OrthoLoC: UAV 6-DoF Localization and Calibration Using Orthographic Geodata

Oussema Dhaouadi[1,2,3*] **Riccardo Marin**[2,3] **Johannes Meier**[1,2,3]
Jacques Kaiser[1] **Daniel Cremers**[2,3]

[1]DeepScenario [2]TU Munich [3]Munich Center of Machine Learning
oussema.dhaouadi@tum.de

Abstract

Accurate visual localization from aerial views is a fundamental problem with applications in mapping, large-area inspection, and search-and-rescue operations. In many scenarios, these systems require high-precision localization while operating with limited resources (e.g., no internet connection or GNSS/GPS support), making large image databases or heavy 3D models impractical. Surprisingly, little attention has been given to leveraging orthographic geodata as an alternative paradigm, which is lightweight and increasingly available through free releases by governmental authorities (e.g., the European Union). To fill this gap, we propose OrthoLoC, the first large-scale dataset comprising 16,425 UAV images from Germany and the United States with multiple modalities. The dataset addresses domain shifts between UAV imagery and geospatial data. Its paired structure enables fair benchmarking of existing solutions by decoupling image retrieval from feature matching, allowing isolated evaluation of localization and calibration performance. Through comprehensive evaluation, we examine the impact of domain shifts, data resolutions, and covisibility on localization accuracy. Finally, we introduce a refinement technique called AdHoP, which can be integrated with any feature matcher, improving matching by up to 95% and reducing translation error by up to 63%. The dataset and code are available at: https://deepscenario.github.io/OrthoLoC.

1 Introduction

Visual localization for Unmanned Aerial Vehicles (UAVs) is essential for digital-twin modeling [60, 74], surveillance [29], search-and-rescue [51], and infrastructure inspection [34], yet faces unique challenges not addressed by ground-level localization systems. While ground-level approaches [56, 71, 70] benefit from similar viewpoints between images [59, 49, 57], aerial applications encounter dramatic perspective differences and require scalability over large areas [69, 72].

Current UAV localization algorithms rely on retrieving the closest match from a database of posed images [72, 77], which is inaccurate, or on 3D models of the scene [69, 66], which are memory and computationally expensive. In limited resources settings, as it is often the case for connectivity-limited environments, this can result in accuracy degradation. Recent approaches like LoDLoc [78] improve storage efficiency by using Level-of-Detail (LoD) but still assume unchanged environments, perform poorly in building-sparse areas such as highways, and its initialization depends on positioning sensors.

In contrast, a compelling solution involves geodata, such as orthographic aerial views (Digital Orthophotos (DOPs)) and elevation maps (Digital Surface Models (DSMs)). These provide a reliable, lightweight source for localizing UAV images, as shown in Figure 1. Such data is increasingly accessible through free releases from European government geoportals [46, 17], and where public access is limited, can be synthesized using photogrammetric tools [20]. Geodata are scalable and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

*Corresponding Author.

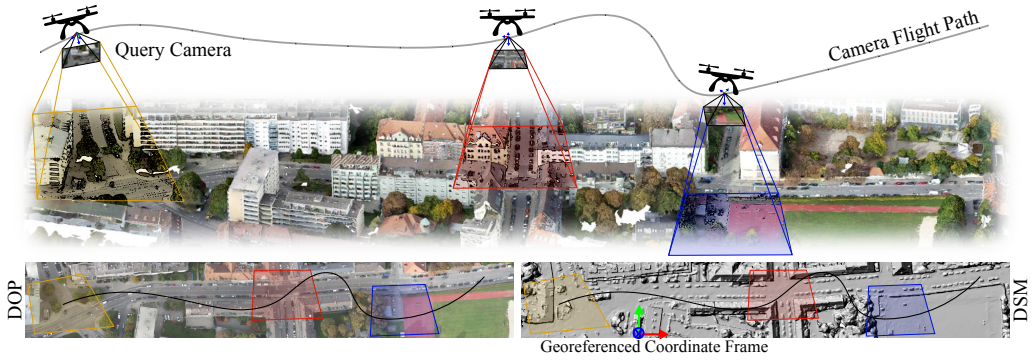


Figure 1: **Georeferenced UAV Localization / Calibration with Orthographic Geodata.** Our framework bridges the aerial-to-orthographic domain gap. It enables precise 6-DoF localization and calibration using only DOP and DSM geodata. This approach works even in GNSS-denied environments without requiring expensive 3D models or image databases.

better suited for low-resource settings. For example, covering an area of approximately 0.265 km² would require a 3D model of around 8 GB [69], whereas geodata requires about 30 times less memory. Surprisingly, no existing UAV localization approach seems to fully leverage these data sources. We believe this is mainly due to the absence of aligned cross-domain datasets and the lack of full-pose paired large-scale benchmarks specifically designed for localization using these types of geodata.

To fill this gap, we capture and release the Orthographic Aerial Localization and Calibration Dataset (OrthoLoC). It comprises 5 main modalities such as UAV imagery, DOPs, DSMs, 3D point maps, and 3D meshes with a total of 16.4K images captured in 47 regions in 19 cities across 2 countries. Our dataset is the first to offer three key advantages: (1) paired UAV-geodata structure that decouples pose estimation from image retrieval, eliminating confounding error sources in the evaluations; (2) precise 6-DoF poses obtained through multi-view georeferenced photogrammetric reconstruction; and (3) additional reference data sources to increase the domain gaps in the dataset.

We have evaluated state-of-the-art methods on this novel localization and calibration task in a comprehensive benchmark. Additionally, we introduce a method-agnostic refinement technique called *Adaptive Homography Preconditioning (AdHoP)* that further improves localization and calibration accuracy. The technique exploits the uniform structure of DOPs to perform homography-based warping by assuming quasi-planar surfaces common in built environments.

Our evaluation reveals several insights. First, state-of-the-art matching algorithms can generalize to aerial perspectives but struggle with the substantial domain gap between perspective UAV imagery and orthographic reference data. Second, our *AdHoP* technique significantly reduces the perspective disparity, improving all metrics across the tested methods, particularly achieving up to 95% and 63% enhancements in matching and translation accuracy, respectively. Third, camera calibration in aerial settings presents unique challenges due to fundamental geometric ambiguities that affect parameters estimation. Finally, reference data characteristics including domain shifts, data resolutions, and covisibility. significantly impact localization performance, with higher resolution geodata providing improvement in accuracy.

The main contributions of this paper are: (1) OrthoLoC, the first UAV dataset providing alignment with geodata across multiple modalities and locations; (2) a unified benchmarking framework for UAV localization and calibration that integrates with state-of-the-art matching algorithms and includes our AdHoP technique for addressing perspective disparities; and (3) benchmarking results for camera localization and calibration and an analysis of performance factors including cross-domain challenges, data resolution effects, and covisibility.

2 Related Work

2.1 UAV Localization Datasets

The advancement of UAV localization research has been hampered by dataset limitations. Most existing collections fail to support comprehensive 6-DoF evaluation due to several shortcomings. Datasets such as University-1652 [77] and DenseUAV [18] provide only *partial pose information*

(typically 2-DoF or 3-DoF), insufficient for applications requiring complete 6-DoF estimation. Collections derived from Google Earth [73, 9, 52] predominantly feature nadir views, exhibiting *limited viewpoint diversity* that fails to capture oblique perspectives common in practical UAV operations. Several datasets incorporate *synthetic data*—either entirely synthetic environments [40, 32] or synthetically rendered views [68, 66]—introducing domain gaps that affect generalization to real-world scenarios.

Most importantly, existing datasets lack *integrated geodata resources* crucial for evaluating localization methods leveraging lightweight orthographic representations. While the concurrent AnyVisLoc [72] dataset includes orthographic geodata, its primary pose evaluation focus is on 3-DoF rather than full 6-DoF. Additionally, it presents a misalignment between its low-resolution satellite imagery and aerial photogrammetry data, which compromises effective evaluation of cross-domain geodata-based localization. We illustrate this misalignment in the supplementary material.

In contrast, OrthoLoC provides complete 6-DoF ground-truth poses with calibration information, diverse viewpoints across multiple altitudes and angles, real-world imagery from different geographic environments, carefully aligned high-resolution geodata, and paired structure that facilitates isolated evaluation of localization algorithms, independent of retrieval errors. This comprehensive design establishes a foundation for decoupled evaluation of UAV localization and calibration methods using lightweight orthographic references, filling a critical research gap.

2.2 Visual Localization

Image retrieval-based localization. Image retrieval methods [2, 1, 50, 61] use global descriptors to match query images against geo-tagged databases. CNN-based approaches such as NetVLAD [1] and Dislocation [2] are efficient but struggle with large viewpoint and illumination changes in UAV imagery. Recent works [25, 31, 26] mitigate these issues through view synthesis and self-supervised learning, yet performance drops under extreme perspective shifts. Chen et al. [10] introduced ComplexUAV, a high-resolution UAV dataset covering diverse terrains, along with a contrastive learning framework that improves retrieval robustness and generalization. Nonetheless, retrieval-based methods remain insufficient for accurate 6-DoF UAV localization, motivating alternatives that leverage geodata directly.

Matching-based localization. Structure-based methods typically build a 3D model using Structure from Motion (SfM) techniques [53] and establish 2D–3D correspondences, either using mesh models [7, 76, 45, 69] or dense depth maps [66]. Pose estimation is then performed via PnP algorithms [28, 24, 44, 36] coupled with RANSAC optimization [14, 13, 5, 6, 3]. Recent advances in feature matching have produced three primary categories of matchers: dense matchers (e.g., DKM [21], ROMA [23]), semi-dense matchers (e.g., LoFTR [57], eLoFTR [65], XoFTR [62]), and sparse matchers (e.g., SuperGlue [49], DeDoDe [22], XFeat [48]). Geometry-aware techniques such as MASt3R [38] and DUSt3R [64] further improve matching by integrating geometric constraints. While these state-of-the-art matchers provide robust performance across many scenarios, their effectiveness with orthographic geodata remains unexplored until now.

UAV-specific localization. Aerial vehicle positioning systems have evolved from 2-DoF to 6-DoF approaches to address specific challenges. Early CNN-based techniques employed multiscale block attention [82] and capsule networks [80], while recent transformer-based frameworks integrate semantic guidance [81] and relation-aware global attention [18, 58, 39] to address scale variations and urban uncertainties. However, most of these methods target only 2-DoF or 3-DoF localization rather than full 6-DoF pose estimation required for advanced applications.

For extended pose estimation, several approaches have emerged with increasing degrees of freedom. For 4-DoF estimation, methods align UAV observations with rendered or autoencoded satellite imagery [47, 4]. 5-DoF methods employ dual Siamese networks with visual odometry and Kalman filtering [55]. Recent 6-DoF frameworks leverage curriculum learning [30], viewpoint-robust feature extraction [12], attention-based architectures [27], visibility-aware registration [11], and photorealistic synthetic data [66, 68]. While all these methods depend on complete 3D models that require extensive manual effort to create, our approach utilizes widely accessible geodata for 6-DoF pose estimation and camera calibration. Benchmarking results demonstrate accurate localization despite temporal gaps between geodata acquisition and UAV flight. This simplifies deployment by utilizing

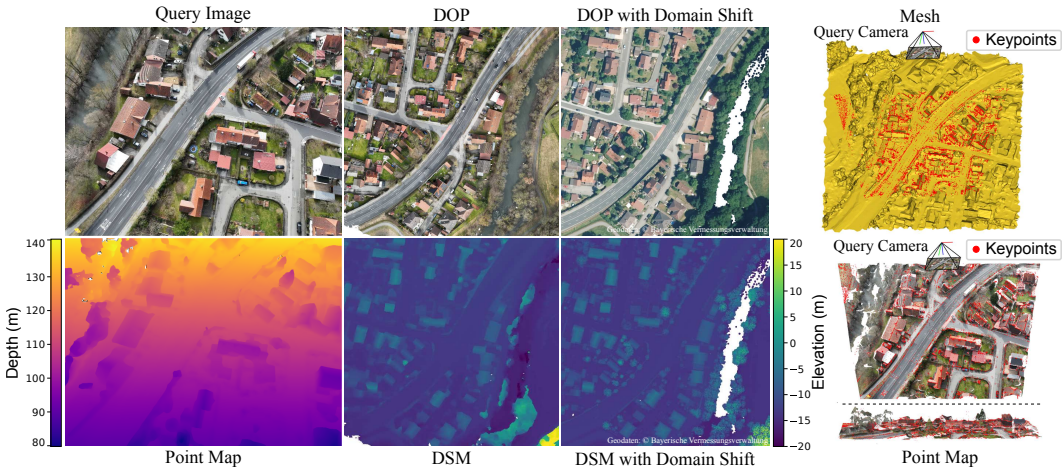


Figure 2: **Data Modalities in OrthoLoC.** Each sample includes a query image, a point map (represented as a depth map), a local mesh, visible 3D keypoints, and photogrammetrically reconstructed DOP/DSM. The dataset also includes an augmented version of DOP/DSM derived from secondary sources, introducing domain gaps for increased variability.

standardized, government-provided resources rather than requiring custom 3D reconstruction for each operational area.

3 The OrthoLoC Dataset

We introduce a comprehensive UAV localization dataset that addresses key limitations in existing benchmarks. Our dataset comprises 16.4k real UAV images spanning 47 locations across 19 cities in Germany and the United States, captured in diverse environmental contexts including urban, suburban, rural, and highway scenes. Each sample provides a query image with precise ground-truth 6-DoF pose, camera intrinsics, and rich 3D scene representations: point maps, 3D keypoints, local meshes, and aligned 2.5D geodata rasters derived from multiple sources. Figure 2 illustrates the data modalities in our dataset. Figure 3 presents the complete creation pipeline. Dataset details are provided in the supplementary material.

3.1 Data Acquisition and Processing

Data collection employed commercial drones equipped with Global Positioning System (GPS). For each location, we performed 3D scene reconstruction using SfM and Multi-View Stereo (MVS) techniques to generate camera poses, dense point clouds, and textured meshes. The reconstructions were georeferenced using Real-Time Kinematic (RTK) measurements or manually annotated Ground Control Points (GCPs) to ensure precise spatial alignment.

From these reconstructions, we generated orthographic DOPs via camera renderings and DSMs through rasterization at 5 cm/pixel resolution. We complemented these with SIFT [42] keypoints extracted from the DOP and lifted to 3D using corresponding DSM elevations, providing reliable landmarks for pose verification.

3.2 Data Pairing

To recover the pose and intrinsic parameters, visual localization methods often require solving image retrieval before running the proper estimation algorithm. These two steps are coupled, making it difficult to disentangle the contribution of each component. Hence, we pair the query with reference data to isolate the contribution of different components and evaluate localization algorithms independent of retrieval performance.

To achieve this, we establish precise correspondences by ray-tracing from each query image with known camera parameters onto the 3D mesh model and exact cropping regions in the DOP and DSM that geometrically align with the query viewpoint. To quantify how positional uncertainty affects

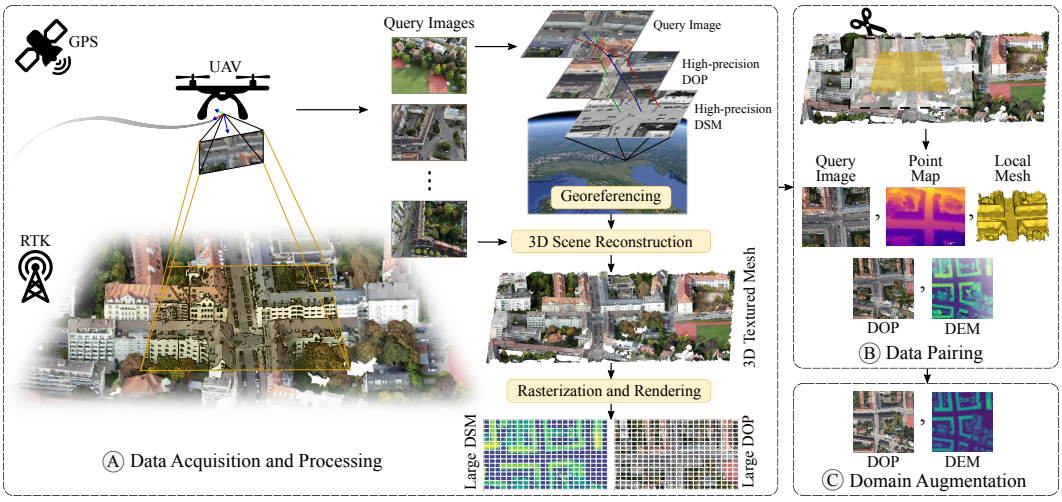


Figure 3: **Dataset Creation Pipeline.** First, (A) data acquisition involves UAV imagery collection. This data, combined with georeferencing techniques like GCPs and RTK, reconstructs a georeferenced 3D textured mesh. Subsequently, geodata is derived through rasterization and orthographic rendering. Then, (B) data pairing identifies regions of interest for each query image via raycasting. These areas undergo random expansion, followed by cropping geometric elements to form samples. Finally, (C) the data is augmented with geodata from external sources, where spatial alignment is verified.

localization accuracy, we extend the reference area beyond the visible query region through spatial perturbations by applying random offsets of 0-10 meters that simulate realistic retrieval imprecision.

