Leaving the City: A Large-Scale Aerial Dataset for Cross-Season Localization in Unstructured Environments

Michael Schleiss[1,3], Henry Hölzemann[2], Fahmi Rouatbi[2], Torsten Fiolka[2], Thomas Pany[1], Roger Förstner[1], and Daniel Cremers[3]

[1] Universität der Bundeswehr München, Neubiberg, Germany
[2] Fraunhofer FKIE, Wachtberg, Germany
[3] Technical University of Munich, Munich, Germany

Abstract. Long-term aerial localization requires matching live flight imagery against archival reference maps, demanding feature representations that are invariant to severe appearance changes and perceptual aliasing. However, current benchmarks report only aggregate accuracy over predominantly man-made environments, masking severe terrain-dependent performance gaps. As a result, localization performance over unstructured natural landscapes—where self-similar textures and drastic seasonal changes dominate—remains effectively unmeasured. To address this, we introduce *Leaving the City*, the first large-scale aerial benchmark designed to isolate and quantify terrain-dependent localization gaps. Captured via a microlight aircraft, our dataset comprises 1,379 km of flight trajectories flown repeatedly to capture distinct seasonal variations. We pair high-frame-rate imagery and inertial measurements with semantic terrain masks, multi-year-old orthophotos, and precise 6-DoF ground truth. Evaluating state-of-the-art matchers through our terrain-stratified protocol reveals a systematic bias: methods that succeed on man-made surfaces degrade sharply over natural terrain undergoing strong appearance change. By exposing where current methods fail, our benchmark provides a rigorous foundation for developing robust, all-terrain aerial localization. The dataset and code are publicly available.[4]

Keywords: Visual Localization · Aerial Dataset · Terrain-Aware Evaluation

1 Introduction

Visual localization has made rapid progress across a range of well-studied scenarios: indoor rooms, city streets, road corridors, and landmarks [30]. On these structured scenes — where walls, roads, and tracks provide a persistent geometric anchor — state-of-the-art methods report high accuracy, even under long-term appearance changes caused by season and weather [26, 27, 38]. Because these datasets are

[4] Project page: `https://hlzmnhnry.github.io/publications/ltc`

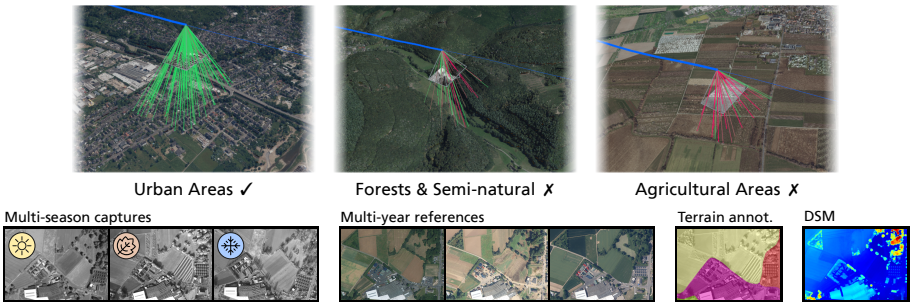


Fig. 1: Our dataset contains over 1.6 million multi-season aerial images with terrain-aware evaluation and annotations, digital surface models, and multi-year reference orthophotos, captured across six microlight flights spanning very large-scale trajectories.

dominated by scenes with man-made structures, their strong results may reflect the reliable geometry of the scene rather than the genuine robustness of the underlying methods. Consequently, localization performance on unstructured natural terrain — such as forests, cropland, and grassland — where such anchors are scarce and visual appearance is less stable, is not yet separately characterized. For aerial visual localization, this evaluation gap is particularly critical. Unlike street-level datasets that are inherently tied to road networks, aerial imagery frequently captures unstructured landscapes. While prior work has recognized that feature matching and place recognition degrade over such natural terrain [31], no existing benchmark provides the means to measure this gap systematically or track progress toward closing it. Existing aerial benchmarks [14, 36] report only aggregate metrics without stratifying by terrain type, making it impossible to detect whether methods that succeed on man-made structures also generalize to natural landscapes.

We introduce *Leaving the City*, a large-scale aerial benchmark that makes this distinction explicit. By assigning semantic land-cover labels to every pixel, our benchmark decomposes feature detection, matching, and localization performance by terrain type at the pixel level — measuring for the first time the effect of underlying land cover (see Figure 1). Our dataset comprises 1,379 km of flight trajectories, captured by a microlight aircraft and reflown across seasons over the same geographic extent — isolating seasonal variation as a controlled variable. Each image is paired with multi-year reference orthophotos, making it possible to disentangle the effects of reference age and terrain type on matching performance. Semantic terrain masks distinguish artificial surfaces, forests, agricultural areas, and water, enabling terrain-stratified evaluation. Sub-meter 6-DoF ground truth is verified against independent reference data.

In summary, our contributions are:

– **A large-scale aerial localization benchmark.** We present *Leaving the City* (LTC), a large-scale benchmark comprising over 1,379 km of repeated

Dataset	Year	Queries	Distance	Terrain	Orthophoto	DSM	6-DoF GT	Multi-season	Multi-year ref.	Terrain annot.	VPR	VL	VIO
University-1652 [41]	2020	90 k	—	●	✓	–	–	–	–	–	✓	–	–
VPAIR [31]	2022	2.7 k	107 km	● ■ ▲ ♦	✓	✓	✓	–	–	–	✓	✓	–
Alto [11]	2022	15 k	410 km	● ■ ▲ ♦	✓	–	✓	–	–	–	✓	–	–
SUES-200 [42]	2023	10 k	—	●	✓	–	–	–	–	–	✓	–	–
DenseUAV [12]	2024	18 k	—	●	✓	–	–	–	✓	–	✓	–	–
UAV-VisLoc [39]	2024	6.7 k	—	● ■ ▲ ♦	✓	–	–	◇	–	–	✓	–	–
AnyVisLoc [40]	2025	18 k	—	● ■ ▲	✓	✓	–	◇	◇	–	✓	✓	–
ViLD [2]	2025	90 k	395 km	■ ▲	✓	–	–	–	–	–	✓	–	–
OrthoLoC [14]	2025	16 k	—	● ■	✓	✓	✓	–	◇	–	–	✓	–
AerialVL [17, 22]	2025	18 k	70 km	● ■	✓	–	▽	–	–	–	✓	✓	✓
LTC *(Ours)*	2026	**1.65 M**	**1379 km**	● ■ ▲ ♦	✓	✓	✓	✓	✓	✓	✓	✓	✓

◇ = variation present but not isolated
▽ = claimed but not evaluated
— = no continuous trajectories
● Artificial ■ Agricultural ▲ Forest ♦ Water

Table 1: Aerial localization datasets with orthophoto map references. Multi-season and multi-year columns imply revisitation of the same geographic area. Terrain annotation implies dense per-pixel land-cover masks.

multi-season flight trajectories with orthophoto references, semantic terrain annotations, and accurate 6-DoF ground truth.

