GEOSYNC: UNIFIED REPRESENTATION LEARNING FOR END-TO-END CROSS-VIEW GEOLOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose GeoSync, a unified framework for cross-view image geolocalization that jointly learns satellite image retrieval and fine-grained localization from a shared cross-view representation. Existing approaches typically separate these tasks, using global image-level matching to retrieve a satellite reference and a dedicated localizer to estimate the position within it. This decoupling limits interaction between retrieval and localization while adding computation during inference and increasing system cost. GeoSync learns a shared representation that captures global scene understanding for retrieval and local feature correlations for precise localization. We further introduce consistency supervision across positive and semi-positive satellite references and a sub-meter refinement module to improve accuracy and robustness. Localization-aware re-ranking then uses localization evidence to refine retrieved candidates, allowing localization to reinforce retrieval. Experiments on VIGOR and DReSS demonstrate strong performance in both retrieval and fine-grained localization, yielding accurate and fast end-to-end geolocalization from a single shared representation and maintaining robustness as the target moves toward the boundary of the satellite reference.

1 INTRODUCTION

Cross-view image geolocalization aims to estimate the geographic location of a ground image from satellite imagery. The goal is to predict a precise geographic position, which is important for applications such as navigation, autonomous systems, and localization when conventional positioning is unavailable or unreliable (Wilson et al., 2024; Zhu et al., 2021; Fervers et al., 2023). This matters because Global Navigation Satellite System (GNSS) is blocked in dense urban areas and can be jammed or spoofed, which is a real risk for navigation and autonomous systems (Wang et al., 2025).

Existing work follows two main directions. Cross-view retrieval methods learn global representations that match a ground image against a geo-tagged database of satellite imagery, which in principle can cover any location on earth (Workman et al., 2015; Hu et al., 2018; Shi et al., 2019; Liu & Li, 2019; Shi et al., 2020; Yang et al., 2021; Zhang et al., 2023; 2024; Deuser et al., 2023; Zhu et al., 2022; Zhang & Zhu, 2024). While effective for large-scale search, retrieved satellite image only provides a coarse location (i.e. the center of the satellite image). In contrast, fine-grained methods estimate precise location or pose within a given satellite image by exploiting geometric alignment or local correspondences (Xia et al., 2022; Wang et al., 2023; Xia et al., 2024; Shi et al., 2023; Shi & Li, 2022; Lentsch et al., 2023; Song et al., 2023; Xia & Alahi, 2025; Abu Lehyeh et al., 2026; Xia et al., 2026). Although several of these methods also recover the camera orientation (Xia et al., 2024; Xia & Alahi, 2025), in our scope we assume that the orientation is known and focus on estimating the end-to-end 2D position. However, these fine-grained methods assume the satellite reference is already given, which leaves the preceding large-scale retrieval search outside their scope. A complete end-to-end system therefore has to cascade a retrieval model with a fine-grained one, since neither line alone produces a precise global position. The cascade is intuitive, but it treats the two tasks as independent problems when they are in fact two scales of a single one: both ask where the ground image was taken, one across a database and one within a tile. A cascade cannot exploit this. Its interface is strictly one-directional, so once retrieval commits to a tile, localization is confined to refining a position inside that choice, and the geometric evidence it recovers is never allowed to revise the decision. A retrieval error propagates unchecked in the final estimate, and the two tasks never reinforce each other, although each produces evidence the other needs. The cascade also carries two trained models and two inference stages, which is costly wherever latency or memory is constrained.

Recent work has begun to close this gap. UnifyGeo (Song et al., 2026) places retrieval and metric localization in one hierarchical framework and coordinates them with a learned re-ranking module.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The coupling is architectural: after a partially shared backbone, each task is carried by its own learnable heads and decoder, so the tasks share early layers but not the features they are finally read from. In contrast, we ask a different question: *can retrieval and fine-grained localization be learned from a single geometric representation, so that the two tasks mutually reinforce each other through the representation itself?*

We propose GeoSync, unified and efficient framework for precise end-to-end cross-view geolocalization that exposes interpretable signals of prediction reliability. A single backbone supports both global retrieval and fine-grained localization. The shared representation is supervised at complementary spatial scales: global contrastive learning encourages cross-view discriminability, while position correlation and residual refinement impose precise geometric constraints. This formulation allows us to study the interaction between retrieval and localization directly rather than treating one as a fixed upstream stage. In particular, we show that localization reinforces retrieval both implicitly through the shared representation and explicitly through Localization-Aware Re-ranking (LAR). LAR reuses geometric evidence already produced by the localization branch to construct a verification score for retrieved satellite candidates, requiring no additional training or separately learned re-ranking module. The resulting improvement in retrieval directly results in improved end-to-end geographic localization. GeoSync also provides two complementary forms of interpretable confidence. A candidate-level verification signal measures cross-view agreement at the predicted location and indicates whether the retrieved satellite image is consistent with the query, while the concentration of the localization distribution indicates how precisely the model has identified the position within a correct satellite image. We show that these signals separate confident from uncertain predictions, providing a simple way to filter errors.

Our main contributions can be summarized as fourfold. (1) a unified shared representation that jointly supports retrieval and fine-grained localization, together with controlled experiments showing how the two objectives reinforce each other, improving both retrieval and end-to-end geolocalization; (2) A localization-aware re-ranking mechanism that reuses fine-grained geometric evidence back into retrieval without additional training or a separately learned re-ranking module, leading to improved end-to-end localization; (3) a hierarchical confidence analysis that separates candidate verification from within-tile localization confidence and demonstrates their usefulness for selective and failure-aware geolocation; and (4) An end-to-end system that answers a query with a single encoder pass, running faster than competitive baselines and at higher accuracy.

2 Related Work

Cross-view image retrieval. Cross-view geo-localization is commonly formulated as image retrieval, where a ground-level image is matched to a geo-referenced satellite image. Early methods such as CVUSA (Workman et al., 2015), CVM-Net (Hu et al., 2018), and SAFA (Shi et al., 2019) learn global cross-view representations for retrieval. Later methods improve the representation and matching process in different ways. L2LTR (Yang et al., 2021) uses layer-to-layer transformers for cross-view matching, GeoDTR (Zhang et al., 2023) learns disentangled geometric layout correspondences, and Sample4Geo (Deuser et al., 2023) improves retrieval through geographically and visually hard-negative sampling. VIGOR (Zhu et al., 2021) further shows that the query may be spatially offset from the center of its reference tile. AuxGeo (Xia et al., 2025) extends this setting with DReSS, a dataset designed to study large decentrality across diverse cities, and proposes auxiliary supervision to improve retrieval under such offsets. These methods primarily provide a coarse location through satellite tile retrieval, rather than precise position estimation within the tile.

Fine-grained cross-view localization. Fine-grained cross-view localization estimates the fine-grained position of a ground image within a given satellite reference. Global descriptor-based methods (Xia et al., 2022; Lentsch et al., 2023; Xia et al., 2024) compare the query against satellite representations at different candidate poses. Geometry-based methods instead transform the ground view before alignment. HC-Net (Wang et al., 2023) uses a differentiable spherical transform followed by correlation-aware homography estimation. FG^2 (Xia & Alahi, 2025) constructs a 3D point representation, selects features along the height dimension to form a BEV point plane, and estimates pose from sparse correspondences. Loc^2 (Xia et al., 2026) learns ground-satellite image-plane correspondences and lifts the matched ground points with monocular depth before metric pose estimation. These methods achieve precise within-reference localization, but require the satellite reference to be provided in advance and therefore do not address reference selection.

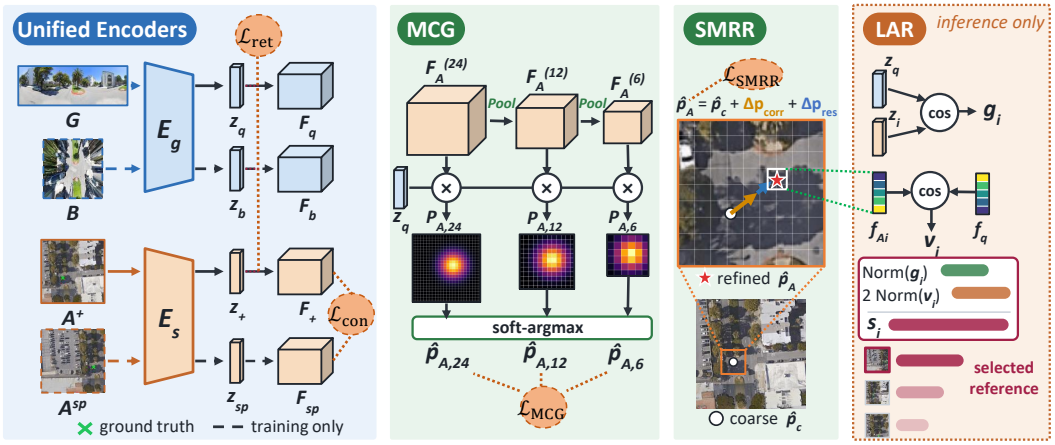

Figure 1: Overview of GeoSync. (a) *Shared encoding.* Ground panorama G, training-only BEV projection B, and primary/semi-positive satellite tiles A^+ and A^{sp} are encoded by E_g and E_s into global CLS embeddings z and patch features F. The query embedding z_q is used for both retrieval and localization. (b) *Coarse-to-fine localization.* The Multi-Scale Correspondence Grid (MCG) matches z_q with satellite features at three scales (24×24, 12×12, 6×6) and applies soft-argmax to obtain a coarse position $\hat{\mathbf{p}}_c$. Sub-Meter Residual Refinement (SMRR) refines it within a local window by a correlation displacement and a residual correction, giving $\hat{\mathbf{p}}_A$. The same geographic point is supervised in A^+ and A^{sp}. (c) *Localization-Aware Re-ranking (LAR).* The top-K candidates from global similarity g_i are localized individually and verified by local feature agreement v_i at their predicted positions. The combined score s_i selects the final tile and location.

