ICLR
International Conference On Learning Representations

Open-YOLO 3D: Towards Fast and Accurate Open-Vocabulary 3D Instance Segmentation

Mohamed El Amine Boudjoghra[1,2], Angela Dai [2], Jean Lahoud[1],
Hisham Cholakkal[1], Rao Muhammad Anwer[1,3], Salman Khan[1,4], Fahad Shahbaz Khan[1,5]

[1]Mohamed Bin Zayed University of Artificial Intelligence [2]Technical University of Munich [3]Aalto University
[4]Autralian National University [5]Linkoping University

ICLR 2025

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Aalto University

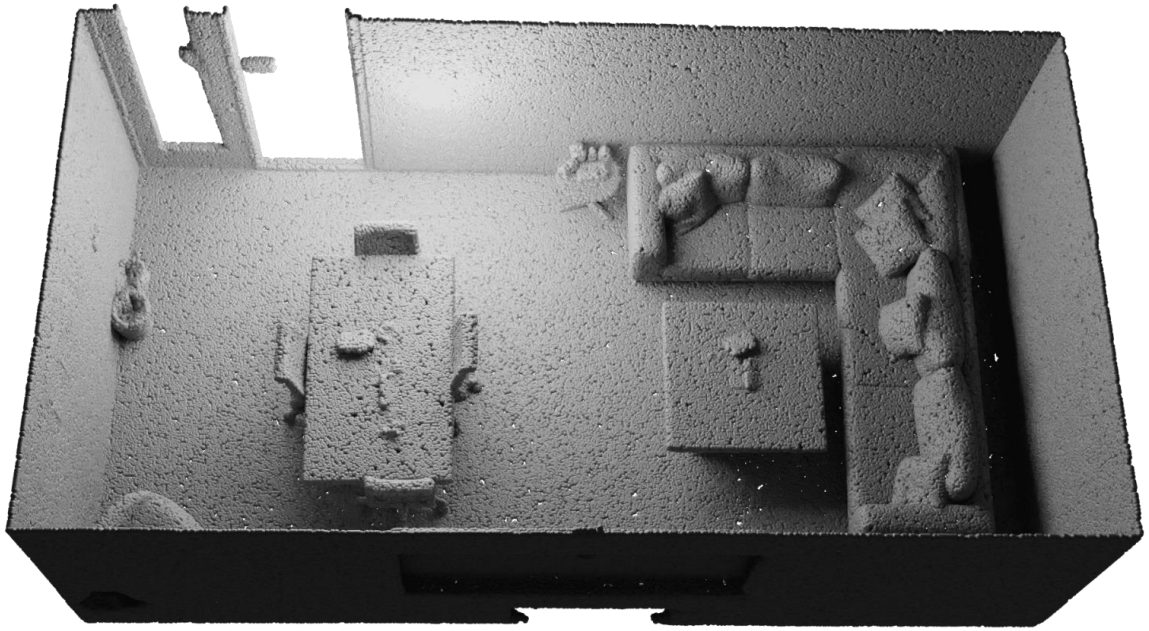
Australian National University

LINKÖPING UNIVERSITY

Task Introduction

- **Input**: 3D point cloud scene + a text prompt

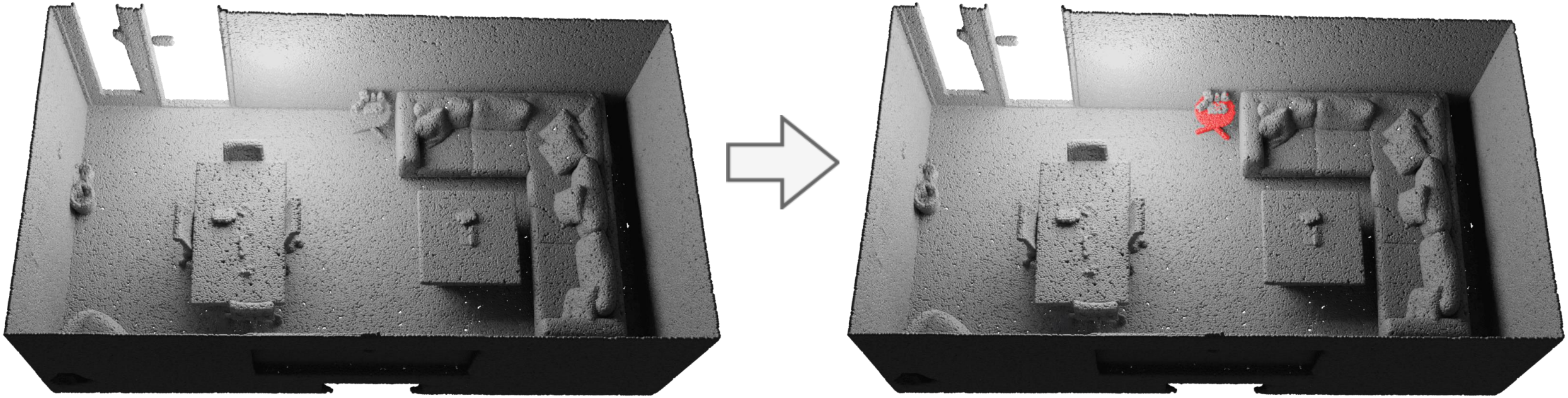
“Round table”



