Exploring the best way for UAV visual localization under Low-altitude Multi-view Observation Condition: a Benchmark

Supplementary Material

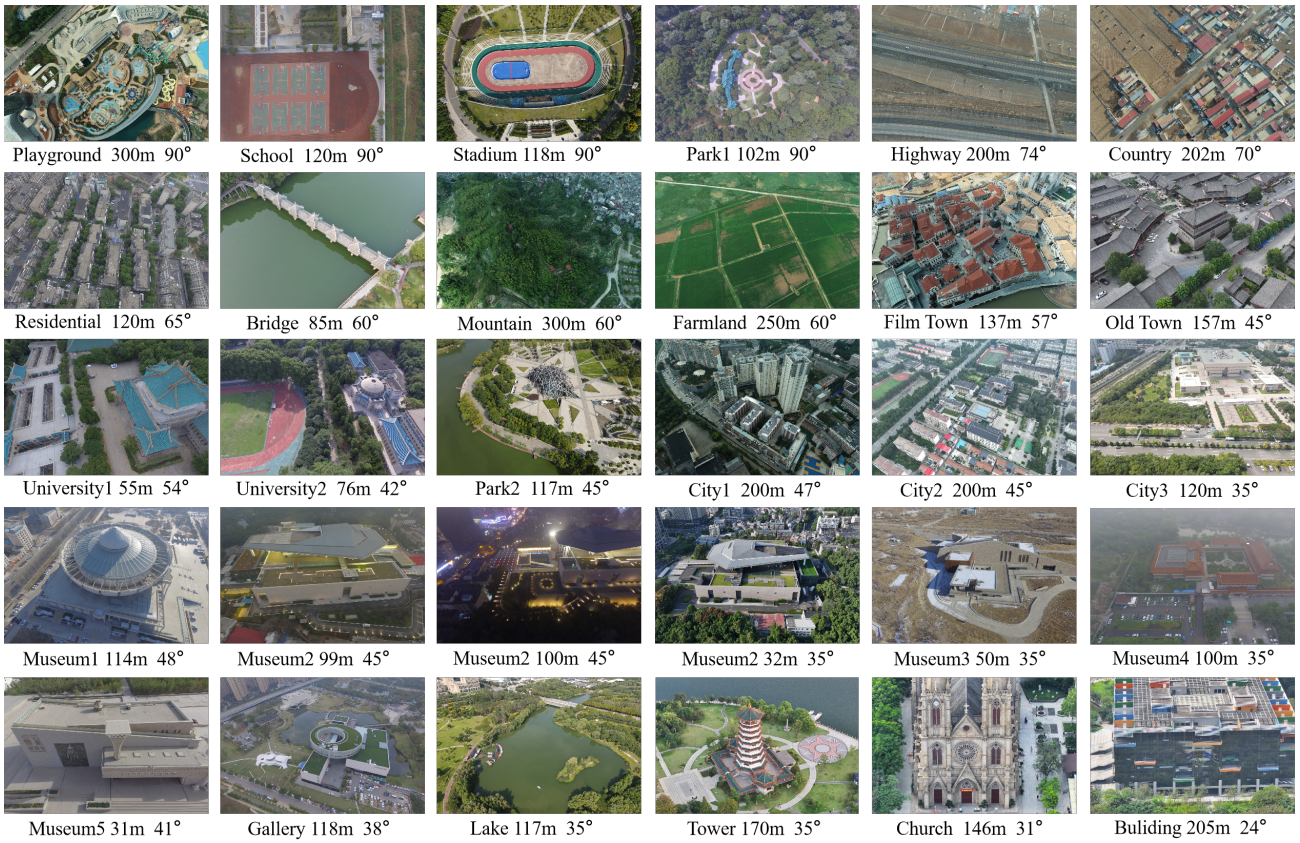


Figure 8. More examples of the UAV images in AnyVisLoc dataset. Each UAV image shows its scene, altitude and pitch angle below.

