STD-GS: Exploring Frame-Event Interaction for SpatioTemporal-Disentangled Gaussian Splatting to Reconstruct High-Dynamic Scene
Supplementary Material

Hanyu Zhou[1, 2]*, Haonan Wang[1]*, Haoyue Liu[1], Yuxing Duan[1], Luxin Yan[1]†, Gim Hee Lee[2]†
[1] National Key Lab of Multispectral Information Intelligent Processing Technology
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2] School of Computing, National University of Singapore
{hy.zhou, gimhee.lee}@nus.edu.sg yanluxin@hust.edu.cn

In this supplementary material, we provide the detailed description of building coaxial frame-event dataset in Sec. 1. Then, we further present the appearance-motion clustering approach in Sec. 2. Next, we also provide several analysis experiments about the proposed method in Sec. 3, including continuous tracking of dynamic objects in Sec. 3.1, deblur preprocessing in Sec. 3.2, Gaussian fusion strategy in Sec. 3.3, analysis of clustering loss in Sec. 3.4, weight sensitivity in Sec. 3.5, efficiency of scene reconstruction in Sec. 3.6, and description of limitation in Sec. 3.7. Finally, we provide the qualitative comparison results on various datasets from Sec. 4.1 to Sec. 4.4.

1. Coaxial Frame-Event Dataset

The prerequisite for utilizing frame images and event stream to reconstruct dynamic scenes via Gaussian splatting is to obtain the pixel-aligned frame and event data with gyroscope. To this end, we build an optical coaxial frame-event device with inertial measurement unit (IMU) in Fig. 1, and collect the paired frame-event data via two steps, including time synchronization and spatial calibration.

Regarding the issue of time synchronization, we utilize microcontroller to generate two pulses with different frequencies but same timestamp as external trigger to synchronize the time between frame and event cameras, including 30 Hz for frame camera and 1MHz for event camera. Note that the external pulse with the frequency of 30 Hz is also used to set the timestamp of the gyroscope from IMU.

Regarding the issue of spatial calibration, we divide this step into two sub-steps, *i.e.*, frame-event spatial calibration and frame-IMU spatial calibration. As for frame-event spatial calibration, we first set up a physically coaxial optical device with a beam splitter for frame and event cameras,

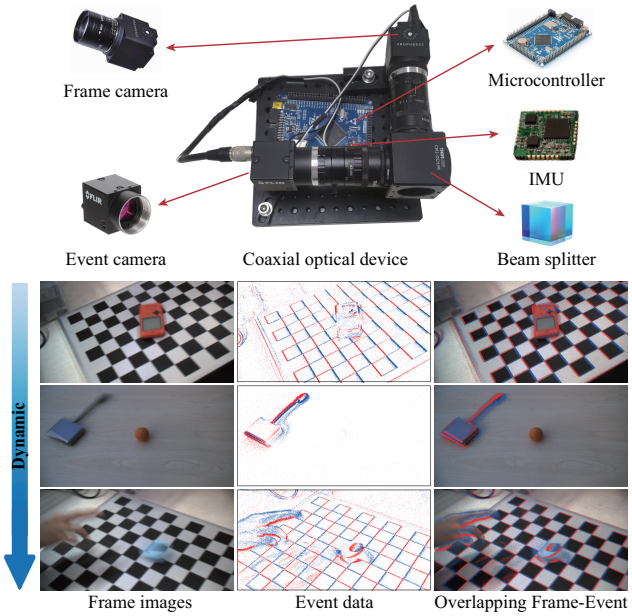


Figure 1. Collection device and examples of the proposed dataset CoFED. As for collection device, we build an optical coaxial frame-event device with inertial measurement unit (IMU), where we use microcontroller to achieve time synchronization between different sensors. As for dataset, the proposed dataset CoFED includes different dynamic scenes with pixel-aligned frame-event data.

which allows the same light to pass through the same lens and enter different cameras, thus achieving the overall field of view alignment. Next, we further perform a standard stereo rectification between frame data and event data, and then fine tune the slight calibration errors via pixel offset [1]. As for frame-IMU spatial calibration, given a period, we integrate IMU data to obtain the odometry, and perform a typical SLAM algorithm [2] to estimate the frame-based visual odometry. Then, we construct the consistency constraint between the two odometries to optimize the extrinsic parame-

*These authors contributed equally.
†Corresponding authors.

