HOW TO REDUCE LOCALIZATION AMBIGUITY? GEOMETRY-SEMANTIC CONSTRAINED BEV REPRESENTATION LEARNING FOR SATELLITE-GROUND LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Satellite-ground localization estimates the planar position and yaw orientation of a ground camera within a geo-referenced satellite image of its surroundings. The predominant approach to this task maps features from ground-view images and satellite references into a shared bird's-eye-view (BEV) space and then establishes spatial correspondences between them. Although effective, this approach still faces ambiguity in feature placement and descriptor matching. When mapping ground-view features into BEV space, insufficient depth constraints allow the same feature to be assigned to different distances along the viewing direction, creating geometric ambiguity in BEV feature placement. Meanwhile, similar appearances at different locations create descriptor matching ambiguity, and existing descriptor learning lacks explicit semantic supervision to distinguish these locations. To reduce these ambiguities, we propose GeoSem-BEV, a geometry-semantic constrained BEV representation learning method. First, we geometrically constrain ground-view BEV feature placement through radial depth supervision for distance assignment and vertical height supervision for height aggregation. Then, shared explicit semantic supervision further promotes consistent semantic predictions across views and helps distinguish different locations with similar semantics. These constraints jointly improve feature placement and descriptor discriminability, thus improving the satellite-ground localization performance of state-of-the-art models by a large margin. Qualitative and quantitative results demonstrate the effectiveness of GeoSem-BEV in enhancing these models. On VIGOR with unknown orientation, GeoSem-BEV reduces mean orientation error relative to the corresponding state-of-the-art method by 37.2% and 38.1% in the cross-area and same-area settings, respectively. The corresponding errors were reduced by 10.8% and 15.6% on DReSS-D. On KITTI-CVL, GeoSem-BEV reduces the same-area mean orientation error by 26.8% relative to the corresponding baseline under $\pm 10^{\circ}$ orientation noise.

1 INTRODUCTION

Satellite-ground localization estimates the planar position and yaw orientation of a ground camera within a geo-referenced satellite image, supporting applications in vehicle localization and navigation, aerial-platform geolocation, and remote-sensing image registration for urban mapping and monitoring (Hu & Lee, 2020; Shi & Li, 2022; Zheng et al., 2020; Shan et al., 2014; Yu et al., 2021). Given a coarse Global Navigation Satellite System (GNSS) estimate of the surrounding area, satellite-ground localization can match ground observations with corresponding content in a satellite image to further refine the camera pose (Shi & Li, 2022). However, substantial appearance differences between street-level ground images and overhead satellite observations make accurate spatial correspondence estimation challenging.

In recent years, satellite-ground localization has received increasing attention, and mapping ground-view features into a shared bird's-eye-view (BEV) space has become a predominant approach. This shared top-down coordinate system reduces the viewpoint gap between ground and satellite views

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.

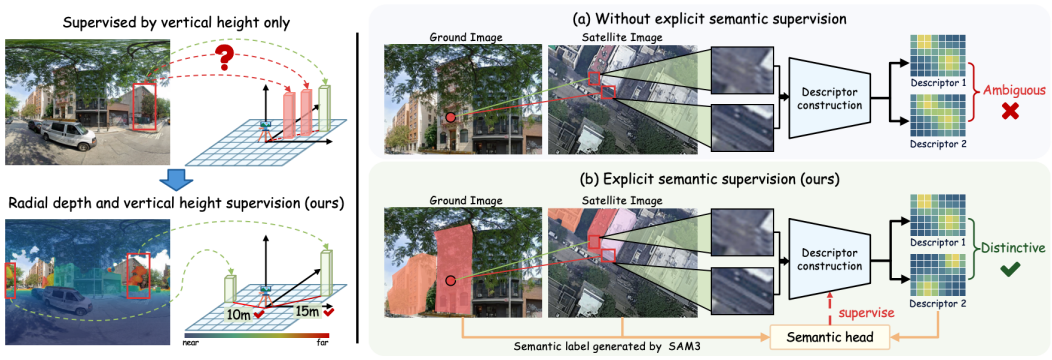


Figure 1: Two ambiguities in BEV-space feature matching. Left: Insufficient radial-depth constraints leave BEV feature placement ambiguous along viewing rays. (a) Upper right: Similar appearances at different locations can produce ambiguous descriptors without explicit semantic supervision. (b) Lower right: Explicit semantic supervision helps distinguish them.

and makes their spatial correspondences explicit. Existing methods construct and match BEV representations through geometric projection, learned feature transformation, displacement fields, or dense 3D feature sampling (Shi et al., 2023; Fervers et al., 2023; Song et al., 2023; Wang et al., 2023; Xia & Alahi, 2025). Despite their effectiveness, these methods share a common challenge: the construction of the ground-view BEV representation remains insufficiently constrained along viewing rays. Loc^2 highlights this issue by observing that BEV representation construction can introduce ray-directional distortions and lose height information. It therefore matches features directly in the image planes, which we refer to as image-space feature matching, before lifting matched points using depth. This pipeline requires a monocular depth prediction model at inference to determine the spatial coordinates of matched points (Xia et al., 2026). Unlike Loc^2, ViewBridge continues to improve BEV representation construction through visible-surface modelling and height consistency constraints (Xia et al., 2025a). Its supervision constrains vertical height, specifying where scene content lies along the vertical axis. However, it ignores the ambiguity in radial depth, which determines how far that content lies from the camera along the viewing ray. Consequently, the same image feature may contribute to BEV locations at different distances along the viewing ray, leaving geometric ambiguity in feature placement, as illustrated in the upper-left part of Fig. 1.

Besides geometric ambiguity in feature placement, similar appearances at different locations can create descriptor matching ambiguity. For example, visually similar regions at different locations may produce similar descriptors, allowing incorrect locations to receive high matching scores, as illustrated in Fig. 1(a). Meanwhile, drastic viewpoint differences can make the same location appear visually different in ground and satellite images, causing correct correspondences to receive low matching scores. These appearance differences make semantic cues useful for associating corresponding locations. Prior work uses semantic layouts for ground-to-GIS matching (Castaldo et al., 2015) or semantic map elements for localization (Sarlin et al., 2023a). Because these cues come from external maps, map errors can add uncertainty to cross-view matching (Castaldo et al., 2015). Within BEV matching, existing methods learn descriptors mainly from appearance, matching, and pose supervision, without explicit semantic supervision (Xia & Alahi, 2025; Xia et al., 2025a).

To reduce these ambiguities, we propose GeoSem-BEV, a geometry-semantic constrained BEV representation learning method that jointly constrains feature placement and descriptor learning within BEV-space feature matching. First, we impose geometric constraints on how ground-view features are assigned to and aggregated on the BEV grid. Radial depth supervision constrains their distance from the camera along viewing rays, while vertical height supervision constrains the aggregation of features sampled at different heights. We then introduce shared explicit semantic supervision for ground- and satellite-view descriptors. SAM3 (Carion et al., 2026) generates semantic masks for paved surfaces, buildings, and vegetation as training targets. A shared semantic head predicts semantic classes from ground- and satellite-view descriptors, allowing these targets to supervise descriptor learning in both views, as illustrated in Fig. 1(b). This semantic supervision promotes consistent semantic predictions at corresponding locations. To further distinguish spatially different locations with similar semantics, we additionally select same-class hard negatives outside the correct spatial

neighbourhood and penalize their high matching scores. Together, shared semantic supervision and same-class hard negatives improve descriptor discriminability, while geometric constraints improve feature placement. We evaluate GeoSem-BEV across ViewBridge (Xia et al., 2025a), FG^2 (Xia & Alahi, 2025), and DenseFlow (Song et al., 2023) on VIGOR (Zhu et al., 2021), KITTI-CVL (Shi et al., 2022), and DReSS-D (Xia et al., 2025b) to assess whether the proposed constraints improve localization across BEV matching frameworks. On VIGOR with unknown orientation, GeoSem-BEV reduces mean orientation error relative to the corresponding state-of-the-art method by 37.2% and 38.1% in the cross-area and same-area settings, respectively. The corresponding errors were also reduced by 10.8% and 15.6% on DReSS-D. These results demonstrate that GeoSem-BEV improves satellite-ground localization accuracy across different matching frameworks while providing more discriminative correspondence evidence.

