

AirZoo: A Unified Large-Scale Dataset for Grounding Aerial Geometric 3D Vision

Xiaoya Cheng^{1†}, Rouwan Wu^{1†}, Xinyi Liu^{1†}, Zeyu Cui^{1†}, Yan Liu^{2†}
 Na Zhao³, Yu Liu¹, Maojun Zhang¹, and Shen Yan^{1*}

¹ National University of Defense Technology, Changsha, China
 {chengxy, wurouwan97, liuxinyi24, czy806, jasonyuliu, mjzhang,
 yanshen12}@nudt.edu.cn

² National Key Laboratory of Advanced Guidance and Control Technology,
 Changsha, China
 cowisee633@gmail.com

³ Singapore University of Technology and Design, Singapore
 na_zhao@sutd.edu.sg

† Equal contribution.

<https://nudt-sawlab.github.io/AirZoo/>

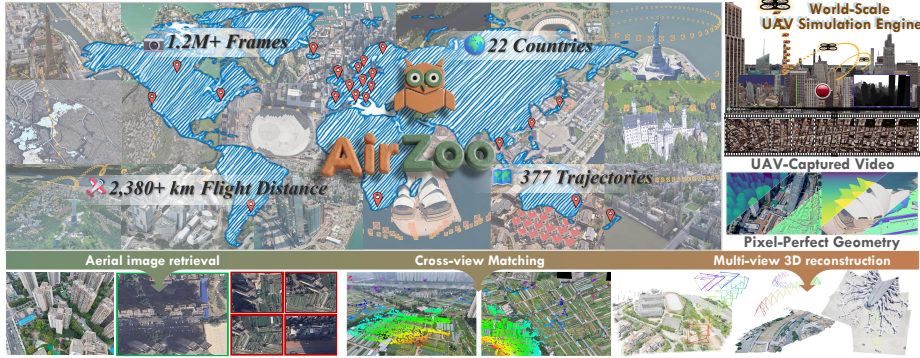


Fig. 1: Overview of AirZoo. (Top) A million-scale synthetic UAV dataset with pixel-perfect geometry featuring extensive flight trajectories across diverse global regions. (Bottom) Fine-tuning on AirZoo empowers SoTA models to achieve superior performance in aerial geometric 3D vision tasks, including **Aerial image Retrieval**, **Cross-view Matching**, and **Multi-view 3D Reconstruction**.

Abstract. Despite the rapid progress in data-driven 3D vision, aerial geometric 3D vision remains a formidable challenge due to the severe scarcity of large-scale, high-fidelity training data. Existing benchmarks, predominantly biased toward ground-level or object-centric views, do not account for complex viewpoint transformations and diverse environmental conditions in UAV-based sensing. To bridge this critical gap, we propose **AirZoo**, a unified large-scale dataset and benchmark for grounding aerial geometric 3D vision. AirZoo possesses three appealing properties: **1) Scalable Generation Pipeline:** Leveraging freely

* Corresponding author.

available, world-scale photogrammetric 3D meshes, it renders vast outdoor environments with customizable UAV flight trajectories and configurable weather/illumination. **2) Comprehensive Scene Diversity:** It provides extensive global coverage (22 countries, 95 base flight sequences, and 377 weather-conditioned trajectories), systematically encompassing both highly structured urban landscapes and complex unstructured natural environments. **3) Rich Geometric Annotations:** Each frame provides synchronized, pixel-level metric depth and precise 6-DoF georeferenced poses, essential for geometry-aware learning. Through three rigorous evaluation tracks—aerial image retrieval, cross-view matching, and multi-view 3D reconstruction—we demonstrate that AirZoo serves as a powerful "pre-training engine." Extensive experiments on both public and newly collected real-world benchmarks reveal that fine-tuning on AirZoo yields substantial performance gains for state-of-the-art models (e.g., MegaLoc, RoMa, VGGT, and Depth Anything 3), establishing a new performance upper bound for aerial spatial intelligence.

Keywords: Aerial Geometric 3D Vision · Global 3D Rendering · Cesium for Unreal · Large-scale Aerial Dataset · Retrieval, Matching, and Reconstruction

1 Introduction

Geometric 3D vision, encompassing fundamental tasks such as image-based retrieval, feature matching, and multi-view reconstruction, serves as the cornerstone of spatial intelligence and is indispensable for autonomous robotics and mixed reality. Recently, the advent of foundation models with large-scale dataset training has yielded unprecedented performance milestones. However, these advancements are largely tethered to a terrestrial-centric bias, as mainstream training corpora are predominantly harvested from ground-level [7, 13, 25, 29, 41, 57] or object-centric [6, 12, 28, 53] perspectives. This inherent bias creates a severe domain gap when these models are deployed on UAV platforms. Unlike ground-based conditions, UAVs operate in a highly unconstrained 6-DoF space, introducing unique geometric challenges: drastic viewpoint shifts from oblique to nadir, the frequent replacement of horizontal facades with roofs, and extreme scale variations due to changing flight altitudes. Consequently, there is an urgent imperative to curate a large-scale, geometry-aware aerial dataset for these foundation models to unlock the potential of aerial spatial intelligence.

Existing UAV datasets struggle to simultaneously deliver the global scale and dense geometric supervision required for robust 3D vision (see Tab. 1). While many existing UAV datasets offer photorealism [8, 43, 54, 61], they often lack the scale and complete geometric ground truth (e.g., 6-DoF poses, metric depth) needed for robust training. Even recent efforts providing geometric data [9, 11, 16, 19, 24, 35, 48, 51] are limited to a few manually crafted city models and lack viewpoint variety, making them insufficient for sequential geometric supervision. Alternatively, while some works use freely available, global-scale 3D models like Google Maps [2, 59] to achieve unbounded spatial coverage, they are mainly

designed for image retrieval tasks. Consequently, they only provide photometric data, lacking the underlying 3D geometric ground truth entirely.

