Sector-Level Cross-View Geo-Localization with Implicit Orientation via Azimuthal Scanning

Yiru Li, Le Wu, Yingchen Tan, and Yingying Zhu ✉

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
https://1203ll.github.io/SectorGeo/
2350271010@email.szu.edu.cn, wule2023@email.szu.edu.cn, 2410103063@mails.szu.edu.cn, zhuyy@szu.edu.cn

Abstract. Cross-View Geo-Localization (CVGL) is crucial for navigation in GPS-restricted environments. However, bridging the severe perspective gap between ground and aerial views remains challenging. Existing approaches either rely on coarse image-level retrieval, suffer from unstable optimization in pose regression, or require expensive fine-grained annotations. These limitations suggest that current formulations of cross-view geo-localization are insufficient for reliable fine-grained localization. To address this limitation, we introduce a new task, Sector-Level Cross-View Geo-Localization (SLCVGL), which aims to perform fine-grained sector-level geo-localization of a limited Field-of-View (FoV) ground image within an omnidirectional aerial image, while simultaneously inferring its ground-view orientation without manually annotated pose labels. To operationalize this task, we propose an Azimuthal Scanning framework. It decomposes aerial feature maps into multiple azimuthal sectors, instead of relying on holistic matching. By aligning ground features with these sectors, the model explicitly retrieves a sector-level sub-region and implicitly estimates orientation, eliminating the need for continuous pose regression supervision. This design also reduces interference from unobserved overhead regions. To benchmark the proposed task, we repurpose standard panoramic datasets into a limited-FoV sector-level protocol with random FoV cropping and soft labeling. Extensive experiments on CVUSA and CVACT demonstrate improvements over state-of-the-art methods, with a 30.17% R@1 improvement on CVACT under the 70° FoV setting, while maintaining strong robustness across varying FoVs.

Keywords: Cross-View Geo-Localization · Image Retrieval · Limited Field-of-View

1 Introduction

Cross-view Geo-localization (CVGL) has emerged as an effective vision-based alternative, providing robust localization when GPS signals are obstructed or perturbed by noise [5,7,13,32,37,39]. The core objective of CVGL is to ascertain the

✉ Corresponding author.

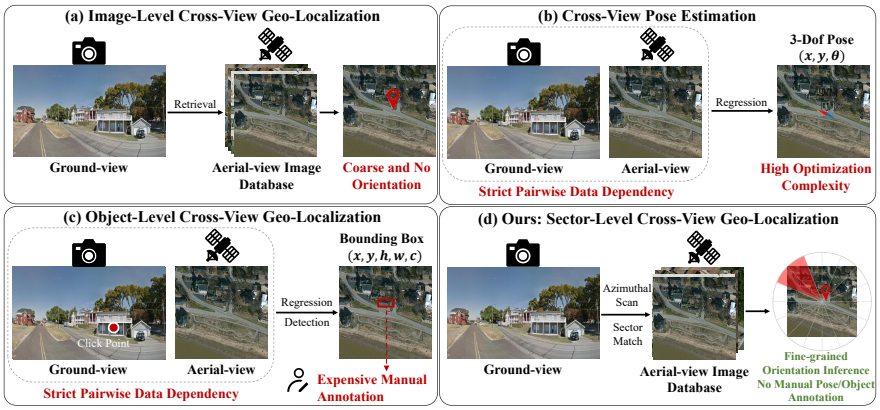


Fig. 1: Comparison of different cross-view geo-localization tasks. (a) Image-Level Cross-View Geo-Localization focuses on coarse-grained matching but remains orientation-agnostic. (b) Cross-View Pose Estimation attempts to regress the 3-DoF pose but suffers from high optimization complexity and strict pairwise data dependency, limiting its scalability in large-scale retrieval. (c) Object-Level Cross-View Geo-Localization achieves fine granularity but relies on expensive manual annotations and rigid instance-level correspondences. (d) Ours introduces a fine-grained scanning mechanism that explicitly retrieves the sector-level sub-region and implicitly determines the orientation without requiring explicit pose labels or labor-intensive supervision.

geo-localization of a ground query image by matching it against a geo-referenced database of aerial imagery. However, despite remarkable strides in deep representation learning, this task remains a formidable challenge. The primary difficulty stems from the drastic perspective gap between ground and aerial views, which manifests as severe geometric deformations and significant variations in visual appearance [10, 20, 28, 33, 42]. Consequently, overcoming these challenges is critical for unlocking advanced applications such as autonomous navigation [16, 21], augmented reality [2, 3], and precise search and rescue missions [18].

Existing research generally formulates the problem into three distinct tasks. Image-Level Cross-View Geo-Localization (Fig. 1a) enables efficient coarse-scale matching, yet remains limited by a significant granularity gap, as it lacks explicit modeling of spatial position and orientation. In contrast, fine-grained tasks such as Cross-View Pose Estimation (Fig. 1b) and Object-Level Cross-View Geo-Localization (Fig. 1c) attempt to address this limitation, but introduce substantial practical challenges. The former explicitly regresses pose parameters under a pairwise alignment formulation, which limits its scalability to large-scale retrieval settings. The latter relies on expensive instance-level annotations and assumes explicit instance-level correspondence between cross-view targets, which hinders scalability in large-scale real-world deployments.

Consequently, existing tasks remain trapped between coarse image-level retrieval and costly fine-grained localization. To break this impasse, we argue that fine-grained spatial cues can be extracted from inherent geometric constraints

rather than relying on expensive manual annotations. We thus revisit the geometric inclusion relationship between ground and aerial views. The disparity in their FoVs implies that a limited-FoV ground image corresponds to a spatial sub-region of the omnidirectional aerial scene. Building upon this sub-region relationship, the limited-FoV of a ground image further constrains the corresponding aerial region to a continuous azimuthal range. From a top-down perspective, such an angular constraint naturally translates into a wedge-shaped *sector*. Based on this observation, we propose a new task named Sector-Level Cross-View Geo-Localization (SLCVGL) (Fig. 1d). The objective is to identify the specific aerial sector whose features optimally align with the limited-FoV ground query. The spatial position of the matched sector provides fine-grained localization, while its azimuthal span implicitly indicates the orientation of the ground image. This design enables simultaneous sector-level localization and orientation inference, without manually annotated pose labels or labor-intensive instance-level labels.

