CoVisFormer: Co-Visibility-Aware Correspondence Learning for Cross-View Geo-Localization

Anonymous authors
Paper under double-blind review

Abstract

Estimating 3-DoF pose between a ground image and an aerial image is vital for a variety of computer vision applications. However, achieving robust cross-view geo-localization remains challenging because the limited cross-view interaction results in erroneous matches in non-overlapping regions, and fixed rules for inlier selection limit the effective rejection of outliers. To deal with these issues, a novel cross-view geo-localization method is proposed by integrating cross-view interaction with dynamic inlier detection, including a co-visibility-aware cross-view contextual feature learning (C3FL) module and a dynamic inlier detection-based pose regression (DIDP) module. The proposed CoVisFormer enjoys several merits. First, the proposed C3FL module can leverage cross-view interaction to focus matching on overlapping regions. Second, the DIDP module can adaptively update according to the current input, enabling effective outlier rejection. Extensive experimental results on two challenging benchmarks show that our proposed method significantly outperforms state-of-the-art cross-view geo-localization methods.

1 Introduction

Cross-view geo-localization between a ground-level query image and an aerial image of its surrounding area is a fundamental task with broad applications in visual localization, autonomous driving, and outdoor mobile robotics (Shi & Li, 2022; Fervers et al., 2023b; Park et al., 2026). Because of its substantial practical value, this task has attracted increasing attention from both academia and industry over the past decades. However, robust and accurate cross-view geo-localization remains challenging due to the severe viewpoint gap, scale differences, and appearance variations between ground and aerial imagery.

To address the above challenges, numerous cross-view geo-localization methods have been proposed. Specifically, global descriptor-based methods (Lentsch et al., 2023; Xia et al., 2023; Park et al., 2026) typically encode each sampled candidate location in the aerial image into a global descriptor and match it against the global descriptor of the ground-level image to determine the camera location. For example, CCVPE (Xia et al., 2023) constructs orientation-aware global descriptors and jointly predicts position and orientation in a coarse-to-fine manner. However, due to the highly abstract nature of global descriptors, these methods have difficulty in explicitly identifying which objects or regions in the images actually contribute to the localization decision, and their localization accuracy is also limited by the sampling density. To achieve lower localization errors, local feature-based methods (Fervers et al., 2023b;a; Shi et al., 2023a; Song et al., 2023; Wang et al., 2023c; Xia & Alahi, 2025; Xia et al., 2026) have been introduced. These methods typically obtain bird's-eye-view (BEV) representations from ground-view images via homography transformation or Transformer-based architectures, and achieve more accurate localization through dense correlation (Fervers et al., 2023b;a; Shi et al., 2023a) or pixel-level matching (Wang et al., 2023c; Song et al., 2023; Wang et al., 2023b; 2024; Barroso-Laguna et al., 2024; Xia & Alahi, 2025; Xia et al., 2026). Taking the current state-of-the-art method FG^2 (Xia & Alahi, 2025) as an example, it employs a Transformer to construct the BEV representation of the ground-level image and performs pixel-level matching with the aerial representation to establish correspondences, followed by post-

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.

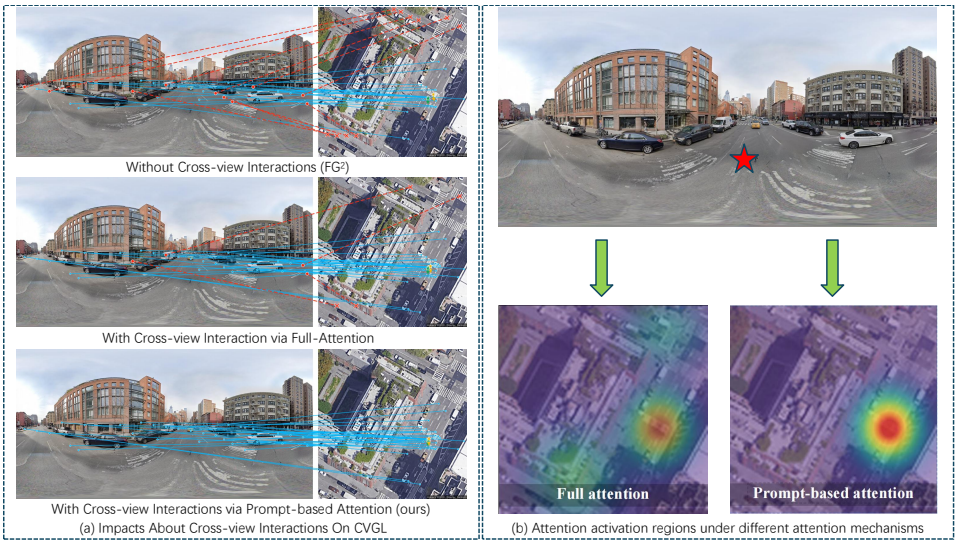

Figure 1: Illustration of our motivation. (a) shows the quality of correspondences under different cross-view interaction strategies, including no interactions (Xia & Alahi, 2025), full attention, and our prompt-based attention. Blue lines indicate geometrically correct matches; red dashed lines with crosses denote incorrect matches. (b) shows the comparison between our proposed prompt-based attention and full attention, a selected ground region (star) and aerial attention responses, where full attention would aggregate features from irrelevant areas.

processing algorithms such as Procrustes–RANSAC (Schönemann, 1966; Fischler & Bolles, 1981) to estimate the camera pose.