Joint retrieval and localization. UnifyGeo (Song et al., 2026) combines retrieval and metric localization in one hierarchical multi-task framework with re-ranking. It shares early encoder stages, then adds task-specific heads, an attention aggregator, and a learned localization decoder to mitigate task conflict. We instead view them as one problem at two scales, and study mutual reinforcement between them. A single learned representation supports both tasks, from which we derive an explicit re-ranking mechanism that enhances retrieval and improves end-to-end localization. Unlike UnifyGeo, which assumes perfectly centered references, we use off-center references and show that the learned representation remains translation-consistent and robust to the resulting position shifts. We further provide interpretable confidence estimates that allow uncertain localization predictions to be filtered before use, a capability not available in UnifyGeo. Together, these properties yield a consistent end-to-end gain over UnifyGeo in accuracy and efficiency.

3 Methodology

Given a ground panorama G and a geo-tagged satellite gallery $\mathcal{A} = \{A_i\}_{i=1}^{M}$, we want the precise position of G. Existing pipelines answer this in two steps: a retrieval model selects a satellite tile, and a second model localizes the query inside it. GeoSync treats both steps as one question asked at two scales, *which* tile contains the query and *where* inside that tile, and answers both from a single representation. The framework has three parts (Figure 1). A shared encoder gives each image a global embedding for retrieval and a patch grid for localization (Section 3.1). A coarse-to-fine localizer reads the query position from the patch grid of a satellite tile: the Multi-Scale Correspondence Grid (MCG) gives a coarse estimate, and Sub-Meter Residual Refinement (SMRR) refines it (Section 3.2). Localization is trained on both centered and off-center tiles, so that any tile covering the query can serve as a reference (Section 3.3). At inference, Localization-Aware Re-ranking (LAR) localizes the query in each retrieved candidate and uses the agreement at the predicted position to re-rank them (Section 3.4). For any tile A covering the query, $\mathbf{p}_A \in \mathbb{R}^2$ denotes the query position in A. During training, each query comes with a primary tile A^+, where it lies near the center, and a semi-positive tile A^{sp}, which covers the same point off-center.

3.1 Shared Encoding and Retrieval

GeoSync uses two independently parameterized DINOv3 (Siméoni et al., 2025) ViT-S+/16 encoders, E_g for ground images and E_s for satellite images. Each encoder outputs a global CLS embedding

$z \in \mathbb{R}^d$ and a 24×24 grid of patch features $F \in \mathbb{R}^{24\times24\times d}$:

$$\begin{aligned}(z_q, F_q) &= E_g(G), \qquad (z_b, F_b) = E_g(B),\\ (z_+, F_+) &= E_s(A^+), \quad (z_{sp}, F_{sp}) = E_s(A^{sp}).\end{aligned} \tag{1}$$

The CLS embedding summarizes the whole view, while the patch grid keeps where things are. B is a bird's-eye-view projection of G that follows AuxGeo (Xia et al., 2025); it serves as an intermediate view between ground and satellite and is used only during training. For any two feature vectors x and y, $\mathrm{sim}(x, y) = \bar{x}^\top \bar{y}$ is their cosine similarity, with $\bar{x} = x/\|x\|_2$, and $f_A(\mathbf{p})$ is the satellite feature bilinearly sampled from F_A at position $\mathbf{p}$.

Retrieval is trained with a symmetric InfoNCE loss over a batch of N samples,

$$\mathcal{L}_{\mathrm{NCE}}\left(x, \{y_j\}_{j=1}^N\right) = -\log \frac{\exp\left(\mathrm{sim}(x, y^+)/\tau\right)}{\sum_{j=1}^N \exp\left(\mathrm{sim}(x, y_j)/\tau\right)}, \tag{2}$$

where y^+ is the matching sample and τ is a temperature. The retrieval objective aligns the ground, BEV, and satellite embeddings:

$$\mathcal{L}_{\mathrm{ret}} = \mathcal{L}_{\mathrm{NCE}}(z_q, z_+) + \lambda_b \left[\mathcal{L}_{\mathrm{NCE}}(z_q, z_b) + \mathcal{L}_{\mathrm{NCE}}(z_b, z_+)\right]. \tag{3}$$

The same query embedding z_q is then used for localization: it is compared with satellite *embeddings* to find the tile, and with satellite *patch features* to find the position inside it.

3.2 Coarse-to-Fine Localization

Given a candidate tile A, GeoSync estimates the query position in two steps. MCG finds where to look in the tile, and SMRR refines this estimate below the patch resolution. Neither step uses a learned decoder: both operate directly on the shared features, with only a small MLP in SMRR.

Multi-Scale Correspondence Grid (MCG). The query embedding z_q is compared with every cell of the tile's patch grid, pooled to $r \times r$ cells, $F_A^{(r)} = \mathcal{P}_r(F_A)$ with $r \in \mathcal{R} = \{24, 12, 6\}$:

$$P_{A,r}(u \mid z_q) = \frac{\exp\Big(\mathrm{sim}(z_q, F_A^{(r)}(u))/\tau_m\Big)}{\sum_{u'} \exp\Big(\mathrm{sim}(z_q, F_A^{(r)}(u'))/\tau_m\Big)}, \qquad \hat{\mathbf{p}}_{A,r}(z_q) = \sum_u P_{A,r}(u \mid z_q)\, \mathbf{c}_r(u), \tag{4}$$

where τ_m is a temperature and $\mathbf{c}_r(u)$ is the center of cell u. The softmax turns cross-view similarity into a probability map over the tile, and the soft-argmax turns this map into a continuous position. Coarser levels cover more context per cell, while the finest level gives the most precise estimate. All levels are supervised, for z_q and, during training, z_b:

$$\mathcal{L}_{\mathrm{MCG}}(A) = \frac{1}{2} \sum_{r\in\mathcal{R}} \alpha_r \left(\|\hat{\mathbf{p}}_{A,r}(z_q) - \mathbf{p}_A\|_2 + \|\hat{\mathbf{p}}_{A,r}(z_b) - \mathbf{p}_A\|_2\right), \tag{5}$$

where α_r weights each level. At inference, the coarse estimate is $\hat{\mathbf{p}}_c = \hat{\mathbf{p}}_{A,24}(z_q)$. Since MCG uses the retrieval embedding z_q as its query, $\mathcal{L}_{\mathrm{MCG}}$ also trains the retrieval representation. Retrieval alone only asks z_q to match the tile as a whole; MCG additionally asks it to match the satellite features at the query location rather than anywhere else in the tile. This finer signal improves top-1 retrieval over a retrieval-only model which shows implicit reinforcement in our model (Section 4.3).

Sub-Meter Residual Refinement (SMRR). A single grid cell still covers several meters on the ground. SMRR samples satellite features at a small set of offsets $\Delta\mathbf{c}_j$ around $\hat{\mathbf{p}}_c$ and compares them with a local query descriptor f_q, obtained by mean-pooling F_q:

$$\pi_j = \mathrm{softmax}_j\left(\frac{\mathrm{sim}(f_q, f_A(\hat{\mathbf{p}}_c + \Delta\mathbf{c}_j))}{\tau_f}\right), \qquad \Delta\mathbf{p}_{\mathrm{corr}} = \sum_j \pi_j\, \Delta\mathbf{c}_j, \tag{6}$$

where τ_f is a temperature. The displacement $\Delta\mathbf{p}_{\mathrm{corr}}$ moves the estimate toward the best-matching point in the window, and a small MLP predicts a bounded residual $\Delta\mathbf{p}_{\mathrm{res}}$:

$$\hat{\mathbf{p}}_A = \hat{\mathbf{p}}_c + \Delta\mathbf{p}_{\mathrm{corr}} + \Delta\mathbf{p}_{\mathrm{res}}. \tag{7}$$

We detach $\hat{\mathbf{p}}_c$, so refinement does not change the MCG distribution, and supervise $\hat{\mathbf{p}}_A$ with the Huber loss (Gokcesu & Gokcesu, 2021), which is quadratic for small errors and linear beyond a threshold β to limit the effect of outliers:

$$\mathcal{L}_{\mathrm{SMRR}}(A) = \frac{1}{2} \sum_{d \in \{x,y\}} \ell_\beta\left(\hat{p}_A^{\,d} - p_A^d\right), \qquad \ell_\beta(e) = \begin{cases} \dfrac{e^2}{2\beta}, & |e| \le \beta, \\ |e| - \dfrac{\beta}{2}, & |e| > \beta. \end{cases} \tag{8}$$

