CLIP-based Image Geolocation using Hierarchical Feature Learning and RAG

Akshay Raman
New York University
New York, USA
ar8692@nyu.edu

Aman Gupta
New York University
New York, USA
ag9900@nyu.edu

Satyanarayana Chillale
New York University
New York, USA
sc9960@nyu.edu

Srikanth Balakrishna
New York University
New York, USA
sb9558@nyu.edu

Prithviraj Murthy
New York University
New York, USA
prithviraj.murthy@nyu.edu

***Abstract*—Worldwide Geolocation aims to determine the precise location of images taken anywhere on Earth. Traditional deep learning methods are impractical due to the immense variation in geographical landscapes and prediction on a global scale. In this project, we train a CLIP-based image geolocation model on the MediaEval-16 dataset and explore various inference approaches with a focus on efficiency and scalability. Current approaches (GeoCLIP) perform inference by searching through an entire gallery of GPS coordinates. To overcome this limitation, we propose an improved inference method that leverages hierarchical feature clustering at multiple geographical scales. By organizing the GPS gallery into a tree structure, we drastically reduce the search space. Our inference approach achieves comparable performance to GeoCLIP while being 100x more efficient than previous methods, making it more practical for large-scale geolocation tasks. Furthermore, we experimented with retrieval-augmented generation (RAG)-based inference using multiple Large Language Models (LLMs), such as GPT-4o and Mistral. We trained our model on the MP-16 Pro dataset and evaluated its performance on the IM2GPS3k benchmark, demonstrating its effectiveness. The code is available on GitHub: Hierarchical GeoCLIP, G3**

I. Introduction

Geolocalization from visual data is a challenging and intricate task requiring systems to predict geographic coordinates solely from visual cues, such as landscapes, vegetation, architectural structures, and environmental details. Games like GeoGuessr highlight human capabilities to analyze such cues intuitively; however, replicating this process in machine learning models has proven difficult, particularly for unseen or ambiguous images. The diversity of global environments, coupled with challenges in lighting variations and obstructions, creates a complex generalization problem for machine learning systems.

Recent advancements in deep learning and vision-language models (VLMs) have paved the way for improved geolocation performance. Traditional approaches leveraging Convolutional Neural Networks (CNNs), such as AlexNet and ResNet, demonstrate the ability to extract fundamental geographic features but often struggle with fine-grained predictions. On the other hand, Vision-Language models, such as CLIP (Contrastive Language-Image Pretraining), have enabled improved semantic understanding by aligning textual and visual features.

In this work, we build on the GeoCLIP framework [1], a CLIP-inspired Image-to-GPS retrieval approach that enforces alignment between images and their corresponding GPS locations. Unlike GeoCLIP that relies on flat GPS galleries, we introduce a hierarchical tree structure that organizes GPS coordinates into multi-level geographic groups (e.g., neighborhoods, cities, countries). This approach drastically reduces the potential search space and more scalable for large GPS galleries.

We also explore RAG-based inference techniques and experiment with various large language models (LLMs), such as GPT-4o, Mistral, and LLaMA. By incorporating text embeddings into the inference pipeline, we assess the role of textual information in prediction performance. To further enhance accuracy, we add neighborhood parameters to the textual information, enabling the model to make use of fine-grained geographical context.

We summarize our main contributions as follows:

- Trained a CLIP-based image geolocation model on the MediaEval-16 Dataset (4M+ images).
- Designed a novel inference approach using hierarchical feature clustering that drastically reduces the search space and improves overall efficiency.
- Explore RAG-based inference techniques that leverage LLMs to incorporate textual information.

II. Related Work

Image geolocation has emerged as an essential task in computer vision, enabling applications such as navigation, crime tracking, and environmental monitoring. Numerous studies have tackled this problem using diverse approaches, ranging from dataset creation to advanced geolocation methodologies.

Large-scale datasets like ***MP16-Pro*** [2] and benchmarks dataset like ***IM2GPS*** [3] and have been instrumental in advancing geolocation research. The ***IM2GPS*** dataset serves as a benchmark for evaluating geolocation models, providing test images from diverse global environments to assess model generalization and accuracy. Additionally, the ***MP16-Pro*** dataset facilitates training by offering extensive geo-tagged images enriched with textual geographical descriptions. This multimodal design supports models in learning the interplay between visual, textual, and GPS data, making it a valuable resource for geolocation tasks.