Despite the remarkable progress achieved by these methods, they still tend to produce a large number of erroneous matches in non-overlapping regions between aerial and ground views, and their pose solvers rely on fixed rules for inlier selection. To make cross-view geo-localization more robust to real-world challenges and better suited to practical applications, the following two issues need to be carefully addressed. **(1) How to avoid a large number of erroneous matches in non-overlapping regions?** In cross-view geo-localization, aerial and ground-level images often contain extensive non-overlapping areas. Existing methods, such as FG^2 (Xia & Alahi, 2025) and LoC^2 (Xia et al., 2026), typically extract features from the two views independently. Due to the lack of effective information interaction between the two branches, the model has limited awareness of co-visible regions, which leads to numerous incorrect matches in non-overlapping areas, as illustrated by the no-interaction case in the first row of Fig. 1(a). Although attention mechanisms can capture co-visible information by modeling cross-view interactions, full attention tends to introduce substantial noise when aggregating contextual information, and may still produce erroneous matches in non-overlapping regions, as illustrated by the full-attention case in the second row of Fig. 1(a) and its response map in the left panel of Fig. 1(b). Therefore, an effective attention mechanism needs to be proposed to enable cross-view information interaction, enhance awareness of co-visible regions, and reduce erroneous matches in non-overlapping areas, as illustrated by the prompt-based case in the third row of Fig. 1(a). **(2) How to achieve adaptive inlier detection for pose regression?** In real-world scenarios with variations in illumination, viewpoint, and scale, the matched correspondences in cross-view geo-localization often exhibit diverse inlier patterns. Existing methods, such as FG^2 (Xia & Alahi, 2025) and LoC^2 (Xia et al., 2026), typically employ RANSAC (Fischler & Bolles, 1981) to identify inliers for pose regression. However, RANSAC relies on predefined rules, such as fixed error thresholds and inlier criteria, and therefore lacks the flexibility to adapt to different inlier patterns. Therefore, it is necessary to design a detector capable of dynamically identifying inliers to enable accurate pose regression.

Motivated by the above observations, we propose **CoVisFormer**, a co-visibility-aware correspondence learning framework for cross-view geo-localization. It consists of a co-visibility-aware cross-view contextual feature learning (C3FL) module and a dynamic inlier detection-based pose regression (DIDP) module. In the C3FL module, we introduce two sets of learnable prompts as bridges

for cross-view information interaction. Specifically, the prompts interact with features from the two views through attention mechanisms to effectively aggregate intra-view contextual information. The updated prompts from the two views are then concatenated and further processed by self-attention to capture long-range cross-view dependencies. Subsequently, the refined prompts are used to enhance the feature representations of both views through similarity-based aggregation. Finally, correspondences are established from the enhanced features using dual-softmax matching and mutual nearest-neighbor filtering. In the DIDP module, we aim to adaptively capture diverse inlier patterns among the established correspondences. Specifically, we integrate the geometric locations and matching confidence of each correspondence to construct its representation. We then introduce a set of detector prototypes, which interact with these correspondence representations through attention to capture the inlier patterns of the current input and generate dynamic inlier detectors. These detectors are subsequently used to identify reliable inliers, from which two MLP heads separately regress the camera position and orientation.

Our main contributions are summarized as follows: (1) We propose **CoVisFormer**, a co-visibility-aware correspondence learning framework for cross-view geo-localization. It introduces a co-visibility-aware cross-view contextual feature learning module that leverages learnable prompts to facilitate cross-view context interaction and enhance correspondence learning, together with a dynamic inlier detection-based pose regression module that adaptively identifies reliable inliers from matched correspondences and accurately regresses the camera pose. (2) We conduct extensive experiments on the VIGOR and KITTI datasets. The results demonstrate that our method consistently outperforms existing state-of-the-art approaches in terms of mean localization error across both same-area and cross-area settings, under both known- and unknown-orientation conditions.

2 Related Work

In this section, we briefly review **global descriptor-based** and **local feature-based** cross-view geo-localization methods.

Global descriptor-based methods (Lentsch et al., 2023; Xia et al., 2023; Park et al., 2026) typically sample candidate locations in the aerial image and then generate a global descriptor for each candidate location, which is used to find matches with the global descriptor of the ground image to estimate the camera pose. For example, SliceMatch (Lentsch et al., 2023) extracts aerial descriptors by projecting ground slices onto the aerial feature map and aggregating features from the corresponding regions. Despite their effectiveness, the extreme visual differences between ground and aerial images still limit the cross-view consistency of global descriptors. To alleviate this inconsistency, VIRD (Park et al., 2026) jointly addresses horizontal and vertical misalignment through polar transformation and context-enhanced positional attention, respectively. Nevertheless, these methods are inherently constrained by a trade-off between localization accuracy and computational efficiency. Moreover, their reliance on global descriptors often overlooks salient local regions that provide important cues for precise localization. Therefore, local feature-based methods have received growing attention in recent years.

Local feature-based methods (Fervers et al., 2023b;a; Shi et al., 2023a; Song et al., 2023; Wang et al., 2023c; Xia & Alahi, 2025; Xia et al., 2026; Kim et al., 2026) commonly transform ground features into a BEV representation through homography warping or attention mechanisms. Subsequently, localization can be performed through either dense correlations or explicit pixel-level correspondences. Dense correlation-based methods generally estimate camera poses by computing correlations between ground BEV features and aerial features. For example, Uncertainty-Aware CVGL (Fervers et al., 2023b) constructs a ground BEV representation using self-attention and cross-attention, and then performs correlation with aerial features for camera pose estimation. Although dense correlation methods have achieved promising results, occluded regions may interfere with pixel-wise alignment between the two BEV representations. Therefore, pixel-level correspondence-based methods have been proposed, which estimate camera pose by explicitly establishing dense (Wang et al., 2023c; Song et al., 2023; Wang et al., 2023b; 2024) or sparse (Barroso-Laguna et al., 2024; Xia & Alahi, 2025; Xia et al., 2026) correspondences within the regions that are truly matched between the ground and aerial images. For instance, HC-Net (Wang et al., 2023c) uses a spherical transformation to reduce the viewpoint gap and constructs a 4D correlation volume by computing all-pairs feature similarities between the transformed ground and aerial features. It then samples

