FROSSv2: Improving Real-Time 3D Semantic Scene Graph Generation Model

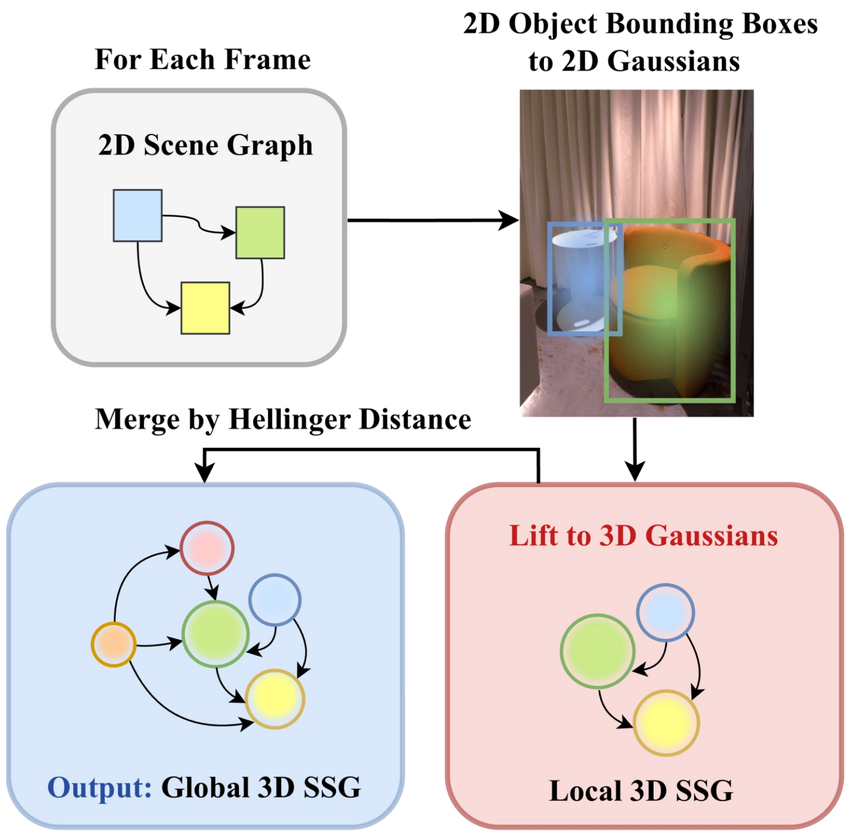
Hao-Yu Hou, Tzu-Yun Huang, Yu-Shin Jian
Department of Computer Science, University of Toronto

Introduction

- 3D Semantic Scene Graphs (SSGs) are graph data structures abstracting a 3D scene.
- They provide efficient **scene representations** for robotics, VR / AR, and more.
- **Efficiency** and **accuracy** remain challenging.

Related Work

- Kim et al. [1] introduced the first 3D SSG generation method.
- Wald et al. [2] proposed 3DSSG, the first large scale benchmark for 3D SSG evaluation.
- Wu et al. [3] proposed a fast and incremental prediction method based on point cloud input.
- **FROSS** [4] achieved **real-time incremental** SSG generation by representing objects as **3D Gaussians**.



Problems with FROSS [4]

(a) Incorrect Object Position
If the center of bounding box is **occluded**, the estimated **depth** will be incorrect.

