Beyond 2D Matching: A Unified Single-Stage Framework for Geometry-Aware Cross-View Object Geo-Localization

Liyao Wang[1*], Ruipu Wu[1*], Haojun Xu[1*], Lei Shi[2]
Linjiang Huang[1†], and Si Liu[1]

[1] Beihang University, Beijing, China
[2] Meituan, Beijing, China
{bastien_wu,xuhaojun123,ljhuang,liusi}@buaa.edu.cn, shilei53@meituan.com
https://cipual.github.io/GAGeo-project-page/

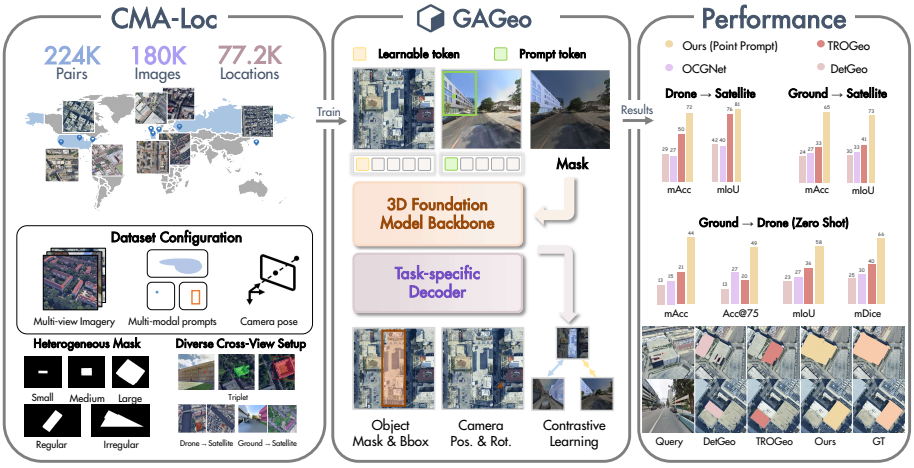

Fig. 1: Overview of CMA-Loc dataset and GAGeo framework. We introduce CMA-Loc, a large-scale building dataset for advancing cross-view geo-localization, featuring 224K instance pairs and 180K images across 77.2K locations. It incorporates multi-view imagery, multi-modal prompts, and camera pose to simulate complex real-world scenarios. Leveraging CMA-Loc, we propose GAGeo, a geometry-aware cross-view object geo-localization framework that adapts a 3D Foundation Model backbone with task-specific decoders in a single-stage manner, achieving state-of-the-art performance in drone-to-satellite, ground-to-satellite, and zero-shot ground-to-drone setups.

Abstract. Cross-view object geo-localization (CVOGL) aims to locate a target object from a query view (e.g., ground or drone) within a

[*] Equal contribution.
[†] Corresponding author.

geo-tagged reference image (e.g., satellite). Existing approaches heavily rely on 2D appearance matching and are constrained by limited datasets lacking geometric metadata, diverse prompts, and standard field-of-view imagery. To address these intertwined challenges, we first introduce CMA-Loc, a large-scale, high-fidelity building dataset comprising over 220,000 ground-satellite and drone-satellite pairs. It provides multi-modal prompts (points, boxes, masks) and camera poses to enable flexible target referring and explicit spatial modeling. Furthermore, we propose a novel single-stage Geometry-Aware Geo-localization framework (GAGeo), built upon the permutation-equivariant 3D foundation model π^3. By seamlessly integrating visual features, referring prompts, and learnable task tokens, our model adapts the inherited 3D prior to jointly predict bounding boxes, segmentation masks, and camera poses in a single forward pass. Additionally, we introduce a contrastive loss that utilizes the satellite view as a universal anchor, implicitly aligning ground and drone representations to enable zero-shot ground-to-drone localization without requiring triplet training data. Extensive experiments demonstrate that our approach significantly outperforms state-of-the-art methods, exhibiting exceptional generalization ability in unseen scenes and novel cross-view setups.

Keywords: Cross-view object geo-localization · Geometry-aware detection and Segmentation · Zero-shot generalization

1 Introduction

In real-world applications ranging from disaster monitoring [12, 39] to drone navigation [26], there is a growing need to ground specific target objects onto geo-tagged reference imagery (e.g., satellite views), thereby decoupling spatial positioning from reliance on external signals. Inherently, this task requires building a spatial bridge that projects local, ego-centric observations onto a global coordinate system, seamlessly shifting the paradigm from recognizing "what" is in the scene to precisely localizing "where" it resides in the physical world.

Driven by this practical need, researchers have curated various cross-view datasets, evolving from coarse-grained image-level retrieval [5, 24, 38, 44, 46] to recent object-centric datasets [32, 40]. However, current research is severely hindered by coupled limitations in both existing datasets and methodologies. On the data side, existing resources rely exclusively on panoramic images that introduce severe geometric distortions, failing to reflect the standard field-of-view (FoV) of practical cameras. Furthermore, they are restricted to simple point-coordinate prompts, lack geometric metadata (e.g., camera positions and rotations), and suffer from limited scale (merely 12,000 pairs for CVOGL [32]) and scene variability. Beyond these data deficiencies, prevailing research conventionally treats CVOGL as a 2D matching problem. Consequently, most existing frameworks [6, 14, 29, 30, 32, 36, 40] rely on appearance-based cues, failing to fully exploit the inherent 3D geometric structure. Lacking such 3D awareness, these

models struggle to bridge the spatial gap, leaving them vulnerable to perceptual ambiguity and visual inconsistencies, which limit their robust generalization.

To tackle these intertwined challenges, we propose a comprehensive framework comprising a high-fidelity building dataset, **CMA-Loc**, and a novel Geometry-Aware GEO-localization approach **GAGeo**. First, to bridge the data gap, CMA-Loc is constructed via two tailored pipelines. To construct ground-satellite pairs, we collect imagery from Google Street View [1] and Google Earth [10]. We then associate OpenStreetMap-derived [11] satellite footprints with SAM3-generated [2] ground masks based on geometric cues like orientation and scale. For drone-satellite pairs, we utilize a Cesium rendering pipeline to deterministically extract and pair cross-view masks via unique building IDs. Finally, to improve boundary quality across the dataset, all non-SAM-generated masks are refined using SAM3. From these refined masks, we derive bounding boxes and inner point prompts as additional referring prompts, followed by rule-based filtering to reduce low-quality or information-sparse cases. Crucially, CMA-Loc also provides accurate camera poses for each pair, unlocking the capability for explicit spatial modeling and geometric supervision. Through this rigorous construction pipeline, we ultimately curate a large-scale dataset comprising $112,063$ ground-satellite and $111,704$ drone-satellite instance pairs across eight global cities, resolving the data scale and scene diversity constraints of prior works. Additionally, in CMA-Loc's test set, we incorporate a manually annotated subset, which comprises ground-drone-satellite triplets from University-1652 [44], serving as a dedicated dataset to evaluate models' zero-shot and cross-domain generalization.

