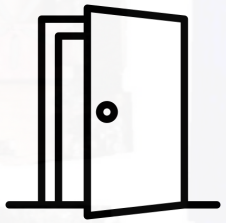


OpenScene

3D Scene Understanding with Open Vocabularies

Songyou Peng

ETH Zurich and Max Planck Institute for Intelligent Systems

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Arxivores @ Stability.ai

June 15, 2023

Who Am I?

- 4th Year PhD Student
  - Marc Pollefeys
  - Andreas Geiger

ETH zürich

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

- Internships during PhD
  - 2021: Michael Zollhoefer
  - 2022: Tom Funkhouser

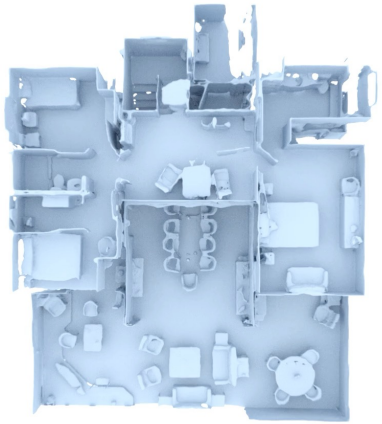
Meta

Google Research

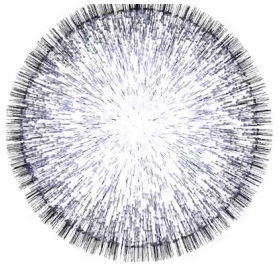
- Graduate this fall ☺

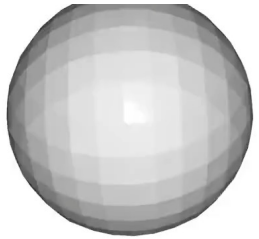
pengsongyou.github.io

My PhD Topics: Neural Scene Representations for 3D reconstruction and 3D scene understanding

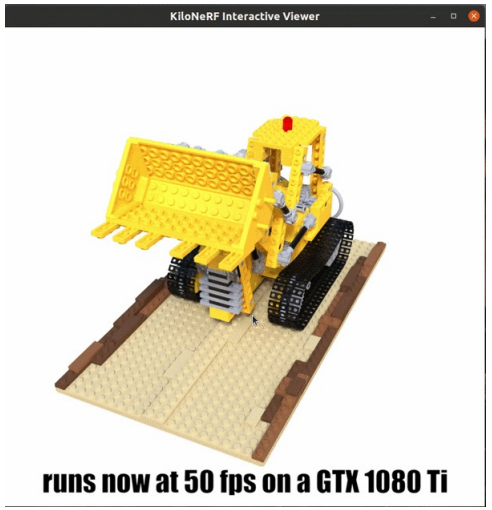


Convolutional Occupancy Nets
ECCV 2020 (Spotlight)

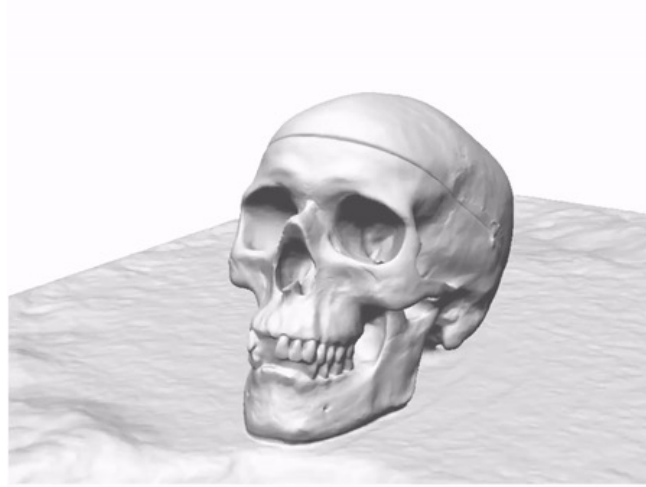




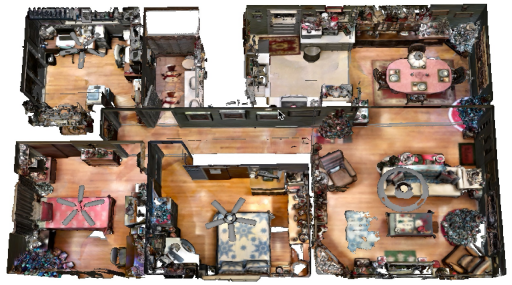
Shape As Points
NeurIPS 2021 (Oral)



KiloNeRF
ICCV 2021

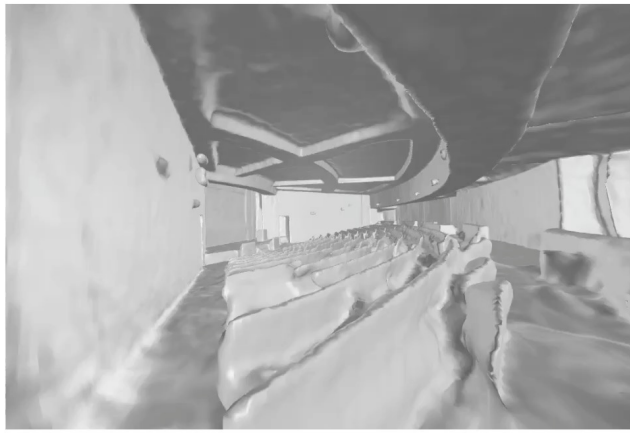


Ours
UNISURF
ICCV 2021 (Oral)

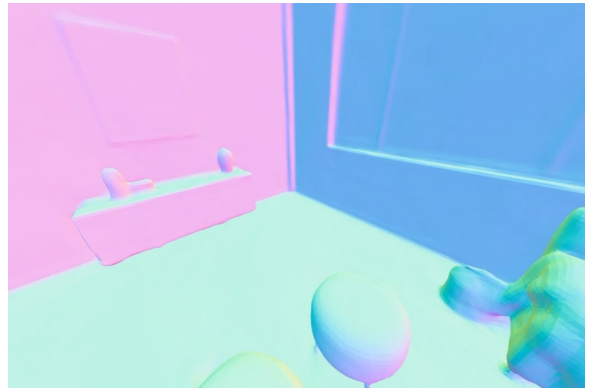


NICE-SLAM
CVPR 2022

OpenScene
CVPR 2023

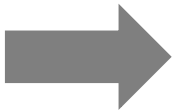


Ours
MonoSDF
NeurIPS 2022



NICER-SLAM
arXiv 2023

Input Images



3D Reconstruction

Input 3D Geometry

Traditional Semantic Segmentation

Only train and test on a few common classes

Input 3D Geometry

- Affordance prediction
- Material identification
- Physical property estimation
- Rare object retrieval
- Activity site prediction
- Fine-grained semantic segmentation
- Many more…

