AEROSCR: RECOVERING STABLE SCENE-COORDINATE DECODABILITY UNDER SCALE AND ROTATION SHIFTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent scene coordinate regression (SCR) methods have narrowed the robustness gap to structure-based localization through stronger local representations and improved generalization. Yet SCR remains vulnerable when query observations differ substantially from the mapping conditions. We investigate this gap in near-nadir aerial relocalization, where altitude and heading changes produce large shifts in image scale and in-plane orientation. Controlled experiments show that descriptor variation can preserve relative correspondence discrimination while destabilizing scene-coordinate predictions, revealing different robustness requirements for correspondence retrieval and coordinate decoding. Motivated by this distinction, AeroSCR expands decoder training coverage across scale changes, canonicalizes large heading changes in descriptor space, and learns tolerance to residual yaw errors. Direct metric supervision and retrieval-gated GeoCells further adapt SCR to weak aerial geometry and large-area maps. Across six scenes from AnyVisLoc, AeroSCR reaches 90.8% success at 5 m/5°, essentially matching HLoc at 90.2%, with 4.0× less scene-specific storage and 8.2× lower mean online latency. On the roughly 10 km^2 Oslo-Large scene, it achieves 86.2% success with 6.0× less storage and 3.1× lower latency than HLoc.

1 INTRODUCTION

Accurate alignment of aerial images with a pre-built metric map enables georeferenced perception, target geolocation, and map-based navigation and planning in UAV applications (Burri et al., 2015; Wu et al., 2024; Cheng et al., 2026). With sufficiently accurate GNSS/INS measurements and calibrated sensor geometry, aerial imagery can be directly georeferenced (Colomina & Molina, 2014); visual relocalization provides a complementary map-relative constraint when GNSS positioning is unavailable, degraded, or otherwise insufficient (Chang et al., 2023; Wu et al., 2024; Cheng et al., 2026). Structure-based localization remains a widely adopted paradigm for accurate visual relocalization, typically estimating query poses by matching local features against an explicit 3D reference map (Sarlin et al., 2019). This explicit map representation and correspondence-based inference introduce scene-specific storage and online matching overhead. Despite recent efforts to reduce these costs (Laskar et al., 2024; Sattler & Kukelova, 2026; Omama et al., 2026), these methods still rely on scene-specific reference features and online correspondence search. Scene coordinate regression (SCR) offers a compact alternative by implicitly encoding the scene in a scene-specific decoder that maps local image features directly to metric 3D coordinates, rather than relying on an explicit feature map and online correspondence search (Wang et al., 2024).

SCR enables compact scene-specific representations and efficient inference, but an accuracy and robustness gap to structure-based localization remains. Recent SCR methods have narrowed this gap through large-scale query pre-training and stronger local representations, including descriptors originally developed for feature matching (Bruns et al., 2025; Jiang et al., 2025; Mao et al., 2026). However, compared with structure-based localization, SCR can degrade sharply when query conditions extend beyond those covered by the mapping views. Near-nadir aerial localization presents a representative case of this problem, where differences in altitude and heading induce large changes in apparent scale and in-plane orientation (Wu et al., 2024; Dhaouadi et al., 2025). These observa-

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.

tion shifts motivate a closer examination of the robustness requirements imposed on the same local representation by correspondence search and scene-coordinate decoding.

Figure 1 evaluates an explicit nearest-neighbor feature bank and a point-wise SCR head under controlled scale and yaw shifts using the same frozen DeDoDe-B representation. The SCR head is trained only on the mapping views, without scene-specific scale or yaw augmentation. Under a $\pm 2.5\%$ scale shift, positive-correspondence R@1 remains at 95.5–95.7%, and the explicit NN bank achieves 100% pose success at 5 m/5°, whereas SCR succeeds on only 0–7.1% of queries. A similar pattern appears under yaw variation. At a 10° rotation, R@1 remains 90.7% and the NN bank retains 100% pose success, while SCR pose success drops to 0%. Only at larger rotations does correspondence ranking deteriorate substantially, at which point NN-bank localization begins to fail.

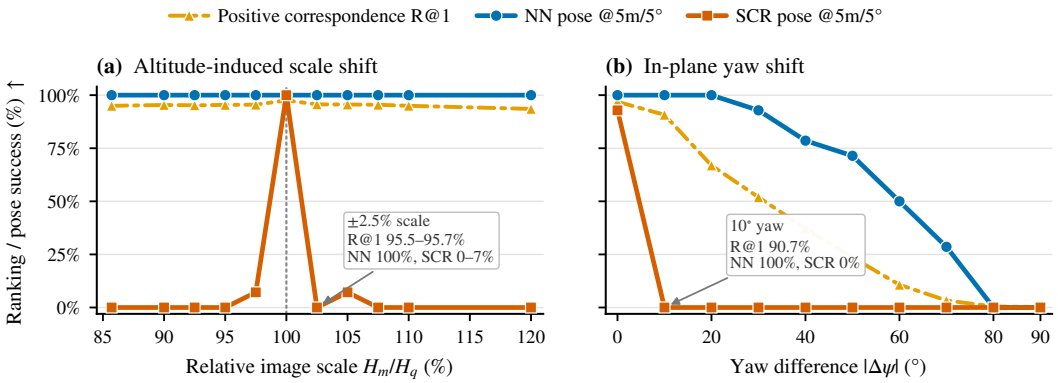

Figure 1: Controlled comparison of relative correspondence discrimination and scene-coordinate decoding under scale and yaw shifts. Both use the same frozen DeDoDe-B representation. Query altitude is varied at matched map locations for the scale probe, while in-plane orientation is varied along fixed projected tracks for the yaw probe. R@1 reports positive-correspondence ranking; NN and SCR pose success are evaluated at 5 m/5°.

These controlled probes isolate a representation-level distinction in how the same local representation is used downstream. Correspondence retrieval relies on relative discrimination, requiring the correct reference to remain more similar than competing candidates. SCR instead requires consistent absolute coordinate decoding. A scene-specific decoder is learned from a finite set of mapping observations. At test time, it must still map descriptors of the same scene point observed from previously unseen viewpoints to the same metric 3D coordinate. Descriptor variation can therefore leave the correct correspondence top-ranked while causing the decoder to predict inconsistent 3D coordinates for the same scene point.

This analysis suggests that robust scene-coordinate decoding does not require descriptors to be fully invariant to observation changes; instead, relevant variation can be covered during decoder training, or transformed descriptors can be aligned toward representations observed during training. Based on this principle, we introduce AeroSCR for near-nadir aerial relocalization. Real-image crop-resize augmentation broadens scale variation during training, while descriptor steering canonicalizes large heading changes and residual training handles the remaining yaw variation. Complementing these components, direct metric supervision stabilizes coordinate learning under weak aerial geometry, while retrieval-gated GeoCells extend the compact representation to large scenes.

Across six AnyVisLoc scenes, AeroSCR essentially matches HLoc at 5 m/5° (90.8% vs. 90.2%) under the orientation-prior protocol, while using 4.0× less storage and 8.2× lower latency. On Oslo-Large, it retains 86.2% success with 6.0× less storage and 3.1× lower latency than HLoc.

Our contributions are threefold:

1. We identify a representation-level distinction between relative correspondence discrimination and stable scene-coordinate decoding through controlled scale and rotation probes.
2. We introduce AeroSCR, combining real-image scale augmentation, descriptor-space yaw canonicalization, and local residual training for robust near-nadir aerial relocalization.

3. We validate AeroSCR on public real imagery, cross-altitude and cross-heading tests, and kilometer-scale scenes, with scene-storage and online-latency comparisons.

2 Related Work

2.1 Explicit Feature Matching and Efficient Map Representations

Structure-based localization stores scene-specific 3D points and local descriptors and matches query features to this map for pose estimation (Sattler et al., 2017; Sarlin et al., 2019). Prior work reduces storage and matching costs through 3D-point pruning, descriptor compression and quantization, learned map reduction, and query-side acceleration (Cao & Snavely, 2014; Camposeco et al., 2019; Yang et al., 2022; Laskar et al., 2024; Sattler & Kukelova, 2026; Omama et al., 2026), but still relies on an explicit feature map and online correspondence search. SCR instead encodes scene geometry in a compact coordinate predictor.

2.2 Scene Coordinate Regression

Scene coordinate regression (SCR) predicts metric 3D coordinates directly from image features. Earlier methods established differentiable pose estimation and improved scalability through expert models and compact scene-specific heads (Brachmann et al., 2017; Brachmann & Rother, 2018; 2019; 2022; Brachmann et al., 2023). ACE-G improves generalization through large-scale query pre-training, R-SCoRe adopts strong pretrained local descriptors such as DeDoDe, and CoLoR further trains the representation for cross-view consistency (Bruns et al., 2025; Jiang et al., 2025; Mao et al., 2026). AeroSCR instead targets scale and heading variation through decoder training coverage and descriptor canonicalization.