To overcome the limitations of pure 2D appearance matching, our methodology shifts the paradigm by harnessing the powerful geometric priors embedded in 3D foundation models [25, 33, 43]. Specifically, we adopt π^3 [35] as our framework's backbone. A primary motivation for this choice is that π^3 processes input views in a permutation-equivariant manner; it establishes a shared representation space without anchoring to any initial frame's perspective, thereby reducing viewpoint biases. Leveraging this property, we propose a unified single-stage framework capable of processing both ground-to-satellite and drone-to-satellite inputs. The model first fuses visual tokens from a frozen DINOv2 [27] with high-fidelity prompt tokens from a SAM2-pretrained encoder. Unlike previous paradigms that rely on cumbersome two-stage networks for detection and segmentation (e.g., TROGeo [40]) or append heavy, isolated prediction heads to a backbone, our design is remarkably streamlined. We seamlessly integrate learnable tokens and prompt tokens as additional inputs for π^3, augmenting its alternating local and global attention layers with a specialized masking scheme designed for multi-task decoding. Through a single forward pass, this single-stage structure yields diverse multi-task predictions, including bounding boxes, segmentation masks, camera position and rotation.

Furthermore, since collecting perfectly aligned ground-drone-satellite triplets in the real world is notoriously difficult, we introduce a contrastive learning objective optimized purely on available pairs. Conceptually akin to ImageBind [9], which binds diverse modalities to a shared latent space using images as the

central anchor, we utilize the satellite view as a universal intermediate bridge. Benefiting from the view-agnostic nature of π^3, aligning mask-pooled object representations within ground-satellite and drone-satellite pairs implicitly pulls the ground and drone representations closer together. This strategy circumvents the reliance on scarce triplet training data, unlocking the capability for zero-shot ground-to-drone geo-localization.

Extensive experiments demonstrate that our framework sets a new state-of-the-art in the cross-view object geo-localization task, significantly outperforming existing approaches. Notably, in the ground-to-satellite setup, our method achieves a significant improvement of 34.38% mAcc on the object detection task and 33.36% mIoU on the object segmentation task. Meanwhile, it also exhibits robust generalization to unseen dataset and novel ground-to-drone setup, evidently outperforming previous best-performing methods. Furthermore, comprehensive ablation studies validate the effectiveness of each proposed strategy.

2 Related Works

Cross-View Image Geo-Localization. Early cross-view geo-localization formulated the task as an image retrieval problem [23,42]. Despite the availability of diverse multi-view datasets [5,24,38,44,46], existing methods predominantly rely on holistic, 2D appearance-based metric learning [14,29,36]. To address extreme viewpoint discrepancies, subsequent works incorporated geometric transformations [16,30] and spatial partitioning [4,34]. However, these approaches remain constrained by rigid geometric priors and lack a robust 3D prior, fundamentally limiting their spatial perception in complex environments.

Cross-View Object Geo-Localization. Shifting from image-level retrieval, cross-view object geo-localization (CVOGL) targets specific objects using point prompts. Current methods typically frame this as cross-view detection or segmentation [32, 40, 47], relying on complex multi-stage architectures [40, 41] or perspective-specific algorithms [15, 21]. Fundamentally confined to 2D feature alignment, these methods struggle with extreme viewpoint variations and repetitive visual patterns. In contrast, our work proposes a unified framework that processes multi-modal prompts and generates multi-modal outputs in a single forward pass, overcoming the structural bottlenecks of pure 2D alignment.

Bridging Extreme Views via 3D Geometric Foundation Models. Framing cross-view grounding as a 3D problem is highly effective, as corresponding pixels across perspectives inherently project to the same 3D point [19]. Recently, feed-forward 3D geometric foundation models [25, 33, 35] have demonstrated an emergent understanding of extreme-view geometry [43], providing a robust scaffold for complex scenarios. Building on this observation, we introduce π^3 [35] to the CVOGL task. This novel geometry-aware paradigm distills intrinsic 3D topologies that remain invariant across heterogeneous perspectives, fundamentally resolving extreme view discrepancies and visual ambiguities.

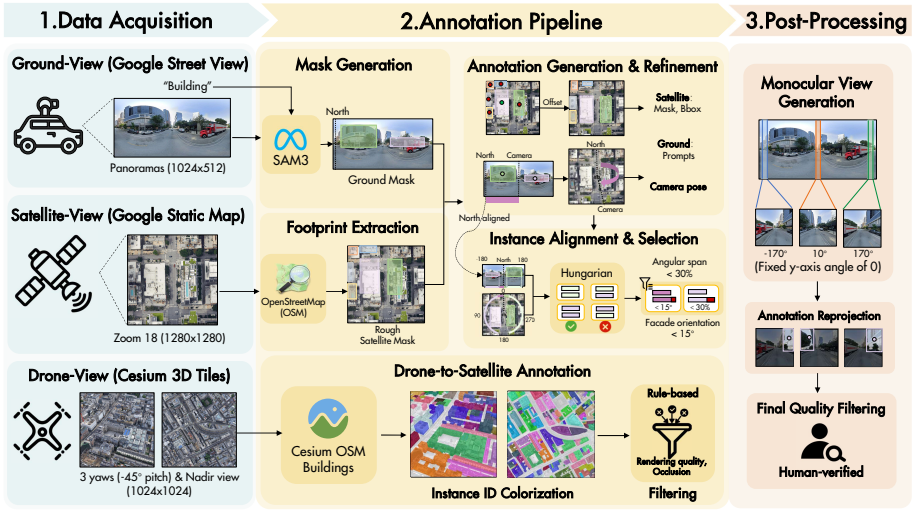

Fig. 2: Illustration of the pipeline construction process for CMA-Loc. We present two specialized workflows for generating **ground-to-satellite** and **drone-to-satellite** instance pairs for cross-view geo-localization. The process integrates Google Street View panoramas with Cesium-synthesized drone views and nadir satellite imagery, ensuring precise geometric alignment across diverse cross-view perspectives.