Our main contributions are as follows: **(i)** We analyse two sources of localization ambiguity in BEV-space feature matching: geometric ambiguity caused by insufficient constraints on feature placement along viewing rays, and descriptor matching ambiguity arising from similar appearances at different locations. **(ii)** We propose GeoSem-BEV, a geometry-semantic constrained BEV representation learning method. Geometric constraints guide BEV feature placement through radial depth and vertical height supervision. Shared explicit semantic supervision promotes cross-view semantic consistency, while same-class hard negatives further improve descriptor discriminability. **(iii)** We evaluate GeoSem-BEV by applying it to three state-of-the-art BEV localization models, ViewBridge, FG^2, and DenseFlow, on VIGOR, KITTI-CVL (Shi et al., 2022), and DReSS-D. On VIGOR with unknown orientation, GeoSem-BEV reduces mean orientation error relative to the corresponding state-of-the-art method by 37.2% and 38.1% in the cross-area and same-area settings, respectively, and achieves the lowest mean translation error among compared methods in both settings. On DReSS-D and KITTI-CVL, it also achieves competitive results, improving orientation and localization metrics over the corresponding baseline models.

2 Related Work

2.1 Satellite-Ground Localization

Satellite-ground localization has progressed from retrieving a matching reference image to estimating the pose of a ground camera within that image. Early cross-view methods mainly addressed ground-to-aerial image retrieval by learning shared representations (Lin et al., 2013; 2015; Workman et al., 2015). Subsequent retrieval models improved cross-view representation learning with convolutional global descriptors (Hu et al., 2018) and transformer-based spatial context (Zhu et al., 2022). Later studies moved beyond image retrieval to estimate the position and orientation of a ground camera from overhead imagery (Vo & Hays, 2016; Shi & Li, 2022). VIGOR relaxed the one-to-one retrieval assumption by evaluating camera locations beyond satellite-image centres (Zhu et al., 2021), while subsequent methods explored differentiable geometric alignment (Shi & Li, 2022), pose-dependent descriptor comparison (Lentsch et al., 2023), and convolutional cross-view pose estimation (Xia et al., 2023). These developments shifted the focus from image retrieval toward fine-grained pose estimation and explicit spatial correspondence. They also support applications such as vehicle localization with temporal filtering (Hu & Lee, 2020) and drone-view target localization and navigation (Zheng et al., 2020). For fine-grained pose estimation, mapping ground-view and satellite features into a shared BEV space has become a major approach.

Within BEV-space feature matching, earlier methods estimate pose distributions (Fervers et al., 2023) or homographies from cross-view correlations (Wang et al., 2023). Correspondence-based approaches make spatial associations explicit and recover pose from matched coordinates. DenseFlow predicts displacement fields between the two views (Song et al., 2023), while FG^2 matches BEV descriptors and samples correspondences for geometric alignment (Xia & Alahi, 2025). ViewBridge further improves BEV representation construction through visible-surface modelling and matching-score refinement (Xia et al., 2025a). Loc^2 identifies distortions and information loss introduced during BEV transformation and explores an image-space feature matching route, which matches features in the image planes before lifting the matched points using depth (Xia et al., 2026). These studies motivate further investigation of the BEV-space feature matching route, where the construction of ground-view BEV representations and the learning of matching descriptors remain

subject to ambiguity. Our work focuses on reducing geometric ambiguity in feature placement and descriptor matching ambiguity within this BEV-based framework.

2.2 Geometric Constraints for BEV Representation Construction

BEV representation construction maps image features to a top-down grid, with geometric constraints guiding their spatial placement. Related work in autonomous-driving perception explores depth-based lifting and attention-based feature aggregation. Lift-Splat-Shoot distributes image features over depth hypotheses before aggregating them on a BEV grid (Philion & Fidler, 2020). BEVDepth further introduces explicit depth supervision to improve this transformation for 3D object detection (Li et al., 2023). BEVFormer uses spatial cross-attention and temporal self-attention to construct BEV representations for 3D detection and map segmentation (Li et al., 2024). Although these methods target multi-camera perception, they provide two relevant principles for BEV construction: assigning image features to spatial locations and aggregating features over the BEV grid.

Satellite-ground localization applies similar construction operations under a different camera geometry. DenseFlow projects ground features under a fixed camera-height assumption, while GGCVT combines geometric projection with learned attention (Song et al., 2023; Shi et al., 2023). FG^2 samples ground-view features at 3D queries and learns to aggregate them over height; ViewBridge further introduces visible-surface modelling and explicit height consistency constraints (Xia & Alahi, 2025; Xia et al., 2025a). These methods constrain how features are distributed along the vertical axis, yet they do not directly supervise radial depth along viewing rays. Consequently, the same image feature may contribute to BEV locations at different distances from the camera, leaving geometric ambiguity in BEV feature placement.

2.3 Semantic Information for Cross-View Matching

Semantic information provides cues beyond visual appearance for associating corresponding content across viewpoints. Semantic Cross-View Matching uses semantic categories and their spatial layout to associate street-level observations with GIS maps (Castaldo et al., 2015). OrienterNet aligns image-derived BEV features with semantic elements in public maps for image localization (Sarlin et al., 2023a). SNAP learns neural maps from ground and overhead imagery, with semantic structure emerging during map learning without explicit semantic labels (Sarlin et al., 2023b). For cross-view remote-sensing applications, SG-BEV further combines satellite and ground observations for building-attribute segmentation (Ye et al., 2024). These studies demonstrate that semantic information can complement appearance and geometry in cross-view association. However, methods relying on external semantic maps may inherit errors, missing entries, or temporal changes in the reference-side annotations, which can introduce uncertainty into matching (Castaldo et al., 2015).

For local BEV feature matching, semantic information is also relevant to descriptor ambiguity caused by similar appearances at different locations. Prior work has studied this issue in joint cross-view retrieval and offset calibration by modelling semantic ambiguity through offset uncertainty (Feng et al., 2025). General descriptor-learning methods use supervised contrastive objectives or hard-negative mining to improve class-level discrimination (Khosla et al., 2020; Mishchuk et al., 2017). However, these objectives alone do not establish spatial correspondence, since different buildings or road segments may share the same semantic class. Existing BEV matching methods primarily learn descriptors through appearance, matching, and pose supervision, without explicitly supervising the semantic content encoded by the descriptors. This motivates explicit semantic supervision for ground- and satellite-view descriptors, which provides scene-content information beyond appearance while preserving the distinction between spatially different locations.

3 Method

3.1 Overview

Given a ground-view image G and a geo-referenced satellite image S covering its surroundings, satellite-ground localization estimates the planar camera pose $\hat{T} = (\hat{t}_x, \hat{t}_y, \hat{\theta})$, where $\hat{t}_x$ and $\hat{t}_y$ denote translation in the satellite reference frame and $\hat{\theta}$ denotes yaw orientation. Let T^* denote

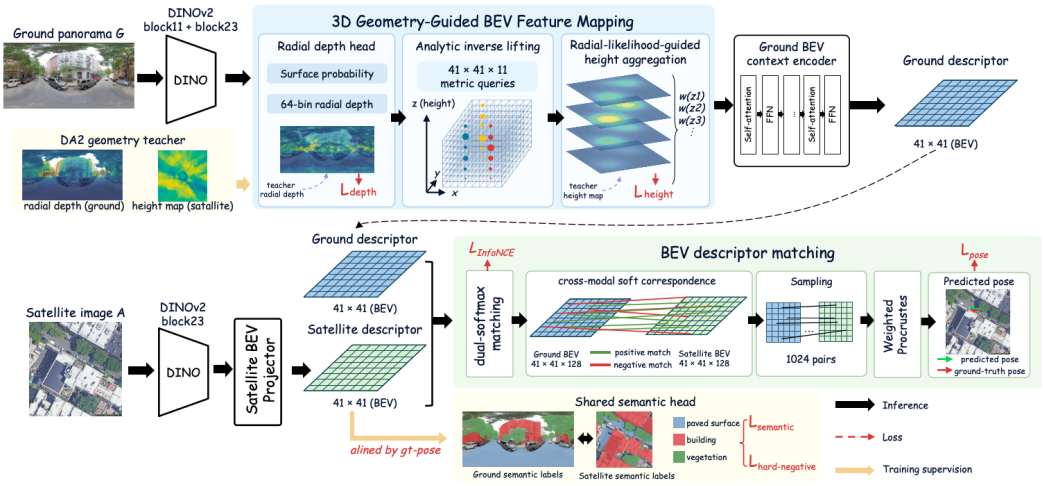


Figure 2: GeoSem-BEV with FG^2 as the base model. Geometric supervision constrains radial assignment and height aggregation of ground-view features. A shared semantic head supervises both descriptor fields, and same-class hard negatives penalize spatially incorrect matches.

