# GeoBridge: A Semantic-Anchored Multi-View Foundation Model Bridging Images and Text for Geo-Localization

Supplementary Material

## A. Overview

This appendix supplements the proposed GeoBridge and our datasets GeoLoc with details excluded from the main paper due to space constraints. The appendix is organized as follows:

- Section B. Supplementary experiments to verify effectiveness.
- Section C. Construction and preprocessing details of the GeoLoc dataset.
- Section D. Implementation details of instruction formatting for constructing the descriptions in the GeoLoc dataset.
- Section E. Additional examples of model inference results.
- Section F. Datasheets for GeoLoc dataset.
- Section G. Discussion on limitations and societal impact.

## B. Supplementary Experiment

### B.1. Text Description Validation

Our method does not rely on GPT-4o [4], it uses GPT-4o only as a tool to generate cross-view textual descriptions. To validate this, we also employ text generated by the open-source models Qwen3 [9] and Gemini3 [7], as well as ensemble text produced by Gemini3 combining outputs from Qwen3 and GPT-4o, to serve as semantic anchors. As shown in Table 1, using text from different models leads to consistent improvements.

### B.2. Public Subset Construction

To further verify that our method does not depend on any proprietary data source, we constructed a three-view subset from the publicly available CVUSA [5] street–satellite geo-localization dataset and generated the corresponding semantic anchors. Specifically, we re-cropped the drone images to $750 \times 750$ pixels based on the coordinate points and applied the same filtering procedure, yielding 481 image sets, of which 81 were reserved for testing. We obtained the corresponding text descriptions in the same way and conducted experiments with GeoBridge. The results are presented in Table 2. Owing to the limited size of the dataset, only the projection layer was fine-tuned. The retraining results on this subset consistently show that our method outperforms the baseline.

## C. GeoLoc Construction and Processing

After acquiring the GeoLoc data, we performed systematic, multi-stage cleaning and quality control. Compared with the description in the main paper, this section provides more fine-grained operational details. It is worth noting that each step of the pipeline is manually monitored, and the overall process requires approximately 150 hours of human effort.

### C.1. Drone Data Acquisition and Cropping

The original drone imagery used in GeoLoc is illustrated in Fig. 1. It covers a wide range of scene types, including regions with prominent structural features (Fig. 1a, 1b), locations with highly similar ground patterns (Fig. 1c, 1d), and samples captured from different flight altitudes, resulting in varying perspectives (Fig. 1e, 1f). It also contains large-scale natural environments with limited discriminative structure, such as extensive water bodies, forests, grasslands, and deserts (Fig. 1g, 1h, 1i, 1j, 1k). In addition, due to diverse acquisition conditions, some images exhibit degraded quality issues such as low illumination (Fig. 1l).

In the initial stage of data construction, we manually inspected all original drone images and removed those with poor visual quality or dominated by large natural areas lacking clearly identifiable ground features (*e.g.*, Fig. 1g, 1h, 1i, 1j 1k, and 1l). This manual screening process required approximately 20 hours in total.

To obtain drone-view images with precise geometric structure and strict correspondence to street-view imagery, we applied a two-stage cropping and alignment procedure to the original drone data.

First, since the availability of street-view imagery at specific geographic coordinates was unknown a priori, we applied a sliding-window cropping operation over the original drone imagery using a preset $80 \times 80$ pixel window. For each cropped region, we queried the Google Street View Static to obtain candidate street-view coordinates. For locations where a valid street-view image was found, the associated coordinates were then used as the center for a reverse cropping operation on the original drone image. In this way, the street-view viewpoint is approximately centered within the corresponding drone sub-image, yielding more precise spatial alignment between the two views.

Furthermore, to enhance scene diversity and scale robustness, we did not crop patches based on a fixed image resolution. Instead, we defined cropping windows according to different ground coverage areas. This design enables us to capture multi-scale perspectives at varying reso-

Table 1. Further validation of the text description on the GeoLoc dataset. Best results are shown in bold; second-best results are shown in underlined.

| Method | D2S | | S2D | | P2S | | S2P | | D2P | | P2D | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| No semantic anchor | 38.20 | 43.76 | 34.63 | 47.82 | 6.43 | 14.12 | 6.95 | 15.98 | 7.16 | 16.26 | 4.90 | 12.25 |
| Qwen3 [9] | **46.07** | 47.26 | 44.72 | 45.49 | 30.56 | 35.16 | 38.97 | 43.69 | 30.24 | 36.59 | 32.05 | 32.25 |
| Gemini3 [7] | 38.39 | 47.85 | 37.47 | 46.38 | 24.58 | 37.64 | 28.12 | 43.71 | 25.51 | 29.89 | 24.95 | 36.33 |
| Gemini3 [7] (Qwen3[9] + GPT-4o [4]) | 42.91 | **53.42** | 41.68 | 41.89 | 31.81 | **47.33** | 32.41 | **45.52** | 31.81 | 38.71 | 31.79 | 43.34 |
| GeoBridge | 45.05 | 49.05 | **44.81** | **48.76** | **38.87** | 42.10 | **39.20** | 41.96 | **41.22** | **43.54** | **41.15** | **43.41** |

Table 2. Comparison on the CVUSA subdataset (D for Drone, P for Street-View Panorama, and S for Satellite). The best results are shown in bold, the second-best results are underlined.

| Method | D2S | | S2D | | P2S | | S2P | | D2P | | P2D | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP | R@1 | AP |
| Sample4Geo [2] | 13.70 | 22.22 | 7.40 | 16.34 | 7.41 | 14.81 | 8.64 | 12.35 | 13.58 | 17.68 | 12.35 | 17.37 |
| MEAN [1] | 12.35 | 17.37 | 13.58 | 19.60 | - | - | - | - | 4.41 | 7.27 | 5.23 | 8.94 |
| MCCG [6] | 10.27 | 14.22 | 14.17 | 17.89 | - | - | - | - | 2.63 | 6.31 | 5.58 | 7.16 |
| panorama-BEV [10] | - | - | - | - | 4.94 | 8.64 | 11.11 | 17.28 | 6.17 | 14.81 | 9.88 | 15.60 |
| AuxGeo [8] | - | - | - | - | 3.07 | 18.93 | 7.48 | 21.80 | 13.74 | 14.95 | 10.00 | 12.49 |
| GeoBridge | **49.38** | **62.46** | **49.39** | **63.17** | **28.40** | **43.68** | **28.39** | **42.42** | **20.99** | **31.52** | **16.05** | **31.03** |

lutions and flight altitudes while maintaining comparability in geographic scale, which is beneficial for learning more generalizable cross-view representations. We obtain drone subimages covering ground areas of approximately $80 \times 80$, $100 \times 100$, $120 \times 120$, $150 \times 150$, and $180 \times 180$ $(m^2)$.