3 Dataset

In this section, we introduce **CMA-Loc** (**C**ross-view **M**ulti-prompt **A**nnotated **Loc**alization dataset), a building-focused dataset designed to advance cross-view geo-localization by providing large-scale, multi-prompt geometric supervision. Spanning 77,200 geographic locations across 8 diverse cities (e.g., Tokyo, London, etc.), the dataset provides 112,063 ground-satellite and 111,704 drone-satellite instance pairs. By spanning diverse architectural styles and cultural environments, CMA-Loc significantly exceeds the scale of existing datasets. The following sections detail the data collection process, the annotation pipeline, and dataset statistics.

3.1 Data Collection

The data collection process of CMA-Loc follows a query-reference paradigm, matching ground- and drone-view query images against universal satellite reference images via the pipeline shown in Fig. 2.

Ground-to-Satellite Image Collection. For the ground-view queries, following the protocol established by Omnicity [20], we collect Google Street View panoramas (1024×512) sampled at 65m intervals. To ensure geometric alignment, we record the GPS coordinates, panorama IDs, and north-aligned headings for each site. Corresponding satellite imagery is sourced from Google Earth at zoom level 18, providing 1280×1280 reference images.

Drone-to-Satellite Image Synthesis. Drone-to-satellite data is synthesized via a Cesium-based rendering pipeline, leveraging Google Photorealistic 3D Tiles for urban reconstruction. We sample these environments at a 0.005 spatial grid interval. At each location, we simulate three drone views with a fixed -45° pitch and yaw angles spaced at 120° intervals with a random initial offset. Corresponding satellite references are generated via nadir orthographic rendering (-90° pitch) at the same coordinates. All images are rendered at 1024×1024 resolution, with camera Euler angles and grid coordinates stored as metadata.
Ground-Drone-Satellite Image Triplet Preparation. To validate models' zero-shot geo-localization capability in unseen cross-view setups and environments, we carefully select 1245 triplets from the University-1652 dataset [44] to prepare raw ground-drone-satellite image triplets.

3.2 Data Annotation

Fig. 2 depicts our automated annotation pipeline, which generates multi-modal prompts and geometric supervision without prohibitive manual labor.
Ground-to-Satellite Annotation. The ground-to-satellite annotation pipeline proceeds in three streamlined stages as follows:

1) Mask Generation and Refinement. We first extract initial building masks from ground panoramas using SAM3 [2], filtering them by object size and occlusion. For satellite imagery, we generate high-fidelity masks by prompting SAM with OpenStreetMap (OSM) [11] footprints: using each footprint's center as a positive prompt and the five nearest neighbors as negative prompts. After correcting inherent OSM-SAM misalignments via global offset estimation, the masks are filtered by area, confidence, and IoU to remove unreliable or low-information cases. Finally, bboxes and interior points are extracted from these refined masks to serve as multi-modal prompts.

2) Geometric Instance Alignment. Next, we establish cross-view correspondences using spatial metadata. By leveraging north-aligned rotation angles, we compute each building's facade orientation and angular span. A ground-satellite pair is retained only if the facade orientation difference is less than 15° and the angular span discrepancy is within 30%, which filters cases where occlusion or limited visible facade length makes the correspondence ambiguous. Additionally, camera pixel coordinates and north-relative orientations are explicitly recorded in the satellite images to provide rich pose annotations.

3) Monocular View Synthesis. Finally, to align with real-world deployment, we transform panoramas into standard monocular views.
Drone-to-Satellite Annotation. For the synthesized drone-to-satellite data, the annotation process is deterministic. We leverage Cesium OSM Buildings to generate instance masks via unique RGB encoding, ensuring exact cross-view correspondence and avoiding heuristic matching. Finally, to ensure dataset quality, image pairs are filtered to remove poor renderings, invalid masks, or significant occlusions (e.g., dense trees).
Ground-Drone-Satellite Triplet Annotation. We construct a specialized evaluation subset by annotating the ground-drone-satellite triplets. Concretely,

Table 1: Comparison of cross-view localization datasets. G, D, S, Pt. denote ground, drone, satellite, and point, respectively. Mono. Ground View refers to standard FOV.

	CVUSA	VIGOR	Univ-1652	CVOGL	CVOGL-Seg	**CMA-Loc**
# Instance Pairs	–	–	–	12,478	12,478	223,767
# Satellite Images	44,416	105,214	951	5,836	5,836	77,200
# Ground Images	44,416	90,618	2,921	5,279	5,279	93,144
# Drone Images	–	–	51,355	5,279	5,279	10,086
Cross-View Setup	G→S	G→S	G ⇵ D↔S	G/D→S	G/D→S	G/D→S
Annotation Grain	Image	Image	Image	Object	Mask	Mask
Prompt Type	–	–	–	Point	Point	Pt. / Bbox / Mask
Mono. Ground View	○	○	●	○	○	●
Orientation Info	●	●	○	○	○	●
Triplet Eval	○	○	●	○	○	◐

Initial point and bounding box prompts are first generated using Gemini [8], which are then subjected to rigorous manual verification to ensure their fidelity. Finally, high-fidelity segmentation masks are produced using SAM3 [2], utilizing the verified bounding boxes as prompts.

3.3 Dataset Statistics

Both the drone-to-satellite and ground-to-satellite tasks localize targets via cross-view correspondence, inherently sharing a unified problem formulation. Unlike CVOGL, which addresses these tasks in isolation, we introduce a single, comprehensive dataset to investigate both paradigms jointly. As summarized in Tab. 1, CMA-Loc offers several distinct advantages over existing datasets:

First, it features a much larger data scale, being approximately 18× larger than previous datasets. Second, it supports rich, multi-modal prompts, including points, bounding boxes, and masks, overcoming the limitations of datasets like CVOGL [32] that rely solely on sparse point prompts. Third, it provides geometric annotations, including precise observer positions and camera poses, delivering the explicit supervision necessary for spatial understanding.

Beyond scale and annotation richness, CMA-Loc spans 8 globally distributed cities with highly distinct architectural styles. This geographic diversity introduces substantial variations in object scale, shape, and occlusion, thereby significantly elevating the challenge of robust detection and segmentation. Furthermore, to better reflect real-world deployment conditions, CMA-Loc discards impractical panoramic ground images in favor of standard monocular views, enhancing the benchmark's practical applicability.

