ANCHORFUSE: ROBUST ANCHOR-CENTERED TEMPORAL FUSION FOR MULTIMODAL AERIAL-GROUND PLACE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal aerial-ground place recognition localizes a ground platform by matching camera–LiDAR observations to georeferenced aerial imagery. However, descriptor-level temporal aggregation does not explicitly distinguish evidence relevant to the queried location from observations of neighboring locations, and retrieval can degrade when context is unavailable. We propose AnchorFuse, an anchor-centered framework that explicitly separates the retrieval target from its temporal context. An anchor observation defines both the positive aerial references and the coordinate frame for fusion, keeping the retrieval target fixed as auxiliary observations change. Building on this formulation, Anchor-Aligned Cross-Frame Fusion (AACF) transforms local camera–LiDAR features into the anchor frame and selects nearby evidence according to spatial proximity, temporal distance, and learned reliability. A gated residual integrates the selected evidence into the query descriptor. To accommodate missing context, Robust Anchor-Memory Consistency (RAMC) jointly supervises clean and corrupted queries against the same retrieval target and aligns corrupted descriptors with a stop-gradient clean target. Context corruption preserves the multimodal anchor, and RAMC adds no inference-time branch. AnchorFuse achieves R@1 of 44.28% on KITTI360-AG and 78.35% on nuScenes-AG. On KITTI360-AG, RAMC improves R@1 over AACF by 17.03 percentage points when two non-anchor frames are removed, compared with 1.49 points on clean queries.

1 INTRODUCTION

Aerial-ground place recognition localizes a ground platform by retrieving matching images from a georeferenced aerial database. Learned descriptors support scalable retrieval (Arandjelovic et al., 2016; Berton et al., 2022; Izquierdo & Civera, 2024; Ali-bey et al., 2024), but ground views emphasize facades and nearby objects under occlusion, whereas aerial images capture spatial layout. Combining visual and geometric evidence can help bridge this discrepancy.

Cross-view methods improve matching through embeddings, geometry, and local correspondence (Hu et al., 2018; Zhu et al., 2022; Zhang et al., 2024; Wang et al., 2024a; Xia & Alahi, 2025), including relaxed query–reference centering (Zhu et al., 2021). Camera–LiDAR methods combine appearance and geometry (Komorowski et al., 2021; Zhou et al., 2024); AGPlace extends camera–LiDAR fusion to single-frame aerial retrieval (Figure 1(A)) (Wang et al., 2025). Temporal methods exploit sequences in same-view (Milford & Wyeth, 2012; Regmi & Shah, 2021) and cross-view localization (Shi et al., 2022; Vyas et al., 2022; Zhang et al., 2023; Pillai et al., 2024; Gao et al., 2026), motivating neighboring observations as additional evidence.

However, temporal context contains both evidence about the queried location and structures farther along the route. Descriptor-level aggregation (Figure 1(B)) combines observations after local evidence has been compressed into global descriptors, limiting explicit selection around the queried location. Missing context presents a second challenge: retrieval can degrade although the target remains fixed.

To select evidence relevant to the queried location, we introduce AnchorFuse (Figure 1(C)). The *anchor* defines the positive aerial set, fusion coordinates, and local queries; neighboring observations supply auxiliary evidence. Anchor-Aligned Cross-Frame Fusion (AACF) aligns this evidence to the

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.

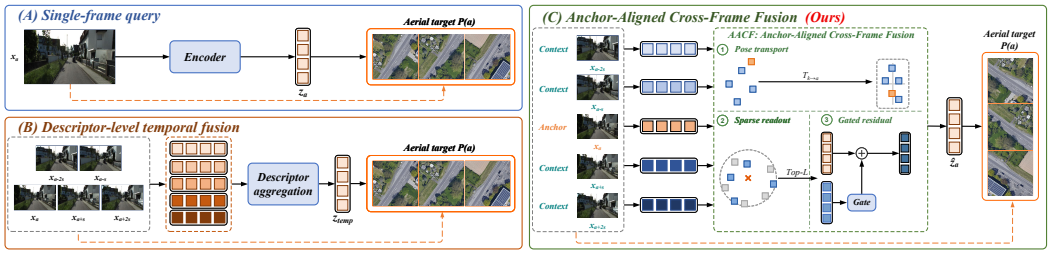

Figure 1: Temporal-query formulations and target assignment. (A) Single-frame retrieval encodes only the anchor observation $\mathbf{x}_a$. (B) Descriptor-level temporal fusion aggregates observations from multiple poses before retrieval, leaving the observation that defines $\mathcal{P}(a)$ implicit. (C) AACF explicitly designates $\mathbf{x}_a$ as the anchor: it defines $\mathcal{P}(a)$, while pose-aligned local context refines the global query descriptor through sparse readout and a gated residual. Orange dashed paths indicate the corresponding target assignment.

anchor frame, selects nearby entries using spatial proximity, temporal distance, and learned reliability, and refines the query descriptor through a gated residual.

To address missing context, Robust Anchor-Memory Consistency (RAMC) removes non-anchor frames during training. Retrieval supervision maintains discrimination among aerial candidates, while consistency aligns corrupted descriptors with a stop-gradient clean target. Aerial descriptors remain reusable, and RAMC adds no inference branch (Figure 2). AnchorFuse achieves the highest R@1 in the external comparison on both benchmarks; robustness gains are largest on KITTI360-AG under missing temporal or camera context.

Our contributions are summarized as follows:

- We formulate multimodal temporal retrieval around an anchor that jointly defines the retrieval target, fusion coordinates, and local queries for auxiliary evidence.
- We design AACF to refine the query descriptor with local camera–LiDAR evidence selected by geometry, temporal distance, and reliability.
- We introduce RAMC to combine retrieval supervision for corrupted-context queries with a clean descriptor target, improving robustness to missing frames and cameras on KITTI360-AG.

2 Related Work

2.1 Cross-View Geo-Localization

Cross-view geo-localization associates ground observations with aerial references. Early data-driven matching (Lin et al., 2013; Workman et al., 2015) led to learned aggregation, orientation reasoning, view synthesis, and transformer matching (Hu et al., 2018; Liu & Li, 2019; Toker et al., 2021; Shi et al., 2020; Zhu et al., 2022). Later methods improve layout correspondence, feature transport, and hard-negative sampling (Zhang et al., 2024; Yuan et al., 2024; Deuser et al., 2023). SALAD and BoQ aggregate visual-place descriptors through optimal transport and learnable queries (Izquierdo & Civera, 2024; Ali-bey et al., 2024). We instead examine local evidence across ground observations for retrieval at a designated anchor.

2.2 Multimodal Aerial-Ground Place Recognition

Point-cloud methods encode 3D scans, voxels, and range representations (Uy & Lee, 2018; Komorowski, 2021; Ma et al., 2022); camera–LiDAR fusion combines appearance and geometry for ground-level retrieval (Komorowski et al., 2021; Zhou et al., 2024). Cross-source methods study geometric correspondence across sensor domains (Guan et al., 2023). AGPlace matches ground camera–LiDAR observations to aerial images (Wang et al., 2025). We extend its single-observation backbone with pose-aligned neighboring evidence while retaining its frozen aerial encoder.

2.3 Temporal Modeling and Robustness

SeqSLAM matches sequences under appearance variation (Milford & Wyeth, 2012); Regmi & Shah (2021) combine geo-temporal features with GPS smoothing. In cross-view retrieval, GAMa matches

ground-video clips to aerial images and combines clip predictions (Vyas et al., 2022), while attention and sequential dropout accommodate variable-length sequences (Zhang et al., 2023). ConGeo addresses orientation and field-of-view variation (Mi et al., 2024), and robustness benchmarks evaluate image corruptions (Zhang & Zhu, 2024). Our setting preserves a multimodal anchor while perturbing non-anchor temporal and sensor context: AACF organizes local evidence, and RAMC supervises corrupted retrieval against a clean descriptor target.

3 METHOD

AnchorFuse reuses descriptors from the frozen AGPlace aerial encoder. Its ground branch combines AACF query refinement with RAMC training for missing context. Here *context* denotes non-anchor observations; *memory* denotes local entries available to AACF, including anchor entries.

3.1 ANCHOR-CENTERED QUERY FORMULATION

Let $\mathbf{x}_t = (\mathbf{I}_t, \mathbf{P}_t, \mathbf{T}_t^w)$ denote an observation at time t, comprising camera images, a LiDAR scan, and a local-odometry pose. The odometry frame w need not coincide with the global map frame. For reference time t_0, we construct a temporal window with T slots:

$$\mathcal{X}(t_0; a, s) = \{\mathbf{x}_{t_0+o_k}\}_{k=1}^{T}, \qquad o_k = (k-a)s, \qquad k = 1, \dots, T, \tag{1}$$

where s is the sampling stride and a is the anchor slot. We relabel the sampled observations as $\{\mathbf{x}_k\}_{k=1}^{T}$, with $\mathbf{x}_a = \mathbf{x}_{t_0}$. The default window has odd length and a central anchor.

Anchor and context. The query separates the anchor observation from its context:

$$\mathcal{Q}_a = (\mathcal{G}_a, \mathcal{C}_a), \qquad \mathcal{G}_a = \mathbf{x}_a, \qquad \mathcal{C}_a = \{\mathbf{x}_k : k \in \{1, \dots, T\},\ k \neq a\}. \tag{2}$$

The anchor pose defines the coordinate frame for local fusion. Its location ℓ_a determines the retrieval target in the aerial database $\mathcal{D}^A = \{(\mathbf{A}_j, \ell_j^A)\}_{j=1}^{N_D}$, where $\mathbf{A}_j$ is a reference image at location ℓ_j^A. For positive radius r_+, the positive-reference indices are

$$\mathcal{P}(a) = \left\{ j : \mathrm{dist}(\ell_a, \ell_j^A) \leq r_+ \right\}, \tag{3}$$

where dist measures geographic distance. Context can change the query representation without changing $\mathcal{P}(a)$.

