AeroGrid100: A Real-World Multi-Pose Aerial Dataset for Implicit Neural Scene Reconstruction

Qingyang Zeng, Adyasha Mohanty
Harvey Mudd College, Claremont, CA
bozeng@g.hmc.edu, admohanty@g.hmc.edu

***Abstract*—Aerial scene reconstruction, view synthesis, and robotics applications increasingly demand large-scale drone datasets with comprehensive spatial-angular coverage to enable robust neural rendering and embodied intelligence systems. However, existing UAV datasets typically provide only nadir views and lack precise pose annotations, limiting their utility for advanced multi-view learning tasks and requiring computationally expensive Structure-from-Motion pipelines to obtain camera poses. To address these limitations, we introduce AeroGrid100, a novel multi-pose aerial image dataset tailored for neural scene reconstruction, view synthesis, and aerial robotics research. Captured using a high-resolution DJI Air 3 drone across diverse semi-urban environments, the dataset contains over 15,000 images from 100 geospatial anchors, each systematically sampled at five altitudes and across a grid of yaw and pitch angles. AeroGrid100 provides ground-truth 6-DoF camera poses, eliminating the need for Structure-from-Motion pipelines. The dataset's dense spatial-angular coverage can enable rapid progress on research topics such as reinforcement learning-based path planning in discrete orientation spaces and geometry-aware vision-language modeling.**

I. Introduction

Unmanned aerial vehicles (UAVs) increasingly operate in challenging environments such as urban canyons, poor weather, and GPS-denied zones, demanding robust localization and scene understanding. As GPS suffers from multipath issues in such settings, sensor fusion with onboard cameras and environment-aware models is gaining traction. In this context, implicit neural representations have emerged as a powerful alternative to traditional geometric modeling methods. Unlike voxel grids or mesh-based reconstructions, which suffer from memory inefficiency and discretization artifacts [21], implicit representations like neural implicit surfaces [19], signed distance functions [13], and occupancy networks [11] offer continuous and compact scene encoding at arbitrary resolution. Among these, Neural Radiance Fields (NeRFs) [12] stand out as a leading method for photorealistic scene reconstruction and novel view synthesis [4], mapping 5D inputs (spatial location and viewing direction) to view-dependent radiance and volume density. Their ability to generate high-fidelity reconstructions from sparse image sets makes them suitable for aerial applications, where acquiring dense viewpoints may be infeasible.

Recent efforts have extended NeRF frameworks to aerial scenarios. On the data collection side, UAV-specific protocols have been developed to capture diverse multi-view imagery within constrained flight envelopes [22]. Algorithmic adaptations such as Mega-NeRF [18] and Block-NeRF [17] have tackled the scalability challenges of large outdoor scenes by spatially partitioning environments, allowing for parallelized training and rendering. Applications of NeRFs in aerial domains have expanded to include semantic segmentation, instance-level modeling, and vision-language integration [23]. However, the application of these methods to real-world urban and semi-urban environments remains limited by the lack of structured datasets that offer comprehensive spatial and angular coverage along with accurate pose metadata.

Several datasets have attempted to address parts of this gap. UAVid [10] provides annotated nadir-view imagery for semantic segmentation but lacks the pose and viewpoint diversity needed for NeRF-based modeling. Indoor datasets like ICL-NUIM [6] and outdoor datasets such as ETH Zurich MAV [2] offer accurate pose tracking but are constrained by narrow camera motion and limited angular variability. OpenDroneMap [1] and PIX4D datasets [15] support aerial mapping workflows but do not include dense 6-DoF pose annotations or structured view sampling. FPV-NeRF [22] assembles drone footage into simulated NeRF trajectories but is limited by its lack of systematic multi-angle sampling and reliance on post-processed egocentric video. Similarly, DroNeRF [14] optimizes drone trajectories for object-level NeRF training but focuses on small-scale scenes and sparse coverage.

In parallel, simulation-based datasets for aerial vision-language navigation (VLN)—such as UrbanScene3D [8], AerialVLN [9], CityNav [7], OpenUAV [20], and OpenFly [5]—have advanced language-grounded UAV perception. These datasets emphasize semantic navigation and instruction following, but typically suffer from narrow angular coverage, simplified vehicle dynamics, and idealized lighting, limiting their utility for photorealistic reconstruction. For instance, OpenFly [5] provides over 100K simulated UAV trajectories but is optimized for language tasks rather than high-fidelity 3D modeling. Similarly, the AVDN dataset [3] uses fixed-altitude nadir satellite imagery and lacks angular variation and real-world depth complexity.