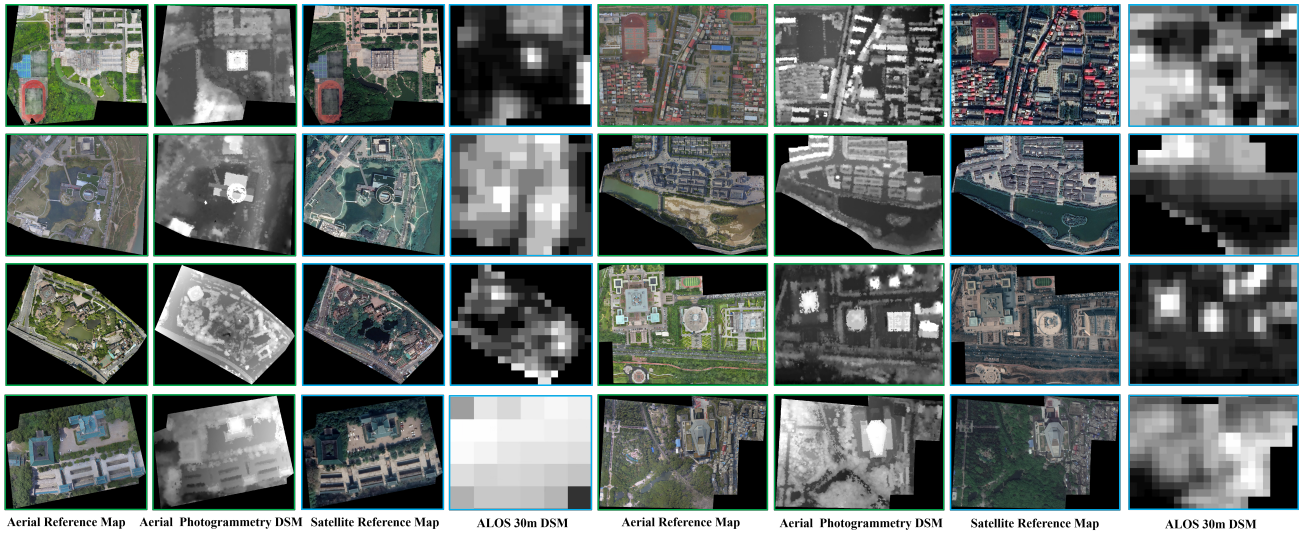


Figure 9. More examples of the reference maps in AnyVisLoc dataset. The coverage area of these reference maps is **smaller than 1** km^2.

