AerialMetric: Benchmarking and Adapting UAV Monocular Metric Depth Estimation in the Real World

Zhongqiang Song[1], Guanying Chen[1⋆], Yuqi Zhang[2,3], Yin Zou[1], Chuanyu Fu[1], Zhiyuan Yuan[1], Chuan Huang[4,2], Shuguang Cui[3,2], and Xiaochun Cao[1]

[1]Sun Yat-sen University, Shenzhen Campus, China [2]FNii–Shenzhen, China
[3]SSE, CUHKSZ, China [4]SIAS, USTC, China

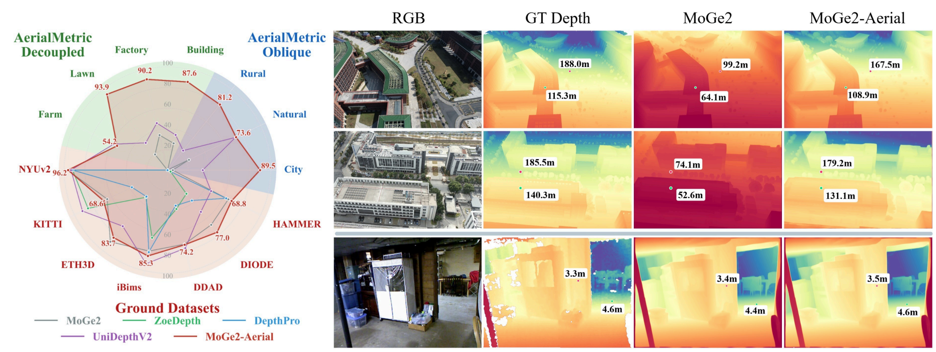


Fig. 1. By benchmarking existing methods on the AerialMetric dataset, we identify a significant domain gap in aerial depth estimation (the value reported in the left figure is δ_1, the higher the better). Fine-tuning the model with our dataset resolves this issue, delivering robust metric depth estimation on both aerial and ground-level scenes.

Abstract. This paper addresses the problem of monocular metric depth estimation in aerial UAV imagery. Although recent data-driven methods have achieved remarkable progress in ground-level scenarios, models trained primarily on street-view and indoor datasets exhibit significant domain gaps when applied to aerial viewpoints. To tackle these challenges, we introduce *AerialMetric*, a benchmark dataset designed to evaluate and facilitate the adaptation of monocular metric depth estimation under UAV aerial viewpoints. The dataset consists of four complementary subsets collected from different sources, jointly covering real-world photogrammetry data, controlled aerial acquisition settings, photorealistic synthetic scenes, and in-the-wild Internet imagery. In total, *AerialMetric* provides 52K real-world and 16K synthetic image–depth pairs with reliable metric ground truth. Based on this dataset, we conduct systematic evaluations of existing state-of-the-art models under aerial settings and

⋆ Corresponding author.

investigate the impact of viewpoint, altitude, and camera parameters on metric depth prediction. In addition, by fine-tuning representative metric depth model on our dataset, we establish a comprehensive aerial benchmark and achieve state-of-the-art performance across diverse aerial imagery. Our dataset, code, and model weight are publicly available at `https://kuieless.github.io/AerialMetric-ECCV2026-page/`.

Keywords: Monocular Depth · Metric Depth · UAV Datasets

1 Introduction

The rapid advancement of aerial robotics technologies has accelerated the large-scale deployment of autonomous Unmanned Aerial Vehicles (UAVs) across diverse applications, *e.g.*, aerial delivery services, infrastructure inspection, environmental monitoring, and public safety. These emerging applications demand reliable 3D perception of the surrounding environment [30, 41]. Among various perception tasks, monocular metric depth estimation plays a central role by aiming to recover scene geometry with absolute metric scale from a single image.

In recent years, data-driven approaches for depth estimation have achieved remarkable progress, largely driven by the availability of large-scale annotated datasets. Deep neural networks trained on diverse depth data have demonstrated strong performance in metric depth prediction settings [26, 37, 38, 63, 64, 74, 75, 77]. However, despite their strong performance, state-of-the-art metric depth estimation models suffer from a significant domain gap between ground-level data and aerial UAV imagery, resulting in limited generalization to aerial viewpoints, where significantly different viewing geometries and bird's-eye perspectives dominate.

Currently, there is a lack of real-world, diverse, and variable-controlled aerial datasets with metric depth, which limits effective model adaptation and prevents reliable learning of absolute scene scale. In addition, the absence of an evaluation benchmark makes it difficult to systematically assess existing methods under diverse UAV imaging conditions. Despite rapid advances in depth estimation architectures, aerial metric depth estimation remains substantially behind its ground-level counterpart (see Fig. 1).

To bridge this gap, we introduce **AerialMetric**, a large-scale benchmark for monocular metric depth estimation in aerial scenarios. Our dataset consists of four complementary components, as shown in Fig. 2. *AerialMetric-Oblique* is constructed from publicly available real-world oblique aerial imagery, from which we process and curate high-quality image-depth pairs with reliable metric ground truth, comprising 47K samples for training and evaluation. *AerialMetric-Decoupled* is collected by ourselves through systematic photogrammetry acquisition, where each scene is captured under controlled variations of viewing angles, fields of view, and flight altitudes (four angles, two FOVs, and two heights). This subset enables a disentangled analysis of how aerial imaging configurations affect metric depth estimation. *AerialMetric-Synthetic* contains 16K photorealistic image-depth pairs for training rendered using Unreal Engine and Google Earth, providing controlled, multi-angle views of scenarios that are difficult to

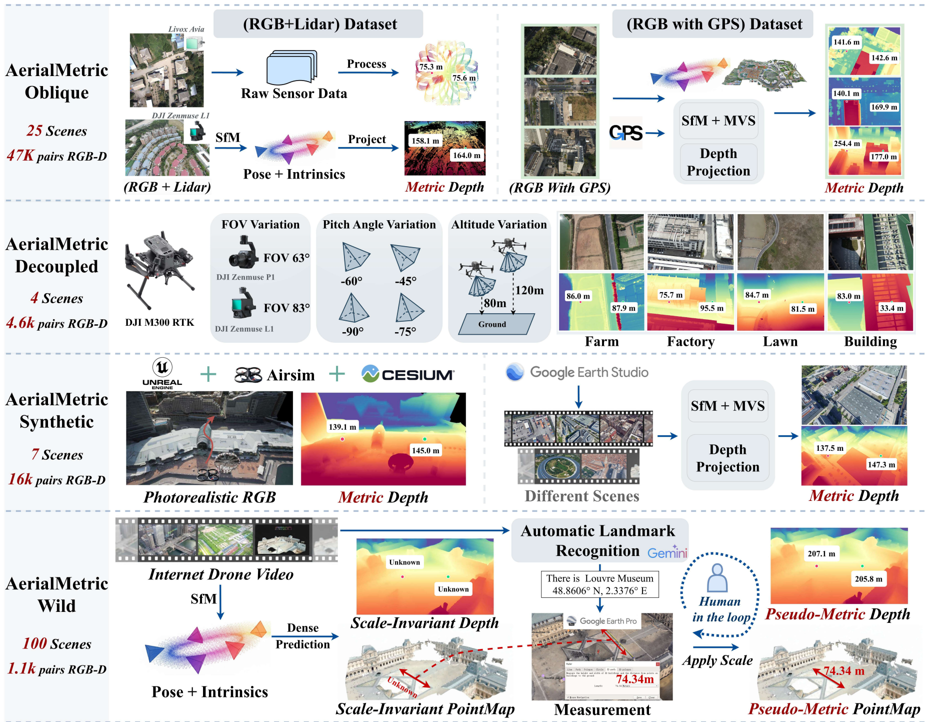


Fig. 2. Data collection pipeline of the proposed AerialMetric dataset.

capture in real-world scenes. *AerialMetric-Wild* consists of diverse in-the-wild aerial images collected from the Internet, serving as a challenging evaluation set for cross-domain generalization.

