OpenSU3D: Open World 3D Scene Understanding using Foundation Models

Rafay Mohiuddin[1*], Sai Manoj Prakhya[2], Fiona Collins[1], Ziyuan Liu[2] and André Borrmann[1]

***Abstract*— In this paper, we present a novel, scalable approach for constructing open set, instance-level 3D scene representations, advancing open world understanding of 3D environments. Existing methods require pre-constructed 3D scenes and face scalability issues due to per-point feature representation, additionally struggle with contextual queries. Our method overcomes these limitations by incrementally building instance-level 3D scene representations using 2D foundation models, and efficiently aggregating instance-level details such as masks, feature vectors, names, and captions. We introduce fusion schemes for feature vectors to enhance their contextual knowledge and performance on complex queries. Additionally, we explore large language models for robust automatic annotation and spatial reasoning tasks. We evaluate our proposed approach on multiple scenes from ScanNet [1] and Replica [2] datasets demonstrating zero-shot generalization capabilities, exceeding current state-of-the-art methods in open world 3D scene understanding. Project page: https://opensu3d.github.io/**

I. INTRODUCTION

Recent advancements in AI, particularly in open world understanding of 2D images, are largely attributed to pre-trained foundation models [3]–[5] and large vision language models [6], [7]. However, extending these breakthroughs to 3D environments remains a challenge. While innovative, current 3D understanding methods [8]–[12] have not yet achieved the performance levels seen in 2D. Addressing this gap is critical for robotics applications, as it could transform how robots perceive, interact, and reason within the three-dimensional world.

Our work builds upon advancement in several research areas. Foundation models [3], [5], [13] integrate visual and textual information into a unified representation, enabling multimodal understanding. [14], [15] and grounding approaches [16], [17], based on SAM [4] provide promptable, open-vocabulary segmentation capabilities. Large language models [18], [19] and large vision language models [7], [20] have significantly advanced natural language and multimodal understanding. Traditional 3D scene understanding approaches [21]–[24] created 3D metric semantic maps but were limited by closed-set paradigms. In open-vocabulary 3D scene understanding, early works [9], [11], [12], [25] faced computational and scalability challenges, while later methods [26], [27] adopted instance-centric approach but were limited by non-incremental processing. Global 3D spatial reasoning remains challenging, while conventional scene graph based approaches [23], [24] show limitations, early works [8], [10], [28] leveraging LLMs for 3D reasoning, shows promising results. Recent related works [27], [29]–[34] have also proposed method to tackle these challenges. Unlike [32], [33], we ensure semantic consistency through geometric overlap in 3D space. [27], [31] are non-incremental and [29], [30] struggle with contextual queries. In contrast to [32]–[34], we explore in-context learning leveraging large context length of LLM for complex spatial reasoning.

In this work, we present a novel approach for open-set 3D scene representations enabling instance recall, segmentation, annotation, and spatial reasoning. Our method leverages 2D foundation [3], [4], [16], [35] and large language models [7] to extract instance-level information from RGB images and efficiently associate to 3D space. Following are the key contributions:

1. An incremental and scalable approach for creating open-set 3D scene representation using 2D foundation models.
2. Improved feature fusion formulation, enabling instance identification through contextual queries.
3. Complex spatial reasoning using large language models in conjunction with our open-set 3D scene representation.

II. METHOD

Given RGB-D image sequence with poses, the steps are:

A. Per-Image Feature Extraction

From images $\mathcal{I}$ a subset $\mathcal{I}'$ is sampled with stride s to minimize computational redundancy. For each image, GroundedSAM [16] obtains 2D masks M, bounding boxes BB, and prediction scores S_{pred}. Crops of each instance based on $bb \in BB$ are passed to GPT-4V [7] for names N and captions C. Each instance is assigned a unique ID, modifying 2D masks with these IDs and adding a border px around each mask, updating $\mathcal{M}$ to $\mathcal{M}'$. Feature vectors are extracted using the CLIP [3] encoder in two stages: a global feature vector f_{G} from the entire image and instance-specific feature vectors $F = \{f_{\text{MS}}\}$ by fusing feature vectors from crops at multiple scales S_r. M', and instance-level metadata including IDs, names $n \in N$, captions $c \in C$, prediction scores $s_{\text{pred}} \in S_{\text{pred}}$, fused feature vectors $f_{\text{MS}} \in F$, and global feature vector f_{G}, are stored for each image in $\mathcal{I}'$.