3D Scene Understanding Tasks w/o Labels

OpenScene

3D Scene Understanding with Open Vocabularies

CVPR 2023

Songyou Peng

Kyle Genova

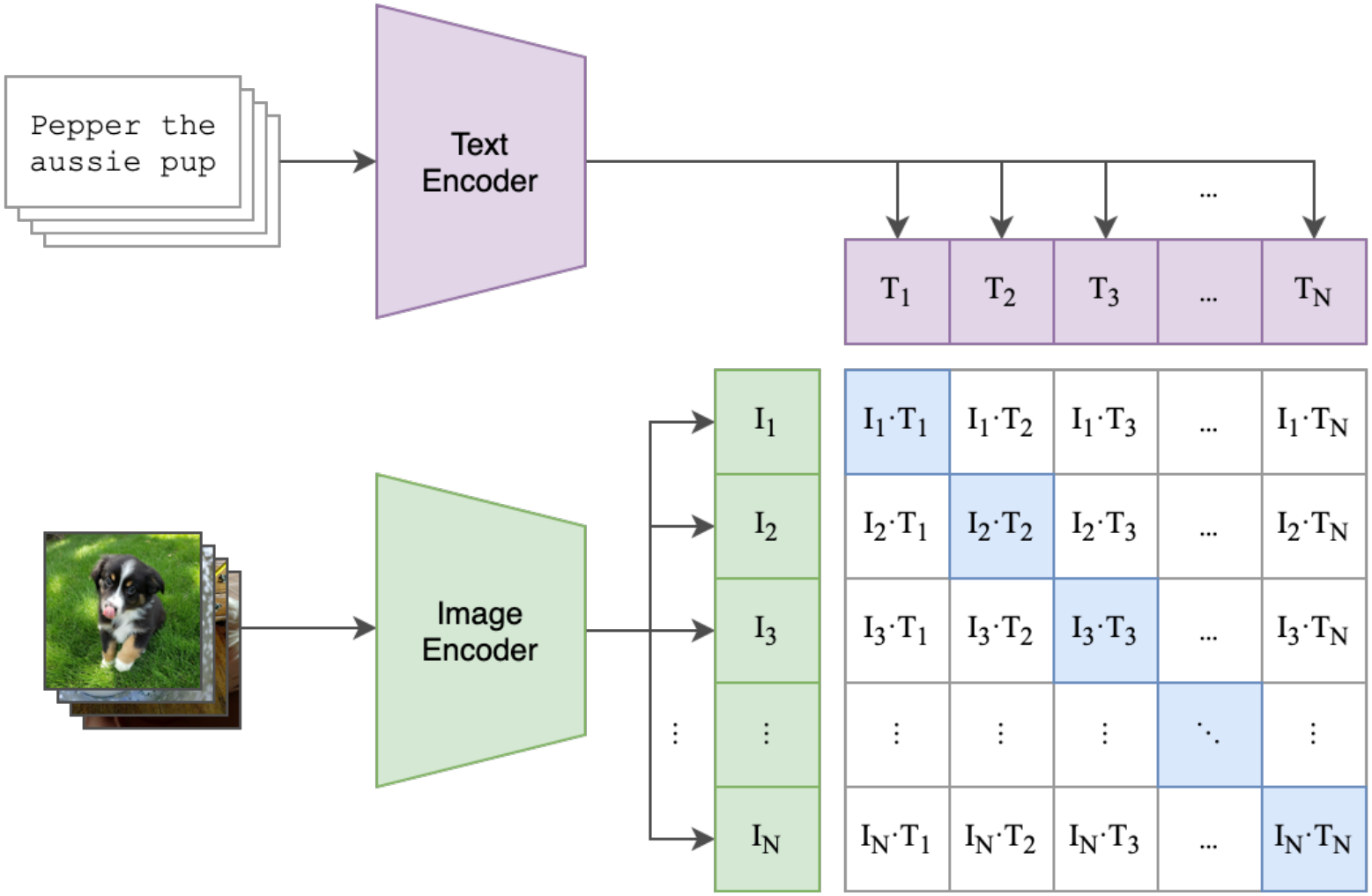
Chiyu “Max” Jiang

Andrea Tagliasacchi

Marc Pollefeys

Tom Funkhouser

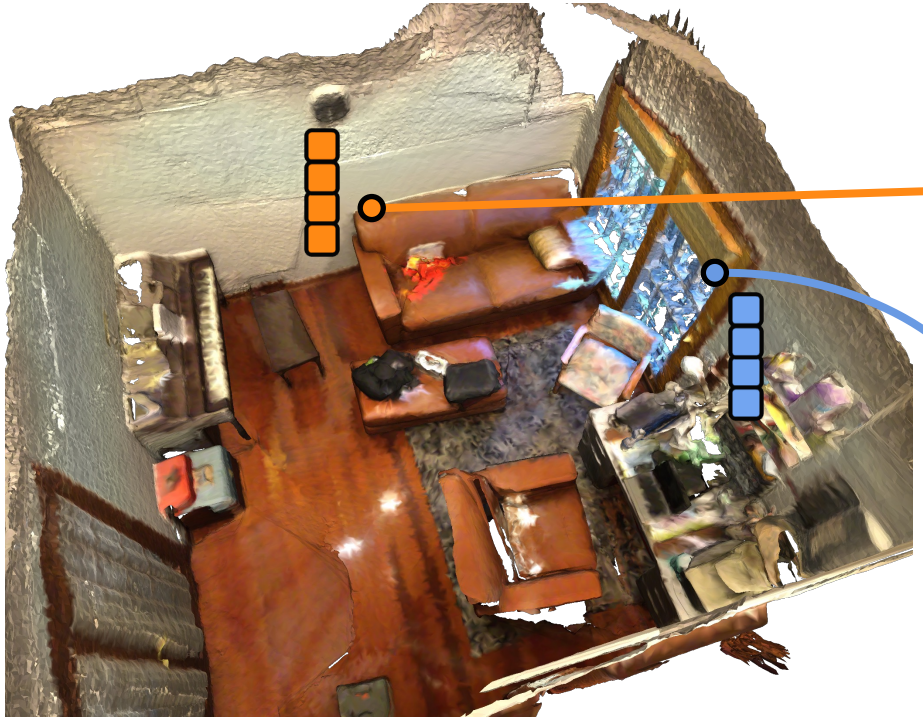
Key Idea: Co-embed 3D features with CLIP features



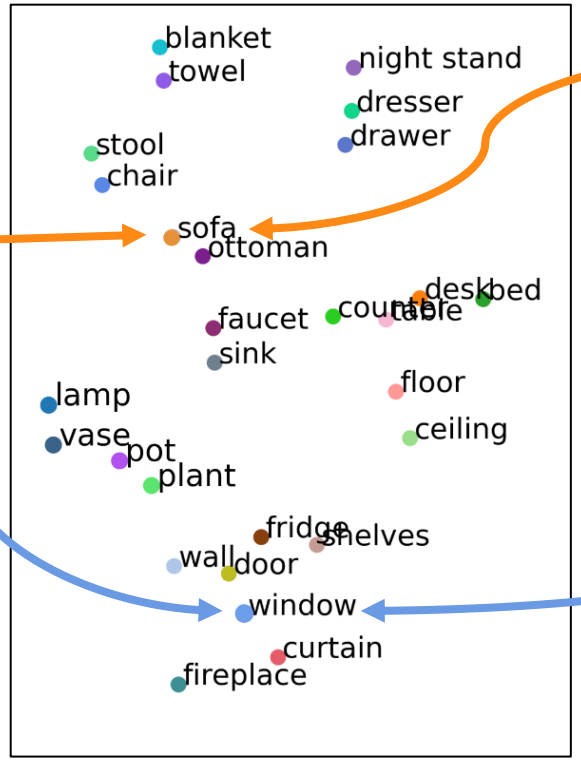
CLIP: Contrastive Language-Image Pre-Training

Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021

Key Idea: Co-embed 3D features with CLIP features



3D Geometry

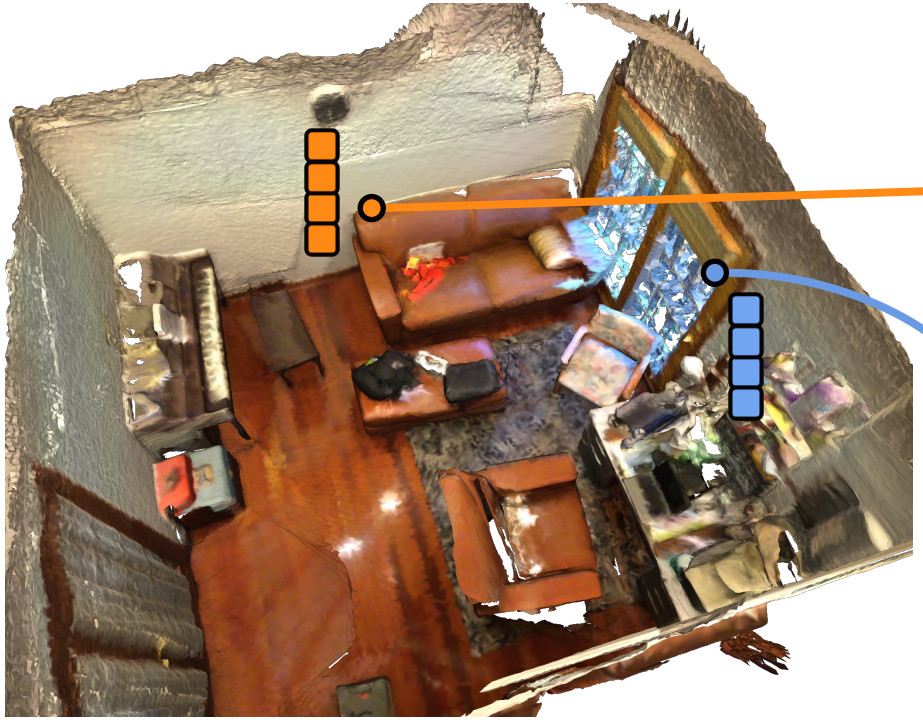


CLIP Text Features

(visualize with T-SNE)

