

OpenScene

3D Scene Understanding with Open Vocabularies

Input 3D Point Cloud

“fan” - Object

“made of metal” - Material

Songyou Peng

ETH Zurich and Max Planck Institute for Intelligent Systems

ETH zürich

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS



Zero-shot Semantic Segmentation

“anyth” - Property

“where to sit” - Affordance

June 09, 2023

Who Am I?

- 4th Year PhD Student
 - Marc Pollefeys
 - Andreas Geiger
- Internships during PhD
 - 2021: Michael Zollhoefer
 - 2022: Tom Funkhouser
- Open to collaborate, feel free to reach out!

ETH zürich

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

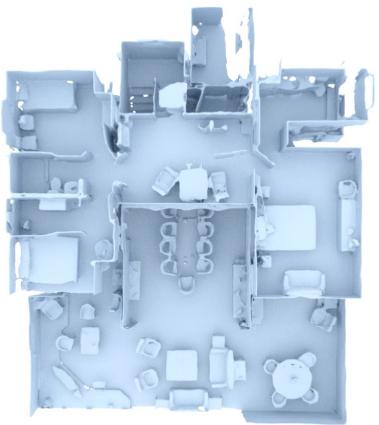


Meta
Google Research

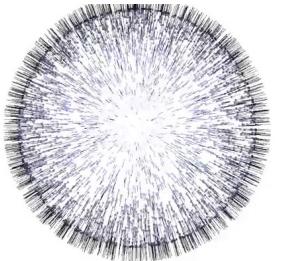


pengsongyou.github.io

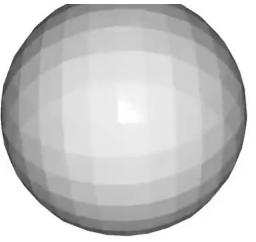
My PhD Topics: Neural Scene Representations for 3D reconstruction and 3D scene understanding



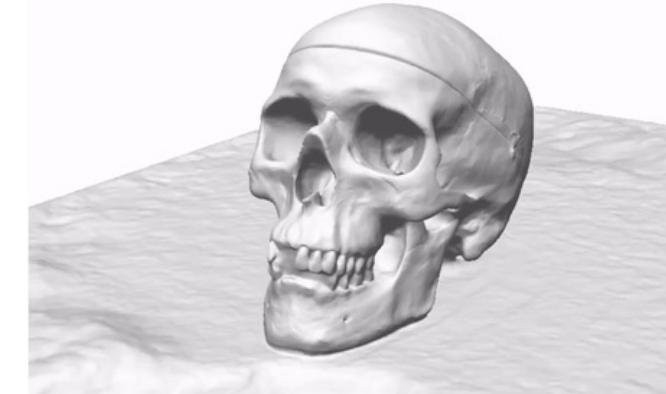
Convolutional Occupancy Nets
ECCV 2020 (Spotlight)



Shape As Points
NeurIPS 2021 (Oral)



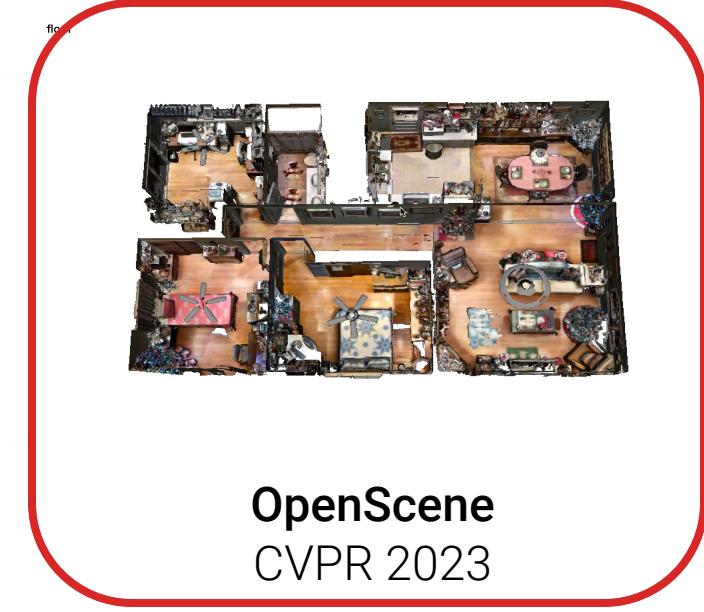
KiloNeRF
ICCV 2021



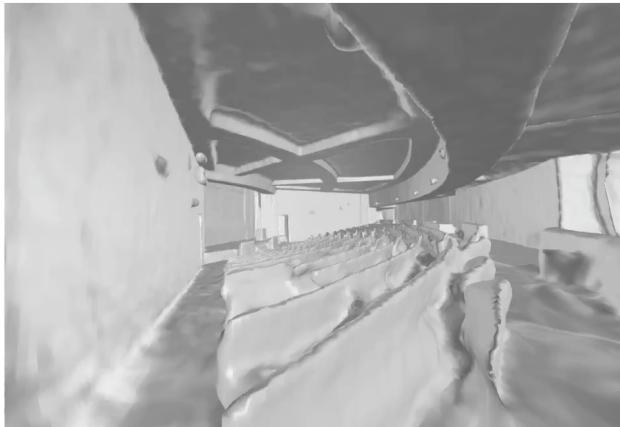
Ours
UNISURF
ICCV 2021 (Oral)



