

Instant Gaussian Stream: Fast and Generalizable Streaming of Dynamic Scene Reconstruction via Gaussian Splatting

Supplementary Material

A. Overview

With in the supplementary, we provide:

- Details of metrics calculation in Sec. B.
- Details of experiment settings in Sec. C.
- More ablation study in Sec. D.
- More limitations and future work in Sec. F.
- More Results and in Sec. G.
- Pseudocode and demo videos including:

- IGS_code.zip
- IGS-s_testview.mp4
- IGS-l_testview.mp4
- IGS_freeview.mp4

The complete code, pretrained weights, and the training dataset we constructed will be released as open-source after the review process is completed.

B. Details of the metrics calculation

As mentioned in Sec. 5 of the main paper, all metrics are averaged over the full 300-frame sequence, including frame 0, along with previous methods[2, 7]. Specifically:

Storage: The storage required for IGS includes the Gaussian primitives for frame 0 and each key frame, as well as the residuals for each candidate frame. Since each candidate frame is generated by applying motion from the previous key frame using AGM-Net, we only need to store the corresponding displacement (du) and rotation ($drot$), along with the mask of points with motion. We report the average storage requirements over the 300 frames.

Train: In line with previous methods[2, 7], we report the training time, which refers to the average time required to construct an Free-Viewpoint Video from a multi-view video sequence. This includes the time for constructing the Gaussian primitives for frame 0, generating candidate frames using AGM-Net, and refining the key frames. The total time is averaged over all 300 frames, which corresponds to our per-frame reconstruction time.

C. More implementation details

The reconstruction quality of Gaussian primitives for the first frame in each scenario is summarized in Tab. C1. For the N3DV scenes, we set the SH degree to 3, whereas for Meeting Room, it was set to 1 to mitigate overfitting caused by sparse viewpoints. During the Max Points Bounded Refinement process, all scenarios used the same learning rate settings. Specifically, the learning rate for position and rotation was set to ten times that in 3DGS, while the other

parameters were kept consistent with 3DGS.

The Max Points Number N_{\max} was determined based on the number of Gaussians in the initial frame of each scene. Specifically, N_{\max} was set to 150,000 for N3DV, 40,000 for the Meeting Room dataset.

Table C1. Reconstruction results of Gaussian models for the first frame in each scenario.

| Scene | PSNR \uparrow (dB) | Train \downarrow (s) | Storage \downarrow (MB) | Points Num |
|------------------|-------------------------|---------------------------|------------------------------|---------------|
| N3DV[3] | | | | |
| cur roasted beef | 33.96 | 287 | 36 | 149188 |
| sear steak | 34.03 | 287 | 35 | 143996 |
| Meeting room[2] | | | | |
| trimming | 30.36 | 540 | 3.9 | 37432 |
| vrheadset | 30.68 | 540 | 4 | 38610 |

D. More ablation study

In our experiments, we also explored incorporating additional modules into AGM-Net. However, the results showed that these modules did not achieve the expected improvements. The ablation studies are detailed as follows:

Attention-Based View Fusion: During the Projection-Aware Motion Feature Lift, we consider assigning different weights to features from different viewpoints instead of using the simple averaging method described in Eq.2 of the main paper. Specifically, for an anchor, the features obtained from each viewpoint are concatenated with the embedding of the corresponding viewpoint’s pose. These N_v features were then processed through self-attention, followed by a Softmax operation to compute the weights for aggregating the multi-view features. The experimental results, as shown in Tab. D2, indicate that adding this module doesn’t yield improvements on the test scenes of N3DV. This is likely because N3DV features forward-facing scenes, where differences between camera viewpoints are not significant. However, for 360° scenes, this module could be a promising direction for future work.

