RflyPano: A Panoramic Benchmark for Ultra-low Altitude UAV Localization Powered by RflySim

Dun Dai[1*], **Ze Lu**[1*], **Xunhua Dai**[2], **Quan Quan**[1, 3†]

[1]School of Automation Science and Electrical Engineering, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Central South University, Changsha, China
[3]Tianmushan Laboratory, Hangzhou, China
d_dai_3@buaa.edu.cn, luze@buaa.edu.cn, dai.xh@csu.edu.cn, qq_buaa@buaa.edu.cn

Abstract

Ultra-low altitude UAVs (below 120 meters) are gaining importance in the booming low-altitude economy, where GNSS signals are often unreliable or unavailable. Vision-based localization emerges as a promising alternative; however, existing benchmarks are not designed for ultra-low flight and typically adopt pinhole cameras with limited field of view, making them less effective in handling occlusions and repetitive textures near the ground. To address these limitations, we introduce the first panoramic UAV localization dataset tailored for ultra-low altitude scenarios. Built on a four-fisheye-camera system in the high-fidelity RflySim platform, our dataset captures diverse conditions — including day/night cycles, extreme weather, and dynamic obstacles — and contains over hundreds of thousands of frames. It is further enhanced with real-world UAV panoramic data to narrow the sim-to-real gap and will be continuously updated for broader applicability. Comprehensive experiments confirm the effectiveness and transferability of our dataset, establishing it as a robust benchmark for future research in vision-based UAV localization.

The project is publicly available at: —
https://github.com/DUNDAI1998/RflyPano

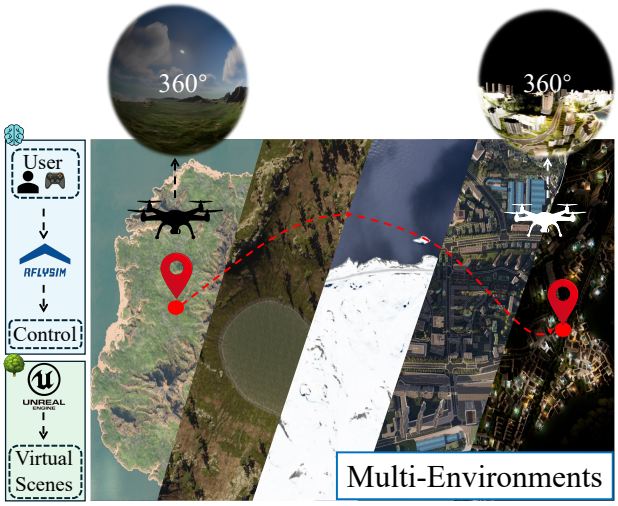

Figure 1: We propose the first UAV panoramic dataset with a virtual environment powered by RflySim.

Introduction

Uncrewed Aerial Vehicles (UAVs) have gained significant attention in both academic research and industrial applications (Quan 2017). In recent years, the low-altitude economy has become prosperous. For example, current research shows that the low-altitude aviation market will reach 1.5 billion USD by the end of 2025 (Huang 2024). Particularly, ultra-low altitude operations avoid conflicts with commercial air traffic and avoid regulated airspace in most countries. This unique capability allows for more precise monitoring, real-time intervention, and cost-effective solutions, making it an ideal choice for numerous applications in both urban and rural settings.

Given the growing significance of UAVs in the low-altitude economy, ensuring their safe flight is crucial. As the general solution, GNSS-based localization struggles in environments with weak or unavailable satellite signals, such as cities with high-rise buildings or areas subject to hostile interference (Balamurugan, Valarmathi, and Naidu 2016). Vision-based localization, using onboard cameras and inertial sensors, has emerged as an alternative solution. However, a major challenge is the lack of datasets covering ultra-low altitude, multi-environment scenarios from the UAV's perspective. Some pinhole-view datasets (Zheng, Zheng, and Yang 2020; Wu et al. 2024; Xu et al. 2024; Ji et al. 2025) address this gap, offering new possibilities for diverse views and scales.

Nevertheless, the pinhole views have a more limited field of view, particularly at ultra-low altitudes, where occlusion effects and repetitive textures become more pronounced, resulting in unreliable localization. To overcome these limitations, panoramic views offer a promising alternative. By capturing a 360° horizontal field of view, panoramic cameras significantly expand the observable area. This broader perspective allows UAVs to simultaneously perceive multiple directions, increasing the likelihood of identifying dis-

*These authors contributed equally.
†Corresponding author.

tinctive landmarks and reference points. As a result, issues caused by partial occlusions or repetitive textures are mitigated, enhancing the robustness and accuracy of visual localization. Consequently, panoramic UAV provide richer visual cues and facilitate more reliable localization methods and attracts increasing research interest (Lin et al. 2025), and major drone and camera manufacturers like DJI and Insta360 are also embracing panoramic vision.