To boost the research on aerial geometric 3D vision, we present **AirZoo**, a large-scale synthetic dataset designed to bridge the critical ground-to-aerial data gap. Our dataset has several appealing properties: **1) Scalable Generation Pipeline:** We develop a fully automated simulator based on a custom AirSim-Cesium-Unreal Engine pipeline. The pipeline can automatically generate diverse environmental conditions, including varying weather and lighting, ensuring high-fidelity rendering that reflects the complexity of the real world. **2) Comprehensive Scene Diversity:** Unlike previous works limited to closed worlds, our collection spans 22 countries with 95 base flight sequences rendered into 377 weather-conditioned trajectories. This massive scale covers a wide range of operational UAV environments, ranging from dense urban centers to rural landscapes, to ensure robust generalization. **3) Rich Geometric Annotations:** Each flight sequence is captured with synchronized sensors, providing RGB images, metric depth maps, and absolute geo-coordinates. This compels the network to learn features grounded in stable 3D geometry rather than transient textures.

To systematically evaluate the utility of AirZoo, we propose a comprehensive evaluation framework spanning three core aerial vision tasks: a) aerial image retrieval, b) cross-view matching, and c) multi-view 3D reconstruction. Beyond standard public benchmarks, we introduce AirZoo-Real, a real-world dataset designed to test the limits of current algorithms. By fine-tuning a suite of foundation models—including MegaLoc [3], RoMa [10], VGGT [45], and Depth Anything 3 [26]—on the synthetic diversity of AirZoo, we demonstrate substantial performance gains across all tracks. Most notably, our results reveal that AirZoo imparts critical domain-invariant geometric knowledge, empowering these models with robust zero-shot generalization capabilities when deployed in complex, unseen real-world UAV scenarios.

Our main contributions are summarized as follows:

- We introduce AirZoo, a million-scale synthetic UAV dataset covering diverse global regions with pixel-perfect geometric supervision (metric depth, verified poses).
- We propose a scalable AirSim-Cesium-Unreal simulator that enables automated, high-fidelity rendering of global 3D tiles with varying weather and lighting.
- We finetune SoTA methods (MegaLoc, RoMa, VGGT, Depth Anything 3) on aerial retrieval, matching, and reconstruction tasks, demonstrating that training on AirZoo significantly improves the performance.

2 Related Work

UAV-Specific Datasets. Constructing a large-scale aerial benchmark with precise geometry is non-trivial. Existing datasets struggle to simultaneously balance

Table 1: Comparison with large-scale UAV visual geometry datasets. **Regions** denotes geographically distinct environments. (Note that University-1652 and SUES-200 are limited to isolated landmarks, while AirZoo covers city-scale regions). **Condition** refers to environmental variations. **Sequence** denotes continuous flight trajectories. **Pose** indicates precise 6-DoF trajectory. ✓ indicates available, ✗ unavailable.

Dataset	Images	Regions	Altitude (m)	Pitch	Type	Condition	Pose	Depth	Sequence
University-1652 [59]	37.8k	1652	-	Varied	Syn.	✗	✗	✗	✗
SUES-200 [61]	24.1k	200	150–300	Varied	Real	✗	✗	✗	✓
DenseUAV [8]	27k	14	80–100	Fixed	Real	✓	✗	✗	✗
UAV-VisLoc [54]	6.7k	11	400–2k	Fixed	Real	✓	✗	✗	✓
UAVD4L [51]	6.8k	1	50–300	Varied	Real	✗	✓	✓	✗
Game4Loc [19]	33.7k	1	80–650	Varied	Syn.	✗	✓	✗	✗
MatrixCity [24]	67k	2	100–450	Oblique	Syn.	✓	✓	✓	✗
UAVScenes [48]	120k	4	80–130	Fixed	Real	✗	✓	✓	✓
OrthoLoC [9]	16.4k	47	50–300	Varied	Real	✗	✓	✓	✗
AirZoo (Syn)	1.2M	95	0–800	Varied	Syn.	✓	✓	✓	✓
AirZoo (Real)	9.4k+	2	0–350	Varied	Real	✓	✓	✓	✓

global scale, environmental diversity, and dense geometric supervision. Specifically, expansive aerial collections and large-scale retrieval benchmarks [2, 8, 43, 54, 59, 61] offer high photorealism or expansive coverage, but they fundamentally lack complete geometric ground truth. To circumvent this, geometry-centric datasets [9, 11, 16, 19, 24, 35, 48, 51] provide dense annotations, yet they are severely restricted to a few localized scenes or specific city models with limited viewpoint variety. To bridge this critical gap, we propose AirZoo to unlock the geometric potential of world-scale 3D maps. By introducing an automated pipeline that simulates realistic UAV flight trajectories to render continuous sequences from global 3D tiles, AirZoo uniquely pairs immense geographic variety with synchronized dense geometry. This sequence-level design provides the essential spatial context required for robust 3D vision, significantly boosting zero-shot generalization on real-world aerial tasks.

Aerial Image Retrieval. Cross-view image retrieval establishes the similarity between UAV and satellite imagery to achieve coarse UAV localization. Due to the high cost of collecting real-world data, University-1652 [59] utilized Google Earth to simulate circle flights, constructing the first cross-view UAV-satellite dataset and employing contrastive learning for discriminative feature learning. With the advent of foundation models [33], AnyLoc [20] combined large-scale pre-trained models with training-free aggregation, proposing a unified method for various sub-tasks. However, given the scarcity of cross-view data and ground truth, most current approaches still adopt methodologies from ground-based place recognition, such as SALAD [18] and MegaLoc [3], which benefit from more extensive training data.

Cross-view Matching. Cross-view matching aims to establish dense correspondences or feature associations between images captured from drastically different viewpoints. Previous works have largely focused on the ground-to-aerial domain, utilizing datasets such as AerialMegadepth [43] and BlendedMVS [56]. However, state-of-the-art matching [10, 40] methods are predominantly trained on ground-

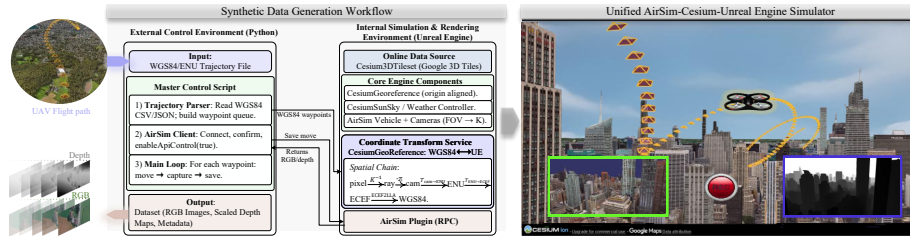


Fig. 2: Overview of the AirSim-Cesium-Unreal simulator. Python issues WGS84 trajectories, Cesium for Unreal streams Google 3D Tiles into UE, and AirSim controls synchronized RGB, depth, and georeferenced pose capture.

to-ground datasets like MegaDepth. Consequently, these models suffer from a significant domain gap when applied to aerial-to-satellite matching. Due to the lack of dedicated training on top-down satellite imagery, their capability to handle the extreme geometric distortions and scale variations in aerial-satellite pairs is severely limited.

