Joint 2D-3D Segmentation and Association in Street-level Imaging

Amir Melnikov

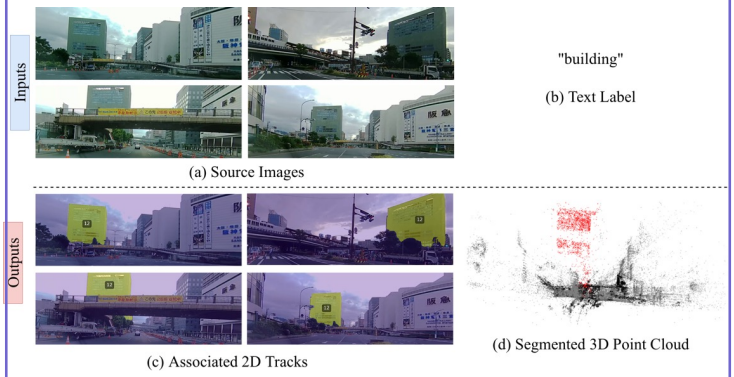
Department of Systems and Control Engineering, School of Engineering, Institute of Science Tokyo, Japan

Institute of SCIENCE TOKYO

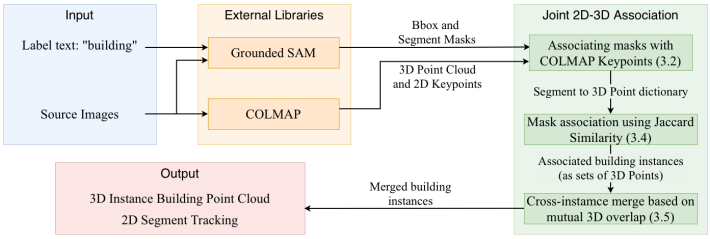
1. INTRODUCTION

Research Objectives

- To demonstrate a non-sequential image processing pipeline for object segmentation in 2D and 3D.
- Associate 2D texture information and 3D geometrical information.
- To utilize existing SOTA frame detection and segmentation tools for data processing without retraining.



2. PROPOSED METHODOLOGY



2.1. The framework comprises five modular stages:

1) **Semantic Extraction:** Grounded-SAM generates zero-shot 2D masks (M) based on text prompts.
2) **Geometric Projection:** 3D points reconstructed via SfM (COLMAP) are reprojected onto the 2D image plane to establish a link between pixels and 3D coordinates.
3) **Similarity Check:** Each mask is compared to masks in other images for shared 3D Point IDs.
4) **Cross-View Association:** Masks are grouped into instances based on similarity check result.
5) **Global Aggregation:** Redundant clusters are merged through a final geometric consistency check.

2.2. Association Logic

Instance similarity is quantified using the **Jaccard Similarity Coefficient** (J) applied to 3D point sets:

$$J(Sa, Sb) = \frac{|Sa \cup Sb|}{|Sa \cap Sb|}$$

This formulation ensures that association is grounded in physical world-space rather than transient pixel-space appearance.

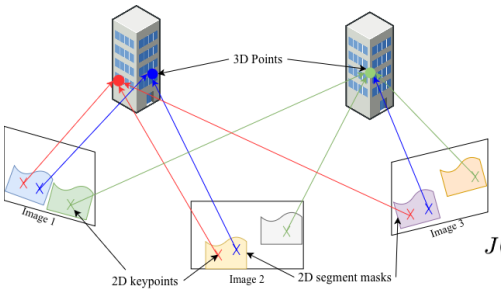


Image 2 Segment 1 (Sa)	Image 3 Segment 1 (Sb)
Point 1	Point 1
Point 2	Point 2
Point 3	Point 3
Point 4	Point 6
Point 5	Point 7