B. 2D to 3D Fusion & Tracking

We initiate the fusion and tracking module by initializing an empty 3D point cloud for the complete 3D scene $\mathcal{P}_{\text{scene}} \in \mathbb{R}^{x,y,z,\text{ID}}$, and a global hash table $\mathcal{Q}$ for tracking the unique IDs, defined as: $\mathcal{Q} : \mathcal{Q} \mapsto \{\text{ID} \in \texttt{uniq}(\text{ID} \in \mathcal{P}_{\text{scene}}) : \{\text{ID} \in \{M'\}\}\}$.

* Corresponding author rafay.mohiuddin@tum.de
[1] Chair of Computational Modeling & Simulation, Technical University of Munich, 80333 Arcisstraße 21, Germany
[2] Intelligent Cloud Technologies Lab, Huawei Munich Research Center, 80992 Riesstraße 25, Germany

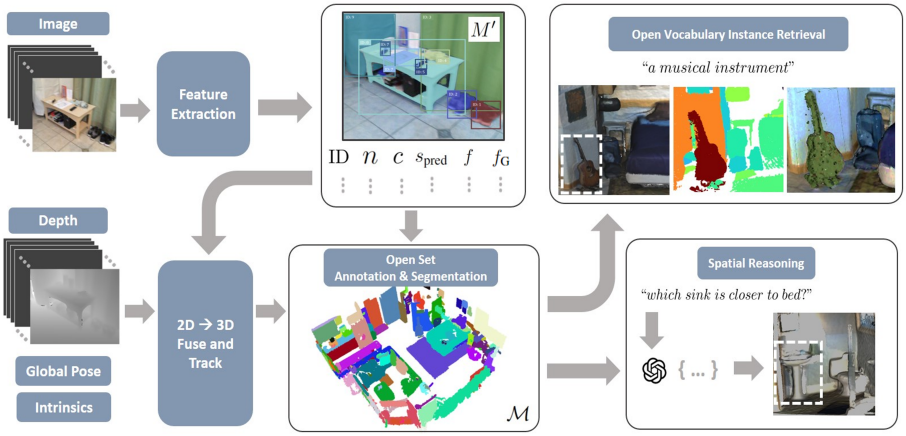

Fig. 1: **Open World 3D Scene Understanding Pipeline.** Our method takes a sequence of RGB-D images and constructs a 3D scene representation for open vocabulary instance retrieval, open set annotation, segmentation, and spatial reasoning.

For each image I', we retrieve the depth map D, global poses T, updated masks M', and camera intrinsic K. Each pixel $(u, v) \in I'$ is back-projected into 3D space, assigning a semantic label from M' and aggregating into $\mathcal{P}_{\text{frame}}$:

$$\mathcal{P}_{\text{frame}} = \left\{ T\left(D(u,v) \cdot K^{-1} \begin{pmatrix} u & v & 1 \end{pmatrix}^{\top} \right), M'(u,v) \right\} \tag{1}$$

Index pairs $\{(\mathbf{i}_{\text{frame}}, \mathbf{i}_{\text{scene}})\}$ are determined to identify corresponding points between $\mathcal{P}_{\text{frame}}$ and $\mathcal{P}_{\text{scene}}$. We sample $\mathcal{P}'_{\text{scene}}$ from $\mathcal{P}_{\text{scene}}$ using geometric bounds of $\mathcal{P}_{\text{frame}}$ to limit search space. A KDTree search using Euclidean distance function $d(\cdot,\cdot)$ to matches points $\mathbf{p} \in \mathcal{P}_{\text{frame}}$ with $\mathbf{q} \in \mathcal{P}'_{\text{scene}}$. If $d(\mathbf{p}, \mathbf{q}) < \epsilon$, we group indices to obtain index pairs $\{(\mathbf{i}_{\text{frame}}, \mathbf{i}_{\text{scene}})\}$, ensuring constant computation per update. To update and track IDs, similar to [36] approach, we obtain unique IDs $\{\text{ID}_f\}$ for each segment and the total point count $\{c_{\mathcal{P}_f}\}$ in $\mathcal{P}_{\text{frame}}$. Using index pairs, we get overlapping points from $\mathcal{P}_{\text{scene}}$, deriving segment IDs $\{\text{ID}_s\}$ and their point counts $\{c_{\mathcal{P}_s}\}$. The overlap ratio is evaluated as:

$$\text{OverlapRatio} = \frac{\max(\{c_{\mathcal{P}_s}\})}{\min(c_{\mathcal{P}_f}, \max(\{c_{\mathcal{P}_s}\}))} \tag{2}$$