We introduce the first panoramic UAV localization dataset, built upon the high-fidelity UAV simulation platform RflySim (Dai, Tu, and Quan 2025). By leveraging Unreal Engine 5 (UE5), we construct photorealistic synthetic environments featuring dynamic lighting, seasonal variation, and complex terrains. This not only ensures high-quality data but also significantly reduces the cost and difficulty of real-world data collection. Moreover, UE5 enables the simulation of rare or hazardous scenarios, such as extreme weather conditions. RflySim further enhances realism and flexibility by supporting both Software-in-the-Loop (SIL) and Hardware-in-the-Loop (HIL) simulations, enabling precise control of single UAVs and swarms. Compared to other simulators like AirSim (Shah et al. 2018), it offers broader support for vehicle types, richer sensor options, a distributed architecture, and seamless MATLAB/Simulink integration. A key feature of our dataset is the inclusion of fisheye camera simulations to enable panoramic views, a common practice in synthetic panoramic training (Murrugarra-Llerena, da Silveira, and Jung 2022). Unlike AirSim, which lacks native support for such lenses, RflySim allows flexible sensor configurations and higher frame rates, enabling dense, wide-FoV data collection tailored for UAV navigation and localization.

The main contributions of the paper are as follows:

- We present a novel ultra-low-altitude panoramic UAV dataset covering diverse conditions and scenes. We also include real-world validation data to facilitate studies on sim-to-real transfer.
- We develop a virtual environment based on RflySim that supports attitude and position control, high-level decision-making, etc. It also provides user-friendly APIs for customizing scenes and trajectories with automatic data recording.
- We assess our synthetic dataset through comparisons with real-world data, enabling evaluation across diverse localization scenarios and offering opportunities for future research.

Related Work

UAV Localization Dataset

Early widely-used UAV localization datasets(Glocker et al. 2013; Kendall and Cipolla 2017; Sturm et al. 2012; Schöps et al. 2017) primarily provide pinhole camera images along with extrinsic labels. However, due to cost constraints and the fact that these datasets were not specifically designed for UAVs, the types of scenes are limited. Furthermore, the absence of real-world environmental variations, such as changes in lighting, weather, and seasonal conditions, restricts the robustness of models trained on these datasets. In contrast, many recent UAV localization datasets — particularly those with downward or oblique views — employ labeled satellite imagery as reference data for cross-view localization. For example, Zheng, Zheng, and Yang (2020) propose the first 3-DoF UAV-view dataset using university satellite images, while Wu et al. (2024) extend this idea to 6-DoF estimation with a large-scale urban dataset built on a textured 3D real-world model.

With rising virtual image quality and concerns over safety and cost, many studies have turned to virtual environments for scalable data collection, repeatable testing, and simulating rare or hazardous scenarios. Early work, such as Mueller et al. (2018) uses UE4 to simulate UAV-view images with ground-truth poses. Li et al. (2023) later release an RGB-D localization dataset of urban pinhole images, while Ji et al. (2025) leverage modern video games to generate diverse scenes across flight altitudes.

Regarding panoramic imagery, early work such as Huang and Yeung (2022) present large-scale 3-DoF outdoor video sequences, while Huang et al. (2024) improve upon this with 360-degree images and structured reference-query splits. Synthetic data has also become mainstream in training UAV localization models (Trugillo Da Silveira and Jung 2017; Lai et al. 2019; Murrugarra-Llerena, da Silveira, and Jung 2022), and Liu et al. (2024a) propose a platform tailored for panoramic localization. Notably, Sun et al. (2021) pioneer a UAV-view panorama for aerial segmentation.

However, despite promising progress in panoramic localization, existing efforts remain ground-centric. To the best of our knowledge, no panoramic dataset has been developed specifically for UAV localization — motivating us to take a step forward.

UAV Vision Localization

The current methods for UAV localization can be categorized into two main approaches: 1) image retrieval, and 2) regression-based.

Image Retrieval is a widely used technique in visual-based UAV localization, where the UAV's position is estimated by matching key features extracted from real-time captured images with a pre-existing map or database of reference images. Traditional methods like NetVLAD (Arandjelović et al. 2018) use deep learning for image retrieval by encoding image features into compact vectors, but struggle under challenging conditions, such as rapid motion, where image quality is degraded. OpenIBL (Ge et al. 2020) integrates local descriptors into a global framework, yet it fails to maintain accuracy in dynamic environments with motion blur, such as those caused by vertical oscillations. Similarly, CosPlace (Berton, Masone, and Caputo 2022) improves robustness by considering spatial context, but this becomes distorted under high motion, leading to poor performance in unstable flight scenarios. DINO (Caron et al. 2021) and Swin Transformer (Liu et al. 2022), based on the Transformer architecture, improve feature matching by capturing long-range dependencies.

Regression-based localization is a technique that directly estimates the UAV's position and orientation (6-DoF pose) by learning a mapping from input data, typically images, to

Configuration	Dual Fisheye (Side)	Dual Fisheye (Top-Bottom)	Multi-Pinhole	Proposed
Hor./Ver. coverage	Near-full/Narrow	360°/180°	360°/Narrow	360°/Near-full
Image distortion	High	High	Low	Moderate
Feature completeness	Medium	Low	High	High
Hardware complexity	Low	Low	High	Moderate
Sketch				

Table 1: Comparison of the proposed Four Fisheyes and other widely-used configurations in panorama construction.

output coordinates using machine learning models. Unlike feature matching approaches that rely on keypoint extraction and matching, regression-based methods aim to predict the UAV's pose in a continuous manner based on the entire image. PoseNet (Kendall, Grimes, and Cipolla 2015) is the first to employ a CNN-based architecture to predict the absolute pose. Shavit, Ferens, and Keller (2021) propose a method that allows multi-scene estimation. Alternatively, Melekhov et al. (2017) estimate the relative pose from a pair of pinhole images, and Tu et al. (2024) later extend this approach to panoramic views. Unlike feature matching, regression-based methods require large-scale datasets with a wide variety of images to train the model effectively. The dataset should cover different environments, lighting conditions, UAV attitudes, and altitudes to ensure that the model generalizes well.