$$J(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|}$$

3. METRICS AND EVALUATION

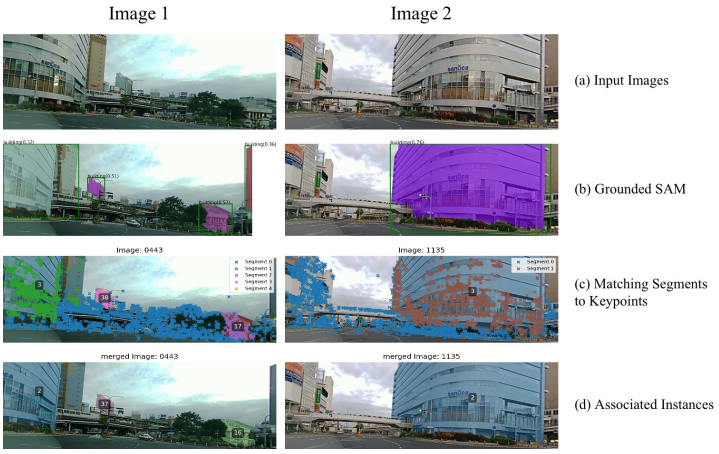
- **Rationale:** Conventional MOT metrics (e.g., IDF1, MOTA) are unsuitable for wide-baseline tasks as they heavily penalize instances entering/exiting frames and assume dense temporal continuity.
- **Coverage (C):** Coverage measures how consistently a tracker retrieves the correct building identity across all frames where that building appears in.

$$\text{Coverage} = \frac{\textit{Correctly Identified Frames}}{\textit{Total Identifiable Frames in GT}}$$

- **Adjusted Coverage (C_{adj}):** To isolate the performance of the association logic from upstream segmentation failures, this metric removes frames where the detector produced no mask:

$$\text{Adjusted Coverage} = \frac{\textit{Correctly Identified Frames}}{\textit{Total Identifiable Frames in GT} - \textit{misdetections}}$$

4. EXPERIMENTAL RESULTS AND ANALYSIS



Setup

- **Datasets:** Evaluated on Dataset 1 (Sannomiya Station, Kobe, Japan) and CityScapes (Zurich).
- **Baselines:** Compared against a naive IoU Tracker and the state-of-the-art SAM2+MOTRv2 video segmentation feature-based 2D tracker.

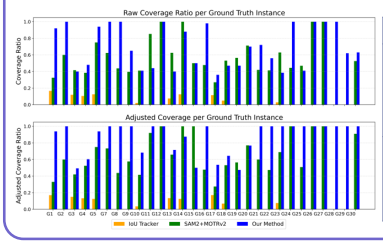


Table 1: Average Coverage and Adjusted Coverage

Average Coverage Ratio	Our Algo	0.66
	IoU	0.03
	SAM2+MOTRv2	0.53
Average Adjusted Coverage Ratio	Our Algo	0.83
	IoU	0.05
	SAM2+MOTRv2	0.606

5. CONCLUSIONS AND FUTURE WORK

- **Contribution:** This work demonstrates that integrating multi-view geometric consistency effectively bridges the gap between 2D semantics and 3D space, enabling robust instance association in complex urban environments.
- **Limitations:** The framework currently relies on the computational intensity of full Structure-from-Motion (SfM) reconstruction and the quality of upstream foundation models in detection and segmentation.
- **Future Directions:**
  - Additional datasets review, in conjunction with Aerial mapping
  - Pipeline application: extraction of single 3D object reconstruction from segmented frames using ZipNERF
  - Verification of method with additional 2D baselines

6. REFERENCES

[1] **Kirillov, A., et al.**: "Segment Anything," *ICCV*, pp. 4015-4026 (2023).
[2] **Ravi, N., et al.**: "SAM 2: Segment Anything in Images and Videos," *arXiv:2408.00714* (2024).
[4] **Liu, S., et al.**: "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," *ECCV* (2024).
[5] **Cordts, M., et al.**: "The Cityscapes Dataset for Semantic Urban Scene Understanding," *CVPR*, pp. 3213-3223 (2016).
[6] **Schönberger, J. L., et al.**: "Structure-from-Motion Revisited," *CVPR*, pp. 4104-4113 (2016).
[7] Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively Segmenting Instances and Semantics in Point Clouds. (CVPR) (2019).
[8] **Liu, S., et al.**: "Grounded SAM: Assembling Open-World Models for RL Segmentation," *arXiv:2303.15343* (2023).
[9] **Bernardin, K. & Stiefelhagen, R.**: "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP JIVP*, Art. 246309 (2008).