the corresponding ground-truth pose. The pose is recovered from cross-view correspondences between the ground observation and the satellite reference. In the BEV-space feature matching route, ground-view features are first assigned to a top-down grid and then matched with satellite features in the shared coordinate system. However, insufficient constraints on this assignment can create radial placement ambiguity, while insufficient descriptor discrimination can favour visually similar but spatially incorrect matches. GeoSem-BEV addresses these limitations by constraining both BEV feature placement and descriptor learning. The method is designed as a framework-agnostic enhancement: each evaluated model retains its original correspondence estimator, pose solver, and base localization objective. GeoSem-BEV adds supervision at two complementary stages of the shared pipeline. Geometry supervision acts during BEV representation construction, constraining ground-view feature placement on the BEV grid. Semantic supervision acts on the resulting descriptor fields, encouraging descriptors from corresponding ground and satellite locations to encode compatible scene content while separating spatially different candidates.

Figure 2 illustrates GeoSem-BEV with FG^2 as the base model (Xia & Alahi, 2025). The following formulation describes the geometric and semantic constraints used across the evaluated BEV matching frameworks. The pipeline extracts ground-view and satellite-view features, samples ground features at metric 3D queries, and aggregates them into a 2D BEV grid. Radial-depth and vertical-height supervision constrain feature placement and height aggregation, after which projection heads produce descriptors for correspondence estimation. During training, SAM3-derived targets supervise a shared semantic head, while pose-aligned consistency and same-class hard negatives improve descriptor discrimination. At inference, only the ground and satellite images are required. Framework-specific implementations are described in Appendix B.6.

3.2 Geometry-Constrained BEV Feature Construction

The geometric constraint addresses the construction of a ground-view BEV representation before correspondence estimation. For each image feature, the viewing ray determines a direction, but the feature can still be assigned to multiple distances along that ray. We therefore distinguish radial depth from vertical height: radial depth specifies the distance assignment along a viewing ray, whereas vertical height specifies how samples at different heights contribute to one BEV cell. Radial-depth supervision and vertical-height supervision jointly constrain these two parts of BEV feature placement.

Sampling ground-view features. Let $q_{ijk} = (x_i, y_j, z_k)$ be a fixed metric 3D query in the vertical column associated with BEV cell (x_i, y_j). The ground coordinate frame is centred at the camera,

with two horizontal axes and a vertical axis. The camera projection $\pi(q_{ijk})$ identifies the image location from which a feature is sampled:

$$F_{3D}(q_{ijk}) = \text{sample}(F_G, \pi(q_{ijk})). \tag{1}$$

Each BEV cell receives candidate appearance features at the queried heights. Sampling establishes a viewing direction for each candidate, but does not determine whether a surface exists at that distance.

Radial assignment and height aggregation. A depth head predicts a surface probability $p_s(u, v)$ and a discrete conditional distribution $p_r(b \mid u, v)$ over radial depth bins with centres d_b. Let $\widetilde{p}_r(r \mid u, v)$ denote its linearly interpolated value at distance r. For a query at radial distance $r_{ijk} = \|q_{ijk}\|_2$, its radial likelihood is

$$\ell_{ijk} = p_s(\pi(q_{ijk}))\, \widetilde{p}_r(r_{ijk} \mid \pi(q_{ijk})). \tag{2}$$

Queries beyond the supported depth range receive zero likelihood. This distinguishes candidate locations along the same viewing ray using the predicted surface distance. The radial likelihood is combined with the height aggregation weights to obtain the contribution weight w_{ijk}. The ground BEV feature is

$$F_{\text{BEV}}(x_i, y_j) = \sum_k w_{ijk} F_{3D}(q_{ijk}). \tag{3}$$

Radial predictions therefore constrain where sampled ground-view features contribute to the BEV grid. The aggregation weights also receive vertical-height supervision, as described below.

Geometry supervision. Frozen monocular depth teachers provide pseudo-targets for radial distance and relative height during training. Ground-view targets supervise surface presence, radial depth, expected depth, and uncertainty, while the ground-truth pose T^* aligns satellite height targets with the ground BEV grid. The geometry objective combines these terms:

$$\mathcal{L}_{\text{geom}} = \mathcal{L}_{\text{surface}} + \mathcal{L}_{\text{radial-dist}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{uncertainty}} + \mathcal{L}_{\text{height}}. \tag{4}$$

At inference, the image-based student predicts the geometric weights without running the teacher. Thus, the teacher supplies training-time geometric targets, while the student produces the constrained BEV representation from the input images alone at inference.

3.3 Explicit Semantic Supervision for BEV Descriptors

After geometric supervision constrains ground-view feature placement, semantic supervision constrains descriptor content. Matching and pose supervision alone can still assign high scores to visually similar regions at different locations.

Semantic targets and shared prediction. The ground BEV feature map and satellite feature map are encoded into descriptor fields D_G and D_S. We use SAM3 (Carion et al., 2026) to generate masks for paved surfaces, buildings, and vegetation in both images. Ground and satellite semantic masks are converted into soft targets on their respective BEV grids. The resulting soft class targets Y_G and Y_S are fixed supervision signals, distinct from the student's predicted class distributions.

A shared head h predicts class probabilities $P_v = \text{softmax}(h(D_v))$, for $v \in \{G, S\}$. The soft cross-entropy loss $\mathcal{L}^v_{\text{CE}}$ compares these predictions with the SAM3-derived targets Y_v over valid BEV cells. Gradients through the shared head teach the matching descriptors to encode semantics together with appearance.

We warp P_G with the ground-truth pose and enforce Jensen–Shannon consistency with P_S at corresponding locations:

$$\mathcal{L}_{\text{semantic}} = \tfrac{1}{2}(\mathcal{L}^G_{\text{CE}} + \mathcal{L}^S_{\text{CE}}) + \mathcal{L}_{\text{JS}}(P_G^{T^*}, P_S). \tag{5}$$

Here, $P_G^{T^*}$ denotes the warped ground-view class probabilities.

Distinguishing same-class locations. Same-class hard negatives are selected outside the positive spatial neighbourhood to distinguish locations with similar semantics. The directional hard-negative

loss contrasts the positive correspondence with these spatially incorrect candidates using the matching logits. The reverse direction is defined analogously, and we average both directions:

$$\mathcal{L}_{\mathrm{HN}} = \tfrac{1}{2}(\mathcal{L}^{\mathrm{HN}}_{G\to S} + \mathcal{L}^{\mathrm{HN}}_{S\to G}). \tag{6}$$

Together, these terms align descriptors of corresponding content and distinguish visually similar regions at different locations. The semantic head and SAM3 targets are used only during training.

3.4 Correspondence Sampling and Training

The descriptor fields retain their BEV coordinates and are processed by each framework's correspondence estimator, which produces candidate matches with nonnegative confidence weights after the proposed geometric and semantic supervision. This separation keeps the proposed representation-learning constraints independent of the implementation-specific form of correspondence estimation.