- **A terrain-stratified evaluation protocol.** We introduce an evaluation protocol that decomposes localization performance by semantic terrain class, making terrain-dependent failure measurable and trackable beyond aggregate metrics.
- **A multi-task baseline suite.** We provide benchmark results for visual localization, visual place recognition, and both visual and visual-inertial odometry using representative state-of-the-art methods, establishing baselines for future work.

2 Related Work

Localizing perspective aerial imagery against orthophoto references is a cross-domain problem, bridging the geometric and temporal gap between onboard cameras and orthographic map products. Several established aerial visual-inertial odometry and SLAM benchmarks provide sequential imagery with inertial data but no orthophoto map references [8, 15, 23, 34, 37]. At a larger scale, UAVScenes [36] offers rich multi-sensor trajectory data but evaluates only same-domain matching, while LASED [28] provides extensive temporal coverage with multi-year reference data but no onboard imagery. Table 1 focuses on the cross-domain challenge of localizing perspective imagery against orthophoto and DSM map products, where viewpoint and modality gaps are central.

Several datasets adapt the ground-to-satellite geo-localization paradigm to the aerial-to-orthophoto setting. University-1652 [41], SUES-200 [42], DenseUAV [12],

and UAV-VisLoc [39] formulate the task as image retrieval, learning global embeddings that bridge the domain gap. ViLD [2] stands out with extensive flights over predominantly natural terrain, but remains retrieval-only. These benchmarks provide no metric pose, no continuous trajectories, and limited or no temporal variation.

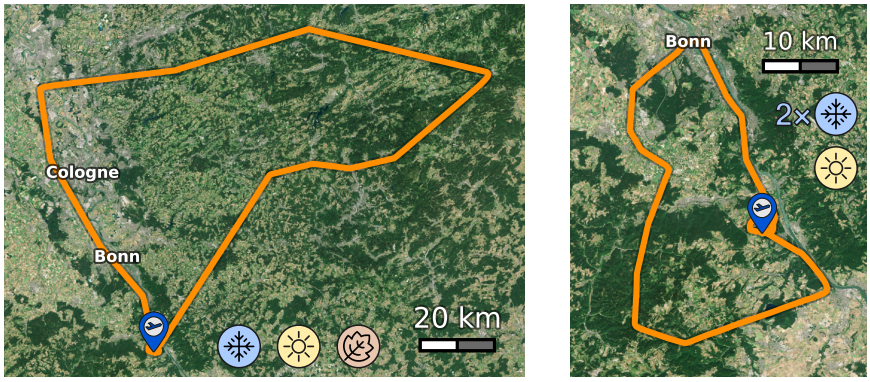
A second group targets 6-DoF pose estimation or trajectory-level evaluation beyond retrieval. VPAIR [31] evaluates a coarse-to-fine pipeline against rendered orthophoto and depth references, including a per-terrain VPR breakdown. AnyVisLoc [40] benchmarks matching against both aerial photogrammetry and satellite references. Most similar to ours, OrthoLoC [14] provides an in-depth analysis of the cross-domain gap between onboard imagery and orthophoto references, but lacks multi-season revisitation, multi-year reference data, and terrain annotations. AerialVL [17, 22] is the only prior dataset to cover retrieval, matching, and VIO, but it does not evaluate full 6-DoF poses and relies on single-point GNSS ground truth. No existing dataset stratifies localization performance by terrain type. In contrast, LTC combines controlled multi-season revisitation, per-frame terrain annotations, and sub-meter 6-DoF ground truth across all three tasks—cross-domain retrieval, metric localization, and visual-inertial odometry—while significantly exceeding existing aerial localization datasets in scale.

3 The *Leaving the City* Dataset

Leaving the City is a long-term aerial localization dataset totaling 1,379 km continuous trajectories, reflown across seasons and paired with orthophoto and DSM references, with per-pixel terrain annotations that enable stratified evaluation by land-cover class. Each flight image is paired with multi-year orthophotos and a digital surface model from publicly available governmental geodata as cross-domain reference targets. Per-pixel terrain annotations decompose the benchmark into four semantic superclasses—artificial surfaces, agricultural areas, forests, and water—so that the effect of seasonal appearance change on localization can be isolated per terrain type, a diagnostic axis no existing aerial dataset supports.

3.1 Campaign Overview

The dataset consists of six campaigns flown along two predefined routes over western Germany by a manually piloted microlight aircraft between February 2022 and January 2024 (Fig. 2). A short route (∼125 km) and a long route (∼335 km) are each reflown across seasons so that seasonal appearance change is the only variable between campaign pairs while the underlying scene geometry remains fixed. The short route covers winter (February 2022, two flights on the same day, one clockwise and the other one counterclockwise) and summer (June 2022); the long route covers summer (June 2022), autumn (October 2022), and winter (January 2024). Both routes cover a mix of urban, agricultural, and forested terrain. At cruise altitudes typically between 300 m and 500 m above ground, with excursions up to 800 m, median re-flight deviations between campaigns remain



(a) Large flight trajectory (>300km) **(b)** Small trajectory (>100km)

Fig. 2: Overview of the flight trajectories with indicators of seasonal coverage.

	Small Trajectory			Large Trajectory		
Dataset Statistic	22-02-23 11h	22-02-23 15h	22-06-10 10h	22-06-14 10h	22-10-19 14h	24-01-29 12h
CLC (●■▲♦) in %	29/26/36/8	29/26/35/10	31/27/33/10	28/29/40/3	28/29/40/3	27/29/41/3
GNSS Types in %	0/100	0/100	0/100	1/99	0/100	2/98
Flight Duration	58 min	58 min	49 min	2 h 8 min	2 h 25 min	2 h 30 min
Trajectory Length	126 km	127 km	117 km	333 km	336 km	340 km
Images (lost images)	154k (2)	153k (10)	125k (0)	365k (1)	417k (0)	432k (0)
ΔPosition XY [m]	*Ref.*	37.2/75.5	21.9/37.3	*Ref.*	19.8/39.3	18.6/23.0
ΔPosition Z [m]	*Ref.*	36.8/39.1	31.2/39.9	*Ref.*	22.9/64.7	28.9/69.6
ΔVelocity XY [m/s]	*Ref.*	8.7/2.7	5.8/2.8	*Ref.*	6.9/2.7	8.9/3.1
ΔVelocity Z [m/s]	*Ref.*	1.0/0.9	1.3/1.2	*Ref.*	0.3/1.9	1.5/1.4

Table 2: Overview of campaign statistics in terms of extent, quality, diversity, and re-flight deviations (median/std) against reference campaigns (22-02-23 11h, 22-06-14 10h). Land cover distributions refer to the four superclasses of the CORINE Land Cover.

below 40 m horizontally (cf. Tab. 2), well within a single-frame footprint, such that corresponding images across seasons largely observe the same scene.

