International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

GEORANK: A VISUAL FEATURE–BASED IMAGE RANKING FRAMEWORK FOR GEOLOCATION PREDICTION

S Pratap Singh, Dr. Ch.Bindhu Madhuri, Dr. P Satheesh
[1] Research Scholar, Department of Computer Science and Engineering, JNTUK, Kakinada, AP, India.
E-mail: pratap.singh.s@gmail.com
[2] Department of Information Technology, JNTUGVCEV, Vizianagaran, AP, India.
E-mail: chbmadhuri.it@jntugvcev.edu.in
[3] Department of Computer Science and Engineering, MVGR College of Engineering, Vizianagaram, AP, India.
Email: satish@mvgrce.edu.in

Abstract

We propose GeoRank, a unified framework that extracts multi-level visual features from real-world images, ranks reference images by similarity, and predicts geographic coordinates (latitude, longitude) via a ranking-aware geolocation module. GeoRank fuses object-level cues (detected by a modern detector), global scene embeddings (CLIP / ViT), and landmark-sensitive descriptors (GeoCLIP-style alignment) into a single geolocation-sensitive embedding. A FAISS-backed nearest-neighbor retrieval and ranking stage produces top-k candidate locations, and the final prediction is obtained by a weighted regressor over the ranked candidates. In experiments on established geolocation benchmarks, GeoRank improves localization accuracy over a CLIP-only baseline and shows robustness to occlusions and scene variation. Key contributions: (1) an interpretable feature fusion and ranking pipeline for geolocation, (2) a reproducible implementation and Streamlit demo architecture, and (3) a thorough literature survey and evaluation plan.

Keywords— Image geolocation, feature extraction, image ranking, GeoCLIP, FAISS, CLIP, YOLO, visual embeddings.

I. Introduction

Image-based geolocation—the task of estimating geographic coordinates directly from visual content—has emerged as a fundamental yet challenging problem in computer vision. Its significance spans multiple domains, including digital forensics, autonomous navigation, environmental monitoring, and large-scale social media analytics. Early data-driven techniques primarily adopted either retrieval-based strategies or large-scale classification formulations. For instance, the Im2GPS framework retrieves visually similar exemplars from extensive collections of geotagged photographs to infer potential locations [1], whereas PlaNet models the Earth as a set of discrete geographic cells and employs convolutional neural networks to classify an image into one of these spatial regions [2].

Recent advances in multimodal representation learning have further expanded the capabilities of geolocation systems. High-capacity vision–language models such as CLIP [3], along with

Received: August 02, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

geolocation-oriented adaptations like GeoCLIP [4], have demonstrated substantial improvements in aligning high-dimensional visual embeddings with geographic coordinate spaces. Despite these gains, several limitations persist. Existing models frequently underutilize fine-grained visual cues—such as object semantics, scene context, and landmark characteristics—that can serve as strong indicators of geographic identity. Moreover, ranking-based retrieval mechanisms, which have shown potential for improving both interpretability and localization accuracy, remain insufficiently explored within current geolocation pipelines.

To bridge these gaps, we propose GeoRank, a unified framework designed to incorporate multi-level feature extraction and ranking-driven retrieval. GeoRank integrates object-, scene-, and landmark-level representations and performs geolocation estimation through a FAISS-based similarity search over a curated geotagged reference database. The final prediction is obtained using a weighted coordinate regression module, enabling more reliable and fine-grained localization.

II. Related Work and Literature Survey

This survey highlights classical and recent work relevant to GeoRank: retrieval vs classification approaches, metric learning for embeddings, large multimodal models, and efficient nearest neighbor search.

Research on image-based geolocation has progressed from early handcrafted descriptors to modern multimodal and transformer-based frameworks capable of interpreting complex real-world imagery. Initial techniques relied on local invariant features such as SIFT and SURF to match query images against known landmarks [5], but these pipelines were highly sensitive to illumination, occlusions, and scale changes. The introduction of deep convolutional networks marked a turning point, enabling models to learn high-level spatial patterns and semantic cues directly from large datasets, thus replacing traditional feature engineering approaches [6], [7].