Occlusion-Aware Projection: We also attempted to account for occlusion effects during the Projection-Aware Motion Feature Lift by considering how anchor points might be obscured during projection. Specifically, we employ point rasterization[6], ensuring that each pixel corresponds to only one visible anchor point. The experimental results, shown in Tab. D2, reveal that this approach doesn’t

improve performance. Since we project anchor points, which are much sparser compared to Gaussian points, significant occlusion effects are rare. Moreover, using rasterization for projection reduces the accuracy of feature extraction.

Table D2. More ablation study results.

| Method | PSNR(dB) \uparrow |
|---------------------------------|---------------------|
| Add-Attention-based view fusion | 33.58 |
| Add-Occlusion aware projection | 33.50 |
| Ours-s | 33.62 |

E. Mode discussion

E.1. Frame Jittering

As shown in the supplementary video, frame jittering in our method mainly occurs in static background areas. Comparing adjacent frames Fig. E1, we observe that key-frame optimization causes disturbances in the background, while no such issue arises between adjacent candidate frames. This suggests that key-frame optimization deforms Gaussians in the background, particularly with floaters Fig. F3. In the moving foreground, AGM-Net prevents jitter seen in 3DGStream by smoothing point deformation. A potential solution is to segment the scene into foreground and background and apply the segmentation mask during key-frame optimization. A more robust first-frame reconstruction in sparse views could also help.

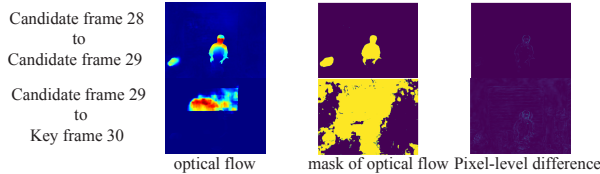


Figure E1. The difference between renderings of Adjacent frames

E.2. The impact of the number of anchor points

We tested performance with varying numbers of anchor points, shown in Fig. E2. The number of anchor points has little impact on performance.

F. More limitations and future work

There are additional limitations that constrain the performance of IGS, which also present opportunities for future research directions.

First, the performance of streaming-based dynamic scene reconstruction is influenced by the quality of static reconstruction in the first frame[7]. Poor reconstruction in the first frame, such as the presence of excessive floaters around moving objects as shown in Fig. F3, can degrade

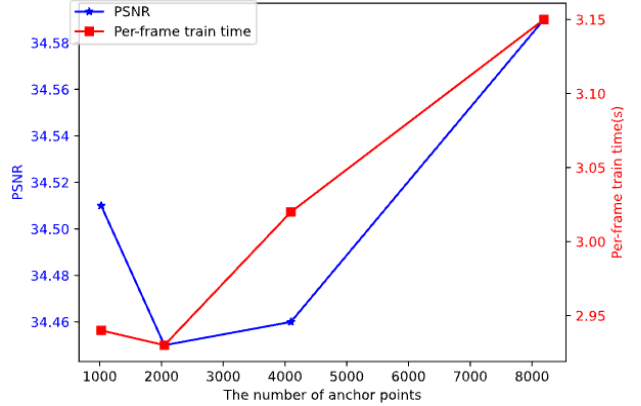


Figure E2. The impact of the number of anchor points

the performance of AGM-Net. Although addressing static reconstruction is beyond the scope of our work, adopting more robust static reconstruction methods could enhance the results of dynamic scene reconstruction. Second, AGM-Net has been trained on four sequences from the N3DV indoor dataset. The limited size of the training data constrains its generalization capability. Training on larger-scale multi-view video sequences is a promising direction for future improvements. Notably, our method only relies on view synthesis loss for supervision, making it easier to incorporate large-scale datasets without requiring annotated ground truth. Finally, our current approach injects depth and view conditions into the embeddings of an optical flow model to enable awareness of 3D scene information. Leveraging more accurate long-range optical flow[9] or scene flow[5, 8] methods could further improve our results.