3.2 Data Acquisition

The sensor payload (Fig. 3) is mounted under the wing of a microlight aircraft. A global-shutter grayscale camera (1600×1100 px, 12 mm lens) captures images at 50 Hz, while an inertial navigation system (INS) provides IMU measurements at 200 Hz and RTK-GNSS-based 6-DoF ground truth at 200 Hz. Four campaigns maintained exclusively centimeter-level GNSS solutions (RTK fixed/float), whereas the remaining two contain only short non-RTK GNSS intervals.

All sensors are hardware-triggered from a common GNSS-synchronized clock, achieving sub-millisecond accuracy for more than 99% of frames. Calibrations are refined per campaign following [18]. Full sensor specifications and synchronization details are provided in the supplementary material.

Campaign	Translation error (m)			Rotation error (deg)		
	3D	XY	Z	Roll	Pitch	Yaw
2022-02-23 11h	1.58/2.05	1.38/1.96	0.55/0.89	0.11/0.21	0.12/0.25	0.09/0.09
2022-02-23 15h	2.03/2.02	1.76/1.92	0.55/0.92	0.12/0.18	0.15/0.24	0.14/0.18
2022-06-10 10h	2.22/3.01	1.99/2.90	0.65/1.17	0.12/0.21	0.14/0.23	0.28/0.18
2022-06-14 10h	2.06/2.46	1.62/2.03	0.80/1.67	0.11/0.18	0.10/0.21	0.46/0.34
2022-10-19 14h	2.27/3.15	1.77/2.63	0.99/2.05	0.12/0.25	0.11/0.25	0.49/0.30
2024-01-29 12h	2.16/3.09	1.80/2.73	0.85/1.76	0.11/0.23	0.23/0.28	0.36/0.29

Table 3: Deviations between VGT and INS-based camera poses (median/std).

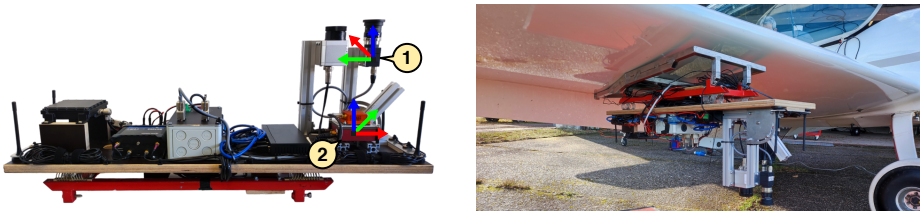
3.3 Reference Data

Each flight image is paired with publicly available governmental orthophotos and a digital surface model (DSM) provided by the state surveying agencies of North Rhine-Westphalia (NRW) and Rhineland-Palatinate (RLP).[5] The orthophotos are delivered as color tiles in UTM 32N at a uniform ground sampling distance of 20 cm (resampled from 10 cm for NRW, native for RLP). Each tile covers four to seven acquisition years between 2011 and 2025. Since the flights span 2022–2024, the temporal offset to the reference data ranges from near zero to over a decade. This temporal variation is intentional rather than removed: it reflects the practical use case of localizing current flight imagery against archival map products. To separate this effect from terrain-dependent performance, we report additional analyses stratified by reference age and season distance. The digital surface models are photogrammetrically derived from the same aerial surveys (bDOM) and provide per-pixel elevation that lifts 2D image correspondences to 3D for pose estimation.

3.4 Ground Truth Verification

The primary ground truth is the RTK-GNSS/INS trajectory at 200 Hz. To verify its accuracy independently, we estimate a separate 6-DoF visual ground-truth

[5] Orthophotos and surface models: © Geobasis NRW, dl-de/zero-2-0; © GeoBasis-DE / LVermGeoRP (2011–2025), dl-de/by-2-0, `www.lvermgeo.rlp.de`.



(a) Sensor setup on a damped platform

(b) Payload mounted under "Dolphin"

Fig. 3: Overview of the sensor payload used for data collection. The main system components are a FLIR BFS camera ('1' in (a)) and an SBG Ellipse2-D INS ('2' in (a)).

trajectory (VGT) from feature correspondences between camera frames and the orthophoto references from Sec. 3.3, lifted to 3D via the bDOM and refined in a factor graph. Median translation differences between VGT and INS-based camera poses are approximately 2 m; rotational deviations remain below 1° (Tab. 3). Because VGT verification is more challenging in unstructured regions, we additionally analyze the navigation system's uncertainty across terrain classes, finding no terrain-correlated increase. A body-frame analysis shows no systematic spatial bias across campaigns. A complementary reprojection test using manually annotated ground control points yields mean errors of 3 px to 5 px across all campaigns.

Details of the VGT pipeline, body-frame error analysis, timestamp synchronization verification, per-terrain accuracy statistics, and reprojection analysis are provided in the supplementary material.

3.5 Terrain Annotations

Per-pixel land-cover masks are derived by projecting Germany's CLC5 land-cover data into each camera frame using the ground-truth poses, calibrated camera geometry, and the precomputed ground footprint. CLC5 follows the CORINE Land Cover (CLC) nomenclature [9], but is generated from Germany's higher-resolution LBM-DE land-cover model [7]. We aggregate the 44 CLC classes into four superclasses: *artificial surfaces*, *agricultural areas*, *forest and semi-natural areas*, and *water bodies*. Per-campaign class frequencies are listed in Tab. 2. These annotations enable a terrain-stratified evaluation protocol. Each query frame is assigned to the superclass with the largest pixel share in its ground footprint. Optionally, a minimum share threshold can be imposed to restrict evaluation to frames with a clear terrain majority, discarding ambiguous boundary regions (cf. Fig. 4). All localization and odometry metrics are then computed per class, isolating and comparing terrain effects across methods.

4 Experimental Evaluation

We evaluate LTC with representative methods for visual and visual-inertial odometry (VO/VIO), visual localization (VL), and visual place recognition (VPR), using terrain-stratified metrics throughout.

4.1 Visual and Visual-Inertial Odometry

To perform a terrain-based evaluation for odometry estimation, we evaluate two representative visual-inertial pipelines: VINS-Mono [29] (referred to as 'VINS'), which relies on KLT-based sparse feature tracking, and ORB-SLAM [10], which performs descriptor-based feature matching. For pure VO, we evaluate ORB-SLAM as a classical feature-based method and DROID-SLAM [33] as a modern learning-based approach.