3.3 Spatially Consistent Training

Retrieval does not guarantee a centered tile: any tile that covers the query can be retrieved, and the query may sit near its border. A localizer trained only on centered tiles expects the query near the center and fails on such tiles (Section 4.4). We therefore supervise localization on both A^+ and A^{sp}, each with its own ground truth, and ask the two tiles to describe the same geographic point with the same feature:

$$\mathcal{L}_{\mathrm{loc}} = \frac{1}{2}\left[\mathcal{L}_{\mathrm{MCG}}(A^+) + \mathcal{L}_{\mathrm{MCG}}(A^{sp})\right] + \frac{\lambda_s}{2}\left[\mathcal{L}_{\mathrm{SMRR}}(A^+) + \mathcal{L}_{\mathrm{SMRR}}(A^{sp})\right] + \lambda_c \mathcal{L}_{\mathrm{con}}, \tag{9}$$

$$\mathcal{L}_{\mathrm{con}} = 1 - \mathrm{sim}(f_{A^+}(\mathbf{p}_{A^+}),\, f_{A^{sp}}(\mathbf{p}_{A^{sp}}))\,. \tag{10}$$

These terms make a satellite feature depend on *what* is at a location rather than *where* it falls in the tile, so every tile covering the query becomes a usable reference, not only the best-centered one. The complete end-to-end objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{ret}} + \lambda_{\mathrm{loc}} \mathcal{L}_{\mathrm{loc}}, \tag{11}$$

with loss weights given in Appendix A. Trained this way, a single shared model keeps localization accuracy close to a localization-only model while improving top-1 retrieval (Section 4.3).

3.4 Localization-Aware Re-ranking

Training couples retrieval and fine-grained localization through the shared embedding. LAR couples them at inference as well and show explicit reinforcement: retrieval proposes candidates, and localization checks them. Satellite embeddings and patch grids are encoded offline once, so each query needs a single forward pass through E_g. Global retrieval computes $g_i = \mathrm{sim}(z_q, z_i)$ and keeps the top-K candidates. Each candidate is localized with MCG and SMRR to obtain $\hat{\mathbf{p}}_i$, and verified at that position with the SMRR descriptor, $v_i = \mathrm{sim}(f_q, f_{A_i}(\hat{\mathbf{p}}_i))$. The candidates are re-ranked by

$$s_i = \mathrm{Norm}(g_i) + \gamma\, \mathrm{Norm}(v_i), \tag{12}$$

where $\mathrm{Norm}(\cdot)$ normalizes each score over the K candidates of a query and $\gamma = 2$ favors local verification (Appendix A.5). The candidate with the largest s_i and its position form the prediction. LAR needs no training and makes localization explicitly reinforce retrieval. Because candidate features are precomputed and localization reuses the shared representation, re-ranking even twenty candidates adds almost no time (Section 4.2). And because localization also works on off-center tiles (Section 3.3), LAR can promote any tile that covers the query, not only the best-centered one.

Hierarchical confidence. GeoSync gives confidence at two levels. At the retrieval level, v_i measures query-satellite agreement at the predicted position and indicates whether the retrieved tile is correct. At the fine-grained location level, we use the negative entropy of the finest MCG map, $c_i^{\mathrm{loc}} = \sum_u P_i(u) \log P_i(u)$ with $P_i(u) = P_{A_i,24}(u \mid z_q)$. A sharp map gives high c_i^{loc} and a confident position, while a diffuse or multi-modal map signals positional uncertainty (Section 4.4).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate GeoSync on VIGOR and DReSS under same-area (S) and cross-area (C) protocols. Dataset details and evaluation settings are provided in Appendix B.1.

Metrics. We evaluate end-to-end geolocalization by retrieving a satellite reference and localizing the query within it, reporting median error and R@1m, R@5m, and R@10m. We additionally evaluate

Table 1: **End-to-end geolocalization.** The model first selects a satellite candidate from the retrieval pool and then estimates the query position within that candidate. Median error is reported in meters, together with the percentage of predictions within $1/5/10$ m. Best in **bold**.

	Same Area				Cross Area			
Method	Med.↓	R@1m↑	R@5m↑	R@10m↑	Med.↓	R@1m↑	R@5m↑	R@10m↑
VIGOR								
AuxGeo → HC-Net	1.30	39.52	81.82	86.85	2.08	22.46	67.47	70.80
AuxGeo → FG^2	1.28	39.86	83.28	89.35	1.97	24.07	68.26	72.28
UnifyGeo	1.59	29.58	83.79	88.44	1.85	24.87	70.46	72.65
GeoSync + LAR (K=10)	1.21	40.70	89.64	93.44	1.64	27.10	79.97	82.14
GeoSync + LAR (K=20)	**1.21**	**40.75**	**89.82**	**93.62**	**1.63**	**27.24**	**80.38**	**82.56**
DReSS								
AuxGeo → HC-Net	23.92	7.70	37.26	44.33	74.90	2.04	19.18	25.01
AuxGeo → FG^2	24.38	1.81	23.76	38.50	89.59	0.62	10.43	20.47
GeoSync + LAR (K=10)	2.11	23.00	67.50	71.03	47.75	8.41	42.81	47.71
GeoSync + LAR (K=20)	**2.04**	**23.59**	**68.97**	**72.52**	**8.52**	**9.25**	**45.90**	**50.47**

Table 2: **Inference computational efficiency on VIGOR cross-area.** Latency is measured on one H200 at batch size 1, satellite gallery features are precomputed offline. Lower is better.

Method	Params	FLOPs	Latency	Peak Mem.
AuxGeo → HC-Net	99.8 M	99.0 G	24.61 ms	978 MB
AuxGeo → FG^2	453.5 M	5599.6 G	103.10 ms	2043 MB
UnifyGeo	57.7 M	114.2 G	7.95 ms	689 MB
GeoSync	**57.4 M**	**90.9 G**	**6.28 ms**	**322 MB**
+ LAR (K=20)	**57.4 M**	**90.9 G**	6.36 ms	354 MB

retrieval and fine-grained localization separately. Retrieval uses Recall@k ($k \in 1, 5, 10$), measuring whether a valid reference is retrieved within the top k, and Hit Rate, which also considers whether the retrieved tile contains the query location. Fine-grained localization uses the ground-truth satellite reference and reports mean error and median error.

Baselines. For retrieval, we compare TransGeo, FRGeo, Sample4Geo, AuxGeo, and UnifyGeo; for fine-grained localization, HC-Net, FG^2, CCVPE, Loc^2, and UnifyGeo. For end-to-end evaluation, we use retrieval-localization cascades (AuxGeo→HC-Net and AuxGeo→FG^2) and the UnifyGeo pipeline. AuxGeo is used in the cascades as a strong retrieval baseline, while HC-Net and FG^2 represent dedicated strong localization methods. Since most localization methods report only VIGOR, we retrain HC-Net and FG^2 on DReSS using their official code, these results are marked with †.

4.2 End-to-End Geolocalization: Accuracy and Computational Efficiency

We first evaluate GeoSync end to end, where the model must retrieve a satellite reference and then localize the query inside it. This is the setting the unified framework is designed for, so we report both what it achieves and what it costs.

Accuracy. Table 1 shows that GeoSync achieves the best reported end-to-end results across both datasets, outperforming the cascaded pipelines and UnifyGeo. The gains are largest in the harder cross-area setting: on VIGOR, GeoSync reaches a median error of 1.63m and improves R@5m by more than ten points over the strongest cascade. On DReSS, increasing the candidate set from K=10 to K=20 reduces cross-area median error by nearly sixfold. This demonstrates LAR's power to pull the correct reference into top-1 from a larger candidate set, improving end-to-end localization beyond cascades restricted by a single top-1 retrieval decision.

Inference computational efficiency. Because retrieval and localization share one representation in GeoSync, each query requires a single backbone pass, after which the same features support retrieval, LAR re-ranking, and fine-grained localization. GeoSync answers a query in 6.28ms, faster than UnifyGeo (7.95ms) and the HC-Net and FG^2 cascades at 24.61ms and 103.10ms, respectively, while using fewer parameters and less peak memory (Table 2). UnifyGeo's learned re-ranking and intermediate modules add computation, whereas LAR reuses the shared representation. Re-ranking

Table 3: **Cross-view retrieval.** S and C denote same-area and cross-area evaluation. LAR re-ranks the top-K retrieved candidates using the localization verification score, and requires no additional training. Best in **bold**, second best underlined. R@10 and R@1% are reported in Appendix 8.

	VIGOR-S			VIGOR-C			DReSS-S			DReSS-C		
Method	R@1	R@5	Hit	R@1	R@5	Hit	R@1	R@5	Hit	R@1	R@5	Hit
TransGeo	–	–	–	–	–	–	19.86	40.29	–	4.09	11.45	–
FRGeo	–	–	–	–	–	–	27.19	46.17	–	8.07	18.23	–
Sample4Geo	77.86	95.66	89.82	61.70	83.50	69.87	51.40	72.78	55.89	28.23	48.19	30.65
AuxGeo	80.34	96.25	93.78	63.94	84.98	76.25	54.70	76.09	59.20	32.44	53.18	34.90
UnifyGeo	**82.80**	94.92	89.12	67.58	81.60	72.27	–	–	–	–	–	–
GeoSync	79.11	94.73	92.26	63.35	83.73	75.15	55.12	73.79	59.02	34.97	58.24	37.17
+ LAR (K=10)	80.65	95.84	94.87	68.72	86.47	83.16	68.24	77.66	72.42	46.81	58.24	49.47
+ LAR (K=20)	80.87	**96.44**	**95.09**	**69.09**	**88.02**	**83.60**	**69.72**	**80.77**	**73.97**	**48.14**	**61.91**	**50.89**

Table 4: **Fine-grained localization** with the ground-truth satellite reference. Position error is reported in meters, lower is better. † retrained by us on DReSS with official code; – denotes unreported or unreleased settings. Best in **bold**, second best underlined.