Finally, as detailed in Sec. 3.2, the evaluation protocol of CMA-Loc goes beyond standard in-domain testing. Its evaluation set contains not only ground-satellite and drone-satellite pairs aligned with the training distribution, but also

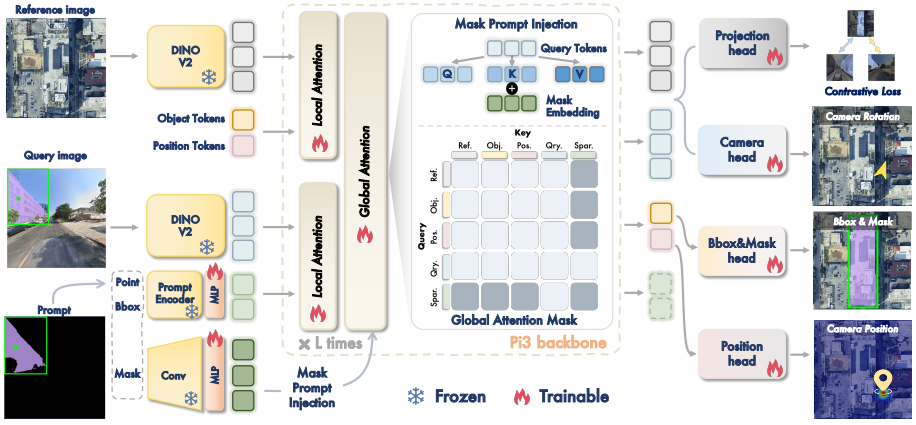

Fig. 3: Overview of GAGeo, which integrates multi-modal geometric prompts and task-specific tokens into a unified, single-stage, multi-task transformer framework.

perfectly aligned ground-drone-satellite triplets meticulously curated to validate the models' cross-view generalization capabilities.

4 Method

In this section, we first formally define the task formulation for cross-view object grounding and localization. Then, Section 4.1 briefly reviews the preliminary architecture of contemporary 3D Foundation Models (3DFMs). Building upon this foundation, Section 4.2 details our proposed **G**eometry-**A**ware Cross-View **Geo**-Localization framework (**GAGeo**), which seamlessly integrates multi-modal geometric prompts and task-specific tokens into a unified, single-stage transformer framework. Finally, Sec. 4.3 presents the joint multi-task optimization objectives employed to train the entire framework end-to-end.

Task Formulation. Given a query image I_q (e.g. from ground or drone) and a reference image I_r (e.g. from satellite), our objective is to learn a mapping function $\mathcal{F}(I_q, I_r, p) \rightarrow y_r$ that identifies and locates the target object in I_r based on a spatial referring prompt p defined in I_q. In our framework, the referring prompt p is a multi-modal geometric prompt that can be represented as a point $\mathbf{p}_q = (x, y)$, a bounding box $\mathbf{b}_q = (x, y, w, h)$, or a binary mask $\mathbf{m}_q \in \{0, 1\}^{H \times W}$ that indicate the exact location of the target object in I_q. The model processes the cross-view visual context to output the target's state y_r in the reference domain, which is manifested as a predicted bounding box $\hat{\mathbf{b}}_r$ for object-level localization or a predicted mask $\hat{\mathbf{m}}_r$ for pixel-level segmentation.

4.1 Preliminary

3DFM Architecture. Contemporary 3DFMs [25, 33, 35] typically share a unified architectural framework. An encoder ε first maps input images into patch-

level embeddings. Specifically, given two input images, I_1 and I_2, each divided into $N_p = H_p \times W_p$ patches, the tokens are processed by a shared transformer backbone that interleaves *local* and *global* attention blocks. Local attention blocks process each view independently: $\mathbf{T}^{(i)}_{\text{frame}} \in \mathbb{R}^{N_p \times D}$ where $i \in \{1, 2\}$. Conversely, global attention blocks facilitate cross-view interaction by aggregating tokens from both images via concatenation:

$$\mathbf{T}_{\text{global}} = [\mathbf{T}^{(1)}_{\text{frame}}, \mathbf{T}^{(2)}_{\text{frame}}] \in \mathbb{R}^{2N_p \times D}. \tag{1}$$

Both variants employ standard self-attention. For layer l and head h, queries $\mathbf{Q} = f_Q(\mathbf{T}_*)$, keys $\mathbf{K} = f_K(\mathbf{T}_*)$, and values $\mathbf{V} = f_V(\mathbf{T}_*)$ are derived via learned linear projections, yielding the attention weights:

$$\mathbf{A}^{(l)}_h = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_h}), \tag{2}$$

where d_h is the head dimensionality. Finally, the processed tokens are fed into task-specific decoding heads, such as a camera head for pose estimation and a position head for localization.

4.2 Geometry-Aware CVOGL Pipeline

Extending the π^3 framework [35], our pipeline augments the input sequence with task-specific tokens and referring prompt tokens tailored for CVOGL, rather than treating the query and reference images as standard stereo pairs. Critically, while conventional 3DFMs anchor their spatial representations to the first frame, our approach leverages the permutation-equivariant design of π^3 [35]. By deliberately discarding order-dependent components, such as frame-wise positional embeddings and camera tokens, the model reduces view-order inductive biases. This inherited permutation equivariance is important for generalization across unseen cross-view setups (e.g., ground-to-drone). More importantly, it enables the model to support joint training across diverse cross-view setups within a unified framework, effectively leveraging heterogeneous cross-view pairs.

Image and Prompt Encoding. We employ a frozen DINOv2 encoder to extract patch-level tokens $T_r, T_q \in \mathbb{R}^{N_p \times D}$ from both views. Concurrently, referring prompts are processed using a pretrained SAM2 module, which generates two types of representations: a prompt encoder produces sparse tokens $T_p \in \mathbb{R}^{N_e \times D}$ ($N_e = 1$ for points and $N_e = 2$ for box corners), while a CNN encoder extracts dense embeddings $E_d \in \mathbb{R}^{N_p \times D}$ for mask prompts.

Prompt Injection and Cross-View Interaction. During this stage, we leverage the inherent advantages of the π^3 backbone. Specifically, its interleaved local and global attention mechanisms provide geometry-aware representation capabilities and cross-view information exchange. This architectural design naturally allows us to seamlessly integrate task-specific learnable tokens and referring prompts directly into the sequence processing. Consequently, our framework forms a fully transformer-based single-stage paradigm [7,31], unifying feature extraction, cross-view interaction, and task prediction within a single forward pass.

Compared to previous cross-view methods that rely on cumbersome two-stage networks for detection and segmentation (e.g., TROGeo [40]) or append isolated prediction decoders to the backbone, our streamlined approach is simpler and more efficient.

Concretely, we employ an asymmetric token injection strategy tailored for our task formulation. Since the referring prompt is defined on the query image, while the target predictions (e.g., detection, segmentation) are executed on the reference image, we decouple their token assignments. Specifically, for the reference stream, we introduce learnable task tokens $T_l \in \mathbb{R}^{(N_{obj}+N_{pos})\times D}$, comprising object tokens T_{obj} and position tokens T_{pos}, which are concatenated with the reference tokens to yield $\overline{T}_r = [T_r, T_l]$ for generating the task results.