To address these limitations, we present **AeroGrid100**, a real-world UAV dataset designed for neural rendering and spatial reasoning tasks in semi-urban environments. Captured using a DJI Air 3 drone, AeroGrid100 includes 17,100 high-resolution images collected across a 10×10 grid of geospatial anchor points. Each point is sampled at five altitudes and from multiple yaw-pitch configurations: 64 inner points

with eight yaw angles, 32 edge points with five orientations, and four corners with three viewpoints. Every image is paired with precise 6-DoF camera poses, aligned with the OpenGL convention used in NeRF pipelines, thus removing the need for structure-from-motion processing via tools like COLMAP [16]. Unlike datasets driven by continuous drone motion, AeroGrid100 uses a pose-centric sampling strategy, where each $(x, y, z, \text{yaw}, \text{pitch})$ configuration is treated as a discrete observation. This enables both flexible construction of synthetic flight paths and parallelized analysis of scene coverage. As such, AeroGrid100 supports robust benchmarking for NeRF reconstruction, view synthesis, and discrete path planning, while also enabling future extensions in semantic segmentation, reinforcement learning, and multimodal perception for aerial robotics.

II. Methodology

A. Data Collection and Image Specifications

Images were collected using a DJI Air 3 drone equipped with a dual-lens camera system. For consistency, only the wide-angle camera with a 24 mm equivalent focal length was used. Each image has a resolution of 4032×2268 pixels and is saved in high-quality JPEG format.

TABLE I: Summary of AeroGrid100 Dataset Statistics

Total images captured	17,100
Grid size	10×10
Altitudes per grid point	5 (20m, 40m, 60m, 80m, 100m)
Yaw and pitch configuration	Up to 8 yaw $\times$ 5 pitch(see Fig. **??**)
Image resolution	4032×2268
Camera type	DJI Air 3 (Wide-Angle)
Metadata format	JSON with full extrinsics
Flight log availability	Complete drone trajectory included

To ensure high data quality, all flights were conducted during daytime hours—typically between 10:00 AM and 6:00 PM—under clear skies with minimal wind. These conditions helped minimize shadows, lighting variability, and motion artifacts. The drone operated in stable flight modes throughout the data collection process, and all captured images were reviewed post-flight to filter out frames with excessive motion blur or exposure imbalance.

B. Pose Generation

Each image is accompanied by a JSON metadata file that includes the full camera-to-world extrinsic matrix, aligned with NeRF's OpenGL coordinate system, as well as GPS coordinates and altitude at the moment of capture. Camera poses were derived from the drone's onboard GNSS and IMU readings; orientation data was used to compute yaw, pitch, and roll angles, which were then converted into a camera-to-world transformation matrix compatible with NeRF conventions. All pose metadata is stored in a structured JSON format, enabling direct integration with NeRF-based frameworks. Additionally, the complete flight log is provided, offering a detailed record of the drone's trajectory and behavior during data collection to support reproducibility and facilitate further analysis of flight dynamics and scene coverage.

C. Geolocation and Grid Partitioning

The proposed dataset, AeroGrid100, was collected within a well-defined urban region located in Claremont, California. The sampling strategy was carefully designed to ensure systematic coverage of the area, capturing the scene from diverse altitudes and perspectives to support robust aerial reconstruction and rendering tasks.

The survey region, shown in Figure 2a, covers approximately 0.209 square kilometers (equivalent to 51.7 acres) and is geographically bounded by the following GPS coordinates: the top-left corner at $(34.10536, -117.71315)$, the top-right at $(34.10530, -117.70821)$, the bottom-left at $(34.10127, -117.71312)$, and the bottom-right at $(34.10123, -117.70807)$. Each edge of the quadrilateral region measures approximately 0.458 kilometers (0.285 miles), forming a near-rectangular coverage zone ideal for structured aerial sampling. The selected area represents a mixed-use urban environment. It includes academic buildings, roads, parking areas, and natural elements such as trees, grass, and landscaped vegetation. In addition, the presence of pedestrians and occasional vehicles introduces dynamic scene elements that more accurately reflect real-world conditions. As such, it provides a challenging and valuable testbed for evaluating view-dependent neural rendering methods, generalization capabilities, and robustness of 3D scene reconstruction algorithms under realistic aerial constraints.