2.3 Aerial Visual Localization

Compact localization is valuable for aerial platforms with limited onboard resources and large reference maps (Nitsche et al., 2020; Dhaouadi et al., 2025; Wu et al., 2024). Metric UAV localization commonly registers queries against textured 3D models, rendered views, or synthetic references (Yan et al., 2023; Wu et al., 2024; Cheng et al., 2026), which can incur substantial storage or online rendering costs (Dhaouadi et al., 2025; Cheng et al., 2026). Recent methods explore lighter representations including low-LoD city models, orthographic geodata, and TDOM/DSM maps (Zhu et al., 2024; 2025; Dhaouadi et al., 2025; Liu et al., 2026), at the cost of projection or cross-view alignment challenges. AeroSCR instead uses compact scene-coordinate heads over prior aerial mapping imagery, avoiding online rendering and local query-to-reference matching.

2.4 Rotation-Robust and Equivariant Representations

Near-nadir aerial queries induce large in-plane rotations. Local features address rotation through canonical orientation assignment, learned invariance, or equivariant representations (Lowe, 2004; Liu et al., 2019; Pautrat et al., 2020; Lee et al., 2023), while steerable descriptors model transformations under rotations and more general affine warps (Bökman et al., 2024a;b). These approaches are mainly designed for correspondence; AeroSCR instead uses steering to canonicalize heading variation before scene-coordinate decoding, leaving bounded residual variation to the decoder.

3 Relative Discrimination vs. Absolute Scene-Coordinate Decoding

We formalize the representation-level distinction observed in Fig. 1 by examining the requirements imposed by relative correspondence discrimination and scene-coordinate decoding on the same local representation. Let $d = g(I, u) \in \mathbb{R}^D$ denote a local descriptor extracted at image location u, and let $h_\theta : \mathbb{R}^D \to \mathbb{R}^3$ denote a scene-specific coordinate decoder. We consider observation changes that preserve the identity of the physical point while changing its image appearance and descriptor.

3.1 RELATIVE DISCRIMINATION IN FEATURE MATCHING

Feature matching relies on relative discrimination among candidates. For a nearest-neighbor matcher, let d_q^τ denote a transformed query descriptor, d^+ its correct reference descriptor, and $\mathcal{N}_q$ the competing descriptors. A sufficient condition for retrieving the correct correspondence is

$$\mathrm{sim}(d_q^\tau, d^+) > \max_{d^- \in \mathcal{N}_q} \mathrm{sim}(d_q^\tau, d^-). \tag{1}$$

Descriptor-learning objectives based on positive–negative distances, margins, or ranking encourage relative discrimination (Tian et al., 2017; Mishchuk et al., 2017; He et al., 2018). The descriptor representation therefore need not remain invariant under the observation change; it only needs to preserve sufficient relative discrimination for the correct correspondence. At large transformations, however, this relative discrimination can disappear, causing correspondence to fail.

3.2 STABLE SCENE-COORDINATE DECODABILITY

Let $\mathcal{T}_{\mathrm{map}}$ and $\mathcal{T}_q$ denote the mapping and query observation conditions. For a physical scene point $X^\star$, the corresponding descriptor sets are

$$\mathcal{D}_c(X^\star) = \{ g(I^\tau, u^\tau) \mid \tau \in \mathcal{T}_c \}, \qquad c \in \{\mathrm{map}, q\}. \tag{2}$$

The decoder $h_{\theta_{\mathrm{map}}}$ is learned from the mapping descriptors and then applied unchanged to query descriptors. Stable scene-coordinate decodability over the query conditions requires

$$\sup_{d \in \mathcal{D}_q(X^\star)} \left\| h_{\theta_{\mathrm{map}}}(d) - X^\star \right\|_2 \le \delta_X. \tag{3}$$

Here, δ_X denotes the tolerated scene-coordinate error.

Mapping supervision constrains $h_{\theta_{\mathrm{map}}}$ over $\mathcal{D}_{\mathrm{map}}(X^\star)$, whereas $\mathcal{T}_q$ need not be contained in $\mathcal{T}_{\mathrm{map}}$. Query descriptors outside this observed set may remain sufficiently discriminative for correspondence retrieval but yield unreliable coordinates from the fixed decoder. Across many points, these errors reduce geometric consistency and the inlier support available to the pose solver.

This distinction motivates the design of AeroSCR in Section 4.

4 METHOD

4.1 OVERVIEW

Figure 2 summarizes AeroSCR. Training combines real-image scale augmentation, descriptor-space yaw canonicalization with residual perturbations, and direct metric supervision. At inference, AeroSCR can use coarse orientation cues from the onboard state estimate to reduce the orientation search space. When available, a bounded yaw observation restricts the steering hypotheses, while an approximate gravity direction is supplied to the pose solver. Descriptors are extracted once and reused across yaw hypotheses and, for large scenes, retrieval-selected GeoCell heads for pose estimation.

4.2 COVERAGE-BALANCED REAL-IMAGE SCALE AUGMENTATION

Image-scale augmentation has been used in recent SCR systems to improve robustness to scale variation (Jiang et al., 2025). For near-nadir imagery with similar intrinsics, camera-to-terrain distance largely determines the apparent scale of ground structures. We therefore use crop-resized real mapping images to approximate this dominant scale change. The training scale range is derived from the expected query depth range. Rather than sampling this range uniformly, we allocate the augmentation budget across native, interior, and boundary scales.

Let D_m denote a representative mapping depth and $D_q \in [D_q^{\min}, D_q^{\max}]$ the expected query effective-depth range. Under fixed intrinsics,

$$s \approx \frac{D_m}{D_q}, \qquad [s_{\min}, s_{\max}] = \left[\frac{D_m}{D_q^{\max}}, \frac{D_m}{D_q^{\min}} \right]. \tag{4}$$

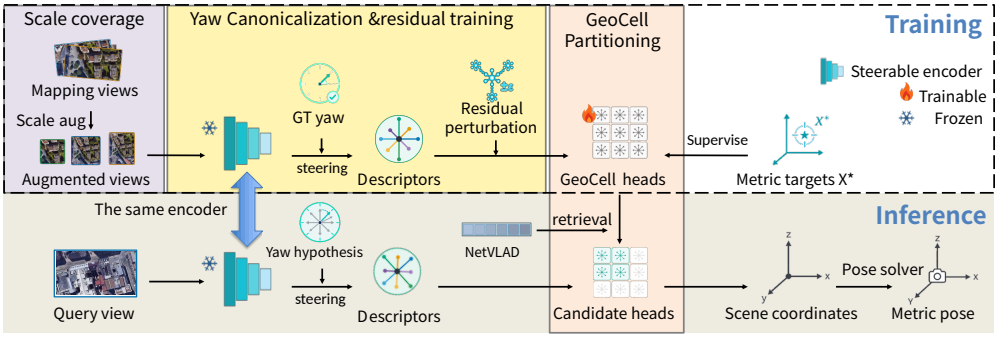

Figure 2: Overview of AeroSCR. Training combines coverage-balanced scale augmentation, yaw canonicalization with local residual training, and geometry-anchored scene-coordinate heads, with GeoCell partitioning used for large scenes. At inference, extracted descriptors are reused across yaw hypotheses and, when applicable, retrieval-selected GeoCell heads for pose estimation.

This interval defines the scale range covered during training and is specified at the scene level, without requiring query depth at inference. Let $C_{s,o}$ denote a crop-resize transformation parameterized by s and o. For a mapped observation (I_i, u_i),

$$d_i^{(s,o)} = g(C_{s,o}(I_i), C_{s,o}(u_i)) \,. \tag{5}$$

The augmented observation retains the same metric scene-coordinate target $X_i^\star$ as the corresponding native mapping observation. It is therefore used as an additional training sample for the shared decoder h_θ, extending supervision beyond the native image scale.

Coverage-balanced sampling. With a fixed augmentation budget, sampling continuously over a broad scale range can leave its extremes sparsely represented. We therefore allocate training views among the native scale, log-uniform interior intervals, and explicit boundary scales. This preserves coverage near the original mapping condition. The expected scale extremes are also explicitly represented during training. The same image-space transformation is applied consistently to the RGB image, camera intrinsics, and pixel-aligned geometric supervision. The exact ranges and quotas are dataset-specific; Appendix B provides the complete schedules and implementation details.

4.3 YAW CANONICALIZATION AND LOCAL RESIDUAL INVARIANCE

Near-nadir heading changes mainly induce in-plane rotations, causing descriptors of the same landmark to vary with observation heading. When mapping views cover only a narrow heading range, the scene-coordinate decoder receives limited supervision for unseen rotations. AeroSCR handles this variation in two stages, canonicalizing large heading changes before decoding and training the decoder to tolerate only the remaining bounded yaw residual.