Camera Configuration for Panorama

Table 1 summarizes representative camera configurations commonly used for panoramic imaging. Dual-fisheye setups are typically categorized into side-by-side and top-bottom configurations. The side-by-side configuration, as described in Benseddik, Morbidi, and Caron (2020), is compact but tends to introduce noticeable distortion due to limited coverage. The top-bottom configuration (Zhang et al. 2016) better captures a 360° field of view; however, it may introduce higher distortion in the horizontal plane — where dense scene content often appears — potentially affecting feature extraction.

An alternative approach is the multi-pinhole setup (Caesar et al. 2020), which offers low-distortion imaging. However, because each pinhole camera covers a limited field of view, multiple cameras are typically required to achieve full panoramic coverage. This increases system complexity and may pose challenges for resource-constrained platforms such as UAVs. Additionally, the stitching process becomes more demanding as the number of cameras increases.

In our work, we adopt a four-fisheye camera configuration (Liu et al. 2024b), which provides a balance between hardware simplicity and spatial coverage. It enables near-complete panoramic capture with moderate distortion and fewer seams, helping preserve image features more effectively and facilitating downstream visual tasks.

Problem Formulation

In the proposed environment, the scene coordinate system is the North-East-Down (NED) coordinate frame. To simplify, the UAV-camera rigid body coordinate is written as the UAV coordinate. The arrangement of the system is illustrated in Figure 2.

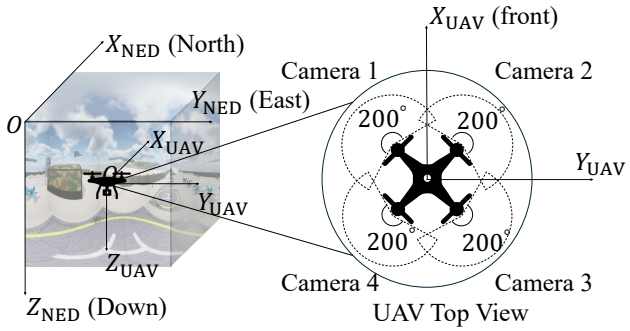

Figure 2: Four fisheye cameras are installed at the edges of the UAV frame.

In detail, four fisheye cameras are configured, each with a field of view of 200° and a resolution of 640 × 640 pixels. The angle between each camera is 90°, lying in the $X_{\text{UAV}}-O_{\text{UAV}}-Y_{\text{UAV}}$ plane. The panoramic image is constructed by stitching the four fisheye images to a panorama with a resolution of 1280 × 640 pixels.

Coordinate System

The following two right-hand coordinate systems(Quan 2017) are utilized in the dataset:

- NED coordinate system $\{O_{\text{NED}} - X_{\text{NED}}Y_{\text{NED}}Z_{\text{NED}}\}$. The origin is situated at a specific point within the environment, for instance, the take-off point, with the X_{NED} axis oriented towards the north and the Y_{NED} axis directed towards the east. The system is designed to describe the global pose of the UAV.
- Body coordinate system $\{O_{\text{UAV}} - X_{\text{UAV}}Y_{\text{UAV}}Z_{\text{UAV}}\}$. The origin is fixed at the center of mass of the UAV. The X_{UAV} axis points to the nose direction in the symmetric

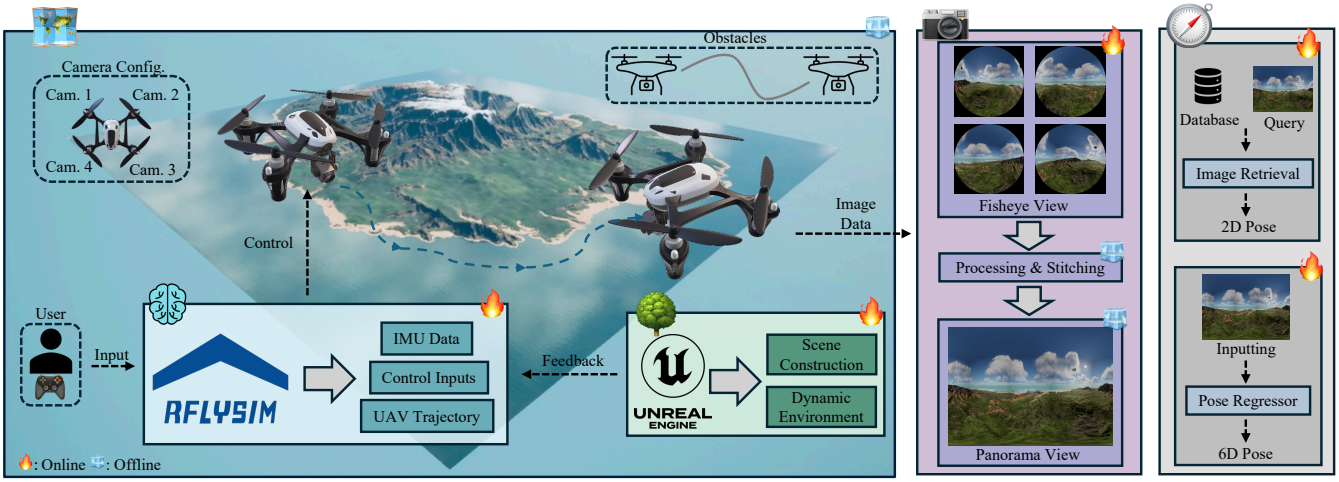

Figure 3: The overall pipeline of the proposed work consists of multiple integrated components. We first construct a virtual environment based on UE5, which simulates diverse realistic conditions for UAV navigation. The RflySim platform offers robust real-time control capabilities, enabling the simulation of actual flight dynamics. High-quality pose and IMU data are continuously recorded during flight. The camera module is responsible for capturing fisheye images and stitching them into panoramic images. Finally, we conduct experiments on standard localization tasks, demonstrating the dataset's suitability for general navigation applications.

plane of the UAV, while the Z_{UAV} axis points downward. It is employed to describe the pose of the cameras.

