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
https://github.com/shiyoung77/OVIR-3D

Abstract

This work presents OVIR-3D, a straightforward yet effective method for open-vocabulary 3D object instance retrieval without using any 3D data for training. Given a language query, the proposed method returns a ranked set of 3D object instance segments based on the feature similarity of the instance and the language query. This is achieved by fusing text-aligned 2D region proposals from multiple views into 3D space. The 2D region proposal network can leverage 2D datasets, which are more readily available and typically larger than 3D datasets benefiting the method's performance. The proposed fusion process can be performed in real-time for most indoor 3D scenes and does not require additional training in 3D space. Experiments on public datasets show the effectiveness of the method and its potential for applications in robot navigation and manipulation.

1. Introduction

There has been recent progress in open-vocabulary 2D detection and segmentation methods [5, 34, 16] that rely on pre-trained vision-language models [24, 10, 31]. However, their counterparts in the 3D domain have not been extensively explored. One reason is the lack of large 3D datasets with sufficient object diversity for training open-vocabulary models. Early approaches for dense semantic mapping [20, 26, 19, 21] project multi-view 2D detections to 3D using closed-set detectors but cannot handle arbitrary language queries. More recently, OpenScene [23] achieves open-vocabulary 3D semantic segmentation by projecting text-aligned pixel features to 3D points and distilling 3D features from the aggregated 2D features. During inference, given a text query, a heatmap of the point cloud will be generated based on the similarity between point features and the query feature. Nevertheless, manual thresholding is required for object search to convert the heatmap to a binary mask and it lacks the ability to separate instances from the same category. This limits its use in robotic applications, such as automated manipulation and navigation.

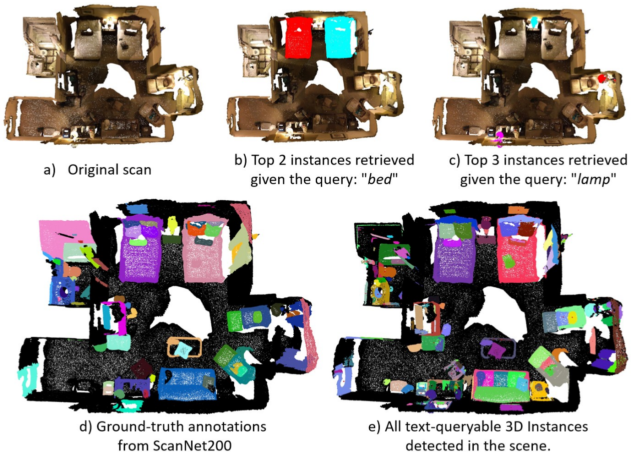
a) Original scan
b) Top 2 instances retrieved given the query: "*bed*"
c) Top 3 instances retrieved given the query: "*lamp*"
d) Ground-truth annotations from ScanNet200
e) All text-queryable 3D Instances detected in the scene.

Figure 1: **Examples of open-vocabulary 3D instance retrieval.** (a-c) Given a 3D scan (e.g., scene0645 from ScanNet [1]) and a text query (e.g., "bed", "lamp"), the proposed method retrieves a set of 3D instances ranked based on their semantic similarity to the text query. (d-e) Instances that are not even in the ground-truth annotations can also be detected and queried by the proposed method, such as the cushions on the sofa.

This work focuses on open-vocabulary 3D instance retrieval, aiming to return a ranked set of 3D instance segments based on their semantic similarities to a given text query. Instead of fusing pixel-level information and then either grouping them into instances by thresholding at inference time [23] or training an object identification model with additional ground truth data [28], the proposed method directly fuses instance-level information into the scene without additional training, so that given a text query such as "lamp", a robot can immediately locate the top-related object instances and perform required tasks.

The proposed method first generates 2D object region proposals and their corresponding text-aligned features by querying a 2D open-vocabulary detector with an extensive vocabulary. Data association and periodic filtering and merging of 3D instances are performed to improve instance masks and remove noisy detections. Finally, a post-processing step handles isolated objects. Extensive experiments on real scans demonstrate the effectiveness of the proposed method, which offers an efficient 2D-to-3D instance fusion module ($\sim$ 30 fps for a scene in ScanNet[1]) and an open-vocabulary 3D instance retrieval method with near-instant inference time for a text query.

The main contributions of this work are: (i) an efficient 2D-to-3D instance fusion module given text-aligned region proposals, which results in (ii) an open-vocabulary 3D instance retrieval method that ranks 3D instances based on semantic similarity given a text query.

