ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Sat-RoMa: Cross-Scale Dense Matching for Multi-Temporal UAV-to-Orthophoto Registration

Maciej Krupka[1], Jan Węgrzynowski[1,2], and Piotr Skrzypczyński[1]

***Abstract*—Reliable Global Navigation Satellite System (GNSS) signals are increasingly denied or jammed in real-world applications, such as search and rescue operations. In such scenarios, Unmanned Aerial Vehicles (UAVs) must rely on downward-facing cameras for absolute localization against reference satellite maps. While Visual Inertial Odometry (VIO) is highly accurate locally, it inevitably accumulates drift over time. Localizing a drone image against a pre-existing satellite map (e.g., Google Earth) via homography estimation is a viable solution, but it is severely challenged by seasonal variations, construction, and vegetation changes. In this paper, we propose *Sat-RoMa*, an end-to-end robust dense feature matcher adapted from the state-of-the-art RoMa architecture. By utilizing a frozen, pre-trained DinoV3 encoder specifically tuned for satellite imagery, and formulating the task as matching a small drone image to a $4\times$ larger reference map, Sat-RoMa explicitly handles scale discrepancies and temporal appearance changes. Preliminary results demonstrate that Sat-RoMa significantly outperforms baselines like LoFTR and LightGlue, achieving an 11.2% scale error compared to over 100% for existing methods, paving the way for robust GPS-denied UAV navigation.**

I. INTRODUCTION

In an increasing number of critical applications—such as search and rescue operations in jammed or contested areas—there is a strict requirement for a GPS-denied navigation system. To achieve absolute positioning without GNSS, it is highly desirable to utilize a downward-facing camera to localize the Unmanned Aerial Vehicle (UAV) against a georeferenced map.

While state-of-the-art Visual Inertial Odometry (VIO) and Visual SLAM systems [1]–[3] provide excellent local state estimation, they inevitably experience drift over long trajectories. For a drone flying hundreds of kilometers, even a 1% drift rate results in unacceptably large absolute translation errors. Consequently, there is a critical need to acquire periodic measurements against a global reference map to cancel this unavoidable visual odometry drift. To accomplish this, one must compute a precise homography estimation between the live drone camera feed and reference satellite imagery.

A similar line of work exists in cross-view geo-localization and image retrieval. Methods based on contrastive learning, metric learning, or masked reconstruction objectives [4], [5] can be used to select and retrieve relevant map tiles. However, as stand-alone models, these retrieval networks output a similarity score or classification; they cannot provide the precise relative position and orientation (6-DoF or homography) measurement required for continuous metric UAV localization.

Fig. 1. The severe challenges of satellite image matching. **Top row:** An urban parking lot environment exhibiting structural and moving-object changes over three years (2020-2023). **Bottom row:** An agricultural scene demonstrating drastic appearance changes due to seasonal vegetation cycles and crop harvesting (2017-2022).

Matching real-time drone imagery to archived satellite imagery poses a near-impossible task for even the newest deep-learning-based feature matchers [6], [7]. This difficulty stems from severe appearance variations caused by seasonal changes, vegetation differences, new constructions (roads, buildings), and the presence or absence of moving objects (e.g., cars) that act as unreliable visual references (as illustrated in Fig. 1).

To solve these challenges, we propose to end-to-end train a state-of-the-art image matcher—adapted from RoMa [8]—to make it robust and invariant to these temporal changes. To explicitly align the matcher with the task of drone localization, the architecture is designed to handle extreme scale differences, where the reference satellite image is $4\times$ larger than the live drone image query. Furthermore, we replace the standard DinoV2 backbone with a dedicated DinoV3 encoder pre-trained on satellite images. By keeping this encoder frozen during training, we inherit its large-scale generalization capabilities while focusing the matching heads on the geometric alignment task.

Contributions: Our main claims and contributions are as follows: (1) We highlight that current state-of-the-art image matchers are not dedicated to satellite imagery, which limits their direct use for GPS-denied navigation. (2) We propose **Sat-RoMa**, a matching architecture designed to match small live drone images with much larger global reference maps. (3) By leveraging a frozen, pre-trained DinoV3 encoder and cross seasonal training we achieve strong robustness

[1]The authors are with the Institute of Robotics and Machine Intelligence, Poznan University of Technology, Poznan, Poland. maciej.krupka@put.poznan.pl

[2]Jan Węgrzynowski is also with the IDEAS Research Institute, Warsaw, Poland.

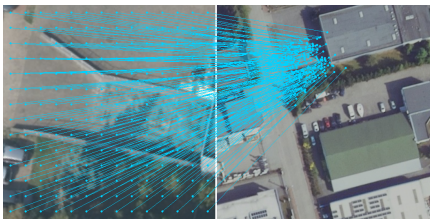

Fig. 2. Qualitative visualization of **Sat-RoMa**. The network successfully establishes dense, accurate patch-space matches between a small UAV camera patch (left) and a much larger, differently scaled reference satellite map (right), despite differences in lighting and time of capture.