Multi-view visual geometry estimation. Traditional pipelines and early learning methods [4, 22, 23, 36–38, 44] heavily rely on stable correspondences and struggle in ill-posed conditions. Recently, feed-forward Transformers [5, 26, 45, 47, 55, 58] shifted the paradigm by directly predicting point maps to jointly recover depth and poses. Leveraging massive training data and unified architectures, these methods have achieved state-of-the-art performance and remarkable generalization across most standard scenarios. However, since their training datasets mainly consist of ground-level images [7, 13, 25, 29, 41, 57], deploying these general-purpose models on UAVs reveals a severe domain gap. They struggle with unique aerial characteristics such as drastic oblique-to-nadir viewpoint shifts and unstructured terrains. Consequently, scalable and robust geometry estimation tailored for UAVs remains an open challenge. In this work, we leverage our constructed dataset to empower these foundation models, effectively unlocking their potential for aerial applications.

3 The AirZoo Dataset

In this section, we describe the data generation framework and the statistics of AirZoo.

3.1 Data Generation and Processing

Global Region Selection. We utilize Cesium for Unreal, a plugin that streams Google 3D Tiles directly into Unreal Engine 5 (UE5). These tiles provide the highest level of detail for global-scale 3D photogrammetry models currently available. We select regions across six continents, covering urban, rural, and natural terrains for subsequent UAV trajectory simulations. Together, these curated regions establish the bases for our subsequent UAV trajectory simulations.

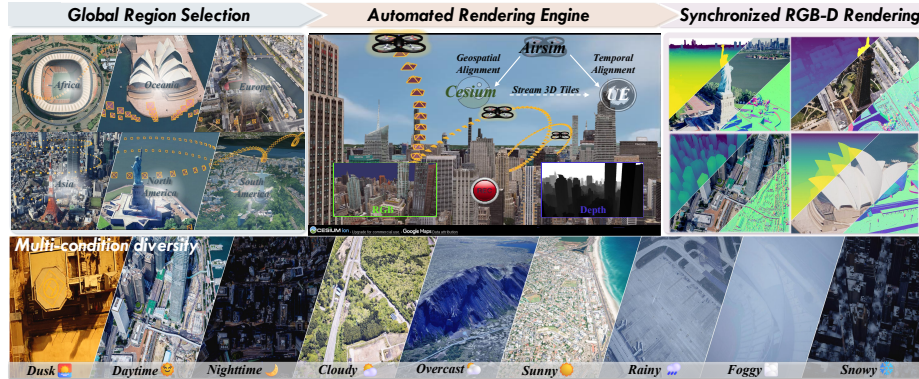


Fig. 3: Construction pipeline and properties of AirZoo. Global region selection (top-left), simulation (top-middle), synchronized RGB-D rendering (top-right), and multi-condition diversity (bottom).

Automated Rendering Pipeline. We develop a custom AirSim-Cesium-Unreal simulator for continuous UAV trajectories over global terrains (Fig. 2). The system integrates three core components:

- **Unreal Engine (UE)** provides physically based rendering and supports dynamic weather and illumination.
- **Cesium for Unreal** streams Google 3D Tiles into UE and anchors the simulation world to geographic coordinates via Cesium Georeference.
- **AirSim** simulates the multi-rotor UAV and virtual cameras, executing WGS84 waypoints with reproducible pose control.

A Python controller orchestrates the acquisition loop across these components. It parses predefined trajectories, sends waypoint commands through the AirSim RPC interface, and performs on-the-fly WGS84-to-UE coordinate conversion. For each waypoint, AirSim’s Steppable Clock coordinates tile loading and sensor capture, allowing the simulator to move the UAV to the target pose and record $\mathcal{M} = \{RGB, depth, intrinsics, pose\}$ in a synchronized step after the streamed tiles are stabilized. This controlled loop provides deterministic captures with reliable geometric annotations for downstream learning.

Synchronized RGB-D Rendering. Driven by the aforementioned pipeline, we record continuous 30 fps flight trajectories rather than isolated keyframes, generating temporally consistent data along each trajectory. As illustrated in Fig. 3 (top-right), each rendered frame provides an RGB image coupled with pixel-aligned absolute depth (in meters) and calibrated camera intrinsics. By extracting these complete modalities, our engine overcomes the limitations of previous works [2, 59] that rely solely on photometric screenshots, enabling true geometry-aware learning. Furthermore, we leverage UE5’s dynamic volumetric systems to simulate diverse environmental conditions. As shown in the condition examples of Fig. 3, we systematically randomize the time of day and weather effects for each scene.

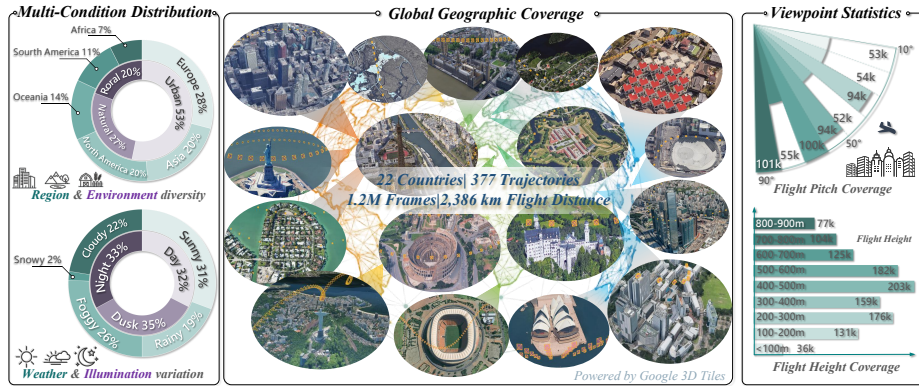


Fig. 4: Comprehensive statistics of AirZoo. (Left) Region, environment, and weather distributions. (Center) Global coverage. (Right) Altitude and pitch envelopes.