Large-rotation canonicalization. We use a rotation-equivariant local representation. Descriptors extracted under different headings are steered toward a common canonical orientation before scene-coordinate decoding. Specifically, AeroSCR adopts the released SO2-Spread-B DeDoDe-B descriptor–steerer pair of Bökman et al. (2024b), built on DeDoDe (Edstedt et al., 2024). Both the descriptor and steerer are pretrained under Setting B and kept frozen during AeroSCR training.

For a descriptor d, we write the normalized steering operator as

$$S_\psi(d) = \mathcal{N}(\exp(\psi A)\,\mathcal{N}(d)) \,, \tag{6}$$

where A is the pretrained generator matrix and $\mathcal{N}(\cdot)$ denotes ℓ_2 normalization. The released descriptor–steerer pair approximately satisfies

$$g(R_\psi I, R_\psi u) \approx S_\psi(g(I, u)) \,. \tag{7}$$

This approximate equivariance allows a descriptor observed at a different heading to be rotated in feature space toward the canonical representation seen by the scene-coordinate decoder.

Let $\psi^{\mathrm{gt}}_{i\to c}$ denote the signed rotation from a training observation to the canonical mapping heading, computed from its calibrated camera pose. The canonicalized training descriptor is

$$d_i^c = S_{\psi^{\mathrm{gt}}_{i\to c}}(g(I_i, u_i))\,. \tag{8}$$

This aligns descriptors observed at different training headings toward a shared canonical representation. At inference, the exact query rotation is unknown. We therefore evaluate a set of steering hypotheses before scene-coordinate decoding and pose evaluation; when a bounded yaw observation is available, it is used to restrict this hypothesis set.

Local residual invariance. The yaw hypothesis available at inference is affected by sensor error or discrete hypothesis spacing, leaving a bounded rotation residual after canonicalization. We expose the scene-coordinate decoder to this local residual during training. For each observation, sample

$$\epsilon_i \sim \mathcal{U}[-\epsilon_{\max}, +\epsilon_{\max}], \qquad \tilde{\psi}_{i\to c} = \psi^{\mathrm{gt}}_{i\to c} + \epsilon_i. \tag{9}$$

where $\epsilon_{\max}$ specifies the training residual-yaw range. The perturbed canonical descriptor is

$$d_i^{c,\epsilon_i} = S_{\tilde{\psi}_{i\to c}}(g(I_i, u_i))\,. \tag{10}$$

The perturbed descriptor retains the same metric scene-coordinate target $X_i^\star$. Residual-steered descriptors are therefore included in training with the same coordinate supervision, restricting the decoder's learned yaw tolerance to a local neighborhood around the canonical orientation.

4.4 GEOGRAPHIC CONDITIONING AND GEOMETRIC PROCESSING

Geographic conditioning. In large scenes, visually similar local descriptors may correspond to distant geographic regions. Recent large-scale SCR methods address this ambiguity by conditioning a shared decoder on a global scene representation (Wang et al., 2024; Jiang et al., 2025; Mao et al., 2026). When the global condition is uncertain, multiple conditioning hypotheses may need to be evaluated at inference. For large scenes, AeroSCR instead partitions the aerial map into overlapping GeoCells and uses NetVLAD retrieval (Arandjelovic et al., 2016) to activate only a small set of local scene-coordinate heads. Each selected cell is evaluated under the yaw hypotheses allowed by the orientation prior, replacing broad global-conditioning search with explicit geographic routing. Routing details are provided in Appendix D.2.

Metric coordinate supervision. Near-nadir mapping typically has small baseline-to-height ratios, resulting in narrow triangulation angles and weak depth conditioning (Hartley & Sturm, 1997). We therefore complement reprojection supervision with direct metric coordinate targets, following coordinate-supervised SCR formulations (Brachmann & Rother, 2018; Brachmann et al., 2023):

$$\mathcal{L} = \mathcal{L}_{\mathrm{reproj}} + m_i \lambda_{\mathrm{geo}}\, \rho_{\mathrm{geo}}\left(\|\hat{X}_i - X_i^\star\|_2\right), \tag{11}$$

where m_i indicates whether a reliable metric target $X_i^\star$ is available. Target construction, robust loss, supervision coverage, and scheduling are detailed in Appendix B.

Pose estimation. For each evaluated scene-coordinate head under each yaw hypothesis, the gravity-aware upright two-point (UP2P) minimal solver (Kukelova et al., 2011) is embedded in MSAC (Torr & Zisserman, 2000) to estimate a pose from the predicted 2D–3D correspondences. Each candidate pose is refined with unconstrained 6-DoF optimization. Refined hypotheses are ranked by robust reprojection cost, with geometric support and RANSAC inlier count as secondary criteria. Solver parameters and numerical details are provided in Appendix F.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and splits. We evaluate AeroSCR in four settings. AGL/AMSL, derived from the public UAV-photogrammetry dataset (Agüera-Vega et al., 2024), provides a real-image evaluation with

variations in flight altitude and heading. We further evaluate on six of the 24 scenes in AnyVisLoc (Ye et al., 2026) that meet our eligibility criteria. These criteria cover near-nadir imagery, camera-model consistency, sufficient mapping support, and reconstruction connectivity. In addition, Oslo-Center provides controlled scale and yaw changes at matched locations, while Oslo-Large extends the evaluation to a 3.2×3.2 km region for large-area retrieval-gated localization. Scene-specific SCR training uses mapping observations only.

Table 1: Mapping and query observations used in the main evaluation. N_m and N_q denote the total numbers of mapping and query images, respectively, in each evaluation setting. For AnyVisLoc, the counts are summed over the six scenes satisfying the eligibility criteria described in Sec. 5.1.

	Mapping		Query	
Dataset	N_m	Observations	N_q	Observations
AGL/AMSL	44	AGL flight, 9.3° heading span	206	Cross flights, 294.5° heading span
AnyVisLoc	199	Near-nadir UAV views	974	Natural altitude / heading variation
Oslo-Center	65	600 m, co-oriented	370	300–1000 m, fixed / random heading
Oslo-Large	697	600 m, co-oriented	4,450	300–1000 m, fixed / random heading

Baselines and efficiency. We compare with HLoc (Sarlin et al., 2019) using SuperPoint (DeTone et al., 2018) and LightGlue (Lindenberger et al., 2023), and the SCR baselines ACE-G (Bruns et al., 2025), GLACE (Wang et al., 2024), and R-SCoRe (Jiang et al., 2025). All SCR methods use metric scene-coordinate targets from the same mapping-only mesh; native baselines retain their original models, losses, and augmentation settings. R-SCoRe+ applies AeroSCR's scale/yaw treatment, metric supervision, and pose estimation to R-SCoRe's shared global-conditioned decoder; details are provided in Appendix D.4. Efficiency follows the unified protocol in Appendix E.

Evaluation metrics. Localization succeeds when both translation and rotation errors fall below the stated thresholds. Main comparisons report localization success, scene-specific storage, and online latency, with median translation and rotation errors included where appropriate.

Orientation-prior protocol. Unless otherwise stated, random-heading experiments use perturbed orientation observations with $\pm 60^\circ$ yaw and 10° gravity uncertainty. HLoc uses the yaw observation, while AeroSCR and R-SCoRe+ use the corresponding yaw and gravity observations; method-specific hypothesis construction is detailed in Appendix D.1. ACE-G, GLACE, and native R-SCoRe retain their original inference pipelines.

5.2 AGL/AMSL Evaluation

Table 2: Public AGL/AMSL evaluation on 206 real aerial queries. Methods in the orientation-prior setting use the perturbed orientation observations defined in Section 5.1. Detailed results by yaw gap are reported in Appendix D.1.

Setting	Method	Success (%) ↑		Median error ↓		Efficiency	
		@5m/5°	@1m/1°	t (m)	r (°)	Mean (ms) ↓	Map (MB) ↓
Native inference	HLoc	13.6	**13.6**	115.78	127.19	852	47.9
	ACE-G	**21.8**	3.4	**33.06**	**39.49**	**60**	12.6
	GLACE	13.1	11.2	93.78	107.21	84	**9.5**
	R-SCoRe	13.6	13.6	128.16	121.95	794	43.0
Orientation prior	HLoc	**100**	**100**	**0.06**	**0.05**	1748	48.3
	R-SCoRe+	**100**	99.0	0.12	0.11	1145	42.8
	AeroSCR	**100**	99.5	0.15	0.15	**118**	**10.6**

Table 2 shows that, under native inference, all baselines remain below 22% success at 5 m/5°. With the orientation prior, HLoc, R-SCoRe+, and AeroSCR all recover near-saturated localization perfor-

mance, reaching 100% at 5 m/5° and at least 99.0% at 1 m/1°. HLoc yields the most precise poses, while AeroSCR retains 99.5% success at 1 m/1° with substantially lower scene storage and online latency: 10.6 MB and 118 ms, compared with 48.3 MB and 1.75 s for HLoc.

