MWAG: Multi-Season Wide-Area Air Ground Dataset for 3D Scene Reconstruction and Novel View Synthesis

Kshitij Singh Minhas [1], **Qiao Wang** [1], **Niluthpol Chowdhury Mithun** [1], **Ben Southall** [1], **Supun Samarasekera** [1], **Rakesh Kumar** [1]

[1]SRI International
firstname.lastname@sri.com

Abstract

3D scene reconstruction has seen significant advancements through methods like Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS), and is increasingly using generative AI models to improve the quality of reconstruction and novel view synthesis. However, existing datasets often lack the scale and diversity in complex, real-world outdoor environments. As a result, many of the state-of-the-art methods, including the generative models used in them, may not work well with real-world data. To further advance research in 3D reconstruction, neural rendering, and generative models on complex real-world data, we introduce the MWAG (Multi-season Wide-area Air Ground) dataset. MWAG features over 9,000 high-resolution RGB images captured from ground and aerial views across multiple seasons and differing environmental conditions. Each image includes highly accurate geodetic pose metadata, enabling tasks such as sparse-view 3D reconstruction, cross-view synthesis, and precise localization. The dataset supports multiple challenges in neural rendering, including handling environmental variations, training efficiency over expansive areas, and integrating multi-modal data for improved model completeness. We describe in detail our steps in data acquisition, processing, and alignment, to help the community create more diverse and challenging datasets to develop better methods and models in the future. Our dataset will be available at mwag.sri.com.

1 Introduction

Reconstructing accurate and detailed 3D representations of expansive outdoor scenes from a set of unposed images remains a challenging problem in computer vision and graphics research, with significant implications for applications such as autonomous navigation, urban planning, environmental monitoring, and virtual tourism. Recent neural rendering and 3D reconstruction techniques, such as Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023), have revolutionized 3D scene modeling by enabling high-quality novel view synthesis. However, the effectiveness of these methods heavily depends on the quality and density of input images, and upon the range of illumination and environmental (seasonal, weather) variation seen in the inputs.

Existing benchmark datasets for 3D reconstruction primarily focus on small-scale, localized, or object-centric scenes with densely captured images, often from ground-level cameras, and typically lack ground-truth geodetic poses. While suitable for controlled experiments, these datasets are insufficient for modeling in-the-wild environments characterized by complex geometries, sparse data, diverse viewpoints, and environmental variations. Addressing such challenges requires datasets that incorporate multi-altitude perspectives (e.g., aerial and ground views) and capture seasonal variations. Moreover, precise ground-truth geo-poses are essential not only for robust geo-localization and accurate evaluation of 3D reconstruction methods but also for enabling several real-world applications.

While NeRF and 3DGS perform well with dense inputs, it remains a major challenge to achieve high-quality 3D reconstruction and novel view synthesis with sparse-view inputs, and the research community is increasingly using generative AI models to improve the performance. One such successful approach uses a latent diffusion model (LDM) (Rombach et al. 2022) to create augmented views near input views and incorporate them in NeRF training (Wu et al. 2024). Others (Gao et al. 2024) directly synthesize novel views on the ground from aerial images (synthetic data) using LDM and ControlNet (Zhang, Rao, and Agrawala 2023). However, due to the limitation of datasets described earlier, these methods may not work well with real-world wide-area data which come with much higher diversity.

To bridge this gap, we introduce the MWAG (Multi-season Wide-area Air Ground) dataset. MWAG offers a diverse collection of aerial and ground images of outdoor areas, captured across seasons, with highly accurate ground-truth poses. It serves as a comprehensive testbed for benchmarking state-of-the-art 3D modeling techniques in challenging real-world scenarios. With accurate ground-truth poses, it is easy to create sparse-view test cases from subsets of the dataset. MWAG also supports critical tasks such as camera calibration under diverse conditions, a prerequisite for achieving precise 3D reconstructions.

2 Prior Work

There is a tight interplay between neural representation based 3D reconstruction and the challenges existing datasets provide. Challenging datasets are required to expose short-

Figure 1: Image grid showing few samples from MWAG, with top row showing ground images, and bottom row showing aerial images from similar locations. Images on the left show data from autumn and the right show data from winter.

comings of current modeling methods in order to spur innovation. When NeRF (Mildenhall et al. 2021) was introduced, work primarily focused on reconstructing indoor 3D objects by having dense images from all angles around a single object. Similar work (Barron et al. 2022) was also shown to work on 3D objects outdoors, but outdoor objects were small in size (bench, bicycle etc.). Block-NeRF (Tancik et al. 2022) first came out with a wide-area dataset and methodology to build NeRFs of a city-wide scale but had a limitation of using only ground images. The limitation of relying on ground-only views is that gathering such a dataset is not always feasible (or is very resource-consuming). Whereas capturing aerial data is relatively more efficient where just one fly-through with a drone can capture a vast amount of area. Reconstructions created from air and ground views must ingest multi-altitude data to create a unified model. This can further be extended to render ground views with air only images and can also aid work which enables 3D model building from sparse views (Truong et al. 2023) using diffusion (Zou, Zhang, and Liu 2024).