	VIGOR-S		VIGOR-C		DReSS-S		DReSS-C	
Method	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
CCVPE	3.60	1.36	4.97	1.68	–	–	–	–
Loc^2	3.06	1.59	3.43	1.90	–	–	–	–
HC-Net	2.65	1.17	3.35	1.59	11.09†	3.29†	15.19†	5.16†
FG^2	**1.95**	**1.08**	**2.41**	**1.37**	12.42†	6.77†	15.50†	9.64†
UnifyGeo	2.47	1.45	3.08	1.42	–	–	–	–
GeoSync	2.14	1.18	2.72	1.46	**3.52**	**1.58**	**6.24**	**2.53**

twenty candidates adds only 0.1ms and 32 MB, while reducing cross-area median error to 1.63m and improving R@1m by three points. Thus, GeoSync achieves efficient end-to-end performance by reusing shared features without additional computational stages.

4.3 Retrieval and Fine-Grained Localization: Mutual Reinforcement

Having established end-to-end performance, we next isolate the two capabilities that must be supported by a unified geolocalization system: global candidate retrieval and precise localization within a satellite reference.

Cross-view retrieval. Table 3 reports retrieval on all four protocols. Adding a localization objective does not cost global discriminability: without any re-ranking, GeoSync is already on par with dedicated retrieval architectures, and it is the strongest method on DReSS, where candidates are geographically dispersed and global descriptors are least reliable. LAR further improves retrieval without any retraining. The gains are especially clear in the challenging cross-area settings, indicating stronger generalization under geographic shift.

Fine-grained localization. Table 4 supplies the correct satellite image to every method and therefore isolates fine-grained localization. On VIGOR, GeoSync remains competitive with specialized fine-grained localizers, achieving 2.72 m / 1.46 m in cross-area settings. On DReSS, GeoSync clearly outperforms our HC-Net reproduction, reducing the median error from 5.16 m to 2.53 m in cross-area. FG^2 and HC-Net degrade more on DReSS, where lower-resolution satellite imagery provides less fine-grained detail and greater decentrality makes correspondence-based localization more difficult. We attribute this performance in part to our SMRR module, which refines the predicted location, and to our consistency module, which encourages similar representations for positive and semi-positive satellite references. This makes the geometric representation more translation-consistent, reducing sensitivity to target position, particularly under DReSS's larger decentrality.

Mutual task reinforcement. Table 5 shows how retrieval and localization reinforce each other when learned jointly. Retrieval-only training provides strong candidate selection but poor fine-grained localization, while localization-only training does the opposite, *showing that neither objective alone is sufficient for an end-to-end system. Joint training preserves accurate fine-grained localization*

Table 5: **Reinforcement between retrieval and fine-grained localization** on VIGOR. Each row uses the same trained model for both evaluations. Retrieval is evaluated on the standard candidate set, while localization is evaluated with the ground-truth satellite reference to isolate fine-grained accuracy. LAR only changes candidate ranking, leaving localization unchanged. Best in **bold**.

	Retrieval				*Fine-grained localization*			
	Same Area		Cross Area		Same Area		Cross Area	
Training variant	R@1↑	Hit↑	R@1↑	Hit↑	Mean↓	Med.↓	Mean↓	Med.↓
Retrieval-only	76.13	**95.09**	62.17	76.66	13.00	12.16	12.81	12.17
Localization-only	6.46	6.60	5.57	5.75	2.19	**1.14**	2.89	**1.42**
GeoSync (joint)	79.11	92.26	63.35	75.15	**2.14**	1.18	**2.72**	1.46
+ LAR (K=10)	80.65	94.87	68.72	83.16	**2.14**	1.18	**2.72**	1.46
+ LAR (K=20)	**80.87**	**95.09**	**69.09**	**83.60**	**2.14**	1.18	**2.72**	1.46

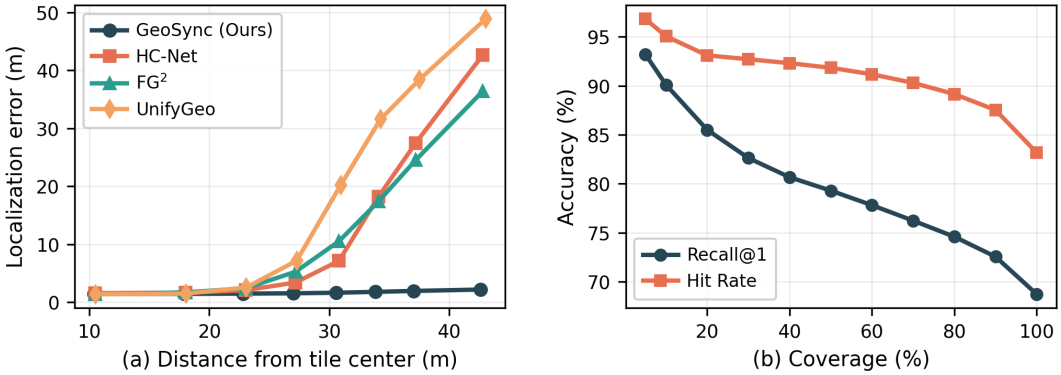

Figure 2: **Robustness and selective prediction on VIGOR cross-area.** (a) Queries are grouped by the physical distance between the ground-truth location and the center of the selected satellite reference, each point is the median localization error within an equal-count bin. GeoSync stays nearly flat as the displacement grows, while all other methods degrade sharply. (b) Predictions ranked by the verification score v, with the top $x\%$ retained, retrieval accuracy improves smoothly as coverage shrinks, showing the model's own confidence signal can filter out unreliable retrieval predictions. Full numerical values, including the localization side, are in Appendix 10.

while adding strong retrieval. This demonstrates that the shared representation supports both objectives without sacrificing fine-grained accuracy. LAR further exposes an explicit reinforcement at inference by using localization to refine retrieval. On VIGOR cross-area, LAR10 improves R@1 from 63.35% to 68.72% and Hit Rate from 75.15% to 83.16%. Thus, retrieval enables localization, while localization provides local evidence that resolves retrieval ambiguity. Together, the results demonstrate mutual reinforcement between the two tasks.

4.4 Reference Displacement Robustness, Confidence and Component Ablation

Robustness beyond the reference-centered assumption. Fine-grained localization is usually evaluated with the query close to the center of its satellite reference, an assumption a retrieval-driven system cannot guarantee. We therefore group VIGOR-C queries by the physical distance between the ground-truth location and the reference-tile center and report the median localization error in Figure 2(a). GeoSync remains nearly flat as displacement grows, while competing methods degrade steadily. We attribute this robustness to consistency learning and semi-positive supervision, which expose the same location through differently positioned references. Thus, retrieval does not need to return the best-centered reference, any reference containing the query can support accurate localization, highlighting why Hit Rate is an important metric for our pipeline.

Hierarchical confidence. GeoSync provides confidence at two levels. First, the verification score measures whether the retrieved satellite reference is correct by evaluating cross-view agreement at the predicted location. Figure 2(b) shows this directly by varying the percentage of predictions retained: keeping only the most confident predictions substantially improves retrieval accuracy. At 5% coverage, verification raises R@1 from 68.72% to 93.22%. Second, confidence in the fine-grained

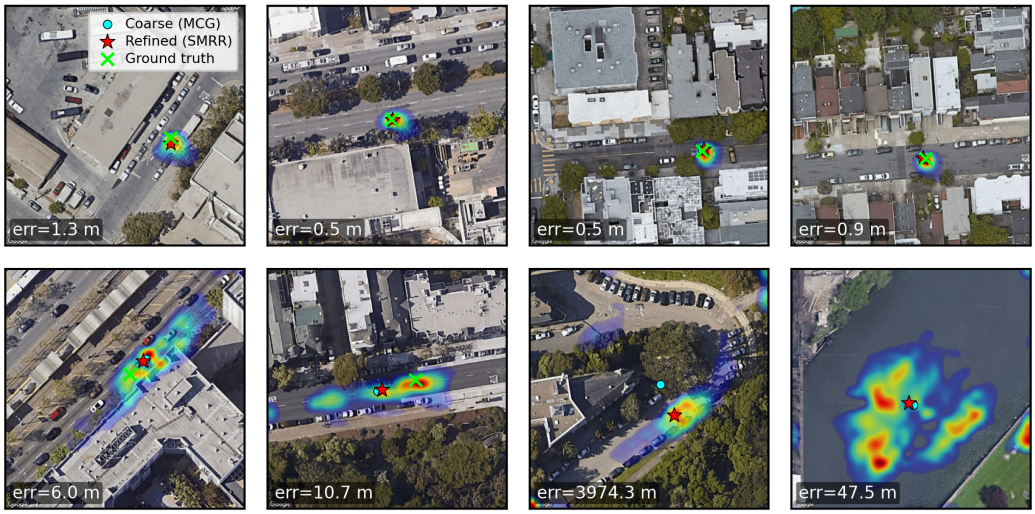

Figure 3: **Qualitative confidence.** Each example shows the retrieved satellite tile with the localization distribution overlaid. *Top:* confident predictions with a concentrated distribution and low localization error. *Bottom:* uncertain predictions with diffuse and larger errors.

