THOTH: Task-Aware Semantic Memory for Aerial Missions

Joseph Letobar[1†], Arianna Lee[2†], Jylen Tate[3†], Clint Johnson[4]

***Abstract*—Long-duration aerial missions produce visual streams that are difficult for human operators or downstream systems to inspect and query after collection. We present THOTH (Task-aware, Hierarchical, Open-vocabulary, Traceable Heatmap), a system that converts aerial video into a persistent, grounded semantic representation of the observed scene. Given a natural-language task, a vision-language model proposes open-vocabulary scene labels and assigns each a priority score. A segmentation model grounds these labels into masks, which are visualized as inspectable priority maps and accumulated within a global knowledge graph. Graph nodes preserve grounded entities and their attributes, while spatial and open-vocabulary semantic edges capture relationships across the scene.**

We demonstrate the complete system on long-range aerial imagery through four case studies across two datasets, each addressing a distinct natural-language task. For each natural-language query, a large language model reasoned over the serialized knowledge graph, correctly identifying the task-relevant region or route and citing supporting entities and relationships. Corresponding high-priority imagery was then surfaced for visual verification.

***Index Terms*—semantic mapping, aerial robotics, vision-language models, open-vocabulary perception, knowledge graphs**

I. Introduction

Robots deployed for extended missions, including disaster response and remote inspection, collect far more visual data than any human operator or downstream system can directly consume. This motivates semantic compression of observations into a task-aware, persistent representation. Similarly, humans do not represent visual experiences as an exhaustive record of every retinal input (e.g., the precise color and texture of every passerby's shoes). Instead, attention selectively prioritizes information relevant to current goals, making certain aspects of a scene more salient, such as noticing those same shoes when considering buying a pair [1].

Within this representation, we follow Ruan et al. [2] in defining semantic information as the identity of entities and the explicit and implicit relationships between them in a scene. Work on intermittent supervision has shown that query-driven summaries substantially outperform generic summaries or raw playback in terms of retrieval accuracy [3], demonstrating that query specificity is necessary to support targeted search. Retrieval-augmented memory systems such as ReMEmbR [4] demonstrate a persistent, indexed memory that is queried on demand but lacks object identity or inter-object relational structure.

Open vocabulary object detection, segmentation, and language grounding models can compute transient per-frame object relations and narrate them in natural language [5] or structure them into object-relationship graphs that generalize across viewpoint and modality but have not been evaluated beyond ground-level, indoor settings [6]. Recent LLM-driven structures like SPINE [7] have enriched nodes with semantic attributes but keep relational structure thin by restricting edges to traversability or visibility between nodes. A stream of transient, per-frame model outputs or a sparse relational graph is not a persistent representation an operator or downstream planner can inspect, revisit, or query after the fact.

Recent aerial systems address this representation problem only partially. HALO [8] performs real-time monocular metric-semantic mapping, generates a task-relevancy representation, and executes language-conditioned exploration at altitude but maintains no relationship structure between detected entities, preventing graph-based querying or reasoning over the entire scene. USS-Nav [9] constructs an online hierarchical spatio-semantic scene graph on edge UAV hardware, integrating open-vocabulary object semantics into incremental topological mapping. However, the system is geared towards navigation efficiency, with graph reuse for scene-level query treated as a secondary capability rather than a primary output.

In this paper, we present THOTH (Task-aware, Hierarchical, Open-vocabulary, Traceable Heatmap), which turns visual streams into persistent, grounded, and inspectable semantic artifacts that scaffold downstream reasoning by operators, query systems, or other modules.

Our contributions include:

- **Task-aware semantic memory for aerial missions**
- **Queryable long-horizon scene comprehension**
- **Grounded, inspectable post-mission artifacts**

II. Method

THOTH is a task-aware system that converts aerial video into a structured representation. As shown in Fig. 1, the system combines open-vocabulary scene understanding (stage 1), promptable segmentation (stage 2), semantic mapping (stage 3a), and a persistent knowledge-graph representation of the scene (stage 3b) within a unified pipeline. A given natural

[1]Joseph Letobar is with the Department of Computer Science, University of Central Florida, Orlando, FL, USA. joseph.letobar@ucf.edu.

[2]Arianna Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. alee755@gatech.edu.

[3]Jylen Tate is with the Department of Computer Science, The University of Alabama, Tuscaloosa, AL, USA. jylentate@gmail.com.