PlaNet [4] reframes geolocation as a classification problem, dividing the Earth's surface into multi-scale geographical cells

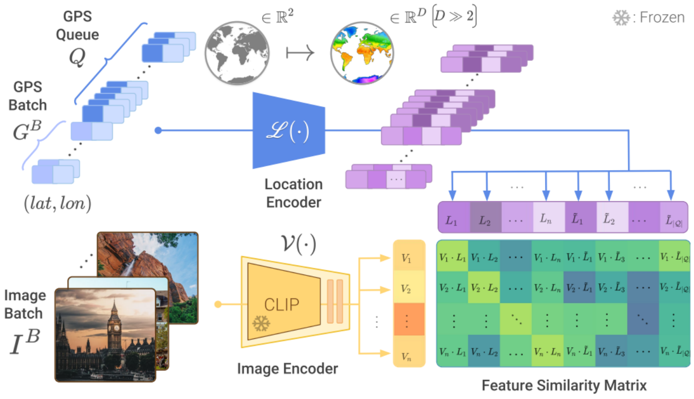

GPS
Queue
Q
GPS
Batch
G^B
(lat, lon)
$\in \mathbb{R}^2$
$\mapsto$
$\in \mathbb{R}^D \; [D \gg 2]$
: Frozen
$\mathscr{L}(\cdot)$
Location
Encoder
L_1 L_2 ... L_n $\tilde{L}_1$ $\tilde{L}_2$ ... $\tilde{L}_{|Q|}$
Image
Batch
I^B
$\mathcal{V}(\cdot)$
CLIP
Image Encoder
V_1 V_2 $\vdots$ V_n
$V_1 \cdot L_1$ $V_1 \cdot L_2$ ... $V_1 \cdot L_n$ $V_1 \cdot \tilde{L}_1$ $V_1 \cdot \tilde{L}_3$ ... $V_1 \cdot \tilde{L}_{|Q|}$
$V_2 \cdot L_1$ $V_2 \cdot L_2$ ... $V_2 \cdot L_n$ $V_2 \cdot \tilde{L}_1$ $V_2 \cdot \tilde{L}_3$ ... $V_2 \cdot \tilde{L}_{|Q|}$
$V_n \cdot L_1$ $V_n \cdot L_2$ ... $V_n \cdot L_n$ $V_n \cdot \tilde{L}_1$ $V_n \cdot \tilde{L}_3$ ... $V_n \cdot \tilde{L}_{|Q|}$
Feature Similarity Matrix

Fig. 1: GeoCLIP Model Architecture

and training convolutional neural networks (CNNs) to assign images to these cells. This hierarchical approach captures global and regional visual semantics, achieving robust geolocation performance across diverse landscapes.

Methods like **IM2City** [5] and **ETHAN** [6] complement these approaches by incorporating reasoning capabilities and contextual prompts, enabling fine-grained geolocation predictions. Additionally, techniques such as hierarchical classification and geocell partitioning, as employed by **PIGEON** [7], utilize clustering algorithms and semantic boundaries to balance class distributions and refine predictions.

Models such as CLIP and its adaptations, including **GeoCLIP** [1], have significantly advanced image geolocation by aligning images, text, and GPS data within a shared embedding space. GeoCLIP extends CLIP's capabilities by introducing a GPS encoder to transform geographical coordinates into high-dimensional representations, enabling the alignment of continuous and discrete geographical features. By leveraging positional encodings, GeoCLIP effectively bridges the gap between visual and geospatial data, achieving competitive geolocation performance.

Recent advancements in geolocation include **retrieval-augmented generation (RAG)** [8] frameworks, exemplified by **G3** [8]. G3 addresses the limitations of retrieval- and generation-based methods by introducing Geo-diversification, and Geo-verification. These techniques align images with textual descriptions and GPS data, generate robust candidate predictions through diverse prompts, and verify predictions using learned multimodal representations. G3 achieves state-of-the-art performance on well-established datasets such as **IM2GPS3k** and **YFCC4k** [9], demonstrating its robustness and scalability.

III. Methodology

A. GeoCLIP

GeoCLIP builds on the CLIP framework, which enables the alignment of images and text within a shared embedding space. It extends the original architecture by incorporating a GPS encoder to transform geographic coordinates (latitude, longitute) into high-dimensional representations. This joint embedding of visual, textual and geographic features forms the backbone of GeoCLIP, allowing effective cross-modal alignment and retrieval.