Table 6: **Component ablation** on VIGOR. *Primary* uses the best-centered positive satellite tile, while *Random* evaluates on a mixture of positive and semi-positive tiles. Errors are in meters. The final variant removes semi-positive supervision together with the cross-tile consistency loss.

	Same Area						Cross Area					
	Primary			*Random*			*Primary*			*Random*		
Configuration	Mean$^{\downarrow}$	Med.$^{\downarrow}$	R@1m$^{\uparrow}$	Mean$^{\downarrow}$	Med.$^{\downarrow}$	R@1m$^{\uparrow}$	Mean$^{\downarrow}$	Med.$^{\downarrow}$	R@1m$^{\uparrow}$	Mean$^{\downarrow}$	Med.$^{\downarrow}$	R@1m$^{\uparrow}$
MCG	2.39	1.41	34.01	3.18	1.70	27.67	3.15	1.80	22.57	4.03	2.01	19.60
+ SMRR	2.12	1.17	42.36	**2.69**	**1.32**	**37.09**	2.80	1.47	30.48	**3.62**	**1.66**	**25.81**
w/o semi-positives	**1.92**	**1.15**	**43.14**	17.36	11.48	12.81	**2.27**	**1.41**	**32.26**	15.96	8.68	10.85

location is reflected by the sharpness of the localization heatmap. Figure 3 shows that confident predictions produce a sharp, concentrated response around the predicted position, while uncertain cases produce diffuse or multiple plausible responses. The latter cases often occur in scenes with repeated or ambiguous structure, where the ground view provides weaker evidence for distinguishing nearby locations. Thus, verification indicates confidence in the retrieved satellite reference, while heatmap sharpness indicates confidence in the predicted position within that reference.

Component ablation. Table 6 isolates the localization components under centered (Primary) and spatially displaced (Random) references. Adding SMRR consistently improves localization: on VIGOR cross-area, median error drops from 1.80m to 1.47m under Primary and from 2.01m to 1.66m under Random, while R@1m increases from 22.57% to 30.48% and from 19.60% to 25.81%, respectively. The gains under Random show that SMRR remains effective under spatial displacement. Removing semi-positive supervision and cross-tile consistency further degrades performance in this setting, confirming their role in improving robustness. Refer to Appendix C.4.

5 Conclusion

We propose GeoSync, a unified framework for cross-view image geolocalization that learns retrieval and fine-grained localization from a shared cross-view representation. By combining multi-scale localization, sub-meter refinement, and spatially consistent supervision, GeoSync maintains accurate localization even when the query is far from the center of the satellite reference. Localization-Aware Re-ranking further allows fine-grained geometric evidence to refine retrieval, demonstrating that the two tasks can mutually reinforce each other. Experiments on VIGOR and DReSS show state-of-the-art end-to-end accuracy with substantially lower inference cost than cascaded pipelines. GeoSync also provides simple, interpretable confidence signals at both the reference and fine-grained localization levels, offering a practical basis for identifying uncertain predictions.

Reproducibility Statement

GeoSync is built from standard components: two DINOv3 encoders, a parameter-free multi-scale correlation (MCG), a small refinement (SMRR) module, and a training-free re-ranking rule (LAR). Section 3 describes each component and its training objective. Appendix A lists all hyperparameters, the training configuration, and the implementation details of every module, and Appendix B describes the datasets, the evaluation protocol for each baseline, and how inference cost is measured. All experiments use the public VIGOR and DReSS benchmarks with their official splits. To support reproducibility, we will publicly release our code, trained model weights, configuration files, and evaluation scripts, so that all results reported in this paper can be reproduced.

References

Ayesh Abu Lehyeh, Xiaohan Zhang, Ahmad Arrabi, Waqas Sultani, Chen Chen, and Safwan Wshah. Geoflow: Real-time fine-grained cross-view geolocalization via iterative flow prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5369–5378, June 2026.

Fabian Deuser, Konrad Habel, and Norbert Oswald. 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.

Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. 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.

Kaan Gokcesu and Hakan Gokcesu. Generalized huber loss for robust learning and its efficient minimization for a robust statistics. *arXiv preprint arXiv:2108.12627*, 2021.

Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7258–7267, 2018.

Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. Slicematch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17225–17234, 2023.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Yujiao Shi and Hongdong Li. 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.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32:10090–10100, 2019.

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2682–2697, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Zhenbo Song, Jianfeng Lu, Yujiao Shi, et al. Learning dense flow field for highly-accurate cross-view camera localization. *Advances in Neural Information Processing Systems*, 36:70612–70625, 2023.

Zhuo Song, Ye Zhang, Kunhong Li, Longguang Wang, and Yulan Guo. A unified hierarchical framework for fine-grained cross-view geo-localization over large-scale scenarios. *arXiv preprint arXiv:2505.07622*, 2026.

Ke Wang, Chongqiang Shen, Xingcan Li, and Jianbo Lu. Uncertainty quantification for safe and reliable autonomous vehicles: A review of methods and applications. *IEEE Transactions on Intelligent Transportation Systems*, 26(3):2880–2896, 2025. doi: 10.1109/TITS.2025.3532803.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *Advances in Neural Information Processing Systems*, 36:5301–5319, 2023.

Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132(4):1350–1392, 2024.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, et al. Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025.

Zimin Xia and Alexandre Alahi. Fgˆ 2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6362–6372, 2025.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision*, pp. 90–106. Springer, 2022.

Zimin Xia, Olaf Booij, and Julian F. P. Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3813–3831, 2024. doi: 10.1109/TPAMI.2023.3346924.

Zimin Xia, Chenghao Xu, and Alexandre Alahi. Loc^2: Interpretable cross-view localization via depth-lifted local feature matching. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29009–29020, 2021.

Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7251–7259, 2024.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3480–3488, 2023.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3640–3649, June 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. 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*, pp. 1162–1171, 2022.

APPENDIX OVERVIEW

The appendix is organized as follows. Appendix A gives the full implementation details of GeoSync: backbone and inputs, training configuration, loss weights, the localization modules (MCG and SMRR), and the LAR scoring rule. Appendix B describes the datasets, the evaluation protocol for each baseline, and how inference cost is measured. Appendix C reports extended quantitative results that support the main paper: complete retrieval numbers, mutual reinforcement on DReSS, the two levels of confidence, the full localization ablation, and the center-bias and confidence analysis on DReSS. Appendix D shows qualitative results, covering typical successes and failures, how LAR recovers correct tiles, and what the refinement step changes. Finally, the LLM usage statement is given at the end of the appendix.

A IMPLEMENTATION DETAILS

A.1 BACKBONE AND INPUTS

GeoSync uses two DINOv3 ViT-S+/16 encoders with a hidden size of d=384 and a patch size of 16 (Section 3.1). The ground encoder E_g processes both the ground panorama and its BEV projection, and the satellite encoder E_s processes both the primary and the semi-positive tile. Because the same encoder produces the global embedding used for retrieval and the patch features used for localization, each gallery tile only needs to be encoded once, offline, for both tasks.

Ground panoramas are resized to 384×768 and satellite tiles to 384×384, so each satellite tile yields a native 24×24 patch grid. MCG further pools this grid to 12×12 and 6×6 (Appendix A.4). Ground images keep the 2:1 aspect ratio of the panorama and yield a 24×48 patch grid.

A.2 TRAINING CONFIGURATION

All models are trained for 100 epochs with batch size 128 and evaluated every 4 epochs. We use AdamW with a base learning rate of 10^{-3}, weight decay 0.01, a cosine schedule with a 1-epoch warmup, gradient value clipping at 100, and mixed-precision training with an initial gradient scale of 2^{10}. Label smoothing of 0.1 is applied inside the retrieval contrastive loss. Data augmentation uses rotation with probability 0.75 and horizontal flipping with probability 0.5. Hard negatives are refreshed at every evaluation cycle: for each query, we draw a pool of 128 nearby tiles using GPS proximity and current embedding similarity, and keep the 64 most confusable ones.

Layer-wise learning-rate decay. A practical finding of this work is that *how* DINOv3 is fine-tuned matters a great deal. Instead of updating all layers with the same learning rate, we use layer-wise learning-rate decay (LLRD): the last transformer block is trained with the base learning rate, and each block closer to the input uses 0.75 times the rate of the block above it. The first of the 12 blocks therefore trains at about 4% of the base rate and the patch embedding at about 2%, while the final norm layer and the SMRR head use the full base rate. In this way, the early layers, which hold the general visual features learned during pre-training, change only slightly, while the later layers adapt freely to the cross-view tasks.

This choice has a large effect. We compared LLRD with full fine-tuning, LoRA adapters (rank 8 on the attention query and value projections), and freezing the first half or three quarters of the blocks. LLRD gave the best retrieval accuracy by a clear margin: R@1 of 63.35 on VIGOR cross-area, versus 49.12 for full fine-tuning at the same training budget. Updating all layers equally appears to overwrite pre-trained features that transfer well to both ground and satellite views, while freezing layers prevents the adaptation that the cross-view gap requires. All reported results use LLRD.