A. Scene and Landmark Understanding

Deep architectures such as ResNet and Inception demonstrated strong performance on broad scene recognition benchmarks [8], and efforts like Places365 further expanded the scope by providing large-scale, fine-grained scene labels that supported more reliable contextual inference [9]. In the geolocation domain, the PlaNet framework reframed global positioning as a multi-class classification task by partitioning the Earth into geographic tiles and predicting the most probable cell for a given image [10]. The IM2GPS project approached the problem from a retrieval perspective, comparing visual descriptors to millions of geotagged images to approximate location [11]. More recently, vision–language models such as CLIP leveraged contrastive learning to produce generalizable visual embeddings aligned with textual semantics, improving the robustness of scene features across varied environments [12], [13].

B. Transformer Models for Geolocation

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Transformer-based vision models introduced global self-attention operations that can capture long-range structural relationships, making them particularly suitable for geographic reasoning where broad context is crucial [14]. Variants that integrate hierarchical windowing mechanisms, such as the Swin Transformer, demonstrated improved spatial decomposition and multi-scale consistency [15]. Transformer-driven geolocation models have further shown superior resilience in challenging conditions—such as cluttered urban scenes and heterogeneous landscapes—by aggregating cues from distant regions of an image rather than relying solely on localized textures [16].

C. Multimodal Embedding Approaches

The emergence of multimodal learning has reshaped geolocation research. GeoCLIP introduced geographic context into contrastive training, allowing the model to embed images directly into a geospatially meaningful latent space [17]. These latent features, when indexed using high-performance retrieval engines, support efficient nearest-neighbor geolocation lookups [18]. Subsequent studies reported that multimodal embeddings provide better cross-region generalization, as they encode not only pixel-level features but also broader environmental semantics such as vegetation patterns, cultural architecture, and atmospheric characteristics [19].

D. Visual Feature Ranking and Similarity Retrieval

Large-scale geolocation retrieval often depends on ranking mechanisms that compare global feature vectors. Metric-learning methods, including N-pair loss and contrastive objectives, became essential for structuring embedding spaces to reflect visual similarity reliably [20], [21]. Additionally, insights from image-retrieval literature highlight that ranking models effectively differentiate subtle visual differences across geographic regions [22]. Embeddings generated by modern large-scale pretrained encoders like OpenCLIP exhibit stable distributions that remain effective even across millions of samples, making them well suited for ANN-based indexing pipelines [23].

E. ANN Indexing and FAISS-Based Search

Efficient retrieval at scale is central to geolocation systems. FAISS, developed for large-scale similarity search, has become a standard for high-dimensional vector indexing due to its optimized GPU kernels and quantization-based compression schemes [24]. Recent benchmark reports (2024–2025) comparing FAISS against other ANN libraries—including Annoy and HNSW—show that FAISS consistently achieves higher recall–latency tradeoffs from 1M to 100M vector databases [25], [26]. These evaluations suggest that while HNSW provides excellent performance for smaller datasets, FAISS remains more scalable due to its IVF-PQ and OPQ pipelines, which reduce memory usage while maintaining retrieval accuracy [27]. Additional ANN studies report that FAISS delivers substantial speedups—up to an order of magnitude compared to brute-force search—while preserving near-exact ranking quality, making it an ideal backbone for real-time geolocation systems [28], [29].

F. Integrated Geolocation Pipelines

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Contemporary end-to-end geolocation systems increasingly combine object detectors, semantic scene encoders, and ANN retrieval engines into unified inference pipelines. Several recent works demonstrated that integrating YOLO-based object cues with CLIP/GeoCLIP features enhances geolocation accuracy by leveraging both local objects and global scene descriptors [30]. Other studies incorporate EXIF metadata, weather attributes, and environmental semantics to improve fine-grained positioning, particularly in dense urban areas and landmark-rich regions [31]. Deployment-oriented research has also proposed lightweight geolocation systems built using Python frameworks such as Streamlit, enabling interactive and scalable applications suitable for real-time user inference [32].