The image encoder consists of a pre-trained vision transformer, which we use as the backbone and keep it frozen. This model is augmented by two trainable linear layers, which are necessary for fine-tuning the image encoder.

The location encoder consists of two important components: (1) RFF Module and (2) MLP Module. The Random Fourier Features (RFF) module transforms low-dimensional GPS data (2 coordinates) into higher dimensions that capture high frequency information. This is done by applying a sinusoidal-based positional encoding to the GPS coordinates using frequencies sampled by a gaussian distribution. The resulting encoding is then passed to a multi-layer perceptron (MLP) which can be tuned during training. Additionally, by varying the σ parameter in RFF, we obtain multiple high-dimensional representations that correspond to different geographical scales which are aggregated to obtain the final location embedding.

For inference, the original geoclip paper matches the features of a query image against a gallery of 100K GPS embeddings. The algorithm then greedily selects the GPS embedding with the highest similarity. This approach is not very efficient or scalable as the image embedding is compared against *all* the GPS coordinates in the gallery.

B. Hierarchical feature clustering

We enhance the inference mechanism used by GeoCLIP by introducing a hierarchical image gallery. Unlike traditional flat galleries, which organize GPS coordinates as a single collection, our hierarchical structure groups them by geographic levels.

Our hierarchical feature clustering approach for inference is designed to improve the efficiency of GeoCLIP's original inference method while maintaining competitive accuracy. Instead of comparing query image embeddings against all the GPS embeddings in the gallery, we organize the GPS embeddings into a tree structure based on their hierarchical features. Our approach mimics human reasoning by first narrowing to a general area and then zooming in on specifics. It

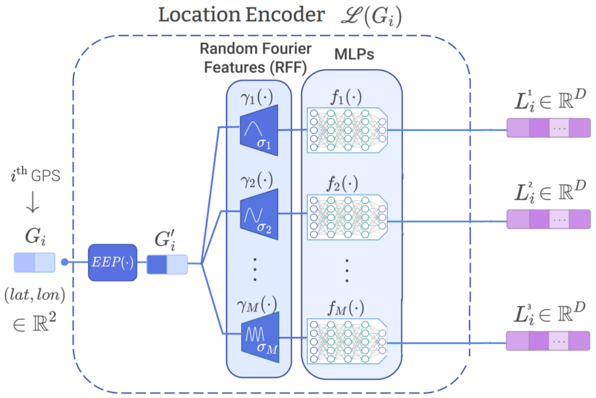


Fig. 2: Modified Location Encoder

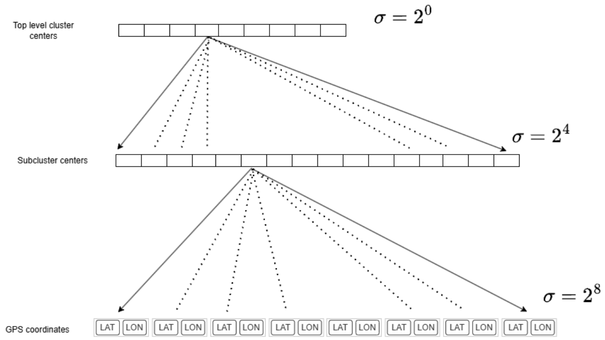


Fig. 3: GPS Gallery structure

reduces the number of candidate embeddings during similarity computation, speeding up the process.

During the encoding process, we generate embeddings for GPS coordinates using three separate RFF + MLP capsules, each configured with a different sigma value. These embeddings capture features at different geographic scales. The embeddings from each capsule are clustered separately to create hierarchical clusters. The GPS gallery is structured as a hierarchical tree, with cluster centers serving as nodes at each level. During inference, The query image is processed by the ImageEncoder to produce an embedding. Then, the query embedding is compared with the top-level cluster centers to identify the most similar region. Within the identified region, the process is repeated with subclusters to narrow down the candidate GPS coordinates further. Finally, the predicted GPS coordinate is selected from the most similar subcluster.