The selected correspondences provide paired metric BEV coordinates g_n and s_n, with nonnegative confidence weights c_n. Ground coordinates are relative to the camera, and satellite coordinates are relative to the reference image centre. The evaluated pipelines recover the planar pose by minimizing the weighted alignment error:

$$\hat{T} = \operatorname*{arg\,min}_{T\in \mathrm{SE}(2)} \sum_{n=1}^{N} c_n \|s_n - T(g_n)\|_2^2. \tag{7}$$

Here, $\mathrm{SE}(2)$ denotes planar rigid transformations consisting of two-dimensional translation and yaw rotation. This objective uses the rigid-alignment form of Procrustes analysis (Umeyama, 1991).

Training retains each pipeline's localization and matching objective $\mathcal{L}_{\mathrm{base}}$ and adds the proposed supervision:

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_g \mathcal{L}_{\mathrm{geom}} + \lambda_s \mathcal{L}_{\mathrm{semantic}} + \lambda_h \mathcal{L}_{\mathrm{HN}}. \tag{8}$$

The coefficients λ_g, λ_s, and λ_h balance geometric supervision, semantic supervision, and same-class hard-negative learning, respectively. The base localization and matching objective is retained for each framework, while the three additional terms provide the proposed supervision.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets and protocols. We evaluate on VIGOR (Zhu et al., 2021), KITTI-CVL (Shi et al., 2022), and DReSS-D (Xia et al., 2025b). VIGOR contains ground panoramas and aerial images from four US cities, with same-area and cross-area splits. The comparison reports unknown-orientation results for both cross-area and same-area splits. DReSS-D evaluates decentrality, where the ground camera need not lie at the centre of its satellite reference tile. Its same-area and cross-area unknown-orientation results are provided in Appendix A.1. KITTI-CVL is a limited-field-of-view satellite-ground localization benchmark built from KITTI driving imagery and satellite references. The complete KITTI-CVL protocol and results are provided in Appendix A.2.

Metrics. We report mean and median translation error in metres and orientation error in degrees. VIGOR additionally uses joint localization recall at $(1\,\mathrm{m}, 5^\circ)$, $(3\,\mathrm{m}, 10^\circ)$, and $(5\,\mathrm{m}, 20^\circ)$. KITTI-CVL reports orientation recall within 1° and 5°. Recall is expressed as a percentage; lower errors and higher recalls indicate better performance.

4.2 Implementation Details

All evaluated variants use the same GeoSem-BEV supervision while retaining the correspondence and pose-estimation components of their underlying frameworks. We refer to the GeoSem-BEV implementations of ViewBridge, FG^2, and DenseFlow as GS-ViewBridge, GS-FG^2, and GS-DenseFlow, respectively. Training and evaluation configurations are summarized in Appendix B.5.

The matching coefficient is $\beta = 100$ for unknown orientation and $\beta = 1$ for known orientation. Sem-FG^2 adds semantic supervision with uniform height aggregation, Geo-FG^2 adds geometric

Table 1: VIGOR unknown-orientation test results. Lower is better. GeoSem-BEV implementations are shown in black bold; best, second-best, and third-best distinct values within each area split are marked in red bold, blue bold, and black bold, respectively. Ties share the same rank. FG^2 reports original single-stage results; published baselines follow the cited comparisons. For each GeoSem-BEV implementation, percentages in parentheses after every metric report the relative change from its corresponding base model, computed as $(E_{\mathrm{GS}} - E_{\mathrm{base}})/E_{\mathrm{base}} \times 100\%$.

	Cross-area				Same-area			
	Localization (m)↓		Orientation (°)↓		Localization (m)↓		Orientation (°)↓	
Method	Mean	Median	Mean	Median	Mean	Median	Mean	Median
SliceMatch	7.220	3.310	25.970	4.510	6.490	3.130	25.460	4.710
CCVPE	5.410	**1.890**	27.780	13.580	**3.740**	**1.420**	12.830	6.620
DenseFlow	7.670	3.670	17.630	2.940	4.970	1.900	11.200	1.590
GS-DenseFlow	5.647 (-26.38%)	2.344 (-36.13%)	15.466 (-12.27%)	2.157 (-26.63%)	4.677 (-5.90%)	1.906 (+0.32%)	**9.340** (-16.61%)	1.589 (-0.06%)
FG^2	10.020	8.140	31.410	5.450	8.950	7.320	15.020	2.940
GS-FG^2	4.662 (-53.47%)	2.364 (-70.96%)	13.850 (-55.91%)	**1.241** (-77.23%)	3.823 (-57.28%)	1.849 (-74.74%)	9.723 (-35.27%)	**0.988** (-66.39%)
Loc^2	**4.230**	2.090	**11.670**	2.210	3.940	**1.780**	9.540	2.000
ViewBridge	**3.790**	**2.010**	**10.590**	**2.040**	**3.080**	**1.570**	**6.020**	**1.200**
GS-ViewBridge	**3.288** (-13.25%)	**2.041** (+1.54%)	**6.655** (-37.16%)	**0.994** (-51.27%)	**2.566** (-16.69%)	**1.570** (+0.00%)	**3.724** (-38.14%)	**0.784** (-34.67%)

supervision, and GS-FG^2 combines both with same-class hard negatives. The same geometric and semantic constraints are incorporated into DenseFlow and ViewBridge while retaining their original correspondence and pose-estimation components. Depth predictions and SAM3 masks are used only to construct training targets; inference uses only the ground and satellite images.

4.3 Quantitative Results

Across the reported datasets, GeoSem-BEV consistently improves the most challenging unknown-orientation evaluations. On VIGOR, all three implementations reduce both mean localization and orientation errors in the cross-area and same-area settings, with the largest overall gains obtained by GS-ViewBridge and GS-FG^2. On DReSS-D, GS-FG^2 reduces both errors in the reported same-area and cross-area settings, while GS-ViewBridge provides its clearest gains in orientation estimation (Appendix A.1). On KITTI-CVL, the limited-field-of-view evaluation shows comparable improvements in orientation and selected localization measures for the enhanced FG^2 and ViewBridge models, supporting the transfer of GeoSem-BEV beyond panoramic ground observations. Together, these results show that the proposed constraints improve different BEV matching frameworks, while the magnitude and type of the gain depend on the underlying representation and evaluation setting.

VIGOR. Table 1 compares the GeoSem-BEV implementations of DenseFlow, FG^2, and ViewBridge with SliceMatch (Lentsch et al., 2023), CCVPE (Xia et al., 2023), DenseFlow (Song et al., 2023), FG^2, Loc^2, and ViewBridge under unknown orientation. Published baselines follow the VIGOR comparisons reported in Loc^2 and ViewBridge (Xia et al., 2026; 2025a), while FG^2 uses its original single-stage results (Xia & Alahi, 2025) and the GeoSem-BEV implementations are evaluated on the official unknown-orientation test splits.

On VIGOR with unknown orientation, all three implementations reduce mean localization and orientation errors relative to their corresponding base models. GS-ViewBridge has the lowest mean localization and orientation errors in both area splits (Table 1). Relative to DenseFlow, GS-DenseFlow reduces cross-area mean translation error from 7.670 m to 5.647 m and mean orientation error from 17.630° to 15.466°, while GS-FG^2 reduces the corresponding errors from 10.020 m to 4.662 m and from 31.410° to 13.850° relative to FG^2. The percentages in parentheses in Table 1 report relative changes obtained by each GeoSem-BEV implementation over its base model.

4.4 Ablation Study

Table 2 compares geometry and semantic configurations on VIGOR cross-area with unknown orientation, using FG^{2†} as the baseline. Sem-FG^2 uses semantic supervision with uniform height aggregation, Geo-FG^2 uses geometric supervision without semantic supervision, and GS-FG^2 combines both. The ablation reveals an important interaction between the two constraints. Adding semantic supervision alone in Sem-FG^2 degrades all reported metrics relative to the reproduced baseline, indicating that semantic supervision alone cannot compensate for insufficiently constrained BEV feature placement. By contrast, Geo-FG^2 improves the baseline across the four pose-error metrics, showing that more accurate feature placement provides a stronger basis for correspondence learning. Once

Table 2: Ablation on VIGOR cross-area with unknown orientation. All results use RANSAC. Best values in each column are bold. FG^{2†} denotes our reproduced FG^2 baseline.

