# UAV Visual Localization in GNSS-Denied Environments

Tai-Cyuan Wang [1], Lai-Han Tsou [2], Jyun-Ping Jhan [3], Ru-Shan Tang [4]

[1] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST, Taiwan - m11452003@mail.ntust.edu.tw
[2] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST , Taiwan - p66111016@gs.ncku.edu.tw
[3] Graduate Institute of A.I. Cross-disciplinary Technology, NTUST, Taiwan - jpjhan@mail.ntust.edu.tw
[4] Systems Development Center, NCSIST, Taiwan - randydalph@gmail.com

**Keywords:** UAV, Visual Localization, Image Matching, GNSS-Denied Environment

**Abstract**

Navigating Unmanned Aerial Vehicles (UAVs) in Global Navigation Satellite System (GNSS)-denied environments requires reliable autonomous localization techniques. This study proposes a vision-based localization framework utilizing satellite true orthophotos and Digital Surface Models (DSMs) as absolute geospatial references. The algorithmic pipeline integrates deep learning architectures—specifically SuperPoint and LightGlue—to establish robust image-to-map feature correspondences. The matched correspondences are used to estimate camera exterior orientation parameters through collinearity-based spatial resection with an Iteratively Reweighted Least Squares (IRLS) approach. To validate the proposed methodology, a multi-altitude dataset (100–250 m) was acquired across structurally diverse terrains, including dense building, high vegetation, and bare ground areas. Experimental evaluations demonstrate that the framework achieves meter-level absolute positioning accuracy and stable pose estimation. Analyses further reveal that matching robustness and localization success rates depend heavily on terrain texture and flight altitude; geometrically structured urban scenes at moderate-to-high altitudes consistently yield reliable correspondences, whereas low-texture environments and lower flight altitudes present persistent challenges for continuous visual tracking.

## 1. Introduction

In Global Navigation Satellite System (GNSS)-denied environments, the autonomous navigation and operational safety of Unmanned Aerial Vehicles (UAVs) face severe challenges. Consequently, developing visual localization techniques that operate independently of external positioning signals has become critically important (Yao et al., 2024).

The fundamental principle of visual positioning is to estimate the camera’s exterior orientation parameters (EOPs) using reference data with known geospatial information. Common types of reference geospatial data include satellite imagery and Digital Surface Models (DSMs). Satellite imagery provides planimetric (2D) information, while DSMs provide elevation data. Therefore, the accurate identification of conjugate points through image-to-map matching between UAV images and reference satellite imagery is a key factor in determining the success of the localization process.

### 1.1 Related Work

Image matching is a fundamental task in computer vision and photogrammetry and is widely applied in Structure-from-Motion (SfM) and image registration. Traditional local feature methods, including SIFT (Lowe, 2004), SURF (Bay et al., 2008), and ORB (Rublee et al., 2011), have shown strong performance in single-view or richly textured scenes. However, in matching UAV and satellite imagery, they often fail to establish reliable correspondences due to radiometric differences, temporal gaps, and large-scale variations.

In recent years, deep learning techniques have improved the robustness of image-to-map matching. SuperPoint (DeTone et al., 2018) employs a self-supervised learning framework to extract highly repeatable feature points in complex and dynamically varying scenes. To alleviate the computational burden of deep learning-based feature matching, LightGlue (Lindenberger et al., 2023) was proposed. By incorporating an adaptive attention mechanism and an early-stopping strategy, LightGlue accounts for the spatial distribution of feature points while achieving an effective balance between matching reliability and computational efficiency. These characteristics make it particularly well suited for deployment on resource-constrained UAV platforms. As illustrated in Figure 1 traditional methods fail to establish reliable correspondences in the challenging UAV-to-satellite matching scenario, whereas SuperPoint and LightGlue achieve more robust and accurate matching results.

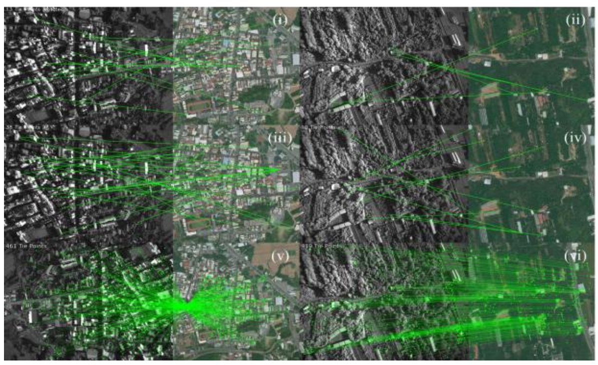


Figure 1 Visual assessment of matching reliability under severe appearance variations: (i)-(ii) ORB; (iii)-(iv) SIFT; (v)-(vi) SuperPoint and LightGlue.