3.2 ANCHOR-ALIGNED CROSS-FRAME FUSION

The ground encoder produces $(\mathbf{z}_a, \mathcal{S}_{1:T}) = E_\theta(\mathcal{Q}_a)$: a full-window base descriptor and per-observation local features. Subscript a identifies the target, not an anchor-only input. AACF refines $\mathbf{z}_a$ using local memory from this window, indexed by $\mathcal{K} = \{1, \dots, T\}$.

Cross-Frame Pose Transport. Pose transport makes local evidence geometrically comparable across observations. For $k \in \mathcal{K}$, let $\mathcal{V}_k$ index sampled LiDAR points with valid camera projections. Each coordinate $\mathbf{x}_{k,n}$ is paired with its sampled image feature. Anchor points define local queries and remain eligible as zero-offset memory entries. With local-odometry pose $\mathbf{T}_k^w$, coordinates are transported by

$$\mathbf{T}_{k\to a} = (\mathbf{T}_a^w)^{-1}\mathbf{T}_k^w, \qquad \bar{\mathbf{x}}_{k,n}^a = \mathbf{R}_{k\to a}\mathbf{x}_{k,n} + \mathbf{t}_{k\to a}, \qquad k \in \mathcal{K}. \tag{4}$$

Here $\mathbf{R}_{k\to a}$ and $\mathbf{t}_{k\to a}$ denote rotation and translation. Each memory value $\bar{\mathbf{e}}_{k,n}$ adds a projected image feature to a geometric embedding of $[\bar{\mathbf{x}}_{k,n}^a, |k-a|]$. A sigmoid MLP predicts reliability $\rho_{k,n} \in (0,1)$ from image features, temporal offset, distance to the anchor origin, and projection validity (architectures in Appendix A.1).

Sparse Geometry-Aware Context Readout. Readout is restricted to the physical neighborhood of each anchor coordinate $\mathbf{x}_{a,u}$ by radius r:

$$\mathcal{C}_u = \{(k,n) : k \in \mathcal{K},\ n \in \mathcal{V}_k,\ \|\bar{\mathbf{x}}_{k,n}^a - \mathbf{x}_{a,u}\|_2 \leq r\}. \tag{5}$$

Within this neighborhood, spatial proximity, temporal distance, and learned reliability determine the ranking:

$$s_{u,k,n} = \tau_r^{-1}\left[-\lambda_s d_{u,k,n} - \lambda_t |\Delta t_k| + \log\max(\rho_{k,n}, \epsilon)\right], \tag{6}$$

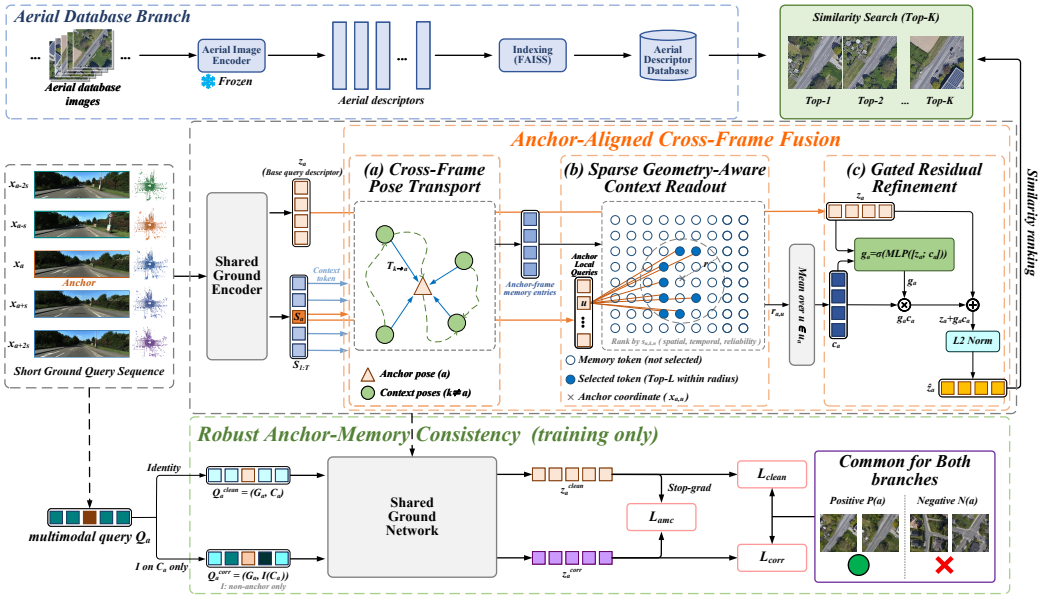

Figure 2: Overview of AnchorFuse. The aerial encoder is frozen. The base descriptor $\mathbf{z}_a$ uses the full query window; a identifies its target. AACF applies (a) pose transport, (b) anchor-local readout from memory including anchor entries, and (c) gated refinement. RAMC shares network parameters and retrieval references across clean/corrupted passes, with stop-gradient only on the clean consistency target. Only the upper query path is used at inference.

where $d_{u,k,n} = \|\bar{\mathbf{x}}^a_{k,n} - \mathbf{x}_{a,u}\|_2$ and $\Delta t_k = k - a$ measure spatial distance and temporal offset in slots. Coefficients λ_s, λ_t weight these penalties, τ_r is the softmax temperature, and $\epsilon = 10^{-6}$ stabilizes the logarithm. No query–key dot product is used.

To bound the number of entries read, let $\mathcal{N}_u$ contain up to L highest-scoring candidates in $\mathcal{C}_u$. For a nonempty neighborhood, softmax weights give the local readout

$$\alpha_{u,k,n} = \frac{\exp(s_{u,k,n})}{\sum_{(i,m)\in\mathcal{N}_u} \exp(s_{u,i,m})}, \qquad \mathbf{r}_{a,u} = \sum_{(k,n)\in\mathcal{N}_u} \alpha_{u,k,n}\bar{\mathbf{e}}_{k,n}. \tag{7}$$

If $\mathcal{N}_u$ is empty, we set $\mathbf{r}_{a,u} = \mathbf{0}$.

Gated Residual Refinement. We average readouts over valid anchor-local queries $\mathcal{U}_a = \mathcal{V}_a$. A query-dependent scalar gate controls the magnitude of the resulting correction:

$$\begin{aligned} \mathbf{c}_a &= |\mathcal{U}_a|^{-1} \sum_{u\in\mathcal{U}_a} \mathbf{r}_{a,u}, \qquad g_a = \sigma(\mathrm{MLP}([\mathbf{z}_a, \mathbf{c}_a])), \\ \tilde{\mathbf{z}}_a &= \mathbf{z}_a + g_a\mathbf{c}_a, \qquad \hat{\mathbf{z}}_a = \frac{\tilde{\mathbf{z}}_a}{\|\tilde{\mathbf{z}}_a\|_2}. \end{aligned} \tag{8}$$

If $\mathcal{U}_a$ is empty, we set $\mathbf{c}_a = \mathbf{0}$, leaving the descriptor unchanged before normalization. Here $[\cdot,\cdot]$ denotes concatenation and σ is the sigmoid function.



Anchor a: local query / entry 3 Context $a + 2s$: selected entries 6–8

Figure 3: Local evidence selected by AACF. Labels link the anchor query to selected context entries after pose transport. Across all five frames, seven of eight selected entries originate from context and carry 74.8% of this local readout's weight. The weights do not measure causal effects on retrieval: AACF and anchor-only memory both give first-positive rank 1.

3.3 Robust Anchor-Memory Consistency

To improve retrieval with incomplete context, RAMC supervises corrupted queries and aligns their descriptors with the clean query descriptor.

Non-Anchor Context Corruption. RAMC creates a clean query $\mathcal{Q}_a^{\text{clean}} = (\mathcal{G}_a, \mathcal{C}_a)$ and a corrupted query

$$\mathcal{Q}_a^{\text{corr}} = (\mathcal{G}_a, \mathcal{I}(\mathcal{C}_a)), \tag{9}$$

where intervention $\mathcal{I}$ acts only on non-anchor context. Both branches share $\mathcal{P}(a)$ and the same sampled aerial negatives $\mathcal{N}(a)$. Training uses non-anchor frame removal as specified in Appendix A.1. Additional context corruptions are evaluated at test time.

