Pose-Constrained UAV Cross-View Geo-Localization with Satellite Imagery for GNSS Correction

Nikolas Potamianos[1], Alexandros Gkillas[1], and Aris S. Lalos[1]

Industrial Systems Institute, Athena Research Center, Patras Science Park, Greece
{potamianos.nick,gillas,lalos}@athenarc.gr

Abstract. This work tackles the problem of correcting noisy GNSS measurements using cross-view alignment between nadir aerial imagery and georeferenced satellite maps. We preprocess each incoming aerial frame into a stabilized nadir view by compensating roll and pitch using inertial attitude estimates and, when the onboard camera introduces strong lens distortion, by applying an appropriate rectification step so that the downstream stages remain camera-agnostic. We then estimate per-frame horizontal translation and yaw corrections by matching transformer-based dense descriptors between the aerial view and satellite tiles. The matching stage is spatially regularized. Instead of searching the entire satellite tile, each aerial pixel is matched only within a local window around its pose-predicted landing point. We then solve a robust geometric refinement over the regularized matches and convert the per-frame correction back to geodetic coordinates to obtain a corrected GNSS stream suitable for VIO or SLAM fusion. The resulting visual–inertial fusion reaches accuracy close to RTK-derived reference measurements across the evaluated datasets. We also release a new RTK-geotagged nadir aerial imagery and video dataset captured over a site in Western Greece.

Keywords: GNSS correction · Cross-view alignment · UAV navigation · Aerial imagery · Satellite imagery · Visual–inertial fusion.

1 Introduction

Reliable global positioning is central to UAV navigation, mapping and inspection, where aerial products such as orthomosaics, surface models and topographic maps must be accurately aligned to a common world frame across flights and over time [21, 17]. In navigation, global position also provides the absolute reference required for long-range operation and for limiting drift in onboard state estimation. This reference is most often provided by commodity Global Navigation Satellite System (GNSS) receivers, which offer lightweight and globally available position measurements [9]. However, low-altitude flight near buildings, infrastructure, vegetation or partially occluded sky can induce multipath and non-line-of-sight (NLOS) reception, leading to errors of several meters and, in challenging urban environments, even tens of meters [13, 19]. High-precision

GNSS workflows, including real-time kinematic (RTK), post-processed kinematic (PPK) and PPP-RTK, can provide centimeter-level positioning under favorable conditions, but their reliability still depends on satellite visibility, correction delivery, ambiguity fixing and proper system integration [21, 17, 19]. As a result, the same environments that degrade standard GNSS can also make RTK-based solutions intermittent or unreliable.

From an estimation perspective, reliable map-referenced UAV pose has been pursued through local state estimation and GNSS-aided fusion. Visual odometry, visual-inertial odometry and SLAM estimate platform motion from onboard sensors, but their trajectories remain locally referenced and accumulate drift over long distances [7]. GNSS-aided visual-inertial systems reduce this drift by fusing local motion estimates with GNSS measurements [3, 10]. However, their performance is still governed by the quality of the GNSS input. When the GNSS measurement is biased, the fused trajectory can be anchored to an incorrect map-referenced position [22, 23]. Correcting the GNSS measurement before fusion is therefore complementary to SLAM and visual-inertial estimation, rather than a replacement for them. A complementary line of work uses georeferenced satellite maps as global visual references for UAV localization. Existing methods align onboard UAV imagery with satellite maps to recover absolute pose in GNSS-degraded or GPS-denied settings [8, 12, 2]. For example, Patel et al. [12] register UAV images against rendered Google Earth views selected from nearby candidate poses, while Bianchi and Barfoot [2] learn compact satellite-image representations for UAV-to-satellite localization. These methods demonstrate the value of satellite imagery for UAV pose estimation, but they are primarily designed for image-based localization or filtered pose recovery rather than for correcting an existing GNSS measurement. In addition, approaches that rely on pre-rendered or pre-selected reference views can require a reference database around the expected operating region. More broadly, retrieval-based cross-view methods depend on selecting the correct satellite candidate or local reference region [26, 5]; therefore, visually ambiguous satellite regions can lead to large localization errors, especially in repetitive or weakly distinctive scenes, as also emphasized by recent large-scale UAV localization benchmarks such as UAV-VisLoc [24] and ALTO [4].

The present paper adopts a different formulation. Rather than localizing the UAV from scratch against a satellite image database, satellite imagery is used to correct an already available GNSS measurement, even when that measurement is biased. This GNSS measurement also provides a coarse spatial prior: it retrieves a local satellite tile and defines the region where the drone image is expected to align. Within this local region, cross-view alignment estimates the residual latitude, longitude and yaw correction. The output is a timestamp-aligned corrected GNSS stream suitable for downstream fusion. This formulation is related to automotive cross-view localization, where ground-level or vehicle-mounted images are aligned with satellite maps to estimate vehicle pose [16, 6, 1]. However, the geometry of nadir UAV correction is substantially different. Automotive methods mainly address forward-facing images and large perspective changes between query and reference. In contrast, the drone image in our setting is approximately top-down and therefore closer to the satellite image viewpoint. The dominant difficulties are therefore not perspective transfer, but two factors specific to the nadir UAV regime: residual roll and pitch deviations that perturb the apparent ground-plane geometry of the drone image, and geographic ambiguity in images with repeated patterns, such as forest canopy, road markings, or crop rows.