Based on the proposed *AerialMetric* dataset, we conduct a comprehensive analysis of existing metric depth estimation methods under aerial settings. Our study reveals that current state-of-the-art models exhibit limited robustness and degraded performance when applied to aerial imagery. In addition, leveraging the high-quality metric depth supervision provided by AerialMetric-Oblique and AerialMetric-Synthetic, we fine-tune a representative metric depth estimation model, which achieves enhanced performance under this challenging aerial setting, demonstrating that our dataset effectively improves the generalization capability of monocular metric depth estimation methods.

In summary, the key contributions of this paper are as follows:

- We introduce AerialMetric, an aerial monocular metric depth dataset comprising four complementary subsets, significantly expanding the scale, diversity, and geometric coverage of aerial depth data.
- We establish a comprehensive benchmark for monocular aerial metric depth estimation, systematically analyzing the impacts of viewing angle, flight altitude, and camera field-of-view.

– By adapting representative state-of-the-art models on AerialMetric dataset, we achieve state-of-the-art performance across diverse UAV deployment scenarios, demonstrating that our dataset effectively improves the generalization capability of monocular metric depth estimation methods.

2 Related Work

Monocular Depth Estimation Monocular depth estimation has advanced significantly with deep learning and large-scale visual models. Early deep learning-based methods [14] focused on multi-scale feature learning. DPT [50] leverages Vision Transformers to advance monocular depth estimation. The MiDaS series [4, 51] enables robust zero-shot relative depth estimation across scenes using large mixed datasets. The Depth Anything series [38, 74, 75] enhances zero-shot generalization with massive unlabeled data and cross-view consistency. The Lotus series [23, 24] leverages diffusion priors to achieve strong generalization and high-precision geometry estimation.

Metric depth is essential for grounded 3D tasks. To unify relative and metric predictions, ZoeDepth [3] and Metric3Dv2 [26, 77] achieve zero-shot metric depth estimation. MoGe2 [63, 64] and MapAnything [31] improve open-domain metric reconstruction. InfiniDepth [78] uses neural implicit fields for arbitrary-resolution depth prediction. Despite the remarkable success in terrestrial scenes, existing methods still struggle when deployed in UAV scenes.

Compared to depth estimation from ground perspectives, UAV-based monocular depth estimation [17, 43, 56, 79] remains challenging, due to the scarcity of dedicated, large-scale, variable-controlled, real-world aerial depth datasets.

General Datasets with Depth The development of monocular depth estimation is deeply rooted in the availability of datasets. NYUv2 [44] and ScanNet [9] provide abundant structured data captured by RGB-D sensors, laying the foundation for diverse 3D tasks. KITTI [20] and Waymo [57] are built for autonomous driving scenarios, equipped with LiDAR data. nuScenes [6] further supplements millimeter-wave radar to achieve full-view coverage, improving outdoor generalization. Argoverse [7,67] provides depth supervision through LiDAR point clouds and stereo image pairs. DAP [40] extends the depth estimation task to panoramic views, enabling zero-shot generalization in panoramic scenes. Although these general datasets have advanced monocular depth estimation, they are almost limited to ground-level perspectives.

UAV Datasets UAV data presents challenges for visual perception due to its unique viewpoints and wide field-of-view. Existing UAV datasets like VisDrone [82] and UAVid [42] focus on detection [12,28], tracking [36,69], and segmentation [62,76], but lack 3D geometric ground truth. UAVLight [13] focuses on illumination variations, and OpenFly [19] focuses on vision-and-language navigation. Neither of them directly supports high-precision 3D tasks.

To support a broader range of 3D perception tasks [59, 79, 80], UAV datasets have expanded into two complementary directions: geometry-centric benchmarks for semantic 3D understanding and depth perception, such as STPLS3D [8], DDOS

Table 1. Comparison between existing datasets and our proposed AerialMetric dataset. Our dataset comprises four complementary subsets designed to cover diverse scenarios, including real-world photogrammetry, controlled aerial acquisitions with decoupled flight parameters, photorealistic synthetic scenes, and in-the-wild imagery.

Dataset	Type	Raw Data	Depth GT	3D Assets	Area (m^2)	Scenes	#Images	Flight Parameters (Decoupled)		
								Pitch	Altitude	FOV
University-1652 [81]	Synthetic	RGB	None	-	-	1652	5.0×10^4	-	-	-
UAVLight [13]	Real	RGB	None	Point Cloud+Mesh	$>6.1 \times 10^5$	18	4385	Oblique	-	-
Mill 19 (Mega-NeRF [59])	Real	RGB	None	-	$>1.3 \times 10^5$	2	$>3.6 \times 10^3$	-	-	-
VisDrone [82]	Real	RGB	None	-	-	14	$>2.7 \times 10^5$	-	-	-
Mid-Air [18]	Synthetic	RGB-D	Metric	-	-	2	$>4.2 \times 10^5$	-	-	-
DublinCity [84]	Real	RGB+LiDAR	Sparse Lidar	Point Cloud	$>2.0 \times 10^6$	1	8504	Oblique	-	-
Hessigheim [34]	Real	RGB+LiDAR	Sparse Lidar	Point Cloud+Mesh	$>1.9 \times 10^5$	1	-	Oblique	-	-
MegaDepth [37]	Real	RGB-D	Relative	Point Cloud	-	196	$>1.3 \times 10^5$	-	-	-
MARS-LVIG [35]	Real	RGB+LiDAR	Sparse LiDAR	Point Cloud	-	4	-	-	-	-
UAVScenes [65]	Real	RGB+LiDAR	Sparse LiDAR	Point Cloud+Mesh	-	21	1.2×10^5	-	-	-
OpenDroneMap [47]	Real	RGB	None	Point Cloud+Mesh	-	41	15061	Oblique	-	-
Esri Sample Drone Datasets [15]	Real	RGB+LiDAR	Sparse LiDAR	Point Cloud+Mesh	-	8	-	Oblique	-	-
UrbanBIS [73]	Real	RGB	None	Point Cloud+Mesh	1.1×10^7	6	1.1×10^5	Oblique	-	-
UrbanScene3D [39]	Real+Synthetic	RGB+LiDAR	Sparse Lidar	Point Cloud+Mesh	1.36×10^8	16	$>1.3 \times 10^5$	-	-	-
GauU-Scene V2 [71]	Real	RGB+LiDAR	Sparse LiDAR	Point Cloud	6.7×10^6	6	4693	-	-	-
OccuFly [21]	Real	RGB-D	Metric	Point Cloud	$>1.9 \times 10^5$	9	$>2.0 \times 10^4$	-	3 Levels	-
AerialMetric-Oblique	Real	RGB/RGB+LiDAR	Metric	Point Cloud+Mesh	6.7×10^6	25	$>4.7 \times 10^4$	Oblique	-	-
AerialMetric-Decoupled	Real	RGB	Metric	Point Cloud+Mesh	2.9×10^5	4	>4600	4 Angles	2 Levels	2 Settings
AerialMetric-Synthetic	Synthetic	RGB-D	Metric	Mesh Partly	2.4×10^7	7	$>1.6 \times 10^4$	4 Angles	3 Levels	3 Settings
AerialMetric-Wild	Real	RGB	Pseudo-Metric	Point Cloud	-	100	>1100	-	-	-

[33], and FIReStereo [10]; and localization/navigation benchmarks, including NTU VIRAL [46], CrossLoc [72], GraCo [83], MUN-FRL [58], and UAVD4L [68].