## C.2. Basic validity screening

After obtaining the initial cropped drone sub-images, we first perform deduplication based on spatial coverage. For any pair of sub-images whose ground coverage overlaps by more than $50\%$, we retain only one representative sample and discard the others as duplicates.

We then conduct basic quality checks to ensure that each image contains a sufficient number of valid pixels and adequate spatial resolution. Subimages that are too small to clearly depict meaningful ground structures are removed. In addition, some images may contain large pure-black or pure-white regions due to sensor failures, rendering errors, or cropping at the image borders. For each image, we compute the proportion of pure-black and pure-white pixels. Images in which the proportion of pure-black or pure-white pixels exceeds $1\%$ are considered invalid and discarded. Representative examples of such removed samples are shown in Fig. 2.

## C.3. Multi-Metric Image Quality Filtering

We applied a three-stage quality filtering process to further refine the drone imagery.

### C.3.1. BH-Gate: Filtering Blur and Low-Texture Images

This module primarily targets blurry images (*e.g.*, motion blur, defocus), low-texture scenes caused by haze, cloud cover, or uneven illumination, as well as images with severe compression artifacts that lead to substantial detail loss. BH-Gate combines global pixel variance with an image sharpness measure to detect the absence of meaningful spatial detail. As illustrated in Fig.3, when an image exhibits extremely low texture variation, it is deemed to contain insufficient visual information and is directly discarded.

### C.3.2. C-Gate: Filtering Images with Insufficient Global Contrast

Although some images pass the first-stage screening, they can still exhibit slight blurring or brightness drift. C-Gate targets such cases by identifying images with an insufficient grayscale dynamic range. These are typically overexposed or underexposed images whose grayscale values are highly concentrated, as well as mildly blurred images that retain only weak structural cues. To this end, we analyze the dispersion of the grayscale distribution (*i.e.*, global contrast). As illustrated in Fig. 4, such images usually fail to provide informative ground textures and contribute little to cross-view and cross-modal learning, and are therefore further filtered out.

### C.3.3. UN-Gate: Filtering Uniform-Texture and Noisy Pseudo-Texture Images

After the first two stages of filtering, there may still remain visually near-uniform yet structurally uninformative scenes (*e.g.*, large expanses of water, sky, grassland, desert, or snow), as well as sensor-induced pseudo-textures or noise patterns (including white noise, mosaic artifacts, stripe noise, and random texture blocks). To handle these challenging cases, UN-Gate further combines information entropy, variance range, and the proportion of saturated pixels

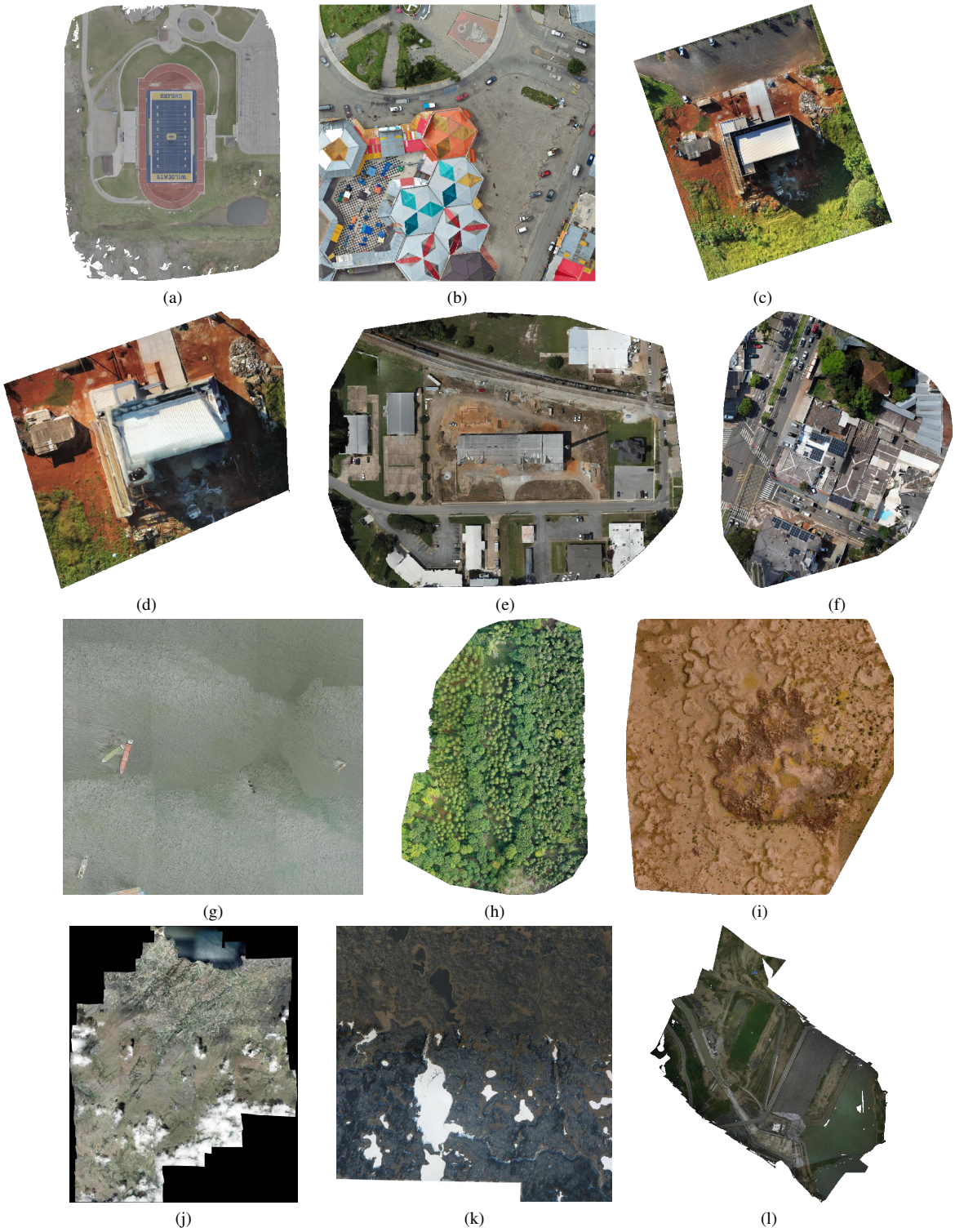


Figure 1. Examples of original drone images.