local correlation neighborhoods according to the current homography and iteratively refines the homography for camera pose estimation. Differently, FG^2 (Xia & Alahi, 2025) instead constructs a ground BEV representation via attention mechanisms, establishes local correspondences with aerial features based on pairwise descriptor similarity, and recovers the 3-DoF pose through Procrustes alignment (Schönemann, 1966) with RANSAC (Fischler & Bolles, 1981). By exposing explicit local correspondences, pixel-level correspondence-based methods can achieve high localization accuracy and provide an intuitive explanation of the evidence used for pose estimation. Nevertheless, the large non-co-visible regions between ground and aerial images still induce many false correspondences. Moreover, subsequent pose estimation typically depends on a geometric solver with predefined rules. In contrast, our co-visibility-aware cross-view contextual feature learning module captures co-visible region information through interaction between the two views. Our dynamic inlier detection-based pose regression module also adapts to diverse inlier patterns and regresses the camera pose.

3 METHOD

In this section, we present our proposed method by a co-visibility-aware cross-view contextual feature learning module and a dynamic inlier detection-based pose regression module for cross-view geo-localization. The overall architecture is illustrated in Fig. 2.

3.1 OVERVIEW

As shown in Fig. 2, our proposed model mainly consists of a co-visibility-aware cross-view contextual feature learning (C3FL) module and a dynamic inlier detection-based pose regression (DIDP) module. Given a ground-level image and a geo-referenced aerial image, we first extract their original image features via a feature extractor inspired by FG^2 (Xia & Alahi, 2025). Next, we use an off-the-shelf monocular depth estimation model to predict a depth map for the ground-level image, which is used to transform the ground image features into a bird's-eye-view (BEV) representation. Then, we send the ground bev and aerial features into the proposed C3FL module to obtain cross-view contextual features descriptors. Specifically, we first define two sets of prompts in the C3FL module, which can separately interact with ground bev features and aerial features via an attention operation to obtain updated prompts. The updated prompts are concatenated and refined through a self-attention mechanism to exchange cross-view context. We then update the ground bev features and aerial features via their similarities with the updated concatenated prompts. After obtaining the cross-view contextual feature descriptors, we use dual-softmax and mutual nearest neighbor to establish ground-to-aerial correspondences. These correspondence sets are sent to the DIDP module to obtain the camera pose. First, we fuse the geometry and confidence of each correspondence to obtain the matching features. A set of prototypes are then defined and interact with these features via a cross attention mechanism, which produces adaptive inlier detectors. Next, we use these detectors to identify reliable inliers, from which two MLP heads separately regress the camera pose.

3.2 CO-VISIBILITY-AWARE CROSS-VIEW CONTEXTUAL FEATURE LEARNING

In order to capture cross-view long-range contexts effectively, we adopt a prompt-based attention mechanism in the proposed co-visibility-aware cross-view contextual feature learning (C3FL) module. Given ground BEV features $\mathbf{F}_G \in \mathbb{R}^{HW\times C}$ and aerial features $\mathbf{F}_A \in \mathbb{R}^{HW\times C}$, we first design two sets of prompts $\mathbf{Q}_G \in \mathbb{R}^{N\times C}$ and $\mathbf{Q}_A \in \mathbb{R}^{N\times C}$ to separately interact with $\mathbf{F}_G$ and $\mathbf{F}_A$ via the attention operation, where prompts are initialized with a set of learnable parameters. Specifically, for a view $v \in \{G, A\}$, keys and values arise from image features $\mathbf{F}_v$, and queries arise from the prompts $\mathbf{Q}_v$. Formally,

$$\mathbf{Q} = \mathbf{Q}_v\mathbf{W}_v^Q, \qquad \mathbf{K} = \mathbf{F}_v\mathbf{W}_v^K, \qquad \mathbf{V} = \mathbf{F}_v\mathbf{W}_v^V. \tag{1}$$

$$\mathbf{Q}_v^* = \mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}. \tag{2}$$

where $\mathbf{W}_v^Q, \mathbf{W}_v^K, \mathbf{W}_v^V \in \mathbb{R}^{C\times C}$ are linear projections. We use multi-head attention with H_P heads, where the per-head dimension is $d = C/H_P$. The scaling factor d in $\sqrt{d}$ denotes the per-head query/key dimension. The outputs of all heads are concatenated and projected back to $\mathbb{R}^{N\times C}$,