OpenDroneMap [47] provides an open-source photogrammetry platform that facilitates large-scale aerial surveying and mapping. Mid-Air [18] and SynDrone [52] have constructed UAV datasets via synthetic data generation to be applicable to depth estimation tasks. ClaraVid [2] provides a synthetic aerial benchmark with dense depth and semantic annotations.

Recent UAV benchmarks have evolved from synthetic data toward real-world scenes. UseGeo [45], WildUAV [16], and UAVid-3D-Scenes [17] provide real UAV imagery with LiDAR- or photogrammetry-derived depth supervision. DublinCity [84], UrbanBIS [73], H3D [34], and GauU-Scene [70,71] provide urban UAV datasets with point clouds or LiDAR ground truth and semantic annotations. MARS-LVIG [35] and UAVScenes [65] offer multi-sensor and semantically annotated data. UrbanScene3D [39] and FlyAwareV2 [1] build real-synthetic hybrid benchmarks for outdoor 3D modeling. AerialMegaDepth [61] addresses aerial-ground viewpoint differences, and OccuFly [21] supports semantic scene completion with metric depth and occupancy annotations.

However, existing UAV benchmarks are constrained by inconsistent data modalities, synthetic-to-real domain gaps, and limited flight scenarios, hindering unified 3D perception. We introduce *AerialMetric*, a large-scale aerial depth benchmark designed to overcome these limitations and address scale ambiguity for UAV 3D sensing (see Table 1).

3 The AerialMetric Dataset

Overview Existing metric depth datasets remain largely constrained to horizon-level views and therefore fail to cover the drastic altitude and pitch variations, and wide depth ranges inherent in aerial imagery. To address this distribution

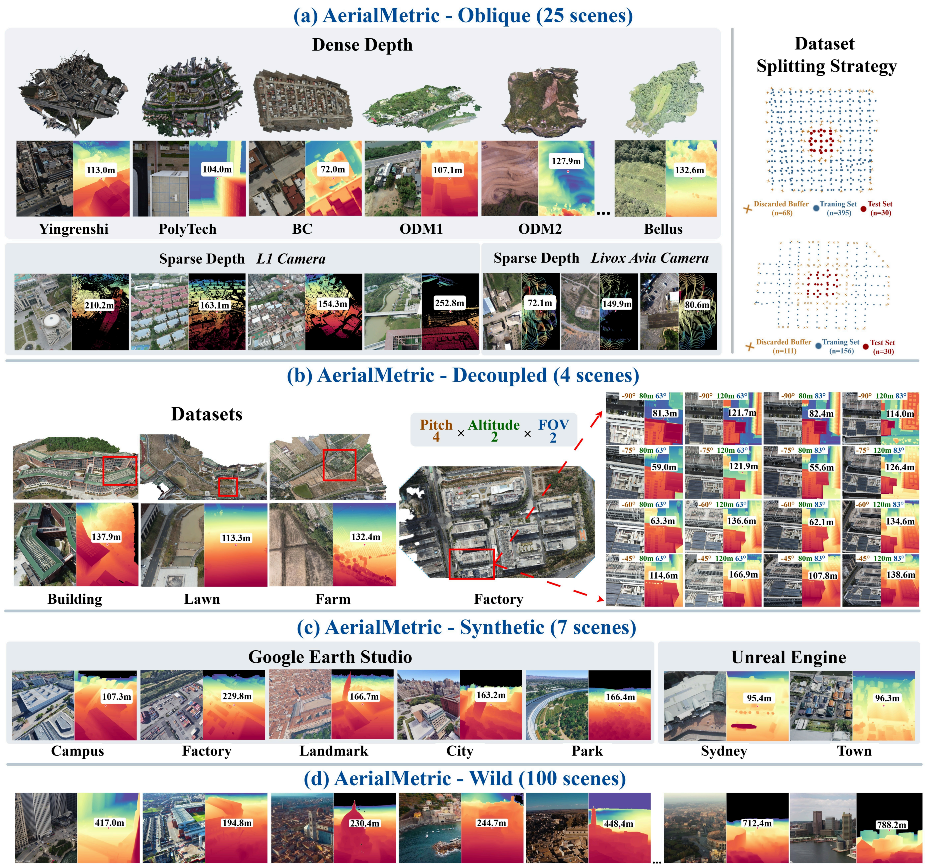


Fig. 3. Visualization of the introduced AerialMetric dataset.

shift, we introduce *AerialMetric*, a large-scale, high-quality aerial depth dataset built through a hybrid construction pipeline that achieves geometric accuracy, parameter decoupling, and scene generalization.

The dataset comprises four complementary components (see Fig. 3): (1) *AerialMetric-Oblique*, an oblique photography dataset curated from publicly available photogrammetry sources, featuring high-quality GPS-positioned imagery (a subset of which includes associated LiDAR data); (2) *AerialMetric-Decoupled*, an RTK-equipped UAV benchmark performing rigorous variable-decoupled sampling across altitudes, pitch angles, and fields of view to enable quantitative evaluation of capture variations; (3) *AerialMetric-Synthetic*, synthetic data generated by utilizing Google Earth Studio and simulation engines (UE + AirSim) to cover long-tail scenarios involving extreme viewpoints and complex lighting conditions, while enabling decoupled capture variations; and (4) *AerialMetric-Wild*, which consists of annotated internet drone videos with metric depth maps for in-the-wild evaluation through a human-in-the-loop scale calibration pipeline that leverages known landmark dimensions to resolve monocular scale ambiguity.

3.1 AerialMetric-Oblique: UAV Photogrammetry Datasets

Data Sources and Aggregation To construct a high-fidelity aerial depth benchmark, AerialMetric-Oblique aggregates and standardizes six prominent urban-scale photogrammetry datasets: *UrbanBIS* [73], *GauU-SceneV2* [71], *UAV-Scenes* [65], *UrbanScene3D* [39], *ODM* [47], and *ESRI* [15]. We construct over 47K image-depth pairs from 25 sub-datasets, spanning diverse scene categories: *City*, *Rural*, and *Natural*. To mitigate high inter-frame redundancy in the *UAVScenes* subset, we apply a 5-frame downsampling interval following the authors' practice.

To accommodate heterogeneous sensor modalities, we design a processing pipeline that categorizes depth recovery into active LiDAR and passive photogrammetry paradigms.

Ground-truth Depth from LiDAR For the *UAVScenes* and *GauU-SceneV2* datasets, we process the raw point clouds captured by LiDAR into per-frame metric depth maps. *UAVScenes* provides single-frame point clouds, enabling direct projection onto the image plane. Since *GauU-SceneV2* does not provide per-frame LiDAR data parsing functionality, we use the professional commercial software DJI Terra [11] to generate a unified point cloud and project it onto individual image views. A critical challenge in LiDAR projection is ray penetration artifacts, where rays from background points erroneously pass through sparse foregrounds and manifest as noise in the resulting depth maps. To address this, we reconstruct a continuous mesh from the data and validate projected depths against this mesh surface. Points exhibiting a discrepancy greater than 1.0 m relative to the mesh are discarded as occlusion noise.

Ground-truth Depth from Photogrammetry For *UrbanBIS*, *Urban-Scene3D*, *ODM*, and *ESRI*, we utilize the RGB images and reconstruct 3D meshes via DJI Terra at the highest native resolution to preserve high-frequency details that would otherwise be lost during downsampling. Metric depth maps are then derived via pose-guided mesh rasterization, ensuring pixel-level correspondence between RGB images and depth values.

Dataset Splits. To enable rigorous zero-shot evaluation, we implement a strict train-test split that prevents geographic data leakage. As illustrated in Fig. 3 (a), we exclude the outermost cameras and prune over 4.4K frames within a 50 m buffer zone of the test regions. The final test set comprises 1.4K images curated from over 45 distinct geographical regions across the 25 sub-datasets. The remaining data is allocated for training. Detailed statistics and split protocols are provided in the supplementary material.

