Real World Semantic Occupancy Prediction for Advanced Air Mobility

Sai Bharadhwaj Matha

Introduction

Accurate scene perception is crucial to the functionality and safety of autonomous robotic systems, as it directly impacts navigation efficiency and decision-making. In ground-based autonomous driving, deep learning-based semantic occupancy grids have advanced perception by combining spatial geometry with semantic understanding of the scene.

This research extends and enhances these advancements to Unmanned Aerial Vehicles (UAVs) by developing a novel semantic occupancy dataset from monocular aerial imagery and benchmarking state-of-the-art semantic occupancy prediction models by adapting them to the aerial domain. The objective is to address the distinct challenges posed by aerial navigation and contribute to enhanced autonomy for UAVs operating in complex real-world environments.

Research Overview

Current semantic occupancy datasets, such as SemanticKITTI [1] and SSCBench [2], are tailored to autonomous driving and predominantly depend on LiDAR data. However, LiDAR has notable limitations: it produces sparse data, suffers from increased noise at long ranges, and incurs high costs. Moreover, existing datasets typically offer only surface-level occupancy (hollow object representation), limited semantic classes, visibility unawareness, and a lack of panoptic or instance-level information.

To overcome these shortcomings, this research proposes a comprehensive pipeline for constructing a high-resolution semantic scene completion (SSC) dataset suitable for aerial robotic applications. The pipeline consists of the following key stages:

1. **Data Collection:** Monocular RGB images captured from altitudes ranging between 5 and 60 meters, encompassing diverse terrains and urban environments.

2. **3D Reconstruction:** Dense 3D point clouds are reconstructed using classical Structure from Motion (SfM) and Multi-View Stereo (MVS) methods, ensuring geometric and photometric consistency without relying on learned priors.

3. **2D-3D Semantic Lifting:** Manual annotations/Custom-trained semantic segmentation models generate masks that are fused with the reconstructed point clouds, resulting in semantically enriched 3D representations.

4. **Voxelization:** A novel voxelization technique densifies object interiors, such as buildings and vehicles, rather than just surfaces, supporting instance-aware voxel grids and panoptic segmentation aiding advanced path-planning.

5. **Aggregation and Hole-Filling:** Class-specific voxel grids are merged into unified scene voxel grids. A tailored surface mesh based novel hole-filling algorithm is applied, particularly to correct ground-level voids.

6. **Ground Extrusion:** A ground extrusion algorithm models real-world topography logically, improving alignment with practical UAV operations.

7. **Frustum Ground Truth Generation:** Using calibrated camera poses, a frustum-aligned voxel grid is created for each image to serve as ground truth. Voxels are assigned both semantic labels and surface-type categories, expanding on SemanticKITTI's taxonomy with seven additional classification types.

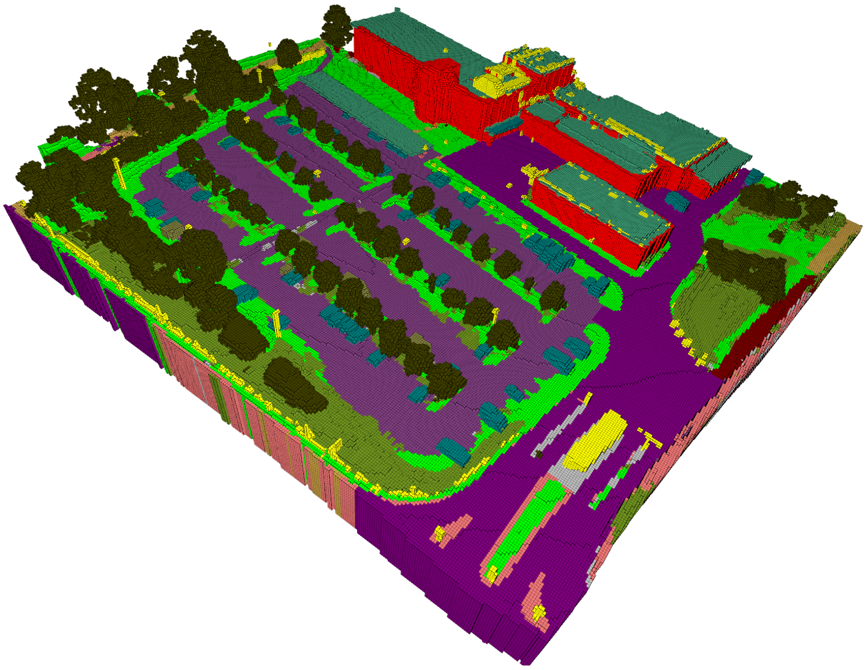

Figure 1: A dense voxelized 3D semantic representation of an example scene.