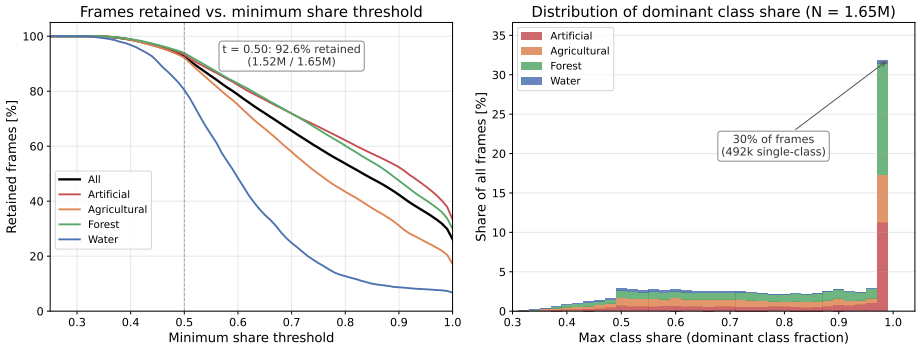


Fig. 4: Sensitivity of the terrain stratification to the minimum share threshold. **Left:** Fraction of frames retained per terrain class as the threshold increases. At t=0.50, 92.6% of all frames are retained (1.52 M of 1.65 M). **Right:** Distribution of the dominant class share across all frames. 30% of frames (492 k) are pure single-class ($\geq$0.99), confirming that the majority of the dataset has well-separated terrain coverage.

T. Init.	Method	2022-02-23 11h					2022-02-23 15h					2022-06-10 10h				
		ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.	ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.	ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.
VIO Ground	VINS-Mono [29]	281.5	1.0	0.04	0.5	100	433.3	0.5	0.05	0.5	100	529.3	1.0	0.04	0.5	100
VIO In flight	VINS-Mono [29]	271.5	0.6	0.05	0.5	100	306.8	0.6	0.02	0.7	100	311.5	0.9	0.04	0.5	100
VIO In flight	ORB-SLAM [10]	884.9	1.8	0.05	1.9	100	772.1	3.5	0.04	2.7	100	1338.1	2.7	0.05	4.0	100
VO In flight	ORB-SLAM [10]	435.2	1.1	2.13	56.3	78	600.4	1.4	1.74	56.5	80	490.2	1.6	1.78	62.9	82
VO In flight	DROID [33]	1268.6	4.7	2.08	56.8	78	1691.7	3.6	1.76	57.1	80	287.9	0.8	1.79	64.4	82
T. Init.	Method	2022-06-14 10h					2022-10-19 14h					2024-01-29 12h				
		ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.	ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.	ATE	MPD	ØRPE$_r$	ØRPE$_t$	Cov.
VIO Ground	VINS-Mono [29]	6834.5	5.6	0.03	0.9	100	4578.4	3.3	0.02	1.1	100	3923.1	3.3	0.05	2.9	100
VIO In flight	VINS-Mono [29]	6733.4	5.6	0.02	0.9	100	3831.6	2.8	0.02	0.7	100	2675.3	2.3	0.03	4.5	100
VIO In flight	ORB-SLAM [10]	4974.2	4.1	0.04	5.3	100	6955.8	5.2	0.02	5.1	100	18578.3	10.3	0.04	16.4	100
VO In flight	ORB-SLAM [10]	10568.2	8.8	1.86	61.8	100	4199.5	3.7	1.00	55.6	100	3875.6	2.7	0.75	55.7	100
VO In flight	DROID [33]	7439.5	10.6	1.83	68.1	51	4961.8	4.7	1.01	55.8	100	10425.8	10.4	0.69	55.1	100

Table 4: Campaign-wise quantitative evaluation of visual and visual-inertial odometry on the LTC dataset. The upper and lower blocks correspond to short and long flights, respectively. Coverage below 100% indicates premature estimation termination.

Experimental Setup To ensure fair comparability, all methods are executed on the campaigns with identical start and end points. All experiments are conducted using images with 25 Hz, and IMU data with 200 Hz, in the case of VIO. We further distinguish between initialization on the ground and in flight since recordings during takeoff exhibit severe motion blur, which severely complicates both initialization and continued state estimation.

For in-flight analysis, we adapt visual-inertial initialization of both ORB-SLAM and VINS based on ground-truth information, since at high velocities combined with insufficiently rich motion patterns, nonlinear system parameters for initialization become poorly observable. Detailed justification of these decisions, and an outline of the changes made, is provided in the supplementary material.

Evaluation Metrics We report the absolute trajectory error (ATE) as the RMSE of the translational absolute pose error (APE). In addition, we report the median relative pose error (RPE) for both translation and rotation over a one-second interval [16]. Further, we evaluate the maximum positional deviation (MPD), defined as $\mathrm{APE}_{max}/|T|$, relating maximum APE to trajectory length $|T|$. Before evaluation, all estimates are aligned to ground truth using Umeyama alignment [35]. For VIO, we apply $SE(3)$ alignment; for VO, $Sim(3)$ is used.

If a method does not produce estimates for a full sequence, metrics are reported up to the last valid estimate. Further, we report the coverage, i.e., the ratio between estimated poses and possible frame poses. Campaign-level baseline results are given in Tab. 4 and form the basis for the terrain-stratified evaluation.

Experimental Results Neither ORB-SLAM (VO/VIO) nor DROID-SLAM (VO) was able to initialize on the ground or to maintain tracking during takeoff. In contrast, VINS successfully initialized on the ground and remained operational. Consequently, for VINS we report ground and in-flight initialization results; for ORB- and DROID-SLAM, only in-flight results are provided.

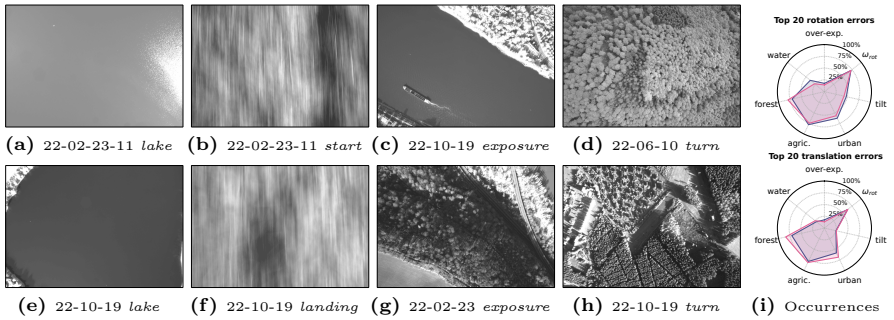
VINS outperforms ORB-SLAM in ATE and MPD in the majority of cases. Major differences appear in RPE_t: VINS achieves median values below 1.5 m in nearly all sequences, whereas ORB-SLAM yields stronger short-term drift accumulation. For VINS, in-flight initialization yields slightly lower ATE than ground initialization in most campaigns. However, differences remain moderate, confirming robust estimation even after substantial motion blur during takeoff.