3.2 AerialMetric-Decoupled: Variable Decoupled Dataset

To systematically evaluate the impact of imaging geometry on monocular depth estimation, we construct AerialMetric-Decoupled, a strictly controlled, variable-decoupled benchmark. Unlike conventional datasets that entangle multiple confounding factors,

AerialMetric-Decoupled disentangles the effects of camera pose (pitch angle, flight altitude) and field of view (FOV) via orthogonal sampling. As detailed in Table 2, we utilize an RTK-equipped UAV platform to sample across four distinct scenes. This dataset is designed exclusively for testing, providing a controlled evaluation protocol free from training bias.

Acquisition Platform Data acquisition is conducted using a DJI Matrice 300 RTK platform equipped with dual payloads: (1) Zenmuse L1 (LiDAR/RGB) and (2) Zenmuse P1 (Full-frame RGB). The RTK system provides centimeter-level absolute spatial alignment, ensuring precise control over camera positioning and enabling accurate metric depth ground truth. The captured images feature resolutions of 8192×5460 (P1 camera) and 5472×3648 (L1 camera).

Variable-Decoupled Acquisition Protocol

We strictly decouple four core variables through orthogonal sampling: (1) *Pitch Angle*: varying from nadir to oblique perspectives ($\theta \in \{-90°, -75°, -60°, -45°\}$); (2) *Relative Altitude*: captured at 80m and 120m above ground level to assess scale sensitivity; (3) *FOV*: leveraging the distinct sensor characteristics (L1 at 83° HFOV, P1 at 63° HFOV); and (4) *Scene Types*: encompassing four distinct semantic environments: *Building*, *Lawn*, *Farmland*, and *Industrial Factory*. For each scene, we captured 16 flight trajectories with strictly controlled combinations of FOV, pitch, and altitude variables, yielding approximately 4,600 image-depth pairs.

Table 2. UAV acquisition parameters for the AerialMetric-Decoupled dataset.

Parameter	Specification			
Positioning System	**Real-Time Kinematic (RTK)** (Centimeter-level accuracy)			
UAV Platform	DJI Matrice 300 RTK			
Sensor & FOV	L1: 83° (LiDAR Multi-modal Payload) P1: 63° (Full-frame RGB Payload)			
Relative Altitude	80 m, 120 m (Dual-scale assessment)			
Gimbal Pitch	$-90°$ (Nadir), $-75°$, $-60°$, $-45°$ (Oblique)			
Image Resolution	L1: 5472×3648; P1: 8192×5460			
Flight Speed	10 m/s			
Trigger Interval	Spatial equidistant, 10 m/frame			
Exposure Strategy	Automatic			
Scene Category	*Building*	*Lawn*	*Farm*	*Factory*
Area ($\times 10^4$ m^2)	2.6	7.2	9.0	10.0

Ground Truth Curation To mitigate geometric degradation at reconstruction boundaries, we extend flight trajectories with high-overlap oblique captures surrounding the core test zones. We employ DJI Terra for 3D reconstruction, processing the captured images to generate high-fidelity geometric models. Depth maps are derived only from the central regions through projection onto the reconstructed geometry, discarding boundary artifacts. Additionally, dynamic entities are manually annotated and explicitly excluded from evaluation to ensure static scene consistency.

3.3 AerialMetric-Synthetic: Variable Controlled Rendering

To overcome the physical constraints of real-world UAV collection, we construct a two-component synthetic dataset for training data augmentation.

Google Earth Studio Dataset We render multi-view sequences in Google Earth Studio (GES) across five diverse scenarios (City, Campus, Factory, Landmark, and Park) for training. We employ a rigorously decoupled parameter grid: three altitude bins in (70-300m), four pitch angles from -90° to -45°, and three FOVs, producing 2,000 RGB images at 3840×2160 resolution. Since GES lacks native depth export, we reconstruct 3D meshes using DJI Terra and project metric depth via pose-guided rasterization.

Unreal Engine Dataset To efficiently scale training data, we integrate Unreal Engine, AirSim, and Cesium to stream Google Earth 3D Tiles. This simulator directly extracts native Z-buffers, by passing photogrammetric reconstruction. The framework generates extreme trajectories (e.g., low-to-medium-altitude UAV flights, acute side-views) with randomized altitudes (80–200 m), pitches, and FOVs. The pipeline produces approximately 16K RGB-D pairs at 3840×2160 resolution, broadening spatial distribution and improving generalization to out-of-distribution wild viewpoints.

3.4 AerialMetric-Wild: Reliable Metric Pseudo-Ground Truth

Existing metric depth datasets lack open-domain diversity due to their reliance on specialized sensors and controlled acquisition environments. To address this limitation, we construct AerialMetric-Wild by exploiting unlabeled Internet drone videos through a human-in-the-loop *Scale Recovery Framework*. This dataset is designed exclusively for in-the-wild evaluation, featuring authentic drone trajectories, diverse FOVs, and unconstrained viewing conditions.

Multi-View Relative Depth Estimation Instead of relying on single-image estimation, we explicitly leverage the temporal multi-view nature inherent to drone tour sequences. We employ COLMAP [54] for initial camera pose estimation and Depth Anything 3 (DA3) [38] for dense depth prediction. This sequence-level conditioning yields a shape-reliable but unscaled relative point cloud $\mathcal{P}_{rel}$, capturing geometric structure without metric constraints.

Human-in-the-loop Metric Scale Calibration Monocular reconstruction recovers geometry only up to an unknown scale. To obtain metric depth, we calibrate each sequence using a small set of human-annotated landmark pairs. Annotators select semantically clear point pairs in an interactive interface, forming a constraint set $\mathcal{K}$. For each pair, we measure its distance in the unscaled point cloud $\mathcal{P}_{rel}$ and assign a real-world distance from geographic tools (*i.e.*, Google Earth Pro). We then estimate a global scale factor by averaging the ratio of real-world distance to reconstructed distance across all selected pairs, i.e., $s = \mathrm{avg}_k\left(D_{gt}^{(k)}/d_{rel}^{(k)}\right)$. Applying this single factor to all points maps $\mathcal{P}_{rel}$ into metric space and yields $\mathcal{P}_{metric}$.

Since manual selection, reference measurements, and depth inference are all imperfect, $\mathcal{P}_{metric}$ should be treated as pseudo-metric supervision. Despite these limitations, the dataset remains valuable for evaluating method performance in in-the-wild scenarios.

Dataset Composition We curated 100 high-fidelity sequences, each with $\sim$11 frames, from over 600 raw UAV flight clips through rigorous filtering, excluding content with ambiguous coordinates, extreme scales, or unreliable

DA3 reconstructions. The final dataset demonstrates substantial geographical diversity, spanning more than 80 landmarks across 50 cities in approximately 30 countries. Detailed statistics and the complete processing pipeline are provided in the supplementary material.

4 Experiments

4.1 Evaluation Protocol and Implementation Details

Evaluation Protocol. We benchmark representative metric depth estimators on our aerial datasets, including *AerialMetric-Oblique*, *AerialMetric-Decoupled*, and *AerialMetric-Wild*. Specifically, we evaluate the state-of-the-art methods, including ZoeDepth-NK [3], DepthPro [5], UniDepthV1-L [49], UniDepthV2-L [48], MoGe2-L [64], and Metric3Dv2-L [26]. For evaluation, predicted depth is clipped to 0.001–400 m, and invalid pixels are excluded. We report AbsRel (the lower the better) and threshold accuracy scores δ_1, δ_2 (the higher the better). All metrics are computed per image and averaged over each dataset. We use the publicly available code of the existing methods for testing.

