OrthoTrack: Continuous 6-DoF UAV Trajectory Estimation Anchored in Public Orthophotos

Oussema Dhaouadi[1,2,3,4,⋆], Zuria Bauer[1], Johannes Meier[2,4]
Olaf Wysocki[3], Marc Pollefeys[1,5], and Daniel Cremers[2,4]

[1] ETH Zurich [2] TU Munich [3] University of Cambridge [4] MCML [5] Microsoft
https://orthotrack.ethz.ch

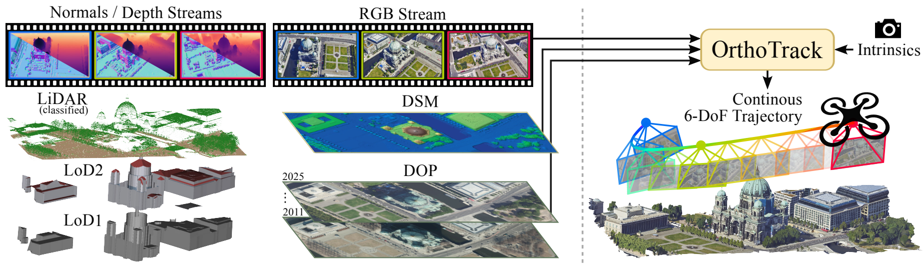
Abstract. Continuous 6-DoF pose estimation is essential for autonomous UAV operations. Yet, existing visual odometry and SLAM methods accumulate drift and yield only relative, up-to-scale trajectories. Single-frame geo-localization, in turn, discards temporal continuity and remains too slow for real-time use. We present OrthoTrack, a training-free system that estimates continuous 6-DoF UAV trajectories using only publicly available orthophotos and surface models as a map prior. OrthoTrack matches keyframes against the orthophoto and lifts correspondences to metric 3D via the surface model. It then propagates these map-anchored correspondences to intermediate frames with optical flow, producing absolute, metrically scaled poses at every frame without GPS or post-hoc alignment. We also introduce the MovingDrone Dataset, a large-scale benchmark pairing photorealistic UAV sequences with dense 6-DoF ground truth and co-registered multi-modal geodata including multi-temporal orthophotos. On MovingDrone and real-world benchmarks, OrthoTrack runs in real time on a single GPU. It outperforms all baselines by a large margin, even those receiving oracle scale and alignment. By relying on publicly available geodata, OrthoTrack enables deployment to new regions without site-specific adaptation.

Keywords: UAV localization · 6-DoF pose estimation · geodata-based tracking · benchmark dataset

1 Introduction

Autonomous Unmanned Aerial Vehicles (UAVs) are increasingly deployed for infrastructure inspection [11], emergency response [20], and environmental monitoring [8], all requiring accurate, continuous 6-degrees of freedom (DoF) pose estimation. While Global Navigation Satellite System (GNSS) provides coarse position, it offers no orientation and degrades or fails in urban canyons and signal-denied environments. Vision-based localization is a natural alternative, yet existing paradigms have fundamental limitations. For instance, Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM) [2, 25, 26] track

⋆ Corresponding author: oussema.dhaouadi@tum.de



(a) MovingDrone Dataset (b) 6-DoF Trajectory Estimation

Fig. 1: OrthoTrack Estimates Continuous 6-DoF UAV Trajectories From Public Geodata. (a) Our MovingDrone Dataset pairs photorealistic UAV sequences with dense 6-DoF ground truth and multi-modal geodata supporting diverse aerial perception tasks (LiDAR, LoD1/2 meshes, DSM, multi-year DOPs, depth, and normals). (b) OrthoTrack matches UAV frames against the DOP, lifts correspondences to 3D via the DSM, and outputs metrically scaled poses at every frame.

features but accumulate unbounded drift without global anchors. Image retrieval [23,40,42] provides limited metric accuracy and requires large databases. Mesh-based methods [19, 39] achieve high accuracy but depend on dense 3D models that are costly to maintain and prohibitive for continuous video. Furthermore, Level of Detail (LoD) city-model approaches [43,44] are compact and privacy-preserving yet restricted to structured urban geometry.

A practical alternative is to localize against publicly available Digital Orthophotos (DOPs) and Digital Surface Models (DSMs), which are metrically accurate, globally registered geodata that government agencies publish at centimeter resolution [4, 40]. Recent dense matchers [6, 33] have made cross-view matching between oblique UAV imagery and top-down orthophotos feasible, achieving sub-meter accuracy on individual frames [4, 40]. However, extending this to continuous video exposes two shortcomings: each dense match is too computationally expensive for frame-rate operation, and independent frames can produce isolated outliers with no mechanism for temporal recovery.

To address both shortcomings, we present OrthoTrack, a training-free system for continuous 6-DoF UAV trajectory estimation from publicly available DOP and DSM data (Fig. 1). The key insight is that geodata-based localization and optical flow tracking are complementary: the former provides drift-free absolute anchoring, the latter temporal continuity at frame-rate speed. OrthoTrack geolocalizes sparse keyframes against the DOP and lifts correspondences to metric 3D positions via the DSM. It then propagates these anchored 2D-3D correspondences to intermediate frames using optical flow. Every pose is thereby globally registered and metrically scaled by construction, without the post-hoc alignment that VO/SLAM methods require.

We also introduce the MovingDrone Dataset, a large-scale benchmark for continuous UAV localization with dense 6-DoF ground truth (Fig. 1). To our knowledge, it is the first to combine continuous video with per-frame metric poses and multi-temporal geodata. Prior datasets offer only single-frame evaluation [4] or at most two map years [40]. We render photorealistic video from a governmentgrade textured mesh, yielding pixel-perfect poses with a genuine cross-domain gap since the mesh and reference geodata stem from independent surveys. It covers 21 Berlin regions with DOP imagery from up to 14 years, and the pipeline extends to any city with publicly available geodata.

Our contributions are threefold. (1) OrthoTrack, a training-free, matcher- and flow-agnostic system for continuous 6-DoF UAV trajectory estimation that produces metrically scaled absolute poses without GPS, task-specific training, or post-hoc alignment. (2) The MovingDrone Dataset, a large-scale benchmark with dense ground truth, co-registered multi-modal geodata, and a scalable generation pipeline. (3) State-of-the-art results against VO/SLAM, 3D foundation models, and UAV localization baselines at real-time throughput.

2 Related Work

UAV Sparse Localization. Single-image UAV localization estimates the 6-DoF camera pose of an individual aerial frame against a pre-existing map. Structure-based approaches match against Structure from Motion (SfM) point clouds or textured meshes [19, 39]. UAVD4L [36] renders synthetic views from a textured city model for coarse retrieval and 6-DoF estimation, but such reconstructions are expensive. LoD city-model approaches exploit structured building geometries: LoD-Loc [44] aligns neural wireframes against LoD3 models, while LoD-Loc v2 [43] uses silhouette alignment on more widely available LoD1/2 data. OrthoLoC [4] localizes against DOPs and DSMs, demonstrating that lightweight public geodata suffices for accurate 6-DoF estimation. All these methods process frames independently and do not exploit temporal continuity. The cross-view matching they rely on has been advanced by dense matchers [5–7], especially with cross-domain training [33] and aerial-ground data [29]. OrthoTrack adopts RoMa v2 [6] for this step.

