Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance

Phuc Nguyen[1*] **Tuan Duc Ngo[1,4*]**

Evangelos Kalogerakis[4] Chuang Gan[2,4] Anh Tran[1] Cuong Pham[1,3] Khoi Nguyen[1]

[1]VinAI Research [2]MIT-IBM Watson AI Lab [3]Posts & Telecom Inst. of Tech. [4]UMass Amherst

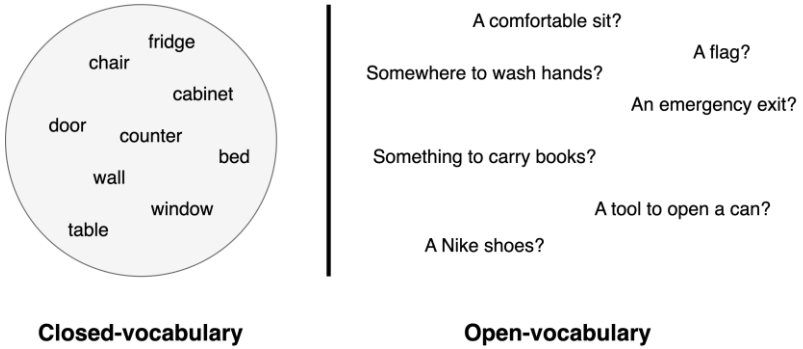
[*]*Equal contribution*

TL;DR: We propose **Open3DIS** addressing 3D Instance Segmentation with Open-Vocabulary queries

Motivations

- **3D point cloud instance segmentation (3DIS)** methods require extensive training data while limited to closed-set categories.

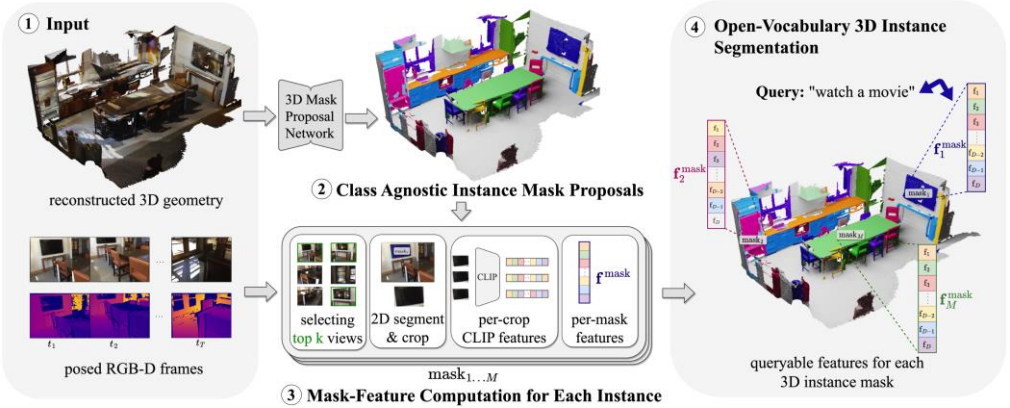
Schult et al *"Mask3D: Mask Transformer for 3D instance segmentation"*



Motivations

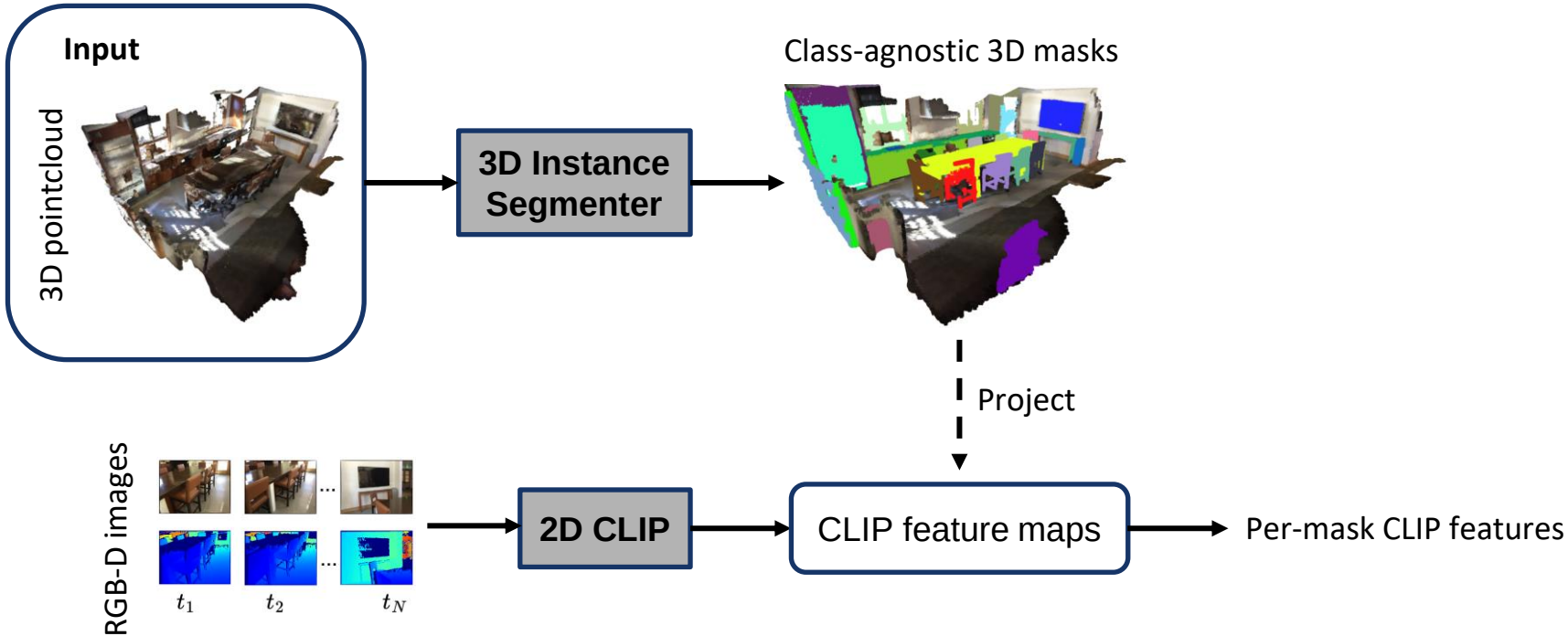
- **3D point cloud instance segmentation (3DIS)** methods require extensive training data while limited to closed-set categories.

- Recent **open-vocab 3DIS** methods struggle with small or ambiguous instances, particularly those from uncommon classes