Figure 2. Examples of basic validity screening.

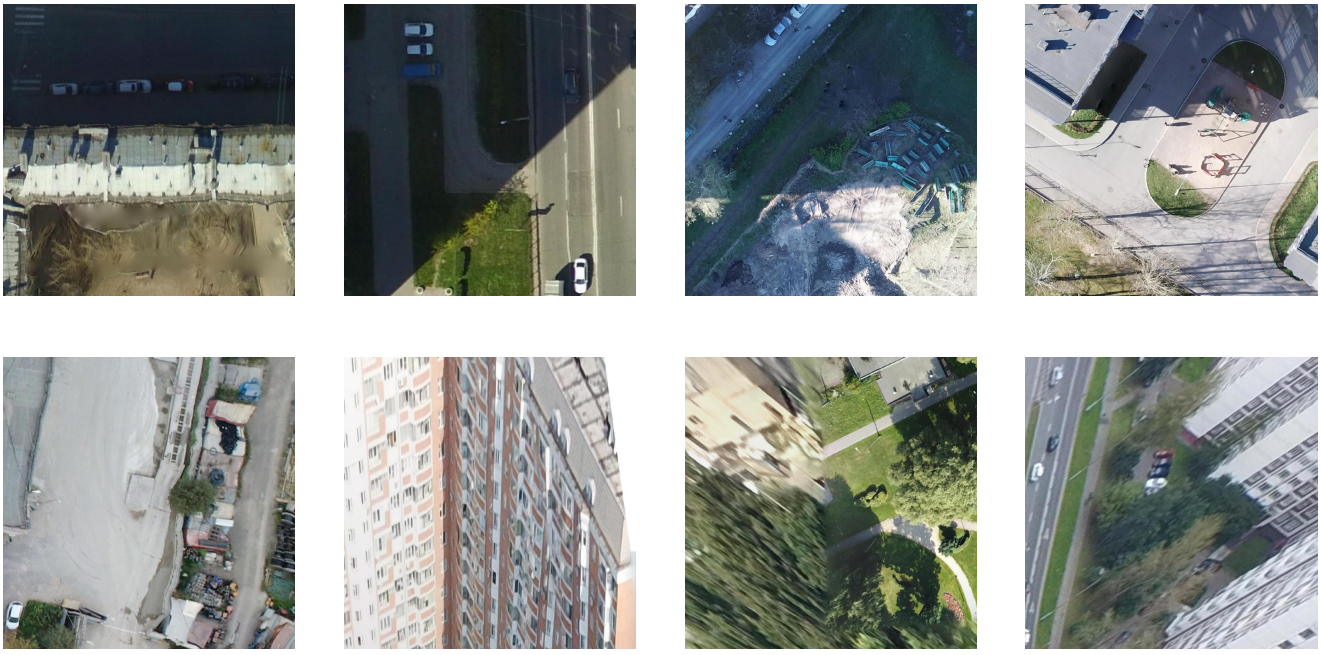


Figure 3. Examples of blurry drone subimages.

to assess the information content of each image, distinguish pseudo-texture noise, and detect images that are close to white-noise patterns. Based on these metrics, we determine whether an image falls into the categories of no effective semantics or noisy pseudo-texture. As shown in Fig.5, images that exhibit highly uniform textures or textures dominated by noise are discarded at this stage.