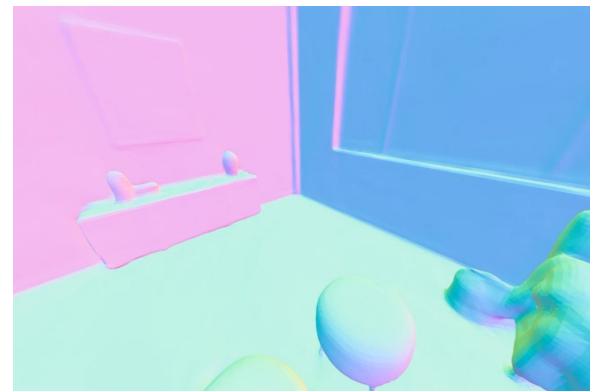
NICE-SLAM
CVPR 2022



OpenScene
CVPR 2023



Ours
MonoSDF
NeurIPS 2022



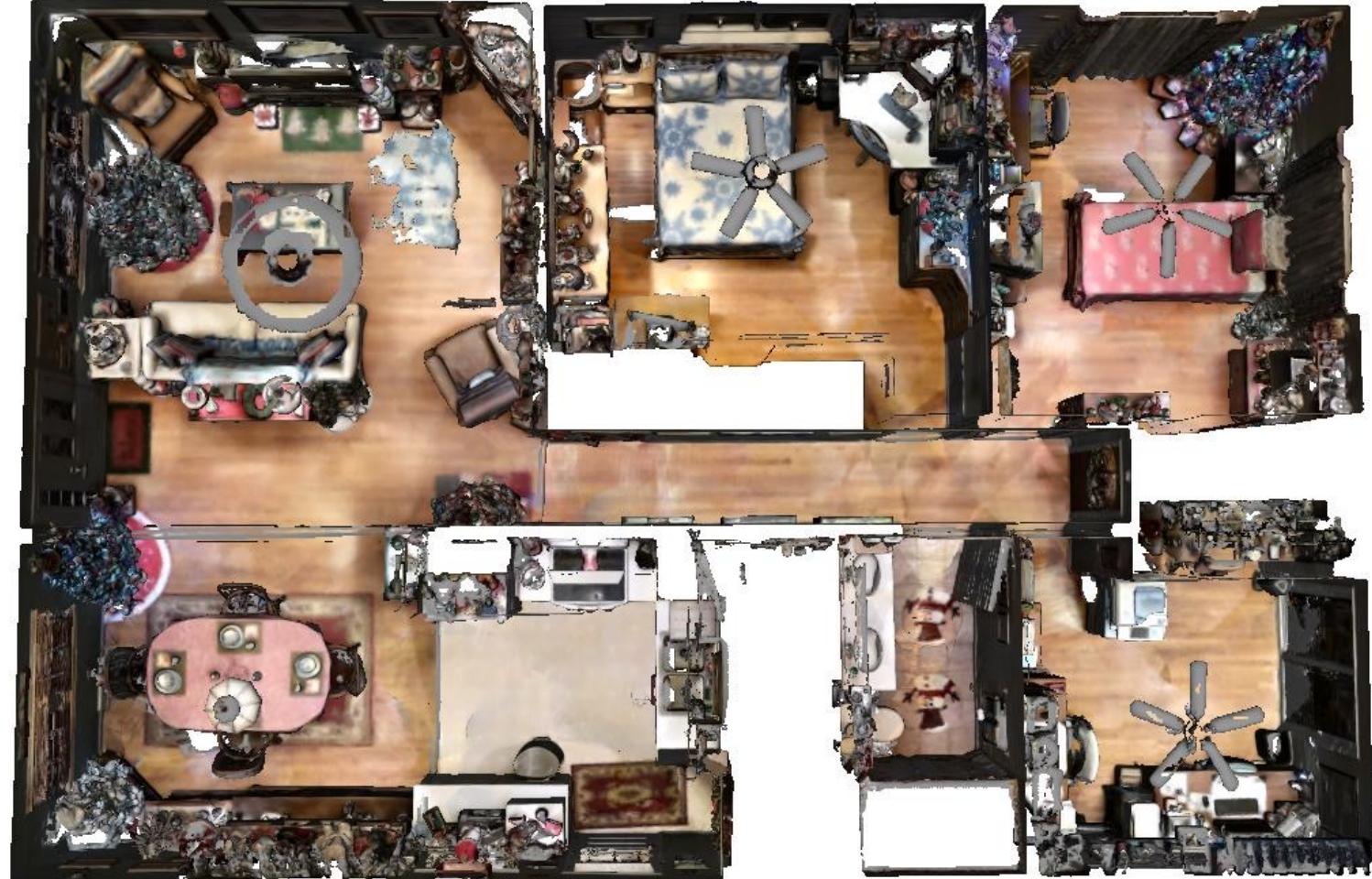
NICER-SLAM
arXiv 2023



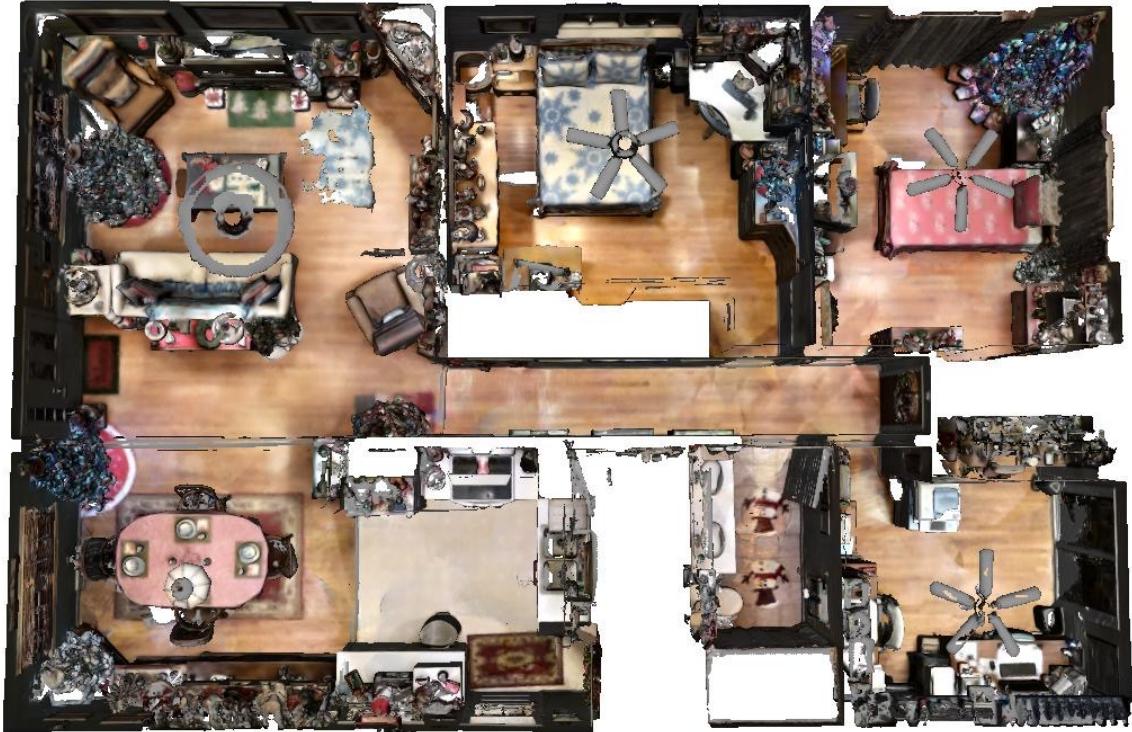
⋮



Input Images



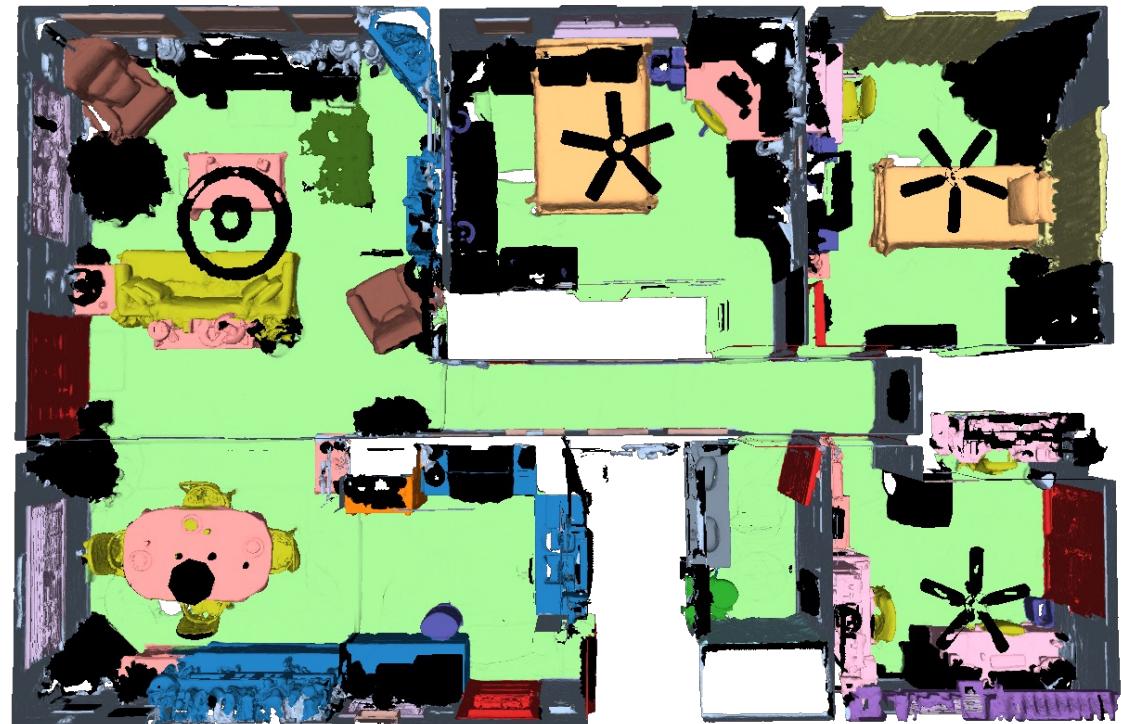
3D Reconstruction



Input 3D Geometry