To operationalize the new task, we introduce a new framework, termed **SectorGeo**, built upon an Azimuthal Scanning mechanism. Instead of treating the aerial image as a monolithic block, we model it as a continuous search space. We decouple the aerial feature map into overlapping azimuthal sectors, simulating azimuthal scanning around the aerial image center. During the matching process, the ground image is compared against these generated aerial sectors through the Joint Azimuthal Alignment and Relational Retrieval Stream. In this way, identifying the correct aerial image anchors the global position, while pinpointing the best-matching sector yields a finer-grained spatial localization and simultaneously reveals the discrete azimuthal orientation of the ground query.

To rigorously evaluate the proposed task, we observe that 360°-to-360° evaluation protocols are designed for coarse image-level retrieval, rendering them incapable of quantifying the precision of sector-level localization and orientation inference. Accordingly, we repurpose standard panoramic benchmarks (CVUSA and CVACT) into a rigorous sector-level evaluation protocol. To address the inherent label ambiguity arising from overlapping sectors and continuous orientation discretization, we introduce a Ground-Truth Soft Labeling mechanism. This provides smooth optimization targets derived directly from geometric priors, enabling both robust training and fine-grained benchmarking without requiring additional manual annotations. Our contributions are summarized as follows:

- To address the challenge of bridging coarse image-level retrieval and expensive fine-grained methods, we introduce a new task, Sector-Level Cross-View Geo-Localization (SLCVGL). Given a limited Field-of-View (FoV) ground image, the task requires identifying the corresponding aerial sector and implicitly estimating its orientation. By exploiting the geometric inclusion relationship between views, SLCVGL transforms fine-grained localization and continuous orientation estimation into a discrete sector-level retrieval problem, enabling explicit localization and implicit orientation inference without the need for pairwise alignments or labor-intensive annotations.
- To operationalize this task, we introduce a sector-based Azimuthal Scanning framework. The aerial feature map is decomposed into overlapping azimuthal

sectors, which are structurally aligned with the limited-FoV ground features. This alignment suppresses interference from unobserved regions. Crucially, by predicting the azimuthal sector of the ground query through discrete feature matching, our framework avoids the optimization challenges inherent in traditional pose regression.
- We establish an evaluation protocol by repurposing panoramic benchmarks and introducing a Ground-Truth Soft Labeling mechanism. This mechanism addresses label ambiguity arising from overlapping azimuthal sectors and continuous orientation angles by providing smooth optimization targets.
- Extensive experiments on the repurposed CVUSA and CVACT benchmarks demonstrate that SectorGeo achieves substantial improvements over state-of-the-art methods. Moreover, it maintains strong consistency between image retrieval and sector-level alignment, validating the effectiveness of SLCVGL.

2 Related Work

2.1 Image-Level Cross-View Geo-Localization

Image-Level Cross-View Geo-Localization is generally treated as an image retrieval task [1, 11, 17, 23, 29, 34, 35, 39, 40, 42, 43]. Given a ground-level query image, the objective is to retrieve the most relevant aerial reference image from a large-scale database. Early approaches relied on Convolutional Neural Networks (CNNs) to bridge the drastic domain gap. Liu et al. [14] pioneered the integration of orientation information into neural networks to assist feature alignment, while Shi et al. [23] employed Optimal Feature Transport theory to address spatial misalignment. Wang et al. [26] further emphasized the importance of local patterns, demonstrating that attending to salient regions significantly facilitates retrieval. With the advent of Vision Transformers (ViTs), the focus shifted toward capturing global dependencies. Yang et al. [32] proposed a layer-to-layer Transformer architecture to model long-range correlations. Ye et al. [34] proposed a Panorama-BEV Co-Retrieval Network to leverage complementary views. Most recently, Li and Zhu introduced PLGeo [11], a patch-level framework designed to explicitly overcome severe orientation discrepancies in retrieval tasks. Despite these advancements, standard geo-localization remains a coarse-grained task. It retrieves an aerial image covering a large physical area but fails to provide the specific coordinate or orientation within that image.

2.2 Cross-View Pose Estimation

Cross-View Pose Estimation aims to regress the precise 3-DoF pose (latitude, longitude, and orientation) of the ground image [4, 9, 30]. Shi et al. [22] retrieves a candidate image and then utilizes feature matching to estimate geometric offsets, achieving meter-level accuracy. SliceMatch [9] introduced a geometry-guided aggregation module, applying circular convolution to aerial feature slices to robustly estimate azimuth. To enhance reliability, Xia et al. [31] combined

metric localization with uncertainty estimates, allowing the model to quantify the confidence of its predicted pose. While pose estimation methods can achieve high localization accuracy, they suffer from several inherent limitations. First, they rely on precise pose annotations, which are costly and difficult to obtain at large scale. Second, learning continuous pose regression across different viewpoints is inherently challenging, often leading to unstable optimization.

2.3 Object-Level Cross-View Geo-Localization

Object-Level Cross-View Geo-Localization focuses on locating an object observed in a ground query within the aerial reference images [8, 12, 24, 37, 38, 41]. Sun et al. [24] refined feature alignment for target detection. Recently, Zhang and Zhu [37] proposed breaking rectangular shackles, moving beyond bounding boxes to perform cross-view object segmentation, achieving pixel-level localization. The primary bottleneck of this task is the high cost of supervision. Training these models requires laborious manual annotations (e.g., bounding boxes).

Existing tasks trade off between localization granularity, flexibility, and supervision cost. To bridge these gaps, we propose a task, called SLCVGL. By decoupling aerial features into directional sectors, our method enables fine-grained sub-region retrieval and implicit orientation estimation, bypassing the requirement for expensive annotations or rigid pose assumptions.

3 Method

3.1 Problem Statement

The goal of this research is to tackle a more complex and practical localization challenge: identifying the corresponding aerial image for a limited-FoV query while simultaneously retrieving its optimally matched azimuthal sector, thereby explicitly identifying a finer-grained sector within the aerial image and implicitly inferring the discrete camera orientation.