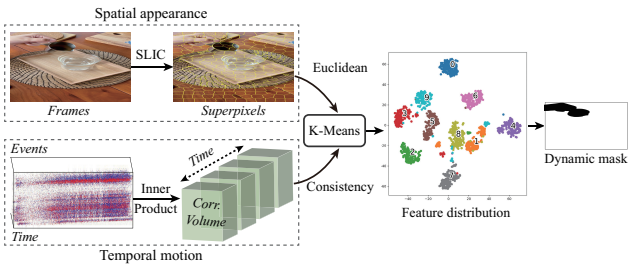


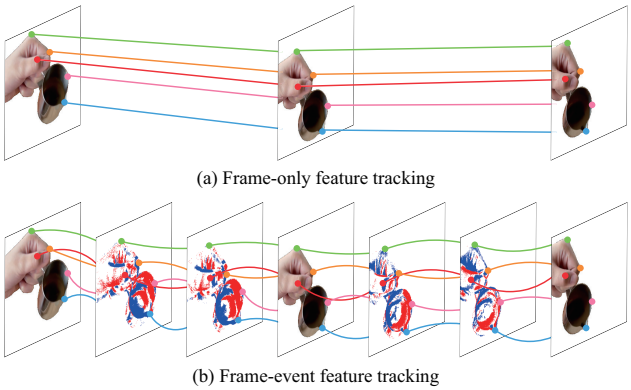
Figure 2. Illustration of appearance-motion clustering. We first segment frames into superpixels as spatial appearance features, and perform inner product on events to obtain correlation volumes as temporal motion features. We then use K-Means to cluster these spatiotemporal features to estimate the mask of the dynamic object.

ters between frame and IMU, which can be used to transform the gyroscope sensor of IMU to the coordinate system of frame camera. In this way, we can obtain the spatiotemporal pixel-aligned frame images, event stream, and gyroscope data. Furthermore, we utilize the coaxial optical device to collect the pixel-aligned frame-event dataset with gyroscope, which covers real complex scenes with various dynamic patterns, *e.g.*, low-dynamic motion and high-dynamic motion.

Regarding the issue of collected data, CoFED is a new dynamic scene reconstruction dataset with pixel-aligned frame and event data. This dataset consists of 55 sequences from various dynamic scenes, where each sequence contains about 200 images. As for the diversity of dynamic patterns, we adjust frame rate and exposure time to capture the dynamic objects with different speeds, thus obtaining low-dynamic and high-dynamic patterns.

2. Description of Appearance-Motion Cluster

In Fig. 2, we illustrate the detailed process of the appearance-motion clustering in dynamic scene disentanglement module. Within a short period, given frame images and corresponding event stream as input, we use SLIC algorithm [3] to segment frame images into multiple superpixels as spatial features containing appearance information, and perform inner product on event stream to obtain multiple correlation volumes as temporal features containing motion information. Then, we measure the global similarity of spatial features via Euclidean distance, and the local similarity of temporal features via the motion consistency prior. Next, we introduce K-Means [4] to cluster these spatiotemporal features of dynamic scenes into different feature sets with specific labels. Finally, we enforce softmax on these feature sets, and integrate these labels to form masks of static background and dynamic objects. Therefore, the proposed appearance-motion clustering approach could effectively distinguish the spatiotemporal features of dynamic scenes, which are disentangled into dynamic objects and static background.



(a) Frame-only feature tracking

(b) Frame-event feature tracking

Figure 3. Visualization of continuous tracking. Dynamic object shows discontinuous linear motion between frame images, while event data enhances the continuous non-linear motion of the object.

Deblurring preprocessing	PSNR ↑	SSIM ↑	LPIPS ↓
Ours w/o deblurring	26.74	0.84	0.30
Ours w/ frame-only DeblurGAN	27.02	0.85	0.28
Ours w/ frame-event GEM	**27.68**	**0.87**	**0.26**

Table 1. Discussion on various deblurring preprocessing strategies.