In summary, a dataset sample consists of the tuple $(I, \mathbf{P}, \mathbf{R}^{\text{DOP}}, \mathbf{R}^{\text{DSM}}, \mathbf{K}, \mathbf{T}, \mathcal{V}, \mathcal{F}, \mathcal{S})$, where $I \in \mathbb{R}^{H \times W \times 3}$ is the UAV image, $\mathbf{P} \in \mathbb{R}^{H \times W \times 3}$ is the point map, $\mathbf{R}^{\text{DOP}} \in \mathbb{R}^{H^{\text{DOP}} \times W^{\text{DOP}} \times 3}$ is the orthophoto raster, $\mathbf{R}^{\text{DSM}} \in \mathbb{R}^{H^{\text{DSM}} \times W^{\text{DSM}}}$ is the elevation raster, $\mathbf{K} \in \mathbb{R}^{3 \times 3}$ is the camera intrinsic matrix, $\mathbf{T} \in SE(3)$ is the camera pose, $\mathcal{V} \in \mathbb{R}^{N \times 3}$ represents the mesh vertices, $\mathcal{F} \in \mathbb{N}^{M \times 3}$ defines the mesh faces, and $\mathcal{S}$ is the set of 3D keypoints. All geometric elements are transformed into a local coordinate system to preserve privacy while maintaining precise geometric relationships.

3.3 Domain Augmentation

Solving UAV localization requires robustness to natural changes in scenes due to time passing. Typically, reference geodata may have been collected months or years before a UAV flight, creating significant domain gaps that cannot be easily addressed through simple data augmentation or domain adaptation techniques. These gaps are particularly challenging because they involve both appearance and structural changes that vary unpredictably across locations and seasons.

We can divide these challenges into two categories: (1) visual domain gaps in DOPs through appearance changes (color shifts, illumination variations, seasonal differences) while maintaining structural consistency; and (2) structural domain gaps in DSMs through geometric modifications (construction changes, vegetation growth, infrastructure evolution).

Including real-world domain gaps in our dataset is essential because synthetic alternatives cannot replicate the complex natural variations occurring over time. Our dataset provides three sample categories: minimal to no domain gap (i.e., same-domain) samples that include geodata from the 3D reconstruction, visual domain gaps only (i.e., cross-domain DOP), and both visual and structural disparities (i.e., cross-domain DOP and DSM). Cross-domain samples were created by incorporating open geodata from European locations and visually verifying alignment with same-domain samples.

3.4 Comparison with Existing Datasets

OrthoLoC presents the first UAV localization dataset for 6-DoF pose estimation using governmental geodata (DOPs and DSMs) as the only reference. This eliminates costly posed image databases, meshes, or point clouds, enabling real-time localization without preprocessing.

Table 1: **Comparison of Existing UAV Localization Datasets.**
Legend: *Country codes*: Switzerland (CH), China (CN), United States (US), Germany (DE); *Geographic*: Urban (U), Suburban (SU), Rural (R), Campus (C), Highway (H); *UAV images*: Real (Re), Synthetic (Sy); *View*: top-down (nadir), angled (oblique), mixed views (both); *Altitude*: $\leq$150 m (low), >150 m (high), mixed altitudes (both); *3D*: Depth (D), Point Map (PM), Level of Detail (LoD); *Task*: Image Retrieval (IR); *Platform*: + indicates georeferencing techniques (RTK, GCP); *XD*: cross-domain (reference data are from external sources).

	Geographic Coverage			UAV Data					Reference Data				
Dataset	**Country**	**Scene**	**#Loc**	**Imgs (Re+Sy)**	**View**	**Alt**	**3D**	**Platform**	**Amount**	**Type**	**3D**	**XD**	**Task**
Unpaired													
MatrixCity [40]$_{2023}$	-	U	1	0+519k	oblique	low	D	virtual	✗	✗	✗	✗	6-DoF
CrossLoc [68]$_{2022}$	CH	U	2	4.5k+19.5k	both	low	D/PM	drone+	1	✗	✗	✗	6-DoF
AirLoc [69]$_{2023}$	CN	U	1	2.7k+0	both	low	✗	drone+	✗	✗	Mesh	✗	6-DoF
UAVD4L [66]$_{2024}$	CN	U	2	0.9k+18k	both	low	D	drone+	1	✗	Mesh/DSM	✗	6-DoF
Swiss-EPFL [78]$_{2024}$	CH	U	2	2.2k+14.7k	both	low	✗	drone+	2	✗	LoD	✗	6-DoF
UAVD4L-LoD [78]$_{2024}$	CN	U	2	3.7k+18k	both	low	✗	drone+	1	✗	LoD	✗	6-DoF
UAV-VisLoc [67]$_{2024}$	CN	U	11	6.7k+0	nadir	high	✗	drone+	11	DOP	✗	✗	IR
GTA-UAV [32]$_{2025}$	-	U	1	0+33k	nadir	both	✗	virtual	✗	✗	✗	✗	6-DoF
AnyVisLoc [72]$_{2025}$	CN	U,R,SU	25	18k+0	both	both	✗	drone+	25	DOP	DSM	✓	3-DoF
Paired													
University-1652 [77]$_{2020}$	US	C	39	701+50.2k	oblique	both	✗	web	951	Images	✗	✓	IR
DenseUAV [18]$_{2023}$	CN	C	14	9k+0	nadir	low	✗	drone	18k	Images	✗	✓	3-DoF
SUES-200 [79]$_{2023}$	CN	U	200	40k+0	both	high	✗	drone	200	DOP	✗	✓	IR
ALTO [15]$_{2022}$	US	U,R,SU	1	15.4k+0	nadir	high	✗	aircraft+	16.5k	DOP	LiDAR	✗	6-DoF
VPAIR [52]$_{2022}$	DE	U,R,SU	1	2.7k+0	nadir	high	✗	aircraft+	2.7k	Images	Depth	✓	6-DoF
OrthoLoC (Ours)	US,DE	U,R,SU,H	47	16.4k+0	both	both	PM	drone+	16.4k	DOP	DSM	✓	6-DoF

Our dataset spans 47 locations across 2 countries with 16.4 k real UAV images, paired multi-modal data (DOPs, DSMs, and 3D reconstructions), diverse viewpoints from nadir to oblique perspectives, and high-precision ground-truth achieving approximately 5 cm median error via GCPs evaluation. Over 4 k governmental orthoimages and surface models enable robust domain adaptation assessment. OrthoLoC uniquely provides aligned DOP+DSM pairs with accurate 6-DoF poses across multiple altitudes and privacy-preserving georeferencing decoupling.

As shown in Table 1, existing datasets suffer from (1) restricted geographic coverage [68, 52, 15, 40, 69, 66, 78], (2) synthetic data dependency [77, 40, 32, 68, 78], or (3) incomplete pose information [77, 79, 18, 67]. Our geometric consistency analysis reveals significant projection errors in CrossLoc [68], UAVD4L [66], and AnyVisLoc [72], which provides only 3-DoF poses with misaligned reference data. Assessment details are in the supplementary material.

4 Localization with Orthographic Geodata

Unlike traditional approaches that rely on image retrieval or 3D models, we explore the novel paradigm of UAV localization using 2.5D orthographic geodata. No existing methods are directly applicable to this scenario, as previous work has not leveraged the combination of DOPs and DSMs for UAV pose estimation. This section presents the problem formulation, our benchmarking framework, and our refinement technique.

4.1 Problem Formulation

Goal. Given an orthophoto raster $\mathbf{R}^{\text{DOP}}$, an elevation raster $\mathbf{R}^{\text{DSM}}$, and a query UAV image I taken from an arbitrary viewpoint, we aim to *determine the georeferenced 6-DoF pose* $\mathbf{T}$ *of the camera (localization) and, optionally, its intrinsic parameters* $\mathbf{K}$ *(calibration).*

Challenges. The key challenge is bridging two fundamentally different projection models: perspective projection for UAV imagery and orthographic projection for geodata. This difference creates a domain gap that is particularly pronounced in oblique views where perspective distortion is significant. Additionally, another domain gap arises from the visual and structural discrepancies between the query and reference data caused by differences in acquisition time. We provide the mathematical principles for both projection types, with particular focus on deriving a formulation for nadir orthographic projection in the supplementary material.

Benchmarking framework. Given the absence of existing methods that directly tackle UAV localization using orthographic geodata, we propose a comprehensive benchmarking framework

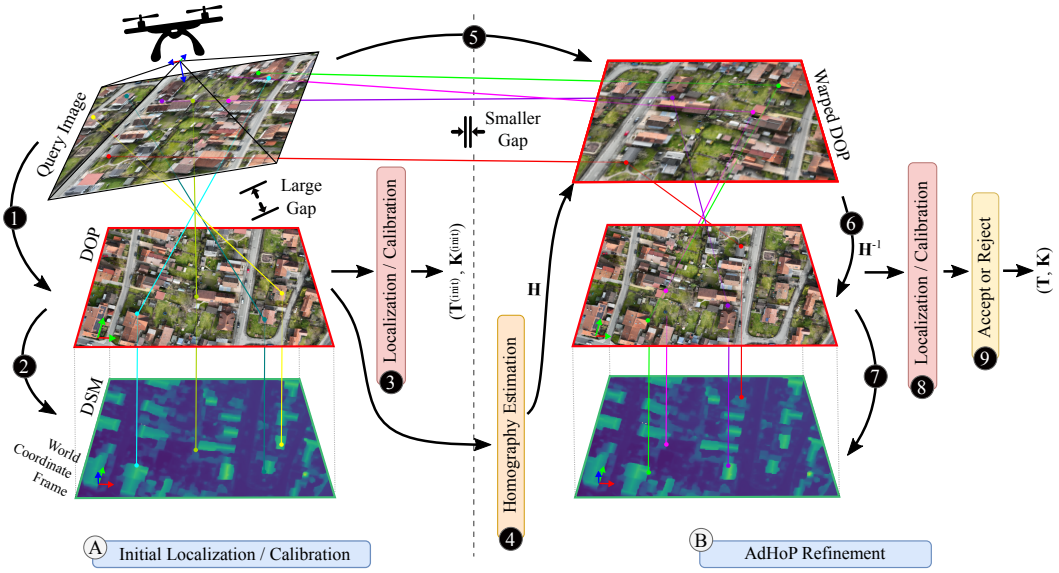


Figure 4: **UAV 6-DoF Localization and Calibration with AdHoP:** (A) *Initial Localization / Calibration:* We match features between the query image and DOP (1), lift the correspondences to 3D using the DSM (2), and compute an initial pose and optional intrinsics (3). (B) *AdHoP Refinement*: Using the initial 2D-2D correspondences, we estimate a homography to warp the DOP (4), thereby reducing perspective differences. This enables enhanced feature matching on the warped orthophoto (5). The new correspondences are then mapped back to the original unwarped coordinate space (6), lifted to 3D using the DSM (7), and used to compute refined camera parameters (8). The refinement is accepted only when it reduces the reprojection error (9).

to evaluate various combinations of matching algorithms as backbones. Our framework is entirely backbone-agnostic, enabling integration with any feature matching method, as illustrated in Figure 4 and detailed in the following subsections.

4.2 Initial Camera Calibration / Localization

We establish 2D-2D correspondences between the query image I and the orthophoto $\mathbf{R}^{\text{DOP}}$ using state-of-the-art matching methods such as GIM+DKM [54], RoMA [23], SuperGlue [49], and LoFTR [57]. Each 2D point matched in $\mathbf{R}^{\text{DOP}}$ is lifted to a 3D point using the corresponding elevation value from $\mathbf{R}^{\text{DSM}}$, providing the necessary 3D-2D correspondences for pose estimation (details in the supplementary material). Next, we filter correspondences by excluding matches with low confidence scores (below 0.5), invalid 3D points (missing data values), and points outside the field of view. Our calibration approach employs a two-stage optimization strategy. In the first stage, we use an initial guess of the focal length to estimate the camera pose by optimizing reprojection errors using RANSAC-EPnP [37] and a 5-pixel inlier threshold. In the second stage, we use this pose to initialize a Levenberg-Marquardt optimization that jointly refines camera intrinsics and extrinsics. For pure localization tasks, we only perform the first stage, as intrinsics are assumed to be known.

4.3 AdHoP Refinement

Perspective differences between query and reference images are a major challenge in UAV localization, especially for oblique viewpoints. Our geodata-based approach addresses this with Adaptive Homography Preconditioning (AdHoP), a method-agnostic refinement technique that exploits the approximate planarity of many aerial elements (roads, building roofs, fields). Formally, AdHoP estimates a homography matrix $\mathbf{H} \in \mathbb{R}^{3\times3}$ from initial 2D–2D correspondences using normalized Direct Linear Transform (DLT) with RANSAC. We adopt this straightforward formulation to avoid the complexity and biases of learning-based methods, requiring no training, dataset dependencies, or ad-hoc domain assumptions while providing a transparent and general baseline. The homography warps the orthophoto to better match the query perspective, enabling a second round of feature matching with improved similarity. The new matches are mapped back using $\mathbf{H}^{-1}$, lifted to 3D via

Table 2: **Quantitative Localization Results on OrthoLoC Test Sets.** Rankings between matchers are highlighted as first, second, and third. **Bold** values indicate the best performance comparing without/with AdHoP. RI indicates a rotation-invariant matcher (matching performed with 4 rotated versions, selecting the one with most correspondences). Abbreviations: SuperPoint (SP), SuperGlue (SG), LightGlue (LG), Minima (MM).

Matcher	RI	ME [px]↓	TE [m]↓	RE [°]↓	RPE [px]↓	1m-1° [%]↑	3m-3° [%]↑	5m-5° [%]↑	Speed [s]↓
SP+SG [19, 49]	✗	**2.2** / **2.2**	0.36 / **0.35**	**0.15** / **0.15**	**2.8** / **2.8**	63.9 / **64.4**	77.4 / **77.6**	78.7 / **78.9**	**0.2** / 0.3
SP+LG [19, 41]	✗	**2.0** / **2.0**	**0.37** / **0.37**	0.16 / **0.15**	**2.9** / **2.9**	64.0 / **64.2**	77.0 / **77.4**	78.8 / **79.0**	**0.1** / 0.2
DeDoDe [22]	✗	**1.2** / **1.2**	0.42 / **0.39**	0.18 / **0.16**	3.6 / **3.2**	27.5 / **28.2**	33.3 / **33.6**	35.6 / **35.7**	**0.3** / **0.3**
XFeat [48]	✗	257.0 / **38.1**	1.58 / **0.96**	0.74 / **0.45**	13.0 / **7.8**	42.7 / **50.8**	57.4 / **63.0**	61.2 / **65.1**	**0.1** / 0.2
XFeat+LG [48, 41]	✗	4.3 / **3.2**	0.57 / **0.48**	0.25 / **0.20**	4.7 / **3.8**	42.5 / **45.7**	54.4 / **56.3**	56.3 / **57.3**	**0.1** / 0.3
LoFTR [57]	✗	317.2 / **312.9**	121.56 / **118.77**	109.49 / **107.22**	1451.9 / **1384.7**	18.0 / **21.0**	23.3 / **25.6**	23.9 / **26.3**	**0.1** / 0.2
MM+LoFTR [33, 78]	✗	**266.9** / 269.1	87.17 / **84.69**	98.89 / **97.81**	902.4 / **841.4**	14.5 / **18.2**	21.5 / **23.1**	22.5 / **23.7**	**0.3** / 0.6
eLoFTR [65]	✗	329.5 / **311.9**	124.29 / **117.53**	109.25 / **102.50**	1552.2 / **1471.1**	19.0 / **22.9**	24.0 / **27.6**	24.8 / **28.4**	**0.1** / 0.2
XoFTR [62]	✗	291.7 / **285.9**	113.65 / **113.15**	**107.24** / 107.65	1322.4 / **1275.2**	19.7 / **21.5**	23.8 / **24.8**	24.1 / **25.4**	**0.1** / 0.2
DKM [21]	✓	8.8 / **2.7**	3.83 / **1.40**	1.93 / **0.63**	31.9 / **11.9**	33.6 / **42.2**	44.2 / **49.8**	45.6 / **50.4**	**0.8** / 1.7
XFeat* [48]	✗	222.2 / **9.2**	1.07 / **0.66**	0.48 / **0.30**	8.8 / **5.4**	48.8 / **59.8**	67.3 / **72.3**	70.4 / **73.7**	**0.1** / 0.2
GIM+DKM [54, 21]	✓	1.5 / **1.3**	0.40 / **0.32**	**0.12** / **0.12**	3.1 / **2.6**	74.1 / **75.4**	86.6 / **87.9**	87.4 / **88.4**	**1.3** / 2.6
DUSt3R [64]	✓	5.0 / **4.9**	**3.45** / 3.68	**1.47** / 1.53	**25.8** / 27.3	3.6 / **6.4**	33.6 / **33.8**	**51.7** / 49.8	**1.5** / 2.1
MASt3R [38]	✓	2.4 / **2.3**	0.61 / **0.60**	0.28 / **0.26**	5.0 / **4.8**	62.4 / **63.5**	81.4 / **82.0**	84.2 / **84.5**	**2.2** / 3.4
RoMa [23]	✓	21.6 / **2.4**	1.47 / **0.75**	0.67 / **0.32**	12.5 / **6.2**	44.4 / **54.6**	56.1 / **65.1**	59.2 / **66.8**	**1.1** / 2.1
MM+RoMa [33, 23]	✓	70.8 / **4.6**	3.63 / **1.21**	1.92 / **0.55**	34.2 / **9.9**	38.6 / **47.9**	48.9 / **58.0**	51.6 / **59.5**	**1.1** / 2.1

the DSM, and used to refine pose estimation. The refinement is accepted only if it reduces mean reprojection error.