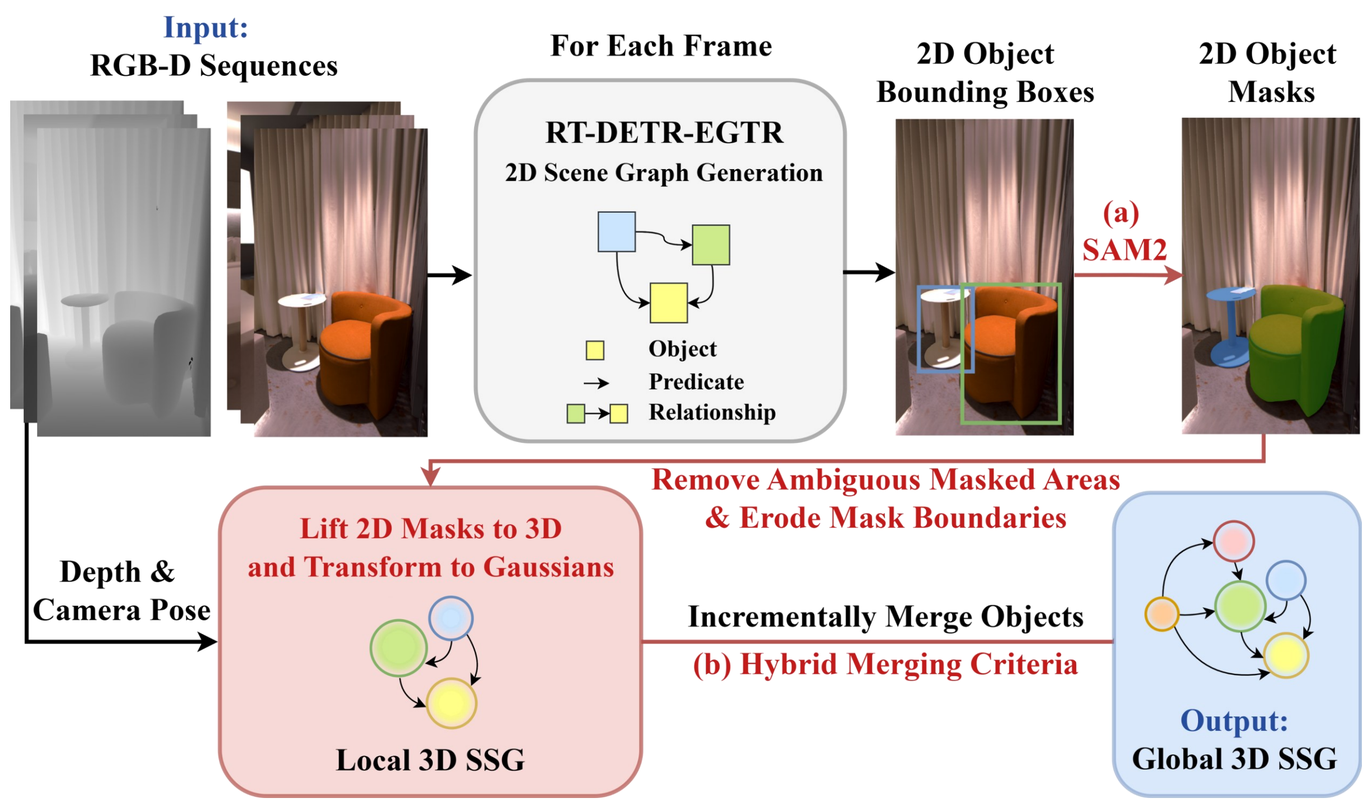
(b) Inaccurate Variance on Depth Axis
Directly lifting a 2D Gaussian to 3D leaves the **depth variance undefined**.

(c) Merging Criteria
The update algorithm uses only the **similarity between Gaussians** in the **same class**.

References

[1] Ue-Hwan Kim et al., 3-D Scene Graph: A Sparse and Semantic Representation of Physical Environments for Intelligent Agents. In *IEEE Trans. Cybernetics*, 2019.
[2] Johanna Wald et al., Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions. In *CVPR*, 2020.
[3] Shun-Cheng Wu et al., Incremental 3D Semantic Scene Graph Prediction from RGB Sequences. In *CVPR*, 2023.
[4] Hao-Yu Hou et al., FROSS: Faster-than-Real-Time Online 3D Semantic Scene Graph Generation from RGB-D Images. In *ICCV*, 2025.
[5] Ravi Nikhila et al., SAM 2: Segment Anything in Images and Videos. In *arXiv*, 2024.

Methodology



We address the aforementioned issues by incorporating the **Segment Anything Model 2 (SAM 2)** [5] and a **hybrid merging criteria**.

(a) Using SAM 2 for more accurate 3D Gaussian estimation

1. SAM 2 generates segmentation masks from 2D bounding boxes.
2. Overlap mask regions from different classes are removed, and mask boundaries are eroded.
3. Back-project masks to 3D, and compute means and covariances.

(b) Hybrid criteria for incremental object merging

Merge two objects if their weighted sum of Gaussians' Hellinger distance and class-distribution dot product is below a threshold.

Experimental Results

- **Quantitative Comparison on the 3DSSG [2] Dataset**

Method	Recall (%)		
	Object	Predicate	Relationship
FROSS (Baseline)	62.4	33.0	27.9
FROSS + SAM 2 w/o post-processing	64.3	39.2	32.6
FROSS + SAM 2 w/ post-processing	69.2	40.9	34.8
FROSSv2 (SAM 2 + Hybrid, Ours)	**71.2**	**46.5**	**39.6**

- **Qualitative Comparison on the ReplicaSSG [4] Dataset**

(a) The predicted 'chair' (yellow) is correctly lifted.
(b) Can better capture the shape of flat objects, such as windows.
(c) Rather than projecting Gaussians onto object surfaces, FROSSv2 better approximates each object's shape.