[4]Clint Johnson is a Chief Scientist and Principal Research Scientist at the Georgia Tech Research Institute, Atlanta, GA, USA. clint.johnson@gtri.gatech.edu.

[†]Equal contribution

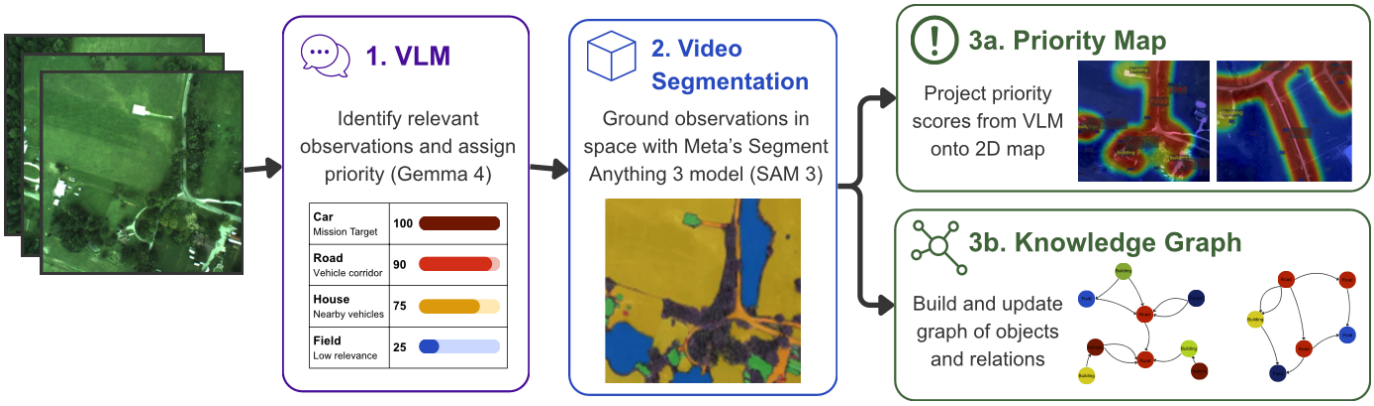

Fig. 1. THOTH system architecture, illustrated using the task *"Find cars."*

language task provides context that determines an observation's priority. The vision-language model (VLM) examines incoming imagery and proposes semantic labels alongside their priority scores. Then, Segment Anything Model 3 (SAM 3) [10] grounds these proposed labels into masks, visualized as *priority maps* and accumulated within a global *knowledge graph*.

This modular design separates semantic interpretation from visual grounding and long-term memory representation, aligning with 3D scene-graph systems such as Hydra [11] and 3D Dynamic Scene Graphs [12], which likewise decouple perception from graph construction. Neither incorporates task-awareness, and both restrict edges to a closed, largely geometric vocabulary. THOTH differs from these systems by targeting aerial imagery and inferring open-vocabulary semantic relationships directly via a VLM. THOTH also differs from task-conditioned systems like Clio [13], which uses relevance to decide which detections are kept in the final scene graph. THOTH instead uses the task to condition VLM outputs and visual grounding, supporting targeted scene comprehension while preserving the resulting semantic and visual evidence.

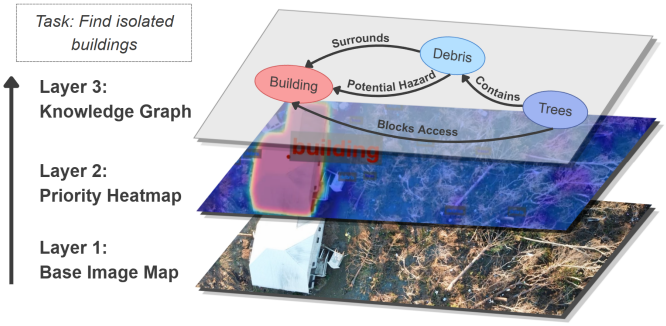

Fig. 2. Conceptual representation of the system hierarchy: base imagery from RescueNet [14], task-aware priority maps, and relational knowledge graphs.

A. Open-Vocabulary Scene Understanding

Vision-language models have been utilized broadly across the robot autonomy stack and open-vocabulary visual recognition [15]. In contrast to closed-set metric-semantic methods whose categories are fixed at training time [11], [12], [16], [17], our method uses a VLM to propose open-vocabulary labels with an assigned task-aware priority score, with higher scores indicating greater relevance to the specified task.