Retrieval and Clean-Teacher Consistency. The two passes share network parameters. Their normalized AACF descriptors, $\mathbf{z}_a^{\text{clean}}$ and $\mathbf{z}_a^{\text{corr}}$, receive triplet supervision against shared aerial references (Sec. 3.4):

$$\begin{aligned} \mathcal{L}_{\text{clean}} &= \mathcal{L}_{\text{ret}}(\mathbf{z}_a^{\text{clean}}, \mathcal{P}(a), \mathcal{N}(a)), \\ \mathcal{L}_{\text{corr}} &= \mathcal{L}_{\text{ret}}(\mathbf{z}_a^{\text{corr}}, \mathcal{P}(a), \mathcal{N}(a)). \end{aligned} \tag{10}$$

Corrupted-query retrieval supervision maintains discrimination among aerial candidates. The consistency loss aligns the corrupted descriptor with the clean descriptor:

$$\mathcal{L}_{\text{amc}} = 1 - \cos\big(\text{sg}(\mathbf{z}_a^{\text{clean}}), \mathbf{z}_a^{\text{corr}}\big), \tag{11}$$

where sg stops gradients only through the consistency target; the clean pass remains supervised by $\mathcal{L}_{\text{clean}}$. The RAMC objective is

$$\mathcal{L}_{\text{RAMC}} = \mathcal{L}_{\text{clean}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{amc}}\mathcal{L}_{\text{amc}}. \tag{12}$$

3.4 Training Objective and Inference

For a sampled training tuple, let $j^+ \in \mathcal{P}(a)$ denote its positive aerial reference and $j_n^- \in \mathcal{N}(a)$ its M negatives. With query descriptor $\mathbf{z}^q$, positive aerial descriptor $\mathbf{z}_{j^+}^A$, and negative aerial descriptors $\{\mathbf{z}_{j_n^-}^A\}_{n=1}^M$, the retrieval term is

$$\mathcal{L}_{\text{ret}}\big(\mathbf{z}^q, \mathcal{P}(a), \mathcal{N}(a)\big) = \frac{1}{M}\sum_{n=1}^{M}\Big[\mu + d(\mathbf{z}^q, \mathbf{z}_{j^+}^A) - d(\mathbf{z}^q, \mathbf{z}_{j_n^-}^A)\Big]_+, \tag{13}$$

where $d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2$, μ is the triplet margin, and $[x]_+ = \max(x, 0)$. We average this loss over the mini-batch. Alongside RAMC, two auxiliary terms regularize local camera–LiDAR features and the global descriptor under context masking:

$$\mathcal{L} = \mathcal{L}_{\text{RAMC}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}. \tag{14}$$

The local term $\mathcal{L}_{\text{local}}$ is a reliability-weighted symmetric InfoNCE loss over valid camera–LiDAR pairs. The auxiliary term $\mathcal{L}_{\text{mask}} = 1 - \langle \text{sg}(\mathbf{z}_a^{\text{clean}}), \mathbf{z}_a^{\text{mask}} \rangle$ uses the normalized descriptor $\mathbf{z}_a^{\text{mask}}$ from a separate masked-context pass: all non-anchor images are zeroed and their LiDAR points removed, with batch-normalization statistics fixed. It adds consistency alone; RAMC combines consistency with retrieval supervision after frame removal. Appendix A.1 specifies masking and loss weights; Table 11 evaluates these auxiliaries with RAMC fixed.

At inference, only the AACF query path is evaluated and matched against the precomputed aerial descriptors. The corruption, masking, and teacher branches are used only during training.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on two satellite-reference benchmarks using a 25 m positive radius. KITTI360-AG is derived from KITTI-360 for camera–LiDAR to aerial-image retrieval (Wang et al., 2025; Liao et al., 2022); its test set contains 7,147 ground queries and 3,589 aerial references. nuScenes-AG contains 6,008 ground queries and 6,663 aerial references, with six cameras and LiDAR

for each ground observation. Within each benchmark, the main comparisons use the same query set, aerial database, and retrieval criterion.

Evaluation metrics. We report Recall@K (R@K), $K \in \{1, 5, 10\}$, the percentage of queries whose top-K results contain a reference from the anchor-defined positive set in Eq. (3). In Figures 3, 5, and 6, *first-positive rank* denotes the rank of the highest-ranked reference in $\mathcal{P}(a)$. Unless stated otherwise, the best and second-best results in comparative tables are shown in **bold** and underlined, respectively. Tabulated recall differences are computed from the displayed values. Additional temporal recall metrics and paired statistics are reported in the supplement; Figures A1–A3 provide the corresponding qualitative diagnostics.

Implementation. We use a center-anchored window of five observations with stride two. Full-context experiments retain only anchors with valid observations at every offset, and no temporal window crosses a data-split boundary. The aerial branch remains fixed, and its descriptors are precomputed and shared across variants. AACF uses a 6 m readout radius and at most $L = 8$ neighbors per anchor feature; relative poses are obtained from local odometry. Matched-backbone variants are trained independently for four epochs under the same schedule. Temporal mean uses the trained single-observation model without additional training. Remaining scoring, loss, and optimization settings are provided in the supplement.

Baselines. The single-frame base is trained on anchor observations. Temporal mean is a parameter-free inference-time control: it averages the five individually ℓ_2-normalized observation descriptors and normalizes the result again. The learnable temporal variants comprise learned frame weighting, AACF, and AnchorFuse.

Table 1: Comparison with existing single-observation and sequential methods on KITTI360-AG and nuScenes-AG using satellite references. Methods are grouped by sensing modality and temporal setting.

		KITTI360-AG			nuScenes-AG		
Method	T	R@1	R@5	R@10	R@1	R@5	R@10
Single-observation RGB							
Sample4Geo (Deuser et al., 2023)	1	27.99	40.82	50.06	73.71	81.40	85.12
ConGeo (Mi et al., 2024)	1	32.25	37.57	41.32	70.09	77.44	80.31
TransGeo (Zhu et al., 2022)	1	22.04	31.90	38.02	70.06	74.83	76.15
GeoDTR+ (Zhang et al., 2024)	1	34.19	46.82	52.71	74.38	79.24	81.26
SALAD (Izquierdo & Civera, 2024)	1	26.75	34.65	38.34	54.39	68.43	74.45
BoQ (Ali-bey et al., 2024)	1	27.72	36.66	42.35	62.37	69.47	72.65
SinGeo (Chen et al., 2026)	1	35.79	52.23	56.81	75.24	83.69	84.55
Sequential RGB							
CVIS/SeqGeo (Zhang et al., 2023)	5	25.06	35.24	42.74	45.61	60.55	67.83
GAReT (Pillai et al., 2024)	5	29.51	46.17	53.78	58.77	75.05	78.73
TrajLoc (Gao et al., 2026)	5	38.65	57.57	68.56	71.55	86.68	90.08
Single-observation RGB+LiDAR							
MinkLoc++ (Komorowski et al., 2021)	1	27.18	39.32	47.01	69.98	81.61	86.85
AdaFusion (Lai et al., 2022)	1	27.33	41.08	49.04	71.51	82.86	87.80
LCPR (Zhou et al., 2024)	1	26.81	42.74	49.94	57.21	73.26	79.86
UMF (García-Hernández et al., 2024)	1	26.78	41.80	47.21	68.31	81.84	86.52
PRFusion++ (Wang et al., 2024b)	1	29.95	45.23	52.58	73.94	85.21	89.74
MSSPlace-LI (Melekhin et al., 2025)	1	24.73	36.84	44.92	70.75	81.64	87.09
BdFusion (Liang et al., 2025)	1	30.21	45.89	53.18	74.18	85.63	90.07
AGPlace (Wang et al., 2025)	1	35.06	49.24	55.74	76.00	85.82	88.91
Sequential RGB+LiDAR							
GSPR (Qi et al., 2025)	3	38.17	55.43	63.28	76.92	87.31	89.94
SeqMamba-MPR (Zhang et al., 2025)	3	37.64	54.92	62.87	76.41	86.82	89.76
AnchorFuse	5	**44.28**	**62.11**	**68.84**	**78.35**	**87.97**	**90.58**

Baseline results are obtained from matched reimplementations or task-specific adaptations under common query sets, aerial databases, and retrieval criteria within each benchmark. T denotes the number of ground observations per query.

4.2 Comparison with Existing Methods

Table 1 compares methods with different inputs: AnchorFuse uses five RGB+LiDAR observations, AGPlace one, and visual sequential methods RGB alone. Single-observation methods use the center frame; sequential methods use the indicated T. GAReT uses its visual stream, and TrajLoc uses only video inputs. Matched-input comparisons follow below.

Satellite-reference results. AnchorFuse achieves the highest R@1/5/10 among the evaluated adaptations on both benchmarks (Table 1), reaching 44.28% R@1 on KITTI360-AG and 78.35% on nuScenes-AG. Its KITTI360-AG margin over TrajLoc, the strongest sequential RGB comparator, is 5.63 points. Matched RGB+LiDAR controls below distinguish temporal-modeling gains from differences in sensor inputs.

Parameter-free temporal controls. Five-frame averaging improves R@1 for all eight RGB+LiDAR methods without additional training: by 0.89–1.46 points on KITTI360-AG and 0.70–2.65 on nuScenes-AG (Table 5). This establishes a parameter-free baseline for assessing learned temporal fusion.

4.3 Temporal Modeling

With the same five-observation input, AACF exceeds learned frame weighting by 3.86 R@1 points on KITTI360-AG, and AnchorFuse exceeds temporal mean by 8.06 points (Table 2). Figure 5 illustrates the corresponding retrieval improvement. An independent five-seed replication with retraining yields positive R@1 differences in 5/5 seed-matched pairs for both temporal mean versus the single-frame base and AnchorFuse versus AACF (Table A2).

Temporal window. The symmetric $T = 5$, stride-two window performs best on both benchmarks (Table 4) but requires future frames. Past-only retrieval removes this delay at an R@1 cost of 8.56 points on KITTI360-AG and 0.77 on nuScenes-AG.

4.4 Modality and AACF Ablations

Query modality. Under the same evaluation protocol, the multimodal variant reaches 44.28% R@1, compared with 26.72% for RGB alone and 11.25% for LiDAR alone (Table 8). The unimodal variants are trained independently.

Anchor-local AACF exceeds global-context refinement by 3.33 R@1 points (42.79 vs. 39.46%; Table 7), supporting local evidence selection in this architecture. Removing pose transport, reliability, sparse readout, or adaptive gating lowers recall at all three cutoffs. Table 9 tests anchor designation with the same five observations and positive set.

4.5 Objective Ablations and Robustness

We evaluate robustness under six perturbations that remove context frames, cameras, or LiDAR while preserving the anchor observation. We report R@1 on unperturbed queries (Clean), its mean over the six perturbations (Mean Corr.), and the minimum over them (Worst). Figure 6 complements these

Table 2: Comparison of temporal aggregation and RAMC on KITTI360-AG and nuScenes-AG. Learnable variants use the same ground-backbone architecture and are trained independently. Temporal mean uses the trained single-observation model without retraining; RAMC is used only during training.