A.3 LOSS WEIGHTS AND HYPERPARAMETERS

Table 7 lists all values used for the objectives in Section 3. The query-satellite retrieval term carries the full weight, since it is the objective used at test time. The two BEV terms act as auxiliary constraints between the three views of the same location and are kept small, so they regularize rather than dominate training. Each retrieval term has its own learnable temperature.

Table 7: **Hyperparameters of GeoSync**, grouped by the component they belong to. Symbols follow Section 3.

Component	Hyperparameter	Value
Retrieval (Eq. 2)	temperature τ	learnable, one per term
	BEV weight λ_b	0.1
MCG	temperature τ_m	0.035
	level weights $(\alpha_{24}, \alpha_{12}, \alpha_6)$	$(0.5,\ 0.3,\ 0.2)$
SMRR	local window	9×9 patch cells
	temperature τ_f	0.07
	residual MLP	$81 \to 64 \to 2$, $\tanh$-bounded to 2 px
	SMRR weight λ_s	0.35
	Huber threshold β	1.0
Localization	consistency weight λ_c	0.5
	localization weight λ_{loc}	0.1
LAR	score normalization $\text{Norm}(\cdot)$	z-score over the top-K
	verification weight γ	2

Role of semi-positive supervision. Semi-positive supervision and the consistency term $\mathcal{L}_{\text{con}}$ (Section 3.3) do more than regularize. Removing both leaves accuracy on centered tiles unchanged or slightly better, but localization collapses when the query is evaluated on an arbitrary tile that covers it: on VIGOR cross-area, R@1m drops from 25.81% to 10.85% under the *Random* reference protocol (Table 11). These terms are therefore what makes the model resistant to center bias, rather than a source of general localization accuracy.

A.4 Coarse Localization (MCG) and Refinement (SMRR)

MCG. MCG correlates the query embedding z_q with every cell of the satellite patch grid at three resolutions (24×24, 12×12, 6×6), with softmax temperature τ_m=0.035. Each resolution produces its own soft-argmax estimate, and the three are supervised with level weights α_{24}=0.5, α_{12}=0.3, and α_6=0.2, so the finest, most precise level receives the largest weight while the coarser levels provide a stable signal over a wider context. At inference, the 24×24 estimate gives the coarse position $\hat{\mathbf{p}}_c$.

SMRR. SMRR takes $\hat{\mathbf{p}}_c$ as a fixed anchor and looks only at a 9×9 window of satellite patch features around it. It correlates these features with the local query descriptor f_q using temperature τ_f=0.07, and adds two residuals to the anchor: the soft-argmax displacement $\Delta\mathbf{p}_{\text{corr}}$ of this local correlation map, and a learned residual $\Delta\mathbf{p}_{\text{res}}$ from a two-layer MLP that reads the same 9×9 correlation window (81 inputs, 64 hidden units, output bounded by $\tanh$ to at most 2 pixels). The gradient of SMRR is stopped before the coarse estimate, so the two stages interact only through the forward pass. With everything else fixed, adding SMRR improves R@1m by 6 to 9 points and reduces median error by 17 to 22% on both VIGOR splits and under both reference protocols (Table 11).

A.5 LAR Scoring and Fusion Weight

LAR re-ranks the top-K candidates with signals produced as a byproduct of localizing the query in each candidate, without any additional training (Section 3.4). To choose the scoring rule, we first extracted eight per-candidate signals: the global similarity g_i; the peak value and negative entropy of the 24×24 MCG map; the ratio between its two largest modes; the agreement between the three MCG estimates; the size of the SMRR displacement; the peak of the SMRR correlation map; and the verification score v_i at the refined position.

Each signal is z-scored over the K candidates of a query. The final rule, $s_i = \text{Norm}(g_i) + 2\,\text{Norm}(v_i)$, keeps only the two signals that matter most. We compared it with unweighted reciprocal rank fusion over five of the eight signals, and with a logistic regression fitted on a held-out validation split to learn weights for all eight. The learned weights (verification $+2.01$, global $+0.91$, peakedness $+0.68$, all others much smaller) are close to our 2:1 ratio, and the learned model im-

proves Hit Rate by only half a point while losing just over one point of R@1. We therefore use the simpler two-signal rule, which needs no fitting and no held-out data.

Increasing the number of re-ranked candidates from $K{=}10$ to $K{=}20$ gives a small further gain on all protocols (Table 3), since a wider pool recovers more correct tiles but also adds more plausible distractors.

B Experimental Setup

B.1 Datasets

We evaluate GeoSync on VIGOR (Zhu et al., 2021) and DReSS (Xia et al., 2025), each under same-area (S) and cross-area (C) protocols. VIGOR contains ground panoramas and satellite tiles from four U.S. cities. Its cross-area protocol trains on two cities and evaluates on the other two, measuring cross-city generalization within the United States. DReSS is designed to study decentrality in cross-view geolocalization and covers more than 400 km^2 in each of eight cities worldwide. Its reference imagery is built over these large areas, so the query is not necessarily close to the center of its satellite reference. Its cross-area protocol evaluates on Johannesburg, Rio, Taipei, and Vancouver spanning four continents, a much broader generalization setting than VIGOR. We chose these two datasets because they support end-to-end evaluation, covering both retrieval and fine-grained localization on the same queries.

B.2 Baselines and Evaluation Protocol

Cascades. To build strong cascaded baselines, we combine leading models from each task. AuxGeo is among the strongest retrieval methods on both VIGOR and DReSS, and HC-Net and FG^2 are among the strongest fine-grained localizers on VIGOR. In each cascade, AuxGeo selects the top-1 satellite tile and HC-Net or FG^2 localizes the query within it. On VIGOR we use the official checkpoints; on DReSS we retrain HC-Net and FG^2 with their official code and default settings (marked † in the tables).

UnifyGeo. UnifyGeo's reported localization results are not computed end to end. We therefore evaluate its released VIGOR checkpoints under the same protocol as all other methods, using its default re-ranking over $k{=}5$ candidates. For end-to-end results, the top-1 tile after its re-ranking is used for localization, and the error is measured against the true query position. For consistency and a fair comparison, all UnifyGeo numbers in Tables 1, 2, 3, 4, and 8 are these reproduced results. No DReSS checkpoints are available, so UnifyGeo is reported on VIGOR only.

End-to-end error. In all end-to-end results, the error is the distance between the predicted position and the query's true GPS location, whether or not the selected satellite tile covers the query.

B.3 Inference Cost Measurement

Table 2 measures the online cost of answering one query on VIGOR cross-area. All methods run on the same H200 GPU with batch size 1 in fp32, and latency is averaged over repeated timed runs after warm-up. Satellite gallery features are precomputed offline for every method, so the comparison measures only the per-query cost. For GeoSync this is one pass of E_g followed by retrieval, MCG, SMRR, and, when enabled, LAR over the stored patch grids. Re-ranking twenty candidates adds under 0.1 ms and 32 MB of peak memory, because it reuses features that are already stored. UnifyGeo is timed with its default re-ranking over $k{=}5$ candidates.

C Additional Results

C.1 Extended Retrieval Results

Table 8 extends Table 3 with R@10 and R@1%. Since LAR only re-orders the top-K candidates, R@1% is unchanged by re-ranking, and R@10 is unchanged for $K{=}10$. The gains of LAR are concentrated at R@1 and Hit Rate, which are the metrics that decide which tile the end-to-end system localizes in.

Table 8: **Complete cross-view retrieval results**, extending Table 3 with R@10 and R@1%. Best in **bold**.

Dataset	Method	R@1	R@5	R@10	R@1%	Hit
VIGOR-S	Sample4Geo	77.86	95.66	97.21	99.61	89.82
	AuxGeo	80.34	96.25	**97.57**	**99.67**	93.78
	UnifyGeo	**82.80**	94.92	96.57	99.47	89.12
	GeoSync	79.11	94.73	96.49	99.52	92.26
	+ LAR (K=10)	80.65	95.84	96.49	99.52	94.87
	+ LAR (K=20)	80.87	**96.44**	97.17	99.52	**95.09**
VIGOR-C	Sample4Geo	61.70	83.50	88.00	**98.17**	69.87
	AuxGeo	63.94	84.98	88.98	98.02	76.25
	UnifyGeo	67.58	81.60	86.02	97.26	72.27
	GeoSync	63.35	83.73	87.71	97.24	75.15
	+ LAR (K=10)	68.72	86.47	87.71	97.24	83.16
	+ LAR (K=20)	**69.09**	**88.02**	**89.76**	97.24	**83.60**
DReSS-S	TransGeo	19.86	40.29	49.39	98.60	–
	FRGeo	27.19	46.17	53.82	98.01	–
	Sample4Geo	51.40	72.78	78.52	98.29	55.89
	AuxGeo	54.70	76.09	81.46	**98.74**	59.20
	GeoSync	55.12	73.79	78.73	98.03	59.02
	+ LAR (K=10)	68.24	77.66	78.73	98.03	72.42
	+ LAR (K=20)	**69.72**	**80.77**	**82.07**	98.03	**73.97**
DReSS-C	TransGeo	4.09	11.45	16.26	77.86	–
	FRGeo	8.07	18.23	24.07	82.43	–
	Sample4Geo	28.23	48.19	56.03	92.93	30.65
	AuxGeo	32.44	53.18	60.98	**93.82**	34.90
	GeoSync	34.97	58.24	60.23	92.27	37.17
	+ LAR (K=10)	46.81	58.24	60.23	92.27	49.47
	+ LAR (K=20)	**48.14**	**61.91**	**64.55**	92.27	**50.89**

Table 9: **Interaction between retrieval and fine-grained localization** on DReSS. Each row uses the same trained model for both evaluations. Retrieval is evaluated on the standard candidate set, while localization is evaluated with the ground-truth satellite reference to isolate within-tile accuracy. LAR only changes the candidate ranking, so its localization results are unchanged. Best in **bold**.