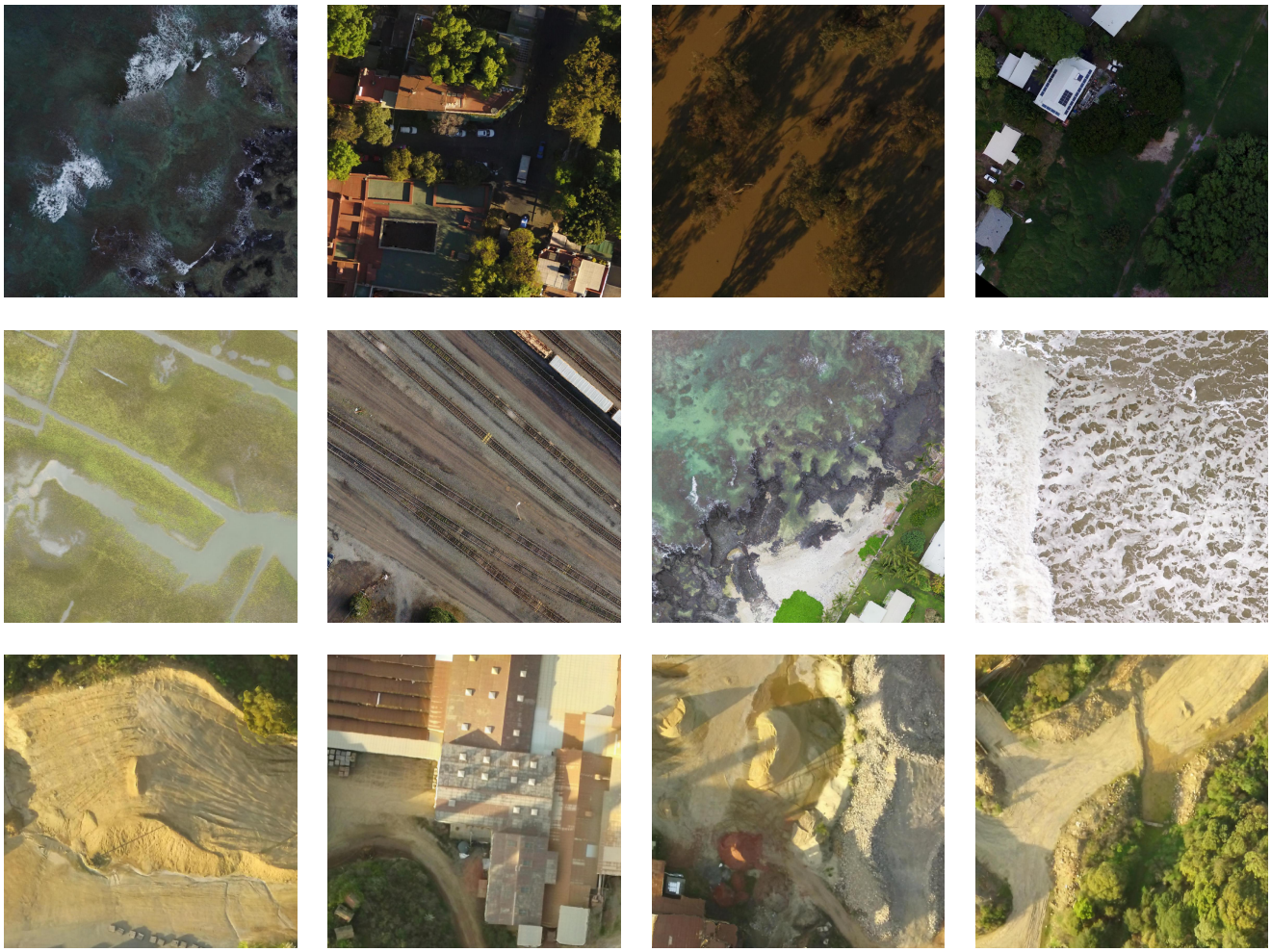


Figure 4. Examples of low global-contrast drone subimages

### C.4. Tri-view Alignment

After completing the filtering of all drone sub-images, we download the corresponding street-view panoramic images and satellite images using the candidate coordinates associated with the retained drone samples.

When downloading street-view images, we query the Google Street View Static API at the specified panorama location, with the camera heading fixed to face north, the pitch set to eye level, and the field of view set to 120°. Under this configuration, the API returns a street-view image rendered from the high-resolution panorama at that location, facing a consistent direction. For satellite imagery, we retrieve Google Maps satellite images covering the geographic extent corresponding to the latitude–longitude bounds of each retained drone subimage. Examples of the resulting tri-view correspondence (drone, street-view, and satellite) are shown in Fig. 6.

## D. Instruction Details for the GeoLoc Dataset

To construct high-quality cross-view semantic descriptions for GeoLoc, we design a unified instruction protocol that guides a large language model to generate consistent, viewpoint-agnostic textual annotations for each tri-view set (drone, satellite, and street-view images). The goal of these instructions is to ensure that the resulting descriptions serve as reliable semantic anchors, capturing stable structural cues shared across views while avoiding viewpoint-specific bias. The specific details are shown in Fig. 7, 8, 9, and 10.

The instruction set is crafted to ensure that the generated descriptions meet the following requirements:

- Cross-view semantic consistency: the text must focus on structural elements that remain stable across drone, satellite, and street-view perspectives. These include roads, intersections, bridges, buildings, parks, rivers, land parcels, plazas, and distinctive landmarks. The model is instructed

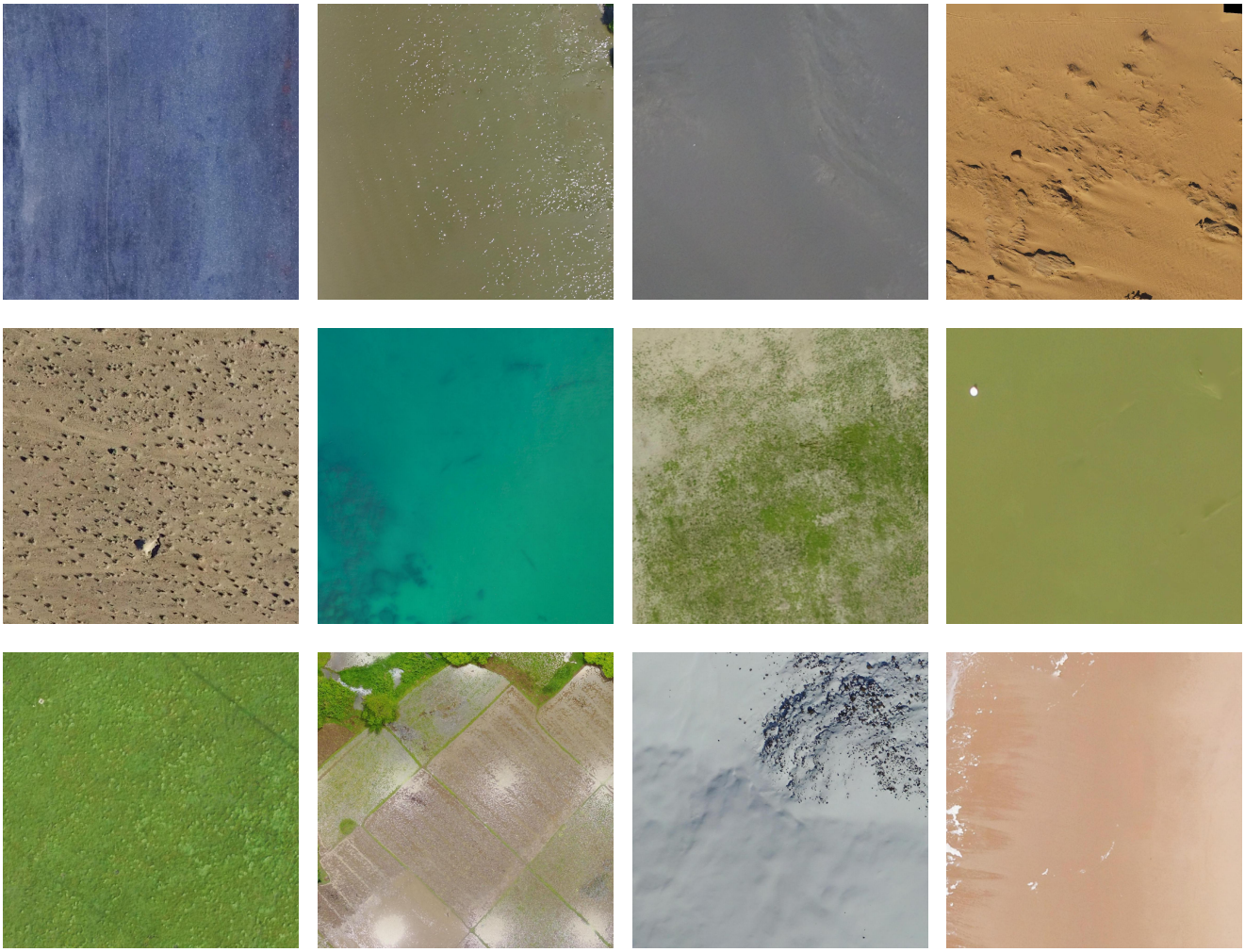


Figure 5. Examples of uiform-texture and noisy pseudo-texture drone subimages

to ignore features that vary across viewpoints.