C. RAG Inference Method

1) Retrieval-Augmented Generation (RAG)-Based Inference: To address the limitations of purely retrieval-based methods, we incorporate RAG for inference. The retrieval phase identifies top candidate locations from the hierarchical gallery, aligning input images with similar entries in the shared embedding space. The generation phase augments these candidates by leveraging large language models (LLMs), including GPT-4o, Mistral, and LLaMA. Using multiple prompts tailored to different geographic contexts, the LLMs generate location predictions that complement the retrieved candidates.

Our inference pipeline includes Geo-verification, where retrieved and generated candidates are ranked and refined based on multimodal similarity metrics. This process ensures robust predictions even for ambiguous or unseen input images, effectively addressing generalization challenges.

2) Incorporating Neighborhood-Level Context: To enhance street-level prediction accuracy, we incorporate neighborhood-level parameters into the textual descriptions used during inference. By encoding fine-grained geographical context, such as landmarks and local features, into the text embeddings, we enable the model to better differentiate among visually similar but geographically distinct locations. This additional layer of context significantly improves the precision of geolocation predictions at finer granularities.

3) Text Embeddings and Experimental Comparisons: We assess the role of text embeddings in geolocation performance by experimenting with models that include and exclude textual information in the inference pipeline. This comparative analysis highlights the contribution of textual context in refining predictions, particularly in ambiguous cases where visual features alone are insufficient for accurate localization.

IV. Experiments

A. Dataset Preparation

To explore the potential to improve the prediction of the location at the street level with additional neighborhood information, along with city, country, continent. We worked with the MP16-Pro dataset, an extension of the MediaEval Placing Task 2016 dataset. The MP16-Pro dataset comprises 4.72 million geotagged images from Flickr, enriched with multi-level geographical descriptions such as neighborhood, city, county, state, region, country, and continent.

***Precompute CLIP Embeddings**:* To improve speed-up, we precomputed the embeddings for all training images and used them instead of doing inference on the frozen CLIP model. This reduced the epoch time from 12 hours to under 3 minutes.

Given the size of the data, with images stored as a 400 GB tar file, it was not feasible to load the entire tar file directly into memory. To address this issue, we tranformed the dataset into the WebDataset format [], which enabled on-the-fly extraction and streaming of images. This approach allowed us to construct a custom dataset and DataLoader for efficient training.

B. Model Training and Evaluation

***Hyperparameter Tuning and Training**:* We trained our model on the entire dataset by precomputing CLIP embeddings as well as data-limited settings (20% of the data). To optimize performance under these conditions, we performed extensive hyperparameter tuning. Key parameters, such as learning rate, batch size, and dropout rates, were adjusted to ensure effective learning with limited data. Additionally, we experimented with various configurations of the hierarchical retrieval structure and RAG prompts to identify the most effective settings for our task.

Evaluation: The model was evaluated using the IM2GPS3K test dataset. The evaluation metric measured the percentage of predictions falling within specific distance thresholds (1 km, 25 km, 200 km, 750 km, and 2500 km) from the ground truth.

C. Hierarchical Feature Clustering

In our experiments, we explored multiple clustering algorithms to organize the GPS gallery, including K-means, DBSCAN, and HDBSCAN. For K-means, we evaluated different approaches to optimize the clustering process, using the elbow method and metrics like the Silhouette score and Davies-Bouldin index.

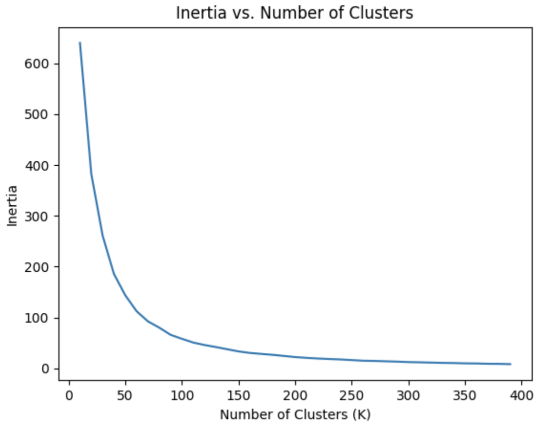


Fig. 4: Using Elbow method, 150-200 clusters was determined to be the ideal range.

We also tested DBSCAN, but it struggled to cluster the GPS coordinates effectively, likely due to the uneven geographical distribution of the data, as shown in Figures **?? ?? ??**.

To structure the GPS gallery, we arranged the clusters into tree-like hierarchies with one and two levels. For two-level trees, we analyzed the effects of different configurations, such as having more clusters at the top level and fewer at the sublevel, and vice versa. These variations allowed us to study the trade-offs between hierarchical depth and clustering granularity.

