Beyond Fixed Forms:

Query-Aware Refinement for Open-Vocabulary 3D Instance Segmentation

Zhenghao Zhang, Jie Hu

Paper and Code available here

TL;DR

Previous:
1. Segment whole 3D scene
2. Retrieve the most suitable fixed-shape mask

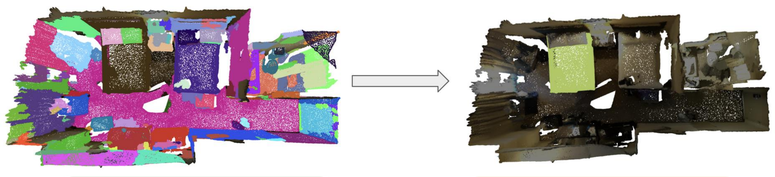
Mask incorrect/ different granularity => Can't find / Bad masks

Ours:
1. Segment whole 3D scene
2. Query-Aware mask from 2D
3. Refine fixed masks from step 1

Good masks! This is what I want.

Background:

Previous Open-Voc 3D instance segmentation



Stage 1: zero-shot 3D instance mask prediction

Fixed-shape masks with associated featuers

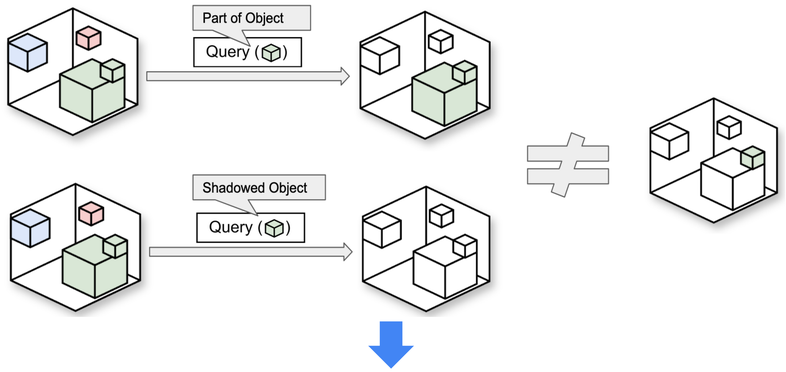
Stage 2: open-vocabulary semantic queries

Choose from **fixed masks**

Text queries are not used to their **full extent**: Current methods treat the query phase as **purely a retrieval process**

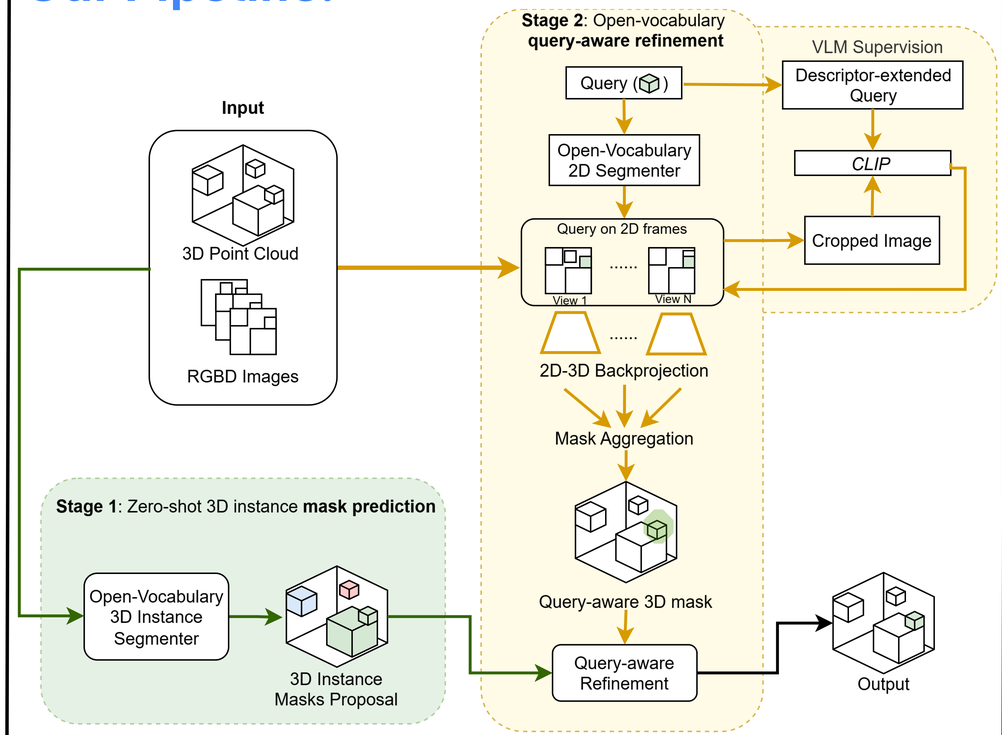
Motivation:

Want to find exactly what you want?
NO, YOU CAN'T



We want a process for *semantic-based*, *query-time* mask *refinement*

Our Pipeline:



Even if the model shadows small objects, there's still **a second chance** for it to be rediscovered utilizing the semantic information entailed by the text prompts.

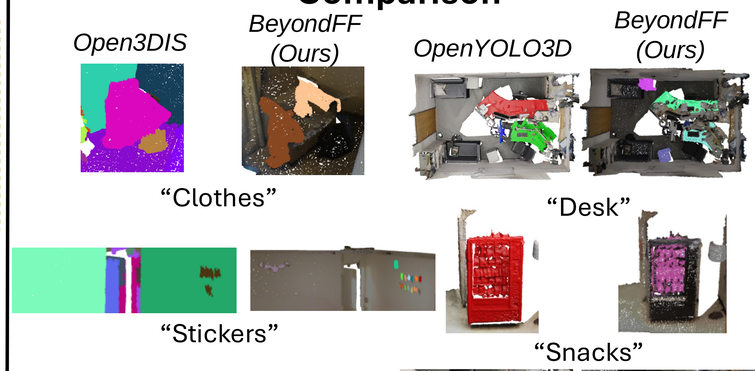
Quantitative Results:

Scannet200 classes*	Methods	AP	AP50	AP25
Overall	Open3DIS	23.7	28.2	31.2
	Open3DIS + BeyondFF	**27.4**	**33.3**	**39.6**
Head	Open3DIS	27.0	32.5	35.5
	Open3DIS + BeyondFF	**29.4**	**36.5**	**42.0**
Common	Open3DIS	21.3	24.9	26.5
	Open3DIS + BeyondFF	**26.9**	**32.0**	**38.6**
Tail	Open3DIS	22.4	26.6	31.1
	Open3DIS + BeyondFF	**25.4**	**30.8**	**37.8**
Base	Open3DIS	23.6	28.7	32.4
	Open3DIS + BeyondFF	**27.8**	**34.6**	**43.2**
Novel	Open3DIS	23.7	28.0	30.8
	Open3DIS + BeyondFF	**27.2**	**32.8**	**38.3**

* Results tested on 120 classes, distributed as follows: Head:40, Com-mon:40, Tail:40, Base:90, Novel:30

Qualitative Results:

Comparison



Arbitrary Text Query