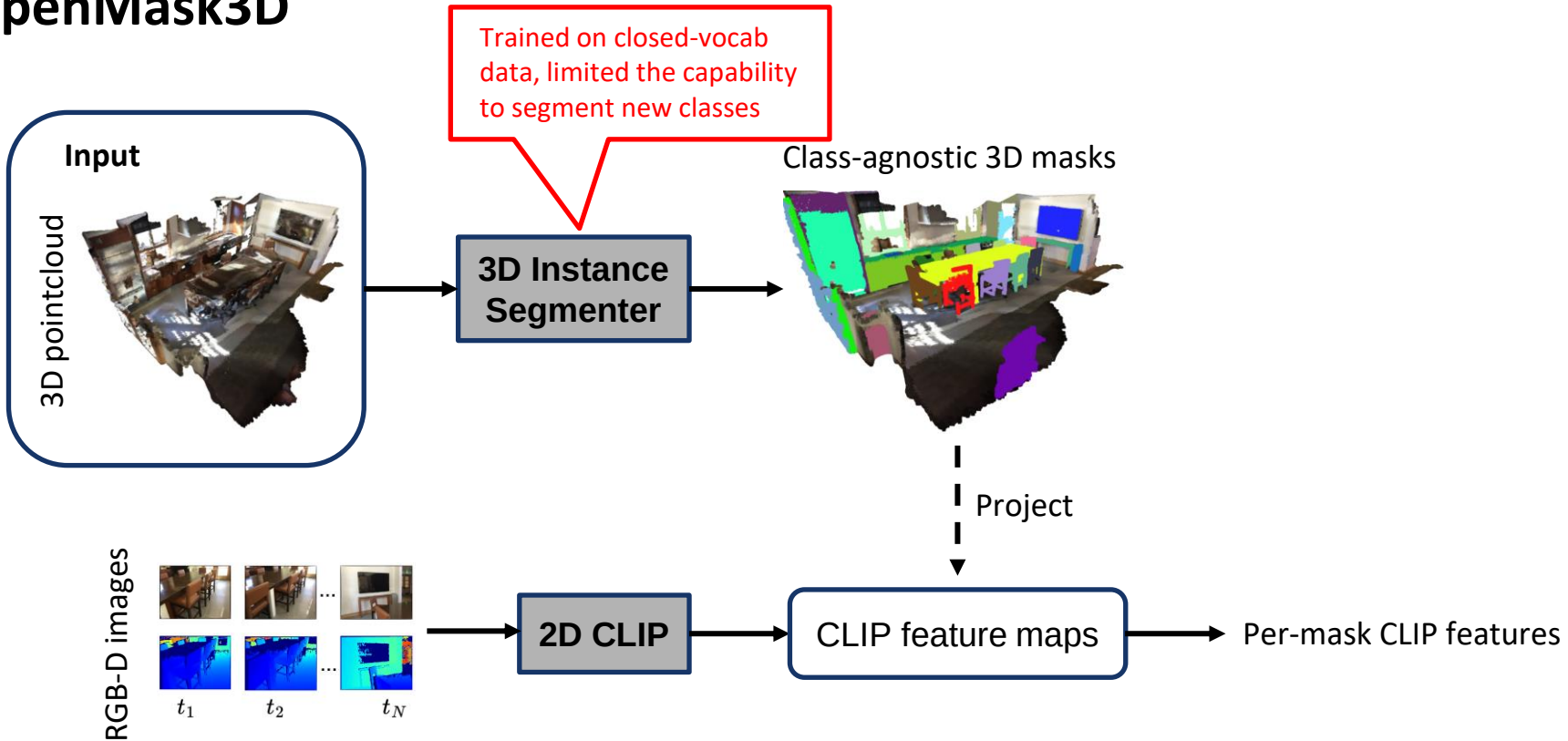


Takmaz et al "*OpenMask3D: Open-vocabulary 3D Instance Segmentation*"

OpenMask3D


Input
3D pointcloud
3D Instance Segmenter
Class-agnostic 3D masks
Project
RGB-D images
t_1
t_2
t_N
2D CLIP
CLIP feature maps
Per-mask CLIP features

OpenMask3D

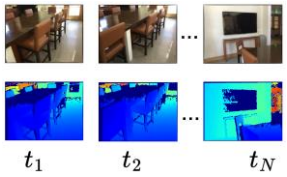

Trained on closed-vocab data, limited the capability to segment new classes
Input
3D pointcloud
3D Instance Segmenter
Class-agnostic 3D masks
Project
RGB-D images
t_1
t_2
t_N
2D CLIP
CLIP feature maps
Per-mask CLIP features