(a) Academic buildings and walkways

(b) Urban road networks and parking lots

(c) Vegetation and landscaped areas

(d) Dynamic elements: pedestrians and vehicles

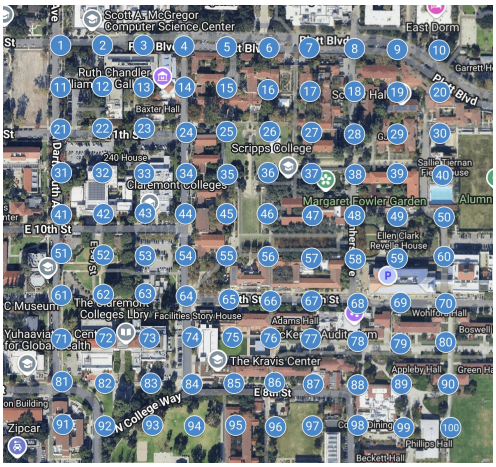
Fig. 1: Sample scenes from the *AeroGrid100* dataset showing a wide range of urban features, including built structures, transportation infrastructure, natural textures, and dynamic elements.

D. Sampling Strategy

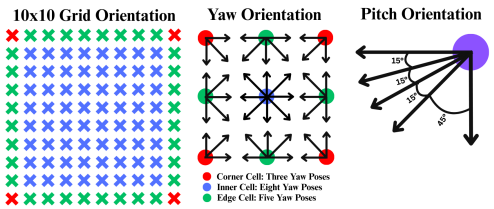
The dataset is constructed over a 10×10 grid, yielding 100 uniformly distributed geospatial anchor points that span the entire survey region. At each point, images are captured from multiple altitudes and orientations to maximize spatial and angular diversity.

Each grid point is sampled at five fixed altitudes: 20 m,

40 m, 60 m, 80 m, and 100 m. These heights were selected to capture both low-angle and elevated views of the environment for robust 3D reconstruction. At each altitude, images are captured by systematically varying the drone's yaw and pitch angles. Five fixed pitch angles (-10, -20, -30, -40, -50) are used uniformly across all yaw directions and grid locations to ensure consistent vertical sampling. Yaw angles vary by spatial location to optimize coverage and minimize redundancy. As shown in Figure 2a(b), inner grid points (blue) are sampled with 8 yaw angles at 45° intervals. Edge points (green) use 5 inward-facing yaw angles. Corner points (red) are sampled at 3 yaw angles aligned with the adjacent edges and inward diagonal. This structured sampling strategy yields **17,100 images** with precise camera-to-world pose metadata, offering high spatial-angular coverage across the survey region. Combining adaptive yaw and fixed pitch angles results in efficient image acquisition for evaluating multi-view 3D reconstruction and view synthesis under realistic aerial constraints. The sampling pattern is shown in Figure 2b.



(a) Sampling grid with fixed geospatial anchor points.



(b) Spatial variation as a result of varying the drone's yaw and pitch orientations.

Fig. 2: AeroGrid100 coverage and environment illustration.

III. Results

Our final NeRF reconstruction demonstrates high-fidelity scene synthesis across the full 10×10 spatial grid. As shown in Fig. 3, the reconstructed scenes capture a broad range of architectural textures and geometric features with impressive detail. The tiled rooftops, solar panels, tree canopies, and building contours are well preserved, indicating the robustness of our data collection and pose estimation pipeline. Finer-grained elements, such as shadows cast by trees and rooftop structures, further validate the photometric accuracy of the reconstruction.

While some finer occlusions and thin structures (e.g., small tree branches, window grilles) are less pronounced or slightly blurred, the overall spatial coherence remains intact. Notably, the dataset enables diverse scene generation, from densely vegetated courtyards to wide urban layouts, suggesting its suitability for downstream tasks such as aerial path planning, semantic segmentation, and view synthesis in urban-suburban environments. The full AeroGrid100 dataset is publicly accessible online here.

Fig. 3: Sample frame from the final reconstructed 3D scene showcasing architectural diversity and geometric fidelity.