Visual Odometry and SLAM. Monocular VO and SLAM recover camera motion from frame-to-frame correspondences but produce trajectories with unknown metric scale. While loop closure and relocalization, as in ORB-SLAM3 [2], can reduce accumulated drift, they require revisiting previously observed geometry, which is rare in typical UAV forward-flight scenarios. Recent learning-based methods address some of these shortcomings: DROID-SLAM [25] performs differentiable dense bundle adjustment and DPVO [26] offers a competitive speed-accuracy trade-off via recurrent patch updates, yet both still suffer from unbounded drift without loop closures and lack absolute metric scale. OrthoTrack addresses both by periodically re-anchoring to DSM-derived georeferenced 3D positions, producing globally consistent, metrically scaled trajectories.

Feed-Forward 3D Foundation Models. Recent feed-forward models predict 3D geometry directly from images, bypassing classical multi-view pipelines. While designed for general-purpose 3D understanding, they can in principle estimate camera trajectories from video, making them relevant baselines for UAV pose estimation. DUSt3R [31] pioneered pairwise point-map prediction from uncalibrated image pairs; follow-ups extend the paradigm to dense matching

Table 1: Comparison With Existing UAV Localization Benchmarks.

Dataset	Year	Data	Scene	#Loc	Frames	View	Alt.	fps	Depth	6-DoF	DOP	DSM	LiDAR	Mesh	LoD
University-1652 [42]	2020	Re+Sy^R	C	72	50k	O	B	–	✗	✗	$1^\dagger$	✗	✗	✗	✗
VPAIR [21]	2022	Re	M	1	2.7k	N	H	1	✓	✓	1	✗	✓	✗	✗
ALTO [3]	2022	Re	M	1	15.4k	N	H	$20^\S$	✗	✓	1	✗	✓	✗	✗
CrossLoc [38]	2022	Re+Sy	U/N	2	24k	B	L	–	✓	✓	✗	✗	✗	✗	✓‡
UAVD4L [36]	2024	Re+Sy	U	2	19k	B	L	0.5	✓	✓	✗	✓	✗	✓	✓‡
UAV-VisLoc [37]	2024	Re	M	11	6.7k	N	H	0.2	✗	✗	$1^\dagger$	✗	✗	✗	✗
GTA-UAV [10]	2025	Sy^G	M	1	33k	N	B	–	✗	✓	✗	✗	✗	✗	✗
OrthoLoC [4]	2025	Re	M	47	16.4k	B	B	–	✓	✓	2	✓	✗	✓	✗
AnyVisLoc [40]	2025	Re	M	25	18k	B	B	–	✗	✗	$1+1^\dagger$	✓	✗	✗	✗
UAVScenes [32]	2025	Re	M	4	120k	B	L	10	✗	✓	✗	✗	✓	✓	✗
MovingDrone (Ours)	2026	Sy^R	M	21	343k	B	B	30	✓	✓	≤ 14	✓	✓	✓	✓

Re=real, Sy^R=synthetic (photorealistic), Sy^G=synthetic (game engine); U=urban, C=campus, M=mixed, N=nature; N=nadir, O=oblique, B=both; L=low ($\leq$150 m), H=high (>150 m).
†Satellite tiles, not surveyed orthophotos. ‡Datasets extended to UAVD4L-LoD and Swiss-EPFL with LoD models in [43, 44]. §Full dataset never publicly released; only sparse competition subsets available.

(MASt3R [14]), feed-forward multi-view inference (MUSt3R [1]), joint camera-depth-track prediction (VGGT [30]), and metric-scale reconstruction (Pow3R [9], DA3 [16], MapAnything [12], Pi3 [34]). Despite their generality, all these models produce local-frame reconstructions that require a similarity transform for georeferenced alignment, limiting their applicability to absolute UAV pose estimation. In contrast, OrthoTrack directly anchors every prediction to the geodata reference frame, avoiding post-hoc alignment entirely.

UAV Localization Datasets. Existing benchmarks cover only a subset of the requirements for evaluating continuous map-based localization (Tab. 1). Several benchmarks [3,10,21,37,42] target geo-localization or place recognition but provide either no metric 6-DoF poses [37,42] or only sparse imagery [4,10,38,40,42] unsuitable for sequential tracking. Among datasets with 6-DoF annotations, CrossLoc [38] ships no geodata, UAVD4L [36] pairs mesh-derived geodata with only 0.5 fps imagery, and AnyVisLoc [40] provides single images with two map resolutions. Closest to our setting, OrthoLoC [4] pairs DOPs/DSMs with 6-DoF poses across 47 locations but supplies only sparse, independent frames, while UAVScenes [32] offers dense video with Light Detection and Ranging (LiDAR) but ships neither DOPs nor DSMs. To our knowledge, MovingDrone is the first dataset to combine dense video with a jointly aligned, multi-temporal suite of geodata modalities.

3 Method

Given a UAV video and a georeferenced map consisting of a DOP and a DSM, OrthoTrack estimates a 6-DoF pose for every frame (Fig. 2). The system first localizes itself on the map without any external sensor. It then enters a loop that alternates between two modes: at keyframes, a dense cross-view match against the DOP produces a globally anchored pose. Between keyframes, optical flow propagates the existing correspondences for lightweight pose updates. An

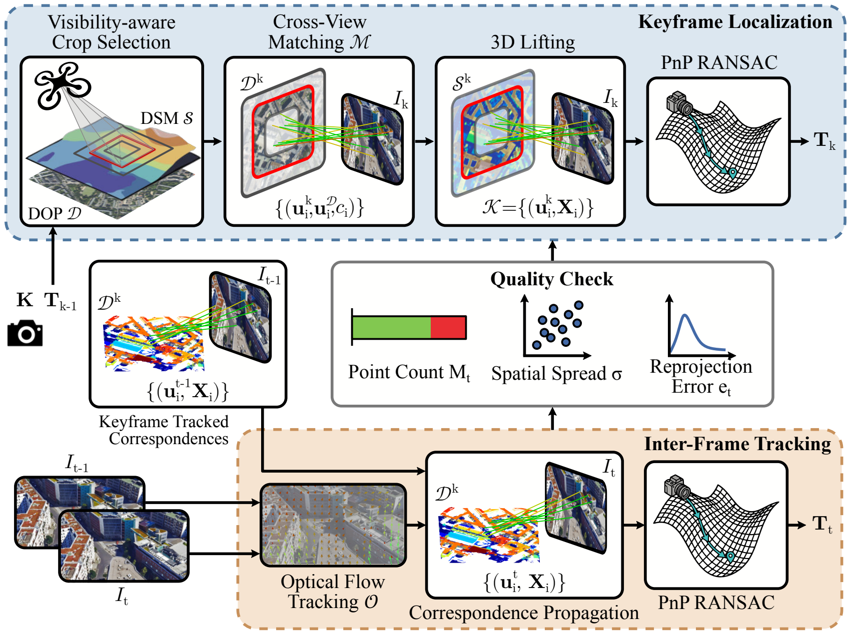


Fig. 2: OrthoTrack Pipeline. Keyframe localization (top): a visibility-aware crop selects DOP $\mathcal{D}^k$ and DSM $\mathcal{S}^k$ from the previous pose. The dense matcher $\mathcal{M}$ matches keyframe $\mathbf{I}_k$ against $\mathcal{D}^k$, correspondences are lifted to 3D via $\mathcal{S}^k$, and PnP-RANSAC recovers the keyframe pose $\mathbf{T}_k$. Inter-frame tracking (bottom): optical flow $\mathcal{O}$ propagates keyframe-anchored correspondences from previous frame $\mathbf{I}_{t-1}$ to current frame $\mathbf{I}_t$ for a lightweight PnP solve. The adaptive quality monitor (center) triggers re-localization when point count, spatial spread, or reprojection error degrades.

adaptive quality monitor triggers new keyframes when tracking degrades. We begin with notation and then detail each component.