For the query stream, prompt integration employs a modality-dependent strategy. Sparse prompt tokens T_p are directly concatenated with the query tokens to form $\overline{T}_q = [T_q, T_p]$. In contrast, for dense mask embeddings E_d, we diverge from the standard SAM-style approach [28], which adds these embeddings directly to the input image tokens. To ensure better generalization and preserve the symmetric input structure expected by the π^3 backbone, we utilize E_d strictly as a modulation signal within the global attention blocks. Specifically, during global attention, E_d is added element-wise exclusively to the *keys* ($\mathbf{K}$) of the query image tokens. This key-side modulation avoids directly altering the layer's output feature distribution; instead, it refines the attention weights, effectively guiding the reference tokens to aggregate features from the mask-related regions of the query image during global interaction.

Within global attention blocks, we regulate cross-view interaction via a directional isolation mask. Specifically, sparse prompt tokens T_p can attend to query tokens T_q to absorb spatial priors, but are strictly isolated from the reference stream (T_r and T_l), and vice versa. This isolation is essential: since geometric prompts are inherently query-view specific, exposing them to reference tokens provides no contextual benefit and risks corrupting reference features.

Through this stage, we can obtain the updated tokens $\hat{T}_q$, $\hat{T}_r$, $\hat{T}_l$, corresponding to the query features, the reference features and the task tokens, respectively.

Lightweight Task-Specific Decoding. The output task tokens $\hat{T}_l$ are partitioned into object tokens $\hat{T}_{obj}$ and position tokens $\hat{T}_{pos}$ for lightweight decoding. For object prediction, $\hat{T}_{obj}$ is processed by two parallel heads: a DETR-style [3] 3-layer MLP $\mathcal{D}_{det}$ for bbox regression, and a SAM-style [17] hyper-network $\mathcal{D}_{seg}$ that dynamically generates convolution kernels from $\hat{T}_{obj}$, which are applied to bilinearly upsampled reference features $\hat{T}_r^{\uparrow}$ for mask prediction. For the estimation of the observer's position (camera position) in the query image, a position head $\mathcal{D}_{pos}$ computes a spatial heatmap H via a dot product between $\hat{T}_{pos}$ and upsampled reference features $\hat{T}_r^{\uparrow}$. Finally, following π^3 [35], we use a camera head $\mathcal{D}_{cam}$ to process the updated tokens $\hat{T}_q$ and $\hat{T}_r$ to estimate the orientation of the query camera relative to the reference coordinate frame.

4.3 Optimization Objective

The proposed framework is trained end-to-end using a joint multi-task objective. To ensure robust convergence and training stability, we apply deep supervision across intermediate decoder layers $\mathcal{K} = \{4, 11, 17\}$. The overall loss function is formulated as follows, where w_k is the weight for the k-th layer:

$$\mathcal{L}_{total} = \sum_{k \in \mathcal{K}} w_k \left(\mathcal{L}_{grd}^{(k)} + \mathcal{L}_{pos}^{(k)} + \mathcal{L}_{rot}^{(k)} \right) + \mathcal{L}_{cl}. \tag{3}$$

For more details, please refer to the supplementary material.

Grounding Loss ($\mathcal{L}_{grd}$). We employ Hungarian matching to assign the optimal prediction to the ground truth for object tokens T_{obj}. The matched token is supervised via:

$$\mathcal{L}_{grd} = \lambda_{cls}\mathcal{L}_{focal} + \lambda_{box}(\mathcal{L}_{L1} + \mathcal{L}_{giou}) + \lambda_{mask}(\mathcal{L}_{bce} + \mathcal{L}_{dice}), \tag{4}$$

which combines a binary focal loss for classification, L_1 and GIoU losses for bounding box regression, and BCE and Dice losses for mask generation.

Camera Pose Loss ($\mathcal{L}_{pos} + \mathcal{L}_{rot}$). The query camera's pixel position ($\mathcal{L}_{pos}$) is supervised using a focal loss variant [18] between the predicted heatmap and a Gaussian-augmented ground truth heatmap. For camera's orientation ($\mathcal{L}_{rot}$), we minimize the geodesic error on the SO(3) manifold:

$$\mathcal{L}_{rot} = \arccos\left(tr(\hat{R}^T R_{gt})/2 - 0.5 \right), \tag{5}$$

where $\hat{R}$ and R_{gt} denote the predicted and GT rotation matrices, respectively.

Contrastive Learning Loss ($\mathcal{L}_{cl}$). To bridge the cross-view representation gap, we utilize a contrastive loss [13] to align mask-pooled object features, using the satellite view as an anchor. We first project $\hat{T}_q$ and $\hat{T}_r$, then average pool them via the target object masks to obtain z_r and z_q. The paired query embedding is treated as the positive z_q^+, while other query embeddings in the MoCo queue form negatives z_q^-:

$$\mathcal{L}_{cl} = -\log \frac{\exp(z_r \cdot z_q^+ / \tau)}{\exp(z_r \cdot z_q^+ / \tau) + \sum_{z_q^- \in Q} \exp(z_r \cdot z_q^- / \tau)}, \tag{6}$$

where z_r is the reference-view anchor, Q denotes the negative sample queue, and $\tau = 0.07$.

5 Experiments

To demonstrate the superiority of GAGeo, we first detail our experimental setup in Sec. 5.1. Then, we present a comprehensive comparison with state-of-the-art approaches across multiple CVOGL tasks in Sec. 5.2 along with detailed ablation studies in Sec. 5.3. Finally, we present qualitative comparison results in Sec. 5.4.

5.1 Experimental Setup

Datasets. Departing from isolated training protocols, *all trainable compared methods are jointly trained across ground/drone-to-satellite setups in CMA-Loc* following their original implementations with dataset-specific anchor or scale adaptation. For evaluation, we use the CMA-Loc test set, containing 4,197 seen and 4,220 unseen instance pairs. Zero-shot ground-to-drone generalization is assessed using 1,245 annotated ground-drone-satellite triplets (see Sec. 3.2). Furthermore, we evaluate out-of-distribution scene generalization for object detection and segmentation on CVOGL [32] and CVOGL-Seg [40] across ground/drone-to-satellite setups.
Evaluation Metrics. We use Acc@K ($K \in \{75\%, 50\%\}$) and mAcc (the average accuracy over IoU thresholds from 0.5 to 0.95 with a 0.05 interval) to evaluate the cross-view object detection task. For the cross-view object segmentation task, we use mIoU, mDice, AAE, and ME (see the supplementary material for details).
Implementation Details. During training, the DINOv2 encoder and the SAM prompt encoder are kept frozen. Meanwhile, the π^3 backbone, together with the camera head, are initialized from pretrained weights. All other components are trained from scratch. (see further details in the supplementary material).