### 1.2 Objectives

The key challenge of UAV visual positioning lies not only in establishing reliable image-to-map correspondences, but also in ensuring robustness under varying ground sampling distances (GSDs) and terrain textures. Earlier studies on geolocalization have demonstrated that viewpoint discrepancies can be partially mitigated through learned image-to-map representations between query images and georeferenced reference images (Workman et al., 2015). However, significant differences in acquisition time, severe perspective distortions, and variations in spatial resolution between UAV aerial imagery and satellite remote sensing imagery make image matching highly challenging (Zhu et al., 2023).

Although recent deep learning-based methods, such as SuperPoint and LightGlue, have improved matching reliability, their performance in terms of matching robustness and localization accuracy across diverse and complex terrains, as well as under different flight altitudes, remains insufficiently investigated. Therefore, this study aims to validate a practical image-to-map visual localization pipeline. The UAV imagery acquired at different flight altitudes is matched with satellite true orthophoto and DSM. The EOPs are then estimated via spatial resection (Dhaouadi et al., 2025). The localization accuracy is evaluated by comparing the estimated EOPs with those derived from SfM processing.

Rather than introducing a new feature matching architecture, this study focuses on systematically investigating how factors such as flight altitude, terrain texture, reference satellite tile size, and keypoint budget influence matching robustness, localization success rate, and EOP accuracy. In addition, because UAV operations in GNSS-denied environments do not provide direct online accuracy feedback, this study further incorporates a reliability-oriented verification criterion to assess whether the estimated pose is reliable for practical visual localization.

## 2. UAV System and Data Acquisition

Reliable visual localization requires the synergistic integration of high-performance imaging hardware and high-precision geospatial reference data. This section presents the UAV platform configuration, describes the characteristics of the study area and dataset, and defines the evaluation criteria and experimental conditions used to assess the reliability and accuracy of image-to-map matching under varying parameters and environmental conditions.

### 2.1 UAV System

The data acquisition was performed using a Vertical Take-Off and Landing (VTOL) UAV platform, as illustrated in Figure 2. The system integrates an industrial-grade monochrome camera, which provides higher spectral response and sharper image gradients than RGB sensors equipped with Bayer filters. The monochrome sensor has an image resolution of 1440 × 1080 pixels, with a pixel size of 3.45 μm and a focal length of 1.8 mm, resulting in a wide field of view (FOV) of approximately 100° × 80° to increase ground coverage, and a GSD of approximately 19 cm at a flight altitude of 100 m. Although this configuration achieves a relatively low GSD, it remains significantly higher than that of typical optical satellite imagery. Therefore, four different flight altitudes were designed in the actual flight missions to evaluate the effects of varying spatial resolutions and ground coverage on subsequent image matching performance.

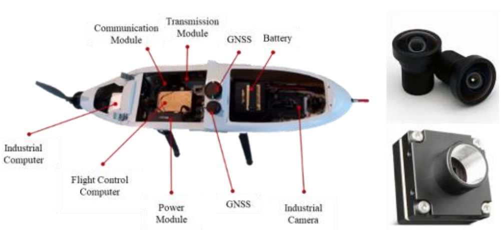


Figure 2. VTOL and monochrome industrial camera.

### 2.2 Flight Mission and Study Area

The flight missions were conducted in the Wuri District of Taichung City, Taiwan. As illustrated in Figure 3, the study area features topographic variations, with elevations ranging from 24 m to 244 m. To establish an absolute spatial reference for localization, a 50-cm resolution satellite true orthophoto and a 1-m resolution Digital Surface Model (DSM), covering a total extent of 25 km², were utilized as reference data. The UAV imagery was acquired in September 2025, whereas the reference orthophoto consists of a multi-temporal mosaic derived from imagery captured between December 2018 and January 2024. These datasets supply the requisite planimetric and elevation information for the derivation of virtual Ground Control Points (GCPs) and the subsequent execution of spatial resection.

The flight mission was designed to encompass multiple altitudes and representative terrain categories. Figure 3 presents the reference geospatial data, flight trajectories (e.g., indicated by the green line in panel ii), and sample scenes from High Vegetation Areas (HVA), Dense Building Areas (DBA), and Bare Ground Areas (BGA). The flight area covers approximately 6.712 km², with a total flight trajectory length of 52.03 km. Operations were conducted across four distinct altitudes, ranging from 100 m to 250 m, yielding an average flight distance of 13.01 km per altitude level. Table 1 summarizes the distribution of UAV images across these flight altitudes and terrain categories. The complete dataset comprises 3,770 images, consisting of 2,293 from HVA, 843 from DBA, and 634 from BGA. This multi-scale configuration provides a structured framework for evaluating the influence of terrain texture and flight altitude on image matching robustness and localization performance.