The second difficulty is particularly important because matching between drone-image features and satellite-image features is a central stage in cross-view localization pipelines, both in UAV and automotive settings. Existing methods rely either on classical feature matching or on learned descriptors, but in both cases the matching is ultimately driven by visual similarity [15, 18, 11]. This becomes problematic in satellite images with repeated patterns, where forest canopy, road markings, crop rows or nearly identical rooftops may produce many locally similar candidate regions. When a matcher is allowed to compare a drone-image pixel with candidates across the satellite tile, nearby drone-image pixels can be assigned to distant and unrelated neighborhoods in the satellite image. The resulting correspondence field loses spatial consistency: nearby drone pixels no longer map to nearby satellite locations, or they collapse onto an incorrect repeated structure. Although these matches may be plausible according to descriptor similarity, they are wrong in geographic space. Once such an inconsistent correspondence field is used for geometric pose refinement, the estimated correction can converge to the wrong satellite region instead of the true alignment. Fig. 3 illustrates this failure, where pixels from the drone image are matched to an incorrect region of the satellite tile.

In light of this, we propose a novel cross-view correction architecture for satellite-aided GNSS refinement. The pipeline first converts each incoming drone image into a stabilized nadir view by compensating residual roll and pitch from inertial attitude estimates, with lens rectification applied when required by the camera model. It then uses the GNSS measurement as a coarse spatial prior to predict where each drone-image pixel should lie in the satellite image. During matching, each pixel is allowed to search only within a local metric window around this predicted location, rather than across the entire satellite tile. This pose-conditioned search reduces matches to distant but visually similar regions while preserving enough freedom to estimate the remaining translation and yaw error. The resulting correspondences define constraints between drone-image pixels and satellite-image locations. The refined drone pose is obtained by finding the SE(2) transformation whose projected drone pixels best agree with these matched satellite locations. This robust optimization is solved with Levenberg–Marquardt and estimates the residual horizontal translation and yaw correction. The contributions of this work are:

– **Satellite-Aided GNSS Correction as Measurement-Level Cross-View Alignment.** We formulate drone-to-satellite cross-view alignment as measurement-level GNSS correction rather than global re-localization from scratch. Unlike prior UAV-to-map methods that use a coarse prior only to seed a local search and then output an absolute UAV pose, the proposed architecture keeps the available GNSS measurement as the primary quantity and estimates only a residual position and yaw correction, producing a timestamp-aligned corrected GNSS stream for downstream VIO/SLAM fusion.

- **Pose-Conditioned Spatial Regularization for Dense Drone-to-Satellite Matching.** A pose-conditioned matching strategy is introduced to address a common limitation of descriptor-based matching: in repetitive satellite imagery, standard matching can assign nearby drone-image pixels to distant but visually similar satellite regions. The proposed strategy restricts each pixel search to a local metric window around its pose-predicted satellite location, preserving spatial consistency for pose refinement. In the ablation study, this reduces the corrected RMSE from 16.36 m to 2.33 m, an 85.7% improvement over unconstrained appearance-based matching.
- **Corrected GNSS Stream from the Proposed Cross-View Module for Visual–Inertial SLAM Fusion.** The corrected GNSS stream produced by the proposed cross-view module is evaluated as a global-position input to VINS-Fusion. Compared with native GNSS fusion, it substantially improves trajectory accuracy and reaches performance close to using RTK-derived reference measurements.
- **RTK-Geotagged Nadir Drone Dataset.** A new RTK-geotagged nadir image and video dataset is introduced, captured with a DJI Mavic 3 Enterprise over a site in Western Greece. It covers buildings, roads, vegetation and nearby residential/rural areas, supporting evaluation of satellite-aided UAV localization and GNSS correction under realistic low-altitude conditions.

2 Proposed Methodology

Fig. 1 shows the proposed cross-view GNSS correction pipeline. Given a nadir UAV image, GNSS measurement, initial yaw, and IMU roll/pitch estimates, the method first produces a stabilized nadir view through attitude compensation and optional lens rectification. Dense UAV–satellite correspondences are then computed using pose-conditioned spatial regularization, which limits matching to local metric windows around the predicted satellite locations. A robust SE(2)[1] refinement finally estimates the residual translation and yaw correction and outputs the corrected GNSS measurement.

2.1 Preprocessing and satellite retrieval

Direct cross-view matching is affected by platform attitude and, for wide-FOV cameras, lens distortion. We therefore preprocess each frame into an approximately perspective nadir view with consistent ground-plane geometry. The preprocessing includes optional lens rectification, applied only when distortion is

[1] SE(2) denotes the two-dimensional special Euclidean group, i.e., planar rigid-body transformations composed of a rotation and a translation. In this work, it corresponds to yaw correction and horizontal East–North translation correction.

