Towards Robust Cross-Seasonal Image-to-Map Matching against Publicly Available Prior Maps

***Abstract*—Publicly available remote-sensing maps provide a global prior for outdoor robot localization, but exploiting them requires robustness to cross-seasonal appearance changes and strong query-to-reference scale asymmetry. We propose Sat-RoMa, which replaces RoMa's ConvNet-based fine-level refinement with a geometry-aware homography refinement operating directly on the multimodal coarse matching distribution. Instead of committing early to a single correspondence, we extract per-patch Gaussian-mixture hypotheses and resolve them via global RANSAC consensus. On cross-seasonal orthophoto pairs at $4{:}1$ asymmetry, this refinement halves corner (16.9→7.2 m), scale (24.0→10.5%), and rotation errors (14.6° →9.0°), and improves Success@5/10/20 m from 48/67/82% to 67/81/91%. Both variants outperform prior dense and classical matchers. The results show that preserving and resolving multimodality provides a simple, training-free alternative to learned fine-level refinement.**

I. INTRODUCTION

A wide range of publicly available remote-sensing products, including national orthophoto programmes, open GIS basemaps, and commercial satellite archives, can serve as common prior maps for outdoor robotics. Together they cover most populated regions, are regularly updated, and often span more than a decade of repeat acquisitions over the same ground. For an outdoor robot, this provides a pre-built absolute reference: at essentially every location, a registered overhead view already exists. Using this prior for absolute localization reduces largely to *image-to-map matching*: establishing dense correspondences between the robot's live view and a georeferenced reference tile, and recovering the metric transformation.

State-of-the-art visual-inertial odometry and visual SLAM [1]–[3] deliver excellent local estimates but accumulate drift over long trajectories [4]. Even with strong systems reporting sub-percent drift over kilometre-scale runs, a UAV flying hundreds of kilometres would accumulate translation errors far beyond operational tolerance. Periodic matches against a georeferenced prior cancel this drift, but only if each match yields a precise metric transformation. Cross-view geo-localization and image retrieval [5], [6] address appearance change but return similarity scores or tile classifications rather than relative pose; they are complementary to, not substitutes for, dense image-to-map matching.

Current dense matchers do not simultaneously satisfy two key requirements in this setting. **(P1) Temporal robustness.** Query and prior may be separated by months, years, or a full seasonal cycle. As shown in Fig. 1, scenes can gain or lose buildings and vegetation, undergo crop cycles, and include transient objects such as parked cars—conditions not encountered during training on near-simultaneous Internet imagery. **(P2) Cross-scale robustness.** The image-to-map asymmetry (a $4:1$ linear ratio per axis, equivalently $16:1$ in ground area between reference tile and query) reflects the deployment pattern: a wide-area prior is preloaded once, while each observation covers only a fraction of it. Generic dense matchers such as LoFTR [7], LightGlue [8], and RoMa [9], trained on near-isoscale imagery, produce poorly calibrated correspondences in this regime (Sec. III).

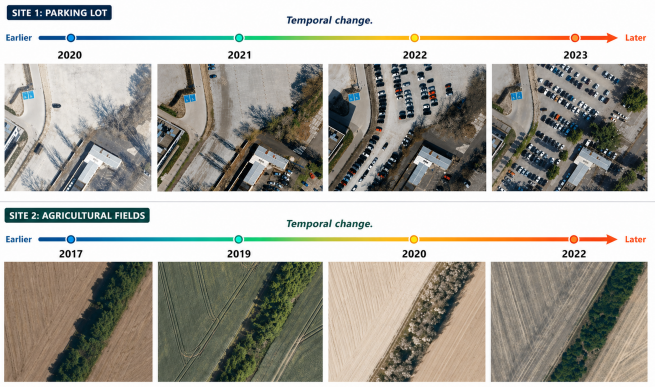

Fig. 1. The challenges of matching overhead prior maps across time. **Top row:** an urban parking-lot environment with structural and moving-object changes over three years (2020–2023). **Bottom row:** an agricultural scene with drastic appearance changes from seasonal vegetation cycles and crop harvesting (2017–2022).