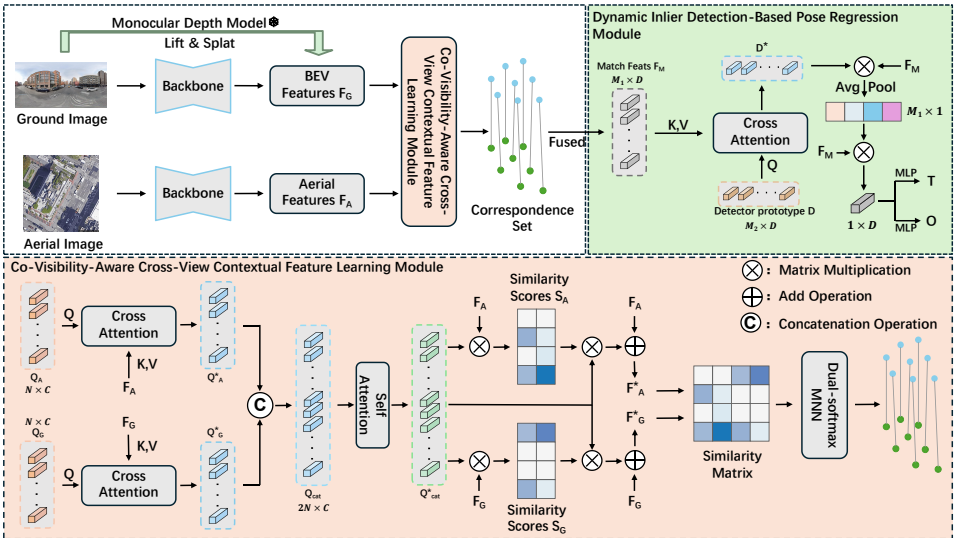

Figure 2: The architecture of our CoVisFormer consists of two major components, a co-visibility-aware cross-view contextual feature learning (C3FL) module and a dynamic inlier detection-based pose regression (DIDP) module. The ground-level image I_G and aerial image I_A are first sent into a feature extractor to obtain features. Next, we use a monocular depth estimation model to transform the ground features into the corresponding bev features $\mathbf{F}_G$, while the aerial branch produces features $\mathbf{F}_A$. Then, in the C3FL module, we define two sets of prompts $\mathbf{Q}_G$ and $\mathbf{Q}_A$ to separately interact with $\mathbf{F}_G$ and $\mathbf{F}_A$. After that, we concatenate updated prompts and enable cross-view interactions via a self-attention mechanism. Finally we use the updated prompts to produce cross-view contextual features. Next, we use the updated ground bev features and aerial features to establish cross-view correspondences and transform them into matching features $\mathbf{F}_M$. Then, in the DIDP module, we design a set of prototypes $\mathbf{D}$ to interact with $\mathbf{F}_M$ to produce adaptive inlier detectors, and we use these detectors to identify reliable inliers. Finally, we use two MLP heads to separately regress camera position and orientation from these inliers. For more details, please refer to the text.

yielding $\mathbf{Q}_v^* \in \mathbb{R}^{N\times C}$. To enable cross-view interactions, we concatenate the two updated prompts as $\mathbf{Q}_{\text{cat}} = [\mathbf{Q}_A^*; \mathbf{Q}_G^*] \in \mathbb{R}^{2N\times C}$ and apply self-attention to obtain $\mathbf{Q}_{\text{cat}}^* \in \mathbb{R}^{2N\times C}$. Then, we update original features $\mathbf{F}_G$ and $\mathbf{F}_A$ by fusing $\mathbf{Q}_{\text{cat}}^*$ to obtain cross-view contextual features. To this end, for a view $v \in \{G, A\}$, we calculate similarity scores $\mathbf{S}_v \in \mathbb{R}^{2N\times HW}$ between $\mathbf{F}_v$ and $\mathbf{Q}_{\text{cat}}^*$:

$$\mathbf{S}_v = \mathbf{Q}_{\text{cat}}^* \mathbf{F}_v^\top, \qquad \mathbf{F}_v^* = \mathbf{F}_v + \mathbf{S}_v^\top \mathbf{Q}_{\text{cat}}^*, \tag{3}$$

where $\mathbf{F}_v \in \mathbb{R}^{HW\times C}$. The above operations (Eqs. 1, 2, and 3) constitute the prompt-based attention mechanism. After obtaining the updated cross-view contextual features $\mathbf{F}_G^*$ and $\mathbf{F}_A^*$, we establish ground-to-aerial correspondences via dual-softmax and mutual nearest neighbors. Then, we use a confidence threshold τ to filter these matching pairs and retain at most M_1 of them, padding the set to M_1 pairs when fewer are retained. We write the resulting set as $\{(\mathbf{x}_k^G, \mathbf{x}_k^A, c_k)\}_{k=1}^{M_1}$, where $\mathbf{x}_k^G, \mathbf{x}_k^A \in \mathbb{R}^2$ are the metric ground and aerial coordinates, and c_k is the matching confidence.

3.3 Dynamic Inlier Detection-Based Pose Regression

To adaptively identify reliable inliers from matched correspondences and accurately regress the camera pose, we introduce an attention-based inlier detector in the dynamic inlier detection-based pose regression (DIDP) module. Given the correspondences established by the C3FL module, we first fuse the ground coordinate, aerial coordinate, displacement, and log-confidence of each pair into a match representation $\mathbf{z}_k \in \mathbb{R}^D$ via a pair encoder $\phi : \mathbb{R}^7 \to \mathbb{R}^D$:

$$\mathbf{z}_k = \phi\big([\mathbf{x}_k^G, \mathbf{x}_k^A, \mathbf{x}_k^A - \mathbf{x}_k^G, \log(c_k + \epsilon)]\big), \tag{4}$$

where $\epsilon > 0$ stabilizes the logarithm. Then we stack the pair representations to obtain the match features $\mathbf{F}_M = [\mathbf{z}_1; \ldots; \mathbf{z}_{M_1}] \in \mathbb{R}^{M_1\times D}$. After that, we define a set of learnable detector prototypes $\mathbf{D} \in \mathbb{R}^{M_2\times D}$ to capture the inlier patterns of the current correspondence set via the attention

operation. Specifically, keys and values arise from $\mathbf{F}_M$, and queries arise from $\mathbf{D}$. Formally,

$$\mathbf{Q}_D = \mathbf{D}\mathbf{W}_D^Q, \qquad \mathbf{K}_M = \mathbf{F}_M\mathbf{W}_M^K, \qquad \mathbf{V}_M = \mathbf{F}_M\mathbf{W}_M^V, \tag{5}$$

$$\mathbf{D}^* = \mathrm{Attention}(\mathbf{Q}_D, \mathbf{K}_M, \mathbf{V}_M) = \mathrm{softmax}\left(\frac{\mathbf{Q}_D\mathbf{K}_M^\top}{\sqrt{d}}\right)\mathbf{V}_M. \tag{6}$$