TABLE I
QUANTITATIVE COMPARISON ON 180 CROSS-SEASON SATELLITE PAIRS.

Method	Reproj. Err. (px) ↓	Scale Err. (%) ↓	Rot. Err. (°) ↓
SAT-RoMa (Ours)	**42.3**	**11.2**	**11.1**
LoFTR [6]	265.6	519.9	31.2
LightGlue [7]	314.5	2937	45.5
RoMa [8]	390.8	6189	82.4

to seasonal and temporal changes and outperform existing baselines.

II. METHODOLOGY

To address the limitations of generic matchers when applied to remote sensing, we introduce Sat-RoMa. The core framework builds upon the principles of RoMa (Robust Dense Feature Matching) [8], but introduces critical modifications for the UAV-to-satellite paradigm, where experiments on satellite pairs approximate nadir UAV imagery.

A. Architecture and Training

In Sat-RoMa, we replace the standard backbone with a DinoV3 vision transformer that has been specifically pretrained on massive datasets of satellite and remote sensing imagery. Crucially, during our training phase, we keep this DinoV3 encoder **frozen**. This design choice prevents catastrophic forgetting of the foundational satellite features and enforces generalization, ensuring the model remains robust to the seasonal and structural variations shown in Fig. 1.

We construct a specialized dataset where image pairs consist of satellite tiles captured at the same geographic coordinates but across different seasons and years. Sat-RoMa is trained end-to-end on this data to output dense warp fields and match confidence maps, yielding precise correspondences as seen in Fig. 2.

III. EXPERIMENTS AND RESULTS

We compare Sat-RoMa to several state-of-the-art image matchers to demonstrate strong performance in the UAV-to-Map localization task.

A. Baselines and Setup

We compare Sat-RoMa against representative state-of-the-art methods, including RoMa [8], LoFTR [6], and LightGlue [7], which serve as strong baselines for dense feature matching. All methods are evaluated under a consistent experimental setup on cross-season satellite image pairs.

B. Discussion of Results

Table I presents a pixel-space comparison across multiple geometric error metrics. Sat-RoMa drastically outperforms both LoFTR and LightGlue across all evaluated dimensions.

Most notably, Sat-RoMa achieves a **Mean Reprojection Error of 42.3 pixels**, reducing the error of LoFTR (265.5 px) and LightGlue (314.4 px) by approximately 6–7×, highlighting a substantial improvement in geometric alignment accuracy.

The most significant advantage of our method is observed in the **Scale Error**. Because the reference map is $4\times$ larger than the query image, generic matchers struggle immensely to estimate the scale transformation, resulting in a 519.9% scale error for LoFTR and a catastrophic 2937% error for LightGlue. SAT-RoMa, constrained by the pretrained satellite features and trained specifically for this asymmetric matching, achieves a remarkably low scale error of just **11.2%**. Furthermore, SAT-RoMa reduces rotational errors to 11.1° compared to LoFTR's 31.2° and LightGlue's 45.5°. While LoFTR and LightGlue produce dense correspondences, the resulting homographies exhibit geometric distortions that render them unsuitable for metric localization—RoMa's rotation error alone exceeds 82°, effectively producing random orientations. In contrast, SAT-RoMa consistently recovers geometrically faithful transformations, reducing reprojection error by $6.3\times$ over the next best method.

IV. CONCLUSION

In this work, we presented Sat-RoMa, a novel approach to dense feature matching tailored specifically for cross-scale and cross-temporal drone-to-satellite localization. By modifying the RoMa architecture to utilize a frozen, satellite-pretrained DinoV3 backbone, and training it on a dataset featuring severe seasonal changes and $4\times$ scale disparities, we overcome the limitations of standard matchers. Our method consistently reduces scale and reprojection errors compared to state-of-the-art baselines like LoFTR and LightGlue, suggesting its potential viability as a drift-correction mechanism for GPS-denied UAV navigation.

REFERENCES

[1] C. Campos *et al.*, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.

[2] S. Leutenegger, "OKVIS2: Realtime scalable visual-inertial SLAM with loop closure," *arXiv preprint arXiv:2202.09199*, 2022.

[3] A. Fontan *et al.*, "VSLAM-LAB: A comprehensive framework for visual SLAM methods and datasets," in *Proc. IEEE/RSJ IROS*, 2025.

[4] Q. Wu *et al.*, "Camp: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.

[5] U. Mall, B. Hariharan, and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *Proc. CVPR*, 2023.

[6] J. Sun *et al.*, "Loftr: Detector-free local feature matching with transformers," in *Proc. CVPR*, 2021.

[7] P. Lindenberger *et al.*, "Lightglue: Local feature matching at light speed," in *Proc. ICCV*, 2023.

[8] J. Edstedt *et al.*, "Roma: Robust dense feature matching," in *Proc. CVPR*, 2024.