Our proposed Open3DIS

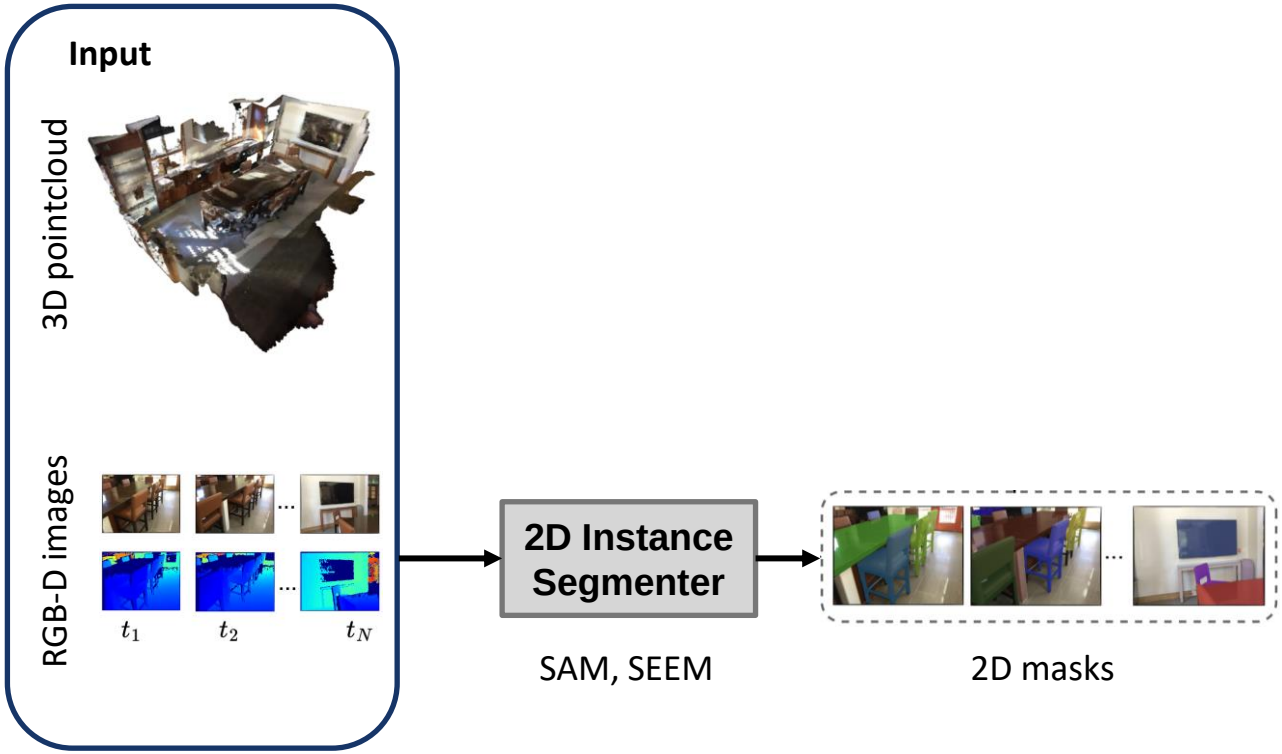
Input

3D pointcloud

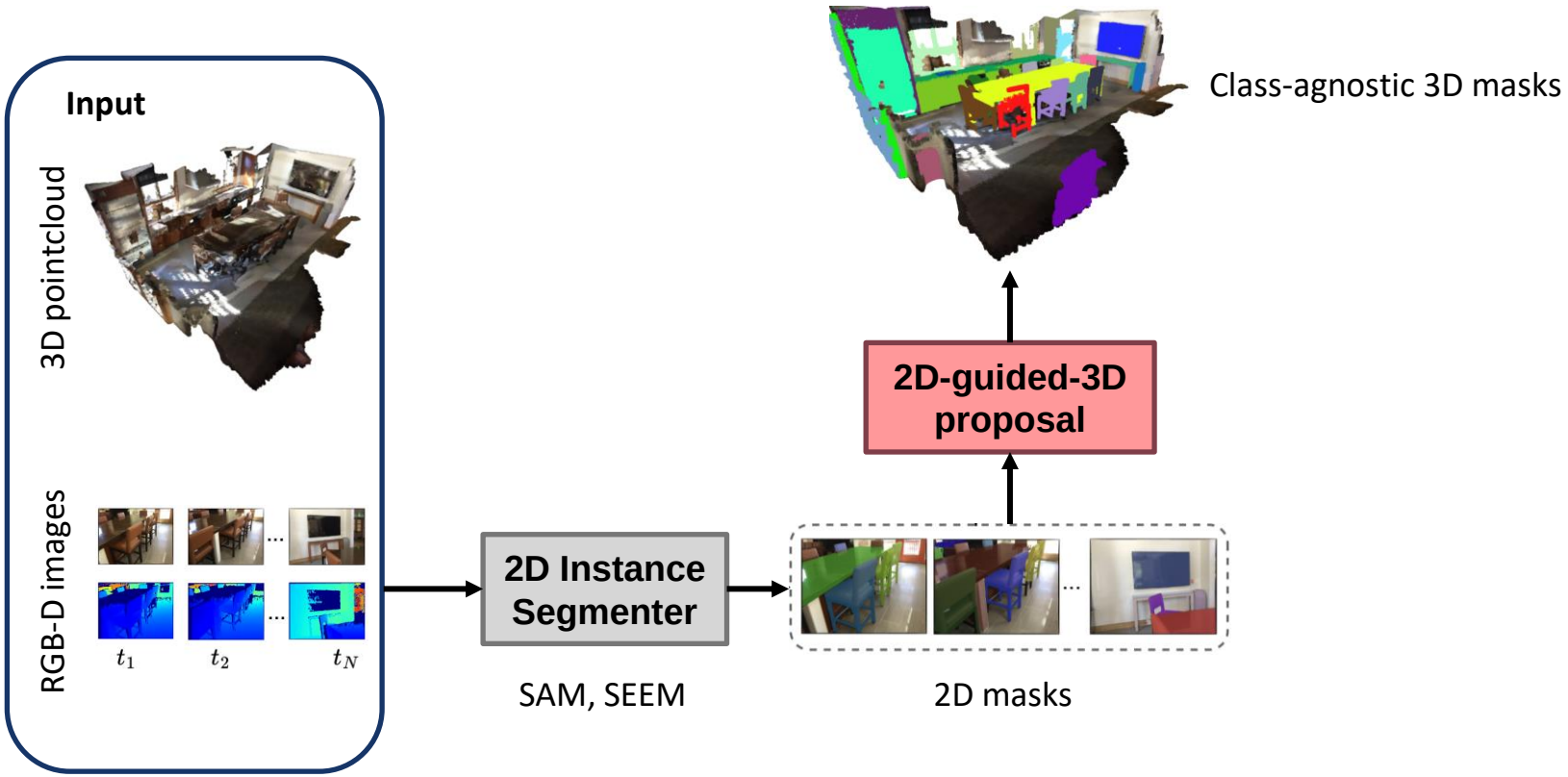
RGB-D images



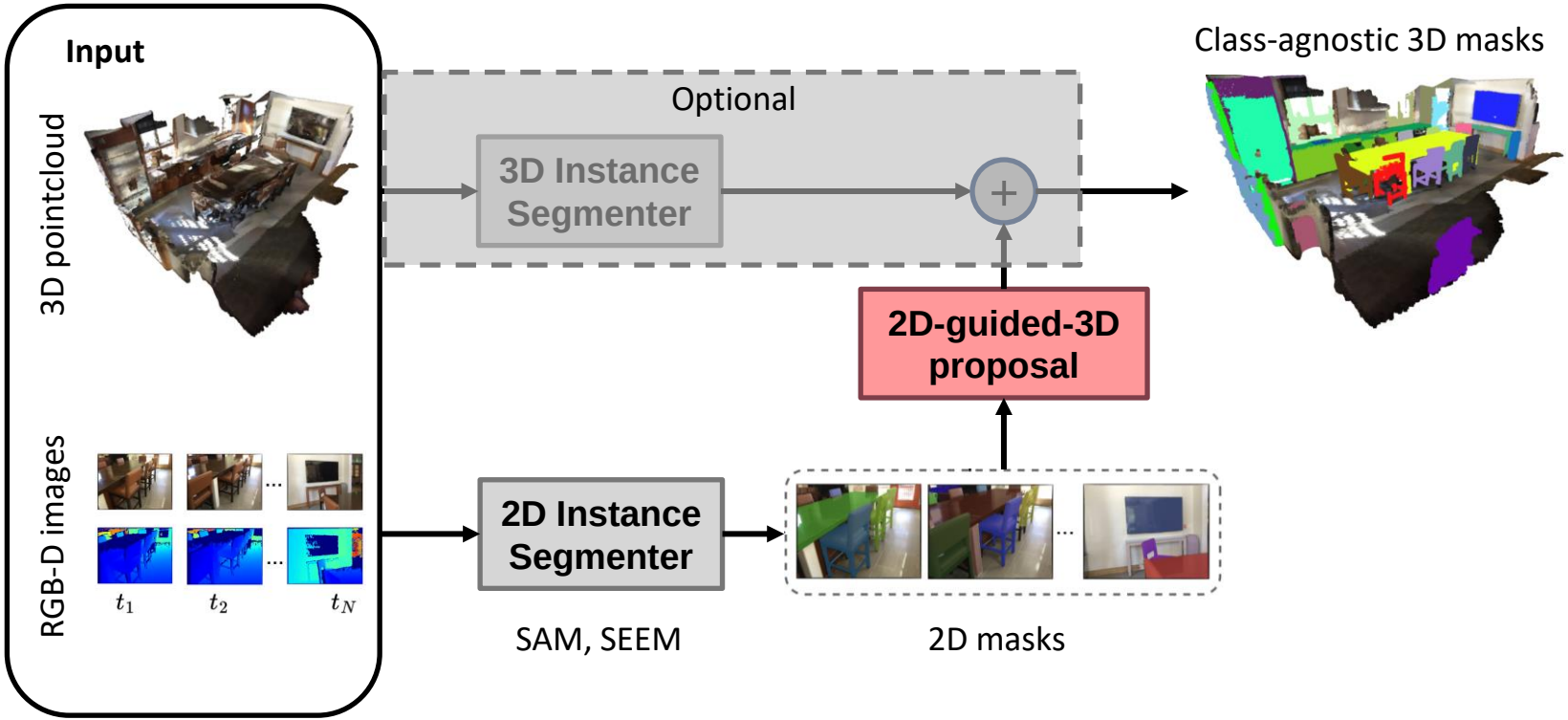
Our proposed Open3DIS



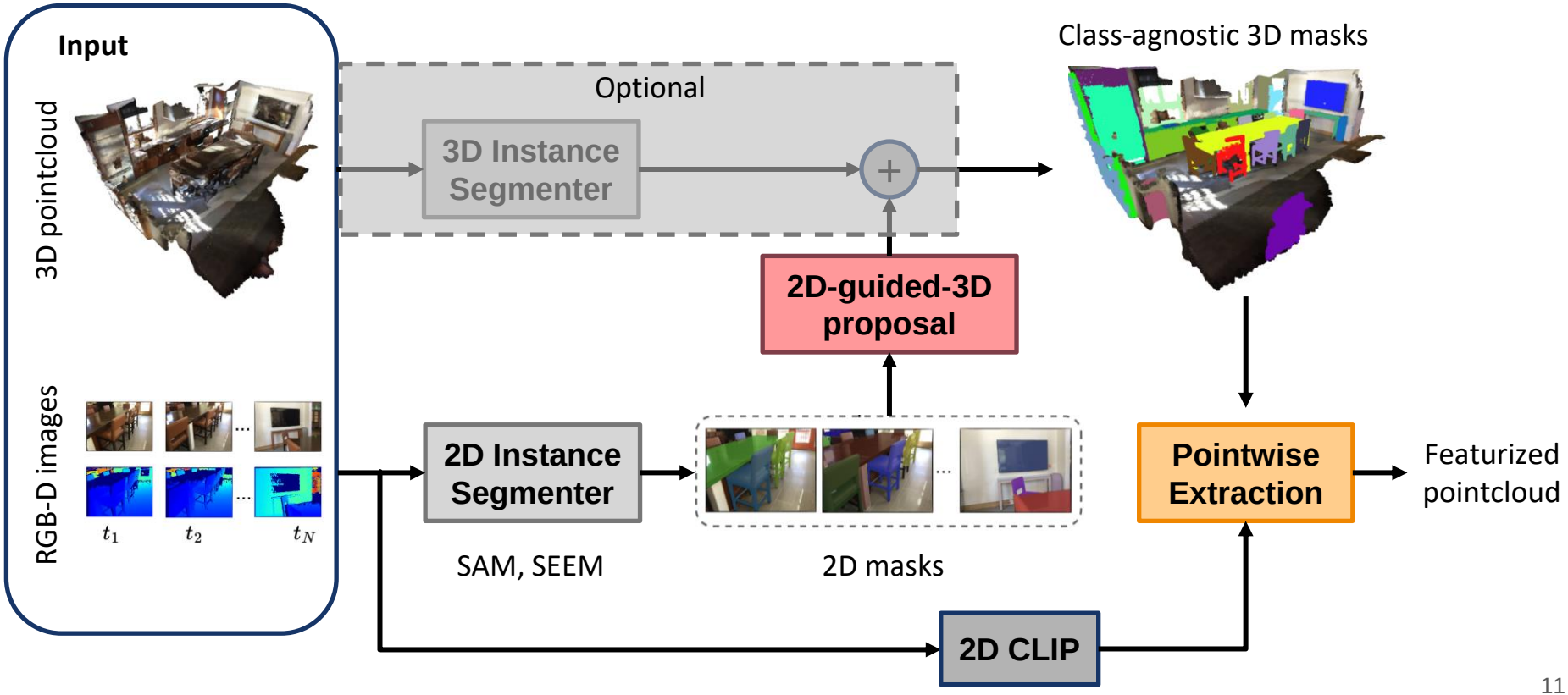
Our proposed Open3DIS



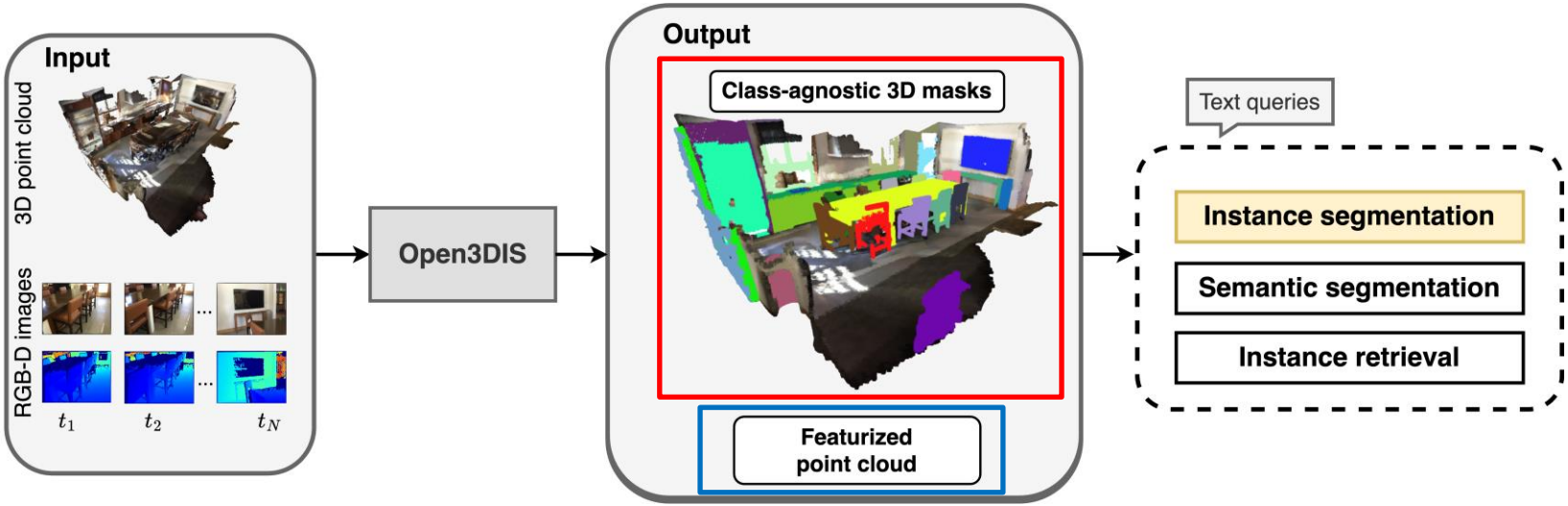
Our proposed Open3DIS