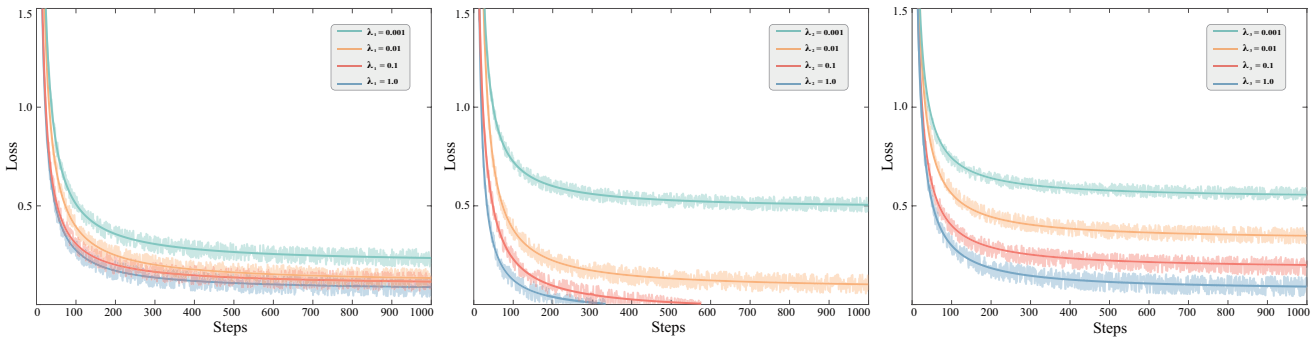
3. Discussion

3.1. Visualization of Continuous Tracking

In Fig. 3, we visualize the continuity of the track of dynamic object based on frame images and event data. Assuming that the dynamic object has a non-linear motion pattern, the object shows discontinuous linear motion in the frame image in Fig. 3 (a), while the event data can enhance the continuous non-linear motion process of the object in Fig. 3 (b). Therefore, the motion continuity at the imaging level is conducive to the continuous tracking of the dynamic object, thereby improving the accuracy of the initial position of the subsequent object Gaussian representation.

3.2. Impact of Various Deblurring Preprocessing

In the dynamic scene disentanglement module of the proposed framework, deblurring serves as a preprocessing strategy and plays a role in dynamic scene reconstruction. To validate the effectiveness of deblurring as a preprocessing strategy, we analyze the impacts of unimodal frame-only deblurring (*e.g.*, DeblurGAN [5]) and multimodal frame-event deblurring (*e.g.*, GEM [6]) methods on dynamic scene reconstruction in Table 1. First, deblurring preprocessing can indeed improve scene reconstruction performance to some extent. The main reason is that high-dynamic objects can bring in potential motion blur in frame imaging, which affects the matching and representation of spatiotemporal features. Second, multimodal frame-event deblurring has a significantly better impact on dynamic scene reconstruction



(a) Weight sensitivity of $\mathcal{L}_{recon}$ (b) Weight sensitivity of $\mathcal{L}_{clu}$ (c) Weight sensitivity of $\mathcal{L}_{consis}$

Figure 4. The weight sensitivity of main model losses.

Fusion solution	PSNR ↑	SSIM ↑	LPIPS ↓
Ours w/ additive	27.32	0.86	0.26
Ours w/ shadow-based weight	**27.68**	**0.87**	0.26

Table 2. Discussion on object Gaussian fusion solutions.

Metric	Object DBI [7] ↓		Background DBI [7] ↓		PSNR ↑	SSIM ↑
	Spatial	Temporal	Spatial	Temporal		
w/o $\mathcal{L}_{clu}$	0.57	0.48	0.39	0.73	26.03	0.80
w/ $\mathcal{L}_{clu}$	0.24	0.15	0.12	0.27	27.68	0.87

Table 3. Impact of $\mathcal{L}_{clu}$ on clustering process and performance.

compared to unimodal frame-only deblurring. This is because frame images with low frame rate struggle to capture the motion trends of high-dynamic objects, while event data can capture continuous motion, facilitating the restoration of motion blur for scene representation. Therefore, multimodal frame-event deblurring preprocessing strategy can help improve dynamic scene reconstruction.

3.3. Choice of Fusion Solution

In Table 2, we compare the influences of different Gaussian fusion solutions (*e.g.*, additive fusion and shadow-based weighting fusion) on the performance of high-dynamic scene reconstruction. We can observe that shadow-based weighting fusion strategy significantly outperforms direct additive fusion. This is because shadows reflect the occlusion conditions of dynamic objects and static background in a novel view, and the corresponding weights are learned by a neural network to adaptively adjust the fusion between object Gaussians and background Gaussians.

