OVIR-3D: Open-Vocabulary 3D Instance Retrieval Without Training on 3D Data

Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, Kostas Bekris
Computer Science @ Rutgers University

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Code Available!

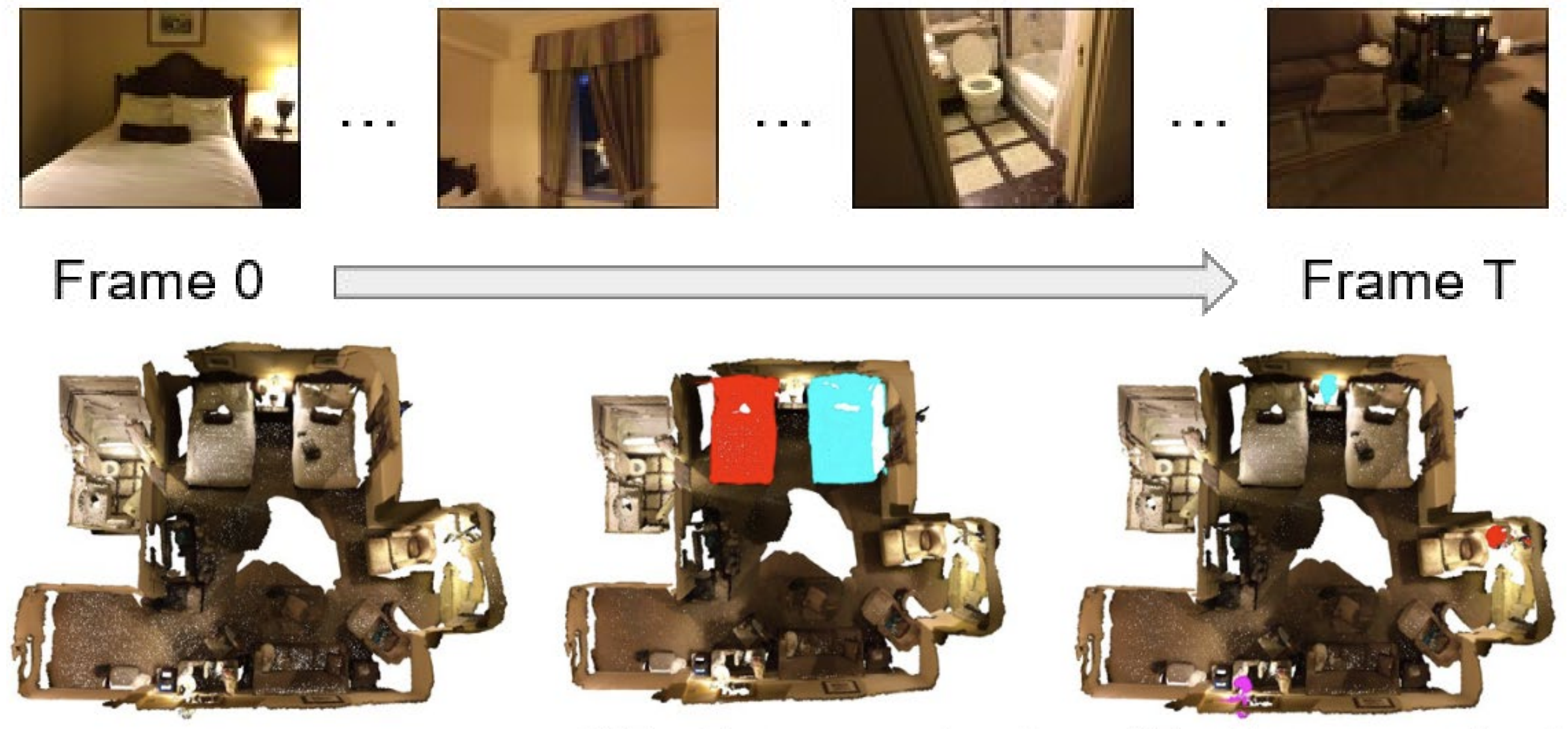
Motivation

		Closed-Set	Open-Vocabulary
Instance Segmentation	2D	Pretty much solved e.g., MaskRCNN, (ICCV'17)	New area with exciting progress e.g., Detic, (ECCV'22)
	3D	Towards mature e.g., Mask3D (ICRA'23)	? Missing

Problem Formulation

Input: a 3D point cloud reconstructed from an RGB-D video and a language query.