- Explicit modeling of spatial structures and relationships: the description must convey meaningful spatial relations (*e.g.*, a bridge spans the river, the road curves around the building) while avoiding explicit directional terminology. Instead of using directional words such as north, south, east, west, left, right, front, or back, the text should describe structural relationships directly and unambiguously.
- Fine-grained but concise representation: each description is restricted to a single, compact paragraph. The instruction emphasizes the inclusion of salient structural cues without unnecessary adjectives, narrative elements, or scene-level speculation.
- Landmark naming when applicable: when a location contains identifiable landmarks such as notable buildings, bridges, squares, or monuments, the model is instructed to explicitly name them. This enhances the distinctiveness of the descriptions and improves performance in cross-view retrieval and localization.
- Exclusion of transient or unstable elements: the instruction explicitly prohibits mentioning transient objects (*e.g.*, people, cars, bicycles, animals), weather, temporary decorations, graffiti, or other viewpoint-dependent artifacts. This ensures that descriptions remain stable over time and across viewpoints.

The above instructions also apply when generating descriptions for cross-modal geo-location.

## E. Visualizations of Model Inference Results

We visualize the cross-modal geo-location results. In this setting, we generate a textual description from a single viewpoint and use it to retrieve images from the other viewpoints. Fig. 11, 12, and 13 show the retrieval results for satellite images, drone images, and street-view images using descriptions derived from street-view, satellite, and

Figure 6. Examples of aligned tri-view images.

I will provide three images of the same location: a drone view, a satellite view, and a street view. Please generate a unified text description that:

- Focuses on fine-grained and salient features such as roads, intersections, bridges, buildings, parks, rivers, and landmarks, ensuring consistency across the three images.
- Clearly describes spatial structures and relationships (e.g., The road crosses the river, and tall buildings stand beside it).
- Avoids mentioning viewing angles or using directional terms (east, west, north, south, left, right, front, back, center).
- Highlights unique features that distinguish the location, directly naming landmarks or significant structures when present.
- Keeps the description concise and precise, omitting unnecessary modifiers.
- Excludes transient or random elements such as vehicles, graffiti, or people.



The location features a landscaped area with a distinct hexagonal paving pattern interspersed with green spaces and trees. A curved road encircles part of this area, bordered by buildings with flat rooftops and modern designs. A small red-roofed structure and a green-roofed structure are situated within the landscaped zone, connected by paved walkways. Adjacent to the landscaped area, a multi-lane road runs alongside a prominent multi-story building with a curved facade and a grid of windows. The road is separated from the building by a low wall and a narrow strip of greenery. The area is characterized by its organized layout, combining urban infrastructure with green spaces.



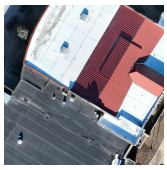
The location features a cluster of rectangular commercial buildings with flat roofs, some of which show signs of wear, such as discoloration and patches. The buildings are aligned along a paved road with marked parking spaces adjacent to the structures. The area includes a mix of brick and light-colored facades, with one prominent curved corner on a brick building and another building featuring a white exterior with a red awning above an entrance. The road intersects with another, forming a junction near the buildings. Additional structures with metal roofs and open parking areas are visible nearby, contributing to the area's urban layout.


The location features a central parking lot with marked spaces, including designated accessible parking. Surrounding the lot are several rectangular commercial buildings with flat roofs, some of which have visible HVAC units. A prominent red brick building with signage for "Ant Street Inn" serves as a distinctive landmark. The area is bordered by paved roads with sidewalks and small landscaped areas, including trees and decorative brickwork. The streets intersect near the parking lot, creating a cohesive urban layout.

Figure 7. Tri-View instruction protocol for generating unified semantic descriptions. The blue text box denotes the instruction prompt; the first column presents the street-view panorama, the second column shows the drone-view image, the third column displays the satellite image, and the fourth column contains the generated textual description.

I will provide one image of the drone view. Please generate a unified text description that:

- Focuses on fine-grained and salient features such as roads, intersections, bridges, buildings, parks, rivers, and landmarks, ensuring consistency across the three images.
- Clearly describes spatial structures and relationships (e.g., The road crosses the river, and tall buildings stand beside it).
- Avoids mentioning viewing angles or using directional terms (east, west, north, south, left, right, front, back, center).
- Highlights unique features that distinguish the location, directly naming landmarks or significant structures when present.
- Keeps the description concise and precise, omitting unnecessary modifiers.
- Excludes transient or random elements such as vehicles, graffiti, or people.



The image depicts a paved area with a parking lot bordered by a red-brick building featuring white accents and a flat roof. The building is accompanied by landscaped greenery, including shrubs and small trees. In the background, another structure with a similar red-brick design is visible, separated by the parking lot. Light poles are evenly distributed across the lot, and the surrounding area includes additional trees and open spaces, suggesting a suburban or institutional setting. The layout is organized, with clear pathways and defined boundaries between structures and greenery.

The image depicts an urban area characterized by intersecting paved roads bordered by sidewalks. A small grassy area with scattered trees is visible near the intersection. A red-brick building with rectangular windows stands adjacent to the road, accompanied by a smaller, light-colored structure. Utility poles with overhead power lines run along the streets, adding to the infrastructure. In the background, additional buildings and open spaces are visible, contributing to the area's mixed-use layout.

The image depicts a wide, paved road with two distinct lanes, bordered by grassy areas on both sides. Adjacent to one side of the road, there is a large parking lot connected to a rectangular building with a flat roof, likely a commercial or institutional structure. The surrounding area features scattered trees and open green spaces, with no prominent landmarks or water bodies visible. The scene is framed by a cloudy sky, adding depth to the setting.

Figure 8. Cross-modal geo-location drone image description instructions, with blue text boxes indicating the prompts.