Several datasets are already available to test and evaluate NeRF-based methods, as mentioned in (Yan et al. 2023), but these datasets either provide ground only or air only data or the datasets are simulated. Here are a few key datasets that currently exist and are used in the field:

Mip-NeRF 360 dataset (Barron et al. 2022) encompasses a collection of 9 scenes, featuring a mix of outdoor and indoor environments. Each scene has a central object or area of complexity, with a detailed background. The captured images offer variable resolutions. To minimize color harmonization issues, outdoor scenes were captured under overcast skies and a large diffuse light was used for indoor scenes to avoid shadows. The dataset does not have ground truth pose. UrbanScene3D (Lin et al. 2022) dataset contains 128,000 high-resolution images and is designed for research in urban scene perception and reconstruction. Images are sourced from synthetic scenes generated by CAD and real urban scenes captured using drones only. While ground truth pose data is unavailable for the entire real scenes, ground-truth meshes are provided for some buildings within those scenes. OMMO (Lu et al. 2023) is a large outdoor multi-modal dataset. It features various scenes containing calibrated images. The images are sourced from YouTube videos and drone captures. Camera poses are provided using COLMAP, and the dataset includes prompt annotations for multi-model NeRF. The dataset does not have ground truth data.

Table 1: Comparison of our dataset with several widely used datasets for evaluating 3D reconstruction methods.

Dataset	Air Data	Ground Data	GT Pose	Wide Area	Seasonal Variation
Mip-NeRF360	x	✓	x	x	x
UrbanScene3D	x	✓	x	✓	x
OMMO	✓	✓	x	✓	✓
Mill 19	✓	x	x	✓	x
Tanks & Temples	x	✓	x	✓	x
Block-NeRF	x	✓	✓	✓	x
NeRF-on-the-go	x	✓	x	x	x
Photo Tourism	x	✓	x	✓	✓
Ours (MWAG)	✓	✓	✓	✓	✓

The Mill 19 dataset (Turki, Ramanan, and Satyanarayanan 2022) uses a drone to capture two big areas around CMU. The first area is a 500 m x 250 m industrial building. The other area is near a construction site with debris. The first area has 1940 pictures, and the second area has 1768 pictures with high resolution. PixSFM (Lindenberger et al. 2021) was used to improve camera poses, but there is no 3D ground truth available for this dataset as well. Tanks and Temples (Knapitsch et al. 2017) offers a variety of ground-only indoor and outdoor scenes with different complexities and sizes. Captured under real-world conditions using high-resolution video sequences and also includes ground truth camera poses.

NeRF On-the-go dataset (Ren et al. 2024) is a collection of 12 outdoor and indoor sequences. These casually recorded scenes focus on a very small region with varying degrees of transients. The Photo Tourism dataset (Snavely, Seitz, and Szeliski 2006) comprises several 3D scenes of famous landmarks. Each scene includes a diverse collection of user-uploaded images taken at various dates, times of day, and with different cameras and exposure settings.

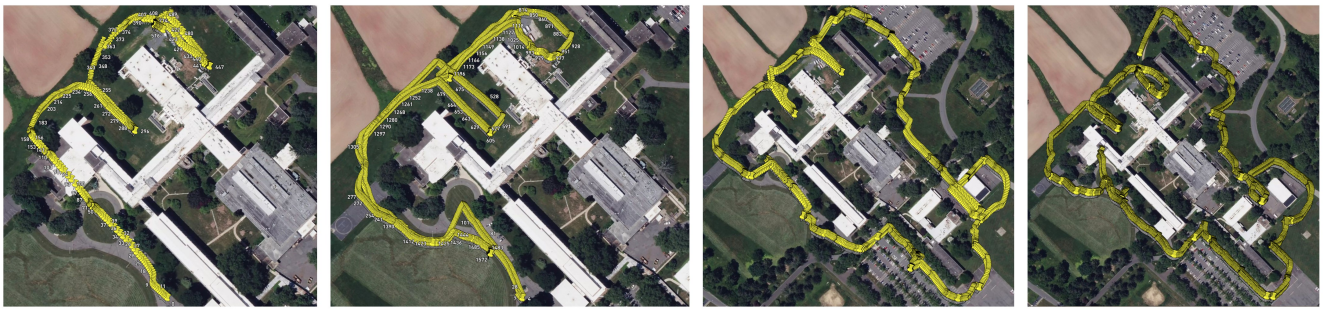

Figure 2: Left-most image marks the path of data collection done using the drone in autumn (MWAG-01). Second from the left shows the locations of the images captured from ground around the same time (MWAG-02). Notice we try to view the building from various sides so a detailed 3D model can be built. Right two images show MWAG-03 and MWAG-04 which are air and ground collects done in the winter. Notice these paths cover a much bigger area and the full campus of buildings.