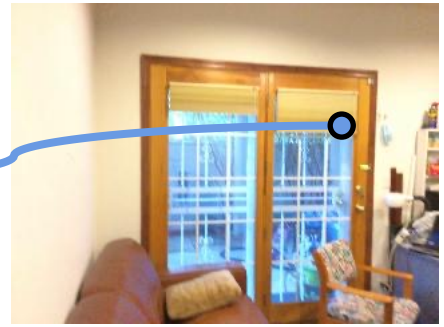
RGB Images

Key Idea: Co-embed 3D features with CLIP features



3D Geometry

CLIP Text Features
(visualize with T-SNE)

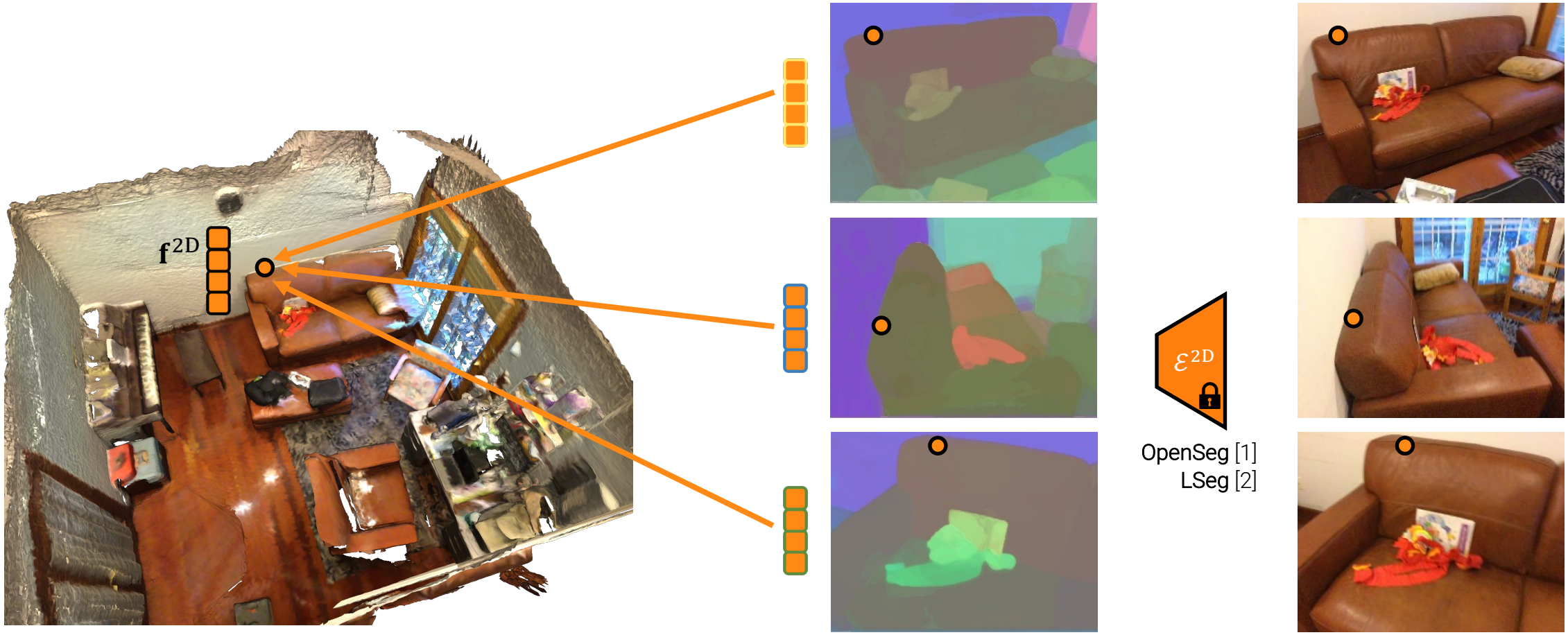


RGB Images

Note: bold word embeddings are approximate

How to Learn Such Text-Image-3D Co-Embeddings?

Step 1: Multi-view Feature Fusion



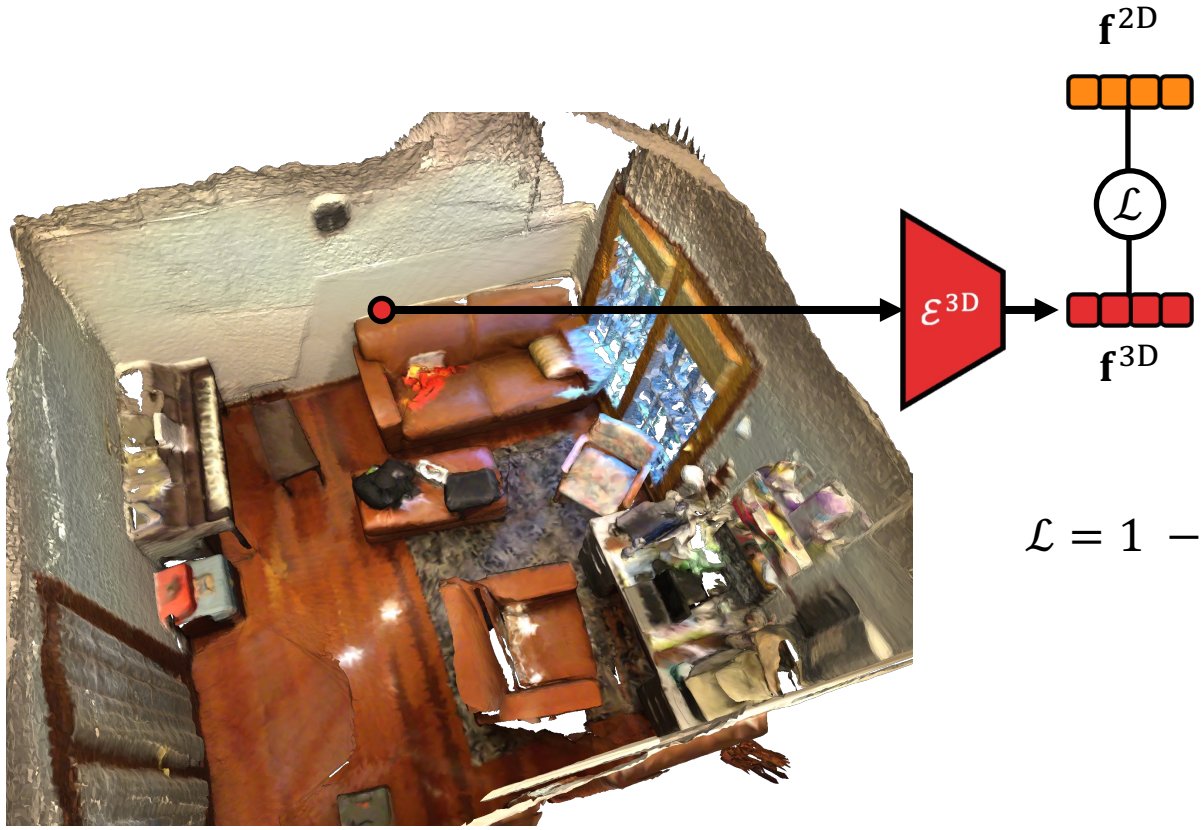
3D Geometry

Per-pixel Features
(visualize with PCA)