I will provide a single street-view panorama image. Please generate a unified text description that:
- Focuses on fine-grained and salient features such as roads, intersections, bridges, buildings, parks, rivers, and landmarks, ensuring consistency across the three images.
- Clearly describes spatial structures and relationships (e.g., The road crosses the river, and tall buildings stand beside it).
- Avoids mentioning viewing angles or using directional terms (east, west, north, south, left, right, front, back, center).
- Highlights unique features that distinguish the location, directly naming landmarks or significant structures when present.
- Keeps the description concise and precise, omitting unnecessary modifiers.
- Excludes transient or random elements such as vehicles, graffiti, or people.

The scene features a paved road intersecting with another road, marked by clear white lane markings. On one side, a multi-story residential building with balconies and a beige facade is enclosed by a white metal fence. Adjacent to it, a gray building with angular windows and a covered entrance houses a small storefront or service area, with bicycles parked nearby. Overhead, utility wires are suspended between poles, adding to the urban infrastructure. The area is surrounded by additional low-rise buildings and sparse greenery, including a small tree near the intersection.

The scene features an urban street flanked by modern and industrial-style structures. On one side, a contemporary building with large glass panels and concrete framing houses commercial spaces. Adjacent to it, a multi-story industrial building is partially obscured by stacked shipping containers painted in various colors, some adorned with graffiti. The street is divided by painted lane markings, with protective orange barriers placed near the containers. The area reflects a mix of functional and modern architectural styles, with the containers adding a distinctive industrial character.

The image depicts an urban street scene with two prominent buildings. On one side, a modern structure features a curved facade with horizontal louvered panels and glass accents, accompanied by landscaped greenery and a paved walkway. Adjacent to it, a beige multi-story building with rounded edges and rows of rectangular windows dominates the view. A row of bicycles is parked near its base, and the road runs between the two buildings, bordered by sidewalks and small trees. Additional tall buildings are visible in the background, contributing to the dense cityscape.

Figure 9. Cross-modal geo-location street-panorama image description instructions, with blue text boxes indicating the prompts.

I will provide one image of the satellite view. Please generate a unified text description that:
- Focuses on fine-grained and salient features such as roads, intersections, bridges, buildings, parks, rivers, and landmarks, ensuring consistency across the three images.
- Clearly describes spatial structures and relationships (e.g., The road crosses the river, and tall buildings stand beside it).
- Avoids mentioning viewing angles or using directional terms (east, west, north, south, left, right, front, back, center).
- Highlights unique features that distinguish the location, directly naming landmarks or significant structures when present.
- Keeps the description concise and precise, omitting unnecessary modifiers.
- Excludes transient or random elements such as vehicles, graffiti, or people.

The image depicts a cluster of rectangular buildings with flat roofs arranged in a geometric pattern, surrounded by green spaces and mature trees. Pathways connect the buildings and traverse the grassy areas, creating a network of walkways. The structures are interspersed with landscaped areas, including small patches of open grass and shrubs. The layout suggests a planned complex, possibly residential or institutional, with the buildings positioned close to one another.

The image depicts a large rectangular building with a flat roof, featuring skylights and ventilation structures. Adjacent to the building, a curved road runs alongside a grassy area with scattered trees and vegetation. A smaller structure, resembling a house, is situated near the road within the greenery. The surroundings include open spaces and patches of forested land, creating a contrast between the built environment and natural elements.

The image depicts a large rectangular building with a red-tiled roof and a central courtyard. The structure features evenly spaced triangular dormers along the roof and a circular tower at one corner. The courtyard includes a circular fountain surrounded by sparse vegetation and paved open space. Dense greenery surrounds the building, with trees forming a natural boundary along one side. A pathway runs adjacent to the building, connecting it to the surrounding area.

Figure 10. Cross-modal geo-location satellite image description instructions, with blue text boxes indicating the prompts.

drone images, respectively. These examples demonstrate that GeoBridge effectively captures key semantic cues in the text descriptions and accurately localizes the corresponding regions in candidate images from other views, substantially improving the ranking of the correct matches.

## F. Datasheets

In this section, we document essential details about the proposed datasets and benchmarks following the CVPR Dataset and Benchmark guidelines and the template provided by Gebru *et al.* [3].

### F.1. Motivation

The questions in this section are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests. The latter may be particularly relevant for datasets created for research purposes.

1. *"For what purpose was the dataset created?"*
   **A:** The GeoLoc dataset is created to address limitations of existing cross-view geo-localization benchmarks and to enable robust multi-view and cross-modal localization under realistic conditions. Previous datasets largely follow a two-view, satellite-centric paradigm, with limited geographic diversity, restricted multi-height and multi-condition coverage, and no complementary drone or street-view perspectives. In contrast, GeoLoc establishes a multi-view, multi-modal foundation that supports robust localization when satellite imagery is missing or outdated, facilitates air–ground multi-sensor fusion and closed-loop evaluation, and improves model generalization across diverse geographic regions and urban morphologies.
2. *"Who created the dataset (e.g., which team, research group) and on behalf of which entity?"*
   **A:** The dataset was created by the following authors:
   - Zixuan Song, Jing Zhang, Di Wang, Zidie Zhou, Wenbin Liu, Haonan Guo, En Wang, Bo Du
3. *"Who funded the creation of the dataset?"*
   **A:** The dataset creation was funded by the affiliations of the authors involved in this work.

### F.2. Composition

Most of the questions in this section are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for their chosen tasks. Some of the questions are designed to elicit information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions. Questions that apply only to datasets that relate to people are grouped together at the end of the section. We recommend taking a broad interpretation of whether a dataset relates to people. For example, any dataset containing text that was written by people relates to people.