Task Introduction

- **Input**: 3D point cloud scene + a text prompt
- **Output**: a mask for the target instances that align with the prompt

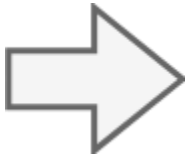
“Round table”



Task Introduction

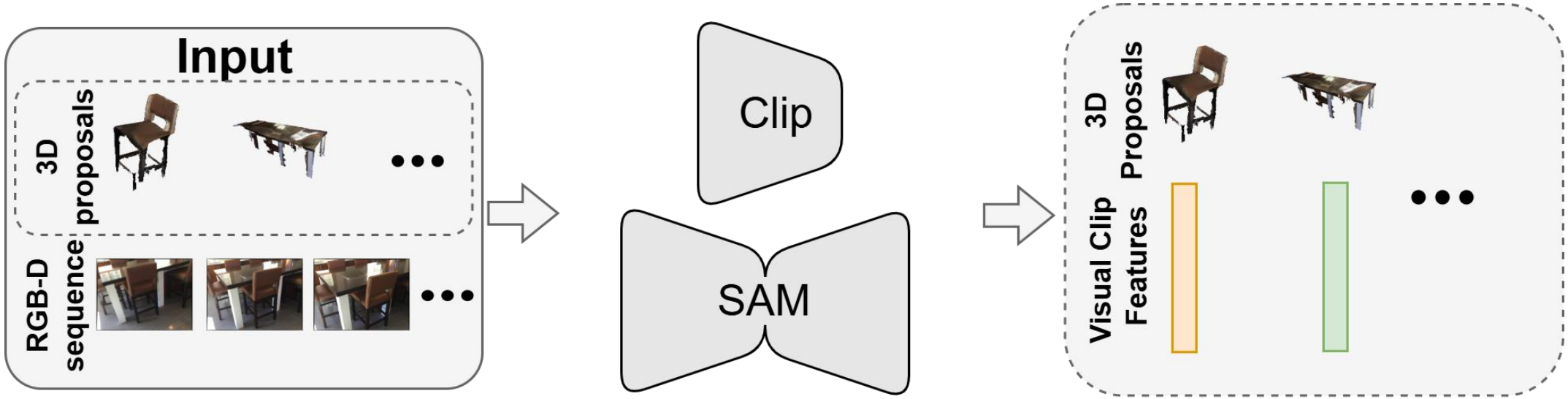
- **Input**: 3D point cloud scene + a text prompt
- **Output**: a mask for the target instances that align with the prompt

“Couch”



Previous works

- OpenMask3D [1] uses SAM to generate 2D crops from 2D projections of the 3D proposal and CLIP to generate **quarriable clip features** for 3D proposals.
- Open3DIS [1] uses SAM to aggregate 3D proposal and CLIP to generate **quarriable clip features** for 3D proposals.