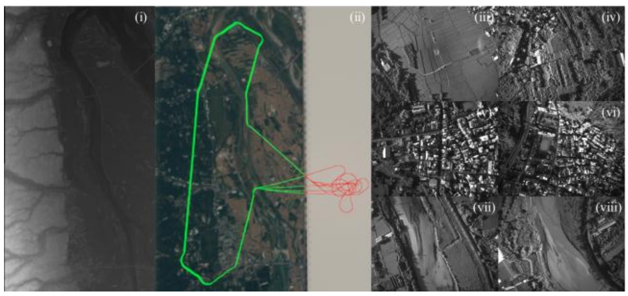


Figure 3. (i) DSM; (ii) Ture Orthophoto & Flight Trajectory; (iii)(iv) HVA; (v)(vi) DBA; (vii)(viii) BGA.

| Altitude | HVA | DBA | BGA |
|---|---|---|---|
| 250 m | 493 | 171 | 176 |
| 200 m | 608 | 198 | 156 |
| 150 m | 607 | 212 | 135 |
| 100 m | 585 | 262 | 167 |
| Total | 2293 | 843 | 634 |

Table 1. Number of images in each category & flight altitude.

### 2.3 Evaluation criteria and experimental conditions

To systematically evaluate matching performance and localization accuracy, the following experimental conditions are analyzed:

(a) **Algorithmic Parameters:** Evaluating the influence of reference satellite tile size and keypoint extraction limits on localization success.


https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-55-2026  

(b) **Altitude Changes:** Assessing the impact of varying flight altitudes (100–250 m) and the resulting GSD changes on feature repeatability.

(c) **Terrain Heterogeneity:** Evaluating localization performance across the terrain categories summarized in Table 1 to identify texture-dependent limitations in matching and EOP’s estimation.

## 3. Methodology

As illustrated in Figure 4, the proposed visual localization framework and its reliability assessment consist of four major components: (1) UAV image acquisition and data processing, (2) visual localization, including image preprocessing, image matching and spatial resection, (3) reliability and accuracy assessment under different conditions, and (4) conduct a global search to validate the robustness. Detailed descriptions are provided in the following sections.

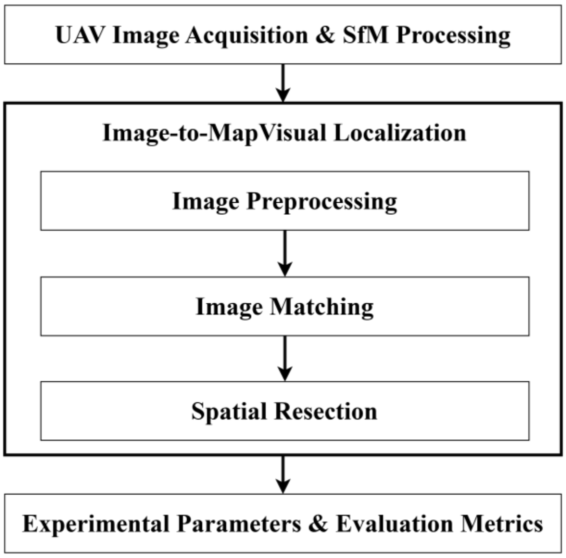


Figure 4. Image-to-map Visual localization.

### 3.1 UAV Image Acquisition and Data Processing

The multi-altitude image acquisition and terrain characteristics are detailed in Section 2.2. To validate the accuracy of the proposed image-to-map visual localization framework, a precise estimation of the image exterior orientation parameters (EOPs) is essential. Eight well-distributed and visually distinctive features (e.g., clear road markings) were selected from the aerial imagery as ground control points (GCPs), and their coordinates were measured utilizing real-time kinematic (RTK) GNSS equipment. Subsequently, the acquired UAV imagery and the GCP measurements were integrated into a Structure-from-Motion SfM pipeline using Agisoft Metashape. Through a bundle block adjustment within the software, highly accurate EOPs were derived to serve as the ground truth for evaluating the proposed framework. As illustrated in Figure 5, the UAV trajectories and the measured GCPs establish the geometric basis for this validation

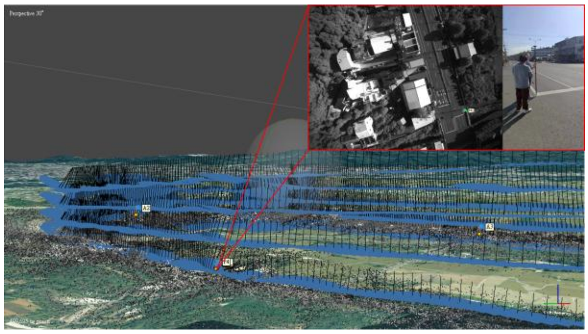

Figure 5. UAV trajectories and GCP measurements. The enlarged view highlights the aerial imagery and corresponding GCP, along with the RTK measurement procedure.