5.2 Main Comparisons

As shown in Table 2 and 3, GAGeoconsistently surpasses existing SOTA methods across all metrics on the CMA-Loc test set, irrespective of the prompt modality. Notably, this superiority is achieved as *all trainable compared methods are fully trained on our dataset*. Furthermore, our approach effectively bridges the performance gap between seen and unseen environments. Remarkably, GAGeoyields higher segmentation accuracy in a single forward pass than previous methods relying on an additional SAM Prompt Stage (SPS) [40], although its inference cost is still dominated by the 3DFM backbone. This robust performance extends to the zero-shot setting on CVOGL-Seg (Tab. 4), where our method nearly doubles the mAcc and mIoU of the prior SOTA without any domain-specific fine-tuning, thereby highlighting its exceptional transferability. Finally, as shown in Tab. 5, GAGeoalso outperforms the SPS-enhanced baseline in the unseen ground-to-drone setup, validating its generalization to novel viewpoint setup.

5.3 Ablation Study

To validate the design choices of GAGeo, we conduct detailed ablation studies on CMA-Loc seen test sets, with all experiments evaluated using point prompts.
Loss ablations. We evaluate the loss components by treating the grounding loss as the baseline, and incrementally incorporating deep supervision, contrastive learning, and camera pose loss to quantify their cumulative contributions, as shown in Tab. 6.

Table 2: Comparison on cross-view object detection task. The best and second-best results are marked in **bold** and underline, respectively.

Method	Drone → Satellite						Ground → Satellite					
	Seen			Unseen			Seen			Unseen		
	mAcc↑	Acc@75↑	Acc@50↑	mAcc↑	Acc@75↑	Acc@50↑	mAcc↑	Acc@75↑	Acc@50↑	mAcc↑	Acc@75↑	Acc@50↑
RK-Net [22]	7.47	1.97	27.66	7.02	1.44	28.36	0.11	0.05	0.44	0.15	0.00	0.73
L2LTR [37]	5.58	1.42	20.47	4.07	0.89	16.66	0.20	0.05	0.93	0.19	0.00	1.15
TransGeo [45]	8.71	2.29	32.05	7.15	1.44	29.35	0.33	0.10	1.77	0.13	0.00	0.78
SAFA [30]	6.90	1.69	25.82	6.03	1.19	24.44	0.40	0.10	1.62	0.36	0.05	1.70
Sample4Geo [6]	8.80	2.29	32.55	7.50	1.54	30.19	1.23	0.25	5.75	0.50	0.14	2.11
DetGeo [32]	31.18	33.15	54.08	28.64	31.63	46.65	46.04	55.21	59.23	24.13	29.04	32.20
OCGNet [15]	29.51	32.28	49.18	27.19	29.40	43.43	47.08	56.14	60.76	26.77	32.52	36.19
TROGeo [40]	46.86	48.67	81.27	50.30	51.86	84.63	51.58	60.76	69.50	32.66	39.22	45.46
Ours (Point)	68.48	77.01	92.86	71.74	82.80	93.95	**71.50**	80.70	89.69	65.30	72.75	81.28
Ours (Bbox)	74.47	85.90	97.21	77.03	89.19	97.62	**71.50**	80.94	**89.78**	66.70	**74.72**	82.48
Ours (Mask)	**76.00**	**88.32**	**97.71**	**79.16**	**91.77**	**98.26**	71.41	**81.04**	89.39	**67.04**	**74.72**	**82.94**

Table 3: Comparison on cross-view object segmentation task. "+ SPS" denotes combination with SAM Prompt Stage [40]. The best and second-best results are marked in **bold** and underline, respectively.

Method	Drone → Satellite								Ground → Satellite							
	Seen				Unseen				Seen				Unseen			
	mIoU↑	mDice↑	AAE↓	ME↓	mIoU↑	mDice↑	AAE↓	ME↓	mIoU↑	mDice↑	AAE↓	ME↓	mIoU↑	mDice↑	AAE↓	ME↓
Sample4Geo [6]	38.18	53.73	3407.4	18.18	37.61	53.05	4483.2	20.85	23.59	35.50	14400.8	53.87	15.69	24.65	11538.6	89.03
Sample4Geo + SPS	50.09	61.58	3155.0	16.79	53.09	63.81	4147.4	19.16	19.02	27.10	15946.0	55.34	12.15	17.81	13224.0	88.39
DetGeo [32]	36.27	45.91	3882.5	55.68	32.20	39.42	4533.9	103.62	33.47	42.91	15674.4	78.32	21.38	26.45	11738.0	147.70
DetGeo + SPS	46.14	52.35	1786.3	55.16	41.83	45.67	2307.3	103.00	48.86	53.59	3936.1	78.57	29.69	31.89	5661.0	147.06
OCGNet [15]	33.11	42.06	4382.4	71.85	30.15	37.19	5192.3	105.34	33.76	43.39	16628.9	78.42	23.04	28.76	13463.0	137.24
OCGNet + SPS	43.06	48.70	1855.4	71.52	39.68	43.46	2520.8	104.49	50.07	54.91	3675.9	78.56	32.91	35.37	5395.1	136.26
TROGeo [40]	46.86	60.69	2934.0	46.21	49.85	63.21	3200.2	45.01	43.03	52.89	7452.9	115.08	29.80	36.98	8382.4	173.62
TROGeo + SPS	69.09	77.99	1175.2	17.15	75.70	82.68	1283.6	16.90	57.18	62.96	3291.9	59.81	40.71	43.95	4551.8	113.74
Ours (Point)	79.39	87.50	793.2	6.16	81.47	88.83	810.8	5.57	**77.37**	**84.01**	1758.1	**19.18**	72.68	78.63	2691.4	31.04
Ours (Bbox)	82.28	89.69	530.7	4.35	84.12	90.86	570.1	4.02	77.36	83.99	**1734.2**	19.57	73.82	79.76	**2528.7**	29.40
Ours (Mask)	**83.19**	**90.35**	**488.2**	**4.05**	**85.11**	**91.58**	**461.2**	**3.61**	77.21	83.82	1770.9	19.74	**74.07**	**80.03**	2533.4	**28.85**