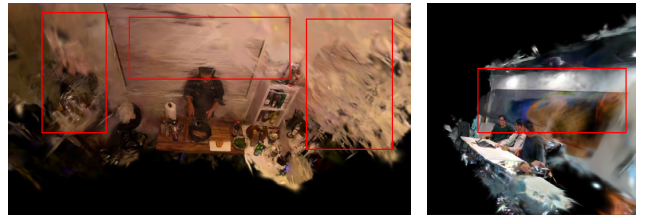


Figure F3. Bad Case in first-frame reconstruction: Due to sparse viewpoints, floaters are present around moving objects, which negatively impact our streaming performance and lead to issues such as background jitter.

G. More results

The per-scene comparison results on the N3DV dataset against previous SOTA methods [1, 2, 4, 7, 10–12] are shown in Tab. G3. Further qualitative comparisons with 3DGStream[7] are illustrated in Fig. G4.

Additionally, we provide videos showcasing the re-

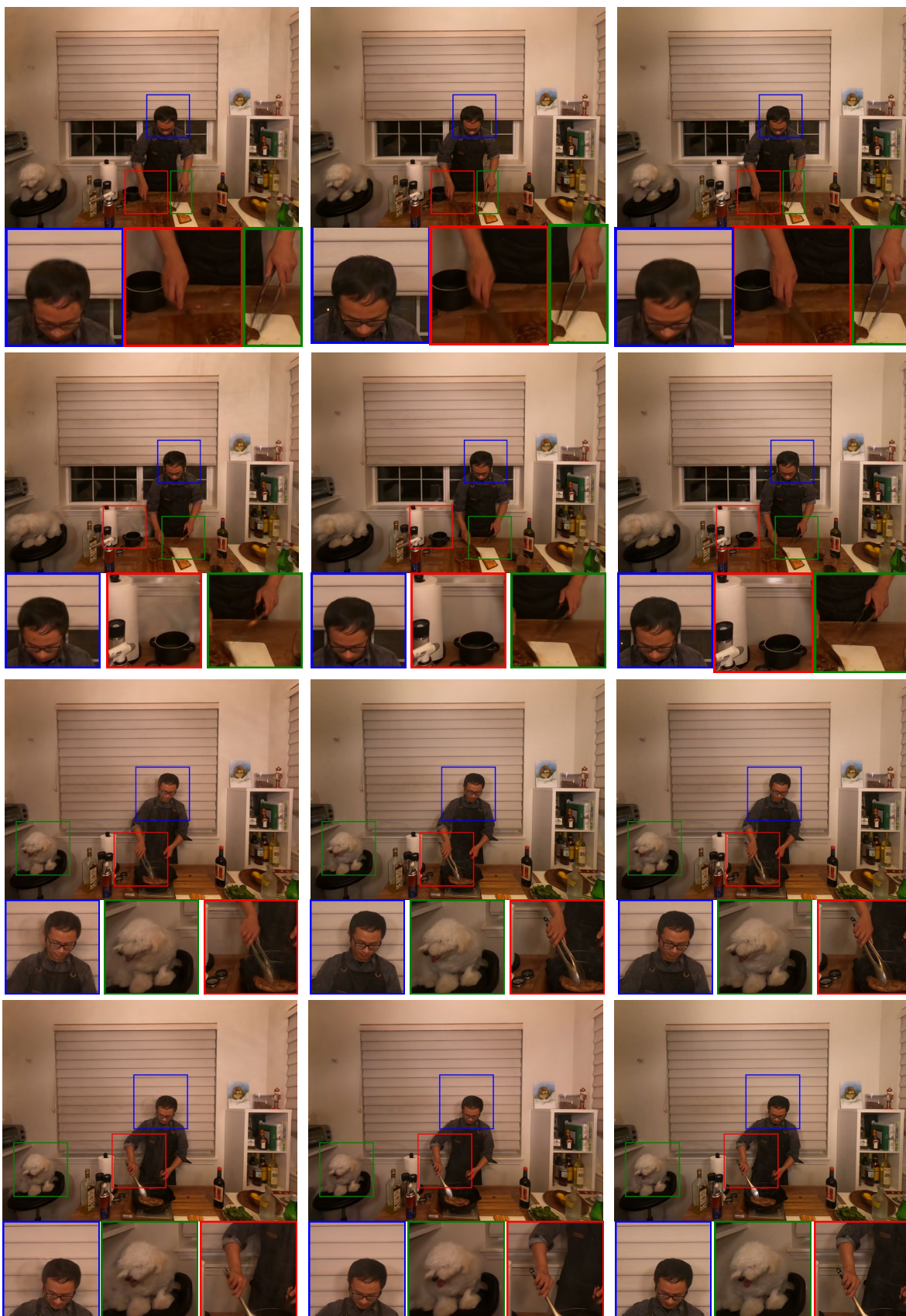
Table G3. Per-scene results on N3DV

| Method | cut roasted beef | | | sear steak | | |
|-------------------|------------------|--------------|--------------|--------------|--------------|--------------|
| | PSNR(dB)↑ | DSSIM↓ | LPIPS↓ | PSNR(dB)↑ | DSSIM↓ | LPIPS↓ |
| Offline training | | | | | | |
| Kplanes[1] | 31.82 | 0.017 | - | 32.52 | 0.013 | - |
| Realtime-4DGS[12] | 33.85 | - | - | 33.51 | - | - |
| 4DGS[10] | 32.90 | 0.022 | - | 32.49 | 0.022 | - |
| Spacetime-GS[4] | 33.52 | 0.011 | 0.036 | 33.89 | 0.009 | 0.030 |
| Saro-GS[11] | <u>33.91</u> | 0.021 | <u>0.038</u> | 33.89 | <u>0.010</u> | 0.036 |
| Online training | | | | | | |
| StreamRF[2] | 31.81 | - | - | 32.36 | - | - |
| 3DGStream[7] | 33.21 | - | - | 33.01 | - | - |
| 3DGStream[7]† | 32.39 | <u>0.015</u> | 0.042 | 33.12 | 0.014 | 0.036 |
| Ours-s | 33.62 | 0.012 | 0.048 | <u>34.16</u> | <u>0.010</u> | 0.038 |
| Ours-l | 33.93 | 0.011 | 0.043 | 34.35 | <u>0.010</u> | <u>0.035</u> |

