From Street to Orbit: Training-Free Cross-View Retrieval via Location Semantics and LLM Guidance

Jeongho Min[1, *], Dongyoung Kim[2, *], Jaehyup Lee[3, †]
[1]Ulsan National Institute of Science and Technology (UNIST), South Korea
[2]Electronics and Telecommunications Research Institute (ETRI), South Korea
[3]Kyungpook National University, South Korea
jeongho.min@unist.ac.kr, dongyoung.kim@etri.re.kr, jaehyuplee@knu.ac.kr

Abstract

Cross-view image retrieval, particularly street-to-satellite matching, is a critical task for applications such as autonomous navigation, urban planning, and localization in GPS-denied environments. However, existing approaches often require supervised training on curated datasets and rely on panoramic or UAV-based images, which limits real-world deployment. In this paper, we present a simple yet effective cross-view image retrieval framework that leverages a pretrained vision encoder and a large language model (LLM), requiring no additional training. Given a monocular street-view image, our method extracts geographic cues through web-based image search and LLM-based location inference, generates a satellite query via geocoding API, and retrieves matching tiles using a pretrained vision encoder (e.g., DINOv2) with PCA-based whitening feature refinement. Despite using no ground-truth supervision or finetuning, our proposed method outperforms prior learning-based approaches on the benchmark dataset under zero-shot settings. Moreover, our pipeline enables automatic construction of semantically aligned street-to-satellite datasets, which is offering a scalable and cost-efficient alternative to manual annotation. All source codes will be made publicly available at https://jeonghomin.github.io/street2orbit.github.io/.

1. Introduction

Cross-view image retrieval, especially *Street-to-Satellite Retrieval*, has emerged as a critical task in various applications such as autonomous navigation, urban planning, and location-based services [5, 18, 23, 29, 32]. Given a street-level image, the goal is to accurately retrieve its corresponding satellite view image of the same location, despite large variations in viewpoint, scale, and appearance. This functionality is particularly valuable in GPS-denied or weak-signal environments such as urban canyons, remote areas,

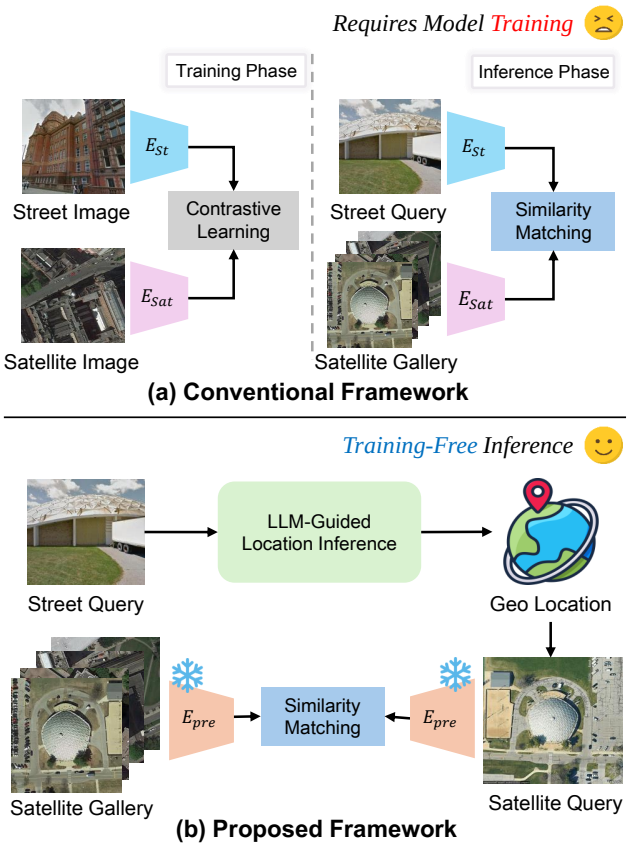

Figure 1. Comparison between conventional contrastive learning-based cross-view retrieval (top) and our proposed training-free, LLM-guided framework (bottom). Unlike prior work requiring separate encoders and contrastive training, our approach uses a single pretrained vision encoder (e.g., DINOv2) for both views, enabling direct retrieval in a shared feature space.

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

or disaster zones, where satellite imagery can serve as the only reliable reference for geo-localization [32].

This cross-view retrieval remains inherently challenging due to the extreme domain gap and disparity between ground and satellite viewpoints. Viewpoint distortion, lighting changes, and background clutter introduce a substantial misalignment in the visual feature space [18, 23].

To overcome these challenges, prior works adopt supervised learning approaches trained on curated datasets with strong geometric alignment (see Figure 1 (a)). These methods typically utilize panoramic or UAV-captured images [12, 30, 35], enabling dual encoders or transformer-based alignment modules to learn cross-view correspondence with the assumption that the model would be given the ideal input conditions, such as full 360-degree panoramas or low-altitude drone imagery.

