GACLoc: Ground-Anchor Geometry Modeling for Cross-View Localization

Anonymous authors
Paper under double-blind review

Abstract

Cross-view geo-localization estimates the planar camera pose of a ground image with respect to geo-referenced satellite imagery. Recent correspondence-based cross-view localization methods enable interpretable camera pose recovery through explicit ground–satellite matches. However, learning such correspondences remains challenging because training data usually provide only camera-pose supervision, without pixel-level ground–satellite correspondence annotations. Cross-view correspondence further exhibits an inherent geometric asymmetry: under a known ground sampling distance (GSD), equal-sized satellite tokens correspond to approximately fixed metric regions, whereas perspective ground tokens correspond to depth-dependent regions whose size, shape, and orientation vary substantially. Moreover, in repetitive urban scenes, visual similarity alone is insufficient to establish reliable correspondences, as incorrect satellite locations can receive high matching scores. To address these issues, we propose GACLoc: Ground-Anchor Geometry Modeling for Cross-View Localization (GACLoc), a geometry-aware correspondence framework with two key components: Anchor Geometry Representation (AGR) and Geometry-Aware Local Correspondence Refinement (GALCR). AGR models the metric region geometry of ground anchors and uses it to construct geometry-based candidates and geometric features, while GALCR fuses visual and geometric features to obtain more reliable soft correspondences for weighted analytic pose recovery. Experiments demonstrate competitive localization accuracy under challenging settings, including cross-area evaluation and unknown orientation.

1 Introduction

Cross-view localization estimates the planar 3-DoF pose of a ground camera by matching a ground image against a geo-referenced satellite image covering the camera's surrounding area, typically recovering its 2D metric position and yaw within the local satellite reference. As a visual complement to Global Navigation Satellite Systems (GNSS), it is valuable for autonomous driving and mobile robotics in urban environments where GNSS measurements can be degraded by occlusion, multipath propagation, and signal attenuation. However, ground images are captured under perspective projection, whereas satellite images are approximately orthographic, leading to substantial differences in viewpoint and spatial geometry between the two views. These differences make reliable cross-view correspondence particularly challenging.

Existing cross-view localization methods can be broadly categorized into implicit and explicit correspondence-based approaches. Implicit methods estimate camera pose from global or dense representations, including global descriptor matching (Lentsch et al., 2023), BEV-transformed features and dense localization responses (Xia et al., 2024; Shi et al., 2023; Song et al., 2023; Wang et al., 2023), as well as orientation-invariant representations for large heading variations (Park et al., 2026). Although effective, such approaches typically embed localization evidence in global response structures or transformed feature representations, providing only limited correspondence-level interpretability. More recently, correspondence-based methods such as FG^2 and Loc^2 explicitly establish ground–satellite local matches and recover camera pose with analytic geometric solvers, offering a more interpretable alternative (Xia & Alahi, 2025; Xia et al., 2026). FG^2 constructs sparse correspondences in the BEV plane, whereas Loc^2 matches local ground and satellite features, uses

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.

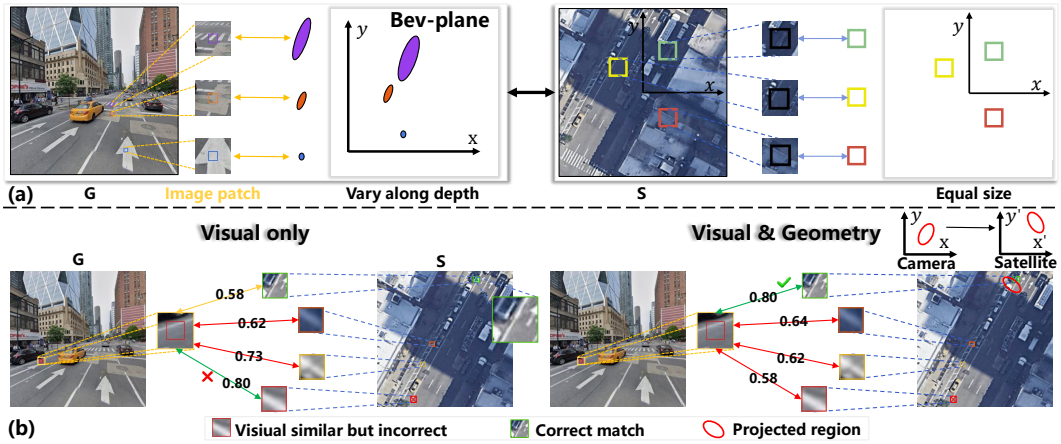

Figure 1: (a) Perspective ground patches map to BEV regions whose size and shape vary with depth, whereas satellite patches correspond to fixed metric regions under a known GSD. (b) Visual similarity alone can produce ambiguous and incorrect ground–satellite matches in repetitive scenes. Incorporating the projected anchor-region geometry provides complementary geometric evidence for distinguishing visually similar candidates and establishing more reliable correspondences.

monocular depth to map matched ground tokens to 2D metric coordinates in the ground frame, and recovers camera pose from the resulting metric correspondences.

Although explicit correspondence modeling has improved localization performance and interpretability, it still leaves two important limitations insufficiently addressed. First, in cross-view localization, equal-sized tokens in a perspective ground image can correspond to substantially different physical regions as scene depth varies, whereas equal-sized tokens in an approximately orthographic satellite image correspond to nearly uniform metric regions under a known ground sampling distance (GSD), as illustrated in Figure 1(a). Existing local correspondence-based methods largely overlook this asymmetry by representing the geometry of each matched ground token as a single 2D metric point, thereby overlooking the fact that the region of a ground token changes in size, shape, and orientation with scene depth. Second, existing local correspondence-based methods typically establish local matches primarily from visual feature similarity, without explicitly incorporating the geometry of the region of ground tokens into correspondence scoring. However, repetitive urban structures make visual matching inherently ambiguous. A ground token may retrieve multiple satellite regions with similar appearance, so correspondences established solely from visual similarity can be unreliable: a high visual similarity score does not necessarily indicate a correct match. As illustrated in Figure 1(b) left, several incorrect satellite candidates can exhibit high visual similarity to the same ground token. This ambiguity motivates the use of geometric information as complementary evidence for correspondence refinement.

To address these two limitations, we propose GACLOC, a geometry-aware correspondence-based framework for cross-view localization that follows a two-stage coarse-to-fine pipeline. In Stage I, coarse ground–satellite correspondences are established based on visual feature similarity to estimate a coarse camera pose. The top-N ground tokens with the highest visual matching scores are selected as *ground anchors*, and, for each ground anchor, the top-N_1 satellite candidates ranked by visual feature similarity are retained as visual candidates. In Stage II, we explicitly construct the geometry of each ground-anchor region to retrieve geometry-based candidates, combine them with the Stage I visual candidates to form the final candidate set, construct geometric features for the anchor and each anchor–candidate pair, and fuse these geometric features with visual features to rerank the candidates, favoring those that are geometrically compatible and visually plausible. To obtain the geometry-based candidates and geometric features, we introduce the *Anchor Geometry Representation* (AGR) module. AGR back-projects the metric-depth pixels within each ground anchor onto the ground plane and summarizes the resulting anchor region by its centroid and covariance, from which anchor-level geometric features are constructed. Using the coarse pose estimated in Stage I, the anchor region is projected into the satellite coordinate frame. AGR then retrieves the top-N_2 geometry-based satellite candidates with the smallest Mahalanobis distances to the projected anchor region and

constructs anchor–candidate geometric features that describe their relative geometric relationship. To identify candidates that are both geometrically compatible and visually plausible, we further introduce the *Geometry-Aware Local Correspondence Refinement* (GALCR) module. GALCR fuses the visual features with the anchor-level and anchor–candidate geometric features provided by AGR to rerank the retained satellite candidates. The most likely candidates are then aggregated into a soft correspondence for each ground anchor. The refined correspondences are assigned learned reliability weights and passed to a weighted analytic pose solver for final pose recovery, as illustrated in Figure 1(b) right. Overall, GACLoc complements visual matching with explicit anchor-region geometry, addressing both the geometric information loss of point-based representations and the ambiguity of visual correspondence estimation.

Our main contributions are summarized as follows: (1) We propose GACLoc, a new geometry-aware correspondence framework for cross-view localization that goes beyond point-only local matching by explicitly incorporating ground-anchor region geometry into correspondence estimation, yielding more reliable correspondences for analytic pose recovery. (2) We introduce two key modules: *Anchor Geometry Representation* (AGR) for modeling ground-anchor region geometry, and *Geometry-Aware Local Correspondence Refinement* (GALCR) for refining ambiguous local correspondences with geometric and visual cues. (3) Extensive experiments on VIGOR and KITTI demonstrate the strong effectiveness of GACLoc. On the VIGOR unknown-orientation cross-area setting, it reduces the mean localization and orientation errors by 21.5% and 18.7%, respectively. On KITTI under $\pm 180^\circ$ heading perturbation, it reduces the same-area mean and median localization errors by 24.9% and 31.3%, respectively.

