OpenLex3D: A New Evaluation Benchmark for Open-Vocabulary 3D Scene Representations

Christina Kassab[*1], Sacha Morin[*2,4], Martin Büchner[*3], Matías Mattamala[1], Kumaraditya Gupta[2,4], Abhinav Valada[3], Liam Paull[2,4,5], Maurice Fallon[1]

[1]University of Oxford [2] Université de Montréal [3]University of Freiburg
[4]Mila - Quebec AI Institute [5]Canada CIFAR AI Chair

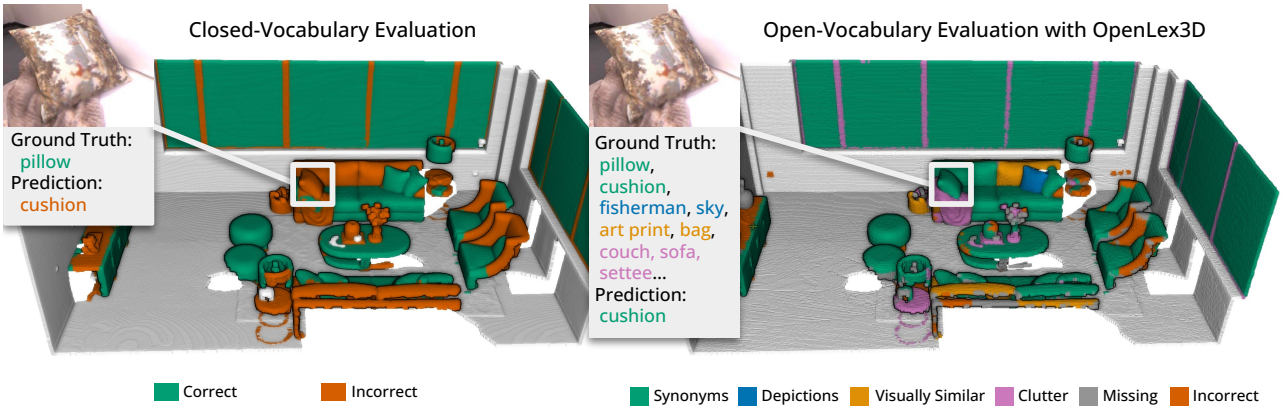

Fig. 1: **The OpenLex3D evaluation benchmark enables detailed analysis of open-vocabulary 3D scene representations compared to closed-vocabulary evaluation methods.** We compare the same open-vocabulary representation when assessed under closed-vocabulary semantics (left) and using OpenLex3D labels (right). In contrast to closed-vocabulary methods where a prediction must match the exact ground truth label, OpenLex3D provides a manifold of label categories of varying precision: *synonyms* being the most precise; *depictions*, which include, e.g., printed images on objects; *visually similar*, which refer to objects with comparable appearance; and *clutter*, which accounts for label perturbation due to imprecise segmentation.

***Abstract*— We present a new evaluation benchmark for open-vocabulary scene representations, generating novel label sets for widely used RGB-D datasets. This enables detailed analysis of failure cases and potential improvements. The benchmark is publicly available at: `https://openlex3d.github.io`.**

I. Introduction

Evaluating fixed-class models is relatively straightforward—the predicted class label of each point-wise prediction can be compared with the point's ground truth label as shown in Fig. 1. In contrast, assessing the performance of open-vocabulary models is more challenging and is not yet well defined by a benchmark. Published works on open-vocabulary representations [13], [12] have typically used closed-set semantic segmentation labels and metrics—despite this underlying mismatch. This defeats the purpose and flexibility of open-vocabulary predictions by constraining the model assessment to a limited set of evaluation classes [12], [8].

Furthermore, relying on evaluation benchmarks designed for fixed-class semantics overlooks the nuance of real-world language labeling, which rarely lends itself to binary classification. Examples of this include synonyms or other descriptive words that encompass more general classes. For example, a couch might also be referred to as a "sofa" or "seating". Prior work has proposed using existing ontologies like WordNet (a large English language lexical database) [7] to mitigate these ambiguities in language benchmarks, though differentiating between *similarities* and *associations* remains an open question [4].