We define $\mathcal{I}^q = \{\mathbf{I}_i^q\}_{i=1}^N$ as a set of N ground query images and $\mathcal{I}^a = \{\mathbf{I}_j^a\}_{j=1}^M$ as a gallery of M geo-referenced aerial images. Each query $\mathbf{I}_i^q$ is a sub-region randomly cropped from a North-aligned panoramic image with a specific FoV $\phi \in \{70°, 90°, 135°, 180°\}$. Geometrically, a ground image captures a limited horizontal perspective, covering only a fraction of the surrounding environment. Consequently, when mapped onto the top-down aerial footprint, a limited-FoV query inherently corresponds to an azimuthal sector, rather than the entire 360° aerial image. Formally, given a query $\mathbf{I}^q$ with an unknown orientation θ and a known FoV ϕ, the task is to jointly identify the optimal aerial reference image index j^* and the regional sector index p^*:

$$(j^*, p^*) = \underset{\substack{j\in\{1,\dots,M\}\\ p\in\{1,\dots,P\}}}{\arg\max}\ \mathcal{F}\big(\mathbf{I}^q, \mathrm{Sector}(\mathbf{I}_j^a, p, \phi)\big)\,. \tag{1}$$

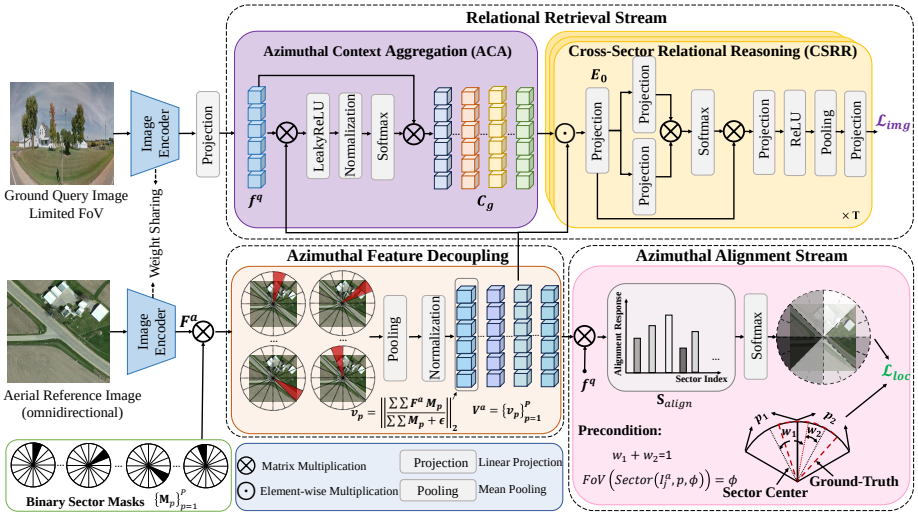


Fig. 2: Overview of the proposed framework. Given a limited-FoV ground query and a geo-referenced aerial image, the model performs azimuthal scanning by decomposing aerial features into multiple directional sectors and aligning them with the ground representation. The optimal sector correspondence jointly determines the matched aerial reference, the sector-level sub-region, and the estimated camera orientation.

where $\text{Sector}(I_j^a, p, \phi)$ represents the p-th azimuthal sector of the aerial image I_j^a with an angular span ϕ. In our protocol, all candidate sectors are generated by rotating a sector-shaped scanning window around the center of the aerial image. This center-based mask generation serves as a consistent scanning convention for defining sector candidates, rather than a requirement that the ground camera center must exactly coincide with the aerial image center. Here, $\mathcal{F}$ denotes the overall cross-view matching objective, and P is the total number of candidate sectors. The arg max operator essentially formulates a global search, identifying the specific index pair (j^*, p^*) that maximizes the matching similarity across all M aerial references and their P directional sectors. *By uniformly discretizing the 360° azimuth into P sectors, the optimally matched index p^* intrinsically defines the geographical heading of the sector, which directly yields the estimated discrete orientation $\hat{\theta}$ of the ground camera.*

3.2 Feature Extraction and Representation

As shown in Fig. 2, we employ a dual-branch architecture based on ConvNeXt-Base [15] to extract features from both views. By sharing parameters, the model maps ground and aerial images into a unified embedding space, thereby mitigating the inherent domain gap and facilitating cross-view alignment.

Ground-view Branch. Given a limited-FoV ground query $\mathbf{I}^q$, the backbone network serves as a global encoder. It directly extracts a high-level semantic

representation, which is then L_2 normalized to obtain the final global descriptor $\mathbf{f}^q \in \mathbb{R}^C$, where C denotes the feature dimension.

Aerial-view Branch. Unlike the ground branch, the aerial image $\mathbf{I}^a$ requires the preservation of spatial information to support directional alignment. To this end, we remove the final global pooling layer and the fully connected head, utilizing the backbone as a dense feature extractor to produce a spatial feature map $\mathbf{F}^a \in \mathbb{R}^{C\times H\times W}$, where $H \times W$ represents the spatial resolution.

Azimuthal Feature Decoupling. To bridge the gap between the global query vector $\mathbf{f}^q$ and the spatial feature map $\mathbf{F}^a$, we perform an azimuthal scanning process by rotating a sector-shaped mask around the aerial image center. This fixed center is used only to define a unified set of directional sector candidates. Specifically, we generate a set of P binary sector masks $\{\mathbf{M}_p\}_{p=1}^{P}$, where each mask $\mathbf{M}_p \in \{0,1\}^{H\times W}$ defines a sector-shaped region with an angular span ϕ matching the FoV of the query. These masks are generated by rotating a sector-shaped scanning window clockwise in the circular azimuthal space with a constant angular step $\Delta\theta$, yielding $P = \lfloor 360^\circ/\Delta\theta \rfloor$ candidate sector positions. Each sector has an angular span of ϕ, matching the FoV of the ground query. In this way, the aerial feature map is scanned over a full azimuthal cycle, while the degree of overlap between adjacent sectors is determined by the relationship between ϕ and $\Delta\theta$.

To aggregate the spatial features within each sector, we employ a masked average pooling strategy. Since the backbone feature map already encodes high-level spatial semantics, the sector-level pooling acts as a robust regional summarization rather than destroying the underlying spatial structure. Let $\mathbf{F}^a(h,w) \in \mathbb{R}^C$ denote the feature vector at spatial location (h,w), and $\mathbf{M}_p(h,w) \in \{0,1\}$ denote the corresponding scalar mask value. The p-th azimuthal descriptor $\mathbf{v}_p \in \mathbb{R}^C$ is computed by spatially integrating the masked features:

$$\mathbf{v}_p = \left\| \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{F}^a(h,w)\mathbf{M}_p(h,w)}{\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbf{M}_p(h,w) + \epsilon} \right\|_2, \tag{2}$$

where ϵ is a small constant, and $\|\cdot\|_2$ denotes L_2 normalization. This process yields a sequence of directional candidates $\mathbf{V}^a = \{\mathbf{v}_p\}_{p=1}^{P}$, effectively transforming the aerial feature into an orientation-aware feature sequence that structurally corresponds to discrete geographic headings. Since sectors may be partially truncated by image boundaries, normalization by the masked area ensures that each descriptor remains scale-invariant with respect to sector size.

3.3 Joint Azimuthal Alignment and Relational Retrieval

We propose a dual-path framework that decouples the task into: (i) a Relational Retrieval Stream for image-level matching, and (ii) an Azimuthal Alignment Stream for fine-grained sector localization and implicit orientation estimation.

