Indoor Scene Understanding via 2D and 3D Semantic Segmentation: Integrating Depth for Geometry-Aware Reconstruction

Karthik Pythireddi [1] Arya Bakthiar [2]

[1]Electrical Engineering, Stanford University
[2]Computer Science, Stanford University

Introduction

Indoor Scene understanding plays a crucial role in applications like robotics and augmented reality. A Key challenge is capturing both appearance and geometric context to accurately interpret complex environments. We explore how combining the 2D Segmentation with geomtric cues can enhance per-frame scene interpretation and voxel-level semantic labeling.

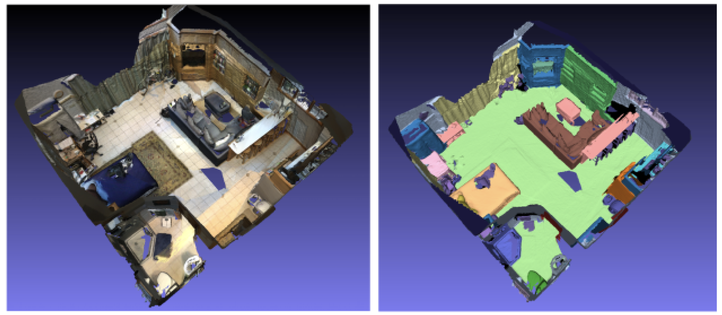

Figure 1. Mesh reconstruction of a ScanNet apartment scene.

To study this:

- We begin by benchmarking off-the-shelf 2D models (**DeepLabV3+** and **Segformer-B0**) in a *zero-shot setting* on the **ScanNet** dataset.
- We then refine one of the 2D models by incorporating depth as an additional input channel, effectively leveraging **RGB-D** information for improved 2D segmentation.
- Finally, we build a **3D volumetric representation** from all RGB-D frames and train a **3D U-Net** to perform voxel-wise semantic segmentation using either *ground-truth* or *self-calibrated* depth.

Our work demonstrates the benefit of **depth-aware modeling** in enhancing semantic segmentation for both **2D and 3D**.

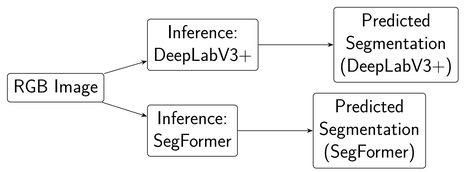

Figure 2. Inference pipeline using DeepLabV3+ and SegFormer to produce segmentation masks from RGB input.

Background and Related Work

Our Work builds on the principle of combining appearance and geometry for indoor scene understanding.

- We enhance 2D Segmentation by incorporating **Depth** as an additional input.
- Prior work shows this improves boundary segmentation [Silberman et al., 2012].
- For 3D, voxel-based models like 3D U-Net [Milletari et al., 2016] and fusion frameworks like SemanticFusion [McCormac et al., 2017] use volumetric context to improve semantic accuracy.

Dataset and Pre-Processing

Our primary data set comprises the ScanNet Dataset [Dai et al., 2017],a large-scale collection of RGB-D video sequences from indoor environments. We selected three distinct apartment scenes to serve as our main training and evaluation environment, designated as **Apartment 1 (5578 frames)**, **Apartment 2 (1988 frames)** and **Apartment 3 (4439 frames)**.Each frame in these scenes consists of corresponding RGB images, depth data, camera pose information, camera intrinsics, and semantic labels. Figure 3 shows an example of RGB-D frames from Apartment 1 with their corresponding overlay segmentation