3.2 Dataset Statistics

Scale and Modalities. AirZoo represents the largest and most geographically diverse UAV geometric dataset to date, structured as continuous video sequences rather than isolated image collections. As shown in Fig. 4, it spans 377 trajectories across 22 countries, yielding over 1.2 million high-resolution (1600×1200) frames derived from nearly 2,400 km of cumulative flight trajectories. These regions cover six continents and encompass a wide range of environmental diversity, including urban, rural, and natural landscapes (summarized in the top-left of Fig. 4).

As illustrated in the left panel of Fig. 4, the sequences are annotated with semantic tags covering various weather presets (*Sunny*, *Cloudy*, *Rainy*, *Foggy*, *Snowy*) and times of day (*Day*, *Sunset*, *Night*). Along these sequences, our simulated acquisition flights operate within the 0–800 m UAV altitude envelope. To provide oblique-to-nadir coverage for existing methods, we dynamically sweep the camera gimbal pitch from 10° (oblique) to 90° (nadir) across the sequences (detailed in the right panel of Fig. 4).

Geometric Consistency. For seamless integration with standard geometric pipelines (e.g., COLMAP), we unify data structure and provide a standard pin-hole camera model with per-frame intrinsics (f_x, f_y, c_x, c_y) and precise 6-DoF poses logged in both WGS84 (longitude, latitude, altitude) and ECEF Cartesian coordinate systems. The full pixel-to-WGS84 coordinate transformation chain is detailed in the supplementary material.

To validate spatial fidelity, we back-project points from a source frame using rendered depth and pose, then re-project them into a neighboring target frame. Bidirectional projection yields a median relative depth error of 0.066% with P90 at 0.174% and P95 at 0.380% (Fig. 5), confirming pixel-level depth and pose fidelity. We further visualize multi-view correspondences and covisibility across

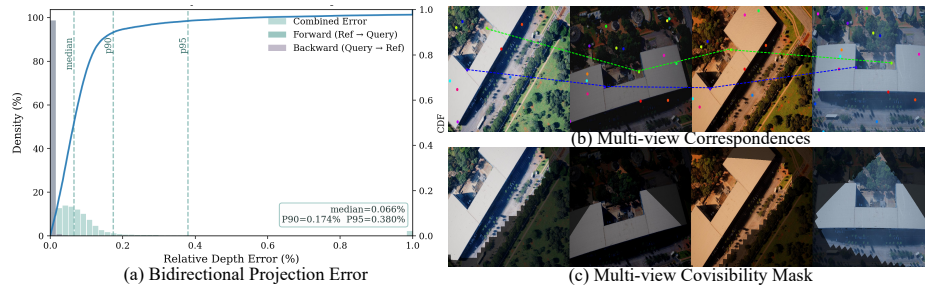


Fig. 5: Geometric verification. (a) Projection-error CDF with 0.066% median error. (b,c) Multi-view correspondences and covisibility from one region across four weather trajectories.

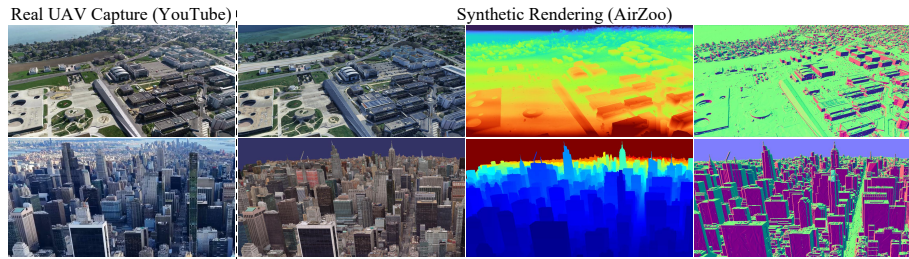


Fig. 6: Real vs. synthetic comparison at the same viewpoint.

weather trajectories in Fig. 5(b,c), showing stable geometric alignment under varying appearance conditions.

Photometric Realism. AirZoo uses Google Photorealistic 3D Tiles as the data source, rendering from real-world 3D meshes textured with high-resolution imagery [15]. To inspect visual realism, we collect real UAV views from public DJI clips on YouTube, and reproduce the same viewpoints in AirZoo for synthetic rendering. As shown in Fig. 6, the real and synthetic views are comparable in scene layout, object scale, and dominant structural patterns. Although lighting and shading differences remain visible cues for distinguishing the two, AirZoo additionally provides synchronized geometry that is difficult to obtain from real online videos.

3.3 Dataset Usage and Evaluation Protocol

We enforce zero spatial overlap between the AirZoo training set and all test benchmarks (Table 2). Specifically, the synthetic AirZoo training split covers 19 countries, while Brazil, the USA, and New Zealand are held out for AirZoo-Test. The evaluation protocol combines AirZoo-Real with public UAV benchmarks, allowing us to test whether synthetic geometric diversity transfers to unseen real regions across retrieval, matching, and reconstruction.

Table 2: Train/test datasets with zero spatial overlap between Part 1 and Part 2.

Dataset	Type	Regions / Cities	Usage	Task	Results
Part 1, Training (Synthetic)					
AirZoo (Train)	Synth.	Global (19 countries)	Train	All	-
Part 2, Zero-Shot Test					
AirZoo-Real	Real	Changsha (China)	Test	All	Tabs. 4, 6, 7
UAV-VisLoc [54]	Real	11 regions (China)	Test	Retrieval	Tab. 3
AerialExtreMatch [52]	Real	Changsha (China)	Test	Matching	Tab. 5
UAVScenes [48]	Real	HK, Singapore	Test	Reconstruction	Tab. 7
UrbanScene3D [27]	Real	Shenzhen (China)	Test	Reconstruction	Tab. 7
AirZoo-Test	Synth.	Brazil, USA, NZ	Test	Reconstruction	Tab. 7

4 Experiments

We evaluate AirZoo as a training dataset on three representative UAV tasks: aerial image retrieval, cross-view matching, and multi-view 3D reconstruction. For each task, we keep the backbone architecture fixed and compare its original pre-trained version with the same model fine-tuned on AirZoo, so the performance gap reflects the data contribution rather than architectural changes.