Localization

6-DoF UAV localization aims to estimate the pose of the UAV relative to the NED coordinate system. In this context, the pose refers to the UAV-camera rigid body coordinate relative to the virtual scene coordinate. Specifically, the 6-DoF pose is characterized by seven parameters: $\mathbf{T} = [\mathbf{x}^\top \ \mathbf{q}^\top]^\top \in \mathbb{R}^7$, where:

- $\mathbf{x} \in \mathbb{R}^3$ denotes the position, which accounts for the 3-DoF, representing the UAV's displacement along the X_{NED}, Y_{NED}, and Z_{NED} axes.
- $\mathbf{q} \in \mathbb{R}^4$ represents the rotation of the UAV in the form of a unit quaternion, capturing the 3-DoF in orientation. It is written as $\mathbf{q} = [q_w \ q_x \ q_y \ q_z]^\top$, providing a singularity-free representation of 3D rotation. These parameters define the unique rotation matrix $\mathbf{R}$.

Camera Model

We provide both a panoramic image and the corresponding four fisheye images as reference data, along with their corresponding pose labels for each frame. To convert four fisheye images into a single equirectangular panorama, we model each fisheye camera using an n-th order polynomial fisheye model, which is a radial polynomial function that maps the incident angle θ to the image radius r:

$$r(\theta) = \sum_{i=0}^{n} k_i \theta^{2i+1}.$$

where k_i are polynomial coefficients used to model the radial distortion.

To project the unit direction vector $\mathbf{d} = [d_x \ d_y \ d_z]^\top \in \mathbb{S}^2$ onto the fisheye image, we compute $\theta = \arccos(d_z)$, then evaluate $r(\theta)$ and map it to fisheye pixel coordinates (u, v) relative to the distortion center. Given a panoramic image of size $W \times H$, for each pixel (x, y), we compute the corresponding longitude and latitude:

$$\lambda = 2\pi \cdot \frac{x}{W} - \pi, \quad \phi = \frac{\pi}{2} - \pi \cdot \frac{y}{H}$$

and convert them into a unit viewing direction:

$$\mathbf{d}_{\text{scene}} = \begin{bmatrix} \cos\phi \sin\lambda \\ \sin\phi \\ \cos\phi \cos\lambda \end{bmatrix}.$$

This direction is rotated into the local camera frame using the inverse of its extrinsic rotation and mapped to the fisheye image as described above. We sample the corresponding pixel value via bilinear interpolation and average contributions from all valid cameras to form the final panorama.

RflyPano Dataset

We present a panoramic UAV view dataset in virtual environments with the RflySim platform, as shown in Figure 3, which consists of four fisheye camera images and corresponding stitched panoramic views captured at each frame, along with ground truth. Table 2 gives an overall insight into the proposed benchmark.

The ground truth label includes the following parameters: timestamp, position (x, y, z), velocity (v_x, v_y, v_z), orientation (roll, pitch, yaw in radians), and angular velocity

Aspect	Description
Simulation Platforms	RflySim (for UAV control and data logging in SIL mode) and UE5 (for photorealistic rendering)
Time Conditions	Day, night, and dusk with low illumination
Seasonal Variation	Summer and snowy winter scenes influencing the color and texture of environments
Dynamic Environment	Moving clouds, grass, trees, water, and dynamic light
Control Interface	Gamepad-based flight control and logging interface
Sim-to-real validation	Panoramic real-world dataset captured using a four-fisheye configured camera

Table 2: Overview of dataset simulation configurations and features.

$(\omega_x, \omega_y, \omega_z)$, all recorded from the RflySim to simulate IMU data.

Dataset Statistics

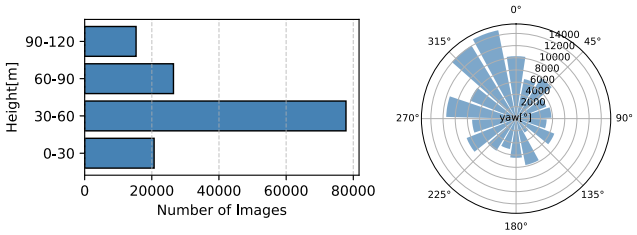

Figure 4: Summary Statistics Data of RflyPano. (Left) Pertains to Height; (Right) Concerns Yaw.

The dataset comprises 117 trajectories over 50 kilometers collected from 13 scenes, totaling over 140,000 panoramic images. Each image is annotated with rich metadata, including a 6-DoF pose, four corresponding fisheye views, a timestamp, and linear and angular velocities in the local NED coordinate system. The NED frame origin is defined at the starting point of each trajectory. Trajectory lengths vary depending on scene complexity. For example, urban environments with richer visual features typically result in longer and denser trajectories. All data instances are consistently labeled with pose and motion information to support accurate localization and analysis.