Relational Retrieval Stream (RRS). The primary objective of the relational retrieval stream is to determine the global similarity between the limited-FoV ground query and the aerial reference images by evaluating the fine-grained

correspondences between the query and individual azimuthal sectors. Geometrically, a direct image-to-image comparison is suboptimal because the narrow-angle query only corresponds to a specific azimuthal sub-region of the omnidirectional aerial image, while the remaining sectors act as noise. To address this, we propose a hierarchical process that calculates sector-wise similarities and refines them into a robust image-level retrieval score.

Azimuthal Context Aggregation (ACA). To bridge the gap between the omnidirectional aerial sequence and the limited-FoV ground view image, we employ the ACA mechanism to perform semantic reconstruction. This module utilizes each aerial azimuthal candidate $\mathbf{V}^a \in \mathbb{R}^{P\times C}$ to adaptively aggregate cues from the ground descriptor $\mathbf{f}^q$. Specifically, an affinity matrix $\mathbf{A} \in \mathbb{R}^{P\times 1}$ is first computed via matrix multiplication:

$$\mathbf{A} = \text{LeakyReLU}\left(\mathbf{V}^a \mathbf{f}^q\right). \tag{3}$$

Here, the matrix multiplication calculates the raw geometric cosine alignment in the latent space. The application of LeakyReLU introduces non-linearity while preventing zero-gradients for diametrically opposed directions, ensuring training stability. The affinity vector $\mathbf{A}$ is then L_2 normalized and transformed into attention weights via a smoothed Softmax. Although $\mathbf{f}^q$ is a single global descriptor, the attention weights modulate its contribution differently for each azimuthal sector, yielding sector-specific semantic emphasis. The reconstructed ground context $\mathbf{C}_g \in \mathbb{R}^{P\times C}$, which represents the ground-level information semantically aligned to each aerial heading, is derived as:

$$\mathbf{C}_g = \text{Softmax}(A/\|A\|_2)(\mathbf{f}^q)^\top. \tag{4}$$

The model allocates semantic information of the query across the P aerial headings. By scaling the ground feature with the distributed probabilities, the model inherently amplifies the ground signal along the most likely matching directions while suppressing its magnitude along highly mismatched headings, thereby yielding a highly discriminative, direction-specific context representation.

Cross-Sector Relational Reasoning (CSRR). We perform an element-wise fusion between the reconstructed context $\mathbf{C}_g$ and the original azimuthal sequence $\mathbf{V}^a$ using the Hadamard product $\odot$. The fused representation is then projected through a linear transformation $\mathbf{W}_t$ and L_2 normalized to form the initial similarity embedding $\mathbf{E}_0$:

$$\mathbf{E}_0 = \left\|(\mathbf{C}_g \odot \mathbf{V}^a)\mathbf{W}_t\right\|_2, \tag{5}$$

where $\mathbf{W}_t$ is a learnable projection matrix. This embedding encapsulates the raw matching strength between the ground observations and each individual aerial sector in the latent space. Relying strictly on isolated sector-wise comparisons is prone to local visual ambiguities, as distinct geometric locations may share similar local patterns. To ensure a robust match, it is crucial to verify the holistic spatial configuration of the 360° aerial image. To capture these global contextual dependencies and suppress isolated false positives, we employ T iterative CSRR blocks (indexed by $t = 1, \ldots, T$). Each block treats the similarity embeddings of

the azimuthal sectors $\mathbf{E}_{t-1}$ as nodes in a graph and updates their states through a dynamic self-attention mechanism:

$$\mathbf{M}_t = \mathrm{Softmax}\left((\mathbf{E}_{t-1}\mathbf{W}_Q)(\mathbf{E}_{t-1}\mathbf{W}_K)^\top\right), \quad \mathbf{E}_t = \mathrm{ReLU}\left((\mathbf{M}_t\mathbf{E}_{t-1})\mathbf{W}_g\right), \quad (6)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_g$ are learnable projection matrices and $\mathbf{M}_t$ represents the inter-sector affinity. Finally, this refined sequence of orientation-aware similarity embeddings is converted into a reliable global retrieval score. Specifically, the score $S_{ret} \in \mathbb{R}$ is computed by applying average pooling over the refined node features $\mathbf{E}_T$ along the sector dimension, followed by a linear projection $\mathbf{W}_e$:

$$S_{ret} = \left(\frac{1}{P}\sum_{p=1}^{P}\mathbf{E}_T(p)\right)\mathbf{W}_e, \quad (7)$$

where $\mathbf{W}_e \in \mathbb{R}^{C\times 1}$ is a learnable projection matrix. The relational retrieval stream addresses the viewpoint asymmetry between queries and omnidirectional references. By modulating ground context via aerial headings and enforcing structural consensus, the network suppresses regional noise and visual ambiguities. Consequently, S_{ret} ensures that the top candidates share strong structural correspondences with the limited-FoV ground query images.

Azimuthal Alignment Stream (AAS). The azimuthal alignment stream aims to identify the specific azimuthal sector within the aerial image that corresponds to the ground-level view. By establishing a sector-wise correspondence between the ground descriptor $\mathbf{f}^q \in \mathbb{R}^C$ and the candidate orientation-aware sequence $\mathbf{V}^a = \{\mathbf{v}_p\}_{p=1}^P$, we treat orientation estimation as a discrete sector localization task. The alignment response $\mathbf{S}_{align} \in \mathbb{R}^P$ is computed via:

$$\mathbf{S}_{align} = \mathbf{V}^a\mathbf{f}^q. \quad (8)$$

The AAS serves as a geometric anchor. By localizing the optimal sector index $p^* = \arg\max(\mathbf{S}_{align})$, the orientation of the query image is implicitly determined by the heading of the matched sector. This stream operates on cosine similarity without contextual reasoning, providing a lightweight geometric prior for orientation estimation instead of a discriminative verification signal.

The core strength of this design lies in the reciprocal synergy between the Azimuthal Alignment Stream and the Relational Retrieval Stream. While the Azimuthal Alignment Stream provides implicit orientation supervision, the Relational Retrieval Stream refines this direction awareness by forcing the model to distinguish fine-grained structural details during scoring. As these tasks are optimized simultaneously, the discriminative features from the retrieval process back-propagate to enhance the precision of sector localization, resulting in a more robust and orientation-aware representation.