Our proposed Open3DIS



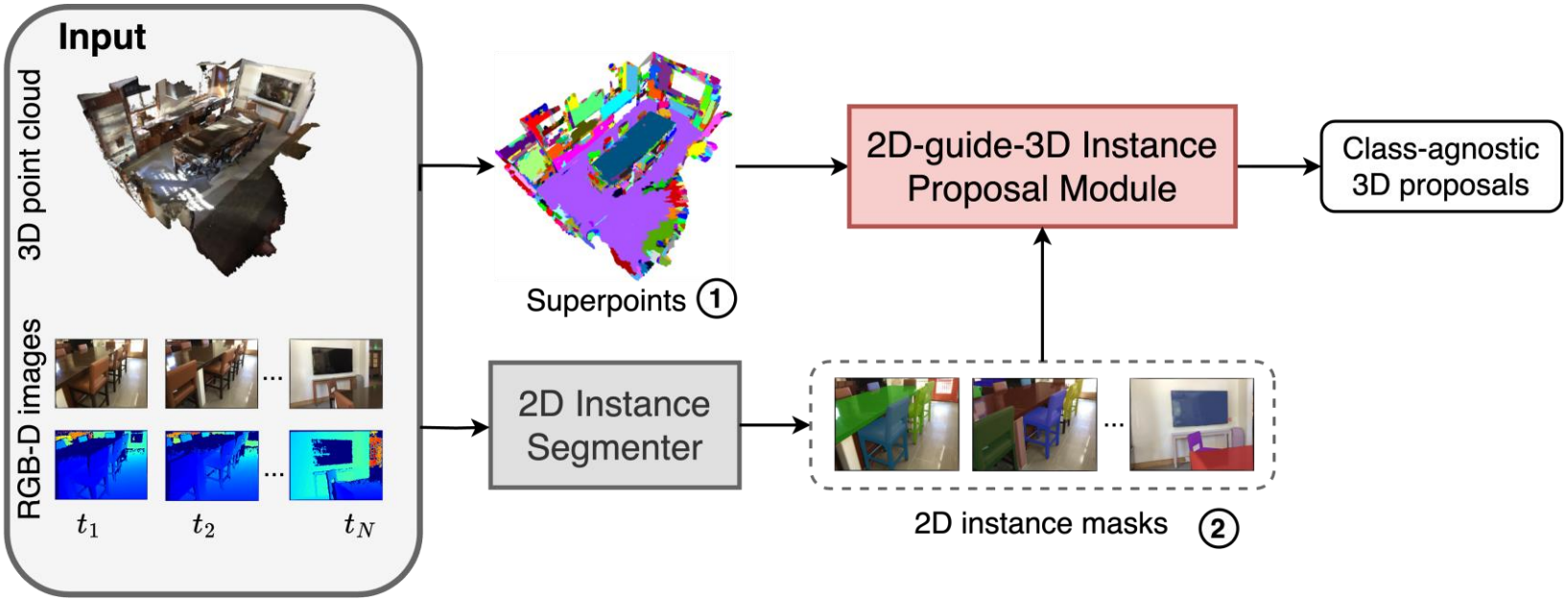
Our proposed Open3DIS



Given a **3D point cloud** with the corresponding **RGB-D sequences**, Open3DIS generates a set of **class-agnostic 3D proposals** and a **featurized point cloud** for open-vocabulary queries

How to obtain 3D instance proposals?

