

EgoGaussian: Dynamic Scene Understanding from Egocentric Video with 3D Gaussian Splatting

Supplementary Material

1. Static Reconstruction Pipeline

A graphical illustration of our static reconstruction pipeline is provided in [Figure 6](#). For each static clip from the egocentric video, we get two Gaussian sets for static background and dynamic object respectively, where the masks projected from object Gaussians are used to complete the previously occluded parts of background.

2. More Experiment Results

2.1. Free-Viewpoint Rendering

Our compact scene representation allows us to render novel views of the dynamic scene from arbitrary viewpoints. We show such renderings in the supplementary video.

2.2. Novel View Synthesis

The supplementary video shows the sequences from which the images originate separately. The video results suggest that the deformation-field-based method is not able to track the rigid movement of objects in egocentric video and tends to overfit to the training views.

2.3. Reconstruction of Dynamic Object

We also compare our methods with Deformable 3DGS [1] and 4DGS [2] on datasets of frames that exclusively contain the targeted object to exclude the effects from both static background and human body. The datasets are preprocessed to crop out everything except the dynamic object; gradient updates are also disabled on hands that may obscure the object during interaction.

As illustrated in the supplementary video, the deformation-field-based methods encounter difficulties tracking and modeling the dynamic object—Def-3DGS [3] can only model the very initial movement while 4DGS [2] does not yield any meaningful rendering. Our method is able to accurately track and therefore reconstruct the object with the sequential pose estimation.

2.4. Pose Estimation with Larger Time Gap

In Table 2, we show that we are still able to model the dynamic objects in the scene accurately with a larger time gap in pose estimation. This exchanges a slight performance cost for a significant reduction in training time. We further demonstrate this qualitatively in video results and visualized object trajectories. For simple movements where translation is dominant, the object reconstruction quality and estimated trajectory after pose interpolation remain similar;

while for more complicated movements involving rotation, we observe degradation in reconstruction quality.

Although our method can track the dynamic object even across larger time gaps, we cannot properly reconstruct the background, likely due to much sparser information in the presence of strong camera motion. This applies not only to our method but also to both 4DGS and Deformable 3DGS. See examples of this in [Figure 7](#).

2.5. Failure Cases

We mainly observe two types of reconstruction failure. The first type is when the interacted object moves out of sight, which sometimes happens in egocentric video for cameras with a wider field of view, the pose estimation can get stuck in a local minima and lose track of the object in the subsequent frames. The second type arises when the camera registration quality is poor, which is often due to featureless surfaces and insufficient overlap between video frames. The inaccurate camera poses result in a static reconstruction that lacks any geometric understanding of the scene, thereby causing the dynamic reconstruction to fail.

References

- [1] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Ming Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1
- [2] Guanjin Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 1
- [3] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 1

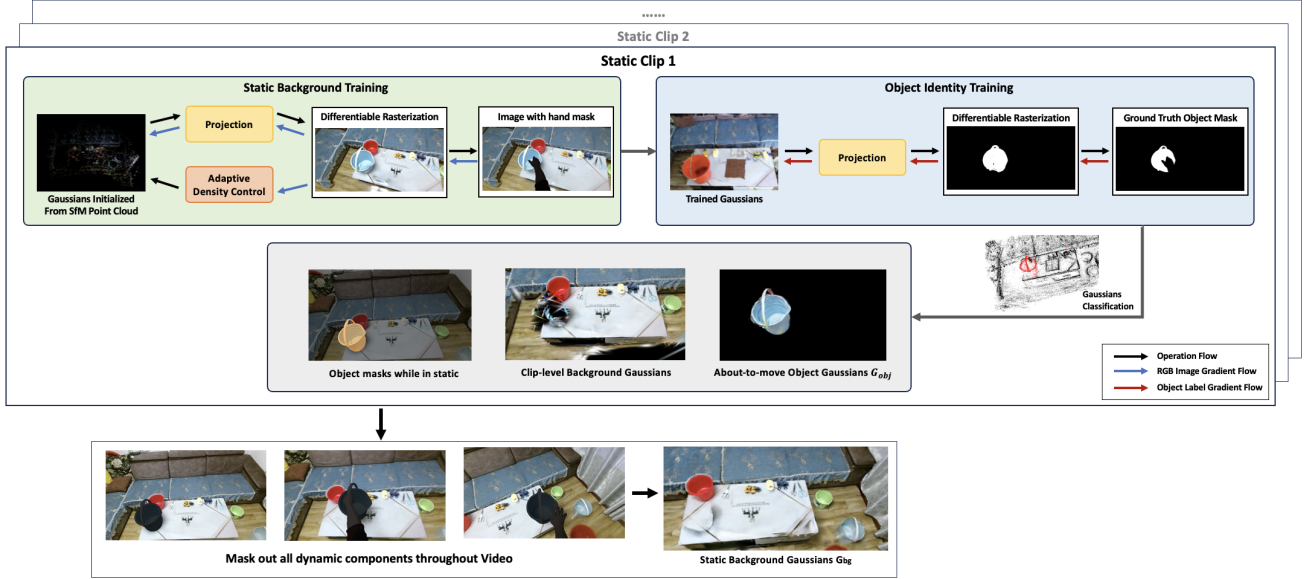


Figure 6. **Static Reconstruction Pipeline.** We use this pipeline to jointly reconstruct and segment dynamic object from static background from a static egocentric video clip.

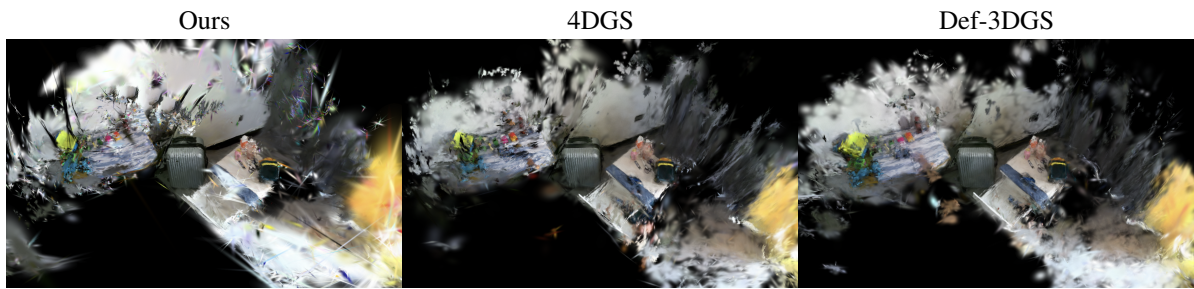


Figure 7. **Comparison with SOTA methods on background reconstruction with larger time gap** All methods suffer from floaters due to sparse training viewpoints