3 Data Acquisition Framework

Our images are captured using a high-resolution Teledyne FLIR Blackfly S camera (rgb images, global shutter 20 MP imager). For this camera we used a wide field of view (180 degree FoV) lens to capture greater areas of the scene in each image. In addition to video, we collected synchronized (via hardware clock pulse) data from an Intel RealSense d435i camera which has a stereo camera pair (black and white images, global shutter camera) and a Bosch IMU, and captured highly accurate (centimeter precision) GPS RTK data. The sensors were integrated into a rigid package that could be carried by a person on a backpack for ground data collection, and flown on a drone for air collection. In addition to intrinsics calibration for the camera lenses, we performed extrinsic calibration between the IMU, FLIR camera, RealSense stereo cameras, and GPS unit.

The collected data were processed using CamSLAM simultaneous localization and mapping framework (SLAM). CamSLAM is composed of a powerful visual-inertial odometry backbone using an error-state Extended Kalman Filter for sensor fusion, and a very efficient and lightweight parallel mapping engine utilizing a keyframe-based pose graph data structure and binary descriptors for feature matching and indexing (Oskiper, Samarasekera, and Kumar 2017). We ingested 200 Hz raw IMU data, 15 Hz stereo camera pair data, and 10 Hz GPS RTK data into the CamSLAM software which produced highly accurate poses for the RealSense stereo pair. Poses for the FLIR images (collected at 1.5Hz, owing to the large image size and limited recording bandwidth) are then computed using a rigid body transformation with the calibrated extrinsic parameters. Key points to note here when collecting high speed very precise data is to have a hardware synchronization between devices (the two cameras, IMU and GPS in our case). Most data collection hardware rely on software synchronization of different data streams which lead to inaccurate poses.

To address the challenges of data integrity and adaptability for diverse 3D reconstruction tasks, we incorporated an advanced pipeline for quality assessment and metadata validation. This process leverages automated scripts for post-capture calibration, ensuring consistency between ground and aerial data sets. Furthermore, our acquisition framework integrates various sensor data (e.g., GPS and IMU cross-validation) to minimize errors in geodetic pose estimation. We also developed custom data annotation protocols to label artifacts such as transient objects, extreme lighting conditions and PII. This rigorous approach ensures the dataset is robust for ablation studies and model-based data improvement techniques, aligning directly with emerging needs in data-driven research highlighted in this workshop's themes.

4 The MWAG Dataset

The MWAG dataset consists of 9092 high-resolution images collected over 4 sequences. MWAG-01 and MWAG-02 contain 1086 aerial images and 1734 ground images respectively. These were collected around a building surrounded by trees and fields in autumn. These datasets were captured during late afternoon early evening and has some images which are under-exposed. The area covered by this data is 200m by 200m, where we cover a building from 3 sides. Figure 2 (left two images) shows a map of the ground and aerial collection and sample images are shown in Figure 1 (columns marked as Autumn). The focus of this dataset is to capture a building from all sides on the ground and air so that a 3D model of the building can be made using neural rendering methods.

MWAG-03 and MWAG-04 are similar dataset but covering a much larger area during the winter (snow on ground). These datasets have 2579 aerial and 3693 ground images respectively. These sets cover an area of 350 meters by 250 meters and cover the whole campus of office buildings from all 4 sides. These sets were collected right after snowfall and white cover can be seen on all the vegetation surrounding the building. This could promote future work on embeddings based seasonal change generation for neural renderings. These datasets were collected in bright and direct sunlight which leads to some images having over-saturation.

Every image has a corresponding json-formatted metadata file, including: highly accurate pose data (latitude, longitude, height above ground, 3 camera angles); timestamps; calibrated parameters for an OpenCV lens distortion model; a label denoting whether the image is collected from the ground or the air; and mask information for any transient

Figure 3: Image grid showing samples of real world effects such as shadows, over exposure, under exposure and transients in the scene

object in the image. All Personally Identifiable Information (PII) has been removed - namely faces and license plates are blurred in all images.

5 Applications and Experiments

In this section we will propose possible vectors along which sub-datasets can be created from our main dataset. We provide scripts to create many such sub-datasets from our main dataset. Test metrics would include comparison of reconstructed camera locations, and also measure image quality metrics for rendered outputs against a hold-out set of posed images from the main MWAG corpus.