Such requirements are difficult to satisfy in practice, by highly relying on ideal input assumptions and paired supervision, limiting real-world deployment. Panoramic or UAV imagery requires specialized equipment and constrained acquisition setups, making it unsuitable for general users. In contrast, monocular images from smartphones are far more accessible, lightweight, and practical for real-world deployment. They also reduce computational overhead, enabling efficient inference on mobile and edge devices.

However, monocular photos captured in the wild present additional challenges. They often contain occlusions, poor lighting, or ambiguous structures, which degrade the performance of models trained on curated and aligned inputs. In addition, constructing large-scale, diverse paired datasets for supervised training remains costly and time-consuming.

To overcome these limitations, we propose a training-free, LLM-guided cross-view image retrieval framework that requires no feature-level alignment or supervised training (see Figure 1 (b)). Our approach takes a single street-view photo as input, infers its geolocatable semantics using a large language model (specifically, Mistral 7B [8]), converts the inferred location to latitude and longitude coordinates through geocoding, and generates the corresponding satellite query via map APIs from a pre-indexed satellite gallery. The retrieved satellite image is then encoded using a large-scale pretrained vision encoder [4, 6, 17, 33] without any additional fine-tuning, while PCA-whitening is applied to suppress low-level artifacts such as lighting and texture biases in the embedding space, thereby improving retrieval robustness [2, 16, 34]. Through extensive experiments, we demonstrate that this embedding refinement consistently improves zero-shot performance not only in the satellite domain but also for general image retrieval tasks.

Moreover, our framework is designed to flexibly leverage spatial metadata, coordinate systems, and geocoding APIs, making it highly practical and generalizable for unconstrained monocular street photos in real-world scenarios. In particular, our method achieves state-of-the-art performance in the challenging Street-to-Satellite retrieval task of the University-1652 benchmark [35], despite relying solely on single-view queries without drone imagery or paired supervision. Furthermore, the proposed pipeline naturally enables automatic construction of Street-to-Satellite image pairs, allowing scalable dataset generation without manual annotation. This provides a practical foundation to overcome the limitations of existing benchmarks and facilitates broader data accessibility for future learning-based cross-view research.

Our main contributions are summarized as follows:

- We present a training-free and real-world photo compatible retrieval framework that infers location semantics from monocular street-view images using LLMs, without requiring paired supervision.
- We introduce a lightweight embedding refinement using PCA-whitening, which improves retrieval robustness in zero-shot settings.
- Our method achieves state-of-the-art performance on University-1652, surpassing existing models under both zero-shot and retrieval-based localization settings, while also enabling scalable dataset construction through automatic street-satellite pairing.

2. Related Works

2.1. Cross view image retrieval

Cross-view image retrieval aims to retrieve images of the same location captured from drastically different perspectives, typically street-level views against top-down satellite images. To tackle this challenging task, prior studies have proposed various methods that utilize cross-view visual features to retrieve the most semantically relevant image pairs [5, 27, 29, 31].

Early efforts modeled the problem using Siamese CNN architectures [9] that minimized feature-space distances between paired ground-to-satellite images [12, 21, 22]. However, the drastic domain shift, which includes viewpoint distortions, occlusions, and inconsistent scale, limits the performance of purely CNN-based models.

To address this, LPN [27] proposed part-based attention mechanisms that divide ground-level inputs into spatial regions and aggregate them using learned weights, effectively emphasizing informative subregions. While this improved feature discriminability, the fundamental gap between ground and aerial perspectives remained substantial.

Subsequent work introduced intermediate UAV imagery to bridge the domain gap. For instance, PLCD [32] jointly trains on drone-to-satellite and ground-to-satellite retrieval tasks, treating drone views as a natural intermediary. This