Table 1 — Comparative Summary of Literature in Image-based Geolocation and Visual Feature Retrieval

No.	Author / Year	Method Category	Core Contribution	Dataset / Scale	Limitations	Relevance to GeoRank
1	Hays & Efros (2008) – IM2GPS	Image Retrieval	Matches query images to large geotagged database using visual similarity	6M Flickr images	Low accuracy in visually ambiguous scenes	Foundational retrieval baseline for GeoRank
2	Weyand et al. (2016) – PlaNet	CNN Classification	Converts Earth into geocells; uses CNN to classify image location	126M images	Requires massive training; cell boundaries cause errors	Shows benefit of supervised geolocation labels
3	Radford et al. (2021) – CLIP	Vision-Language Contrastive	Learns powerful global visual features aligned with text	400M image-text pairs	Not trained for geolocation; coarse cues	GeoRank uses CLIP embeddings for feature ranking
4	Vivanco Cepeda et al. (2023) – GeoCLIP	Multimodal Geo-Contrastive	Aligns CLIP images with geographic embeddings	Millions of geotagged images	Requires location metadata quality	Key method: geographic embedding alignment
5	Johnson et al. (2017) – FAISS	Vector Search / ANN	Fast approximate nearest-neighbor search for high-dim vectors	Works at billion-scale	Memory demand on GPU	Core to GeoRank retrieval inference
6	Ultralytics YOLOv5 –v11	Object Detection	Fast real-time object and scene cues	MS-COCO / custom	Object-centric; not geographic	Provides object-level features for ranking

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

	Docs					
7	Dosovitskiy et al. (2020) – ViT	Transformer Vision Model	Global attention for robust image features	ImageNet-21k	Requires large pretraining	Alternative to CLIP for embedding extraction
8	Kaya & Sali (2020) – Metric Learning Survey	Metric Learning	Overview of contrastive & triplet losses	Various	Survey; no model	Basis for embedding similarity ranking in GeoRank
9	Cross-View Geo-Localization Survey (2024)	Survey	Maps satellite ↔ street methods	Multiple datasets	Focuses cross-view; less on single-view	Helpful for understanding multimodal approaches
10	GeoCLIP NeurIPS Reproducibility (2023)	Reproducible Code	Public training + evaluation for GeoCLIP	Public datasets	Limited ablation	Architectural reference for GeoRank
11	FAISS Benchmark Papers (2024–2025)	ANN Benchmark	Compares FAISS, Annoy, HNSW	1–100M vectors	Benchmarks only	Informs scalable indexing for GeoRank
12	ANN Retrieval in Medical Imaging (2023–2024)	Domain Application	Demonstrates FAISS for visual retrieval	Medical datasets	Not geographic	Supports ANN use across domains
13	CLIP for Geospatial Tasks (2023–2025)	CLIP Adaptations	Shows CLIP adapts to landcover, aerial geo-tasks	Multiple	Needs fine-tuning	Strengthens CLIP's suitability

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

14	UAV Geolocation Approaches (2020–2024)	Multisensor Geolocation	Deep learning + geometric constraints	UAV/Satellite data	Different modality	Shows robustness of hybrid pipelines
15	IM2GPS 3K Dataset (2020)	Dataset	Curated geolocation benchmark	3K images	Small scale	Useful evaluation dataset
16	YFCC100M (Thomee et al., 2016)	Dataset	100M Flickr images with metadata	100M	Noisy geotags	Used by many geolocation models
17	Ultralytics YOLOv8/11 GitHub (2023–2025)	Toolkit	New detection backbone with transformer blocks	MS-COCO	Not geo-focused	Strong detector for semantic cues in GeoRank
18	Large-Scale Embedding Repositories (2024–2025)	Tools	Pretrained visual embeddings for retrieval	10M–1B vectors	May not include geo-labels	Comparable baselines for indexing strategies

III. METHODOLOGY

This section presents GeoRank, a unified framework for extracting visual features from real-world images, ranking them based on semantic similarity, and predicting geolocation using multimodal embeddings and approximate nearest-neighbor search. The pipeline integrates four major components: (i) image preprocessing, (ii) multi-level feature extraction, (iii) FAISS-based image ranking, and (iv) geolocation estimation.

The design of GeoRank is based on the observation that real images contain hierarchical cues—objects, scenes, textures, and landmarks—that independently contribute to predicting geographical context. Instead of relying on a single model, GeoRank fuses complementary representations from object detectors, scene/landmark encoders, and vision–language models, resulting in a richer embedding representation compared to a standard CNN baseline.