D. RAG

Embedding Retrieval: For each query image, the RAG method retrieved the top-20 candidate coordinates based on similarity. Top-20 negative candidates were also retrieved to improve robustness of the inference.

Candidate Pool and Selection: We experimented with varying numbers of RAG candidates and found that candidates in the range of 7-10 achieved the best accuracy. We conducted the inference with GPT-4o and mistral models.

Prompts: Below Listing 1 is the prompt provided to the large multimodal models (LMMs) for geolocation prediction.

```
You are a geo-localization expert. Given an image,
you must predict its GPS coordinates.
For reference:
- Similar images have coordinates: {candidates_gps}.
- Dissimilar images have coordinates: {reverse_gps}.

Output your best guess in the following strict JSON
format:
{"location":{"latitude": float, "longitude": float}}

If you cannot make a prediction, use the default
output:
{"location": {"latitude": null, "longitude": null}}

Provide no additional text or error messages.
```

Listing 1: Prompt for LMMs

E. Implementation Details

Below are the final implementation details.

- **Clip Model**: openai/clip-vit-large-patch14
- **Optimizer**: AdamW
- - Learning Rate: 6×10^{-5}
- **Batch Size**: 256
- **Hardware**: 1-4 NVIDIA V100 GPUs
- **Prediction Model**: GPT-4o and Mistral
- **Tools**: Faiss for embeddings storage, clip-retrieval library

V. Results and Analysis

We compare the performance of various models across hierarchical geolocation metrics, including accuracy at street, city, region, country, and continent levels. In addition, we highlight the effectiveness of RAG inference with GPT-4o and mistral vision models. Table I shows the comparison of our model with state-of-the-art models. Our models (RAG with GPT-4o and RAG with Mistral) was evaluated on a subset of 500 images from the IM2GPS3K test set due to resource constraints.

Model training

We conducted ablation studies on GeoCLIP to evaluate different visual encoders, including ViT-B-32, ViT-L-14, and ViT-SO400M-14-SigLIP-384. SigLIP performed best across all geographical scales, likely because of its pretraining on large-scale, diverse datasets. We also experimented with various batch sizes for training GeoCLIP, testing sizes of 128, 256, 512, 1024, and 2048. While increasing the batch size generally led to slightly higher performance across most spatial resolutions, the performance gains at the 2500 km resolution were inconclusive. This may be because larger batch sizes improve the representation of fine-grained details by providing more diverse negative samples in contrastive learning, but at broader spatial scales (like 2500 km), the model may already capture sufficient information, leading to diminishing returns. Additionally, the global-scale features are less dependent on fine-grained diversity within batches.

TABLE I: RAG: Evaluation Result Comparisons on IM2GPS3K Test Set

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
w/o PlaNet	8.5	24.8	34.3	48.4	64.6
w/o Geo-CLIP	14.11	34.47	50.65	69.67	83.82
w/o G3	16.65	40.94	55.56	71.24	84.68
RAG with GPT-4o	**15.01**	**32.53**	**60.06**	**72.5**	**85.08**
RAG with Mistral	**13.32**	**27.23**	**55.46**	**62.5**	**80.6**

TABLE II: Hierarchical features: Cluster Size Comparison

GPS Gallery Tree Sizes	Continent 2500km	Country 750km	Region 200km	Street 1km	% of coordinates considered
100000 (original)	0.753	0.568	0.351	0.084	100%
200 (one-level)	0.618	0.390	0.188	0.047	0.7%
800 (one-level)	**0.710**	**0.481**	**0.253**	**0.049**	**1 %**
1000 (one-level)	0.701	0.480	0.248	0.050	2%
20, 100 (two-level)	0.676	0.430	0.200	0.034	0.2%
100, 20 (two-level)	0.636	0.398	0.177	0.029	0.2%
200, 10 (two-level)	**0.676**	**0.429**	**0.200**	**0.033**	**0.3%**