Model Adaptation Details. To validate adaptation on existing methods, we use MoGe2 [64] as base model, and we adopt LoRA [25] for parameter-efficient fine-tuning. The fine-tuned model is denoted as MoGe2-Aerial. For training data, we use the *AerialMetric-Oblique* training split together with *AerialMetric-Synthetic*. We additionally include ground-domain data (*i.e.*, Hypersim [53], the MVS-Synth dataset introduced in DeepMVS [27], and TartanAir [66]) to better preserve generalization on ground-level scenes.

MoGe2-Aerial is optimized using AdamW, with input images resized to 1K resolution and cropped to dimensions divisible by 14. It is trained with a batch size of 32 for 1,800 steps. The learning rates for the encoder and decoder are set to 1×10^{-6} and 5×10^{-5}, respectively, with polynomial decay. Discussion of fine-tuning strategy and data composition are provided in the ablation study.

4.2 Benchmarking Metric Depth Estimation Methods

Evaluation on AerialMetric-Oblique and AerialMetric-Decoupled. Table 3 presents the metric depth prediction results with and without ground-truth camera intrinsics. Across both protocols, zero-shot transfer from ground-domain pretraining to aerial imagery is weak, with several baselines showing very high AbsRel and near-zero δ_1, indicating a severe domain gap. In zero-shot evaluation, UniDepthV2 [48] is the strongest baseline, but still shows clear gaps and instability across scenes.

Adapting existing method with our *AerialMetric* consistently improves performance. Table 3 shows that MoGe2-Aerial significantly outperforms zero-shot MoGe2 [64] on all aerial datasets. For example, on *AerialMetric-Oblique-City* without ground-truth intrinsic, δ_1 increases from 5.1% to 89.3%.

Table 3. Quantitative evaluation of metric depth estimation on aerial datasets.

Method	AerialMetric-Oblique						AerialMetric-Decoupled							
	City		Natural		Rural		Building		Factory		Farm		Lawn	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
Metric depth map (w/o GT intrinsics)														
ZoeDepth [3]	97.1	0.0	93.0	0.0	92.8	0.0	95.4	0.0	96.3	0.0	96.6	0.0	95.2	0.0
DepthPro [5]	97.8	0.0	92.2	0.0	96.9	0.0	97.8	0.0	97.6	0.0	97.3	0.0	96.7	0.0
UniDepthV1 [49]	83.9	0.0	80.3	0.0	86.0	0.0	87.4	0.0	81.1	0.0	89.5	0.0	84.6	0.0
UniDepthV2 [48]	31.0	34.1	19.2	73.0	28.1	24.2	24.1	35.0	22.0	46.7	**17.2**	**57.9**	24.4	33.8
MoGe2 [64]	48.4	5.1	45.9	23.0	56.7	0.3	29.0	26.6	26.8	34.8	71.0	0.0	32.1	19.5
MoGe2-Aerial	**10.3**	**89.3**	**16.8**	**73.6**	**14.2**	**81.2**	**11.9**	**87.6**	**11.4**	**90.2**	22.8	54.2	**9.8**	**94.0**
Metric depth map (w/ GT intrinsics)														
UniDepthV1 [49]	85.1	0.0	82.3	0.0	87.2	0.0	88.4	0.0	82.4	0.0	90.3	0.0	85.6	0.0
UniDepthV2 [48]	29.6	34.4	17.9	76.8	18.4	57.6	21.0	43.5	18.2	61.0	**10.9**	**91.0**	18.4	57.4
DepthAnything 3 [38]	96.6	0.0	84.9	0.0	91.3	0.0	87.0	0.0	93.3	0.0	93.1	0.0	93.9	0.0
Metric3DV2 [26]	75.4	0.0	68.5	0.1	73.2	0.1	68.5	0.0	63.9	0.0	81.1	0.0	73.3	0.0
MoGe2 [64]	36.2	18.1	43.4	25.1	42.3	8.0	19.1	59.7	12.3	83.4	61.9	0.8	18.5	61.0
MoGe2-Aerial	**8.8**	**93.8**	**14.3**	**78.5**	**11.8**	**85.3**	**10.9**	**90.3**	**13.9**	**85.7**	17.8	71.2	**11.5**	**90.7**

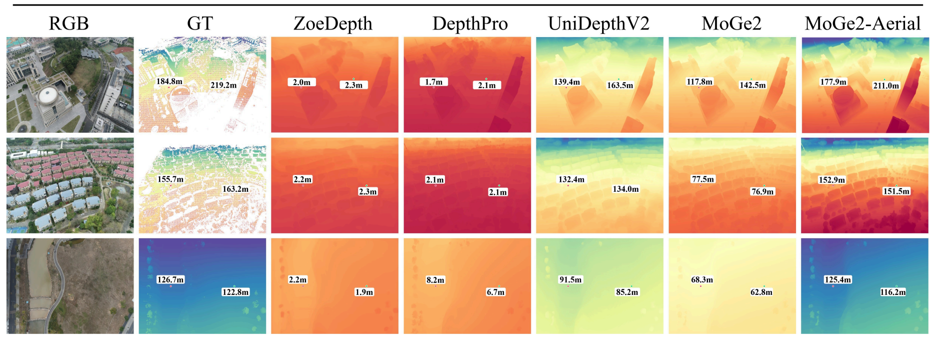


Fig. 4. Qualitative comparison with representative methods on aerial scenes.

The qualitative results in Fig. 4 support these trends. Compared with ZoeDepth [3], DepthPro [5], UniDepthV2 [48], and MoGe2 [64] our adapted model produces cleaner boundaries, fewer local depth inversions, and more coherent global scale, especially in distant roads, and vegetation regions.

Hard scene analysis. The *Farm* subset in *AerialMetric-Decoupled* remains the most challenging setting. MoGe2-Aerial exhibits lower accuracy on this subset compared to others (*e.g.*, $\delta_1 = 54.2\%$ in Table 3). This performance bottleneck is likely attributable to the prevalence of repetitive vegetation textures and a lack of strong geometric anchors. Interestingly, UniDepthV2 [48] achieves the highest performance on the *Farm* category. We hypothesize that this occurs because UniDepthV2's pre-training corpus encompasses a data distribution resembling agricultural environments.

Robustness Under Decoupled Camera Factors We systematically evaluate the robustness of our method against variations in camera pitch, altitude, and FOV. As illustrated by the radar summaries in Fig. 5, MoGe2-Aerial exhibits an expansive and highly isotropic performance footprint across diverse scene types (*Building*, *Factory*, *Farm*, and *Lawn*). This demonstrates remarkable resilience to

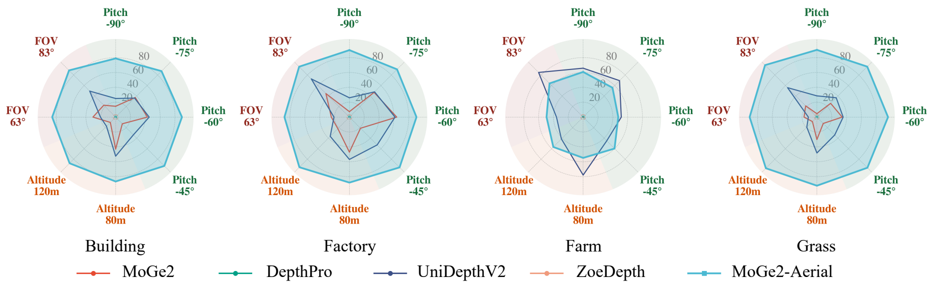
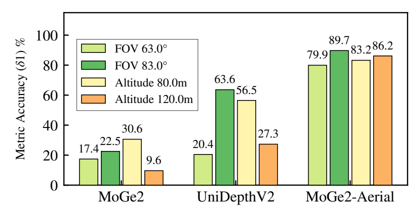


Fig. 5. Radar summary of parameter robustness on AerialMetric-Decoupled. Larger radius indicates better performance under each decoupled setting.