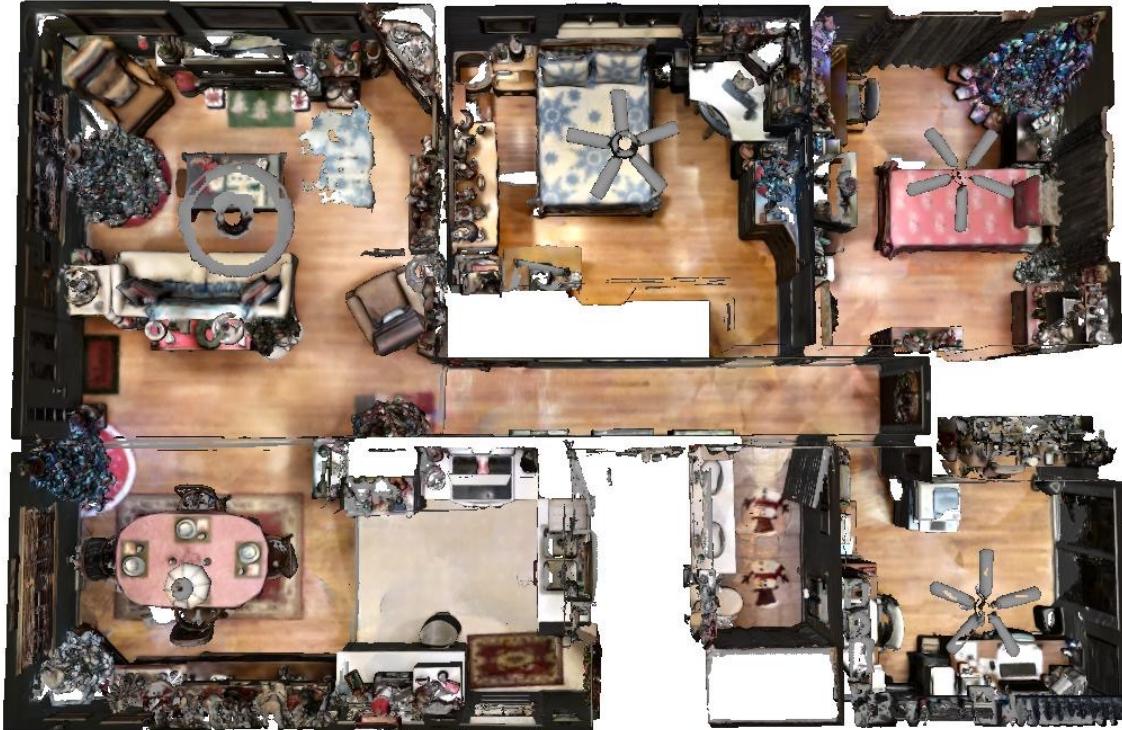
Legend:

- wall
- floor
- cabinet
- bed
- chair
- sofa
- table
- door
- window
- counter
- curtain
- toilet
- sink
- bathtub
- other
- unlabeled



Traditional Semantic Segmentation

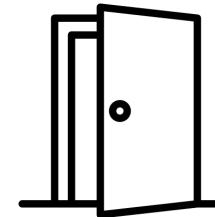
Only train and test on a few common classes