If the overlap ratio $\geq \rho$, we replace and update IDs, creating $\mathcal{P}'_{\text{frame}}$ which is concatenated to $\mathcal{P}_{\text{scene}}$. To retain constant sparsity, points from $\mathcal{P}'_{\text{scene}}$ may be deleted. The updated IDs are appended to $\mathcal{Q}$. If the overlap ratio does not meet the threshold, a new entry is added.

1) Post Processing: The point cloud $\mathcal{P}_{\text{scene}}$ with updated IDs, corresponding tracked overlapping IDs $\mathcal{Q}$, along with per-image metadata, is processed into instance-centric map $\mathcal{M} = \{(\mathcal{P}, n, c, f_{\text{MV}}, bb_{3D}, (x_c, y_c))_i | i \in \texttt{uniq}(\text{ID} \in \mathcal{P}_{\text{scene}})\}$.

For each 3D object $\mathcal{P}_i$, DBSCAN clustering reduces noise and generates fine-grained 3D masks. We compute the 3D bounding box $bb_{3D,i}$ and centroid $(x_c, y_c)_i$. For each multiview image of $\mathcal{P}_i$, names N', captions C', prediction scores S'_{pred}, and feature vectors F' are retrieved via $\mathcal{Q}$ for aggregation. The label $n_i \in N'$ and caption $c_i \in C'$ with the highest S'_{pred} are assigned to $\mathcal{P}_i$. Alternatively, the top m names from N' are refined via LLM [7] using the prompt: *"assign a single name to the object based on a given list of names"*, yielding label n'_i. Multiview feature vector f_{MV_i} is obtained through fusion (Sec. II-B.2) of top m feature vectors based on S'_{pred}.

2) Feature Fusion: Given feature vectors $\{f\}_k$ from multi-scale crops of an image instance and $\{f_{MS}\}_m$ from multiview images of a 3D instance, a simple fusion scheme aggregates these vectors as follows:

$$f_{\text{MS}} = \frac{1}{k} \sum_{i=1}^{k} f_i \tag{3}$$

$$f_{\text{MV}} = \frac{1}{m} \sum_{i=1}^{m} f_{\text{MS}_i} \tag{4}$$

The fusion schemes in Eq. 3 and Eq. 4 lack contextual information, leading to suboptimal performance for contextual queries (Table IV). [26] and our ablations (Sec. III-B.1) show that while multiscale crops improve accuracy, larger crops reduce object recall. To address this, we propose a modified multiscale fusion scheme (Eq. 5), using weighted aggregation where f_1 is feature vector from best-fit crop and $\varepsilon \approx 1\text{e}-8$.

$$f_{\text{MS}} = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{f_1 \cdot f_i}{\max(\|f_1\|_2 \cdot \|f_i\|_2, \varepsilon)} \right) \cdot f_i \tag{5}$$

For multiview feature fusion, inspired by per-pixel representation in [12], we incorporate global feature vector f_{G} from the entire image into our per-instance representation, defined as:

$$f_{\text{MV}} = \frac{1}{m} \sum_{i=1}^{m} f_{\text{MS}_i} + \left(\frac{f_{\text{MS}_i} \cdot f_{\text{G}_i}}{\max(\|f_{\text{MS}_i}\|_2 \cdot \|f_{\text{G}_i}\|_2, \varepsilon)} \right) \cdot f_{\text{G}_i} \tag{6}$$

C. Instance Retrieval & Segmentation

Query $\mathcal{K}$ is processed using the CLIP [3] text encoder to obtain the feature vector $f_{\mathcal{K}}$, cosine similarity scores $\{S_{\text{score}}\}$ are computed with all $f_{\text{MV}} \in \mathcal{M}$. The segmentation mask of the 3D instance with the highest similarity score is retrieved as the most likely response to $\mathcal{K}$.