3.1 Preliminaries

Notation. We denote the UAV video as $\{\mathbf{I}_t\}_{t=1}^T$ with known pinhole intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$. All poses are world-to-camera: $\mathbf{T}_t = [\mathbf{R}_t \mid \mathbf{t}_t] \in \mathrm{SE}(3)$ maps a world point $\mathbf{X}$ to the camera frame as $\mathbf{R}_t\mathbf{X} + \mathbf{t}_t$, with projection $\mathbf{u} = \pi(\mathbf{K}(\mathbf{R}_t\mathbf{X} + \mathbf{t}_t))$ where $\pi([x, y, z]^\top) = [x/z, y/z]^\top$. The camera center is $\mathbf{C}_t = -\mathbf{R}_t^\top \mathbf{t}_t$. All coordinates use a metric Universal Transverse Mercator (UTM) frame.
Map Representation. The reference map consists of two georeferenced rasters: (1) the DOP $\mathcal{D} : \mathbb{R}^2 \to \mathbb{R}^3$, an orthorectified aerial photograph providing top-down RGB at each position (x, y), and (2) the DSM $\mathcal{S} : \mathbb{R}^2 \to \mathbb{R}$ encoding surface elevation $z = \mathcal{S}(x, y)$. Together, any DOP pixel at (x, y) corresponds to the 3D point $\mathbf{X} = [\xi(x), \eta(y), \mathcal{S}(x, y)]^\top$, where ξ and η define the pixel-to-UTM affine transformation.
Dense Feature Matching. We employ a pre-trained dense matcher $\mathcal{M}$ (e.g., RoMa-v2 [6]) that, given images $\mathbf{A}$ and $\mathbf{B}$, produces N correspondences

$\{(\mathbf{u}_i^A, \mathbf{u}_i^B, c_i)\}$ with confidence $c_i \in [0, 1]$. In our pipeline, $\mathbf{A}$ is the UAV frame and $\mathbf{B}$ is a DOP crop.

Problem Formulation. Given video $\{\mathbf{I}_t\}$, a georeferenced map $(\mathcal{D}, \mathcal{S})$, and intrinsics $\mathbf{K}$, we aim to estimate the 6-DoF pose $\mathbf{T}_t$ for every frame in metric world coordinates, without GPS or IMU.

3.2 First-Frame Initialization

OrthoTrack is fully self-initializing and requires neither GPS nor IMU at any stage. The only input beyond the video and intrinsics is a search area over the reference map, which can span the entire available DOP coverage. We match the first UAV image against the full DOP at reduced resolution to obtain cross-view correspondences whose DOP-side positions we convert to UTM. The median of these UTM positions serves as an approximate center around which we seed a multi-scale refinement: several DOP crops (we use 3 crops) of increasing extent are matched independently at full resolution, and the merged 2D–3D correspondences are passed to a single PnP solve for the initial pose.

When the coarse match fails to produce a reliable position (fewer than N_{gs} correspondences), we fall back to a coarse-to-fine pyramid tiling strategy: we partition the search area into overlapping tiles with 50% overlap and match each tile independently. The search terminates early once a tile yields N_{stop} PnP inliers, and its position seeds the same multi-scale refinement.

3.3 Keyframe Localization

Each keyframe yields an absolute pose following the single-frame localization principle of OrthoLoC [4], which matches a UAV frame against a DOP crop and solves PnP from DSM-lifted correspondences. However, OrthoLoC assumes a pre-cropped DOP/DSM pair with high co-visibility, which is unavailable in a sequential tracking setting. We address this with a visibility-aware crop selection that handles oblique views, a two-phase matching strategy with multi-scale fallback for difficult frames, and an adaptive triggering mechanism that embeds the localizer within a tracking loop.

Visibility-Aware Crop Selection. Matching the full DOP at every keyframe is prohibitively expensive and degrades precision, since the matcher must compress a large area into a fixed-resolution input. Instead, we crop a local region around the expected camera position to preserve spatial detail. A naive nadir-centered crop fails for oblique views where the visible ground may lie far from the point directly below the UAV. Given the previous pose $\mathbf{T}_{k-1}$, we project a dense grid of DSM points into the image, retaining those that (i) are in front of the camera, (ii) fall within image bounds, and (iii) are not occluded, as determined by a coarse depth buffer that discards points behind nearer surfaces. Let $\mathcal{V}$ denote the surviving visible points with camera-frame depths d_j. For oblique views ($\theta > 15°$), we compute the crop center as an inverse-depth-weighted mean

$$\bar{\mathbf{X}} = \frac{\sum_{j \in \mathcal{V}} d_j^{-1}\, \mathbf{X}_j}{\sum_{j \in \mathcal{V}} d_j^{-1}}, \tag{1}$$

which biases the crop toward the near-ground region that occupies the largest image area and where the matcher resolves detail best, rather than toward distant terrain that projects to few pixels. For near-nadir views ($\theta \leq 15°$), where depth variation is small, we simply use the bounding-box center of $\mathcal{V}$. The resulting crop defines the local DOP region $\mathcal{D}^k$ and its co-registered DSM region $\mathcal{S}^k$, both used in the subsequent matching and lifting steps. We set the crop size proportional to the spatial extent of $\mathcal{V}$.

Cross-View Matching and 3D Lifting. We match the local DOP crop $\mathcal{D}^k$ against the UAV frame using $\mathcal{M}$, yielding correspondences $\{(\mathbf{u}_i^k, \mathbf{u}_i^{\mathcal{D}^k}, c_i)\}$. Each DOP-side match is converted to a UTM coordinate via the known georeferencing transform and lifted to 3D: $\mathbf{X}_i = [\xi(x_i), \eta(y_i), \mathcal{S}(x_i, y_i)]^\top$ using bilinear interpolation of the full-resolution DSM. We discard correspondences with $c_i < c_{\min}$ and, if fewer than $N_{\min}$ points remain, relax the threshold to $c'_{\min}$ to retain sufficient coverage. As a fast path, we first attempt PnP on this single crop. If it yields enough inliers, we accept the pose immediately. Otherwise, we match additional crops of increasing extent centered on the initial matches, merge the resulting correspondences, and re-run PnP.

PnP Pose Estimation. We pass the resulting 2D–3D correspondences to PnP-RANSAC (SQPnP solver [27]), which jointly rejects outliers and recovers $\mathbf{T}_k$. To avoid conditioning issues with large UTM coordinates, we subtract the 3D centroid before solving and restore it in the recovered translation. Because every 3D point is sampled from the georeferenced DSM, the resulting pose is metric and globally anchored by construction.