4.1 Aerial Image retrieval

Aerial image retrieval is formulated as cross-view geo-localization: given a UAV query image, the goal is to retrieve the most relevant geo-tagged satellite tile. This setting is challenging because viewpoint, scale, illumination, and seasonal appearance can vary simultaneously. Existing methods still suffer from a clear domain gap. Several approaches [17, 30, 51] transfer models trained on ground-level data, and zero-shot pipelines such as AnyLoc [20] rely on frozen features. Game4Loc [19] is designed for aerial retrieval but is trained in relatively limited synthetic environments.

Training on AirZoo. We adopt MegaLoc [3] as the base retriever and fine-tune it on AirZoo. MegaLoc is pre-trained on four ground-view datasets [1, 7, 42, 50], providing strong generic representations but remains suboptimal for nadir-to-oblique matching. During fine-tuning, we use weighted InfoNCE [19], where positive similarity is modulated by continuous overlap ratios between queries and satellite tiles.

Evaluation Benchmark. We evaluate on two real-world aerial benchmarks: the public UAV-VisLoc [54] and AirZoo-Real (i.e., our newly collected real-flight set). In both settings, UAV images serve as queries, and the retrieval gallery is built from Google Maps [14] satellite imagery at 0.3m ground resolution. Together, the benchmark covers 13 geographic regions and over 1,000 query images across diverse lighting conditions and flight altitudes. To reduce scale ambiguity caused by altitude variation, we construct the gallery with multi-scale cropping (tile sizes 512×512 and 1024×1024) and a 50% overlap ratio.

Table 3: Geo-localization results on the UAV-VisLoc dataset. Our AirZoo fine-tuned MegaLoc consistently improves over the original MegaLoc across all evaluation metrics (R@1, R@3, R@5, and AP@5). Improvements over MegaLoc are shown in gray, and the best results are highlighted in **bold**.

Method	R@1	R@3	R@5	AP@5
AnyLoc [20]	17.01	29.51	36.57	12.60
SALAD [18]	25.72	40.03	46.77	17.95
Game4Loc [19]	18.68	30.75	41.79	13.20
MegaLoc [3]	26.91	45.04	52.68	19.09
MegaLoc (Ours)	28.64 (+1.73)	46.33 (+1.29)	53.74 (+1.06)	19.78 (+0.69)

Table 4: Geo-localization results on the AirZoo-Real dataset. Our AirZoo fine-tuned MegaLoc consistently improves over the original MegaLoc across all evaluation metrics (R@1, R@3, R@5, and AP@5). Improvements over MegaLoc are shown in gray, and the best results are highlighted in **bold**.

Method	R@1	R@3	R@5	AP@5
AnyLoc [20]	4.52	9.35	10.45	2.94
SALAD [18]	7.36	19.18	29.48	7.73
Game4Loc [19]	12.67	23.59	34.18	10.66
MegaLoc [3]	6.46	16.52	22.16	6.33
MegaLoc(Ours)	18.66 (+12.21)	40.70 (+24.18)	50.53 (+28.38)	17.22 (+10.89)

Evaluation Metrics. We use Recall@K ($K \in \{1, 3, 5\}$) as the primary metric, where a query is counted as correct if at least one positive tile appears in the top- K retrieved results. We also report AP@5 to evaluate ranking quality among the top returned candidates.

Results. Quantitative comparisons on UAV-VisLoc and our AirZoo-Real are reported in Tab. 3 and Tab. 4, respectively (with per-scene results in the supplementary). We observe that 1) AirZoo fine-tuning consistently improves top-rank retrieval quality across both benchmarks; 2) the gain on UAV-VisLoc is moderate since its query-reference pairs are predominantly nadir-dominant and primarily differ in scale rather than viewpoint, making the cross-view gap less pronounced; and 3) the gain on our real-flight benchmark is substantially larger, where arbitrary viewpoints coupled with stronger environmental variations create a more challenging cross-view domain gap that better reflects real-world deployment scenarios. These results confirm that AirZoo is particularly effective for improving robustness in realistic UAV geo-localization under extreme cross-view conditions beyond near-nadir settings.

4.2 Cross-view Matching

Cross-view matching between orthophoto references and oblique UAV imagery is a core step for UAV pose estimation. The task is difficult because large viewpoint changes, altitude variation, and weather differences jointly affect geometric correspondence quality. Most existing matchers are not designed for extreme aerial



Fig. 7: Qualitative cross-view geo-localization comparisons. Each row corresponds to a real-world UAV query, and each column shows the Top-1 retrieved satellite tile from different methods. Correct and incorrect predictions are highlighted with green and pink borders, respectively. Fine-tuning on AirZoo yields more stable retrievals across challenging viewpoints and environmental changes.

gaps. ELoFTR [49] and RoMa [10] are primarily trained on ground-view datasets such as MegaDepth [25]. RoMa (GIM) [39] leverages massive video supervision to enlarge data scale; however, such video-derived data lacks long-term temporal spans and diverse environmental dynamics, failing to capture complex illumination.

Training on AirZoo. We adopt RoMa [10] as the base matcher and fine-tune it on a curated mixture of the MegaDepth [25] and AirZoo datasets. The network weights are initialized from the publicly available RoMa checkpoint pre-trained on outdoor scenes [10].

Evaluation Benchmark. We evaluate on two complementary benchmarks: the public AerialExtreMatch [52] dataset and AirZoo-Real (i.e., our newly collected real-flight benchmark). Combined, they span three geographic locations and more than 500 UAV query images. For each query, we use its GPS prior to crop spatially corresponding reference tiles from Digital Surface/Orthophoto Models (DSM/DOM), and then perform cross-view matching on these candidates.

Table 5: 6-DoF localization results on the AerialExtreMatch dataset. Gray values indicate improvements over the previous baseline.

Method	(5m, 1°) ↑	(10m, 1°) ↑	(20m, 2°) ↑
LoFTR [40]	66.67	66.67	84.47
ELoFTR [49]	81.82	81.82	85.61
DUST3R [47]	1.52	3.78	16.29
MASt3R [32]	76.14	76.52	87.50
RoMA [10]	95.83	95.83	96.59
RoMA (GIM) [39]	94.32 (−1.51)	94.32 (−1.51)	97.35 (+0.76)
RoMA (Ours)	96.21 (+0.38)	96.21 (+0.38)	98.11 (+1.52)

Table 6: Translation results on the AirZoo-Real dataset. Gray values indicate improvements over the previous baseline.