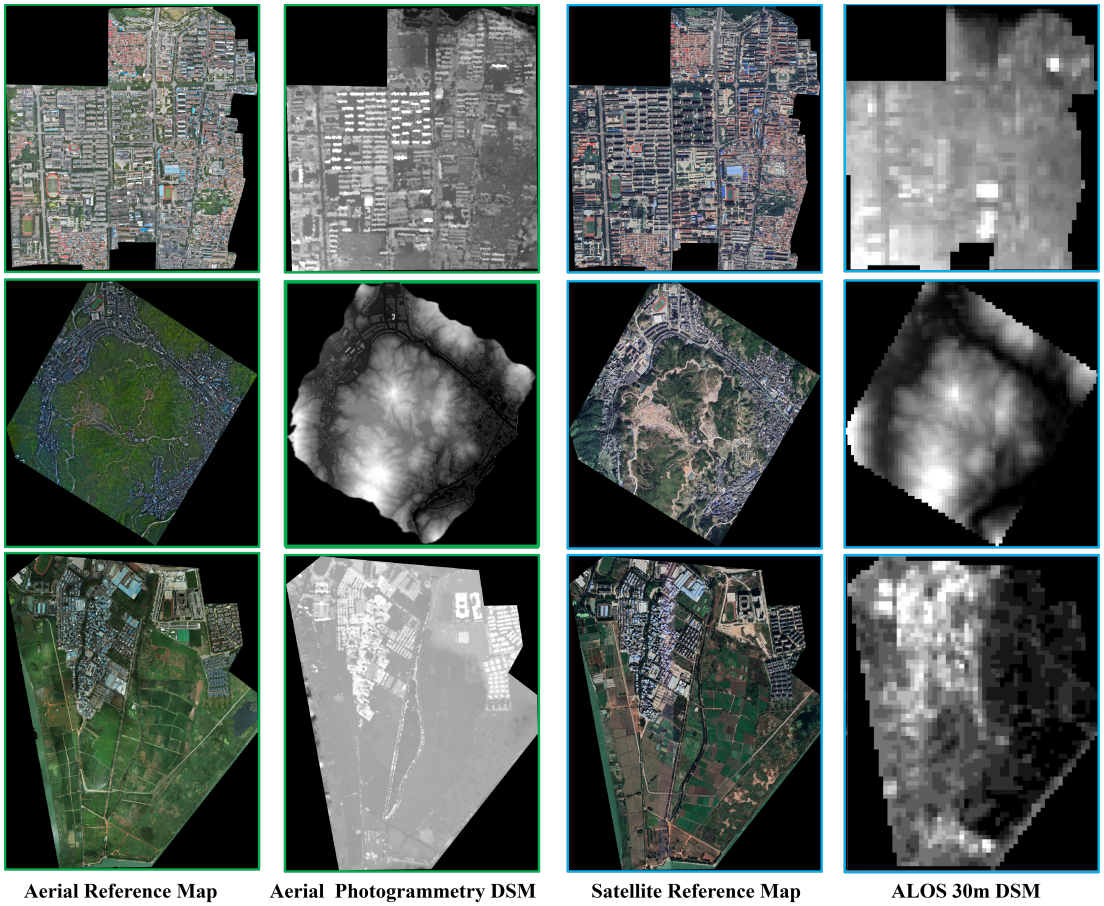


Figure 10. More examples of the reference maps in AnyVisLoc dataset. The coverage area of these reference maps is **larger than 1** km^2.

Table 8. Details of AnyVisLoc Dataset. **Img.Num.**: the number of UAV images. **A.M.SR**: the Spatial Resolution of Aerial Map. **S.M.SR**: the Spatial Resolution of Satellite Map. **A.D.SR**: the Spatial Resolution of Aerial DSM map. **S.D.SR**: the Spatial Resolution of Satellite DSM map. **Phan.**: Phantom. **Mav.**: Mavic. **Area**: the coverage area of region.

Region	Scene Description	Reference Map Location	Img.Num.	Drone Type	A.M.SR(m)	S.M.SR(m)	A.D.SR(m)	S.D.SR(m)	Area(km^2)
1	Highway;Country	105.9448°E 36.8788°N	1846	Phan. 4 RTK	0.32	0.546	1.982	30	9.01
2	Mountain;Town	119.4947°E 27.4627°N	1833	Phan. 4 RTK	0.076	0.14	0.152	30	3.02
3	City1	102.6890°E 25.0593°N	1831	Phan. 4 RTK/Mav.2/Mini 4 pro	0.084	0.142	1.142	30	3.96
4	City2	118.4672°E 36.6810°N	1824	Mav.3/Mav.2	0.064	0.26	1.805	30	2.34
5	Museum1;Park1	112.9877°E 28.2145°N	1132	Phan. 4/Mav.3 pro	0.046	0.142	0.569	30	0.25
6	Park2	113.0000°E 28.2210°N	1106	Phan. 4/Mav.3 pro/Mav.2	0.046	0.143	0.377	30	0.43
7	Farmland;Town	102.6901°E 24.9234°N	1291	Phan. 4 RTK	0.074	0.283	0.591	30	2.54
8	Museum2	118.4603°E 36.6794°N	909	Mav.3/Mav.2	0.064	0.26	1.805	30	1.24
9	University1	114.3603°E 30.5424°N	699	Phan. 4	0.037	0.139	0.45	30	0.07
10	Museum3	88.1644°E 39.0156°N	651	Phan. 4	0.025	0.129	0.05	30	0.04
11	Playground;Town	112.9110°E 28.0819°N	626	Phan. 4 RTK	0.1	0.142	1.285	30	0.98
12	Museum4	113.7390°E 23.0450°N	563	Mav. Air	0.02	0.284	0.384	30	0.16
13	City3	112.5270°E 37.8661°N	533	Mav.3/Mav.2	0.043	0.133	1.226	30	0.24
14	Town	118.4752°E 36.6890°N	487	Phan. 3/Mav.3 Pro	0.061	0.26	0.937	30	0.19
15	Church	113.2552°E 23.1151°N	300	Mav.3 Pro	0.03	0.285	0.72	30	0.09
16	Museum5	106.2308°E 38.4845°N	295	Phan. 4	0.042	0.132	0.763	30	0.13
17	Park	117.2925°E 31.7312°N	289	Mav.2	0.049	0.132	1.115	30	0.74
18	Countryside	114.2158°E 33.3006°N	273	Phan. 3/Phan. 4	0.16	0.271	2.423	30	1.6
19	University2	113.3888°E 23.0400°N	256	Mav.3 Pro	0.045	0.142	0.655	30	0.23
20	City4	112.9847°E 28.2268°N	252	Phan. 3	0.054	0.138	0.974	30	0.18
21	School	114.2592°E 33.2714°N	252	Phan. 3	0.051	0.132	0.909	30	0.25
22	University3	114.3579°E 30.5421°N	220	Phan. 4/Phan. 3	0.022	0.142	0.225	30	0.01
23	Stadium	120.7660°E 37.7871°N	196	Phan. 3	0.056	0.132	1.058	30	0.09
24	Gallery	112.9446°E 28.1223°N	163	Phan. 4	0.067	0.278	0.83	30	0.2
25	Museum6	106.0204°E 38.7396°N	125	Phan. 4	0.037	0.131	1.227	30	0.14