Output: a ranked set of 3D instance segments



a) Original scan

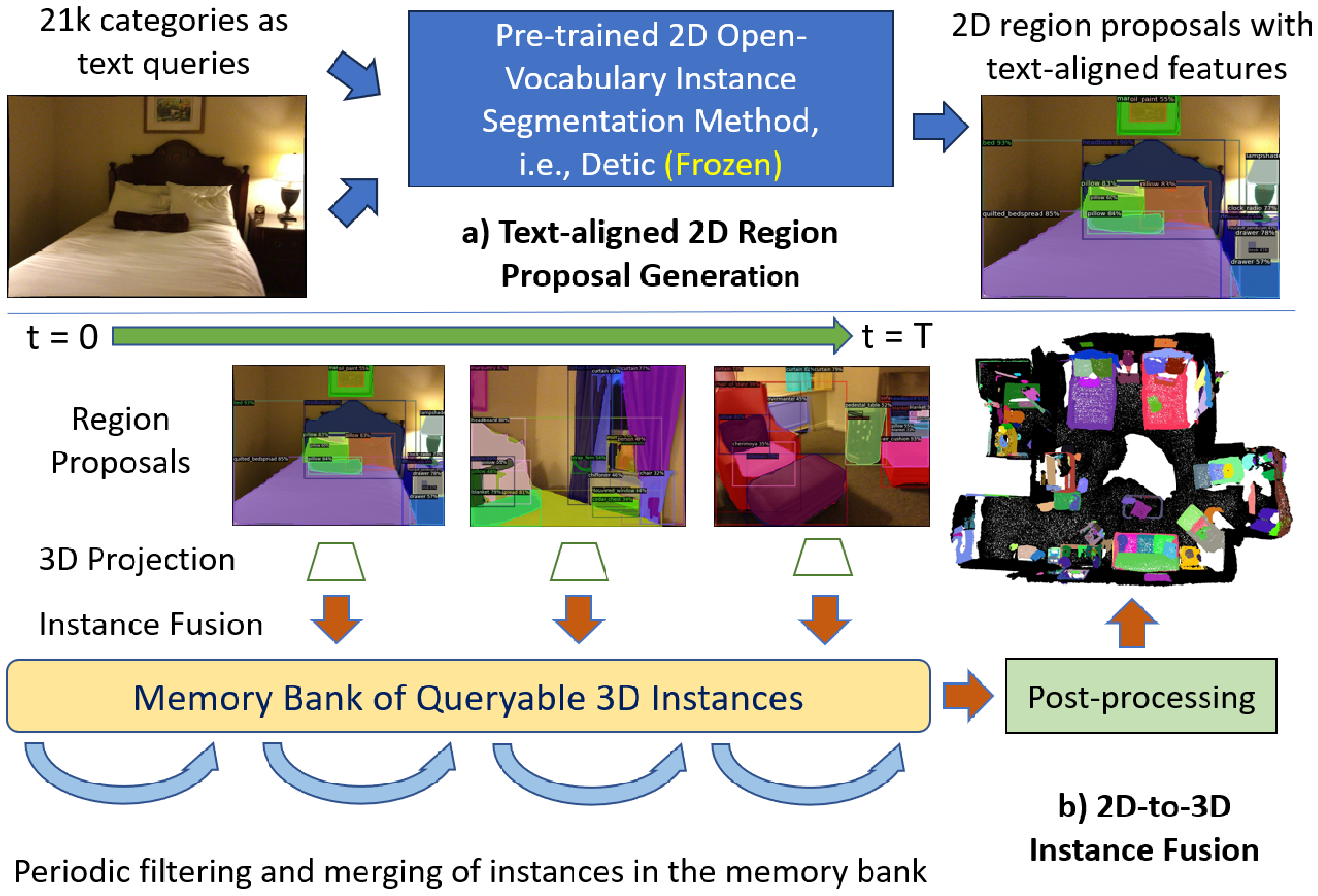
b) Top 2 instances retrieved given the query: *"bed"*

c) Top 3 instances retrieved given the query: *"lamp"*

Key Takeaways

Annotated 3D data with sufficient object diversity is hard to acquire. This problem could instead be viewed as 3D fusion problem from text-aligned 2D region proposals, which can make use of pretrained 2D models.

Overall Pipeline



21k categories as text queries

Pre-trained 2D Open-Vocabulary Instance Segmentation Method, i.e., Detic (Frozen)