Method	median ↓	acc@5m ↑	acc@10m ↑	acc@20m ↑
LoFTR [40]	244.20	3.04	6.89	8.13
ELoFTR [49]	229.64	3.93	7.73	9.40
DUST3R [47]	214.57	2.46	7.80	14.15
MASt3R [32]	183.75	12.70	20.06	24.12
RoMA [10]	5.87	44.95	55.58	57.30
RoMA (GIM) [39]	3.22 (−2.65)	77.71 (+32.76)	84.14 (+28.56)	86.93 (+29.63)
RoMA (Ours)	3.03 (−2.84)	79.78 (+34.83)	85.35 (+29.77)	86.93 (+29.63)

Evaluation Metrics. We report recall-based localization metrics. On AerialExtreMatch, which provides full 6-DoF annotations, we measure pose recall at (5m, 1°), (10m, 1°), and (20m, 2°). On the AirZoo-Real dataset, where the ground truth comes from precise RTK positions, we report translation recall at 5m, 10m, and 20m, together with median translation error.

Results. Quantitative evaluations on the AerialExtreMatch and AirZoo-Real datasets are reported in Tab. 5 and Tab. 6, with qualitative comparisons in Fig. 8. We observe that 1) AirZoo fine-tuning consistently improves performance over the original RoMa in 6-DoF and translation-only evaluations; 2) the advantage remains under both strict and relaxed error thresholds, indicating stable geometric generalization rather than threshold-specific gains; and 3) improvements are especially clear in translation-only localization, where the AirZoo-trained model produces fewer mismatches and more spatially distributed correspondences on static scene regions. AirZoo provides effective cross-view supervision for robust UAV matching and downstream pose estimation under challenging viewpoint and environmental variations.

4.3 Multi-view 3D Reconstruction

Multi-view 3D reconstruction from UAV imagery aims to recover camera poses and dense geometry from image sequences. Compared with ground-view capture, UAV data introduces larger altitude changes, stronger oblique views, and wider illumination variations. Recent transformer-based systems, including CUT3R [46],

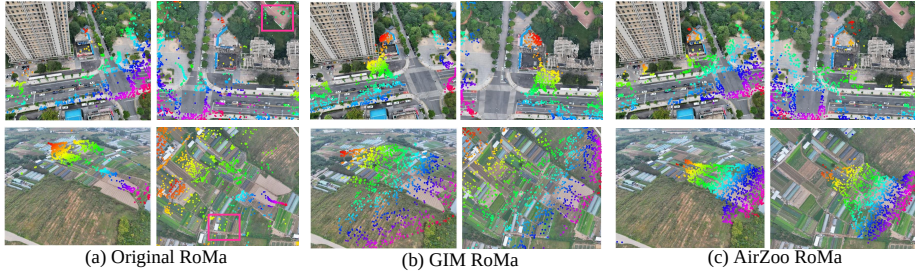


Fig. 8: Qualitative cross-view matching results on AirZoo-Real. Pink boxes highlight mismatched regions. The original RoMa produces many incorrect correspondences under aerial cross-view conditions. RoMa (GIM), trained on large-scale aerial video data, alleviates this issue to some extent. Our RoMa, trained specifically for cross-view matching, produces more uniformly distributed correspondences concentrated on static scene structures.

MapAnything [21], VGGT [45], and DA3 [26], are mainly trained on ground-level datasets such as CO3D [34] and RealEstate10K [60]. As a result, their direct transfer to UAV trajectories remains challenging.

Training on AirZoo. We fine-tune VGGT [45] and DA3 [26] on AirZoo. We select 16 trajectories from Brazil, the USA, and New Zealand for validation/testing, and use the remaining 361 trajectories for training on 8 A100 GPUs. During training, we sample clips with overlap ratios between 50% and 75%, and variable sequence lengths (2–24 frames for VGGT, 2–10 frames for DA3). This strategy increases exposure to diverse baselines and motion patterns.

Evaluation Benchmark. We evaluate on a synthetic test set constructed by selecting 16 trajectories from AirZoo, as well as on AirZoo-Real (i.e., our newly collected real-flight set). The real-flight dataset contains 9,430 images captured over four areas (school, driving school, substation, and plaza) during three time windows (06:00–08:00, 12:00–14:00, and 18:00–20:00), with additional low-light data (22:00–24:00) for two scenes. During acquisition, the UAV pitch angle is maintained between 30° and 45° at approximately 160 m altitude, with RTK ensuring accurate extrinsics. Depth maps are obtained by rendering from 3D models built via oblique photogrammetry.

Evaluation Metrics. Following the standard evaluation protocol in [26], we report distance-thresholded 3D reconstruction metrics using the F1 score computed from precision and recall based on Chamfer Distance, with a threshold of $d = 10$.

Results. Quantitative evaluations on UAVScenes [48], UrbanScene3D [27], AirZoo test set, and AirZoo-Real are reported in Tab. 7, with qualitative comparisons in Fig. 9. We observe that 1) AirZoo fine-tuning consistently improves performance over both VGGT and DA3 across all test sets; 2) the discrepancy between real-world and synthetic results reveals that domain gap remains a