This supplementary document provides additional details and experimental results supporting the main paper. It includes extended visual examples of the AnyVisLoc dataset and its complete statistics (Sec. 8). We also provide further details on ground-truth accuracy (Sec. 9), implementation specifics (Sec. 10), image retrieval settings (Sec. 11), and the use of prior information (Sec. 12). Additional evaluations include an analysis of scene-dependent performance (Sec. 13), an ablation study on satellite DSM data (Sec. 14), and training comparisons that further demonstrate the effectiveness of our dataset (Sec. 15).

8. Further AnyVisLoc Dataset Details

More examples: **More examples:** Fig. 8 presents additional UAV image examples from the AnyVisLoc dataset, which includes data captured under diverse conditions, including different scenes, altitudes, viewpoints, weather conditions, seasons, and illumination conditions. Fig. 9 and Fig. 10 provide further examples of reference maps in the AnyVisLoc dataset. Additionally, our dataset covers various yaw angles, exhibiting a relatively even distribution (see Fig. 11). These diverse data characteristics enable a comprehensive evaluation of UAV AVL approaches.

More information: Tab. 8 provides additional details of the dataset, including the scene description, the geographic location of each reference map, the number of UAV images, the DJI drone types used, the spatial resolutions of the reference maps, and the coverage area of each region.

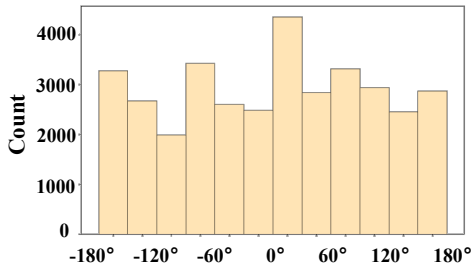


Figure 11. Distribution of Yaw Angle in AnyVisLoc Dataset.

9. Ground Truth Accuracy

The aerial reference map was reconstructed using highly overlapping images captured by the drone, which was equipped with a high-precision GNSS receiver (standard or RTK). Modern Structure-from-Motion techniques were employed, with the precise onboard GNSS data serving as initial positions. The resulting model was further refined through ground control point correction and global camera parameter optimization to ensure metric accuracy. Furthermore, high-precision registration was performed between the satellite and aerial reference maps to ensure the georeferencing accuracy of the satellite reference map.

10. Further Implementation Details

Drone-to-satellite geo-localization models [16, 50, 63, 65] are always trained with square images, whereas UAV images often have an aspect ratio (e.g., DJI Mavic 3 images have a ratio of 1.77:1). To maintain the model's input shape, we cropped the largest square region from the center of UAV images to serve as the query image (the corresponding gallery image is also square). During the image matching process, we used the rectangular UAV images and their corresponding square reference images.