3.4. Analysis of Clustering Loss

In Table 3,We analyze the impact of L_{clu} on the clustering process and performance of Clusterformer, where we use Davies-Bouldin Index [7] (DBI,↓) as the quantitative indicator to evaluate the clustering performance. We can observe that L_{clu} loss can effectively improve the clustering ability of Clusterformer and promote overall performance.

Method	Input	PSNR ↑	SSIM ↑	Time ↓	FPS ↑	Storage ↓
E2NeRF [8]	F+E	19.42	0.60	2 days	0.04	112 MB
E2GS [9]	F+E	20.01	0.62	50 min	140.00	28 MB
TiNeuVox [10]	F	18.11	0.54	28 min	1.50	48 MB
3DGS [11]	F	17.08	0.54	10 min	170.00	10 MB
4DGS [12]	F	18.29	0.55	8 min	82.00	18 MB
Ours	F+E	23.57	0.69	15 min	70.00	24 MB

Table 4. Efficiency of scene reconstruction on Event-HyperNeRF.

3.5. Weight Sensitivity of Model Losses

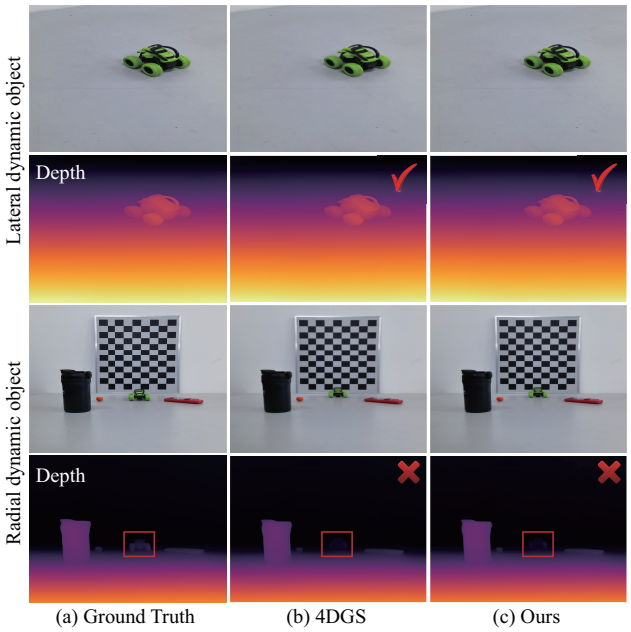
To choose the optimal weight parameters, we conduct the study on the weight sensitivity of the typical losses in Fig. 4, such as $\mathcal{L}_{recon}$, $\mathcal{L}_{clu}$ and $\mathcal{L}_{consis}$. In Fig. 4 (a), the reconstruction loss $\mathcal{L}_{recon}$ is robust to the training of the whole framework. In Fig. 4 (b), the clustering consistency loss $\mathcal{L}_{clu}$ is sensitive to the training process. As the weights increase, the training curve collapses. In Fig. 4 (c), the larger the weight of the spatiotemporal consistency loss $\mathcal{L}_{consis}$, the more rapidly the training framework converges. Therefore, we set the weights as $[\lambda_1, \lambda_2, \lambda_3]$=[1.0, 0.01, 1.0].

3.6. Efficiency of Scene Reconstruction

In Table 4, we compare the scene reconstruction efficiency of different methods on the Event-HyperNeRF dataset, including PSNR, SSIM, training time, rendering speed, and model size. First, overall, the Gaussian splatting methods require less training time for modeling scenes, less inference time for rendering novel views, and smaller model size. Second, the proposed multimodal method does have slightly more computational resources than the unimodal methods (*e.g.*, 3DGS [11] and 4DGS [12]) but achieves better scene reconstruction performance, while the computa-

Dataset	HyperNeRF [13]				CED [14]				N3DV [15]				PanopticSports [16]			
	PSNR ↑	SSIM↑	Time↓	FPS↑	PSNR↑	SSIM↑	Time↓	FPS↑	PSNR ↑	SSIM↑	Time↓	FPS↑	PSNR↑	SSIM↑	Time↓	FPS↑
4DGS [12]	25.20	0.71	**1.0 h**	34	24.07	0.79	**0.4 h**	134	31.19	0.94	9.0 h	33	27.20	0.91	0.8 h	40
E-D3DGS [17]	25.53	0.70	1.3 h	**140**	24.68	0.81	1.1 h	**210**	31.27	0.94	**2.1 h**	**75**	26.97	0.90	**0.6 h**	**120**
Ours (S/J)	**26.35**	**0.76**	1.5/1.9 h	30	**28.36**	**0.90**	0.5/0.8 h	120	**32.42**	**0.95**	7.8/9.1 h	31	**30.34**	**0.93**	1.1/1.4 h	48

Table 5. Quantitative results on original public datasets. "S" is separate two-phase training, "J" is joint training.