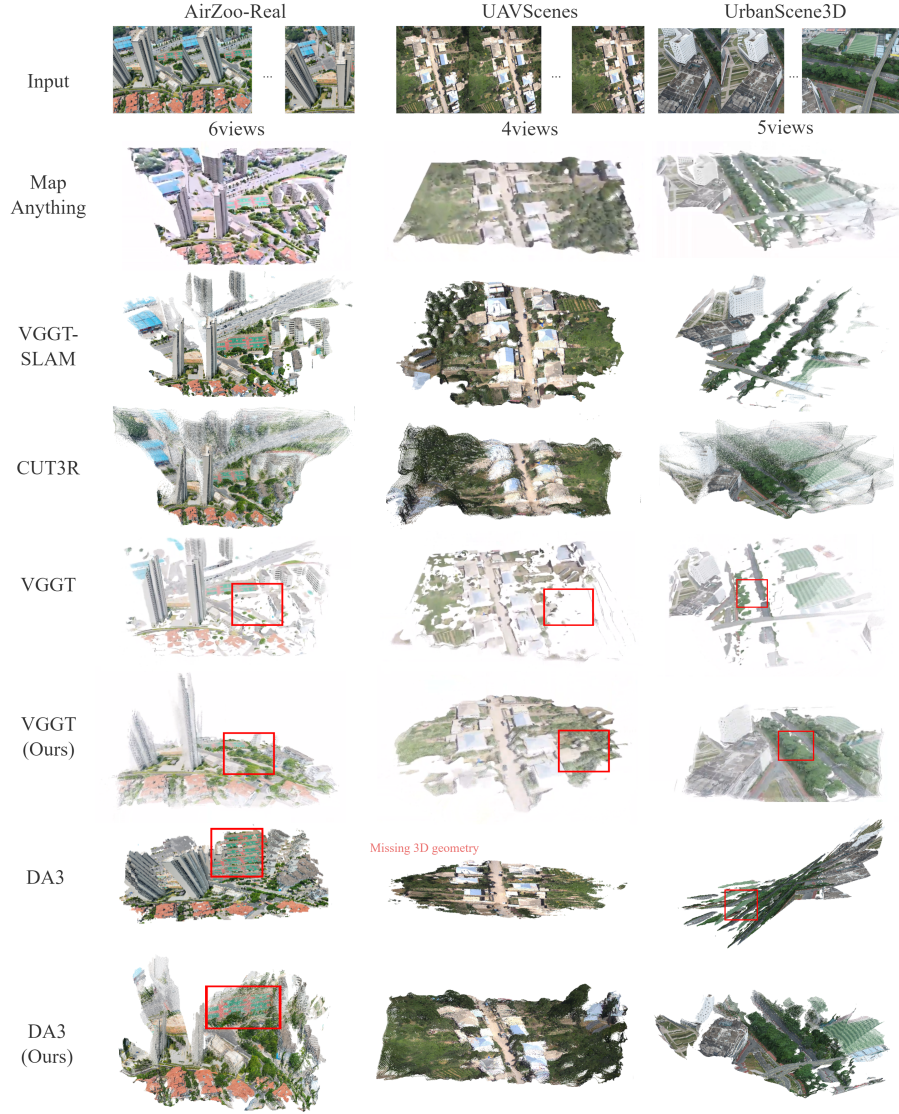


Fig. 9: Qualitative results on AirZoo-Real, UAVScenes and UrbanScene3D. Red boxes highlight missing or erroneous regions in the reconstruction produced by the original model. For each column, the reconstruction result is shown with a side view on the left and a top-down view on the right. Compared with the baselines, our fine-tuned model produces cleaner and more complete geometry.

substantial challenge for real-flight trajectories, where AirZoo training provides notable improvements. We also note that while VGGT-SLAM achieves the best

Table 7: Reconstruction results on AirZoo-Test, AirZoo-Real, UAVScenes and UrbanScene3D datasets. Our AirZoo fine-tuned VGGT and DA3 consistently improve over the original methods across all testsets. Improvements over original methods are shown in gray, and the best results are highlighted.

Method	UAVScenes [48] F1↑	UrbanScene3D [27] F1↑	AirZoo-test F1↑	AirZoo-Real F1↑
CUT3R [46]	83.51	39.64	64.99	48.88
MapAnything [21]	90.30	49.73	83.35	45.00
VGGT-SLAM [31]	86.42	45.32	75.65	53.95
VGGT [45]	86.20	43.38	67.67	42.48
VGGT(Ours)	86.98(+0.78)	52.61(+8.23)	77.32(+9.65)	43.67(+1.19)
DA3-Large [26]	87.38	47.83	54.10	49.97
DA3-Large (Ours)	91.42 (+2.93)	55.91 (+ 8.08)	86.09 (+31.99)	53.07 (+3.10)

performance on AirZoo-Real, this is largely due to its BA-based optimization, giving it an accuracy edge over purely feed-forward baselines. AirZoo provides effective supervision for robust multi-view 3D reconstruction from UAV sequences under challenging viewpoint, illumination, and environmental variations.

5 Conclusion

This paper presents AirZoo, a unified large-scale dataset and benchmark for grounding aerial geometric 3D vision. The proposed benchmark introduces three key contributions namely 1) a scalable generation pipeline that leverages world-scale photogrammetric meshes to render customizable UAV trajectories, 2) comprehensive scene diversity spanning 377 weather-conditioned trajectories across 22 countries, and 3) rich geometric annotations providing synchronized pixel-level metric depth and precise 6-DoF geo-referenced poses. Extensive evaluations across three rigorous geometric tracks demonstrate that AirZoo serves as a powerful pre-training engine, yielding substantial performance gains for state-of-the-art models on real-world benchmarks. We believe this work not only advances the field of aerial geometric perception by bridging the critical gap in high-fidelity training data, but also establishes a new performance upper bound for aerial spatial intelligence.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) Young Scientists Fund (Grant Nos. 62406331 and 62503491). The authors would like to thank Yipeng Liu, Yuxuan Cheng, Wanchang Li, Qing Chang for dataset construction, Xia Li for rendering engine debugging. We sincerely thank [Cesium for Unreal](#), [AirSim](#), and [Google Earth](#) for providing the data platform, simulation framework, and data source.