Inference Pipeline. During inference, we leverage this dual-path design to form an efficient coarse-to-fine localization pipeline. For a given query, we first compute the image-level retrieval scores S_{ret} to rank all aerial references in the global database. Subsequently, the azimuthal alignment stream $\mathbf{S}_{align}$ is

directly evaluated on the top-K aerial candidates to estimate the most likely corresponding sector, thereby simultaneously yielding the sector-level location and the discrete orientation with limited additional computational overhead.

3.4 Training Objective

Given a mini-batch of size B, we jointly optimize the relational retrieval and azimuthal alignment objective.

Relational Retrieval Loss. Let $S_{ret}(i,j)$ denote the retrieval score between the i-th query and the j-th reference produced by the relational retrieval stream. For a mini-batch of size B, the pairwise scores form a similarity matrix $\mathbf{S}_{ret} \in \mathbb{R}^{B\times B}$. We adopt InfoNCE loss [19] with learnable temperature parameter τ:

$$\mathcal{L}_{img} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(S_{ret}(i,i)/\tau\right)}{\sum_{j=1}^{B}\exp\left(S_{ret}(i,j)/\tau\right)}. \tag{9}$$

Azimuthal Alignment Loss. For the azimuthal alignment stream, the model outputs azimuthal logits $\mathbf{S}_{align}^{(i)} \in \mathbb{R}^{P}$ for the i-th query, where each element corresponds to the predicted score of one azimuthal sector. Instead of using a hard one-hot label, we construct a soft target distribution $\mathbf{y}^{(i)} \in \mathbb{R}^{P}$ by assigning non-zero weights $w_1^{(i)}$ and $w_2^{(i)}$ to the two neighboring sector indices $\{p_1^{(i)}, p_2^{(i)}\}$ based on the linear inverse angular distance to the ground-truth orientation θ_{gt}, satisfying $w_1^{(i)} + w_2^{(i)} = 1$. The alignment loss is defined:

$$\mathcal{L}_{loc} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{p=1}^{P} y_p^{(i)}\,\log\left(\mathrm{Softmax}\left(\mathbf{S}_{align}^{(i)}\right)_p\right). \tag{10}$$

Overall Objective. The final loss is a weighted combination:

$$\mathcal{L} = \beta\mathcal{L}_{loc} + (1-\beta)\mathcal{L}_{img}, \tag{11}$$

where β balances the weights of azimuthal alignment and relational retrieval.

4 Experiments

4.1 Datasets

CVUSA and CVACT. CVUSA [27] comprises 35,532 training and 8,884 testing pairs, covering rural environments. CVACT [14] offers a similar scale with 35,532 training and 8,884 testing pairs, but focuses on dense urban structures.
CVUSA-C and CVACT-C. To evaluate the robustness against real-world environmental challenges, we validate our model on two corruption robustness benchmarks, CVUSA-C [36] and CVACT-C [36].

4.2 Sector-Level Geo-Localization Protocol

Ground-Truth Soft Labeling. While random FoV cropping provides the necessary limited-FoV queries, it introduces a critical challenge for our specific sector-level task: a randomly sampled continuous orientation θ rarely aligns perfectly with a discrete sector center. To resolve this and successfully repurpose the dataset for orientation estimation, we introduce a Ground-Truth Soft Labeling mechanism. Given the discretization of the orientation space into P azimuthal sectors, we formulate the alignment target as a soft distribution. For a query with orientation θ, we identify the two closest discrete sector indices $\{p_1, p_2\}$ and assign their weights $\{w_1, w_2\}$ via linear interpolation based on angular distance [6, 25], such that $w_1 + w_2 = 1$. This soft-labeling strategy preserves the continuous geometric information of the camera heading and provides a smoother optimization landscape compared to traditional one-hot encodings.

Metrics. We evaluate the performance using two metrics: **Image-Level Recall (R@K)** and **Sector-Level Recall (S-R@K)**. **R@K** follows the standard retrieval protocol, measuring the percentage of queries whose ground-truth aerial image is ranked within the top K candidates ($K \in \{1, 5, 10\}$). **S-R@K** is a stricter joint metric specifically tailored to the proposed sector-level task. For each query, we first determine whether its ground-truth aerial image is retrieved within the top K candidates according to the image-level retrieval ranking. If so, we further evaluate the predicted azimuthal sector associated with this ground-truth aerial image. A query is counted as correct under S-R@K only when the ground-truth aerial image is retrieved within the top K candidates *and* the predicted sector index exactly matches the closest ground-truth sector, defined as the sector with the maximum soft-label weight.

4.3 Implementation Details

Model Architecture. We use Sample4Geo [1] as baseline and adopt ConvNeXt-Base [15] as backbone. To preserve the distinct geometric properties of each view, we apply specific resizing strategies. The aerial reference images are resized to 384×384. For the ground view, the original full-view panoramas are resized to 140×768. During training and inference, the limited-FoV queries are dynamically cropped from these panoramas based on the specified FoV angle ϕ.

Training Protocol. Our model is implemented in PyTorch and trained on a single NVIDIA A100-SXM4-40GB GPU. The training spans 40 epochs with a batch size of 64. We employ the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.05. The learning rate follows a cosine decay schedule ending at 1×10^{-5}.

Baseline Adaptation. To ensure a fair comparison under the proposed sector-level protocol, all methods are retrained on the repurposed CVUSA and CVACT datasets with the same limited-FoV query generation strategy. Since most existing methods are originally designed for image-level retrieval, we equip them with the same Azimuthal Feature Decoupling strategy described in Sec. 3.2 to obtain sector-level aerial descriptors. For each ground query and aerial

Table 1: Quantitative comparison on the repurposed CVUSA and CVACT benchmarks. We report performance across varying limited-FoVs ($\phi \in \{70°, 90°, 135°, 180°\}$) with unknown orientations. **Bold** and underline indicate the best and second-best performance, respectively. All comparison methods are implemented according to publicly available source code settings to ensure a fair comparison.