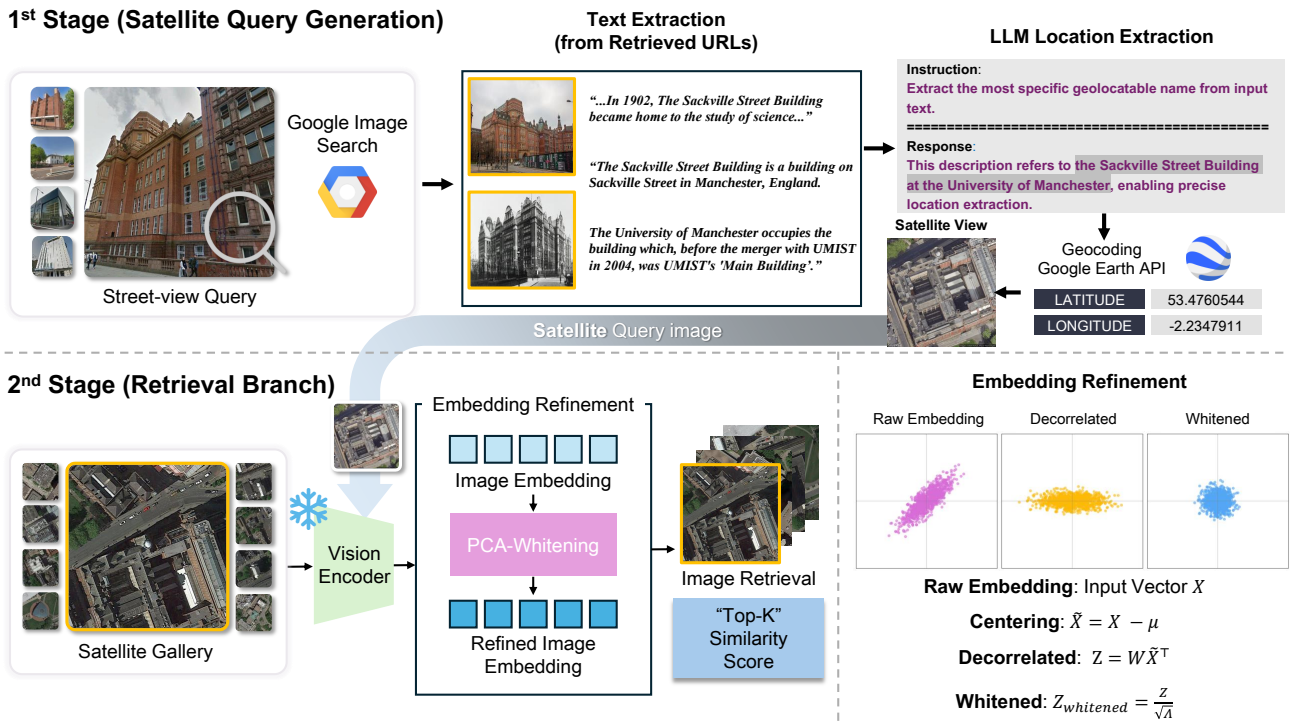

Figure 2. Overview of our proposed training-free cross-view retrieval pipeline. Given a street-level image, we first extract textual context using Google Image Search, then infer the most specific geolocatable name via a large language model(LLM). This name is geocoded into coordinates to retrieve a satellite tile, which is embedded using a pretrained vision encoder (e.g., DINOv2). The resulting feature is whitened and matched against a satellite gallery using similarity-based retrieval. Our method enables semantically aligned cross-view matching without any supervised training.

tri-view strategy reduces the viewpoint discrepancy and improves retrieval robustness.

Beyond tri-view architectures, drone-centric retrieval frameworks have recently emerged as a distinct research direction that fully substitutes street-level inputs with UAV imagery. DWDR [28] enhances robust matching by using diverse rotation augmentations and local patches of Drone images. MCCG [20] enforces cross-view consistency through a multi-classifier scheme based on ConvNeXt backbones, while Sample4Geo [5] simplifies the retrieval pipeline by adopting symmetric contrastive learning using GPS-aligned positives and visually hard negatives, without relying on polar transformations or dense alignment.

However, retrieving corresponding satellite views from a single, narrow field-of-view street image remains extremely challenging due to severe viewpoint and scale disparities. To mitigate this, recent studies such as MFRGN [29] and Panorama-BEV [31] attempt to bridge the domain gap by leveraging panoramic street images or transforming them into Bird's-Eye View (BEV) representations. These methods reduce the perspective discrepancy by expanding the field of view or generating intermediate geometric representations better aligned with satellite perspectives.

Although effective, these methods typically assume access to panoramic inputs, low-altitude drone imagery, or well-aligned assumptions about well-aligned training data that often do not hold in practical scenarios. In contrast, our approach embraces single-view, in-the-wild monocular photos and removes the need for supervised training or explicit geometric conversion, making it more scalable and applicable to real-world deployments.

2.2. Image Encoder

Progress in vision-language modeling and self-supervised learning has led to powerful image encoders that generalize well across domains and tasks. In particular, transformer-based encoders [4, 6, 13, 14] have shown strong capabilities in learning semantically rich representations.

ViT (Vision Transformer) [6] processes images as tokenized patches and models long-range dependencies via self-attention. These global features make ViT-based backbones a natural fit for high-level retrieval and recognition tasks. CLIP [17] introduces contrastive learning over image-text pairs at scale, resulting in a joint embedding space that supports zero-shot classification and retrieval. However, its effectiveness deteriorates in domains like satellite or aerial imagery, where the distribution of content diverges significantly from that of natural images.