### 3.2 Image Preprocessing

A preprocessing stage is applied to both aerial and satellite data sources to mitigate scale and rotation discrepancies prior to feature matching.

**3.2.1 Aerial Image Resampling:** Deep-learning feature extractors typically lack inherent scale and rotation invariance. Consequently, incoming aerial images are rotationally rectified using prior heading estimates to align with the north-oriented reference map. To address scale discrepancies, the aerial images are dynamically resampled to match the spatial resolution of the satellite orthophoto. This process is guided by the ratio between the estimated Ground Sample Distance (GSD) of the UAV imagery at flight altitude $H$ and the fixed GSD of the satellite reference. This resampling process reduces metric-scale inconsistency between the UAV and satellite images and improves the stability of subsequent image-to-map matching.

**3.2.2 Reference Satellite Tile Size:** To compare the reliability and computational efficiency associated with different satellite tile sizes for matching, the satellite orthophotos were uniformly divided into fixed-size spatial tiles. The features of these patches are pre-extracted offline and structured into a georeferenced database. During the initialization phase, this tiled architecture accelerates the retrieval process by matching the query image against the database to identify the probable geographic region (as illustrated in Figure 6).

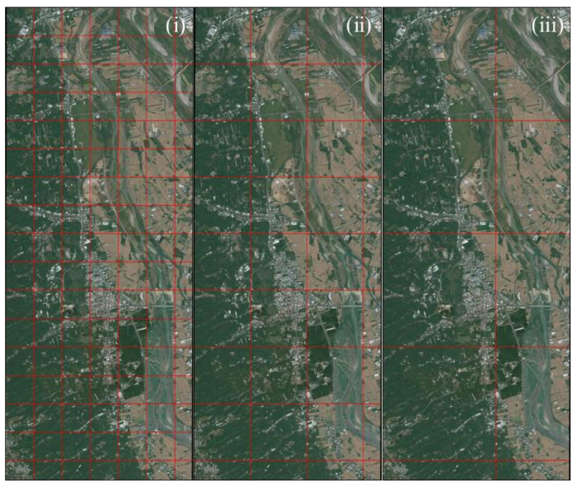


Figure 6. Satellite data with different Tile Size: (i) 1000 × 1000; (ii) 2000 × 2000; (iii) 4000 × 4000 pixels.


https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-55-2026  

### 3.3 Image Matching

To conduct the image-to-map visual localization, a fully neural feature matching pipeline is adopted. Rather than relying on traditional descriptors, this pipeline sequentially integrates self-supervised keypoint detection, context-aware matching, and a global retrieval strategy to ensure robust geometric correspondences.

**3.3.1 Local Feature Extraction**: As the initial step in the pipeline, SuperPoint is used to extract salient keypoints and their corresponding descriptors from the input images. Its self-supervised design enables robust and repeatable feature detection under varying illumination and viewpoint conditions. To balance robustness and efficiency, at most 2,048 keypoints are retained per image. For the reference database, satellite tiles are processed offline and stored as reusable feature files to avoid redundant computation during localization.

**3.3.2 Neural Feature Matching via LightGlue:** Given the extracted features, LightGlue is employed to establish image-to-map correspondences. Its transformer-based attention layers effectively capture visual and geometric consistency, enabling robust matching under viewpoint and scale variations. A reliable correspondence set is then obtained by selecting high-confidence matches.

### 3.4 Spatial Resection and Pose Estimation

Once reliable image correspondences have been established, each matched reference point is projected onto the DSM to obtain its georeferenced 3D coordinate. The original 2D–2D correspondences are therefore converted into 2D–3D observations, enabling absolute camera pose estimation through photogrammetric spatial resection.

**3.4.1 Coordinate Transformation and Distortion Correction**: Before solving the collinearity equations, the observed image coordinates are corrected for radial and decentring lens distortions. Concurrently, the 2D pixel coordinates of the matched points on the reference satellite orthophoto are projected onto the corresponding DSM to acquire their global 3D ground coordinates $(X, Y, Z)$ within the TWD97 coordinate system.