11. Further Image Retrieval Details

Estimated Geo-Location of the Current Frame: As shown in Fig. 12, we use the latitude, longitude, and altitude of the camera (x, y, z), as well as pitch angle(θ) and yaw angle(ψ) to estimate the geo-location of the current frame , denoted as (X_g, Y_g). The formula is Eq. (4). These data can be acquired from the flight metadata.

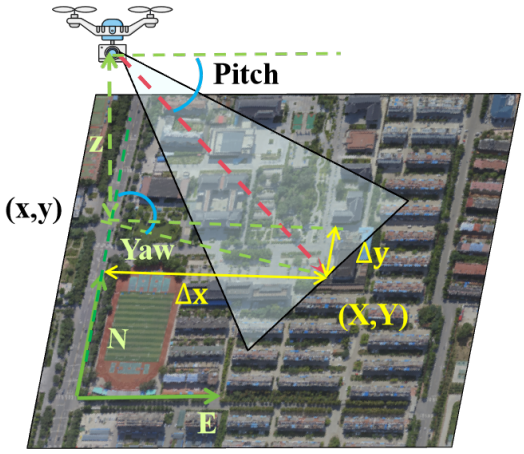


Figure 12. The Schematic Diagram of UAV Imaging Process. The ground is assumed as a flat surface. The red arrow represents the camera's optical axis.

$$\begin{aligned} \mathrm{X_g} &= x + \Delta x = x + \frac{z}{\tan(\theta)} \cdot \cos(\psi - \pi/2) \\ \mathrm{Y_g} &= y + \Delta y = y - \frac{z}{\tan(\theta)} \cdot \sin(\psi - \pi/2) \end{aligned} \tag{4}$$

Sampling Interval: In the image retrieval task, sub-images are cropped as gallery images from the reference map according to a certain sampling interval. A smaller sampling interval results in a shorter distance between the predicted position and the ground truth, but also significantly increases the computational cost. In the experiments of this paper, the sampling interval is set to 50% of the gallery image width w_g. Here, w_g is determined by $w_g = w_q \cdot (r_{UAV}/r_{map})$, where r_{UAV} and r_{map} are the

spatial resolution of the UAV image(see Eq. (7)) and the reference map, respectively. w_q denotes the width of query (UAV) image .

PDM@K: The spatial distance of image retrieval error is $d_i = \sqrt{(X_q - X_g)^2 + (Y_q - Y_g)^2}$, where (X_q, Y_q) denotes the retrieved position. Assuming r is the spatial resolution of the reference map, and w_i is the width of the gallery image(reference map tiles), then the ratio of the pixel distance to the image's width $R_i = d_i/(r \cdot w_i)$. The formula of PDM@K is Eq. (5).

$$PDM@K = \sum_{i=1}^{K} \frac{(K-i+1)\cdot e^{-\lambda\cdot(R_i-\alpha)}}{1+e^{-\lambda\cdot(R_i-\alpha)}} / \sum_{i=1}^{K}(K-i+1) \quad (5)$$

In Eq. (5), $\frac{e^{-\lambda\cdot(R_i-\alpha)}}{1+e^{-\lambda\cdot(R_i-\alpha)}}$ is the retrieval score of the i-th sample in the retrieval result order. λ and α are scaling parameters. In this paper, λ is set to 6 and α is set to 0.9. The reason for designing the function and selecting parameters in this manner is to make the relationship between R_i and localization accuracy more closely align with the actual distribution (as shown in Fig. 4 of the main text). Considering the geographical continuity of reference images in the UAV AVL task, we employ a weighted averaging approach to fully consider the positional distribution of the top K images. The weight assigned to the i-th sample is $(K-i+1)$.

Normalization: Since the AnyVisLoc dataset contains 25 regions, and the number of images for different regions and their corresponding gallery images varies, to ensure a fair contribution of different regions when calculating the final PDM@K metric, we adopted a normalization strategy. The formula is Eq. (6), where N_R is the number of regions(in our dataset, $N_R = 25$), N_i is the number of UAV images in the i-th region.

$$PDM@K^{norm} = \frac{1}{N_R}\sum_{i=1}^{N_R}\sum_{j=1}^{N_i}\frac{PDM@K_j}{N_i} \quad (6)$$

12. Prior Information Utilization

12.1. UAV Image Scale and Rotation Estimation

Current UAV platforms generally carry sensors that can provide information about pitch/yaw angles and altitude (e.g. gyroscopes and altimeters). In this paper, these prior information is used to align the UAV image and the reference map to the similar rotation and scale, reducing the search space for image retrieval and matching and improving localization accuracy. The details are given below.