3.4 Inter-Frame Tracking

Once a keyframe pose is established, subsequent frames are processed online by propagating the existing 2D–3D correspondences forward via optical flow $\mathcal{O}$, avoiding the cost of repeated cross-view matching.

Correspondence Propagation. Let $\mathcal{C}_k = \{(\mathbf{u}_i^k, \mathbf{X}_i)\}$ be the set of 2D–3D correspondences established at keyframe k. For each subsequent frame $t > k$, we track each 2D keypoint from $\mathbf{I}_{t-1}$ to $\mathbf{I}_t$ using $\mathcal{O}$. We apply a forward-backward consistency check and discard points whose round-trip displacement exceeds τ_{fb} pixels. Because the 3D points are anchored in the static map, the propagated pairs $\{(\mathbf{u}_i^t, \mathbf{X}_i)\}$ directly yield a pose via PnP-RANSAC.

Keyframe Triggering. In principle, keyframes could be triggered at fixed intervals based on pre-defined thresholds. However, the optimal interval varies with scene content and camera motion, making a fixed schedule either wasteful during stable flight or insufficient during rapid changes. Instead, we trigger a new keyframe only when tracking quality degrades. Let M_t be the number of surviving correspondences at frame t, M_k the count at keyframe k, and $\Delta t = t - k$ the elapsed frames. We trigger a new keyframe when any of the following conditions is met: (1) $M_t < N_{\min}$ (insufficient points for reliable PnP), (2) $M_t < \alpha\, M_k$ (large fraction of correspondences lost since the last keyframe), (3) $\min(\sigma_x, \sigma_y) < \sigma_{\min}$ where σ_x and σ_y are the spatial standard deviations of $\mathbf{u}^t$ (spatial collapse: re-

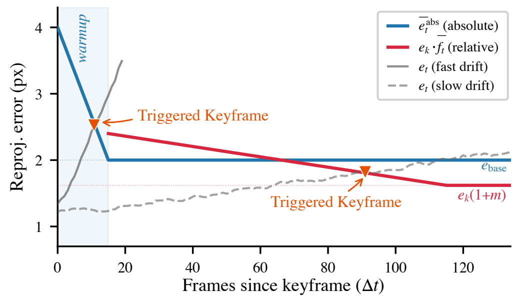


Fig. 3: Adaptive Reprojection Thresholds. The absolute threshold (blue) starts relaxed during warmup, then settles at e_{base}. The relative threshold (red) tightens over D frames. Rapid drift is caught by the absolute bound, slow drift by the tightening relative bound.

maining points cluster in a small image region, making PnP ill-conditioned), or (4) reprojection error e_t exceeds an adaptive threshold.

A fixed reprojection threshold for condition (4) is problematic: cross-view matching produces a variable initial error at each keyframe, so a tight threshold triggers an infinite re-triggering loop, while a loose one misses slow drift. We therefore combine two adaptive thresholds that adjust to each keyframe's matching quality (Fig. 3), making the criterion robust across scenes without per-dataset tuning. The first is an absolute threshold that relaxes the bound during a warmup window after each keyframe:

$$\bar{e}_t^{\text{abs}} = e_{\text{base}} + \delta \cdot \max\bigl(0,\ 1 - \tfrac{\Delta t}{G}\bigr), \tag{2}$$

where e_{base} is the steady-state bound, δ the initial tolerance, and G the warmup duration in frames. The second is a relative threshold, active for $\Delta t \geq G$, that detects slow drift by comparing e_t to the keyframe's baseline error e_k. It triggers when $e_t > \bar{f}_t \cdot e_k$, where

$$\bar{f}_t = 1 + (f_0{-}1)(1{-}\rho) + m\,\rho, \quad \rho = \min\bigl(1,\ \tfrac{\Delta t - G}{D}\bigr), \tag{3}$$

decreases from f_0 to $1{+}m$ over D frames. Since this threshold scales with e_k, it adapts to each keyframe's matching quality: a noisier keyframe permits more absolute error, while the decaying factor ensures gradual drift eventually triggers re-localization.

4 The MovingDrone Dataset

We present MovingDrone, a large-scale UAV localization benchmark that renders photorealistic video from the Berlin VirtualCityMap [28], a high-precision textured photogrammetric mesh reconstructed from real aerial imagery (cross-validated to 0.18 m median vertical accuracy) that covers Germany's largest city. We select 21 regions to span Berlin's full architectural and landscape diversity, from dense historical centers and modern high-rises to parks, waterways, airports, and stadiums. The fully automated pipeline can generate arbitrarily many

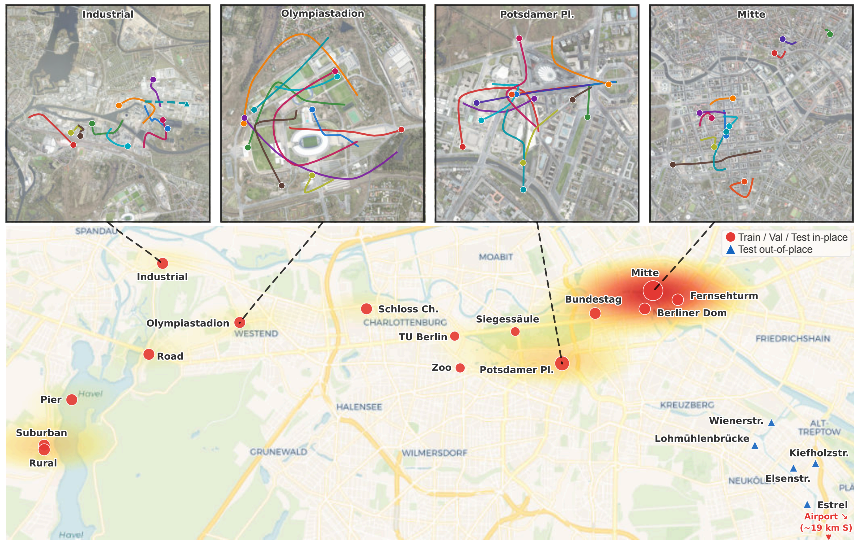


Fig. 4: MovingDrone Locations. 194 sequences across 21 Berlin regions. Insets show representative trajectories over satellite imagery. The heatmap encodes sequence density.

sequences anywhere within the city's coverage at negligible cost, circumventing the airspace restrictions, privacy regulations, and expensive RTK-GNSS supervision that make real UAV data collection impractical at metropolitan scale. Because the mesh preserves authentic textures, lighting, and geometric detail from the original survey, the synthetic-to-real domain gap is substantially smaller than in game-engine datasets such as GTA-UAV [10]. The underlying reconstruction stems from a comprehensive campaign, yielding a watertight, gap-free surface, unlike meshes in other benchmarks [32, 36] that suffer from holes in areas due to insufficient image coverage. Each sequence is paired with the same publicly available geodata that a practitioner would use in the field, ensuring that every label is by construction spatially consistent with operational map assets at sub-pixel accuracy.