The attitude angles exhibit characteristic distributions (Figure 4). Roll centers around $0°$ within $[-25°, 20°]$. Pitch is concentrated near $-8°$ in the range $[-25°, 25°]$, reflecting forward thrust requirements. Yaw spans the full $[-180°, 180°]$ to cover all flight directions.

Data Collection

The pipeline of dataset collection and the following localization process is as illustrated in Figure 3. The simulation platforms used are RflySim and UE5. By encapsulating RflySim's control and logging APIs, our framework allows users to operate UAVs conveniently via a gamepad interface, while the data is automatically recorded. This design eliminates the need for users to have prior knowledge of RflySim and facilitates the seamless integration of custom scenarios into the simulation environment. The images collected cover a variety of environments with urban and in-the-wild settings (see Figure 5).

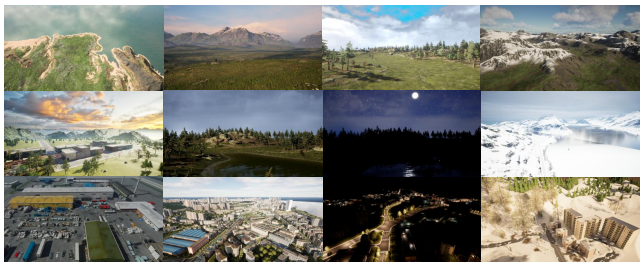

Figure 5: Overview of scenes from the RflyPano dataset, covering a wide range of high-resolution conditions.

To evaluate the sim-to-real gap of the proposed dataset, we are collecting a set of real-world panoramic images using cameras with the four-fisheye configuration mounted on an ultra-low-altitude UAV (see Figure 6) with a developed custom panoramic UAV (see Figure 7). Due to cost constraints, the current real-world dataset is captured under "Daytime" and "Nighttime" conditions only. The ground truth of 6-DoF pose, along with the linear and angular velocity, is derived from PX4's fusion of data collected by the IMU, GNSS, barometer, and magnetometer.

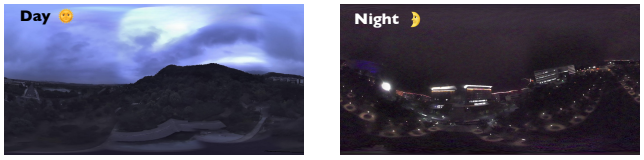

Figure 6: Samples of the daytime (in-the-wild) and nighttime (urban) real-world data captured by the custom UAV.

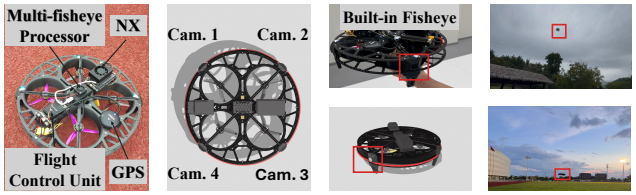

Figure 7: Fully custom integrated four-fisheye-camera UAV system with high-compute on-board devices for assessing sim-to-real gap.

Modeling of Dynamic Obstacles with RflySim

As illustrated in Figure 8, these obstacles can follow randomized or predefined trajectories, generating realistic motion patterns and occlusions that challenge visual localization systems. The simulation supports up to eight simultaneously moving obstacles, enabling the construction of complex and realistic scenes that closely mirror real-world flight conditions. Besides, the benchmark also provides a gamepad-based control and logging interface that streamlines UAV operation and data collection without requiring expertise in UAV control.

RflySim therefore provides a versatile and high-fidelity environment for UAV research, bridging the gap between virtual simulation and real-world deployment while enabling controlled studies of environmental dynamics.

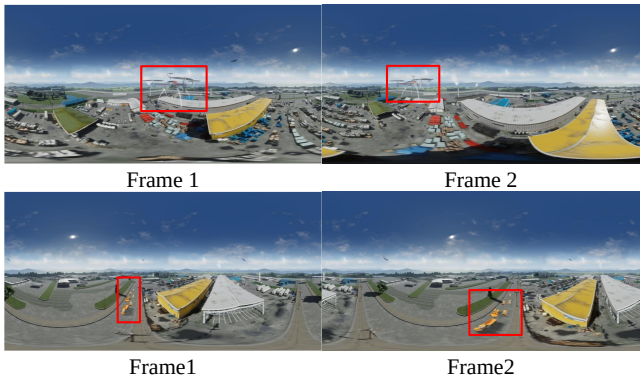

Figure 8: Example of dynamic obstacles.

Experiments

Experiment Settings

We evaluate our dataset under two fundamental UAV localization scenarios: 1) **Image Retrieval**, and 2) **Regression-based Methods**. Each scenario is tested under four representative environmental conditions:

1. **Daytime**, representing standard visual navigation conditions with well-lit scenes;
2. **Nighttime**, introducing severe challenges due to low visibility and limited texture information;
3. **Repetitive Structure** (RS), capturing highly repetitive and reflective structures in in-the-wild environments;
4. **Low-light conditions** (LI), simulating weather-induced illumination variations during UAV daytime cruising.

In addition, our dataset includes scenarios where the UAV experiences climbs and descents during flight. We classify the severity through differential calculations, categorizing them into sharp and mild scales of climb and descent to simulate the entire flight process of the UAV.