In our experiments, we demonstrate that combining different matching algorithms with AdHoP significantly improves localization and calibration accuracy, with GIM+DKM+AdHoP emerging as the most effective combination across diverse scenarios. This approach highlights the practical advantages of integrating geodata into localization pipelines.

5 Experimental Results

In this section, we introduce the evaluation metrics (Section 5.1) and present our benchmarking results. We evaluate both localization (Section 5.2) and calibration performance (Section 5.3) using state-of-the-art feature matchers as backbones. We then analyze some factors affecting performance across different scenarios (Section 5.4). For complete experimental results and additional analyses, please refer to the supplementary material.

5.1 Evaluation Metrics

We report several metrics: Matching Error (ME) in pixels as the median distance between ground-truth and estimated matching coordinates; Translation Error (TE) in meters and Rotation Error (RE) in degrees for pose accuracy; Reprojection Error (RPE) in pixels for keypoint reprojection errors; recall percentages at thresholds 1m-1°, 3m-3°, and 5m-5°; and Relative Focal Length Error (RFE) in percent for calibration accuracy. We also report computation time in seconds.

5.2 Camera Localization

We report localization performance when using state-of-the-art feature matching approaches with and without our AdHoP strategy in Table 2. GIM+DKM [54, 21] achieves the highest performance across the majority of metrics. SP+SG [19, 49] and SP+LG [19, 41], along with GIM+DKM, all achieve precise localization below 40 cm and 0.16 degrees. However, the sparse matchers (SP-based) have notably lower recall compared to GIM+DKM, indicating they successfully localize fewer images. Semi-dense approaches like LoFTR [57] and XoFTR [62] perform poorly, with recall below 19.7%. Our intuition is that these approaches suffer from limited training datasets or architectural constraints that prevent handling large domain shifts.

Integrating AdHoP substantially improves performance across all matchers. We observe an average matching improvement of about 30%, yielding translational and rotational error reductions of 20% each. The best-performing GIM+DKM [54, 21] with AdHoP reduces translation error by 20%, from 0.40 m to 0.32 m. Previously underperforming methods show even more dramatic improvements:

XFeat* [48] matching error decreases by 95.86%, DKM [21] reduces translation error by 63%, and RoMa [23] increases 1m-1° recall by 23% while reducing translation error by half. We illustrate in Figure 5 the impact of AdHoP in reducing errors.

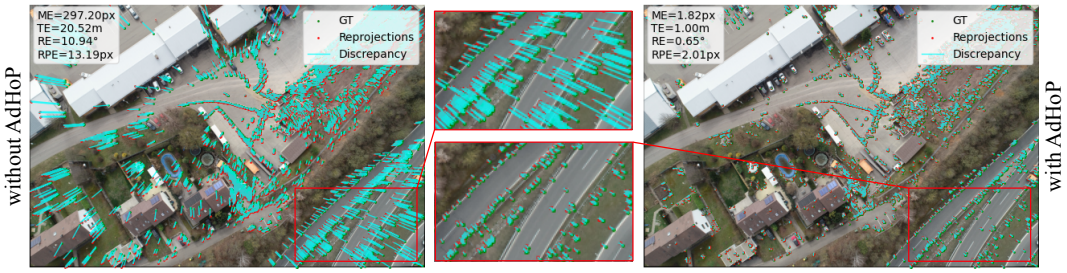


Figure 5: **Localization Without and With AdHoP.** xFeat* [48] matching results showing 3D keypoint projections in green (using the ground-truth pose) and red (using the estimated pose). Blue lines indicate projection discrepancies between estimated and ground-truth positions.

5.3 Camera Calibration

Our calibration experiments reveal a fundamental challenge in estimating camera intrinsics from UAV imagery due to geometric ambiguity between focal length and translation estimation. We provide mathematical proof of this ambiguity and detailed calibration benchmarking results in the supplementary material. Despite this challenge, the combination of GIM+DKM [54, 21] with our AdHoP technique achieves the best focal length estimation with just 1.6% relative error with a translation error of 2.09 m. However, the recall remains relatively low at 21.8%, highlighting the inherent difficulty of the calibration task.

5.4 Performance Factors

Domain shift. Using cross-domain DOPs affects algorithms differently, with varying robustness to appearance changes. Even the best-performing method, GIM+DKM [54, 21] with AdHoP, shows a threefold increase in translational error under these conditions. Further domain shift, combining both cross-domain DOPs and DSMs, causes significant degradation across all methods. For instance, GIM+DKM [54, 21] with AdHoP experiences a sevenfold increase in translation error, rising from 0.16 m to 1.12 m. These findings highlight the need for localization algorithms robust against both visual and geometric domain shifts. Complete results are provided in the supplementary material.

Resolution and covisibility impact. Performance remains robust when scaling images to 30% of original size ($\sim$ 300 pixels, with highest geodata resolution at 13 m/pixel), with degradation occurring only at lower resolutions. Additionally, localization accuracy depends heavily on the covisibility ratio between query and reference images, dropping significantly when less than 20% of the elements seen in the query image are visible in the reference data. These findings have important implications for real-world UAV deployment, where error-prone upstream tasks like image retrieval may result in extraction of incomplete reference data. A detailed analysis of these factors is in the supplementary.

6 Conclusion

We presented a novel paradigm for UAV visual localization using widely available geodata. To support this approach, we introduced OrthoLoC, a diverse large-scale UAV localization and calibration dataset spanning multiple environments and regions. Our benchmarking framework matches UAV imagery with orthophotos, which are lifted to 3D using 2.5D elevation models to solve pose estimation via PnP. Our proposed AdHoP technique consistently enhances various matching algorithms, yielding significant improvements in both pose estimation and camera calibration performance.

Our benchmarking demonstrated that standard 2.5D geodata proved sufficient for accurate 6-DoF pose estimation in outdoor UAV localization. Our evaluations revealed that dense matchers, specifically GIM+DKM [54, 21] with AdHoP, achieved 75.4% recall at 1m-1° threshold, though with limited robustness to domain shifts. Camera calibration performance remained challenging due to inherent

geometric ambiguities between focal length and translation estimation. We also demonstrated that higher resolution data significantly improved localization accuracy, confirming that low-resolution reference data (such as satellite imagery) limited performance. Moreover, we found that the area coverage of geodata—typically determined by upstream tasks like region of interest detection through image retrieval—critically affected correspondence distribution and reliable pose estimation.

Limitations and future work. Calibration remains affected by translation and focal length ambiguities, which could be addressed by training end-to-end networks for improved localization. While our framework shows strong performance, it requires geodata with at least 20% covisibility and does not yet scale to large rasters without region detection, a key factor for real-world deployment. Moreover, AdHoP improves partially incorrect correspondences but fails when matches are completely corrupted. This highlights the potential of more advanced homography estimators, particularly learning-based approaches, to reduce reliance on initial matching and increase robustness.

Acknowledgement. This work is a result of the joint research project STADT:up. The project is supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), based on a decision of the German Bundestag. The author is solely responsible for the content of this publication.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] R. Arandjelović and A. Zisserman. Dislocation: Scalable descriptor distinctiveness for location recognition. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12*, pages 188–204. Springer, 2015.

[3] D. Barath, J. Matas, and J. Noskova. Magsac: marginalizing sample consensus. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10197–10205, 2019.

[4] M. Bianchi and T. D. Barfoot. UAV localization using autoencoded satellite images. *CoRR*, abs/2102.05692, 2021.

[5] E. Brachmann and C. Rother. Expert sample consensus applied to camera re-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7525–7534, 2019.

[6] E. Brachmann and C. Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4322–4331, 2019.

[7] J. Brejcha, M. Lukáč, Y. Hold-Geoffroy, O. Wang, and M. Čadík. Landscapear: Large scale outdoor augmented reality by matching photographs with terrain models using learned descriptors. In *European Conference on Computer Vision*, pages 295–312. Springer, 2020.

[8] D. Cernea. OpenMVS: Multi-view stereo reconstruction library. `https://github.com/cdcseacave/openMVS`, 2020.

[9] C.-H. Chang, C.-N. Chou, and E. Y. Chang. Clkn: Cascaded lucas-kanade networks for image alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2213–2221, 2017.

[10] J. Chen, G. Wen, H. Jian, and X. Fan. A visual localization benchmark for uavs in complex multi-terrain environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[11] S. Chen, X. Wu, M. W. Mueller, and K. Sreenath. Real-time geo-localization using satellite imagery and topography for unmanned aerial vehicles. In *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 2275–2281. IEEE, 2021.

[12] Y. Chen and J. Jiang. An oblique-robust absolute visual localization method for gps-denied uav with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2023.

[13] O. Chum and J. Matas. Optimal randomized ransac. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1472–1482, 2008.

[14] O. Chum, J. Matas, and J. Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pages 236–243. Springer, 2003.

[15] I. Cisneros, P. Yin, J. Zhang, H. Choset, and S. Scherer. Alto: A large-scale dataset for uav visual place recognition and localization. *arXiv preprint arXiv:2207.12317*, 2022.

[16] E. Cledat, L. V. Jospin, D. A. Cucci, and J. Skaloud. Mapping quality prediction for rtk/ppk-equipped micro-drones operating in complex natural environment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:24–38, 2020.

[17] E. Commission. Commission implementing regulation (eu) 2023/138 laying down a list of specific high-value datasets and the arrangements for their publication and re-use, 2023. Accessed: 2025-02-20.

[18] M. Dai, E. Zheng, Z. Feng, L. Qi, J. Zhuang, and W. Yang. Vision-based uav self-positioning in low-altitude urban environments. *IEEE Transactions on Image Processing*, 33:493–508, 2023.

[19] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*,

pages 224–236, 2018.
[20] DJI. Dji terra. https://enterprise.dji.com/dji-terra.
[21] J. Edstedt, I. Athanasiadis, M. Wadenbäck, and M. Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023.
[22] J. Edstedt, G. Bökman, M. Wadenbäck, and M. Felsberg. Dedode: Detect, don't describe—describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision (3DV)*, pages 148–157. IEEE, 2024.
[23] J. Edstedt, Q. Sun, G. Bökman, M. Wadenbäck, and M. Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024.
[24] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng. Complete solution classification for the perspective-three-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 25(8):930–943, 2003.
[25] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 369–386. Springer, 2020.
[26] A. Gordo, J. Almazan, J. Revaud, and D. Larlus. End-to-end learning of deep visual representations for image retrieval. *International Journal of Computer Vision*, 124(2):237–254, 2017.
[27] W. Hanyu, S. Qiang, D. Zilong, C. Xinyi, and X. Wang. Absolute pose estimation of uav based on large-scale satellite image. *Chinese Journal of Aeronautics*, 37(6):219–231, 2024.
[28] B. M. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *International journal of computer vision*, 13:331–356, 1994.
[29] J. C. Hodgson, S. M. Baylis, R. Mott, A. Herrod, and R. H. Clarke. Precision wildlife monitoring using unmanned aerial vehicles. *Scientific reports*, 6(1):22574, 2016.
[30] B. Hu, L. Chen, R. Chen, S. Bu, P. Han, and H. Li. Curriculumloc: Enhancing cross-domain geolocalization through multi-stage refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
[31] M. Humenberger, Y. Cabon, N. Guerin, J. Morat, V. Leroy, J. Revaud, P. Rerole, N. Pion, C. de Souza, and G. Csurka. Robust image retrieval-based visual localization using kapture. *arXiv preprint arXiv:2007.13867*, 2020.
[32] Y. Ji, B. He, Z. Tan, and L. Wu. Game4loc: A uav geo-localization benchmark from game data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3913–3921, 2025.
[33] X. Jiang, J. Ren, Z. Li, X. Zhou, D. Liang, and X. Bai. Minima: Modality invariant image matching. *arXiv preprint arXiv:2412.19412*, 2024.
[34] S. Jordan, J. Moore, S. Hovet, J. Box, J. Perry, K. Kirsche, D. Lewis, and Z. T. H. Tse. State-of-the-art technologies for uav inspections. *IET Radar, Sonar & Navigation*, 12(2):151–164, 2018.
[35] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013.
[36] L. Kneip, D. Scaramuzza, and R. Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR 2011*, pages 2969–2976. IEEE, 2011.
[37] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *International journal of computer vision*, 81:155–166, 2009.
[38] V. Leroy, Y. Cabon, and J. Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.
[39] Q. Li, X. Yang, J. Fan, R. Lu, B. Tang, S. Wang, and S. Su. Geoformer: An effective transformer-based siamese network for uav geo-localization. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.
[40] Y. Li, L. Jiang, L. Xu, Y. Xiangli, Z. Wang, D. Lin, and B. Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023.
[41] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023.
[42] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.
[43] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.
[44] F. Moreno-Noguer, V. Lepetit, and P. Fua. Accurate non-iterative o(n) solution to the pnp problem. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. Ieee, 2007.
[45] V. Panek, Z. Kukelova, and T. Sattler. Meshloc: Mesh-based visual localization. In *European Conference on Computer Vision*, pages 589–609. Springer, 2022.
[46] E. Parliament and C. of the European Union. Directive (eu) 2019/1024 on open data and the re-use of public sector information, 2019. Accessed: 2025-02-20.
[47] B. Patel, T. D. Barfoot, and A. P. Schoellig. Visual localization with google earth images for robust global pose estimation of uavs. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 6491–6497. IEEE, 2020.

[48] G. Potje, F. Cadar, A. Araujo, R. Martins, and E. R. Nascimento. Xfeat: Accelerated features for lightweight image matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2682–2691, 2024.

[49] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020.

[50] T. Sattler, M. Havlena, K. Schindler, and M. Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1582–1590, 2016.

[51] J. Scherer, S. Yahyanejad, S. Hayat, E. Yanmaz, T. Andre, A. Khan, V. Vukadinovic, C. Bettstetter, H. Hellwagner, and B. Rinner. 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*, pages 33–38, 2015.

[52] M. Schleiss, F. Rouatbi, and D. Cremers. Vpair–aerial visual place recognition and localization in large-scale outdoor environments. *arXiv preprint arXiv:2205.11567*, 2022.

[53] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[54] X. Shen, Z. Cai, W. Yin, M. Müller, Z. Li, K. Wang, X. Chen, and C. Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024.

[55] A. Shetty and G. X. Gao. Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1827–1833. IEEE, 2019.

[56] Y. Shi, L. Liu, X. Yu, and H. Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019.

[57] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021.

[58] J. Sun, R. Yan, B. Zhang, B. Zhu, and F. Sun. A cross-view geo-localization method guided by relation-aware global attention. *Multimedia Systems*, 29(4):2205–2216, 2023.

[59] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii. Inloc: Indoor visual localization with dense matching and view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7199–7209, 2018.

[60] A. To, M. Liu, M. Hazeeq Bin Muhammad Hairul, J. G. Davis, J. S. Lee, H. Hesse, and H. D. Nguyen. Drone-based ai and 3d reconstruction for digital twin augmentation. In *International conference on human-computer interaction*, pages 511–529. Springer, 2021.

[61] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015.

[62] Ö. Tuzcuoğlu, A. Köksal, B. Sofu, S. Kalkan, and A. A. Alatan. Xoftr: Cross-modal feature matching transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4275–4286, 2024.

[63] P. UAB. Pixpro photogrammetry software. `https://www.pix-pro.com/`.

[64] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[65] Y. Wang, X. He, S. Peng, D. Tan, and X. Zhou. Efficient loftr: Semi-dense local feature matching with sparse-like speed. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21666–21675, 2024.

[66] R. Wu, X. Cheng, J. Zhu, Y. Liu, M. Zhang, and S. Yan. Uavd4l: A large-scale dataset for uav 6-dof localization. In *2024 International Conference on 3D Vision (3DV)*, pages 1574–1583. IEEE, 2024.