Scale and Diversity. MovingDrone comprises 194 sequences totaling 343,114 frames across 21 regions (Fig. 4). Camera tilt ranges from near-nadir (0°) to highly oblique (83°), with 92% of frames exceeding 20°, reflecting the oblique-view regime common in inspection and monitoring applications. On average, each sequence is paired with 13.19 DOP years (2011–2025), enabling analysis of localization robustness to map age (Fig. 5).

Generation Pipeline. Camera trajectories are designed in Google Earth Studio across 21 Berlin regions covering four archetypes (orbital, linear, oblique descent, multi-turn) at 64–1,183 m altitude, inspired by the trajectories of public drone footages. The exported trajectories are resampled via cubic splines to bound speeds to 10–100 km/h. Although the base trajectories are designed rather than recorded from a physical drone, three physically motivated augmentations inject realistic flight dynamics: (i) multi-band trajectory noise com-

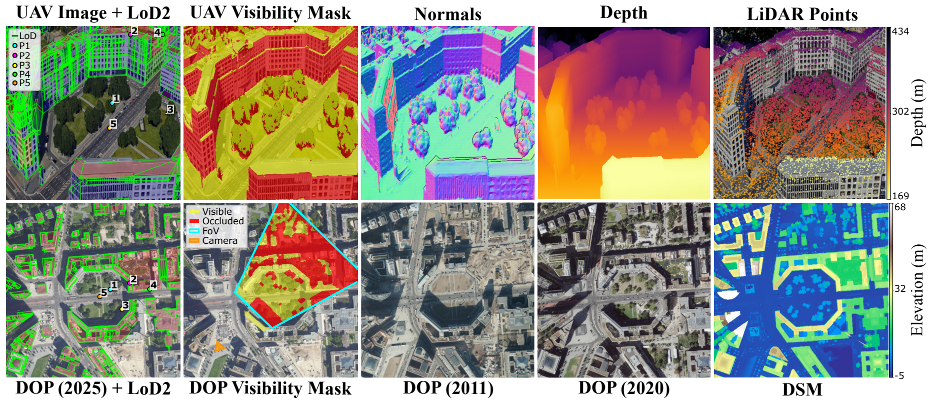


Fig. 5: MovingDrone Multi-Modal Data Per Frame. Each rendered UAV frame is paired with depth, surface normals from the mesh, LiDAR, multi-year DOPs, a DSM, and LoD1/2 building meshes. Comparing the three DOPs reveals the temporal domain gap: lighting, vegetation, and even building footprints change across years.

bining GPS drift with motor vibration, (ii) stochastic wind-gust events, and (iii) velocity-dependent directional motion blur. Photorealistic video is rendered from the Berlin VirtualCityMap [28] textured mesh, with per-frame depth obtained through ray casting for pixel-exact consistency. Each sequence is automatically paired with open geodata from Berlin and Brandenburg portals [13,22] (Fig. 5): multi-year DOPs and a DSM (both at 0.2 m resolution), classified airborne LiDAR (5–20 pts/m^2, ASPRS classes), the textured photogrammetric mesh used for rendering, and CityGML building meshes (LoD1/LoD2). Full details on the generation pipeline, augmentations, and data splits are provided in the supplement.

5 Experiments

In the following, we evaluate OrthoTrack's trajectory estimation accuracy on MovingDrone and the real-world UAVScenes [32] benchmark against VO/SLAM systems, 3D foundation models, and geodata-based localization methods, and validate each design choice through ablation studies.

5.1 Experimental Setup

Datasets. We evaluate on the 10 MovingDrone test sequences (with DOP from year 2025) and all 20 real-world UAVScenes [32] sequences (121 k real-world frames at 10 fps) across Armenia and Hong Kong. Since UAVScenes ships neither DOPs nor DSMs, we render both from the provided LiDAR-reconstructed mesh at 0.10 m GSD. In contrast, MovingDrone uses independently surveyed geodata, producing a genuine cross-domain gap, while the cross-domain gap in UAVScenes is small. No dataset-specific adaptation is applied.

Table 2: 6-DoF UAV Trajectory Estimation on UAVScenes and MovingDrone. Bold: best; underline: second best. Metrics: ATE and TE in meters [m]↓, RE in degrees [°]↓, R@k in percent [%]↑.

Method	Align.	UAVScenes [32]					MovingDrone (Ours)					
		ATE↓	TE↓	RE↓	R@1↑	R@2↑	ATE↓	TE↓	RE↓	R@1↑	R@2↑	FPS↑
VO / SLAM												
Five-Point VO [18]	Sim(3)	76.08	61.98	28.53	0.0	0.2	55.89	45.25	36.76	0.0	0.0	2.0
ORB-SLAM3 [2]	ff+s	84.39	56.48	29.25	**5.6**	8.5	29.20	25.45	10.25	12.1	27.6	**22.3**
ORB-SLAM3 [2]	Sim(3)	36.90	20.58	14.61	1.3	6.3	13.14	9.11	14.30	5.9	22.9	**22.3**
ORB-SLAM3 [2]	G, pw	6.98	1.47	108.83	0.6	3.6	5.24	2.37	48.90	0.8	5.7	**22.3**
ORB-SLAM3 [2]	G+M, pw	6.98	1.47	99.47	2.7	8.7	5.24	2.37	42.82	0.8	7.6	**22.3**
DROID-SLAM [25]	ff+s	147.52	126.52	51.49	3.4	4.4	0.89	0.75	0.14	72.7	89.6	4.7
DROID-SLAM [25]	Sim(3)	103.96	86.81	59.85	0.5	3.0	0.34	0.22	0.47	88.8	89.7	4.7
DROID-SLAM [25]	pw	8.03	2.53	**0.80**	5.4	**32.1**	0.63	0.35	**0.08**	**93.7**	**99.1**	4.7
DROID-SLAM [25]	G, pw	**1.04**	**0.08**	46.63	1.7	6.1	**0.13**	**0.05**	4.57	24.1	37.5	4.7
DROID-SLAM [25]	G+M, pw	**1.04**	**0.08**	35.02	5.0	12.2	**0.13**	**0.05**	4.30	25.7	41.2	4.7
DPVO [26]	ff+s	151.06	128.21	49.78	3.1	3.5	4.01	2.06	0.37	63.8	82.2	4.5
DPVO [26]	Sim(3)	100.92	85.78	53.88	0.1	0.9	2.43	1.04	0.68	72.7	84.1	4.5
Foundation models												
DUSt3R [31]†	Sim(3)	84.14	63.37	37.84	0.0	0.0	53.43	43.60	81.32	0.0	0.0	0.2
DA3-Nested [16]	Sim(3)	58.27	38.74	16.47	**0.2**	**1.1**	**2.74**	**1.83**	**2.57**	**28.7**	**51.4**	11.3
VGGT-SLAM v2 [30]	Sim(3)	**57.17**	**34.57**	**9.47**	0.0	0.0	11.94	9.28	15.88	0.0	0.0	3.1
Pi3 [34]‡	Sim(3)	102.80	80.39	53.75	0.0	0.0	23.98	19.30	39.46	0.2	14.2	**25.2**
UAV localization (GPS/IMU prior)												
LoD-Loc [44]	none			*no LoD models*			17.12	9.52	1.49	4.0	11.5	3.0
LoD-Loc [44] (retrained)	none			*no LoD models*			10.04	5.25	1.10	10.8	30.3	3.3
OrthoLoC [4]*	none	38.24	**1.26**	**0.43**	**23.5**	**93.8**	19.78	0.50	0.08	85.4	95.4	0.5
OrthoTrack – loc. only	none	1.89	1.28	**0.43**	20.5	91.7	4.83	**0.30**	**0.06**	**95.8**	**99.2**	1.9
OrthoTrack – track only	none	102.64	19.53	7.75	12.3	15.0	27.70	16.36	5.20	3.0	7.7	13.4
OrthoTrack (Ours)	none	**1.51**	1.38	0.49	20.6	88.0	**0.67**	0.33	**0.06**	90.9	97.9	**23.8**

Align.: Sim(3) = oracle 7-DoF, ff+s = first-frame anchor + scale, pw = piecewise Sim(3) to geodata anchors, none = absolute poses (no alignment). † windowed; ‡ sometimes windowed; * GIM(DKM)+AdHoP, same DOP/DSM inputs as OrthoTrack.