Some open-vocabulary challenges have sought to evaluate performance in a *functional* manner, by focusing on tasks such as visual question answering and object retrieval [1], [14], [2]. However, methods that focus on querying specific prompts will fail to evaluate the full 3D representation and offer limited insights into overall limitations.

In this work, we aim to overcome these limitations by introducing OpenLex3D, a novel benchmark that evaluates open-vocabulary scene representation methods. OpenLex3D introduces a procedure built on *four different label categories* of description accuracy: *synonyms*, *depictions*, *visual similarity*, and *clutter*. These categories evaluate the performance of a method in capturing the correct class (*synonyms*) while also quantifying different degrees of misclassifications. Our contributions are:

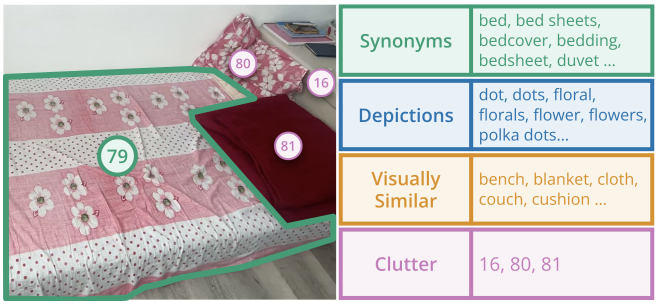

Fig. 2: **OpenLex3D augmented labels example from ScanNet++ [15]**. We provide not only synonyms for the object (bed sheet, duvet) but also labels for various potential failure cases.

Open-set Category Labels. We introduce a new labeling scheme where each object has multiple free-form text labels organized into four categories with different accuracy levels of linguistic description.

Dataset. We provide OpenLex3D labels for 23 scenes from Replica [11], Scannet++ [15] and Habitat-Matterport 3D [9]. Each object has been reviewed by four human annotators, resulting in an average of 11 labels per object.

Two Evaluation Tasks. We provide evaluation on two tasks using the OpenLex3D dataset: semantic segmentation and object retrieval given a text query. We introduce two novel open-set metrics for segmentation and an extended query list for object retrieval. We evaluate state-of-the-art 3D open-vocabulary methods on both tasks.

Benchmark Tookit. We make the OpenLex3D toolkit and ground truth data publicly available at: https://openlex3d.github.io.

II. The OpenLex3D Benchmark

A. Benchmark Design

Our benchmark aims to provide a comprehensive evaluation of 3D scene representations. To achieve this, we propose to change the goal of scene segmentation from determining a single correct label per object, to *determining a descriptive category* instead. The four categories we consider, in decreasing order of description accuracy, are:

Synonyms includes the primary labels for the target object as well as any other equally valid label. For instance, "glasses" and "spectacles".

Depictions describes any images or patterns depicted on the target object. For example, if a pillow features an image of a tree, the label "tree" would fall under this category.

Visually Similar includes objects that appear to be visually similar to the target objects and are likely to be confused for it. For example, visually similar terms for "glasses" could include "sunglasses" or "goggles".

Clutter covers any nearby or surrounding objects. Surrounding object features may "leak" into the features of interest due to 1) co-visibility in the same RGB frames and/or 2) incorrect merging in object-centric representations. This is the only category not defined by word labels but object IDs pointing to the surrounding objects.