D. Spatial Reasoning

For complex spatial reasoning, our approach involves in-context learning, leveraging large context window of LLM [7]. Adopting stratergy of [37] for 3D, $\mathcal{M}' = \mathcal{M} \setminus \{\mathcal{P}, f_{\text{MV}}\}$ is provided to LLM along with a query and system prompt crafted using the following prompting strategy:

1. Use 'Name' & 'Description' to understand object.
2. Use 'ID' to refer object.
3. Use 'Cartesian Coordinates'.
4. Get 'Centroid' & 'Bounding Box' information.
5. Compute 'Euclidean Distance' if necessary.
6. Assume 'Tolerance' if necessary.

III. Experiments

A. Implementation Details

1) Datasets: Scenes from Replica [2] (*room0*, *room1*, *room2*, *office0-office4*) and ScanNet [1] (*scene0000_00*, *scene0034_00*, *scene0164_03*, *scene0525_01*, *scene0549_00*) were used for evaluation. Similar to [12], [33] limited scenes were selected due to extensive manual evaluations.

2) Hyperparameter Settings: We use top $m = 5$ images, $k = 3$ crop levels with scaling ratio $S_r = [0.8, 1, 1.2]$, and stride $s = 40$. GroundedSAM [16] thresholds: IoU 0.4, bounding box 0.25, text 0.25. Mask border padding: $px = 20$ pixels. Overlap ratio: voxel size $\epsilon = 0.02$, threshold $\rho = 0.3$. DBSCAN: epsilon 0.1, minimum cluster 20 points. GPT-4 [7] temperature: 0.

3) Filtering and Post-Processing: Large background objects (walls, ground, roof, ceiling) and those with bounding boxes $> 95\%$ of image area are excluded to prevent their feature vectors from exhibiting similarity to foreground objects. In DBSCAN, clusters with points $\geq 80\%$ of the largest cluster are considered separate instances with unique IDs and attributes. For undetectable objects by GPT-4 [7], instances are labeled using RAM++ [35] with captions: *"an {object} in a scene"*.

B. Results and Discussion

1) Ablation Studies: We conducted ablation studies on Crop Level k, Top Images m, and Crop Ratios S_r using [26]'s setup. Top Images m influence multiview feature fusion (Eq. 6), while Crop Ratio S_r (ratio for scaling sides of crop) and Crop Level k affect multiscale fusion (Eq. 3), with higher k subsequently leading to larger crops. Extreme values of these hyperparameters can degrade performance. Low m, S_r and k reduces redundancy, while high m may include poor images, whereas larger S_r provide more context but may saturate similarity scores as leading to performance deterioration, shown in Table I.

Parameter	Value	Replica [2]				
		mAcc	F-mIoU	AP	AP50	AP25
Top Images (m)	1.0	39.6	43.4	8.7	19.3	27.2
	5.0	**40.8**	**44.7**	**8.9**	**19.6**	**27.7**
	10.0	39.3	44.3	8.7	19.1	27.5
Crop Levels (k)	1.0	35.9	43.6	**9.1**	**19.6**	**27.7**
	3.0	**40.8**	**44.7**	8.9	**19.6**	**27.7**
	5.0	39.4	44.3	8.8	19.4	26.9
Crop Ratio (S_r)	[0.1,1,1.1]	39.9	44.4	**8.9**	19.4	**28.1**
	[0.8,1,1.2]	**40.8**	**44.7**	**8.9**	**19.6**	27.7
	[0.7,1,1.3]	39.9	44.8	**8.9**	19.4	27.3

TABLE I: **Ablation Study of Hyperparameters.** Total images m w.r.t best prediction scores (s_{pred}) for multiview fusion; effect of crop ratio S_r and number of crops k on multiscale fusion.

2) Quantitative Comparison with Baseline Methods: For quantitative evaluation, as in [26], [33], 3D masks were retrieved with ground truth labels and the prompt: *"an {object} in a scene"*. Masks were downsampled to 0.25cm voxel size, followed by nearest neighbor search for intersecting points. Results on the Replica [2] dataset were compared against state-of-the-art models [12], [26], [31], [33], using identical prompts and foundation models with direct feature fusion formulation (Eq. 4, 3) for fair comparision. Our method demonstrates better performance than baselines on quantitative metrics, as shown in Tables II and III.