**3.4.2 Collinearity Equations and Least Squares Optimization:** The geometric relationship linking the 3D ground point $(X, Y, Z)$, the projection center $(X_s, Y_s, Z_s)$, and the corrected image points $(x', y')$ is mathematically modeled by the collinearity condition:

$$x' - x_p = -f\frac{m_{11}(X-X_s)+m_{12}(Y-Y_s)+m_{13}(Z-Z_s)}{m_{31}(X-X_s)+m_{32}(Y-Y_s)+m_{33}(Z-Z_s)} \quad (1)$$

$$y' - y_p = -f\frac{m_{21}(X-X_s)+m_{22}(Y-Y_s)+m_{23}(Z-Z_s)}{m_{31}(X-X_s)+m_{32}(Y-Y_s)+m_{33}(Z-Z_s)} \quad (2)$$

where $f$ is the focal length, and $m_{ij}$ represents the element of the $3 \times 3$ rotation matrix $R$, which is parameterized by the exterior orientation angles $(\omega, \phi, \kappa)$.

To solve for the EOP, the non-linear collinearity equations are linearized via Taylor series expansion around initial approximations to form the design matrix $A$ and the observation vector $L$. The initial position is approximated using the mean coordinates of the matched ground points, while the initial heading angle $\kappa$ is rapidly estimated via the dot and cross products of the corresponding image and ground vectors. A weighted least-squares approach iteratively solves the normal equations to derive the parameter increments $\Delta U$:

Since spatial resection is a non-linear optimization problem, robust initial values are strictly required to ensure convergence. The initial positional coordinates $(X_s, Y_s, Z_s)$ are approximated using the centroid of the matched ground control points and a scale factor derived from the spatial distances between corresponding points. The initial heading angle κ is robustly estimated by computing the dot and cross products of displacement vectors formed by adjacent feature points in both the image and object spaces, while roll ω and pitch ϕ are initialized near zero assuming a near-nadir flight configuration.

$$N = A^T PA, \;\; \Delta U = N^{-1} A^T PL \quad (3)$$

where $P$ denotes the weight matrix. The unknown vector $U$ is iteratively updated until the increments converge within a predefined tolerance.

**3.4.3 Robust Estimation:** Although the neural matcher provides a high-quality initial set of tie points, false correspondences may evade detection. To mitigate the adverse effects of these outliers, an Iteratively Reweighted Least Squares (IRLS) strategy is incorporated into the optimization process. At each iteration, the weight $P_{ii}$ for the $i$-th observation is dynamically adapted using its standardized residual $v_i$ through a piecewise robust estimator:

$$P_{ii}(|\tilde{v}_i|) = \begin{cases} 1 & , \quad if\ c_0 < |\tilde{v}_i| \le c_0 \\ \dfrac{c_0}{|\tilde{v}_i|}\left(\dfrac{c_1 - |v_i|}{c_1 - c_0}\right)^2 & , \quad if\ c_0 < |\tilde{v}_i| \le c_1 \\ 0 & , \quad if\ |\tilde{v}_i| > c_1 \end{cases} \quad (4)$$

where $c_0$ and $c_1$ are threshold constants controlling the confidence interval.

### 3.5 Experimental Parameters and Evaluation Metrics

To analyze the impact of different parameters on visual localization performance, three key factors are considered in this study: (1) number of keypoints, (2) different flight height, (3) terrain type, and (4) satellite image tile size. To evaluate the performance, three metrics are adopted: root mean square error (RMSE), success rate, and computational efficiency. RMSE quantifies the deviation between the estimated pose parameters and the corresponding ground-truth values; the success rate represents the proportion of frames that meet a predefined localization reliability criterion; and computational efficiency reflects the time cost associated with feature matching and pose estimation.

All experiments in this section were conducted on a desktop workstation equipped with an NVIDIA GeForce RTX 5080 (16 GB), an Intel Core Ultra 7 265K processor, and 64 GB of DDR5 memory.

## 4. Results and Analysis

This section presents the experimental results of the proposed visual localization framework. The analysis is organized into four parts, including multi-altitude evaluation, terrain-dependent performance, comprehensive localization analysis, and computational efficiency assessment.

### 4.1 Impact of Maximum Keypoint Quantity

To evaluate the influence of feature density on the proposed localization pipeline, we analyzed the system performance under different maximum keypoint limits. As shown in Figure 7, increasing the number of extracted keypoints produces a denser set of correspondences between the UAV imagery and the reference satellite orthophoto, which generally improves the structural completeness of the matched regions. This indicates that a higher keypoint budget can provide richer geometric information for subsequent pose estimation.

However, a larger number of extracted features does not necessarily lead to proportional gains in localization robustness. The quantitative results in Figure 8 reveal a clear trend of diminishing returns: a noticeable improvement in localization success rate is observed when the keypoint limit is increased from 1,024 to 2,048, whereas further increasing the limit to 3,072 or 4,096 yields only marginal performance gains while significantly increasing the computational burden of feature matching. Because the neural matching stage must process a substantially larger set of descriptors at higher keypoint limits, the per-frame execution time also increases accordingly. Therefore, considering both localization reliability and computational efficiency, a maximum threshold of 2,048 keypoints was adopted as the default configuration for subsequent experiments.