1) Image Retrieval We randomly select panoramic views of the same scene under the same conditions but different aerial trajectories as queries. Specifically, we denote the i^{th} feature from condition c and scene s as $\mathcal{F}^i_{c,s}$.

To evaluate its distinctiveness, we compute the L_2 distance between $\mathcal{F}^i_{c,s}$ and a reference feature set $\mathcal{R}^{\setminus i}_{c,s} = \left\{\mathcal{F}^j_{c,s} \mid j \neq i\right\}$, which includes the remaining panoramic views of the same scene under the same conditions.

We identify the nearest neighbor feature in the reference set as:

$$j^* = \arg\min_{j \neq i} L_2(\mathcal{F}^i_{c,s}, \mathcal{F}^j_{c,s}),$$

and define the final retrieval score as:

$$\mathbf{S}^i_{c,s} = L_2(\mathcal{F}^i_{c,s}, \mathcal{F}^{j^*}_{c,s}).$$

The corresponding 3-D coordinate of the retrieved image is denoted as $\mathbf{P}^{j^*}_{c,s}$. We compute the 2-norm error between the retrieved image's 3-D coordinate and the ground truth coordinate, using a threshold of $\tau = 1.0\,\text{m}$ to determine correctness.

We report Recall@1 (R@1) and Recall@5 (R@5), indicating whether the ground-truth is within the top-1 or top-5 retrieved results, and Precision@5 (P@5), representing the average correct ratio among top-5 candidates.

2) Regression-based Methods The primary goal of our experiments is to evaluate whether general models can learn effective representations from our dataset to handle the diverse conditions it contains. We apply the signature absolute pose estimation model (Kendall, Grimes, and Cipolla 2015), modified with a backbone structure following (Shavit, Ferens, and Keller 2021). For evaluation, we employ metrics such as the mean and median error of rotation and translation. The loss function $\mathcal{L}$ balances position and orientation errors as follows:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2 + \beta \left\| \mathbf{q} - \frac{\hat{\mathbf{q}}}{\|\hat{\mathbf{q}}\|} \right\|_2 \tag{1}$$

where $\mathbf{x}$ and $\mathbf{q}$ are the predicted position and orientation, $\hat{\mathbf{x}}$ and $\hat{\mathbf{q}}$ are the ground truth position and orientation, and β is a scaling factor to balance the two terms.

Image Retrieval Results

Diverse Environment Conditions According to Table 3, our evaluation of the RflyPano dataset highlights both the strengths and limitations of leading image retrieval-based localization models under diverse environmental conditions. While all methods perform well in standard daytime scenarios, they struggle in challenging environments. In "Night" and "Low Illumination" conditions, recall drops significantly, and in "Repetitive Structure" scenes, recall@1 hovers around 60%. The evaluation on real-world data (Daytime) shows results that are broadly consistent with those under the "Daytime" synthetic condition, suggesting a degree of sim-to-real generalization.

Descent/Climb Conditions The Table 4 presents the image retrieval results of RflyPano under Descent/Climb conditions, specifically simulating sudden vertical oscillations during UAV flight in "Climb" and "Descent" scenarios with the general "Daytime" setting. These results highlight the difficulty of handling sudden vertical oscillations, which introduce significant motion blur and noise in the imagery.

Model	Daytime Scene {001, 006, 008}			Night Scene {012, 013}			Repetitive Structure Scene 005			Low Illumination Scene 004			Real-World Data		
	R@1	R@5	P@5	R@1	R@5	P@5	R@1	R@5	P@5	R@1	R@5	P@5	R@1	R@5	P@5
NetVLAD (Arandjelović et al. 2018)	**0.946**	**0.967**	**0.929**	0.598	0.708	0.187	0.602	0.714	0.194	0.575	0.769	0.233	0.844	0.919	0.648
OpenIBL (Ge et al. 2020)	0.935	0.954	0.864	0.577	0.726	0.193	0.605	0.718	0.194	0.566	0.772	0.233	0.904	**0.965**	**0.735**
DINO (Caron et al. 2021)	0.885	0.923	0.724	**0.631**	**0.802**	**0.224**	**0.625**	**0.834**	**0.233**	**0.626**	**0.811**	0.234	0.747	0.914	0.572
CosPlace (Berton, Masone, and Caputo 2022)	0.942	0.959	0.917	0.570	0.751	0.211	0.606	0.728	0.200	0.555	0.802	**0.251**	**0.907**	0.958	0.728
SwinTransV2 (Liu et al. 2022)	0.944	0.965	0.927	0.593	0.732	0.196	0.593	0.704	0.188	0.570	0.791	0.243	0.903	0.962	0.703

Table 3: Image retrieval results of RflyPano under diverse environment conditions.

Model	Sharp Climb			Mild Climb			Sharp Descent			Mild Descent		
	R@1	R@5	P@5	R@1	R@5	P@5	R@1	R@5	P@5	R@1	R@5	P@5
NetVLAD (Arandjelović et al. 2018)	0.575	0.698	0.216	0.731	0.855	0.267	0.710	0.831	0.351	**0.823**	0.867	0.491
OpenIBL (Ge et al. 2020)	0.464	0.678	**0.264**	**0.758**	**0.862**	**0.271**	**0.750**	**0.833**	0.320	0.778	0.833	0.321
DINO (Caron et al. 2021)	**0.589**	0.698	0.254	0.717	0.841	0.253	0.711	0.819	0.344	0.808	**0.882**	0.488
CosPlace (Berton, Masone, and Caputo 2022)	0.500	0.678	0.257	**0.758**	0.839	0.269	0.708	0.812	0.341	0.750	0.833	**0.516**
SwinTransV2 (Liu et al. 2022)	0.547	**0.699**	0.216	0.710	0.841	0.262	0.722	0.807	**0.361**	0.779	0.852	0.361

Table 4: Image retrieval results of RflyPano under Descent/Climb conditions.