2D region proposals with text-aligned features

a) Text-aligned 2D Region Proposal Generation

t = 0 → t = T

Region Proposals

3D Projection

Instance Fusion

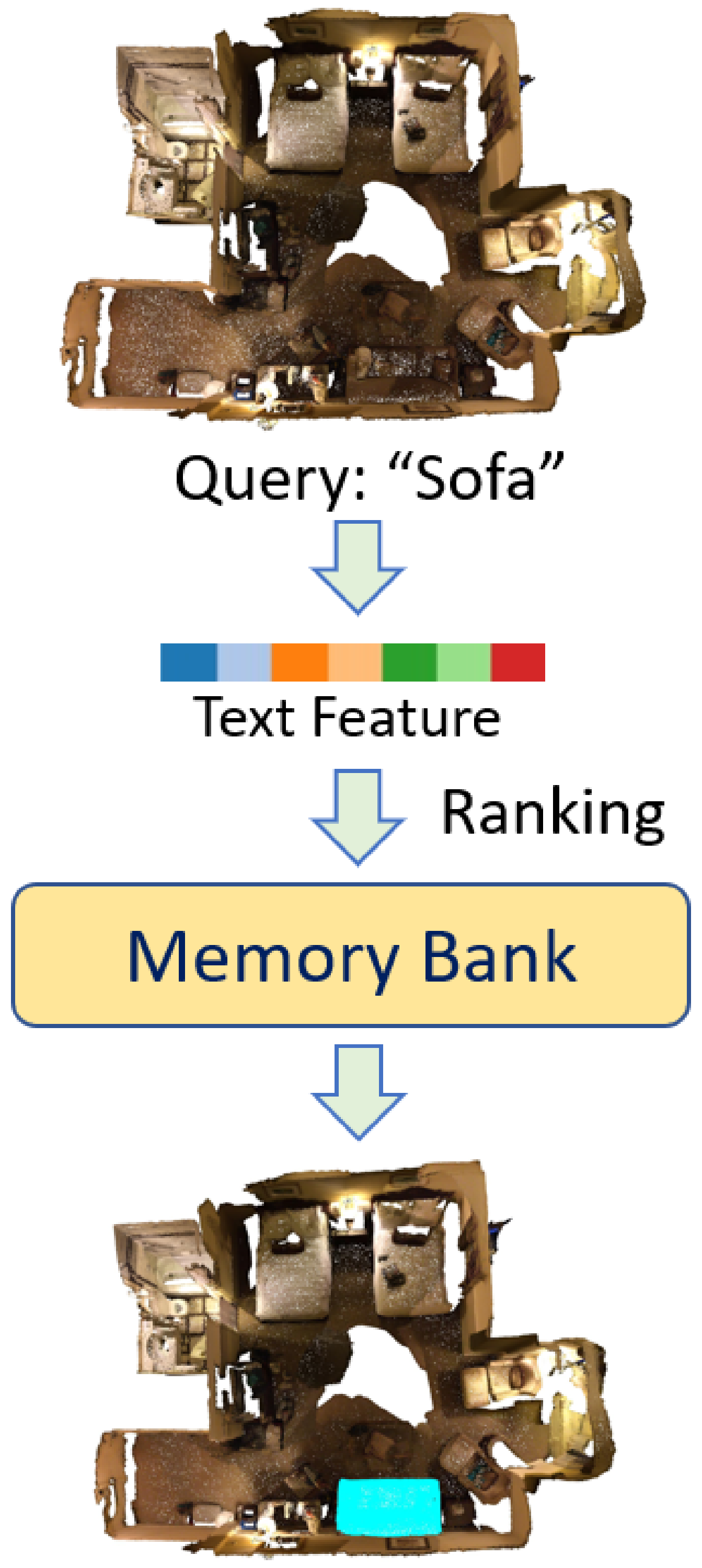
Memory Bank of Queryable 3D Instances

Post-processing

b) 2D-to-3D Instance Fusion

Periodic filtering and merging of instances in the memory bank

Inference



Query: "Sofa"

Text Feature

Ranking

Memory Bank

Top 1 is shown in blue.

Quantitative Results

The proposed method outperforms existing methods on both ScanNet200 (200 classes) and YCB-Video (21 classes) using mAP metric.

	ScanNet200 [25]		YCB-Video [29]	
Method	mAP_{50}	mAP	mAP_{50}	mAP
OpenScene [23]	0.190	0.089	0.333	0.116
Fusion++ [19]	0.253	0.094	0.464	0.120
PanopticFusion [21]	0.370	0.150	0.803	0.393
Ours	**0.443**	**0.211**	**0.848**	**0.465**

Table 1: Results on ScanNet200 [25] and YCB-Video [29]

Ablation Studies

	COCO	ScanNet200	LVIS	ImageNet21k
mAP_{50}	0.228	0.419	0.429	**0.443**
	ImageNet21k - ScanNet200			
mAP_{50}	0.410			

Table 2: Results on ScanNet200 [25] with different input queries to the region proposal network.

	Average	KMeans(16)	KMeans(64)
mAP_{50}	0.428	0.429	**0.443**
	Feature from largest 2D detection		
mAP_{50}	0.380		

Table 3: Results on ScanNet200 [25] with different feature ensemble strategies