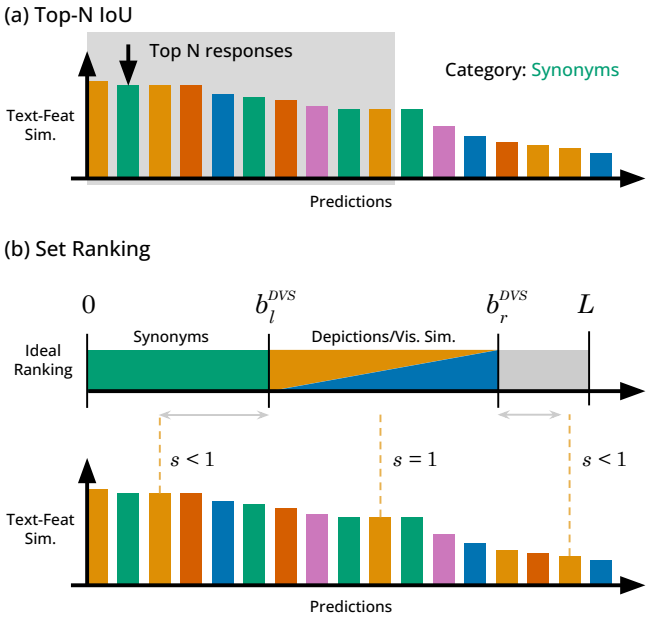

Fig. 3: **Top-N IoU and Set Ranking metrics illustration**. (a) Top-N IoU measures whether any of the top-N responses contain a label from category C. (b) Set Ranking evaluates the ranking of responses, assessing how closely the predicted rankings align with ideal rankings of categories.

An example of these categories is shown in Fig. 2. An ideal 3D scene representation would generate predictions that fall exclusively under the *synonym* category.

We use nouns (including multiple word labels such as "sofa cushion") for our ground truth categories to reduce the ambiguities of sentences and captions. For instance, sentence embedding models [10], are sensitive to variations in word order and struggle to distinguish between sentences with similar structures but different meanings.

B. Evaluation Methodology

We propose two tasks: a semantic segmentation task and an object retrieval task. The first task tests the accuracy of the method in describing different objects in the scene, the latter assesses whether it can identify and segment all instances that best match a given query.

1) Task 1: Open-Set Semantic Segmentation: In the first task, we first compute the cosine similarity of the features against the prompt list embeddings. The prompt lists we introduce in OpenLex3D are built from unique labels across all categories and scenes in each dataset, containing between 1,000 and 3,000 unique words. To evaluate the performance of the 3D representation, we introduce two metrics:

a) Top-N IoU at Category: This metric characterizes the proportion of objects $o \in \mathcal{O}$ in a scene that are correctly classified into a particular category C. We define it as:

$$\text{IoU}^C = \frac{\sum_{o \in \mathcal{O}} \frac{TP_o^C}{n_o}}{\sum_{o \in \mathcal{O}} \left(\frac{TP_o}{n_o} + \frac{FP_o}{n_o} + \frac{FN_o}{n_o} \right)}, \tag{1}$$