Set	Methods	CVUSA						CVACT					
		R@1	R@5	R@10	S-R@1	S-R@5	S-R@10	R@1	R@5	R@10	S-R@1	S-R@5	S-R@10
FoV = 70°	FRGeo [35]	10.03	20.54	34.10	9.41	23.18	32.20	7.82	21.52	30.91	7.20	19.55	27.96
	GeoDTR [40]	4.20	9.23	12.47	2.40	4.94	6.29	3.89	8.47	10.58	2.21	4.56	5.98
	GeoDTR+ [39]	4.78	10.24	14.58	3.67	5.84	10.47	4.24	9.24	12.98	3.24	5.24	9.68
	Sample4Geo [1]	35.33	62.83	73.35	34.20	60.73	70.82	6.96	18.73	26.51	6.24	16.69	23.55
	ConGeo [17]	40.49	64.13	72.81	39.74	62.66	71.04	8.31	20.05	27.02	7.60	18.25	24.57
	PLGeo [11]	44.56	69.35	77.53	28.69	44.17	48.95	11.33	26.47	35.31	9.78	21.97	28.79
	Ours	**63.79**	**84.62**	**89.91**	**59.92**	**78.47**	**82.95**	**41.50**	**64.27**	**72.82**	**39.81**	**61.02**	**68.83**
FoV = 90°	FRGeo [35]	11.82	28.96	39.75	11.31	27.27	38.10	8.36	20.78	29.72	7.74	18.83	26.80
	GeoDTR [40]	6.47	13.54	25.21	4.96	9.58	11.54	5.24	11.57	20.14	4.36	9.24	10.68
	GeoDTR+ [39]	7.01	15.84	26.96	5.78	10.27	13.84	6.16	12.98	21.74	5.47	9.87	11.58
	Sample4Geo [1]	40.48	69.42	79.04	39.29	67.42	76.67	12.02	27.42	36.09	11.12	24.88	32.50
	ConGeo [17]	48.77	72.47	80.15	47.81	70.89	78.34	11.29	25.06	32.99	10.70	23.39	30.59
	PLGeo [11]	53.94	77.96	85.03	40.42	58.34	63.78	19.69	36.47	44.20	15.53	26.55	30.72
	Ours	**72.44**	**89.87**	**93.62**	**66.41**	**81.43**	**84.66**	**49.30**	**71.05**	**77.66**	**40.75**	**56.74**	**61.23**
FoV = 135°	FRGeo [35]	19.80	42.74	55.08	19.30	41.68	53.66	13.44	32.14	43.34	11.47	27.09	36.77
	GeoDTR [40]	11.37	22.46	28.13	5.29	10.31	12.83	9.57	20.58	25.64	5.01	9.87	11.23
	GeoDTR+ [39]	12.85	25.98	30.47	6.89	12.58	16.58	9.89	21.47	26.22	6.31	10.44	13.24
	Sample4Geo [1]	50.40	79.19	87.40	49.08	77.13	85.10	22.53	43.77	53.52	21.10	40.07	48.34
	ConGeo [17]	67.01	86.26	90.52	65.89	84.69	88.71	21.43	40.01	48.28	20.44	37.61	44.85
	PLGeo [11]	55.88	79.92	86.36	51.19	72.97	78.59	27.61	45.88	53.88	24.52	38.80	44.38
	Ours	**81.33**	**94.55**	**96.76**	**77.44**	**89.19**	**91.12**	**52.47**	**73.62**	**80.73**	**42.59**	**57.64**	**61.89**
FoV = 180°	FRGeo [35]	31.42	57.00	67.83	30.07	54.40	64.62	23.48	48.22	58.81	21.67	43.88	53.15
	GeoDTR [40]	12.78	23.50	29.64	5.98	10.96	13.70	11.24	19.68	25.41	5.04	7.89	11.24
	GeoDTR+ [39]	14.89	29.87	32.47	7.48	13.24	19.89	13.47	23.57	27.47	7.21	12.57	15.96
	Sample4Geo [1]	65.47	87.32	92.19	63.28	84.30	88.93	40.36	63.13	70.72	37.24	57.58	64.02
	ConGeo [17]	79.71	92.69	95.09	77.90	90.07	**92.23**	37.92	57.96	64.93	35.78	53.24	59.08
	PLGeo [11]	77.30	93.02	95.89	73.66	88.23	90.79	37.72	58.95	66.83	36.50	56.76	64.24
	Ours	**87.87**	**96.75**	**98.01**	**83.67**	**91.12**	92.02	**57.38**	**76.52**	**82.04**	**46.11**	**59.23**	**66.56**

reference, the predicted sector index is determined by selecting the aerial sector descriptor with the highest similarity to the ground descriptor. Therefore, R@K evaluates image-level retrieval performance, while S-R@K further measures whether the predicted sector associated with the ground-truth aerial image matches the ground-truth sector label.

Hyperparameters. The parameter β in Eq. (11) is set to 0.9 and 0.3 on the repurposed CVUSA and CVACT datasets. The parameter T in Eq. (7) is set to 3 on the repurposed CVUSA and CVACT datasets.

4.4 Comparison with State-of-the-Art Methods

Effectiveness of Image Retrieval. Table 1 presents a comparison on the repurposed CVUSA and CVACT benchmarks. The proposed method consistently outperforms state-of-the-art methods. The performance margin is particularly pronounced in the most constrained scenarios ($\phi \in 70°, 90°$). Specifically, on the urban CVACT dataset, our framework achieves R@1 improvements of 30.17%

and 29.61% over the second-best methods at 70° and 90° settings, respectively. Similar trends are observed on the rural CVUSA dataset. In these restricted settings, the limited FoV of the ground image reduces overlapping context with the aerial image, making holistic matching prone to errors. Our Azimuthal Scanning mechanism decomposes the aerial feature map into directional sectors, aligning ground features with the most relevant sub-regions. This targeted alignment reduces interference from unobserved areas and directly contributes to the observed performance gains, particularly under tight FoV constraints.

Precision of Azimuthal Alignment. Unlike competing methods that exhibit significant performance drops when transitioning from image retrieval to joint sector alignment, our framework demonstrates remarkable resilience. For instance, the performance of PLGeo on CVUSA at 70° drops sharply from 53.94% (R@1) to 40.42% (S-R@1). In contrast, our framework preserves metric consistency. Under the 90° FoV, our S-R@1 reaches 66.41% on CVUSA and 40.75% on CVACT, outperforming the second-best methods by 18.60% and 25.22%, respectively. This consistent joint performance confirms that our model effectively aligns the limited-FoV ground query with its corresponding geometric sub-region rather than merely retrieving visually similar global images. Empowered by the Ground-Truth Soft Labeling, our discrete sector-matching mechanism reliably identifies the correct azimuthal sector, achieving robust directional localization.