where $\mathbf{W}_D^Q, \mathbf{W}_M^K, \mathbf{W}_M^V \in \mathbb{R}^{D\times D}$ are learned linear projections. We use multi-head attention with H_D heads, where the per-head dimension is $d = D/H_D$. The scaling factor d in $\sqrt{d}$ denotes the per-head query/key dimension. The outputs of all heads are concatenated and projected back to $\mathbb{R}^{M_2\times D}$, yielding $\mathbf{D}^* \in \mathbb{R}^{M_2\times D}$. We treat each row $\mathbf{d}_r^* \in \mathbb{R}^D$ of $\mathbf{D}^*$ as an adaptive inlier detector. To evaluate correspondence k, we calculate its inlier probability under each detector and average the probabilities across all M_2 detectors:

$$p_{rk} = \mathrm{sigmoid}\left(\frac{(\mathbf{d}_r^*)^\top\mathbf{z}_k}{\sqrt{D}}\right), \qquad p_k = \frac{1}{M_2}\sum_{r=1}^{M_2} p_{rk}. \tag{7}$$

where $p_{rk}, p_k \in (0,1)$. We then combine p_k with the matching confidence c_k to weight each correspondence. Let $m_k \in \{0,1\}$ indicate whether pair k is valid rather than padded, and let $\beta \geq 0$ control the contribution of matching confidence. We compute the unnormalized weight $\bar{w}_k \geq 0$, normalized weight $w_k \geq 0$, and pooled feature $\mathbf{h} \in \mathbb{R}^D$ as

$$\bar{w}_k = m_k p_k c_k^\beta, \qquad w_k = \frac{\bar{w}_k}{\sum_{q=1}^{M_1}\bar{w}_q + \epsilon}, \qquad \mathbf{h} = \sum_{k=1}^{M_1} w_k\mathbf{z}_k. \tag{8}$$

Finally, we use two independent MLP heads $g_t, g_o : \mathbb{R}^D \to \mathbb{R}^2$ to predict translation $\mathbf{T} = g_t(\mathbf{h}) \in \mathbb{R}^2$ and orientation $\mathbf{O} = g_o(\mathbf{h}) \in \mathbb{R}^2$ from the pooled feature. The predicted camera translation is $\widehat{\mathbf{t}} = \mathbf{T} \in \mathbb{R}^2$. We obtain the yaw $\widehat{\theta} \in [-\pi, \pi)$ from the normalized orientation vector $\mathbf{O}$, interpreted as $[\cos\widehat{\theta}, \sin\widehat{\theta}]$. This representation avoids a discontinuity at $\pm\pi$.

3.4 Training Objectives

After obtaining the correspondence confidence matrix $\mathbf{C}$, the inlier probabilities p_k, and the predicted camera pose, we introduce four objective functions $\mathcal{L}_{\mathrm{match}}, \mathcal{L}_{\mathrm{soft}}, \mathcal{L}_{\mathrm{trans}}, \mathcal{L}_{\mathrm{yaw}} \in \mathbb{R}$ to guide model training. The matching loss encourages high confidence for geometrically consistent correspondences. The soft-inlier loss guides the dynamic detectors to assign higher probabilities to reliable matches. The translation and yaw losses supervise pose regression. To construct matching supervision, let $\theta^* \in [-\pi, \pi)$ and $\mathbf{t}^* \in \mathbb{R}^2$ denote the ground-truth pose, and let $\mathbf{R}(\theta^*) \in \mathbb{R}^{2\times 2}$ be its planar rotation matrix. For a ground-grid coordinate $\mathbf{x}_i^G \in \mathbb{R}^2$, the corresponding aerial coordinate $\mathbf{x}_i^{A,*} \in \mathbb{R}^2$ is

$$\mathbf{x}_i^{A,*} = \mathbf{R}(\theta^*)\mathbf{x}_i^G + \mathbf{t}^*. \tag{9}$$

A ground–aerial cell pair is retained if the following three conditions are satisfied: the transformed ground coordinate lies within the aerial grid, the two cells are mutually nearest in metric distance, and their distance is within a preset radius. Denoting the resulting ground-truth pairs by $\mathcal{G} \subseteq \{1,\ldots,L\}^2$, we define a focal-style positive matching loss:

$$\mathcal{L}_{\mathrm{match}} = -\frac{1}{|\mathcal{G}|}\sum_{(i,j)\in\mathcal{G}} \alpha(1-\mathbf{C}_{ij})^\gamma \log(\mathbf{C}_{ij}+\epsilon), \tag{10}$$

where $\alpha > 0$ weights positive pairs, $\gamma \geq 0$ is the focusing exponent, and $\epsilon > 0$ ensures numerical stability. Then, to supervise the inlier probabilities predicted by DIDP, we derive a soft target for each selected correspondence from its geometric error. With a metric scale $\sigma > 0$, the detached error $e_k \geq 0$ and soft inlier target $y_k \in (0,1]$ are

$$e_k = \left\|\mathbf{R}(\theta^*)\mathbf{x}_k^G + \mathbf{t}^* - \mathbf{x}_k^A\right\|_2, \qquad y_k = \exp\left(-\frac{e_k^2}{2\sigma^2}\right). \tag{11}$$

We apply binary cross-entropy only to valid, non-padded pairs:

$$\mathcal{L}_{\mathrm{soft}} = \frac{1}{\sum_k m_k}\sum_k m_k\ \mathrm{BCE}(p_k, y_k). \tag{12}$$

For pose regression, we penalize the translation error and the wrapped yaw error:

$$\mathcal{L}_{\text{trans}} = \|\widehat{\mathbf{t}} - \mathbf{t}^*\|_2, \qquad \mathcal{L}_{\text{yaw}} = \text{SmoothL1}\big(\text{wrap}(\widehat{\theta} - \theta^*), 0\big). \tag{13}$$

Finally, we combine the four objectives by a weighted sum:

$$\mathcal{L} = \lambda_m \mathcal{L}_{\text{match}} + \lambda_s \mathcal{L}_{\text{soft}} + \lambda_t \mathcal{L}_{\text{trans}} + \lambda_o \mathcal{L}_{\text{yaw}}, \tag{14}$$

where $\mathcal{L} \in \mathbb{R}$ is the total loss and $\lambda_m, \lambda_s, \lambda_t, \lambda_o \geq 0$ are balancing weights for the objectives.