1. *"What do the instances that comprise our datasets represent (e.g., documents, photos, people, countries)?"*
   **A:** Each instance in the GeoLoc dataset corresponds to a single real-world geographic location and consists of four aligned modalities: a low-altitude drone image, a ground-level street-view panorama, an overhead satellite image, and a unified textual description. All four share the same GPS coordinate, with the three visual views offering complementary spatial structure and the text providing a concise, viewpoint-agnostic semantic summary of the scene.
2. *"How many instances are there in total (of each type, if appropriate)?"*
   **A:** The GeoLoc dataset contains a total of 52,679 aligned instances, each consisting of a drone-view image, a street-view panoramic image, a satellite-view image, and a unified textual description. Among these, 47,328 instances are used for training and validation, and 5,351 instances from non-overlapping cities form the held-out evaluation set.
3. *"Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?"*
   **A:** Yes. The GeoLoc dataset is a curated subset drawn from a much larger pool of globally available imagery sources—such as those accessible through Google's APIs. From this broad collection, we retain only locations where drone, street-view, and satellite imagery can all be reliably obtained and strictly aligned.
4. *"Is there a label or target associated with each instance?"*
   **A:** Yes. Each instance is associated with a unified textual description that serves as a semantic label, and all three visual views share the same GPS coordinate, which provides the geographic target for geo-localization tasks.
5. *"Is any information missing from individual instances?"*
   **A:** No, each individual instance is complete.
6. *"Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?"*
   **A:** Yes. The primary relationship made explicit is the shared geographic coordinate: all instances located at different positions are independent, while each instance internally links its four modalities through strict spatial co-location.
7. *"Are there recommended data splits (e.g., training, development/validation, testing)?"*
   **A:** Yes. We recommend using the provided split, where 47,328 instances are allocated for training and validation, and 5,351 instances from non-overlapping cities are reserved as a held-out test set to ensure geographic sep-

The scene features a paved parking lot bordered by a multi-story hotel building with a curved facade and numerous windows. Adjacent to the hotel, there is a landscaped area with trees and shrubs. In the background, modern office buildings with glass and brick exteriors are visible, separated by green spaces and pathways. The area appears organized and urban, with a mix of commercial and landscaped elements. The sky is clear, enhancing the visibility of the structures.

The image depicts a modern multi-story office building with a curved glass facade and brick accents. The structure features a prominent entrance with a canopy and steps leading to the main doors. Adjacent to the building is a large parking lot with organized rows of parked vehicles. A small landscaped area with ornamental grasses and a rectangular structure is situated near the parking lot. In the background, another multi-level parking garage is visible, along with a mix of greenery and distant structures. The scene is characterized by its clean, corporate environment and well-maintained surroundings.

The scene features a paved road intersecting near modern office buildings with glass facades and curved architectural designs. A landscaped area with trimmed grass, small plants, and young trees is situated adjacent to the road. The buildings are symmetrical in style, with multiple stories and large windows, creating a corporate or commercial environment. The area is open and well-maintained, with no visible obstructions or additional structures dominating the view.

Figure 11. Qualitative results for cross-modal geo-location. Using street view descriptions to match satellite perspectives, the top three results are reported; red boxes indicate correct matches.

aration and fair evaluation.

8. *"Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?"*
   **A:** The dataset draws on external sources for the street-view and satellite imagery, specifically Google Street View and Google Maps Satellite.
9. *"Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?"*
   **A:** No. The dataset contains only drone imagery, Google Street View panoramas, Google Maps satellite imagery, and generated textual descriptions. It does not include personal or confidential information. The Google Street View and Google Satellite images used in the dataset are subject to Google's terms of service and licensing policies, and their use follows the corresponding Google usage agreements.
10. *"Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?"*
    **A:** No, GeoLoc does not contain any data with negative information.

The image showcases a circular building with a distinct rooftop design featuring multiple square structures arranged in a grid pattern. Surrounding the building, there is a mix of paved areas, green spaces, and trees. Adjacent to the building, a parking lot contains several parked vehicles, bordered by a grassy area with small structures and pathways. Residential buildings with uniform rooftops line a road that runs alongside the circular structure. A section of the area includes a park-like setting with geometric landscaping and shaded seating areas created by triangular canopies. The layout emphasizes a blend of urban and recreational spaces.

The image depicts a structured urban area with a central building featuring a distinct multi-angled roof surrounded by parking lots filled with vehicles. Adjacent to the building, a landscaped area with trees and small pathways is visible. To one side, there is a larger industrial structure with a white roof and an open yard containing containers and organized materials. The scene is bordered by a road with multiple lanes, separated from the buildings by a line of trees. The spatial arrangement highlights a mix of commercial, industrial, and landscaped elements.

The image depicts a developed area with a cluster of buildings featuring pitched roofs, some of which are equipped with large solar panel installations. Surrounding the buildings are paved parking lots with marked spaces and a few scattered trees providing greenery. A network of narrow roads connects the buildings and parking areas, forming intersections within the complex. Adjacent to the buildings, a fenced residential area is visible, containing gardens, small structures, and open green spaces. Dense tree coverage borders parts of the scene, creating a natural boundary between the residential and developed areas.

Figure 12. Qualitative results for cross-modal geo-location. Using satellite view descriptions to match drone perspectives, the top three results are reported; red boxes indicate correct matches.

### F.3. Collection Process

In addition to the goals outlined in the previous section, the questions in this section are designed to elicit information that may help researchers and practitioners create alternative datasets with similar characteristics. Again, questions that apply only to datasets that relate to people are grouped together at the end of the section.

1. *"How was the data associated with each instance acquired?"*
   **A:** Each instance is created by acquiring a drone-view image with GPS metadata, retrieving the corresponding street-view panorama via the Google Street View Static API, and obtaining the matching satellite image from Google Maps Satellite. All three views share the same coordinates. A unified textual description is then generated using GPT-4.0 under a controlled instruction protocol to ensure consistent, viewpoint-agnostic semantics.
2. *"What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?"*
   **A:** Data collection used drone-mounted cameras with GPS to capture drone imagery. Street-view and satellite images were obtained through the Google Street View Static API and Google Maps Satellite. The pipeline also included manual curation and automated procedures for cropping, alignment, filtering, and GPT-4.0–based text generation.
3. *"If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?"*

The image depicts a modern building complex with a white rooftop, surrounded by landscaped areas featuring paved walkways, small trees, and patches of grass. Adjacent to the building, a row of bicycles is neatly arranged along the edge of the structure. The building exhibits a geometric design with sections of colorful facades in yellow and blue tones. Curved pathways weave through the landscaped grounds, connecting open spaces with planted areas. The scene is characterized by organized spatial planning and a mix of built and natural elements.

The image shows a large rectangular building with a flat, light-colored roof featuring multiple mechanical units and ventilation systems arranged in structured patterns. A curved glass canopy extends from one side of the building, connecting to a paved area below. Adjacent to the building is a landscaped strip with sparse vegetation and evenly spaced light poles. A paved road runs parallel to the building, marked with yellow dashed lines and bordered by sidewalks. The structure exhibits a clean, industrial design with distinct geometric elements.

