3D Scene Understanding on a City-Scale Level (3DV Project Proposal)

Nicolas S. Blumer
ETHZ / University of Zurich
nblume@student.ethz.ch

Qingxuan Chen
ETHZ / University of Zurich
qingxuan.chen@uzh.ch

Marco Zamboni
ETHZ
mzamboni@student.ethz.ch

Valentin Bieri
ETHZ
bieriv@student.ethz.ch

Abstract

Recent studies sucessfully lifted embeddings generated by 2D vision-language foundation models to 3D space, enabling direct language-based interaction with 3D scenes. Whereas most studies focus on understanding controlled, small environments, we aim to adapt these methods to city-scale scenes. Taking aerial imagery as a starting point, we hope to produce a (partial) representation of a city that enables language-based analysis of urban infrastructure and inventory as well as socio-economical dynamics.
We can validate extracted information against the publicly available city registries and public benchmarks, hoping to demonstrate high usability across a wide range of applications.

1. Related Work

Recent progress in large-scale pre-training of vision-language models (VLMs) allows a new level of image understanding, boosting the state of the art in a wide range of tasks such as image classification [5, 14], object detection [1], and open or closed set segmentation [2, 12].
Multiple studies have carried these successes to the field of 3D vision, enabling a new generation of open-vocabulary scene understanding applications. Most remarkably, VLM features allow open vocabulary queries on 3D scenes by object or material. This is typically achieved by reconstructing a 3D scene from a set 2D images, then analyzing and embedding the 2D images using a VLM, and finally fusing the extracted 2D information into the 3D representations [3, 8, 10, 11]. If desired, segmentation masks can be computed during this pipeline [12] or at query time [10].
Importantly, existing work limits the scope of the represented environment to scenes at the scale of rooms, or single landmarks [3, 10, 12]. We would be - to the best of our knowledge - the first to extract meaningful information from semantically enriched embeddings at the scale of a city by text guidance.

2. Methodology

We identify the following main building blocks of our pipeline:
1. 3D scene construction
2. 2D image embedding
3. Fusion of 2D embeddings into 3D scene
4. Evaluation

In the following, we briefly discuss the anticipated implementation of each building block, its necessary assumptions, and potential challenges.

2.1. 3D scene construction

Hopefully we will be able to use already available 3d scenes such as STPLS3D[2] or Google Maps data. A key challenge is that for these datasets the relevant 2D images might be unavailable or hard to register; for this reason, we may have to build our own 3D scenes from 2D datasets using aerial imagery as a starting point. Regardless, it may be beneficial to add imagery from multiple perspectives (such as street-level views) even if we can use a large point collection. Our embeddings may benefit from the additional information. Examples of large street-level data collections would be Mapillary or KataView (by OpenStreetMap). Importantly, note that these images do not require ground truth labels. Their purpose is uniquely to improve the scene representation and its semantics.

2.2. 2D image embedding

Ideally, we can reuse the OpenScene[10] or OpenMask[12] pipeline, including the used CLIP-based embeddings or SAM[7] segmentation masks.
However, CLIP[11] was trained on street-level images, not necessarily aerial images. Hence the features may not generalize. Depending on our chosen input perspective, we

could try to find a VLM that is trained on a related task such as RemoteCLIP[8], RSGPT[6], GRAFT[9], or even EarthGPT[13].

2.3. Fusion of 2D embeddings into 3D scene

We plan to distill the semantic embeddings into the 3D representation following either OpenScene[10] or OpenMask3D[12]. Given our ambitions of scale, the 3D mask-based approach appears particularly appealing, as it does not require explicit storage of embeddings for every point. Whether or not reasoning at instance level provides sufficient detail will have to be determined.

2.4. Evaluation

The research of this project is directed towards two main questions:

Firstly, does the language guidance provide new insights into urban scenes that were not easily extractable with conventional closed-set approaches? If successful, our model may be able to estimate e.g. the age of buildings, the type of and wealth of city districts, the number of churches, or the presence of construction sites. The relevant ground truth can be either extracted from a city registry or annotated by hand (in the case of somewhat rare objects). As these would be mostly new applications, we would not necessarily benchmark our method against existing methods but mainly demonstrate that the task is solved to some extent by providing qualitative results.

Secondly, can we improve performance on already partially solvable tasks by incorporating additional information? The combination of multiple perspectives (e.g. multiple satellites or street level and aerial imagery) may enhance results in remote sensing or urban scene understanding tasks. As a first step (and given the successful incorporation of multiple perspectives), we would like to benchmark our method on the ISPRS Potsdam dataset [4] of satellite image segmentation (buildings, roads, trees, etc) in a zero- or few-shot setting.

Alternatively, we can compare with respect to the Semantic Terrain Points Labeling - Synthetic 3D (STPLS3D) dataset, which offers a synthetic urban environment with 20 predefined classes for semantic segmentation tasks[2].

3. Milestones

1. Obtain or build a sparse city reconstruction with registered 2d images.
2. Apply or adapt either of the fusion methods to enrich it with semantic embeddings.
3. Enable extraction of information about concrete, tangible objects (buildings, roads, parks, etc.).
4. Enable extraction of abstract semantic information (year of the building, condition, etc.).

References

[1] Kyle Buettner and Adriana Kovashka. Investigating the role of attribute context in vision-language models for object recognition and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5474–5484, 2024. 1

[2] Meida Chen, Qingyong Hu, Zifan Yu, Hugues THOMAS, Andrew Feng, Yu Hou, Kyle McCullough, Fengbo Ren, and Lucio Soibelman. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 1, 2

[3] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7010–7019, 2023. 1

[4] International Society for Photogrammetry and Remote Sensing. 2d semantic labeling contest - potsdam. 2

[5] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model, 2024. 1

[6] Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023. 2

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[8] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *arXiv preprint arXiv:2306.11029*, 2023. 1, 2

[9] Utkarsh Mall, Cheng Perng Phoo, Meilin Kelsey Liu, Carl Vondrick, Bharath Hariharan, and Kavita Bala. Remote sensing vision-language foundation models without annotations via ground remote alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[10] 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

[11] 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

[12] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-Vocabulary 3D Instance Segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[13] Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multi-modal large lan-

guage model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822*, 2024. 2

[14] Ronglai Zuo, Fangyun Wei, and Brian Mak. Natural language-assisted sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14890–14900, 2023. 1