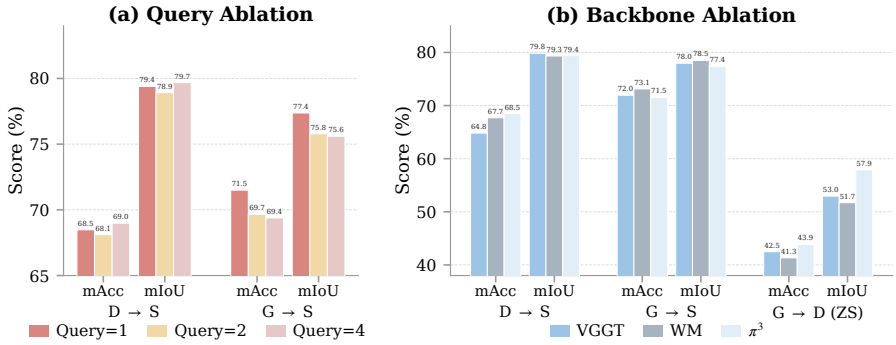

Fig. 4: Ablation studies of GAGeo. (a) Number of learnable tokens (Query): Evaluating the impact of query quantity on the D→S and G→S setups. **(b) Backbone architectures:** Comparing our proposed π^3 backbone against representative 3DFMs (VGGT [33] and World Mirror [25]). ZS denotes zero-shot.

Table 4: Zero-shot detection and segmentation performance on CVOGL-Seg datasets (Test set). Note that for segmentation metrics, the comparison methods are equipped with the SAM Prompt Stage (SPS) [40].

Method	Drone → Satellite							Ground → Satellite						
	Detection			Segmentation				Detection			Segmentation			
	mAcc↑	Acc@75↑	Acc@50↑	mIoU↑	mDice↑	AAE↓	ME↓	mAcc↑	Acc@75↑	Acc@50↑	mIoU↑	mDice↑	AAE↓	ME↓
Sample4Geo [6]	0.70	0.21	2.88	8.73	11.90	3554.1	112.40	0.42	0.10	1.75	5.10	7.14	3585.1	133.47
DetGeo [32]	12.09	11.41	21.79	19.14	21.71	5377.1	172.24	6.15	5.76	10.89	10.12	11.43	4809.6	208.39
OCGNet [15]	12.75	12.13	23.84	20.94	23.71	6893.0	159.25	8.77	8.74	16.14	14.09	15.76	7073.2	180.67
TROGeo [40]	7.10	2.77	21.89	24.30	28.66	8823.8	137.85	3.51	1.54	9.76	10.53	12.25	8175.8	193.37
Ours	**29.43**	**30.55**	**50.27**	**46.79**	**53.38**	**2191.3**	**102.38**	**19.37**	**19.28**	**35.21**	**33.02**	**38.36**	**3027.6**	**113.47**

Table 6: Ablation study (seen test set) with **point** prompts. Each row cumulatively adds the indicated module. Gray denotes the full model.

Task	Method	mAcc↑	Acc@50↑	Acc@75↑	mIoU↑	mDice↑	AAE↓	ME↓
D → S	(a) Base	62.86	89.97	71.06	77.19	85.82	861.1	7.22
	(b) + Deep Supervision	66.38	91.85	75.64	78.71	86.96	797.3	6.39
	(c) + Contrastive	67.63	92.11	76.92	**79.51**	**87.55**	**738.9**	6.18
	(d) + Camera Pose	**68.48**	**92.86**	**77.01**	79.39	87.50	793.2	**6.16**
G → S	(a) Base	67.41	86.64	77.01	74.86	81.78	2099.4	22.23
	(b) + Deep Supervision	70.06	88.41	79.57	76.58	83.29	1780.0	19.54
	(c) + Contrastive	71.33	89.47	80.69	**77.50**	**84.12**	**1748.8**	19.47
	(d) + Camera Pose	**71.50**	**89.69**	**80.70**	77.37	84.01	1758.1	**19.18**

Notably, Tab. 5 confirms that our contrastive learning loss narrows the ground-to-drone feature gap by using satellite object embeddings as anchors.

Table 5: Zero-shot performance on Ground → Drone.

Method	mAcc↑	mIoU↑
Sample4Geo [6]	0.43	12.07
DetGeo [32]	12.60	22.83
OCGNet [15]	15.16	27.16
TROGeo [40]	20.92	35.81
Ours (w/o CL)	38.38	57.00
Ours (Point)	**43.86**	**57.88**

Number of learnable tokens. Fig. 4(a) shows that a single task token is sufficient to capture the target information. Increasing queries to 2 or 4 degrades performance in $G \rightarrow S$ without consistent gains in $D \rightarrow S$, as additional tokens introduce redundant noise for single-object localization.

Impact of different backbones. As shown in Fig. 4(b), despite its lighter architecture (18 *vs.* 24 layers), π^3 delivers competitive performance and superior zero-shot G→D generalization. This suggests that the inherited geometric prior and the permutation-equivariant design help reduce view-specific biases.

5.4 Qualitative Results

Fig. 5 visualizes the performance across three cross-view setups on CVOGL-Seg and CMA-Loc dataset. Our method yields precise segmentation masks aligned with the ground truth, even under the unseen ground-to-drone setup. In contrast,

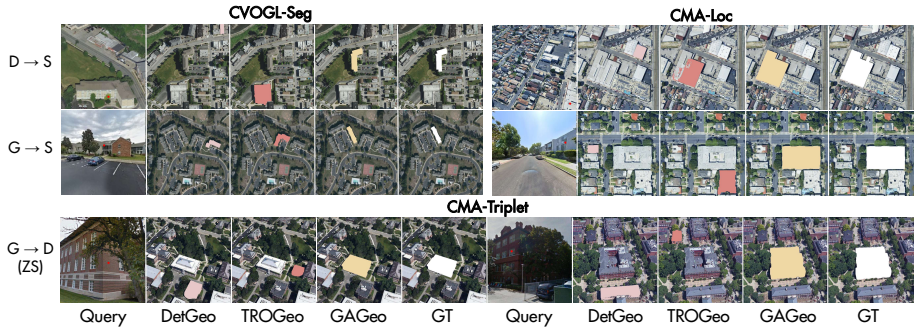

Fig. 5: Qualitative Results on CVOGL-Seg [40] and CMA-Loc datasets.

existing methods like TROGeo [40] and DetGeo [32] frequently yield fragmented or misaligned predictions, further validating the robustness of GAGeo.