[67] W. Xu, Y. Yao, J. Cao, Z. Wei, C. Liu, J. Wang, and M. Peng. Uav-visloc: A large-scale dataset for uav visual localization. *arXiv preprint arXiv:2405.11936*, 2024.

[68] Q. Yan, J. Zheng, S. Reding, S. Li, and I. Doytchinov. Crossloc: Scalable aerial localization assisted by multimodal synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17358–17368, 2022.

[69] S. Yan, X. Cheng, Y. Liu, J. Zhu, R. Wu, Y. Liu, and M. Zhang. Render-and-compare: Cross-view 6-dof localization from noisy prior. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2171–2176. IEEE, 2023.

[70] S. Yan, Y. Liu, L. Wang, Z. Shen, Z. Peng, H. Liu, M. Zhang, G. Zhang, and X. Zhou. Long-term visual localization with mobile sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17245–17255, 2023.

[71] H. Yang, X. Lu, and Y. Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

[72] Y. Ye, X. Teng, S. Chen, Z. Li, L. Liu, Q. Yu, and T. Tan. Exploring the best way for uav visual localization under low-altitude multi-view observation condition: a benchmark. *arXiv preprint arXiv:2503.10692*, 2025.

[73] A. Yol, B. Delabarre, A. Dame, J.-E. Dartois, and E. Marchand. Vision-based absolute localization for unmanned aerial vehicles. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3429–3434. IEEE, 2014.

[74] J. Yoon, Y. Kim, S. Lee, and M. Shin. Uav-based automated 3d modeling framework using deep learning for building energy modeling. *Sustainable Cities and Society*, 101:105169, 2024.
[75] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2002.
[76] Z. Zhang, T. Sattler, and D. Scaramuzza. Reference pose generation for long-term visual localization via learned features and view synthesis. *International Journal of Computer Vision*, 129(4):821–844, 2021.
[77] Z. Zheng, Y. Wei, and Y. Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020.
[78] J. Zhu, S. Yan, L. Wang, z. shengYue, Y. Liu, and M. Zhang. Lod-loc: Aerial visual localization using lod 3d map with neural wireframe alignment. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 119063–119098. Curran Associates, Inc., 2024.
[79] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023.
[80] Y. Zhu, B. Sun, X. Lu, and S. Jia. Geographic semantic network for cross-view image geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.
[81] J. Zhuang, X. Chen, M. Dai, W. Lan, Y. Cai, and E. Zheng. A semantic guidance and transformer-based matching method for uavs and satellite images for uav geo-localization. *Ieee Access*, 10:34277–34287, 2022.
[82] J. Zhuang, M. Dai, X. Chen, and E. Zheng. A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization. *Remote Sensing*, 13(19):3979, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions of this work are clearly summarized in the abstract and the final paragraph of the introduction. They align well with the overall scope of the work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses its limitations in the conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are clearly stated, including the orthographic projection formulation and the calibration ambiguity, which are fully detailed and proven in the supplementary material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All steps for dataset creation and benchmarking are clearly described to ensure reproducibility. We also provide complete code and scripts to enable others to replicate our results and validate our findings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the dataset and code with clear instructions for easy use.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We used image matchers with their default settings from original repositories and provide all necessary details, including parameter settings, in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We followed standard localization metrics, which typically exclude error bars, and omitted them due to the high computational cost of full benchmarking.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Average execution times are reported in the results tables, and hardware details are provided in the supplementary material. CPU memory usage is not included due to the shared cluster setup, but sufficient information is given to support reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research strictly follows privacy policies and conforms with the NeurIPS Code of Ethics. Drone flights were conducted at heights preventing identification of individuals or sensitive details. We downscaled 4K images to full HD to protect privacy while maintaining quality. Georeferenced 3D reconstructions include random offsets to remove exact locations. Data from geoportals complies with copyright terms (CC BY 4.0 or CC BY-ND 4.0) and is clearly linked in the project page and repository.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This paper focuses on advancing UAV localization, essential for applications such as digital twins in infrastructure management, urban planning, and autonomous systems. These advancements facilitate efficient monitoring, planning, and critical operations like search-and-rescue and environmental inspections.

We recognize potential ethical concerns, particularly related to surveillance and use in GPS-denied or sensitive areas. However, robust sensor-based systems are expected to be favored in these contexts, reducing the risk of misuse of vision-based localization.

Our research is intended to broadly enhance UAV localization knowledge, supporting developments that prioritize societal benefits while carefully considering ethical implications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our research poses no significant risk of misuse. We rely exclusively on self-collected drone imagery compliant with aviation regulations and official data from European geoportals, avoiding any web-scraped content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: A subset of our dataset includes cross-domain samples sourced from geoportals under CC BY 4.0 or CC BY-ND 4.0 licenses for research use. All licensing details and source links are clearly documented on the project page and GitHub repository.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: All new assets are thoroughly documented, with detailed instructions and explanations provided in the supplementary materials, project page, and GitHub repository to ensure accessibility and ease of use.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing experiments were conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No crowdsourcing experiments were conducted.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We did not use LLMs as part of the core methodology in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix

Table of Contents

A Mathematical Formulation **22**
- A.1 General Camera Projection Model . . . 22
- A.2 Perspective Projection . . . 22
- A.3 Nadir Orthographic Projection . . . 22
- A.4 Lifting 2D Coordinates in DOP to 3D Coordinates using DSM . . . 23
- A.5 Optimization of Camera Parameters . . . 24

B OrthoLoC Dataset Details **24**
- B.1 Dataset Statistics . . . 24
- B.2 Samples Diversity . . . 25
- B.3 Comparison with Existing Datasets . . . 25
- B.4 Dataset Creation Pipeline . . . 25

C Additional Experimental Details and Results **30**
- C.1 Experiment Setting . . . 30
- C.2 Qualitative Assessment of AdHoP Performance . . . 30
- C.3 Camera Calibration Results . . . 32
- C.4 Proof of Focal Length and Translation Ambiguity . . . 32
- C.5 Domain Shift Analysis . . . 35
- C.6 Resolution Analysis . . . 36
- C.7 Covisibility Analysis . . . 36
- C.8 Utilizing Multi-Modal Data for Ground-Truth Geometry-Aware Correspondences 37
- C.9 Are 2.5D Rasters Sufficient for Accurate Localization? . . . 38

A Mathematical Formulation

This section establishes the mathematical foundation for our approach to localization and calibration using orthographic geodata. We present the camera projection models and the 3D lifting operation from a DOP raster to 3D coordinates using a DSM raster.

A.1 General Camera Projection Model

The general form of projecting a 3D point onto a camera image plane is given by:

$$\begin{aligned} \pi : \mathbf{P} &\mapsto \mathbf{p}\,, \\ \lambda \tilde{\mathbf{p}} &= \mathbf{K}\mathbf{\Pi}\mathbf{T}\tilde{\mathbf{P}}\,, \end{aligned} \tag{1}$$

where $\tilde{\mathbf{P}} \in \mathbb{R}^4$ is a 3D point in homogeneous coordinates, $\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix with focal lengths f_x, f_y and principal points c_x, c_y, $\mathbf{T} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \in \mathbb{R}^{4\times4}$ is the extrinsic matrix representing the world-to-camera pose with $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$, $\mathbf{\Pi} \in \mathbb{R}^{3\times4}$ is a projection matrix, $\tilde{\mathbf{p}} \in \mathbb{R}^3$ is the projected point in homogeneous image coordinates, and λ is a scale factor. Note that for simplicity, we ignore distortion effects in the following.

Our work leverages two distinct projection models: perspective projection from UAV cameras and orthographic projection (specifically nadir view) used in raster geodata. We illustrate both projections in Figure 6.

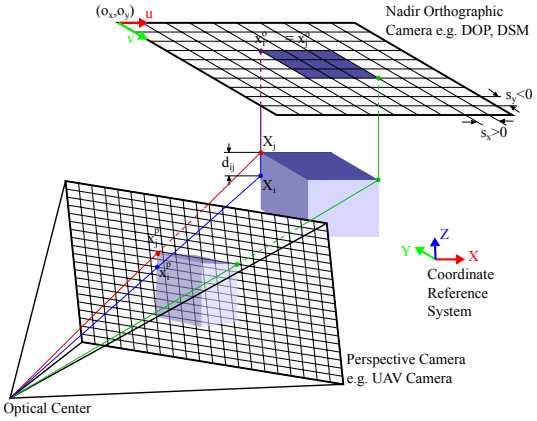


Figure 6: **Comparison of Projection Models Used in This Work.** The perspective projection, commonly employed in UAV cameras, features rays converging at a single camera center, resulting in perspective effects where parallel lines in the real world appear to converge in the image. In contrast, the orthographic nadir projection, often applied to raster geodata, uses parallel vertical rays that preserve scale relationships and spatial accuracy in the world.

A.2 Perspective Projection

A characteristic of perspective projection π_p is the intersection of all rays at a single point (camera center), accounting for perspective effects that make parallel lines in the world intersect in the image plane. In this case, $f_x > 0$, $f_y > 0$, λ corresponds to the depth of the 3D point with respect to the camera, and $\mathbf{\Pi} = [\mathbf{I}_3 \quad \mathbf{0}]$ with $\mathbf{I}_3$ being the 3×3 identity matrix and $\mathbf{0}$ a column vector of zeros. We use perspective projection for modeling UAV imagery.

A.3 Nadir Orthographic Projection

In orthographic nadir projections, all rays passing through the camera are parallel and strictly vertical, yielding a camera center at infinity. Projection lines are perpendicular to the XY plane, resulting

in parallel rays without perspective effects. This characteristic makes orthographic projection the standard representation for geodata such as satellite imagery and digital elevation models.

A nadir orthographic camera is defined by an origin (o_x, o_y) (the top left raster position in common reference frames) with scales s_x and s_y defining the metric grid cell size. Unlike perspective cameras, orthographic cameras maintain the same distance relationships in pixel coordinates as in world coordinates, scaled by s_x and s_y.

For the nadir orthographic projection function π_o, we can derive the closed-form formulation yielding $\lambda = 1$, $\mathbf{R} = \mathbf{I}_3$, $\mathbf{t} = [-o_x \quad -o_y \quad 0]^\top$, $f_x = 1/s_x$, $f_y = 1/s_y$, and $c_x = c_y = 0$. The projection matrix is $\mathbf{\Pi} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, which eliminates the Z component, effectively collapsing 3D points onto a plane.

Proof. In raster geodata such as DOPs and DSMs with (o_x, o_y) describing the XY position of the origin in predefined 3D geographic reference coordinate system (e.g., UTM) and pixel size (s_x, s_y) in metric space, a pixel coordinate (x, y) can be mapped to its X and Y coordinates in that 3D reference coordinate system using a simple linear transformation:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} s_x x + o_x \\ s_y y + o_y \end{bmatrix} . \tag{2}$$

We can extract x and y as:

$$\begin{aligned} x &= \frac{X - o_x}{s_x} , \\ y &= \frac{Y - o_y}{s_y} . \end{aligned} \tag{3}$$

In matrix form, this gives:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{s_x} & 0 & 0 & -\frac{o_x}{s_x} \\ 0 & \frac{1}{s_y} & 0 & -\frac{o_y}{s_y} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} . \tag{4}$$

Decomposing this compact projection matrix into the form $\mathbf{K\Pi T}$ yields:

$$\mathbf{K\Pi T} = \begin{bmatrix} \frac{1}{s_x} & 0 & 0 \\ 0 & \frac{1}{s_y} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & -o_x \\ 0 & 1 & 0 & -o_y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \tag{5}$$

In standard geospatial data conventions, the origin is typically at the top-left corner of the raster, with $s_x > 0$ and $s_y < 0$. This negative s_y accounts for the fact that the y-axis in pixel coordinates increases downward, while in world coordinates it increases upward. Given this convention, we have $f_x = 1/s_x > 0$ and $f_y = 1/s_y < 0$. This negative focal length has no direct physical counterpart in traditional optical systems. Nevertheless, this mathematical abstraction effectively represents the orthographic projection found in aerial mapping data, while enabling us to represent both orthographic and perspective projections in a unified formulation.

A.4 Lifting 2D Coordinates in DOP to 3D Coordinates using DSM

Our framework leverages orthographic geodata in the form of DSM and DOP rasters, denoted as $\mathbf{R}^{\text{DSM}} \in \mathbb{R}^{W^{\text{DSM}} \times H^{\text{DSM}}}$ and $\mathbf{R}^{\text{DOP}} \in \mathbb{R}^{W^{\text{DOP}} \times H^{\text{DOP}} \times 3}$, respectively. These rasters are characterized by scales $(s_x^{\text{DSM}}, s_y^{\text{DSM}})$, $(s_x^{\text{DOP}}, s_y^{\text{DOP}})$, and origins $(o_x^{\text{DSM}}, o_y^{\text{DSM}})$, $(o_x^{\text{DOP}}, o_y^{\text{DOP}})$. Utilizing the linear

correspondence between these representations, we derive 3D scene points from 2D coordinates in the DOP raster as:

$$\mathbf{P}_i = \left[\mathbf{p}_i^{\text{DOP}\top} \quad \mathbf{R}^{\text{DSM}}(f(\mathbf{p}_i^{\text{DOP}})) \right]^\top, \tag{6}$$

where $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a linear transformation mapping coordinates between the rasters. This transformation is defined as:

$$\tilde{\mathbf{p}}_i^{\text{DSM}} = \begin{pmatrix} \frac{s_x^{\text{DOP}}}{s_x^{\text{DSM}}} & 0 & \frac{o_x^{\text{DOP}} - o_x^{\text{DSM}}}{s_x^{\text{DSM}}} \\ 0 & \frac{s_y^{\text{DOP}}}{s_y^{\text{DSM}}} & \frac{o_y^{\text{DOP}} - o_y^{\text{DSM}}}{s_y^{\text{DSM}}} \\ 0 & 0 & 1 \end{pmatrix} \tilde{\mathbf{p}}_i^{\text{DOP}}, \tag{7}$$

where $\tilde{\cdot}$ denotes homogeneous coordinates. In our dataset, both rasters share the same scales and origins, simplifying f to the identity function. This formulation establishes 3D-2D correspondences between 2.5D orthographic geodata and UAV imagery.

A.5 Optimization of Camera Parameters

Given a query image $I \in \mathbb{R}^{W^I \times H^I \times 3}$ captured by a UAV, georeferenced camera calibration involves estimating camera parameters within a geospatial reference frame by minimizing a reprojection loss:

$$\begin{aligned} \mathbf{T}^*, \mathbf{K}^* &= \arg\min_{\mathbf{T},\mathbf{K}} \mathcal{L}_{\text{reproj}}, \\ \mathcal{L}_{\text{reproj}} &= \sum_i \rho \left(\| \pi_p(\mathbf{P}_i, \mathbf{K}, \mathbf{T}) - \mathbf{p}_i^I \|_2 \right), \end{aligned} \tag{8}$$

where $\mathbf{P}_i \in \mathbb{R}^3$ represents the 3D scene points, $\mathbf{p}_i^I \in \mathbb{R}^2$ corresponds to their associated 2D points in the image I, and $\rho(\cdot)$ is a robust cost function, specifically the Huber loss, which mitigates the impact of outlier correspondences. The calibration process involves optimizing the intrinsic and extrinsic parameters using the Levenberg-Marquardt [43, 75] algorithm. For initialization, we assume the focal length is $f_x = f_y = \max(W^I, H^I)$ and derive the initial extrinsic parameters through RANSAC-EPnP [37], employing a 5-pixel inlier threshold. In localization tasks, the intrinsic parameters $\mathbf{K}$ remain fixed, and only the extrinsic parameters $\mathbf{T}$ are estimated.

B OrthoLoC Dataset Details

B.1 Dataset Statistics

Our dataset consists of 16,427 samples with raster sizes of 1024×1024 pixels and query image sizes of 1024×682 or 1024×767 pixels.

From the 51 locations in our dataset, 48 were split into training (13K samples) and validation (1.5K samples) sets to facilitate future work focused on training models using our data. Representative samples from these 48 locations were used to create an in-Place test set, reflecting performance on previously seen environments. The remaining 3 locations were reserved for an out-Place test set, designed to evaluate generalization to novel environments. The dataset samples are further categorized into three types: same domain, cross-domain within DOP, and cross-domain between DOP and DSM.

Table 3: **Distribution of Samples Across Dataset Splits.**

Sample Type	All	Train	Val	Test In-Place	Test Out-Place
Same domain	10,923	9,255	1,030	142	496
DOP cross-domain	4,698	3,764	421	17	496
DOP & DSM cross-domains	806	328	38	17	423
All types	16,427	13,347	1,489	176	1,415

As shown in Table 3, our dataset is well-distributed across different sample types, with the majority being same-domain samples (10,923 samples). The training set contains 13,347 samples (81.3%), while validation and testing sets comprise 1,489 (9.1%) and 1,591 (9.7%) samples respectively.