The image depicts a residential area characterized by two elongated buildings with dark roofs and multiple windows. Between the buildings, there is a paved area used for parking, bordered by trees and small landscaped spaces. Adjacent to the buildings, a green lawn with neatly trimmed hedges and scattered shrubs is visible. A road runs along the top edge of the image, intersecting with the parking area and lined with additional greenery. The spatial arrangement highlights the organized layout of the residential structures and their integration with open spaces and vegetation.

Figure 13. Qualitative results for cross-modal geo-location. Using drone view descriptions to match street perspectives, the top three results are reported; red boxes indicate correct matches.

**A:** Please refer to the details listed in the main text Section 4.

### F.4. Preprocessing, Cleaning, and Labeling

The questions in this section are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks. For example, text that has been converted into a "bag-of-words" is not suitable for tasks involving word order.

1. *"Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?"*
   **A:** Yes. Extensive preprocessing and cleaning were performed, including multi-stage image quality filtering (removing blurry, low-texture, low-contrast, and uniform-texture images), spatial deduplication based on coverage overlap, elimination of images with large invalid pixel regions, and multi-scale cropping with precise spatial alignment. All tri-view samples were retained only when drone, street-view, and satellite images were reliably co-located. Textual descriptions were generated using GPT-4.0 under a controlled instruction protocol, serving as the semantic labels for each instance.
2. *"Was the 'raw' data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?"*
   **A:** Yes, raw data is accessible.
3. *"Is the software that was used to preprocess/clean/label the data available?"*
   **A:** Yes, the necessary software used to preprocess and clean the data is publicly available.

### F.5. Uses

The questions in this section are intended to encourage dataset creators to reflect on tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers make informed decisions, thereby avoiding potential risks or harms.

1. *"Has the dataset been used for any tasks already?"*
   **A:** No.
2. *"Is there a repository that links to any or all papers or systems that use the dataset?"*
   **A:** Yes, we will provide such links in the GitHub and the Huggingface repository.
3. *"What (other) tasks could the dataset be used for?"*
   **A:** Beyond geo-localization, the GeoLoc dataset can support a wide range of tasks, including multi-view representation learning, cross-modal retrieval, vision–language grounding, UAV navigation and path planning, urban scene understanding, air–ground sensor fusion, and benchmarking models for semantic alignment across heterogeneous viewpoints and modalities.
4. *"Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"*
   **A:** No.
5. *"Are there tasks for which the dataset should not be used?"*
   **A:** N/A.

### F.6. Distribution

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. *"Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?"*
   **A:** No. The datasets will be made publicly accessible to the research community.
2. *"How will the dataset be distributed (e.g., tarball on website, API, GitHub)?"*
   **A:** We will provide GeoLoc in the GitHub and the Huggingface repository.
3. *"When will the dataset be distributed?"*
   **A:** We will create a repository to release the data once the paper is officially published, ensuring compliance with the anonymity principle.
4. *"Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?"*
   **A:** Yes, the dataset will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.
5. *"Have any third parties imposed IP-based or other restrictions on the data associated with the instances?"*
   **A:** Yes. The street-view and satellite images included in the dataset originate from Google Street View and Google Maps Satellite, which are subject to Google's terms of service and usage policies. These modalities must be used in compliance with Google's licensing restrictions.
6. *"Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?"*
   **A:** No.

### F.7. Maintenance

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

1. *"Who will be supporting/hosting/maintaining the dataset?"*
   **A:** The authors of this work serve to support, host, and maintain the datasets.
2. *"How can the owner/curator/manager of the dataset be contacted (e.g., email address)?"*
   **A:** The curators can be contacted via the email addresses listed on our paper or webpage.
3. *"Is there an erratum?"*
   **A:** There is no explicit erratum; updates and known errors will be specified in future versions.
4. *"Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?"*
   **A:** Future updates (if any) will be posted on the dataset website.
5. *"Will older versions of the dataset continue to be supported/hosted/maintained?"*
   **A:** Yes.
6. *"If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?"*
   **A:** Yes, we will provide detailed instructions for future extensions.

## G. Limitation and Potential Societal Impact

In this section, we discuss the limitations and potential societal impact of this work.

### G.1. Potential Limitations

Despite its scale and multi-view design, GeoLoc has several inherent limitations. First, the dataset is constrained by the availability of Google Street View and Google Maps Satellite imagery, which biases geographic coverage toward regions with strong mapping infrastructure and may underrepresent rural or geopolitically restricted areas. Sec-

ond, the multi-stage quality filtering removes scenes with extreme degradation, low texture, or severe environmental noise, potentially reducing diversity in highly challenging conditions. Third, the drone imagery comes from specific acquisition settings and may not capture the full variability of UAV platforms, sensors, or flight trajectories used in real-world operations. Finally, the textual descriptions are generated by GPT-4.0 following a controlled instruction protocol, which may introduce stylistic regularities or semantic biases that affect downstream language–vision tasks.

### G.2. Potential Negative Societal Impact

GeoLoc, like other geo-referenced vision datasets, may pose potential risks if misused. The dataset contains imagery tied to real-world geographic locations, and improper use could enable unauthorized geo-identification or raise privacy concerns, especially in sensitive or restricted regions. Although all street-view and satellite images come from publicly accessible Google services and adhere to Google's usage policies, the combination of multi-view data may still inadvertently reveal structural or environmental details that could be exploited for surveillance or adversarial location inference. Furthermore, models trained on GeoLoc might be adapted for applications beyond their intended scope, such as invasive tracking or monitoring without consent. Care should therefore be taken to use the dataset responsibly, ensuring compliance with local regulations, Google's licensing terms, and ethical guidelines for geographic and visual data.

## References


[1] Zhongwei Chen, Zhao-Xu Yang, and Hai-Jun Rong. Multilevel embedding and alignment network with consistency and invariance learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–15, 2025. 2

[2] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 2

[3] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. 10

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 2

[5] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 1

[6] Tianrui Shen, Yingmei Wei, Lai Kang, Shanshan Wan, and Yee-Hong Yang. Mccg: A convnext-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1456–1468, 2024. 2

[7] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 2

[8] Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, Yingying Zhang, Jiangwei Lao, Jingdong Chen, Ming Yang, and Yongjun Zhang. Cross-view geo-localization with panoramic street-view and vhr satellite imagery in decentrality settings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 227:1–11, 2025. 2

[9] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 2

[10] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. In *European Conference on Computer Vision*, pages 74–90. Springer, 2024. 2