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

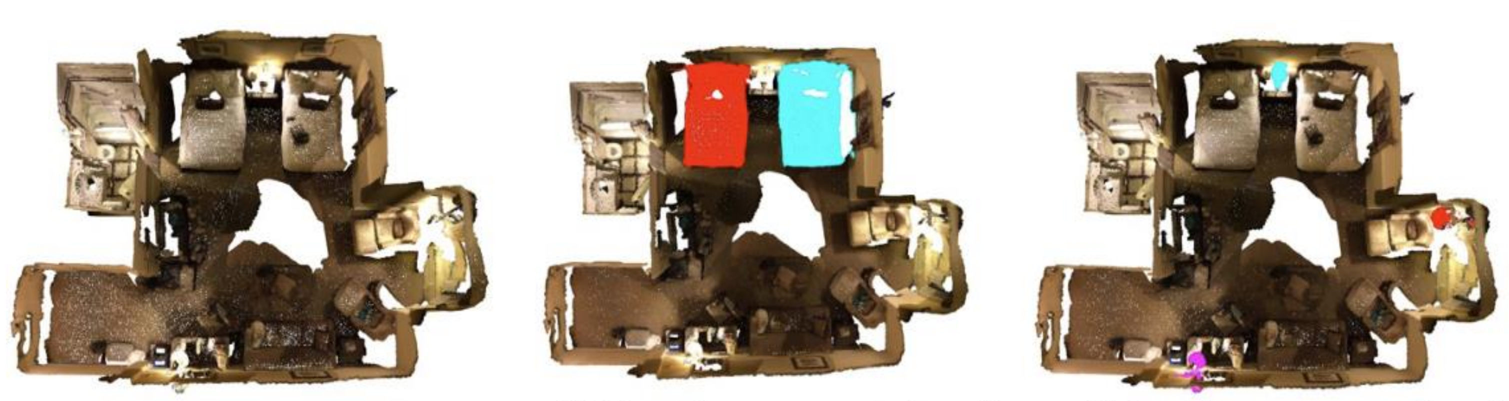
Shiyang Lu[1], Haonan Chang[1], Eric Jing[1], Yu Wu[2], Abdeslam Boularias[1], Kostas Bekris[1]
[1]Rutgers University [2]Wuhan University

RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

To appear at CoRL'23 and ICCV-W'23

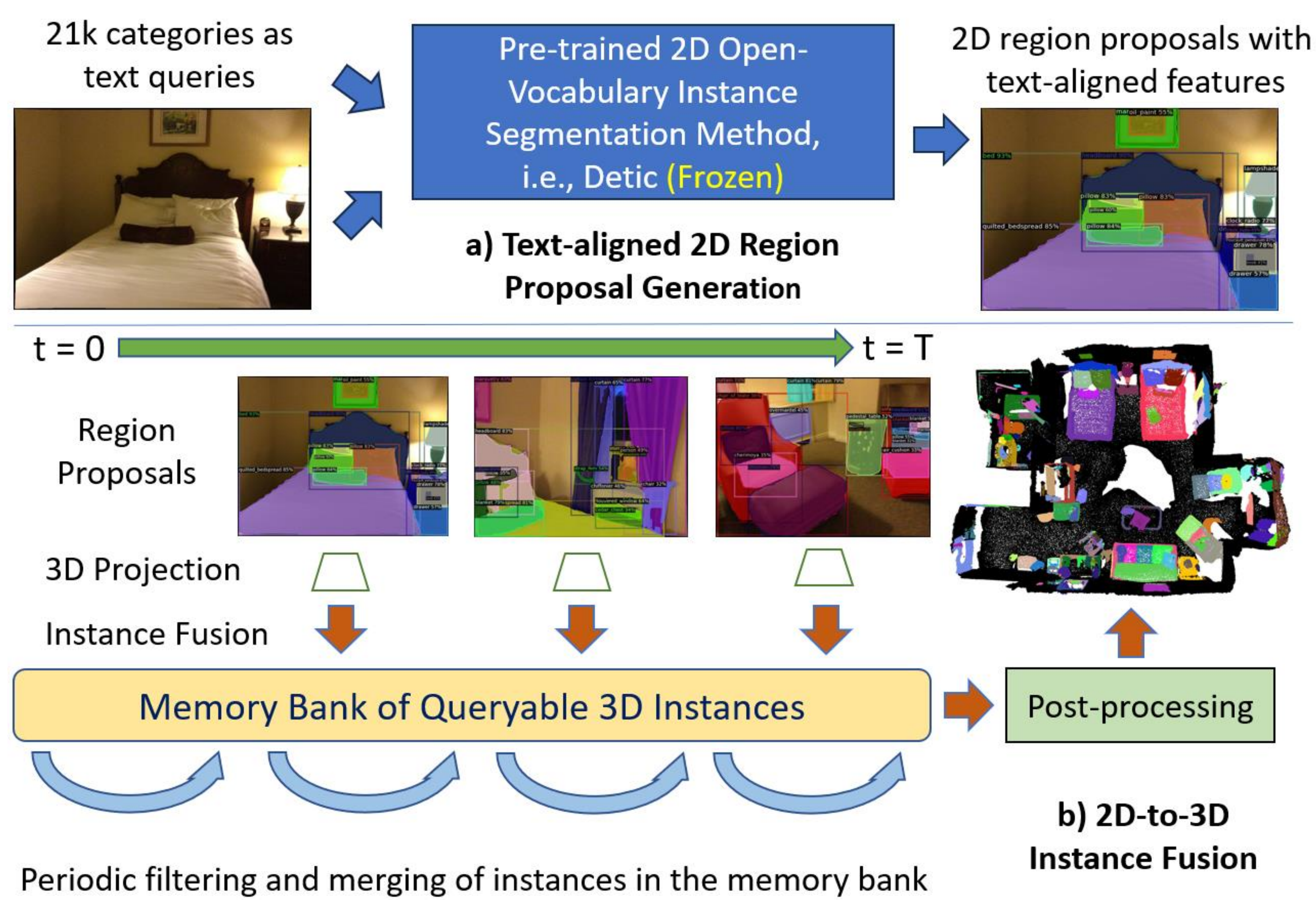
## Introduction

Recent progress on open-vocabulary (language-driven, without a predefined set of categories) 3D segmentation addresses the problem mainly at the semantic level. Nevertheless, robotic applications, such as manipulation and navigation, often require 3D object geometries at the instance level. This work provides a solution for open-vocabulary 3D instance retrieval, which returns a ranked set of 3D instance segments given a 3D point cloud reconstructed from an RGB-D video and a language query.

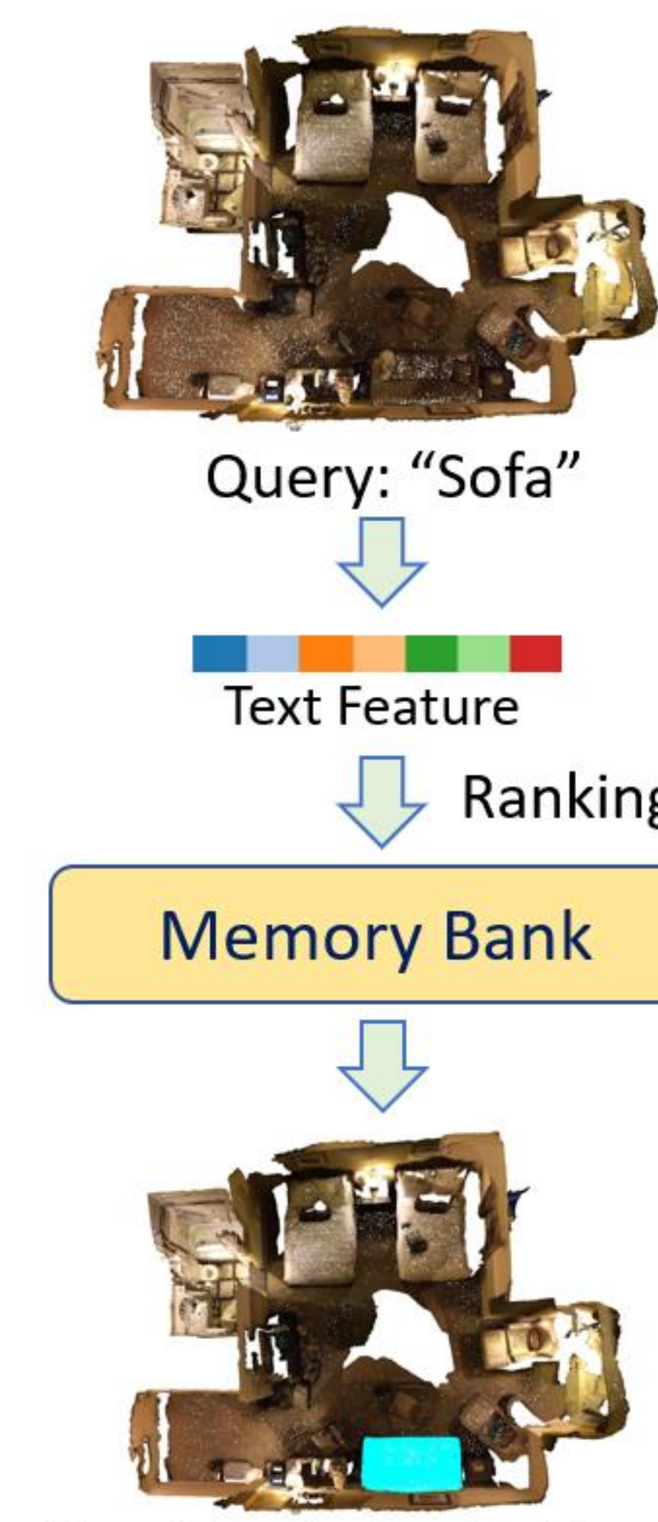


a) Original scan b) Top 2 instances retrieved given the query: "*bed*" c) Top 3 instances retrieved given the query: "*lamp*"

## Overall Pipeline




### Inference




## Key Takeaways

Directly training an open-vocabulary 3D segmentation model is hard due to the lack of annotated 3D data with enough category varieties. Instead, this work views this problem as a 3D fusion problem from language-guided 2D region proposals, which could be trained with extensive 2D datasets, and provides a straightforward yet effective method to project and fused 2D instance information in the 3D space for fast retrieval.

## 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 | 0.378 | 0.136 |
| **Ours** | **0.443** | **0.211** | **0.801** | **0.427** |

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