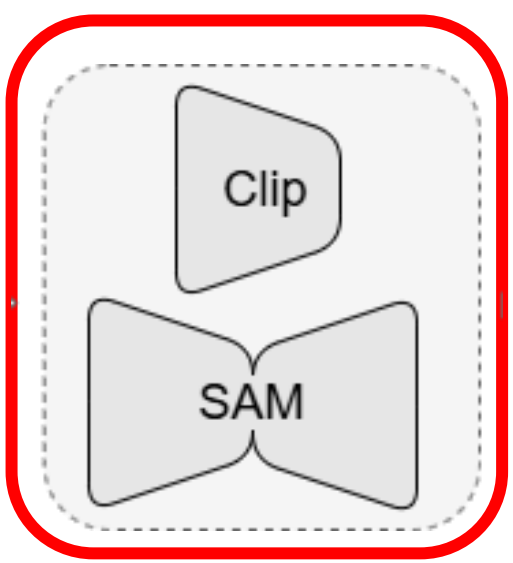


[1] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. NeurIPS 2023

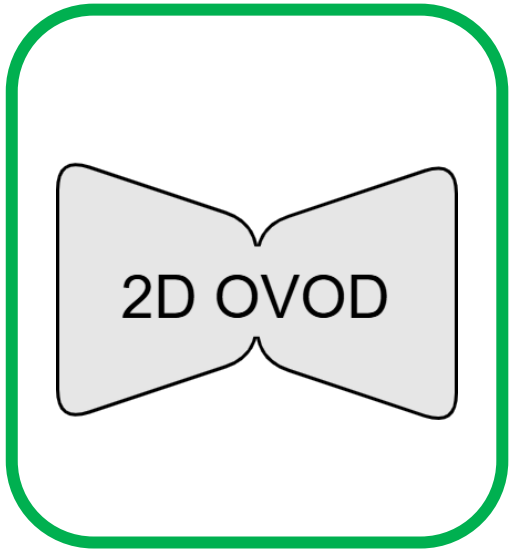
[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

Motivation

- Replacing SAM and CLIP which are very slow with and Object detector which is real time.
- Querying 3D instances with only an Open-Vocabulary Object Detector (OVOD).



SLOW!