In the VO setting, ORB-SLAM outperforms DROID-SLAM in most campaigns. Notably, ORB-SLAM often achieves lower ATE than its VIO variant. While the comparison is not fair due to $Sim(3)$ alignment, it indicates sensitivity of VIO to initialization quality and inertial parameter estimation in our large-scale aerial setting. Rotational drift is significantly larger in VO than in VIO, highlighting the stabilizing role of inertial data for attitude estimation.

Impact of Trajectory Scale. Comparing small and large trajectory estimates clearly reveals scale effects. On the small trajectory, VINS achieves MPD values $\leq 1\%$, while they increase substantially (2%–6%) for the long trajectory, reflecting cumulative bias effects for extended flights. Large-scale sequences further expose extreme metric values in specific campaigns, demonstrating phenomena that are not visible in short benchmark sequences, such as gradual observability degradation, environmental homogeneity, and long-term bias evolution.

Open Challenges. The dataset demonstrates strong performance of tightly coupled VIO methods on moderate scales. At the same time, it exposes fundamental challenges in large-scale aerial scenarios in unstructured environments:

1. **Robust initialization for fixed-wing aerial platforms:** Reliable VIO initialization remains challenging for fixed-wing aircraft due to high forward velocities and limited motion diversity, which weaken the observability of scale, velocity, and gravity.



(a) 22-02-23-11 *lake* **(b)** 22-02-23-11 *start* **(c)** 22-10-19 *exposure* **(d)** 22-06-10 *turn*

(e) 22-10-19 *lake* **(f)** 22-10-19 *landing* **(g)** 22-02-23 *exposure* **(h)** 22-10-19 *turn* **(i)** Occurrences

Fig. 5: Overview of typical visual failure cases for VO and VIO, next to an accumulated analysis of top 20 failure cases across campaigns for ■ VINS-Mono and ■ ORB-SLAM.

2. **Very large-scale drift mitigation:** While odometry methods perform well on short sequences, very long trajectories expose cumulative bias effects and structural limitations. Drift-reduction strategies that do not depend on loop closure are essential for realistic operational large-scale deployment.
3. **Robust VO in natural terrain:** Existing approaches are implicitly tuned for structured scenes. Robust navigation in unstructured natural terrain remains underexplored, despite its importance for long-range aerial operation beyond built environments.

Failure Case Analysis To contextualize high errors and extreme metric values observed in the preceding section, we perform a systematic analysis of recurring failure patterns. For each in-flight VIO estimate, rotational and translational RPE are aggregated per minute. For every estimate, the top 20 highest-error minute segments are selected and collected across campaigns.

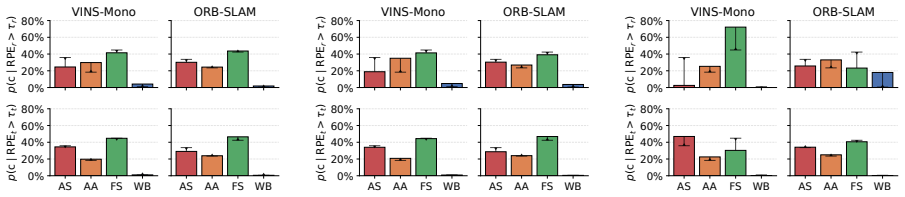
Within each segment, we evaluate the presence of four factors: (i) terrain type, (ii) off-nadir viewing (tilt $> 10°$), (iii) high angular velocity ($|\omega| > 0.15\,\mathrm{rad/s}$), and (iv) partial overexposure (>10% of pixels with intensity > 245). The terrain type is derived from the pixel-level CLC masks and a class is considered present if it covers at least 20% of pixels in any frame of the segment. The resulting occurrence statistics are visualized in Fig. 5i, with examples shown in Fig. 5a–Fig. 5h.

Terrain-dependent Effects. Across both rotational and translational RPE, urban, forest, and agricultural regions dominate the high-error segments, largely reflecting their overall prevalence in the dataset (cf. Tab. 2). Water bodies, however, are clearly overrepresented among failure cases, particularly for rotational errors and for VINS. For ORB-SLAM, forest regions occur disproportionately often, consistent with repetitive, self-similar vegetation structures that yield many detectable features but less distinctive correspondences.

Photometric and motion-related Factors. Overexposure appears in approximately 15–20% of the top-error segments and affects both KLT-based tracking and

M.	Campaign	Artificial surfaces				Agricultural areas				Forest + semi-natural				Water bodies			
		#feat	∅TL	∅Den.	∅NN	#feat	∅TL	∅Den.	∅NN	#feat	∅TL	∅Den.	∅NN	#feat	∅TL	∅Den.	∅NN
VINS-Mono [29]	2022-02-23 11h	293 k	47.0	421.2	36.9	801 k	10.6	225.5	37.4	348 k	40.7	329.2	37.7	120 k	8.5	81.9	39.7
	2022-02-23 15h	235 k	56.2	394.8	37.2	626 k	13.8	228.8	37.0	303 k	46.6	349.9	36.6	48 k	35.3	110.7	39.8
	2022-06-10 10h	254 k	45.1	398.6	36.9	711 k	9.3	195.1	38.0	270 k	42.4	353.2	37.8	96 k	16.5	116.3	37.6
	2022-06-14 10h	713 k	44.4	421.0	36.9	751 k	24.3	171.9	38.1	835 k	47.2	331.7	37.8	65 k	19.3	176.0	37.4
	2022-10-19 14h	518 k	65.6	392.9	37.4	566 k	35.6	167.2	37.9	533 k	88.2	366.4	37.3	31 k	32.7	127.7	38.8
	2024-01-29 12h	519 k	69.8	409.3	37.0	554 k	42.3	188.4	38.3	485 k	96.4	313.5	37.2	20 k	46.7	116.0	39.6
	Average	422 k	55.4	406.8	37.1	668 k	21.3	186.0	37.9	462 k	62.3	338.3	37.4	63 k	19.8	124.3	38.8
		#MPs	∅Ob.	∅Den.	∅NN	#MPs	∅Ob.	∅Den.	∅NN	#MPs	∅Ob.	∅Den.	∅NN	#MPs	∅Ob.	∅Den.	∅NN
ORB-SLAM [10]	2022-02-23 11h	412 k	36.0	411.5	10.4	343 k	24.3	256.4	14.4	517 k	23.9	283.3	13.4	19 k	27.0	74.5	10.7
	2022-02-23 15h	448 k	36.1	424.7	9.7	370 k	25.4	281.1	13.7	547 k	25.6	345.1	12.1	38 k	29.6	101.2	14.4
	2022-06-10 10h	389 k	34.1	415.9	10.6	294 k	23.4	246.3	15.2	478 k	23.3	341.7	13.8	33 k	28.6	110.4	11.1
	2022-06-14 10h	801 k	33.8	303.2	12.3	583 k	29.5	179.0	15.6	726 k	29.6	200.8	17.9	34 k	29.1	143.5	12.2
	2022-10-19 14h	968 k	37.1	361.8	10.8	727 k	33.0	220.6	14.0	1139 k	36.2	330.4	13.7	36 k	33.3	176.0	10.1
	2024-01-29 12h	890 k	39.2	351.2	10.8	747 k	35.6	229.4	14.4	959 k	41.1	261.7	14.3	27 k	35.7	139.1	11.0
	Average	651 k	36.4	360.6	10.9	511 k	30.1	223.2	14.5	728 k	32.0	278.8	14.4	31 k	30.7	130.3	11.7

Table 5: Terrain-stratified evaluation of visual-inertial feature tracks per flight campaign using the baseline trajectories reported in Tab. 4, including campaign-wise averages.