Method	Replica [2]	
	mAcc	F-mIoU
ConceptFusion [12]	24.2	31.3
ConceptFsuion+SAM [12]	31.5	38.7
ConceptGraph [33]	40.6	36.0
ConceptGraph-Detector [33]	38.7	35.4
OpenSU3D (Ours)	**42.6**	**40.9**

TABLE II: **Comparison of open-vocabulary segmentation results** with ConceptGraph [33] setup.

Method	Replica [2]		
	AP	AP50	AP25
OpenMask3D [26]	**13.0**	18.4	24.2
OpenMask3D+Segment3D [31]	-	18.7	-
OpenSU3D (Ours)	8.9	**19.6**	**27.7**

TABLE III: **Comparison of open-vocabulary segmentation results** with OpenMask3D [26] setup.

3) Qualitative Comparison with Baseline Methods: The quantitative evaluation focused on closed vocabulary assessments and recall accuracy this does not reflect real-world open vocabulary needs. Additionally, mask proposal-based evaluations [31] may not fully capture true performance against closed-set ground truth. To address this, we provide qualitative comparisons with baseline works (as shown in Fig. 2), assessing segmentation mask recall for given open vocabulary queries. Our approach resulted improves 2D-to-3D associations and similarity score distribution with proposed multiscale and multiview feature fusion (Eq. 5, 6).

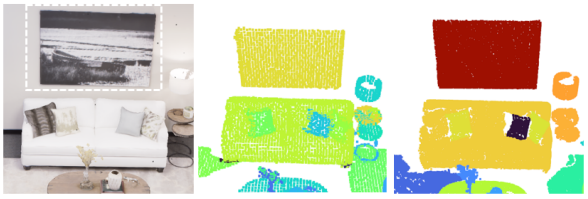

Fig. 2: **Text-Instance Similarity Heatmaps.** Cosine similarity for text queries using ConceptGraph [33], OpenMask3D [27] and our method. ■: max, ■: min similarity.

4) Assessment of Feature Fusion Schemes: s A comprehensive qualitative assessment with over 1,000 queries containing instances, affordances, properties, and relative queries was made. Performance was evaluated based on CLIP [3] instance retrieval (Sec. II-C) for four fusion schemes: *Scheme 1* represents direct aggregation of multiscale (Eq. 3) and multiview features (Eq. 4), *Scheme 2* utilizes (Eq. 3) with updated multiview features (Eq. 6), *Scheme 3* utilizes (Eq. 4) with updated multiscale features (Eq. 5) with increased crop expansion ratios ($S_r = [1, 2, 4]$), and *Scheme 4* represents combination of both updated multiview and multiscale fusion formulations (Eq. 5, 6). As shown in Table IV and Fig. 3 for instance, property, and affordance queries, performance the updated fusion formulations in *Schemes 2, 3,* and *4* improved the recall of instance masks and similarity score distribution, with *Scheme 4* performing the best overall.

Feature Fusion	Replica [2]				ScanNet [1]			
	Inst.	Aff.	Prop.	Rel.	Inst.	Aff.	Prop.	Rel.
Scheme 1	0.8	0.7	0.7	0.3	0.8	**0.8**	0.7	0.4
Scheme 2	0.8	0.7	**0.9**	0.5	**0.9**	0.7	**0.8**	0.6
Scheme 3	**0.9**	**0.9**	**0.9**	**0.6**	**0.9**	**0.8**	0.7	0.6
Scheme 4	0.8	**0.9**	**0.9**	**0.6**	**0.9**	0.7	0.7	**0.7**

TABLE IV: **Evaluation of feature fusion schemes**. Accuracy of fusion schemes for retrieval with "Inst." (instance), "Aff." (affordance), "Prop." (property), and "Rel." (relative) text queries, as assessed by a human evaluator.