This end-to-end pipeline enables the generation of high-fidelity training data for semantic occupancy prediction tasks in aerial robotics. The ground truth voxel grid per sample contains 3.1 million voxels (192 × 128 × 128) with a voxel size of 0.5 meters.

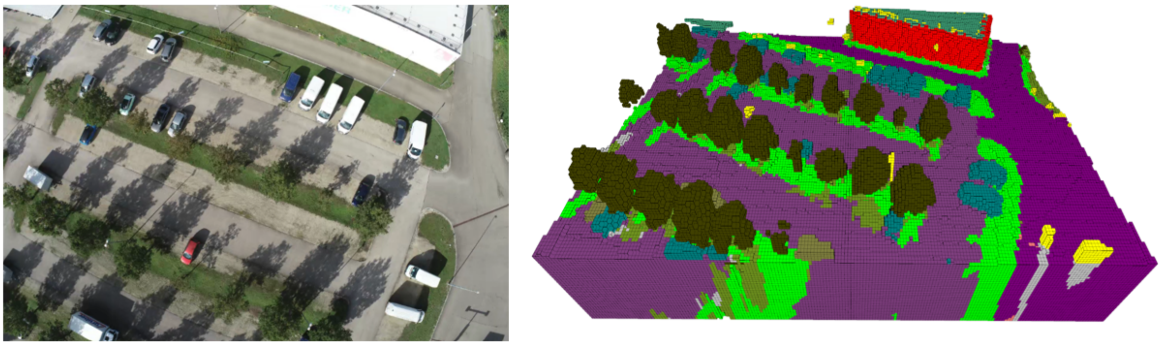

Figure 2: Sample monocular RGB image and its corresponding frustum-aligned ground truth voxel grid. The frustum extends beyond the field of view, enabling prediction models to hallucinate unseen regions.

Model Training and Evaluation

To validate the quality of the dataset, state-of-the-art semantic occupancy prediction models like **MonoScene** [3], **Symphonize 3D** [4], **CGFormer** [5] were trained on the generated aerial dataset.

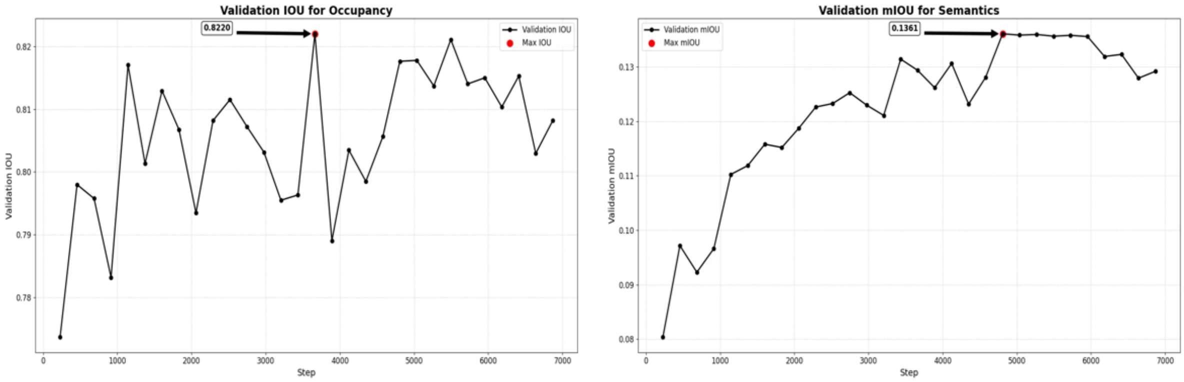

Figure 3: Validation performance metrics: IoU for occupancy prediction and mIoU for semantic segmentation.