The poor performance of these well-established models emphasizes the value of our dataset in simulating real-world UAV challenges, offering a crucial benchmark for developing more robust models capable of handling such unpredictable flight dynamics.

Regression-Based Method Results

Figure 9 demonstrates the model's convergence during training with diverse environment conditions and demonstrates that the model achieves stable and effective learning, indicating its potential to generalize well across diverse environment conditions for reliable UAV localization. The stable convergence of the model, despite the complexities introduced by different environments, highlights the robustness of the dataset. It not only includes typical conditions but also tests the model's performance in more challenging situations, such as low light, repetitive structures, and varying terrain. Table 5 shows the numerical results of the pose esti-

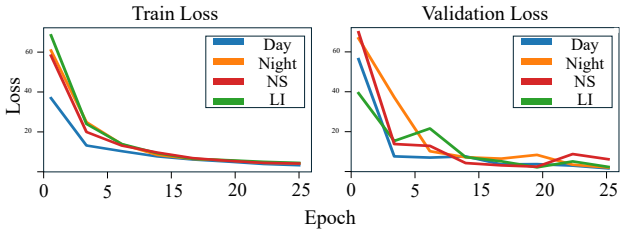

Figure 9: Train (Left) and Validation (Right) Loss of PoseNet across Different Environmental Conditions.

mation. While the mode performs well in typical conditions, such as daytime (with median rotation errors of $3.501°$), it faces challenges in more complex environments, like night, where rotation errors increase to $8.106°$.

Experiments suggest that although general models can learn useful information in standard conditions, they still struggle with the variability introduced by different environments and flight scenarios, specifically, during conditions like sharp climbs and descents, model performance drops. This indicates that further research is needed to improve model adaptability to dynamic flight maneuvers and challenging environmental conditions.

Exp. Setting	Rotation (°)			Position (m)		
	Mean	Med.	R@10° (%)	Mean	Med.	R@3m (%)
Diverse Environment Conditions						
Daytime	5.653	3.501	0.833	2.144	1.926	0.764
Night	9.102	8.106	0.693	2.263	2.082	0.884
RS	6.273	4.927	0.636	2.169	2.055	0.749
LI	8.452	6.256	0.819	1.693	1.449	0.629
Climb/Descent Conditions						
Sharp Climb	13.107	8.485	0.546	2.227	1.903	0.786
Mild Climb	5.092	4.564	0.894	1.935	1.545	0.842
Sharp Descent	15.769	11.951	0.475	1.486	1.294	0.916
Mild Descent	13.257	7.951	0.522	2.272	1.909	0.844

Table 5: The mean and median rotation/position errors in (m/°) over diverse conditions.

Conclusion

To the best of our knowledge, we present the first panoramic UAV localization dataset, which enables research beyond traditional pinhole camera models under ultra-low-altitude conditions. In addition, the environment includes multiple features to support dynamic and realistic UAV simulations. Using the dataset, we evaluate common localization tasks and show that general models can learn useful representations. Evaluation on real-world data suggests the potential of synthetic data for real-world applications, while revealing challenges that inform future improvements in model robustness. We have open-sourced both the dataset and the virtual environment and will continue to maintain and expand them. We note that the current real-world data for assessing the sim-to-real gap is limited by cost and policy, and extending it remains an important direction for future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62573021, Grant 62450127, Grant 62406345, and the Hunan Provincial Natural Science Foundation of China (2025JJ50341).

References

Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2018. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1437–1451.

Balamurugan, G.; Valarmathi, J.; and Naidu, V. P. S. 2016. Survey on UAV Navigation in GPS Denied Environments. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 198–204.

Benseddik, H.-E.; Morbidi, F.; and Caron, G. 2020. PanoraMIS: An Ultra-Wide Field of View Image Dataset for Vision-Based Robot-Motion Estimation. *The International Journal of Robotics Research*, 39(9): 1037–1051.

Berton, G.; Masone, C.; and Caputo, B. 2022. Rethinking Visual Geo-Localization for Large-Scale Applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4878–4888.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11618–11628. Los Alamitos, CA, USA: IEEE Computer Society.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Dai, X.; Tu, J.; and Quan, Q. 2025. RflySim ToolChain: a rapid development and validation toolchain for intelligent unmanned swarm systems. *Journal of Systems Engineering and Electronics*, 36(4): 1–19.

Ge, Y.; Wang, H.; Zhu, F.; Zhao, R.; and Li, H. 2020. 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*, 369–386. Springer.

Glocker, B.; Izadi, S.; Shotton, J.; and Criminisi, A. 2013. Real-Time RGB-D Camera Relocalization. In *2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 173–179. IEEE.