Fig. 6. Robustness analysis of depth estimation models (MoGe2, UniDepthV2, and MoGe2-Aerial) across different Field of Views (FOV) and altitudes. Results represent the average δ_1 on the AerialMetric-Decoupled dataset.



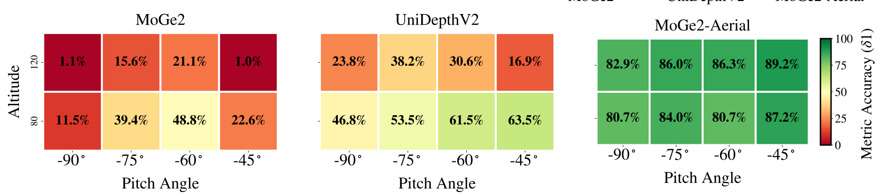


Fig. 7. Heatmap visualization of metric accuracy across altitudes and pitch angles.

aerial-specific camera factors, whereas prior foundation models yield irregular and contracted footprints, indicating severe sensitivity to these parameters.

Figure 6 further details the impact of FOV and altitude shifts. Zero-shot foundation models (e.g., MoGe2 [64], UniDepthV2 [48]) experience a precipitous drop in metric accuracy when the altitude increases from 80m to 120m. Although a wider FOV (83° compared to 63°) partially mitigates this degradation by capturing broader spatial context, the overall baseline performance remains inadequate. In contrast, MoGe2-Aerial effectively resolves the scale ambiguity prevalent in high-altitude captures, maintaining consistently high accuracy (over 80%) across all FOV and altitude configurations.

Furthermore, the heatmaps in Fig. 7 highlight the joint influence of altitude and camera pitch. Baseline models exhibit severe, non-monotonic degradation regarding pitch, suffering catastrophic failures at both strictly nadir ($-90°$) and oblique views up to ($-45°$) viewing angles. This vulnerability stems directly from the domain gap between ground-centric training data and aerial perspectives. Conversely, our fine-tuned MoGe2-Aerial demonstrates exceptional stability, retaining uniformly high accuracy across all evaluated pitch angles and altitudes without succumbing to the domain shifts that hinder zero-shot approaches.

Table 5. Quantitative evaluation of metric depth estimation on ground datasets.

Method	NYUv2		KITTI		ETH3D		iBims		DDAD		DIODE		HAMMER	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
Metric depth map (w/o GT intrinsics)														
ZoeDepth [3]	11.0	91.9	17.0	85.4	57.1	33.7	17.4	67.2	38.9	38.6	39.3	29.3	94.3	3.23
DepthPro [5]	10.7	91.9	23.5	38.3	38.5	32.8	15.9	81.5	33.4	35.3	31.9	37.7	39.1	63.0
UniDepthV2 [48]	10.2	93.4	**9.1**	**91.4**	20.7	69.5	**9.0**	**93.1**	18.4	**77.5**	42.9	51.7	38.1	46.8
MoGe2 [64]	**6.9**	**96.7**	17.6	64.7	**10.0**	**88.8**	14.6	80.5	**15.6**	73.7	16.7	67.9	**26.0**	65.3
MoGe2-Aerial	7.1	96.2	16.3	68.7	13.1	86.7	14.7	85.4	16.2	74.3	**15.1**	**77.0**	31.7	**68.9**
Metric depth map (w/ GT intrinsics)														
UniDepthV2 [48]	7.5	96.2	5.6	97.5	15.0	85.2	7.8	94.7	14.1	89.2	40.9	67.0	37.7	47.1
DepthAnything 3 [38]	7.1	97.0	7.3	96.0	12.5	85.5	**7.5**	**96.7**	12.4	87.3	15.2	76.6	**27.4**	66.0
Metric3Dv2 [26]	7.2	96.5	**5.3**	**98.0**	11.8	88.8	9.96	94.1	**9.21**	**93.7**	49.1	1.98	35.7	44.3
MoGe2 [64]	**6.0**	**97.2**	8.6	93.8	**9.3**	**93.7**	10.4	92.0	13.0	85.6	**14.9**	80.1	29.8	**72.5**
MoGe2-Aerial	6.0	97.1	8.1	94.3	12.9	85.5	10.9	91.8	12.9	86.7	15.1	**81.1**	35.8	72.0

Generalization to in-the-wild videos. We further evaluate existing methods on *AerialMetric-Wild*, which contains unconstrained internet aerial videos with large appearance variation and wide depth ranges. This setting is particularly challenging, and existing zero-shot models still show high AbsRel and low threshold accuracy (see Table 4). After fine-tuning on *AerialMetric*, performance improves markedly. Compared with zero-shot MoGe2 [64], our adapted model reduces AbsRel from 55.26 to 21.34 in the 0–400 m range, while improving δ_1 from 9.5 to 53.7, respectively. These results demonstrate that *AerialMetric* can enhance both robustness and generalization in monocular aerial depth estimation.

4.3 Evaluation on Ground-Domain Datasets

To verify that aerial adaptation does not overfit only to *AerialMetric-Oblique* and *AerialMetric-Decoupled*, we additionally report cross-domain performance on seven ground benchmarks (NYUv2 [44], KITTI [20], ETH3D [55], iBims [32], DDAD [22], DIODE [60], HAMMER [29]) to measure transfer and forgetting after aerial adaptation.

Table 4. Result on AerialMetric-Wild.

Method	Range: 0–400m			Range: 0–800m		
	AbsRel	δ_1	δ_2	AbsRel	δ_1	δ_2
ZoeDepth [3]	93.56	0.27	1.03	94.10	0.26	0.99
DepthPro [5]	98.06	0.00	0.02	98.18	0.00	0.03
UniDepthV2 [48]	52.88	27.65	61.34	55.97	22.92	55.02
Metric3Dv2 [26]	82.38	4.90	13.90	85.31	4.82	11.18
MoGe2 [64]	55.26	9.50	21.40	55.56	9.70	21.60
MoGe2-Aerial	**21.34**	**53.7**	**85.1**	**22.28**	**52.3**	**83.9**

Table 5 shows that MoGe2-Aerial remains competitive on ground benchmarks and improves multiple datasets over zero-shot MoGe2 [64] (e.g., KITTI, and iBims, DDAD), supporting that the adapted model preserves generalization beyond aerial domains.

4.4 Ablation Study

Fine-tuning Strategy Analysis. To determine the most effective adaptation approach, we evaluate several fine-tuning strategies: full fine-tuning, partial

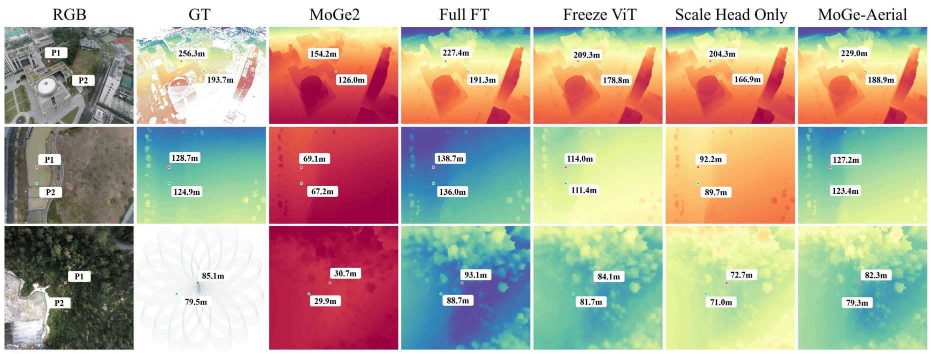


Fig. 8. Qualitative comparison of fine-tuning strategies.

Table 6. Fine-tuning strategies.