2. Related Work

2D Open-Vocabulary Detection and Segmentation Several methods have been proposed for 2D open-vocabulary object detection and segmentation utilizing large vision-language pre-trained models like CLIP [24], ALIGN [10], and LiT [32][14, 4, 30, 5, 34, 18, 15]. For 2D semantic segmentation, LSeg[14] aligns pixel features with segment label embeddings, while OpenSeg [4] uses image-level supervision. GroupViT [30] performs hierarchical spatial grouping, and ViLD [5] aligns class-agnostic region proposals with text label features. Detic [34] addresses long-tail detection, and OWL-ViT [18] transfers image-text models to object detection. The proposed method employs Detic [34] as a backbone detector for 2D object localization, providing pixel-level instance segmentation and text-aligned features.

3D Scene Understanding Early work on 3D semantic mapping [20, 26, 21, 19, 13] demonstrates impressive results with closed-set 2D object detection and segmentation. However, they are not designed to fit open-vocabulary detectors and cannot be queried with language without revision. Recent efforts focus on open-vocabulary 3D scene understanding [28, 23, 33, 8, 9, 11, 3]. Methods like OpenScene [23] project features from 2D open-vocabulary segmentation models to 3D reconstructions, while ConceptFusion [9] fuse multi-modal features. LeRF [11] combines multi-scale CLIP features for open-vocabulary queries. These methods lack instance-level segmentation. PLA [3] aims for instance segmentation but only has limited demonstrations on furniture-level instances. In contrast, the proposed method focuses on instance-level, open-vocabulary 3D segmentation without manual 3D annotation.

3. Problem Formulation

A 3D scan $\mathcal{X}^N$ represented by N points is reconstructed from an RGB-D video $\mathcal{V} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_T\}$ given known camera intrinsics C and camera poses P_t, where $\mathcal{I}_t$ is the video frame at time t. The objective in open-vocabulary 3D instance retrieval is to return a list of K ranked instances represented as binary 3D masks $\mathcal{M}^N = \{m_i | i \in [1, K]\}$ over the 3D scan $\mathcal{X}^N$, given a text query Q and the desired number of instances K to be retrieved, where the ranking is based on the semantic similarity between the 3D instance and the text query.

4. Method

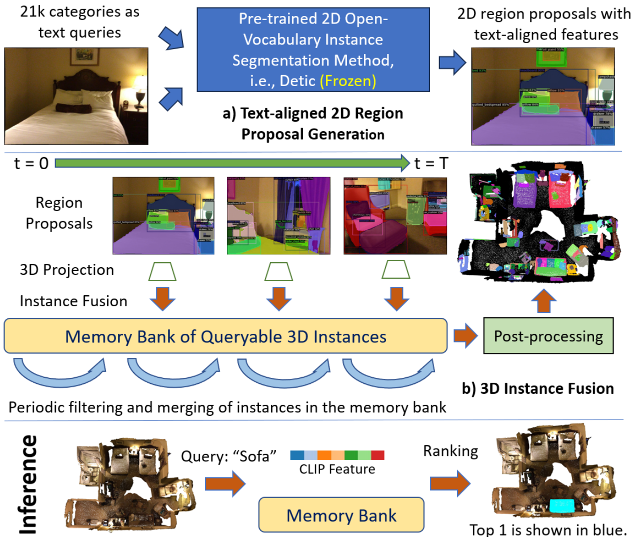
21k categories as text queries
Pre-trained 2D Open-Vocabulary Instance Segmentation Method, i.e., Detic (Frozen)
2D region proposals with text-aligned features
a) Text-aligned 2D Region Proposal Generation
t = 0
t = T
Region Proposals
3D Projection
Instance Fusion
Memory Bank of Queryable 3D Instances
Post-processing
b) 3D Instance Fusion
Periodic filtering and merging of instances in the memory bank
Inference
Query: "Sofa"
CLIP Feature
Ranking
Memory Bank
Top 1 is shown in blue.