5.3 AnyVisLoc Evaluation

Table 3: AnyVisLoc evaluation across six real UAV scenes. Accuracy and scene-specific storage are computed per scene and averaged equally across the six scenes, while mean online latency is computed over all 974 evaluation queries pooled across scenes.

Setting	Method	Success (%) ↑			Efficiency	
		@5m/5°	@2m/2°	@1m/1°	Mean (ms) ↓	Avg. Map (MB) ↓
Native inference	HLoc	**78.2**	**75.3**	**69.0**	342	42.5
	ACE-G	21.5	6.9	1.9	53	12.6
	GLACE	22.5	9.8	4.1	**49**	**9.5**
	R-SCoRe	68.3	55.0	36.3	695	42.8
Orientation prior	HLoc	90.2	**87.1**	**80.3**	905	42.5
	R-SCoRe+	86.4	67.4	41.7	1222	42.6
	AeroSCR	**90.8**	79.9	60.4	**111**	**10.6**

Table 3 shows that AeroSCR leads the evaluated SCR methods at all three thresholds. Under the orientation prior, it reaches 90.8% at 5 m/5°, essentially matching HLoc at 90.2%, while HLoc remains stronger at tighter thresholds. AeroSCR reduces latency from 905 ms to 111 ms and storage from 42.5 MB to 10.6 MB relative to HLoc; R-SCoRe+ improves over R-SCoRe but remains below AeroSCR.

5.4 Oslo-Center: Controlled Scale and Yaw Shifts

Table 4 isolates scale and heading shifts on Oslo-Center. AeroSCR remains robust under fixed and random heading, with its largest gains over SCR baselines at the tighter threshold. HLoc remains more accurate, while AeroSCR provides comparable 5 m/5° robustness with lower storage and latency.

Table 4: Localization results on Oslo-Center and Oslo-Large. Fixed-yaw results isolate altitude and geographic scaling effects, while random-yaw results evaluate heading variation under the Section 5.1 orientation-prior protocol.

Dataset	Setting	Method	Success (%) ↑		Median error ↓		Efficiency	
			@5m/5°	@2m/2°	t (m)	r (°)	Mean (ms) ↓	Map (MB) ↓
Oslo-Center	Fixed yaw	HLoc	**100**	**97.6**	**1.01**	**0.07**	991	105.2
		ACE-G	11.9	1.1	17.97	1.80	**46**	12.6
		GLACE	52.7	11.9	4.83	0.46	53	**9.5**
		R-SCoRe	92.7	63.2	1.63	0.14	491	43.3
		R-SCoRe+	98.1	55.1	1.90	0.19	452	43.2
		AeroSCR	99.7	78.1	1.49	0.13	124	18.6
	Random yaw + prior	HLoc	**98.7**	**87.0**	**1.19**	**0.09**	2329	105.2
		R-SCoRe+	92.2	40.8	2.30	0.24	1204	43.2
		AeroSCR	97.8	59.7	1.78	0.18	**129**	**18.6**
Oslo-Large	Fixed yaw	HLoc	**98.2**	**88.7**	**1.16**	**0.09**	371	1052.8
		AeroSCR	92.8	54.3	1.87	0.19	**82**	**176.1**
	Random yaw + prior	HLoc	**90.5**	**76.3**	**1.30**	**0.11**	904	1052.8
		AeroSCR	86.2	43.2	2.23	0.23	**289**	**176.1**

5.5 OSLO-LARGE: LARGE-SCALE LOCALIZATION

On Oslo-Large, Table 4 shows that AeroSCR reaches 86.2% at 5 m/5° versus 90.5% for HLoc under random heading, while using 6.0× less scene-specific storage and 3.1× lower mean online latency. Retrieval recalls the query's core cell for 99.1% and 95.0% of fixed- and random-heading queries while activating only 3.29 and 3.89 of nine heads on average, respectively; routing details are provided in Appendix D.2. HLoc remains stronger at tighter localization thresholds. The R-SCoRe+ control does not reach usable convergence at this scale under the evaluated shared global-conditioned configuration; diagnostics are provided in Appendix D.4.

5.6 COMPONENT ABLATIONS AND SENSITIVITY

Scale augmentation and metric supervision. Table 5(a) shows that scale augmentation drives the cross-altitude gain, raising 5 m/5° success from 67.6% to 99.7%; removing metric supervision lowers 2 m/2° success from 78.1% to 69.5%. Figure 3 shows similar aggregate accuracy for matched i.i.d. LogUniform and coverage-balanced sampling, but the latter gains 20.8 and 9.5 points at 900 and 1000 m and is used by default.

Yaw canonicalization and residual training. C_8 denotes eight-way discrete steering; Continuous (GT) uses the pose-derived continuous angle. Table 5(b) shows that canonicalization drives the large-heading gain: C_8 and Continuous (GT) raise 5 m/5° success from 11.4% to 90.8% and 98.4%. Residual training narrows the gap between discrete C_8 and Continuous (GT), raising C_8 to 97.8% at 5 m/5° and 53.5% at 2 m/2°. Appendix D.3 analyzes residual-yaw sensitivity; Appendix D.1 shows that wider yaw search increases computation without systematic accuracy loss.

Table 5: Component ablations on Oslo-Center.

(a) Scale aug. & metric sup.

Scale aug.	Metric sup.	@5m/5°	@2m/2°
×	✓	67.6	31.9
✓	×	98.1	69.5
✓	✓	**99.7**	**78.1**

(b) Yaw canonicalization & residual

Steering	Residual	@5m/5°	@2m/2°
None	×	11.4	6.5
None	✓	15.9	9.5
C_8	×	90.8	45.1
C_8	✓	97.8	53.5
Continuous (GT)	×	98.4	56.8
Continuous (GT)	✓	**98.6**	**61.6**

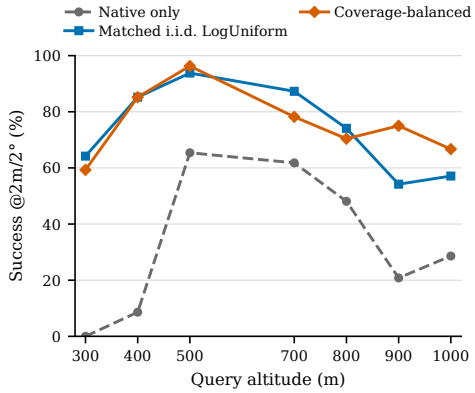

Figure 3. Scale robustness across altitude on Oslo-Center under matched augmentation budgets and scale ranges.

6 CONCLUSION

We show that descriptor variation that preserves relative correspondence discrimination does not necessarily preserve stable scene-coordinate decodability. AeroSCR addresses this gap by expanding decoder training coverage across scales and aligning heading-induced descriptor variation to a locally tolerant canonical representation, while retaining compact inference. Across six AnyVisLoc scenes, AeroSCR essentially matches HLoc at 5 m/5° under the orientation-prior protocol while using 4.0× less storage and 8.2× lower mean online latency. On the roughly 10 km^2 Oslo-Large scene, it retains 86.2% success with 6.0× less storage and 3.1× lower latency than HLoc.

Limitations and Future Work. The current study focuses on near-nadir imagery and calibrated mapping and query imagery with matched camera intrinsics, and uses an approximate gravity direction for pose initialization. Future work will extend AeroSCR to cross-camera and focal-length variation, strongly oblique imagery, and reduced dependence on external orientation cues.

AI USE STATEMENT

In this work, we used generative AI tools to assist with research-methodology and experiment-design feedback, implementation and debugging, result interpretation, and manuscript writing and revision. We additionally used generative AI tools to edit scientific figures, search and organize literature, and format citations and tables. All AI-assisted outputs were reviewed by the authors, and AI-generated suggestions were accepted only after checking them against the project source, experimental artifacts, or cited papers, as applicable. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 defines the method, and Section 5.1 specifies the datasets, splits, baselines, metrics, orientation-prior protocol, and latency setup. The appendix documents training and augmentation settings, additional analyses, latency and storage accounting, and pose-solver parameters in Appendices B, D, E, and F. We plan to include anonymized configuration and evaluation scripts with the supplementary material.

REFERENCES

Francisco Agüera-Vega, Ezequiel Ferrer-González, Patricio Martínez-Carricondo, Julián Sánchez-Hermosilla, and Fernando Carvajal-Ramírez. Influence of the inclusion of off-nadir images on UAV-photogrammetry projects from nadir images and AGL (above ground level) or AMSL (above mean sea level) flights. *Drones*, 8(11):662, 2024. doi: 10.3390/drones8110662. URL https://doi.org/10.3390/drones8110662.