3D instance proposal module



3D Superpoints

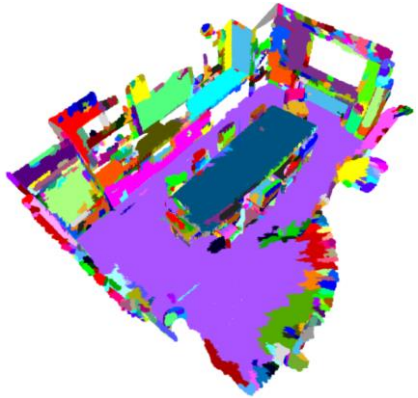
Original images

Super-pixels

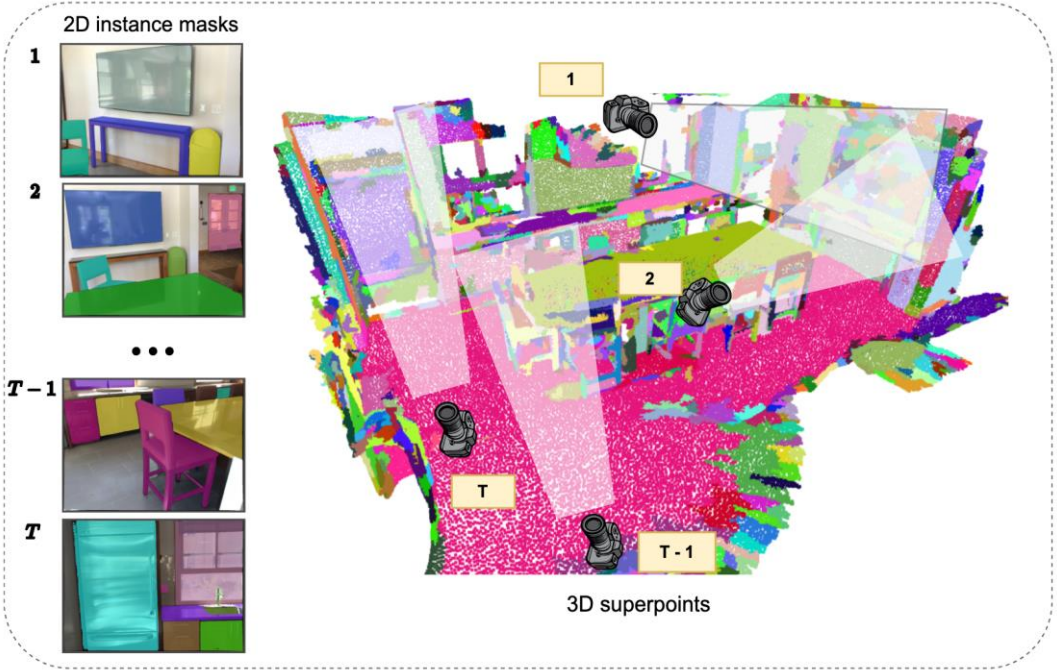
3D Superpoints

Original point cloud

Superpoints



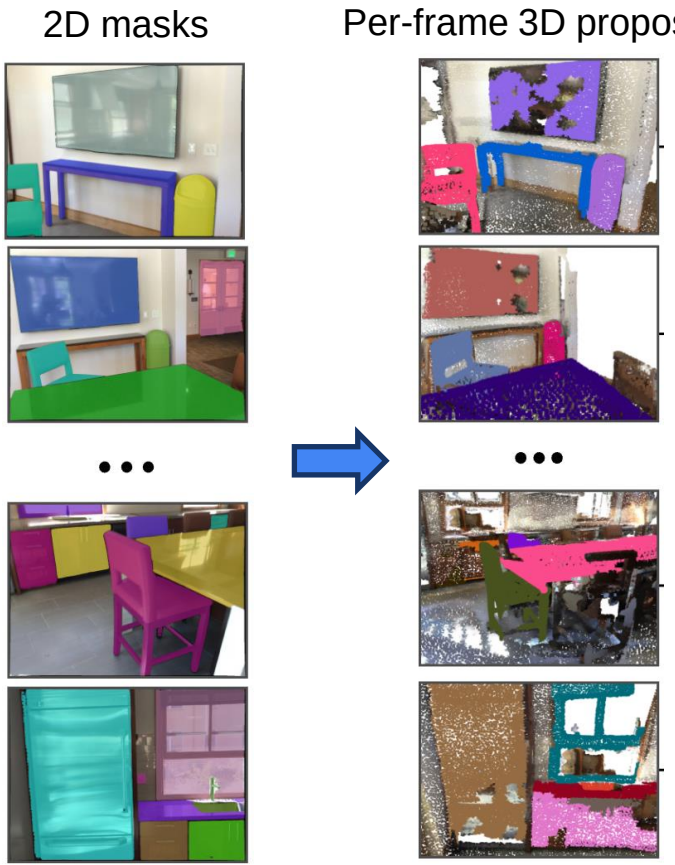
Per-frame superpoints merging



1. Use SAM to obtain a set of **2D masks** of every RGB frame

2. For each 2D mask *m,* we project **superpoints** onto their image planes and calculate the IoUs with predicted **2D masks**

Per-frame superpoints merging

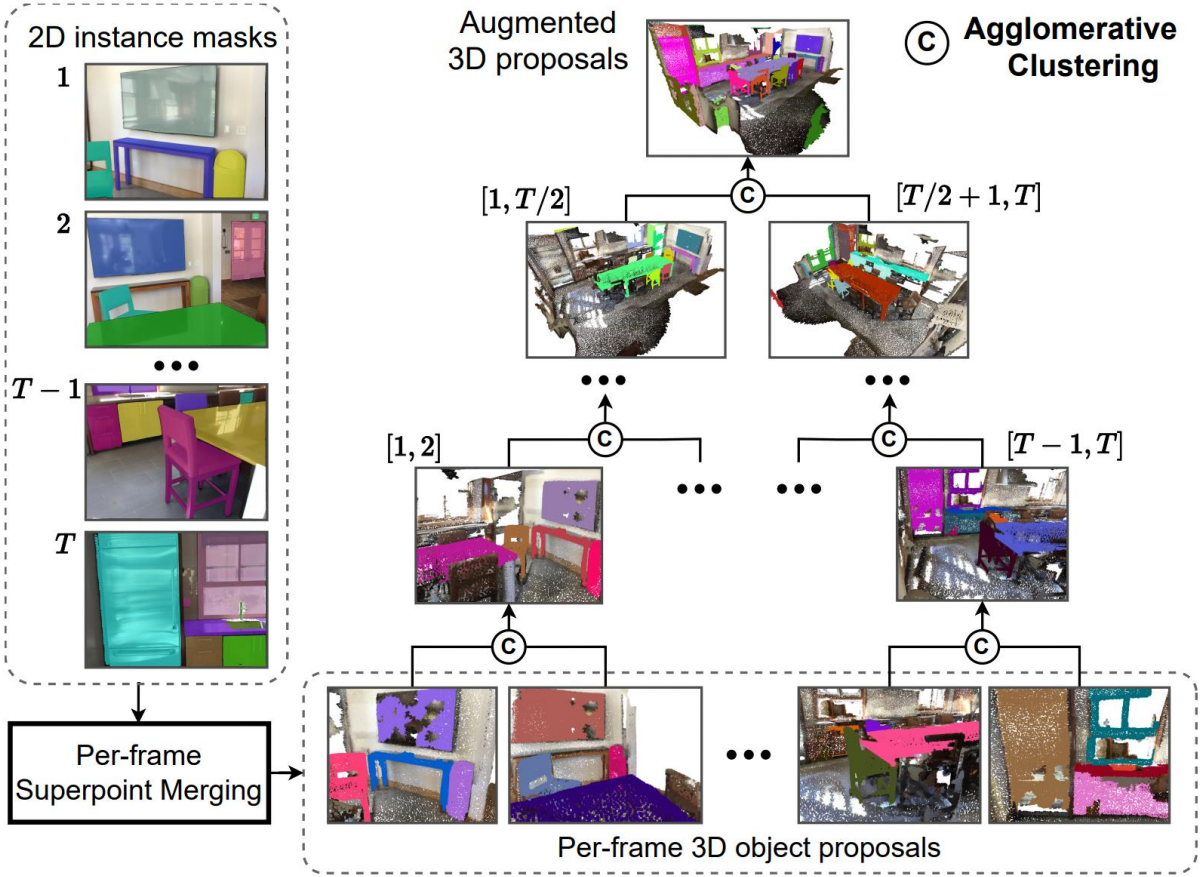


1. Use SAM to obtain a set of **2D masks** of every RGB frame

2. For each 2D mask *m,* we project **superpoints** onto their image planes and calculate the IoUs with predicted **2D masks**