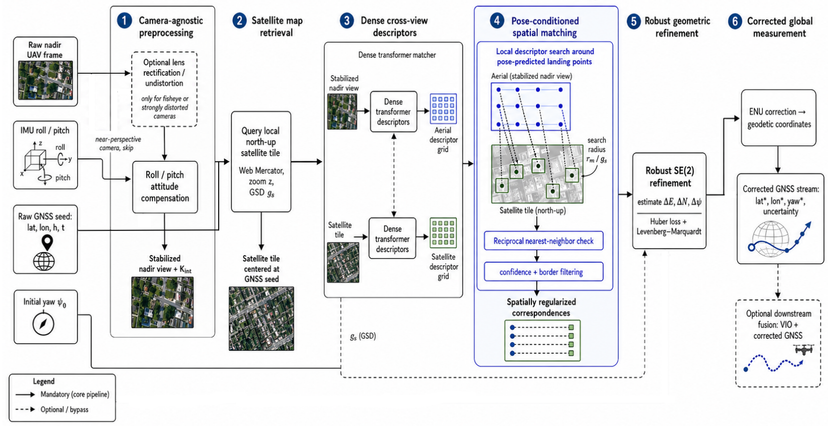

Fig. 1. Overview of the satellite-aided GNSS correction pipeline.

non-negligible, and attitude compensation using inertial roll and pitch estimates. The result is a stabilized nadir image and a corresponding perspective intrinsic matrix K_{flat}, used downstream for scale estimation and metric-to-pixel mapping.

Stabilized nadir view generation. Let I_{raw} be the input UAV image. If the camera is fisheye or strongly distorted, rectification is included in the inverse warp through the calibrated projection model $\pi(\cdot; K, D)$ where K and D denote the raw camera intrinsic matrix and distortion parameters, respectively; otherwise the distortion parameters D are omitted. The stabilized nadir image I_d is obtained by mapping each output pixel $\tilde{\mathbf{p}} = [u, v, 1]^\top$ to the original image as $\mathbf{x} = \pi(R_{\text{comp}}^\top K_{\text{flat}}^{-1} \tilde{\mathbf{p}}; K, D)$ and sampling $I_d(u, v) = I_{\text{raw}}(\mathbf{x})$. The compensation rotation is $R_{\text{comp}} = R_x(-\theta) R_y(\phi)$, where ϕ and θ are platform roll and pitch. This optional rectification handles lens distortion, while roll/pitch compensation removes attitude-induced ground-plane tilt and preserves yaw for satellite alignment, keeping the mapping close to the planar homography assumed by the later image-to-map refinement.

Satellite tile retrieval. For each stabilized drone image at timestamp t_k, the nearest valid GNSS measurement is used to retrieve a local north-up satellite tile centered at the corresponding (lat, lon). The GNSS measurement is used only as a coarse spatial prior that defines the local search region, not as the final corrected position. The retrieved tile has size $W_s \times H_s$ and known ground sampling distance g_s. Satellite tiles are retrieved through the Google Maps and Azure Maps APIs at zoom levels 18–19, yielding $g_s \approx 0.4$–0.8 m/pixel.

Ground sampling distance. At latitude φ and map level z, the satellite ground sampling distance (GSD) is approximated by

$$g_s(\varphi, z) = \frac{2\pi R_\oplus \cos\varphi}{256 \cdot 2^z} \quad [\text{m/pixel}], \tag{1}$$

where $R_\oplus$ denotes the Earth radius, assuming Web Mercator tiles.

2.2 Spatially regularized cross-view correspondence

Given the stabilized drone image and the satellite tile centered at the GNSS seed, cross-view correspondences are estimated using dense transformer-based descriptors. The GNSS-based pose prior is used to spatially constrain the descriptor search: each drone-image pixel is matched only within a local metric window around its predicted location in the satellite image. This pose-conditioned search reduces geographically inconsistent matches in repetitive satellite regions while preserving the appearance-based matching capability of the descriptor model.

Feature extraction and reciprocal matching. Let I_d be the stabilized nadir UAV image from Sec. 2.1, and let I_s denote the satellite tile. The matcher computes dense descriptor tensors $\mathbf{F}_d \in \mathbb{R}^{H_d \times W_d \times C}$ and $\mathbf{F}_s \in \mathbb{R}^{H_s \times W_s \times C}$ (and, when available, confidence maps $\mathbf{q}_d, \mathbf{q}_s$). To obtain correspondences efficiently, we sample a subset of aerial pixel locations on a regular grid with stride s_{stride} (subsampling factor). For each sampled location $\mathbf{p}_i = (u_i, v_i)$ in I_d, we do not search the whole satellite tile; instead, we restrict the candidate region to a local window $\Omega_s^{(i)} \subset \Omega_s$ centered at the predicted landing point $\tilde{\mathbf{p}}'_i$ obtained from the current affine prior (Eq. (2) evaluated at $\boldsymbol{\theta} = \mathbf{0}$), with a radius $r_{\text{px}} = r_{\text{m}}/g_s$ derived from an expected metric search radius r_{m}. Within this window we find the most similar descriptor,

$$\mathbf{p}'_i \;=\; \arg\max_{\mathbf{p} \in \Omega_s^{(i)}} \; \langle \mathbf{F}_d(\mathbf{p}_i), \mathbf{F}_s(\mathbf{p}) \rangle,$$

This spatially regularized search prevents the matcher from selecting visually plausible but geographically distant candidates in satellite tiles with repeated structure, such as parallel rooftops, tree rows and highway markings. Without this constraint, a small number of large matching errors can dominate the error tail, as quantified in Sec. 3.6. We then enforce a reciprocal consistency check, where a candidate match $(\mathbf{p}_i \leftrightarrow \mathbf{p}'_i)$ is retained only if $\mathbf{p}_i$ is also the best match when searching back from $\mathbf{p}'_i$ to the aerial image. This mutual nearest neighbor criterion substantially reduces ambiguous matches in repetitive satellite textures.