Metrics. We report six metrics, all computed per sequence and averaged: ATE (RMSE of 3D position errors), TE (median translation error), RE (median rotation error), and joint recall R@k (percentage of frames with TE$<k$ m and RE$<k°$). All timing is measured on a single NVIDIA L40S GPU.

Baselines. We compare in Tab. 2 against monocular VO/SLAM systems, recent 3D foundation models, and geodata-based localization methods. All baselines and models in the ablation studies use official code with default hyperparameters.

Evaluation Protocol. VO/SLAM systems receive raw video under oracle Sim(3) alignment and first-frame anchoring with scale correction (ff+s). Foundation models not fitting to GPU memory use sliding-window inference with per-window Sim(3) stitching. OrthoLoC [4] and LoD-Loc [44] receive simulated sensor priors (noisy GPS + orientation). Piecewise-aligned DROID-SLAM (pw) aligns a pre-computed trajectory every 50 frames via Sim(3) using the same matcher and pose solver as OrthoTrack keyframes.

Implementation Details. We use RoMa-v2 [6] Precise as the dense matcher, selected for its strong cross-domain performance, and GPU-accelerated Lucas-Kanade [17] for inter-frame optical flow. The design choice is justified in Tab. 3. The PnP-RANSAC pose solver uses SQPnP [27]. The remaining hyperparameters and timing details are in the supplementary.

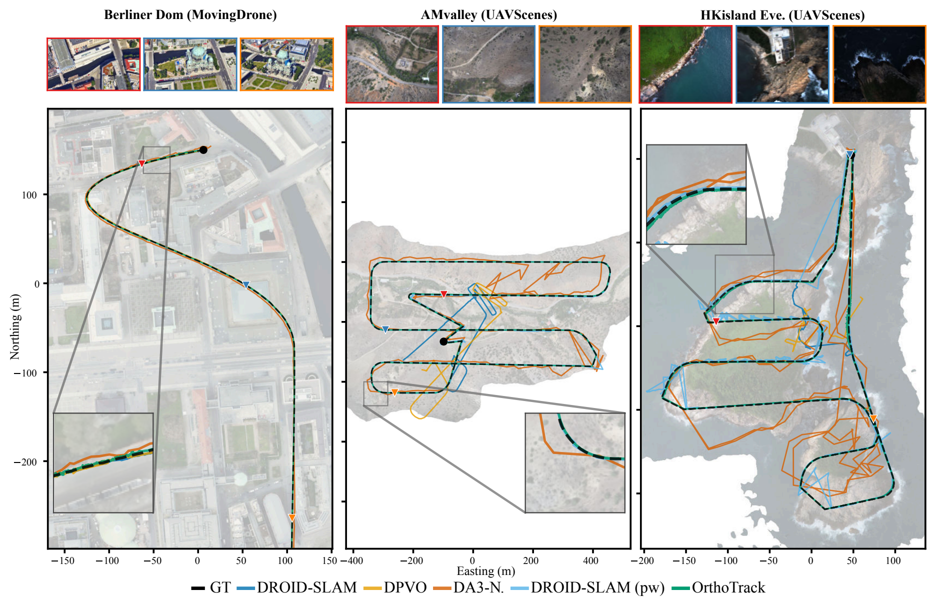


Fig. 6: Trajectory Comparison. VO/SLAM baselines receive oracle Sim(3) alignment yet drift or collapse on long flights. Piecewise-anchored DROID-SLAM (pw) reduces drift but accumulates errors between anchors. OrthoTrack produces absolute poses without alignment and remains accurate.

5.2 Comparison with State of the Art

Tab. 2 presents results on MovingDrone and on real-world UAVScenes [32]. OrthoTrack is the only method to maintain sub-metre ATE on both benchmarks without oracle alignment or sensor priors, while operating at real-time speed. Representative trajectories are shown in Fig. 6 and per-frame localization quality is visualized in Fig. 7.

Among VO/SLAM methods, DROID-SLAM [25] achieves the lowest ATE on MovingDrone under oracle Sim(3) (0.34 m), yet collapses to 104 m on UAVScenes (a 69× gap to OrthoTrack). We attribute this to the near-planar scene geometry at high altitude: monocular parallax cannot resolve depth when the baseline-to-depth ratio is negligible, causing the reconstructed trajectory to compress and accumulate drift that no post-hoc alignment can recover. Piecewise geodata anchoring (pw) even slightly outperforms OrthoTrack on MovingDrone (0.63 vs. 0.67 m ATE). On UAVScenes, however, the VO collapse propagates between anchors (8.03 vs. 1.51 m, >5× above OrthoTrack). Providing piecewise alignment with oracle GNSS (G) and magnetometer (M) significantly improves the VO baselines by continuously re-anchoring the scale and heading, allowing DROID-SLAM to reach 1.04 m ATE on UAVScenes. However, the addition of the magnetometer (+M) does not significantly reduce the overall rotation error (RE). This is because a magnetometer provides only a global yaw reference and cannot correct accumulated drift in pitch and roll, leaving the 3D orientation degraded. Furthermore, this piecewise approach relies on perfect continuous sensor

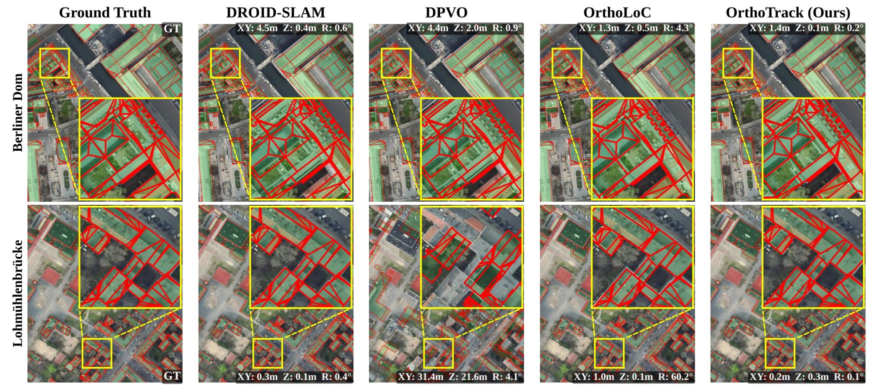


Fig. 7: Per-Frame Localization Quality. LoD2 building wireframes (red) overlaid on the orthophoto are shifted by each method's position error, making misalignment directly visible.

priors, which are often unavailable or unreliable in real-world aerial scenarios. DPVO [26] and ORB-SLAM3 [2] exhibit the same degeneracy, confirming a fundamental limitation of relative-pose methods on aerial scenes. Moreover, DROID runs at 4.7 FPS vs. OrthoTrack's 23.8 FPS, making it 5× slower even before accounting for the alignment oracle it requires.