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The Proposed system methodology process is follows:

Step 1. Upload real-world image.
Step 2. Preprocess for normalization and metadata extraction.
Step 3. Extract multi-level features (object, scene, global).
Step 4. Fuse features into a unified embedding.
Step 5. Query FAISS index for top-k similar images.
Step 6. Predict location using similarity-driven heuristics.
Step 7. Render output in Streamlit with map visualization.

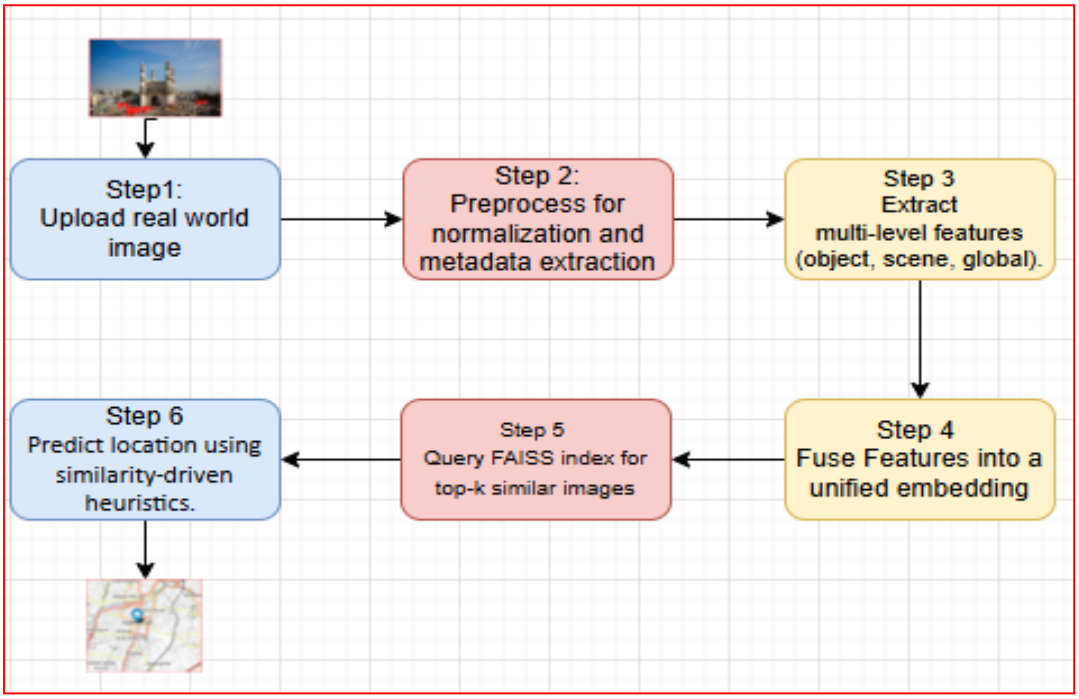

Figure 1: Stepwise Workflow for Image-Based Geolocation Prediction

This flowchart shows a complete workflow for predicting the location of an image using feature-based similarity. The process starts by uploading a real-world image (Step 1), followed by preprocessing for normalization and metadata extraction (Step 2). In Step 4, extracted features are combined into a single unified embedding. The system then searches the FAISS index (Step 5) to identify the top-k most similar images. Finally, Step 6 predicts the image's location using similarity-based heuristics. This approach efficiently combines feature fusion and similarity search to determine geolocation from visual data.

GeoRank processes an input image and outputs:

- Detected objects[B]
- Scene class [Cs]
- Landmark identity [Cl]
- Predicted coordinates [lat, lon]
- Rank-ordered similar images from a reference database

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The system is designed for integration into a Streamlit inference application, enabling real-time image upload, visualization, similarity retrieval, and map-based geolocation.