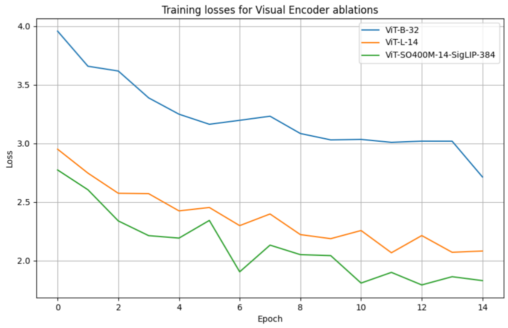


Fig. 5: Training losses for Visual Encoder ablations

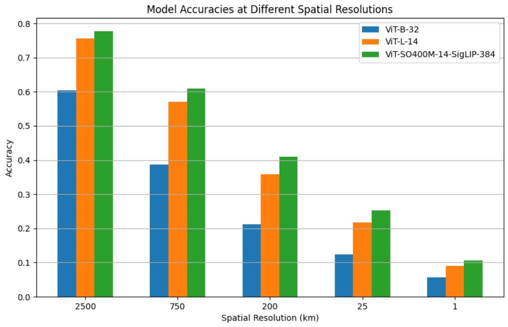


Fig. 6: Model Accuracies at Different Spatial Resolutions

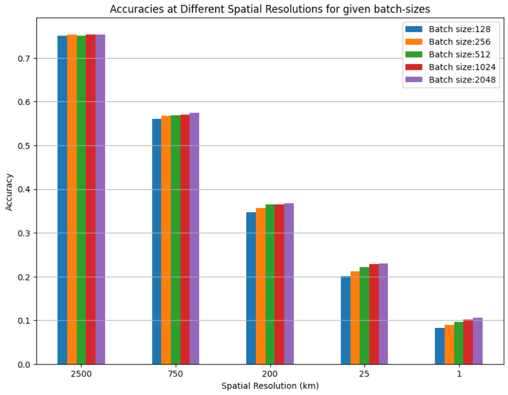


Fig. 7: Accuracies at Different Spatial Resolutions for given batch-sizes

Hierarchical Feature clustering

As shown in Table II, One-level clustering with 800 clusters offers the best trade-off between accuracy and efficiency, closely matching the original implementation while considering only 1% of the GPS points. Additional levels of clustering can significantly enhance scalability when dealing with much larger GPS galleries. For a given clustering level, adding an additional layer of clustering (e.g., moving from one-level with 200 clusters to two-level with 200 top-level and 10 sub-level clusters) can maintain performance at larger geographical resolutions (e.g., 2500 km: 0.618 to 0.676) while significantly reducing the number of candidates considered (from 0.7% to 0.3%). This highlights the potential of multi-level clustering to balance accuracy and efficiency at broader scales. Figures 8, 9 visualize a given hierarchically clustered GPS gallery with the black points representing the cluster centers.

Retrieval-Augmented Generation (RAG)

The Retrieval-Augmented Generation (RAG) inference based on GPT-4o achieved slightly better results, likely due to the candidate references and the model's enhanced vision decoding capabilities. In contrast, inference with Mistral model produced lower accuracy, as it failed to predict locations for certain images with people or image takes from a further distance. We also noticed a slight improvement in street level accuracy when the neighbour information was added to training.

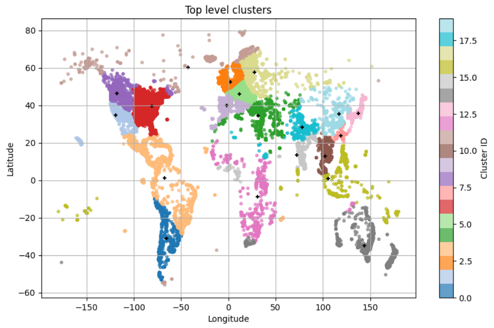


Fig. 8: Level-1 clusters

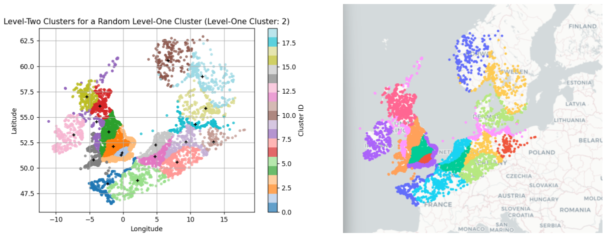


Fig. 9: Level-2 clusters

VI. Conclusion

In this project, we trained a CLIP-based image geolocation model on the MediaEval-16 dataset. We presented enhanced inference approaches by building upon GeoCLIP and G3 frameworks. Our method addressed the limitations of searching through a large gallery of GPS coordinates during inference by employing an efficient hierarchical clustering-based method to perform coarse-to-fine geolocalization. Our inference approach achieved comparable performance to GeoCLIP while being 100x more efficient than previous methods, making it more practical for large-scale geolocation tasks. Finally, we explored retrieval-augmented generation (RAG)-based inference techniques with multiple LLMs, such as GPT-4o, Llama and Mistral, to leverage contextual and semantic reasoning for robust geolocalization. Our approach achieves competitive performance even when trained on a smaller subset of the dataset.