Huang, H.; Liu, C.; Zhu, Y.; Hui, C.; Braud, T.; and Yeung, S.-K. 2024. 360Loc: A Dataset and Benchmark for Omnidirectional Visual Localization with Cross-Device Queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, H.; and Yeung, S.-K. 2022. 360VO: Visual Odometry Using A Single 360 Camera. In *2022 International Conference on Robotics and Automation (ICRA)*, 5594–5600.

Huang, X. 2024. The Small-Drone Revolution is Coming — Scientists Need to Ensure it Will be Safe. *Nature*, 637: 29–30.

Ji, X.; Li, Y.; Wang, J.; Zhang, L.; Wang, Y.; and Wang, X. 2025. Game4Loc: A UAV Geo-Localization Benchmark from Game Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kendall, A.; and Cipolla, R. 2017. Geometric Loss Functions for Camera Pose Regression with Deep Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5974–5983.

Kendall, A.; Grimes, M.; and Cipolla, R. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2938–2946.

Lai, P. K.; Xie, S.; Lang, J.; and Laganière, R. 2019. Real-Time Panoramic Depth Maps from Omni-Directional Stereo Images for 6 DoF Videos in Virtual Reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 405–412.

Li, Y.; Jiang, L.; Xu, L.; Xiangli, Y.; Wang, Z.; Lin, D.; and Dai, B. 2023. MatrixCity: A Large-Scale City Dataset for City-Scale Neural Rendering and Beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3205–3215.

Lin, X.; Ge, X.; Zhang, D.; Wan, Z.; Wang, X.; Li, X.; Jiang, W.; Du, B.; Tao, D.; Yang, M.-H.; and Qi, L. 2025. One Flight Over the Gap: A Survey from Perspective to Panoramic Vision. arXiv:2509.04444.

Liu, H.; Zhao, L.; Peng, Z.; Xie, W.; Jiang, M.; Zha, H.; Bao, H.; and Zhang, G. 2024a. A Low-Cost and Scalable Framework to Build Large-Scale Localization Benchmark for Augmented Reality. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2274–2288.

Liu, P.; Feng, C.; Xu, Y.; Ning, Y.; Xu, H.; and Shen, S. 2024b. OmniNxt: A Fully Open-Source and Compact Aerial Robot with Omnidirectional Visual Perception. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10605–10612.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; Wei, F.; and Guo, B. 2022. Swin Transformer V2: Scaling Up Capacity and Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11999–12009.

Melekhov, I.; Ylioinas, J.; Kannala, J.; and Rahtu, E. 2017. Relative Camera Pose Estimation Using Convolutional Neural Networks. In *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings 18*, 675–687. Springer.

Mueller, M.; Casser, V.; Lahoud, J.; Smith, N.; and Ghanem, B. 2018. Sim4CV: A Photo-Realistic Simulator for Computer Vision Applications. *International Journal of Computer Vision*, 126.

Murrugarra-Llerena, J.; da Silveira, T. L. T.; and Jung, C. R. 2022. Pose Estimation for Two-View Panoramas Based

on Keypoint Matching: A Comparative Study and Critical Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 5202–5211.

Quan, Q. 2017. *Introduction to Multicopter Design and Control*. Springer Singapore.

Schöps, T.; Schönberger, J. L.; Galliani, S.; Sattler, T.; Schindler, K.; Pollefeys, M.; and Geiger, A. 2017. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2538–2547.

Shah, S.; Dey, D.; Lovett, C.; and Kapoor, A. 2018. AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In *Field and Service Robotics: Results of the 11th International Conference*, 621–635. Springer.

Shavit, Y.; Ferens, R.; and Keller, Y. 2021. Learning Multi-Scene Absolute Pose Regression with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2713–2722. Los Alamitos, CA, USA: IEEE Computer Society.

Sturm, J.; Engelhard, N.; Endres, F.; Burgard, W.; and Cremers, D. 2012. A Benchmark for the Evaluation of RGB-D SLAM Systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*.

Sun, L.; Wang, J.; Yang, K.; Wu, K.; Zhou, X.; Wang, K.; and Bai, J. 2021. Aerial-PASS: Panoramic Annular Scene Segmentation in Drone Videos. In *2021 European Conference on Mobile Robots (ECMR)*, 1–6.

Trugillo Da Silveira, T. L.; and Jung, C. R. 2017. Evaluation of Keypoint Extraction and Matching for Pose Estimation Using Pairs of Spherical Images. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 374–381.

Tu, D.; Cui, H.; Zheng, X.; and Shen, S. 2024. PanoPose: Self-Supervised Relative Pose Estimation for Panoramic Images . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20009–20018. Los Alamitos, CA, USA: IEEE Computer Society.

Wu, R.; Cheng, X.; Zhu, J.; Liu, X.; Zhang, M.; and Yan, S. 2024. UAVD4L: A Large-Scale Dataset for UAV 6-DoF Localization. In *Proceedings of the International Conference on 3D Vision (3DV)*.

Xu, W.; Yao, Y.; Cao, J.; Wei, Z.; Liu, C.; Wang, J.; and Peng, M. 2024. UAV-VisLoc: A Large-Scale Dataset for UAV Visual Localization. *arXiv preprint arXiv:2405.11936*.

Zhang, Z.; Rebecq, H.; Forster, C.; and Scaramuzza, D. 2016. Benefit of Large Field-of-View Cameras for Visual Odometry. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 801–808.

Zheng, Z.; Zheng, L.; and Yang, Y. 2020. University-1652: A Multi-View Multi-Source Benchmark for Drone-Based Geo-Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1988–1997.