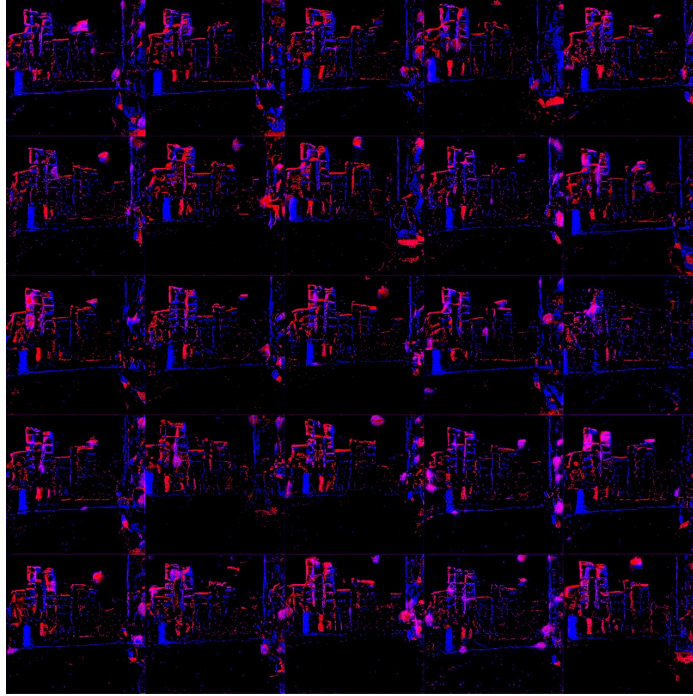


Figure 6. Qualitative comparison of generated LFEs under out-of-training-scale settings. Starting from 2×2 input views, 3×3 views are generated using *Guo et al.* [6] and the proposed method (“Ours”). Our method demonstrates superior angular consistency and detail preservation across all settings.

Generating 25 views

Guo et al.



Ours

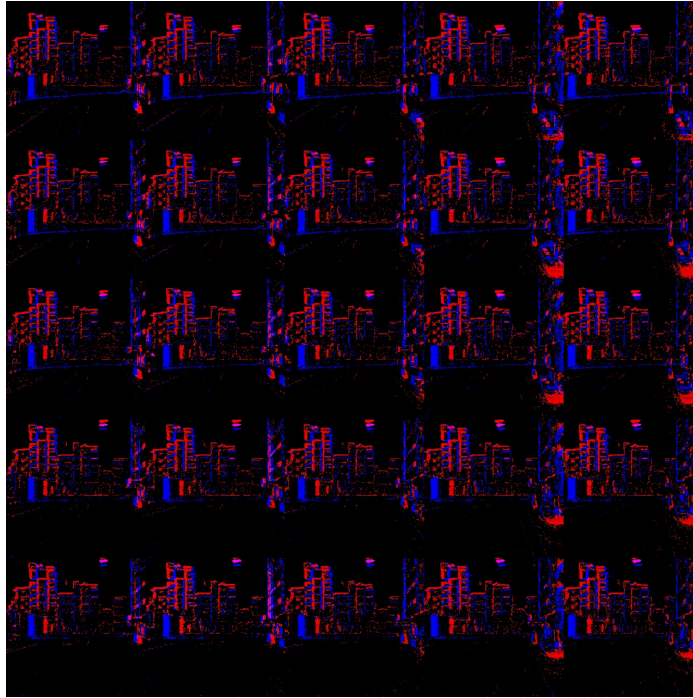
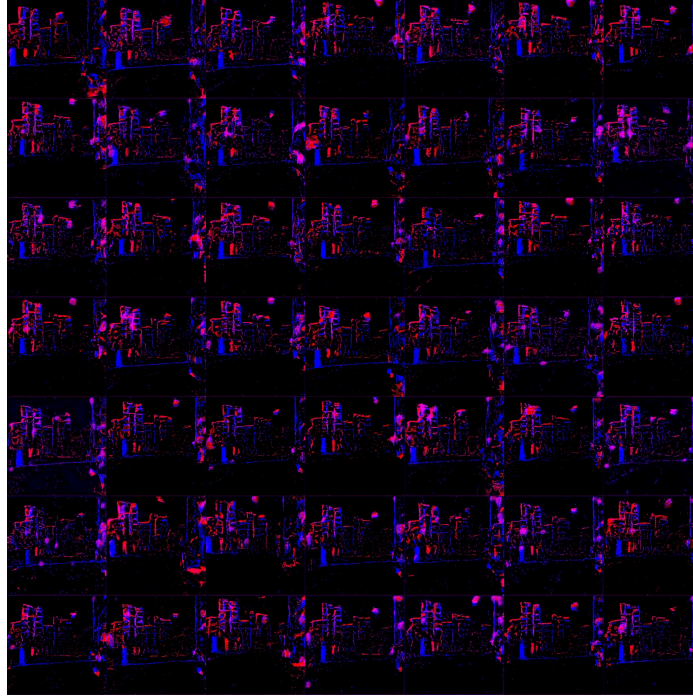


Figure 7. Qualitative comparison of generated LFEs under out-of-training-scale settings. Starting from 2×2 input views, 5×5 views are generated using *Guo et al.* [6] and the proposed method (“Ours”). Our method demonstrates superior angular consistency and detail preservation across all settings.

Generating 49 views

Guo et al.