The training outcomes significantly outperformed those on traditional datasets like SemanticKITTI. In Figure 3, the results of the training on Monoscene with very limited training data are shown. The dataset during the initial training included ground truths from only two scenes, which is now expanded to 20 scenes, see the section below.

- **Occupancy IoU:** 82.2% (compared to 42.51% on SemanticKITTI hidden test set)

- **Semantic mIoU:** 13.6% (vs. 11.08% on SemanticKITTI)

These results are particularly noteworthy, given that the feature extractors were not extensively trained on aerial data, which highlights the strength and generalizability of the proposed dataset. But, there is a need for significant architectural changes and an increase in the scale of the dataset, as evident from the predictions shown in Figure 4.

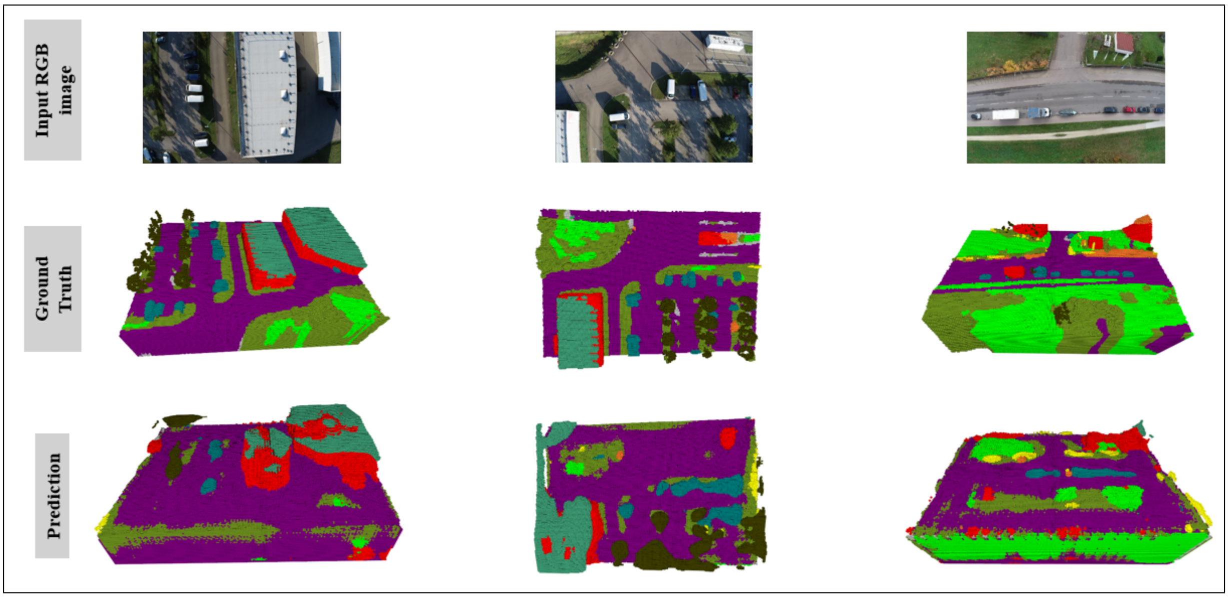

Figure 4: Sample prediction results from the validation set showing semantic occupancy output.

Ongoing Developments and Future Directions

Since the initial submission of this thesis, substantial progress has been made in both dataset construction and model development, further advancing the scope of semantic scene completion for aerial robotics.

1. **Dataset Expansion and Multimodal Integration:** The dataset, which originally contained only two annotated scenes, has been scaled to **20 diverse scenes** comprising over **50,000 data samples**. This expansion covers more than **one million square meters** of heterogeneous environments, including urban, semi-urban, and natural terrains with varying topographical complexity. The voxel resolution is increased to 0.3 meters, making a frustum with 288×224×224 voxels, which makes it 6 times larger than the SemanticKITTI or SSCBench frustum voxel grid.
Furthermore, the dataset now includes **thermal imagery**, enabling multimodal semantic occupancy modeling. Work is in progress to generate **thermal occupancy ground truths**, making this the **first dataset** to integrate thermal modalities in

both autonomous driving and aerial perception, an unexplored area in the literature.