To bridge this gap, RemoteCLIP [11] retrains the CLIP architecture on domain-specific satellite and UAV datasets,

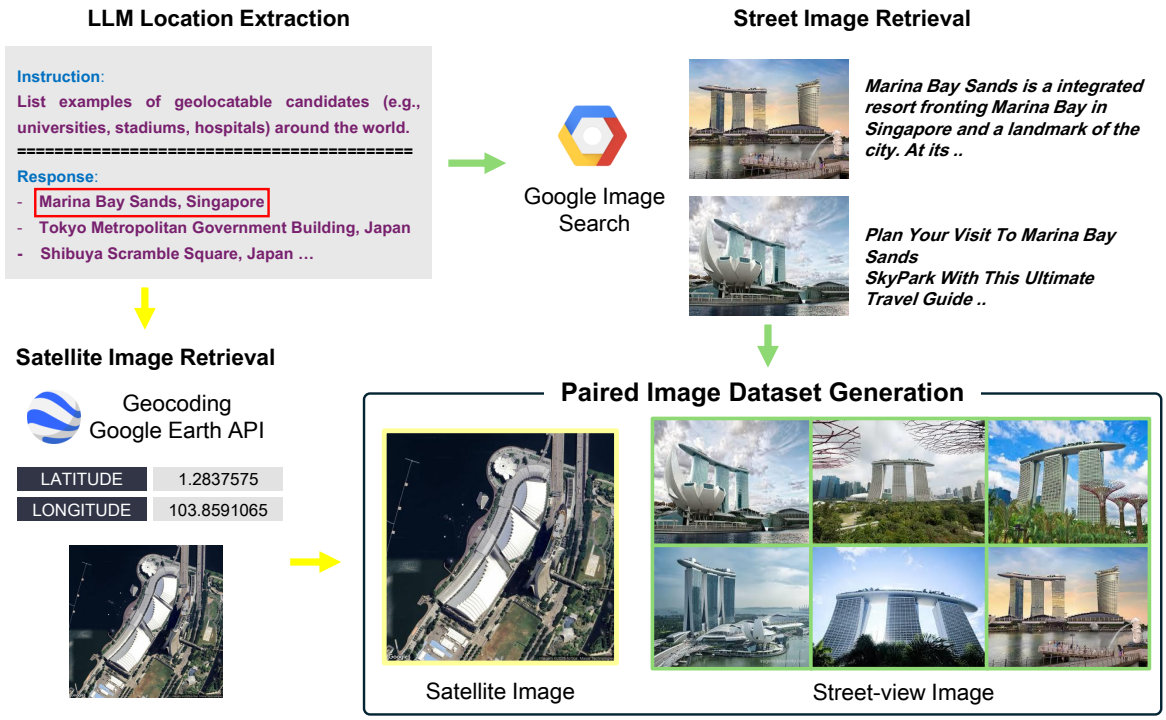

Figure 3. Overview of our dataset generation pipeline. We prompt an LLM to produce a list of globally recognizable locations (e.g., "Marina Bay Sands", "Tokyo Metropolitan Government Building"). For each location, we retrieve street-view images via Google Image Search and obtain geographic coordinates using a Google Geocoding API. These coordinates are used to generate corresponding satellite tiles through the Google Static Maps API. This process produces high-quality, semantically aligned street-to-satellite image pairs at scale.

improving both semantic fidelity and alignment in remote sensing contexts. In addition, DINO [4] introduces a ViT-based self-distillation framework that iteratively transfers knowledge between student and teacher networks, yielding robust frozen features that generalize well across tasks. Building on this, DINOv2 [15] scales up the training with a larger curated dataset and incorporates techniques such as SwAV [3] centering and the KoLeo regularizer [19] to achieve semantic representations comparable to weakly-supervised models, despite being purely self-supervised. Meanwhile, SigLIP [33] modifies the CLIP architecture by replacing the contrastive loss with a binary classification loss, mitigating label imbalance and improving generalization to diverse image-text pairs. SigLIP v2 [25] extends this framework with deeper ViT backbones, longer pretraining, and fine-grained tuning, achieving stronger multimodal zero-shot performance.

In this work, we leverage these state-of-the-art pretrained vision encoders to encode satellite imagery in a training-free setup. Its frozen representations, combined with coordinate-guided search and language-driven semantic inference, allow for robust retrieval in the absence of paired supervision or domain-specific fine-tuning.

3. Method

We propose a novel training-free, language-guided cross-view retrieval framework designed to handle real-world monocular photos without requiring any paired supervision. As illustrated in Figure 2, our pipeline consists of three main components: (i) Satellite query Generation via location semantics, (ii) Visual Embedding and Similarity-based Retrieval, and (iii) Embedding Refinement for robust matching. Furthermore, this pipeline naturally enables large-scale Photo-to-Satellite dataset construction with minimal manual effort.

3.1. Satellite Query Image Generation via Location Semantics

Location-Centric Semantic Extraction Given an unconstrained street-view image, we leverage the Google Image Search interface to retrieve surrounding contextual information such as page metadata, descriptive captions, text content, and related keywords. This multimodal signal forms a noisy but information-rich description from which geospatial semantics can be inferred. For batch queries, descriptive texts across all candidate scenes are concatenated into a single prompt to increase contextual coherence. At this time, url's metadata, web document body text, and text information around the image collected from the search result are

used together. If there are several Image queries, the information corresponding to each photo scene is collected in the description at once.