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



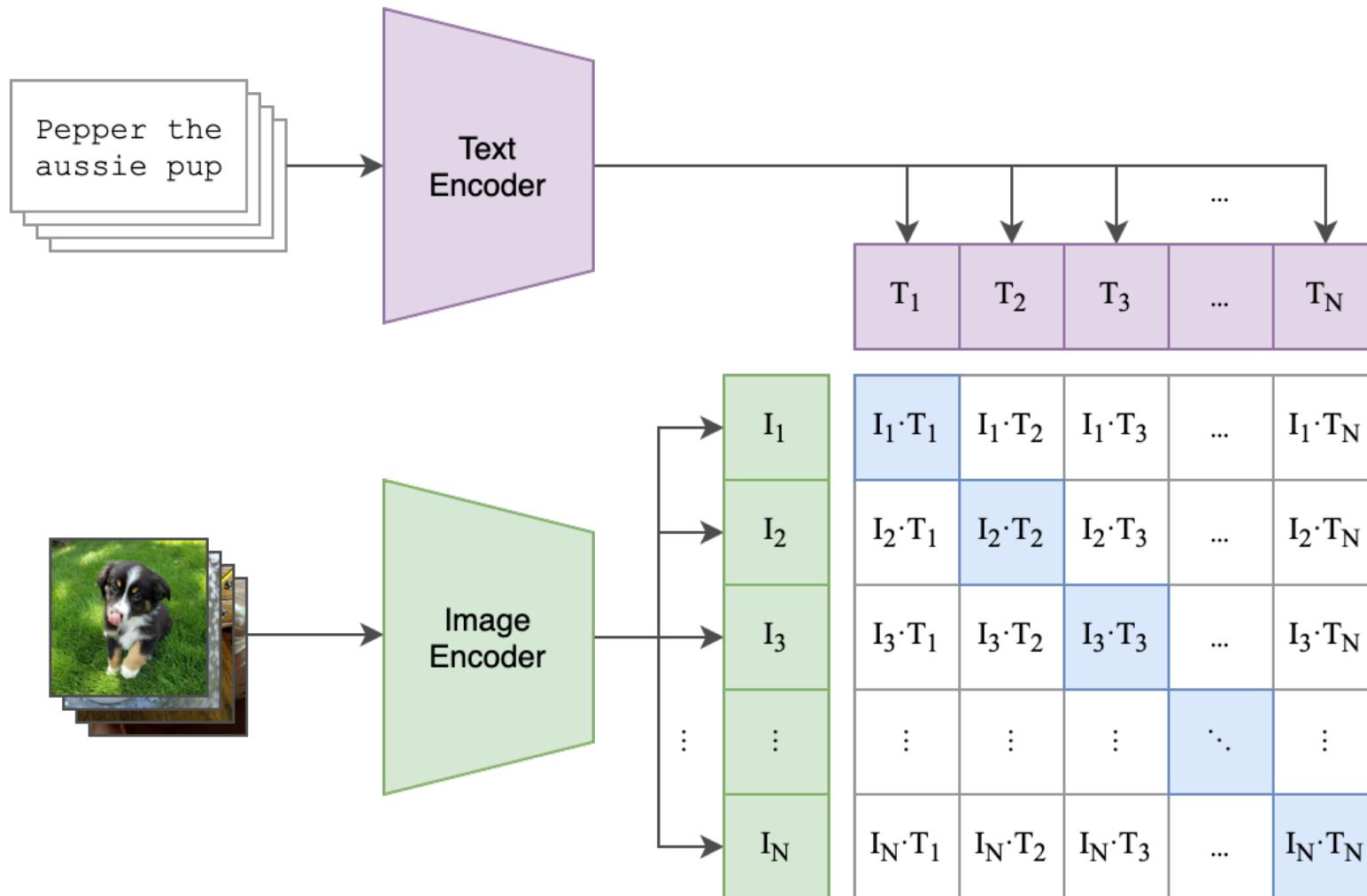
Marc Pollefeys



Tom Funkhouser



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

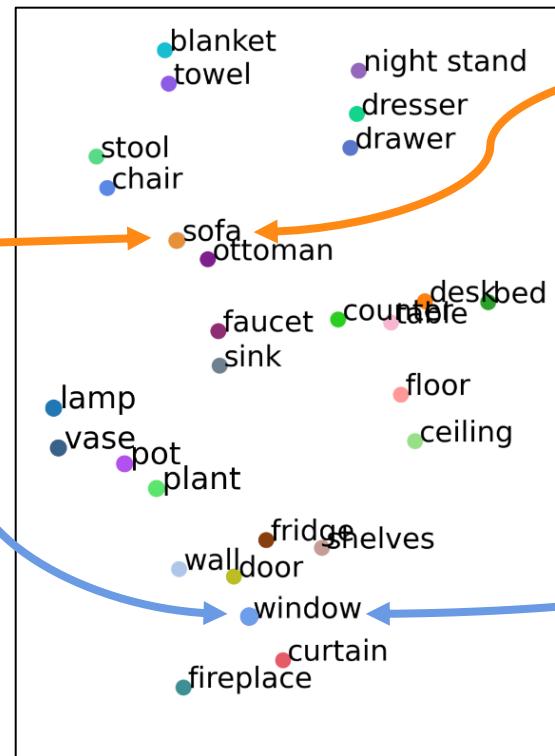


CLIP: Contrastive Language-Image Pre-Training

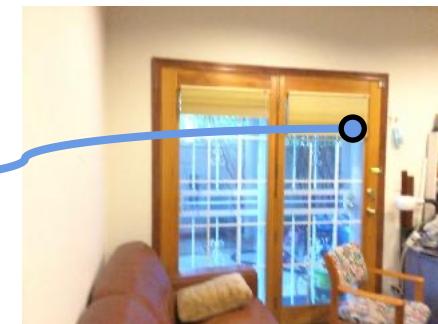
Key Idea: Co-embed 3D features with CLIP features



3D Geometry



CLIP Text Features
(visualize with T-SNE)



RGB Images

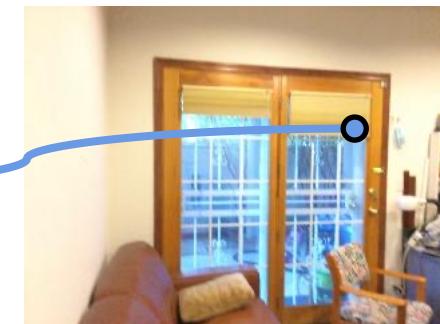
Key Idea: Co-embed 3D features with CLIP features



3D Geometry



CLIP Text Features
(visualize with T-SNE)

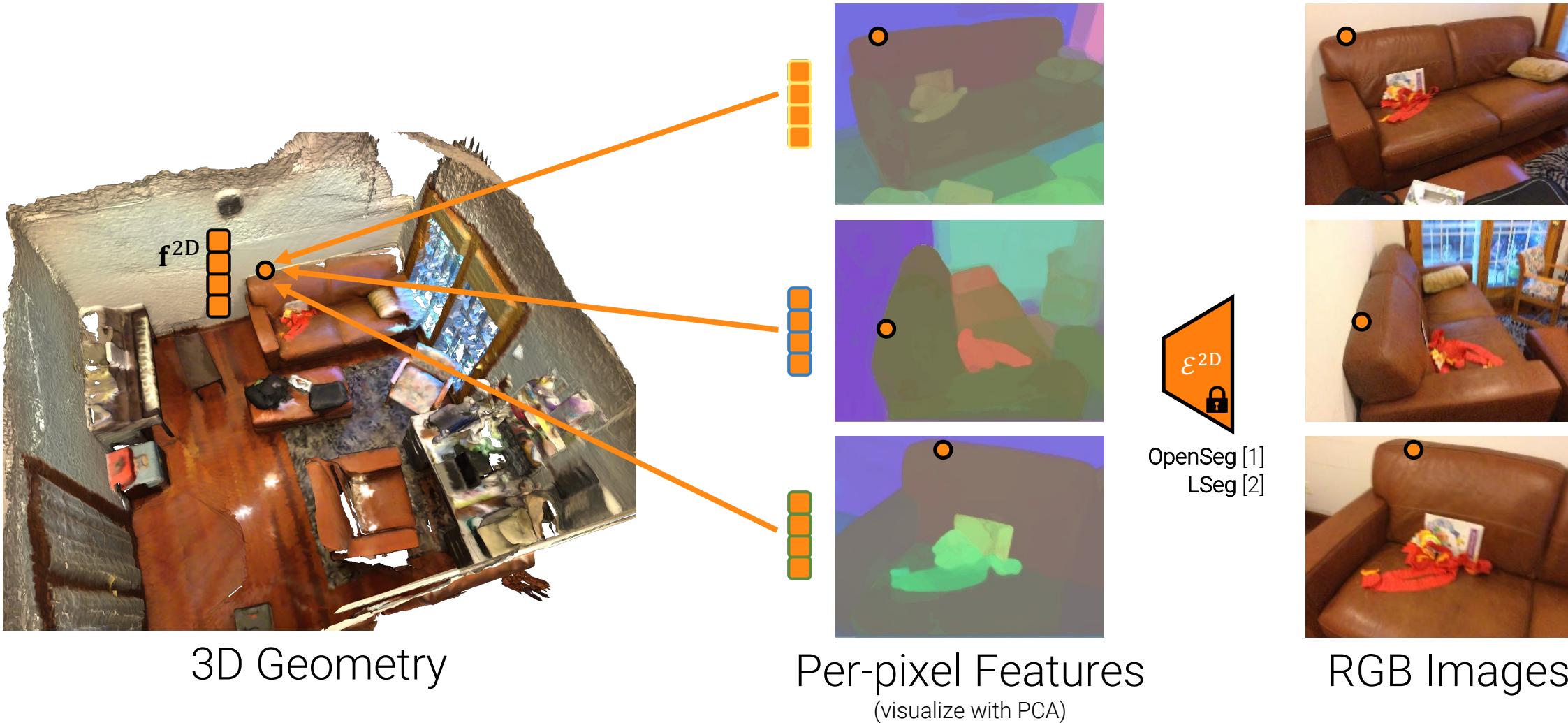


RGB Images

Note: bold word embeddings are approximate

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

Step 1: Multi-view Feature Fusion



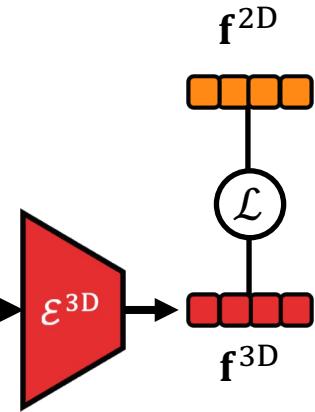
[1] Ghiasi, Gu, Cui, Lin: [Scaling Open-Vocabulary Image Segmentation with Image-Level Labels](#). ECCV 2022

[2] Li, Weinberger, Belongie, Koltun, Ranftl: [Language-driven Semantic Segmentation](#). ICLR 2022