Scale Estimation: As shown in Fig. 13, We use the camera's pitch angle(θ), φ (field of view), altitude(z), and the image size (width and height) to roughly estimate the spatial resolution r in the UAV image. The formula is Eq. (7).

$$r = \frac{z}{\sin(\theta)} \cdot \tan(\frac{\varphi}{2}) \cdot \frac{1}{\sqrt{(w^2+h^2)}} \quad (7)$$

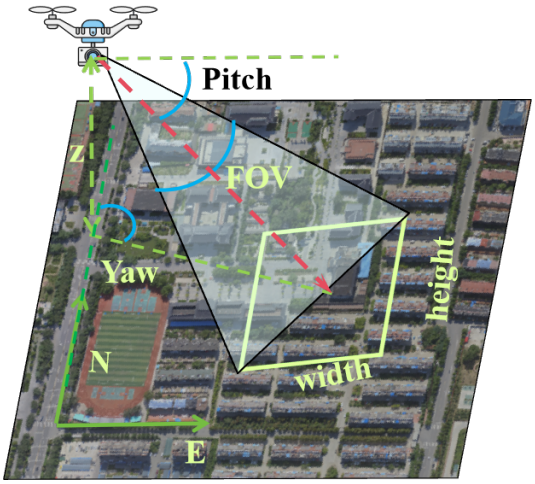


Figure 13. The Schematic Diagram of UAV Image Scale and Rotation Estimation. The ground is assumed as a flat surface. The red arrow represents the camera's optical axis, and the parallelogram area on the ground indicates the camera's field of view.

Rotation Estimation: Since the UAV cameras are attached to the gimbal, the camera's roll angle can be fixed at 0. Therefore, we can directly use the yaw angle to rotate the reference map to match the yaw angle of the UAV image.

12.2. Prior Altitude Noise

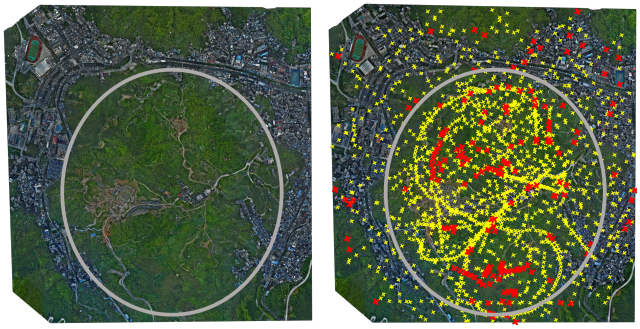
As seen in Eq. (7), **Altitude noise** also affects scale estimation, but its impact on localization accuracy is smaller(see Tab. 9). The main reason is that the standard deviation of the altitude noise is small relative to the absolute flight altitude, and this proportion decreases with increasing flight altitude. For a drone flying at 200m, a 30m altitude error introduces a 15% error in scale estimation. Existing image retrieval and matching algorithms are robust to this level of noise.

Table 9. Impact of Prior Altitude Noise on Localization Accuracy.

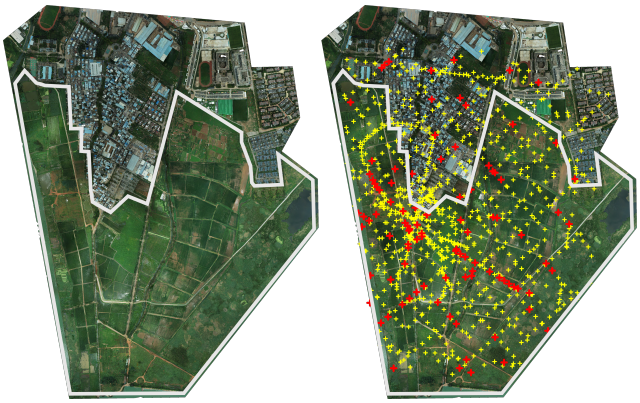
Noise Std	A@5m	A@10m	A@20m
0m	74.6	87.6	94.2
3m	74.4 (↓ 0.2)	87.8	94.3
5m	74.5 (↓ 0.1)	87.7	94.1
7m	75.1 (↑ 0.5)	87.9	94.1
10m	74.0 (↓ **0.6**)	87.2	94.3
20m	75.1 (↑ 0.5)	87.7	93.3
30m	74.4 (↓ **0.2**)	87.4	93.7