LLM-based Location Inference To extract the most specific geolocatable name (e.g., landmark, buildings, or administrative regions), we prompt a large language model [1, 7, 24] with the aggregated textural context. The prompt is designed to prioritize physical place names over abstract or event-based descriptors. This LLM-based inference effectively filters out irrelevant noise while capturing semantically meaningful references suitable for geocoding.

We employ **Mistral 7B** [8] for this task due to its strong performance on reasoning and language understanding benchmarks, surpassing even larger models such as LLaMA 2 13B and 34B [24] across tasks like commonsense QA, math, and reading comprehension. Additionally, its use of grouped-query attention (GQA) and sliding window attention (SWA) ensures fast inference and efficient memory usage—an essential factor in real-time geolocation pipelines. As an open and lightweight model, Mistral 7B offers an optimal balance of performance, efficiency, and deployability in practical applications.

This step plays a critical role in disambiguating verbose or indirect location mentions. For example, as illustrated in Figure 2, a complex input such as:

> "In 1902, The Sackville Street Building became home to the study of science..."
> "The Sackville Street Building is a building on Sackville Street in Manchester, England. ..."

is interpreted by the LLM as:

> *"This description refers to the **Sackville Street Building** at the University of Manchester, enabling precise location extraction."*

Such inference allows our system to confidently retrieve accurate geocoordinates via geocoding APIs, even when the input descriptions are indirect or historical. Through this semantic understanding, we ensure that the generated satellite queries faithfully reflect the true geographic intent behind the original image or description.

Geocoding and Satellite Image Generation. The location string inferred by the LLM is submitted to the Google Maps Geocoding API to convert it into geographic coordinates (i.e., latitude and longitude). Based on these coordinates, we retrieve a satellite image centered at the inferred location via Google Earth Engine and the Google Static Maps API with a fixed zoom level and resolution. The resulting satellite tile serves as the retrieval query proxy that bridges the domain gap between the original street-view image and satellite-view, and is used as input in the retrieval branch.

3.2. Retrieval Branch with Visual Embedding and Similarity Matching

Feature Encoding and Retrieval. The satellite query image, along with a pre-indexed satellite gallery, is processed using the same frozen vision encoder. In our experiments, we adopt DINOv2 as the backbone, extracting 768-dimensional global features. For retrieval, cosine similarity is computed between the query embedding and gallery entries, and top-k nearest candidates are selected based on similarity scores.

Embedding Refinement via Whitening. Although DINOv2 provides semantically rich features, raw embeddings may still be affected by low-level artifacts such as illumination or texture biases. To address this issue, we newly introduce to apply a lightweight PCA-whitening procedure at inference time. Specifically, given an embedding matrix $X \in \mathbb{R}^{n \times d}$, we perform mean-centering:

$$\tilde{X} = X - \mu,$$

followed by projection via the top-k principal components $W \in \mathbb{R}^{k \times d}$:

$$Z = W\tilde{X}^\top.$$

and final whitened normalization:

$$Z_{\text{white}} = \Lambda^{-\frac{1}{2}} Z,$$

where $\Lambda \in \mathbb{R}^{k \times k}$ contains the principal eigenvalues. The resulting representation Z_{white} consists of decorrelated components with unit variance.

3.3. Scalable Photo-to-Satellite Dataset Construction

As illustrated in Figure 3, our framework enables automatic construction of aligned street-to-satellite pairs by leveraging LLM-based scene inference and satellite tile fetching. Given a set of seed place names (e.g., "Marina Bay Sands", "Colosseum"), we use the Google Image API to collect diverse real-world street-view photos. These are then paired with satellite images generated via the same geocoding pipeline described in Section 3.1.

Note that this automated pairing process not only removes the need for human annotation but also supports scalable expansion across diverse regions. The resulting dataset contains semantically aligned image pairs that can serve as supervision for future learning-based models, enabling broader generalization beyond limited benchmark regions.

4. Experiments

4.1. Dataset

We evaluate out method on the University-1652 benchmark [35], a large-scale dataset designed for cross-view

Model	Drone	R@1	R@5	R@10	R@1%
LPN [27]	✗	0.43	1.59	2.87	3.02
LPN [27]	✓	1.28	3.84	6.59	6.98
Swin-B + DWDR [14, 28]	✗	0.35	2.33	3.88	4.30
MCCG [20]	✗	0.93	3.30	6.01	6.55
Sample4Geo [5]	✗	1.35	4.76	7.63	8.02
PLCD [32]	✓	6.86	14.39	18.50	19.15
Ours (DINOv2-Large)	✗	22.84	33.48	37.93	37.64
+ Refinement (256D)	✗	**25.57**	**37.21**	**40.66**	**39.94**

Table 1. **Retrieval performance on University-1652 (Street→Satellite).** Top-k recall metrics are reported. '✓' indicates drone images were used in training, '✗' indicates only ground-satellite views were used.

geo-localization. It contains 1,652 buildings from 72 global universities, captured from three viewpoints: street, drone, and satellite. The training dataset comprises 701 buildings from 33 universities, while the remaining 951 buildings from 39 universities are used for test, ensuring no overlap.