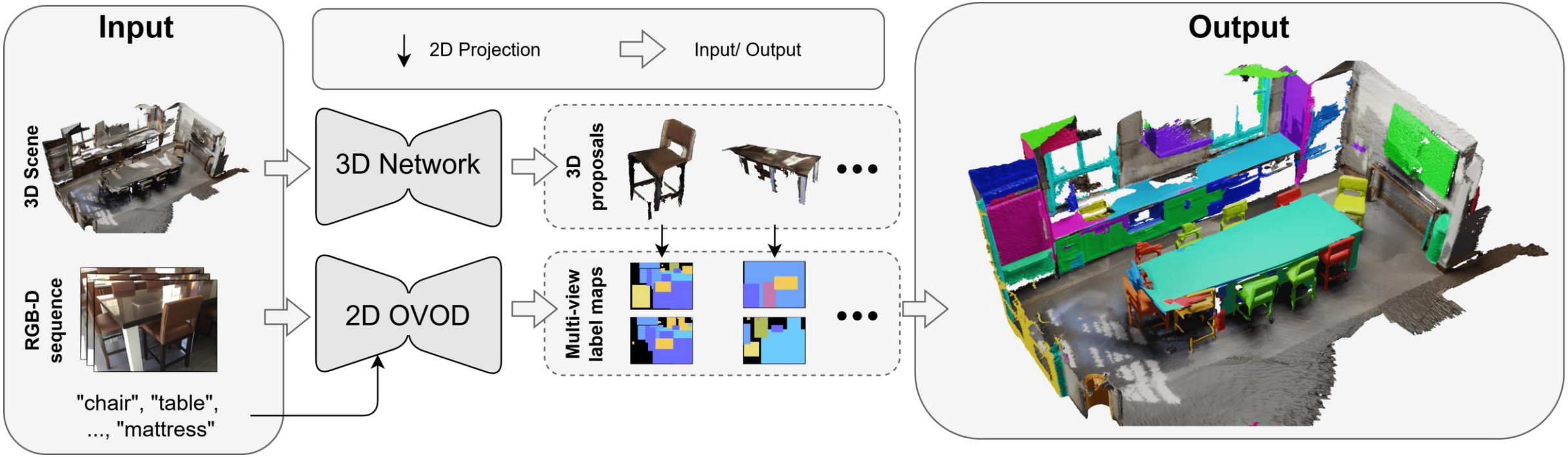
FAST

[1] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. NeurIPS 2023

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. PMLR 2021

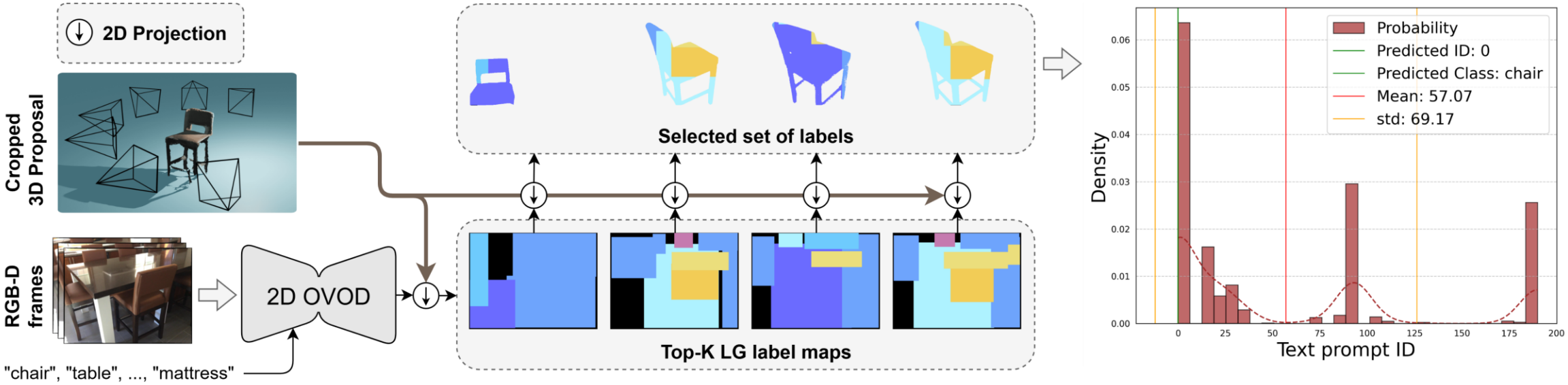
Methodology

- We use a pretrained 3D proposal network to generate class-agnostic 3D instance masks
- An Open-Vocabulary Object Detector (OVOD) as prior to query instances with text

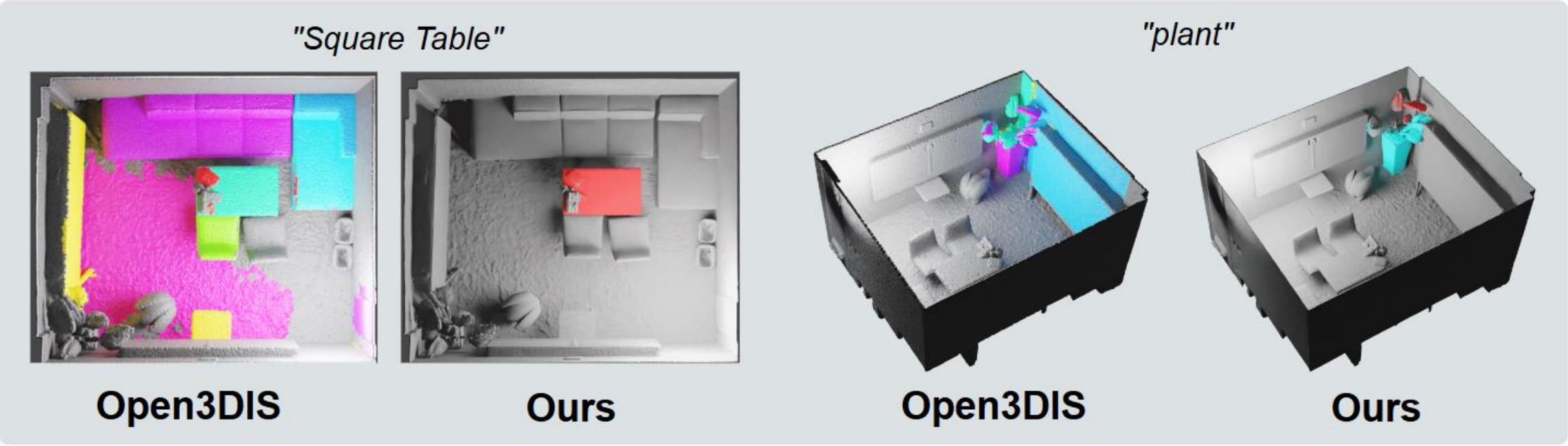


Methodology (Multi-View Prompt Distribution)