		Components				KITTI360-AG			nuScenes-AG		
Method	T	Mean	Learn. W.	AACF	RAMC	R@1	R@5	R@10	R@1	R@5	R@10
Single-observation base	1					35.06	49.24	55.74	76.00	85.82	88.91
Temporal mean	5	✓				36.22	50.51	57.79	77.00	85.68	88.51
Learned frame weighting	5		✓			38.93	52.85	60.71	77.21	86.38	88.80
AACF	5			✓		42.79	59.59	67.64	77.65	87.78	90.45
AnchorFuse	5			✓	✓	**44.28**	**62.11**	**68.84**	**78.35**	**87.97**	**90.58**

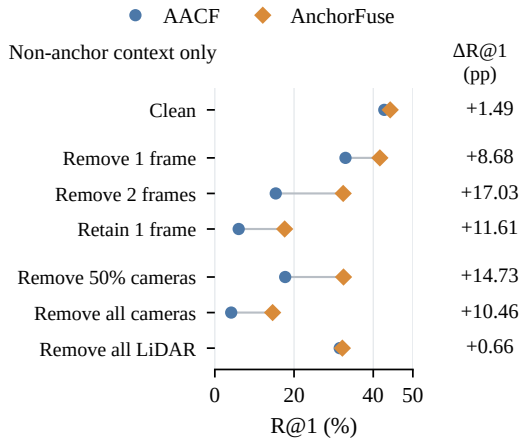

Figure 4: Context robustness on KITTI360-AG with the anchor preserved. Right: R@1 gains from RAMC (pp).

Table 3: Mean query latency (ms) and peak GPU memory (GB) with inputs preloaded. Data loading is excluded.

Method	Params (M)	ms	GB
Single-frame	17.179	13.37	0.27
Temporal mean	17.179	30.72	1.16
AACF	17.511	35.16	1.16
AnchorFuse	17.511	33.60	1.16

Table 4: Temporal-window comparison, R@1 (%). Delay: maximum future offset in frames.

Window	T	s	Delay	KITTI	nuScenes
Symmetric	3	2	2	28.54	77.75
Symmetric	5	1	2	40.84	78.08
Symmetric	5	2	4	**44.28**	**78.35**
Past-only	5	2	0	35.72	77.58
Symmetric	7	2	6	40.11	77.96

Table 5: Parameter-free temporal averaging. R@1 gains (pp) over $T = 1$ using the same trained RGB+LiDAR retriever.

Method	KITTI360-AG	nuScenes-AG
MinkLoc++	+1.11	+0.70
AdaFusion	+1.41	+1.25
LCPR	+0.89	+1.84
UMF	+1.04	+2.65
PRFusion++	+0.98	+1.35
MSSPlace-LI	+1.46	+1.72
BdFusion	+1.21	+1.23
AGPlace	+1.16	+1.00

Table 6: Additional context corruptions, R@1 (%). Anchor and target unchanged.

Condition	AACF	Ours	Δ (pp)
Pose (Level 3)	27.82	31.47	+3.65
All temporal outliers	4.67	23.37	+18.70



Anchor of the shared five-frame query



Anchor retained; remove context $a - 2s$ and $a + s$



Mean: rank 1 | Ours: rank 1 | Ours: rank 2 | Ours: rank 3

Figure 5: Temporal retrieval. First-positive rank improves from > 20 for temporal mean to 3 for AnchorFuse; top-1 remains incorrect. Green marks a positive within 25 m.



w/o RAMC: rank 1 | with RAMC: rank 1

First-positive rank

	Clean	Drop
w/o RAMC	3	> 20
with RAMC	3	1

Figure 6: Matched RAMC example. The same two-frame removal preserves the anchor and positive set. RAMC yields a positive top-1 result after removal.

aggregate metrics with a controlled two-frame removal that keeps the anchor, positive set, and AACF inference architecture fixed.

RAMC objective decomposition. Adding clean-target consistency improves mean corrupted recall by 2.85 points to 28.50% and worst-case recall by 7.24 points to 14.59% (Table 10). Symmetric consistency gives only 5.82% worst-case recall. These single-run comparisons support combining corrupted-query supervision with a stop-gradient clean target.

Auxiliary objective ablation. With RAMC fixed, $\mathcal{L}_{\text{local}}$ alone gives the highest clean recall, 44.42%. Adding $\mathcal{L}_{\text{mask}}$ gives the highest mean and worst-case corrupted recall, 28.50% and 14.59%, at a small cost in clean recall (Table 11).

Robustness to missing context. RAMC is trained with frame removal and evaluated with missing frames, cameras, or LiDAR (Figure 4). KITTI360-AG gains over AACF are 1.49 R@1 points on clean queries, 17.03 under two-frame removal, 14.73 under half-camera removal, and 0.66 under

Table 7: Ablation of context formulation and AACF components on KITTI360-AG. Mean, Global, and Local denote descriptor averaging, global-context refinement, and anchor-local readout; the AACF columns denote pose transport, reliability weighting, sparse readout, and adaptive gating.

Context formulation			AACF components				Recall		
Mean	Global	Local	Pose	Reliab.	Sparse	Gate	R@1	R@5	R@10
✓							36.22	50.51	57.79
	✓					✓	39.46	55.18	62.63
		✓		✓	✓	✓	41.37	57.84	65.71
		✓	✓		✓	✓	41.72	58.14	66.01
		✓	✓	✓		✓	41.86	58.39	66.21
		✓	✓	✓	✓		41.14	57.63	65.41
		✓	✓	✓	✓	✓	**42.79**	**59.59**	**67.64**

Table 8: Ablation of ground-query modalities on KITTI360-AG.

RGB	LiDAR	R@1	R@5	R@10
✓		26.72	34.60	40.73
	✓	11.25	21.63	27.48
✓	✓	**44.28**	**62.11**	**68.84**

Table 9: Anchor reassignment: old–new descriptor cosine and top-1 shift (m), on fixed windows with identical positive sets.

Method	Cosine	Shift
Mean	1.0000	0.00
Learned weights	0.9986	0.63
AACF	0.9336	12.31
AnchorFuse	0.9461	10.58

Table 10: RAMC objectives on KITTI360-AG. First row: AACF without RAMC. Clean is R@1 (%); Mean Corr./Worst are the mean/minimum over six context-removal conditions.

$\mathcal{L}_{\text{corr}}$	$\mathcal{L}_{\text{amc}}$	Consistency	Clean	Mean Corr.	Worst
			42.79	17.97	4.13
✓			43.82	25.65	7.35
	✓	Stop-grad.	43.37	25.10	8.72
✓	✓	Symmetric	44.10	25.59	5.82
✓	✓	Stop-grad.	**44.28**	**28.50**	**14.59**

Table 11: Auxiliary-loss ablation with RAMC and other training settings held fixed. Implementation details appear in Appendix A.1; metrics follow Table 10.

$\mathcal{L}_{\text{local}}$	$\mathcal{L}_{\text{mask}}$	Clean	Mean Corr.	Worst
		43.61	27.36	13.54
✓		**44.42**	27.88	13.59
	✓	44.12	28.28	14.58
✓	✓	44.28	**28.50**	**14.59**

context LiDAR removal. Severe degradation remains: removing all context cameras yields 14.59% R@1. The two-frame-removal gain is only 0.87 points on nuScenes-AG (Table A9), showing benchmark-dependent benefits.

Additional pose and temporal-outlier corruptions yield gains of 3.65 and 18.70 points, respectively (Table 6); R@1 under outlier replacement remains 23.37%.

4.6 Efficiency

AnchorFuse retains 256-D descriptors, adding 1.93% parameters over the single-frame base. Mean single-query latency is 33.60 ms: 2.51× the base and 9.4% above temporal mean (Table 3), excluding future-observation delay. RAMC adds no inference branch; latency differences from AACF reflect measurement variation. Table A11 additionally reports P95 latency and batch-dependent throughput.

5 Limitations

Our evaluation covers two satellite-reference driving benchmarks, with semantic road-map transfer evaluated only on KITTI360-AG. Both the recall reduction from removing context and the improvement from RAMC are modest on nuScenes-AG, where the intact anchor contains six cameras and LiDAR. The default symmetric window requires future observations, while past-only retrieval is less accurate. The AACF ablations measure changes within a coupled architecture rather than independent component effects. Generalization to other datasets, map representations, and corruption severities requires further evaluation.

6 Conclusion

We presented AnchorFuse, an anchor-centered framework for multimodal aerial-ground place recognition. An explicit anchor defines the retrieval target and fusion coordinates, while AACF selects local evidence from neighboring observations. Experiments on KITTI360-AG and nuScenes-AG support the effectiveness of this formulation. RAMC further improves retrieval under missing temporal and camera context on KITTI360-AG. Future work includes online temporal inference, evaluation under broader corruptions, and transfer to additional map representations.

AI USE STATEMENT

We used OpenAI ChatGPT and Codex to assist with initial manuscript drafting, LaTeX typesetting, software implementation and debugging, and consistency checking. The authors independently conducted the experiments, interpreted the experimental results, prepared all figures, and led the subsequent writing and revision of the manuscript. All AI-assisted outputs were reviewed and verified by the authors. The authors take full responsibility for the final manuscript, code, experiments, figures, and references.

REPRODUCIBILITY STATEMENT

Appendix A.1 specifies implementation details, optimization hyperparameters, retrieval sampling, and auxiliary objectives; Appendix A.2 describes baseline adaptations and evaluation protocols. Appendices B–G provide additional temporal comparisons, repeated-run analyses, qualitative results, ablations, robustness evaluations, and computational-cost measurements. An anonymized implementation, including training/evaluation scripts and configuration files, is provided as supplementary material.

