EgoLifter: Open-world 3D Segmentation for Egocentric Perception

Qiao Gu[1,2], Zhaoyang Lv[2], Duncan Frost[2], Simon Green[2], Julian Straub[2], Chris Sweeney[2]
[1] University of Toronto, [2] Meta Reality Labs

Build diverse and photorealistic 3D digit assets, from egocentric videos of natural everyday activities.

Overview

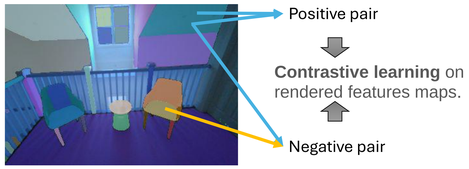
- **Egocentric videos** capture diverse object instances and extensive dynamics.
- **Open-world 3D instance segmentation** without expensive annotation or extra training.
- **A transient prediction module** to remove floaters in reconstruction.
- **A dynamic egocentric video benchmark** for 3D reconstruction and segmentation.

3D Gaussians with Contrastive Instance Feature

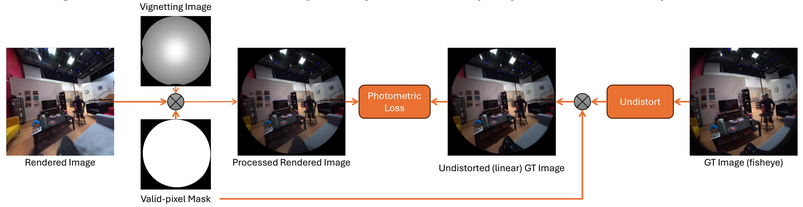
Contrastive Lift [1]: implicitly solving the multi-view association problem of 2D masks.

- 3D position
- 3D covariance
- Color
- Opacity
- **Instance Feature**

3D Gaussian [2]



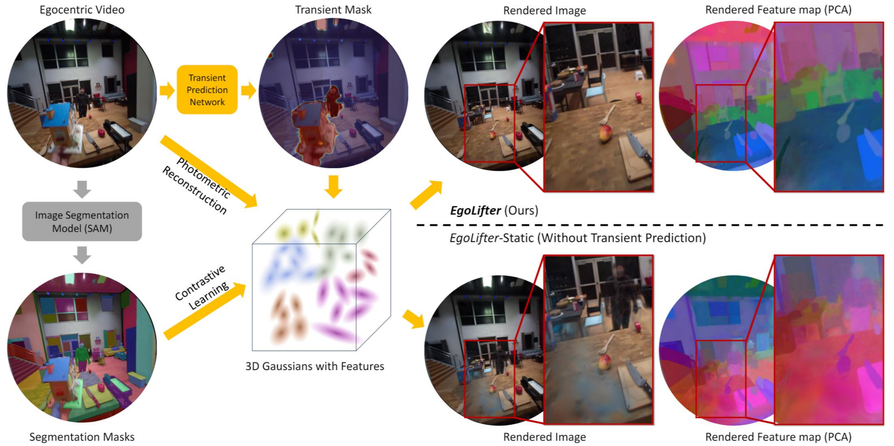
Handle egocentric videos collected by fisheye cameras (Project Aria Glasses).



References
[1] Yash Bhalgat, et al . "Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion." NeurIPS 2023.
[2] Kerbl, Bernhard, et al. "3D Gaussian Splatting for Real-Time Radiance Field Rendering." ACM Trans. Graph 2023.
[3] Ye, Mingqiao, et al. "Gaussian grouping: Segment and edit anything in 3d scenes." ECCV 2024.
[4] Yang, Zeyu, et al. "Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting." ICLR 2024.

Transient Prediction for Floaters Removal

- Egocentric videos present sparse and rapid dynamic phenomena.
- Naïve 3D reconstruction results in many ghostly "floaters" and bad rendering results.
- We proposed a transient prediction module to remove dynamics, which can be trained in a self-supervised manner without extra annotations.