(a) Ground Truth (b) 4DGS (c) Ours

Figure 5. Limitation of the proposed method. The proposed method performs well on reconstructing lateral dynamic object that moves to camera along the x, y-axis, but suffers challenges for the radial dynamic object that moves to camera along the z-axis.

tional resources are less than the multimodal method E2GS [9]. Therefore, the proposed method achieves both real-time and high-precision reconstruction of dynamic scenes.

3.7. Limitation

In Fig. 5, we further illustrate the reconstruction performance of the proposed method for dynamic objects with different motion patterns, including lateral and radial dynamic objects. Note that the former moves to the camera along the x, y-axis, while the latter moves to the camera along the z-axis. We can observe that the proposed method performs well on reconstructing the lateral dynamic object, while suffers challenges for the radial dynamic object. Specifically, the rendered radial dynamic object appears similar to image GT, but the corresponding depth is erroneous. This is because frame and event cameras can only capture x, y-axis motion patterns, while the radial dynamic object makes the two cameras ineffective for this z-axis special motion pattern, thus causing the spatiotemporal features of the object to resemble those of the background. In the future, we will introduce LiDAR to assist the frame camera in modeling the 3D motion field to distinguish the x, y, z-axis spatiotemporal features of dynamic scene.

4. Comparison Experiments

4.1. Comparison on Original Public Dataset

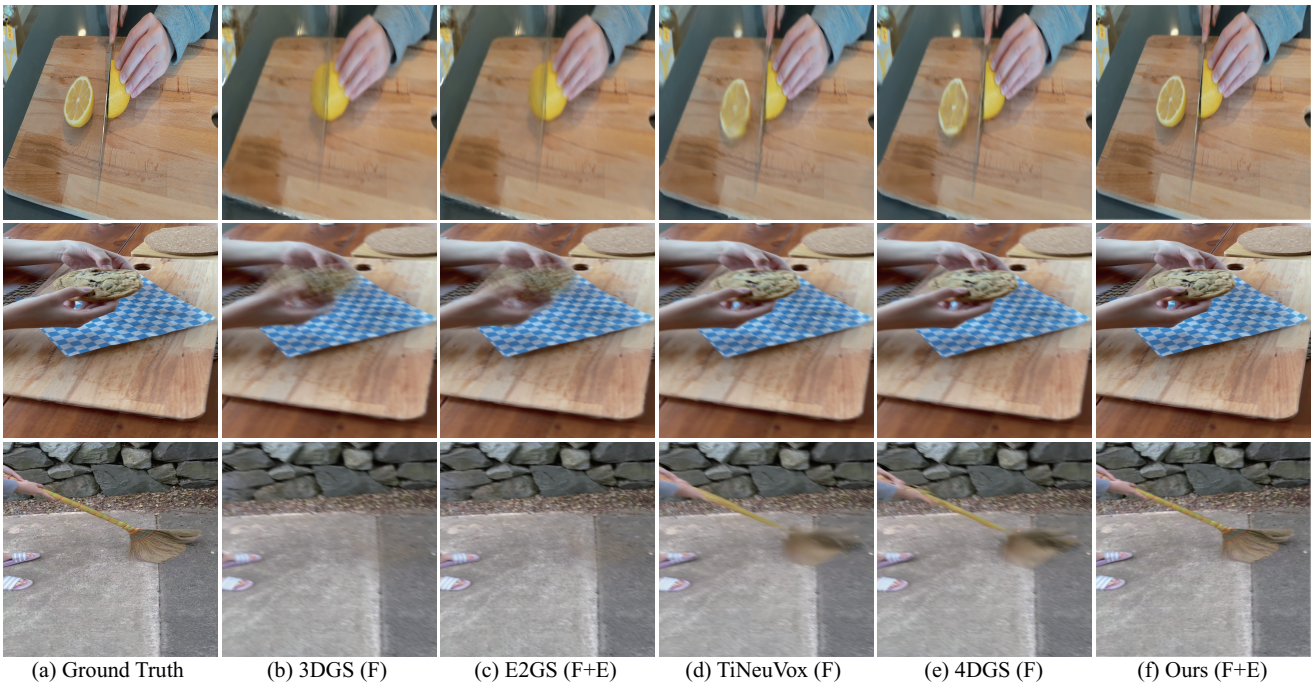
In Table 5,we compare the performance and efficiency metrics (*e.g.*, training time and FPS) of recent methods on general original public datasets (*e.g.*, HyperNeRF [13], CED [14], N3DV [15] and PanopticSports [16]). We can observe that the proposed method still performs better than the competing methods, but the efficiency metrics are not the best. The main reason is that the idea of spatiotemporal disentanglement is conducive to enhancing the spatiotemporal representation of Gaussian for dynamic scenes, regardless of low- or high-dynamic patterns. However, while the frame-event fusion process will consume some additional computing resources, it can ensure the reconstruction performance of the high-dynamic scene. In addition, we also found that the proposed method can be applied to different camera settings. Among the four datasets we used for comparative experiments, HyperNeRF and CED are monocular settings, while N3DV and PanopticSports are multi-view settings.