REFERENCES

Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17794–17803, 2024.

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.

Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4878–4888, 2022.

Yang Chen, Xieyuanli Chen, Junxiang Li, Jie Tang, and Tao Wu. SinGeo: Unlock single model's potential for robust cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19403–19412, 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*, pp. 16847–16856, 2023.

Tianyi Gao, Jiayu Lin, Danielle Beaulieu, and Nathan Jacobs. Trajectory-aware cross-view geo-localization with sequential observations. In *European Conference on Computer Vision*, 2026.

Alberto García-Hernández, Riccardo Giubilato, Klaus H. Strobl, Javier Civera, and Rudolph Triebel. Unifying local and global multimodal features for place recognition in aliased and low-texture environments. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3991–3998, 2024.

Tianrui Guan, Aswath Muthuselvam, Montana Hoover, Xijun Wang, Jing Liang, Adarsh Jagan Sathyamoorthy, Damon Conover, and Dinesh Manocha. CrossLoc3D: Aerial-ground cross-source 3D place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11335–11344, 2023.

Sixing Hu, Mengdan Feng, Rang M. H. 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.

Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17658–17668, 2024.

Jacek Komorowski. MinkLoc3D: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1790–1799, 2021.

Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzciński. MinkLoc++: LiDAR and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2021.

Haowen Lai, Peng Yin, and Sebastian Scherer. AdaFusion: Visual-LiDAR fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4):12038–12045, 2022.

Anbang Liang, Zhipeng Chen, Wen Xiong, Fanyi Meng, Yu Yin, Dejin Zhang, and Qingquan Li. BdFusion: Bi-directional visual-LiDAR fusion for resilient place recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228:408–419, 2025.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3292–3310, 2022.

Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 891–898, 2013.

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*, pp. 5624–5633, 2019.

Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. OverlapTransformer: An efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition. *IEEE Robotics and Automation Letters*, 7(3):6958–6965, 2022.

Alexander Melekhin, Dmitry A. Yudin, Ilia Petryashin, and Vitaly D. Bezuglyj. MSSPlace: Multi-sensor place recognition with visual and text semantics. *IEEE Access*, 13:177098–177110, 2025.

Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. ConGeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, pp. 214–230. Springer, 2024.

Michael J. Milford and Gordon F. Wyeth. SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights. In *2012 IEEE International Conference on Robotics and Automation*, pp. 1643–1649. IEEE, 2012.

Manu S. Pillai, Mamshad Nayeem Rizve, and Mubarak Shah. GAReT: Cross-view video geolocalization with adapters and auto-regressive transformers. In *European Conference on Computer Vision*, pp. 466–483, 2024.

Zhangshuo Qi, Junyi Ma, Jingyi Xu, Zijie Zhou, Luqi Cheng, and Guangming Xiong. GSPR: Multimodal place recognition using 3D Gaussian splatting for autonomous driving. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8864–8871, 2025.

Krishna Regmi and Mubarak Shah. Video geo-localization employing geo-temporal feature learning and GPS trajectory smoothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12126–12135, 2021.

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*, pp. 4064–4072, 2020.

Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. CVLNet: Cross-view feature correspondence learning for video-based camera localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pp. 652–669, December 2022.

Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6488–6497, 2021.

Mikaela Angelina Uy and Gim Hee Lee. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4470–4479, 2018.

Shruti Vyas, Chen Chen, and Mubarak Shah. GAMa: Cross-view video geo-localization. In *European Conference on Computer Vision*, pp. 440–456. Springer, 2022.

Shan Wang, Chuong Nguyen, Jiawei Liu, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Kaihao Zhang, and Hongdong Li. View from above: Orthogonal-view aware cross-view localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14843–14852, 2024a.

Sijie Wang, Qiyu Kang, Rui She, Kai Zhao, Yang Song, and Wee Peng Tay. PRFusion: Toward effective and robust multi-modal place recognition with image and point cloud fusion. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20523–20534, 2024b.

Sijie Wang, Rui She, Qiyu Kang, Siqi Li, Disheng Li, Tianyu Geng, Shangshu Yu, and Wee Peng Tay. Multi-modal aerial-ground cross-view place recognition with neural ODEs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11717–11728, June 2025.

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*, pp. 3961–3969, 2015.

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

Dong Yuan, Frederic Maire, and Feras Dayoub. Cross-attention between satellite and ground views for enhanced fine-grained robot geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1249–1256, 2024.

Jinpeng Zhang, Yunzhou Zhang, Le Zhou, Li Wang, and Sizhan Wang. SeqMamba-MPR: A spatial-temporal Mamba network for place recognition using sequential multi-modal data. *IEEE Transactions on Intelligent Vehicles*, 10(9):4514–4527, 2025.

Qingwang Zhang and Yingying Zhu. Benchmarking the robustness of cross-view geo-localization models. In *European Conference on Computer Vision*, pp. 36–53. Springer, 2024.

Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2914–2923, 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*, 46(12):10419–10433, 2024.

Zijie Zhou, Jingyi Xu, Guangming Xiong, and Junyi Ma. LCPR: A multi-scale attention-based LiDAR-camera fusion network for place recognition. *IEEE Robotics and Automation Letters*, 9(2): 1342–1349, 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*, pp. 3640–3649, 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.

A Additional Experimental Details

A.1 Implementation Details

We use center-anchored, full-context windows with offsets $(-4, -2, 0, +2, +4)$ within a single data split. Aerial descriptors are precomputed and shared, with the aerial branch frozen. Matched-backbone variants use the same initialization procedure and learning-rate schedule; evaluation uses the model obtained at the end of training. In Tables 2, A1, and 5, temporal mean uses the corresponding trained single-observation model: each of the five observation descriptors is ℓ_2-normalized, averaged, and normalized again. No temporal parameters or temporal retraining are introduced.

Local features and readout. AACF uses radius $r = 6\,\text{m}$ and at most $L = 8$ neighbors per local query. We select up to 384 LiDAR points per frame by deterministic strides through input indices, then discard invalid projections. Relative poses are obtained from local odometry. Image features are sampled at the camera projection of each LiDAR point, without learned sampling offsets. Readout uses $\tau_r = 1$, $\lambda_s = 1$, and $\lambda_t = 0.1$, with temporal distances measured in sequence slots.
Image features are linearly projected to $D = 256$ dimensions and added to geometric embeddings from a $4 \to D \to D$ ReLU MLP. The latter takes transported coordinates and absolute temporal offset. For projected image feature $\mathbf{h}^I_{k,n}$, reliability is

$$\rho_{k,n} = \sigma\big(f_\rho\big([\mathbf{h}^I_{k,n}, |k-a|, \|\bar{\mathbf{x}}^a_{k,n}\|_2, v_{k,n}]\big)\big)\,, \tag{15}$$

where $v_{k,n} = 1$ for valid projections and f_ρ is a $(D+3) \to D \to 1$ ReLU MLP. Its radial input measures distance to the anchor origin. The residual gate uses a $2D \to D \to 1$ ReLU MLP followed by sigmoid; zero final-layer weights and bias -2 give an initial gate of 0.119.

Optimization and retrieval sampling. Training uses Adam for four epochs, 4,000 queries per epoch, and cache refresh every 4,000 queries. Backbone and refinement learning rates remain fixed at 10^{-6} and 2×10^{-5}. The default mini-batch contains six queries, with gradients accumulated over three steps. All four auxiliary-loss variants in Table 11 use mini-batches of nine queries and two-step gradient accumulation, giving an effective batch of 18. The auxiliary-loss ablations share the same random initialization, optimization schedule, RAMC objective, retrieval sampling, and evaluation protocol; only the inclusion of the local and masked-context losses varies. All four compute the global descriptor from the full observation window, with the anchor defining local readout coordinates and the retrieval target. Loss weights are $\lambda_{\text{corr}} = 0.5$, $\lambda_{\text{amc}} = 0.05$, $\lambda_{\text{local}} = 0.005$, and $\lambda_{\text{mask}} = 0.01$. Retrieval training uses triplet margin $\mu = 0.1$, $M = 10$ negatives per query, and 1,000 mining candidates. Training positives lie within 10 m, inside the 25 m evaluation region. The triplet loss averages over queries and sampled negatives.

Auxiliary objectives. Symmetric InfoNCE uses matched image–geometry pairs as positives and off-diagonal pairs within each mini-batch as negatives. The pairwise losses are weighted by detached reliability scores and normalized by the sum of these scores. All four auxiliary-loss variants use temperature 0.03. Gradient accumulation does not enlarge this negative pool. Masked-context consistency deterministically masks all four non-anchor observations in the five-frame window. All non-anchor images are set to zero and their LiDAR points removed before encoding, affecting both the global descriptor and local memory. The consistency target is the stop-gradient descriptor of the complete input, and batch-normalization running statistics remain fixed in this branch. This auxiliary objective differs from RAMC, which randomly removes context frames with probability 0.5 per query, as described below.

RAMC training distribution. During training, non-anchor frame removal is applied with probability 0.5 to each query. For each mini-batch, the number of frames to remove is sampled uniformly from $\{1, 2\}$. Non-anchor frames are selected by seeded sampling. Both branches share retrieval references. RAMC uses the specified loss weights from the first epoch. Only the clean query path is used at inference.