Method	Loc. mean ↓ (m)	Loc. median ↓ (m)	Ori. mean ↓ (°)	Ori. median ↓ (°)	R@1m/5° ↑	R@3m/10° ↑	R@5m/20° ↑
FG^{2†}	5.751	2.943	18.908	1.337	13.07	50.11	67.04
Sem-FG^2	6.339	3.335	22.740	1.397	11.06	45.25	62.76
Geo-FG^2	5.014	2.469	15.980	**1.240**	15.57	57.75	74.26
GS-FG^2	**4.662**	**2.364**	**13.850**	1.241	**16.53**	**59.81**	**76.72**

this geometric constraint is present, adding semantic supervision and same-class hard negatives in GS-FG^2 further reduces the mean translation and orientation errors and yields the highest recalls. These results suggest that semantic supervision is most effective after the BEV features have been placed at geometrically meaningful locations.

4.5 Discussion

GeoSem-BEV shows its clearest gains in the more challenging VIGOR unknown-orientation evaluations, where all three adapted models reduce both mean localization and orientation errors. The DReSS-D and KITTI-CVL results further show that the method transfers across datasets and limited-field-of-view observations, while the magnitude and type of improvement vary with the underlying framework and evaluation setting. These differences suggest that geometric ambiguity in BEV feature placement and descriptor matching ambiguity are most consequential when the baseline retains substantial correspondence uncertainty; when the baseline already forms effective BEV features, the added constraints leave less ambiguity to resolve. The ablation likewise shows that semantic supervision is most effective after geometric supervision has improved feature placement.

5 Conclusion

Satellite-ground localization estimates a ground-camera pose by matching ground-view and satellite-view observations across a large viewpoint difference. BEV-space feature matching provides a common coordinate system, but underconstrained radial placement and insufficient descriptor discrimination can leave localization ambiguity. We analyse these two sources of ambiguity and propose GeoSem-BEV to address them through complementary geometric and semantic supervision. Specifically, radial-depth and vertical-height supervision constrain where ground-view features contribute on the BEV grid, reducing geometric ambiguity in feature placement. Shared explicit semantic supervision promotes consistent semantic predictions across views, while same-class hard negatives improve descriptor discrimination and reduce descriptor matching ambiguity. By incorporating GeoSem-BEV into state-of-the-art BEV localization models, including FG^2, DenseFlow, and ViewBridge, the resulting models achieve their most consistent improvements under unknown orientation on VIGOR, supporting the effectiveness of reducing localization ambiguity in BEV-space feature matching. On VIGOR with unknown orientation, GeoSem-BEV reduces mean orientation error relative to the corresponding state-of-the-art method by 37.2% and 38.1% in the cross-area and same-area settings, respectively. On DReSS-D, the corresponding mean orientation errors are reduced by 10.8% and 15.6% relative to the reproduced FG^{2†} baseline. Taken together, the results show that GeoSem-BEV is most effective when the base model retains substantial geometric ambiguity in BEV feature placement or descriptor matching ambiguity. Future work will improve GeoSem-BEV for limited-field-of-view settings, adapt its constraints across matching frameworks, and develop more accurate methods for BEV feature placement and cross-view matching.

Ethics Statement

This work uses public cross-view localization benchmarks and does not collect new personal data or conduct experiments on human participants. Dataset licenses, image-access restrictions, and the intended use of the released benchmarks should be followed when distributing code or derived artifacts.

AI USE STATEMENT

Large language models and other AI tools were used to assist with writing and language polishing, retrieve and identify related literature, support research execution through code development and engineering workflows, and draft manuscript sections from author-provided methods and experimental records. The authors retain responsibility for the research decisions, experimental results, and final manuscript, including the accuracy and originality of the text. AI-assisted literature suggestions and technical content are subject to author verification against original sources, implementation code, and experimental records.

REPRODUCIBILITY STATEMENT

The experiments use Ascend NPUs. Section 4 records the evaluation protocols, metrics, and controlled training settings. Appendix B.6 specifies the model adaptations, while Appendix B.3 and Appendix B.4 detail pseudo-target construction and supervision.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 9–17, 2015.

Mingtao Feng, Fenghao Tian, Jianqiao Luo, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Saeed Mian. Semantic ambiguity modeling and propagation for fine-grained visual cross view geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2978–2986, 2025.

Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21621–21631. IEEE, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sixing Hu and Gim Hee Lee. Image-based geo-localization using satellite imagery. *International Journal of Computer Vision*, 128(5):1205–1219, 2020.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. Da

2
: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1477–1485, 2023.

Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):2020–2036, 2024.

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

Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5007–5015, 2015.

Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30, 2017.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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. Springer, 2020.

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, 2023a.

Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. Snap: Self-supervised neural maps for visual positioning and semantic understanding. *Advances in Neural Information Processing Systems*, 36:7697–7729, 2023b.

Qi Shan, Changchang Wu, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M Seitz. Accurate geo-registration by ground-to-aerial image matching. In *2014 2nd International Conference on 3D Vision*, volume 1, pp. 525–532. IEEE, 2014.

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, 2022.

Yujiao Shi, Xin Yu, Shan Wang, and Hongdong Li. Cvlnet: Cross-view semantic correspondence learning for video-based camera localization. In *Asian Conference on Computer Vision*, pp. 123–141. Springer, 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21459–21469. IEEE, 2023.

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

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 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.

Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pp. 494–509. Springer, 2016.

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

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE international conference on computer vision*, pp. 3961–3969, 2015.

Panwang Xia, Qiong Wu, Lei Yu, Yi Liu, Mingtao Xiong, Xudong Lu, Haoyu Guo, Yongxiang Yao, Junjian Zhang, Xiangyuan Cai, et al. Viewbridge: Revisiting cross-view localization from image matching. *arXiv preprint arXiv:2508.10716*, 2025a.

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

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

Zimin Xia, Olaf Booij, and Julian FP Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3813–3831, 2023.

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*, volume 2026, pp. 105738–105764, 2026.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024.

Junyan Ye, Qiyan Luo, Jinhua Yu, Huaping Zhong, Zhimeng Zheng, Conghui He, and Weijia Li. Sg-bev: Satellite-guided bev fusion for cross-view semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27748–27757. IEEE, 2024.

Kun Yu, Xiao Zheng, Bin Fang, Pei An, Xiao Huang, Wei Luo, Junfeng Ding, Zhao Wang, and Jie Ma. Multimodal urban remote sensing image registration via roadcross triangular feature. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4441–4451, 2021.

Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pp. 1395–1403, 2020.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5316–5325. IEEE, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1152–1161. IEEE, 2022.

APPENDIX

Here we provide supplementary material to support the main paper:

- **A.** Additional Experiments.
  - **A.1.** DReSS-D Results.
  - **A.2.** KITTI-CVL Results.
  - **A.3.** Qualitative Results.
  - **A.4.** GT-Aligned BEV Correspondence Localization.
- **B.** Implementation Details and Pipeline Adaptations.
  - **B.1.** Shared Implementation Details.
  - **B.2.** Coordinate and Sampling Conventions.
  - **B.3.** Geometry Supervision Details.
  - **B.4.** Semantic Supervision Details.
  - **B.5.** Training and Evaluation Configuration.
  - **B.6.** Pipeline Adaptations.
    - **B.6.1.** GS-FG^2.
    - **B.6.2.** GS-ViewBridge.
    - **B.6.3.** GS-DenseFlow.

A ADDITIONAL EXPERIMENTS

A.1 DRESS-D RESULTS

Table 3 reports the DReSS-D unknown-orientation regimes. Published CCVPE and ViewBridge values follow the DReSS-D comparison in ViewBridge (Xia et al., 2025a), which reports this benchmark without RANSAC refinement. We therefore evaluate GS-ViewBridge, the reproduced FG^{2†} baseline, and GS-FG^2 without RANSAC refinement on the official DReSS-D test splits, so that every entry in the table is directly comparable.