VII. Future Work

While our current approach leverages CLIP-based embeddings and RAG with large multimodal models (LMMs), several improvements can further enhance performance and scalability. One promising direction is the incorporation of beam search algorithms during inference, which systematically explores the most likely candidate locations, offering greater geolocation accuracy while mitigating extra computation. Scaling the system to handle significantly larger GPS galleries (1M+ coordinates) is another critical area. Additionally, leveraging fine-grained textual information, such as neighborhood names or county labels, could refine predictions, particularly in urban or text-rich environments. Another avenue is the use of diffusion models, which could generate high-quality synthetic data for underrepresented regions, helping to address dataset imbalances and improve generalization. These models might also be applied to augment geolocation by refining noisy visual inputs or enhancing low-resolution data.

Authors Contribution

Akshay Raman and Srikanth Balakrishna worked on GeoCLIP model training and hierarchical clustering methods. Prithviraj Murthy contributed to the Retrieval-Augmented Generation (RAG) inference using multiple open-source models. Satyanarayana Chillale conducted experiments on the models and RAG inference. Aman Gupta was responsible for data curation and analysis.

References

[1] V. Vivanco Cepeda, G. K. Nayak, and M. Shah, “Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 8690–8701.

[2] P. Jia *et al.*, “Mp16-pro dataset,” https://huggingface.co/datasets/Jia-py/MP16-Pro, 2024, accessed: 2024-12-19.

[3] J. Hays and A. A. Efros, “Im2gps: Estimating geographic information from a single image,” https://graphics.cs.cmu.edu/projects/im2gps/, 2008, accessed: 2024-12-19.

[4] T. Weyand, I. Kostrikov, and J. Philbin, “Planet - photo geolocation with convolutional neural networks,” *ArXiv*, vol. abs/1602.05314, 2016.

[5] M. Wu and Q. Huang, “Im2city: image geo-localization via multi-modal learning,” in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, ser. GeoAI ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 50–61. [Online]. Available: https://doi.org/10.1145/3557918.3565868

[6] A. Names, “Ethan: Enhanced geolocation through large vision-language models,” *CVPR*, 2024.

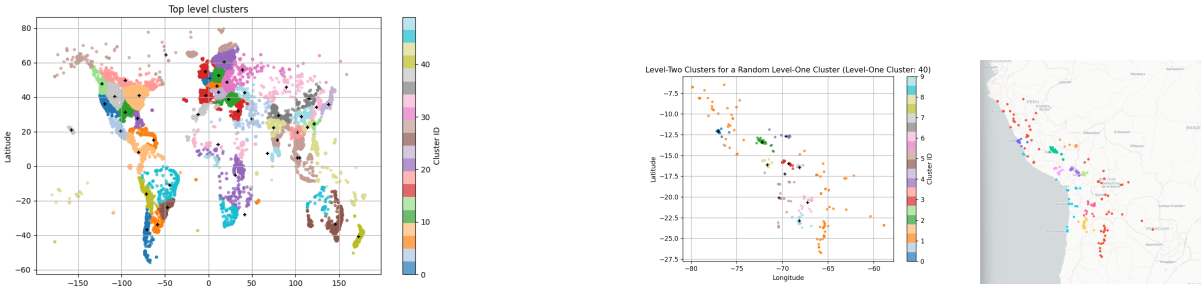
[7] L. Haas, M. Skreta, S. Alberti, and C. Finn, “Pigeon: Predicting image geolocations,” 2024. [Online]. Available: https://arxiv.org/abs/2307.05845

[8] P. Jia *et al.*, “G3: An effective and adaptive framework for worldwide geolocalization using large multi-modality models,” *NeurIPS*, 2024.

[9] Y. Research, “Yfcc4k dataset,” https://multimedia-commons.s3.amazonaws.com/index.html, 2016, accessed: 2024-12-19.

