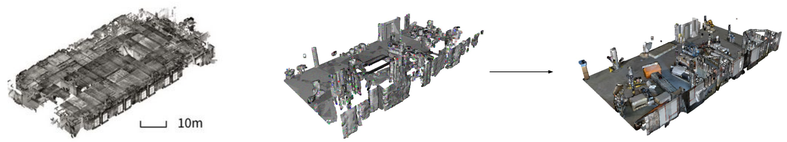
Towards Open Scene Understanding For Construction Analysis

Emily Steiner

Introduction

- Construction tracking and monitoring currently lacks strong quantitative methods to understand the progress and efficiency of a construction site over time.
- This work works towards providing a feedback mechanism by evaluating a 3D open-scene understanding method for the unique challenges of the construction environment.



- Leveraging robust 2D foundational vision models to provide context from multi-view renderings by semantically segmenting a 3D mesh scene.

Related Work

- **PVT3:** closed-vocabulary set 3D semantic segmentation, trained on ScanNet
- **OpenScene:** ensembles CLIP/Lseg features from multi-view projections and 3D model
- **OpenMask3D:** uses CLIP features of cropped views of object assets then projected to 3D masks, relies on 3D masking methods trained on ScanNet
- **CLIP-FO3D:** projects CLIP feature embeddings then trains a 3D CNN to learn to predict the embeddings

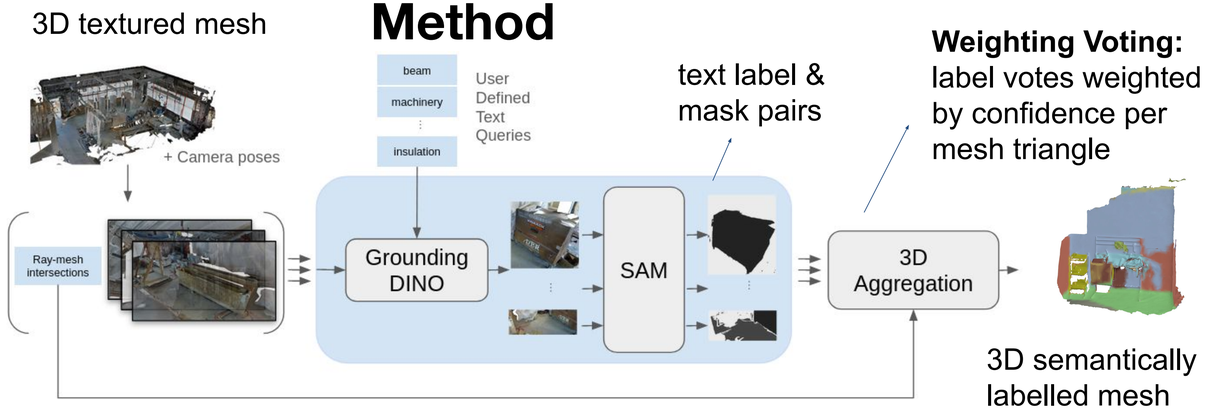
References

[1] Liu et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, Mar. 2023.
[2] A. Kirillov et al. Segment Anything, Apr. 2023.
[3] Ren et al. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks, Jan. 2024.

Method

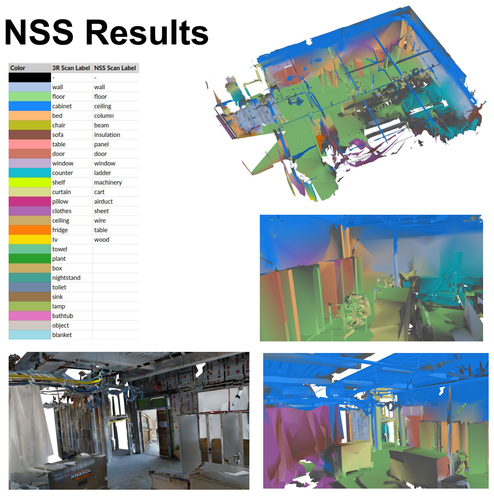
Data
1. Nothing Stands Still
2. 3RScan

Open-vocabulary scene understanding: Proposed method is agnostic to the semantic segmentation categories



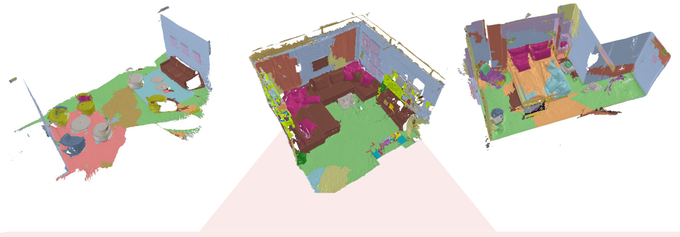
Experimental Results

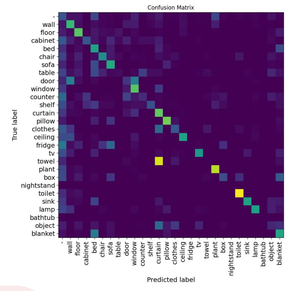
NSS Results



Color	3R Scan Label	NSS Scan Label
	wall	wall
	floor	floor
	cabinet	ceiling
	bed	column
	chair	beam
	sofa	insulation
	table	panel
	door	door
	window	window
	counter	ladder
	shelf	machinery
	curtain	cart
	pillow	airduct
	clothes	sheet
	ceiling	wire
	fridge	table
	tv	wood
	towel	
	plant	
	box	
	nightstand	
	toilet	
	sink	
	lamp	
	bathtub	
	object	
	blanket	

3RScan Results





Average Precision (AP) used to quantitatively evaluate results

Approach	mAP
Baseline (majority voting)	0.3542
Weighted by SAM confidence	0.3867
Weighted by GroundingDINO	0.4027
Full Method (weighted by both)	**0.4482**

Table 1. Ablation Experiments

Conclusions

- Qualitative the 3D construction dataset shows some successful segmentation, but significant improvements are required to properly capture the long-tail concepts of interest in the construction industry.
- Promise with 3RScan segmentation, allowed for method tuning, showed the importance of proper image context

Future Work: Fine-tuning SAM using 2D construction dataset, improving 3D data rendering method compute time, investigating feature embeddings