Step 2: 3D Distillation



3D Geometry

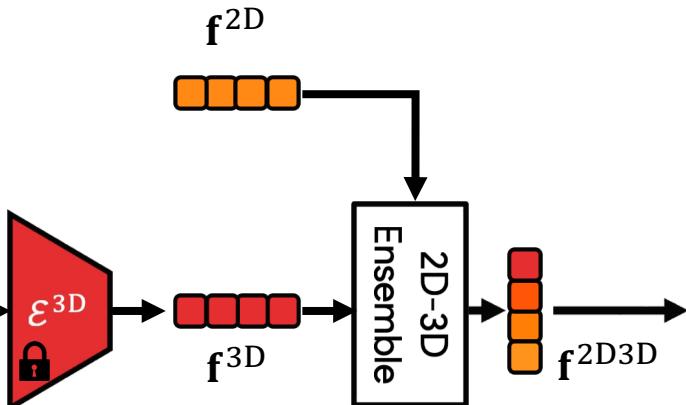


$$\mathcal{L} = 1 - \cos(f^{2D} - f^{3D})$$

Step 3: 2D-3D Ensemble



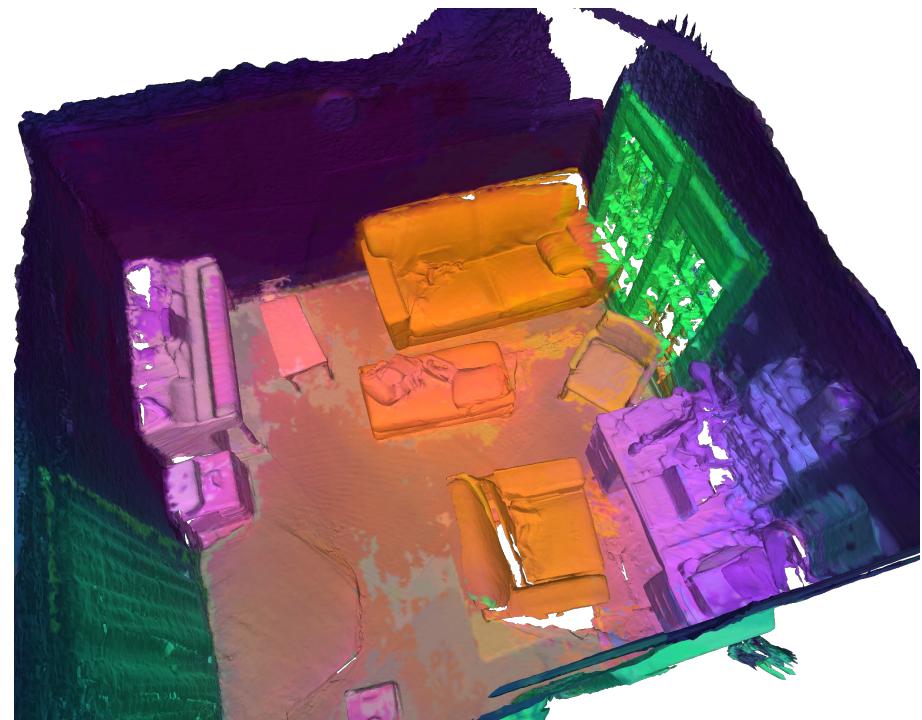
3D Geometry



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

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

Choose the feature with
the highest max score among all prompts

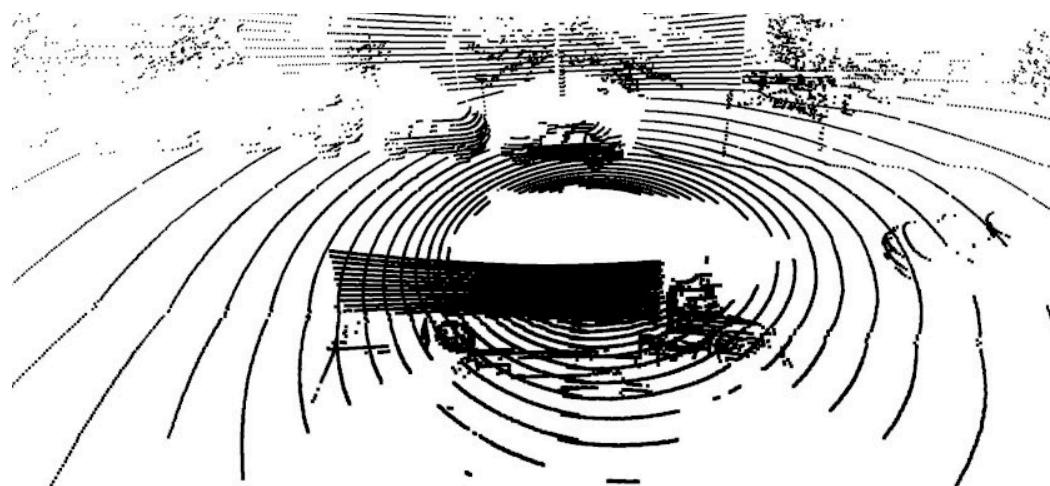


2D-3D Ensemble Features
(visualize with PCA)

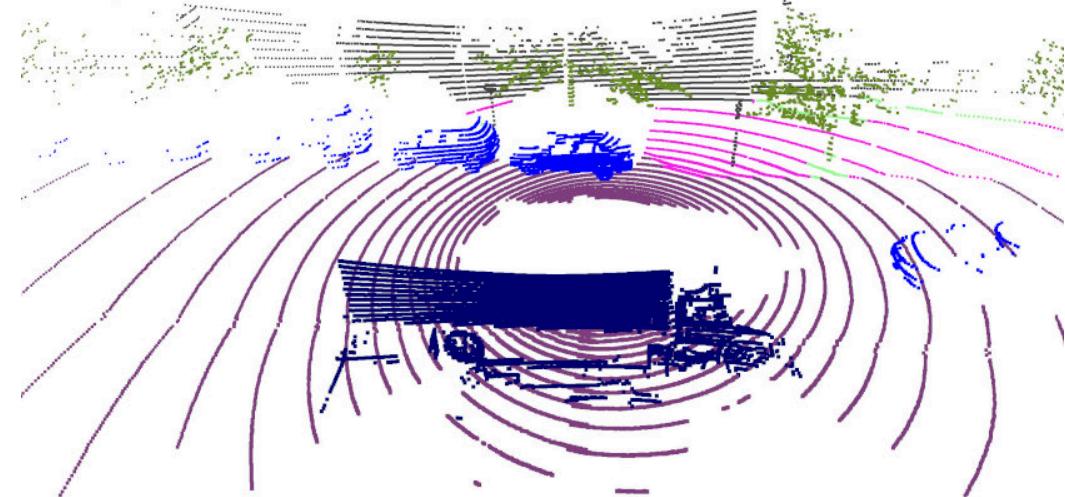
Open-Vocabulary, Zero-shot 3D Semantic Segmentation