Appendix

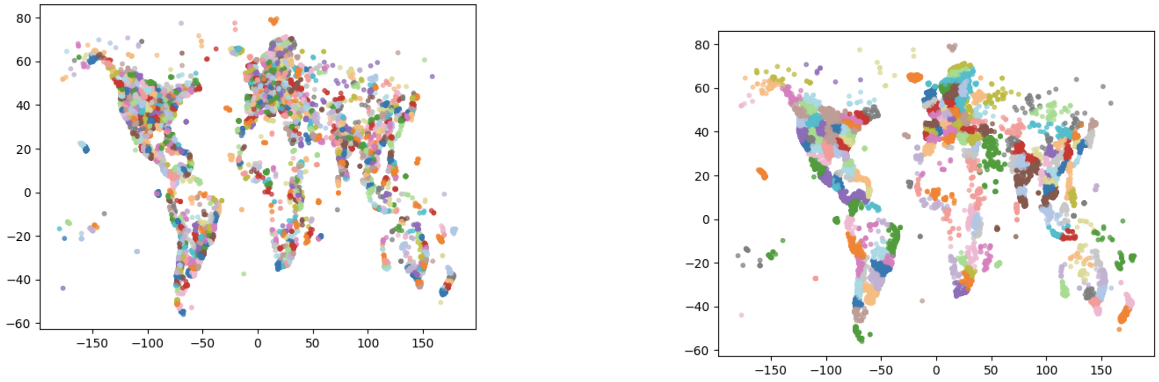
The appendix below includes additional ablation studies involving different clustering algorithms and cluster sizes. We also include some sample model predictions on different settings.



(a) Top-level clusters: K-means 50 clusters

(b) A random sublevel cluster from K-means clustering: 10 sublevel clusters

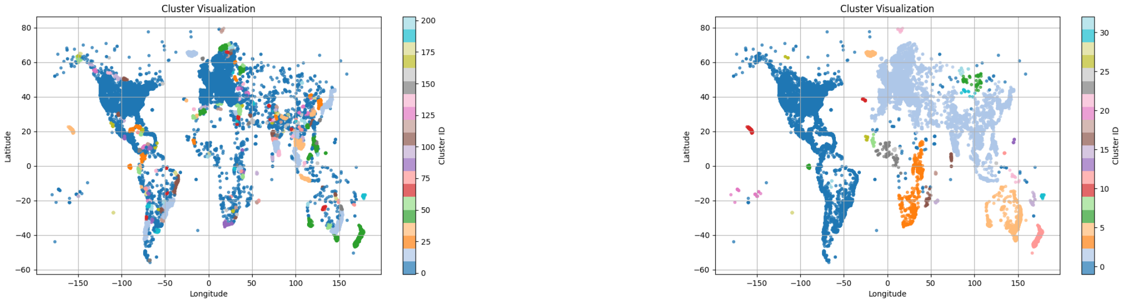
Fig. 10: Hierarchical K-Means clustering samples



(a) Clustering with aggregated GPS embeddings from all 3 location encoder capsules

(b) Clustering with only top-level GPS embeddings

Fig. 11: Significance of hierarchical GPS embeddings. K-Means algorithm successfully clustered the locations using top-level GPS embeddings however failed to cluster using the aggregated embeddings.



(a) Top-level clusters: DBSCAN eps=0.01

(b) Top-level clusters: DBSCAN eps=0.05

Fig. 12: Hierarchical K-Means clustering samples

(a) Sydney Tower

(b) Mountain Region

Actual:	−33.861328, 151.209039
Final:	−33.8708476, 151.2073203
Zero Shot:	−33.870987, 151.208843
5 RAG:	−33.8708476, 151.2073203
10 RAG:	−33.870846, 151.206732
15 RAG:	−33.869844, 151.209296

Geodesic Distance: 1.0678 km

Actual:	46.167286, 7.099698
Final:	45.96096, 6.94477
Zero Shot:	44.0958, 6.8467
5 RAG:	44.4343, 6.6292
10 RAG:	45.96096, 6.94477
15 RAG:	45.920135, 6.869433

Geodesic Distance: 25.8776 km

Fig. 13: Two samples from the IM2GPS3K dataset. For each image, we provide the actual location, the final predicted location, and the RAG-based inference predictions using different numbers of candidate references supplied to the large multimodal models (LMMs).