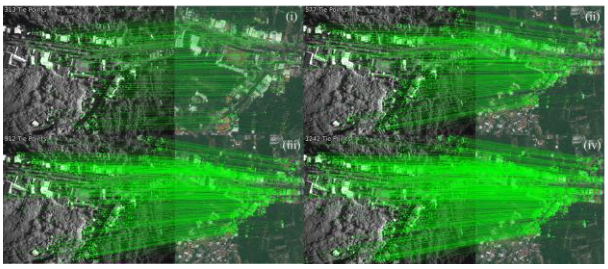


Figure 7. Comparison of feature matching under different numbers of keypoints. (i) 1024 points; (ii) 2048 points; (iii) 3072 points; (iv) 4096 points.

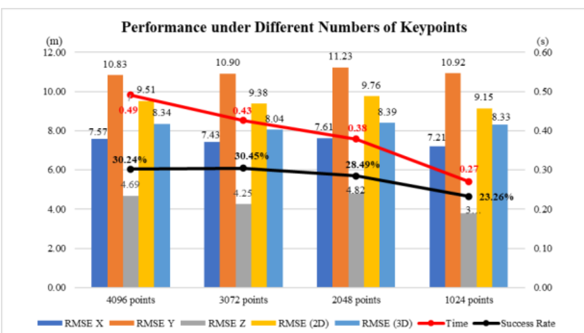


Figure 8. System performance under different numbers of keypoints.

### 4.2 Terrain-Dependent Analysis

As illustrated in Figure 9, the quality of image-to-map feature matching varies substantially across different terrain types. DBA exhibit the most reliable correspondences because building edges, road markings, and other man-made structures provide abundant and highly distinctive geometric features. In contrast, HVA produce fewer stable matches, as the scene is dominated by repetitive and weakly structured natural textures. BGA present the most challenging condition, where successful correspondences are limited to a small number of boundary features or terrain discontinuities, resulting in sparse and unstable matches.

The quantitative results in Figure 10 are consistent with the qualitative observations in Figure 9. DBA achieve the most stable localization performance, reflecting the advantage of well-distributed and repeatable keypoints. HVA show weaker performance because the extracted features are less distinctive and more sensitive to appearance variation. BGA exhibit the lowest robustness, as the limited availability of stable visual structures constrains both successful matching and the subsequent pose estimation. These results confirm that terrain texture and structural distinctiveness are major determinants of localization robustness in the proposed framework.

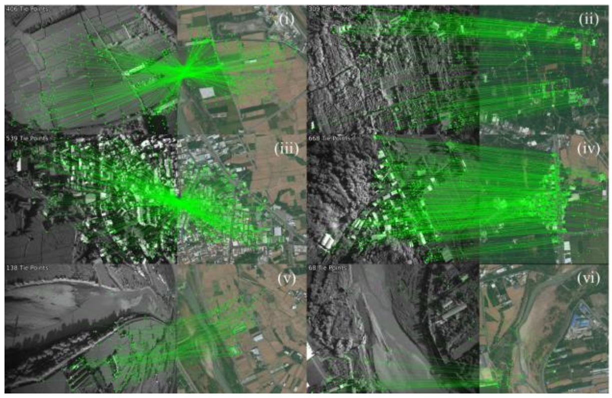


Figure 9. Feature matching results across different terrain types: (i), (ii) HVA; (iii), (iv) DBA; (v), (vi) BGA.

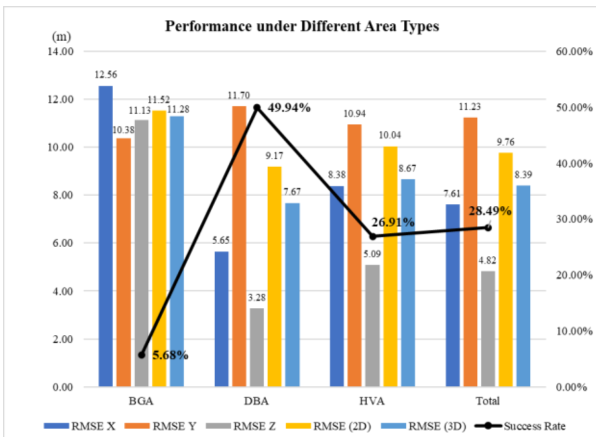


Figure 10. Localization performance across different terrain types.

### 4.3 Multi-Altitude Analysis

As illustrated in Figure 11, flight altitude has a clear influence on image-to-map matching behavior. When the UAV operates at higher altitudes, each frame covers a broader ground footprint and captures a more diverse mixture of buildings, terrain structures, and surface patterns. This broader spatial context generally provides a larger pool of extractable and repeatable keypoints, thereby improving the completeness and stability of the matched correspondences. In contrast, lower-altitude images contain a narrower field of view and more localized scene content, which reduces structural diversity and makes reliable matching more difficult.