RGB Images

[1] Ghiasi, Gu, Cui, Lin: Scaling Open-Vocabulary Image Segmentation with Image-Level Labels. ECCV 2022
[2] Li, Weinberger, Belongie, Koltun, Ranftl: Language-driven Semantic Segmentation. ICLR 2022

Step 2: 3D Distillation

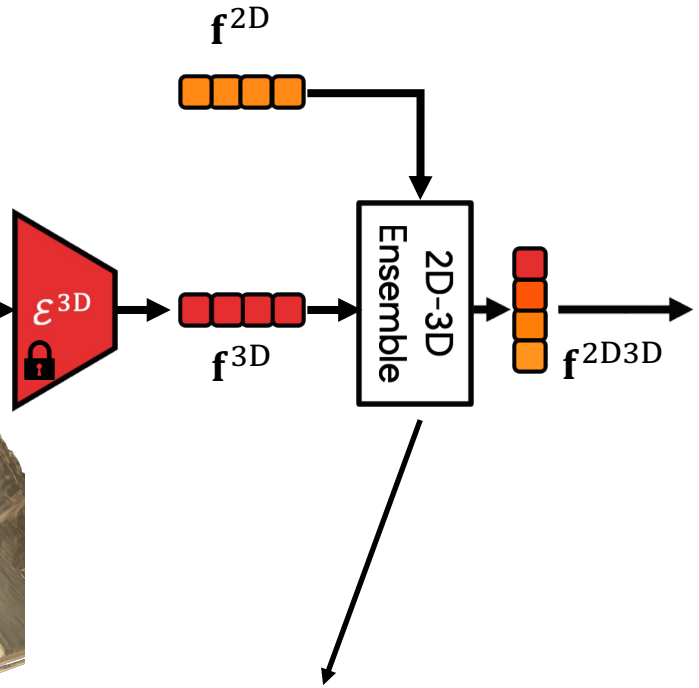


3D Geometry

$$\mathcal{L} = 1 - \cos(\mathbf{f}^{2D} - \mathbf{f}^{3D})$$

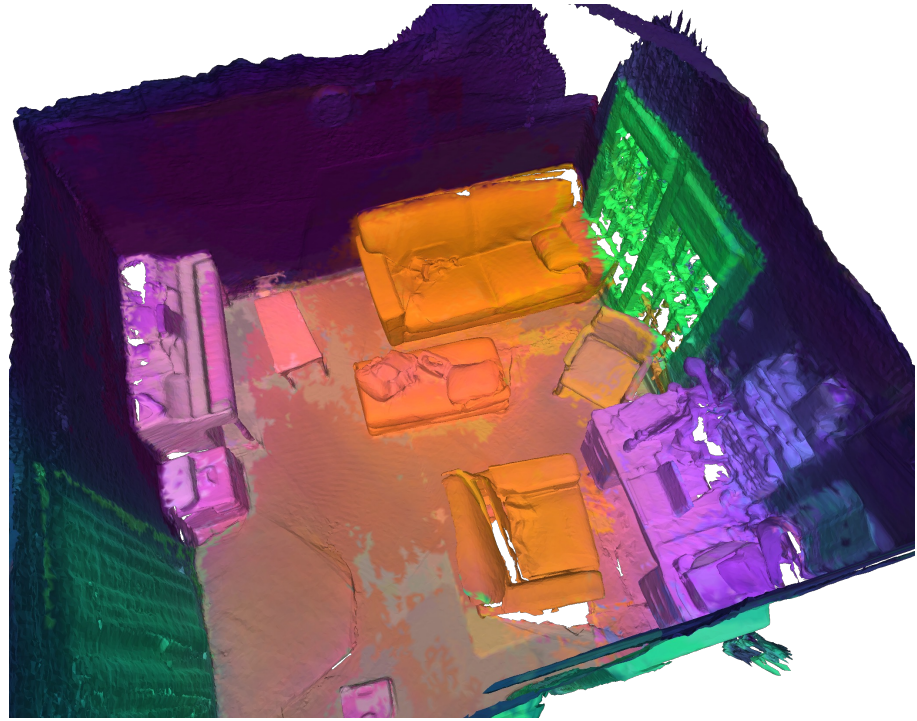
Step 3: 2D-3D Ensemble

3D Geometry



$$\mathbf{s}_n^{2D} = \cos(\mathbf{f}^{2D}, \mathbf{t}_n)$$
$$\mathbf{s}_n^{3D} = \cos(\mathbf{f}^{3D}, \mathbf{t}_n)$$

Choose the feature with
the highest max score among all prompts



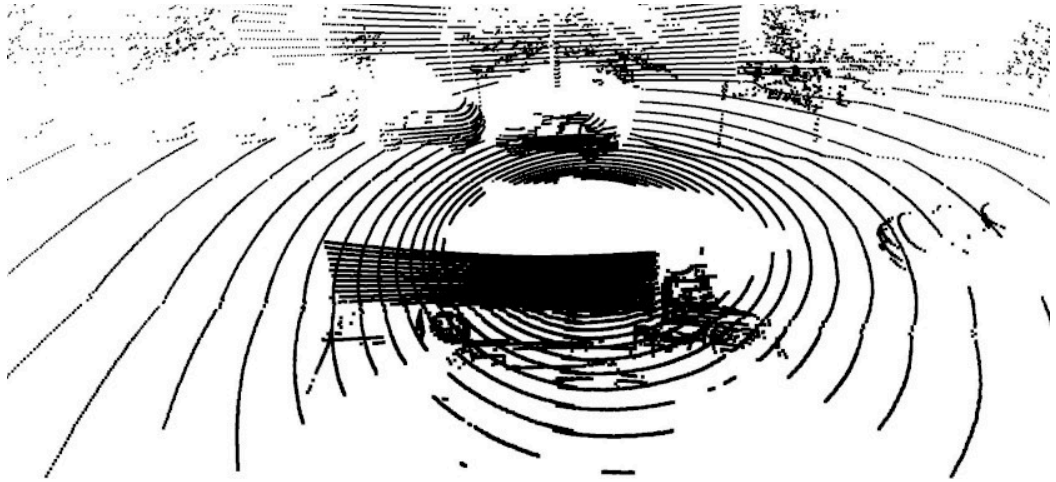
2D-3D Ensemble Features
(visualize with PCA)