A.2 External Baseline Adaptation

External baselines are reimplemented or adapted to the present cross-view retrieval task. Adaptations accommodate the benchmark data, input modalities, temporal windows, and cross-view training and evaluation procedures. Results refer to these task adaptations, rather than exact reproductions of the original studies. Within each benchmark, all methods share the test queries, aerial reference database, positive-reference criterion, and recall computation. Input modalities and observation counts are

specified in the corresponding tables. On KITTI360-AG, evaluation uses 7,147 queries, 3,589 aerial references, and a 25 m positive radius. Each learnable external baseline is trained independently and uses neither AACF nor RAMC; the shared evaluation protocol does not imply identical training procedures across methods.

The image-based baselines (Sample4Geo, ConGeo, TransGeo, GeoDTR+, SALAD, BoQ, and SinGeo) are retrained and evaluated under this protocol; single-image methods use the center observation. TransGeo uses its transferable first stage, while SALAD and BoQ retain their original aggregation modules and objectives. Sequential dropout is omitted for CVIS/SeqGeo. GAReT uses its visual stream, and TrajLoc uses video inputs only.

For each temporal-mean control, single-frame retrieval and five-frame averaging use the same trained model weights. At inference, the five observation descriptors are individually ℓ_2-normalized, averaged, and normalized again. No temporal parameters or additional training are introduced. These paired comparisons thus isolate the effect of descriptor averaging while holding model weights and training fixed, whereas the external comparisons assess adapted methods under a common evaluation protocol. MinkLoc++, AdaFusion, LCPR, UMF, PRFusion++, MSSPlace-LI, and BdFusion adapt their ground-side multimodal models to the fixed aerial database. GSPR and SeqMamba-MPR use sequential multimodal adaptations. The matched AGPlace implementation provides the single-observation base for the matched-backbone comparison.

B Temporal Controls and Repeated-Run Analysis

B.1 Additional Parameter-Free Temporal Averaging Controls

The main paper evaluates parameter-free temporal averaging across single-observation RGB+LiDAR retrievers. Table A1 complements that analysis with RGB retrievers. Each paired control averages five normalized descriptors from a fixed trained retriever and renormalizes the result; it neither introduces temporal parameters nor requires temporal retraining.

On KITTI360-AG, temporal averaging improves R@1 for every RGB retriever, with gains from 0.67 to 2.19 points. Temporal mean with SinGeo achieves the highest R@1 among these controls at 37.05%, remaining 7.23 points below AnchorFuse under the same five-observation budget.

Table A1: Additional parameter-free temporal averaging controls for single-observation RGB retrievers on KITTI360-AG and nuScenes-AG. $T{=}1$ denotes center-observation retrieval, whereas mean $T{=}5$ averages five normalized descriptors from the same trained retriever and renormalizes the result. No temporal parameters or temporal retraining are introduced. Values are R@1 (%).

	KITTI360-AG			nuScenes-AG		
Method	$T{=}1$	Mean $T{=}5$	Δ	$T{=}1$	Mean $T{=}5$	Δ
Sample4Geo	27.99	28.78	+0.79	73.71	74.02	+0.31
ConGeo	32.25	33.05	+0.80	70.09	70.68	+0.59
TransGeo	22.04	24.23	+2.19	70.06	70.92	+0.86
GeoDTR+	34.19	36.09	+1.90	74.38	75.00	+0.62
SALAD	26.75	27.42	+0.67	54.39	58.11	+3.72
BoQ	27.72	29.70	+1.98	62.37	64.68	+2.31
SinGeo	35.79	37.05	+1.26	75.24	76.44	+1.20

Baseline entries use matched reimplementations or task-specific adaptations under common benchmark-specific query sets, aerial databases, and retrieval criteria. Δ is computed directly from the displayed two-decimal values as mean $T{=}5$ minus $T{=}1$.

B.2 Repeated-Run Analysis

We pair five runs by prespecified random seed. Within each comparison, training and evaluation settings remain fixed, with no seed-dependent tuning or best-run selection. Seeds govern stochastic training operations; pretrained initialization follows the same rule within each experimental group. All runs share the queries, aerial database, positive-reference criterion, and retrieval metrics.

Table A2: Independent five-seed replication with retraining and paired R@1 comparisons. Δ is the mean paired difference in percentage points, and *positive pairs* denotes the number of runs with a positive difference. These results are independent of the experiments in Table 2.

Comparison	Δ R@1 (pp)	Positive pairs
Temporal mean – Single-frame base	+1.99	5/5
AnchorFuse – AACF	+0.73	5/5

Table A2 reports an additional, independent five-seed replication with separate retraining. Table 2 reports the primary experiments, not averages over these five replication runs. For each seed, temporal mean and the single-frame base share the same independently retrained single-frame model; AACF and AnchorFuse are retrained separately and paired by seed. Mean gains are computed from the five paired differences and need not equal differences between main-table entries. Statistical inference applies to these replication runs alone.

For temporal mean versus the single-frame base, let d_i be the paired R@1 difference in percentage points for run i. We compute a two-sided 95% Student's t confidence interval as

$$\bar{d} \pm t_{0.975,4} \frac{s_d}{\sqrt{5}}, \tag{16}$$

where s_d is the sample standard deviation of the five paired differences and $t_{0.975,4} = 2.776$. The mean gain is 1.99 percentage points, with a confidence interval of $[1.53, 2.45]$. This interval quantifies uncertainty in the mean gain across seeds; it is not obtained by resampling test queries. AnchorFuse improves over AACF by an average of 0.73 percentage points, with positive differences in 5/5 seed-matched pairs. The positive-pair counts describe the direction of the observed differences across these five seeds and do not establish statistical significance.

Scope of the comparisons. Only aggregate recall is available for learned frame weighting, the nuScenes temporal baselines, and several external temporal controls in Table A1; comparisons involving these results are descriptive and do not establish statistical significance. Reported recall includes only completed evaluations with valid retrieval descriptors.

Multimodal temporal averaging. Table A3 gives the single-observation and five-frame recall values underlying the gains in Table 5.

Table A3: Effect of parameter-free temporal averaging on single-observation RGB+LiDAR retrievers. T=1 uses the center observation; mean T=5 averages and renormalizes five ℓ_2-normalized descriptors from the same trained retriever, without learnable temporal parameters or retraining. Values are R@1 (%), and Δ is mean minus single-observation recall, computed from the displayed values.

	KITTI360-AG			nuScenes-AG		
Method	T=1	Mean T=5	Δ	T=1	Mean T=5	Δ
MinkLoc++ (Komorowski et al., 2021)	27.18	28.29	+1.11	69.98	70.68	+0.70
AdaFusion (Lai et al., 2022)	27.33	28.74	+1.41	71.51	72.76	+1.25
LCPR (Zhou et al., 2024)	26.81	27.70	+0.89	57.21	59.05	+1.84
UMF (García-Hernández et al., 2024)	26.78	27.82	+1.04	68.31	70.96	+2.65
PRFusion++ (Wang et al., 2024b)	29.95	30.93	+0.98	73.94	75.29	+1.35
MSSPlace-LI (Melekhin et al., 2025)	24.73	26.19	+1.46	70.75	72.47	+1.72
BdFusion (Liang et al., 2025)	30.21	31.42	+1.21	74.18	75.41	+1.23
AGPlace (Wang et al., 2025)	35.06	36.22	+1.16	76.00	77.00	+1.00

Temporal-window comparison. Table A4 extends Table 4 with R@5 and R@10 for each window length, stride, and context direction.

Table A4: Ablation of temporal-window length, stride, and context direction on KITTI360-AG and nuScenes-AG. Delay denotes the maximum future offset, measured in frames relative to the anchor.

				KITTI360-AG			nuScenes-AG		
Window	T	Stride	Delay	R@1	R@5	R@10	R@1	R@5	R@10
Symmetric	3	2	2	28.54	44.84	54.33	77.75	87.45	90.15
Symmetric	5	1	2	<u>40.84</u>	58.78	67.04	<u>78.08</u>	<u>87.72</u>	<u>90.38</u>
Symmetric	5	2	4	**44.28**	**62.11**	**68.84**	**78.35**	**87.97**	**90.58**
Past-only	5	2	0	35.72	54.55	63.59	77.58	87.28	89.98
Symmetric	7	2	6	40.11	<u>60.24</u>	<u>67.59</u>	77.96	87.63	90.31

C Qualitative Analysis

Figure A1 shows the selected local evidence, readout weights, and retrieval results for a representative query.