References

1. Ali-Bey, A., Chaib-Draa, B., Giguere, P.: Gsv-cities: Toward appropriate supervised visual place recognition. *Neurocomputing* **513**, 194–203 (2022)
2. Berton, G., Junglas, L., Zacccone, R., Pollok, T., Caputo, B., Masone, C.: Meshvpr: Citywide visual place recognition using 3d meshes. In: *ECCV* (2024)
3. Berton, G., Masone, C.: Megaloc: One retrieval to place them all. In: *CVPR*. pp. 2861–2867 (2025)
4. Brachmann, E., Cavallari, T., Prisacariu, V.A.: Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In: *CVPR*. pp. 5044–5053 (2023)
5. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: *CVPR*. pp. 1050–1060 (2025)
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: *CVPR*. pp. 5828–5839 (2017)
8. Dai, M., Zheng, E., Feng, Z., Qi, L., Zhuang, J., Yang, W.: Vision-based uav self-positioning in low-altitude urban environments. *IEEE TIP* (2023)
9. Dhaouadi, O., Marin, R., Meier, J., Kaiser, J., Cremers, D.: Ortholoc: Uav 6-dof localization and calibration using orthographic geodata. *arXiv preprint arXiv:2509.18350* (2025)
10. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: *CVPR*. pp. 19790–19800 (2024)
11. Fonder, M., Van Droogenbroeck, M.: Mid-air: A multi-modal dataset for extremely low altitude drone flights. In: *CVPRW*. pp. 0–0 (2019)
12. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. *IJCV* **129**(12), 3313–3337 (2021)
13. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)* (2013)
14. Google: Google Maps. <https://www.google.com/maps> (2026), accessed: 2026-03-01
15. Google: Photorealistic 3D Tiles. <https://developers.google.com/maps/documentation/tile/3d-tiles> (2026), accessed: 2026-06-28
16. 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. *arXiv preprint arXiv:2512.20770* (2025)
17. He, M., Chen, C., Liu, J., Li, C., Lyu, X., Huang, G., Meng, Z.: Aerialvl: A dataset, baseline and algorithm framework for aerial-based visual localization with reference map. *IEEE Robotics and Automation Letters* **9**(10), 8210–8217 (2024)
18. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: *CVPR*. pp. 17658–17668 (2024)
19. Ji, Y., He, B., Tan, Z., Wu, L.: Game4loc: A uav geo-localization benchmark from game data. In: *AAAI* (2025)
20. Keetha, N., Mishra, A., Karhade, J., Jatavallabhula, K.M., Scherer, S., Krishna, M., Garg, S.: Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters* **9**(2), 1286–1293 (2023)

21. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zaus, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414* (2025)
22. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: *ICCV*. pp. 2938–2946 (2015)
23. Li, X., Wang, S., Zhao, Y., Verbeek, J., Kannala, J.: Hierarchical scene coordinate classification and regression for visual localization. In: *CVPR*. pp. 11983–11992 (2020)
24. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In: *ICCV* (2023)
25. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: *CVPR*. pp. 2041–2050 (2018)
26. 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)
27. Lin, L., Liu, Y., Hu, Y., Yan, X., Xie, K., Huang, H.: Capturing, reconstructing, and simulating: the urbanscene3d dataset. In: *ECCV*. pp. 93–109. Springer (2022)
28. Liu, L., Xu, W., Fu, H., Qian, S., Yu, Q., Han, Y., Lu, C.: Akb-48: A real-world articulated object knowledge base. In: *CVPR*. pp. 14809–14818 (2022)
29. Loiseau, T., Bourmaud, G.: Rubik: A structured benchmark for image matching across geometric challenges. In: *CVPR*. pp. 27070–27080 (2025)
30. Luo, X., Wan, X., Gao, Y., Tian, Y., Zhang, W., Shu, L.: Jointloc: A real-time visual localization framework for planetary uavs based on joint relative and absolute pose estimation. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 3348–3355. IEEE (2024)
31. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549* (2025)
32. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In: *CVPR* (2025)
33. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
34. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordon, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: *ICCV* (2021)
35. Rizzoli, G., Barbato, F., Caligiuri, M., Zanuttigh, P.: Syndrone-multi-modal uav dataset for urban scenarios. In: *ICCV*. pp. 2210–2220 (2023)
36. Sarlin, P.E., Unagar, A., Larsson, M., Germain, H., Toft, C., Larsson, V., Pollefeys, M., Lepetit, V., Hammarstrand, L., Kahl, F., et al.: Back to the feature: Learning robust camera localization from pixels to pose. In: *CVPR*. pp. 3247–3257 (2021)
37. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: *CVPR*. pp. 4104–4113 (2016)
38. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: *ECCV*. pp. 501–518. Springer (2016)
39. Shen, X., Cai, Z., Yin, W., Müller, M., Li, Z., Wang, K., Chen, X., Wang, C.: Gim: Learning generalizable image matcher from internet videos. In: *ICLR* (2024)
40. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loft: Detector-free local feature matching with transformers. In: *CVPR*. pp. 8922–8931 (2021)
41. Sun, P., Kretschmar, H., Dotiwala, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: *CVPR*. pp. 2446–2454 (2020)

42. Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N.: Megascenes: Scene-level view synthesis at scale. In: ECCV. pp. 197–214. Springer (2024)
43. Vuong, K., Ghosh, A., Ramanan, D., Narasimhan, S., Tulsiani, S.: Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In: CVPR. pp. 21674–21684 (2025)
44. Wang, F., Jiang, X., Galliani, S., Vogel, C., Pollefeys, M.: Glace: Global local accelerated coordinate encoding. In: CVPR. pp. 21562–21571 (2024)
45. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupperecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
46. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: CVPR. pp. 10510–10522 (2025)
47. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR. pp. 20697–20709 (2024)
48. 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. pp. 28946–28958 (2025)
49. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: CVPR. pp. 21666–21675 (2024)
50. Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., Civera, J.: Mapillary street-level sequences: A dataset for lifelong place recognition. In: CVPR. pp. 2626–2635 (2020)
51. Wu, R., Cheng, X., Zhu, J., Liu, Y., Zhang, M., Yan, S.: Uavd4l: A large-scale dataset for uav 6-dof localization. In: 3DV (2024)
52. Wu, R., Huang, Z., He, X., Yan, L., Shen, Y., Peng, S., Zhou, X., Zhang, M.: Aerialextrematch-localization dataset (2026), <https://huggingface.co/datasets/Xecades/AerialExtreMatch-Localization>
53. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR. pp. 803–814 (June 2023)
54. Xu, W., Yao, Y., Cao, J., Wei, Z., Liu, C., Wang, J., Peng, M.: Uav-visloc: A large-scale dataset for uav visual localization. arXiv (2024)
55. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR. pp. 21924–21935 (2025)
56. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blended-mvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR. pp. 1790–1799 (2020)
57. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV. pp. 12–22 (2023)
58. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
59. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACM MM. pp. 1395–1403 (2020)
60. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. ACM TOG **37** (2018), <https://arxiv.org/abs/1805.09817>

61. Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., Hu, W.: Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. IEEE TCSVT (2023)