This stage inherits two limitations directly from the VLM. First, VLMs can hallucinate regions not present in the image. Second, their assigned priority scores may be miscalibrated by not capturing the common sense relationship between the label and task. We note that these limitations are closely tied to the scale and intelligence of the underlying VLM used (e.g., Gemma 4 E2B assigned a dense forest a priority score of 75/100 for a *car search* task). Our system was evaluated using Gemma 4 31B [18].

B. Grounded Semantic Observations and Priority Mapping

Each VLM-proposed label is passed into a segmentation model, which returns the corresponding mask localized in the image. This process enables creating a mission-specific priority map, as shown in Fig. 3, by color-coding each segmented region in accordance with its priority score using OpenCV's JET colormap [19]. The resulting priority map serves as an inspectable artifact, and we motivate further use in Section IV.

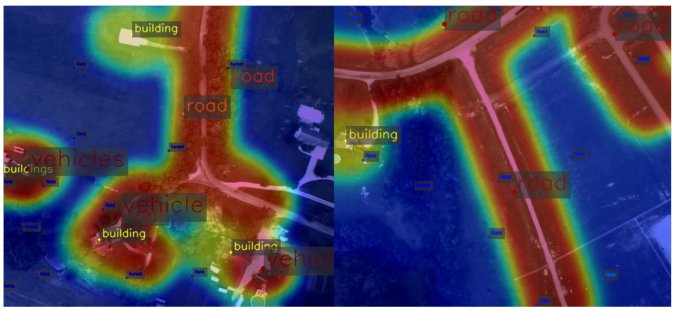

Fig. 3. The illustrated priority map corresponds to a *find cars* task, in which roads receive high priority. Images are from the ALTO dataset [20].

SAM 3 is prompted with the VLM-proposed label to obtain the mask, mirroring the multimodal large language model (MLLM)-to-SAM 3 grounding pattern demonstrated by SAM

3 Agent [10]. Prior open-vocabulary systems often relied on intermediate grounding models such as Grounding DINO [21] or OWL-ViT [22] to convert proposed text labels into spatial prompts before segmentation [6], [23]–[26].

C. Knowledge Graph

THOTH stores observations in a global knowledge graph that represents long-horizon scene memory. Masks are converted into individual nodes with a label, priority score, position, mask, and observation count. This gives an operator or downstream system information about what, where, and when an object has been seen. A node's location is projected onto the global representation to align with a constructed orthomosaic of the entire scene.

Edges represent spatial and semantic relationships between objects. Purely geometric spatial relationships such as directional (left of, right of, above, below), adjacency, containment, and occlusion can be computed directly from object masks and locations but cannot capture the semantic significance of a scene on their own. Recent work on functional 3D scene graphs makes a similar case within indoor robotics, where relationships inferred through VLMs and LLMs capture functional dependencies that spatial-only representations miss [27], leading to substantial improvements over spatial-relationship baselines compared to ConceptGraphs [6].

The VLM used during Section II-A infers relevant open-vocabulary relationships, referred to as *semantic edges* (e.g., blocking access to, surrounded by, potential hazard to), to enhance task-aware scene comprehension.

Individual, per-frame knowledge graphs are merged into the persistent global graph. For every node in a local graph, spatial and semantic proximity determine whether it re-observes an existing global node or represents a new one. Local edges are then remapped and folded into the global graph, where duplicate edges are collapsed.

III. THOTH Evaluation

We demonstrate a downstream capability across the ALTO [20] and RescueNet [14] datasets using four case studies, each with a distinct task and associated question-and-answer pair. After each recorded flight, the resulting knowledge graph containing nodes, spatial edges, and semantic edges was provided to Gemma 4 31B as a structured *scaffold* for reasoning.

We describe one case study in detail and summarize the remaining three more briefly.

A. Case Study Overview

TABLE I
EVALUATION RESULTS
✓ INDICATES A CORRECT ANSWER

Dataset	Task	Question	Selected answer
ALTO	Flood risk	Most urgent flood-risk check after heavy rain?	✓Riverside houses linked to nearby roads
ALTO	Safe landing	Where should the helicopter land?	✓Open clearing beside buildings and a road
RescueNet	Isolated buildings	Which building is hardest for responders to reach?	✓Storm-damaged house surrounded by debris
RescueNet	Road access	Which route should ground responders use?	✓Road beside the residential strip

B. Detailed Case Study

This section provides a detailed account of the ALTO *Find flood risks* case study from Table I while clearly specifying the experimental setup.