Strategy	Oblique		Decoupled		Ground	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
Zero-shot	49.50	7.90	38.38	20.00	**15.34**	76.78
Full FT	**6.68**	**95.60**	13.22	**86.00**	35.55	56.70
Freeze ViT	10.81	87.70	17.19	82.50	37.72	35.80
Scale Head	22.85	53.00	15.22	71.90	43.72	36.00
LoRA ($r = 64$)	12.66	82.90	13.03	84.30	16.02	**80.70**
LoRA ($r = 96$)	12.45	84.40	13.19	83.90	16.30	79.59
LoRA ($r = 128$)	12.94	82.40	**13.00**	84.20	18.80	69.40

Table 7. Analysis of training dataset composition.

Training Data	Aerial				Ground			
	Oblique (Mean)		Decoupled (Mean)		Ground-domain (Mean)		ETH3D	
	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1	AbsRel	δ_1
MoGe2	49.50	7.90	38.38	20.00	**15.34**	76.78	**9.98**	**88.79**
Oblique	**11.42**	**87.4**	16.65	77.45	20.51	77.64	26.37	61.50
Oblique+Synth	12.69	82.30	**13.05**	**84.20**	17.68	**79.65**	19.99	71.54
Oblique+Synth+Ground	12.45	84.40	13.19	83.90	16.30	79.59	13.10	86.67

fine-tuning (freezing the ViT backbone or updating only the scale head), and Low-Rank Adaptation (LoRA) with varying ranks. For LoRA, we inject adapters into all linear layers (`qkv`, `proj`, `fc1`, and `fc2`) with $r \in \{64, 96, 128\}$.

As shown in Table 6, while full fine-tuning achieves the highest scores on aerial datasets, it significantly degrades generalization to ground-level scenes. Partial fine-tuning, either freezing the backbone or isolating updates to the scale head—yields suboptimal performance. In contrast, LoRA provides a superior trade-off; specifically, $r = 96$ offers the best balance across diverse domains. Increasing the rank to $r = 128$ results in a slight performance drop, likely due to over-parameterization or increased sensitivity to distribution noise.

Effect of Training Data Composition. In the single-stage mixed-domain training phase, we balance the influence of various data sources through weighted sampling. The sampling proportions for *AerialMetric-Oblique*, *AerialMetric-Synthetic*, and Ground-domain datasets are maintained at 80%, 15%, and 5%, respectively.

Table 7 shows the effect of multi-domain data composition. Training with only aerial-domain data gives strong gains in *AerialMetric-Oblique* but leads to severe catastrophic forgetting on ground-domain tasks, as evidenced by a substantial performance drop on the ground dataset. While adding *AerialMetric-Synthetic* improves results on *AerialMetric-Decoupled*, it is insufficient to fully recover the robustness on ground-domain performance, with δ_1 results on ETH3D degrade to 71.54. Incorporating a small fraction of ground data effectively mitigates this forgetting issue, limiting the δ_1 degradation on ETH3D to merely 5.1 compared to the initial baseline, ultimately yielding the best overall compromise. Detailed results can be found in the supplementary materials.

5 Conclusion

We present AerialMetric, a large-scale and diverse benchmark for monocular metric depth estimation in aerial UAV imagery. Our dataset combines four complementary subsets, including large-scale real and synthetic image-depth pairs, a controlled acquisition subset for factorized analysis, and in-the-wild aerial imagery for cross-domain evaluation. Built on this benchmark, we conduct systematic studies of how viewpoint, altitude, and camera parameters affect aerial metric depth prediction, and establish a comprehensive evaluation protocol for this setting. By fine-tuning representative state-of-the-art models on AerialMetric, we achieve substantial gains and set new state-of-the-art performance across diverse UAV deployment scenarios.

Future Work An important next step is extending the current monocular benchmark to multi-view metric depth estimation, enabling stronger geometric consistency and broader applicability to UAV mapping and reconstruction pipelines. In addition, although our dataset improves aerial diversity, it still has limited coverage of extreme weather conditions, and future data collection under adverse weather is needed to improve robustness in real-world deployments.

6 Acknowledgements

This work was in part supported by the NSFC (Grant Nos. 62472453, 62411540034), Guangdong Natural Science Foundation (No. 2025A1515010782), Shenzhen Science and Technology Program (No. SYSRD20250529113401002), Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and the Guangdong Province Radio Science Data Center.

References

1. Barbato, F., Caligiuri, M., Zanuttigh, P.: Flyawarev2: A multimodal cross-domain uav dataset for urban scene understanding. Signal Processing: Image Communication p. 117483 (2026)
2. Beche, R., Nedevschi, S.: Claravid: A holistic scene reconstruction benchmark from aerial perspective with delentropy-based complexity profiling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26015–26025 (2025)
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
4. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
5. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2025)
6. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
7. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3d tracking and forecasting with rich maps. In: CVPR (2019)
8. Chen, M., Hu, Q., Yu, Z., Thomas, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L.: Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. arXiv preprint arXiv:2203.09065 (2022)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
10. Dhrafani, D., Liu, Y., Jong, A., Shin, U., He, Y., Harp, T., Hu, Y., Oh, J., Scherer, S.: Firestereo: Forest infrared stereo dataset for uas depth perception in visually degraded environments. IEEE Robotics and Automation Letters **10**(4), 3302–3309 (2025)
11. DJI: Dji terra. `https://enterprise.dji.com/dji-terra`, accessed: 2025-11-3
12. Du, B., Huang, Y., Chen, J., Huang, D.: Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In: CVPR (2023)
13. Du, K., Liao, X., Xia, J., Guo, C., Gu, Y., Guan, Y., Wang, D., Huang, S., Wang, Z.: Uavlight: A benchmark for illumination-robust 3d reconstruction in unmanned aerial vehicle (uav) scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5670–5679 (2026)
14. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. NIPS (2014)
15. ESRI: Create a 3d product with arcgis drone2map. `https://www.esri.com/zh-cn/arcgis/products/arcgis-reality/resources/sample-drone-datasets`, accessed: 2025-10-5
16. Florea, H., Miclea, V.C., Nedevschi, S.: Wilduav: Monocular uav dataset for depth estimation tasks. In: 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP). pp. 291–298. IEEE (2021)
17. Florea, H., Nedevschi, S.: Tandepth: Leveraging global dems for metric monocular depth estimation in uavs. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 5445–5459 (2025)

18. Fonder, M., Van Droogenbroeck, M.: Mid-air: A multi-modal dataset for extremely low altitude drone flights. In: CVPRW (2019)
19. Gao, Y., Li, C., You, Z., Liu, J., Li, Z., Chen, P., Chen, Q., Tang, Z., Wang, L., Yang, P., Tang, Y., Tang, Y., Liang, S., Zhu, S., Xiong, Z., Su, Y., Ye, X., Li, J., Ding, Y., Wang, D., Wang, Z., Zhao, B., Li, X.: Openfly: A comprehensive platform for aerial vision-language navigation. arXiv preprint arXiv:2502.18041 (2025)
20. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
21. Gross, M., Matha, S.B., Fahmy, A., Song, R., Cremers, D., Meeß, H.: Occufly: A 3d vision benchmark for semantic scene completion from the aerial perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21474–21485 (2026)
22. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: CVPR (2020)
23. He, J., Li, H., Sheng, M., Chen, Y.C.: Lotus-2: Advancing geometric dense prediction with powerful image generative model. arXiv preprint arXiv:2512.01030 (2025)
24. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Liu, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 (2024)
25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
26. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. TPAMI **46**(12), 10579–10596 (2024)
27. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR (2018)
28. Huang, Y., Chen, J., Huang, D.: Ufpmp-det: Toward accurate and efficient object detection on drone imagery. In: AAAI (2022)
29. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: On the importance of accurate geometry data for dense 3d vision tasks. In: CVPR (2023)
30. Katkuri, A.V.R., Madan, H., Khatri, N., Abdul-Qawy, A.S.H., Patnaik, K.S.: Autonomous uav navigation using deep learning-based computer vision frameworks: A systematic literature review. Array **23**, 100361 (2024)
31. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: 2026 International Conference on 3D Vision (3DV). pp. 499–509. IEEE (2026)
32. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: ECCV Workshops (2018)
33. Kolbeinsson, B., Mikolajczyk, K.: Ddos: The drone depth and obstacle segmentation dataset. In: CVPR (2024)
34. Kölle, M., Laupheimer, D., Schmohl, S., Haala, N., Rottensteiner, F., Wegner, J.D., Ledoux, H.: The hessigheim 3d (h3d) benchmark on semantic segmentation of high-resolution 3d point clouds and textured meshes from uav lidar and multi-view-stereo. ISPRS Open Journal of Photogrammetry and Remote Sensing **1**, 100001 (2021)
35. Li, H., Zou, Y., Chen, N., Lin, J., Liu, X., Xu, W., Zheng, C., Li, R., He, D., Kong, F., et al.: Mars-lvig dataset: A multi-sensor aerial robots slam dataset for