2.3 Per-frame geometric refinement

Given the matched drone-image and satellite-image pixels from Sec. 2.2, the next step is to refine the drone pose associated with the current GNSS measurement. Each match provides a geometric constraint: a point observed in the stabilized drone image should project to its corresponding location in the satellite tile under the corrected pose. We therefore estimate a residual planar pose correction $\boldsymbol{\theta} = (\Delta\psi, \Delta E, \Delta N)$, where $\Delta\psi$ is the yaw correction and $(\Delta E, \Delta N)$ are horizontal corrections in the local East–North tangent plane.

The residual correction is estimated through a local 2D alignment model. For a candidate $\boldsymbol{\theta}$, the model maps each stabilized drone-image pixel into the satellite tile using the initial yaw, the drone-to-satellite scale ratio, and the East–North translation expressed in satellite pixels. The refined pose is obtained by minimizing the residuals between the projected drone-image pixels and their matched satellite-image locations.

2D alignment model. Let $\mathbf{p}_i = [u_i, v_i]^\top$ denote a pixel in the rectified aerial image and let $\mathbf{p}'_i = [x_i, y_i]^\top$ be its matched pixel in the satellite tile. We model the mapping from aerial pixels to satellite pixels as an anisotropic similarity transform about a pivot point plus a translation induced by the horizontal pose correction:

$$\hat{\mathbf{p}}'_i = \mathbf{c}_s + \mathbf{t}_{\text{px}} + s\, R(\psi_0 + \Delta\psi)\, A_{\text{ratio}}(\mathbf{p}_i - \mathbf{c}_d), \tag{2}$$

where ψ_0 is the initial heading estimate associated with the GNSS seed, $\mathbf{c}_d = [c_x, c_y]^\top$ is the principal point of the rectified aerial view (from K_{flat} in Sec. 2.1) and $\mathbf{c}_s = [W_s/2, H_s/2]^\top$ is the center of the $W_s \times H_s$ satellite tile. The 2×2 rotation matrix is

$$R(\psi) = \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix}. \tag{3}$$

The diagonal matrix A_{ratio} converts aerial pixel displacements into satellite pixel displacements using a ground-sampling-distance (GSD) ratio:

$$A_{\text{ratio}} = \operatorname{diag}\left(\frac{g_d^x}{g_s}, \frac{g_d^y}{g_s}\right), \qquad g_d^x = \frac{h}{f_x}, \;\; g_d^y = \frac{h}{f_y}, \tag{4}$$

where h is the platform height above the imaged ground (Sec. 2.3), (f_x, f_y) come from K_{flat} and g_s is the satellite GSD in meters per pixel. The scalar s is fixed to 1 in all experiments. The metric translation correction $(\Delta E, \Delta N)$ is converted to a pixel shift in the satellite image via

$$\mathbf{t}_{\text{px}} = \begin{bmatrix} \Delta E/g_s \\ -\Delta N/g_s \end{bmatrix}, \tag{5}$$

where the minus sign accounts for the image y axis pointing downward. Using $\mathbf{c}_d$ as the pivot is important in practice because rectification can move the principal point away from the geometric image center.

Altitude and GSD ratio. The drone-image GSD in Eq. (4) depends on the camera height above the observed ground, rather than on the GNSS altitude alone. We estimate this height as $h = h_{\text{plat}} - h_{\text{ref}}$, where h_{plat} is the platform altitude and h_{ref} is a local ground-height reference.

Robust least-squares optimization. We solve for $\boldsymbol{\theta} = (\Delta\psi, \Delta E, \Delta N)$ by minimizing robust reprojection residuals over the matched pixel pairs:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \rho\Big(\|\hat{\mathbf{p}}'_i(\boldsymbol{\theta}) - \mathbf{p}'_i\|_2^2\Big), \tag{6}$$

where $\hat{\mathbf{p}}'_i(\boldsymbol{\theta})$ is given by (2) and $\rho(\cdot)$ is a robust loss, Huber in our experiments, used to reduce the influence of mismatches and locally inconsistent correspondences. The optimization is initialized at $\Delta\psi = \Delta E = \Delta N = 0$, corresponding to the uncorrected GNSS measurement, and solved with Levenberg–Marquardt. The resulting $\boldsymbol{\theta}$ is the residual pose correction that best aligns the stabilized drone image with the satellite tile under the matched correspondences. Table 1 summarizes the steps of the proposed solution.

Table 1. Proposed satellite-aided GNSS correction pipeline.

Step	Operation
Step 1	**Input acquisition.** For each timestamp t_k, read the UAV image, GNSS measurement $(\mathrm{lat}, \mathrm{lon}, h, t_k)$, initial yaw ψ_0, and IMU roll/pitch estimates.
Step 2	**Stabilized nadir preprocessing.** Apply optional lens rectification and compensate roll/pitch to obtain a stabilized nadir image with intrinsics K_{flat}; see Sec. 2.1.
Step 3	**Satellite retrieval.** Use the GNSS seed as a coarse spatial prior to retrieve a north-up satellite tile centered at $(\mathrm{lat}, \mathrm{lon})$. Compute its GSD using Eq. (1).
Step 4	**Pose-conditioned dense matching.** Extract dense descriptors from the stabilized UAV image and satellite tile. For each sampled UAV pixel, predict its satellite location using Eq. (2) and restrict matching to a local window.
Step 5	**Correspondence filtering.** Retain spatially regularized reciprocal matches and reject low-confidence or inconsistent correspondences; see Sec. 2.2.
Step 6	**Metric SE(2) refinement.** Estimate $\boldsymbol{\theta} = (\Delta\psi, \Delta E, \Delta N)$ by minimizing the robust objective in Eq. (6), using the image-to-map model in Eq. (2), the GSD ratio in Eq. (4), and the ENU-to-pixel translation in Eq. (5).
Step 7	**Corrected GNSS output.** Convert $(\Delta E, \Delta N, \Delta\psi)$ back to corrected geodetic position and yaw, producing a timestamp-aligned corrected GNSS stream for standalone evaluation or downstream VIO/SLAM fusion.