Ours

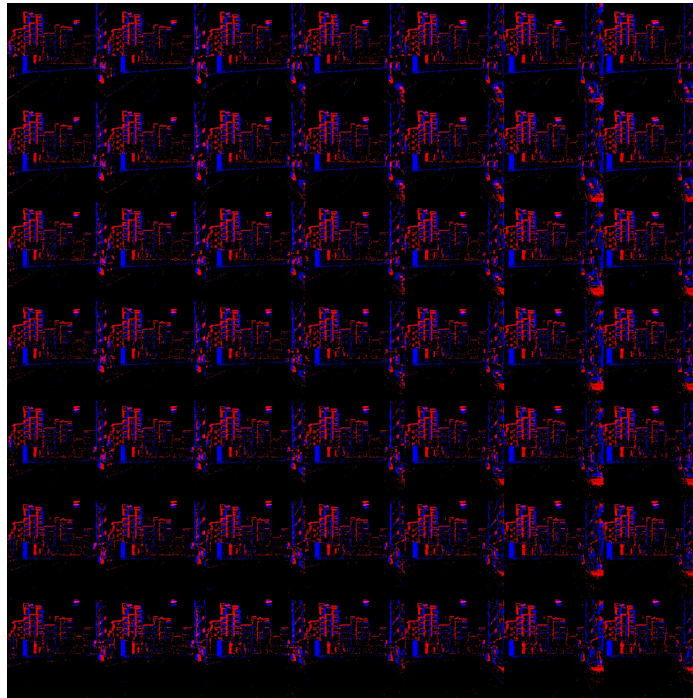
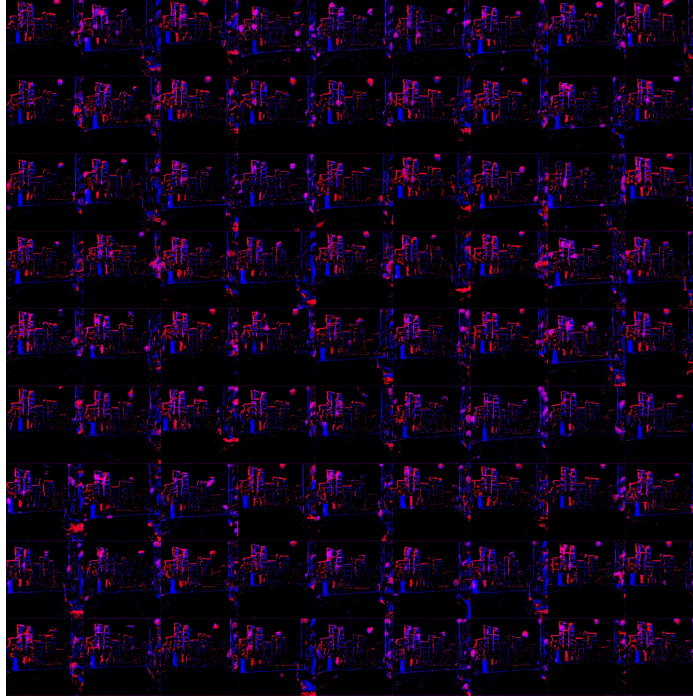


Figure 8. Qualitative comparison of generated LFEs under out-of-training-scale settings. Starting from 2×2 input views, 7×7 views are generated using *Guo et al.* [6] and the proposed method (“Ours”). Our method demonstrates superior angular consistency and detail preservation across all settings.

Generating 81 views

Guo et al.



Ours

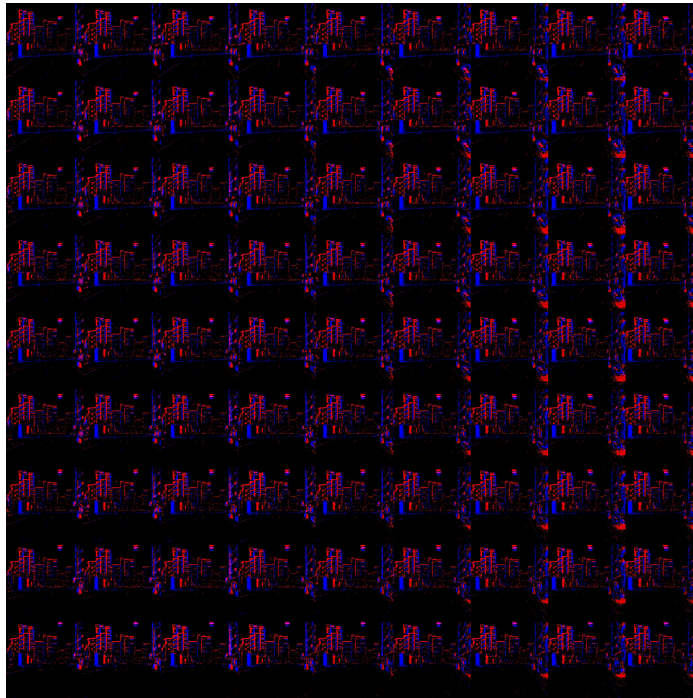


Figure 9. Qualitative comparison of generated LFEs under out-of-training-scale settings. Starting from 2×2 input views, 9×9 views are generated using *Guo et al.* [6] and the proposed method (“Ours”). Our method demonstrates superior angular consistency and detail preservation across all settings.

References

- [1] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 4
- [2] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 4
- [3] Zhen Cheng, Yutong Liu, and Zhiwei Xiong. Spatial-angular versatile convolution for light field reconstruction. *IEEE Transactions on Computational Imaging*, 8:1131–1144, 2022. 4, 5, 6
- [4] A Dosovitskiy, G Ros, F Codevilla, et al. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 1, 2
- [5] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 4, 5
- [6] Mantang Guo, Junhui Hou, Jing Jin, Hui Liu, Huanqiang Zeng, and Jiwen Lu. Content-aware warping for view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9486–9503, 2023. 4, 5, 6, 7, 8, 9, 10
- [7] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5480–5490, 2022. 4
- [8] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023. 4
- [9] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, pages 612–629. Springer, 2022. 4, 5, 6
- [10] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 4, 5
- [11] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 4, 5
- [12] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 1