2 Related Work

Global descriptor-based methods compare the ground-image descriptor with aerial descriptors extracted at different candidate poses (Lentsch et al., 2023; Xia et al., 2024; Park et al., 2026). SliceMatch (Lentsch et al., 2023) constructs pose-dependent aerial descriptors by aggregating aerial features according to the geometry of the ground-camera viewing frustum, and estimates the pose through descriptor similarity. VIRD (Park et al., 2026) further improves cross-view descriptor consistency through a dual-axis transformation, using polar transformation for horizontal alignment and context-enhanced positional attention for vertical alignment, followed by descriptor matching and pose refinement. While global descriptors provide compact representations for efficient pose search, they aggregate local evidence into holistic representations and do not explicitly expose ground–aerial local feature correspondences. **Geometry transformation-based methods** first transform ground features toward a BEV or geometrically aligned representation, and then search for the best alignment or estimate the relative pose from cross-view feature matching (Shi & Li, 2022; Shi et al., 2023; Song et al., 2023; Wang et al., 2023; Xia & Alahi, 2025). GGCVT (Shi et al., 2023) constructs an overhead representation from ground features using geometry-guided cross-view transformation and estimates camera rotation and translation from the aligned representations. DenseFlow (Song et al., 2023) instead estimates dense correspondences between ground BEV and aerial features through a learned flow field and recovers the camera pose from the resulting alignment. Related approaches employ geometric projection, homography estimation, or point-plane matching to reduce the large ground–aerial viewpoint gap (Shi & Li, 2022; Wang et al., 2023; Xia & Alahi, 2025). **Local feature matching** explicitly establishes local cross-view correspondences and recovers camera pose from the matched points. FG^2 (Xia & Alahi, 2025) lifts ground features into a 3D point cloud, selects features along the height dimension to construct a BEV point plane, and recovers the relative pose from sparse BEV–aerial correspondences through Procrustes alignment. Loc^2 (Xia et al., 2026) further avoids feature matching in BEV and directly learns correspondences between the ground and aerial image planes; matched ground points are lifted using monocular depth and the camera pose is recovered through scale-aware Procrustes alignment. However, BEV-based matching can be suboptimal under unknown orientation. As shown in Loc^2 (Xia et al., 2026), transforming ground features into BEV introduces distortions to elevated structures and discards information along the height dimension, weakening discriminative cues from above-ground structures such as building facades. This is particularly problematic in repetitive urban scenes, where ground-plane patterns such as roads can be visually ambiguous, leading to degraded matching and orientation estimation. Our method follows this interpretable local feature matching formulation, but moves beyond representing each selected ground feature as a single point by explicitly modeling its ground anchor region geometry and using this geometry for candidate retrieval and correspondence refinement before analytic pose recovery.

3 METHOD

Given a ground image G and its geo-referenced satellite image A, our goal is to estimate the camera's planar 3-DoF pose (x, y, θ), where (x, y) denotes its metric position and θ denotes its yaw. As illustrated in Figure 2, GACLOC follows a two-stage geometry-aware correspondence framework. In the first stage, following Loc^2 (Xia et al., 2026), we establish coarse ground–satellite correspondences based on visual matching scores, retain top-N ground tokens as *ground anchors*, and estimate a coarse pose $\hat{T}_0$. In the second stage, Anchor Geometry Representation (AGR) uses metric depth predicted by a frozen monocular depth estimator to model the metric geometry of each ground-anchor region on the camera-centric BEV plane, and projects the resulting region geometry into the satellite frame using the coarse pose. Based on the projected region geometry, AGR retrieves geometrically compatible satellite candidates using the Mahalanobis distance, combines them with visually retrieved candidates, and constructs geometric features for each anchor and its retained satellite candidates. Geometry-Aware Local Correspondence Refinement (GALCR) fuses the geometric features constructed by AGR with visual features to rerank the retained satellite candidates for each ground anchor. The top-L candidates are then selected and aggregated according to their refined scores to form a soft satellite correspondence. A lightweight correspondence-weighting head further predicts a confidence score c_i for each soft correspondence, which is directly used as its solver weight, i.e., $w_i = c_i$. The resulting weighted correspondences are finally passed to weighted scale-Procrustes (Umeyama, 1991) to recover the camera pose.

3.1 COARSE CORRESPONDENCE AND ANCHOR SELECTION

Following Loc^2 (Xia et al., 2026), we extract patch-level features using a frozen DINOv2 (Oquab et al., 2024) backbone and map them through independent trainable ground and satellite projectors to ℓ_2-normalized visual descriptors g_i and s_j. Their visual similarity is defined as

$$S_{ij}^{\text{vis}} = \frac{g_i^\top s_j}{\tau_{\text{app}}}, \tag{1}$$

where i and j index ground and satellite tokens, respectively. We apply Dual-Softmax matching followed by local soft-argmax to obtain a continuous coarse satellite correspondence $\hat{y}_i^{(0)}$ for each ground token. We retain N reliable correspondences and refer to their corresponding ground tokens as *ground anchors*. Let $\mu_i \in \mathbb{R}^2$ denote the metric center of ground anchor i in the camera ego frame, whose construction is detailed in Section 3.2. The selected coarse correspondences form

$$\mathcal{C}_0 = \left\{ \left(\mu_i, \hat{y}_i^{(0)}, w_i^{(0)} \right) \right\}_{i \in \mathcal{I}_0}, \qquad \hat{T}_0 = \left(\hat{s}_0, \hat{R}_0, \hat{t}_0 \right) = \text{WeightedScaleProcrustes}\left(\mathcal{C}_0 \right). \tag{2}$$

where $\mathcal{I}_0$ denotes the selected ground-anchor indices and $w_i^{(0)}$ is the coarse correspondence weight. For each selected anchor, we retain the top-N_1 satellite tokens according to S_{ij}^{vis} as its visual candidate set $\mathcal{K}_i^{\text{vis}}$. The N ground anchors, their visual features and metric coordinates, the visual candidate sets $\{\mathcal{K}_i^{\text{vis}}\}$, the visual matching scores S_{ij}^{vis}, and the coarse pose $\hat{T}_0$ are passed to Stage II for geometry-aware correspondence refinement.

3.2 ANCHOR GEOMETRY REPRESENTATION

For each ground anchor i, AGR explicitly models the metric geometry of its associated ground-image patch. Using metric depth predicted by a frozen monocular depth estimator, we back-project the K_i pixels within the anchor region into the camera ego frame and project them onto the horizontal BEV plane. The resulting BEV points, their metric center, and covariance are

$$\mathcal{P}_i = \{p_{ik}\}_{k=1}^{K_i}, \qquad \mu_i = \frac{1}{K_i} \sum_{k=1}^{K_i} p_{ik}, \qquad \Sigma_i = \frac{1}{K_i} \sum_{k=1}^{K_i} (p_{ik} - \mu_i)(p_{ik} - \mu_i)^\top. \tag{3}$$

Here, μ_i specifies the metric center of the anchor region, while Σ_i captures its spatial extent, anisotropy, and orientation. Compared with representing a ground token by a single metric point, Σ_i additionally captures the spatial extent, anisotropy, and orientation of its corresponding physical

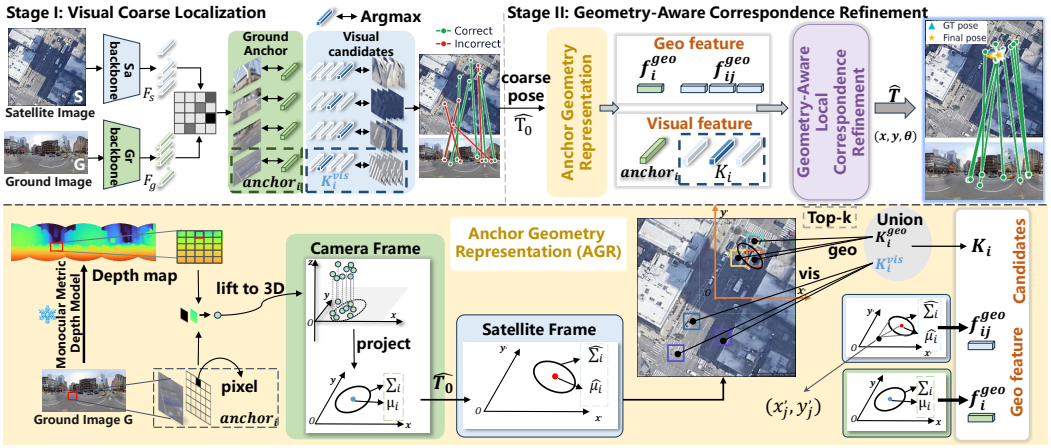

Figure 2: Overview of GACLOC. Stage I uses visual local matching to estimate a coarse pose and retain reliable ground anchors with visual satellite candidates. In Stage II, AGR reconstructs each anchor's metric BEV region from depth and projects it into the satellite frame using the coarse pose. Mahalanobis-based geometry candidates are merged with visual candidates, and AGR constructs anchor-level and anchor–candidate geometric features. GALCR fuses visual and geometric features to rerank the candidates and aggregates the top-L into a soft correspondence. Confidence-weighted correspondences are finally passed to weighted scale-Procrustes for pose recovery.

region. Using the coarse pose $\hat{T}_0$, we project the anchor-region geometry into the satellite metric frame:

$$\hat{\mu}_i = \hat{s}_0 \hat{R}_0 \mu_i + \hat{t}_0, \qquad \hat{\Sigma}_i = \hat{s}_0^2 \hat{R}_0 \Sigma_i \hat{R}_0^\top. \tag{4}$$

Here, $\hat{\mu}_i$ and $\hat{\Sigma}_i$ describe the predicted center and region geometry of anchor i in the satellite frame. We then use the projected region geometry to select geometrically compatible satellite candidates. For a satellite token j, we obtain its metric coordinate y_j in the local satellite frame from its grid location and the known ground sampling distance (GSD) of the geo-referenced satellite image. We then measure its geometric compatibility with anchor i using the Mahalanobis distance

$$d_{ij}^2 = (y_j - \hat{\mu}_i)^\top \hat{\Sigma}_i^{-1} (y_j - \hat{\mu}_i). \tag{5}$$

For each ground anchor i, we construct two complementary satellite candidate sets. $\mathcal{K}_i^{\text{geo}}$ contains the N_2 satellite tokens with the smallest Mahalanobis distances to the projected anchor region, whereas $\mathcal{K}_i^{\text{vis}}$ contains the N_1 tokens with the highest Stage-I visual matching scores. Their union forms the retained candidate set:

$$\mathcal{K}_i^{\text{vis}} = \text{TopK}_{N_1}\left(\{S_{ij}^{\text{vis}}\}_j\right), \qquad \mathcal{K}_i^{\text{geo}} = \text{TopK}_{N_2}\left(\{-d_{ij}^2\}_j\right), \qquad \mathcal{K}_i = \mathcal{K}_i^{\text{vis}} \cup \mathcal{K}_i^{\text{geo}}. \tag{6}$$

This allows geometrically plausible satellite tokens to be retained even when they are not highly ranked by visual similarity. For subsequent correspondence refinement, AGR further constructs geometric features at both the anchor and anchor–candidate levels:

$$f_i^{\text{geo}} = \left[\log a_i,\ \Sigma_{i,xx},\ \Sigma_{i,yy},\ \Sigma_{i,xy},\ u_i,\ r_i\right]. \tag{7}$$

$$f_{ij}^{\text{geo}} = \left[\Delta x_{ij},\ \Delta y_{ij},\ \widetilde{\Delta x}_{ij},\ \widetilde{\Delta y}_{ij},\ d_{ij}^2,\ \log\det\hat{\Sigma}_i,\ S_{ij}^{\text{vis}}\right], \qquad j \in \mathcal{K}_i. \tag{8}$$

The anchor-level feature f_i^{geo} provides a compact description of the geometry and reliability of the anchor region, while the anchor–candidate feature f_{ij}^{geo} encodes how each retained satellite candidate is positioned relative to this region. These features are designed to make the modeled region geometry directly usable for correspondence refinement: the anchor-level statistics summarize the scale, shape, orientation, and uncertainty of the region, whereas the anchor–candidate statistics describe candidate displacement and compatibility under this region-specific geometry. In this way, GALCR can evaluate candidates not only by visual similarity, but also according to their geometric relation to the physical region represented by each ground anchor. Detailed definitions and normalization are provided in Appendix C.2. AGR therefore outputs the retained candidate set $\mathcal{K}_i$, the anchor geometric feature f_i^{geo}, and the corresponding anchor–candidate geometric features $\{f_{ij}^{\text{geo}}\}_{j\in\mathcal{K}_i}$, which are subsequently used by GALCR for local correspondence refinement.

3.3 GEOMETRY-AWARE LOCAL CORRESPONDENCE REFINEMENT

Given the retained candidate set $\mathcal{K}_i$ and the geometric features constructed by AGR, GALCR refines the local correspondence for each ground anchor i by jointly exploiting visual and geometric information. For each anchor i and its retained satellite candidate $j \in \mathcal{K}_i$, we separately encode the anchor-level feature f_i^{geo} and the anchor–candidate feature f_{ij}^{geo}, and fuse them with the ground visual feature g_i and satellite visual feature s_j, respectively:

$$q_i = \mathrm{Linear}_q\left[g_i,\, \mathrm{MLP}_{\mathrm{geo}}(f_i^{\mathrm{geo}})\right], \qquad h_{ij} = \mathrm{Linear}_c\left[s_j,\, \mathrm{MLP}_{\mathrm{pair}}(f_{ij}^{\mathrm{geo}})\right]. \tag{9}$$

Here, q_i is the geometry-aware representation of ground anchor i, while h_{ij} represents satellite candidate j conditioned on its geometric relation to the anchor. This fusion enables the modeled anchor-region geometry to directly rerank candidate matching scores, promoting geometrically compatible candidates that may have relatively weak visual similarity while suppressing visually strong but geometrically inconsistent candidates. A local cross-attention module uses q_i as the query and the candidate representations $\{h_{ij}\}_{j\in\mathcal{K}_i}$ as keys and values. Its candidate-wise outputs are linearly projected to obtain the matching-score corrections:

$$\left\{\Delta S_{ij}^{\mathrm{local}}\right\}_{j\in\mathcal{K}_i} = \mathrm{Linear}_{\mathrm{score}}\left(\mathrm{CA}\left(q_i,\, \{h_{ij}\}_{j\in\mathcal{K}_i}\right)\right). \tag{10}$$

The candidate-wise corrections are added residually to the original visual matching scores to obtain the refined scores S_{ij}^{ref}. For each ground anchor i, we then select the top-L candidates according to S_{ij}^{ref}, forming $\mathcal{J}_i \subseteq \mathcal{K}_i$. Their refined scores are normalized into a local matching distribution p_{ij}, from which the final soft satellite correspondence $\hat{y}_i$ is obtained by probability-weighted aggregation:

$$S_{ij}^{\mathrm{ref}} = S_{ij}^{\mathrm{vis}} + \Delta S_{ij}^{\mathrm{local}}, \qquad \mathcal{J}_i = \mathrm{TopK}_L\left(\{S_{ij}^{\mathrm{ref}}\}_{j\in\mathcal{K}_i}\right). \tag{11}$$

$$p_{ij} = \mathrm{softmax}_{j\in\mathcal{J}_i}\left(S_{ij}^{\mathrm{ref}}/\tau_{\mathrm{ref}}\right), \qquad \hat{y}_i = \sum_{j\in\mathcal{J}_i} p_{ij} y_j. \tag{12}$$

Since the quality of the resulting soft correspondences can vary across anchors, we further employ a lightweight correspondence-weighting head. For each ground anchor i, the head aggregates the refined visual and geometric matching information of its top-L candidates according to the local matching distribution p_{ij}, and combines it with sample-level correspondence context to predict a confidence score c_i. We directly use c_i as the solver weight w_i for the corresponding soft correspondence. Detailed implementation configurations are provided in Appendix D.2.

After GALCR, each retained ground anchor is associated with a soft satellite correspondence $\hat{y}_i$ and a correspondence weight w_i. We select the top-N_f ground–satellite correspondence pairs according to their confidence scores, form the weighted correspondence set $\mathcal{C}_f$, and pass them to the weighted scale-Procrustes solver for final pose recovery:

$$\mathcal{C}_f = \{(\mu_i, \hat{y}_i, w_i)\}_{i\in\mathcal{I}_f}, \qquad \hat{T} = \mathrm{WeightedScaleProcrustes}\left(\mathcal{C}_f\right). \tag{13}$$

3.4 LEARNING OBJECTIVES

Stage I. We first train the coarse correspondence branch following the training strategy of Loc^2 (Xia et al., 2026). Specifically, we use the bidirectional NCE loss $\mathcal{L}_{\mathrm{NCE}}$ to learn cross-view visual correspondences, and follow FG^2 (Xia & Alahi, 2025) to supervise the coarse pose using the Virtual Correspondence Error (VCE) $\mathcal{L}_{\mathrm{coarse}}$. Motivated by our observation that a ground token corresponds to a spatial region rather than a single point, we further introduce a region matching loss $\mathcal{L}_{\mathrm{region}}$ to encourage the predicted visual matching distribution to cover the corresponding satellite region, instead of supervising only a single point correspondence. The Stage-I objective is

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{NCE}} + \mathcal{L}_{\mathrm{region}} + \mathcal{L}_{\mathrm{coarse}}. \tag{14}$$

Stage II. In Stage II, we freeze the Stage-I feature projectors and train GALCR together with the correspondence-weighting branch. We introduce a candidate loss $\mathcal{L}_{\mathrm{cand}}$ to supervise the refined local candidate distribution and encourage geometrically compatible satellite candidates to receive higher matching scores. A confidence loss $\mathcal{L}_{\mathrm{conf}}$ supervises the predicted correspondence confidence c_i, so that more accurate soft correspondences receive larger solver weights w_i. Finally, a pose loss $\mathcal{L}_{\mathrm{pose}}$ applies VCE to the pose recovered by weighted scale-Procrustes, providing end-to-end supervision for the final localization result. The Stage-II objective is

$$\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{cand}} + \mathcal{L}_{\mathrm{conf}} + \mathcal{L}_{\mathrm{pose}}. \tag{15}$$

Detailed target construction and loss configurations are provided in Appendix E.

4 EXPERIMENTS

4.1 DATASETS AND METRICS

VIGOR (Zhu et al., 2021a) contains panoramas and geo-referenced satellite images from four US cities. We use its corrected splits: same-area uses all four cities, while cross-area trains on New York and Seattle and tests on Chicago and San Francisco. For known/unknown orientation, we report mean and median localization error; the unknown setting additionally reports yaw error, following prior work (Xia & Alahi, 2025; Xia et al., 2026). **KITTI** (Geiger et al., 2013) provides forward-facing ground-level images with a limited field of view. We follow Loc^2 (Xia et al., 2026) and use Test1 and Test2 as the same-area and cross-area test sets, respectively. Orientation noise is sampled from $\pm 10^\circ$ or $\pm 180^\circ$ during training and evaluation. We report the mean and median localization and orientation errors.

4.2 IMPLEMENTATION DETAILS

For VIGOR, each panorama is converted into four horizontal perspective views; KITTI uses the forward-facing left color camera. We use frozen DINOv2 ViT-L/14 (Oquab et al., 2024) features and metric depth from Depth Anything 3 (Lin et al., 2026). We train on four RTX 3090 GPUs using AdamW (Loshchilov & Hutter, 2019), batch size 8, learning rate 10^{-4}, and weight decay 10^{-4}. Stage I is trained for 25/50 epochs on VIGOR/KITTI; Stage II is initialized from the trained Stage I model and optimized for 50 epochs, while the Stage-I feature projectors remain frozen throughout Stage-II training. In Stage I, we retain 1,024 ground anchors and the top-8 visual satellite candidates for each anchor. In Stage II, we additionally retrieve 24 geometry-based candidates for each anchor and merge them with the visual candidates. GALCR reranks the retained candidates, aggregates the top-8 candidates into a soft correspondence, and selects the top-256 correspondences according to their predicted confidence for final pose recovery. All loss weights are set to one. Excluding frozen backbones, VIGOR/KITTI have 8.95M/7.37M trainable parameters. Main results apply RANSAC (Fischler & Bolles, 1981) only to the coarse solve; the final solver is weighted scale-Procrustes. Ablations use weighted solvers without RANSAC.

4.3 QUANTITATIVE RESULTS

Table 1: VIGOR test results. GACLOC reports the final pose with RANSAC only in the Stage I coarse solve. Best results within each orientation-and-area block are bold, and second-best results are underlined.

Ori.	Method	Cross-area				Same-area			
		Localization (m)		Orientation (°)		Localization (m)		Orientation (°)	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Unknown	SliceMatch (2023)	7.22	3.31	25.97	4.51	6.49	3.13	25.46	4.71
	CCVPE (2024)	5.41	1.89	27.78	13.58	3.74	1.42	12.83	6.62
	DenseFlow (2023)	7.67	3.67	17.63	2.94	4.97	1.90	11.20	1.59
	FG^2 (2025)	10.02	8.14	31.41	5.45	8.95	7.32	15.02	2.94
	Loc^2 (2026)	4.23	2.09	11.67	2.21	3.94	1.78	9.54	2.00
	VIRD (2026)	4.61	1.55	16.50	**1.17**	3.00	**1.20**	9.52	**0.96**
	GACLOC (Ours)	**3.32**	**1.51**	**9.49**	2.09	**2.97**	1.31	**7.74**	1.86
Known	SliceMatch (2023)	5.53	2.55	–	–	5.18	2.58	–	–
	CCVPE (2024)	4.97	1.68	–	–	3.60	1.36	–	–
	GGCVT (2023)	5.16	1.40	–	–	4.12	1.34	–	–
	DenseFlow (2023)	5.01	2.42	–	–	3.03	**0.97**	–	–
	HC-Net (2023)	3.35	1.59	–	–	2.65	1.17	–	–
	FG^2 (2025)	**2.41**	**1.37**	–	–	**1.95**	1.08	–	–
	Loc^2 (2026)	3.43	1.90	–	–	3.06	1.59	–	–
	VIRD (2026)	3.85	1.47	–	–	2.57	1.14	–	–
	GACLOC (Ours)	2.97	1.42	–	–	2.46	1.20	–	–

VIGOR. Table 1 shows that GACLOC achieves state-of-the-art mean localization error under unknown orientation on both cross-area and same-area evaluation. On cross-area evaluation, it achieves 3.32 m mean localization error and 9.49° mean yaw error, reducing the corresponding Loc^2 results by 21.5% and 18.7%. GACLOC also achieves state-of-the-art mean localization and yaw

errors on same-area evaluation. Under known orientation, FG^2 achieves lower mean localization errors, whereas the median errors of GACLoc remain close to the state of the art. This gap is partly related to the dependence of Stage II on the Stage-I coarse pose when constructing geometry-based candidates: an accurate coarse pose allows geometrically compatible candidates to be effectively retrieved, whereas a larger coarse-pose error can shift the projected anchor regions away from the correct satellite locations and limit the benefit of Stage II. This dependence is further quantified by the Stage-I error stratification in Appendix A.2; see Tables 5 and 6.

Table 2: KITTI results under the Loc^2 evaluation protocol. Baseline results are taken from Loc^2 (Xia et al., 2026). For GACLoc, RANSAC is applied only to the Stage I coarse pose estimation. Best results are bold and second-best results are underlined within each area-and-noise block.

	Ori.	Method	Loc. (m) ↓		Lateral (%) ↑		Long. (%) ↑		Orien. (°) ↓		Orien. (%) ↑	
			Mean	Median	R@1m	R@5m	R@1m	R@5m	Mean	Median	R@1°	R@5°
Cross-area	±180°	SliceMatch	14.85	11.85	24.00	72.89	7.17	33.12	**23.64**	7.96	**31.69**	31.69
		CCVPE	13.94	10.98	23.42	60.46	11.81	42.12	77.84	63.84	3.14	14.56
		Loc^2	11.71	9.11	13.64	50.88	14.04	50.73	55.18	33.23	2.53	12.78
		VIRD	12.30	7.05	**43.61**	81.86	12.88	46.83	25.10	**2.22**	27.65	**69.86**
		GACLoc	**10.71**	**5.52**	36.63	**82.56**	**19.84**	**54.53**	33.99	6.70	10.94	42.97
	±10°	CCVPE	9.16	3.33	44.06	90.23	23.08	64.31	**1.55**	**0.84**	**57.72**	**96.19**
		HC-Net	8.47	4.57	**75.00**	**97.76**	**58.93**	**76.46**	3.22	1.63	33.58	83.78
		DenseFlow	7.97	3.52	54.19	91.74	23.10	61.75	2.17	1.21	43.44	89.31
		FG^2	7.31	4.15	37.89	85.65	21.98	60.77	3.62	2.37	23.03	77.84
		Loc^2	**5.60**	3.01	45.29	92.43	27.01	68.26	3.32	2.12	26.03	80.68
		GACLoc	6.56	**2.65**	55.65	95.08	29.45	71.78	2.71	1.63	33.36	86.29
Same-area	±180°	SliceMatch	7.96	4.39	49.09	98.52	15.19	57.35	4.12	3.65	13.41	64.17
		CCVPE	6.88	3.47	53.30	85.13	25.84	68.49	15.01	6.12	8.96	42.75
		Loc^2	1.85	1.31	62.18	97.80	59.37	**97.48**	9.70	6.17	9.46	41.64
		VIRD	5.47	2.07	79.46	98.30	31.65	75.78	5.16	**1.02**	**49.32**	92.95
		GACLoc	**1.39**	**0.90**	**87.91**	**99.79**	**68.57**	97.24	**1.92**	1.35	39.04	**94.59**
	±10°	CCVPE	1.22	0.62	97.35	99.71	77.13	97.16	0.67	0.54	77.39	**99.95**
		HC-Net	0.80	0.50	**99.01**	99.73	92.20	**99.25**	**0.45**	0.33	**91.35**	99.84
		DenseFlow	1.48	**0.47**	95.47	99.79	87.89	94.78	0.49	**0.30**	89.40	99.31
		FG^2	**0.75**	0.51	95.81	99.66	**92.50**	99.05	0.93	0.66	67.27	98.91
		Loc^2	1.13	0.77	93.59	99.97	71.51	98.17	1.97	1.43	36.68	92.84
		GACLoc	1.01	0.66	96.58	**100.00**	77.60	98.54	1.07	0.76	62.18	98.62

KITTI. Table 2 shows that GACLoc achieves state-of-the-art mean localization error under ±180° orientation noise on both Test1 and Test2. On Test1, it reduces the Loc^2 mean/median localization errors by 24.9%/31.3%. On Test2, the corresponding errors decrease from 11.71/9.11 m to 10.71/5.52 m. Under ±10° orientation noise on Test2, GACLoc yields a higher mean localization error than Loc^2, while achieving a lower median error and better localization recalls, indicating that the degradation is mainly caused by a small number of large-error cases rather than a general loss of localization accuracy. As discussed above, inaccurate Stage-I coarse poses can limit the benefit of Stage II by degrading the retrieval of geometry-based candidates.

Table 3: Ablation studies. Reg. loss denotes the proposed region matching loss; Geo. retrieval denotes the inclusion of geometry-based candidates retrieved by AGR; Geo. fusion denotes the fusion of AGR geometric features with visual features in GALCR for candidate reranking. The visual baseline retains the Stage I architecture but disables region supervision and Stage II.

Variant	Reg. loss	Geo. retrieval	Geo. fusion	Localization (m) ↓		Orientation (°) ↓	
				Mean	Median	Mean	Median
Visual baseline	–	–	–	3.822	2.225	9.424	3.474
+ Region loss	✓	–	–	3.571	1.975	8.310	2.819
+ Geometry-based candidates	✓	✓	–	3.416	1.750	9.215	2.841
+ Geo. feature fusion(GACLoc)	✓	✓	✓	**3.080**	**1.529**	**7.206**	**2.278**

4.4 Visualization

Figure 3 clearly demonstrates the effectiveness of GACLoc in alleviating visual mismatches through geometry-aware refinement. The visualization is conducted on VIGOR under the unknown-

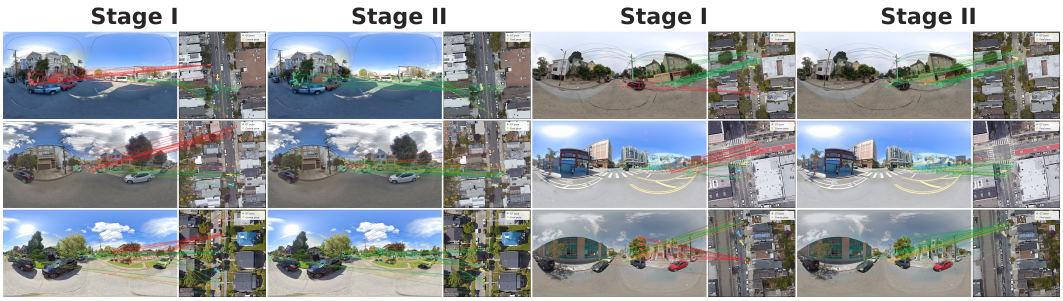

Figure 3: Qualitative comparison on VIGOR. Stage I and Stage II show the coarse and refined correspondences, respectively. Green, yellow, and red lines denote small, moderate, and large matching errors. The yellow star and blue triangle mark the predicted and GT poses. Stage II visibly reduces red mismatches. Additional KITTI results are provided in Appendix F.1.

orientation, same-area setting, without RANSAC in either coarse or final pose estimation. Each row contains two samples, each shown in two columns: Stage I coarse pose on the left and the corresponding Stage II final pose on the right. We visualize the top-64 correspondences ranked by their predicted solver weights. The connecting lines indicate correspondences between ground anchors and satellite candidate tokens, where green, yellow, and red denote accurate, moderately erroneous, and severely erroneous matches, respectively. Stage I often produces visually plausible but incorrect matches in repetitive urban scenes, whereas Stage II uses anchor-region geometry to correct many of these mismatches and move the correspondences toward more accurate satellite locations. The refined results exhibit stronger geometric consistency, demonstrating the effectiveness of the proposed refinement.

4.5 Ablation Studies

All ablations use the VIGOR unknown-orientation, same-area test split without RANSAC in both solvers. Table 3 reports the main ablation results, while Appendix A provides controlled candidate-level diagnostics, operating-boundary analyses, and hyperparameter sensitivity. **Region loss.** As shown in Table 3, introducing the proposed region loss consistently improves all pose metrics, demonstrating the importance of modeling the spatial region represented by each ground token in cross-view correspondence learning. **Geometry-based candidates and feature fusion.** Table 3 shows that introducing geometry-based candidates while using only visual features in GALCR yields only limited improvement. In contrast, when geometric features are further fused with visual features in GALCR to rerank the retrieved candidates, the performance improves substantially. This indicates that geometry-based candidate retrieval and geometry-aware candidate reranking are complementary: the former expands the candidate set with more geometrically plausible hypotheses, while the latter identifies more reliable correspondences from these candidates for final pose recovery. Appendix Table 4 further isolates these two effects: under a fixed checkpoint, geometry-based retrieval improves candidate coverage, whereas enabling feature fusion under the same candidate construction substantially improves the inlier ratio of the correspondences passed to the final solver. In addition, Appendix Table 9 shows that Stage II reduces mean, median, and P90 errors for both translation and orientation, indicating that the gain is not driven only by a small set of outliers.

5 Conclusion

We propose GACLoc, which extends point-based cross-view correspondence matching by explicitly modeling the anchor-region geometry represented by each ground anchor. The proposed geometry-based candidate retrieval and geometry-aware candidate reranking jointly improve correspondence quality for analytic pose recovery. As expected from the coarse-to-fine design, Stage II depends on a reasonably accurate Stage-I pose for geometry-based candidate retrieval, and very large anchor regions remain more difficult to refine. Experiments on VIGOR and KITTI demonstrate strong localization performance under challenging settings, including cross-area generalization and unknown orientation.

AI USE STATEMENT

Generative AI tools were used solely for English grammar checking and language polishing to improve the clarity and readability of the manuscript. All suggested revisions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

This work uses existing public cross-view geo-localization benchmarks and does not collect new data or involve human-subject experiments. The method is intended for benchmarked camera localization. As with other localization systems, deployment should respect applicable privacy, consent, and data-use requirements.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, coordinate conventions, training objectives, localization procedure, and evaluation protocols in the main paper and appendix. The experimental settings, datasets, and evaluation metrics used to obtain the reported results are also specified to facilitate reproducibility.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. PointDSC: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15859–15869, 2021.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A trainable CNN for joint description and detection of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8092–8101, 2019.

Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12426–12434, 2020.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. doi: 10.1177/0278364913491297.

Richard I. Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004. doi: 10.1017/CBO9780511811685.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. doi: 10.1364/JOSAA.4.000629.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753, 2019.

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

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, pp. 71–91, 2024. doi: 10.1007/978-3-031-73220-1_5.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pp. 1–18, 2022. doi: 10.1007/978-3-031-20077-9_1.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, pp. 141261–141285, 2026.

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3D object detection. In *European Conference on Computer Vision*, pp. 531–548, 2022. doi: 10.1007/978-3-031-19812-0_31.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49–55, 1936.

Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004. doi: 10.1023/B:VISI.0000027790.02288.f2.

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Juhye Park, Wooju Lee, Dasol Hong, Changki Sung, Youngwoo Seo, Dongwan Kang, and Hyun Myung. VIRD: View-invariant representation through dual-axis transformation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33693–33702, 2026.

Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D. In *European Conference on Computer Vision*, pp. 194–210, 2020. doi: 10.1007/978-3-030-58568-6_12.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniK3D: Universal camera monocular 3d estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1028–1039, 2025.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026. doi: 10.1109/TPAMI.2025.3628473.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Jérôme Revaud, César De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2D2: Reliable and repeatable detector and descriptor. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4938–4947, 2020.

Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. Back to the feature: Learning robust camera localization from pixels to pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3247–3257, 2021.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17010–17020, 2022.

Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-DoF ground-to-satellite camera localization accuracy via geometry-guided cross-view transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21516–21526, 2023.

Zhenbo Song, Xianghui Ze, Jianfeng Lu, and Yujiao Shi. Learning dense flow field for highly-accurate cross-view camera localization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70612–70625, 2023. doi: 10.52202/075280-3094.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8922–8931, 2021.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Xiaolong Wang, Runsen Xu, Zhuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5301–5319, 2023. doi: 10.52202/075280-0234.

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.

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

Zimin Xia, Chenghao Xu, and Alexandre Alahi. Loc^2: Interpretable cross-view localization via depth-lifted local feature matching. In *International Conference on Learning Representations*, pp. 105738–105764, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b. doi: 10.52202/079017-0688.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9043–9053, 2023.

Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5845–5854, 2019.

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, 2021a.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021b.

A Additional Ablation Studies

Unless stated otherwise, all analyses in this section use the VIGOR unknown-orientation, same-area test split, weighted scale-Procrustes, and no RANSAC. GT-derived quantities are computed only after inference for diagnostic purposes and are never exposed to the model.

A.1 Candidate Retrieval and Feature Fusion Diagnostics

The main ablation changes both the candidate source and the learned Stage-II refinement. To isolate their respective roles, Table 4 reports two controlled comparisons. The first two rows use the same checkpoint trained without geometry feature fusion and a fixed budget of 32 candidates; replacing Visual@32 with Visual-8 + Geometry-24 therefore changes only the candidate source. Center recall

increases from 64.45% to 66.19%, GT-region mass from 58.37% to 63.90%, and the final inlier ratio from 54.41% to 63.47%. This shows that geometry-based candidate retrieval complements visual retrieval by recovering geometrically plausible candidates that may be missed by visual similarity alone.

The second and third rows use the same Visual-8 + Geometry-24 candidate construction. Enabling geometry feature fusion in GALCR increases the Top-256 inlier ratio from 49.59% to 61.55% and the final inlier ratio from 63.47% to 87.12%. The two comparisons separate candidate coverage from correspondence discrimination: geometry-based retrieval expands the set of plausible hypotheses, while feature fusion is important for selecting reliable correspondences for final pose recovery.

Table 4: Controlled diagnostics for geometry-based candidate retrieval and geometry feature fusion on VIGOR unknown-orientation, same-area. The first two rows use the same checkpoint trained without geometry feature fusion and differ only in the test-time candidate set. The last row uses the same Visual-8 + Geometry-24 candidate construction with geometry feature fusion enabled in GALCR. GT-derived quantities are post-hoc diagnostics and are not used for inference.

Setting	Center recall (%)	GT-region mass (%)	Top-256 inlier (%)	Final inlier (%)	Residual (m)	t_{mean} (m)	θ_{mean} (°)
Visual@32, without feature fusion	64.45	58.37	45.52	54.41	4.319	3.691	10.666
Visual-8 + Geometry-24, without feature fusion	66.19	63.90	49.59	63.47	3.453	3.416	9.215
Visual-8 + Geometry-24, with feature fusion (GACLOC)	66.19	63.90	**61.55**	**87.12**	**1.423**	**3.080**	**7.206**

A.2 Dependence on Stage-I Coarse-Pose Quality

Table 5 stratifies samples by Stage-I translation error. Candidate recall decreases from 94.95% for Stage-I errors below 3 m to 35.20% for errors of at least 6 m. Stage II still reduces the mean translation error in all three groups, but the most difficult group retains a final mean error of 10.464 m. These results quantify the operating boundary of the coarse-to-fine design: geometry-based candidate retrieval is most reliable when the Stage-I pose places the projected anchor region near the correct satellite neighborhood, whereas severe coarse-pose errors can shift the region too far for local refinement to fully recover.

Table 5: Performance stratified by Stage-I translation error on VIGOR unknown-orientation, same-area. Candidate recall measures the fraction of anchors whose retained candidate set contains at least one GT-compatible candidate. It is a GT-based post-hoc diagnostic and is not available to the model at inference.

Stage-I error bin	Samples	Candidate recall (%)	Stage-I t (m)	Final t (m)	Gain (m)	Improved (%)
< 3 m	34,241	94.95	1.353	1.188	+0.165	64.61
3–6 m	9,819	70.55	4.209	3.250	+0.959	79.95
≥ 6 m	8,545	35.20	11.726	10.464	+1.262	68.80

To further isolate the effect of coarse translation from other sample difficulty factors, we perform a controlled test-time intervention on all 52,605 VIGOR unknown-orientation, same-area test samples. We keep the model, input images, visual descriptors, selected ground anchors, and Stage-I visual candidates unchanged, while replacing only the pose passed to Stage II with the ground-truth rotation, unit scale, and a translation displaced from the ground truth by $r \in \{0, 2, 6, 10\}$ m. For each sample, the nonzero offsets use the same one of eight deterministic directions, with only the displacement magnitude changed.

Table 6 shows a clear dependence on the controlled coarse translation. As the displacement increases from 0 to 10 m, the final mean translation error increases from 0.996 to 3.948 m and the final mean orientation error from $1.355°$ to $4.393°$. Meanwhile, Geometry Top-24 recall decreases from 98.91% to 3.44%, the Top-256 inlier ratio decreases from 87.92% to 28.84%, and the final solver residual

Table 6: Sensitivity to controlled coarse-translation error on VIGOR under the unknown-orientation, same-area setting. Only the pose passed to Stage II is intervened: rotation is fixed to the ground truth, scale to one, and translation is displaced from the ground truth by the specified radius. Visual Top-8 candidates remain unchanged, while Geometry Top-24 candidates are reconstructed from the intervened pose.

Controlled coarse translation (m)	Final translation (m)		Final orientation (°)		Geometry recall (%)	Top-256 inlier (%)	Final inlier (%)	Solver residual (m)
	Mean	Median	Mean	Median				
0	0.996	0.584	1.355	0.755	98.91	87.92	93.04	1.098
2	1.519	1.265	1.563	0.926	86.98	80.70	90.53	1.281
6	2.358	1.884	2.469	1.583	10.41	43.50	68.72	2.348
10	3.948	3.343	4.393	2.651	3.44	28.84	45.19	3.821

increases from 1.098 to 3.821 m. Since the Visual Top-8 candidates remain unchanged across all four conditions, these aligned changes indicate that inaccurate coarse translation limits Stage II by shifting the projected anchor regions and reducing the recoverable geometry-based candidates.

Importantly, Stage II still corrects a substantial fraction of the imposed nonzero translation error: the 2, 6, and 10 m interventions are reduced to final mean translation errors of 1.519, 2.358, and 3.948 m, respectively. Thus, Stage II has a clear local correction capability, while its attainable final accuracy remains conditioned on the quality of the coarse initialization.

A.3 Behavior across Anchor-Region Areas

We further examine whether the refinement behavior depends on the physical area represented by a ground anchor. Table 7 shows that geometry-based candidates contribute additional GT-region probability mass in all three area groups. Refinement is most reliable for small and medium anchor regions. Very large anchor regions ($A_i > 10\,\mathrm{m}^2$) instead exhibit a negative center gain, a lower final inlier ratio, and a larger residual. This indicates that broad anchor regions remain a difficult operating regime even when their geometry is modeled explicitly.

Table 7: Diagnostics across anchor-region areas for the full model on VIGOR unknown-orientation, same-area. Area intervals are measured in m^2. Positive center gain means that GALCR improves over the visual soft match.

Group	Anchor-region area	Added GT mass (%)	Center gain (m)	Final inlier (%)	Residual (m)
Small	$A_i \leq 1$	32.51	+0.788	92.39	1.081
Medium	$1 < A_i \leq 10$	37.64	+1.763	83.56	1.623
Large	$A_i > 10$	32.92	-3.036	48.15	4.129

A.4 Sensitivity to Anchor-Region Covariance for Geometry-Based Candidate Retrieval

Table 8 changes only the anchor-region covariance used for geometry-based candidate retrieval while retaining the trained geometry encoding in GALCR and the confidence head. Removing geometry-based candidates degrades candidate coverage, solver-input quality, and both pose metrics. Isotropic covariance recovers most of this loss, while the full anchor-region covariance gives the highest GT-region mass and the best final pose by a small margin. The identical outcomes of the unit and equal-area isotropic variants follow from the scale invariance of isotropic Mahalanobis Top-K: multiplying an isotropic covariance by a scalar does not change the candidate ordering for a given anchor. These results indicate that geometry-centered candidate coverage is the dominant effect, while the detailed covariance anisotropy has limited sensitivity at the final-pose level in this setting.

Table 8: Sensitivity to the anchor-region covariance used for geometry-based candidate retrieval on VIGOR unknown-orientation, same-area. All rows use the same trained checkpoint, weighted scale-Procrustes, and no RANSAC. The intervention changes only the covariance used for geometry-based candidate retrieval; GALCR and the confidence head retain the original geometry encoding.

Candidate construction	Center recall (%)	GT-region mass (%)	Top-256 inlier (%)	Final inlier (%)	Residual (m)	t_{mean} (m)	θ_{mean} (°)
Visual candidates only	64.453	58.370	60.163	85.791	1.971	3.367	9.114
Isotropic covariance (unit)	**70.275**	62.869	**61.679**	**87.882**	**1.379**	3.091	7.261
Isotropic covariance (equal-area)	**70.275**	62.869	**61.679**	**87.882**	**1.379**	3.091	7.261
Full anchor-region covariance	66.190	**63.903**	61.553	87.120	1.423	**3.080**	**7.206**

A.5 Coarse-to-Final Error Distribution

Table 9 reports mean, median, and 90th-percentile errors before and after Stage II. All three statistics improve for both translation and orientation. Stage II improves translation for 68.16% and orientation for 62.67% of paired samples. The improvement therefore extends beyond a small set of outliers, while the remaining regressions confirm that local refinement is not uniformly successful for every sample.

Table 9: Coarse-to-final error distribution on VIGOR unknown-orientation, same-area. P90 exposes tail behavior that is not visible from mean and median alone.

Stage	Translation (m)			Orientation (°)		
	Mean	Median	P90	Mean	Median	P90
Stage I coarse	3.571	1.975	8.491	8.310	2.819	16.580
Stage II final	**3.080**	**1.529**	**7.226**	**7.206**	**2.278**	**12.174**

A.6 Local Aggregation and Solver-Budget Analysis

Table 10 varies one budget at a time. With $N_f = 256$, varying $L \in \{4, 8, 16\}$ changes mean localization error by at most 0.023 m, indicating that the method is stable over this range rather than being sharply tuned to $L = 8$. Reducing the final solver budget from $N_f = 256$ to $N_f = 128$ increases the final inlier rate but reduces spatial coverage and slightly degrades the pose estimates. This result shows that a smaller and cleaner correspondence set does not necessarily provide sufficient spatial distribution for stable 3-DoF pose recovery. We use $L = 8$ and $N_f = 256$ by default.

Table 10: Top-K sensitivity on VIGOR unknown-orientation, same-area. L is the number of candidates aggregated within each anchor and N is the number of anchor correspondences passed to the solver. All runs use seed 2026, weighted scale-Procrustes, and no RANSAC.

L	N	t_{mean} (m)	t_{median} (m)	θ_{mean} (°)	θ_{median} (°)	Final inlier (%)	Coverage (%)	Residual (m)
4	256	**3.067**	**1.519**	**7.196**	2.282	86.77	**18.87**	1.468
8	256	3.080	1.529	7.206	**2.278**	87.12	18.42	1.423
16	256	3.090	1.546	7.239	2.329	88.42	18.42	1.338
8	128	3.103	1.564	7.377	2.421	**91.67**	11.93	**1.053**

B Panorama Processing and Projector Architecture

B.1 VIGOR Panorama Processing

To avoid applying a planar backbone directly to strongly distorted equirectangular imagery, we convert each panorama to locally perspective views (Eder et al., 2020). The cubemap conversion uses four horizontal faces in the cyclic order front, right, back, and left. Each face is 518×518

pixels with a 90° horizontal and vertical field of view; the upper and lower faces are discarded. The virtual pinhole intrinsics are $f_x = f_y = 259$ and $c_x = c_y = 259$. Consequently, a DINOv2 ViT-L/14 backbone (Oquab et al., 2024; Dosovitskiy et al., 2021) produces a 37×37 token grid per face. RGB, depth, pixel centers, and camera rays share this projection.

The camera ego frame uses $+X$ forward, $+Y$ right, and $+Z$ upward. The four virtual cameras are co-located and differ only by their 90° yaw rotations. For unknown orientation, a deterministic yaw in $[-180^\circ, 180^\circ)$ is assigned to each panorama before cubemap conversion and is held fixed across epochs. The known-orientation conversion uses the unshifted panorama. KITTI uses one forward-facing left color camera and therefore disables inter-view exchange.

B.2 Ground and Satellite Projectors

The projectors follow PETR-style position-conditioned attention (Liu et al., 2022; Vaswani et al., 2017). Related image-to-BEV representations use explicit lifting, spatial aggregation, or attention-based sampling to inject camera geometry into learned features (Philion & Fidler, 2020; Li et al., 2022; Zhu et al., 2021b). For each ground token, the position input contains normalized image coordinates (u, v), the three-dimensional unit ray in the ego frame, and normalized log depth. A four-band Fourier map (Tancik et al., 2020) and a two-layer MLP encode these six scalars to 256 dimensions. A 1×1 convolution maps the frozen 1,024-D DINOv2 feature to the same width. Query and key use content plus position, whereas value uses content only.

The VIGOR ground projector alternates two intra-view self-attention blocks with two shared adjacent-edge blocks. Each edge block exchanges the outer three token columns across the cyclic front–right–back–left seams. All blocks use eight heads, a 256-D hidden width, an FFN expansion ratio of four, and dropout 0.1. The satellite projector encodes normalized 2D grid coordinates and applies four blocks with the same width and head count. Both branches end with LayerNorm, a linear projection to 128 dimensions, and ℓ_2 normalization (Ba et al., 2016).

C Anchor Geometry Representation Details

C.1 Anchor-Region Geometry

For each ground anchor i, we sample the 7×7 pixels associated with its ground-image token and back-project the valid pixels using metric depth predicted by a frozen monocular estimator (Lin et al., 2026). The use of monocular depth as a transferable geometric prior follows recent progress in robust relative and metric depth estimation (Yang et al., 2024a;b; Piccinelli et al., 2024; 2026; 2025; Yin et al., 2023; Ranftl et al., 2022). The back-projection follows the standard pinhole-camera geometry (Hartley & Zisserman, 2004). Let $m_{ik} \in \{0, 1\}$ indicate whether sample k has valid depth and a valid back-projected coordinate, and let

$$K_i = \sum_{k=1}^{49} m_{ik}. \tag{16}$$

The resulting valid BEV points are denoted by $\mathbf{p}_{ik} \in \mathbb{R}^2$. For a valid anchor, its metric center and covariance are

$$\mathbf{x}_i = \frac{1}{K_i} \sum_{k=1}^{49} m_{ik} \mathbf{p}_{ik}, \qquad \mathbf{C}_i = \frac{1}{K_i} \sum_{k=1}^{49} m_{ik} (\mathbf{p}_{ik} - \mathbf{x}_i)(\mathbf{p}_{ik} - \mathbf{x}_i)^\top. \tag{17}$$

Second-order region statistics are widely used to describe the scale and affine shape of local image regions (Mikolajczyk & Schmid, 2004). We characterize the anchor-region area using the 2σ covariance ellipse:

$$A_i = 4\pi \sqrt{\max\left(\det \mathbf{C}_i, 10^{-6}\right)}. \tag{18}$$

Let d_{ik} denote the valid metric depth samples. Their mean and variance are

$$\bar{d}_i = \frac{1}{K_i} \sum_{k=1}^{49} m_{ik} d_{ik}, \qquad \sigma_{d,i}^2 = \frac{1}{K_i} \sum_{k=1}^{49} m_{ik} (d_{ik} - \bar{d}_i)^2. \tag{19}$$

The depth reliability is

$$r_i = \mathrm{clip}_{[0,1]}\left[\frac{K_i}{49}\exp\left(-\frac{\sigma_{d,i}^2}{\max(\bar{d}_i^2, 10^{-6})}\right)\right], \tag{20}$$

and the anchor-geometry uncertainty is

$$u_i = \log(1 + A_i)(1 - r_i). \tag{21}$$

Invalid anchors are masked in subsequent geometric computation.

Let c denote the metric width of one satellite-token cell, with

$$A_{\mathrm{cell}} = c^2, \qquad v_{\mathrm{cell}} = \frac{c^2}{12}. \tag{22}$$

For numerical stability, we use a covariance stabilization operator $\mathcal{S}(\cdot)$ that symmetrizes its input, adds $10^{-4}\mathbf{I}$, and clamps the resulting eigenvalues to $[10^{-4}, 144]$.

Given the Stage-I coarse pose $\hat{\mathbf{T}}_0 = (\hat{s}_0, \hat{\mathbf{R}}_0, \hat{\mathbf{t}}_0)$, the anchor-region geometry is projected to the satellite frame as

$$\hat{\boldsymbol{\mu}}_i = \hat{s}_0\hat{\mathbf{R}}_0\mathbf{x}_i + \hat{\mathbf{t}}_0, \tag{23}$$

$$\hat{\boldsymbol{\Sigma}}_i = \mathcal{S}\left(\hat{s}_0^2\hat{\mathbf{R}}_0\mathcal{S}(\mathbf{C}_i)\hat{\mathbf{R}}_0^\top + v_{\mathrm{cell}}\mathbf{I}\right). \tag{24}$$

For satellite candidate j, let

$$\boldsymbol{\delta}_{ij} = \mathbf{y}_j - \hat{\boldsymbol{\mu}}_i. \tag{25}$$

Using the Cholesky factorization $\hat{\boldsymbol{\Sigma}}_i = \mathbf{L}_i\mathbf{L}_i^\top$, we define

$$\tilde{\boldsymbol{\delta}}_{ij} = \mathbf{L}_i^{-1}\boldsymbol{\delta}_{ij}, \tag{26}$$

and the corresponding Mahalanobis distance (Mahalanobis, 1936):

$$d_{ij}^2 = \boldsymbol{\delta}_{ij}^\top\hat{\boldsymbol{\Sigma}}_i^{-1}\boldsymbol{\delta}_{ij} = \left\|\tilde{\boldsymbol{\delta}}_{ij}\right\|_2^2. \tag{27}$$

C.2 Geometric Feature Construction

AGR constructs an anchor-level geometric feature and an anchor–candidate geometric relation feature. The anchor-level feature is

$$\mathbf{f}_i^{\mathrm{geo}} = \left[\log\frac{A_i}{A_{\mathrm{cell}}}, \frac{C_{i,xx}}{v_{\mathrm{cell}}}, \frac{C_{i,yy}}{v_{\mathrm{cell}}}, \frac{C_{i,xy}}{v_{\mathrm{cell}}}, u_i, r_i\right]. \tag{28}$$

For numerical stability, the area term is clipped to $[-8, 8]$, the normalized covariance terms to $[-64, 64]$, u_i to $[0, 16]$, and r_i to $[0, 1]$. This feature summarizes the metric scale, shape, orientation, and reliability of the region associated with each ground anchor. Invalid anchors are zero-masked.

The anchor–candidate geometric relation feature is

$$\mathbf{f}_{ij}^{\mathrm{geo}} = \left[\frac{\delta_{ij,x}}{c}, \frac{\delta_{ij,y}}{c}, \tilde{\delta}_{ij,x}, \tilde{\delta}_{ij,y}, d_{ij}^2, \log\det\hat{\boldsymbol{\Sigma}}_i, S_{ij}^{\mathrm{vis}}\right]. \tag{29}$$

The two cell-normalized offsets and the covariance-normalized offsets are clipped to $[-64, 64]$, d_{ij}^2 to $[0, 10^4]$, $\log\det\hat{\boldsymbol{\Sigma}}_i$ to $[-32, 32]$, and S_{ij}^{vis} to $[-100, 100]$. Here

$$S_{ij}^{\mathrm{vis}} = \frac{\langle\bar{\mathbf{g}}_i, \bar{\mathbf{s}}_j\rangle}{\tau_{\mathrm{app}}}, \qquad \tau_{\mathrm{app}} = 0.1, \tag{30}$$

is the temperature-scaled visual matching logit before Dual-Softmax (Sun et al., 2021).

The anchor and pair features are encoded as

$$\mathrm{MLP}_{\mathrm{anchor}}(\mathbf{f}_i^{\mathrm{geo}}) = \mathrm{Linear}_{64\to64}\left(\mathrm{GELU}\left(\mathrm{Linear}_{6\to64}\left(\mathrm{LayerNorm}\left(\mathbf{f}_i^{\mathrm{geo}}\right)\right)\right)\right), \tag{31}$$

and

$$\mathrm{MLP}_{\mathrm{pair}}(\mathbf{f}_{ij}^{\mathrm{geo}}) = \mathrm{Linear}_{128\to128}\left(\mathrm{GELU}\left(\mathrm{Linear}_{7\to128}\left(\mathrm{LayerNorm}\left(\mathbf{f}_{ij}^{\mathrm{geo}}\right)\right)\right)\right). \tag{32}$$

respectively. Both encoders use LayerNorm and GELU nonlinearities (Ba et al., 2016; Hendrycks & Gimpel, 2016).

D Soft Correspondence and Correspondence Weighting

D.1 Anchor-Wise Soft Correspondence

Learned local-feature pipelines have progressively coupled repeatable feature extraction, match reliability, and geometric localization (DeTone et al., 2018; Dusmanu et al., 2019; Revaud et al., 2019; Sarlin et al., 2019; 2021; Leroy et al., 2024; Wang et al., 2024). Our setting differs in that the retained candidates correspond to the same ground anchor across a severe ground–satellite viewpoint change. GALCR refines at most 32 satellite candidates for each ground anchor. A lightweight dustbin head predicts a rejection logit a_i^{dust} to model candidate sets that do not contain a reliable match, following the general practice of explicitly representing unmatched assignments in learned feature matching (Sarlin et al., 2020).

After GALCR refinement, the top $L = 8$ candidates are retained:

$$\mathcal{J}_i = \mathrm{TopK}_L\left(\{S_{ij}^{\mathrm{ref}}\}_{j\in\mathcal{K}_i}\right). \tag{33}$$

The refined candidate logits and dustbin logit are jointly normalized:

$$\tilde{\mathbf{p}}_i = \mathrm{softmax}\left(\left[\left\{S_{ij}^{\mathrm{ref}}/\tau_{\mathrm{ref}}\right\}_{j\in\mathcal{J}_i}, a_i^{\mathrm{dust}}\right]\right), \qquad \tau_{\mathrm{ref}} = 0.1. \tag{34}$$

After removing the dustbin mass, the real candidate probabilities are renormalized as

$$p_{ij} = \frac{\tilde{p}_{ij}}{\sum_{k\in\mathcal{J}_i}\tilde{p}_{ik}}, \qquad j \in \mathcal{J}_i. \tag{35}$$

We convert the discrete candidate distribution into a continuous coordinate by expectation, analogous to differentiable coordinate regression (Nibali et al., 2018):

$$\hat{\mathbf{y}}_i = \sum_{j\in\mathcal{J}_i} p_{ij}\mathbf{y}_j. \tag{36}$$

D.2 Feature Aggregation and Confidence Prediction

Learning correspondence reliability from contextualized match representations has proved effective for geometric matching and registration (Zhang et al., 2019; Bai et al., 2021). Each retained candidate $j \in \mathcal{J}_i$ is represented by a candidate token $\mathbf{z}_{ij}$ constructed from its visual matching information, anchor–candidate geometry, anchor geometry, and local matching statistics. The encoded candidate tokens are aggregated using the same local matching distribution used for soft correspondence estimation:

$$\mathbf{h}_i = \sum_{j\in\mathcal{J}_i} p_{ij}\mathbf{z}_{ij}. \tag{37}$$

A learnable query $\mathbf{q}_0$ attends to the correspondence representations of all valid anchors to obtain sample-level context, following the general set-attention formulation (Lee et al., 2019):

$$\mathbf{q}^{\mathrm{ctx}} = \mathrm{CrossAttn}\left(\mathbf{q}_0, \{\mathbf{h}_i\}_{i=1}^{N}\right). \tag{38}$$

This context is used only for correspondence weighting and does not modify the GALCR matching scores or the soft correspondence coordinates.

For anchor i, the confidence logit is predicted from its local correspondence representation, the sample-level context, their element-wise interaction, and a local quality summary $\mathbf{r}_i^{\mathrm{qual}}$:

$$a_i^{\mathrm{conf}} = \mathrm{MLP}_{\mathrm{conf}}\left(\left[\mathbf{h}_i,\, \mathbf{q}^{\mathrm{ctx}},\, \mathbf{h}_i \odot \mathbf{q}^{\mathrm{ctx}},\, \mathbf{r}_i^{\mathrm{qual}}\right]\right). \tag{39}$$

The confidence score and final correspondence weight are

$$c_i = \sigma(a_i^{\mathrm{conf}}), \qquad w_i = c_i. \tag{40}$$

The confidence MLP is implemented as

$$\mathrm{LayerNorm}(776) \to \mathrm{Linear}(776, 256) \to \mathrm{GELU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(256, 1). \tag{41}$$

The confidence logits rank anchors for the final Top-N_f selection, with $N_f = 256$. This branch predicts correspondence reliability and does not regress camera pose.

E Training Objectives and Target Construction

E.1 Stage-I Region Supervision

For each ground anchor, the GT pose projects its center and covariance into the satellite frame:

$$\boldsymbol{\mu}_i^{\mathrm{gt}} = \mathbf{R}_{\mathrm{gt}}\mathbf{x}_i + \mathbf{t}_{\mathrm{gt}}, \tag{42}$$

$$\boldsymbol{\Sigma}_i^{\mathrm{gt}} = \mathcal{S}\left(\mathbf{R}_{\mathrm{gt}}\mathbf{C}_i\mathbf{R}_{\mathrm{gt}}^{\top} + v_{\mathrm{cell}}\mathbf{I}\right). \tag{43}$$

A discrete Gaussian target is constructed over all satellite tokens:

$$d_{ij,\mathrm{gt}}^2 = (\mathbf{y}_j - \boldsymbol{\mu}_i^{\mathrm{gt}})^{\top}(\boldsymbol{\Sigma}_i^{\mathrm{gt}})^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_i^{\mathrm{gt}}), \tag{44}$$

$$q_{ij} = \frac{\exp(-d_{ij,\mathrm{gt}}^2/2)}{\sum_k \exp(-d_{ik,\mathrm{gt}}^2/2)}. \tag{45}$$

The predicted Stage-I matching distribution is

$$p_{ij} = \mathrm{softmax}_j\left(S_{ij}^{\mathrm{vis}}\right). \tag{46}$$

We estimate the fraction of the Gaussian region observable within the satellite crop as

$$\rho_i = \mathrm{clip}_{[0,1]}\left[\frac{c^2}{2\pi\sqrt{\det\boldsymbol{\Sigma}_i^{\mathrm{gt}}}}\sum_j \exp(-d_{ij,\mathrm{gt}}^2/2)\right], \tag{47}$$

and retain anchors with

$$o_i = \mathbb{I}[\rho_i \geq 0.2]. \tag{48}$$

The reliability weight used for region supervision is

$$\omega_i = \mathrm{stopgrad}\left[\mathbb{I}[\mathrm{valid}_i]\, r_i\, \sigma\left(\frac{4 - \log\max(A_i, 10^{-6})}{2}\right)\right]. \tag{49}$$

The region loss is the observable weighted KL divergence

$$\mathcal{L}_{\mathrm{region}} = \frac{\sum_i o_i\omega_i D_{\mathrm{KL}}\left(q_i \| p_i\right)}{\sum_i o_i\omega_i}. \tag{50}$$

Together with the bidirectional NCE objective (van den Oord et al., 2018) and the coarse-pose VCE used in fine-grained cross-view localization (Xia & Alahi, 2025), the Stage-I objective is

$$\mathcal{L}_{\mathrm{I}} = \mathcal{L}_{\mathrm{NCE}} + \mathcal{L}_{\mathrm{region}} + \mathcal{L}_{\mathrm{coarse}}. \tag{51}$$

E.2 Stage-II Candidate Supervision

For candidate j, we evaluate its GT geometric compatibility using

$$d_{ij,\mathrm{gt}}^2 = (\mathbf{y}_j - \boldsymbol{\mu}_i^{\mathrm{gt}})^{\top}(\boldsymbol{\Sigma}_i^{\mathrm{gt}})^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_i^{\mathrm{gt}}). \tag{52}$$

The positive candidate set is

$$\mathcal{P}_i = \left\{j \in \mathcal{K}_i \mid d_{ij,\mathrm{gt}}^2 \leq 9\right\}. \tag{53}$$

If $\mathcal{P}_i \neq \varnothing$, the target distribution is

$$q_{ij} = \frac{\mathbb{I}[j \in \mathcal{P}_i]\exp(-d_{ij,\mathrm{gt}}^2/2)}{\sum_{k\in\mathcal{P}_i}\exp(-d_{ik,\mathrm{gt}}^2/2)}, \qquad q_{i,\varnothing} = 0. \tag{54}$$

If no retained candidate is positive, or if the GT-projected anchor center lies outside the satellite crop, all target mass is assigned to the dustbin:

$$q_{i,\varnothing} = 1, \qquad q_{ij} = 0, \quad \forall j \in \mathcal{K}_i. \tag{55}$$

Let π_{ij} and $\pi_{i,\varnothing}$ denote the predicted candidate and dustbin probabilities. The candidate loss is

$$\mathcal{L}_{\mathrm{cand}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{j\in\mathcal{K}_i} q_{ij}\log\pi_{ij} + q_{i,\varnothing}\log\pi_{i,\varnothing}\right]. \tag{56}$$

E.3 Correspondence Confidence Supervision

The soft-correspondence error is measured relative to the GT-projected anchor center:

$$e_i = \left\| \hat{\mathbf{y}}_i - \boldsymbol{\mu}_i^{\mathrm{gt}} \right\|_2 . \tag{57}$$

The soft confidence target is

$$z_i = \mathbb{I}[\boldsymbol{\mu}_i^{\mathrm{gt}} \text{ is visible}]\, \sigma\left(\frac{2.5 - e_i}{0.5} \right). \tag{58}$$

The confidence loss is

$$\mathcal{L}_{\mathrm{conf}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{BCEWithLogits}\left(a_i^{\mathrm{conf}}, z_i \right). \tag{59}$$

E.4 Pose Supervision

The N_f highest-confidence correspondences are passed to weighted scale-Procrustes. This analytic recovery belongs to the classical family of orthogonal Procrustes and absolute-orientation solutions (Schönemann, 1966; Kabsch, 1976; Horn, 1987; Umeyama, 1991), while our cross-view solver follows the scale-aware analytic pose-recovery setting of Loc^2 (Xia et al., 2026):

$$\hat{\mathbf{T}} = \mathrm{WeightedScaleProcrustes}\left(\{(\mathbf{x}_i, \hat{\mathbf{y}}_i, w_i)\}_{i \in \mathcal{I}_f} \right), \qquad |\mathcal{I}_f| = N_f. \tag{60}$$

The recovered pose is supervised using Virtual Correspondence Error:

$$\mathcal{L}_{\mathrm{pose}} = \mathrm{VCE}\left(\hat{\mathbf{T}}, \mathbf{T}_{\mathrm{gt}} \right). \tag{61}$$

We evaluate VCE on a 10×10 virtual grid covering $[-5, 5] \times [-5, 5]$ m in the camera frame and average the Euclidean displacement after applying the predicted and GT poses.

The complete Stage-II objective is

$$\boxed{\mathcal{L}_{\mathrm{II}} = \mathcal{L}_{\mathrm{cand}} + \mathcal{L}_{\mathrm{conf}} + \mathcal{L}_{\mathrm{pose}}} \tag{62}$$

with equal weights for all three terms.

GT pose is used only to construct training targets and evaluation metrics; it is never used for candidate generation, correspondence weighting, Top-N_f selection, or test-time pose recovery.

F Additional Qualitative Results

F.1 KITTI Visualization

Figure 4 presents the KITTI qualitative results. Each sample occupies one row, with the Stage I coarse-pose visualization on the left and the corresponding Stage II final-pose visualization on the right, allowing a direct comparison before and after geometry-aware refinement. Within each visualization, the KITTI perspective image is shown on the left and the satellite image on the right. The connecting lines indicate correspondences between perspective-image tokens and satellite candidate points: green denotes correspondences with small errors and good agreement with the ground truth, yellow denotes moderate deviations, and red denotes large errors or clear mismatches. The ground-truth (GT) pose and predicted pose are marked separately in the satellite image, making it possible to directly observe whether Stage II corrects the Stage I coarse prediction toward the GT pose.

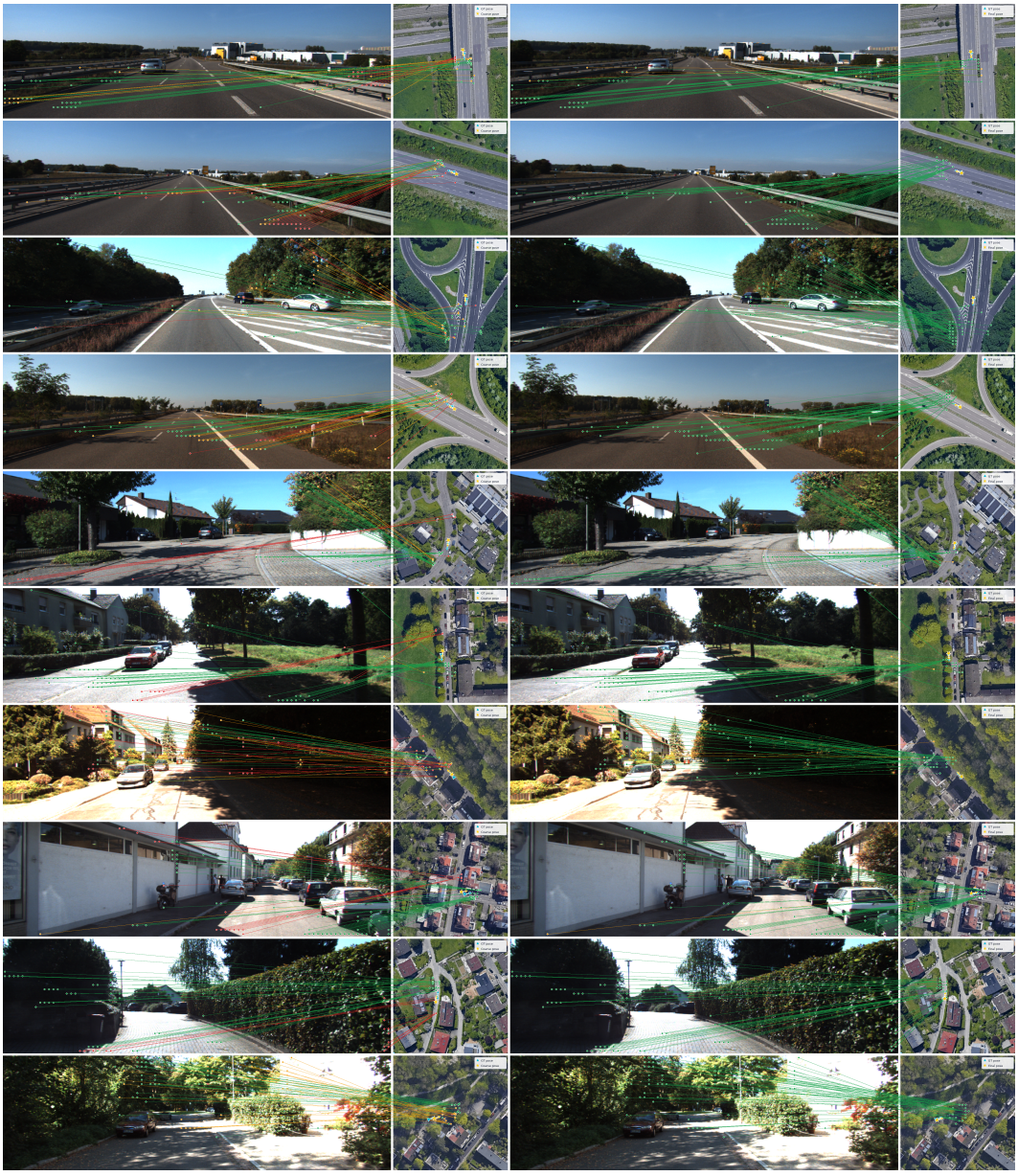

Figure 4: Additional qualitative results on KITTI. Each row compares the Stage I coarse result (left) with the corresponding Stage II refined result (right).