The key observation is that, under cross-seasonal appearance change, the correct correspondence often appears as a secondary mode in the coarse matching distribution, while the most likely mode may be geometrically inconsistent. We adapt the RoMa dense-matching architecture to this regime and *replace its ConvNet-based fine-level warp refinement with a geometry-aware homography refinement operating directly on the multimodal coarse matching distribution*. Rather than committing early to a single correspondence per query, we treat the coarse posterior as a set of candidate matches: an adaptive-window Gaussian-mixture fit extracts multiple modes per query, and a homography RANSAC over their union selects a geometrically consistent subset. The DinoV2-pretrained encoder–decoder is reused unchanged and fine-tuned on cross-seasonal orthophoto pairs at $4:1$ image-to-map asymmetry. The system thus combines an explicit geometric prior (the georeferenced orthophoto map) with an implicit semantic prior from the pretrained encoder, expressed through the multimodal structure of the coarse matching distribution; the training loss remains purely geometric. We call this system *Sat-RoMa*.

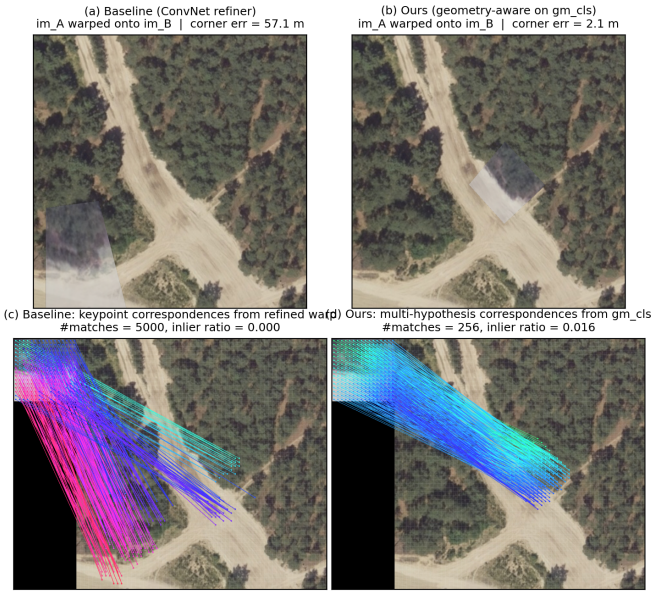

Fig. 2. Qualitative effect of replacing RoMa's ConvNet-based fine-level warp refinement with the proposed geometry-aware homography refinement on the multimodal coarse matching distribution, on a single representative cross-seasonal pair (scale error $75.3 \rightarrow 4.3\%$, rotation $46.7^\circ \rightarrow 4.3^\circ$). **Top row, (a)/(b):** im_A warped onto im_B under the homography recovered by each pipeline. The ConvNet refiner commits each query to a single warp peak; the wrong peak is elected on a coordinated subset of patches, the homography RANSAC inherits the corruption, and the warped query is mis-placed and rotated against the correct clearing (57.1 m corner error). The proposed refinement keeps multiple per-query candidates from the coarse matching distribution and a global RANSAC elects the geometrically consistent mode, recovering a tight alignment (2.1 m corner error). **Bottom row, (c)/(d):** the underlying correspondence sets fed to RANSAC, with im_A inset (top-left) and arrows running to im_B. (c) Baseline samples one correspondence per query from the dense fine-level warp (5000 matches), and the wrong-mode subset visibly fans toward a different ground location. (d) Ours emits a Gaussian-mixture hypothesis set per query patch from the coarse matching distribution (~ 256 candidates total); the RANSAC consensus implicitly chooses one mode per query by global geometric agreement, yielding a coherent bundle.

Contributions. (1) **Delayed geometric commitment from multimodal coarse matches.** We treat RoMa's coarse matching distribution as a multi-hypothesis correspondence source rather than collapsing it to a single match per query. A homography RANSAC over the resulting hypothesis set performs global mode selection, retaining geometrically consistent secondary modes that standard refinement can discard. (2) **Adaptive-window mode extraction.** The Gaussian fit around each posterior mode uses a window that grows iteratively to $n_\sigma \approx 3$ of the fitted covariance, yielding a more faithful mixture decomposition than fixed-window fitters under the diffuse-vs-sharp peaks that overhead pairs produce. (3) On 500 cross-seasonal orthophoto pairs from four held-in locations, the proposed refinement reduces the same network's corner error from 16.9 to 7.2 m (2.3×), scale error from 24.0 to 10.5% (2.3×), and rotation error from 14.6° to 9.0° (1.6×); Success at 5 / 10 / 20 m rises from 48.0 / 67.0 / 82.3% to 66.9 / 81.1 / 90.5%.

II. SYSTEM OVERVIEW

Sat-RoMa reuses the encoder–decoder architecture of RoMa [9] (a DinoV2-pretrained backbone followed by a coarse-to-fine decoder that emits a dense warp field together with a per-pixel certainty map and intermediate per-patch coarse matching distributions). It deviates from the original system in three deliberate choices, two at training time and one at inference time, each tied to one of the targeted properties.