- Leveraging multi-view information for more accurate prediction



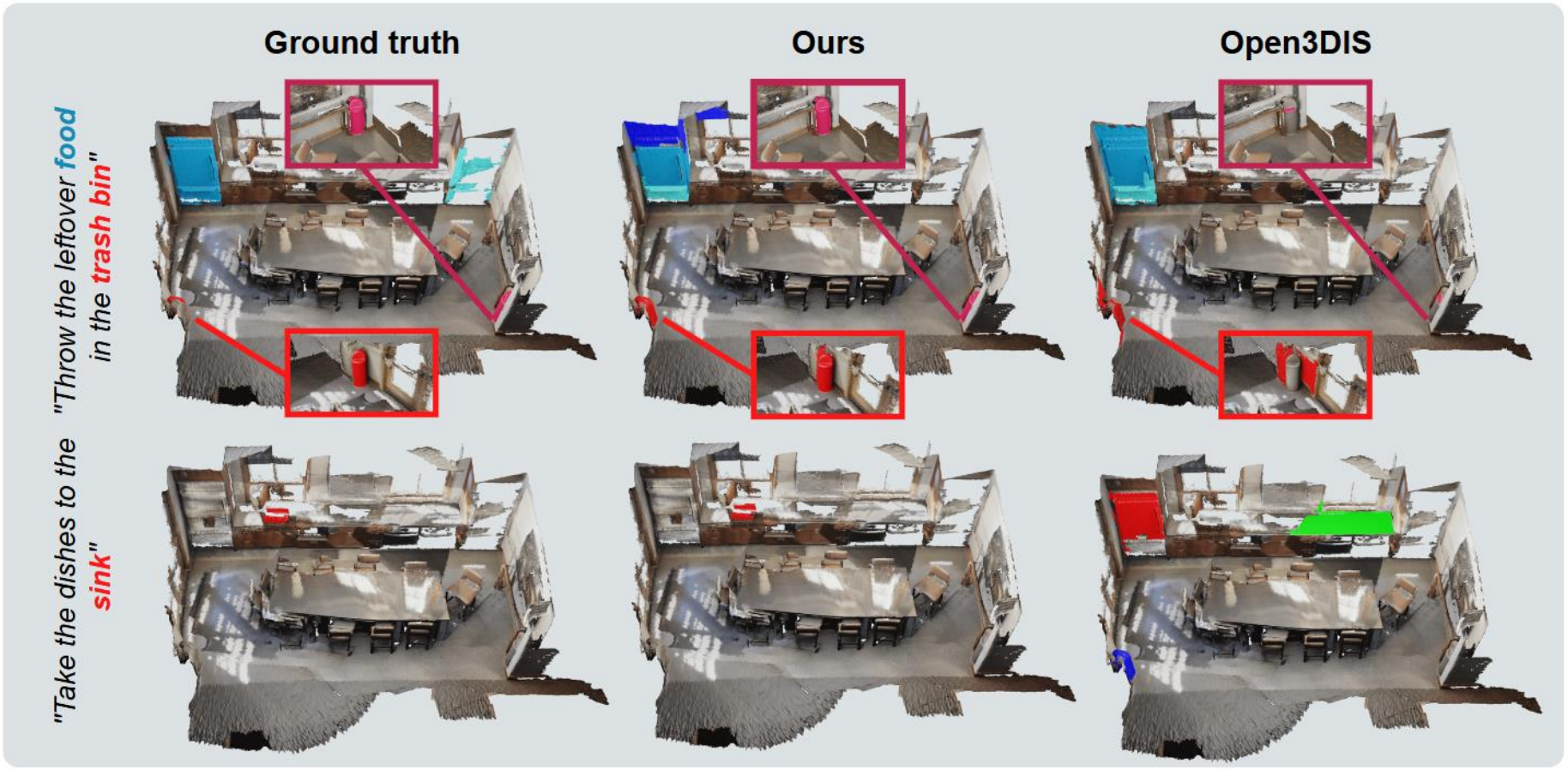
Results (Replica dataset)



[1] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. arXiv 2019

[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

Results (ScanNet dataset)