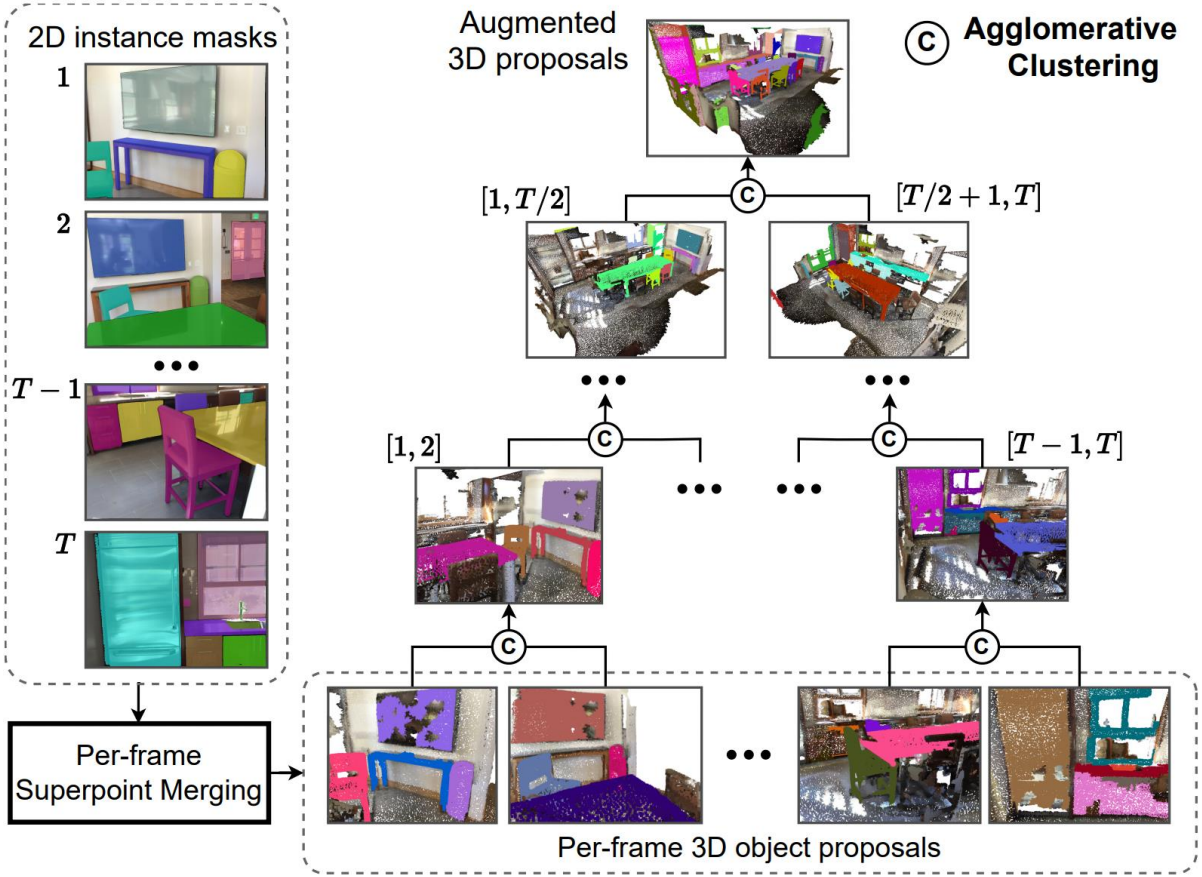
3. We merge all superpoints having sufficient overlap to create an initial **3D proposal of mask** *m*