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5297–5307, 2016. doi: 10.1109/CVPR.2016.572.

Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Affine steerers for structured keypoint description. In *Computer Vision – ECCV 2024*, volume 15144 of *Lecture Notes in Computer Science*, pp. 449–468. Springer Nature Switzerland, 2024a. doi: 10.1007/978-3-031-73016-0_26.

Georg Bökman, Johan Edstedt, Michael Felsberg, and Fredrik Kahl. Steerers: A framework for rotation equivariant keypoint descriptors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4885–4895, June 2024b.

Eric Brachmann and Carsten Rother. Learning less is more – 6d camera localization via 3d surface regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4654–4662, 2018.

Eric Brachmann and Carsten Rother. Expert sample consensus applied to camera re-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7525–7534, 2019.

Eric Brachmann and Carsten Rother. Visual camera re-localization from RGB and RGB-D images using DSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5847–5865, 2022. doi: 10.1109/TPAMI.2021.3070754.

Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. DSAC – differentiable RANSAC for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6684–6692, 2017.

Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using RGB and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5044–5053, 2023.

Leonard Bruns, Axel Barroso-Laguna, Tommaso Cavallari, Aron Monszpart, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. ACE-G: Improving generalization of scene coordinate regression through query pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26751–26761, 2025.

Michael Burri, Helen Oleynikova, Markus W. Achtelik, and Roland Siegwart. Real-time visual-inertial mapping, re-localization and planning onboard MAVs in unknown environments. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1872–1878. IEEE, 2015. doi: 10.1109/IROS.2015.7353622.

Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid scene compression for visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7653–7662, 2019.

Song Cao and Noah Snavely. Minimal scene descriptions from structure from motion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 461–468, 2014.

Yingxiu Chang, Yongqiang Cheng, Umar Manzoor, and John Murray. A review of UAV autonomous navigation in GPS-denied environments. *Robotics and Autonomous Systems*, 170:104533, 2023. doi: 10.1016/j.robot.2023.104533.

Xiaoya Cheng, Long Wang, Yan Liu, Xinyi Liu, Hanlin Tan, Yu Liu, Maojun Zhang, and Shen Yan. PiLoT: Neural pixel-to-3d registration for UAV-based ego and target geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5379–5388, 2026.

Ismael Colomina and Pere Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:79–97, 2014. doi: 10.1016/j.isprsjprs.2014.02.013.

Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 224–236, 2018. doi: 10.1109/CVPRW.2018.00060.

Oussema Dhaouadi, Riccardo Marin, Johannes Meier, Jacques Kaiser, and Daniel Cremers. OrthoLoC: UAV 6-dof localization and calibration using orthographic geodata. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1106. Datasets and Benchmarks Track.

Johan Edstedt, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. DeDoDe: Detect, don't describe – describe, don't detect for local feature matching. In *2024 International Conference on 3D Vision*, pp. 148–157. IEEE, 2024. doi: 10.1109/3DV62453.2024.00035.

Richard I. Hartley and Peter Sturm. Triangulation. *Computer Vision and Image Understanding*, 68 (2):146–157, 1997. doi: 10.1006/cviu.1997.0547.

Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 596–605, June 2018.

Xudong Jiang, Fangjinhua Wang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. R-SCoRe: Revisiting scene coordinate regression for robust large-scale visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11546, 2025.

Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla. Closed-form solutions to minimal absolute pose problems with known vertical direction. In *Computer Vision – ACCV 2010*, volume 6493 of *Lecture Notes in Computer Science*, pp. 216–229. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19309-5_17.

Viktor Larsson and contributors. PoseLib – minimal solvers for camera pose estimation, 2020. URL `https://github.com/vlarsson/PoseLib`.

Zakaria Laskar, Iaroslav Melekhov, Assia Benbihi, Shuzhe Wang, and Juho Kannala. Differentiable product quantization for memory-efficient camera relocalization. In *Computer Vision – ECCV 2024*, volume 15143 of *Lecture Notes in Computer Science*, pp. 470–489. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-73013-9_27.

Jongmin Lee, Byungjin Kim, Seungwook Kim, and Minsu Cho. Learning rotation-equivariant features for visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21887–21897, 2023.

Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17627–17638, 2023. doi: 10.1109/ICCV51070.2023.01616.

Xinyi Liu, Xiaoya Cheng, Rouwan Wu, Zhaochen Wang, Shen Yan, Maojun Zhang, and Yu Liu. PiLoT v2: Pixel-to-orthogonal map alignment for free-view UAV geo-localization. *arXiv preprint arXiv:2606.31098*, 2026. doi: 10.48550/arXiv.2606.31098.

Yuan Liu, Zehong Shen, Zhixuan Lin, Sida Peng, Hujun Bao, and Xiaowei Zhou. GIFT: Learning transformation-invariant dense visual descriptors via group CNNs. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6990–7001, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94.

Xindong Mao, Hang Li, Yuchen Wu, Jiahe Li, Xiao Bai, and Jin Zheng. CoLoR: The devil is in scene coordinate regression for large-scale visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12206–12216, 2026.

Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenović, and Jiří Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4826–4837, 2017.

Matías Nitsche, Facundo Pessacg, and Javier Civera. Visual-inertial teach and repeat. *Robotics and Autonomous Systems*, 131:103577, 2020. doi: 10.1016/j.robot.2020.103577.

Mohammad Omama, Gabriele Berton, Eric Foxlin, and Yelin Kim. AsymLoc: Towards asymmetric feature matching for efficient visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26441–26451, 2026.

Rémi Pautrat, Viktor Larsson, Martin R. Oswald, and Marc Pollefeys. Online invariance selection for local feature descriptors. In *Computer Vision – ECCV 2020*, pp. 707–724. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58536-5_42.

Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12716–12725, 2019.

Torsten Sattler and Zuzana Kukelova. Simple but effective triplet-based compression strategies for compact visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29047–29059, 2026.

Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1744–1756, 2017. doi: 10.1109/TPAMI.2016.2611662.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006 of *Proceedings of SPIE*, pp. 369–386. SPIE, 2019. doi: 10.1117/12.2520589.

Yurun Tian, Bin Fan, and Fuchao Wu. L2-Net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 661–669, July 2017.

Philip H. S. Torr and Andrew Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156, 2000. doi: 10.1006/cviu.1999.0832.

Fangjinhua Wang, Xudong Jiang, Silvano Galliani, Christoph Vogel, and Marc Pollefeys. GLACE: Global local accelerated coordinate encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21562–21571, 2024.

Rouwan Wu, Xiaoya Cheng, Juelin Zhu, Yuxiang Liu, Maojun Zhang, and Shen Yan. UAVD4L: A large-scale dataset for UAV 6-dof localization. In *2024 International Conference on 3D Vision*, pp. 1574–1583. IEEE, 2024. doi: 10.1109/3DV62453.2024.00156.

Shen Yan, Xiaoya Cheng, Yuxiang Liu, Juelin Zhu, Rouwan Wu, Yu Liu, and Maojun Zhang. Render-and-compare: Cross-view 6-dof localization from noisy prior. In *2023 IEEE International Conference on Multimedia and Expo*, pp. 2171–2176. IEEE, 2023. doi: 10.1109/ICME55011.2023.00371.

Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, and Ping Tan. SceneSqueezer: Learning to compress scene for camera relocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8259–8268, 2022.

Yibin Ye, Xichao Teng, Shuo Chen, Leqi Liu, Kun Wang, Xiaokai Song, and Zhang Li. Exploring the best way for uav visual localization under low-altitude multi-view observation condition: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 1731–1741, June 2026.

Juelin Zhu, Shen Yan, Long Wang, Shengyue Zhang, Yu Liu, and Maojun Zhang. LoD-Loc: Aerial visual localization using LoD 3d map with neural wireframe alignment. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119063–119098. Curran Associates, Inc., 2024. doi: 10.52202/079017-3782.

Juelin Zhu, Shuaibang Peng, Long Wang, Hanlin Tan, Yu Liu, Maojun Zhang, and Shen Yan. LoD-Loc v2: Aerial visual localization over low level-of-detail city models using explicit silhouette alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 26610–26621, 2025.

A ARCHITECTURE AND FEATURE PROCESSING

Descriptor steering and PCA. AeroSCR uses the released SO2-Spread Setting-B DeDoDe-B descriptor and steerer of Bökman et al. (2024b). The DeDoDe-L detector, DeDoDe-B descriptor, and steerer are frozen during scene-specific training. Steering is applied to the 256-dimensional descriptor before a fixed PCA projection to 64 dimensions. PCA statistics are estimated separately from the canonicalized training descriptors of each scene; the nine Oslo-Large GeoCells share the same Oslo-Large PCA transform. At inference, a query is encoded once and the resulting descriptors are reused across steering hypotheses and candidate GeoCell heads.