Figure 2: **Pipeline of the proposed method.**

The overall pipeline of the proposed method is illustrated in figure 2. To summarize, given a video frame, the method first generates 2D region proposals $\mathcal{R}^{2D} = \{r_1, .., r_k\}$ with text-aligned features $F^{2D} = \{f_1^{2D}, .., f_k^{2D}\}$ using an off-the-shelf 2D open-vocabulary method trained on large 2D datasets. The 2D region proposals $\mathcal{R}^{2D}$ of each frame $\mathcal{I}_t$ are then projected to the reconstructed 3D point cloud given the camera intrinsics C and poses P_t. The projected 3D regions $\mathcal{R}^{3D}$ are either matched to existing 3D object instances $O = \{o_1, .., o_b\}$ with 3D features $F^{3D} = \{f_1^{3D}, .., f_b^{3D}\}$ stored in the memory bank $\mathcal{B}$, or added as a new instance if not matched with anything. The 2D region to 3D instance matching is based on feature similarity $s_{ij} = cos(f_i^{2D}, f_j^{3D})$ and region overlapping $IoU(r_i^{3D}, o_j)$ in the 3D space. Matched regions are integrated into the 3D instance. To remove unreliable detections and improve segmentation quality, periodic filtering and merging of 3D instances in the memory bank $\mathcal{B}$ is performed every T frames. A final post-processing step removes 3D instances that are too small and separates object instances that are isolated in 3D space but incorrectly merged. During inference time, the text query q will be used to match with a set of representative features of each 3D instance, and the instances O will be ranked based on the similarity and returned. Details of the proposed method are presented below.

4.1. Text-aligned 2D Region Proposal

Learning-based region proposal networks have served as a critical module for many instance segmentation methods, such as MaskRCNN [7]. Directly generating 3D region proposals for open-vocabulary instance retrieval, however, is hard due to the lack of annotated 3D data with enough category varieties, and most existing region proposal networks cannot provide features directly for text query. This work leverages the power of an off-the-shelf open-vocabulary 2D detector, i.e. Detic [34], to generate 2D region proposals $\mathcal{R}^{2D}$ by querying a huge number of categories, i.e. all 21k categories from ImageNet21k [2] dataset. Instead of using the category labels of predicted instances, which could be rather noisy given the vocabulary size, their text-aligned features F^{2D} are extracted before the classification layer. Both region proposals $\mathcal{R}^{2D}$ and their text-aligned features F^{2D} are used for data association in the fusion step.

This work adopts Detic [34] as a region proposal network instead of SAM [12] mainly because it can output text-aligned features for predicted instances, which turns out to be critical for data association in the experiments. In addition, Detic [34] has a much faster inference speed than SAM [12] even when queried with all the categories from ImageNet21k [2] ($\sim$ 10fps on an NVIDIA RTX3090).

4.2. 2D-to-3D Instance Fusion

2D region proposals $R^{2D} = \{r_1^{2D}, .., r_k^{2D}\}$ and their corresponding features $F^{2D} = \{f_1^{2D}, .., f_k^{2D}\}$ for each frame $\mathcal{I}_t$ are first projected to the 3D scan using camera intrinsics C and pose P_t. The projected 3D regions $\mathcal{R}^{3D}$ are either matched to existing 3D object instances $O = \{o_1, .., o_b\}$ with 3D features $F^{3D} = \{f_1^{3D}, .., f_b^{3D}\}$, where b is the number of 3D instances already stored in the memory bank $\mathcal{B}$, or added as a new instance if it is not matched with anything. The memory bank is empty at the beginning.

The matching of 2D region r_i to 3D instance o_j is based on cosine similarity $s_{ij} = cos(f_i^{2D}, f_j^{3D})$ and 3D intersection over union between the projected region r_i^{3D} and visible part of the 3D instance $\hat{o_j}$ in the current frame, i.e., $IoU(r_i^{3D}, \hat{o_j})$. If s_{ij} is greater than a predefined threshold $\theta_s = 0.75$ and the overlapping $IoU(r_i^{3D}, \hat{o_j})$ is also greater than predefined threshold $\theta_{iou} = 0.25$, then they are considered as a match. Matched regions will be aggregated to the 3D instance, i.e., $o_j := o_j \cup r_i^{3D}$ and $f_j^{3D} := f_j^{3D} + f_i^{2D}$. The matching is not restricted to one-to-one as multiple 2D region proposals may correspond to the same instance.

4.3. Periodic 3D Instance Filtering and Merging

The fusion process generates redundant 3D instances when a 2D region proposal fails to match properly, potentially leading to low-quality segmentation and inaccurate data association. To address this, periodic filtering and merging of 3D instances stored in memory bank $\mathcal{B}$ occur every $T = 300$ frames. Filtering is based on the detection rate r_p^{vis} of a point p, where $r_p^{vis} = c_p^{o_i}/c_p^{vis}$, and points with $r_p^{vis} < \theta_{vis} = 0.2$ are removed from instance o_i. If an instance o_i contains fewer than 50 points after filtering, it is filtered entirely.