Results on nuScenes

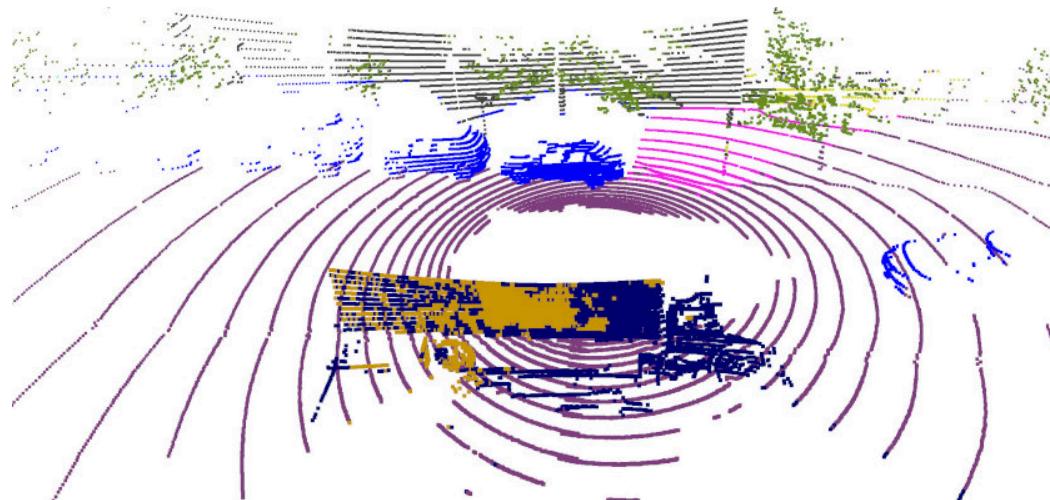
barrier car trailer truck road sidewalk terrain manmade vegetation mseg no mapping unlabeled



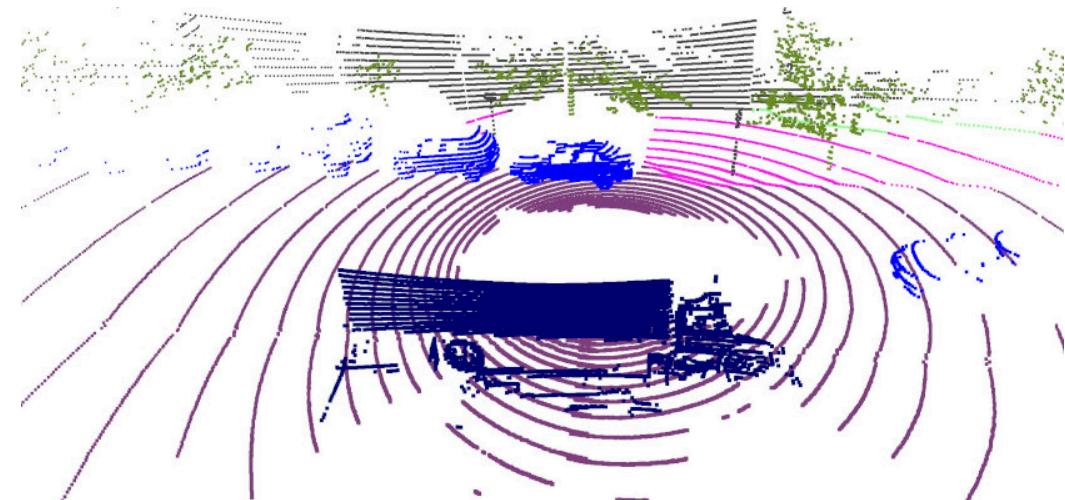
Input Lidar Points



MinkowskiNet (Fully supervised)



Ours (Zero-shot)



GT Label



Input 3D Geometry



Our Zero-shot 3D Segmentation
(20 classes)

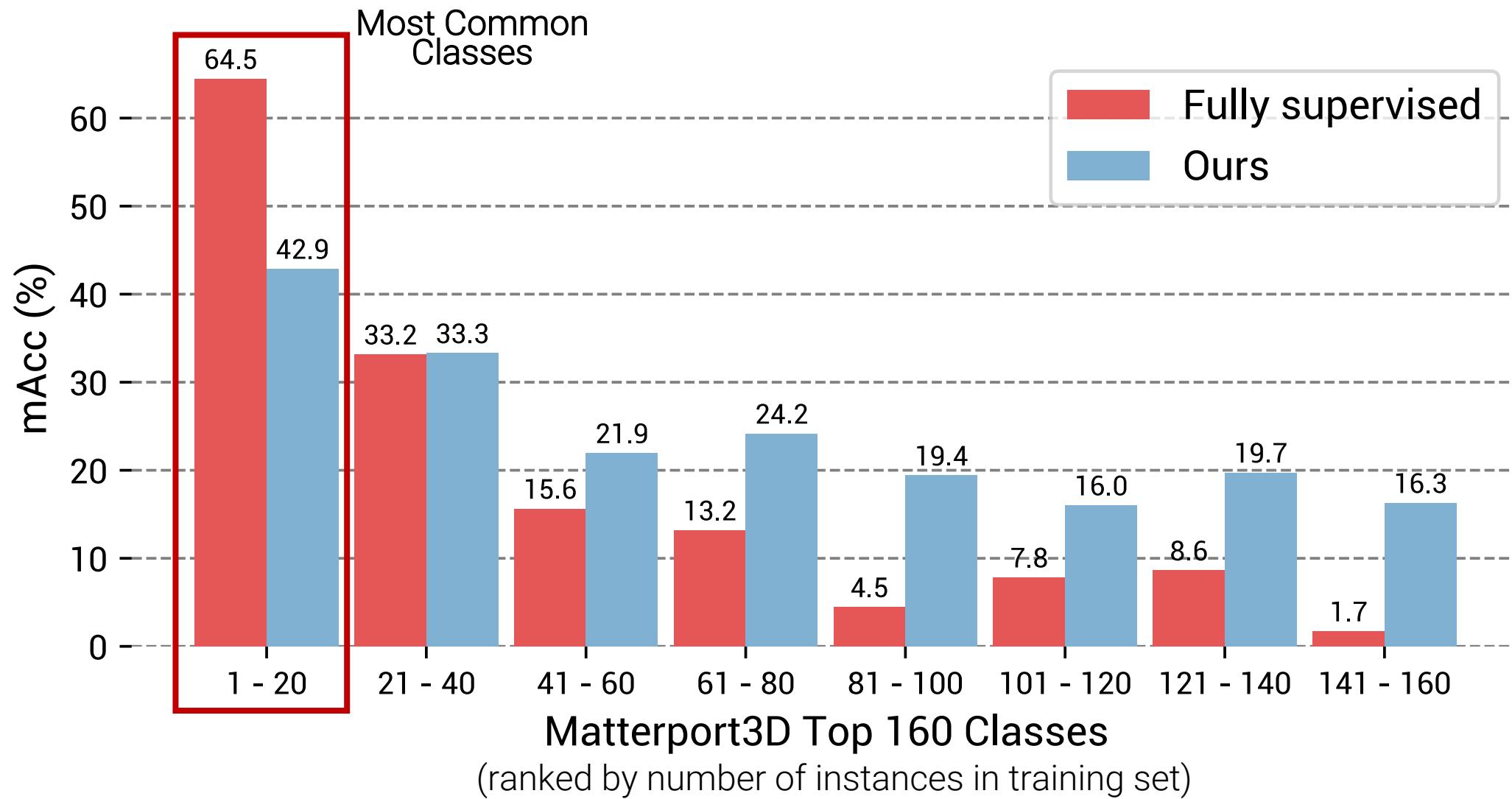
■ wall ■ floor ■ cabinet ■ bed ■ chair ■ sofa ■ table ■ door ■ window ■ bookshelf ■ picture ■ counter ■ desk ■ curtain ■ refrigerator ■ shower curtain ■ toilet ■ sink ■ bathtub ■ other