(a) $\tau_r = 0.1°, \tau_t = 1m$ **(b)** $\tau_r = 0.2°, \tau_t = 2m$ **(c)** $\tau_r = 0.5°, \tau_t = 5m$

Fig. 6: Frequency of dominant terrain classes conditioned on RPE thresholds for ■ Artificial surfaces, ■ Agricultural areas, ■ Forest + semi-nat., and ■ Water bodies. Black markers indicate the unconditioned terrain distribution for each method.

descriptor matching by removing discriminative image content (Figs. 5c and 5g). High angular velocity frequently coincides with both rotational and translational errors, while off-nadir tilt correlates with increased RPE_r (Figs. 5d and 5h).

Systematic Breakdown Scenarios. Recurring VO failures on the small trajectory (cf. coverage in Tab. 4) can be traced to a geographically consistent scene: a lake overflight (Fig. 5a). During this segment, the image shows only water for more than 30 seconds, combining low texture, specular reflections, and partial saturation. Under these conditions, ORB-SLAM and DROID-SLAM lose tracking.

Terrain-based Evaluation To analyze how environmental structure affects large-scale aerial odometry, we associate each tracked feature (VINS) and reconstructed map point (ORB-SLAM) with one of four superclasses from terrain annotation. For this, feature locations and IDs are maintained within every run.

In Tab. 5 we report per campaign and class: (i) total number of entities (#feat / #MPs), (ii) average persistence (∅TL / ∅Ob.), (iii) class-wise feature density per megapixel (∅Den.), and (iv) mean nearest-neighbor distance (∅NN).

Terrain-based Feature Density. Artificial surfaces yield the highest feature densities across terrains, whereas water bodies yield the lowest. This reflects underlying image statistics: urban scenes provide abundant high-contrast edges and corners, while water surfaces are locally homogeneous and affected by reflections. Forests show intermediate densities: vegetation produces numerous gradients but less geometrically stable structure. Agricultural areas typically exhibit lower densities than urban scenes due to large homogeneous regions mixed with repetitive fine-scale texture such as furrows and crops.

Track Persistence vs. Matchability. Distinctive differences emerge in persistence behavior. For VINS, forests often yield long KLT track lengths, in some cases exceeding those in artificial surfaces. Since KLT relies on short-term photometric consistency, locally stable gradients can sustain tracks despite global ambiguity.

In contrast, ORB-SLAM exhibits fewer observations per map point in forests than in artificial areas, despite similar detection rates. This indicates that vegetation produces many detectable points but fewer reliably re-observable correspondences. Repetitive, self-similar structures and viewpoint-dependent appearance changes limit descriptor distinctiveness. Thus, forests can be *feature-rich* yet comparatively *match-poor* for long-horizon association.

Agricultural regions show a complementary effect. For VINS, mean track lengths are noticeably shorter, consistent with local ambiguity and fine-scale repetitive patterns. ORB-SLAM shows a comparable degradation, though typically less severe due to map-point selection and outlier rejection. Overall, optical-flow-based methods are primarily constrained by short-horizon track stability in low-structure regions, whereas descriptor-based feature-matching methods are more sensitive to long-horizon feature distinctiveness.

Error-conditioned Terrain Distributions. Beyond structural statistics, we examine the relationship between terrain composition and RPEs. Each frame is assigned to its RPE interval, and a *dominant terrain class* is defined as the superclass contributing the largest feature share in that frame. We compute class frequencies conditioned on increasing rotational and translational RPE thresholds and compare them to the unconditioned reference distribution (cf. Fig. 6).

For rotational errors, artificial surfaces become underrepresented as the threshold increases, whereas water bodies and agricultural areas are increasingly overrepresented. This indicates that the presence of structured urban imagery significantly prevents large rotational errors. In contrast, water bodies provide insufficient visual constraints, while agricultural regions introduce local ambiguity.

For translational errors, terrain-error correlations are weaker and, for our dataset, more sequence-dependent. However, across moderate thresholds, unstructured classes remain overrepresented, confirming that degraded visual constraints increase the likelihood of short-term translational error spikes.

Implications. The terrain-based evaluation reveals complementary failure characteristics of optical-flow-based tracking and descriptor-based matching under realistic aerial conditions. While both methods benefit strongly from structured

urban imagery, visually homogeneous (water) and repetitive (forest, agricultural) regions are disproportionately associated with increased error segments. These findings indicate that odometry performance is strongly conditioned on the underlying terrain distribution. Benchmarks dominated by structured urban imagery may therefore substantially overestimate real-world robustness, as large portions of natural environments provide weaker and more ambiguous visual constraints. The LTC dataset therefore exposes not only geometric drift phenomena, but also terrain-dependent limitations of visual front-ends for odometry estimation.

4.2 Visual Localization

How does terrain type affect cross-domain matching between perspective aerial imagery and orthophoto references? The VIO evaluation above reveals terrain-dependent differences in feature tracking; here we test whether the same bias persists—and intensifies—when matching across the domain gap to archival orthophotos under seasonal change. We evaluate four state-of-the-art matchers through the terrain-stratified protocol introduced in Sec. 3.5.

Experimental Setup For every sampled query frame, a north-aligned orthophoto crop covering the camera ground footprint is extracted from the governmental reference tiles (Sec. 3.3). Both query and reference are resized to the longest edge of 800 px before matching. 2D correspondences are lifted to 3D via the co-registered digital surface model, and a 6-DoF pose is estimated with PnP inside a RANSAC loop (`PoseLib` [21], reproj. threshold 12 px, max. 10000 iterations). A pose is accepted when at least 20 inliers remain after RANSAC. Orthophotos are north-aligned, but the aircraft heading varies freely; for methods that are not rotation-invariant, each query is therefore matched at four candidate rotations (0°, 90°, 180°, 270°) and the rotation yielding the most inliers is kept. Sampling every 50th frame yields ∼33k queries across all six campaigns.