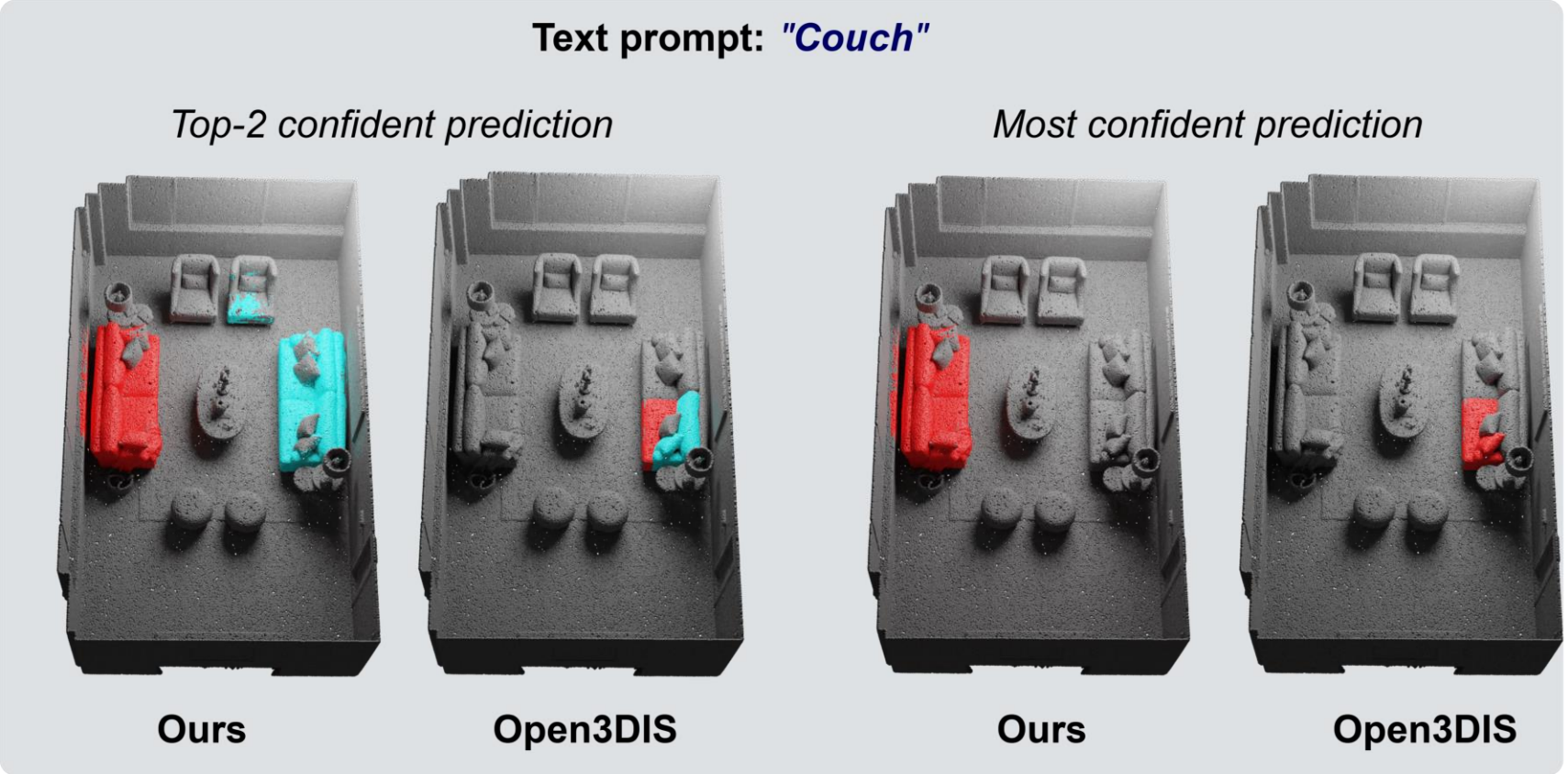


[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. CVPR 2017

[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

Results (Replica dataset)