2. **Panoptic Scene Completion and Instance Awareness:** Beyond semantic occupancy, the dataset now supports **panoptic scene completion (PSC)**, with fine-grained **instance-level annotations** for key object classes. This enhancement allows models to infer both semantic categories and distinguish individual object instances (e.g., vehicles, trucks, buildings). Figure 5 shows a sample panoptic frustum-aligned voxel grid, where different instances are visualized with unique color encodings.

3. **Model Ablations and Novel Architectural Development:** Extensive ablation studies on existing SSC models such as **MonoScene**, **Symphonize 3D**, and **CGFormer** have been conducted to evaluate their adaptation to aerial viewpoints. Insights from these experiments have informed the design of a **novel SSC architecture** specifically optimized for aerial vehicles, addressing key challenges such as large-scale context reasoning, domain adaptation between ground and aerial data, and multimodal feature fusion.

4. **Research Roadmap:** The ongoing work is strategically directed towards a high-impact **CVPR 2026 publication**. By combining large-scale panoptic aerial datasets, multimodal occupancy representations, and a specialized SSC model, this research aims to establish a new benchmark for **Panoptic Scene Completion in Advanced Aerial Mobility**.

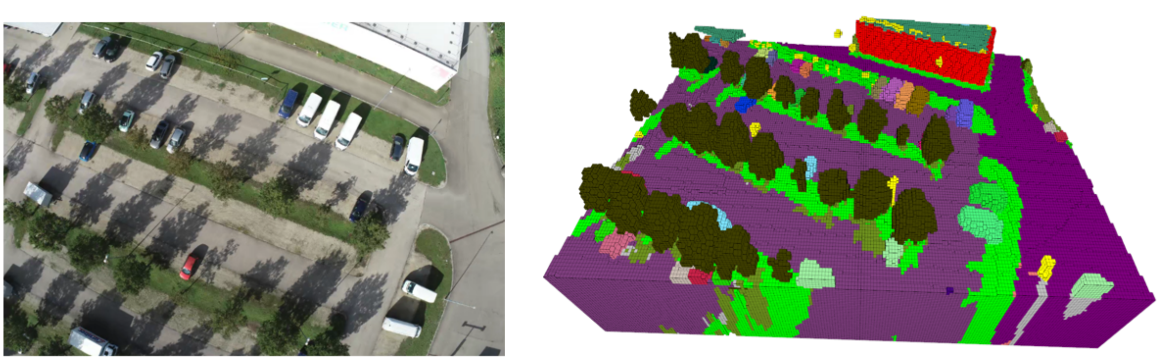

Figure 5: Sample monocular RGB image and its corresponding panoptic frustum-aligned ground truth voxel grid. Vehicle instances are assigned unique colors for improved visualization of instance-level segmentation.

Conclusion

This thesis presents a novel semantic occupancy dataset generation pipeline for UAVs, bridging a major gap in aerial robotics perception research. Through rigorous reconstruction, fusion, and voxelization techniques, it provides a rich and scalable foundation for semantic occupancy prediction. Empirical results with the Monoscene and other SOTA models confirm the dataset's utility and its potential for further advancements in aerial autonomy. The latest advancements increased the scale, incorporated panoptics for PSC, and are working towards a novel SSC model architecture

References

[1] J. Behley, M. Garbade, A. Milioto, *et al.*, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[2] Y. Li, S. Li, X. Liu, *et al.*, "Sscbench: A large-scale 3d semantic scene completion benchmark for autonomous driving," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[3] A.-Q. Cao and R. de Charette, *Monoscene: Monocular 3d semantic scene completion*, 2022. arXiv: `2112.00726 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/2112.00726`.

[4] H. Jiang, T. Cheng, N. Gao, *et al.*, "Symphonize 3d semantic scene completion with contextual instance queries," *CVPR*, 2024.

[5] Z. Yu, R. Zhang, J. Ying, *et al.*, "Context and geometry aware voxel transformer for semantic scene completion," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 1531–1555.