TP_o^C is the number of true positive points in category

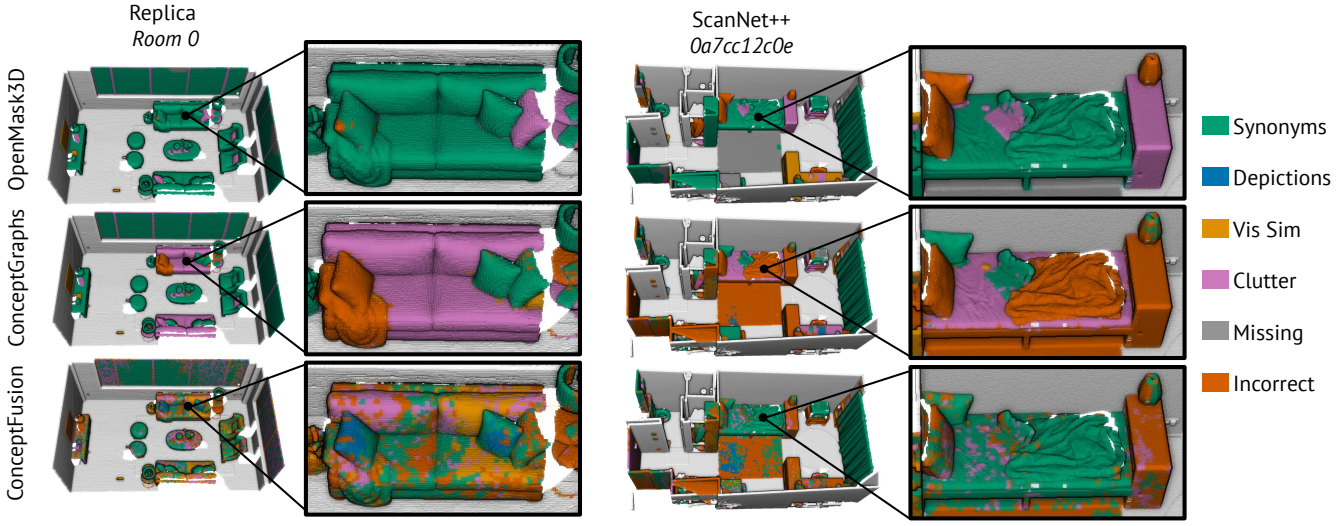

Fig. 4: **Top-5 IoU results for category classification for OpenMask3D [12], ConceptGraphs [3] and ConceptFusion [5] colored by category class.** Object-centric methods that segment in 3D, like OpenMask3D (top), often miss points due to generalization or depth quality issues. Those merging 2D segments tend to merge smaller ones, leading to misclassifications (middle). Dense representations, such as ConceptFusion, produce noisier predictions due to point-level features aggregating information from various context scales. In the highly cluttered environments of ScanNet++ [15], all evaluated methods show reduced performance.

C; TP_o, FP_o, and FN_i are the number of true positive, false positive and false negative points for all categories, normalized by the number of points of the object n_o.

To compute these quantities, we consider a predicted point label as a match for category C if any of the *top-N* responses of the text-feature similarity feature a label in category C (see Fig. 3a). A point is classified as *clutter* if it does not fit into any of the previous categories but shares a label with any category of a neighboring object. If no match is found the point is labeled as *incorrect*. Any points in the ground truth point cloud that have no corresponding points in the predicted cloud are labeled as *missing*.

b) Set Ranking: Our second metric assesses the distribution of the text-feature similarity of each point in the scene representation. For this, we quantify the mismatch of the responses when compared against an *ideal ranking* of category sets. We establish *synonyms* (S) as the first-rank set, while *depictions* and *visually similar* are considered as a joint second-rank set (DVS). The size of the sets is determined by the number of corresponding labels in the ground truth categories for each point. An example is shown in Fig. 3(b).

We first sort the predictions according to the ideal set ranking. We obtain left and right ranking bounds for each set, b_l^C and b_r^C, where C denotes the category set (S or DVS). We then compute a *rank score* s_i for each prediction i, as a function of its rank r_i:

$$s(r_i) = \min\left(1 + \min\left(0, \frac{r_i - b_l^C}{b_l^C}\right), 1 - \max\left(0, \frac{r_i - b_r^C}{L - b_r^C}\right)\right), \tag{2}$$

where L denotes the total number of predictions in the text-feature similarity (given by the evaluation prompt list). If the prediction falls in the right category set, we define $s(r_i) = 1.0$, and $s(r_i) < 1.0$ otherwise. The rank scores are then used to determine set inlier rates R_S and R_{DVS} as:

$$R_C = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \left(\frac{1}{L_C} \sum_{i}^{L_C} \mathbf{1}\left(s(r_i) = 1\right) \right), \tag{3}$$

where L_C denotes the number of labels per point $p \in \mathcal{P}$ for the set C (S or DVS), and $\mathbf{1}(\cdot)$ is an indicator function. In addition, we compute penalty scores that constitute inverse set ranking scores quantifying underscoring of *synonyms* and both under- and overscoring of the labels within the DVS category set, defined as P_S, P_{DVS}, and P_{DVS}. We also report a mean ranking score mR.