4 Experiments

In this section, we first describe the implementation details, datasets, and evaluation metrics. Then, we present results on two public benchmarks and ablation studies of the proposed components. Additional ablations and qualitative results are provided in Appendices A and B.

4.1 Implementation Details

In this work, we implement the proposed model in PyTorch (Paszke et al., 2019). We adopt the same backbone as (Xia & Alahi, 2025) to extract original image features. Following (Xia et al., 2026), we use UniK3D (Piccinelli et al., 2025) for VIGOR and Depth Anything V2 (Yang et al., 2024b) for KITTI to predict their depth maps. In the co-visibility-aware cross-view contextual feature learning module, the number of prompts per view is set to 64, the H_P is set to 8. In the dynamic inlier detection-based pose regression module, the number of detector prototypes is set to 32, the H_D is set to 4, the D is set to 128. We set the matching-confidence exponent to $\beta = 1.0$ and the soft-inlier distance scale to $\sigma = 2.5$ m. We retain mutual nearest neighbor matches with confidence at least $\tau = 0.01$ and select at most 128 correspondences. For the positive matching loss, we use $\alpha = 0.25$ and $\gamma = 2.0$. The loss weights are $(\lambda_m, \lambda_s, \lambda_t, \lambda_o) = (1, 0.1, 1, 1)$. We first train the correspondence modules for 30 epochs using $\mathcal{L}_{\text{match}}$. For the next 70 epochs, we jointly optimize correspondence learning and pose regression using all four losses. We train our model using the AdamW optimizer (Loshchilov & Hutter, 2019). The learning rate is set to 10^{-4}, and the weight decay is 3×10^{-4}. All experiments are conducted on four NVIDIA RTX 3090 GPUs.

4.2 Datasets and Evaluation Metrics

VIGOR. VIGOR (Zhu et al., 2021) is a large-scale cross-view geo-localization dataset comprising ground-level panoramas and aerial orthophotos collected across four major U.S. cities. The main challenge lies in substantial viewpoint and illumination discrepancies. We follow the same procedure as Xia & Alahi (2025) to evaluate our method. For the *evaluation metric*, we report mean and median localization errors in meters and, when orientation is unknown, mean and median orientation errors in degrees.

KITTI. KITTI (Geiger et al., 2013) comprises forward-facing ground-level images captured by a vehicle and corresponding aerial imagery covering the surrounding areas. This dataset is challenging due to substantial viewpoint and scale discrepancies. We follow the same procedure as Xia & Alahi (2025) to evaluate our method. For the *evaluation metric*, we report mean and median localization and orientation errors, together with longitudinal and lateral localization recall and orientation recall at the thresholds specified in Table 2.

4.3 Comparison with State-of-the-Art Methods

Results on VIGOR. Table 1 compares the methods under known- and unknown-orientation settings. CoVisFormer achieves the lowest mean localization errors among the compared methods in both same-area and cross-area splits. Specifically, with known orientation, its mean errors are 1.79 m and 1.99 m, respectively; with unknown orientation, they are 2.27 m and 2.61 m. In the unknown-orientation cross-area split, CoVisFormer also obtains a mean orientation error of $6.09°$, reducing the mean localization and orientation errors of LoC^2 (Xia et al., 2026) by 38% and 48%, respectively. These gains are consistent with the intended roles of the C3FL Module in enhancing co-visible feature interactions and DIDP in weighting reliable correspondences for pose regression. Our advantage is most evident in mean errors, whereas VIRD maintains lower median orientation errors

Table 1: Evaluation results on the VIGOR dataset. Best results are shown in bold and second-best results are underlined. Quantitative comparison under known and unknown orientation settings.

Orien.	Methods	Same-area				Cross-area			
		↓ Localization (m)		↓ Orientation (°)		↓ Localization (m)		↓ Orientation (°)	
		Mean	Median	Mean	Median	Mean	Median	Mean	Median
Known	SliceMatch	5.18	2.58	–	–	5.53	2.55	–	–
	CCVPE	3.60	1.36	–	–	4.97	1.68	–	–
	GGCVT	4.12	1.34	–	–	5.16	1.40	–	–
	DenseFlow	3.03	0.97	–	–	5.01	2.42	–	–
	HC-Net	2.65	1.17	–	–	3.35	1.59	–	–
	FG^2	1.95	1.08	–	–	2.41	1.37	–	–
	LoC^2	3.06	1.59	–	–	3.43	1.90	–	–
	VIRD	2.57	1.14	–	–	3.85	1.47	–	–
	Ours	**1.79**	**0.85**	–	–	**1.99**	**1.05**	–	–
Unknown	SliceMatch	6.49	3.13	25.46	4.71	7.22	3.31	25.97	4.51
	CCVPE	3.74	1.42	12.83	6.62	5.41	1.89	27.78	13.58
	DenseFlow	4.97	1.90	11.20	1.59	7.67	3.67	17.63	2.94
	FG^2	8.95	7.32	15.02	2.94	10.02	8.14	31.41	5.45
	LoC^2	3.94	1.78	9.54	2.00	4.23	2.09	11.67	2.21
	VIRD	3.00	1.20	9.52	**0.96**	4.61	1.55	16.50	**1.17**
	Ours	**2.27**	**0.91**	**4.65**	1.54	**2.61**	**1.12**	**6.09**	1.81

Table 2: Evaluation results on the KITTI dataset. Best results are shown in bold and second-best results are underlined. The "Ori." column denotes the orientation noise used in training and testing, uniformly sampled within $\pm 10°$ or $\pm 180°$.