4.2. Comparison on Synthetic Event Dataset

In Fig. 6, we provide more visualization results of the competing methods on the synthetic event dataset Event-HyperNeRF. First, dynamic scene reconstruction methods significantly outperform static scene reconstruction methods. The main reason is that dynamic scene reconstruction methods can model the temporal information of the scene. Second, multimodal methods outperform unimodal methods, indicating that the complementary knowledge between frame and event modalities helps improve the spatiotemporal feature representation of dynamic scene.

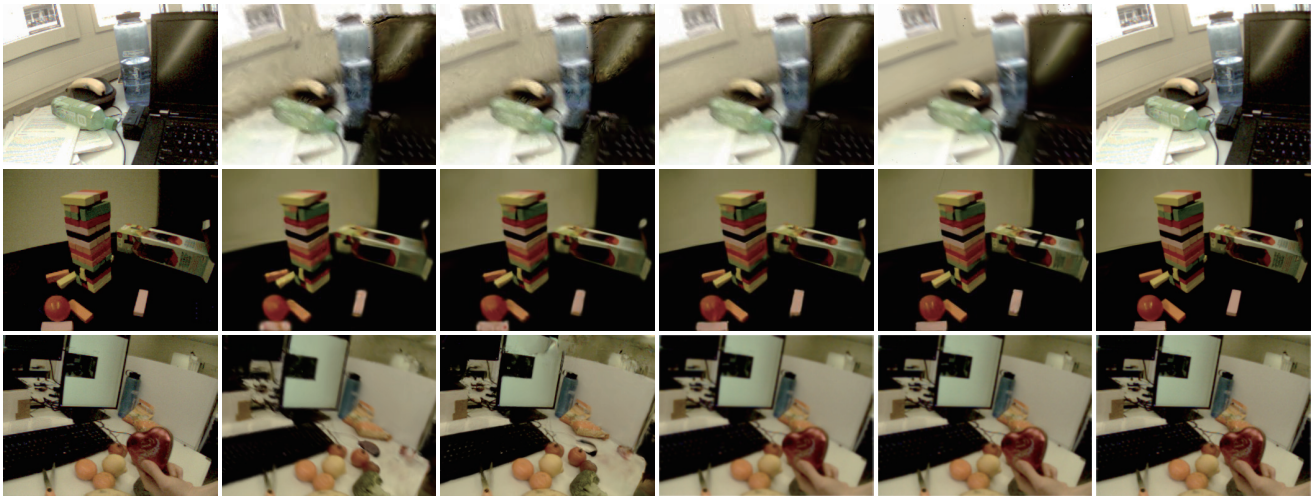
4.3. Comparison on Real Event Dataset

In Fig. 7, we provide additional qualitative results of competing methods on the real event dataset CED. For static scene reconstruction, the multimodal frame-event method significantly outperforms the unimodal frame-only method in rendering the static background. This is because event data helps improve the quality of frame images, thereby enhancing the reconstruction performance of the static parts of the scene. For dynamic scene reconstruction, the proposed multimodal method also significantly outperforms the



(a) Ground Truth (b) 3DGS (F) (c) E2GS (F+E) (d) TiNeuVox (F) (e) 4DGS (F) (f) Ours (F+E)

Figure 6. Visual comparison of novel view synthesis on synthetic Event-HyperNeRF dataset. "F" denotes frame and "E" denotes event.