Table 3: DReSS-D test results under unknown orientation. All results are reported without RANSAC refinement, following the published DReSS-D comparison. CCVPE and ViewBridge values follow that comparison, and FG^{2†} denotes our re-implementation. Best, second-best, and third-best distinct values within each metric are marked in red bold, blue bold, and black bold, respectively; ties share the same rank. Percentages in parentheses after every metric report the relative change from the corresponding base model, computed as $(E_{\mathrm{GS}} - E_{\mathrm{base}})/E_{\mathrm{base}} \times 100\%$.

Method	Cross-area				Same-area			
	Localization (m)↓		Orientation (°)↓		Localization (m)↓		Orientation (°)↓	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
CCVPE	6.050	**2.230**	37.390	10.270	3.010	**1.020**	14.440	7.940
FG^{2†}	6.422	3.814	16.466	**2.411**	5.508	2.850	14.124	2.129
GS-FG^2	**5.926** (-7.72%)	3.396 (-10.96%)	**14.696** (-10.75%)	2.234 (-7.34%)	**4.886** (-11.29%)	2.475 (-13.16%)	**11.921** (-15.60%)	**1.939** (-8.92%)
ViewBridge	**4.220**	**2.370**	**12.580**	**2.600**	**2.940**	**1.490**	**8.470**	**1.620**
GS-ViewBridge	**4.108** (-2.65%)	**2.468** (+4.14%)	**8.810** (-29.97%)	**1.976** (-24.00%)	**3.186** (+8.37%)	**1.707** (+14.56%)	**6.854** (-19.08%)	**1.668** (+2.96%)

The DReSS-D results show that GeoSem-BEV's effect depends on the underlying framework. GS-FG^2 reduces both mean localization and orientation errors in every reported same-area and cross-area setting. Under unknown orientation, its localization errors decrease by 7.7% and 11.3%, and its orientation errors by 10.8% and 15.6%, on cross-area and same-area, respectively. GS-ViewBridge reduces mean orientation error under unknown orientation by 30.0% on cross-area and 19.1% on same-area, and slightly reduces cross-area localization error, while its same-area localization error increases. These results indicate that the constraints can reduce localization ambiguity, while the magnitude and type of improvement vary with the difficulty of the evaluated samples and the matching properties of the base framework.

We next evaluate KITTI-CVL to examine whether these gains transfer from panoramic ground observations to limited-field-of-view imagery.

Table 4: KITTI-CVL test results. Orientation noise is sampled uniformly within $\pm 10^\circ$ during training and testing. Lateral and longitudinal recalls are omitted. Best, second-best, and third-best distinct values within each area and orientation setting are marked in red bold, blue bold, and black bold, respectively; ties share the same rank. For GS-FG^2 and GS-ViewBridge, percentages in parentheses after every metric report relative changes from the corresponding base model. A dash denotes an unreported source metric.

Area	Ori.	Method	Loc. (m)↓		Ori. (°)↓		Ori. (%)↑	
			Mean	Median	Mean	Median	R@1°	R@5°
Cross-area	±10°	GGCVT	–	–	–	–	**98.98**	**100.00**
		CCVPE	9.160	**3.330**	**1.550**	**0.840**	**57.72**	**96.19**
		HC-Net	8.470	4.570	**3.220**	**1.630**	33.58	83.78
		FG^2	**7.310**	4.150	3.620	2.370	23.03	77.84
		GS-FG^2	**7.026** (-3.89%)	4.372 (+5.35%)	3.485 (-3.73%)	2.003 (-15.49%)	28.07 (+21.88%)	80.48 (+3.39%)
		Loc^2	**5.600**	**3.010**	3.320	2.120	26.03	80.68
		ViewBridge	6.700	3.270	3.150	1.990	28.62	81.63
		GS-ViewBridge	**5.880** (-12.24%)	**2.686** (-17.86%)	3.635 (+15.40%)	1.732 (-12.96%)	30.64 (+7.06%)	**85.60** (+4.86%)
Same-area	±10°	GGCVT	–	–	–	–	**99.10**	**100.00**
		CCVPE	1.220	0.620	**0.670**	**0.540**	77.39	**99.95**
		HC-Net	**0.800**	**0.500**	**0.450**	**0.330**	**91.35**	99.84
		FG^2	**0.750**	0.510	0.930	0.660	67.27	98.91
		GS-FG^2	0.830 (+10.67%)	0.582 (+14.12%)	0.795 (-14.52%)	0.590 (-10.61%)	72.65 (+7.99%)	99.55 (+0.65%)
		Loc^2	1.130	0.770	1.970	1.430	36.68	92.84
		ViewBridge	**0.680**	**0.460**	1.100	0.720	61.77	97.97
		GS-ViewBridge	**0.796** (+17.06%)	0.539 (+17.17%)	0.805 (-26.82%)	0.566 (-21.39%)	74.08 (+19.93%)	**99.98** (+2.05%)

A.2 KITTI-CVL Results

The KITTI-CVL benchmark evaluates satellite-ground localization from limited-field-of-view ground observations and satellite references (Shi et al., 2022). We include it to assess whether GeoSem-BEV transfers beyond panoramic ground views while retaining competitive localization and orientation performance. Table 4 reports the KITTI-CVL comparison, with published baseline values collected in Loc^2 (Xia et al., 2026). The GS-FG^2 cross-area and same-area $\pm 10^\circ$ results and the GS-ViewBridge results are evaluated on the official test split without RANSAC.

KITTI-CVL tests whether GeoSem-BEV transfers to limited-field-of-view ground images. In cross-area evaluation, GS-FG^2 reduces both mean localization and orientation errors, while GS-ViewBridge reduces mean localization error. In same-area evaluation, both implementations reduce mean orientation error and improve orientation recall, but their mean localization errors increase. These results support the applicability of GeoSem-BEV to limited-field-of-view scenes and suggest that its constraints can reduce descriptor matching ambiguity, particularly for orientation estimation. However, improved orientation estimates do not guarantee lower localization error, and the localization effect varies with the underlying framework and area split.

A.3 Qualitative Results

Figure 3 provides qualitative examples of the two effects targeted by our supervision. The examples are selected from the VIGOR cross-area unknown-orientation evaluation cases used for the diagnostic visualizations. In the feature-placement examples, adding radial-depth and vertical-height supervision produces projected ground-view content that is more spatially concentrated and better aligned with the corresponding structures in the satellite reference than the unconstrained baseline. In the descriptor examples, explicit semantic supervision concentrates the matching response around the ground-truth location and suppresses responses from visually similar regions at different locations. These observations illustrate how the geometric and semantic constraints address the two forms of localization ambiguity discussed in the main paper.

Figure 4 further compares the correspondence patterns produced by FG^2 and GS-FG^2 on challenging VIGOR samples. In cases where FG^2 produces dispersed or spatially inconsistent correspondences, GS-FG^2 yields more coherent matches and pose estimates closer to the ground truth. These examples illustrate that incorporating GeoSem-BEV can improve correspondence matching on samples that remain difficult for the original model.