Table 4: **Characteristics Across Sample Types and Dataset Splits.**

Characteristic	Stats	All	Sample Type			Dataset Split			
			Same domain	DSP cross	DSP & DSM cross	Train	Val	Test inPlace	Test outPlace
Obliqueness (deg)	Mean	14.6	14.3	14.6	18.6	14.0	14.4	18.6	20.4
	Min	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1
	Max	86.8	86.8	56.3	30.9	86.8	55.8	59.7	56.3
DSM area (m²)	Mean	37,512	36,061	40,860	37,660	36,903	37,477	49,434	41,810
	Min	982	982	11,238	14,691	982	2,394	22,344	11,238
	Max	370,686	370,686	370,686	60,569	370,686	337,931	241,607	370,686
Elevation (m)	Mean	101.92	100	107	104	101	101	111	109
	Min	23	23	72	74	23	24	99	72
	Max	201	201	201	124	154	148	134	147
Scale (cm/pix)	Mean	18.72	18.3	19.8	18.9	18.6	18.7	21.5	20.1
	Min	4.1	4.1	11.3	11.9	4.1	4.8	15.9	11.3
	Max	73.1	73.1	73.1	24.1	73.1	68.1	55.3	73.1
Query visible area (m²)	Mean	18,402	17,329	20,896	18,406	18,061	18,822	23,666	20,524
	Min	550	550	5,518	7,446	550	780	12,080	5,518
	Max	286,785	286,785	286,785	26,678	286,785	271,372	148,775	286,785
Covisibility (%)	Mean	99.99	100.0	100.0	99.9	100.0	100.0	100.0	99.9
	Min	95.7	99.4	99.6	95.7	95.7	98.7	99.9	95.7
	Max	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 4 demonstrates the diversity in our dataset. We observe varying obliqueness angles (0° to 86.8°), elevations (23m to 201m), and scales (4.1cm/pix to 73.1cm/pix) across different sample types and splits. The covisibility remains consistently high (above 95.7%) across all samples, ensuring quality matches between query and reference images.

The test sets feature higher average obliqueness angles and DSM areas compared to the training data, providing more challenging evaluation scenarios. This diversity across all characteristics makes our dataset well-suited for robust model training and evaluation across different geographical conditions.

B.2 Samples Diversity

In addition to the statistics, we illustrate the diversity of our dataset by presenting representative samples from different environments and viewing conditions in Figure 7. We also show local meshes for randomly picked samples from our dataset in Figure 8. These examples showcase the variability in scene content, viewpoint, and domain characteristics that make our dataset particularly challenging and representative of real-world conditions.

B.3 Comparison with Existing Datasets

We performed a detailed analysis of geometric consistency across recent aerial visual localization datasets to assess the accuracy of their ground-truth poses. Our evaluation projects 3D keypoints – extracted from high-precision DOPs and DSMs obtained from open geoportals – onto query images using the provided camera parameters, enabling visual verification of pose quality.

As shown in Figure 9, CrossLoc [68] and UAVD4L [66] exhibit noticeable projection errors, indicating pose inaccuracies. While AnyVisLoc [72] offers cross-domain augmentation using satellite imagery, this data is low resolution and poorly aligned with photogrammetry-based DOPs, limiting its suitability for realistic cross-domain experiments.

In contrast, our OrthoLoC dataset delivers superior geometric consistency with accurately projected keypoints and well-aligned cross-domain data at resolutions comparable to official geoportal sources, enabling reliable cross-domain localization research.

B.4 Dataset Creation Pipeline

The dataset is created by capturing data with drones, building 3D models through photogrammetry with georeferencing, extracting data like orthophotos and elevation rasters, and pairing query images with reference data followed by domain augmentation.

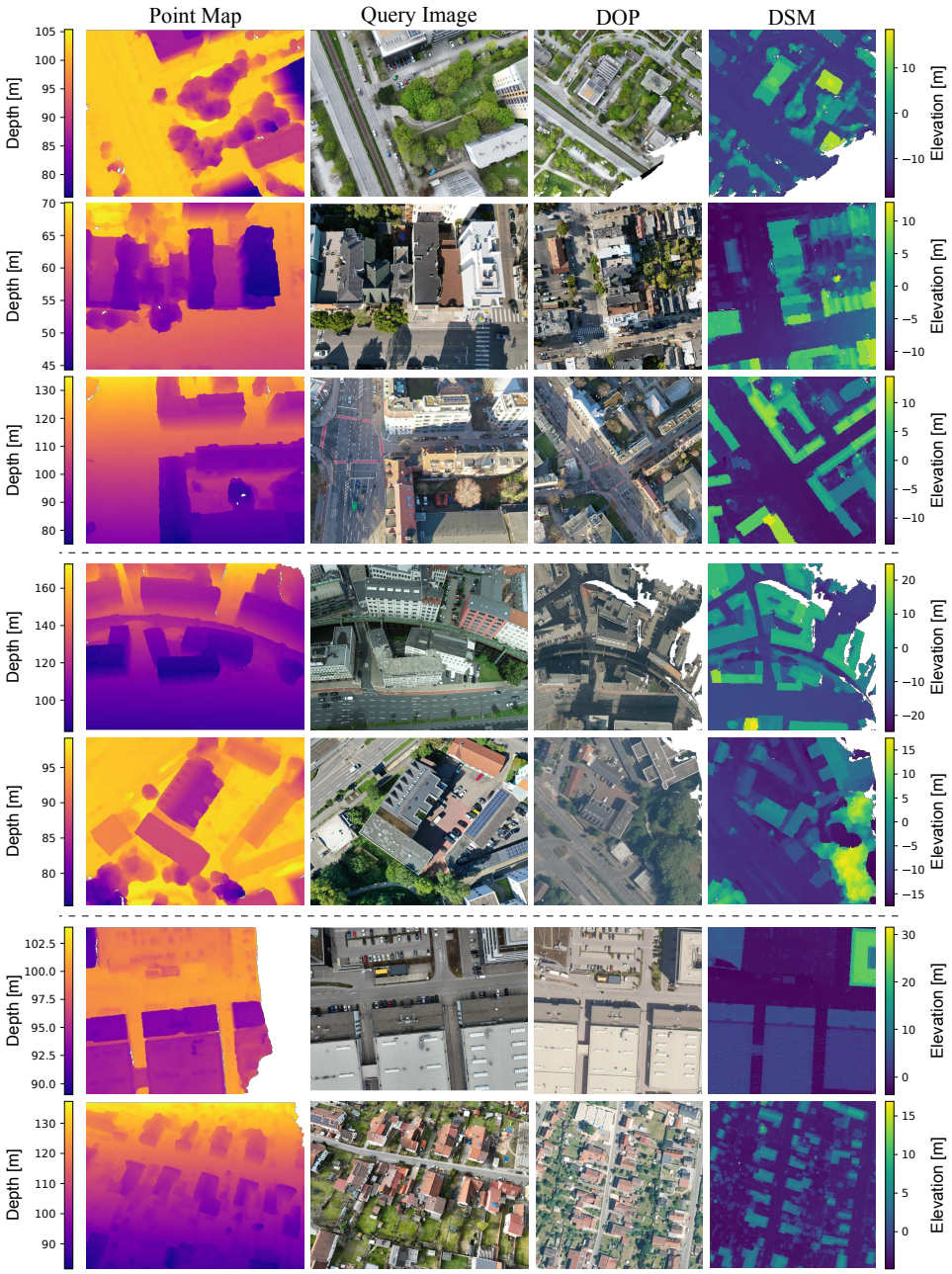


Figure 7: **Dataset Diversity Across Environments and Viewing Conditions.** Our dataset spans diverse scenes (urban, suburban, rural, industrial) and perspectives (nadir, oblique). **(Rows 1-3)** Same domain for query data (query image, point map) and reference data (DOP, DSM); **(Rows 4-5)** DOP domain shifts; **(Rows 6-7)** Combined DOP and DSM domain shifts.[2]

B.4.1 Data Acquisition and Processing

Data collection. Our data collection encompassed 47 locations across 19 regions in Germany and the United States. We utilized a variety of commercial drones equipped with GPS and RTK technology to ensure precise positioning. To facilitate robust photogrammetric reconstruction, we implemented systematic flight paths, following established protocols in aerial mapping.

[2]The geodata used to increase the domain shifts are ontained from different open geoportals. 4^{th} row: Geodata © Geobasis NRW, 5^{th} row: Geodata © HLBG, 6^{th} - 7^{th} rows: Geodata © LDBV Bayern

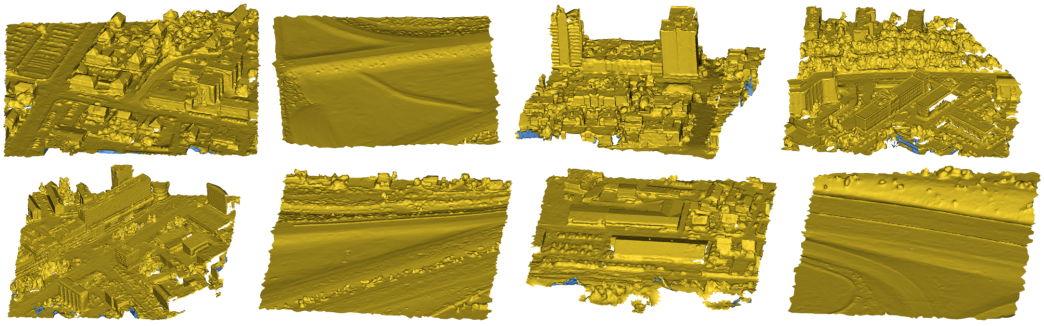
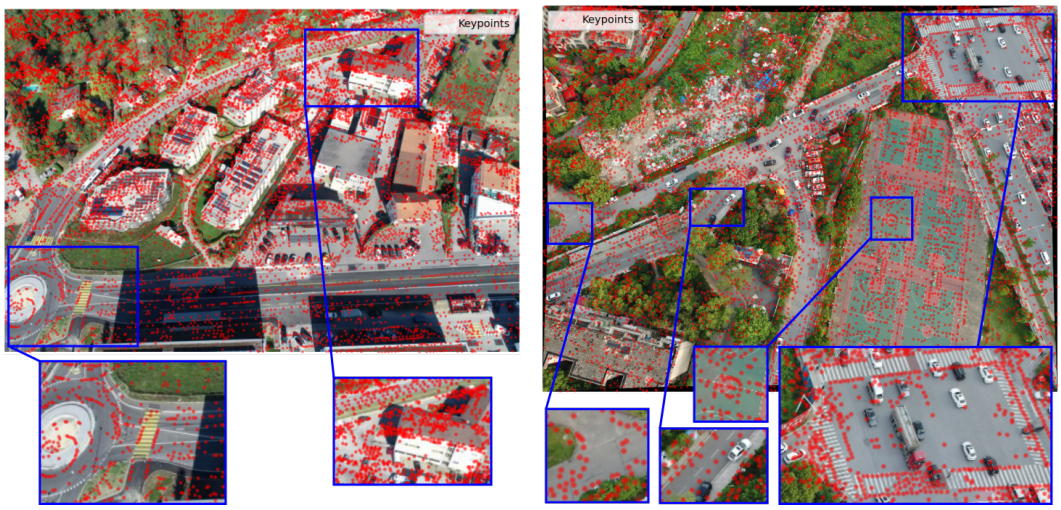
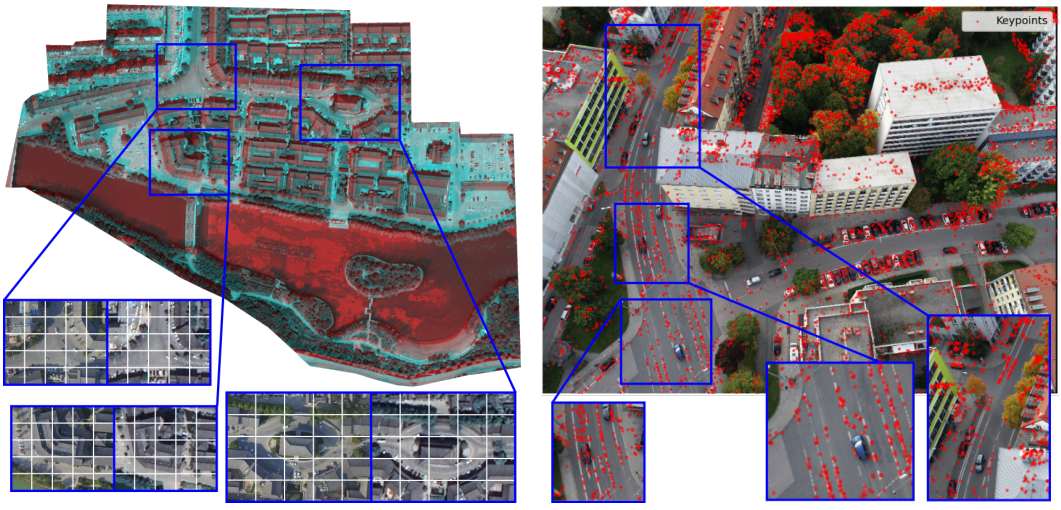


Figure 8: **Examples of Local Meshes in OrthoLoC.**



(a) CrossLoc [68] (b) UAVD4L [66]



(c) AnyVisLoc [72] (d) OrthoLoC (ours)

Figure 9: **Dataset Quality Assessment.** Geometric consistency evaluated by projecting 3D keypoints onto query images or showing alignment between DOP from different sources. Our dataset shows superior consistency with accurate projections.

Geodata were downloaded from public geoportals, with temporal misalignment considered explicitly. The dataset introduces time gaps of 2–8 years between geodata and UAV imagery, matching common update cycles (2–3 years in urban areas, over 5 years in rural regions). This design provides a realistic benchmark for developing methods robust to outdated geodata.

Georeferenced 3D scene reconstruction. At each location, we acquire N flight images I_i ($1 \leq i \leq N$) and begin the pipeline by georeferencing them to constrain the subsequent SfM optimization. We leverage GPS, RTK, or manually annotate GCPs to ensure accurate alignment. We execute a collection of SfM pipelines—including DJI Terra [20], PixPro [63], and COLMAP [53] with MVS [8]—and select the best output for each scene based on bundle-adjustment reprojection errors, GCP RTK residuals, and qualitative keypoint projections, following the procedure used in benchmarked localization methods.

Formally, the pipeline extracts features from the images and constructs a pose graph. SfM computes initial camera poses and a sparse point cloud $\mathcal{P}$, which MVS densifies to refine poses and produce a denser 3D representation. We triangulate using Poisson surface reconstruction [35] and apply texturing to generate a mesh with vertices $\mathcal{V} = \{\mathbf{P}_j\}_j$ and faces $\mathcal{F}$.

We obtain precise 6-DoF ground-truth poses by jointly optimizing the scene geometry $\mathcal{P}$, camera extrinsics $\mathbf{T}_i$, and shared intrinsics $\mathbf{K}$, while incorporating georeferencing constraints from three complementary sources: **(1)** standard GPS for coarse positioning, **(2)** RTK for centimeter-level accuracy, and **(3)** manually annotated GCPs for high-precision alignment. Our GCPs were carefully selected following established best practices similar to those validated in [16]. As shown in fig. 10, we manually chose features with precise, visually distinctive characteristics, such as road marking edges, ensuring optimal visibility and spatial distribution across the mapping area. The corresponding 3D coordinates were obtained from either vehicle-based Mobile Laser Scanning point clouds or high-precision governmental geodata by sampling the DSM at these locations. These 2D–3D correspondences are incorporated into the bundle adjustment during SfM reconstruction as follows:

$$
\begin{aligned}
\mathbf{T}_i, \mathbf{K}, \mathcal{P} &= \underset{\mathbf{T}_i, \mathbf{K}, \mathcal{P}}{\arg\min}\, \mathcal{L}_{\text{reproj}} + \lambda_{\text{GPS}} \mathcal{L}_{\text{GPS}} + \lambda_{\text{GCP}} \mathcal{L}_{\text{GCP}}\,, \\
\mathcal{L}_{\text{reproj}} &= \sum_{i,j} \rho\left(\|\pi_p(\mathbf{P}_j, \mathbf{K}, \mathbf{T}_i) - \mathbf{p}_{ij}\|_2\right)\,, \\
\mathcal{L}_{\text{GPS}} &= \sum_i \|\mathbf{C}_i - \mathbf{C}_i^{\text{GPS}}\|_2^2\,, \\
\mathcal{L}_{\text{GCP}} &= \sum_k \|\mathbf{P}_k - \mathbf{P}_k^{\text{GCP}}\|_2^2\,.
\end{aligned}
\tag{9}
$$

Here, $\mathbf{P}_j \in \mathbb{R}^3$ is a 3D scene point, $\mathbf{p}_{ij} \in \mathbb{R}^2$ is its projection in image I_i through $\pi_p(\cdot)$, $\rho(\cdot)$ is a robust cost function, $\mathbf{C}_i \in \mathbb{R}^3$ is the camera center, and $\mathbf{C}_i^{\text{GPS}}$ is the measured GPS or RTK position. For ground control, $\mathbf{P}_k$ denotes a GCP point, and $\mathbf{P}_k^{\text{GCP}}$ is its reference position. Weighting factors λ_{GPS} and λ_{GCP} are tuned to balance the reliability of each data source. After optimization, the residual GCP errors yield Root Mean Square Error (RMSE) values of 0.023 m, 0.030 m, and 0.042 m in x, y, and z, respectively, with an overall 3D RMSE of 0.051 m.