- Scoring predictions

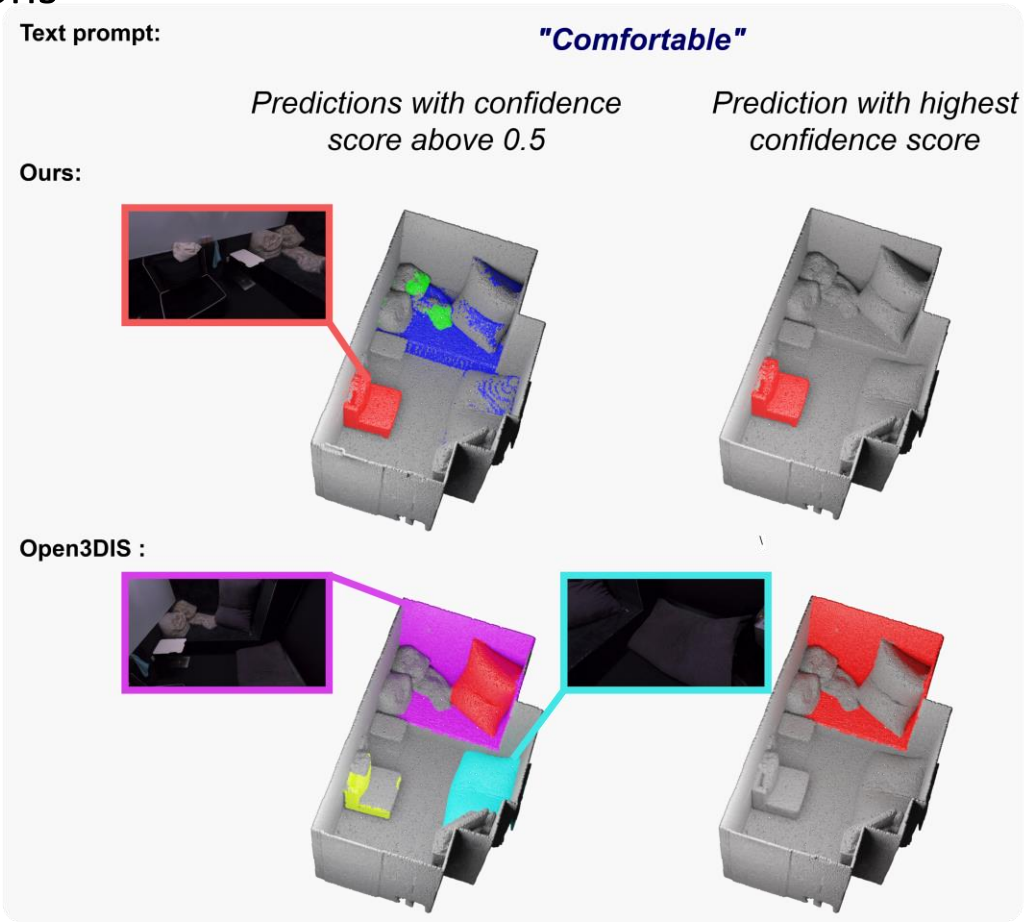


[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. CVPR 2017

[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

Results (Replica dataset)

- Scoring predictions

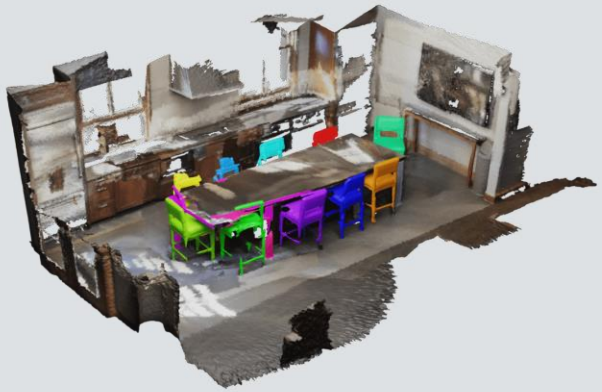


[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. CVPR 2017

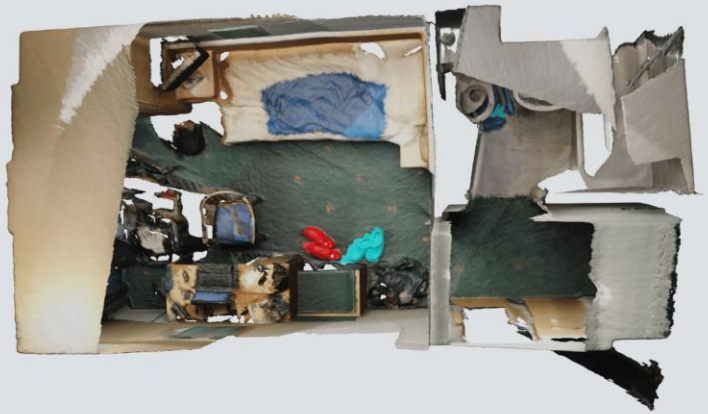
[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