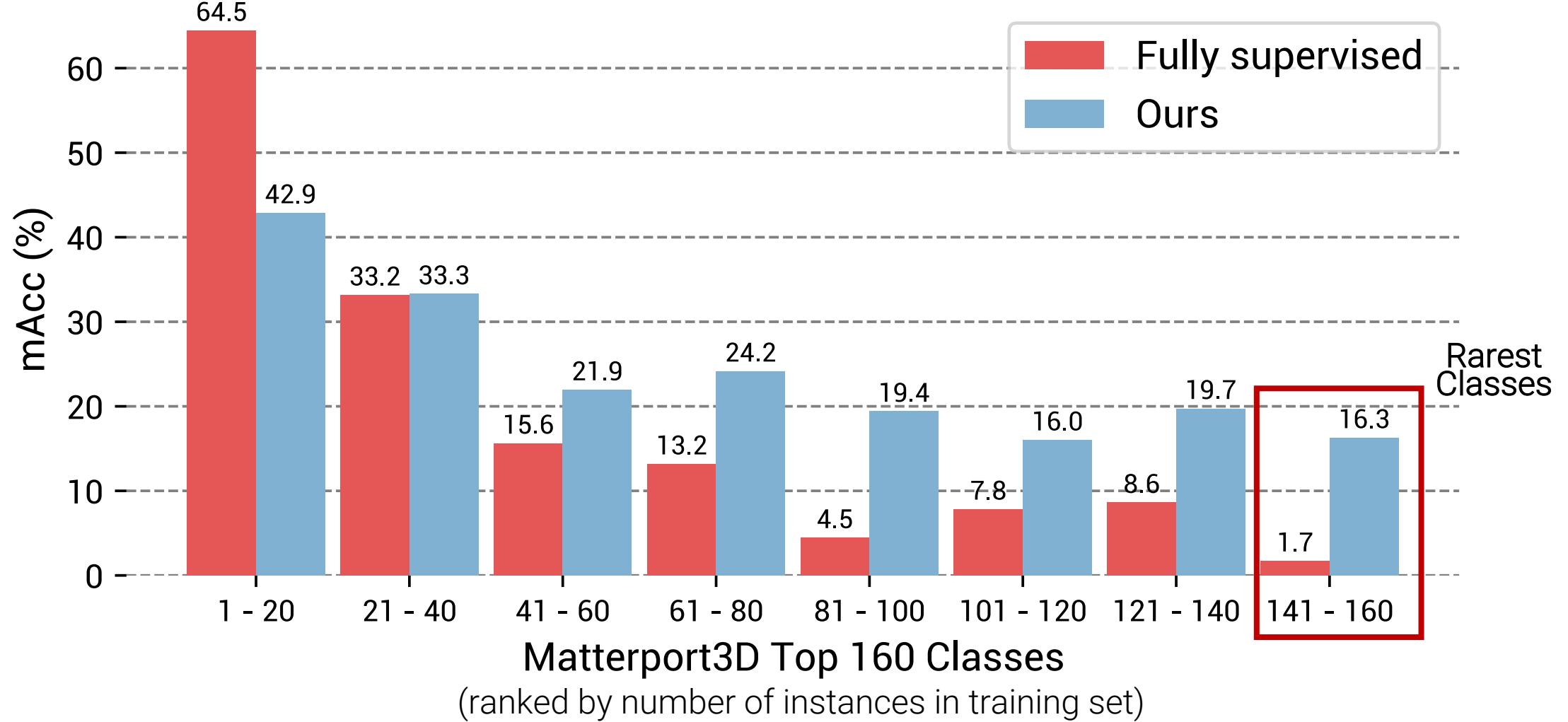
Our Zero-shot 3D Segmentation
(160 classes)