While the optimization includes radial and tangential lens distortion coefficients, these terms are omitted from the equations for simplicity. The dataset provides undistorted images, allowing researchers to focus on focal length estimation, which is the most challenging aspect of UAV camera calibration.

Rasterization and rendering. The 3D mesh reconstruction is converted into two complementary geospatial representations: a DSM matrix $\mathbf{R}^{\text{DSM}}$ and a DOP matrix $\mathbf{R}^{\text{DOP}}$. The DSM is generated by casting rays downward from a planar grid aligned with the XY plane at the maximum elevation. Each grid cell (i, j) corresponds to a geographic position (x, y), and the value at $\mathbf{R}^{\text{DSM}}(i, j)$ represents the highest elevation (z-coordinate) of the mesh intersected by the ray. Cells with no intersections are explicitly marked as invalid. The DOP is created by rendering a nadir-oriented, orthographic view of the textured mesh with a virtual camera aligned along the negative z-axis. The resulting image is georeferenced and resampled to align with the resolution and coordinate system of the DSM.

We also extract a sparse set $\mathcal{S}$ of SIFT [42] keypoints from the DOP, which are lifted to 3D coordinates by mapping their positions to corresponding elevations in the DSM using Equation (6). These keypoints serve as reliable landmarks for pose verification in our evaluations.

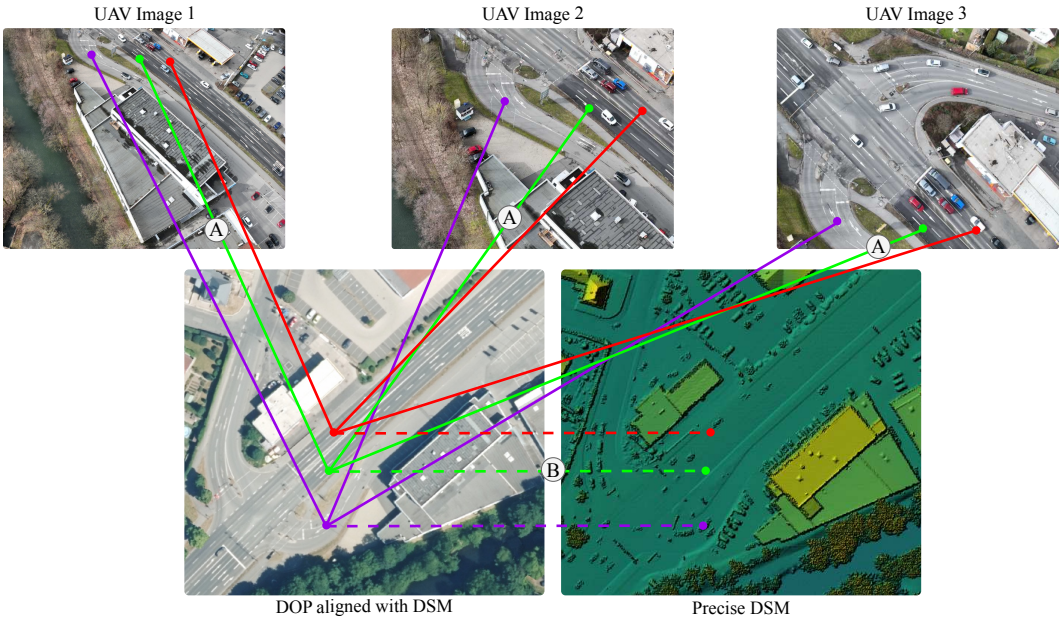


Figure 10: **Manual GCP Selection Procedure.** (A) We manually establish 2D–2D correspondences between the UAV images and a DOP (acquired from high-precision open geoportals), using visually distinctive features such as road markings. The feature positions are extracted in at least three images from the set of collected UAV images. (B) The elevation of the corresponding pixel is then extracted to obtain 2D–3D correspondences.

B.4.2 Data Pairing

Selection of regions of interest. Given known camera parameters $(\mathbf{K}_i, \mathbf{T}_i)$ for each query image I_i, correspondences with geospatial representations are established through a precise geometric approach. To improve computational efficiency, the query image camera is downscaled by a factor of 8, and a grid of image coordinates $\mathbf{p}_j^c$ is generated for the downscaled image. For each coordinate, ray tracing is performed through the camera using $\mathbf{P}_j^c = \pi^{-1}(\mathbf{p}_j^c, \mathbf{K}_i, \mathbf{T}_i, d)$, where $\pi^{-1}(\cdot)$ is the inverse projection function and d is the ray-mesh intersection depth.

Projected points with valid intersections ($d_j < \infty$) are filtered, and an irregular quadrilateral is fitted to the valid points. To introduce variability, stochastic perturbations are applied to the quadrilateral vertices as $\mathbf{P}_j^{c'} = \mathbf{P}_j^c + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{U}(-20\text{m}, 20\text{m})$. The perturbed 3D points $\mathbf{P}_j^{c'}$ are then projected onto the DOP and DSM to define corresponding 2D regions, which are used to crop rasters $\mathbf{R}_i^{\text{DOP}}$ and $\mathbf{R}_i^{\text{DSM}}$ for each sample i.

To enrich data modalities, per-pixel raycasting generates point maps $\mathbf{P}_i \in \mathbb{R}^{W \times H \times 3}$, from which depth maps can derived using the extrinsic parameters. Additionally, visible mesh elements $(\mathcal{V}_i, \mathcal{F}_i)$ and SIFT keypoints $(\mathcal{S}_i)$ are also selected for each viewpoint, increasing the dataset's modalities.

Data anonymization. To preserve geographic privacy while maintaining geometric relationships, we transform each data sample into a local coordinate system. For each sample, we apply a translation transform $\mathbf{v}_i \in \mathbb{R}^3$ defined by randomly selecting a finite 3D point from the visible scene. This translation is consistently applied to all geometric elements:

$$\begin{aligned} \mathbf{P}_i' &= \mathbf{P}_i - \mathbf{v}_i\,, & \mathbf{R}_i' &= \mathbf{R}_i\,, & \mathbf{t}_i' &= \mathbf{t}_i + \mathbf{R}_i\mathbf{v}_i\,, \\ \text{DSM}' &= \text{DSM} - \mathbf{v}_{i,z}\,, & o_x' &= o_x - \mathbf{v}_{i,x}\,, & o_y' &= o_y - \mathbf{v}_{i,y}\,, \\ \mathcal{V}_i' &= \mathcal{V}_i - \mathbf{v}_i\,, & \mathcal{S}_i' &= \mathcal{S}_i - \mathbf{v}_i\,, \end{aligned} \tag{10}$$

where $\mathbf{P}_i$ represents 3D points in the original point cloud, $\mathcal{V}_i$ and $\mathcal{S}_i$ denote visible points and scene points respectively, and $\mathbf{v}_{i,x}$, $\mathbf{v}_{i,y}$, and $\mathbf{v}_{i,z}$ stand for the x, y, and z coordinates of vector $\mathbf{v}_i$.

The camera transformation by translation is derived as follows: First, we convert the world-to-camera pose $\mathbf{T}_i$ (with rotation $\mathbf{R}_i$ and translation $\mathbf{t}_i$) to its inverse camera-to-world pose $\mathbf{T}_i^{-1}$ (with rotation $\mathbf{R}_i^\top$ and translation $-\mathbf{R}_i^\top \mathbf{t}_i$, which represents the camera center). After shifting this camera center by $\mathbf{v}_i$, we obtain a new camera-to-world pose with rotation $\mathbf{R}_i^\top$ and translation $-\mathbf{R}_i^\top \mathbf{t}_i - \mathbf{v}_i$. Converting back to the world-to-camera frame yields the original rotation matrix $(\mathbf{R}_i^\top)^\top = \mathbf{R}_i$ and a new translation vector $-\mathbf{R}^T(-\mathbf{R}_i^\top \mathbf{t}_i - \mathbf{v}_i) = \mathbf{t}_i + \mathbf{R}_i \mathbf{v}_i$.

This transformation suppresses absolute georeferencing while preserving all relative geometric relationships essential for localization evaluation. The random selection of transformation vectors ensures that geographic coordinates cannot be reliably reconstructed from the published dataset, protecting sensitive location information.

C Additional Experimental Details and Results

C.1 Experiment Setting

Our experiments were conducted on a cluster using a computation node equipped with an Intel(R) Xeon(R) Gold 6254 CPU @ 3.10GHz and a single Quadro RTX 8000 GPU with 48GB memory.

For benchmarking, we evaluated algorithms using the provided model weights without fine-tuning to assess inherent robustness. Images were resized to meet each algorithm's requirements, with resulting coordinates transformed back to full resolution before 3D lifting. Rotation-invariant algorithms processed each image in four orientations, selecting the output with the largest correspondence set.

The 6-DoF pose estimation was performed using PnP [28] with LO-RANSAC [14] for outlier rejection, applying a 5-pixel reprojection threshold for inlier selection. For algorithms providing confidence scores, a 0.5 threshold was used to pre-filter correspondences. Optimization was restricted to estimating focal length, assuming a fixed aspect ratio and principal point at the image center, based on empirical validation.

C.2 Qualitative Assessment of AdHoP Performance

This section offers a detailed qualitative evaluation of the AdHoP strategy in various scenarios. While quantitative results in the main paper highlight consistent improvements in localization accuracy, visual analysis of reprojected keypoints provides further insights into the strengths and limitations of the method.

The left column of Figure 11 presents examples of challenging scenarios where AdHoP achieves successful localization. These include cases where AdHoP reduces large initial errors to achieve highly accurate poses with minimal reprojection error, enhances moderately accurate estimates to near-perfect precision, and improves the spatial distribution of correspondences across the image plane, leading to better geometric consistency.

The right column highlights instances where AdHoP struggles to deliver improvements. In some cases, it slightly worsens performance by increasing matching errors or producing less accurate poses, often due to correspondences with worse geometric cues. In other scenarios, the approach fails to improve poorly initialized calibrations, as the homography-based warping introduces distortions that prevent effective matching. These issues are typically observed under extreme domain shifts, highly repetitive patterns, or significant discrepancies between the reference geodata and query images, where existing matchers exhibit poor performance. Notably, the failure modes of AdHoP are captured in the computed projection error, allowing automatic rejection of results when the reprojection error increases.

We also present an example of warped DOP in Figure 12, demonstrating how this transformation significantly improves alignment between the geodata and UAV imagery by reducing domain shifts in appearance. The most accurately warped regions correspond to planar and non-occluded areas. However, some domain shifts persist, primarily in regions where: (1) building facades are missing in the DOPs, (2) areas occluded in the orthographic projection, (3) regions with strong shadows, and (4) inherent appearance differences in DOP from different capture times.

This qualitative analysis supports our quantitative results, showing that AdHoP improves accuracy in most cases.

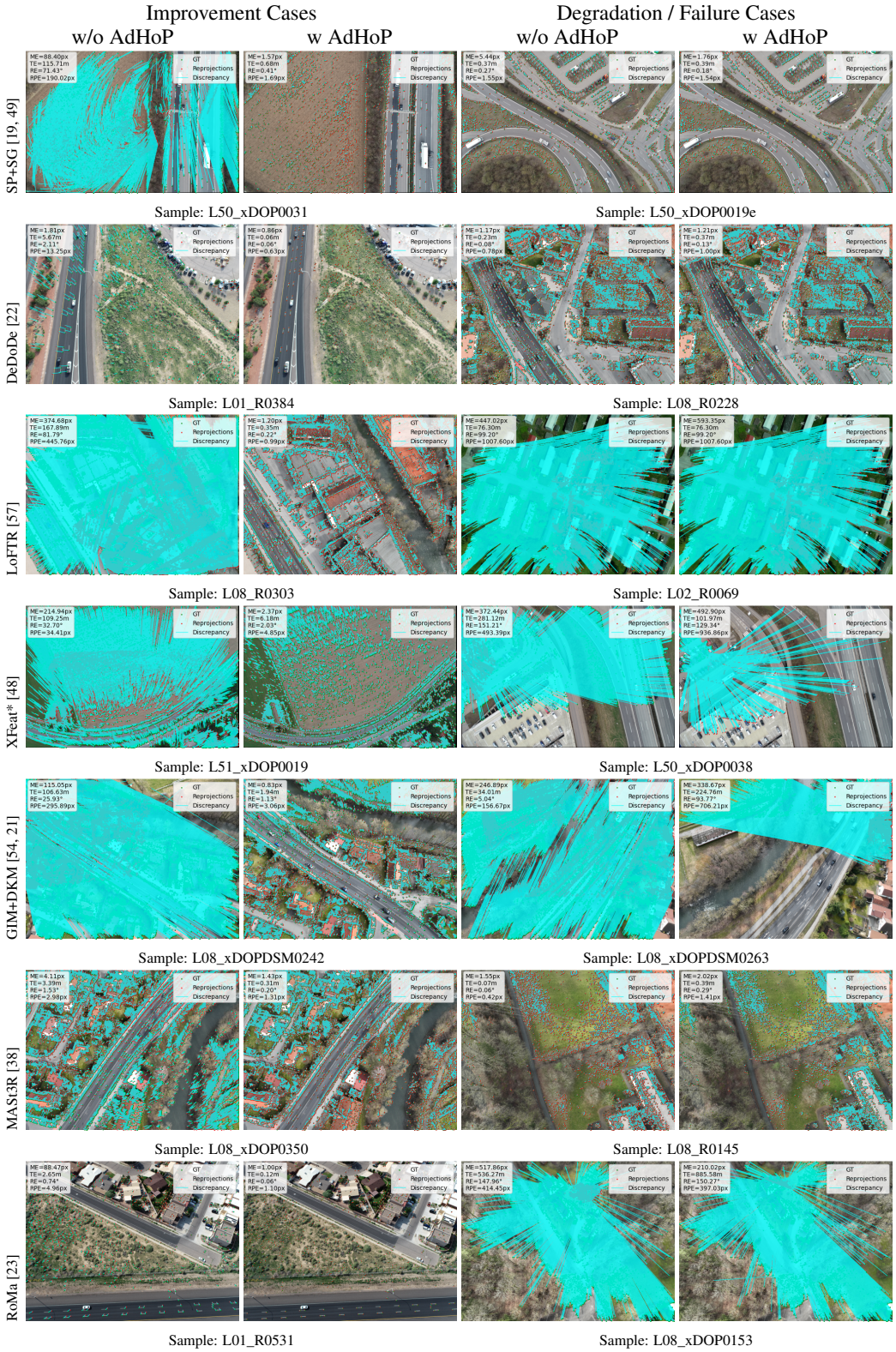
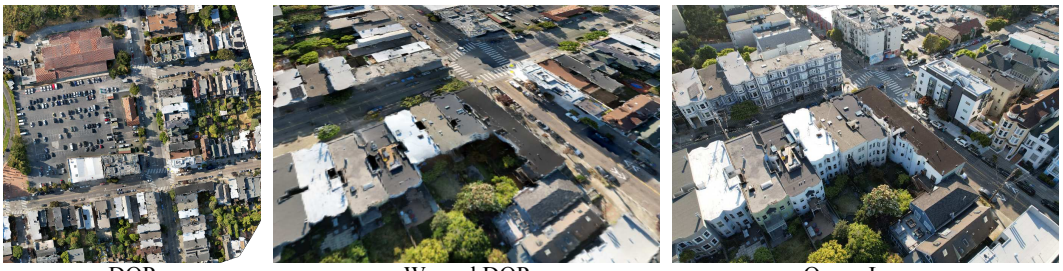


Figure 11: **Qualitative Results of the Localization Using Our Baseline Method.** The left side illustrates successful improvements achieved using AdHoP, while the right side presents degenerate or failure cases. Green and red points denote projections of the 3D keypoints $\mathcal{S}_i$ using the ground-truth and estimated poses, respectively. The blue lines indicate the discrepancies between these projections.



DOP | Warped DOP | Query Image

Figure 12: **Effectiveness of Warped DOP in Addressing Domain Shifts. Left:** The original DOP exhibits significant discrepancies with UAV imagery due to temporal changes, lighting variations, and pronounced viewpoint differences. **Middle:** The warped DOP after applying the computed transformation demonstrates substantially improved alignment with UAV imagery. **Right:** The corresponding UAV query image to match.