The first example vector we propose tests reconstruction with images at varying altitudes. Models can be tested on sets with (i) ground only training images, (ii) air only training images, (iii) combinations of air and ground images. Current methods struggle to integrate multi-altitude data into a single cohesive model; the vector described would allow researchers to establish single-altitude reconstruction performance baselines for their methods in steps (i) and (ii), and then to address the challenge of combined data, and measure its benefits in step (iii). If the test sets for this vector include both air and ground render targets, then steps (i) and (ii) in particular will push GS and NeRF methods to address extrapolation across altitude changes. In Figure 4, we show example renderings of ground only modeled using Nerfacto (Tancik et al. 2023), Splatfacto (Kerbl et al. 2023) and Splatfacto-W (Xu, Kerr, and Kanazawa 2024).

A second vector on which models can be tested is their capability to handle real-world environmental changes. In this vector, same area images are rendered on (i) autumn only images, (ii) autumn and winter images, (iii) autumn only images with over and under sun-exposure, and (iv) autumn and winter images with over and under sun-exposure.

A third vector is to increase the area covered by training data, which may also include disjoint ground subsets that can only be joined using air data. We start off with an area of 50 m by 50 m area going up to 350 m by 250 m area. Models will be challenged here on their training efficiency and model completeness.

Any of these vectors can also be made more challenging by obscuring or spoiling pose data before reconstruction which be used to evaluate and enhance the performance of 3DGS methods which do not rely on pose priors (Smith et al. 2024). Additionally, an image feature based cross-view lo-

Figure 4: Render results of a ground only dataset when tested three baseline: Nerfacto, Splatfacto and Splatfacto-W.

calization model (Mithun et al. 2023) can also be trained on our data which can aid ground agents such as self-driving cars to be localized with a high degree of precision by improving on consumer grade GPS accuracy of 2m.

6 Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior Interior Business Center (DOI/IBC) contract number 140D0423C0075. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Gao, Z.; Teng, W.; Chen, G.; Wu, J.; Xu, N.; Qin, R.; Feng, A.; and Zhao, Y. 2024. Skyeyes: Ground Roaming using Aerial View Images. *arXiv preprint arXiv:2409.16685*.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Knapitsch, A.; Park, J.; Zhou, Q.-Y.; and Koltun, V. 2017. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Lin, L.; Feng, K.; Hu, M.; Jiang, S.; Dong, P.; and Fu, Q. 2022. Capturing, reconstructing, and simulating: the urbanscene3d dataset. In *European Conference on Computer Vision*. Cham: Springer Nature Switzerland.

Lindenberger, P.; Ludescher, F.; Martius, G.; Sturm, J.; and Cremers, D. 2021. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Lu, C.; Xie, Y.; Xu, C.; Xiong, Z.; and Wang, B. 2023. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mithun, N. C.; Minhas, K. S.; Chiu, H.-P.; Oskiper, T.; Sizintsev, M.; Samarasekera, S.; and Kumar, R. 2023. Cross-view visual geo-localization for outdoor augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 493–502. IEEE.

Oskiper, T.; Samarasekera, S.; and Kumar, R. 2017. CamSLAM: Vision Aided Inertial Tracking and Mapping Framework for Large Scale AR Applications. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE.

Ren, W.; Zhu, Z.; Sun, B.; Chen, J.; Pollefeys, M.; and Peng, S. 2024. NeRF On-the-go: Exploiting Uncertainty for Distractor-free NeRFs in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8931–8940.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Smith, C.; Charatan, D.; Tewari, A.; and Sitzmann, V. 2024. FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent. *arXiv preprint arXiv:2404.15259*.

Snavely, N.; Seitz, S. M.; and Szeliski, R. 2006. Photo tourism: exploring photo collections in 3D. In *ACM siggraph 2006 papers*, 835–846.

Tancik, M.; Ceylan, D.; Pradhan, S. K.; Singh, A.; Mildenhall, B.; and Srivastava, A. P. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; et al. 2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–12.

Truong, P.; Andrychowicz, M.; Koska, A.; Fidon, L.; Toth, B.; Dou, Q.; and Glocker, B. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. 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*.

Wu, R.; Mildenhall, B.; Henzler, P.; Park, K.; Gao, R.; Watson, D.; Srinivasan, P. P.; Verbin, D.; Barron, J. T.; Poole, B.; et al. 2024. ReconFusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21551–21561.

Xu, C.; Kerr, J.; and Kanazawa, A. 2024. Splatfacto-W: A Nerfstudio Implementation of Gaussian Splatting for Unconstrained Photo Collections. *arXiv preprint arXiv:2407.12306*.

Yan, Z.; Zhang, J.; Zhao, Z.; Li, H.; and Yao, Y. 2023. NeRFBK: a holistic dataset for benchmarking NeRF-based 3D reconstruction. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48(1): 219–226.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zou, Z.; Zhang, Z.; and Liu, H. 2024. Sparse3d: Distilling multiview-consistent diffusion for object reconstruction from sparse views. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.