A. Geometry-aware homography refinement (P1 and P2)

RoMa's coarse decoder produces, for each query patch on a coarse grid, a categorical distribution over a discretised reference grid plus a no-match bin—a tensor in $\mathbb{R}^{(K^2+1)\times h\times w}$ which after softmax yields the per-query posterior $P(B \mid A_{x,y})$. We refer to this as the *coarse matching distribution*. In standard inference, RoMa feeds this output to a ConvNet-based fine-level warp refiner and samples a single correspondence per query, followed by homography RANSAC. We instead replace this stage with a *geometry-aware homography refinement* that operates directly on the coarse distribution, bypassing the fine-level ConvNet. The method consists of three stages.

Stage 1: adaptive Gaussian-mixture fit per posterior. For each query patch, we detect modes of $P(B \mid A_{x,y})$ via non-maximum suppression on a smoothed heatmap with a small dilation kernel and a low probability floor. Around each peak, we fit a 2D Gaussian $\mathcal{N}(\mu_k, \Sigma_k)$ using image moments in an *adaptively sized* window: the window expands iteratively to cover approximately $n_\sigma \approx 3$ of the fitted Gaussian. This yields tight windows for sharp peaks and wider ones for diffuse blobs, avoiding the systematic misestimation of covariance associated with fixed-window fitting.

Stage 2: multi-hypothesis correspondence set. Stage 1 produces a sparse correspondence set $\mathcal{C}$ in which each query patch contributes multiple candidate matches: $\mathcal{C} = \{(p_A^{(i)}, \mu_B^{(j)}) \mid i \in \text{patches}, j \in \text{modes}(P(B \mid A_i))\}$. For a typical 14×14 grid, a single-match baseline yields 196 pairs, whereas the multi-hypothesis set contains 400–800.

Stage 3: RANSAC consensus on the multi-hypothesis set. A standard homography RANSAC is applied to $\mathcal{C}$. Since each query patch $p_A^{(i)}$ is associated with multiple candidates $\mu_B^{(j)}$, RANSAC implicitly performs per-patch mode selection: correspondences consistent with the global geometry are retained, others rejected as outliers. The resulting homography $\hat{H}$ is reported directly, without further optimisation. The standard RoMa refinement is the degenerate case where Stage 1 produces a single Dirac per query; we report this as the *baseline* variant.

B. Asymmetric 4:1 image-to-map formulation (P2)

Dense matchers in the literature are typically trained on pairs at approximately the same ground-sample distance. We instead sample training pairs at the canonical $4:1$ linear

($16:1$ area) image-to-map asymmetry at matched ground-sample distance, mirroring the deployment geometry of a small live observation localized within a preloaded wide-area prior. This distribution is intended to encourage the decoder toward scale-aware sub-patch correspondence rather than image-wide photometric alignment.

C. Online cross-seasonal pair synthesis

Each sample draws, in order: a location from the six-location mixture; two distinct seasons of that location; a query ground size uniformly from $[13.3,\, 26.6]\,\mathrm{m}$ per edge (corresponding to 5.9–$11.9\,\mathrm{cm/px}$ at $224\,\mathrm{px}$, in the low-altitude UAV regime); and a random similarity homography with rotation in $[-55^\circ, +55^\circ]$, isotropic scale in $[0.65,\, 1.35]$, and translation up to 30% of the tile edge. The reference is cropped at the sampled ground size from one season and rendered at $896{\times}896\,\mathrm{px}$; the query is the central $224{\times}224\,\mathrm{px}$ crop of the warped tile from the other season at matched ground-sample distance.

Appearance variation between query and reference is thus drawn from *real* multi-year orthophoto acquisitions, while the geometric warp is synthetic and exactly known, providing the supervision signal for training. From 4–8 tiles per location, this process yields an effectively unbounded, geometry-labelled dataset; deterministic per-sample seeding (a stable hash of location and global sample id) ensures full reproducibility. Fig. 2 shows a representative output of the trained system.

III. EXPERIMENTAL EVALUATION

A. Experiment design

Training data is drawn from orthophoto stacks at six $2\,\mathrm{km}\times 2\,\mathrm{km}$ locations across a Central European temperate-climate region, spanning distinct land-cover types (two dense urban centres, one peri-urban area, two forested terrains with exposed structural features), each a stack of 4–8 tiles spanning 2014–2024. Pairs are synthesised as in Sec. II. To mirror the deployment task (matching a current observation against a historical prior over the *same* place), evaluation splits by seasonal index rather than location: the query is fixed to the newest season (2024) and candidate references are earlier acquisitions disjoint from training queries. The evaluation therefore shares geographic extent with training; held-out-location generalisation is flagged in Sec. IV.