3 Numerical Results

Fig. 2. Sample nadir frames from the three datasets used in our evaluation. Panel (a) shows a MUN-FRL monocular fisheye frame. Panel (b) shows a UAV-AVL / AnyVisLoc benchmark frame [25] captured from a mixed fleet of DJI platforms. Panel (c) shows a frame from our DJI Mavic 3 Enterprise collection over a site in Western Greece.

3.1 Datasets and sensors

Datasets. We evaluate the proposed pipeline on three aerial datasets. First, we use four MUN-FRL sequences [20], denoted A–D, captured by the NRC Bell 412 ASRA platform over kilometer-scale trajectories in urban, highway, prairie, and waterfront environments. MUN-FRL provides synchronized nadir fisheye images, IMU measurements, GNSS/RTK streams, camera calibration, and high-precision reference trajectories. Second, we use UAV-AVL / AnyVisLoc [25] to evaluate feature-front-end generalization. This dataset contains UAV images captured across different scenes, altitudes, and viewpoints, together with georeferenced 2.5D reference maps derived from aerial photogrammetry.

Finally, we introduce a new RTK-geotagged nadir UAV dataset captured at a site in Western Greece.[2] The site includes buildings, internal roads, vegetation, rural land and nearby residential structures. Data were collected with a DJI Mavic 3 Enterprise equipped with a 4/3-inch CMOS sensor, producing 20 MP images at 5280×3956 pixels and 4K video at 30 fps. RTK ground stations provide centimeter-level geotagging for images and video telemetry. The dataset contains nadir still frames and video sequences for photogrammetry, orthomosaic generation, segmentation, tracking and cross-view localization. The released sample includes three flight segments selected from approximately 45 min of acquisitions: a linear waypoint mission with 108 nadir frames, an area-coverage mission with 135 nadir frames, and a manual session with three nadir stills and three 4K videos totaling 7 min 32 s. Video telemetry includes per-frame latitude, longitude, altitude and gimbal attitude.

3.2 Implementation Details

MASt3R [11] is used as the dense feature extractor for drone-to-satellite matching. It provides dense descriptors and confidence estimates for the local correspondence stage of the proposed correction pipeline. Sec. 3.6 evaluates the effect of replacing MASt3R with sparse alternatives.

Fusion with a visual–inertial SLAM backend. VINS-Fusion [14] is used as a representative VIO+GNSS backend to assess how the corrected GNSS stream affects downstream fusion. All configurations use identical backend parameters, and only the global position input is changed. *VIO only* runs without global updates. *Raw GNSS* reports the original GNSS measurements as a standalone absolute-position baseline. *VIO + native GNSS* fuses the recorded GNSS topic `/fix`. *VIO + corrected GNSS (ours)* fuses the corrected stream produced by the proposed cross-view module. *VIO + RTK reference input* fuses the RTK reference trajectory as a GNSS-like input when available.

Metrics. We measure absolute position accuracy in a local ENU frame anchored at the reference start, using the RTK trajectory as ground truth. For matched timestamps, we report mean position error and position RMSE between the estimated and reference trajectories. For standalone cross-view correction, we additionally report translation error, yaw error and the percentage gain relative to the initial GNSS measurement.

3.3 Cross-view correction performance

Table 2 reports the standalone correction accuracy before downstream fusion, with gain measured as the percentage reduction from the initial GNSS error to the refined mean error. Across the evaluated MUN-FRL sequences, the proposed module reduces the mean translation error by 52.3%–86.7% and the mean yaw error by 61.1%–85.3%. The largest translation gain is obtained on Sequence B,

[2] Dataset available at: `https://drive.google.com/drive/folders/1cvNs2o7mS9U1OXBJIM5dEU7yzs3ZTpK_?usp=sharing`

where the error decreases from 4.06 m to 0.54 m, corresponding to an 86.7% reduction. Sequence D follows with a 76.1% translation gain, while Sequences C and A achieve 72.5% and 52.3%, respectively. For yaw, the gains remain consistently high across all MUN-FRL sequences, reaching 85.3% on Sequence B and 82.6% on Sequence D. On AVL-UAV, the translation error decreases from 4.40 m to 3.06 m, corresponding to a 30.5% gain, while the yaw error decreases from 4.56° to 0.42°, yielding a 90.8% gain. On our DJI Mavic 3 Enterprise dataset, the translation error decreases from 4.39 m to 1.94 m, corresponding to a 55.8% gain, while the yaw error decreases from 4.56° to 0.48°, yielding an 89.5% gain.

Table 2. Standalone cross-view correction quality on the evaluated sequences. Init. mean is the initial GNSS seed error; Ref. med. and Ref. mean are the refined position/yaw errors after Sec. 2.3. Gain is the percentage reduction from Init. mean to Ref. mean. Arrows denote the direction of improvement (↓ lower is better, ↑ higher is better).