C.3 Camera Calibration Results

Table 5: **Quantitative Calibration Results on OrthoLoC Test Sets.** Rankings between matchers are highlighted as first, second, and third. **Bold** values indicate the best performance comparing without/with AdHoP. RI indicates a rotation-invariant matcher (matching performed with 4 rotated versions, selecting the one with most correspondences). Abbreviations: SuperPoint (SP), SuperGlue (SG), LightGlue (LG), Minima (MM).

Matcher	RI	ME [px]↓	RFE [%]↓	TE [m]↓	RE [°]↓	RPE [px]↓	1m-1° [%]↑	3m-3° [%]↑	5m-5° [%]↑	Speed [s]↓
SP+SG [19, 49]	✗	**2.2** / **2.2**	2.6 / **2.4**	3.09 / **2.93**	0.60 / **0.57**	16.4 / **15.9**	**12.1** / 11.1	40.0 / **41.2**	50.7 / **52.7**	**0.2** / 0.3
SP+LG [19, 41]	✗	**2.0** / **2.0**	1.9 / **1.8**	2.49 / **2.30**	0.54 / **0.52**	15.2 / **14.3**	**13.1** / **13.1**	46.1 / **48.1**	54.6 / **57.1**	**0.1** / 0.2
DeDoDe [22]	✗	**1.2** / **1.2**	2.5 / **1.7**	3.40 / **2.11**	0.60 / **0.42**	17.0 / **12.2**	5.3 / **7.4**	17.7 / **21.7**	21.4 / **25.0**	**0.3** / **0.3**
XFeat [48]	✗	257.0 / **60.5**	100.0 / **60.1**	**180.53** / 192.47	**169.09** / 173.66	**648.8** / 809.0	0.0 / **0.1**	0.0 / **1.6**	0.0 / **2.5**	**0.1** / 0.3
XFeat+LG [48, 41]	✗	4.3 / **3.2**	6.1 / **2.7**	7.40 / **3.35**	1.06 / **0.73**	29.1 / **19.5**	4.4 / **6.6**	17.4 / **30.7**	26.0 / **38.5**	**0.1** / 0.3
LoFTR [57]	✗	317.2 / **308.7**	93.2 / **83.5**	170.00 / **154.66**	129.88 / **112.92**	914.2 / **899.4**	0.4 / **5.0**	1.9 / **14.2**	2.7 / **15.5**	**0.1** / 0.3
MM+LoFTR [33, 78]	✗	**261.1** / **261.1**	86.0 / **54.2**	193.17 / **143.00**	112.46 / **91.81**	1035.7 / **950.4**	0.1 / **5.3**	0.9 / **15.0**	1.6 / **17.2**	**0.3** / 0.6
eLoFTR [65]	✗	328.8 / **315.0**	96.0 / **80.7**	188.04 / **160.78**	128.66 / **107.91**	868.8 / **858.1**	0.6 / **5.3**	1.8 / **16.0**	2.6 / **17.6**	**0.1** / 0.2
XoFTR [62]	✗	286.6 / **282.2**	91.9 / **71.6**	180.41 / **146.27**	135.07 / **107.65**	942.8 / **829.1**	0.3 / **4.1**	1.9 / **12.6**	2.5 / **14.5**	**0.1** / 0.2
DKM [21]	✓	49.0 / **2.8**	50.5 / **9.3**	131.70 / **32.22**	108.41 / **4.19**	818.4 / **114.2**	2.8 / **12.3**	12.4 / **30.0**	17.5 / **34.8**	**0.8** / 1.7
XFeat* [48]	✗	222.2 / **10.2**	100.0 / **52.6**	184.95 / **184.39**	**170.90** / 173.72	**660.9** / 789.9	0.0 / **0.2**	0.0 / **2.8**	0.0 / **4.0**	**0.1** / 0.2
GIM+DKM [54, 21]	✓	1.5 / **1.4**	2.4 / **1.6**	3.07 / **2.09**	0.65 / **0.47**	17.8 / **13.1**	16.2 / **21.8**	49.5 / **59.0**	58.2 / **67.8**	**1.3** / 2.6
DUSt3R [64]	✓	5.0 / **4.9**	12.5 / **12.3**	**14.68** / 17.02	**2.39** / 2.82	**75.3** / 86.9	**0.3** / 0.2	**5.7** / 5.5	**12.7** / 11.4	**1.5** / 2.1
MASt3R [38]	✓	2.4 / **2.3**	4.9 / **3.9**	6.10 / **4.72**	0.95 / **0.79**	28.7 / **23.3**	6.0 / **8.4**	31.1 / **38.0**	45.3 / **51.6**	**2.2** / 3.3
RoMa [23]	✓	20.8 / **2.5**	91.9 / **7.0**	150.82 / **10.62**	149.98 / **1.73**	616.0 / **46.2**	1.1 / **5.2**	8.0 / **27.6**	12.5 / **38.8**	**1.1** / 2.1
MM+RoMa [33, 23]	✓	71.1 / **4.3**	99.3 / **13.5**	165.60 / **26.39**	142.71 / **3.70**	646.2 / **97.7**	1.1 / **3.3**	7.1 / **20.7**	11.4 / **30.5**	**1.1** / 2.2

Table 5 presents our full camera calibration results across different feature matching approaches, both without and with our proposed AdHoP strategy. The integration of AdHoP significantly enhances calibration performance across all matchers, with GIM+DKM [54, 21] achieving the best focal length estimation with an RFE of 1.6%. The most dramatic improvement is observed with RoMa [23], where translation errors are reduced from 150.82 m to 10.62 m.

Our experimental results show a strong correlation between focal length errors and translation errors, driven by the inherent mathematical ambiguity of the perspective projection model. In the next section, we provide a mathematical proof of the correlation between focal length and the camera translation.

C.4 Proof of Focal Length and Translation Ambiguity

We denote a 3D point as $\mathbf{P} = [X, Y, Z]^\top$ and its corresponding 2D image point as $\mathbf{p} = [p_x, p_y]^\top$. For simplicity, we assume that the focal lengths in the intrinsic matrix $\mathbf{K}$ are equal, i.e., $f_x = f_y = f$, which is a reasonable assumption for most modern cameras. Let $\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ denote the three rows of the rotation matrix $\mathbf{R}$.

Using the projection model defined in Equation (1), the projection of a single 3D point onto the image plane is expressed as:

$$p_x = f \cdot \frac{\mathbf{r}_1 \cdot \mathbf{P} + t_x}{\mathbf{r}_3 \cdot \mathbf{P} + t_z} + c_x \,, \tag{11}$$

$$p_y = f \cdot \frac{\mathbf{r}_2 \cdot \mathbf{P} + t_y}{\mathbf{r}_3 \cdot \mathbf{P} + t_z} + c_y \,, \tag{12}$$

where $\mathbf{r}_i \cdot \mathbf{P}$ represents the dot product between the i-th row of $\mathbf{R}$ and the 3D point $\mathbf{P}$.

For the reprojection error, we use the Mean Squared Error (MSE) for this proof. Given a set of N corresponding 2D-3D point pairs $\{(\mathbf{p}_i, \mathbf{P}_i)\}_{i=1}^N$:

$$E_{\text{reproj}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|^2 = \frac{1}{N} \sum_{i=1}^{N} \left((p_{x,i} - \hat{p}_{x,i})^2 + (p_{y,i} - \hat{p}_{y,i})^2 \right) \,, \tag{13}$$

where $\hat{\mathbf{p}}_i = [\hat{p}_{x,i}, \hat{p}_{y,i}]^\top$ is the projection of $\mathbf{P}_i$ using the estimated camera parameters.

For notation simplicity, let:

$$\begin{aligned} a_i &= \mathbf{r}_1 \cdot \mathbf{P}_i + t_x \,, \\ b_i &= \mathbf{r}_2 \cdot \mathbf{P}_i + t_y \,, \\ c_i &= \mathbf{r}_3 \cdot \mathbf{P}_i + t_z \,. \end{aligned} \tag{14}$$

Then the projected points become:

$$\hat{p}_{x,i} = f \cdot \frac{a_i}{c_i} + c_x, \quad \hat{p}_{y,i} = f \cdot \frac{b_i}{c_i} + c_y \,. \tag{15}$$

The partial derivative of $\hat{p}_{x,i}$ with respect to f is:

$$\frac{\partial \hat{p}_{x,i}}{\partial f} = \frac{a_i}{c_i} \,. \tag{16}$$

Similarly for $\hat{p}_{y,i}$:

$$\frac{\partial \hat{p}_{y,i}}{\partial f} = \frac{b_i}{c_i} \,. \tag{17}$$

The partial derivative of the reprojection error with respect to f is:

$$\begin{aligned} \frac{\partial E_{\text{reproj}}}{\partial f} &= \frac{2}{N} \sum_{i=1}^{N} \left[(p_{x,i} - \hat{p}_{x,i}) \cdot \left(-\frac{a_i}{c_i} \right) + (p_{y,i} - \hat{p}_{y,i}) \cdot \left(-\frac{b_i}{c_i} \right) \right] \\ &= -\frac{2}{N} \sum_{i=1}^{N} \left[(p_{x,i} - \hat{p}_{x,i}) \cdot \frac{a_i}{c_i} + (p_{y,i} - \hat{p}_{y,i}) \cdot \frac{b_i}{c_i} \right] \end{aligned} \,. \tag{18}$$

Mathematical approximation for $\frac{\partial E_{\text{reproj}}}{\partial f} \propto \frac{1}{f}$. Near the optimal solution, the reprojection errors $(p_{x,i} - \hat{p}_{x,i})$ and $(p_{y,i} - \hat{p}_{y,i})$ become very small. To understand the relationship with f, consider that at near-optimal parameters, we can approximate $p_{x,i} \approx \hat{p}_{x,i}$ and use small perturbations δf around the current estimate of f. The change in projected coordinates would be:

$$\delta \hat{p}_{x,i} = \delta f \cdot \frac{a_i}{c_i} \,. \tag{19}$$

For identical relative changes in focal length ($\frac{\delta f}{f}$), the absolute change in projection is proportional to f, as:

$$\delta \hat{p}_{x,i} = f \cdot \frac{\delta f}{f} \cdot \frac{a_i}{c_i} \,. \tag{20}$$

This means the sensitivity of the projection (and consequently the error gradient) to absolute changes in f scales with $\frac{1}{f}$. For larger f values, the same absolute change has less impact on the projection. Therefore, $\frac{\partial E_{\text{reproj}}}{\partial f} \propto \frac{1}{f}$.

For the translation component t_z, we compute:

$$\frac{\partial \hat{p}_{x,i}}{\partial t_z} = -f \cdot \frac{a_i}{c_i^2} \,. \tag{21}$$

Similarly:

$$\frac{\partial \hat{p}_{y,i}}{\partial t_z} = -f \cdot \frac{b_i}{c_i^2} \,. \tag{22}$$

The partial derivative of the reprojection error with respect to t_z is:

$$\begin{aligned} \frac{\partial E_{\text{reproj}}}{\partial t_z} &= \frac{2}{N} \sum_{i=1}^{N} \left[(p_{x,i} - \hat{p}_{x,i}) \cdot \left(f \cdot \frac{a_i}{c_i^2} \right) + (p_{y,i} - \hat{p}_{y,i}) \cdot \left(f \cdot \frac{b_i}{c_i^2} \right) \right] \\ &= \frac{2f}{N} \sum_{i=1}^{N} \left[(p_{x,i} - \hat{p}_{x,i}) \cdot \frac{a_i}{c_i^2} + (p_{y,i} - \hat{p}_{y,i}) \cdot \frac{b_i}{c_i^2} \right] \end{aligned} \,. \tag{23}$$

Proving that f and t_z are coupled. In aerial imagery, t_z is typically much larger than the variations in scene depth, so $c_i \approx t_z$ for most points. With this approximation:

$$\frac{\partial E_{\text{reproj}}}{\partial t_z} \propto -\frac{f}{t_z^2} \,. \tag{24}$$

The critical insight comes from examining how f and t_z affect the projection. Consider a simplified projection model with $c_i \approx t_z$:

$$\hat{p}_{x,i} \approx f \cdot \frac{a_i}{t_z} + c_x \,. \tag{25}$$

If we simultaneously scale f by a factor α and t_z by the same factor α, the projection remains unchanged:

$$(\alpha f) \cdot \frac{a_i}{(\alpha t_z)} + c_x = f \cdot \frac{a_i}{t_z} + c_x \,. \tag{26}$$

This exact mathematical compensation creates a "valley" in the optimization landscape where different combinations of f and t_z produce nearly identical reprojection errors, making their individual values ambiguous.

For comparison, the partial derivative with respect to t_x is:

$$\frac{\partial \hat{p}_{x,i}}{\partial t_x} = f \cdot \frac{1}{c_i} \,. \tag{27}$$

Comparing these derivatives reveals the key relationships:

$$\frac{\partial E_{\text{reproj}}}{\partial f} \propto \frac{1}{f}, \quad \frac{\partial E_{\text{reproj}}}{\partial t_z} \propto -\frac{f}{t_z^2}, \quad \frac{\partial \hat{p}_{x,i}}{\partial t_x} \propto \frac{f}{c_i} \,. \tag{28}$$

Importance of data variation for robust estimation. The coupling between f and t_z creates an ill-posed optimization problem when 3D points lie approximately on a plane, as is common in aerial imagery. Spatial diversity in point correspondences is crucial for breaking this ambiguity for two key reasons:

- **Depth variation**: Points at different depths create different sensitivity patterns to f and t_z. When the scene contains significant depth variations, the exact compensation relationship between f and t_z breaks down, as the $c_i = \mathbf{r}_3 \cdot \mathbf{P}_i + t_z$ term varies more significantly across points.

- **Geometric constraints**: Points distributed across the image plane, especially toward the borders, experience different projection behaviors than points clustered in the center. The peripheral points are more sensitive to focal length changes, providing stronger constraints during optimization.

Our AdHoP strategy specifically addresses this challenge by encouraging spatially diverse correspondence distributions across the image plane. By ensuring correspondences span different image regions with varying depths, we better constrain the parameter space and reduce the inherent focal length-translation ambiguity, making the optimization more likely to converge to the correct parameter values. This explains why our experimental results show significantly improved focal length and translation estimates when using AdHoP, as the strategy effectively breaks the mathematical coupling that would otherwise lead to ambiguous solutions.

C.5 Domain Shift Analysis

Table 6: **Quantitative Results of Localization on OrthoLoC Test Sets Across Domain Configurations.** We evaluate matchers under three scenarios: same domain (reference and query from identical sources), DOP cross-domain (different orthophoto sources), and DOP+DSM cross-domain (different orthophoto and elevation model sources). For each metric, results are presented as: same domain / DOP cross-domain / DOP+DSM cross-domain. Rankings between matchers are highlighted as first, second, and third. **Bold** values indicate the best performance within each domain configuration group. RI indicates a rotation-invariant matcher (matching performed with 4 rotated versions, selecting the one with most correspondences). Abbreviations: SuperPoint (SP), SuperGlue (SG), LightGlue (LG), Minima (MM).