Merging of two instances o_p, o_q is determined by feature similarity $s_{pq} = cos(f_p^{3D}, f_q^{3D})$ and 3D intersection over union $IoU(o_p, o_p)$, using the thresholds $\theta s = 0.75$ and $\theta iou = 0.25$. Additionally, instances o_p and o_q are merged if $recall(o_p, o_q) = |o_p \cup o_q|/|o_q| \geq \theta_{recall}$ and $s_{pq} \geq \theta_s$, indicating that o_q is mostly contained in o_p and both instances have similar features.

Hyper-parameters are justified through ablation studies in Section 6, and these values remain fixed for experiments in Section 5 across different datasets.

4.4. Post-processing

A simple post-processing step is executed to separate object instances that are isolated in 3D space and filter small segments that are likely to be noise. This is achieved by using DBSCAN [27] to find 3D point clusters in each instance, where the distance parameter eps is set to $10cm$. If an instance o_i has segments not connected in 3D space, DBSCAN will return more than one point cluster and o_i will be separated in multiple instances. Small clusters with less than 50 points are filtered out.

4.5. Inference

During inference time, a text query q is converted to a feature vector f_q using CLIP [24]. Instead of representing each 3D instance with the average feature of associated 2D regions, the K clustering centers by K-Means of associated features, which can be viewed as representative features from a set of viewpoints, are used. The 3D instances are then ranked by the largest cosine similarity s between the text query q and K representative features of an instance.

5. Experiments

Datasets. The first dataset used for the experiment is ScanNet200 [25], which contains a validation set of 312 indoor scans with 200 categories of objects. Uncountable categories ”floor”, ”wall”, and ”ceiling” and their subcategories are not evaluated. The second dataset is YCB-Video [29], which contains a validation set of 12 videos of tabletop scenes. The 3D scans of the tabletop scenes are reconstructed by KinectFusion [22]. The ground truth instance segmentation labels are automatically generated given the object mesh models and annotated 6DoF poses.

Metrics. Standard mean average precision (mAP) metric for instance retrieval is adopted for the evaluation purpose. In particular, mAP_{50} and the overall mAP, i.e $\frac{1}{10}\sum mAP_\theta$, where θ = [0.5:0.05:0.95] are reported. Only annotated object categories in a 3D scene are used as text queries for evaluation. The results were computed for each 3D scene and then averaged for the whole dataset.

Baselines. OpenScene [23], which is the most relevant work to date, is used as the first comparison point. Given an object query, it returns a heatmap of the input point cloud. A set of thresholds θ = [0.5:0.03:0.9] are tested for each category to convert the heatmap into a binary mask and then foreground points are clustered into 3D instances using DBSCAN, similar to the post-processing step in section 4.4. The one with the best overall performance is reported. Furthermore, a series of prior research has focused on semantic

	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]

mapping using closed-vocabulary detectors. Two representative works, Fusion++ [19] and PanopticFusion [21], are used as comparison points with two revisions: 1) Instead of using their whole SLAM system, this work assumes the 3D reconstruction and ground truth camera poses are given, and only tested their data association and instance mapping algorithms. 2) Their backbone detector MaskRCNN [7] is replaced with Detic [34] for open-vocabulary detection, and the mean feature of associated 2D detections for each instance is used to match text queries.

Results. Quantitative results on ScanNet200 [1] and YCB-video [29] dataset are shown in Table 1. The proposed method outperforms all other baselines by a large margin in terms of instance retrieval mAP. It seems that OpenScene[23] does not perform well on this task even with an automatically tuned threshold for each category because fused point features are not distinguishable enough. As a result, grouping points into segments with accurate boundaries by thresholding is rather difficult. The proposed method, on the other hand, directly fuses instance-level information and improves segment quality by periodic merging and filtering. The proposed method outperforms the other two baselines primarily because of the use of instance features for data association as the baselines only consider 3D overlapping, which can easily fail when the 2D detections are noisy, especially in the open-vocabulary setup.

