Vision-Language-Spatial Graph for Object Instance Queries on Construction Sites

Charles Raines†, Jingdao Chen†, and Sudip Mittal†
†Department of Computer Science, Mississippi State University

MISSISSIPPI STATE UNIVERSITY
ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION
ATLANTA 2025

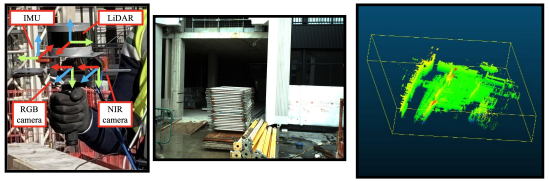
Background

- Tracking the location of equipment and materials on a construction site is important in robot navigation as well as construction monitoring and safety applications
- However, retrieving that information from cluttered images and point clouds is challenging

Objectives

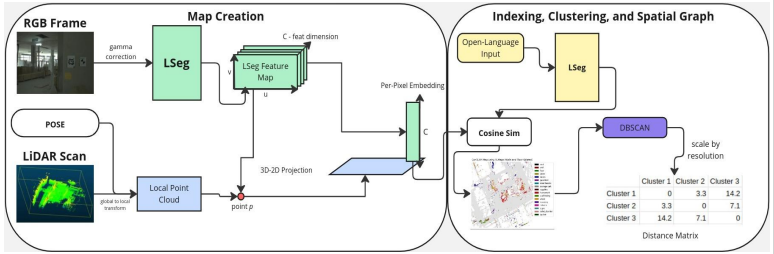
- Use Vision-Language maps to combine vision, language, and spatial features into a single representation of a construction site
- Use clustering and spatial graphs to reason about neighboring relationships between object instances

Dataset



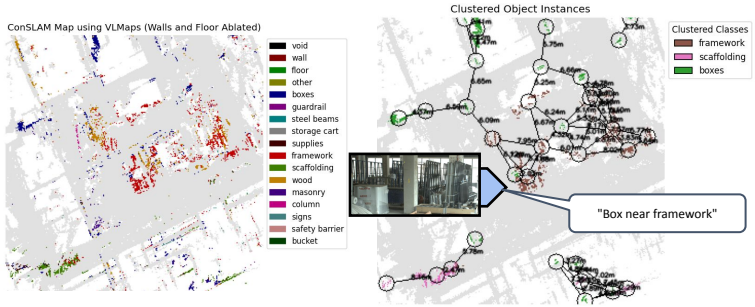
- ConSLAM Dataset: Construction Dataset for SLAM
- Includes data from LiDAR, RGB Camera, and IMU
- Per-Scan LiDAR poses estimated using LIO-SAM

Methodology



Our method proposes using RGB (gamma correction for brightness), laserscan, and pose input to obtain per-pixel embedding vectors. These embeddings are scored using LSeg for both visual and language input, resulting in a semantic map of a construction site after 3D-2D projection. The resulting classes of objects are then clustered using density-based clustering, allowing spatial relationships to be established between individual instances of each class.

Results



Discussion

- Results show that our method is able to reconstruct a rich semantic map from real-world construction site data in a zero-shot manner, as well as create an instance-aware spatial graph of objects of interest.
- Due to the grid-based discretization of the map, multiple LSeg embeddings projected to the same grid cell are handled by averaging, which may result in mixed class overlays and potential noise in the graph creation process.
- Gamma Correction was applied to overcome low brightness in the input images and ensure that effective visual features can be retrieved.
- An exciting future direction of research on this topic would be integrating the spatial matrix with a navigation agent and LLM to enable instance-aware navigation.

References

- Trzeciak, M. et al. (2023). ConSLAM: Periodically Collected Real-World Construction Dataset for SLAM and Progress Monitoring. In: Karlinsky, L., Michaeli, T., Nishino, K. (eds) Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science, vol 13807. Springer, Cham. https://doi.org/10.1007/978-3-031-25082-8_21
- Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., & Ranftl, R. (2022). Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546.
- C. Huang, O. Mees, A. Zeng and W. Burgard, "Visual Language Maps for Robot Navigation," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 10608-10615, doi: 10.1109/ICRA48891.2023.10160969.