IV. Conclusion

We present **AeroGrid100**, a large-scale, real-world aerial dataset specifically designed for multi-view 3D reconstruction and NeRF-based novel view synthesis. Unlike prior datasets that rely on synthetic simulations or nadir-only imagery, AeroGrid100 is constructed entirely from **UAV-captured images** with precise 6-DoF pose metadata, diverse camera orientations, and structured altitude variation. Initial NeRF reconstructions using a subset of the dataset demonstrate the quality and consistency of the collected data. By combining dense spatial sampling with real-world photometric complexity, AeroGrid100 provides a robust foundation for benchmarking and advancing neural radiance field models in realistic outdoor environments. Future work will extend the dataset to full spatial coverage, incorporate dynamic scene elements, and evaluate performance across multiple NeRF architectures.

References

[1] OpenDroneMap Authors. Opendronemap: A command line toolkit to generate maps, point clouds, 3d models and dems from drone, balloon or kite images, 2020. URL https://github.com/OpenDroneMap/ODM. Accessed: [23 October].

[2] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. Achtelik, and R. Siegwart. The zurich urban micro aerial vehicle dataset. *International Journal of Robotics Research*, 36(10):1160–1165, 2017. doi: 10.1177/0278364917715063.

[3] Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Eric Wang. Aerial vision-and-dialog navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3043–3061, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.190. URL https://aclanthology.org/2023.findings-acl.190.

[4] Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. doi: 10.48550/arXiv.2210.00379.

[5] Yunpeng Gao, Chenhui Li, Zhongrui You, Junli Liu, Zhen Li, Pengan Chen, Qizhi Chen, Zhonghan Tang, Liansheng Wang, Penghui Yang, Yiwen Tang, Yuhang Tang, Shuai Liang, Songyi Zhu, Ziqin Xiong, Yifei Su, Xinyi Ye, Jianan Li, Yan Ding, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Openfly: A comprehensive platform for aerial vision–language navigation, 2025. URL https://arxiv.org/abs/2502.18041.

[6] A. Handa, T. Whelan, J.B. McDonald, and A.J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014.

[7] Juhong Lee et al. Citynav: Language-goal aerial navigation dataset with geographic information. *arXiv preprint arXiv:2406.14240*, 2024.

[8] Liqiang Lin, Yilin Liu, Yue Hu, Xingguang Yan, Ke Xie, and Hui Huang. Capturing, reconstructing, and simulating: The urbanscene3d dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2107.04286. Camera-ready version.

[9] Zhenghao Liu et al. Aerialvln: Vision-and-language navigation for uavs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[10] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2020.05.009. URL http://www.sciencedirect.com/science/article/pii/S0924271620301295.

[11] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[12] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[13] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Dipam Patel, Phu Pham, and Aniket Bera. Dronerf: Real-time multi-agent drone pose optimization for computing neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5050–5055, 2023. doi: 10.1109/IROS55552.2023.10342420.

[15] Pix4D S.A. Pix4dmatic example datasets: Industrial, urban, and lidar projects. https://support.pix4d.com/hc/en-us/articles/360052422492, 2024. Accessed May 26, 2025. Example datasets provided for photogrammetry training and evaluation using PIX4Dmatic software.

[16] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, volume 9907 of *Lecture Notes in Computer Science*, pages 501–518. Springer, Cham, 2016. doi: 10.1007/978-3-319-46487-9_31. URL https://doi.org/10.1007/978-3-319-46487-9_31.

[17] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022. doi: 10.1109/cvpr52688.2022.00807.

[18] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12922–12931, 2022. doi: 10.1109/cvpr52688.2022.01258.

[19] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.10689.

[20] Yuan Wang et al. Towards realistic uav vision-language navigation: Platform, benchmark, and methodology. *arXiv preprint arXiv:2410.07087*, 2024.

[21] Zhengren Wang. 3d representation methods: A survey, 2024. URL https://arxiv.org/abs/2410.06475.

[22] Liqi Yan, Qifan Wang, Junhan Zhao, Qiang Guan, Zheng Tang, Jianhui Zhang, and Dongfang Liu. Radiance field learners as uav first-person viewers, 2024. URL https://arxiv.org/abs/2408.05533.

[23] Yuqi Zhang, Guanying Chen, Jiaxing Chen, and Shuguang Cui. Aerial lifting: Neural urban semantic and building instance lifting from aerial imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2403.11812.