2) Task 2: Open-Set Object Retrieval: This task involves segmenting object instances that correspond to a given text-based query in a similar manner to [2]. We generate queries using the synonyms in the OpenLex3D label set, along with combinations of synonyms and their associated depictions. The resulting queries include references to motifs ("polka dots duvet cover"), specific characters ("ironman portrait") and brands ("nike athletic sneaker"). The number of queries ranges from 200 to 1,500 per scene. For evaluation, we use the Average Precision (AP). We report AP_{50} (IoU of 50%), AP_{25} (IoU of 25%), and mAP scores averaged over the IoU range of [0.5 : 0.95 : 0.05].

III. Experiments

A. Implementation

We evaluate four object-centric representations–ConceptGraphs [3], HOV-SG [13], OpenMask3D [12], and a minimal pipeline based on the findings in Kassab2024 [6]. For ConceptGraphs, we benchmark both a CLIP-based (ConceptGraphs) and a GPT-based pipeline (ConceptGraphs (GPT)). We also evaluate two dense representations: OpenScene [8] and ConceptFusion [5]. We exclude floors,

Dataset	Method	$S \uparrow$	$D \downarrow$	$VS \downarrow$	$C \downarrow$	$M \downarrow$	$I \downarrow$
Replica	ConceptGraphs [3]	0.41	0.01	0.11	0.24	0.02	0.22
	ConceptGraphs (GPT) [3]	**0.47**	0.03	**0.05**	0.21	0.02	0.23
	HOV-SG [13]	0.45	**0.00**	**0.05**	0.27	0.07	0.16
	Kassab2024 [6]	0.26	**0.00**	0.06	0.26	0.12	0.30
	OpenMask3D [12]	0.43	0.01	0.07	0.29	0.10	**0.10**
	ConceptFusion [5]	0.32	0.01	0.09	**0.16**	**0.00**	0.41
	OpenScene [8]	0.44	**0.00**	0.06	0.30	0.07	0.13
ScanNet++	ConceptGraphs [3]	0.26	0.02	0.05	0.10	0.13	0.44
	ConceptGraphs (GPT) [3]	**0.43**	0.01	0.04	0.11	0.13	0.28
	HOV-SG [13]	0.40	0.02	0.04	0.16	0.08	0.30
	Kassab2024 [6]	0.11	**0.00**	0.03	0.17	0.38	0.31
	OpenMask3D [12]	0.27	0.01	0.03	0.29	0.13	**0.27**
	ConceptFusion [5]	0.29	0.01	0.03	**0.08**	**0.04**	0.54
	OpenScene [8]	0.16	**0.00**	**0.02**	0.23	0.22	0.36
HM3D	ConceptGraphs [3]	0.27	0.02	0.03	0.12	0.08	0.47
	ConceptGraphs (GPT) [3]	**0.45**	0.01	0.04	0.11	0.09	0.31
	HOV-SG [13]	0.33	0.02	0.04	0.18	0.08	0.36
	Kassab2024 [6]	0.19	0.01	**0.01**	0.15	0.23	0.41
	OpenMask3D [12]	0.31	0.01	0.03	0.13	0.26	**0.26**
	ConceptFusion [5]	0.23	0.01	0.03	**0.09**	0.08	0.57
	OpenScene [8]	0.18	**0.00**	**0.02**	0.16	**0.06**	0.59

TABLE I: **IoU Top 5 Results for Object-Centric and Dense Representations.** Where *S* is the IoU at synonyms, *D* is depictions, *VS* is visually similar, *C* is clutter, *M* is missing and *I* is incorrect. A perfectly performing method would achieve an IoU^S score approaching 1 and an IoU score of 0 for all other categories.

ceilings, and walls from our evaluation. We use the ViT-H-14 CLIP backbone for all methods except OpenScene, which uses the ViT-L OpenSeg backbone.