Each building is associated with street-view images collected from Google Street View and web image search, drone-view images captured along spiral flight trajectories, and high-resolution satellite images. Among the supported retrieval tasks, we focus on the most challenging scenario: Street-to-Satellite retrieval, which involves matching a monocular ground-level image to its corresponding satellite view under significant viewpoint and modality shifts.

4.2. Implementation Details

In this study, we use web-based automation (Selenium) to collect relevant textual metadata from Google Image Search and apply LLM-based location inference using Mistral 7B (v0.3) [8]. The extracted place names are then converted into geo-coordinates via the Google Geocoding API (v1), and satellite query images are generated centered on these coordinates from Google Static Maps API (v1).

For the retrieval stage, we utilize pretrained vision encoders (DINOv2, CLIP, RemoteCLIP, etc.) and perform inference without any additional fine-tuning. All experiments were conducted using only the first image in each query set, assuming one image per folder setting. All embeddings are further normalized through a simple PCA-whitening process and used for final similarity-based matching.

For fair comparison, baseline models based on contrastive learning (LPN, DWDR, MCCG, Sample4Geo, PLCD) were reproduced using the official PyTorch implementations and hyperparameters provided in each paper. All experiments were conducted on a single NVIDIA RTX A6000 GPU under the same hardware environment.

4.3. Evaluation Metric

To evaluate retrieval performance, we use the standard **Recall@k** metric. Given a query image, Recall@k measures the fraction of queries for which a relevant image is retrieved within the top-k candidates:

Model	R@1	R@5	R@10	R@1%
Remote-CLIP (ViT-B/32)	2.30	6.03	9.20	8.19
Remote-CLIP (ViT-L/14)	2.44	5.89	7.90	7.61
Remote-CLIP (RN-50)	1.87	5.60	8.76	8.05
CLIP (ViT-B/32)	7.18	15.37	19.68	19.40
CLIP (ViT-L/14)	10.49	22.99	30.03	28.59
CLIP (RN-50)	4.31	11.49	16.09	15.37
DINOv2-Base	22.27	32.33	36.93	36.64
+ Refinement (512D)	**26.01**	34.20	35.34	35.20
+ Refinement (256D)	23.28	**33.62**	**37.93**	**37.21**
DINOv2-Large	22.84	33.48	37.93	37.64
+ Refinement (512D)	**26.29**	35.92	39.80	39.08
+ Refinement (256D)	25.57	**37.21**	**40.66**	**39.94**

Table 2. **Retrieval performance on University-1652 (Street→Satellite).** Top-k and Top-1% recall are reported. Our method, without any supervised training, achieves state-of-the-art results, and the PCA-whitening refinement "Refinement" further boosts performance across various pretrained encoders.

Recall@k is defined as follows:

$$\text{R@}k = \frac{\text{Number of relevant items in top } k}{\text{Total number of relevant items}} \quad (1)$$

In this equation, the numerator represents the number of relevant items retrieved in the top-k results for each query, and the denominator denotes the total number of relevant items for that query. By varying k, Recall@k provides a comprehensive view of the retrieval system's performance under different ranking thresholds. In this paper, we report **R@1**, **R@5**, **R@10**, and **R1%** as main evaluation metrics.

4.4. Cross-view Model Evaluation Results

Table 1 shows comparison of the retrieval results of the Street-to-Satellite on the University-1652 benchmark. Our proposed training-free framework achieves a Top-1 recall of **22.84%** using DINOv2-Large, outperforming several supervised methods such as LPN [27], DWDR [28], MCCG [20], and Sample4Geo [5], which rely on curated training data and a complex alignment process.

Notably, PLCD [32], which leverages drone images as an intermediate view, achieves 6.86% Top-1 accuracy. In contrast, our method surpasses it by more than 3×, without relying on drone input or any additional training, underscoring the strength of our language-guided query generation and pretrained encoder. This demonstrates that our method can outperform existing supervised approaches even without auxiliary views such as drones.