Open-Vocabulary, Zero-shot

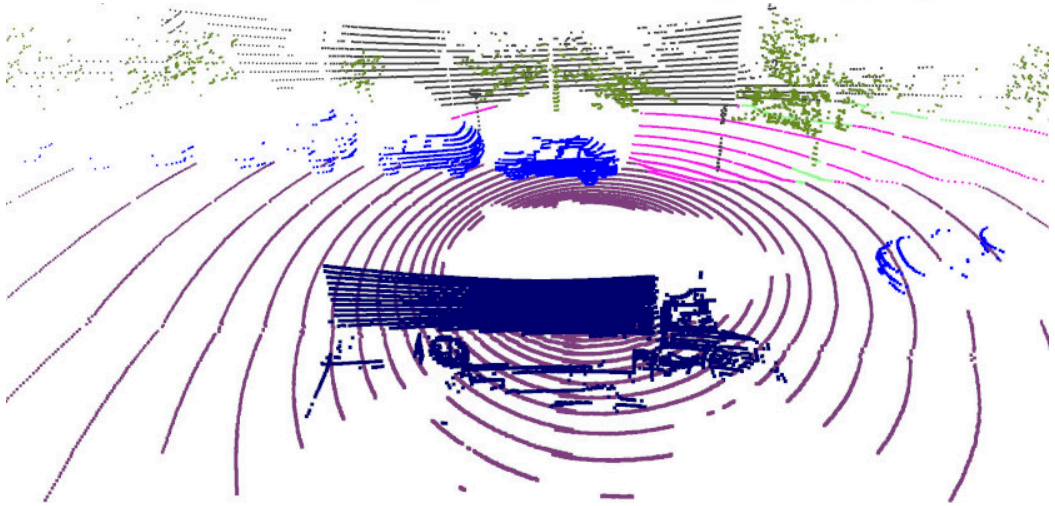
3D Semantic Segmentation

Results on nuScenes

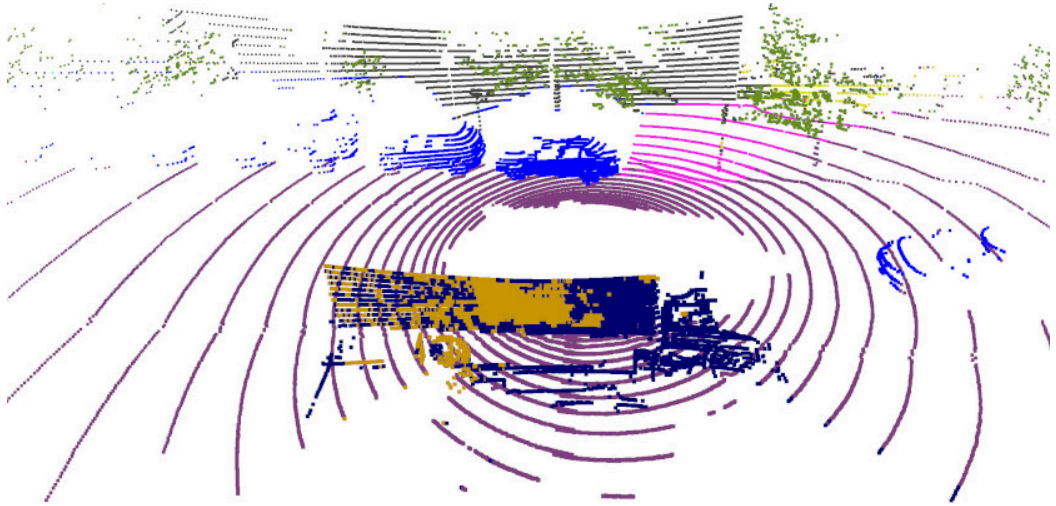
barrier car trailer truck road sidewalk terrain manmade vegetation mseg no mapping unlabeled



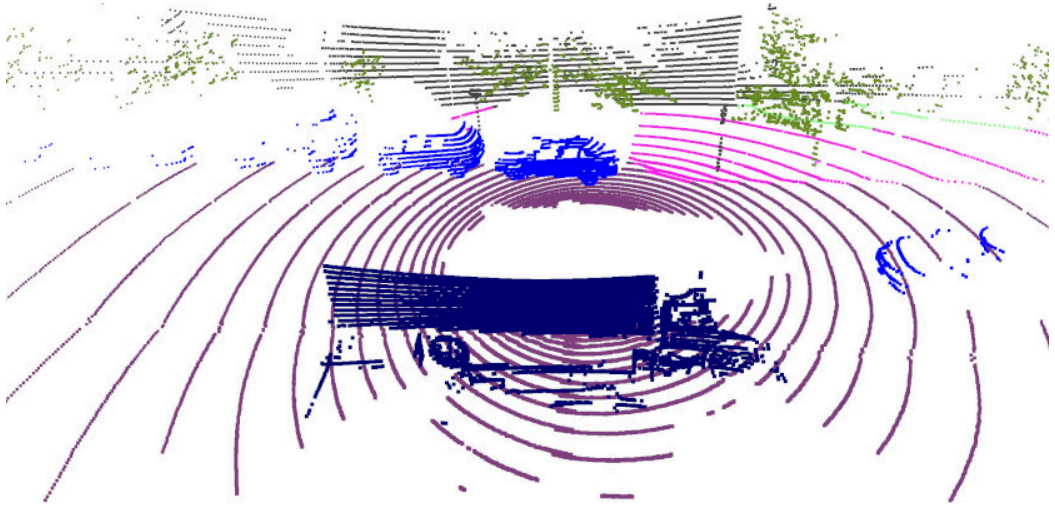
Input Lidar Points



MinkowskiNet (Fully supervised)



Ours (Zero-shot)



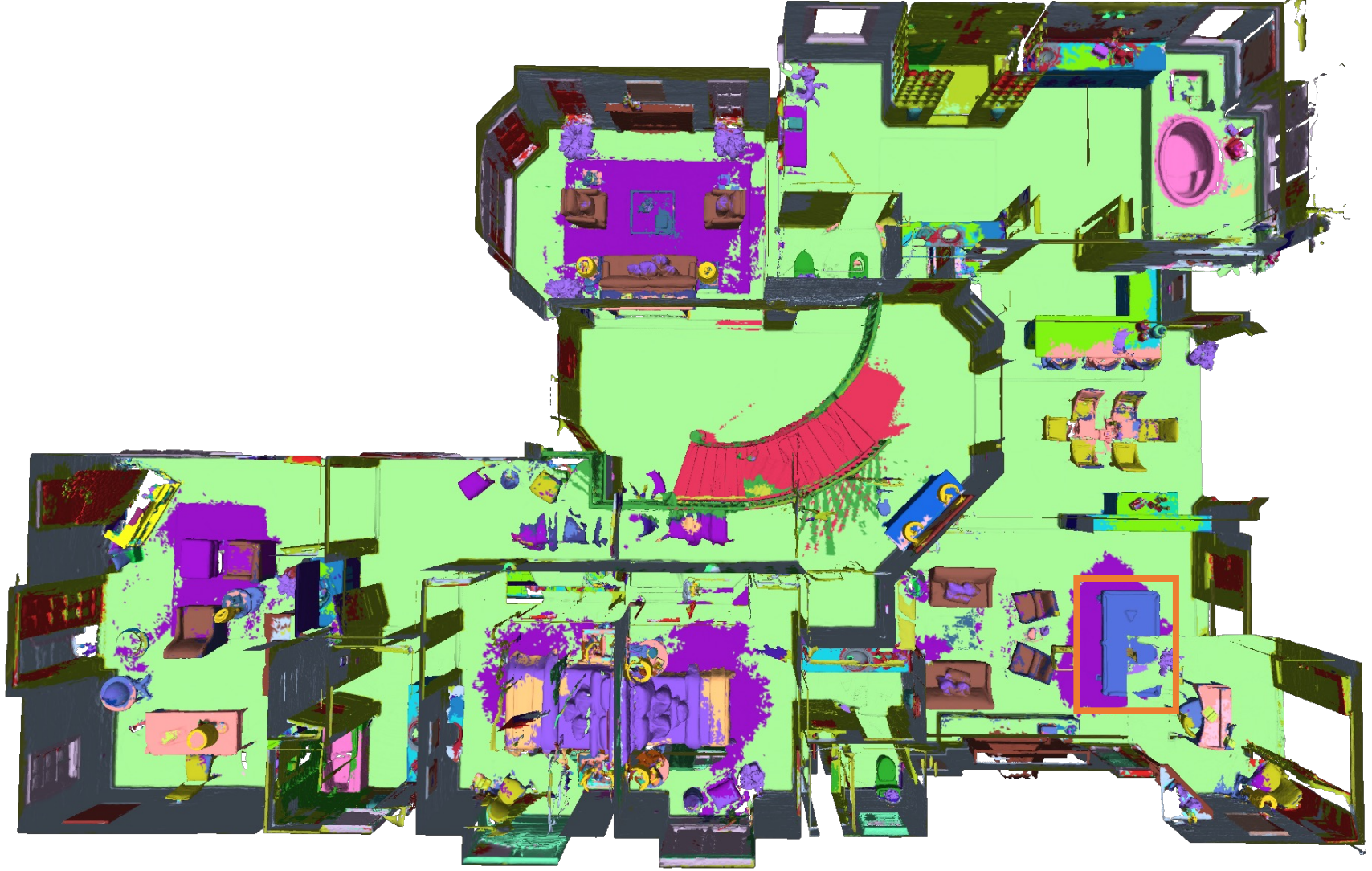
GT Label

Input 3D Geometry

Our Zero-shot 3D Segmentation

(20 classes)

wall floor cabinet bed chair sofa table door window bookshelf picture counter desk curtain refrigerator shower curtain toilet sink bathtub other