B. Open-Set Semantic Segmentation

1) Top-N IoU: We report the Top-5 IoU in Tab. I. Top down views of a selection of the output point clouds colored by category are presented in Fig. 4. ConceptGraphs (GPT) is the top performing method in the *synonyms* category. The GPT prompt generates precise descriptions of the target object, which is then encoded into a highly specific text embedding. In contrast, CLIP, used by other methods, encodes both object-related and broader contextual information from the image, making it more prone to confusion. In general, dense methods produce noisier predictions as they use per-pixel features (Fig. 4). Regarding *depictions* and *visually similar* categories, OpenScene and Kassab2024 consistently yield the best results. This may stem from their distinct feature association strategies. For example, Kassab2024 selects a distinctive feature for each object using Shannon entropy instead of feature merging. This may preserve feature granularity and reduce classification confusion. The *clutter* category has worse IoU results across all methods, suggesting that crop scaling and/or segmentation is critical in improving overall classification performance. This is also apparent in the *missing* category. In general, most methods struggle with ScanNet++ and HM3D, indicating that cluttered, real-world environments still pose challenges for all approaches.

2) Set Ranking Evaluation: In Tab. II, we report set ranking results. The mean results are high, suggesting that *synonyms* tend to score higher in the predicted ranks, and that *depictions* and *visually similar* labels generally score below *synonyms*, described as the ideal ranking in Sec. II-B.1. We observe high R_S scores for ConceptGraphs (GPT), similar to the Top-5 IoU^S, again suggesting that the text embeddings produced from GPT captions are highly specific compared to

Dataset	Method	$mR \uparrow$	$R_S \uparrow$	$P_S \downarrow$	$R_{DVS} \uparrow$	$P_{DVS} \downarrow$	$P_{DVS} \downarrow$
Replica	ConceptGraphs [3]	0.82	0.13	0.14	0.06	0.63	0.23
	ConceptGraphs (GPT) [3]	0.63	**0.21**	0.33	**0.07**	0.52	0.43
	HOV-SG [13]	0.82	0.17	0.14	**0.07**	**0.50**	0.23
	Kassab2024 [6]	0.76	0.10	0.21	0.03	0.54	0.27
	OpenMask3D [12]	0.83	0.17	0.12	0.06	0.51	**0.21**
	ConceptFusion [5]	0.76	0.11	0.21	0.05	0.57	0.28
	OpenScene [8]	**0.85**	0.16	**0.10**	0.05	0.53	**0.21**
ScanNet++	ConceptGraphs [3]	0.80	0.09	0.19	0.03	0.59	0.24
	ConceptGraphs (GPT) [3]	0.66	**0.18**	0.31	0.03	0.60	0.40
	HOV-SG [13]	**0.84**	0.15	**0.14**	**0.04**	0.64	**0.19**
	Kassab2024 [6]	0.72	0.05	0.26	0.01	0.60	0.30
	OpenMask3D [12]	0.79	0.12	0.19	0.02	**0.57**	0.25
	ConceptFusion [5]	0.74	0.10	0.26	0.02	0.63	0.30
	OpenScene [8]	0.77	0.06	0.18	0.01	**0.57**	0.31
HM3D	ConceptGraphs [3]	0.86	0.08	0.13	0.02	0.59	0.20
	ConceptGraphs (GPT) [3]	0.68	**0.15**	0.32	**0.03**	0.59	0.36
	HOV-SG [13]	**0.88**	0.12	0.11	0.02	0.59	**0.19**
	Kassab2024 [6]	0.80	0.06	0.19	0.01	0.62	0.26
	OpenMask3D [12]	0.86	0.10	0.12	0.02	**0.56**	0.20
	ConceptFusion [5]	0.78	0.07	0.20	0.02	0.61	0.27
	OpenScene [8]	0.87	0.05	**0.10**	0.01	**0.56**	0.22

TABLE II: **Set Ranking Evaluation.** For mR, R_S, R_{DVS} being the mean score and the respective inlier rates, higher is better. In contrast, for the underscoring and overscoring penalties P_S, P_{DVS}, P_{DVS}, lower is better.