6 Conclusion

In this paper, we address the intertwined data and methodological bottlenecks in CVOGL through a comprehensive framework. We introduce CMA-Loc, a large-scale, high-fidelity building dataset that overcomes prior limitations by providing diverse referring prompts and camera metadata across standard field-of-view imagery. To bridge the extreme spatial gap, we propose GAGeo, a unified, geometry-aware architecture that adapts the 3D prior of π^3 to simultaneously output multi-task predictions in a single streamlined forward pass. Furthermore, by utilizing the satellite view as a universal bridge in our contrastive learning objective, GAGeoenhances zero-shot ground-to-drone localization without relying on scarce triplet data. Extensive experiments demonstrate that our approach establishes a new state-of-the-art, advancing both detection and segmentation performance while exhibiting robust cross-domain generalization.

Admittedly, the inference efficiency of our framework is currently limited by the architectural complexity of 3D foundation models. Promising avenues for future research involve leveraging established paradigms such as token merging, pruning, and knowledge distillation to mitigate these efficiency constraints.

Acknowledgements

This research is supported in part by the Key Research Program of Hangzhou (No. 2025SZD1A56), the National Natural Science Foundation of China (No. 62461160308, U23B2010, 62576024), the Beijing Natural Science Foundation (No. L231011), the Fundamental Research Funds for the Central Universities (No. 501RCQD2025141003), BeiHang GanWei Project (No. 502GWXM2024141001), the National Science Foundation Support Projects (No. 62425303), the National Key R&D Program of China (No. 2024YFB4707300), and the Beijing Nova Program.

References

1. Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J.: Google street view: Capturing the world at street level. Computer **43**(6), 32–38 (2010)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229. Springer (2020)
4. Dai, M., Hu, J., Zhuang, J., Zheng, E.: A transformer-based feature segmentation and region alignment method for uav-view geo-localization. IEEE TCSVT **32**(7), 4376–4389 (2021)
5. Dai, M., Zheng, E., Feng, Z., Qi, L., Zhuang, J., Yang, W.: Vision-based uav self-positioning in low-altitude urban environments. IEEE TIP **33**, 493–508 (2023)
6. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: ICCV. pp. 16847–16856 (2023)
7. Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., Liu, W.: You only look at one sequence: Rethinking transformer in vision through object detection. NeurIPS **34**, 26183–26197 (2021)
8. Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
9. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: CVPR. pp. 15180–15190 (2023)
10. Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R.: Google earth engine: Planetary-scale geospatial analysis for everyone. Remote sensing of Environment **202**, 18–27 (2017)
11. Haklay, M., Weber, P.: Openstreetmap: User-generated street maps. IEEE Pervasive computing **7**(4), 12–18 (2008)
12. Hänsch, R., Arndt, J., Lunga, D., Gibb, M., Pedelose, T., Boedihardjo, A., Petrie, D., Bacastow, T.M.: Spacenet 8-the detection of flooded roads and buildings. In: CVPR. pp. 1472–1480 (2022)
13. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9729–9738 (2020)
14. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: CVPR. pp. 7258–7267 (2018)
15. Huang, Z., Aryal, J., Nahavandi, S., Lu, X., Lim, C.P., Wei, L., Zhou, H.: Object-level cross-view geo-localization with location enhancement and multi-head cross attention. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2025)
16. Ju, H., Huang, S., Liu, S., Zheng, Z.: Video2bev: Transforming drone videos to bevs for video-based geo-localization. In: ICCV. pp. 27073–27083 (2025)
17. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
18. Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: ECCV. pp. 734–750 (2018)

19. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV. pp. 71–91. Springer (2024)
20. Li, W., Lai, Y., Xu, L., Xiangli, Y., Yu, J., He, C., Xia, G.S., Lin, D.: Omnicity: Omnipotent city understanding with multi-level and multi-view images. In: CVPR. pp. 17397–17407 (2023)
21. Li, Z., Yuan, X., Liu, W., Xu, X.: Vageo: View-specific attention for cross-view object geo-localization. In: ICASSP. pp. 1–5. IEEE (2025)
22. Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., Sebe, N.: Joint representation learning and keypoint detection for cross-view geo-localization. IEEE TIP **31**, 3780–3792 (2022)
23. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: CVPR. pp. 891–898 (2013)
24. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: CVPR. pp. 5624–5633 (2019)
25. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: Worldmirror: Universal 3d world reconstruction with any-prior prompting. arXiv preprint arXiv:2510.10726 (2025)
26. Mithun, N.C., Minhas, K.S., Chiu, H.P., Oskiper, T., Sizintsev, M., Samarasekera, S., Kumar, R.: Cross-view visual geo-localization for outdoor augmented reality. In: 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR). pp. 493–502. IEEE (2023)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
29. Shi, Y., Li, H.: Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In: CVPR. pp. 17010–17020 (2022)
30. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. NeurIPS **32** (2019)
31. Song, H., Sun, D., Chun, S., Jampani, V., Han, D., Heo, B., Kim, W., Yang, M.H.: Vidt: An efficient and effective fully transformer-based object detector. In: ICLR (2022)
32. 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 TGRS **61**, 1–16 (2023)
33. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR. pp. 5294–5306 (2025)
34. 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 TCSVT **32**(2), 867–879 (2021)
35. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
36. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: ICCV. pp. 3961–3969 (2015)
37. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. NeurIPS **34**, 29009–29020 (2021)
38. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: CVPR. pp. 867–875 (2017)

39. Zhang, C., Wang, S.: Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data. In: CVPR. pp. 7839–7849 (2024)
40. Zhang, Q., Zhu, Y.: Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In: ICCV. pp. 8197–8206 (2025)
41. Zhang, X., Cao, S.Y., Bai, X., Li, Y., Shen, Z., Wu, Z., Hu, X., Shen, H.l.: Recurrent cross-view object geo-localization. arXiv preprint arXiv:2509.12757 (2025)
42. Zhang, X., Li, X., Sultani, W., Chen, C., Wshah, S.: Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. IEEE TPAMI **46**(12), 10419–10433 (2024)
43. Zhang, Y., Tung, J., Cai, R., Fouhey, D., Averbuch-Elor, H.: Emergent extreme-view geometry in 3d foundation models. arXiv preprint arXiv:2511.22686 (2025)
44. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACM MM. pp. 1395–1403 (2020)
45. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: CVPR. pp. 1162–1171 (2022)
46. Zhu, S., Yang, T., Chen, C.: Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In: CVPR. pp. 3640–3649 (2021)
47. Zhu, X.L.Y.: Improving cross-view object geo-localization: A dual attention approach with cross-view interaction and multi-scale spatial features. arXiv preprint arXiv:2510.27139 (2025)