4.5 Ablation Study

Effectiveness of Components and Pooling Strategies. To validate the contribution of each component and determine the optimal pooling strategy, we conduct a comprehensive ablation study on the CVUSA dataset, as summarized in Table 2. The full model achieves the strongest image-level retrieval performance, with R@1, R@5, and R@10 reaching 87.87%, 96.75%, and 98.01%, respectively. It also obtains the highest S-R@1 of 83.67%, indicating superior top-1 sector-level localization accuracy. Specifically, removing the CSRR module decreases R@1 from 87.87% to 77.74%, resulting in a substantial drop of 10.13 percentage points. This demonstrates that CSRR is essential for propagating sector-level information across different azimuthal directions and maintaining structural consistency. Similarly, omitting the ACA module reduces R@1 to 84.48%, corresponding to a 3.39 percentage-point drop, which validates its role in aligning ground descriptors with specific aerial headings before relational modeling. Regarding feature aggregation within the Azimuthal Feature Decoupling module, we evaluate Max and Mixed which combine Average and Max pooling for generating the sector-specific descriptors. Standard Average Pooling yields the best top-1 and image-level retrieval performance. The superiority of Average Pooling implies that preserving the global distribution of features within a sector is more beneficial for orientation-aware representations than focusing on isolated local salient peaks.

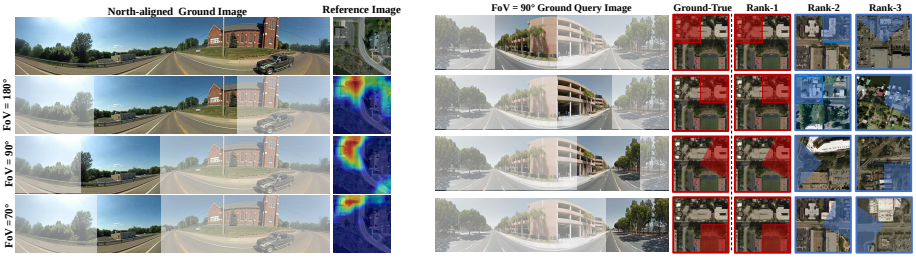
Impact of Azimuthal Step Size $\Delta\theta$. To determine the optimal azimuthal step size $\Delta\theta$ for generating sector-specific descriptors, we evaluate various sampling densities on CVUSA, as summarized in Table 3. The results demonstrate

Table 2: Ablation study of key components and pooling strategies on CVUSA with 180° FoV. 'Avg', 'Max', and 'Mix' represent Average Pooling, Max Pooling, and their combination.

Modules		Pooling	FoV=180°					
CSRR	ACA		R@1	R@5	R@10	S-R@1	S-R@5	S-R@10
	✓	Avg	77.74	92.23	95.97	75.74	89.82	92.37
✓		Avg	84.48	94.81	96.84	82.86	**91.97**	93.55
✓	✓	Max	81.40	94.20	96.66	79.60	91.12	93.09
✓	✓	Mix	81.28	94.44	96.86	79.78	91.75	**93.60**
✓	✓	Avg	**87.87**	**96.75**	**98.01**	**83.67**	91.12	92.02

Table 3: Ablation study on the azimuthal step size $\Delta\theta$ on CVUSA. P represents the number of sampled sectors, calculated as $P = \lfloor 360°/\Delta\theta \rfloor$.

$\Delta\theta(°)$	P	FoV=180°					
		R@1	R@5	R@10	S-R@1	S-R@5	S-R@10
40	9	87.23	96.24	97.65	83.01	89.96	91.53
50	7	**87.87**	**96.75**	98.01	**83.67**	91.12	92.02
60	6	86.92	96.38	97.70	80.13	87.74	88.72
70	5	87.33	96.62	**98.14**	83.54	**91.77**	92.86
80	4	84.56	96.18	97.93	80.95	91.50	**92.98**
90	4	85.94	96.32	97.66	82.56	91.42	92.40
100	3	82.04	95.17	97.12	78.25	90.23	92.06



(a) (b)

Fig. 3: (a) Qualitative visualization of attention heatmaps on the reference aerial images given ground inputs with varying FoVs (180°, 90°, and 70°). (b) Qualitative visualization of the top-3 retrieval results for ground queries with a 90° FoV. The red and blue borders indicate the positive and negative samples, respectively. The overlaid colored masks represent the specific azimuthal sectors localized by the proposed method.

that the evaluated granularity ($\Delta\theta = 50°$, $P = 7$) achieves the highest recall. While the performance remains relatively stable between 50° and 90°, increasing the step size to 100° results in a performance decline. A large step size increases the quantization error between the discretely sampled aerial headings and the actual ground orientation. Conversely, denser sampling generates a finer sequence of directional candidates, ensuring a higher probability that at least one extracted aerial sector geometrically covers the true query FoV. While very fine-grained sampling, such as $\Delta\theta = 40°$, slightly reduces performance compared to the optimal 50° step, likely due to increased redundancy and potential overfitting, indicating that excessively small step sizes are not necessarily beneficial.

4.6 Visualization

Visualization of Heatmaps. Fig. 3a provides heatmap visualizations on reference aerial images to evaluate spatial alignment under limited-FoV query image. A direct spatial correspondence exists between the ground and the high-

activation regions within the aerial heatmap. When the ground image captures specific landmarks such as church buildings or road junctions, the heatmap highlights the exact geographical sector containing these features. As the FoV narrows from 180° to 70°, the activation area on the aerial image becomes more localized and directional. This transition indicates that the azimuthal scanning filters irrelevant overhead information and anchors the orientation of the query by matching the horizontal perspective with the corresponding aerial sector.

Visualization of Retrieval Results. Fig. 3b presents the top-3 retrieval results for ground queries with 90°FoV. The framework successfully retrieves the correct aerial map as the rank-1 candidate. Furthermore, the azimuthal scanning mechanism accurately localizes the specific geometric sub-region corresponding to the ground query. This precise sector localization, highlighted by the overlaid masks in the positive samples, aligns consistently with the ground-truth annotations. For the incorrect rank-2 and rank-3 candidates, the network still extracts plausible structural sectors based on local visual similarities. However, since the image retrieval fails for these negative samples, their corresponding sector predictions are mismatched. This confirms that the proposed architecture effectively couples cross-view image retrieval with fine-grained sector-level localization via explicit geometric correspondences.

5 Conclusion

In this paper, we introduce Sector-Level Cross-View Geo-Localization, a novel task that exploits geometric inclusion relationship between limited-FoV ground queries and omnidirectional aerial imagery. Through the Azimuthal Scanning framework, we reformulate fine-grained sector localization and orientation inference as a discrete sector retrieval task. This allows the model to explicitly retrieve fine-grained sector-level sub-regions and implicitly estimate orientation, without relying on manually annotated pose labels or instance-level supervision. Furthermore, we establish a new sector-level benchmarking protocol equipped with ground-truth soft labeling. Extensive experiments demonstrate that SectorGeo significantly outperforms state-of-the-art methods while maintaining strong robustness across varying FoVs.