We compare two post-coarse refinement modes on the same Sat-RoMa checkpoint — a RoMa-style *baseline* (ConvNet fine-level refinement, then sparse-correspondence RANSAC) and the proposed *multi-hypothesis* variant (adaptive-window Gaussian-mixture fit on the coarse matching distribution, then RANSAC over the hypothesis-set union) — and against six external matchers run at their native input resolutions: LoFTR [7], LightGlue [8], ELoFTR [11], the upstream outdoor RoMa [9], RoMa v2 [10], and a classical RootSIFT [12] + RANSAC pipeline. All external matchers are evaluated using publicly released checkpoints without fine-tuning on the orthophoto distribution; for RoMa v2 in particular only the inference code is publicly available at the time of writing, so fine-tuning on the proposed cross-seasonal pair distribution is not currently feasible without a re-implementation of its training pipeline. Each pair has a ground-truth homography $H^\star$ from rendering metadata; for every method we estimate $\hat{H}$ and report *Succ.@*τ (fraction of pairs with mean reprojection error below $\tau\,\mathrm{m}$, $\tau \in \{5, 10, 20\}$, computed over all pairs), *Corner error* (mean ground-distance between the four query-tile corners under $\hat{H}$ vs $H^\star$, in metres), *Scale error* ($|\sqrt{\sigma_1\sigma_2} - 1|$ in percent, with σ_i the singular values of $J_{\mathrm{rel}} = J_{\hat{H}}(c)\, J_{H^\star}(c)^{-1}$ at the query centre c), and *Rotation error* (absolute rotation of J_{rel} in degrees). Corner / Scale / Rotation means are over pairs with a recovered homography. Neither Sat-RoMa variant applies post-RANSAC optimisation of $\hat{H}$.

B. Results

Table I groups external matchers against the two Sat-RoMa refinement modes. The Sat-RoMa baseline (ConvNet fine-level refinement + RANSAC) already exceeds every external matcher on every metric; the strongest external method (outdoor RoMa at its native $560{\times}560$) reaches 40.8 / 52.6 / 68.0% Success and a $37.9\,\mathrm{m}$ corner error, against 48.0 / 67.0 / 82.3% and $16.9\,\mathrm{m}$ for Sat-RoMa baseline. The remaining external matchers (LoFTR, LightGlue, ELoFTR, RootSIFT) sit below 1% Success at $5\,\mathrm{m}$; their dominant failure mode is collapse onto the wrong reference cell at the asymmetric scale. Replacing the ConvNet refiner with the geometry-aware refinement on the same checkpoint delivers a further large gain.

On P1 — temporal robustness. Every pair is cross-seasonal by construction, so corner error and the Success columns already aggregate over temporal variation. The geometry-aware refinement halves the Sat-RoMa baseline's corner error ($16.9 \rightarrow 7.2\,\mathrm{m}$) and lifts Success at $5\,\mathrm{m}$ from 48.0 to 66.9%. Because both rows use the same checkpoint and coarse matching distribution, the gap isolates the effect of delaying correspondence commitment to the post-coarse geometric stage: RoMa's ConvNet fine-level warp refinement collapses its multimodality early, and any subsequent geometric verifier sees only a single committed correspondence per query. Under cross-seasonal appearance change the dominant coarse mode is sometimes shifted by photometric ambiguity onto the wrong building or field cell while the geometrically correct mode survives as a second peak; preserving that second peak as an explicit hypothesis lets RANSAC consensus elect it.

On P2 — cross-scale robustness. Scale error drops from 24.0 to 10.5% and rotation error from 14.6° to 9.0° on the same network (more-than-halving and one-third reductions, respectively); the same metrics for the strongest external method, outdoor RoMa, are 270.6% and 23.6°. The improvement is consistent with hypothesis-set RANSAC being more robust to locally-ambiguous correspondences in the asymmetric pair: when a query patch can match several reference patches at different reference-side offsets (a building

TABLE I

QUANTITATIVE COMPARISON ON CROSS-SEASONAL ORTHOPHOTO PAIRS AT $4:1$ IMAGE-TO-MAP ASYMMETRY FROM FOUR HELD-IN LOCATIONS. EXTERNAL BASELINES RUN AT THEIR NATIVE RESOLUTION; SAT-ROMA RUNS AT 224 (QUERY)/896 (REFERENCE). THE TWO SAT-ROMA ROWS SHARE THE SAME CHECKPOINT AND DIFFER ONLY IN THE POST-COARSE REFINEMENT.