In addition, applying our PCA-whitening refinement (dimensional reduction from 768 to 256) further boosts the Top-1 score to 25.57, demonstrating that decorrelating and normalizing features enhance retrieval robustness in zero-shot conditions. Overall, these results suggest that our

Model	ILIAS				University-1652 (Street→Satellite)			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
ViT-base	28.10	42.00	49.60	68.40	5.42	12.70	18.12	17.26
+ Refinement (D-512)	**30.80**	**47.00**	**54.40**	**70.60**	**9.27**	19.54	23.82	23.11
+ Refinement (D-256)	28.40	44.00	52.50	69.80	7.99	**19.69**	**24.96**	**23.25**
CLIP-base	47.70	67.60	74.40	84.40	13.12	24.11	**31.95**	**30.67**
+ Refinement (D-512)	53.20	73.10	80.20	90.10	11.13	19.12	24.25	23.54
+ Refinement (D-256)	**57.30**	**76.10**	**82.50**	**91.80**	**14.27**	**26.53**	31.24	**30.67**
DINOv2-base	**54.70**	68.90	75.10	85.60	12.84	24.82	30.96	30.10
+ Refinement (D-512)	53.00	**70.00**	**77.30**	**87.00**	**13.27**	24.68	30.39	29.67
+ Refinement (D-256)	51.30	67.60	74.90	86.80	13.12	**26.68**	**34.38**	**32.95**
SigLIP-base	68.90	83.60	**88.30**	**94.10**	11.41	20.40	29.24	27.96
+ Refinement (D-512)	**70.40**	83.20	87.50	93.50	13.69	24.54	30.81	29.96
+ Refinement (D-256)	**70.40**	**83.70**	**88.30**	**94.10**	**14.12**	**27.25**	**34.52**	**32.67**
SigLIP2-base	**68.80**	81.50	87.10	94.10	19.54	**37.09**	44.94	43.79
+ Refinement (D-512)	68.10	80.80	85.50	92.40	18.40	31.67	37.38	36.66
+ Refinement (D-256)	68.40	**83.50**	**88.40**	**94.90**	**20.97**	36.52	**45.22**	**44.51**

Table 3. **Ablation study on ILIAS and University-1652 datasets.** We evaluate the effect of embedding refinement (PCA-whitening), input variation, and encoder choices on retrieval and localization performance. Top-k retrieval and Top-1% localization accuracy are reported. **Bold** highlights the best result in each block.

Distance Threshold (km)	0.5 km	1.0 km	2.0 km	5.0 km
Accuracy (%)	54.57	60.57	64.14	65.00
Matched Samples (out of 700)	382	424	449	455

Table 4. **The first stage location inference accuracy.** Percentage of queries whose inferred coordinates fall within the specified distance thresholds from the GT in University-1652.

training-free framework provides a practical and efficient solution for cross-view retrieval, well-suited for real-world applications.

4.5. Location Inference Accuracy

The first stage of our proposed framework infers geo-coordinates from a street-view by extracting semantic cues with an LLM and converting them into latitude-longitude pairs via a geocoding API. We assess this step using 700 queries from University-1652, measuring distance between inferred and GT coordinates with Haversine formula [26]. Accuracy is reported across multiple distance thresholds.

As summarized in Table 4, the first stage achieves 54.75% accuracy within 0.5 km and 65% within 5 km. These results demonstrate that the first stage of our framework provides reliable localization performance even without any visual matching.

4.6. Evaluation Across Diverse Visual Encoders

To examine the generalizability of our approach, we evaluate multiple vision encoders under the same retrieval framework, as shown in Table 2. The compared encoders include CLIP [17], RemoteCLIP [11], and DINOv2 [15]. CLIP-based models perform moderately well, with ViT-L/14 outperforming other CLIP variants, including ResNet-50 and ViT-B/32. Interestingly, RemoteCLIP, despite its domain-specific pretraining, underperforms compared to CLIP, suggesting limited cross-view transferability.

In contrast, self-supervised vision encoders such as DINOv2 show superior performance compared to CLIP-based models. Additionally, as shown in Table 2, embedding refinement using PCA-whitening consistently improves performance for both DINOv2-Base and DINOv2-Large models through dimensionality reduction. Especially, the Top-1% score increases up to 39.94, indicating enhanced similarity matching and improved domain generalization in retrieval tasks. These results demonstrate that our framework can be effectively applied across different encoders and consistently improves performance in a training-free manner.

4.7. Ablation Study

Table 3 presents the ablation study results that validate the effectiveness of the proposed PCA-whitening embedding refinement. We conduct experiments on both a general natural image retrieval dataset, **ILIAS** [10], and a cross-view geo-localization benchmark, **University-1652** [35], which focuses on the UAV→Satellite retrieval scenario.

The results show that applying our simple refinement consistently improves retrieval performance across most pretrained vision encoders. Notably, the benefit is more pronounced in **University-1652** benchmark, which involves a

Figure 4. Qualitative results of our training-free retrieval framework. Given a street-level query image (left), our method retrieves the most semantically aligned satellite tile (right) using a pipeline combining Google Image Search, an LLM-based geolocation inference, geocoding, and DINOv2 with PCA-whitening. The red bounding boxes in each satellite image indicate the correctly retrieved region corresponding to the inferred location. The top row shows accurate retrieval of a domed building with distinctive geometry; the second row demonstrates robustness in a greenery scene with minimal context, such as a house; the third row highlights generalization to an urban street with ambiguous visual representations. These samples illustrate our proposed framework's ability to perform reliable cross-view matching across diverse environments, without any kind of supervised training.

significant domain gap between street and satellite views. For example, DINOv2-base encoder achieves a Top-1 Recall of 12.84% without refinement, increasing to 13.27% and 13.12% with refinement (D-512, D-256), respectively.