Four state-of-the-art feature matchers serve as localization backbones: SIFT [25] + LightGlue [24], SuperPoint [13] + LightGlue (SP+LG), LoFTR [32], and SE2-LoFTR [6] using `vismatch` [3]. Performance is measured by recall at coupled (translation, rotation) thresholds of $(1\text{m}, 1°)$, $(2\text{m}, 2°)$, and $(5\text{m}, 5°)$, and by the median translation and rotation error over successfully localized queries.

Terrain-Based VL Results Tab. 6 decomposes localization performance by terrain class, pooled across all campaigns. A consistent ordering emerges across all matchers: artificial surfaces yield the highest recall, followed by forests, agricultural areas, and water bodies. SE2-LoFTR achieves the best overall recall ($R_{5/5}$=34.4%), roughly doubling LoFTR (18.8%) and substantially outperforming both sparse methods. Its rotation-equivariant architecture appears particularly effective at bridging the heading-dependent domain gap between perspective queries and north-aligned orthophotos.

The terrain gap, however, persists regardless of method. Even SE2-LoFTR drops from 66% recall on artificial surfaces to 33% on forests and 9% on agricultural areas. Agricultural terrain is the most challenging class: all four matchers

Method	$R_{1/1}$ [%]				$R_{2/2}$ [%]				$R_{5/5}$ [%]				All	Med. e_t [m]			
	●	■	▲	◆	●	■	▲	◆	●	■	▲	◆	$R_{5/5}$	●	■	▲	◆
SIFT+LG	0.0	0.0	0.0	0.0	5.0	**5.3**	0.0	0.0	45.0	5.3	16.7	0.0	19.2	4.7	32.1	**3.8**	71.6
SP+LG	**5.1**	0.0	0.0	**3.3**	15.3	3.8	1.3	3.3	52.5	7.5	22.7	3.3	24.3	3.9	9.9	5.5	14.4
LoFTR	1.7	0.0	0.0	0.0	11.9	1.9	1.3	0.0	32.2	9.3	18.7	10.0	18.8	5.9	**8.1**	7.6	12.4
SE2-LoFTR	5.0	0.0	**1.3**	0.0	**25.4**	0.0	**5.3**	**6.7**	**66.1**	**9.4**	**33.3**	**20.0**	**34.4**	**3.1**	11.7	4.9	**6.2**

Table 6: Terrain-stratified VL results, aggregated across all campaigns. Each query is assigned to the terrain superclass covering the largest footprint share. $R_{i/j}$: recall at $(i\,\mathrm{m}, j^\circ)$. Columns: ■ Artificial, ■ Agricultural, ■ Forest + semi-nat., ■ Water bodies.

achieve near-zero $R_{1/1}$, and median translation errors exceed 10 m for three of four methods. Water bodies show the most erratic behavior, with the fewest queries and the widest spread in median errors.

At tight thresholds ($R_{1/1}$), only artificial surfaces yield non-trivial recall, indicating that sub-meter localization is currently limited to scenes with strong geometric structure. Dense matchers (LoFTR, SE2-LoFTR) do not close the terrain gap at these tight thresholds—they improve absolute recall across all classes but leave the relative ordering unchanged.

To verify that the terrain gap is not an artifact of preprocessing or temporal bias, we run additional controls using SP+LG on the 2022-02-23 11h campaign. Across image scale, reference age, season distance, and same-day illumination, the artificial-to-natural ordering remains stable: these factors affect absolute success rates, but do not remove the terrain-dependent gap. Full results and protocol details are provided in the supplementary material.

4.3 Visual Place Recognition

To complement the metric visual localization experiments, we evaluate cross-domain visual place recognition, framing the task as image retrieval across the perspective-to-orthophoto domain gap.

Experimental Setup For every sampled query frame (every 50th frame, ∼33k across all six campaigns), a north-aligned orthophoto crop covering the camera ground footprint is extracted at the query's ground-truth position—the same crops used for visual localization. Orthophoto crops form the *database* and perspective aerial frames the *queries*; positives are defined by GNSS proximity within 25 m.

We evaluate four pretrained VPR methods in a zero-shot setting (no fine-tuning on LTC data): CosPlace [4], MixVPR [1], AnyLoc [20] (DINOv2 backbone), and SALAD [19] (DINOv2 backbone). All images are resized and center-cropped to 224×224 px following standard practice [5]. Performance is measured by Recall@K (K=1, 5, 10, 20). Each query is assigned to its dominant terrain superclass, enabling terrain-stratified evaluation.

Method	Recall@1 [%]					Recall@5 [%]					Recall@10 [%]					Recall@20 [%]				
	●	■	▲	◆	All	●	■	▲	◆	All	●	■	▲	◆	All	●	■	▲	◆	All
CosPlace	5.7	6.7	3.1	12.2	5.4	10.6	11.4	6.9	27.2	10.4	13.8	14.4	9.1	34.4	13.4	18.0	18.2	12.3	41.1	17.3
MixVPR	8.1	8.6	3.2	18.5	7.0	16.5	13.5	6.1	31.7	12.6	21.0	15.6	7.9	36.0	15.4	25.9	18.3	10.2	41.0	18.7
AnyLoc	**25.8**	**22.0**	**11.3**	**27.7**	**19.2**	**57.6**	**49.4**	**28.9**	**61.1**	**44.4**	**69.8**	**59.9**	**38.4**	**71.9**	**55.1**	**79.9**	**69.8**	**48.2**	**80.6**	**64.9**
SALAD	16.2	11.3	3.9	17.2	10.2	36.5	21.8	10.1	34.5	22.2	46.1	27.5	14.2	41.9	28.4	55.3	34.5	20.2	50.9	35.8

Table 7: Terrain-stratified cross-domain VPR (aerial query → orthophoto DB). All methods are evaluated zero-shot with pretrained weights. AnyLoc uses aerial-domain VLAD centers. Columns: ■ Artificial, ■ Agricultural, ■ Forest + semi-nat., ■ Water.

Results The results in Tab. 7 reveal a clear performance hierarchy. AnyLoc, leveraging DINOv2 features with aerial-domain VLAD cluster centers, substantially outperforms all other methods, achieving an overall Recall@1 of 19.2% compared to 10.2% for the second-best method, SALAD. CosPlace and MixVPR, both trained on ground-level street-view imagery, struggle with the aerial-to-orthophoto domain gap and remain below 8% Recall@1. This confirms that foundation-model features (DINOv2) generalize more robustly across the perspective-to-orthophoto domain gap than representations learned from ground-level data. Forest and semi-natural areas yield the lowest recall for all methods, mirroring the metric visual localization results. Artificial surfaces achieve comparatively higher retrieval rates, likely due to more salient and unique visual features such as buildings and road junctions.

5 Limitations and Conclusions

Due to the characteristics of the underlying microlight platform, the recorded motion profiles do not include rapid multi-axis maneuvers, resulting in less diverse motion patterns. The terrain annotations are projected land-cover labels rather than image-derived semantic segmentation; small boundary inaccuracies may still occur.