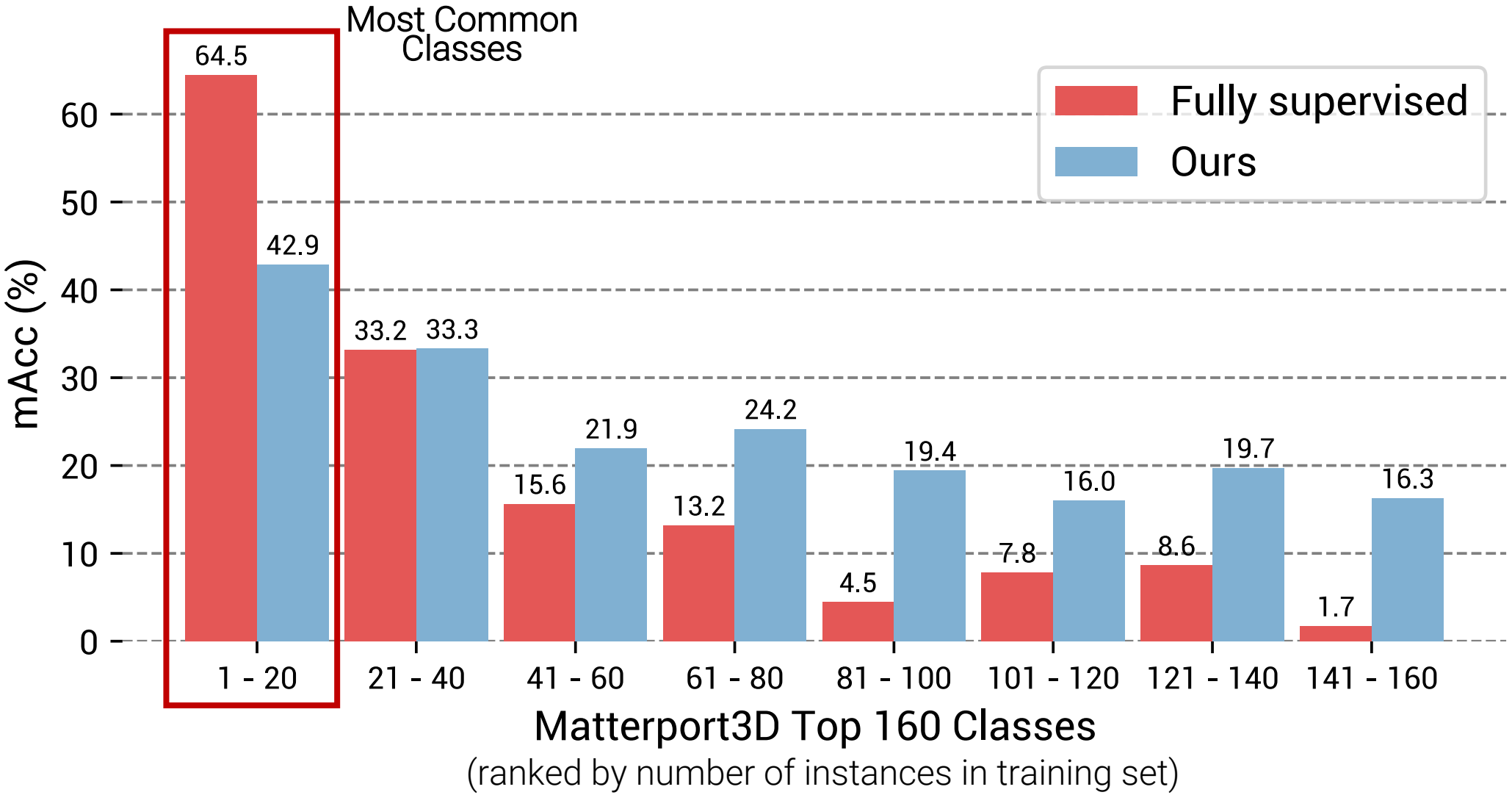


Our Zero-shot 3D Segmentation

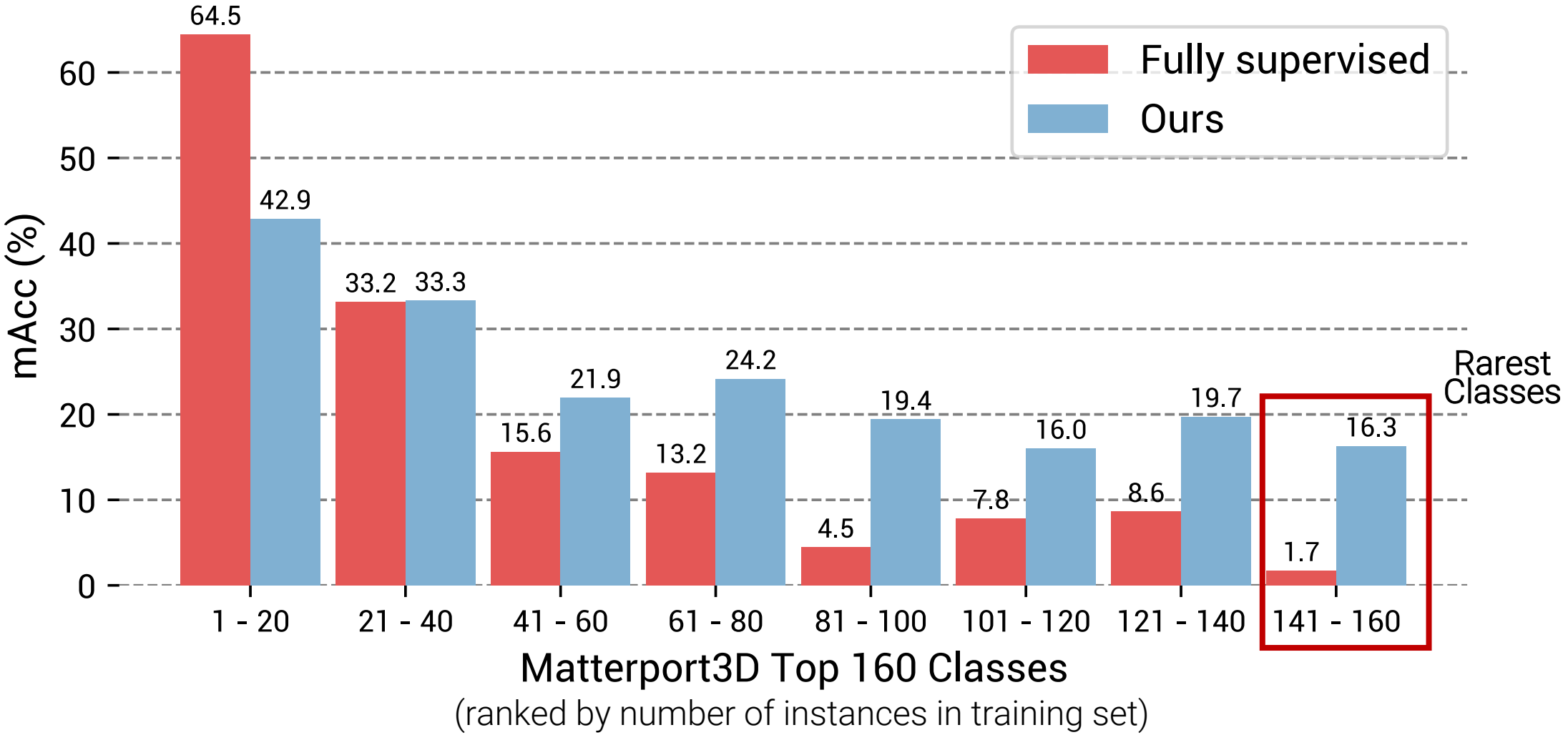
(160 classes)