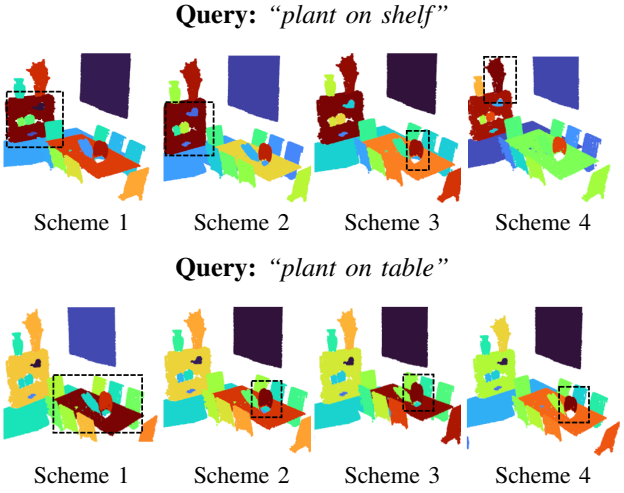

Fig. 3: **Relative Query Similarity Heatmaps.** For a given text query, per-instance cosine similarity heatmaps across feature fusion schemes. ■: max, ■: min similarity.

5) Open Set Annotation and Segmentation: Annotation accuracies for directly assigned labels n using the maximum prediction score S'_{pred} and labels n', where top m labels based on S'_{pred} are refined by LLM (see Sec. II-B.1), were manually verified across Replica [2] and ScanNet [1] scenes. To evaluate open set segmentation, mask merging accuracy is determined by counting and classifying under-merges and over-merges as faulty. Evaluation summarized in Table V shows that LLM labels n' are more accurate and concise than direct labels n, also helping to filter out large instances (Sec. III-A.3) and slightly improving mask merging accuracy.

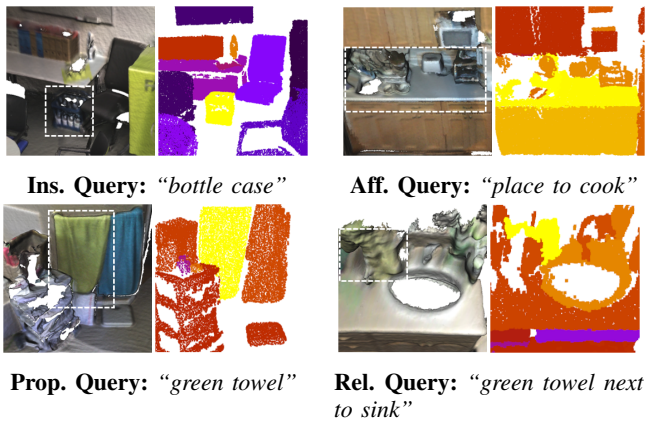

Fig. 4: **Text-Instance Similarity per Query Type.** For a given text query, per-instance cosine similarity heatmaps for different query types. ■: max, ■: min similarity.

Labels	Replica [2]		ScanNet [1]	
	Label Acc.	Merge Acc.	Label Acc.	Merge Acc.
Direct Label (n)	0.83	0.87	0.75	0.85
LLM Label (n')	**0.87**	**0.88**	**0.84**	**0.87**

TABLE V: **Qualitative evaluation of segmentation and annotation accuracy**. For Direct Label (n) and LLM Label (n'), the annotation and merge accuracy of segmentation masks, as assessed by a human evaluator.

6) Complex Spatial Reasoning: To assess spatial reasoning, we posed 70 complex spatial reasoning questions (Example, Fig. 1 *"Which sink is closer to bed"*) across all scenes. The manual assessment showed, with our approach (see II-D) LLM [7] demonstrated effective reasoning ability in 3D space over constructed representation. LLM exhibited higher accuracy in scenes from Replica [2] (0.83), compared to ScanNet [1] (0.68). This decline in performance can be attributed to a comparatively higher incidence of flaws in merging and label assignment in larger ScanNet [1] scenes.

IV. CONCLUSION

In conclusion, this study presents a scalable and incremental framework for constructing open set 3D scene representation for open world 3D scene understanding tasks, addressing the limitations of current methods. By leveraging 2D foundation models, our approach constructs detailed instance level 3D scene representations, efficiently tracking and associating instance-specific information such as feature vectors, names, and captions. The proposed feature fusion schemes encompass contextual information, enhancing performance on relative queries, while large language models improve annotation and enables complex spatial reasoning. The effectiveness of this approach is limited by the capabilities of its underlying foundation models and occasional merging flaws. Comprehensive evaluations show that our method achieves superior zero-shot generalization compared to state-of-the-art solutions. In the future, we plan to explore post-processing methods to refine mask merging, spatio-temporal reasoning in 3D dynamic scenes and extend our approach from indoor to large-scale outdoor environments.