3D instance proposal module



1. We merge point cloud regions from different frames in a bottom-up manner.

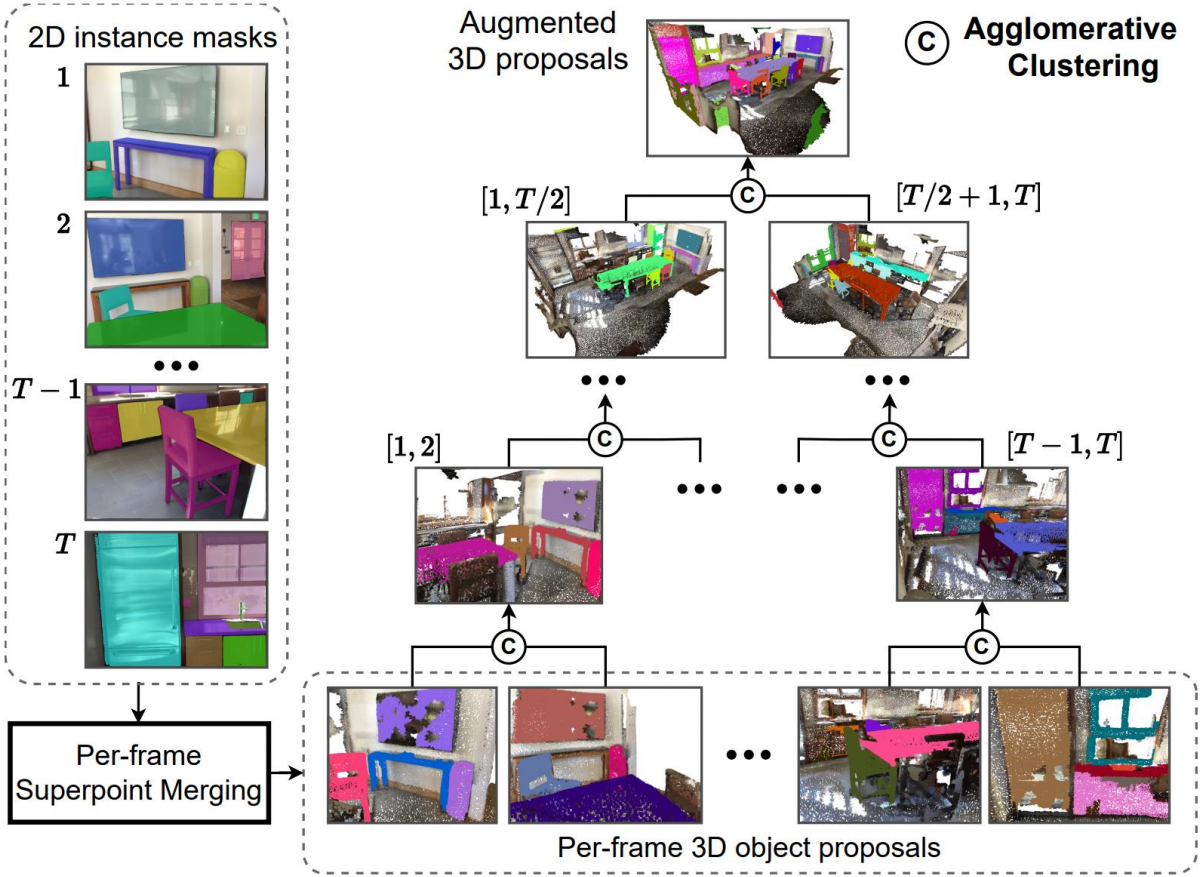
3D instance proposal module



1. We merge point cloud regions from different frames in a bottom-up manner.

2. Agglomerative clustering is chosen to combine proposals from pairs of frames.

3D instance proposal module

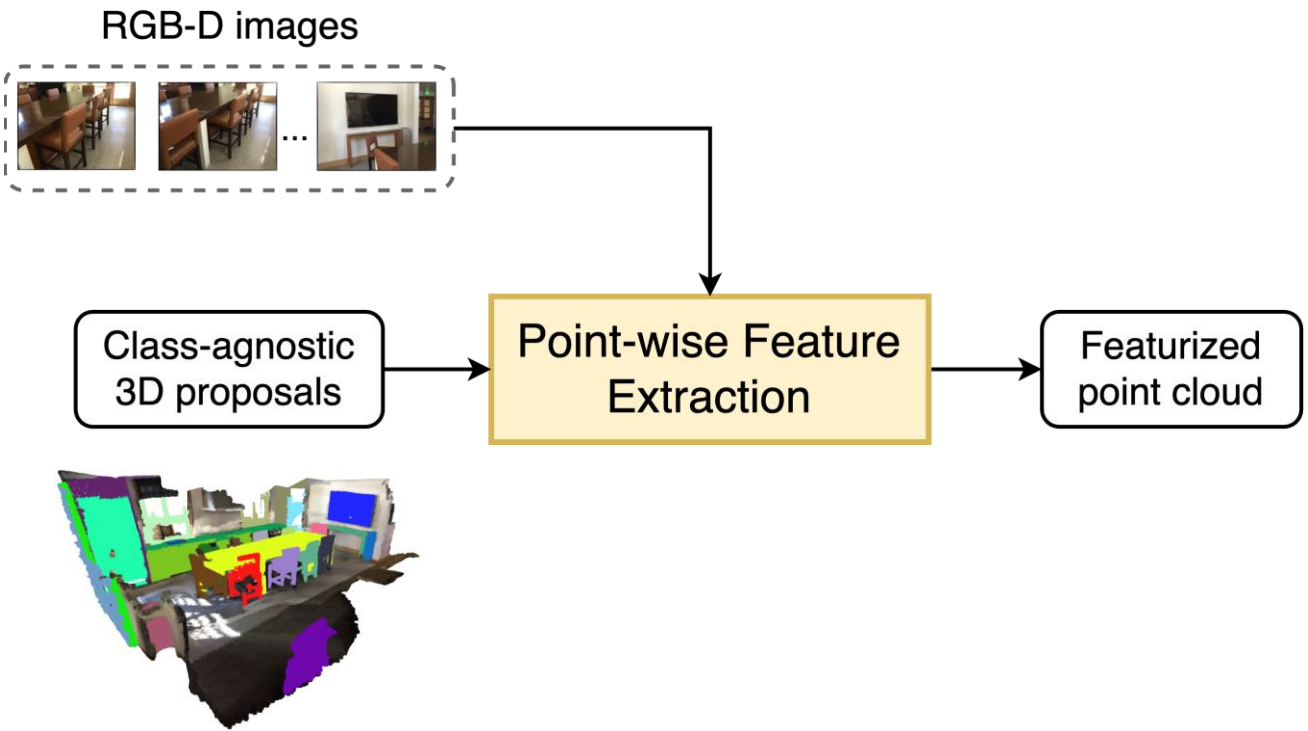


1. We merge point cloud regions from different frames in a bottom-up manner.

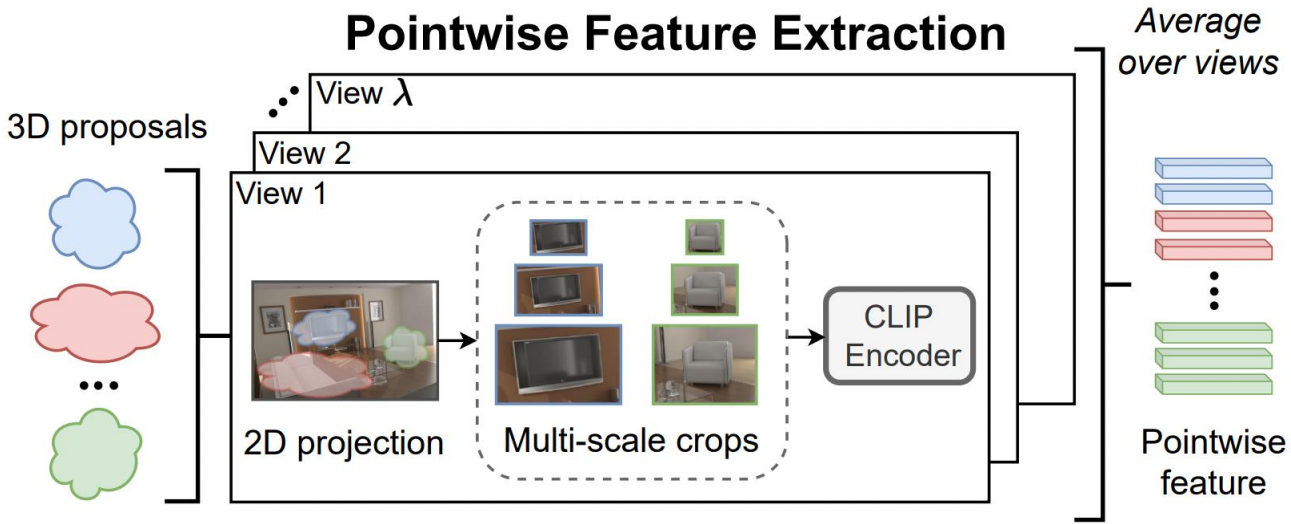
2. Agglomerative clustering is chosen to combine proposals from pairs of frames.

How to obtain featurized point cloud?

Point-wise feature extraction



Point-wise feature extraction



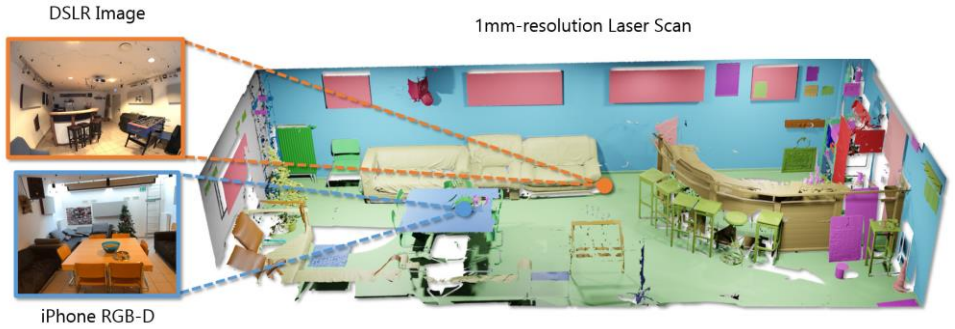
Experiments

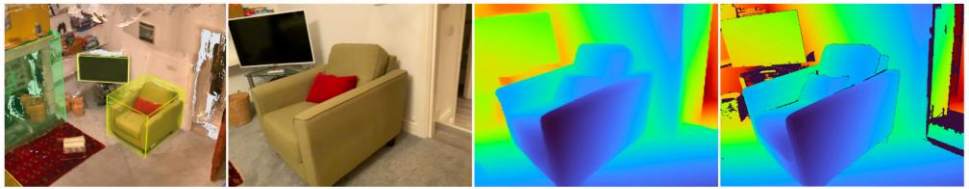
Datasets:

- ScanNet++ and ScanNet200
- ARKitScenes
- S3DIS
- Replica

Metrics:

- AP (Average Precision)
- AR (Average Recall)





Quantitative Results: ScanNet200

Method	Setting	3D Proposal	AP	AP_{50}	AP_{25}	AP_{head}	AP_{com}	AP_{tail}
ISBNet [36]	Fully-sup		24.5	32.7	37.6	38.6	20.5	12.5
Mask3D [42]			26.9	36.2	41.4	39.8	21.7	17.9
OpenScene [37] + DBScan [10]†	Open-vocab	None	2.8	7.8	18.6	2.7	3.1	2.6
OpenScene [37] + Mask3D [42]		Mask3D [42]	11.7	15.2	17.8	13.4	11.6	9.9
SAM3D† [57]		None	6.1	14.2	21.3	7.0	6.2	4.6
OVIR-3D† [33]		None	13.0	24.9	32.3	14.4	12.7	11.7
OpenMask3D [45]		Mask3D [42]	15.4	19.9	23.1	17.1	14.1	14.9
Ours (only 2D)	Open-vocab	None	18.2	26.1	31.4	18.9	16.5	19.2
Ours (only 3D)		ISBNet [36]	18.6	23.1	27.3	24.7	16.9	13.3
Ours (2D and 3D)		ISBNet [36]	**23.7**	**29.4**	**32.8**	**27.8**	**21.2**	**21.8**

Our method surpasses previous OV-3DIS approaches by a large margin, having competitive performance with fully-supervised methods

Quantitative Results: Replica & S3DIS

Method	3D Proposal	AP	AP_{50}	AP_{25}
OpenScene + Mask3D	Mask3D	10.9	15.6	17.3
OpenMask3D	Mask3D	13.1	18.4	24.2
OVIR-3D †	None	11.1	20.5	27.5
Ours	None	**18.1**	**26.7**	**30.5**

