Sparse Multiview Open-Vocabulary 3D Detection

Olivier Moliner [1,2] Viktor Larsson [1] Kalle Åström [1]

[1]Lund University [2]Sony Corporation, Lund Laboratory

Background

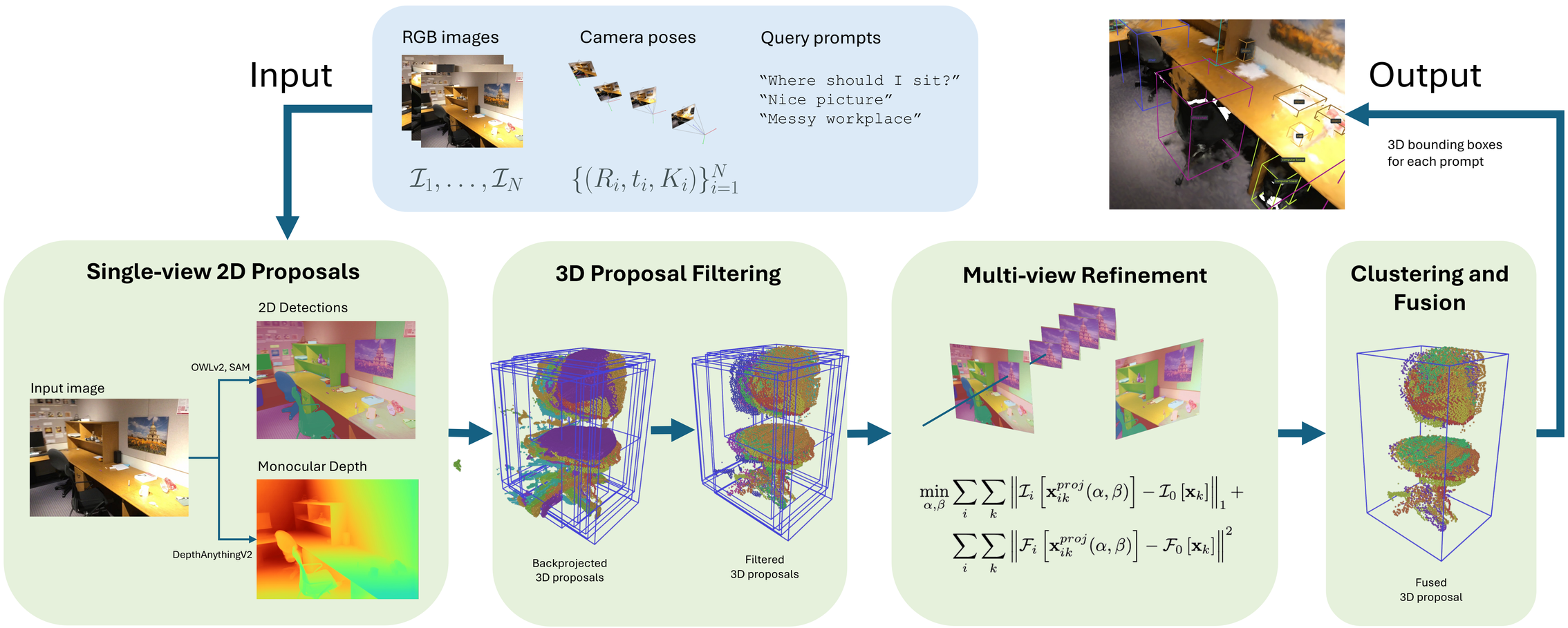
Conventional 3D detection methods often rely on predefined, closed sets of object categories, requiring expensive retraining and new annotations to adapt to new domains. Recent advances in vision-language models have enabled open-vocabulary detection in 2D., but applying these approaches to 3D detection typically requires dense 3D point clouds constructed from RGB-D sensors or extensive multi-view imagery. Existing methods are also constrained by the availability of high-quality 3D data, limiting their practicality in many real-world scenarios.

We propose a novel, training-free framework for open-vocabulary 3D object detection using sparse multi-view RGB images.

Contributions

- We introduce a novel, training-free framework for open-vocabulary 3D object detection using sparse multi-view RGB images.
- We leverage pre-trained, off-the-shelf 2D networks for proposal generation and monocular depth estimation.
- We propose a three-step pipeline involving single-view proposal generation, multi-view refinement using feature consistency, and 3D bounding box fusion.
- Our method achieves competitive results without relying on 3D data and generalizes to unseen, long-tail vocabularies.

Approach



Overview of SMOV3D. Our method takes as input a sparse collection of posed RGB images together with a collection of text query prompts. The pipeline then conists of three steps. **i) Monocular 2D Proposals** For each prompt and image we perform 2D detection yielding a set of masks. These are then lifted to 3D using monocular depth. **ii) Multi-view Refinement** Lifted 3D point clouds are refined by optimizing a multi-view featuremetric loss that combines both photometric and CLIP consistency. **iii) 3D Clustering and Fusion.** The optimized 3D point clouds are aggregated in 3D and greedily fused using a simple heuristic. The output is a collection of 3D bounding boxes. For visualization we show them overlayed on the ground-truth mesh.

Selected Results

Figure 1:**Qualitative results on ScanNet.** Zero-shot 3D object detect detection, using the ScanNet10 categories as prompts.

Method	GT Depth	3D proposal	Mean	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink
OV-3DETIC [2]	✓	3DETR†	12.7	49.0	2.6	7.3	18.6	2.8	14.3	2.4	4.5	3.9	21.1
Object2Scene [4]	✓	L3DET†	24.6	56.3	36.2	16.1	23.0	8.1	23.1	14.7	17.3	23.4	27.9
FM-OV3D [3]	✓	3DETR†	21.5	55.0	38.8	19.2	41.9	23.8	3.5	0.4	6.0	17.4	8.8
OpenIns3D [1]	✓	Mask3D†	**43.7**	**79.5**	**70.5**	**76.9**	15.8	0.0	**53.1**	**40.1**	**41.2**	7.1	53.1
SMOV3D (Ours) RGB-D	✓	-	39.2	78.7	25.3	30.2	69.4	23.8	26.2	4.7	22.4	**58.1**	**53.7**
SMOV3D (Ours) Metric3Dv2†	✗	-	38.3	61.8	33.6	28.4	**71.1**	**37.5**	29.7	1.3	18.5	52.7	48.3
SMOV3D (Ours) DepthAnythingv2	✗	-	30.8	61.3	26.4	18.2	61.3	32.0	23.2	2.2	16.3	37.5	30.0

Table 1:**Open-vocabulary Object Detection on ScanNet10.** We compare our method to point cloud-based methods. "†" denotes a component trained on ScanNet.

Method	Head	Common	Tail
Object2Scene [4]	-	10.1	3.4
OpenIns3D [1] with RGB-D	**25.6**	20.4	16.5
SMOV3D (Ours) RGB-D	21.0	**27.6**	**23.2**
SMOV3D (Ours) Metric3Dv2	18.4	18.6	22.6
SMOV3D (Ours) DepthAnythingv2	11.7	18.9	11.9

Table 2:**Open-vocabulary Object Detection on ScanNet 200.** We present results (mAP_{25}) for the Head, Common and Tail category splits. Our method performs just as well on long-tail classes (Tail) as on the most frequent ones (Head).

References

[1] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. European Conference on Computer Vision, 2024.

[2] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary 3d detection via image-level class and debiased cross-modal contrastive learning. arXiv pre-print, 2022.

[3] Dongmei Zhang, Chang Li, Ray Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang Zhang. Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. In AAAI, 2024.

[4] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. In arXiv, 2023.

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM