STRUCTURAL RADIAL AGGREGATION FOR CROSS-VIEW YAW UNDER POSITION UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Ground-to-satellite localization estimates a street-level camera's position and yaw (heading) from an aerial image. Yaw is the harder part, because a small angle error moves distant scene content far. Position uncertainty turns yaw estimation into aggregating complementary evidence over all candidate poses. Prior work matches each ground-view column to aerial pixels. All positions on the ray through a matched pixel imply the same yaw, so each match votes for a yaw without knowing its distance. Each vote is precise but checks one viewing direction, which leaves many yaws nearly as plausible as the true one at each position. The scene layout is coarser but covers the whole view. Structural Radial Aggregation for Yaw (S-RAY) combines both. A structural score, learned from pose labels only, rates each yaw at each position by how well the directions where the ground view shows structure line up with the aerial structure axes, leaving a few promising yaws per position. A Fourier score of continuous match votes decides among them. Hard-bin voting, which bins every position–match pair separately, is costly and makes the score piecewise constant. Instead, each match adds a learned bump at its implied yaw. Since the implied yaw is a difference of angles, a bump of a few cosines turns scoring all positions and yaws into a few FFT correlations that form no position–match pairs; this scorer is $23\times$ faster than prior hard-bin voting. S-RAY raises yaw recall within 1° over the prior method from 51.07% to 87.65% on KITTI Same Area, from 48.44% to 72.91% on KITTI Cross Area, and from 34.81% to 55.22% on MGL, and has the lowest mean yaw error on both VIGOR splits. As a yaw initialization, it also improves Loc^2 localization on KITTI and MGL.

1 INTRODUCTION

Ground-to-satellite localization finds where a street-level camera stands and which way it faces, its planar position (x, y) and yaw θ (compass heading), by matching its photo to an aerial image of the surrounding area. It supports navigation and robotics where satellite positioning is inaccurate (Brejcha & Čadík, 2017; Shi et al., 2020a; 2023). Without a reliable yaw, the search spans every position and all 360° of yaw, while an ordinary photo shows only a narrow slice of the scene. We target yaw under this position uncertainty, scoring every candidate position and returning the yaw of the best-scoring pose. Once yaw is precise, only a 2-DoF position search remains.

Prior work, LAYS (Kang et al., 2026), makes this search tractable. Each vertical column of the ground photo looks in one direction relative to the camera's optical axis and sees everything along it, near or far, so matching a column to an aerial pixel gives a direction but no distance. For a candidate position, the match fixes the yaw that turns the column toward the pixel, and every position on the ray through the pixel gets the same yaw (*radial invariance*, Fig. 1a), so the vote needs no distance. Off the ray the yaw changes, so matches in different columns agree only near the true position.

Position uncertainty turns yaw estimation into aggregating two complementary kinds of evidence over all candidate poses. Column matches give precise yaw cues, but each checks only one viewing direction. With a narrow view and repetitive streets, many wrong yaws at each position gather nearly as much support as the true one. The scene layout is coarser but covers the whole view, with a few dominant directions, such as the street and building fronts, visible both in the photo and from above.

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.

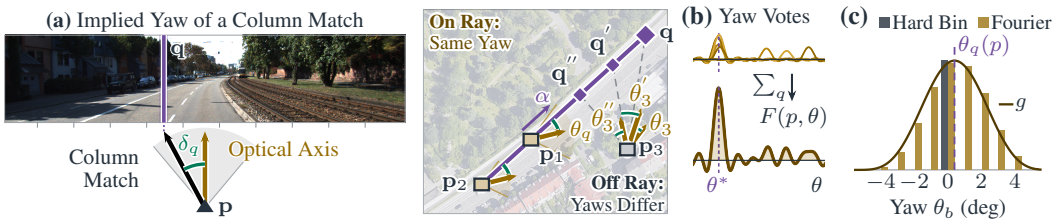

Figure 1: **Radial yaw votes. (a)** A ground-column match fixes its angle δ_q from the camera's optical axis. Positions on the ray through q (p_1, p_2) see q in the same direction $\alpha(q - p)$ and receive the same yaw $\theta_q(p) = \alpha(q - p) - \delta_q$, and so do the other pixels q', q'' on this ray matched to the same column; from a position off the ray (p_3), these pixels imply different yaws $\theta_3, \theta_3', \theta_3''$. **(b)** Fourier votes at one position (top) sum to $F(p, \theta)$. **(c)** For one implied yaw, a hard-bin vote puts all of s_q into the nearest bin; the Fourier vote adds $s_q\, g(\theta_b - \theta_q(p))$ to every bin.

We introduce Structural Radial Aggregation for Yaw (S-RAY), which combines both kinds to score every candidate position and yaw. A *structural score*, trained with pose labels only, compares the viewing directions where the photo shows structure, under each candidate yaw, with the structure axes that a network predicts from the aerial image at the candidate position. It judges the whole view at once, so at each position it keeps only the few yaws aligned with the structure. It cannot pin the yaw down precisely, however, and similar structures such as a parallel road can fit nearly as well.

A *Fourier score* of match votes decides among them. Prior work counts each vote in one discrete yaw bin (*hard bins*). The bin depends on both the position and the match, so every position–match pair is binned separately, at a cost that grows with their number, and a vote's credit in training jumps at bin edges (Fig. 1c). Our module instead adds a learned bump of a few cosines at each match's implied yaw. By angle addition, each cosine is the real part of a product of factors that depend only on the tested yaw, the pixel offset, or the match (Fig. 4). A pair's vote is thus its match factor times its offset factor. The offset factor depends only on where the pixel lies relative to the position, so it is the same ring-shaped kernel around every position. For each cosine, summing these products at every position is a correlation of the match map with this kernel, which an FFT computes as a pointwise product of spectra. No position–match pairs form, and the tested yaw only weights the resulting maps. The full model, structural score included, thus runs faster than the prior method.

We evaluate on KITTI and MGL, with a limited field of view (FoV) and yaw uniform over $\pm 180°$, and on panoramic VIGOR. Yaw recall within $1°$ rises over the prior method from 51.07% to 87.65% on KITTI Same Area, from 48.44% to 72.91% on Cross Area, and from 34.81% to 55.22% on MGL (Tables 1 and 2). S-RAY also has the lowest mean yaw error on both VIGOR splits (Table 3). As a yaw initialization, it lowers the median location error of Loc^2 from 1.31 to 0.68 m on KITTI Same Area, from 9.11 to 6.82 m on Cross Area, and from 8.26 to 6.78 m on MGL (Table 6).

Contributions.

- Under position uncertainty, a structural score learned from pose labels only lines up the photo's viewing directions with aerial structure axes and keeps a few yaws per position.
- Fourier voting replaces discontinuous hard bins, scores all poses with one FFT correlation per cosine without forming position–match pairs, and is $23\times$ faster than LAYS's hard bins.
- S-RAY leads $1°$ yaw recall on KITTI and MGL and mean yaw error on VIGOR, and improves Loc^2 as a yaw initialization.

2 Related Work

From retrieval to fine 3-DoF pose. Cross-view geo-localization began as retrieval, with global descriptors for city- or tile-scale matching (Workman et al., 2015; Hu et al., 2018; Shi et al., 2020b; 2019; Zhu et al., 2022). VIGOR broadened this setting beyond one-to-one image pairs and exact tile centers (Zhu et al., 2021). Recent work estimates the fine pose within an aerial search region: polar

or perspective transforms bring the two views into a common frame, and dense score or uncertainty maps keep several candidate positions (Shi et al., 2020a; Toker et al., 2021; Xia et al., 2024; 2022).

BEV and dense correspondence matching. One route to fine localization projects ground features into a bird's-eye view (BEV) and aligns them with aerial features, learned maps, or a grid of candidate poses (Shi et al., 2023; Wang et al., 2023a; Lentsch et al., 2023; Wang et al., 2024; Song et al., 2023; Wang et al., 2023b). Neural-map systems register image features against a map (Sarlin et al., 2023a;b; Wu et al., 2024); OrienterNet correlates a multi-channel BEV template with the map by FFT once per tested yaw, so its cost and memory grow with the number of yaws. Another route estimates pose from 3D-aware correspondences: FG^2 matches fine-grained features, and Loc^2 lifts local features with depth (Xia & Alahi, 2025; Xia et al., 2026). Recent pose estimators learn view-invariant features (VIRD), estimate an orientation-guided homography (OGH-Net), or refine the pose by second-order optimization (Park et al., 2026; Zhang et al., 2026; Feng et al., 2026).

Yaw under position uncertainty. Early fine-localization benchmarks paired orientation-aligned ground and aerial images, so yaw was given (Zhu et al., 2021; Xia et al., 2022). KITTI localizers then adopted a yaw prior within $\pm 10^\circ$ (Shi & Li, 2022; Shi et al., 2023; Song et al., 2023; Xia & Alahi, 2025), and other protocols bound yaw differently: GeoFlow takes external orientation, GeoDistill a $\pm 45^\circ$ range, and weakly supervised localizers evaluate bounded yaw on limited-FoV images and no yaw on panoramas (Abu Lehyeh et al., 2026; Tong et al., 2025; Shi et al., 2024; Wang et al., 2025). Unknown yaw was long studied mainly on panoramas, where polar transforms, image slices, and rolled descriptors turn it into a circular shift (Shi et al., 2020a; 2019; Lentsch et al., 2023; Xia et al., 2024); SliceMatch and CCVPE report unknown-yaw KITTI only as a secondary setting, where joint regressors can confuse θ and $\theta + 180^\circ$. Only recently has full-range yaw become the main protocol for limited-FoV queries (Park et al., 2026; Xia et al., 2026; Zhang et al., 2026). Orientation-centered methods differ: OGH-Net first estimates a coarse orientation to guide alignment (Zhang et al., 2026), visual compasses read orientation directly from image evidence (Liu et al., 2022; Du et al., 2024), and ARC-Loc intersects bearing rays and defines yaw only modulo 180° (Kim et al., 2026). LAYS matches ground columns to aerial pixels and lets each match vote at every candidate position for its implied yaw in hard bins (Kang et al., 2026). S-RAY keeps this geometry, adds a structural score from all visible directions, and replaces hard bins with continuous votes from FFT correlations. Its ring kernels are the circular harmonics of steerable filters and harmonic networks (Freeman & Adelson, 1991; Worrall et al., 2017), here voting over yaw.

3 METHOD

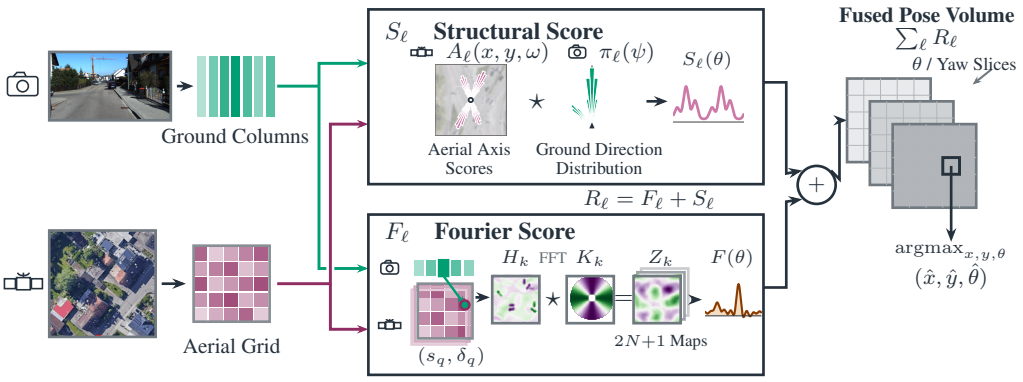

Figure 2: S-RAY overview. Ground-image columns are matched to aerial pixels. The structural score S_ℓ compares the directions visible in the ground image with aerial structure axes. The Fourier score F_ℓ turns match strength and column angle into maps H_k, correlates them by FFT with ring-shaped kernels K_k, and reads a yaw score from the $2N{+}1$ resulting maps. Green is ground and purple aerial evidence. Figs. 1, 3, and 4 give details.

3.1 OVERVIEW

Given an upright ground image and an aerial image of the surrounding area, S-RAY scores every candidate camera position $p = (x, y)$ in the search area and yaw θ on a grid, the *pose grid*, and returns the yaw of the best-scoring cell. Its score $R(p, \theta)$ has one term per kind of evidence (Fig. 2):

$$R(p, \theta) = F(p, \theta) + S(p, \theta). \tag{1}$$

The structural score S checks the layout of the whole view and leaves a few yaws at each position (Sec. 3.3). The Fourier score F decides among them with continuous match votes, which angle addition turns into one FFT correlation per cosine (Sec. 3.4). R peaks where both agree (Fig. 5).

3.2 BACKGROUND: RAY VOTING

Column matching. Following LAYS (Kang et al., 2026), a ground encoder pools each column j of the ground image into one feature, and an aerial encoder gives a feature at every aerial pixel q; both use DINOv2-B (Oquab et al., 2024). Column j looks at a known angle ψ_j from the optical axis, set by its image position and the camera intrinsics (for a panorama, by its index). Every aerial pixel, seen by the camera or not, is matched to the column $j(q)$ whose feature has the largest cosine similarity with it in magnitude; this magnitude is the *match strength* $s_q \geq 0$, and the *column angle* $\delta_q = \psi_{j(q)}$ is the matched column's angle. The magnitude keeps training from collapsing to negative similarities for non-matching pairs, an easier target (Kang et al., 2026). During training, the column is instead sampled in proportion to these magnitudes.

Implied yaw. From candidate position p, the direction to q in the aerial image is $\alpha(q - p)$, where $\alpha(d) = \operatorname{atan2}(d_y, d_x)$ is the direction of an offset d. The yaw at which the matched column points from p toward q, the *implied yaw* $\theta_q(p)$, is this direction minus the column angle:

$$\theta_q(p) = \alpha(q - p) - \delta_q. \tag{2}$$

Moving p along the ray through q leaves $\theta_q(p)$ unchanged, and moving off the ray changes it (*radial invariance*, Fig. 1a). Pixels of a linear structure running away from the camera fall in one column and agree only on their common line; matches in other columns narrow the agreeing positions further.

Hard-bin voting. Each position p takes votes only from the matched pixels in a *ring* around it, $a(d) = \mathbb{I}[r_{\min} \leq \|d\| \leq r_{\max}]$, which is 1 for offsets $d = q - p$ between an inner radius $r_{\min}$ and an outer radius $r_{\max}$ and 0 elsewhere. The outer radius keeps the ring inside the aerial image, and the inner radius leaves out nearby pixels (Appendix C). LAYS counts each vote in a single yaw bin (*hard bins*). With $Q_\Theta(\cdot)$ rounding a yaw to the nearest of the yaw bins Θ, wrapping around the circle, each match adds its full strength to one bin of the *hard-bin score* F_{hard},

$$F_{\text{hard}}(p, \theta) = \sum_q a(q - p)\, s_q\, \mathbb{I}[\theta = Q_\Theta(\theta_q(p))]\,. \tag{3}$$

The bin depends on both the offset $q - p$ and the match, so every position–match pair is binned separately, and the vote's credit jumps at bin edges during training.

3.3 STRUCTURAL SCORE: SCENE-LAYOUT ALIGNMENT

Each match vote judges one viewing direction, which leaves many yaws plausible at each position. The structural score judges the whole view at once. In a photo taken along a street, the road recedes toward the image center and building fronts line its sides. From above, the same street is a band through the camera position with a clear axis. Rotating the photo's viewing directions by the true yaw lines them up with these aerial axes. An alignment term C_ℓ, computed at each level ℓ of a feature pyramid (Sec. 3.5), scores this (Fig. 3) from two inputs learned from pose labels alone.

Aerial axis scores. At each aerial location, a small CNN on the aerial features scores every axis angle bin ω, giving $A_\ell(x, y, \omega)$, which is high when a structure along ω passes near (x, y).

Ground direction distribution. A small 1D CNN scores each ground column. Resampled in column order onto the direction bins ψ of level ℓ and softmax-normalized, the scores give a distribution $\pi_\ell(\psi)$ over the directions where the view shows structure. It is pose-independent (Appendix B.3).

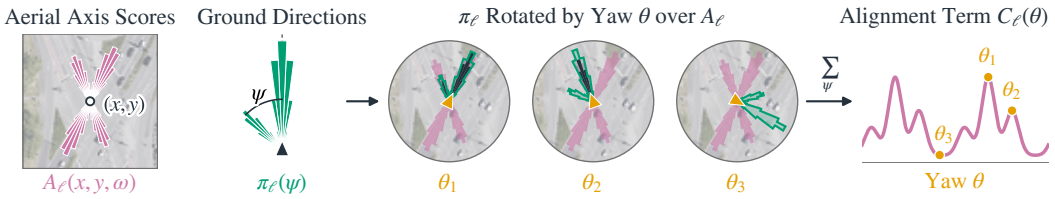

Figure 3: **Structural alignment** at one position (x, y). Yaw θ rotates π_ℓ onto A_ℓ, and $C_\ell(\theta)$ scores agreement (dark): high where the whole view lines up (θ_1), lower where only part does (θ_2).

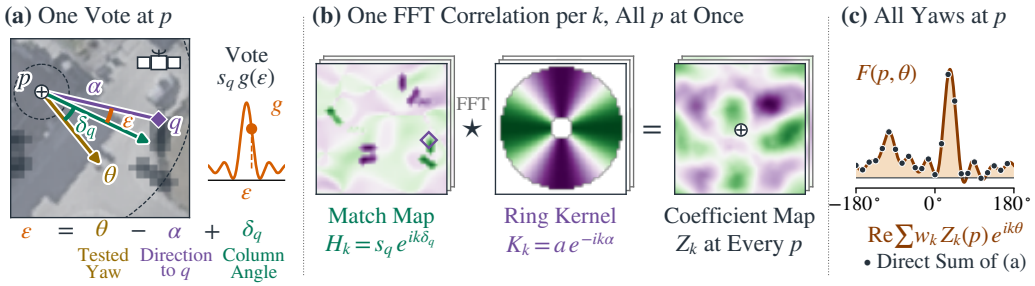

Figure 4: **Fourier scoring. (a)** Match q votes $s_q\, g(\varepsilon)$ for tested yaw θ at position p. In its yaw error ε, the direction α to q depends only on the offset $q - p$, and the column angle δ_q only on match q. Evaluated directly, this repeats for every p, q, and θ. **(b)** For each harmonic k, an FFT correlation of the match map H_k with the fixed ring kernel K_k gives the coefficient map Z_k at every p. **(c)** The coefficients at p give the score at every yaw, equal to the direct sum of (a). Maps show real parts.

Alignment. Following the angle addition of Eq. (2), under yaw θ, bin ψ is read at aerial direction $\theta + \psi$, and C_ℓ weights the axis score there by the photo's evidence in that bin:

$$C_\ell(x, y, \theta) = \sum_\psi \pi_\ell(\psi)\, A_\ell(x, y, \theta + \psi), \tag{4}$$

where ψ runs over the direction bins of level ℓ, angles wrap around the circle, and FFTs give the sum for all yaws at once. Neither input has direction labels. Training through C_ℓ learns how their angle axes correspond. C_ℓ is high only at the few yaws that line up the whole view with the aerial axes, and the matches of F pick the precise one. A long structure such as a street often passes through the camera position and extends both ways, so A_ℓ counts ω and $\omega + \pi$ as one axis. C_ℓ then gives a yaw and its opposite the same score, and F also decides the direction. The structural score S_ℓ adds C_ℓ, weighted by a learned gate λ_ℓ, to the aerial-only pose plausibility score of LAYS (Kang et al., 2026), symmetric in the same way. An auxiliary loss trains C_ℓ at the true position (Appendix B.3).

3.4 Fourier Score: Continuous Match Votes

With 1° bins, implied yaws 0.4° and 0.6° past a bin center land in different bins. The Fourier score replaces the bin indicator of Eq. (3) with a learned kernel g centered at the implied yaw, so no vote is rounded to a bin (Fig. 1c), and the votes of all matches sum to the score (Fig. 1b):

$$F(p, \theta) = \sum_q a(q - p)\, s_q\, g(\theta - \theta_q(p)), \qquad g(\varepsilon) = \sum_{k=0}^{N} w_k \cos(k\varepsilon). \tag{5}$$

Here $\varepsilon = \theta - \theta_q(p)$ is the vote's *yaw error*, the angle from its implied yaw to the tested yaw θ, and g is a sum of cosines of $k\varepsilon$ up to $N = 8$ with weights w_k, shared by all votes. The weights are initialized to concentrate each vote within about $\pm 15°$ of its implied yaw, and are learned (Appendix B.1).

Cost of direct evaluation. Evaluated directly, Eq. (5) adds one vote per tested yaw for every position–match pair. Hard bins add one vote per pair, but their cost still grows with the number of positions times the pixels per ring. Once rounded, the vote no longer splits into offset and match

factors, so every pair needs its own bin index. A sum $\sum_q s_q\, b(q-p)$ with a kernel b of the offset alone is a 2D correlation, which FFTs give for all p at once, but a vote also depends on δ_q and θ.

Separating the three angles. By Eq. (2), the yaw error $\varepsilon = \theta - \theta_q(p) = \theta - \alpha(q-p) + \delta_q$ is a sum of three angles (Fig. 4a). The tested yaw θ is shared by all votes, the direction $\alpha(q-p)$ from p to q depends only on the offset $q-p$, and the column angle δ_q only on match q. Each cosine of g is the real part Re of $e^{ik\varepsilon} = \cos k\varepsilon + i \sin k\varepsilon$, which turns the sum of angles into a product:

$$e^{ik\varepsilon} = \underbrace{e^{ik\theta}}_{\text{tested yaw}}\ \underbrace{e^{-ik\alpha(q-p)}}_{\text{offset } q-p}\ \underbrace{e^{ik\delta_q}}_{\text{match } q}. \tag{6}$$

The tested-yaw factor $e^{ik\theta}$ therefore moves out of the sum over q in Eq. (5). The rest, $Z_k(p)$, is computed once, and each tested yaw only weights it by $e^{ik\theta}$:

$$F(p,\theta) = \mathrm{Re} \sum_{k=0}^{N} w_k\, Z_k(p)\, e^{ik\theta}, \qquad Z_k(p) = \sum_q H_k(q)\, K_k(q-p). \tag{7}$$

The *match map* $H_k(q) = s_q\, e^{ik\delta_q}$ is an image over the aerial pixels, with magnitude s_q and phase $k\delta_q$. The *ring kernel* $K_k(d) = a(d)\, e^{-ik\alpha(d)}$ depends only on the offset d and is precomputed.

All positions and yaws at once. K_k is thus the same around every position and slides over the match map as p varies. Like $\sum_q s_q\, b(q-p)$ above, the sum of products $H_k(q)\, K_k(q-p)$ is a 2D correlation. FFTs compute it as a pointwise product of spectra and give the *coefficient map* Z_k at every position at once (Fig. 4b). $Z_k(p) = \sum_q a(q-p)\, s_q\, e^{-ik\theta_q(p)}$ is the k-th Fourier coefficient of the s_q-weighted implied yaws at p, and $w_k Z_k(p)$ that of the score. Any yaw then costs $O(N)$ (Fig. 4c). Each Z_k is one complex FFT correlation (four real ones for $k \geq 1$, $4N{+}1$ in all, giving $2N{+}1$ real maps). No position–match pair is formed, and no correlation is repeated per yaw bin ($|\Theta| = 360$ at the finest level). Pseudo-code is in Appendix B.2.

3.5 Fusion Across Levels and Training

Both scores are computed at each level ℓ of a coarse-to-fine feature pyramid and added on the pose grid, $R_\ell = F_\ell + S_\ell$ as in Eq. (1). As in prior work (Shi et al., 2023; Kang et al., 2026), each level's score is added to the resampled sum of the coarser levels; the finest sum gives the pose.

Training. At each level, a softmax over yaw turns the summed score at each position p into a yaw distribution $P_\ell(\cdot \mid p)$. The first loss term is $-\log P_\ell(\theta_{\text{gt}} \mid p_{\text{gt}})$, read at the ground-truth position. Since the position is uncertain, a second term rewards the correct yaw at whichever positions score high: it averages $P_\ell(\cdot \mid p)$ with weights $\rho_\ell(p)$, a softmax over positions of each position's best score, and penalizes $-\log \sum_p \rho_\ell(p)\, P_\ell(\theta_{\text{gt}} \mid p)$.

4 Experiments

4.1 Experimental Setup

Datasets and protocols. The main test beds, KITTI (Geiger et al., 2013) and MGL (Sarlin et al., 2023a), have limited-FoV queries, yaw drawn uniformly over $\pm 180°$, and up to ± 20 m initial position error per axis. KITTI's car camera mostly faces the road direction, while MGL's crowd-sourced views often deviate from it. Limited-FoV queries outside driving datasets are less explored in cross-view localization. KITTI uses the official Same Area and Cross Area splits, and Cross Area tests in areas outside its small training set. MGL uses the Amsterdam pinhole views, split, and data preparation of Kang et al. (2026). Panoramic VIGOR (Zhu et al., 2021) uses its Same Area and Cross Area splits with aerial images containing the camera, each panorama shifted by a random yaw in $\pm 180°$.

Baselines. We compare with recent fine-grained cross-view localizers and LAYS (Kang et al., 2026), another yaw estimator (protocols: Appendix D.1). Unmarked values are as published.

Metrics. We report mean and median yaw and location error and yaw recall within 1° and 5°. KITTI adds lateral and longitudinal recall at 1 and 5 m, as commonly reported.

Downstream utility protocol. Loc^2, the most accurate KITTI localizer with public code, estimates the 3-DoF pose, yaw included, from inputs aligned with frozen S-RAY yaw (Appendix D.2).

4.2 Main Results

Limited-FoV KITTI and MGL. S-RAY leads yaw recall within 1° by a wide margin (Tables 1 and 2), raising it over LAYS from 51.07% to 87.65% on KITTI Same Area, from 48.44% to 72.91% on Cross Area, and from 34.81% to 55.22% on MGL, with the lowest median error on every split. MGL, like KITTI Same Area, tests in the city it trains on, but its crowd-sourced views make yaw much harder, and S-RAY still has a median error of 0.87° there. Location is a by-product. Votes are shared along rays, so on KITTI Cross Area, location is constrained laterally (82.83% within 5 m) but poorly along the road (32.22%). MGL localization, hurt by yaw errors, is poor for every method. S-RAY yaw improves it for Loc^2 (Table 6), but this harder setting needs further study.

Margins in mean error and 5° recall are smaller, as the remaining failures are large errors. On KITTI Cross Area, 9.92% of queries have errors near 90° and 7.56% near 180°; in most inspected failures, a wrong position outscores the true one (Appendix A.1). These groups alone cap recall within 5° at 82.5%, and S-RAY reaches 82.37%, so nearly all other errors are below 5°. They also dominate the mean error, on par with the best (23.00° vs. 22.03° for VIRD, whose median is 1.87° vs. our 0.59°). On KITTI Same Area and MGL, S-RAY leads both metrics (1.90°/98.78% and 26.11°/80.47%); LAYS has medians below 2° but means of 46.17° and 44.26°.

Table 1: Yaw comparison on limited-FoV KITTI with 20 m position noise and unknown yaw ($\pm 180^\circ$); full-pose metrics give context. Location error is Euclidean; lateral/longitudinal recalls use the driving frame. *: reproduction reported by VIRD; §: our official-code run, yaw axis only; †: yaw-only; ‡: no public code. **Best**, second, *third* per column and area.

Method	Yaw error (°) ↓		Yaw recall (%) ↑		Lateral (%) ↑		Longitudinal (%) ↑		Location (m) ↓	
	Mean	Med.	R@1°	R@5°	R@1m	R@5m	R@1m	R@5m	Mean	Med.
Same Area										
SliceMatch (Lentsch et al., 2023)	8.71	4.42	11.35	55.82	39.73	87.92	13.63	49.22	9.39	5.41
CCVPE (Xia et al., 2024)	15.01	6.12	8.96	42.75	53.30	85.13	25.84	68.49	6.88	3.47
DenseFlow (Song et al., 2023)*	3.72	0.99	50.62	*90.06*	*73.87*	97.27	14.82	58.50	6.47	4.26
FG^2 (Xia & Alahi, 2025)*	89.64	90.26	0.42	2.73	24.97	75.78	24.57	75.14	5.81	4.26
ARC-Loc (Kim et al., 2026)§	90.61	105.33	12.17	33.08	72.30	96.85	27.83	*78.29*	*4.05*	2.50
Loc^2 (Xia et al., 2026)	9.70	6.17	9.46	41.64	62.18	*97.80*	59.37	97.48	1.85	1.31
LAYS (Kang et al., 2026)†	46.17	0.96	*51.07*	64.75	–	–	–	–	–	–
VIRD (Park et al., 2026)‡	*4.87*	*0.98*	51.20	92.82	79.43	98.41	*29.84*	73.89	5.55	*2.22*
OGH-Net (Zhang et al., 2026)‡	10.17	7.35	13.89	50.18	**86.88**	**99.07**	**86.54**	**99.47**	**0.91**	**0.68**
S-RAY	**1.90**	**0.49**	**87.65**	**98.78**	54.28	97.56	10.71	42.91	9.26	6.52
Cross Area										
SliceMatch (Lentsch et al., 2023)	*23.64*	7.96	*31.69*	31.69	24.00	72.89	7.17	33.12	14.85	11.85
CCVPE (Xia et al., 2024)	77.84	63.84	3.14	14.56	23.42	60.46	11.81	42.12	13.94	10.98
DenseFlow (Song et al., 2023)*	69.51	42.04	10.87	29.57	19.80	47.15	4.35	23.75	25.85	18.84
FG^2 (Xia & Alahi, 2025)*	90.12	90.42	0.65	2.93	9.92	42.48	10.95	43.03	13.58	11.72
ARC-Loc (Kim et al., 2026)§	88.85	87.82	5.78	21.15	*27.49*	*74.30*	6.05	28.06	15.85	12.46
Loc^2 (Xia et al., 2026)	55.18	33.23	2.53	12.78	13.64	50.88	*14.04*	*50.73*	*11.71*	9.11
LAYS (Kang et al., 2026)†	54.94	1.07	48.44	*58.87*	–	–	–	–	–	–
VIRD (Park et al., 2026)‡	**22.03**	*1.87*	30.39	74.87	**45.88**	**85.49**	14.97	52.55	11.12	**5.41**
OGH-Net (Zhang et al., 2026)‡	30.61	13.10	10.12	23.97	25.32	67.45	**30.88**	**55.52**	**10.19**	*9.24*
S-RAY	23.00	**0.59**	**72.91**	**82.37**	36.91	82.83	6.75	32.22	12.59	10.13

Panoramic VIGOR. S-RAY has the lowest Same Area mean and median yaw error ($8.74/0.88^\circ$) and the lowest Cross Area mean (10.32°; Table 3). Only VIRD has a lower Cross Area median (1.17° vs. 1.28°), with a 60% higher mean error. The margins are smaller than with a limited FoV, since matches around a panorama agree only near the true yaw, while a narrow view of a repetitive street leaves many yaws nearly as plausible, which S-RAY resolves.

Table 2: Primary yaw comparison on limited-FoV MGL (Amsterdam) pinhole images with unknown yaw ($\pm 180°$); full-pose metrics give context. †: yaw-only method; §: our official-code run (ARC-Loc: yaw axis only). **Best**, second, *third* per column.

Method	Yaw				Location (m) ↓	
	Mean (°) ↓	Med (°) ↓	$< 1°$ ↑	$< 5°$ ↑	Mean	Med
CCVPE (Xia et al., 2024)§	66.56	16.10	6.55	28.36	16.43	12.79
DenseFlow (Song et al., 2023)§	53.27	*11.75*	*9.38*	*34.80*	13.30	12.34
ARC-Loc (Kim et al., 2026)§	89.21	88.12	4.36	18.52	11.27	9.30
Loc^2 (Xia et al., 2026)§	36.74	14.41	4.99	23.98	**10.28**	**8.26**
LAYS (Kang et al., 2026)†	*44.26*	1.79	34.81	63.29	–	–
S-RAY	**26.11**	**0.87**	**55.22**	**80.47**	*12.03*	*9.51*

Table 3: Unknown-yaw VIGOR: yaw (°) and location (m) errors. *: third-party evaluation; §: our official-code run, yaw axis only; †: yaw-only; ‡: no public code. **Best**, second, *third* per column.

Method	Same Area				Cross Area			
	Yaw (°) ↓		Location (m) ↓		Yaw (°) ↓		Location (m) ↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
SliceMatch (Lentsch et al., 2023)	25.46	4.71	6.49	3.13	25.97	4.51	7.22	3.31
CCVPE (Xia et al., 2024)	12.83	6.62	*3.74*	*1.42*	27.78	13.58	5.41	*1.89*
DenseFlow (Song et al., 2023)	11.20	1.59	4.97	1.90	17.63	2.94	7.67	3.67
FG^2 (Xia & Alahi, 2025)	15.02	2.94	8.95	7.32	31.41	5.45	10.02	8.14
Loc^2 (Xia et al., 2026)	9.54	2.00	3.94	1.78	11.67	2.21	4.23	2.09
ARC-Loc (Kim et al., 2026)§	89.45	88.75	10.62	8.91	89.60	89.25	11.26	9.39
GeoDistill* (Tong et al., 2025)	16.11	3.06	4.73	2.90	24.08	5.12	4.65	2.71
Second-Order Iterative (Feng et al., 2026)	9.44	*1.06*	**2.89**	**1.01**	*14.79*	*1.98*	**3.93**	1.57
LAYS (Kang et al., 2026)†	35.70	1.94	–	–	45.50	5.45	–	–
VIRD (Park et al., 2026)‡	*9.52*	0.96	3.00	1.20	16.50	**1.17**	*4.61*	**1.55**
OGH-Net (Zhang et al., 2026)‡	12.78	2.53	3.98	1.55	20.22	3.72	5.03	2.33
S-RAY	**8.74**	**0.88**	5.71	2.72	**10.32**	1.28	6.05	2.64

Table 4: Ablations on KITTI/VIGOR Same Area.

Variant		KITTI Same Area					VIGOR Same Area				
		Yaw Mean ↓	Yaw Med. ↓	$< 1°$ ↑	$< 5°$ ↑	Loc. Med. ↓	Yaw Mean ↓	Yaw Med. ↓	$< 1°$ ↑	$< 5°$ ↑	Loc. Med. ↓
Module	S only	87.90	89.77	40.47	46.28	11.28	89.71	89.95	26.32	42.30	11.44
	F only	3.55	1.31	40.26	91.94	**5.08**	11.31	2.20	25.29	77.57	**2.47**
Fourier	Hard bin	4.14	0.56	78.69	96.34	6.67	13.63	1.26	41.90	84.01	3.98
	$N=1$	7.85	0.56	82.11	93.93	9.02	14.98	1.01	49.70	84.19	6.86
Ours		**1.90**	**0.49**	**87.65**	**98.78**	6.52	**8.74**	**0.88**	**55.01**	**88.92**	2.72

Table 5: Scorer time over the same poses of one MGL image.

	Fourier (S-RAY)	Hard bin (LAYS)
Time (ms) ↓	**0.83**	19.16

Table 6: Loc^2 errors without and with S-RAY yaw.

S/C: Same/Cross		Loc. (m) ↓ Mean	Loc. (m) ↓ Med.	Yaw (°) ↓ Mean	Yaw (°) ↓ Med.
KITTI S	Loc^2	1.85	1.31	9.70	6.17
	+S-RAY	**0.99**	**0.68**	**4.02**	**2.75**
KITTI C	Loc^2	11.71	9.11	55.18	33.23
	+S-RAY	**9.58**	**6.82**	**29.37**	**9.02**
MGL	Loc^2	10.28	8.26	36.74	14.41
	+S-RAY	**8.95**	**6.78**	**30.75**	**8.53**
VIGOR S	Loc^2	3.94	1.78	9.54	2.00
	+S-RAY	**3.61**	**1.64**	**8.94**	**1.46**
VIGOR C	Loc^2	4.23	2.09	11.67	2.21
	+S-RAY	**4.01**	**2.02**	**8.60**	**1.82**

Downstream localization. With S-RAY yaw, Loc^2's median location error falls from 1.31 to 0.68 m on KITTI Same Area, 9.11 to 6.82 m on Cross Area, and 8.26 to 6.78 m on MGL, with smaller VIGOR gains (Table 6). Its own yaw has a higher median error than S-RAY's.

4.3 Ablation Studies

F and S are complementary. Retrained without the other score (Table 4), F alone (25.29%) or S alone (26.32%) reaches less than half the 55.01% recall within $1°$ of $F + S$ on VIGOR Same Area, and on KITTI Same Area recall rises from 40.26% (F) and 40.47% (S) to 87.65%. Without S, voting trains poorly, as LAYS reports for hard bins (Kang et al., 2026); S narrows each position to a few candidate yaws. S scores a yaw and its opposite equally (Sec. 3.3), so its recall counts only the queries where the tie falls on the true yaw, and its mean error is near $90°$; in Fig. 5, S leaves these two yaws and F picks the true one. S also helps when the view does not follow the road, as in many

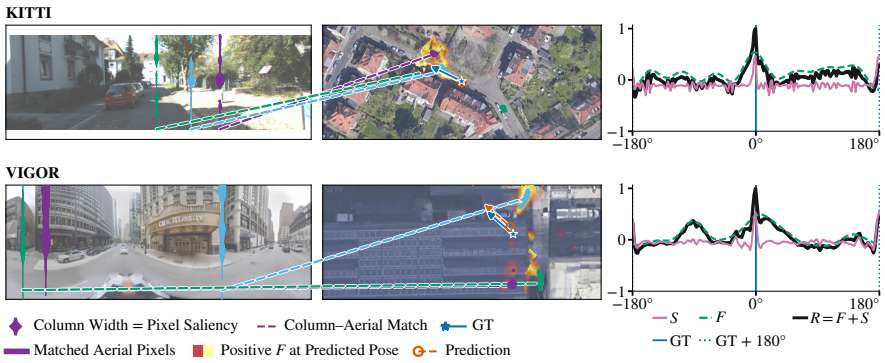

Figure 5: **How the two scores combine** on KITTI (top) and VIGOR (bottom) examples correct in position and yaw. Left: ground columns linked to their matched aerial pixels, over the positive Fourier contribution at the predicted pose. Right: S, F, and $R = F+S$ over yaw at the true position.

MGL views and in VIGOR panoramas rolled by a random yaw, and its ground direction distribution peaks at streets and other structures without direction labels (Appendix Fig. S4). S raises KITTI's location median from 5.08 to 6.52 m, as a layout fits anywhere along a street.

Continuous scoring outperforms hard bins. The hard-bin control keeps everything but counts each vote in its nearest bin. Fourier scoring raises recall within 1° from 78.69% to 87.65% on KITTI and from 41.90% to 55.01% on VIGOR. On KITTI the medians barely differ (0.56° vs. 0.49°), but hard bins leave 3.66% of queries beyond 5° against 1.22%, more than doubling the mean error (4.14° vs. 1.90°). VIGOR's mean falls from 13.63° to 8.74°. Rounding across bin edges splits the true pose's support, so a wrong pose wins more often. With one harmonic (N=1), a vote is the cosine of the yaw error. This score beats hard bins in 1° recall, but recall falls from 87.65% to 82.11% on KITTI and from 55.01% to 49.70% on VIGOR, so the kernel must also be sharp (Appendix B.1).

Scorer cost. Timed alone at the finest level (Table 5), LAYS's hard-bin scorer bins every position–match pair and takes 19.16 ms; Fourier scoring forms no pairs and takes 0.83 ms. Across levels, scoring and pose extraction take 46 ms for S-RAY, structural score included, and 73 ms for LAYS. End to end, S-RAY takes 84 ms per MGL query and LAYS, with the author-provided code, 103 ms (Appendix D.3).

5 Conclusion

S-RAY estimates camera yaw from a ground photo and an aerial image when the camera position is uncertain. An aerial pixel matched to a photo column fixes where the column points from each candidate position, and with the column's angle this implies one yaw, a vote shared along the ray through the pixel. Yaw estimation must aggregate two complementary kinds of evidence over all candidate poses: precise votes, each checking one direction, and the coarser layout of the whole view. A structural score, learned from pose labels only, compares the directions where the photo shows structure, under each yaw, with aerial structure axes predicted at each position, keeping a few promising yaws. The votes decide among them; each adds a learned kernel of a few cosines at its implied yaw instead of a hard bin. By angle addition, a vote is a match factor times an offset factor depending only on relative position, so one FFT correlation per cosine, a pointwise product of spectra, scores all poses without forming any position–match pair. The scorer is 23× faster than the prior hard-bin scorer. S-RAY raises yaw recall within 1° over LAYS on KITTI and MGL, from 48.44% to 72.91% on KITTI Cross Area, has the lowest mean yaw error on VIGOR, and improves Loc^2 localization as a yaw initialization. Position remains a by-product, as votes are shared along rays, and the remaining large errors, near 90° or 180°, mostly come with a wrong position (Appendix A.1). Dedicated localizers and multi-frame evidence could address both.

REPRODUCIBILITY STATEMENT

Sections 3.3–3.5 define the scores and the training objective. Appendix B gives the kernel initialization, scoring pseudo-code, and structural score; Appendix C the architecture, inference, optimization, and training settings; Appendix D the baseline and downstream protocols; and Appendix D.3 the timing setup. All experiments use the public KITTI, MGL, and VIGOR benchmarks with the splits of Sec. 4.

ETHICS STATEMENT

We adhere to the ICLR Code of Ethics. This work uses only public benchmarks under their licenses and involves no human subjects or new data collection.

AI USE STATEMENT

We used LLMs to polish writing, draft text, and find related work, and LLM agents helped implement methods, run experiments, and analyze results. We did not use LLMs to generate data. The authors verified all code, results, and text and take full responsibility for the content.

REFERENCES

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

Jan Brejcha and Martin Čadík. State-of-the-art in visual geo-localization. *Pattern Analysis and Applications*, 20(3):613–637, March 2017. ISSN 1433-755X. doi: 10.1007/s10044-017-0611-1.

Yao Du, Carlos Mateo, and Omar Tahri. A multilayer perceptron-based spherical visual compass using global features. *Sensors*, 24(7):2246, March 2024. ISSN 1424-8220. doi: 10.3390/s24072246.

Mingtao Feng, Xin Li, Jianqiao Luo, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Second-order robust iterative pose optimization for fine-grained cross-view localization. *IEEE Transactions on Image Processing*, 35:3157–3171, 2026. doi: 10.1109/TIP.2026.3673969.

W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(9):891–906, 1991. doi: 10.1109/34.93808.

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.

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 (CVPR)*, pp. 7258–7267, 2018.

Taeho Kang, Nairan Zhang, Yelin Kim, Yujiao Shi, and Youngki Lee. Cross-view yaw estimation in location uncertainty with line-aligning yaw scoring. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer Nature Switzerland, 2026.

Hyeongsik Kim, Mincheol Kim, Heejoon Moon, and Je Hyeong Hong. ARC-Loc: Leveraging azimuthal ray convergence as a geometric cue for direct cross-view localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 205–222. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37023-5_12.

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 (CVPR)*, 2023.

Ying Liu, Junyi Tao, Da Kong, Yu Zhang, and Ping Li. A visual compass based on point and line features for uav high-altitude orientation estimation. *Remote Sensing*, 14(6):1430, March 2022. ISSN 2072-4292. doi: 10.3390/rs14061430.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 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 (CVPR)*, pp. 33693–33702, June 2026.

Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21632–21642. IEEE, June 2023a. doi: 10.1109/cvpr52729.2023.02072.

Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. SNAP: Self-supervised neural maps for visual positioning and semantic understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16989–16999. IEEE, June 2022. doi: 10.1109/cvpr52688.2022.01650.

Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020b.

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 (ICCV)*, pp. 21516–21526, 2023.

Yujiao Shi, Hongdong Li, Akhil Perincherry, and Ankit Vora. Weakly-supervised camera localization by ground-to-satellite image registration. In *European Conference on Computer Vision (ECCV)*, pp. 39–57, 2024.

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.

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 (CVPR)*, pp. 6488–6497, 2021.

Shaowen Tong, Zimin Xia, Alexandre Alahi, Xuming He, and Yujiao Shi. GeoDistill: Geometry-guided self-distillation for weakly supervised cross-view localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25357–25366, 2025. doi: 10.1109/ICCV51701.2025.02352.

Qiwei Wang, Shaoxun Wu, and Yujiao Shi. BevSplat: Resolving height ambiguity via feature-based gaussian primitives for weakly-supervised cross-view localization. *Advances in Neural Information Processing Systems*, 38:156668–156696, 2025.

Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, and Hongdong Li. View consistent purification for accurate cross-view localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8197–8206, 2023a.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14843–14852. IEEE, June 2024. doi: 10.1109/cvpr52733.2024.01406.

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 (NeurIPS)*, pp. 5301–5319, 2023b. doi: 10.52202/075280-0234.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3961–3969, 2015.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7168–7177, 2017. doi: 10.1109/CVPR.2017.758.

Hang Wu, Zhenghao Zhang, Siyuan Lin, Xiangru Mu, Qiang Zhao, Ming Yang, and Tong Qin. MapLocNet: Coarse-to-fine feature registration for visual re-localization in navigation maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

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 (CVPR)*, pp. 6362–6372, 2025. doi: 10.1109/CVPR52734.2025.00596.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. *Visual Cross-View Metric Localization with Dense Uncertainty Estimates*, pp. 90–106. Springer Nature Switzerland, 2022. ISBN 9783031198427. doi: 10.1007/978-3-031-19842-7_6.

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, May 2024. ISSN 1939-3539. 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 (ICLR)*, 2026.

Yangchun Zhang, Xudong Kang, Puhong Duan, Shutao Li, and Ze Song. Orientation-guided homography for fine-grained cross-view localization. *IEEE Transactions on Image Processing*, 35: 6139–6152, 2026. doi: 10.1109/TIP.2026.3697650.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3640–3649, 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 (CVPR)*, pp. 1162–1171, 2022.