	Retrieval				*Fine-grained localization*			
	Same Area		Cross Area		Same Area		Cross Area	
Training variant	R@1↑	Hit↑	R@1↑	Hit↑	Mean↓	Med.↓	Mean↓	Med.↓
Retrieval-only	52.73	57.58	31.01	33.74	33.67	33.85	32.31	32.44
Localization-only	1.70	1.70	1.21	1.22	**3.52**	**1.47**	**6.08**	**2.44**
GeoSync (joint)	55.12	59.02	34.97	37.17	3.53	1.58	6.23	2.54
+ LAR (K=10)	68.24	72.42	46.81	49.47	3.53	1.58	6.23	2.54
+ LAR (K=20)	**69.72**	**73.97**	**48.14**	**50.89**	3.53	1.58	6.23	2.54

C.2 Mutual Reinforcement on DReSS

Table 9 repeats the analysis of Table 5 on DReSS. The pattern holds. Joint training improves retrieval over the retrieval-only model in both settings, raising R@1 from 52.73 to 55.12 (same-area) and from 31.01 to 34.97 (cross-area), and here Hit Rate improves as well. Localization stays close to the localization-only model, within 0.15 m in mean error and about 0.1 m in median error. LAR then adds large retrieval gains on top of the jointly trained model, most notably +13.2 R@1 on DReSS cross-area with K=20. A retrieval-only model cannot localize: its errors of about 33 m correspond to a guess near the tile center.

Table 10: **Two levels of confidence.** Predictions are ranked by confidence and only the most confident fraction is kept. *Retrieval level:* all predictions ranked by the verification score v. *Location level:* predictions with a correctly retrieved tile ranked by MCG sharpness c^{loc}, so the 100% median differs slightly from Table 4. Results use LAR with $K{=}10$.

	Retrieval level (ranked by v) R@1 (%) ↑					Location level (ranked by c^{loc}) Median error (m) ↓				
Kept (%)	100	50	20	10	5	100	50	20	10	5
VIGOR-S	80.65	87.83	92.65	95.68	96.50	1.15	1.06	1.04	1.04	1.01
VIGOR-C	68.72	79.30	85.51	90.11	93.22	1.38	1.28	1.25	1.25	1.21
DReSS-S	68.24	90.50	95.42	96.80	97.85	1.45	1.30	1.26	1.26	1.26
DReSS-C	46.81	70.64	83.48	87.63	90.16	2.16	1.92	1.83	1.83	1.76

C.3 Two Levels of Confidence

GeoSync gives confidence at two levels (Section 3.4): the verification score v tells how confident the model is that the retrieved tile is correct, and the sharpness of the MCG map c^{loc} tells how confident it is about the position inside that tile. Table 10 tests both in the same way. We rank the predictions by confidence and keep only the most confident fraction of them, from all predictions (100%) down to the top 5%. If a confidence signal is useful, the kept predictions should become more accurate as the fraction shrinks.

Both levels behave this way on all four protocols. At the retrieval level, keeping only the queries with high v steadily raises R@1, and the effect is largest under geographic shift: on DReSS cross-area, R@1 rises from 46.81% to 70.64% when only the most confident half of the queries is kept. At the location level, we rank only the queries whose retrieved tile is correct, so that this level reflects within-tile precision alone. Here, higher sharpness steadily lowers the median error, for example from 1.38 m to 1.21 m on VIGOR cross-area. In short, the verification score is the main signal for rejecting wrong tiles, while sharpness separates precise positions from less certain ones within a correct tile.

C.4 Complete Localization Ablation

Tables 11 and 12 extend Table 6 with the full recall curve on VIGOR and DReSS. The two reference protocols differ only in which tile is scored: *Primary* uses the best-centered tile, while *Random* draws uniformly among the four tiles that cover the query, which reflects what a retrieval-driven pipeline actually faces.

Three observations stand out. First, SMRR improves every metric on both datasets and both protocols, so refinement remains useful when the query is off-center. Second, removing semi-positive supervision (with the consistency term that depends on it) slightly helps on centered tiles, but on VIGOR it causes a collapse under *Random* (median 1.66 m → 8.68 m cross-area). This is the expected trade-off: a model that only sees centered tiles learns to expect the query near the center. Third, the same removal hurts much less on DReSS (median 2.66 m → 2.98 m cross-area). DReSS references are already decentered by construction, so even its primary tiles expose the model to off-center queries, which reduces the need for explicit semi-positive supervision.

C.5 Center Bias and Confidence on DReSS

Figure 4 repeats the analysis of Figure 2 on DReSS cross-area.

The left panel shows localization error as a function of the distance between the true location and the tile center. The pattern matches VIGOR: GeoSync's error stays nearly flat toward the tile edge, while HC-Net and FG^2 degrade sharply. Since DReSS spans eight cities on several continents, this confirms that the robustness gained from semi-positive supervision and the consistency term (Appendix A.3) is not specific to VIGOR's cities. The right panel shows the retrieval-level confidence: keeping only the queries with the highest verification score steadily raises both R@1 and Hit Rate, as on VIGOR. UnifyGeo is not included because no DReSS checkpoints are available.

Table 11: **Complete localization ablation on VIGOR.** Each configuration adds one component to the one above it, except the last, which removes semi-positive supervision from + *SMRR*; the cross-tile consistency loss goes with it, since it is defined on the semi-positive branch. *Primary* always scores the best-centered tile, while *Random* draws uniformly over the four tiles covering the query. This table extends Table 6 with the full recall curve.

	Same Area						Cross Area					
	Error (m)↓		Recall (%)↑				Error (m)↓		Recall (%)↑			
Configuration	Mean	Med.	1 m	3 m	5 m	10 m	Mean	Med.	1 m	3 m	5 m	10 m
Primary reference (best-centered tile)												
MCG	2.39	1.41	34.01	79.89	90.68	96.95	3.15	1.80	22.57	73.73	86.97	94.35
+ SMRR	2.12	1.17	42.36	84.24	92.84	97.38	2.80	1.47	30.48	81.65	90.93	95.36
w/o semi-positives	**1.92**	**1.15**	**43.14**	**84.79**	**93.32**	**98.00**	**2.27**	**1.41**	**32.26**	**83.62**	**92.42**	**96.80**
Random reference (uniform over the four covering tiles)												
MCG	3.18	1.70	27.67	71.84	85.81	95.02	4.03	2.01	19.60	68.18	83.24	91.86
+ SMRR	**2.69**	**1.32**	**37.09**	**80.60**	**90.62**	**96.13**	**3.62**	**1.66**	**25.81**	**76.85**	**88.07**	**93.54**
w/o semi-positives	17.36	11.48	12.81	30.39	37.98	48.08	15.96	8.68	10.85	32.68	41.40	52.21

Table 12: **Complete localization ablation on DReSS.** Same configurations, metrics, and reference protocols as Table 11. The cross-area split is cross-continental, so absolute errors are larger throughout, while the ordering of the components is unchanged.

	Same Area						Cross Area					
	Error (m)↓		Recall (%)↑				Error (m)↓		Recall (%)↑			
Configuration	Mean	Med.	1 m	3 m	5 m	10 m	Mean	Med.	1 m	3 m	5 m	10 m
Primary reference (best-centered tile)												
MCG	4.55	2.13	18.90	64.52	80.23	90.81	7.31	3.35	9.59	45.21	65.42	81.92
+ SMRR	3.53	1.58	28.28	**77.86**	**88.89**	**94.39**	6.23	2.54	**14.06**	57.68	76.50	87.48
w/o semi-positives	**3.52**	**1.57**	**28.39**	77.84	88.79	94.29	**6.10**	**2.52**	14.03	**58.14**	**76.76**	**87.49**
Random reference (uniform over the four covering tiles)												
MCG	5.11	2.28	17.70	61.26	77.86	89.21	8.05	3.57	8.98	42.60	62.78	79.78
+ SMRR	**4.00**	**1.64**	**26.86**	**75.69**	**87.21**	**93.23**	**6.94**	**2.66**	**13.25**	**55.49**	**74.45**	**85.74**
w/o semi-positives	5.60	1.85	24.10	68.68	81.14	89.27	8.50	2.98	11.94	50.36	68.53	81.17

D Qualitative Results

D.1 Localization Gallery

Figures 5 and 6 show a broader set of queries from VIGOR and DReSS cross-area, arranged in four rows from the easiest cases to failures.