Foundation models fare similarly: despite oracle alignment, even the strongest variant cannot approach OrthoTrack's performance. This gap is primarily driven by out-of-distribution bias, as these models were trained predominately on ground-level images and struggle with the aerial perspective. Furthermore, because they predict geometry in a local frame, they rely on a similarity transform that breaks down when the reconstruction drifts over long sequences without any mechanism for external re-anchoring.

Among geodata-based methods that use prior sensor data (GPS/IMU) to identify the region of interest, LoD-Loc [44] reaches only 10 m ATE on MovingDrone even after retraining, likely because wireframe alignment is less discriminative than dense texture matching on complex urban scenes or due to degenerate results in regions with limited buildings. OrthoLoC [4] achieves competitive single-frame TE (0.50 vs. 0.33 m on MovingDrone and 1.26 vs. 1.38 m on UAVScenes), confirming per-frame localizer accuracy, yet without temporal consistency sporadic failures propagate directly into ATE (19.78 m and 38.24 m). OrthoTrack cuts ATE by 30× and by 25× on the respective datasets, invoking the matcher at only ∼1% of frames and propagating correspondences via optical flow, effectively filtering isolated outliers without GPS or IMU.

5.3 Ablation Studies

Localization vs. Tracking. The ablation variants in Tab. 2 isolate each component. Localizing (loc. only) every frame via matching, similar to OrthoLoC but with RomaV2 and visibility-aware crop selection and without AdHoP strategy [4], achieves the lowest TE but at prohibitive cost (1.9 FPS). Tracking alone

Table 3: Ablation Studies on MovingDrone. Each section varies one pipeline component while fixing all others. **Bold**: best; underline: second best (per section).

Config	ATE↓	TE↓	RE↓	R@1↑	R@2↑	R@5↑	#KF	FPS_c↑	FPS_s↑
Flow backend (matcher = RoMa v2 Precise, 30 fps)									
LK [17]	0.67	**0.33**	**0.06**	90.9	97.9	99.3	15	**30.3**	**23.8**
RAFT [24]	**0.66**	0.37	0.08	91.9	**98.4**	**99.6**	**10**	11.9	10.0
SEA-RAFT [35]	1.93	0.37	0.08	**93.5**	97.9	99.4	12	17.4	14.1
NeuFlow v2 [41]	0.82	**0.33**	**0.06**	92.9	97.9	98.7	11	24.4	17.6
Keyframe matcher (flow = LK, 30 fps)									
RoMa v2 (Precise)	**0.67**	0.33	**0.06**	90.9	**97.9**	99.3	15	1.3	23.8
RoMa v2 (Base)	0.90	0.49	0.10	86.4	95.9	99.0	13	2.3	**26.3**
RoMa v2 (Fast)	0.88	0.58	0.11	79.9	96.9	99.1	13	**2.6**	25.4
GIM(DKM) [33]	2.56	0.52	0.10	89.1	**97.9**	**99.4**	13	0.7	21.7
L2M [15]	6.23	**0.29**	**0.06**	**94.9**	96.4	98.3	15	0.4	15.0
MASt3R [14]	5.74	2.10	0.46	20.4	56.3	85.6	**11**	0.3	5.5
Input frame rate (matcher = RoMa v2 Precise, flow = LK)									
30 fps (1×)	0.67	**0.33**	**0.06**	90.9	**97.9**	99.3	15	–	**23.8**
10 fps (3×)	**0.58**	0.39	0.08	**91.0**	**97.9**	**99.7**	11	–	19.2
3 fps (10×)	0.68	0.42	0.08	90.6	97.8	99.0	**8**	–	11.1
2 fps (15×)	0.71	0.38	0.07	90.9	97.3	99.4	**8**	–	8.7
Keyframe triggering (matcher = RoMa v2 Precise, flow = LK, 30 fps)									
Adaptive (Ours)	0.67	**0.33**	**0.06**	90.9	97.9	99.3	15	–	23.8
Fixed reproj (e=2 px)	0.79	0.46	0.08	85.2	94.9	**99.7**	9	–	21.4
Fixed interval (K=50)	**0.53**	0.35	**0.06**	**93.3**	**99.3**	**99.7**	28	–	19.0
Fixed interval (K=50, 2 fps)	12.26	1.30	0.07	54.3	71.0	85.8	**3**	–	6.8
Fixed interval (K=150)	0.82	0.41	**0.06**	90.7	96.9	99.3	10	–	**24.2**
Point-count only	1.69	0.93	**0.06**	61.1	80.8	96.0	4	–	22.5

FPS_c: isolated throughput of the varied component (flow tracking or matching rate).
FPS_s: end-to-end system throughput.

(track only) by initializing the first keyframe and iteratively creating a new set tracking points at each frame without global re-anchoring drifts. The full pipeline combines both: flow-based tracking suppresses per-frame outliers while periodic keyframe localization corrects accumulated drift. On UAVScenes, loc-only achieves lower TE (1.28 vs. 1.38 m) but higher ATE (1.89 vs. 1.51 m), confirming that trajectory consistency requires temporal filtering beyond per-frame accuracy.

Optical Flow Backend. Tab. 3 compares four flow backends. LK [17], RAFT [24], and NeuFlow v2 [41] achieve sub-metre ATE. We attribute LK's parity with learned methods to the pipeline's regime: at 30 fps, inter-frame displacements are small enough for a classical pyramid tracker, and keyframe resets bound drift before it affects accuracy. We select LK for its best accuracy-speed trade-off.

Matching Method. Tab. 3 compares matchers spanning dense and standalone methods. RoMa-v2 [6] Precise achieves the best ATE (0.67 m), which we attribute to its training on AerialMegaDepth [29] that targets the air-to-ground domain gap. Five of six matchers produce sub-metre TE, confirming that the pipeline could be designed with any matcher. The end-to-end throughput stays similar because keyframes constitute only ∼1% of frames.

Input Frame Rate. Tab. 3 evaluates robustness from 30 down to 2 fps. Accuracy remains stable because the adaptive monitor self-regulates: at lower rates it triggers re-localization more often (∼0.6% at 30 fps to ∼5% at 2 fps), automatically trading throughput for accuracy.

Keyframe Triggering Strategy. Tab. 3 compares the adaptive dual-threshold against simpler heuristics. Fixing a threshold for the reprojection error does not outperform our adaptive strategy since the reprojection error alone does not sufficiently reflect the quality of the estimated pose. Using a regular interval of keyframes show different results: For small intervals ($K = 50$), the results are slightly better than our adaptive strategy as keyframes are triggered more often. However, this degradates with lower frame rate (2 fps instead of 30 fps), indicating the sensitivity of this parameter to the camera motion. Yet, our adaptive strategy captures both early and slow drifts: the grace ramp prevents re-triggering loops when cross-view reprojection error is initially high, while the decaying relative bound catches slow geometric drift.