	Translation (m) ↓				Yaw (deg) ↓			
Dataset	Init. mean	Ref. med.	Ref. mean	Gain ↑	Init. mean	Ref. med.	Ref. mean	Gain ↑
MUN-FRL								
A	3.90	1.50	1.86	52.3%	3.16	0.94	1.23	61.1%
B	4.06	0.52	0.54	86.7%	2.72	0.17	0.40	85.3%
C	4.62	0.87	1.27	72.5%	3.30	0.74	0.82	75.2%
D	3.56	0.74	0.85	76.1%	2.13	0.28	0.37	82.6%
AVL-UAV	4.40	2.51	3.06	30.5%	4.56	0.35	0.42	90.8%
Our Dataset	4.39	1.55	1.94	55.8%	4.56	0.44	0.48	89.5%

3.4 Comparison with prior cross-view localization results

Tab. 3 compares the proposed single-frame correction with an adapted implementation of Patel et al. [12] on the MUN-FRL sequences. This method is the closest operational baseline among prior UAV satellite-localization approaches, since it registers real UAV images against georeferenced Google Earth views. For a fair comparison, both methods start from the same GNSS initialization. In the adapted baseline, the GNSS measurement replaces the visual-odometry prior used in the original formulation to select nearby rendered reference views. Both methods improve the initial GNSS estimate, but the proposed approach achieves lower translation and yaw errors on all evaluated sequences. The adapted baseline reaches 0.67–1.90 m translation error and 0.48°–1.34° yaw error, while our method reaches 0.54–1.86 m and 0.37°–1.23°, respectively. Beyond accuracy, the proposed method also avoids route-specific pre-rendering. It operates directly on the retrieved satellite tile, whereas the adapted baseline requires a database of rendered Google Earth views around the expected trajectory.

3.5 Visual–inertial fusion with cross-view GNSS correction

We further evaluate whether the corrected GNSS stream improves a standard visual–inertial backend. VINS-Fusion is run with identical parameters in all con-

Table 3. Pose correction accuracy on the evaluated MUN-FRL sequences. Init. mean denotes the initial GNSS seed error; Method [12] and Our method. Gain is the percentage reduction of Our method relative to Method [12].

	Translation (m) ↓				Yaw (deg) ↓			
Sequence	Init. mean	Method [12]	Our method	Gain ↑	Init. mean	Method [12]	Our method	Gain ↑
A	3.90	1.90	1.86	2.1%	3.16	1.34	1.23	8.2%
B	4.06	0.67	0.54	19.4%	2.72	0.57	0.40	29.8%
C	4.62	1.57	1.27	19.1%	3.30	0.94	0.82	12.8%
D	3.56	0.90	0.85	5.6%	2.13	0.48	0.37	22.9%

figurations; only the global position input is changed. We compare VIO-only, native GNSS fusion, fusion with the proposed corrected GNSS stream, and, when available, fusion with an RTK-derived GNSS-like reference. *Each end-to-end localization update requires approximately 1.4 TFLOP of neural computation and runs in about 0.4 s. This rate is sufficient for real-time injection into a SLAM or visual–inertial backend, while the fusion results below show that the corrected stream approaches the accuracy of the RTK-derived reference input.*

Quantitative results are reported in Table. 4. On Run 1, a 1006.91 m trajectory, VIO-only accumulates large drift, while native GNSS fusion remains affected by biased global measurements. Using the corrected GNSS stream reduces the mean error from 36.74 m to 14.58 m and the maximum error from 206.94 m to 26.22 m. This corresponds to a substantially more stable fused trajectory and brings the error range close to the RTK-derived reference configuration. The improvement is stronger on Run 2, where GNSS degradation is more severe. Native GNSS fusion still produces large outliers, with a maximum error of 139.69 m. Replacing it with the corrected GNSS stream reduces the mean error to 2.49 m and the RMSE to 2.88 m, while keeping the 95th-percentile error below 5 m. The fraction of samples within 5 m increases from 54.0% to 95.2%. These results show that the proposed cross-view correction provides a more reliable global input to the fusion backend, especially when the native GNSS stream is degraded.

Table 4. VINS-Fusion results on the evaluated MUN-FRL runs. Sequence A corresponds to Run 1, a 1006.91 m trajectory, while Sequence D is a shorter 392.55 m trajectory with stronger GNSS degradation. Gain columns report the percentage reduction of VIO + corrected GNSS relative to VIO + native GNSS.

Configuration	Mean (m) ↓	RMSE (m) ↓	P95 (m) ↓	Max (m) ↓	< 5 m (%) ↑	Mean gain ↑	RMSE gain ↑
Sequence A / Run 1 (1006.91 m)							
Raw GNSS fed to CrossView	10.71	12.15	18.88	36.39	15.0	–	–
VIO only	101.07	118.74	247.43	285.22	0.0	–	–
VIO + native GNSS (/fix)	36.74	43.12	82.91	206.94	2.9	–	–
VIO + corrected GNSS (ours)	**14.58**	**15.37**	**23.71**	**26.22**	0.0	**60.3%**	**64.4%**
VIO + FRL oracle input (GT-fed)	13.67	14.35	22.43	24.91	1.8	–	–
Sequence D / Run 2 (392.55 m)							
Raw GNSS fed to CrossView	9.89	17.04	40.59	42.40	66.7	–	–
VIO only	26.13	31.14	56.61	67.51	6.4	–	–
VIO + native GNSS (/fix)	8.36	14.27	23.38	139.69	54.0	–	–
VIO + corrected GNSS (ours)	**2.49**	**2.88**	**4.99**	**9.53**	**95.2**	**70.2%**	**79.8%**
VIO + FRL oracle input (GT-fed)	2.82	3.24	5.59	8.19	89.5	–	–

3.6 Ablation studies

We now isolate the contribution of key design choices in the pipeline through controlled ablations covering attitude compensation, matcher choice, and spatial regularization of the matching search.