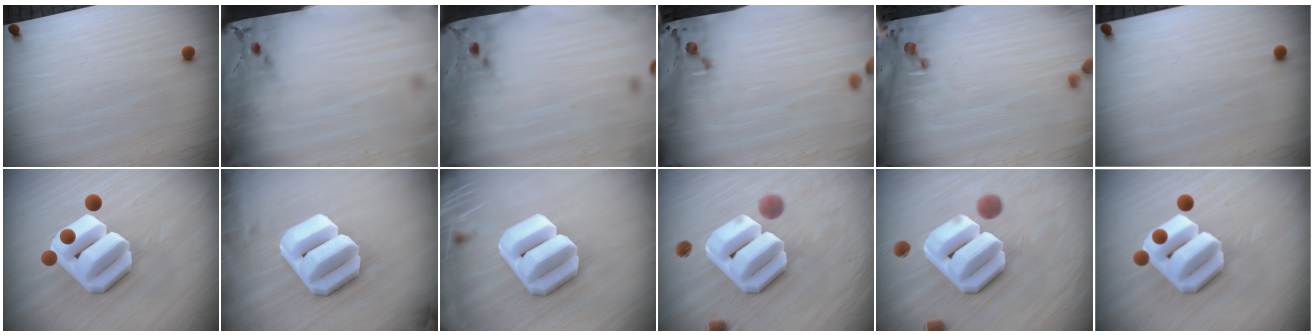
(a) Ground Truth (b) 3DGS (F) (c) E2GS (F+E) (d) TiNeuVox (F) (e) 4DGS (F) (f) Ours (F+E)

Figure 7. Visual comparison of novel view synthesis on real CED dataset. "F" denotes frame and "E" denotes event.

unimodal frame-only method. The main reason is that the proposed method leverages the advantages of both frame and event data in spatiotemporal representation to distinguish between high-dynamic objects and static background, thus facilitating high-dynamic scene reconstruction.

4.4. Comparison on Real High-Dynamic Scenes

We further compare the competing methods on real high-dynamic scenes from the proposed dataset CoFED. As shown in Fig. 8, we can observe that all other competing methods exhibit the positional shifts of high-dynamic objects, while the proposed method effectively render high-dynamic objects with accurate positions. This demonstrates that the proposed multimodal method with spatiotemporal disentanglement fully leverages the superior spatiotemporal knowledge of frames and events to reconstruct high-dynamic objects and static background.



(a) Ground Truth (b) 3DGS (F) (c) E2GS (F+E) (d) TiNeuVox (F) (e) 4DGS (F) (f) Ours (F+E)

Figure 8. Visual comparison of novel view synthesis on the proposed CoFED dataset. "F" denotes frame and "E" denotes event.

References

[1] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16155–16164, 2021. 1

[2] Wei Xu, Yixi Cai, Dongjiao He, Jiarong Lin, and Fu Zhang. Fast-lio2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics*, 38(4):2053–2073, 2022. 1

[3] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. 34(11):2274–2282, 2012. 2

[4] K Krishna and M Narasimha Murty. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(3):433–439, 1999. 2

[5] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8183–8192, 2018. 2

[6] Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Generalizing event-based motion deblurring in real-world scenarios. In *Int. Conf. Comput. Vis.*, pages 10734–10744, 2023. 2

[7] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979. 3

[8] Yunshan Qi, Lin Zhu, Yu Zhang, and Jia Li. E2nerf: Event enhanced neural radiance fields from blurry images. In *Int. Conf. Comput. Vis.*, pages 13254–13264, 2023. 3

[9] Hiroyuki Deguchi, Mana Masuda, Takuya Nakabayashi, and Hideo Saito. E2gs: Event enhanced gaussian splatting. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1676–1682. IEEE, 2024. 3, 4

[10] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3

[12] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20310–20320, 2024. 3, 4

[13] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 4

[14] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 0–0, 2019. 4

[15] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022. 4

[16] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024. 4

[17] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In *European Conference on Computer Vision*, pages 321–335. Springer, 2024. 4