Experiment & Quantitative Results

Evaluations

- **Aria Digital Twin (ADT)**: Provides 2D and 3D object segmentation ground truth.
- **Query-based segmentation**: Segmentation by in-view or cross-view query features.
- **Aria Everyday Activities (AEA) & Ego-Exo4D**: For qualitative results.

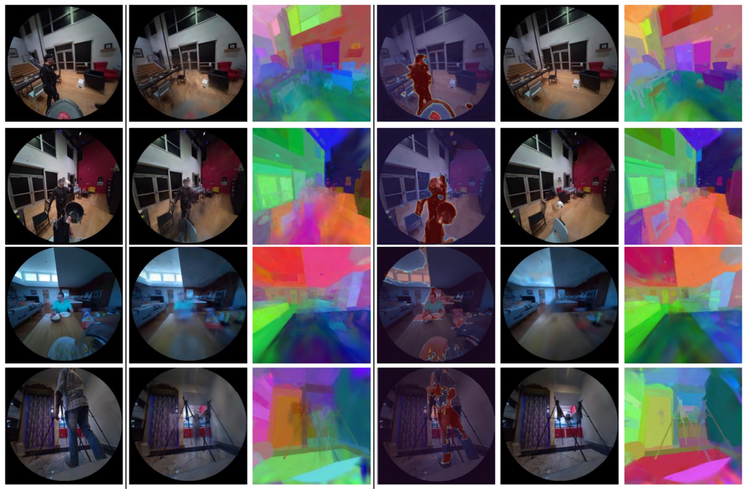
Baselines & Ablations

- **SAM**: Segment-Anything on rendered images; only supports in-view queries
- **Gaussian Grouping [3]**: Uses a video tracker for mask association; learns mask ID.
- ***EgoLifter*-Static**: Disables the transient prediction network.
- ***EgoLifter*-Deform**: Uses a deformable variant of 3DGS [4] to handle dynamics.

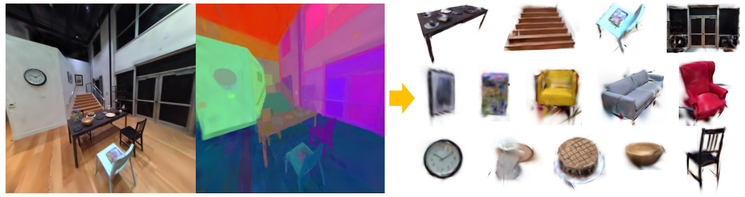
Evaluation	2D mIoU (In-view)			2D mIoU (Cross-view)			PSNR			3D mIoU
Object set	Static	Dynamic	All	Static	Dynamic	All	Static	Dynamic	All	Static
SAM	54.51	32.77	50.69	-	-	-	-	-	-	-
Gaussian Grouping	35.68	30.76	34.81	23.79	11.33	21.58	21.29	14.99	19.97	7.48
EgoLifter-Static	55.67	**39.61**	52.86	51.29	18.67	45.49	21.37	15.32	20.16	21.10
EgoLifter-Deform	54.23	38.62	51.49	51.10	18.02	45.22	21.16	**15.39**	19.93	20.58
***EgoLifter* (Ours)**	**58.15**	37.74	**54.57**	**55.27**	**19.14**	**48.84**	**22.14**	14.37	**20.28**	**23.11**

Qualitative Results

Main Results on Egocentric Datasets



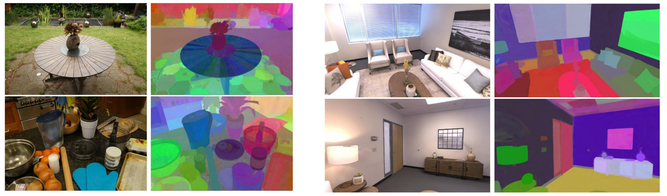
Open-vocabulary Object Reconstruction and Extraction



Rendered Image and Features from *EgoLifter*

Extracted 3D Objects

Results on Non-egocentric Datasets



Rendered RGB | Rendered Feature (PCA) | Rendered RGB | Rendered Feature (PCA)