Replica

Method	B8/N4		B6/N6	
	AP^B_{50}	AP^N_{50}	AP^B_{50}	AP^N_{50}
LSeg-3D [8]	58.3	0.3	41.1	0.5
PLA [8]	59.0	8.6	46.9	9.8
Lowis3D [7]	58.7	13.8	**51.8**	15.8
Ours	**60.8**	**26.3**	50.0	**29.0**

S3DIS

Quantitative Results: ScanNet++

Method	AP	AP_{50}	AP_{25}	AR	AR_{50}	AR_{25}	NOTE
ISBNet [50] (3D)	6.2	10.1	16.2	10.9	16.9	25.2	pretrained Scannet200
SAM3D [78]	7.2	14.2	29.4				
SAM-guided Graph Cut [18]	12.9	25.3	43.6				
Segment3D [26]	12.0	22.7	37.8				
SAI3D [82] (SAM)	17.1	31.1	49.5				
Ours (SAM)	18.5	33.5	44.3	35.6	63.7	82.7	100 frames per scene
Ours (SAM)	20.7	38.6	47.1	40.8	75.7	91.8	all frames per scene

Qualitative Results on ARKitScenes

Picture of a Horse

Qualitative Results on ARKitScenes

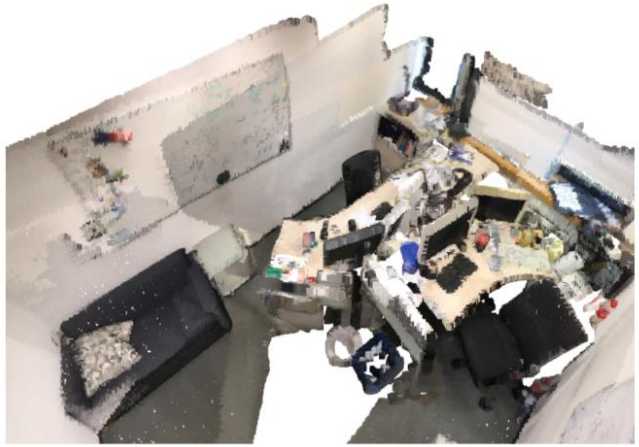
Blue letter M

Qualitative Results on ARKitScenes

Watering plants

Qualitative Results on ScanNet200

Comfortable



Qualitative Results on ScanNet200

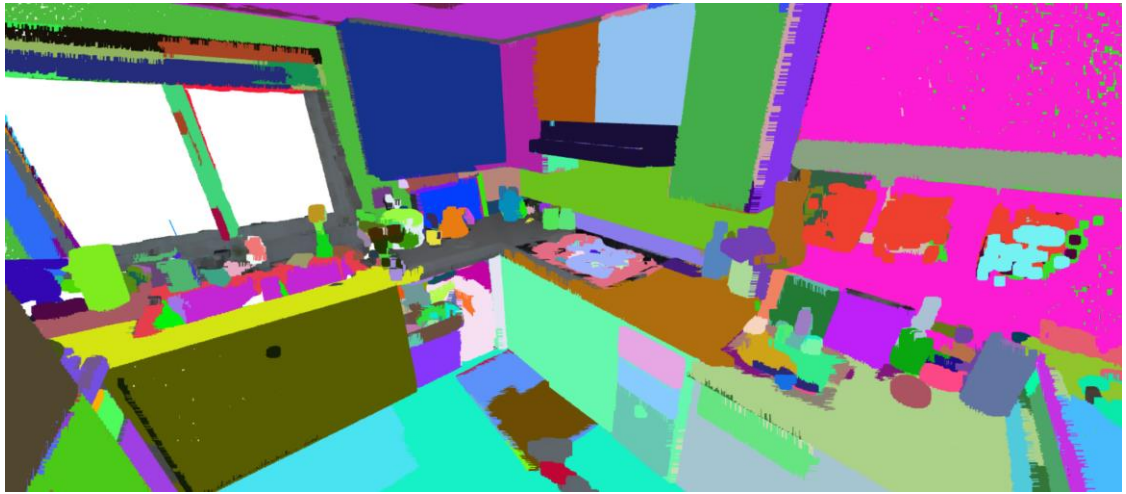
Wash your hand

Qualitative Results on ScanNet200

Chair and Table

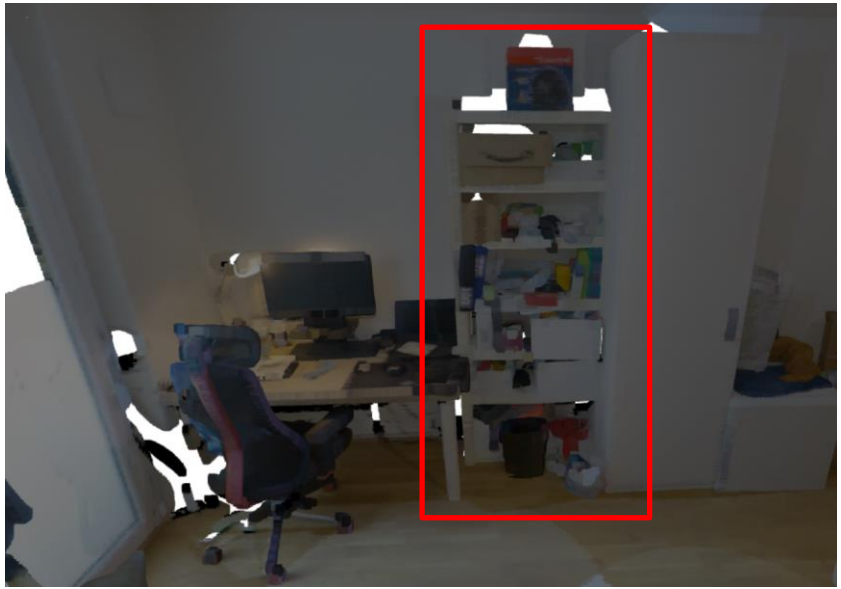
Qualitative Results on ScanNet++

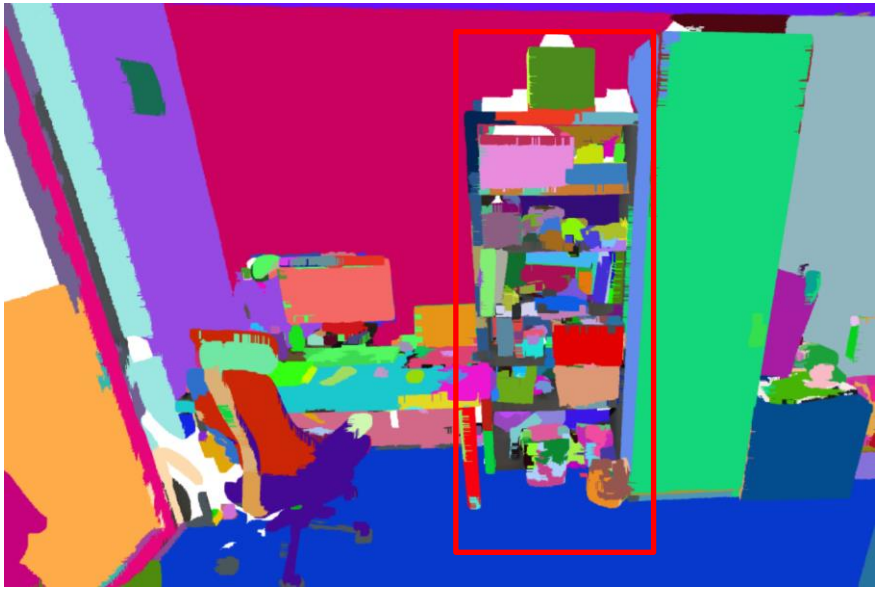
SAM Class-agnostic
3D Instance Segmentation



Qualitative Results on ScanNet++

SAM Class-agnostic 3D Instance Segmentation





Summary

- We introduce **Open3DIS** to address the open-vocabulary 3D instance segmentation task with a novel strategy to generate high-quality 3D proposals from pretrained 2D model.

- Our approach achieves **state-of-the-art performance** on 5 different datasets.

- Future research on adapting **Open3DIS** for other 3D representation such as 3D Gaussian Splatting would be an interesting exploration.

Thank you!