Effect of roll/pitch compensation. We isolate the impact of the roll/pitch compensation step described in Sec. 2.1 on Sequence A of the MUN-FRL dataset. Table 5 shows that attitude compensation substantially improves the corrected trajectory error. Without compensation, the corrected RMSE is 5.124 m. With the proposed roll/pitch compensation, the RMSE decreases to 1.691 m, corresponding to a 67.0% reduction. This confirms that even moderate attitude deviations should be compensated before cross-view matching.

Table 5. Effect of roll/pitch compensation on the MUN-FRL dataset.

Dataset	Comp RMSE (m) ↓	No-comp RMSE (m) ↓	RMSE gain ↑
MUN-FRL	1.691	5.124	67.0%

Fig. 3. Qualitative effect of pose-conditioned spatial regularization on a DJI Mavic 3 Enterprise frame. Unconstrained matching sends several road pixels to a visually similar but incorrect satellite region. Pose-conditioned matching keeps correspondences local to their predicted satellite locations, preserving the road structure for pose refinement.

Effect of the matcher choice. We report the feature front-end ablation in Tab. 6. We replace MASt3R with an adapted CNN-based cross-view method [16] and SIFT, while keeping retrieval, matching regularization, and SE(2) refinement unchanged. MASt3R achieves the best correction, reducing translation error from 4.40 m to 3.06 m and yaw error from 4.56° to 0.42°. The CNN-based baseline provides a smaller gain, reaching 4.16 m translation and 3.00° yaw, while SIFT remains close to the initial GNSS pose. The improvement is most pronounced in yaw, indicating that dense and spatially distributed correspondences are critical for constraining the rotational component of the SE(2) refinement.

Table 6. Effect of feature extractor choice on the UAV-AVL / AnyVisLoc dataset. T and Y denote translation error in meters and yaw error in degrees. Init. is the initial GNSS error, and Ref. is the corrected pose after refinement with each matcher.

Method	Init. T ↓	Ref. T ↓	T gain ↑	Init. Y ↓	Ref. Y ↓	Y gain ↑
MASt3R	4.40	**3.06**	**30.5%**	4.56	**0.42**	**90.8%**
method [16]	4.40	4.16	5.5%	4.56	3.00	34.2%
SIFT	4.40	4.43	-0.7%	4.56	4.50	1.3%

Spatial regularization through pose-conditioned search. We evaluate the effect of the proposed spatial regularization on the DJI Mavic 3 Enterprise dataset. The *Global search* variant follows the common appearance-only matching setting used in prior cross-view methods, where each drone-image pixel can be matched anywhere in the satellite tile based only on visual descriptor similarity. In contrast, the *Spatially regularized* variant restricts matching to the local metric window induced by the current pose prior. All other components are kept fixed. Quantitative results are reported in Tab. 7. Appearance-only global search is sensitive to repeated satellite structures: although the median error is reduced, the mean error increases from 4.397 m to 10.512 m and the RMSE from 4.397 m to 16.360 m, indicating severe outlier failures. This effect is also shown qualitatively in Fig. 3, where visually similar but geographically incorrect regions attract erroneous matches. Spatial regularization removes these failures by enforcing local geometric consistency during matching.

Table 7. Effect of pose-conditioned spatial regularization on our dataset. Gain vs Init reports the percentage reduction from the initial GNSS seed to spatially regularized refinement. Gain vs Global reports the percentage reduction from global-search refinement to spatially regularized refinement.

Metric	Init ↓	Refined (Global) ↓	Refined (Spatial reg.) ↓	Gain vs Init ↑	Gain vs Global ↑
Mean	4.397	10.512	**1.940**	**55.9%**	**81.5%**
Median	4.397	2.482	**1.555**	**64.6%**	**37.3%**
RMSE	4.397	16.360	**2.334**	**46.9%**	**85.7%**

4 Conclusion

This paper introduced a satellite-aided GNSS correction framework that treats drone-to-satellite alignment as measurement-level refinement rather than global re-localization. Starting from an available GNSS measurement, the method retrieves a local satellite tile, stabilizes the nadir drone image, constrains dense matching through pose-conditioned spatial regularization, and estimates the residual horizontal and yaw correction through geometric pose refinement. The resulting corrected GNSS stream is timestamp-aligned and can be used directly or passed to downstream fusion backends. Experiments across MUN-FRL, UAV-AVL and our DJI Mavic 3 Enterprise dataset show consistent reductions in position and yaw error. The ablations confirm that spatial regularization is critical in repetitive satellite imagery, where unconstrained matching can produce geographically inconsistent correspondences, and that roll/pitch compensation and dense features further improve robustness. When injected into VINS-Fusion, the corrected GNSS stream substantially improves the fused trajectory and reaches an accuracy range close to the RTK-derived reference configuration.

Acknowledgments. This work has received funding from the European Union's Horizon Europe research and innovation programme under the GuardAI project, "Robust and Secure Edge AI Systems for Safety-Critical Applications" (GA No. 101168067).