Scene-coordinate decoder. Our implementation is adapted directly from the official R-SCoRe codebase (Jiang et al., 2025), retaining its coarse-to-refined scene-coordinate head and configurable architecture. Following this design, we adjust the head width and coarse-center configuration across scenes while keeping the residual-block layout fixed. AeroSCR takes 64-dimensional PCA-projected descriptors as input and uses three residual blocks before coarse coordinate prediction and two before residual refinement, with an MLP ratio of 2. AGL/AMSL and each AnyVisLoc scene use a single 384-channel head with 44 coarse centers, Oslo-Center uses a single 512-channel head with 50 centers, and Oslo-Large uses nine independent 512-channel GeoCell heads with 200 centers each. Each AnyVisLoc scene is modeled independently with its own PCA transform and coordinate normalization and does not use GeoCell routing. The Oslo-Large heads share the feature extractor,

steerer, and PCA transform, but maintain independent scene-coordinate parameters and coordinate normalizations. No configuration uses a global image code.

B TRAINING AND IMPLEMENTATION DETAILS

Table 6: Per-dataset training configuration for the final AeroSCR models. Exposures denote augmentation views constructed per mapping image. For AnyVisLoc, the exposure count applies independently to each scene-specific single-head training buffer; for Oslo-Large, it applies independently within each GeoCell training buffer.

Dataset	Short side	Exposures	Batch	Steps	Residual yaw	T_c (steps)
AGL/AMSL	480 px	58/image	32,768	60k	$\pm 15^\circ$	30k
AnyVisLoc	480 px	62/image	32,768	60k	$\pm 15^\circ$	30k
Oslo-Center	600 px	62/image	32,768	60k	$\pm 15^\circ$	120k
Oslo-Large	600 px	62/image/cell	32,768	60k	$\pm 15^\circ$	120k

Scale augmentation. For AGL/AMSL, each mapping image contributes 40 native exposures, nine intermediate exposures with $s \sim \mathrm{LogUniform}[1.2, 1.6]$, and nine boundary exposures at $s = 1.75$. The boundary crops use a balanced 3×3 crop-origin grid. For Oslo-Center, Oslo-Large, and AnyVisLoc, each mapping image contributes 24 native exposures, six zoom-out exposures with $s \sim \mathrm{LogUniform}[0.6, 1.0]$, six boundary exposures at $s = 0.6$, 12 zoom-in exposures with $s \sim \mathrm{LogUniform}[1.2, 2.2]$, and 14 boundary exposures at $s = 3.2$. For these datasets, zoom-in crop origins are sampled over the valid image region, while zoom-out exposures use whole-image resizing.

For $s > 1$, a valid crop is resized to the encoder input resolution; for $s < 1$, the whole image is resized. The same image-space transformation is applied to RGB and pixel-aligned geometric supervision, with camera intrinsics updated accordingly while the camera pose remains unchanged. If $H_{s,o}$ denotes the crop-and-resize transformation, we use

$$K' = H_{s,o} K, \qquad T' = T. \tag{12}$$

Descriptors are re-extracted from the transformed RGB image rather than numerically rescaled from previously extracted descriptors.

Residual-yaw augmentation. During buffer construction, each training sample is assigned a residual steering perturbation $\epsilon \sim \mathcal{U}[-15^\circ, 15^\circ]$ around its pose-derived canonical angle. The perturbation affects only descriptor steering; image coordinates, camera intrinsics, camera pose, and metric scene-coordinate targets remain unchanged. The sampled residual is stored with the training sample and is not resampled during optimization. All final main-paper models use this 15° residual radius.

Metric target construction. Scene-coordinate targets are generated from mapping observations only. For each sampled mapping pixel u_i, its calibrated camera ray is intersected with the mapping-only mesh, and a valid ray–mesh intersection defines the metric target $X_i^\star$ in the world coordinate frame. Rays without a valid intersection are marked invalid and excluded from metric-coordinate supervision. Under crop-resize augmentation, image coordinates and intrinsics are transformed consistently, while the corresponding world-space target $X_i^\star$ remains unchanged.

Metric supervision. The metric-coordinate term defined in the main paper (Eq. 11) is applied to both coarse and refined scene-coordinate predictions. Coordinate error is measured by the Euclidean distance to the metric target; invalid targets are masked, errors below 0.1 are ignored, and the remaining errors are clipped at 1000, summed, and normalized by the batch size. The maximum metric loss weight is 1 and follows the cosine schedule

$$\lambda_{\mathrm{geo}}(t) = \frac{1}{2}\left[1 + \cos\left(\pi \min\left(\frac{t}{T_c}, 1\right)\right)\right], \tag{13}$$

Rays without valid mesh intersections are treated as invalid targets and masked from the metric-coordinate loss; no surrogate depth is assigned. AGL/AMSL and AnyVisLoc use $T_c = 30\text{k}$ steps, so the metric term decays to zero halfway through the 60k-step training run. Oslo-Center and Oslo-Large use $T_c = 120\text{k}$, so the metric term remains nonzero throughout their 60k-step runs.

Optimization. All final models are trained for 60,000 optimization steps with batch size 32,768, two-step gradient accumulation, and bfloat16 mixed precision. We use AdamW (Loshchilov & Hutter, 2019) with an initial learning rate of 5×10^{-4}, weight decay 10^{-2}, gradient clipping at 10, and a one-cycle learning-rate schedule (Smith & Topin, 2019) with a peak learning rate of 3×10^{-3} and warm-up fraction 0.04. We use the final checkpoint from each training run; no validation-set-based checkpoint selection is performed.

Ablation controls. The Oslo-Center scale ablations use the same 65 mapping images, scene-coordinate architecture, PCA projection, residual-yaw setting, metric supervision, optimizer, training budget, evaluation set, and pose solver. Native-only training uses identity scale for all 62 exposure slots. The matched i.i.d. control uses 24 native exposures and 38 shifted exposures with $s \sim \text{LogUniform}[0.6, 3.2]$, while the coverage-balanced variant uses the structured 24+6+6+12+14 schedule described above. The variant without dense metric-coordinate supervision retains the same scale buffer and mesh-derived coarse centers but removes the coordinate-loss terms from both coarse and refined predictions. It therefore isolates dense metric-coordinate supervision rather than removing all uses of map geometry.

C DECODABILITY ANALYSIS AND CONTROLLED PROBES

Figures 4 and 5 provide the detailed diagnostics for the controlled scale and yaw experiments summarized in Fig. 1. We report positive-pair cosine similarity, positive-correspondence R@1, the positive–negative similarity margin, coordinate recall within 5 m, and pose success at 5 m/5° for an explicit nearest-neighbor feature bank and a point-wise SCR decoder using the same frozen DeDoDe-B representation.

Paired-height scale probe. We use 14 mapping views rendered at $h = 600$ m with identical camera orientation. The SCR decoder is trained only on these native mapping views, without scene-specific scale augmentation. Query images are rendered at the same horizontal camera locations but at heights spanning 500–700 m, so the underlying locations are preserved while the observation scale changes. We report relative image scale as $s = 600/h_q$.

At the native mapping condition, both explicit matching and SCR localize all queries successfully. Under only a $\pm 2.5\%$ scale change, positive-correspondence R@1 remains 95.5–95.7% and the explicit feature bank retains 98.4–98.6% coordinate recall within 5 m and 100% pose success. In contrast, SCR coordinate recall falls to 22.2–25.1%, with pose success of only 0–7.1%. Thus, the descriptors remain effective for retrieving the correct scene point even when the fixed coordinate decoder no longer maps them reliably to the corresponding metric coordinate.

Pure-yaw probe. To isolate in-plane rotation, we rotate the 14 views from 0° to 90° while updating the camera poses and retaining the same projected 3D tracks. This keeps the physical correspondences fixed and isolates descriptor and decoder sensitivity from detector repeatability. At 10°, positive-correspondence R@1 remains 90.7% and the explicit feature bank still localizes all queries, whereas SCR coordinate recall falls to 1.46% and SCR pose success is zero. With larger rotations, correspondence ranking also deteriorates and explicit matching eventually fails.

Together, the two probes show that stable scene-coordinate decoding can fail under controlled scale and yaw shifts for which relative correspondence discrimination and explicit feature-bank localization remain effective, despite using the same frozen local representation.