Method	**Succ.@5 m** (%) ↑	**Succ.@10 m** (%) ↑	**Succ.@20 m** (%) ↑	**Corner Err.** (m) ↓	**Scale Err.** (%) ↓	**Rot. Err.** (°) ↓
External baselines, $N = 500$						
LoFTR [7] (832^2)	0.0	0.4	13.2	456.6	486.0	34.4
LightGlue [8] (1024^2)	0.4	1.4	16.6	103.2	3032.6	49.3
ELoFTR [11] (832^2)	0.0	0.0	3.3	135.0	614.3	88.4
RoMa outdoor [9] (560^2)	40.8	52.6	68.0	37.9	270.6	23.6
RoMa v2 [10] (precise_hr)	39.2	47.0	56.4	34.2	148.8	33.1
RootSIFT [12]+RANSAC (1024^2)	0.0	0.8	9.2	90.2	762.1	72.2
Sat-RoMa (ours), same checkpoint						
Baseline (ConvNet fine-level refiner)	48.0	67.0	82.3	16.9	24.0	14.6
Multi-hypothesis (Ours)	**66.9**	**81.1**	**90.5**	**7.2**	**10.5**	**9.0**

viewed against repeated rooflines), keeping the full mixture and letting global geometry select the consensus mode is more robust than committing to the single most-likely correspondence per query. The Success curves at 10 m and 20 m (81.1 and 90.5% vs 67.0 and 82.3%) show that the gain holds across the metric range, not only in the tail.

C. Scope of this evaluation

The evaluation covers cross-season orthophoto pairs at $4:1$ asymmetry from six held-in locations. It does not address altitude-diverse UAV imagery, heterogeneous prior sources (satellite vs. low-altitude aerial vs. non-nadir views), occluded or missing prior tiles, or trajectory-level use of Sat-RoMa inside a SLAM stack. The technical contribution evaluated here is the refinement.

IV. DISCUSSION AND CONCLUSION

The main finding is that, for cross-seasonal image-to-map matching, preserving multimodality is more valuable than immediately refining a single dense warp: the multimodality of the coarse matching distribution is collapsed early, and any downstream geometric verifier inherits a single committed correspondence per query. Replacing that stage with a geometry-aware homography refinement that fits a Gaussian mixture to each posterior and resolves the resulting hypothesis set with a global RANSAC suffices to halve the corner error on the same network and lift Success at 5 m by nearly twenty percentage points. **Limitations.** The two design choices (multi-hypothesis correspondence set, adaptive-window mode extraction) are evaluated only jointly; isolating each is left to a follow-up. We report only the homography returned by RANSAC; a downstream Mahalanobis-Huber refinement on the Gaussian covariances is implementable from the same Stage-1 outputs but excluded pending separate validation. External matchers are run as-released, without fine-tuning on the cross-seasonal orthophoto distribution; for RoMa v2 in particular only inference code is publicly available at the time of writing, so a fine-tuned comparison against Sat-RoMa would require re-implementing its training pipeline and is left for future work. Held-out-location generalisation, heterogeneous map sources, incomplete priors, and trajectory-level SLAM use are out of scope and are the natural extensions.

Conclusion. Publicly available prior maps are an under-exploited absolute reference for outdoor robotics; using them requires a matcher that is simultaneously temporally and cross-scale robust on overhead data. Replacing RoMa's ConvNet-based fine-level warp refinement with a geometry-aware homography refinement operating directly on the multimodal coarse matching distribution roughly halves corner, scale, and rotation errors on the same network and lifts Success at 5 / 10 / 20 m from 48 / 67 / 82% to 67 / 81 / 91%. The multimodal structure of the per-query coarse matching distribution is a meaningful resource for image-to-map matching against common prior maps — at no additional training cost.

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] F. Yao, *et al.*, "GNSS-denied geolocalization of UAVs using terrain-weighted constraint optimization," Int. J. Appl. Earth Obs. Geoinf., vol. 135, p. 104277, 2024.

[5] 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.

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

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

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

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

[10] J. Edstedt *et al.*, "RoMa v2: Harder, better, faster, denser feature matching," *arXiv preprint arXiv:2511.15706*, 2025.

[11] Y. Wang, *et al.*, "Efficient LoFTR: Semi-dense local feature matching with sparse-like speed," in *Proc. CVPR*, 2024.

[12] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. CVPR*, 2012.