The quantitative results in Figure 12 further confirm this tendency. Higher flight altitudes generally yield more stable localization performance because the wider scene coverage preserves richer geometric context for image-to-map matching and spatial resection. At lower altitudes, the limited footprint reduces the number of well-distributed correspondences and increases the risk of localization failure. Overall, these results indicate that moderate-to-high flight altitudes are more favorable for robust initialization and stable pose estimation in the proposed UAV visual localization framework.

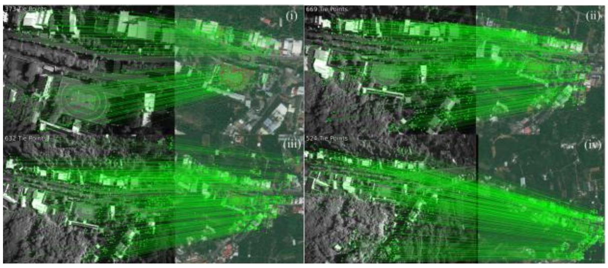


Figure 11. Feature matching results at multiple relative flight altitudes: (i) 100 m; (ii) 150 m; (iii) 200 m; (iv) 250 m.

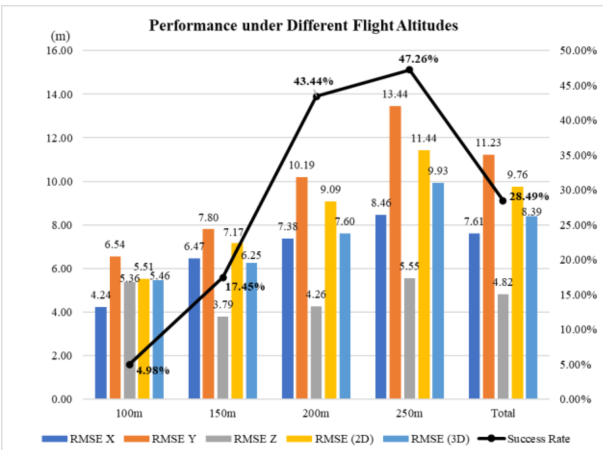


Figure 12. Localization performance under different relative flight altitudes.

## 4.4 Impact of Satellite Image Tile Size

Figure 13 and Figure 14 demonstrate that tile size introduces a clear trade-off between localization robustness and computational efficiency. From the qualitative matching results in Figure 13 smaller tiles generally preserve a tighter spatial context and produce more reliable correspondences between the UAV image and the reference patch. In contrast, larger tiles cover a broader spatial extent and include more irrelevant background content, which reduces the consistency of the matched structures. This tendency is further confirmed by the quantitative results in Figure 14. Smaller tiles (e.g., 1000 × 1000 pixels) yield a higher matching success rate, but they also generate a larger number of candidate patches, which substantially increases the search time during global localization. Conversely, larger tiles (e.g., 4000 × 4000 pixels) reduce the database size and improve retrieval efficiency but significantly decrease the matching success rate.

The degradation in performance with larger tiles can be attributed to the architecture of the neural matcher. LightGlue utilizes self- and cross-attention layers to explicitly encode the relative positional relationships among keypoints. When the spatial extent of the reference tile becomes considerably larger than the field of view of the UAV query image, the resulting scale discrepancy weakens the model’s ability to establish reliable structural consensus. Therefore, considering both matching robustness and computational efficiency, a tile size of 2000 × 2000 pixels was adopted for the subsequent experiments.

Table 2 further compares the ground coverage area of UAV images at different flight altitudes with that of orthophoto tiles under different tile sizes. The results show that the 1000 × 1000 tile provides the most similar spatial coverage to UAV imagery, especially for medium-to-high altitude flights, which explains its superior matching performance. In contrast, the 4000 × 4000 tile covers a substantially larger area than the UAV image footprint and therefore introduces excessive background content and greater structural inconsistency, leading to the weakest performance among the evaluated settings. Considering both robust localization and search efficiency, the 2000 × 2000 tile size was selected as the default setting for the subsequent experiments.

| Category | Condition | Area (km²) |
|---|---|---|
| Camera | Altitude 250 m | 0.37 |
| Camera | Altitude 200 m | 0.23 |
| Camera | Altitude 150 m | 0.13 |
| Camera | Altitude 100 m | 0.06 |
| Orthophoto Tile | 1000 x 1000 px | 0.25 |
| Orthophoto Tile | 2000 x 2000 px | 1.00 |
| Orthophoto Tile | 4000 x 4000 px | 4.00 |

Table 2. Ground Coverage Area of UAV Images and Orthophoto Tiles.