Results (ScanNet dataset)

- Single class retrieval



"Chair"



"Shoe"

"Bed"

Results (ScanNet dataset)

- Single class retrieval

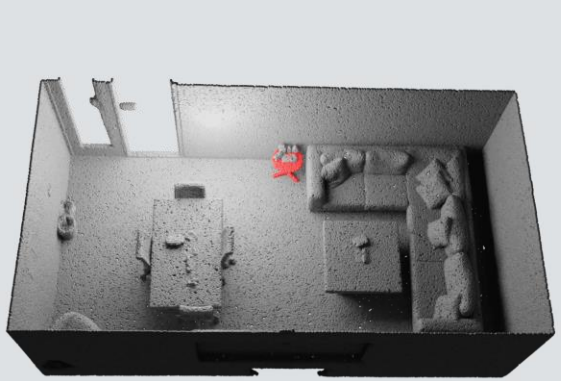
"Television"

"Backpack"

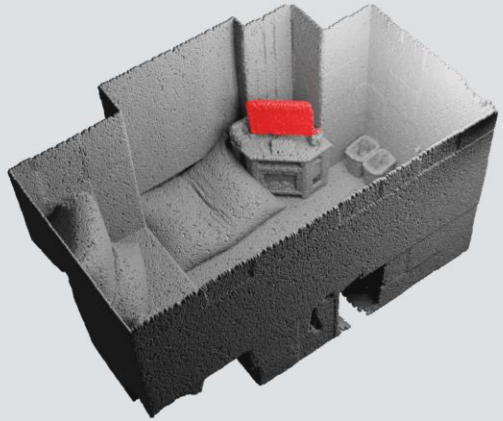
"Refrigerator"

Results (Replica dataset)

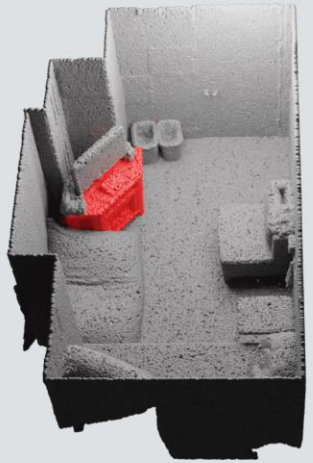
- Single class retrieval



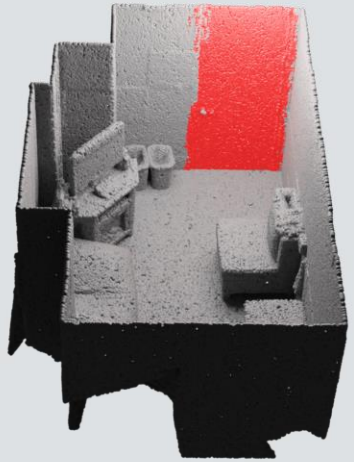
"Round table"



"Television"



"Tv stand"



"Door"

Results

Comparison against baselines on ScanNet200 validation set, when using 3D proposals from a 3D instance segmentation only.

Method	3D proposals from 2D masks	SAM for 3D mask labeling	mAP	mAP50	mAP25	head	comm	tail	time/scene (s)
OpenMask3D	×	✓	15.4	19.9	23.1	17.1	14.1	14.9	553.87
Open3DIS	×	×	18.6	23.1	27.3	24.7	16.9	13.3	57.68
Open-YOLO 3D (Ours)	×	×	**24.7**	**31.7**	**36.2**	**27.8**	**24.3**	**21.6**	21.8

- **x16 speedup against OpenMask3D [1]**
- **x2 speedup and 6.7 mAP gain against state-of-the-art Open3DIS [2]**

[1] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. NeurIPS 2023

[2] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. CVPR 2024

[3] David Rozenberszki, Or Litany, and Angela Dai. Language grounded indoor 3d semantic segmentation in the wild. ECCV 2022

Thank you!

Code

Come visit our poster — #75!

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Aalto University

Australian National University

LINKÖPING UNIVERSITY