A similar trend is observed on the general retrieval benchmark **ILIAS**. For instance, CLIP-base improves from 47.70% Top-1 Recall without refinement to 57.30% with refinement (D-256). Even strong basedline encoders, such as SigLIP and SigLIP2, maintain or slightly improve their Top-1 accuracy above 70% with refinement applied.

These findings suggest that our simple PCA-whitening embedding refinement effectively stabilizes feature representations in zero-shot scenarios, yielding consistent performance gains not only for complex cross-view geolocalization tasks but also for standard natural image retrieval pipelines, without any additional training.

4.8. Qualitative Results

Figure 4 presents qualitative Top-5 retrieval results of our proposed training-free framework applied to the University-1652 dataset. For each street-view query image, our method first infers the location name using an LLM and retrieves the corresponding satellite query through geocoding. The generated query is then matched against the satellite gallery in the retrieval branch, and the Top-k most similar candidates are retrieved based on the refined DINOv2 embeddings.

Due to the nature of our LLM-based satellite query generation pipeline, there may be cases where the generated query slightly differs from the ground-truth satellite image in the dataset. For example, as shown in the fourth row on the left side of Figure 4, the query may include a coordinate shift of the same building, or as in the second and fourth rows on the right side, the satellite query may reflect a different capture condition (e.g., season, weather, or time) than the ground truth, which can lower the Top-k ranking.

Nevertheless, even without any additional training or pairwise supervision, our framework successfully identifies the correct matching region (highlighted with a red bounding box) among the top retrieved results, visually demonstrating strong cross-view retrieval performance.

5. Conclusion

We present a training-free framework for Street-to-Satellite retrieval, integrating LLM-guided location inference, pretrained vision encoders, and geocoding APIs to generate semantically aligned satellite queries. Without training or auxiliary data, our framework achieves competitive performance using a single street-view image dataset.

Extensive experiments show that our approach not only outperforms the existing contrastive learning-based methods but also sets a new state-of-the-art on the University-1652 benchmark in the Street-to-Satellite retrieval. We further show that PCA-whitening consistently enhances retrieval robustness across various pretrained encoders in zero-shot settings, effectively bridging modality gaps. While our proposed method depends on the external APIs and LLMs, it enables scalable and automated construction of Street-to-Satellite image pairs, offering a practical foundation for future learning-based retrieval models.

Acknowledgement

This work was supported by the Research Program of the Electronics and Telecommunications Research Institute (ETRI), funded by [25ZD1120, Development of ICT Convergence Technology for Daegu-GyeongBuk Regional Industry].

This research was supported by the Regional Innovation System & Education(RISE) Glocal 30 program through the Daegu RISE Center, funded by the Ministry of Education(MOE) and the Daegu, Republic of Korea.(2025-RISE-03-001)

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015. 2

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 4

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 3, 4

[5] 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. 1, 2, 3, 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[8] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. Diego de las casas. *Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed*, pages 50–72, 2023. 2, 5, 6

[9] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, pages 1–30. Lille, 2015. 2

[10] Giorgos Kordopatis-Zilos, Vladan Stojnić, Anna Manko, Pavel Suma, Nikolaos-Antonios Ypsilantis, Nikos Efthymiadis, Zakaria Laskar, Jiri Matas, Ondrej Chum, and Giorgos Tolias. Ilias: Instance-level image retrieval at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14777–14787, 2025. 7

[11] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3, 7

[12] 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. 2

[13] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 3

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 6

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 7

[16] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 3, 7

[18] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 1, 2

[19] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018. 4

[20] 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, 2023. 3, 6

[21] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[22] Yuxin Tian, Xueqing Deng, Yi Zhu, and Shawn Newsam. Cross-time and orientation-invariant overhead image geolocalization using deep local features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2512–2520, 2020. 2

[23] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 1, 2

[24] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[25] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 4

[26] Glen Van Brummelen. Heavenly mathematics: The forgotten art of spherical trigonometry. 2012. 7

[27] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021. 2, 6

[28] Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Chenggang Yan, and Yi Yang. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3, 6

[29] Yuntao Wang, Jinpu Zhang, Ruonan Wei, Wenbo Gao, and Yuehuan Wang. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2574–2583, 2024. 1, 2, 3

[30] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 2

[31] 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, 3

[32] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia*, 25:2176–2188, 2022. 1, 2, 6

[33] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2, 4

[34] Bo-Jian Zhang, Guang-Hai Liu, and Zuoyong Li. Image retrieval using unsupervised prompt learning and regional attention. *Expert Systems with Applications*, 247:122913, 2024. 2

[35] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020. 2, 5, 7