13. Impact of Different Scenes on Localization Accuracy

The various scenes in our dataset also enable us to evaluate the performance of localization methods across different environments. For example, we found that the localization accuracy in urban areas is significantly higher than that in natural scenes, such as mountainous areas (see Fig. 14a) and agricultural areas (see Fig. 14b). Natural scenes typically exhibit poor texture and more repetitive patterns compared to urban scenes with artificial structures, making image retrieval and matching more challenging and leading to insufficient robustness in localization.



(a) The Localization Result of Mountainous Scene.



(b) The Localization Result of Agricultural Scene.

Figure 14. The Impact of Different Scenes on Localization Accuracy. For each UAV image frame, if its central field of view is located at (x,y) on the aerial reference map, the point is marked red when the localization error is distributed within the maximum 10% range, otherwise, it is marked yellow. The silver polygon represents areas dominated by mountainous or agricultural fields.

14. Ablation Study of Satellite DSM Data

Although the Digital Surface Model (DSM) data of the satellite reference maps in our AnyVisLoc dataset has a coarser spatial resolution (30 m) than the aerial photogrammetry DSM, it remains crucial for accurate UAV AVL. This section compares the DSM-assisted PnP method (Tab. 6 in the main text) with an approach that assumes the reference area is a planar surface with uniform height. As shown in Tab. 10, even with this coarse satellite DSM, the assisted localization surpasses the planar assumption by 11.2% (A@5m), demonstrating its practical value.

Table 10. Comparison of Localization Accuracy.

Method	A@5m	A@10m	A@20m
Satellite Map with DSM (Our benchmark)	**18.5**	**38.7**	**58.5**
Satellite Map wo DSM (Planar assumption)	7.3	15.7	25.1

15. Training and Evaluation on AnyVisLoc

In this paper, we used the AnyVisLoc dataset to benchmark state-of-the-art image matching models. To leverage the models' optimal performance while testing their generalization capabilities, we adopted off-the-shelf models from existing works. In this section, we go a step further by partitioning the dataset into training and test sets. The training set contains 15 scenes, and the test set contains another 10 distinct scenes. This non-overlapping split is designed to validate model generalization to unseen environments.

To align the training data with the drone-to-map setting of the UAV AVL task, we generated high-precision correspondence labels between drone images and aerial maps using 3D reconstruction techniques (see Fig. 15). This approach provides image pairs and point correspondences captured from distinctly different viewpoints (drone perspective vs. aerial ortho-overhead perspective), which are more targeted to our specific task compared to those found in common matching datasets like MegaDepth and HPatches.

We trained two representative models from scratch: the SuperPoint+LightGlue combination (training only LightGlue) and RoMa. Their performance is compared with off-the-shelf models trained in other studies in Tab. 11. The results show that training on our relatively small AnyVisLoc dataset (10k image pairs) achieves performance comparable to models trained on the large-scale MegaDepth dataset (containing millions of pairs). This indicates that for this specific task, data relevance is more critical than sheer data volume. However, models trained with the GIM framework [44] (utilizing 100 hours of video) achieve higher accuracy, which is expected due to their vast data scale. Nonetheless, our approach offers a favorable balance between performance and data efficiency.