Matcher	RI	ME [px]↓	TE [m]↓	RE [°]↓	RPE [px]↓	1m-1° [%]↑	3m-3° [%]↑	5m-5° [%]↑
SP+SG [19, 49]	✗	**1.5** / 2.8 / 2.8	**0.20** / 0.42 / 1.09	**0.08** / 0.16 / 0.52	**1.6** / 3.2 / 9.2	**96.2** / 75.4 / 41.5	**99.3** / 85.1 / 83.5	**99.5** / 87.4 / 86.3
SP+LG [19, 41]	✗	**1.5** / 2.5 / 2.5	**0.21** / 0.42 / 1.07	**0.07** / 0.17 / 0.50	**1.6** / 3.4 / 9.0	**94.4** / 71.3 / 42.3	**98.1** / 80.4 / 80.4	**98.6** / 83.3 / 83.0
DeDoDe [22]	✗	**1.2** / 2.2 / 2.1	**0.33** / 2.89 / 3.45	**0.13** / 1.18 / 1.56	**2.6** / 23.3 / 28.1	**68.5** / 1.8 / 0.7	**76.2** / 5.5 / 4.8	**79.0** / 6.8 / 6.6
XFeat [48]	✗	**2.6** / 210.4 / 214.9	**0.23** / 16.43 / 27.62	**0.09** / 8.68 / 11.88	**1.9** / 136.0 / 243.8	**92.8** / 28.8 / 15.7	**94.5** / 42.5 / 41.4	**94.7** / 45.8 / 44.5
XFeat+LG [48, 41]	✗	**1.8** / 3.8 / 3.8	**0.19** / 0.75 / 1.36	**0.08** / 0.33 / 0.62	**1.4** / 5.8 / 11.4	**99.1** / 56.0 / 28.3	**99.5** / 67.9 / 67.8	**99.5** / 70.1 / 70.7
LoFTR [57]	✗	**291.3** / 313.4 / 331.4	**110.48** / 123.28 / 129.28	**101.50** / 107.81 / 111.09	**1198.5** / 1505.7 / 1511.8	**30.4** / 18.3 / 10.5	**31.5** / 22.0 / 21.4	**32.6** / 22.4 / 21.8
MM+LoFTR [33, 78]	✗	294.3 / **216.6** / 228.6	90.15 / **82.47** / 84.96	100.86 / **93.99** / 96.14	960.8 / 783.0 / **777.7**	**27.1** / 15.5 / 8.7	**27.9** / 20.2 / 19.6	**28.4** / 20.9 / 20.3
eLoFTR [65]	✗	**285.0** / 330.5 / 334.3	**102.97** / 124.06 / 127.49	**96.61** / 106.57 / 107.73	**1176.8** / 1718.6 / 1745.8	**33.2** / 19.5 / 11.8	**35.4** / 22.8 / 21.8	**35.9** / 23.6 / 23.2
XoFTR [62]	✗	**273.2** / 283.4 / 299.7	**103.26** / 114.13 / 118.62	**100.65** / 108.60 / 114.40	**1254.4** / 1268.0 / 1322.8	**29.5** / 20.1 / 11.8	**29.7** / 22.4 / 20.7	**30.0** / 23.0 / 21.8
DKM [21]	✓	**1.1** / 96.9 / 123.0	**0.20** / 61.30 / 60.74	**0.08** / 63.16 / 69.68	**1.5** / 518.3 / 526.9	**65.0** / 33.1 / 19.8	**66.1** / 40.0 / 37.5	**66.5** / 40.7 / 38.4
XFeat* [48]	✗	**2.2** / 89.1 / 94.4	**0.32** / 1.57 / 1.81	**0.11** / 0.66 / 0.90	**2.6** / 11.3 / 17.1	**95.0** / 44.2 / 26.8	**96.9** / 55.9 / 55.9	**96.9** / 58.1 / 58.2
GIM+DKM [54, 21]	✓	**1.0** / 1.8 / 1.8	**0.16** / 0.48 / 1.10	**0.05** / 0.17 / 0.54	**1.1** / 3.6 / 9.4	**100.0** / 70.4 / 45.5	**100.0** / 80.7 / 78.6	**100.0** / 81.7 / 79.5
DUSt3R [64]	✓	**3.9** / 6.5 / 6.8	**2.43** / 5.27 / 4.99	**1.08** / 2.05 / 2.05	**18.9** / 37.2 / 38.8	**15.2** / 0.6 / 0.5	**59.2** / 16.8 / 16.8	**76.3** / 31.2 / 33.0
MASt3R [38]	✓	**1.6** / 3.2 / 3.2	**0.25** / 1.05 / 1.48	**0.10** / 0.44 / 0.66	**2.0** / 8.0 / 12.5	**99.4** / 47.7 / 30.1	**100.0** / 70.3 / 70.3	**100.0** / 74.8 / 74.0
RoMa [23]	✓	**1.1** / 5.6 / 5.1	**0.18** / 1.59 / 1.84	**0.07** / 0.60 / 0.92	**1.4** / 11.7 / 16.5	**78.7** / 46.2 / 29.5	**79.0** / 55.8 / 55.7	**79.0** / 58.3 / 58.9
MM+RoMa [33, 23]	✓	**1.1** / 28.3 / 33.2	**0.20** / 3.24 / 4.57	**0.07** / 1.44 / 2.24	**1.5** / 35.1 / 46.8	**72.1** / 39.2 / 23.0	**72.4** / 49.1 / 47.5	**72.4** / 51.5 / 50.2

Table 6 presents the quantitative localization results using our baseline with AdHoP, incorporating reference data from different domains.

For same-domain scenarios, the majority of the models reach high performance (except semi-dense matchers and Dust3R [64] with recall 1m-1° below 50%). GIM+DKM [54, 21] and Mast3R [38] demonstrate particularly high accuracy in these conditions with recall 1m-1° 100% and 99.4%, respectively.

When employing DOPs from visually distinct domains, the increased appearance gap between query and reference data leads to noticeable performance degradation. The extent of this degradation varies significantly across matching algorithms, with some exhibiting greater robustness to appearance changes than others. Even the best performing GIM+DKM [54, 21] degrades by 29.6% in cross-domain scenarios. DeDoDe [22] is very sensitive to domain shifts, with recall 1m-1° dropping drastically from 68.5% to 1.8%. Dust3R [64], on the other hand, struggles across all domains, highlighting its limitations in aerial views.

Further increasing the domain shift by additionally incorporating DSMs from different sources generally degrades performance, highlighting the importance of geometry cues for pose estimation. Some algorithms degrade strongly as they find matches primarily on edges or object boundaries where DSMs differ significantly between sources, while others show more resilience by matching features in geometrically stable regions where elevation remains consistent across different DSM sources. An example of performance degradation is illustrated in Figure 13.

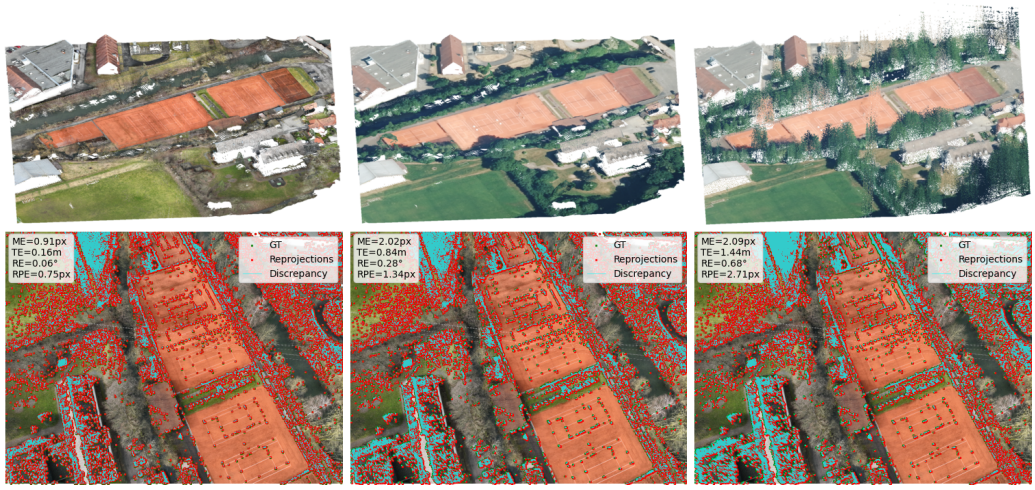


Figure 13: **Example of Domain Shift Impact When Using GIM+DKM [54] and AdHoP.** No shift (left), DOP shift (middle), DOP+DSM shifts (right). Top: colored point clouds (created using DSM with colors from DOP); Bottom: reprojections with green (ground-truth), red (estimated) points and blue discrepancy lines indicating reprojection errros.

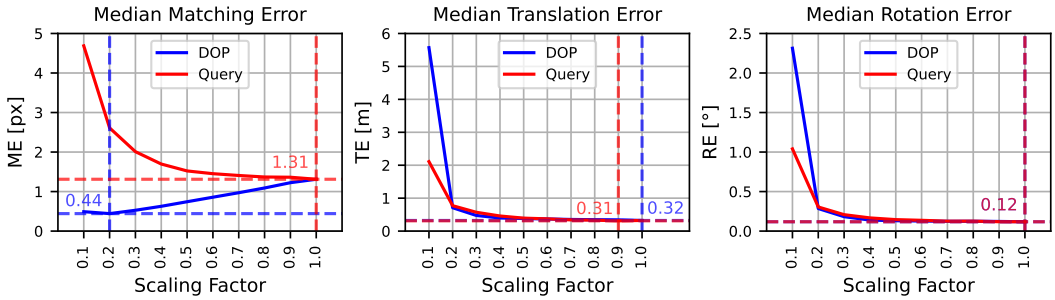


Figure 14: **Resolution Impact on Localization:** Performance of GIM+DKM [54, 21] + AdHoP across varying raster and query image resolutions.

C.6 Resolution Analysis

We evaluate how raster resolution affects our lightweight localization system, an important factor for storage-constrained UAV platforms. Additionally, we analyze the effects of query image resolution on matching performance, with results shown in Figure 14.

Our findings indicate that localization performance remains robust down to 512px raster resolution, with noticeable degradation only at lower resolutions. At 256px, translation error increases by 44% and rotation error by 33% compared to the highest resolution. This suggests that significant storage savings can be achieved with minimal performance impact by using moderately reduced resolution reference data.

Query image resolution shows similar patterns, maintaining adequate performance down to 512px before exhibiting significant degradation. The balance between computational efficiency and localization accuracy becomes particularly important for onboard processing in real-time UAV applications.

C.7 Covisibility Analysis

To assess algorithm robustness in real-world scenarios where perfect image retrieval cannot be guaranteed, we systematically reduce the covisibility ratio between query and reference images by cropping the reference raster. Our experiments reveal that having only a subset of potential correspondences significantly degrades performance, as shown in Figure 15.

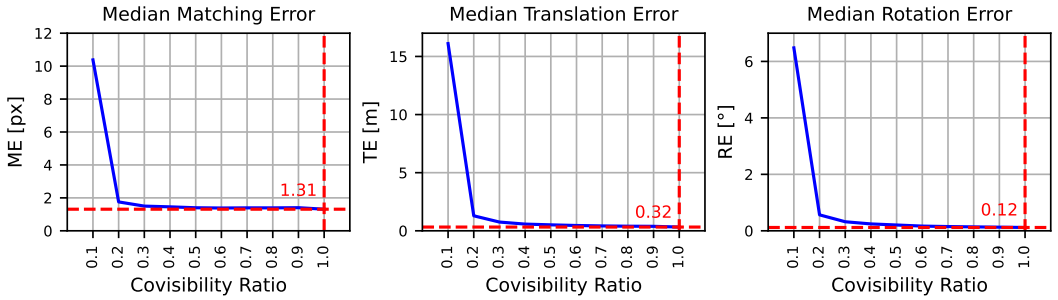


Figure 15: **Covisibility Ratio Impact:** Localization performance of GIM+DKM [54, 21] + AdHoP across different covisibility ratios.

When the covisibility ratio drops below 20%, we observe a sharp increase in both translation and rotation errors. This degradation occurs because the distribution of matched points becomes non-uniform across the image. This non-uniformity causes PnP to overfit to specific image regions, creating an underdetermined problem that compromises localization accuracy.

Figure 16 demonstrates how the same query image produces different localization results depending on whether the reference points are well-distributed or concentrated in a particular area of the image.

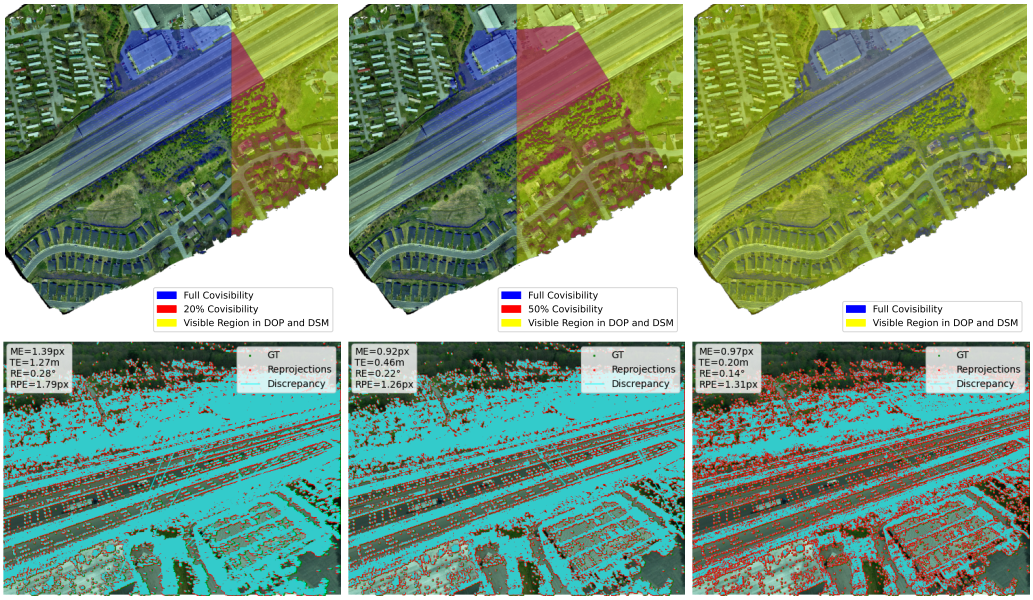


Figure 16: **Covisibility Comparison in Localization Setup.** (left) 20% coverage, (middle) 50% coverage, (right) full coverage. The top row shows query-raster covisibility, while the bottom row displays reprojection of the keypoints $\mathcal{S}_i$ using the estimated pose. Note how the distribution of points significantly affects the quality of calibration.

These findings have important implications for real-world applications, suggesting that image retrieval systems should prioritize maximizing overlap between query and reference images.

C.8 Utilizing Multi-Modal Data for Ground-Truth Geometry-Aware Correspondences

Our dataset enables computing geometry-aware confidences that can guide network training. We show that filtering correspondences based purely on geometric constraints, using ground-truth data, provides perfect pose estimation.

Geometry-aware confidences computation. Given the 3D point maps and DSM in our dataset, we establish ground-truth correspondences with associated confidence values. For each pixel $\mathbf{p}_i^I$ in the query image I, we:

1. Backproject it to a 3D point $\mathbf{P}_i$ using the perspective camera model and the point map.
2. Project $\mathbf{P}_i$ onto the DSM plane using the orthographic projection model.

This process, illustrated in Figure 6, reveals a fundamental limitation in perspective-to-orthographic projection with 2.5D geodata. Unlike full 3D meshes where visible points have one-to-one correspondence with 3D space, raster geodata creates a many-to-one mapping. When ray-casting from the perspective view, multiple points ($\mathbf{p}_i^p$ and $\mathbf{p}_j^p$) can map to the same orthographic position ($\mathbf{p}_i^o = \mathbf{p}_j^o$) because 2.5D representations store only a single height value per (x, y) coordinate. As shown in Figure 6, points along vertical structures (like building facades) in the perspective view map to identical locations in the orthographic view, creating ambiguous correspondences. To address this ambiguity, we introduce a geometry-aware confidence measure α_i for each correspondence using:

$$\alpha_i = \exp(-\gamma \cdot d_i)\,, \tag{29}$$

$$\mathbf{P}_i = \pi_p^{-1}(\mathbf{p}_i^I)\,, \tag{30}$$

$$d_i = \|\mathbf{P}_i - \pi_o^{-1}(\pi_o(\mathbf{P}_i))\|_2\,, \tag{31}$$

where π_p^{-1} is the perspective back-projection, π_o is the orthographic projection, π_o^{-1} is the orthographic back-projection (which assigns the DSM height to the 2D coordinates), and γ is a scaling parameter (we set it to 1). Note that the composition $\pi_o^{-1} \circ \pi_o$ is not an identity due to the dimensional reduction in orthographic projection, as illustrated in Figure 6.

C.9 Are 2.5D Rasters Sufficient for Accurate Localization?

To understand the practical potential of commonly available 2.5D geodata for UAV localization, we investigate how geometric ambiguities affect pose estimation accuracy and whether simple filtering strategies can overcome these challenges.

Table 7 summarizes the localization results achieved using ground-truth correspondences with geometry-aware confidences. Different thresholds τ are used to filter points.

Table 7: **Quantitative Results of Localization on OrthoLoC Test Sets (Same Domain) Using Ground-Truth Matchings.**

Filtering Condition	**TE** [m]↓	**RE** [°]↓	**RPE** [px]↓	**1m-1°** [%]↑	**3m-3°** [%]↑	**5m-5°** [%]↑
$\alpha_i > 0.0$	0.03	0.00	0.2	100.0	100.0	100.0
$\alpha_i > 0.5$	0.03	0.01	0.2	100.0	100.0	100.0
$\alpha_i > 0.95$	0.00	0.00	0.0	100.0	100.0	100.0
$\alpha_i > 0.99$	0.00	0.00	0.0	100.0	100.0	100.0

In the unrestricted 2.5D case ($\alpha_i > 0.0$), all valid points from the 2.5D DSM are used, including those from vertical structures or occluded areas. This approach introduces minor errors, due to ambiguities in the many-to-one mapping of vertical structures. As τ increases, filtering progressively excludes ambiguous points, improving data purity. At higher thresholds (e.g., $\alpha_i > 0.95$), the mapping becomes close to a one-to-one relationship, and pose estimation achieves perfect localization with no observable errors.

These findings demonstrate the sufficiency of 2.5D orthographic geodata for accurate UAV localization when paired with robust geometric filtering. By carefully selecting τ, 2.5D geodata can achieve performance levels comparable to full 3D representations, motivating further research into leveraging 2.5D geodata capabilities.