6. Ablation Studies

Input queries of the 2D region proposal method The proposed method utilizes an open-vocabulary 2D detector as a region proposal method by querying it with an extensive vocabulary. This ablation study tests queries from multiple datasets as input to the region proposal method and displays their impact on the overall performance. In addition to ScanNet200 [25] and ImageNet21K [2], COCO [17] (80 categories), LVIS [6] (1203 categories), and more aggressively, queries with ImageNet21k categories but without ScanNet200 categories are tested. Results of 3D instance retrieval on the ScanNet200 dataset in Table 2 show that an extensive vocabulary is helpful and the region proposal network has certain generalizability, such that even when ScanNet200 categories are completely removed from the ImageNet21k categories, it can still find most regions based on similar categories and the performance on retrieving objects in ScanNet200 only slightly dropped.

	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.

Feature ensemble strategies. Three different feature ensemble strategies are tested to represent a 3D instance based on associated 2D features. The first strategy is to compute the average of all 2D features. The second strategy involves clustering the 2D features from different viewpoints using the K-Means algorithm, and the clustering centers are used to represent each instance. During instance retrieval, the feature similarity is determined as the maximum similarity between the query feature and the clustering centers. The third strategy is to use the feature from the largest associated 2D region. Results of 3D instance retrieval on the ScanNet200 dataset are presented in Table 3. The approach of using multiple features through clustering outperforms simple averaging, while using the feature from the largest associated 2D region yields the poorest results.

	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

Time intervals and visibility threshold for periodic instance filtering and merging. Results of 3D instance retrieval with different time intervals T and visibility threshold θ_{vis} mentioned in section 4.3 are presented in Table 4 and Table 5 respectively. The frame interval $T = 300$ and visibility threshold $\theta_{vis} = 0.2$ yields the best results.

T	1	100	300	500	1000
mAP_{50}	0.340	0.417	**0.443**	0.410	0.412

Table 4: Results on ScanNet200 [25] with different time intervals of periodic filtering and merging

θ_{vis}	0	0.1	0.15	0.2	0.25	0.3
mAP_{50}	0.256	0.386	0.407	**0.443**	0.418	0.408

Table 5: Results on ScanNet200 [25] with different visibility thresholds of periodic filtering

7. Conclusion and Limitations

This work presents OVIR-3D, a straightforward yet effective method for open-vocabulary 3D object instance retrieval without using any 3D data for training. A limitation of the proposed method is that it can miss very small objects (<50 points), as they are likely to be treated as noise and filtered out during fusion. Furthermore, while the proposed method can improve segmentation quality due to multi-view noise filtering, it still relies on a good 2D region proposal network that does not constantly miss certain objects or provide bad segmentation. A promising direction is to integrate this method with a 3D learning-based method to utilize the scarcer but clean 3D annotations.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 2, 4

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3, 4

[3] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Language-driven open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2211.16312*, 2022. 2

[4] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 540–557. Springer, 2022. 2

[5] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*. 1, 2

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 4

[8] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:*, 2022. 2

[9] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *arXiv*, 2023. 2

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[11] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023. 2

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[13] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. *arXiv preprint arXiv:2302.01838*, 2023. 2

[14] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[15] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[16] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022. 1

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[18] Austin Stone Maxim Neumann Dirk Weissenborn Alexey Dosovitskiy Aravindh Mahendran Anurag Arnab Mostafa Dehghani Zhuoran Shen Xiao Wang Xiaohua Zhai Thomas Kipf Neil Houlsby Matthias Minderer, Alexey Gritsenko. Simple open-vocabulary object detection with vision transformers. *ECCV*, 2022. 2

[19] John McCormac, Ronald Clark, Michael Bloesch, Andrew Davison, and Stefan Leutenegger. Fusion++: Volumetric object-level slam. In *2018 international conference on 3D vision (3DV)*, pages 32–41. IEEE, 2018. 1, 2, 4

[20] John McCormac, Ankur Handa, Andrew Davison, and Stefan Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and automation (ICRA)*, pages 4628–4635. IEEE, 2017. 1, 2

[21] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019. 1, 2, 4

[22] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136. Ieee, 2011. 3

[23] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 1, 2, 3, 4

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[25] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In

Proceedings of the European Conference on Computer Vision (ECCV), 2022. 3, 4

[26] Martin Runz, Maud Buffier, and Lourdes Agapito. Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 10–20. IEEE, 2018. 1, 2

[27] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017. 3

[28] Anonymous Submission. Clip-fields: Weakly supervised semantic fields for robotic memory. *RSS 2023*, 2023. 1, 2

[29] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018. 3, 4

[30] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022. 2

[31] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[32] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[33] Junbo Zhang, Runpei Dong, and Kaisheng Ma. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. *arXiv preprint arXiv:2303.04748*, 2023. 2

[34] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 1, 2, 3, 4