Terrain type is a dominant factor in aerial localization failure: across all methods and tasks evaluated on LTC, natural terrain falls dramatically short of artificial surfaces, a gap that aggregate metrics conceal. This bias is not confined to reference map-based localization. In visual-inertial odometry, forests produce abundant features yet fewer reliably re-observable correspondences (Sec. 4.1). In cross-domain visual localization, the same pattern intensifies: even the best-performing matcher (SE2-LoFTR) sees recall at $(5\text{m}, 5°)$ drop from 66% on artificial surfaces to 33% on forests and 9% on agricultural areas (Tab. 6). The consistency of these findings across tracking-based odometry and descriptor-based cross-domain matching points to a shared root cause: visual features on natural terrain lack the geometric distinctiveness needed for robust correspondence under appearance change.

References

1. Ali-bey, A., Chaib-draa, B., Giguère, P.: MixVPR: Feature mixing for visual place recognition. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2998–3007 (2023)
2. Amadei, T., Meinhardt-Llopis, E., Bascle, B., Abgrall, C., Facciolo, G.: Beyond Paired Data: Self-supervised UAV geo-localization from reference imagery alone. arXiv preprint arXiv:2512.02737 (2025)
3. Berton, G., Goletto, G., Trivigno, G., Stoken, A., Caputo, B., Masone, C.: EarthMatch: Iterative coregistration for fine-grained localization of astronaut photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2024)
4. Berton, G., Masone, C., Caputo, B.: Rethinking visual geo-localization for large-scale applications. In: CVPR. pp. 4878–4888 (2022)
5. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: CVPR. pp. 5396–5407 (2022)
6. Bökman, G., Edstedt, J., Felsberg, M., Kahl, F.: Steerers: A framework for rotation equivariant keypoint descriptors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4917–4926 (2024)
7. Bundesamt für Kartographie und Geodäsie: WFS CORINE Land Cover 5 ha, Stand 2018 (WFS CLC5-2018) (2021), https://gdz.bkg.bund.de/index.php/default/wfs-corine-land-cover-5-ha-stand-2018-wfs-clc5-2018.html, accessed: 2026-06-30
8. Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The EuRoC micro aerial vehicle datasets. The International Journal of Robotics Research **35**(10), 1157–1163 (2016)
9. Büttner, G.: Corine land cover and land cover change products. In: Land use and land cover mapping in Europe: practices & trends, pp. 55–74. Springer (2014)
10. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)
11. Cisneros, I., Yin, P., Zhang, J., Choset, H., Scherer, S.: Alto: A large-scale dataset for UAV visual place recognition and localization. arXiv preprint arXiv:2207.12317 (2022)
12. Dai, M., Hu, J., Zhuang, E., Zheng, H.: Vision-based UAV self-positioning in low-altitude urban environments. IEEE Trans. Image Process. **33**, 493–508 (2024)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 224–236 (2018)
14. 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)
15. Fonder, M., Van Droogenbroeck, M.: Mid-Air: A multi-modal dataset for extremely low altitude drone flights. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 553–562 (2019)
16. Grupp, M.: evo: Python package for the evaluation of odometry and SLAM. https://github.com/MichaelGrupp/evo (2017), accessed: 2026-06-30
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). https://doi.org/10.1109/LRA.2024.3441491

18. Hölzemann, H., Schleiss, M.: SCAR: Satellite imagery-based calibration for aerial recordings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 987–996 (2026)
19. Izquierdo, S., Civera, J.: Optimal transport aggregation for visual place recognition. In: CVPR. pp. 17658–17668 (2024)
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 (2024). https://doi.org/10.1109/LRA.2023.3343602
21. Larsson, V., contributors: PoseLib - Minimal Solvers for Camera Pose Estimation (2020), https://github.com/vlarsson/PoseLib, accessed: 2026-06-30
22. Li, C., He, M., Chen, C., Liu, J., Lyu, X., Huang, G., Meng, Z.: GeoVINS: Geographic-visual-inertial navigation system for large-scale drift-free aerial state estimation. IEEE Transactions on Robotics **41**, 5835–5853 (2025). https://doi.org/10.1109/TRO.2025.3613467
23. 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)
24. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17627–17638 (2023)
25. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
26. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017). https://doi.org/10.1177/0278364916679498
27. Olid, D., Fácil, J.M., Civera, J.: Single-view place recognition under seasonal changes. arXiv preprint arXiv:1808.06516 (2018)
28. Papapetros, I.T., Kansizoglou, I., Gasteratos, A.: Visual place recognition for large-scale UAV applications. arXiv preprint arXiv:2507.15089 (2025)
29. Qin, T., Li, P., Shen, S.: VINS-Mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018)
30. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., Kahl, F., Pajdla, T.: Benchmarking 6DoF outdoor visual localization in changing conditions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8601–8610 (Jun 2018). https://doi.org/10.1109/CVPR.2018.00897
31. Schleiss, M., Rouatbi, F., Cremers, D.: VPAIR – aerial visual place recognition and localization in large-scale outdoor environments. arXiv preprint arXiv:2205.11567 (2022)
32. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR. pp. 8922–8931 (2021)
33. Teed, Z., Deng, J.: DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. Advances in Neural Information Processing Systems **34**, 16558–16569 (2021)
34. 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)

35. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(4), 376–380 (2002)
36. Wang, S., Li, S., Zhang, Y., Yu, S., Yuan, S., She, R., Guo, Q., Zheng, J., Howe, O.K., Chandra, L., Srijeyan, S., Sivadas, A., Aggarwal, T., Liu, H., Zhang, H., Chen, C., Jiang, J., Xie, L., Tay, W.P.: UAVScenes: A multi-modal dataset for UAVs. arXiv preprint arXiv:2507.22412 (2025)
37. 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: 2020 IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS). pp. 4909–4916 (2020). https://doi.org/10.1109/IROS45743.2020.9341801
38. Wenzel, P., Yang, N., Wang, R., Zeller, N., Cremers, D.: 4Seasons: Benchmarking visual SLAM and long-term localization for autonomous driving in challenging conditions. International Journal of Computer Vision (2024). https://doi.org/10.1007/s11263-024-02230-4
39. 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 preprint arXiv:2405.11936 (2024)
40. Ye, Y., Teng, X., Chen, S., Liu, L., Wang, K., Song, X., Li, Z.: Exploring the best way for UAV visual localization under low-altitude multi-view observation condition: A benchmark. arXiv preprint arXiv:2503.10692v2 (2025)
41. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACM Int. Conf. Multimedia. pp. 1395–1403 (2020)
42. Zhu, R., Yang, M., Yin, L., Wu, F., Yang, Y.: SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. IEEE Trans. Circuits Syst. Video Technol. **33**(9), 4825–4839 (2023)