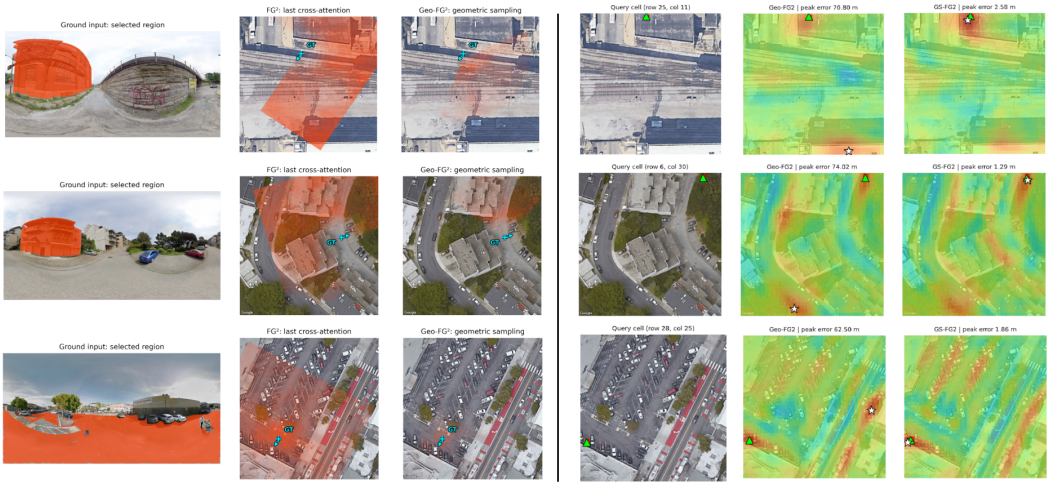
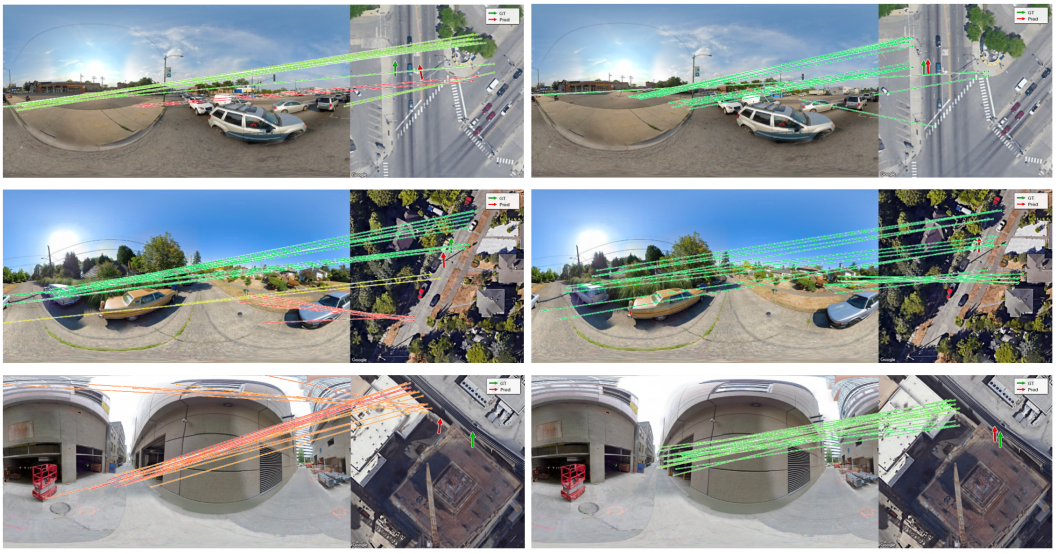


Figure 3: Qualitative visualization of geometry-constrained feature placement and explicitly supervised descriptor learning. In the left block, the orange overlay is the RGB content sampled from the selected ground-image region and projected onto the satellite images; the two columns of satellite images show the projection produced by FG^{2†} and Geo-FG^2, respectively. The cyan $+$ and arrow mark the ground-truth camera position and orientation used for alignment. In the right block, each heatmap shows the normalized descriptor-matching response for one ground-view BEV query over satellite locations: warmer colors indicate higher matching scores. The green triangle marks the ground-truth satellite location corresponding to the query cell, and the white star marks the highest-scoring predicted location; the reported peak error is the distance between these two markers.



(a) FG^2 (b) GS-FG^2(ours)

Figure 4: Qualitative correspondence comparison on challenging VIGOR samples. Each row shows the same ground–satellite image pair for FG^2 (left) and GS-FG^2 (right). Lines visualize selected correspondences; green and red arrows indicate the ground-truth and predicted camera poses, respectively.

A.4 GT-Aligned BEV Correspondence Localization

To directly evaluate whether radial geometry improves the spatial placement of ground-view BEV features, we perform a post-hoc correspondence-localization diagnostic on a fixed 5,000-sample

subset of the VIGOR cross-area validation split with unknown orientation. For each ground-view BEV cell whose ground-truth transformed position lies inside the satellite-view BEV, the ground-truth pose defines its target satellite location. The predicted location is the satellite cell with the highest descriptor-matching score. We report the Euclidean error between these locations and the proportion of valid correspondences within 1, 3, and 5 metres. FG^{2†} and Geo-FG^2 use the same samples, descriptor matcher, and evaluation procedure to assess the geometry-constrained variant.

Table 5: Ground-truth-aligned BEV correspondence localization on 5,000 VIGOR cross-area validation samples with unknown orientation. Correspondence errors are measured from the highest-scoring satellite-view BEV cell to the aligned location. Recalls are percentages. Best values are in bold.

Method	Valid-cell ratio	Corr. error mean ↓ (m)	Corr. error median ↓ (m)	Corr.@1m ↑	Corr.@3m ↑	Corr.@5m ↑
FG^{2†}	0.726	9.095	3.662	9.29	41.20	60.40
Geo-FG^2	0.726	**8.427**	**3.244**	**10.64**	**45.52**	**64.86**

Geo-FG^2 reduces the mean correspondence error from 9.095 m to 8.427 m and the median error from 3.662 m to 3.244 m. It also increases Corr.@1m, Corr.@3m, and Corr.@5m across all three thresholds. A paired sample-level bootstrap gives a mean-error difference of -0.669 m with a 95% confidence interval of $[-0.796, -0.539]$ m. These results support that radial geometry places ground-view features closer to their ground-truth-aligned BEV correspondence locations.

B Implementation Details and Pipeline Adaptations

B.1 Shared Implementation Details

The GS variants share radial geometric supervision, explicit semantic descriptor supervision, and same-class hard negatives. Their feature encoders, aggregation normalization, and correspondence estimators differ. The shared training and evaluation settings are summarized in Appendix B.5; implementation-specific departures are described below.

B.2 Coordinate and Sampling Conventions

The ground-view query grid is indexed by two horizontal coordinates and one height coordinate. A BEV grid contains horizontal cells; each cell has a column of 3D queries at different heights. Satellite pixel offsets are converted to metres using the ground sampling distance, accounting for image resizing and feature-grid resolution. In GS-FG^2, each horizontal axis spans $[-35.5, 35.5]$ m with 41 samples, and the 11 heights span $[-10, 10]$ m relative to the camera. A horizontal metric coordinate x maps to grid coordinate $(x + 35.5)40/71$. Equirectangular image sampling wraps horizontally. Satellite-view sampling uses bilinear interpolation after resizing the satellite feature map to the 41×41 BEV resolution. The main-text pose convention maps ground coordinates to satellite coordinates; the GS-FG^2 implementation internally uses row–column coordinates and the inverse satellite-to-ground transform. Coordinate ordering and transform direction are converted consistently at the interface.

B.3 Geometry Supervision Details

The following target construction and height loss specify GS-FG^2; the separately supervised vertical distributions used by GS-ViewBridge and GS-DenseFlow are identified in Appendices B.6.2 and B.6.3. The frozen ground-view SphereViT teacher from DA^2 (Li et al., 2025) predicts relative radial depth. Its scale is set by dividing the assumed camera height of 2.5 m by the 95th percentile of downward vertical distances in the lower image region. The resulting metric pseudo-depth supervises surface presence, the 64-bin radial distribution over 0–35 m, expected radial depth, and uncertainty. The teacher uses valid downward rays below 0.55 of the image height, extending to all valid downward rays when fewer than 32 candidates remain.

Depth Anything V2 predictions for the satellite view (Yang et al., 2024) are min–max normalized and bilinearly resized to 41×41. A normalized value $h \in [0, 1]$ maps to continuous height-bin index

$10h$, from which we construct a normalized Gaussian soft target h^S_{ijk} with standard deviation 0.75 bins. This is a relative-height pseudo-target over the query bins, not a measured metric elevation. The ground-truth pose T^* aligns this target to the ground-view BEV grid by bilinear sampling, yielding $\hat{h}^G_{ijk}$. For valid overlapping cells Ω, the height loss is normalized by $\log 11$, the entropy of a uniform distribution over the 11 height candidates.

$$\mathcal{L}_{\text{height}} = -\frac{1}{|\Omega| \log 11} \sum_{(i,j)\in\Omega} \sum_{k=1}^{11} \hat{h}^G_{ijk} \log(\max(w_{ijk}, \epsilon)), \qquad \epsilon = 10^{-8}. \tag{9}$$

The target index k corresponds to the same 11 height candidates used in Equation 3. Its gradient therefore updates the surface and radial predictions that generate w_{ijk}.