REFERENCES

[1] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Niessner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, *et al.*, "The replica dataset: A digital replica of indoor spaces." https://arxiv.org/abs/1906.05797, 2019.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8748–8763, PMLR, 2021.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, Gustafson, *et al.*, "Segment anything," in *Proceedings of International Conference on Computer Vision*, 2023.

[5] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, PMLR, 2022.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[7] OpenAI, "Gpt-4 technical report." arXiv preprint arXiv:2303.08774, 2023.

[8] Y. Hong, C. Lin, Y. Du, Z. Chen, J. B. Tenenbaum, and C. Gan, "3d concept learning and reasoning from multi-view images," in *Proceedings of Computer Vision and Pattern Recognition*, 2023.

[9] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. e. a. Funkhouser, "Openscene: 3d scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 815–824, 2023.

[10] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," in *Neural Information Processing Systems*, 2023.

[11] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open vocabulary 3d scene understanding," in *Proceedings of Computer Vision and Pattern Recognition*, 2023.

[12] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, G. Iyer, S. Saryazdi, *et al.*, "Conceptfusion: Open-set multimodal 3d mapping," in *Robotics: Science and Systems*, 2023.

[13] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, PMLR, 2023.

[14] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *International Conference on Learning Representations (ICLR)*, 2022.

[15] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, *et al.*, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, *et al.*, "Grounded sam: Assembling open-world models for diverse visual tasks." http://arxiv.org/abs/2401.14159, Jan. 2024.

[17] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 19769–19782, Curran Associates, Inc., 2023.

[18] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, *et al.*, "Llama: Open and efficient foundation language models," *ArXiv*, vol. abs/2302.13971, 2023.

[20] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

[21] M. Grinvald, F. Furrer, T. Novkovic, J. J. Chung, C. Cadena, R. Siegwart, and J. Nieto, "Volumetric instance-aware semantic mapping and 3d object discovery," *IEEE Robotics and Automation Letters*, 2019.

[22] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: An open-source library for real-time metric-semantic localization and mapping," in *ICRA*, 2020.

[23] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3D scene graph construction and optimization," in *Robotics: Science and Systems (RSS)*, 2022.

[24] S.-C. Wu, J. Wald, K. Tateno, N. Navab, and F. Tombari, "Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7515–7525, 2021.

[25] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023.

[26] A. Takmaz, E. Fedele, R. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," in *Advances in Neural Information Processing Systems* (A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 68367–68390, Curran Associates, Inc., 2023.

[27] Z. Huang, X. Wu, X. Chen, H. Zhao, L. Zhu, and J. Lasenby, "Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation," *European Conference on Computer Vision*, 2024.

[28] J. Yang, X. Chen, S. Qian, N. Madaan, M. Iyengar, D. F. Fouhey, and J. Chai, "Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7694–7701, 2024.

[29] Y. Bhalgat, I. Laina, J. F. Henriques, A. Zisserman, and A. Vedaldi, "N2f2: Hierarchical scene understanding with nested neural feature fields," *arXiv preprint arXiv:2403.10997*, 2024.

[30] C. M. Kim, M. Wu, J. Kerr, K. Goldberg, M. Tancik, and A. Kanazawa, "Garfield: Group anything with radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21530–21539, June 2024.

[31] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, "Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels," 2023.

[32] H. Chang, K. Boyalakuntla, S. Lu, S. Cai, E. P. Jing, S. Keskar, S. Geng, A. Abbas, L. Zhou, K. Bekris, *et al.*, "Context-aware entity grounding with open-vocabulary 3d scene graphs," in *7th Annual Conference on Robot Learning*, 2023.

[33] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024.

[34] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone, "Clio: Real-time task-driven open-set 3d scene graphs," *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8921–8928, 2024.

[35] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, Y. Guo, and L. Zhang, "Recognize anything: A strong image tagging model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1724–1732, June 2024.

[36] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "Sam3d: Segment anything in 3d scenes." https://arxiv.org/abs/2306.03908v1, 2023.

[37] J. Yang, H. Zhang, F. Li, X. Zou, C. Li, and J. Gao, "Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v," 2023.