Dataset	Method	$mAP \uparrow$	$AP_{50} \uparrow$	$AP_{25} \uparrow$
Replica	ConceptGraphs [3]	5.86	11.32	22.39
	ConceptGraphs (GPT) [3]	5.13	10.77	18.19
	HOV-SG [13]	5.76	11.67	**25.30**
	Kassab2024 [6]	1.38	2.87	7.54
	OpenMask3D + NMS [12]	**11.47**	**17.01**	24.02
ScanNet++	ConceptGraphs [3]	1.45	4.36	15.27
	ConceptGraphs (GPT) [3]	1.97	5.54	13.39
	HOV-SG [13]	1.79	4.95	**18.75**
	Kassab2024 [6]	0.40	1.19	3.39
	OpenMask3D + NMS [12]	**4.00**	**6.90**	10.34
HM3D	ConceptGraphs [3]	**5.09**	**8.05**	**11.18**
	ConceptGraphs (GPT) [3]	4.80	7.75	10.76
	HOV-SG [13]	3.44	5.39	7.42
	Kassab2024 [6]	1.03	1.87	3.97
	OpenMask3D + NMS [12]	4.03	5.56	8.35

TABLE III: **Object Retrieval Evaluation.** NMS stands for Non-maximum Suppression and is used to select object masks in the OpenMask3D [12] pipeline.

CLIP image encodings. However, ConceptGraphs (GPT) also consistently overscores the *depictions* and *visually-similar* categories (high P_{DVS}) while underscoring *synonyms* compared to the remaining methods (high P_S).

C. Open-Set Object Retrieval

We report AP results in Tab. III. Overall AP is low and in line with comparable benchmarks [2], highlighting the challenges of this task and the potential for improvement. OpenMask3D leads the Replica and ScanNet++ metrics but fails to generalize to the larger HM3D scenes where ConceptGraphs is the top performing method. Overall, the relative performance of the methods is largely consistent across all three metrics.

IV. Conclusion

We introduced OpenLex3D, a new benchmark for open-vocabulary evaluation that captures real-world language variability across multiple levels of specificity. Our benchmark includes human-annotated labels for three RGB-D datasets—ScanNet++, Replica, and HM3D—enabling comprehensive evaluation across diverse environments.

REFERENCES

[1] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2024.

[2] Francis Engelmann, Ayca Takmaz, Jonas Schult, Elisabetta Fedele, Johanna Wald, Songyou Peng, Xi Wang, Or Litany, Siyu Tang, Federico Tombari, Marc Pollefeys, Leonidas Guibas, Hongbo Tian, Chunjie Wang, Xiaosheng Yan, Bingwen Wang, Xuanyang Zhang, Xiao Liu, Phuc Nguyen, Khoi Nguyen, Anh Tran, Cuong Pham, Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. OpenSUN3D: 1st Workshop Challenge on Open-Vocabulary 3D Scene Understanding. *arXiv preprint arXiv:2402.15321*, 2024.

[3] Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-Vocabulary 3D Scene Graphs for Perception and Planning. In *IEEE Int. Conf. Robot. Autom.*, 2024.

[4] Felix Hill, Roi Reichart, and Anna Korhonen. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[5] 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. *Robot.: Sci. Syst.*, 2023.

[6] Christina Kassab, Matías Mattamala, Sacha Morin, Martin Büchner, Abhinav Valada, Liam Paull, and Maurice Fallon. The Bare Necessities: Designing Simple, Effective Open-Vocabulary Scene Graphs. *arXiv preprint arXiv:2412.01539*, 2024.

[7] George A. Miller. *WordNet: An electronic lexical database*. MIT Press, 1995.

[8] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D Scene Understanding with Open Vocabularies. In *IEEE Int. Conf. Computer Vision and Pattern Recognition*, 2023.

[9] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI. In *Intl. Conf. on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[10] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

[11] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[12] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Intl. Conf. on Neural Information Processing Systems (NeurIPS)*, 2023.

[13] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. *Robot.: Sci. Syst.*, 2024.

[14] Sriram Yenamandra, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton. HomeRobot: Open Vocabulary Mobile Manipulation. *arXiv preprint arXiv:2306.11565*, 2023.

[15] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. In *Intl. Conf. on Computer Vision*, 2023.