OpenScene
3D Scene Understanding with Open Vocabularies
Songyou Peng, Kyle Genova, Chiyu Max Jiang, Andrea Tagliasacchi, Thomas Funkhouser
*Work done during an internship at Google Research

1. Introduction
Problem: Traditional 3D scene understanding only train and test on some fixed common classes
Goal: A zero-shot approach to perform novel 3D scene understanding tasks w/o annotation labels
Key idea: Co-embed 3D features with CLIP image features \(\rightarrow\) naturally also with CLIP text features

2. Method
How to produce text-image-3D co-embedding?
- Multi-view Feature Fusion
- 2D-3D Ensemble
- Average Pooling
- 3D Distillation
- Inference

3. Zero-shot Open-vocabulary Scene Exploration
Input 3D Point Cloud
- "fan" - Object
- "made of metal" - Material
- "kitchen" - Room Type

Zero-shot Semantic Segmentation
- "anything soft" - Property
- "where to sit" - Affordance
- "work" - Activity

4. Additional Applications
ScanNet 3D Semantic Segmentation Benchmarks
- nuScenes

5. More Studies
Robust to Tailed Classes

Ablation Study

Paper, code, and real-time demo are available: pengsongyou.github.io/openscene