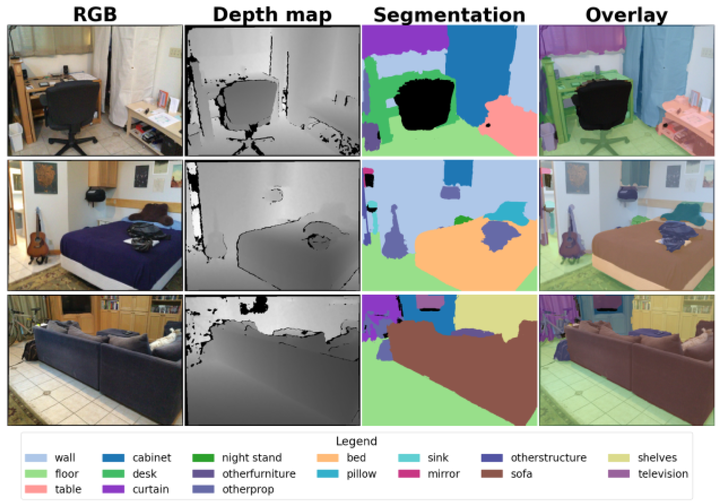

Figure 3. Scene based segmentation across 3 different image scenes

Methods

We evaluate semantic segmentation across three stages:

1. 2D Zero-Shot Inference:

- Applied pre-trained **DeepLabV3+** (COCO) and **SegFormer-B0** (ADE20K) directly on ScanNet RGB frames.
- Segmentation masks generated over 41 NYU40 classes without fine-tuning.

2. 2D SegFormer Fine-Tuning:

- Augmented SegFormer with an additional depth channel (RGB-D input).
- Used depth-aware loss and two-phase training (backbone freeze + unfreeze).

3. 3D Volumetric Segmentation:

- RGB-D frames fused into 4D TSDF volume using Open3D.
- Trained **3D U-Net** on voxelized grid ($4 \times D \times H \times W$).
- Compared performance using GT depth vs. MASt3R-predicted depth.

Experiments and Results

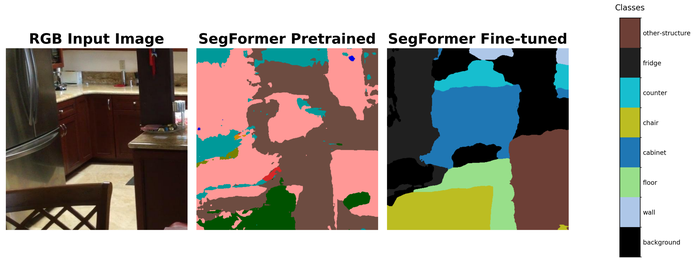

Figure 4. 2D Segmentation fine tuning results

Scene ID	Model	mIoU	F1	Accuracy	Precision	Recall	Dice	DCE Loss
Apartment 1	DeepLabV3+	0.034	0.041	0.095	0.077	0.060	0.057	11.164
	SegFormer	0.039	0.001	0.006	0.004	0.001	0.003	11.761
Apartment 2	DeepLabV3+	0.034	0.054	0.140	0.071	0.079	0.075	9.367
	SegFormer	0.011	0.014	0.069	0.015	0.014	0.018	9.617
Apartment 3	DeepLabV3+	0.033	0.052	0.219	0.060	0.074	0.079	10.732
	SegFormer	0.022	0.012	0.089	0.013	0.011	0.027	11.405

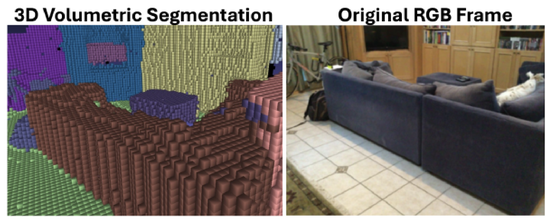

Figure 5. 3D Volumetric Segmentation

Conclusion & Future Work

- **2D Segmentation:** Off-the-shelf models struggled with indoor scenes (low mIoU), but fine-tuning SegFormer with depth significantly improved performance.
- **3D Segmentation:** 3D U-Net on TSDF grids showed that **ground-truth depth consistently outperformed MASt3R-predicted depth**.
- Depth adds clear geometric advantages for both frame-level and volumetric understanding.

Looking ahead, we aim to:

- Scale the pipeline to a broader set of environments and more diverse layouts.
- Explore real-time deployment for embodied agents like legged robots.
- Integrate transformer-based volumetric architectures to exploit multi-modal attention.

Scan to view full references and BibTeX entries
github.com/karthikpythireddi/CS231N_Final_Project