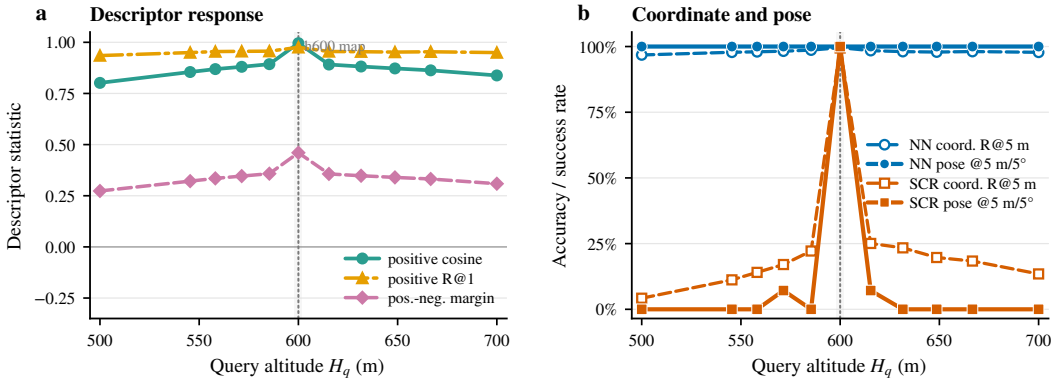

Figure 4: Paired-height scale probe. Fourteen mapping views are rendered at $h = 600$ m, and queries use the same horizontal camera locations at heights from 500 to 700 m. The SCR decoder is trained only at the mapping height, without scene-specific scale augmentation. The left panel reports descriptor statistics and the right panel reports coordinate recall and pose success.

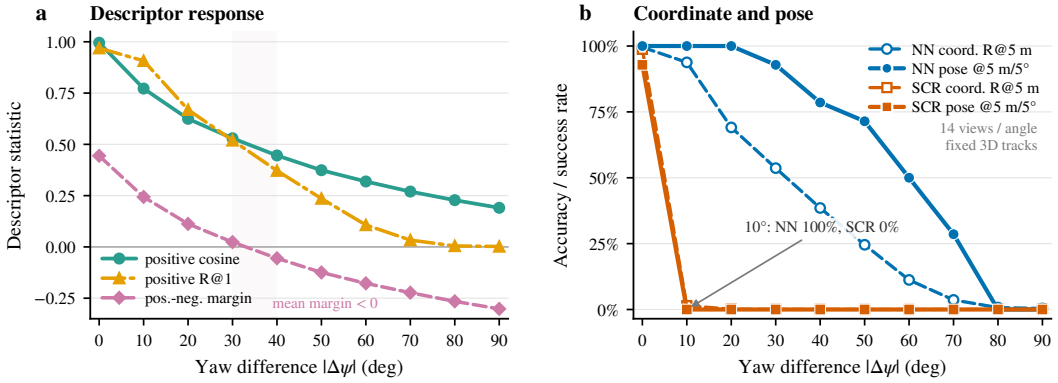

Figure 5: Pure-yaw probe. Fourteen views are rotated from 0° to 90° with updated camera poses while retaining the same projected 3D tracks. The left panel reports descriptor statistics and the right panel reports coordinate recall and pose success.

D ADDITIONAL EVALUATION ANALYSES

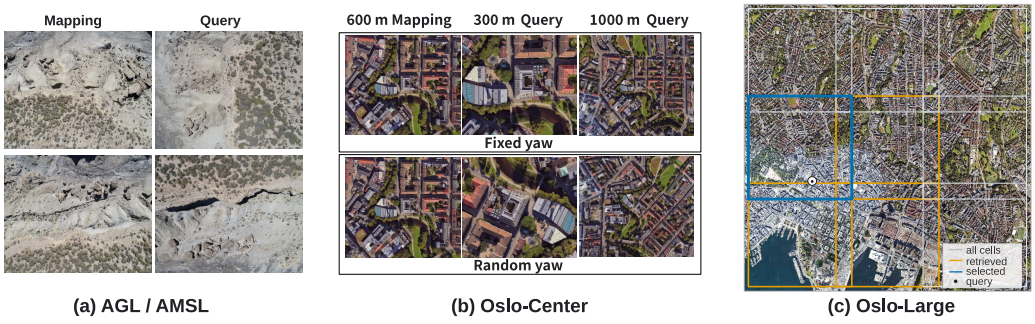

Figure 6: **Evaluation datasets and protocols.** (a) Real AGL/AMSL imagery; (b) controlled Oslo-Center scale/yaw evaluation; (c) retrieval-gated Oslo-Large localization.

D.1 YAW-PRIOR WIDTH SENSITIVITY AND REAL-IMAGE YAW-GAP ANALYSIS

Yaw-prior width sensitivity. We evaluate yaw-prior uncertainty on the same 370-query Oslo-Center random-heading split used in the main experiments. For each query, we recover the normalized yaw perturbation from the default $\pm 60^\circ$ condition and scale it proportionally with the prior half-width W, such that $e_q(W) = W\xi_q$. This preserves the same relative perturbation pattern across $\pm 30^\circ$, $\pm 60^\circ$, and $\pm 90^\circ$, while ensuring that the ground-truth yaw remains inside each bounded interval. AeroSCR uses the same 10° gravity perturbation for all conditions, whereas HLoc does not consume the gravity observation. For bounded priors, HLoc selects compatible C4 image-space rotations from the yaw observation, AeroSCR evaluates steering hypotheses at 15° spacing and supplies the perturbed gravity observation only to UP2P initialization, and R-SCoRe+ evaluates each yaw hypothesis across its 20 retrieved global-conditioning hypotheses. Full-circle evaluation removes the informative yaw observation entirely: AeroSCR evaluates 24 unique steering hypotheses at 15° spacing, while HLoc evaluates all four C4 rotations.

Table 7: Yaw robustness analyses. (a) Controlled yaw-prior width sensitivity on the 370-query Oslo-Center random-heading split. Finite-width conditions use proportionally scaled per-query yaw perturbations with a fixed 10° gravity perturbation for AeroSCR; Full 360° evaluates the complete orientation hypothesis set without an informative yaw observation. (b) Real AGL/AMSL localization at $1\,\mathrm{m}/1^\circ$, decomposed by the yaw difference to the nearest mapping heading.

(a) Oslo-Center yaw-prior width sensitivity

Yaw prior	Method	Hyp./query	@5m/5°	@2m/2°	Mean (ms)
$\pm 30^\circ$	AeroSCR	5.000	98.1	59.2	108.5
	HLoc	1.724	98.9	88.6	1771.0
$\pm 60^\circ$	AeroSCR	9.000	97.8	59.7	126.0
	HLoc	2.311	98.6	87.0	2368.5
$\pm 90^\circ$	AeroSCR	13.000	97.6	60.0	155.4
	HLoc	3.000	98.6	87.8	3097.7
Full 360°	AeroSCR	24.000	98.6	59.2	231.7
	HLoc	4.000	98.6	87.6	4154.8

(b) Real AGL/AMSL yaw-gap robustness

Setting	Method	Overall @1m/1°	$< 45^\circ$	45–135°	$\geq 135^\circ$
No yaw handling	HLoc	13.59	100.00	0.00	0.00
	ACE-G	3.40	17.86	1.65	0.00
	GLACE	11.17	82.14	0.00	0.00
	R-SCoRe	13.59	100.00	0.00	0.00
$\pm 60^\circ$ yaw / 10° gravity	HLoc	100.00	100.00	100.00	100.00
	AeroSCR	99.51	100.00	99.17	100.00

Localization accuracy remains essentially stable as the yaw prior is widened. AeroSCR varies only from 97.6–98.6% at $5\,\mathrm{m}/5^\circ$ and 59.2–60.0% at $2\,\mathrm{m}/2^\circ$, retaining 98.6%/59.2% under full-circle search. The primary effect is computational: AeroSCR latency increases from 108.5 ms with five steering hypotheses to 231.7 ms with 24 full-circle hypotheses. HLoc shows the same accuracy stability, while its mean latency increases from 1.77 s to 4.15 s as the number of evaluated C4 rotations grows. Thus, in this controlled setting, the bounded yaw observation primarily reduces orientation-search cost rather than determining localization accuracy.

Real-image yaw-gap analysis. We retain the AGL/AMSL decomposition by the yaw difference to the nearest mapping heading as a complementary real-image analysis. Without explicit yaw handling, all evaluated methods collapse once the heading gap exceeds 45°, and none succeeds on the $|\Delta\psi| \geq 135^\circ$ subset at $1\,\mathrm{m}/1^\circ$. Under the default $\pm 60^\circ/10^\circ$ orientation condition, HLoc remains at 100% across all yaw-gap bins, while AeroSCR reaches 99.5% overall and 100% on the largest-gap subset.

D.2 GEOCELL ROUTING ON OSLO-LARGE

GeoCell routing. Oslo-Large is partitioned into nine overlapping GeoCells. NetVLAD retrieves the top 20 mapping images, which vote for candidate cells; cells receiving at least 30% of the maximum vote count are evaluated. Core-cell recall is used only as a routing diagnostic and does not affect pose selection.

Table 8: GeoCell routing on Oslo-Large with 4,450 queries per heading setting. Core-cell recall measures whether retrieval activates the query's ground-truth core cell; Heads/query is the mean number of evaluated GeoCell heads out of nine.