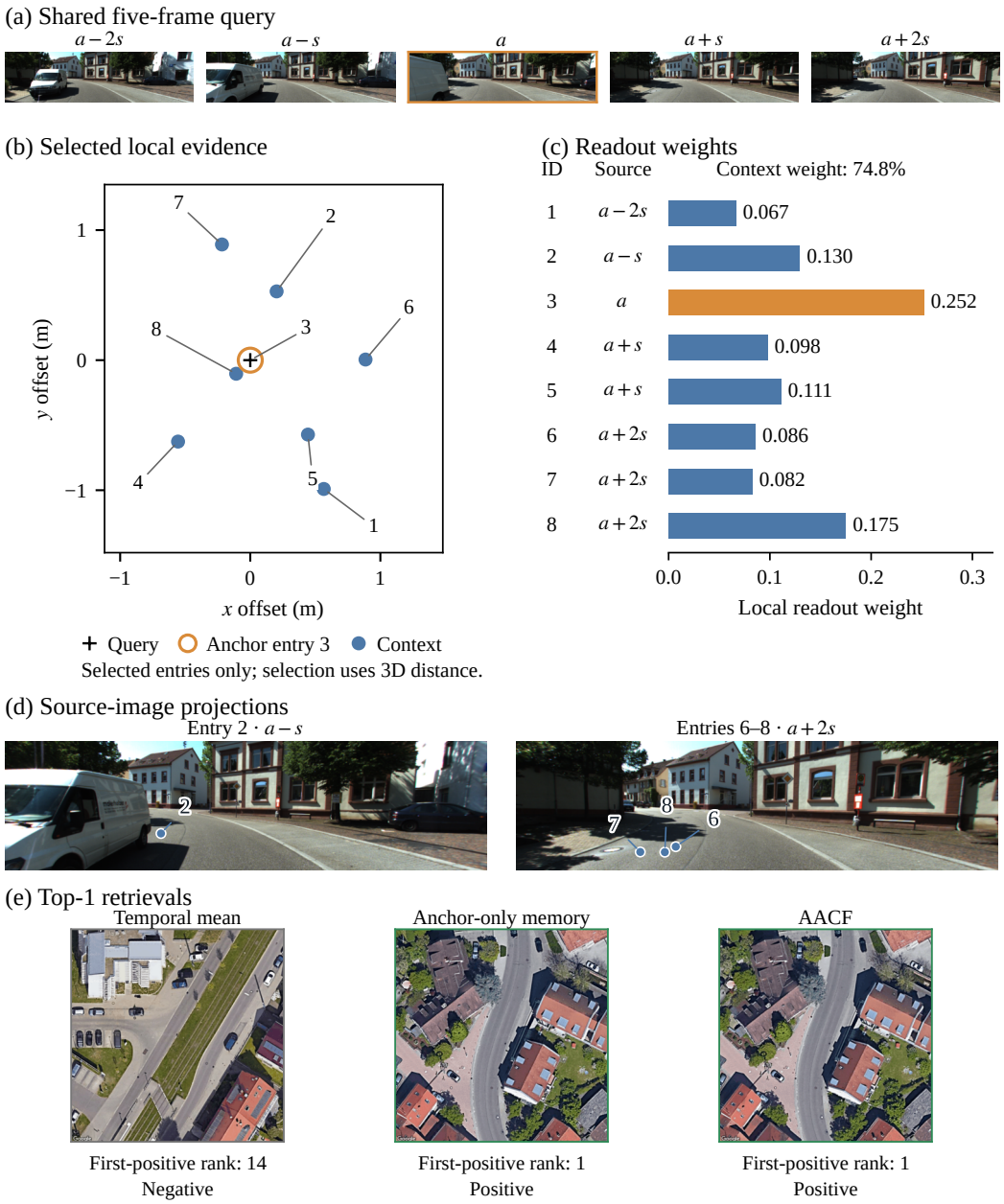

Figure A1: AACF readout for a representative anchor-local query. (b) A magnified x–y view shows the selected entries at their anchor-frame offsets; selection uses full 3D distances with radius 6 m, whose boundary lies outside this view. Anchor entry 3 coincides with the query. (c) Seven context entries carry 74.8% of this local readout weight; this weight distribution does not measure causal effects on retrieval. (d) Calibrated source-image projections for entry 2 and entries 6–8. (e) Top-1 retrievals; AACF and anchor-only memory both have first-positive rank 1 (temporal mean: 14).

Figure A2 compares temporal mean and AnchorFuse on the same five-frame query.

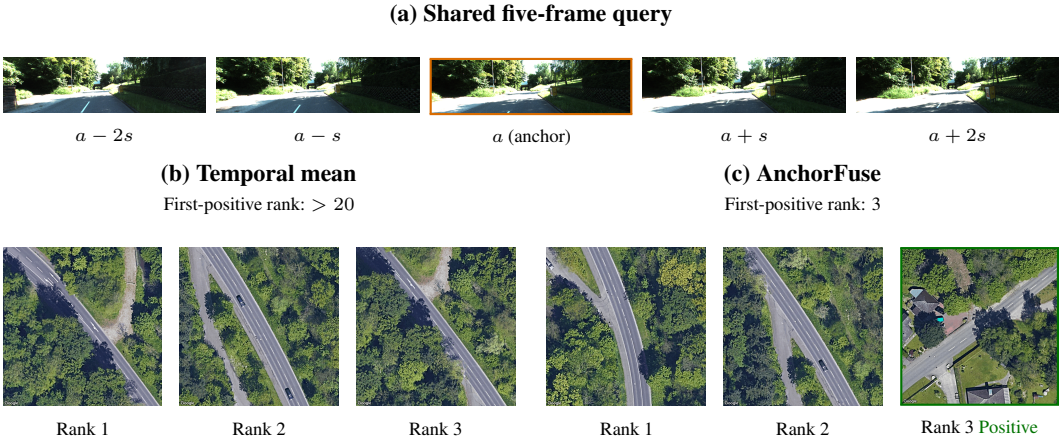

Figure A2: Retrieval for a shared five-frame query. Temporal mean has no positive in its top 20 results, whereas AnchorFuse returns a positive at rank 3; its top-1 result remains incorrect. Orange marks the anchor. Green borders identify positives within 25 m of the anchor.

Figure A3 compares retrieval with and without RAMC before and after removing the same two context frames.



Figure A3: RAMC under matched clean and missing-context inputs. Both models return their first positive at rank 3 on the clean query. Removing $a-2s$ and $a+s$ changes this rank to beyond 20 without RAMC and to 1 with RAMC. The models share the AACF inference architecture, database, and anchor-defined positive set, but differ in training. Green borders mark positives within 25 m. The rank improvement in this example does not establish an aggregate benefit from context removal.

D AACF DIAGNOSTICS

D.1 ANCHOR-REASSIGNMENT SENSITIVITY

We reassign the anchor within a fixed set of five observations, keeping the aerial database and evaluation settings unchanged. We retain only windows whose positive-reference sets are identical across all evaluated anchor designations. Temporal mean averages the same five descriptors and is therefore invariant to reassignment. Learned weighting conditions its frame weights on anchor-relative temporal positions, allowing both the weights and the descriptor to change. AACF and AnchorFuse use the designated anchor for local alignment and context readout. Table A5 reports the resulting descriptor and retrieval changes.

Table A5: Anchor-reassignment sensitivity on fixed five-observation windows with identical positive-reference sets across anchor designations. R@1 is reported in percent for the old, middle, and new anchor designations. Old–new cosine denotes cosine similarity between the corresponding query descriptors; top-1 shift measures displacement between their retrieved reference locations in meters.

Method	Old-anchor R@1	Mid-anchor R@1	New-anchor R@1	Old–new cosine	Top-1 shift (m)
Temporal mean	36.22	36.22	36.22	1.0000	0.00
Learned frame weighting	38.62	38.93	38.79	0.9986	0.63
AACF	40.57	42.79	41.88	0.9336	12.31
AnchorFuse	42.14	44.28	43.39	0.9461	10.58

Temporal mean remains invariant at the reported precision (cosine 1.0000; top-1 shift 0.00 m). Learned weighting changes slightly, whereas AACF and AnchorFuse exhibit lower descriptor similarity and larger top-1 displacement after reassignment. AnchorFuse achieves the highest R@1 under all three anchor designations.

D.2 POST-TRAINING SPARSE-READOUT SENSITIVITY

Table A6 and Figure A4 summarize the test-time sensitivity of AACF sparse readout. The same trained AACF model is used throughout: the readout radius r is varied with $L = 8$, and the neighbor budget L is varied with $r = 6$ m. The default setting is $r = 6$ m and $L = 8$.

Table A6: Post-training sensitivity of AACF sparse readout on KITTI360-AG. Model parameters remain fixed across all settings. Values are recall (%).

Parameter	Test-time setting	R@1	R@5	R@10
Radius ($L = 8$)	$r = 3.0$ m	40.46	58.12	66.34
	$r = 4.5$ m	42.19	59.34	67.73
	$r = 6.0$ m (default)	42.79	59.59	67.64
	$r = 7.5$ m	42.35	59.89	67.92
	$r = 9.0$ m	41.91	59.27	67.58
Neighbor budget ($r = 6$ m)	$L = 2$	40.30	58.43	66.13
	$L = 4$	41.72	59.00	67.38
	$L = 8$ (default)	42.79	59.59	67.64
	$L = 12$	42.27	60.03	67.99
	$L = 16$	41.96	59.73	67.80

The default setting achieves the highest R@1 in both comparisons. Reducing the radius to 3 m or the neighbor budget to two lowers R@1 by 2.33 and 2.49 percentage points, respectively. Increasing r to 7.5 m or L to 12 yields modest gains at R@5 and R@10, but does not improve R@1.

E ADDITIONAL RAMC ROBUSTNESS ANALYSIS

Table A7 summarizes retrieval under missing temporal and sensor context, with the anchor preserved.

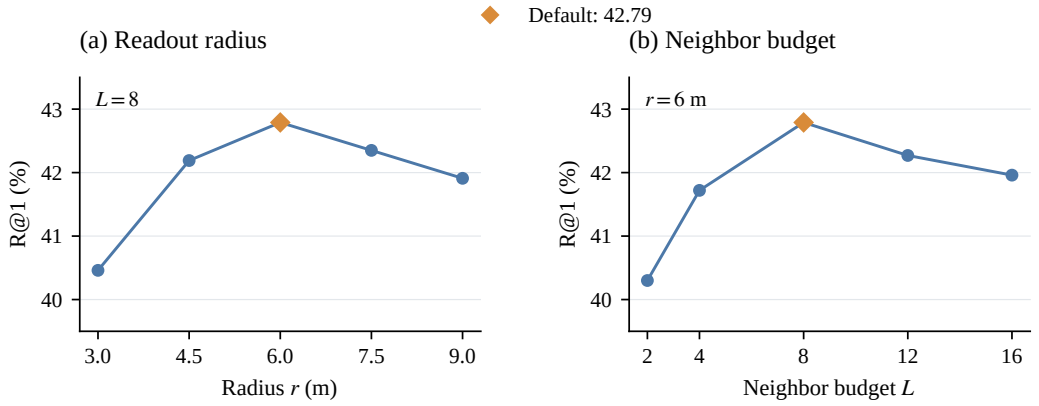

Figure A4: Test-time sparse-readout sensitivity of the same trained AACF model. Left: varying the radius with $L = 8$. Right: varying the neighbor budget with $r = 6$ m. Orange marks the default, which gives the highest R@1 among the evaluated settings; no retraining is performed across settings. Exact R@1/5/10 values appear in Table A6.