construction results for the sear steak test scene from N3DV, including a Test-Viewpoint Video and a Free-Viewpoint Video generated using IGS. For the Free-Viewpoint Video, the viewpoints are uniformly sampled on a sphere to highlight the ability of our IGS to support free-viewpoint interaction with dynamic scenes. The results are available in the video files IGS-s_testview.mp4, IGS-l_testview.mp4 and IGS-freeview.mp4.

References

- [1] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2, 3
- [2] Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Ping Tan. Streaming radiance fields for 3d video synthesis. *Advances in Neural Information Processing Systems*, 35: 13485–13498, 2022. 1, 2, 3
- [3] 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. 1
- [4] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023. 2, 3
- [5] Lukas Mehl, Azin Jahedi, Jenny Schmalfuss, and Andrés Bruhn. M-FUSE: Multi-frame fusion for scene flow estimation. In *Proc. Winter Conference on Applications of Computer Vision (WACV)*, 2023. 2
- [6] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 1
- [7] Jiakai Sun, Han Jiao, Guangyuan Li, Zhanjie Zhang, Lei Zhao, and Wei Xing. 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20675–20685, 2024. 1, 2, 3
- [8] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [9] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: Backward accumulation for long-range optical flow. *arXiv preprint arXiv:2308.13133*, 2023. 2
- [10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2, 3
- [11] Jinbo Yan, Rui Peng, Luyang Tang, and Ronggang Wang. 4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 7871–7880, New York, NY, USA, 2024. Association for Computing Machinery. 3
- [12] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 3



3DStream

IGS(Ours)

GT

Figure G4. Qualitative comparison from the N3DV dataset.