lidar-visual-inertial-gnss fusion. The International Journal of Robotics Research **43**(8), 1114–1127 (2024)
36. Li, Y., Liu, M., Wu, Y., Wang, X., Yang, X., Li, S.: Learning adaptive and view-invariant vision transformer for real-time uav tracking. In: ICML (2024)
37. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018)
38. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
39. Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., Huang, H.: Capturing, reconstructing, and simulating: the urbanscene3d dataset. In: ECCV (2022)
40. Lin, X., Song, M., Zhang, D., Lu, W., Li, H., Du, B., Yang, M.H., Nguyen, T., Qi, L.: Depth any panoramas: A foundation model for panoramic depth estimation. arXiv preprint arXiv:2512.16913 (2025)
41. Loquercio, A., Kaufmann, E., Ranftl, R., Müller, M., Koltun, V., Scaramuzza, D.: Learning high-speed flight in the wild. Science Robotics **6**(59), eabg5810 (2021)
42. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: Uavid: A semantic segmentation dataset for uav imagery. ISPRS journal of photogrammetry and remote sensing **165**, 108–119 (2020)
43. Madhuanand, L., Nex, F., Yang, M., et al.: Deep learning for monocular depth estimation from uav images. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **2** (2020)
44. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
45. Nex, F., Stathopoulou, E., Remondino, F., Yang, M., Madhuanand, L., Yogender, Y., Alsadik, B., Weinmann, M., Jutzi, B., Qin, R.: Usegeo-a uav-based multi-sensor dataset for geospatial research. ISPRS Open Journal of Photogrammetry and Remote Sensing **13**, 100070 (2024)
46. Nguyen, T.M., Yuan, S., Cao, M., Lyu, Y., Nguyen, T.H., Xie, L.: Ntu viral: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint. The International Journal of Robotics Research **41**(3), 270–280 (2022)
47. OpenDroneMap Authors: Odm - a command line toolkit to generate maps, point clouds, 3d models and dems from drone, balloon or kite images. `https://opendronemap.org/odm/datasets/` (2020), accessed: 2025-9-25
48. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Gool, L.V.: UniDepthV2: Universal monocular metric depth estimation made simpler (2025), `https://arxiv.org/abs/2502.20110`
49. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: UniDepth: Universal monocular metric depth estimation. In: CVPR (2024)
50. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
51. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. TPAMI **44**(3), 1623–1637 (2022)
52. Rizzoli, G., Barbato, F., Caligiuri, M., Zanuttigh, P.: Syndrone-multi-modal uav dataset for urban scenarios. In: ICCV (2023)
53. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
54. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)

55. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
56. Shimada, T., Nishikawa, H., Kong, X., Tomiyama, H.: Pix2pix-based monocular depth estimation for drones with optical flow on airsim. Sensors **22**(6), 2097 (2022)
57. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
58. Thalagala, R.G., De Silva, O., Jayasiri, A., Gubbels, A., Mann, G.K., Gosine, R.G.: Mun-frl: A visual-inertial-lidar dataset for aerial autonomous navigation and mapping. The International Journal of Robotics Research **43**(12), 1853–1866 (2024)
59. Turki, H., Ramanan, D., Satyanarayanan, M.: Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In: CVPR (2022)
60. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
61. Vuong, K., Ghosh, A., Ramanan, D., Narasimhan, S., Tulsiani, S.: Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In: CVPR (2025)
62. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 196–214 (2022)
63. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR (2025)
64. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. Advances in Neural Information Processing Systems **38**, 35928–35959 (2025)
65. Wang, S., Li, S., Zhang, Y., Yu, S., Yuan, S., She, R., Guo, Q., Zheng, J., Howe, O.K., Chandra, L., et al.: Uavscenes: A multi-modal dataset for uavs. In: ICCV (2025)
66. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: IROS (2020)
67. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
68. Wu, R., Cheng, X., Zhu, J., Liu, X., Zhang, M., Yan, S.: Uavd4l: A large-scale dataset for uav 6-dof localization. arXiv preprint arXiv:2401.05971 (2024)
69. Wu, Y., Wang, X., Yang, X., Liu, M., Zeng, D., Ye, H., Li, S.: Learning occlusion-robust vision transformers for real-time uav tracking. In: CVPR (2025)
70. Xiong, B., Li, Z., Li, Z.: Gauu-scene: A scene reconstruction benchmark on large scale 3d reconstruction dataset using gaussian splatting. arXiv preprint arXiv:2401.14032 (2024)
71. Xiong, B., Zheng, N., Liu, J., Li, Z.: Gauu-scene v2: Assessing the reliability of image-based metrics with expansive lidar image dataset using 3dgs and nerf. arXiv preprint arXiv:2404.04880 (2024)
72. Yan, Q., Zheng, J., Reding, S., Li, S., Doytchinov, I.: Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In: CVPR (2022)

73. Yang, G., Xue, F., Zhang, Q., Xie, K., Fu, C.W., Huang, H.: Urbanbis: a large-scale benchmark for fine-grained urban building instance segmentation. In: SIGGRAPH (2023)
74. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
75. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NIPS (2024)
76. Ye, C., Zhuge, Y., Zhang, P.: Towards open-vocabulary remote sensing image semantic segmentation. In: AAAI (2025)
77. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: ICCV (2023)
78. Yu, H., Lin, H., Wang, J., Li, J., Wang, Y., Zhang, X., Wang, Y., Zhou, X., Hu, R., Peng, S.: Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. arXiv preprint arXiv:2601.03252 (2026)
79. Yu, K., Li, H., Xing, L., Wen, T., Fu, D., Yang, Y., Zhou, C., Chang, R., Zhao, S., Xing, L., et al.: Scene-aware refinement network for unsupervised monocular depth estimation in ultra-low altitude oblique photography of uav. ISPRS Journal of Photogrammetry and Remote Sensing **205**, 284–300 (2023)
80. Zhang, Y., Chen, G., Chen, J., Cui, S.: Aerial lifting: Neural urban semantic and building instance lifting from aerial imagery. In: CVPR (2024)
81. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACMMM (2020)
82. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. TPAMI **44**(11), 7380–7399 (2021)
83. Zhu, Y., Kong, Y., Jie, Y., Xu, S., Cheng, H.: Graco: A multimodal dataset for ground and aerial cooperative localization and mapping. IEEE Robotics and Automation Letters **8**(2), 966–973 (2023)
84. Zolanvari, S., Ruano, S., Rana, A., Cummins, A., Da Silva, R.E., Rahbar, M., Smolic, A.: Dublincity: Annotated lidar point cloud and its applications. arXiv preprint arXiv:1909.03613 (2019)