Table A7: Robustness to incomplete non-anchor context on KITTI360-AG. The multimodal anchor and its positive set remain fixed; Δ denotes the R@1 difference between AnchorFuse and the model trained without RAMC, computed from the displayed recalls.

Condition	w/o RAMC	AnchorFuse	Δ (pp)↑
Clean	42.79	44.28	+1.49
Incomplete temporal context			
Remove 1 context frame	32.99	41.67	+8.68
Remove 2 context frames	15.38	32.41	+17.03
Retain 1 context frame	6.02	17.63	+11.61
Context modalities removed			
Drop 50% context cameras	17.76	32.49	+14.73
Drop all context cameras	4.13	14.59	+10.46
Drop all context LiDAR	31.54	32.20	+0.66
Summary			
Mean corrupted R@1	17.97	28.50	+10.53
Worst-case R@1	4.13	14.59	+10.46

We evaluate two additional context corruptions: pose perturbation and complete temporal-outlier replacement (Table A8). Each corruption affects only non-anchor context, while the anchor observation and retrieval target remain unchanged.

Pose perturbation. Noise is applied only to the relative transforms from non-anchor frames to the anchor. The three translation components are sampled independently from a zero-mean Gaussian with standard deviation σ_t; yaw noise is sampled from a zero-mean Gaussian with standard deviation σ_ψ. The three severity levels use $(\sigma_t, \sigma_\psi) = (0.25\,\text{m}, 1^\circ)$, $(0.50\,\text{m}, 3^\circ)$, and $(1.00\,\text{m}, 5^\circ)$, respectively. Table A8 reports the highest severity (Level 3), with $\sigma_t = 1.00$ m and $\sigma_\psi = 5^\circ$. The noise transform left-multiplies the original relative transform. No roll or pitch perturbation is applied, and the anchor remains unchanged.

Temporal-outlier replacement. Donor windows are drawn from other queries in the same driving sequence and data split. Candidates must be at least 30 m away in planar query-position distance at severity level one, or 60 m at higher levels. Donors are selected using a fixed random seed. For each selected non-anchor slot, the corresponding donor slot supplies its RGB images, LiDAR points, relative pose, and temporal metadata. Donor poses retain their transforms relative to the donor window's anchor; they are not re-estimated relative to the current anchor. Complete replacement in Table A8 uses the 60 m minimum donor distance and affects all four non-anchor frames while preserving the anchor and its retrieval target. If no eligible donor exists, the query is evaluated without replacement.

Table A8: R@1 under context corruptions on KITTI360-AG. Recall is reported in percent; Δ denotes AnchorFuse minus AACF in percentage points, computed from the displayed recalls. Missing-context recalls and differences reproduce Table A7. All interventions preserve the anchor observation and its positive set.

Condition	AACF	AnchorFuse	Δ (pp)
Clean	42.79	44.28	+1.49
Incomplete temporal context			
Remove 1 context frame	32.99	41.67	+8.68
Remove 2 context frames	15.38	32.41	+17.03
Retain 1 context frame	6.02	17.63	+11.61
Context modalities removed			
Drop 50% context cameras	17.76	32.49	+14.73
Drop all context cameras	4.13	14.59	+10.46
Drop all context LiDAR	31.54	32.20	+0.66
Additional context corruptions			
Pose perturbation (Level 3)	27.82	31.47	+3.65
Complete temporal-outlier replacement ($\geq$ 60 m)	4.67	23.37	+18.70

AnchorFuse improves R@1 under both additional corruptions. The largest gain occurs under outlier replacement, although absolute R@1 remains 23.37%.

F Cross-Benchmark and Domain-Shift Diagnostics

F.1 Additional nuScenes-AG Analysis

Retrieval performance. We evaluate retrieval on nuScenes-AG with complete and partially removed context. The analysis is limited to aggregate recall, as per-query rankings are unavailable. Learned frame weighting provides only modest gains over uniform averaging on nuScenes-AG (Table 2): R@1 increases from 77.00% to 77.21%, while R@5 and R@10 improve by 0.70 and 0.29 percentage points, respectively. Both methods remain below AnchorFuse at all three recall cutoffs.

Table A9: Context-only robustness on nuScenes-AG. Each intervention modifies only non-anchor observations, while the complete six-camera–LiDAR anchor remains available. Values are R@1 (%); Δ denotes AnchorFuse minus AACF in percentage points, computed from the displayed recalls.

Condition	AACF	AnchorFuse	Δ (pp)
Clean	77.65	78.35	+0.70
Remove 2 context frames	77.31	78.18	+0.87
Retain 1 context frame	77.08	78.01	+0.93
Remove all context cameras	76.85	77.80	+0.95
Remove all context LiDAR	77.20	78.06	+0.86
Anchor only	76.95	77.91	+0.96

Context-only interventions. Context removal reduces R@1 by 0.34–0.80 percentage points for AACF and 0.17–0.55 points for AnchorFuse. AnchorFuse retains higher R@1 under every intervention, with gains of 0.86–0.96 points compared with 0.70 points on clean queries. These improvements are modest compared with those on KITTI360-AG.

F.2 Zero-Shot Transfer to Semantic Road Maps

We evaluate satellite-trained models on semantic road-map references without road-map-specific training, fine-tuning, or parameter selection. At test time, we replace the satellite database with geographically aligned road-map references and recompute their descriptors using the frozen aerial encoder. All network parameters, ground-query inputs, reference locations, and the 25 m positive radius remain fixed. Table A10 reports the resulting zero-shot retrieval performance.

Table A10: Zero-shot transfer from satellite imagery to semantic road maps on KITTI360-AG. Only the reference imagery changes; all model parameters and retrieval settings remain fixed. Values are recall (%).

Method	R@1	R@5	R@10
Single-frame base	31.15	45.61	53.71
Temporal mean	32.27	46.08	55.37
AACF	38.21	55.17	63.48
AnchorFuse	39.76	57.09	65.02

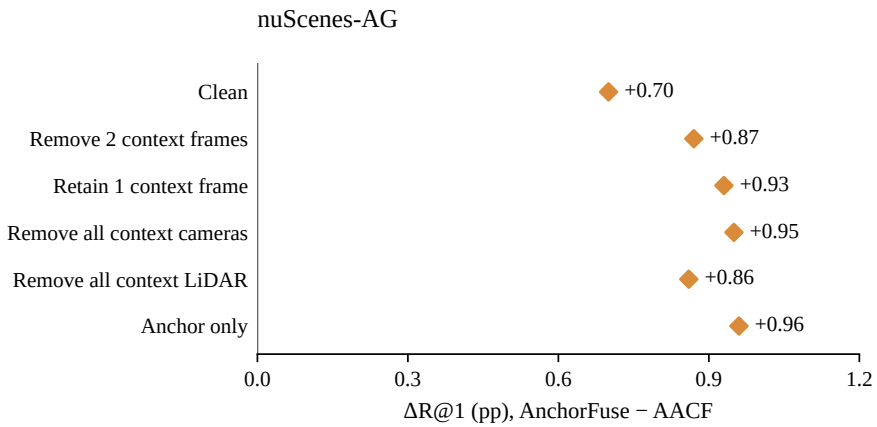

Figure A5: R@1 gains from RAMC on nuScenes-AG under context-only interventions. Each point shows AnchorFuse minus AACF in percentage points; absolute recalls appear in Table A9. The complete six-camera–LiDAR anchor remains available. Gains remain below one percentage point under all evaluated conditions.

AnchorFuse achieves the highest recall at all three cutoffs, improving R@1 over AACF and the single-frame base by 1.55 and 8.61 percentage points, respectively. The ordering of the four methods is preserved across the satellite and road-map reference domains.

G Computational Cost

Table A11 extends the mean single-query latency reported in Table 3 with 95th-percentile (P95) latency and throughput at batch sizes one and eight. Latency is measured per batch; at batch size one it is also the per-query latency. Measurements use an NVIDIA Tesla V100-SXM3 GPU with 32 GB of memory, FP32 precision, PyTorch 2.0.0, and CUDA 11.8. All models run in inference mode with inputs preloaded onto the GPU. Timing excludes data loading and offline aerial-database descriptor computation. After 50 warm-up passes, we measure 200 CUDA-synchronized forward passes.

Table A11: Inference latency and throughput at batch sizes one and eight. Mean and P95 denote forward latency per batch (ms); throughput is measured in queries/s. Batch-one means match Table 3.

	Batch 1			Batch 8		
Method	Mean (ms)	P95 (ms)	Query/s	Mean (ms)	P95 (ms)	Query/s
Single-frame base	13.37	34.82	74.79	20.31	51.37	393.89
Temporal mean	30.72	75.91	32.55	89.47	158.26	89.42
AACF	35.16	77.28	28.44	109.23	171.44	73.24
AnchorFuse	33.60	74.63	29.76	104.68	165.02	76.42
Past-only AnchorFuse	33.43	73.84	29.91	103.91	159.38	76.99

Throughput is computed as $1000B/t_B$, where B is batch size and t_B is mean batch latency in milliseconds. Increasing the batch size from one to eight raises AnchorFuse throughput from 29.76 to 76.42 queries/s, while mean batch latency rises from 33.60 to 104.68 ms. Thus, batching increases throughput at the cost of higher per-batch latency. AACF and AnchorFuse share the same inference architecture because RAMC is used only during training; their small latency differences reflect measurement variation.