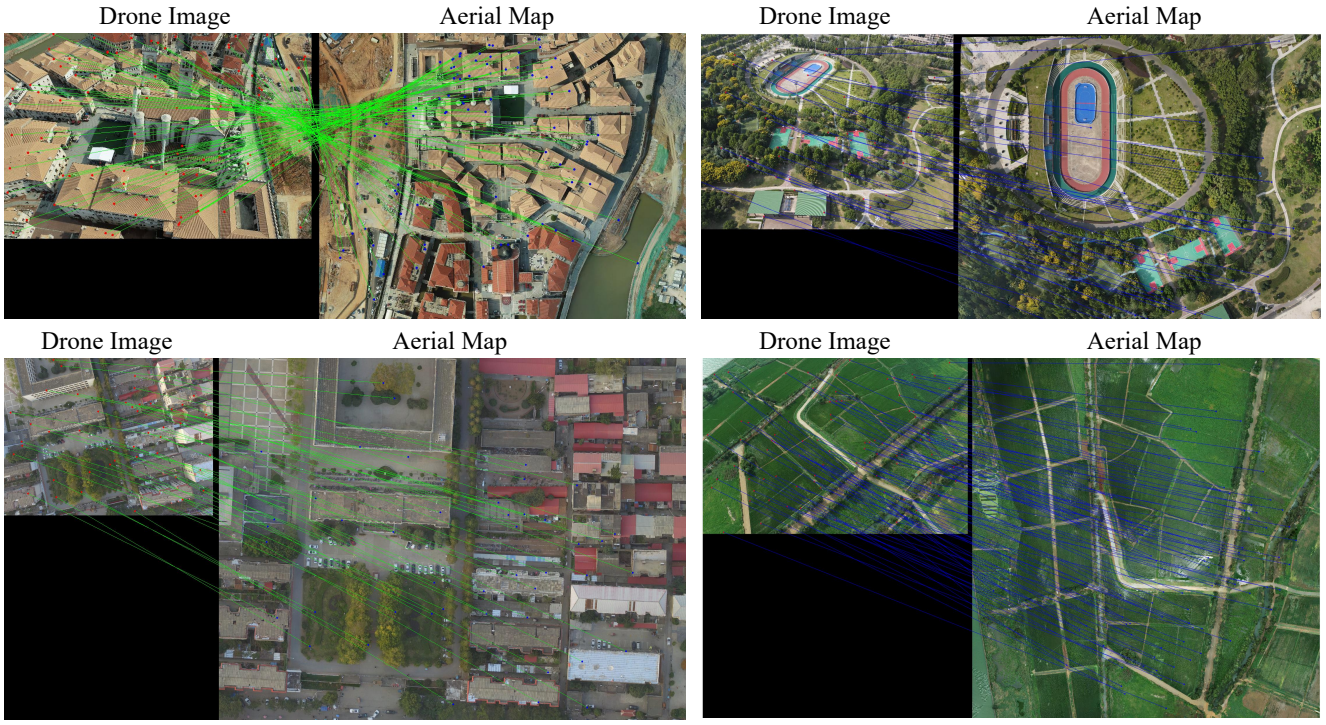


Figure 15. Example of Dense Correspondences between UAV Images and Aerial Maps. These point pairs are automatically generated via the 3D reconstruction pipeline and are used as ground-truth supervision for training the image matching models.

Table 11. Comparison of models trained on different datasets. The results demonstrate that models trained on our task-specific AnyVisLoc dataset achieve competitive performance with high data efficiency.

Model	**Training Dataset**	**Training Dataset Scale**	**A@5m**	**A@10m**	**A@20m**
SP [15]+LG [32]	AnyVisLoc (Ours)	∼10 thousand pairs	<u>41.8</u>	77.4	91.6
SP [15]+LG [32]	MegaDepth	10 million pairs	<u>41.8</u>	<u>80.2</u>	95.1
SP [15]+LG$_{MINIMA}$ [32, 40]	MegaDepth	10 million pairs	41.6	79.3	**95.7**
SP [15]+LG$_{GIM}$ [32, 44]	Internet Videos	100 hours video	**47.6**	**84.7**	**95.7**
RoMa [20]	AnyVisLoc (Ours)	∼10 thousand pairs	59.1	89.4	96.4
RoMa [20]	MegaDepth	10 million pairs	<u>60.1</u>	89.5	<u>96.5</u>
RoMa$_{MINIMA}$ [20, 40]	MegaDepth	10 million pairs	59.5	<u>89.7</u>	96.4
RoMa$_{GIM}$ [20, 44]	Internet Videos	100 hours video	**62.5**	**91.7**	**97.4**