Further evaluation on UAVD4L [36], additional analysis and details are provided in the supplementary material.

6 Conclusion

We present OrthoTrack, a training-free pipeline that matches UAV frames against publicly available orthophotos and surface models, yielding drift-free, metrically scaled 6-DoF poses without GPS or training. We also introduce MovingDrone, a multi-modal benchmark of 194 sequences across 21 Berlin regions whose scalable generation pipeline and paired geodata support broader aerial perception research. On MovingDrone, OrthoTrack achieves 0.67 m ATE and 0.33 m median translation error without any alignment or sensor prior. On the external real-world UAVScenes benchmark it reduces ATE by 69× over the best sim(3)-aligned VO system. Unlike VO/SLAM systems that drift or foundation models bound to a local frame, OrthoTrack anchors every pose in public geodata, enabling deployment to new regions without site-specific data.

Acknowledgments. This research was partially funded by the ETH Foundation Project 2025-FS-352, the SNSF Advanced Grant 216260.

References

1. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3d reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1050–1060 (2025)
2. Campos, C., Elvira, R., Gómez, J.J., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. IEEE Trans. Robot. **37**(6), 1874–1890 (2021)
3. 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)
4. Dhaouadi, O., Marin, R., Meier, J.M., Kaiser, J., Cremers, D.: Ortholoc: UAV 6-dof localization and calibration using orthographic geodata. In: Adv. Neural Inform. Process. Syst. (2025)

5. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: Dkm: Dense kernelized feature matching for geometry estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17765–17775 (2023)
6. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching. arXiv preprint arXiv:2511.15706 (2025)
7. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19790–19800 (2024)
8. Hodgson, J.C., Baylis, S.M., Mott, R., Herrod, A., Clarke, R.H.: Precision wildlife monitoring using unmanned aerial vehicles. Scientific reports **6**(1), 22574 (2016)
9. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1071–1081 (2025)
10. Ji, Y., He, B., Tan, Z., Wu, L.: Game4loc: A uav geo-localization benchmark from game data. In: AAAI. vol. 39, pp. 3913–3921 (2025)
11. Jordan, S., Moore, J., Hovet, S., Box, J., Perry, J., Kirsche, K., Lewis, D., Tse, Z.T.H.: State-of-the-art technologies for uav inspections. IET Radar, Sonar & Navigation **12**(2), 151–164 (2018)
12. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
13. Landesvermessung und Geobasisinformation Brandenburg (LGB): Opengeodata.brandenburg - data portal of geobasis-de. Online geodata platform (2025), `https://data.geobasis-bb.de`, data licensed under Datenlizenz Deutschland Namensnennung 2.0
14. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: Eur. Conf. Comput. Vis. pp. 71–91. Springer (2024)
15. Liang, Y., Hu, Y., Shao, W., Fu, Y.: Learning dense feature matching via lifting single 2d image to 3d space. In: Int. Conf. Comput. Vis. pp. 6621–6631 (2025)
16. 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)
17. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: AAAI. vol. 2, pp. 674–679 (1981)
18. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26**(6), 756–770 (2004)
19. Panek, V., Kukelova, Z., Sattler, T.: Meshloc: Mesh-based visual localization. In: Eur. Conf. Comput. Vis. pp. 589–609. Springer (2022)
20. Scherer, J., Yahyanejad, S., Hayat, S., Yanmaz, E., Andre, T., Khan, A., Vukadinovic, V., Bettstetter, C., Hellwagner, H., Rinner, B.: An autonomous multi-uav system for search and rescue. In: Proceedings of the first workshop on micro aerial vehicle networks, systems, and applications for civilian use. pp. 33–38 (2015)
21. Schleiss, M., Rouatbi, F., Cremers, D.: Vpair–aerial visual place recognition and localization in large-scale outdoor environments. arXiv preprint arXiv:2205.11567 (2022)
22. Senate Department for Urban Development, Building and Housing (Senatsverwaltung für Stadtentwicklung, Bauen und Wohnen), Berlin: Geoportal berlin (geospatial data infrastructure berlin). `https://gdi.berlin.de` (2026), accessed: 2026-02-24

23. Sun, J., Liu, K., Zhang, C., Chen, C., Shen, J., Vong, C.M.: Cross-view uav geo-localization with precision-focused efficient design: A hierarchical distillation approach with multi-view refinement. arXiv e-prints pp. arXiv–2510 (2025)
24. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Eur. Conf. Comput. Vis. pp. 402–419. Springer (2020)
25. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In: Adv. Neural Inform. Process. Syst. vol. 34, pp. 16558–16569 (2021)
26. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. In: Adv. Neural Inform. Process. Syst. vol. 36 (2024)
27. Terzakis, G., Lourakis, M.: A consistently fast and globally optimal solution to the perspective-n-point problem. In: Eur. Conf. Comput. Vis. pp. 478–494. Springer (2020)
28. virtualcitymap berlin 3d city model. `https://berlin.virtualcitymap.de/` (2025), `https://berlin.virtualcitymap.de/`
29. Vuong, K., Ghosh, A., Ramanan, D., Narasimhan, S., Tulsiani, S.: Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21674–21684 (2025)
30. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
31. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20697–20709 (2024)
32. 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: Int. Conf. Comput. Vis. pp. 28946–28958 (2025)
33. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21666–21675 (2024)
34. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
35. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: Eur. Conf. Comput. Vis. pp. 36–54. Springer (2024)
36. Wu, R., Cheng, X., Zhu, J., Liu, X., Zhang, M., Yan, S.: Uavd4l: A large-scale dataset for uav 6-dof localization. In: Int. Conf. 3D Vis. (2024)
37. 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)
38. Yan, Q., Zheng, J., Reding, S., Li, S., Doytchinov, I.: Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17358–17368 (2022)
39. Yan, S., Cheng, X., Liu, Y., Zhu, J., Wu, R., Liu, Y., Zhang, M.: Render-and-compare: Cross-view 6-dof localization from noisy prior. In: Int. Conf. Multimedia and Expo. pp. 2171–2176. IEEE (2023)
40. Ye, Y., Teng, X., Chen, S., Li, Z., Liu, L., Yu, Q., Tan, T.: Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. arXiv preprint arXiv:2503.10692 (2025)
41. Zhang, Z., Gupta, A., Jiang, H., Singh, H.: Neuflow-v2: Push high-efficiency optical flow to the limit. In: IEEE/RSJ Int. Conf. Intell. Robots Syst. pp. 2479–2485. IEEE (2025)

42. 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)
43. Zhu, J., Peng, S., Wang, L., Tan, H., Liu, Y., Zhang, M., Yan, S.: Lod-loc v2: Aerial visual localization over low level-of-detail city models using explicit silhouette alignment. In: Int. Conf. Comput. Vis. pp. 26610–26621 (2025)
44. Zhu, J., Yan, S., Wang, L., shengYue, z., Liu, Y., Zhang, M.: Lod-loc: Aerial visual localization using lod 3d map with neural wireframe alignment. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Adv. Neural Inform. Process. Syst. vol. 37, pp. 119063–119098. Curran Associates, Inc. (2024)