Condition	Core-cell recall (%)	Heads/query	@5m/5°	@2m/2°
Fixed yaw	99.12	3.29	92.81	54.31
$\pm 30^\circ$ yaw / 2° gravity	95.87	3.78	86.97	44.52
$\pm 60^\circ$ yaw / 10° gravity	95.03	3.89	86.16	43.19

Retrieval recalls the query's core cell for 99.1% of fixed-yaw queries and 95.0% under the $\pm 60^\circ/10^\circ$ orientation condition, while activating only 3.29 and 3.89 of nine heads on average, respectively. The retrieval gate therefore reduces the number of scene-coordinate heads evaluated per query while maintaining high routing recall.

D.3 RESIDUAL-YAW TRAINING-RADIUS SENSITIVITY

Residual-yaw radius. We vary the training residual radius $\epsilon_{\max}$ on Oslo-Center while keeping all other training and evaluation settings fixed. Fixed evaluates the native mapping orientation, Oracle uses continuous ground-truth-yaw steering, and Deployable uses the bounded-yaw protocol with 15° steering spacing. The 15° radius used in the main experiments was fixed before this sensitivity sweep. No held-out validation split was used to select $\epsilon_{\max}$, so we retain this pre-specified setting rather than selecting a radius from the test sweep.

Table 9: Sensitivity to the residual-yaw training radius $\epsilon_{\max}$ on Oslo-Center. All rows use the same training and evaluation configuration except for the residual augmentation radius. Fixed evaluates the native mapping orientation, Oracle uses continuous ground-truth-yaw steering, and Deployable uses the bounded-yaw protocol with 15° steering spacing.

	Fixed		Oracle		Deployable	
$\epsilon_{\max}$	@2m/2°	@1m/1°	@2m/2°	@1m/1°	@2m/2°	@1m/1°
0°	74.59	22.16	56.76	14.59	57.03	13.51
7.5°	80.27	18.11	61.89	14.05	58.92	12.70
10°	80.27	22.97	60.81	12.16	62.97	14.05
15°	78.11	20.54	61.62	9.73	59.73	11.89
22.5°	76.76	22.43	60.81	9.46	60.81	8.65

At 2 m/2°, every nonzero residual radius improves over $\epsilon_{\max} = 0^\circ$ in all three evaluation settings. The optimum varies across protocols, with Deployable peaking at 10°, while performance remains similar over moderate nonzero radii. At 1 m/1°, the trend is less consistent, indicating that increased local yaw tolerance does not necessarily improve the strictest pose precision. The main-paper 15° setting is retained because it was fixed before this sweep rather than selected from these test results.

D.4 CONTROLLED R-SCORE+ BASELINE

R-SCoRe+ control. R-SCoRe+ evaluates whether AeroSCR's observation-side treatment can reproduce its gains while preserving R-SCoRe's shared global-conditioning mechanism. It adopts the same scale augmentation, SO(2) descriptor steering, residual-yaw training, metric coordinate supervision, and gravity-aware pose estimation as AeroSCR, while retaining R-SCoRe's global con-

ditioning and shared scene-coordinate decoder. On AnyVisLoc, this configuration is instantiated independently for each of the six scenes. The fixed- and random-heading Oslo-Center evaluations use the same final 60k-step checkpoint.

Table 10: Controlled R-SCoRe+ evaluation on Oslo-Center. Global × yaw reports the number of conditioning combinations evaluated per query.

Condition	@5m/5°	@2m/2°	Median t (m)	Median r (°)	Mean (ms)	P95 (ms)	Global × yaw
Fixed yaw	98.11	55.14	1.90	0.19	451.80	634.71	20×1
Random + prior	92.16	40.81	2.30	0.24	1204.14	1284.68	20×9

The global-conditioning search is evaluated once per yaw hypothesis, increasing the number of joint hypotheses from 20 under fixed yaw to 180 under the orientation-prior protocol.

Oslo-Large convergence. Applying the same R-SCoRe+ configuration to all 697 Oslo-Large mapping conditions completes 60k training steps without numerical failure but does not obtain usable convergence. Under fixed yaw, it reaches 0.09% at 5 m/5° and 0% at 2 m/2°, with median errors of 103.57 m/10.58°. The final training inlier fraction remains approximately 4.6%, with a median reprojection error of approximately 328 px. Retrieval and oracle-candidate checks indicate that neither retrieval nor final hypothesis selection is the dominant failure mode. Since the same configuration converges normally on Oslo-Center, we treat this result as a limitation of the evaluated shared global-conditioned configuration at the Oslo-Large scale, rather than evidence that global conditioning is intrinsically non-scalable.

D.5 AnyVisLoc Per-Scene Evaluation

The AnyVisLoc evaluation uses six eligible scenes, each with an independently defined mapping/query split. Scene eligibility was determined from dataset geometry and reconstruction properties, independently of localization performance. Table 11 reports the scene composition together with per-scene localization success at 5 m/5°. The final row reproduces the equal-weight scene average reported in the main paper; mean online latency is instead pooled over all 974 queries, as detailed in Appendix E.

Table 11: Per-scene AnyVisLoc localization success at 5 m/5°. Mapping and query counts are shown for each of the six eligible scenes. The final row reports the equal-weight mean across scenes used in the main paper.

			Native inference				Orientation prior		
Scene	Map	Query	HLoc	ACE-G	GLACE	R-SCoRe	HLoc	R-SCoRe+	AeroSCR
S1	36	101	88.1	22.8	48.5	72.3	99.0	92.1	96.0
S2	43	235	72.3	9.4	15.7	60.9	85.5	66.4	84.3
S4	26	314	66.9	21.3	18.8	74.5	75.5	80.6	83.1
S5	21	124	87.1	26.6	16.1	83.1	95.2	91.9	83.9
S6	41	135	71.9	13.3	6.7	48.1	97.0	93.3	99.3
S7	32	65	83.1	35.4	29.2	70.8	89.2	93.8	98.5
Mean	–	–	78.2	21.5	22.5	68.3	90.2	86.4	90.8

E Efficiency Measurement and Asset Accounting

Latency. All methods use the same query-input and online timing protocol. Query RGB images are isotropically resized to a height of 480 pixels while preserving aspect ratio, and camera intrinsics are scaled by the same factor. Each method retains its native checkpoint, retrieval budget, hypothesis count, solver, numerical precision, and other inference settings. Timing starts with query-image decoding and ends when the final pose or localization failure is returned. It includes preprocessing and transfer, feature extraction, retrieval and routing, all evaluated candidate branches, feature

matching or scene-coordinate decoding, pose estimation and refinement, and final hypothesis selection. Model, map, and database loading, offline map or feature construction, metric computation, and result serialization are excluded. Measurements use batch size one on a single RTX 4090 with CUDA synchronization around each query. The first 10 queries are used for warm-up, after which the complete query set is evaluated once; failed queries remain in both the latency statistics and the accuracy denominator. For AnyVisLoc, timing is evaluated separately for each method–scene run with 10 warm-up queries excluded from measurement. The reported mean latency is then computed over all 974 timed evaluation queries pooled across the six scenes, rather than by averaging the six scene-level means. Main tables report mean per-query latency.

Scene storage. We report decimal MB for scene-specific assets required online and exclude shared pretrained weights, temporary query caches, logs, training checkpoints not loaded at inference, and evaluation outputs. HLoc includes database local and global descriptors together with the COLMAP map. ACE-G includes its learned scene embeddings and scene configuration, while GLACE and R-SCoRe include their scene-specific regressors and associated scene codes or projections. R-SCoRe+ follows the same scene-specific storage accounting as R-SCoRe, with pretrained descriptor and steerer weights excluded. AeroSCR includes its scene-coordinate head and PCA projection together with retrieval assets where applicable; on Oslo-Large, this additionally includes all nine GeoCell heads, GeoCell metadata, and database NetVLAD descriptors.

F POSE-SOLVER DETAILS

For each evaluated scene-coordinate head under each yaw hypothesis, the gravity-aware upright two-point solver (UP2P) (Kukelova et al., 2011) is embedded in MSAC (Torr & Zisserman, 2000) to obtain an initial pose from the predicted 2D–3D correspondences. On Oslo-Large, the evaluated heads correspond to retrieval-selected GeoCells. For the AnyVisLoc and Oslo-Large evaluations, MSAC uses a 10-pixel reprojection threshold and between 64 and 500 iterations, with at most 5,000 scene-coordinate correspondences passed to the pose solver per query. Each candidate pose is subsequently refined with unconstrained 6-DoF optimization using PoseLib (Larsson & contributors, 2020). Refined hypotheses are ranked by Geman–McClure robust reprojection cost, with geometric support and RANSAC inlier count used as secondary criteria. Gravity is used only for UP2P initialization; roll and pitch are released during final refinement, so the perturbed gravity observation does not constrain the final pose.