Area	Ori.	Methods	↓ Loc. (m)		↑ Lateral (%)		↑ Long. (%)		↓ Orien. (°)		↑ Orien. (%)	
			Mean	Median	R@1m	R@5m	R@1m	R@5m	Mean	Median	R@1°	R@5°
Same-area	±180°	SliceMatch	7.96	4.39	49.09	98.52	15.19	57.35	4.12	3.65	13.41	64.17
		CCVPE	6.88	3.47	53.30	85.13	25.84	68.49	15.01	6.12	8.96	42.75
		VIRD	5.55	2.22	79.43	98.41	29.84	73.89	4.87	**0.98**	**51.20**	92.82
		LoC^2	1.85	1.31	62.18	97.80	59.37	97.48	9.70	6.17	9.46	41.64
		Ours	**0.85**	**0.56**	**90.19**	**99.12**	**90.51**	**99.20**	**1.51**	1.02	48.99	**97.21**
Cross-area		SliceMatch	14.85	11.85	24.00	72.89	7.17	33.12	23.64	7.96	**31.69**	31.69
		CCVPE	13.94	10.98	23.42	60.46	11.81	42.12	77.84	63.84	3.14	14.56
		VIRD	11.12	5.41	**45.88**	**85.49**	14.97	55.25	22.03	**1.87**	30.39	**74.87**
		LoC^2	11.71	9.11	13.64	50.88	14.04	50.73	55.18	33.23	2.53	12.78
		Ours	**7.01**	**2.74**	35.80	77.30	**36.17**	**76.60**	**16.53**	3.10	19.06	64.79
Same-area	±10°	CCVPE	1.22	0.62	97.35	99.71	77.13	97.16	0.67	0.54	77.39	**99.95**
		HC-Net	0.80	0.50	**99.01**	99.73	92.20	99.25	**0.45**	0.33	**91.35**	99.84
		DenseFlow	1.48	0.47	95.47	99.79	87.89	94.78	0.49	**0.30**	89.40	99.31
		FG^2	0.75	0.51	95.81	99.66	92.50	99.05	0.93	0.66	67.27	98.91
		LoC^2	1.13	0.77	93.59	**99.97**	71.51	98.17	1.97	1.43	36.68	92.84
		Ours	**0.72**	**0.43**	92.34	98.65	**97.19**	**99.60**	0.67	0.52	78.51	99.82
Cross-area		CCVPE	9.16	3.33	44.06	90.23	23.08	64.31	**1.55**	**0.84**	**57.72**	**96.19**
		HC-Net	8.47	4.57	**75.00**	**97.76**	**58.93**	76.46	3.22	1.63	33.58	83.78
		DenseFlow	7.97	3.52	54.19	91.74	23.10	61.75	2.17	1.21	43.44	89.31
		FG^2	7.31	4.15	37.89	85.65	21.98	60.77	3.62	2.37	23.03	77.84
		LoC^2	5.60	3.01	45.29	92.43	27.01	68.26	3.32	2.12	26.03	80.68
		Ours	**4.76**	**1.89**	39.59	82.79	56.70	**93.17**	1.92	1.37	37.93	93.54

in both unknown-orientation splits, owing to its polar transformation (Park et al., 2026). Beyond the quantitative results, the six VIGOR examples in Fig. 3 show coherent cross-view correspondences, illustrating the model's ability to establish robust matches.

Results on KITTI. Table 2 evaluates localization with a limited field of view. CoVisFormer shows its largest gains under unknown orientation ($\pm 180°$). In the same-area split, it reduces the lowest baseline mean localization error from 1.85 m to 0.85 m and raises longitudinal R1m from 59.37% to 90.51%. In the cross-area split, its mean localization and orientation errors are 7.01 m and $16.53°$, compared with the lowest baseline errors of 11.12 m and $22.03°$. These results are consistent with the value of reliable correspondence selection when image overlap is limited and the heading is uncertain. With a $\pm 10°$ orientation prior, CoVisFormer also achieves the lowest mean and median localization errors in both area splits, although its orientation metrics are not consistently the best.



Figure 3: Qualitative cross-view localization results for six ground–aerial pairs. Each ground panorama is shown beside its corresponding aerial image, with green lines connecting matched points across the two views. In the aerial images, green triangles (GT) mark ground-truth camera locations, while yellow stars (Ours) mark CoVisFormer estimates.

4.4 Ablation Studies

To assess the contribution of each component in our proposed method, we compare four variants on KITTI-Ori180: model [A] uses neither the C3FL module nor the DIDP module, [B] adds the DIDP module, [C] adds the C3FL module, and [D] uses both. without the DIDP module, pose is estimated using the Procrustes and RANSAC procedure of FG^2 (Xia & Alahi, 2025). The component ablations report mean translation error in the same-area and cross-area splits. Additional hyperparameter and attention comparisons appear in Appendix A.

Table 3: Effectiveness of each component on KITTI-Ori180. We report the mean translation error in the same-area and cross-area settings.

Model	C3FL	DIDP	Same-area (m)↓	Cross-area (m)↓
[A]	✗	✗	8.86	11.63
[B]	✗	✓	3.97	8.76
[C]	✓	✗	1.43	8.05
[D]	✓	✓	**0.85**	**7.01**

Effects of the C3FL and DIDP modules. As shown in Table 3, adding C3FL to the baseline ([A] vs. [C]) reduces the same-area error from 8.86 m to 1.43 m and the cross-area error from 11.63 m to 8.05 m; it further improves the DIDP-equipped variant ([B] vs. [D]), lowering the errors from 3.97 m to 0.85 m and from 8.76 m to 7.01 m. Similarly, adding DIDP to the baseline ([A] vs. [B]) decreases the same-area error from 8.86 m to 3.97 m and the cross-area error from 11.63 m to 8.76 m, and it also boosts the C3FL-equipped variant ([C] vs. [D]), reducing the errors from 1.43 m to 0.85 m and from 8.05 m to 7.01 m. These controlled comparisons confirm that both cross-view contextual feature learning and dynamic inlier detection contribute to localization accuracy.