Acknowledgements

This work was supported in part Guangdong Basic and Applied Basic Research Foundation under Grant 2026A1515011137, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, in part by Shenzhen Science and Technology Program under Grant JCYJ20240813142510014 and Grant 20220810142553001, and in part by the National Natural Science Foundation of China under Grant U22A2097.

References

1. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16847–16856 (2023)
2. Devagiri, J.S., Paheding, S., Niyaz, Q., Yang, X., Smith, S.: Augmented reality and artificial intelligence in industry: Trends, tools, and future challenges. Expert Systems with Applications **207**, 118002 (2022)
3. Dudley, J., Yin, L., Garaj, V., Kristensson, P.O.: Inclusive immersion: a review of efforts to improve accessibility in virtual reality, augmented reality and the metaverse. Virtual Reality **27**(4), 2989–3020 (2023)
4. Elhashash, M., Qin, R.: Cross-view slam solver: Global pose estimation of monocular ground-level video frames for 3d reconstruction using a reference 3d model from satellite images. ISPRS Journal of Photogrammetry and Remote Sensing **188**, 62–74 (2022)
5. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Statewide visual geolocalization in the wild. In: European Conference on Computer Vision. pp. 438–455. Springer (2024)
6. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. arXiv preprint arXiv:1803.07728 (2018)
7. Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M.: Soft exemplar highlighting for cross-view image-based geo-localization. IEEE transactions on image processing **31**, 2094–2105 (2022)
8. Hu, S., Li, Y., Li, Y., Zhu, Y.: Seeing the unseen: Mask-driven positional encoding and strip-convolution context modeling for cross-view object geo-localization. arXiv preprint arXiv:2510.20247 (2025)
9. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.: 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)
10. Li, G., Qian, M., Xia, G.S.: Unleashing unlabeled data: A paradigm for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16719–16729 (2024)
11. Li, Y., Zhu, Y.: Plgeo: A patch-level framework to overcome orientation discrepancies in cross-view geo-localization. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6057–6065 (2025)
12. Li, Y., Zhang, Q., Zhu, Y.: Rethinking cross-view object geo-localization: Towards many-to-many real-world localization. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2025)
13. Lindenberger, P., Sarlin, P.E., Hosang, J., Pollefeys, M., Lynen, S., Trulls, E.: Scaling image geo-localization to continent level. Advances in Neural Information Processing Systems **38**, 134057–134092 (2026)
14. Liu, L., Li, H.: 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)
15. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
16. Liu, Z., Cao, Y., Chen, J., Li, J.: A hierarchical reinforcement learning algorithm based on attention mechanism for uav autonomous navigation. IEEE Transactions on Intelligent Transportation Systems **24**(11), 13309–13320 (2022)

17. Mi, L., Xu, C., Castillo-Navarro, J., Montariol, S., Yang, W., Bosselut, A., Tuia, D.: Congeo: Robust cross-view geo-localization across ground view variations. In: European Conference on Computer Vision. pp. 214–230. Springer (2024)
18. Nasar, W., Da Silva Torres, R., Gundersen, O.E., Karlsen, A.T.: The use of decision support in search and rescue: A systematic literature review. ISPRS International Journal of Geo-Information **12**(5), 182 (2023)
19. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
20. Ouyang, S., Zhu, Y.: Cvgl: Causal learning and geometric topology. Advances in Neural Information Processing Systems **38**, 78247–78266 (2026)
21. Pore, A., Li, Z., Dall'Alba, D., Hernansanz, A., De Momi, E., Menciassi, A., Gelpi, A.C., Dankelman, J., Fiorini, P., Vander Poorten, E.: Autonomous navigation for robot-assisted intraluminal and endovascular procedures: A systematic review. IEEE Transactions on Robotics **39**(4), 2529–2548 (2023)
22. Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H.: Accurate 3-dof camera geo-localization via ground-to-satellite image matching. IEEE transactions on pattern analysis and machine intelligence **45**(3), 2682–2697 (2022)
23. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11990–11997 (2020)
24. Sun, Y., Ye, Y., Kang, J., Fernandez-Beltran, R., Feng, S., Li, X., Luo, C., Zhang, P., Plaza, A.: Cross-view object geo-localization in a local region with satellite imagery. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–16 (2023)
25. Tulsiani, S., Malik, J.: Viewpoints and keypoints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1510–1519 (2015)
26. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2021)
27. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3961–3969 (2015)
28. Wu, L., Bo, L., Ouyang, S., Zhu, Y.: Mrgeo: Robust cross-view geo-localization of corrupted images via spatial and channel feature enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10691–10699 (2026)
29. Xia, P., Wan, Y., Zheng, Z., Zhang, Y., Deng, J.: Enhancing cross-view geo-localization with domain alignment and scene consistency. IEEE Transactions on Circuits and Systems for Video Technology **34**(12), 13271–13281 (2024)
30. Xia, Z., Booij, O., Kooij, J.F.: Convolutional cross-view pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 3813–3831 (2023)
31. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.: Visual cross-view metric localization with dense uncertainty estimates. In: European Conference on Computer Vision. pp. 90–106. Springer (2022)
32. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems **34**, 29009–29020 (2021)
33. Ye, J., Lin, H., Ou, L., Chen, D., Wang, Z., Zhu, Q., He, C., Li, W.: Where am i? cross-view geo-localization with natural language descriptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5890–5900 (2025)

34. Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., He, C.: Cross-view image geo-localization with panorama-bev co-retrieval network. In: European Conference on Computer Vision. pp. 74–90. Springer (2024)
35. Zhang, Q., Zhu, Y.: Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7251–7259 (2024)
36. Zhang, Q., Zhu, Y.: Benchmarking the robustness of cross-view geo-localization models. In: European Conference on Computer Vision. pp. 36–53. Springer (2024)
37. Zhang, Q., Zhu, Y.: Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8197–8206 (2025)
38. Zhang, X., Cao, S.Y., Yu, Z., Wu, Z., Zhang, X., Bai, X., Yu, B., Shen, H.L.: Geoformer: Boosting object distinguishing and prompt understanding for cross-view object geo-localization. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025)
39. Zhang, X., Li, X., Sultani, W., Chen, C., Wshah, S.: Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
40. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 3480–3488 (2023)
41. Zhang, Z., Sun, Y., Ding, L., Zhang, H., Wang, Y., Ding, W.: Halo: Hierarchical adaptive feature learning and cross-view interaction for object-level geo-localization. IEEE Transactions on Geoscience and Remote Sensing (2025)
42. Zhu, S., Shah, M., Chen, C.: 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)
43. Zhu, S., Yang, T., Chen, C.: 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)