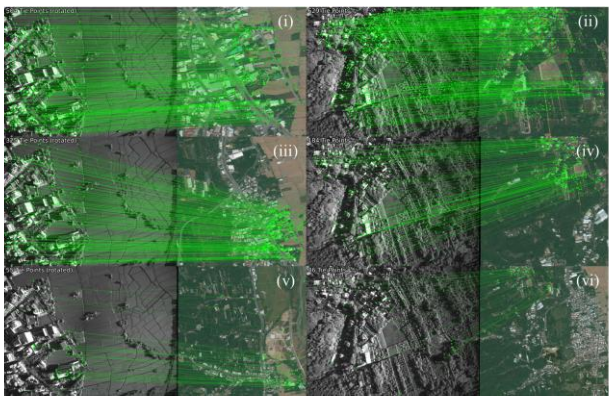


Figure 13. Comparison of feature matching under varying reference tiles.

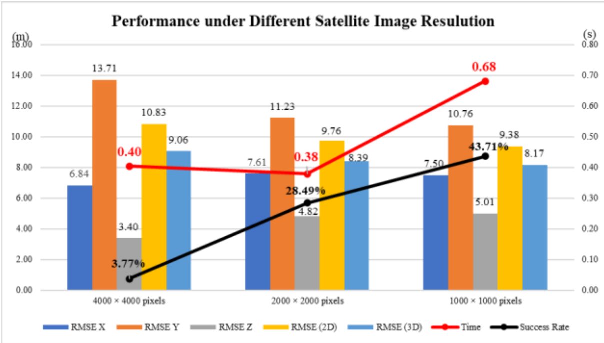


Figure 14. System performance under different satellite image tile resolutions.

## 5. Conclusion

This study presents a practical image-to-map visual localization pipeline for UAV operations in GNSS-denied environments by integrating satellite true orthophotos, DSMs, SuperPoint, LightGlue, and IRLS-based spatial resection. Experimental evaluation on a multi-altitude dataset covering dense building, high vegetation, and bare ground areas demonstrates that the proposed framework can achieve meter-level absolute positioning accuracy under favorable conditions.

The results further show that localization performance is strongly influenced by terrain texture, flight altitude, keypoint budget, and satellite tile size. Dense building areas provide the most reliable correspondences and the highest localization robustness, while high vegetation and bare ground remain more challenging because of weaker or less distinctive visual structures. In addition, images acquired at moderate-to-high flight altitudes generally yield more stable localization results because they contain broader spatial context and more well-distributed features. Future work may consider integrating short-term visual odometry or image-to-map domain adaptation techniques to further improve localization robustness in texture-poor scenes and under larger appearance discrepancies between UAV imagery and satellite references.

## Acknowledgements

This work was supported by the National Chung-Shan Institute of Science and Technology under the project entitled “Geographical Image Positioning of Unmanned Aerial Vehicles” (Project No. NCSIST-ACOM-114-172C001). The authors would like to express their sincere gratitude for this support.

## References

Bay, H., Tuytelaars, T., Van Gool, L., 2008. SURF: Speeded-Up Robust Features. Computer Vision and Image Understanding, 110(3), 346–359. DOI: 10.1016/j.cviu.2007.09.014.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pp. 224–236. doi.org/10.1109/CVPRW.2018.00060.

Dhaouadi, O., Marin, R., Meier, J., Kaiser, J., Cremers, D., 2025. OrthoLoC: UAV 6-DoF Localization and Calibration Using Orthographic Geodata. *arXiv preprint*, arXiv:2509.18350. doi.org/10.48550/arXiv.2509.18350.

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17581–17592. doi.org/10.1109/ICCV51070.2023.01616.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110. doi.org/10.1023/B:VISI.0000029664.99615.94

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An Efficient Alternative to SIFT or SURF. In: *2011 International Conference on Computer Vision (ICCV)*, pp. 2564–2571. doi.org/10.1109/ICCV.2011.6126544

Workman, S., Souvenir, R., Jacobs, N., 2015. Wide-Area Image Geolocalization With Aerial Reference Imagery. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3961–3969. doi.org/10.1109/ICCV.2015.451.

Yao, F., Lan, C., Wang, L., Wan, H., Gao, T., Wei, Z., 2024. GNSS-denied geolocalization of UAVs using terrain-weighted constraint optimization. *International Journal of Applied Earth Observation and Geoinformation*, 135, 104277. doi.org/10.1016/j.jag.2024.104277.

Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., 2023. 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(4), 1824–1837. doi.org/10.1109/TCSVT.2023.3249204.


https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-55-2026 | 