5 Conclusion

In this work, we proposed CoVisFormer, a 3-DoF cross-view geo-localization framework comprising a co-visibility-aware cross-view contextual feature learning (C3FL) module and a dynamic inlier detection-based pose regression (DIDP) module. With these two well-designed modules, our proposed method can learn cross-view contextual feature descriptors and enable identifying reliable inliers for camera pose estimation. Extensive experimental results on VIGOR and KITTI datasets demonstrate the superiority of our proposed method.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language editing, organization, and preparation of the manuscript. They were not used to generate experimental measurements, alter datasets, or determine the reported quantitative results. The authors reviewed the complete AI-assisted text, checked technical descriptions against the implemented method and experimental records, and take responsibility for all claims and artifacts in this work.

ETHICS STATEMENT

This work uses existing public cross-view localization benchmarks and does not involve human-subject experimentation. Cross-view localization can support safer autonomous navigation and robust positioning where GNSS is unreliable, but it may also enable location tracking or surveillance. Any deployment should therefore follow applicable privacy, data-governance, and safety requirements and should assess failure modes before use in safety-critical systems.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model components, tensor dimensions, correspondence construction, pose representation, and all training objectives. Section 4 documents the datasets, evaluation protocols, default architecture settings, and staged training schedule. The final submission will include configuration files and source code as anonymous supplementary material.

REFERENCES

Axel Barroso-Laguna, Sowmya Munukutla, Victor Adrian Prisacariu, and Eric Brachmann. Matching 2d images in 3d: Metric relative pose from metric correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4852–4863, 2024.

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

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16847–16856, 2023.

Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. C-BEV: Contrastive bird's eye view training for cross-view image retrieval and 3-dof pose estimation. *arXiv preprint arXiv:2312.08060*, 2023a.

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

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

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*, pp. 7258–7267, 2018.

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 *European Conference on Computer Vision*, 2026.

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

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with MASt3R. In *European Conference on Computer Vision*, 2024.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5624–5633, 2019.

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

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*, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Matteo Piccinelli et al. UniK3D: Universal camera monocular 3d estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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

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

Paul-Edouard Sarlin, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21632–21642, 2023.

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

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

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4064–4072, 2020.

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

Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2682–2697, 2023b.

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, 2023.

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

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

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

Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. BiFuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5448–5460, 2023a.

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*, pp. 8197–8206, 2023b.

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

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

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

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. Cross-view matching for vehicle localization by learning geographically local representations. *IEEE Robotics and Automation Letters*, 6(3):5921–5928, 2021.

Zimin Xia, Olaf Booij, Marco Manfredi, and Julian F. P. Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *European Conference on Computer Vision*, pp. 90–106, 2022.

Zimin Xia, Olaf Booij, and Julian F. P. Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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*, 2026.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024b.

Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021.

Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1162–1171, 2022.

A Additional Ablation Studies

We further examine the numbers of prompts and inlier detectors and compare prompt-based attention with full attention on KITTI-Ori180. All results report mean translation error in the same-area and cross-area splits.

Table 4: Ablation studies on KITTI-Ori180. We report the mean translation error under the same-area and cross-area settings.

(a) Number of prompts.

# Prompts	Same-area (m)↓	Cross-area (m)↓
16	1.01	7.48
32	0.90	7.05
64	**0.85**	**7.01**
128	0.88	7.03

(b) Number of inlier detectors.

# Detectors	Same-area (m)↓	Cross-area (m)↓
8	0.98	7.35
16	0.91	7.04
32	**0.85**	**7.01**
64	0.93	7.22

Table 5: Comparisons between full attention and our prompt-based attention on KITTI-Ori180. We report the mean translation error in the same-area and cross-area settings and model size.

Model	Same-area (m)↓	Cross-area (m)↓	Model Size (MB)↓
Full Attention	1.12	7.60	11.96
Our Prompt-based Attention	**0.85**	**7.01**	**9.98**

Number of prompts. To examine the effect of prompt count in the C3FL module, we vary the number N of prompts per view. As shown in Table 4a, increasing N from 16 to 64 reduces the same-area and cross-area mean localization errors to 0.85 m and 7.01 m, respectively. Increasing N further to 128 brings no additional improvement. This suggests that 64 prompts per view provide sufficient capacity to capture useful cross-view context. We therefore use $N = 64$ in the reported model.

Number of inlier detectors. To examine the effect of detector count in the DIDP module, we vary the number of learnable detector prototypes. Table 4b shows that increasing the count from 8 to 32 reduces the mean localization errors in both area splits, whereas using 64 detectors increases both errors. Thus, 32 detectors appear sufficient to model useful variations in the inlier patterns, and we use this setting in the reported model. Although the prompt and detector counts affect performance, the overall changes across the tested settings are modest, suggesting that the C3FL and DIDP modules are reasonably robust to these hyperparameters.

Prompt-based attention versus full attention. Table 5 compares the two interaction strategies with the remaining configuration fixed. Prompt-based attention reduces the same-area error from 1.12 m to 0.85 m and the cross-area error from 7.60 m to 7.01 m, while the reported model size decreases from 11.96 to 9.98 MB. These results corroborate the advantage of prompt-based attention over full attention, consistent with the objective of suppressing irrelevant cross-view aggregation.

B Additional Qualitative Results

Figure 4 presents six additional KITTI examples of cross-view correspondences and camera localization estimates.



Figure 4: Additional qualitative results on KITTI. Each ground image is shown beside its corresponding aerial image, with green lines connecting matched points across the two views. In the aerial images, green triangles (GT) mark ground-truth camera locations, while yellow stars (Ours) mark CoVisFormer estimates.