wall, cabinet, bed, pot, bathtub, dresser, stand, clock, tissue box, furniture, soap, cup, hanger, urn, paper towel dispenser, toy
door, curtain, night stand, desk, book, rug, drawer, stove, tv stand, air conditioner, thermostat, ladder, candlestick, decorative plate, lamp shade, foot rest
ceiling, table, toilet, box, air vent, ottoman, container, washing machine, shoe, fire extinguisher, radiator, garage door, light, pool table, car, soap dish
floor, plant, column, coffee table, faucet, bottle, light switch, shower curtain, heater, curtain rod, kitchen island, piano, scale, jacket, toilet brush, cleaner
picture, mirror, banister, counter, photo, refridgerator, purse, bin, headboard, printer, paper towel, board, bag, bottle of soap, drum, computer
window, towel, stairs, bench, toilet paper, bookshelf, door way, chest, bucket, telephone, sheet, rope, display case, water cooler, whiteboard, knob
chair, sink, stool, garbage bin, fan, wardrobe, basket, microwave, candle, blanket, glass, ball, toilet paper holder, tea pot, range hood, paper
pillow, shelves, vase, fireplace, railing, pipe, chandelier, blinds, flower pot, handle, dishwasher, excercise equipment, tray, stuffed animal, candelabra, projector

Comparison


Most Common Classes
Fully supervised
Ours
mAcc (%)
0
10
20
30
40
50
60
64.5
42.9
33.2
33.3
15.6
21.9
13.2
24.2
4.5
19.4
7.8
16.0
8.6
19.7
1.7
16.3
1 - 20
21 - 40
41 - 60
61 - 80
81 - 100
101 - 120
121 - 140
141 - 160
Matterport3D Top 160 Classes
(ranked by number of instances in training set)

Comparison


Fully supervised
Ours
mAcc (%)
0
10
20
30
40
50
60
64.5
42.9
33.2
33.3
15.6
21.9
13.2
24.2
4.5
19.4
7.8
16.0
8.6
19.7
1.7
16.3
1 - 20
21 - 40
41 - 60
61 - 80
81 - 100
101 - 120
121 - 140
141 - 160
Rarest Classes
Matterport3D Top 160 Classes
(ranked by number of instances in training set)

Ablation

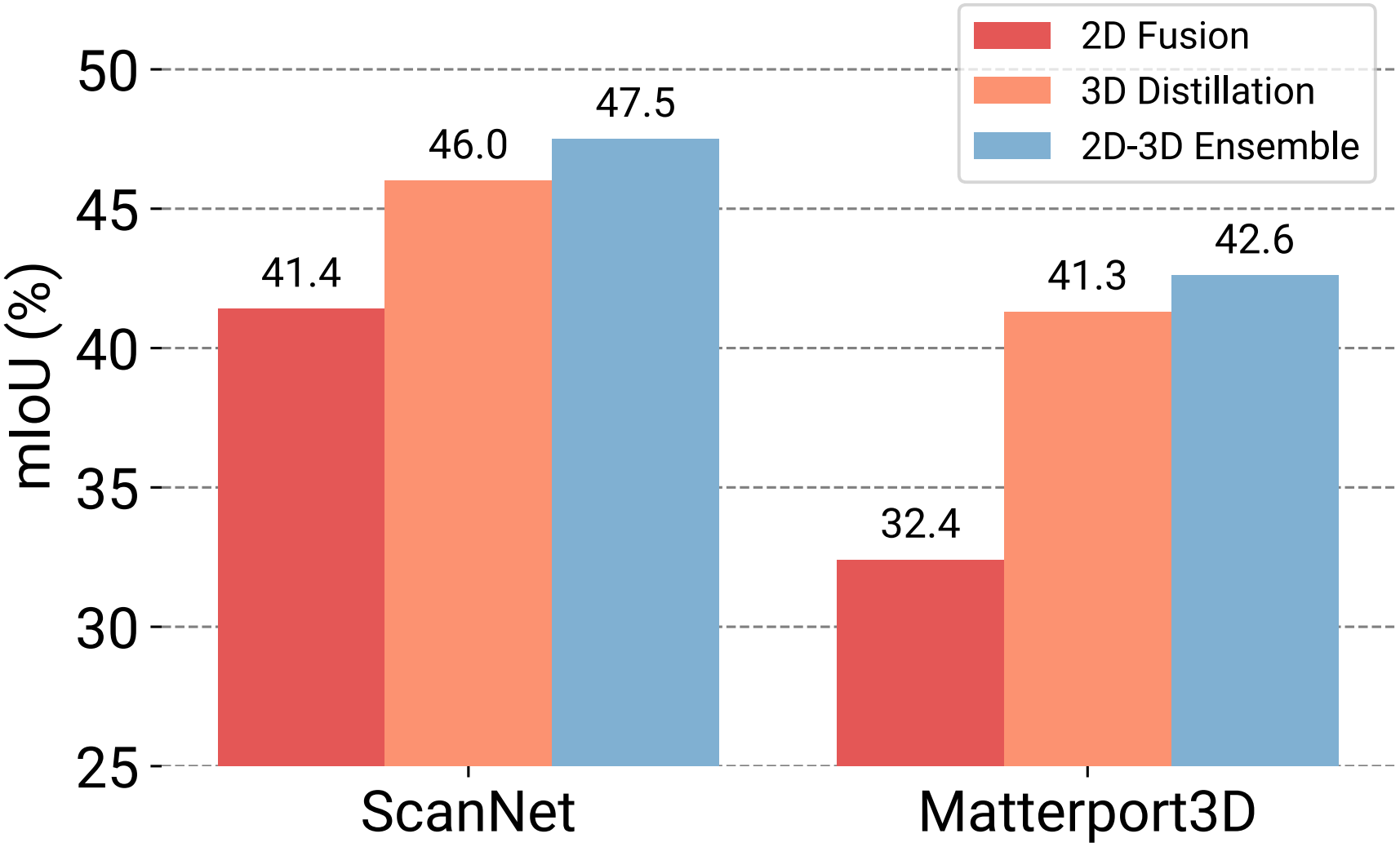

2D Fusion
3D Distillation
2D-3D Ensemble
mIoU (%)
50
45
40
35
30
25
41.4
46.0
47.5
32.4
41.3
42.6
ScanNet
Matterport3D