References

1. Anagnostopoulos, C., Gkillas, A., Piperigkos, N., Lalos, A.S.: Personalized federated learning for cross-view geo-localization. In: 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6 (2024). https://doi.org/10.1109/MMSP61759.2024.10743630
2. Bianchi, M., Barfoot, T.D.: UAV localization using autoencoded satellite images. IEEE Robotics and Automation Letters **6**(2), 1761–1768 (2021). https://doi.org/10.1109/LRA.2021.3060397
3. Cao, S., Lu, X., Shen, S.: GVINS: Tightly coupled GNSS-visual-inertial fusion for smooth and consistent state estimation. IEEE Transactions on Robotics **38**(4), 2004–2021 (2022). https://doi.org/10.1109/TRO.2021.3133730
4. Cisneros, I., Yin, P., Zhang, J., Choset, H., Scherer, S.: ALTO: A large-scale dataset for UAV visual place recognition and localization (2022)
5. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16847–16856 (October 2023)
6. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Uncertainty-aware vision-based metric cross-view geolocalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21621–21631 (June 2023)
7. Gkillas, A., Lalos, A.S., Markakis, E.K., Politis, I.: A federated deep unrolling method for lidar super-resolution: Benefits in SLAM. IEEE Transactions on Intelligent Vehicles **9**(1), 199–215 (2024). https://doi.org/10.1109/TIV.2023.3331533
8. Goforth, H., Lucey, S.: GPS-denied UAV localization using pre-existing satellite imagery. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 2974–2980. IEEE (2019). https://doi.org/10.1109/ICRA.2019.8793558
9. Khalife, J., Maaref, M., Kassas, Z.M.: Opportunistic autonomous integrity monitoring for enhanced UAV safety. IEEE Aerospace and Electronic Systems Magazine **38**(5), 34–44 (2023). https://doi.org/10.1109/MAES.2022.3178664
10. Lee, W., Geneva, P., Yang, Y., Huang, G.: Tightly-coupled GNSS-aided visual-inertial localization. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 9484–9491. IEEE (2022). https://doi.org/10.1109/ICRA46639.2022.9811362
11. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with MASt3R. In: Computer Vision – ECCV 2024. pp. 71–91 (2024). https://doi.org/10.1007/978-3-031-73220-1_5

12. Patel, B., Barfoot, T.D., Schoellig, A.P.: Visual localization with google earth images for robust global pose estimation of UAVs. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 6491–6497. IEEE (2020). https://doi.org/10.1109/ICRA40945.2020.9196606
13. Phillips, C., Broumandan, A., O'Keefe, K.: A deep learning approach for the classification of multipath ranging errors in challenging urban environments. In: Proceedings of the 37th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2024). pp. 2555–2566 (2024). https://doi.org/10.33012/2024.19743
14. Qin, T., Cao, S., Pan, J., Shen, S.: A general optimization-based framework for global pose estimation with multiple sensors (2019)
15. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4938–4947 (June 2020)
16. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17010–17020 (2022)
17. Štroner, M., Urban, R., Seidl, J., Reindl, T., Brouček, J.: Photogrammetry using UAV-mounted GNSS RTK: Georeferencing strategies without GCPs. Remote Sensing **13**(7), 1336 (2021). https://doi.org/10.3390/rs13071336
18. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (June 2021)
19. Tavasci, L., Nex, F., Gandolfi, S.: Reliability of real-time kinematic (RTK) positioning for low-cost drones' navigation across global navigation satellite system (GNSS) critical environments. Sensors **24**(18), 6096 (2024). https://doi.org/10.3390/s24186096
20. Thalagala, R.G., Gunawardena, S.M., De Silva, O., Jayasiri, A., Gubbels, A., Mann, G.K.I., Gosine, R.G.: MUN-FRL: A visual-inertial-lidar dataset for aerial autonomous navigation and mapping. The International Journal of Robotics Research **43**(12), 1853–1866 (2024). https://doi.org/10.1177/02783649241238358
21. Tomaštík, J., Mokroš, M., Surový, P., Grznárová, A., Merganič, J.: UAV RTK/PPK method—an optimal solution for mapping inaccessible forested areas? Remote Sensing **11**(6), 721 (2019). https://doi.org/10.3390/rs11060721
22. Wen, W., Hsu, L.T.: Towards robust GNSS positioning and real-time kinematic using factor graph optimization. In: 2021 International Conference on Robotics and Automation (ICRA). pp. 5884–5890. IEEE (2021). https://doi.org/10.1109/ICRA48506.2021.9562037
23. Wen, W., Zhang, G., Hsu, L.T.: GNSS outlier mitigation via graduated nonconvexity factor graph optimization. IEEE Transactions on Vehicular Technology **71**(1), 297–310 (2022). https://doi.org/10.1109/TVT.2021.3130909
24. Xu, W., Yao, Y., Cao, J., Wei, Z., Liu, C., Wang, J., Peng, M.: UAV-VisLoc: A large-scale dataset for UAV visual localization (2024)
25. Ye, Y., Teng, X., Chen, S., Li, Z., Liu, L., Yu, Q., Tan, T.: Exploring the best way for UAV visual localization under low-altitude multi-view observation condition: A benchmark (2025). https://doi.org/10.48550/arXiv.2503.10692
26. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1162–1171 (June 2022)