In the top row, the model is confident and correct: the coarse prediction, the refined prediction, and the ground truth overlap at sub-meter precision. The second row shows correct tiles where the final position is off by a few meters, usually because of local ambiguity such as a repeated facade or a long straight road. The third row contains borderline cases, where the prediction is still inside the correct tile but tens of meters away, often in open or repetitive areas such as highways, parking lots, or riverbanks that offer little to anchor on.

The bottom row shows outright failures, where retrieval selects a wrong tile. We sample these across the full range of severity rather than showing only the worst ones. Some land nearby, often on a visually similar block one or two streets away, while others land on a distant tile that happens to look similar from above. On VIGOR, the cross-area test set spans two cities, so the most severe failures are separated by roughly the distance between them. Across both datasets, the failures are understandable ones: parking lots, straight highway segments, and repetitive residential blocks are exactly the scenes where the ground view carries the least distinctive information.

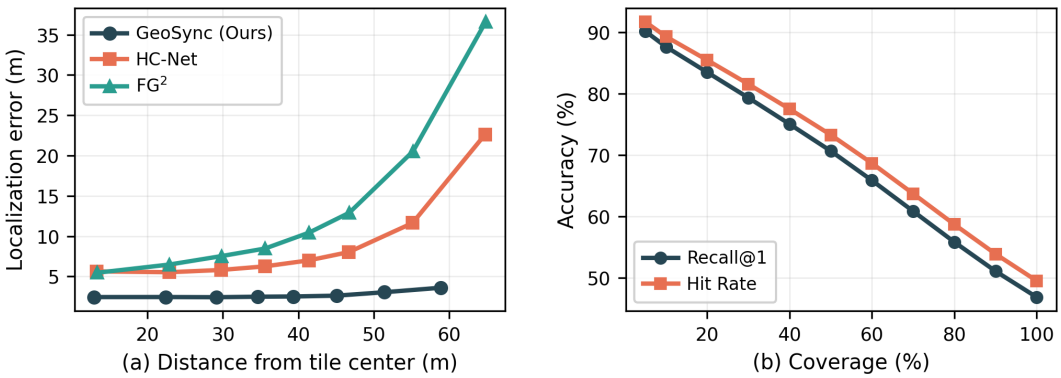

Figure 4: **Center bias and retrieval-level confidence on DReSS cross-area.** (a) Median localization error as a function of the distance from the satellite tile center. GeoSync stays close to flat across the tile, while HC-Net and FG^2 grow sharply once the query moves away from the center. (b) R@1 and Hit Rate when only the most confident fraction of queries is kept, ranked by the verification score v. Both improve steadily as fewer queries are kept.

D.2 How LAR Recovers Correct Tiles

Figures 7 and 8 show what LAR does when it re-ranks the top-10 candidates. Each row shows one query's candidates in their original retrieval order, with the correct tile outlined in green and the tile selected by LAR outlined in red.

For each dataset, we show three queries whose correct tile starts at different ranks: at rank 10, the lowest possible with $K{=}10$; at a middle rank; and just below the top. In every case, the tiles ranked above the correct one are visually plausible, with similar rooftops, street layouts, and block structure. The correct tile only stands out once each candidate is localized and the predicted position is checked against the query with the verification score v (Section 3.4). This is the intuition behind LAR: global similarity is a good filter but a weak final judge, because it cannot see the fine-grained agreement that localization reveals.

D.3 From Coarse Correlation to Refinement

Figures 9 and 10 zoom into the two localization stages of Appendix A.4. The top row shows the full tile with the MCG probability map, and the bottom row zooms into the region around the coarse and refined points.

We deliberately select queries where the coarse estimate is several meters off, so that the effect of refinement is visible. In these cases, SMRR moves the prediction most of the way toward the ground truth, often reducing the error by a factor of three or more (for example, from 8.7 m to 1.9 m, or from 12.8 m to 3.6 m). This reflects the division of labor between the two stages: MCG uses the context of the whole tile to find the right region, and SMRR uses local detail that the pooled coarse map cannot represent to close the remaining gap. The average effect of SMRR over all queries is reported in Table 11.

LLM Usage Statement

We used a large language model (LLM) only to polish the writing of this paper, limited to grammar correction, wording refinement, and improving readability. The LLM played no role in designing the method, implementing or running the experiments, analyzing the results, or producing any scientific content. All technical ideas, experiments, and conclusions are entirely the authors' own, and the LLM served purely as a language editing tool.



Figure 5: **Qualitative localization on VIGOR cross-area**, from confident and precise (top row) to failures (bottom row). Cyan circles mark the coarse MCG prediction, red stars the final SMRR prediction, and green crosses the ground truth (shown only when the selected tile covers the query). Reported errors are end-to-end errors against the query's true GPS position.



Figure 6: **Qualitative localization on DReSS cross-area**, with the same layout as Figure 5. The failure row shows more geographic diversity than on VIGOR, since the DReSS cross-area test set spans four cities on different continents.

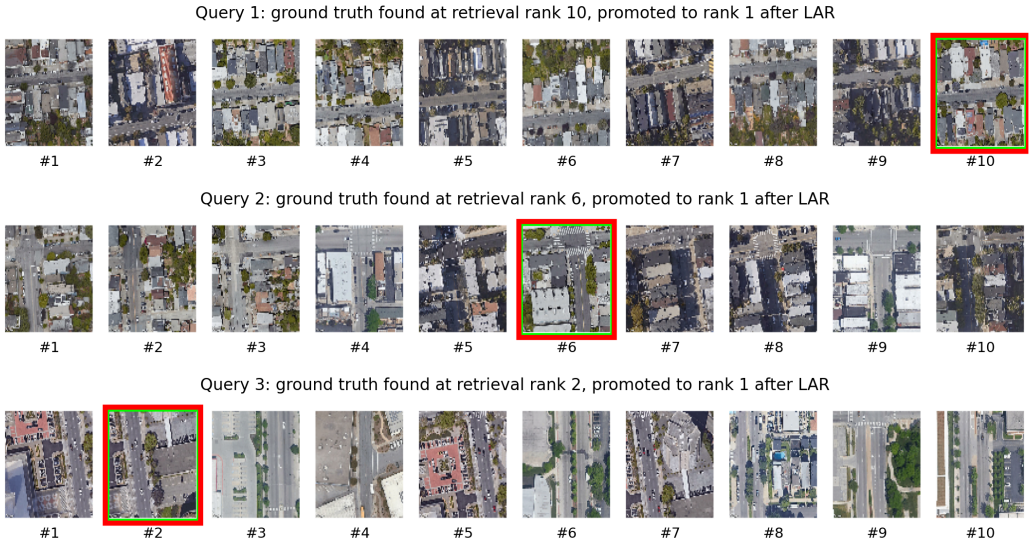

Figure 7: **LAR re-ranking on three VIGOR cross-area queries.** Each row shows the top-10 retrieved tiles in their original order. The correct tile is outlined in green and the tile selected by LAR in red; when the two coincide, LAR has recovered the correct tile from a lower initial rank.

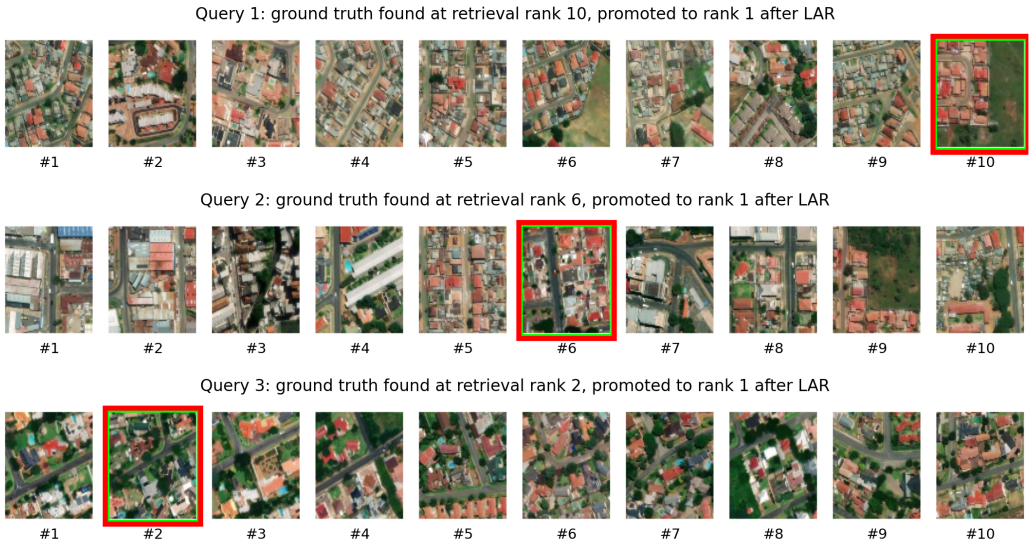

Figure 8: **LAR re-ranking on three DReSS cross-area queries**, ranging from a near miss to the lowest possible starting rank.



Figure 9: **Coarse-to-fine localization on four VIGOR cross-area queries.** Top: the full satellite tile with the MCG probability map and the coarse prediction. Bottom: a zoomed crop showing the coarse point, the SMRR point, and the ground truth at a scale where the refinement is visible.



Figure 10: **Coarse-to-fine localization on four DReSS cross-area queries**, with the same layout as Figure 9.