B.4 Semantic Supervision Details

For view $v \in \{G, S\}$, Y_v and P_v denote the soft semantic targets and predicted class probabilities. Let $\mathcal{V}_v$ contain the valid BEV cells and $\mathcal{C}_v$ the classes with nonzero target mass. With cell index n and class index c, the macro-balanced soft cross-entropy is

$$\mathcal{L}^v_{\text{CE}} = \frac{1}{|\mathcal{C}_v|} \sum_{c\in\mathcal{C}_v} \frac{-\sum_{n\in\mathcal{V}_v} Y_{v,nc} \log P_{v,nc}}{\sum_{n\in\mathcal{V}_v} Y_{v,nc}}, \qquad v \in \{G, S\}. \tag{10}$$

The ground-truth pose T^* bilinearly warps P_G to the satellite grid, producing $P_G^{T^*}$ over the valid overlap $\mathcal{V}_{T^*}$. With $M_n = (P^{T^*}_{G,n} + P_{S,n})/2$, the consistency term is

$$\mathcal{L}_{\text{JS}}(P_G^{T^*}, P_S) = \frac{1}{|\mathcal{V}_{T^*}|} \sum_{n\in\mathcal{V}_{T^*}} \frac{1}{2} \left[D_{\text{KL}}(P^{T^*}_{G,n} \| M_n) + D_{\text{KL}}(P_{S,n} \| M_n) \right]. \tag{11}$$

For direction $X \to Y$, with $(X, Y) \in \{(G, S), (S, G)\}$, let $\mathcal{A}_{X\to Y}$ contain the valid anchors. The positive logit s^+_a is defined by the bilinear soft positive, and $s^-_{a,n}$ denotes the matching logit of negative candidate n. The set $\mathcal{N}_a$ contains the top 32 same-class candidates outside the one-cell neighbourhood. The anchor weight α_a is the product of DA2-derived visibility and semantic confidence. The directional loss is

$$\mathcal{L}^{\text{HN}}_{X\to Y} = \frac{\sum_{a\in\mathcal{A}_{X\to Y}} \alpha_a \left[\log\left(e^{s^+_a} + \sum_{n\in\mathcal{N}_a} e^{s^-_{a,n}} \right) - s^+_a \right]}{\sum_{a\in\mathcal{A}_{X\to Y}} \alpha_a}, \qquad X, Y \in \{G, S\}. \tag{12}$$

Implementation-specific candidate confidence thresholds and back-end logits are given in Appendices B.6.1–B.6.3.

B.5 Training and Evaluation Configuration

The controlled GS-FG^2 experiments use frozen DINOv2 features (Oquab et al., 2023) with 1024 channels, a $41 \times 41 \times 11$ query grid, 64 radial depth bins, and 128-dimensional descriptors. The pose estimator samples 1024 correspondences. Training uses Adam (Kingma & Ba, 2014) with learning rate 10^{-4} and zero weight decay on Ascend NPUs. Feature computation uses BF16; matching probabilities, losses, and pose solving use FP32. The main VIGOR cross-area unknown-orientation experiment uses 25 epochs, seed 0, global batch size 24, and random rotations within $[-180°, 180°]$. Evaluation uses the official test split of 53,694 samples. Depth predictions and SAM3 masks are precomputed or cached, and ground-truth poses align cross-view training targets. GS-FG^2 and its ablations use the same RANSAC protocol unless stated otherwise.

B.6 Pipeline Adaptations

The following descriptions specify the evaluated implementations rather than asserting identical architectures across back ends. We first describe GS-FG^2, the baseline used for the detailed supervision analysis, followed by GS-ViewBridge and GS-DenseFlow.

B.6.1 GS-FG^2

GS-FG^2 retains the FG^2 descriptor projection, dual-softmax matching with a dustbin, and correspondence-based pose solver (Xia & Alahi, 2025). Frozen DINOv2 features are sampled once on a $41 \times 41 \times 11$ grid spanning $[-35.5, 35.5]$ m horizontally and $[-10, 10]$ m vertically. Unlike GS-ViewBridge, this implementation has no independent vertical head: $w_{ijk} = \ell_{ijk}$. Satellite height targets directly supervise these likelihood weights. The aggregated appearance features pass through six planar self-attention/FFN blocks and a 128-dimensional projector. No radial-mass gate is applied after these blocks or to the matching coupling.

The shared semantic head supervises the final normalized descriptors, and same-class hard negatives act on the descriptor matching logits. The base objective is $\mathcal{L}_{\mathrm{VCE}} + \beta \mathcal{L}_{\mathrm{InfoNCE}}$, with unit coefficients on the three added objectives. The VIGOR unknown-orientation configuration uses $\beta = 100$. The RANSAC and direct Weighted Procrustes evaluations are reported separately.

B.6.2 GS-VIEWBRIDGE

GS-ViewBridge retains the frozen DINOv2 encoder, satellite projector, planar context modules, similarity refinement, adaptive dustbin, and soft-coupling back end of the ViewBridge implementation (Xia et al., 2025a). The ground branch replaces learned panorama cross-attention with analytic sampling at a $41 \times 41 \times 11$ metric grid. Ground block11 and block23 features feed the radial head; block23 provides the sampled appearance. The radial head predicts a surface probability, a 64-bin distribution over 0–35 m, and uncertainty.

A separate vertical head predicts v_{ijk}, normalized across the 11 heights of each BEV column. With the radial likelihood ℓ_{ijk} from Equation 2, the aggregation weight is $w_{ijk} = v_{ijk}\ell_{ijk}$. The joint weights are not renormalized over height. Their sum $m_{ij} = \sum_k w_{ijk}$ also gates the ground features before and after the six planar context blocks. Satellite height supervision acts on v_{ijk}. The resulting 128-dimensional descriptors enter ViewBridge's similarity refinement and adaptive dustbin. Ground mass additionally weights correspondence sampling and alignment. The model samples 1024 pairs for Weighted Procrustes; RANSAC is enabled only for the corresponding evaluation protocol.

The shared semantic head is LayerNorm–Linear$(128, 64)$–GELU–Linear$(64, 3)$. The hard-negative objective reads the refined matching logits, uses bilinear soft positives, excludes a Chebyshev radius of one cell, and selects 32 same-class candidates with class confidence at least 0.5. The base objective combines the pose loss and the existing InfoNCE loss; the geometry, semantic, and hard-negative terms have unit coefficients. On VIGOR and DReSS-D, the pose-loss coefficient is 1, and the InfoNCE coefficient is 100 for unknown orientation and 1 for known orientation.

B.6.3 GS-DENSEFLOW

GS-DenseFlow constructs geometry-constrained ground BEV representations using ResNet18 features (He et al., 2016) and retains the RAFT-based flow estimator (Teed & Deng, 2020) used by DenseFlow (Song et al., 2023). The adapted builder samples ground features at 3D queries, predicts radial and vertical distributions, and forms their product $j_{ijk} = \ell_{ijk} v_{ijk}$. It normalizes the appearance aggregation by the column mass: $w_{ijk} = j_{ijk} / \max(\sum_k j_{ijk}, \epsilon)$, with $\epsilon = 10^{-6}$. The unnormalized mass is retained separately as a geometric confidence signal. Thus, this implementation uses the same geometric assignment principle but a different aggregation normalization from GS-ViewBridge and GS-FG^2.

Semantic supervision is applied to the actual ground and satellite feature maps used by RAFT to construct correlations. Its shared head maps 256-dimensional descriptors through a 64-dimensional hidden layer to three classes. Same-class hard negatives are selected from the actual level-zero correlation field. The flow estimator and confidence prediction remain in the pipeline, and geometric mass weights the confidence used for pose alignment. The original DenseFlow training objective is retained, with geometry, semantic, and hard-negative terms added at unit weight.