The model was asked:

> After heavy rain, which area would be the most urgent flood-risk check?
> **A**. A riverside cluster of houses connected to the nearby road network.
> **B**. A large pond beside parking lots, with drainage toward nearby buildings.
> **C**. Waterfront houses beside a road running along the shore.
> **D**. A forest-edge quarry with standing water beside a building group.

Distractors were made to be plausible but unsupported by the scene, requiring scene reasoning rather than wording alone.

The model selected **A**, the correct answer. Its reasoning identified the following evidence:

> **River 0 area:** This is the highest-priority zone. *buildings_46* is explicitly marked *vulnerable_to_river_flooding*, while *building_32*, *buildings_45*, *buildings_43*, and *buildings_47* are connected to *river_0* through *potential_flood_source_for* relationships.

The knowledge graph occupied 680 KiB, a 246× reduction relative to the corresponding raw-frame representation.

The cited nodes were traced back to their source frames and visualized in a COLMAP-based 3D reconstruction, as shown in Fig. 4 [28], [29].

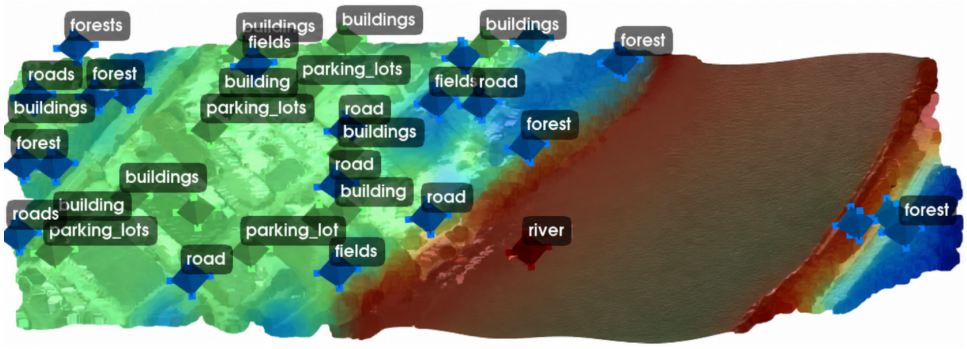

Fig. 4. 3D reconstruction providing an operator-inspectable summary of the high-priority *flood-risk* region.

C. Candidate Choices

We additionally report the other candidate areas considered by the model for the detailed case study, demonstrating that the scene contained multiple plausible flood-risk regions and providing insight into the evidence used to distinguish among them, as shown in Fig. 5:

River 1 area: The graph contains a vegetated riverbank and roads around *river_1*, but no explicit relationships connecting the river to vulnerable buildings.

Pond and quarry systems: The graph contains 2 ponds and drainage relationships, including *pond_0* draining into *quarry_0* and *quarry_1*, but no explicit flood-risk relationships connecting these features to buildings or roads.

Road 21 drainage area: *pond_5* and *pond_6* are connected to *road_21* through *adjacent_to_drainage_path*. This represents a plausible rain-related risk, but the evidence is weaker than the explicit River 0 vulnerability relationships.

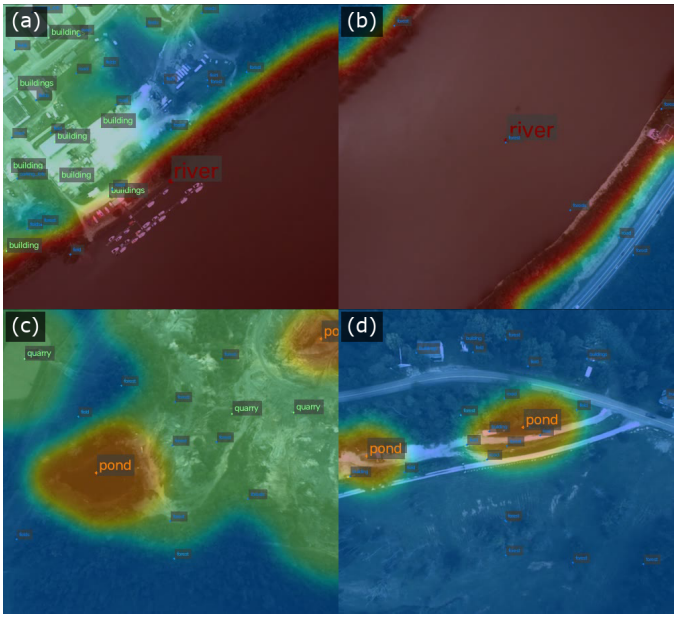

Fig. 5. Qualitative visualizations of the candidate flood-risk regions: **(a)** the selected River 0 area, **(b)** River 1, **(c)** the pond-and-quarry system, and **(d)** the Road 21 drainage area.