Image-based 3D Scene Query

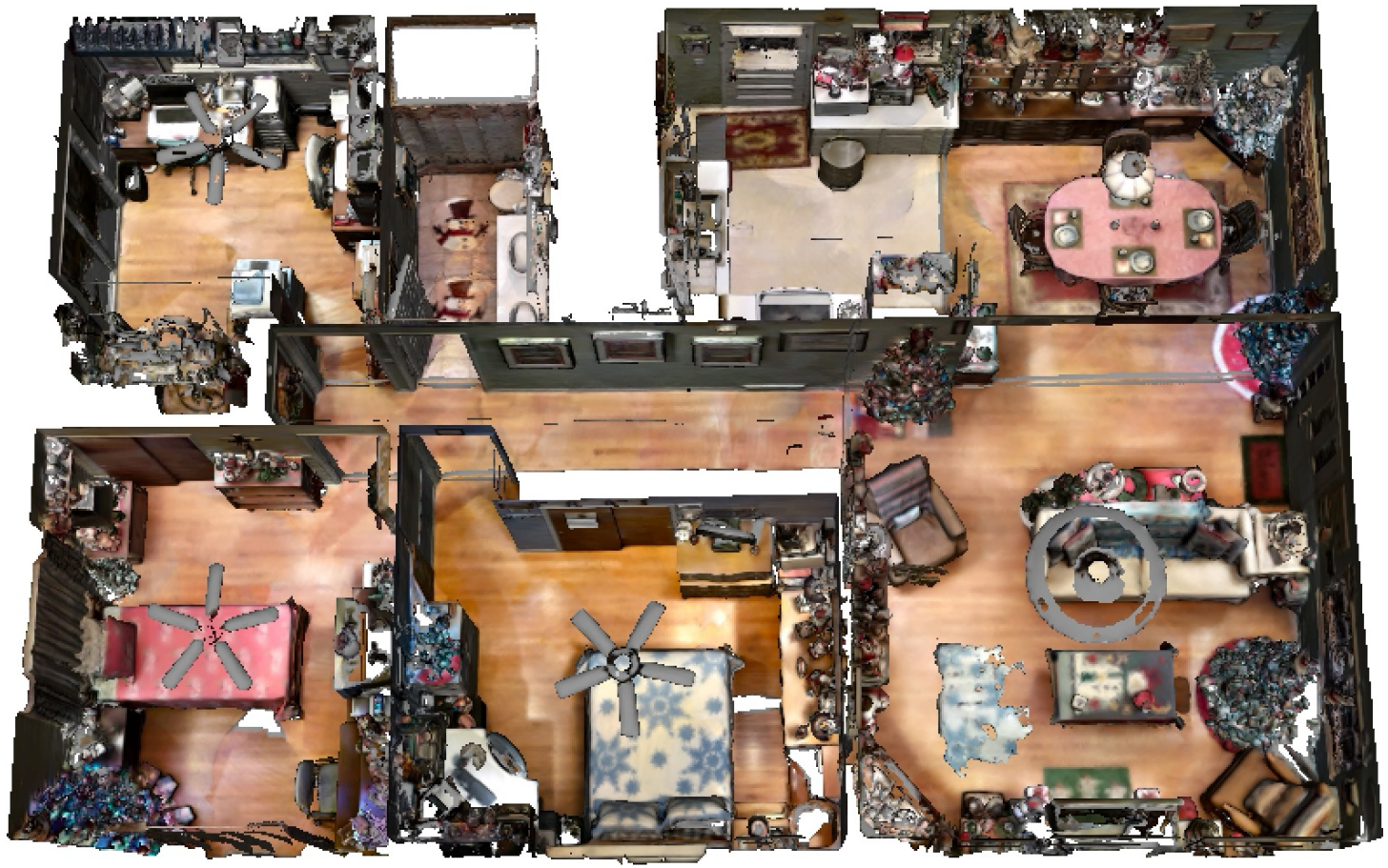
Image Queries

Given 3D Geometry

Interactive Demo

Open-vocabulary 3D Scene Exploration

Text queries:



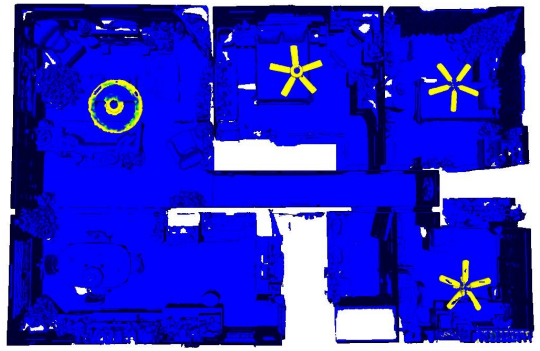
Take-home Message

- We enable a **wide range of applications** by open-vocabulary queries
- This can hopefully influence how people train 3D scene understanding systems in the future
- Our real-time demo already shows the **possibility to directly apply to AR/VR**

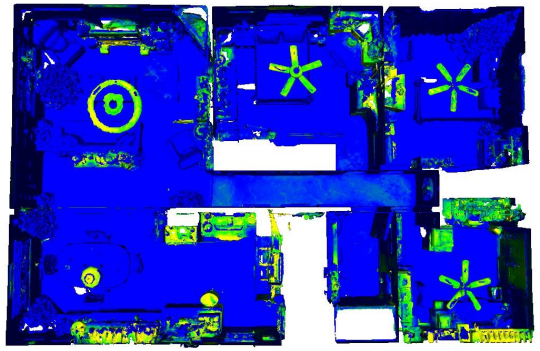
OpenScene

3D Scene Understanding with Open Vocabularies

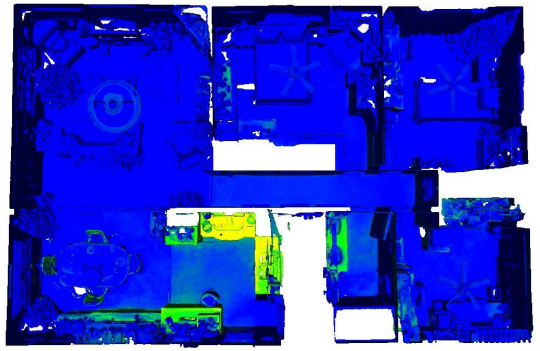
Input 3D Point Cloud



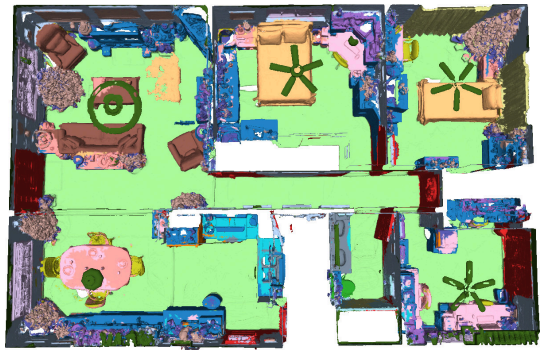
“fan” - Object



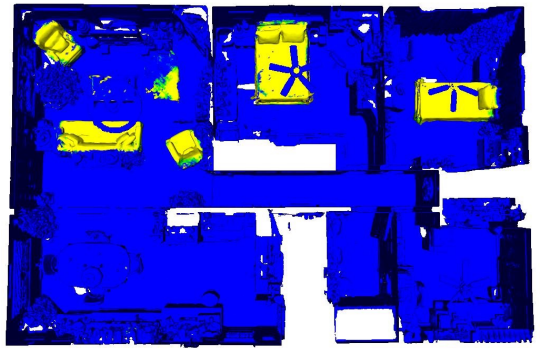
“made of metal” - Material



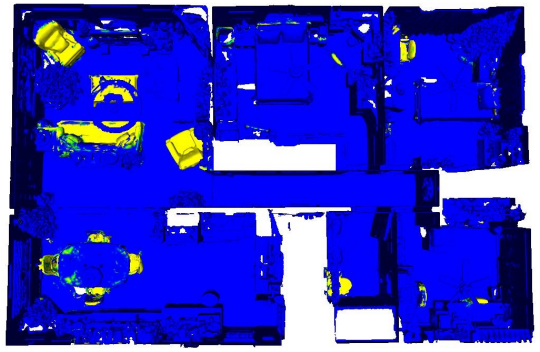
“kitchen” – Room Type



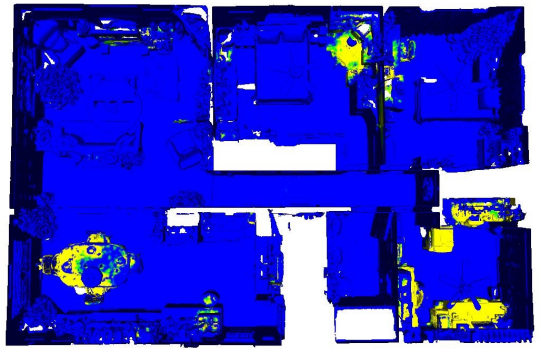
Zero-shot Semantic Segmentation



“anything soft” - Property



“where to sit” - Affordance



“work” - Activity

pengsongyou.github.io/openscene