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

Comparison



Comparison



Ablation

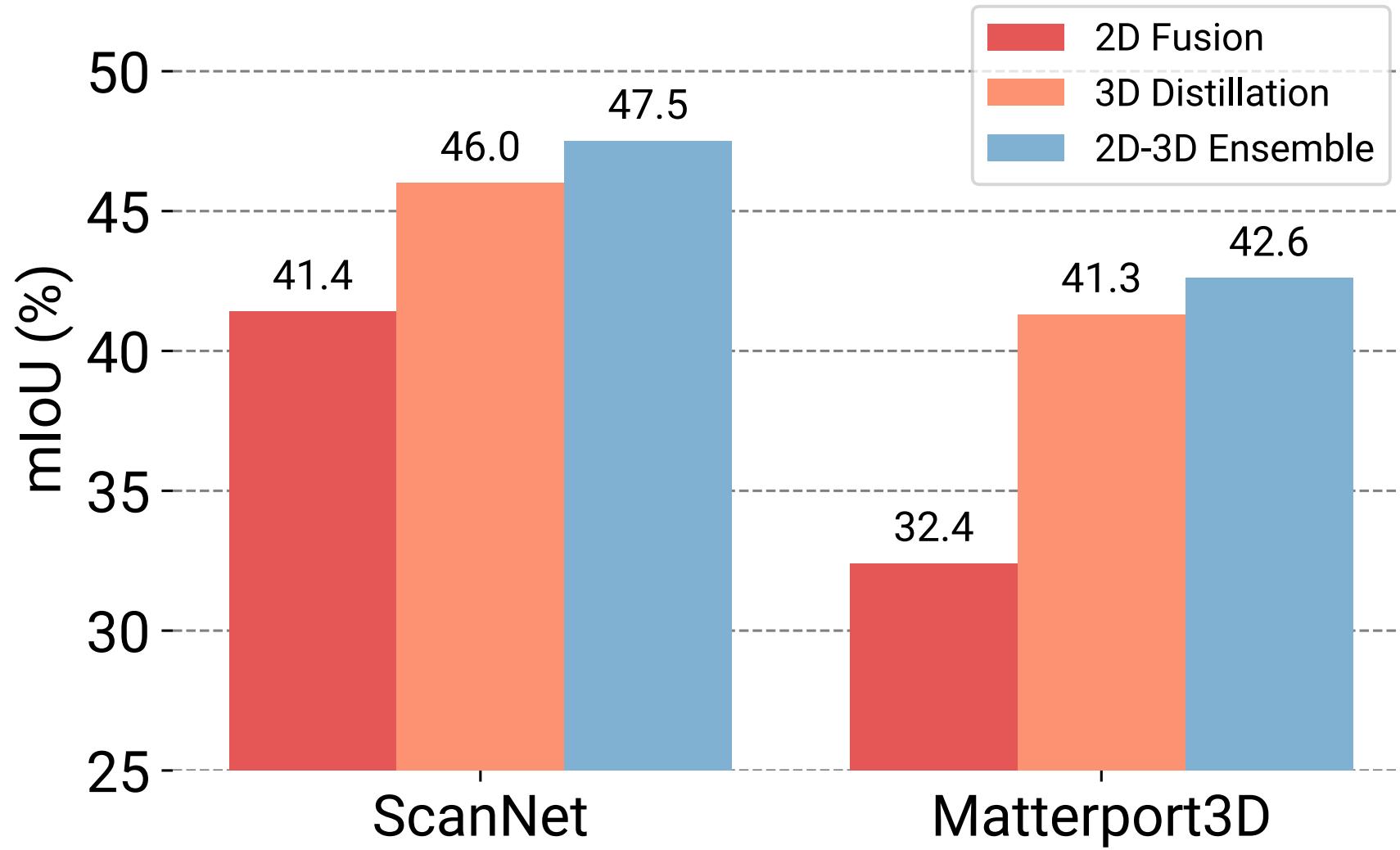


Image-based 3D Scene Query



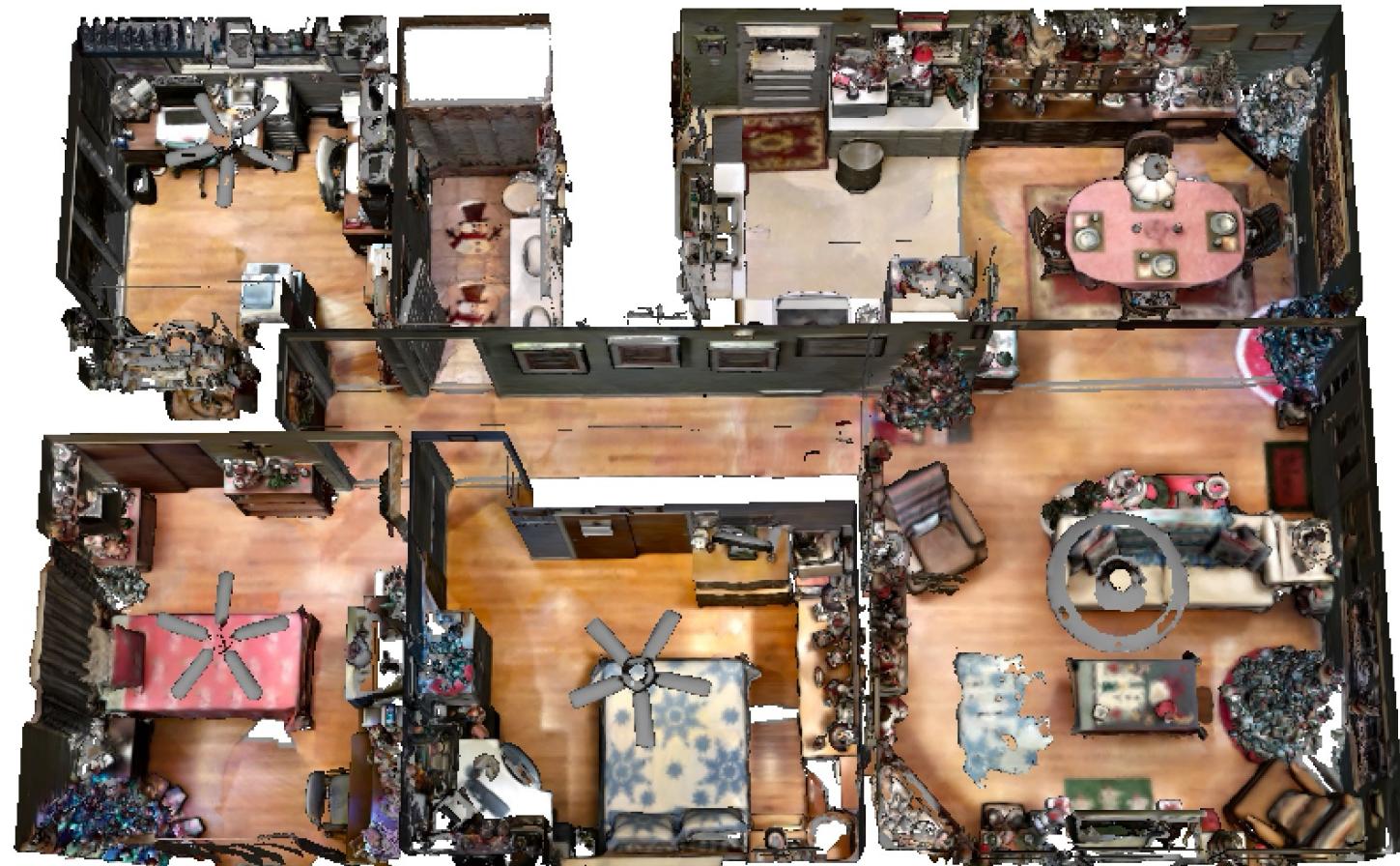
Image Queries

Given 3D Geometry

Interactive Demo

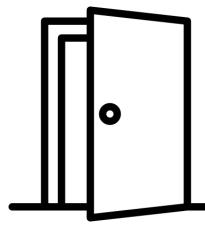
Open-vocabulary 3D Scene Exploration

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**

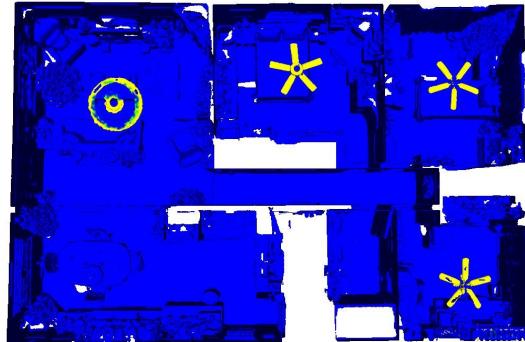


OpenScene

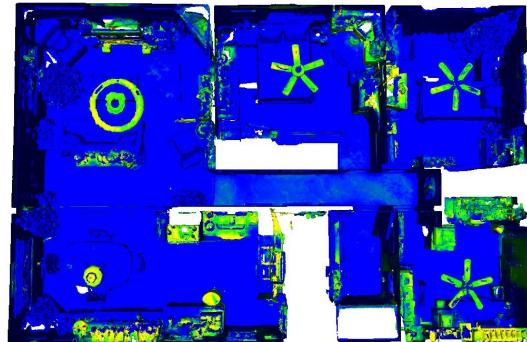
3D Scene Understanding with Open Vocabularies



Input 3D Point Cloud



“fan” - Object



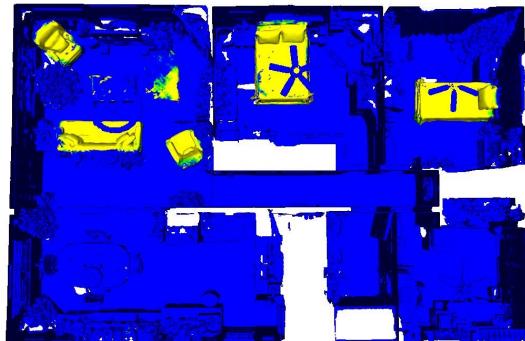
“made of metal” - Material



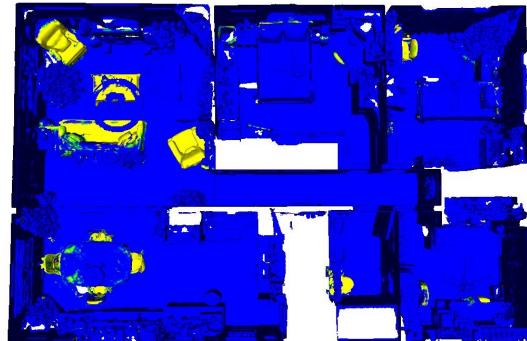
“kitchen” - Room Type



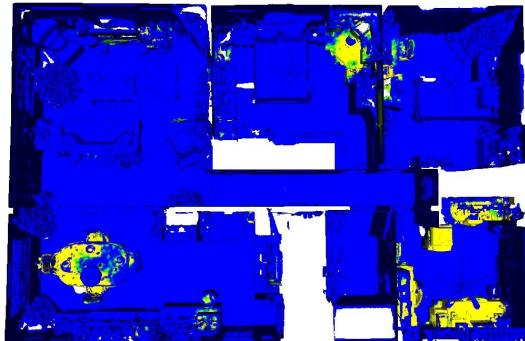
Zero-shot Semantic Segmentation



“anything soft” - Property



“where to sit” - Affordance



“work” - Activity

pengsongyou.github.io/openscene