D. Ablation Study

TABLE II
GRAPH ABLATION RESULTS OVER THE FOUR QUALITATIVE QUESTIONS

Memory representation	Correct
Full graph (nodes + spatial + semantic edges)	4/4 (100%)
Semantic edges only	3/4 (75%)
Nodes only (no edges)	3/4 (75%)
Spatial edges only	2/4 (50%)
No graph (question and choices only)	0/4 (0%)

As shown in Table II, the complete graph achieved the best performance, while every ablation degraded accuracy. Nodes alone were sufficient for some questions, but other cases appeared to benefit from explicit edges. Spatial and semantic edges therefore provide complementary evidence: either edge type alone was less reliable than their combination, while the no-graph condition failed completely, indicating that the questions were not answerable from wording alone.

While limited to a small case study, these results support our fused representation design. A larger and more rigorous evaluation is needed to establish a broader benchmark. We hypothesize that, given the broad priors of the foundation models used, the observed capability will extend beyond these four cases to a wider range of tasks and scenes.

IV. Discussion

A. Priority Map

The priority map provides an intermediate, task-aware representation of visual observations within the hierarchical processing structure (Fig. 2), turning raw camera streams into a perceptual abstraction filtered through the given task. While the current evaluation uses it primarily as an inspectable artifact, this intermediate representation enables downstream systems to operate on already-processed, task-relevant information rather than repeatedly interpreting the original visual stream.

For example, in a closed-loop autonomous system, a goal-directed agent could allocate its motion toward regions with the highest expected task priority. A *priority-seeking controller* could repeatedly select the most valuable available observation according to the current task, potentially resulting in more efficient use of the agent's trajectory compared to coverage-driven exploration strategies.

B. Knowledge Graph

As we target aerial scenes, we do not currently model temporal or motion relationships, which may be more necessary in ground-based scenes to capture object motion and interactions that aerial resolution often treats as negligible.

Adapting an existing knowledge graph to a new task, or to newly available information (e.g., in a *find-a-car* task, the operator later learns that the vehicle crashed in a dense forest), requires deciding which parts of the graph can be reused and which must be recomputed. Spatial edges can be recomputed from stored masks and map coordinates, whereas semantic edges require VLM inference over the scene for the new task. In the evaluation shown in Section III-D, each task was processed from the flight data anew. Therefore, adapting an existing graph would require rerunning semantic inference rather than simply updating priority scores.

We additionally hypothesize that future compression should reduce redundancy while preserving enough scene context to answer unanticipated queries. Redundancy-aware node clustering and bounded-memory policies are promising directions for scaling persistent semantic memory.

V. Conclusion

We presented THOTH, a task-aware semantic memory system for long-horizon aerial video. By combining open-vocabulary scene understanding, promptable segmentation, and knowledge graph construction, THOTH preserves entities and their relationships as a semantic scaffold for post-mission reasoning. Our evaluation shows a large language model can query this memory to identify task-relevant regions and surface supporting visual evidence. These results point toward persistent semantic memory extending robot understanding beyond transient, per-frame perception. Future work will target longer missions, larger benchmarks, and closed-loop task-guided observation.

References

[1] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S004269890400392X

[2] T. Ruan, H. Wang, R. Stolkin, and M. Chiou, "A taxonomy of semantic information in robot-assisted disaster response," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, Nov. 2022, pp. 285–292.

[3] K. Katuwandeniya, L. Tian, and D. Kulić, "What did the robot do in my absence? video foundation models to enhance intermittent supervision," *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3222–3229, Apr. 2025.

[4] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, "ReMEmbR: Building and reasoning over long-horizon spatio-temporal memory for robot navigation," 2024. [Online]. Available: https://arxiv.org/abs/2409.13682

[5] Z. Wang, B. Liang, V. Dhat, Z. Brumbaugh, N. Walker, R. Krishna, and M. Cakmak, "I can tell what i am doing: Toward real-world natural language grounding of robot experiences," in *Proceedings of the 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270. PMLR, 2025, pp. 1863–1890. [Online]. Available: https://proceedings.mlr.press/v270/wang25g.html

[6] 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)*. IEEE, 2024, pp. 5021–5028.

[7] Z. Ravichandran, V. Murali, M. Tzes, G. J. Pappas, and V. Kumar, "SPINE: Online semantic planning for missions with incomplete natural language specifications in unstructured environments," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2025, pp. 13 714–13 721.

[8] Y. Tao, D. Ong, F. Cladera, J. Hughes, C. J. Taylor, P. Chaudhari, and V. Kumar, "Halo: Language-conditioned overhead monocular aerial exploration and navigation," *IEEE Robotics and Automation Letters*, vol. 11, no. 7, pp. 8204–8211, 2026.

[9] W. Gai, Y. Gao, Y. Zhou, Y. Xie, Z. Liu, Y. Wu, X. Zhou, F. Gao, and Z. Meng, "USS-Nav: Unified spatio-semantic scene graph for lightweight UAV zero-shot object navigation," 2026. [Online]. Available: https://arxiv.org/abs/2602.00708

[10] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. S. Coll-Vinent, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. HAZRA, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollar, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "SAM 3: Segment anything with concepts," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=r35clVtGzw

[11] 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 XVIII*, New York, NY, USA, Jun. 2022.

[12] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, "3D dynamic scene graphs: Actionable spatial perception with places, objects, and humans," in *Robotics: Science and Systems*, 2020.

[13] 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, Oct. 2024.

[14] M. Rahnemoonfar, T. Chowdhury, and R. Murphy, "RescueNet: A high resolution UAV semantic segmentation dataset for natural disaster damage assessment," *Scientific Data*, vol. 10, no. 1, p. 913, 2023.

[15] R. Firoozi, J. Tucker, S. Tian, A. Majumdar, J. Sun, W. Liu, Y. Zhu, S. Song, A. Kapoor, K. Hausman, B. Ichter, D. Driess, J. Wu, C. Lu, and M. Schwager, "Foundation models in robotics: Applications, challenges, and the future," *The International Journal of Robotics Research*, vol. 44, no. 5, pp. 701–739, 2025. [Online]. Available: https://doi.org/10.1177/02783649241281508

[16] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3D scene graph: A structure for unified semantics, 3D space, and camera," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 5664–5673.

[17] 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 (CVPR)*, Jun. 2021, pp. 7515–7525.

[18] Gemma Team, "Gemma 4 technical report," Jul. 2026.

[19] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.

[20] I. Cisneros, P. Yin, J. Zhang, H. Choset, and S. Scherer, "ALTO: A large-scale dataset for UAV visual place recognition and localization," 2022. [Online]. Available: https://arxiv.org/abs/2207.12317

[21] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Computer Vision–ECCV 2024*, ser. Lecture Notes in Computer Science, vol. 15105. Springer Nature Switzerland, 2024, pp. 38–55.

[22] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, "Simple open-vocabulary object detection with vision transformers," in *Computer Vision–ECCV 2022*, ser. Lecture Notes in Computer Science, vol. 13670. Springer Nature Switzerland, 2022, pp. 728–755.

[23] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "VoxPoser: Composable 3d value maps for robotic manipulation with language models," in *Proceedings of the 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 540–562. [Online]. Available: https://proceedings.mlr.press/v229/huang23b.html

[24] P. Nguyen, T. D. Ngo, E. Kalogerakis, C. Gan, A. Tran, C. Pham, and K. Nguyen, "Open3DIS: Open-vocabulary 3d instance segmentation with 2d mask guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2024, pp. 4018–4028.

[25] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024. [Online]. Available: https://arxiv.org/abs/2401.14159

[26] X. Wang, W. He, X. Xuan, C. Sebastian, J. P. Ono, X. Li, S. Behpour, T. Doan, L. Gou, H.-W. Shen, and L. Ren, "USE: Universal segment embeddings for open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2024, pp. 4187–4196.

[27] C. Zhang, A. Delitzas, F. Wang, R. Zhang, X. Ji, M. Pollefeys, and F. Engelmann, "Open-vocabulary functional 3D scene graphs for real-world indoor spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 19 401–19 413.

[28] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4104–4113.

[29] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision–ECCV 2016*. Springer Intervaltional Publishing, 2016, pp. 501–518.