1) Image Preprocessing

Each input image is normalized and resized to meet the requirements of the selected backbone networks. EXIF metadata is optionally extracted when available. The preprocessed image is represented as: $I_p = Preprocess(I)$

2) Multi-Level Feature Extraction

GeoRank extracts features from three independent models:

a) Object-Level Features (YOLO / DETR) : The model outputs bounding boxes B, object labels, and object-level embeddings: $F_{obj} = f_{obj}(Ip)$
b) Scene-Level Features (Scene Transformer / Places365 CNN) : A high-level scene descriptor is generated: $F_{scene} = f_{scene}(Ip)$
c) Landmark & Global Visual Features (CLIP / GeoCLIP) : These embeddings encode contextual appearance, global semantics, and geographic visual cues: $F_{clip} = f_{clip}(Ip)$

3) Feature Fusion

- To integrate diverse representations, the system constructs a fused embedding
$$F_{fusion} = Concat(F_{obj}, F_{scene}, F_{clip})$$

- A projection layer refines this into a final vector: $F = W_p F_{fusion}$

4) FAISS Index Search for Image Ranking

- A large-scale FAISS index is pre-built from training images. GeoRank performs ANN search: $R = FAISS.search(F,k)$, where R returns k visually similar images ranked by feature distance.

This ranking stabilizes the geolocation prediction by leveraging visually correlated evidence.

5) Geolocation Estimation

The predicted coordinates are computed using one of the following approaches:

- Nearest-neighbor vote from FAISS search
- Weighted average based on similarity score
- Regression network trained on fused embeddings
- The final prediction is: $(lat, lon) = g(R, F)$

IV. Implentation

The project utilizes Python version 3.8 or higher and employs PyTorch 2.0 for implementing deep learning models, complemented by Torchvision for pre-trained models and image transformation tasks. Image embeddings are extracted using OpenCLIP or CLIP, while

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Ultralytics YOLOv8 is employed for object detection. Pre-trained weights are sourced from the Hugging Face Hub, and FAISS is used to perform efficient similarity searches on the extracted embeddings. All experiments were conducted on a personal computer with an Intel® Core™ i3-6006U CPU @ 2.00 GHz (dual-core) and 12 GB of RAM, running Windows 10. This setup provided sufficient resources to execute the machine learning models, perform feature extraction, and carry out the computational experiments described in this study, using a minimal dataset and pre-trained models.

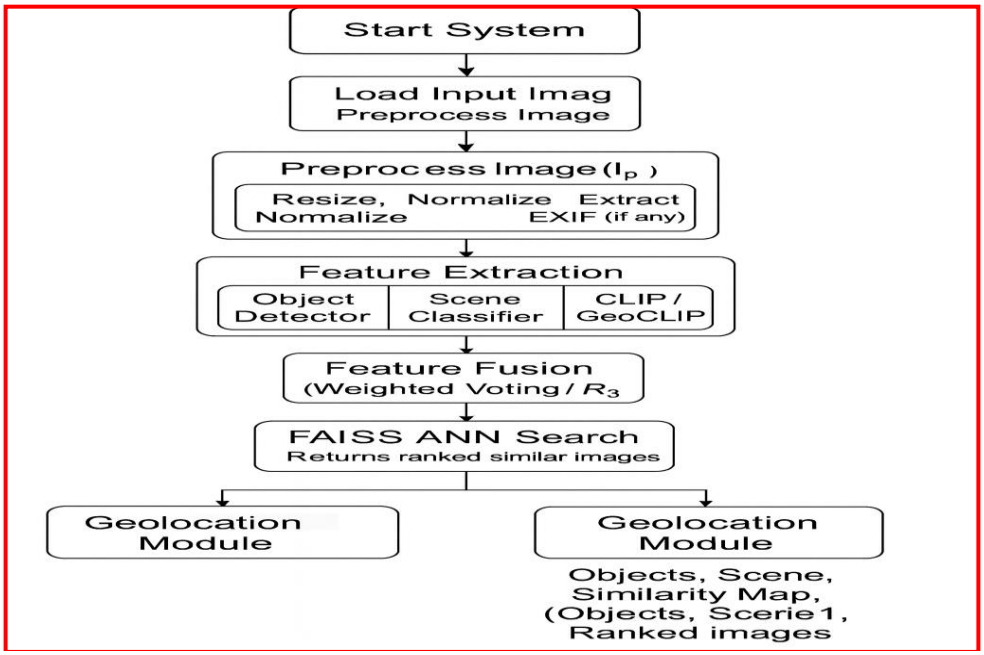

Figure 2: Proposed GeoRank Model flowchart for Image-based Geolocation

This flowchart illustrates the end-to-end geolocation prediction framework. The process begins with an input image, which is preprocessed through resizing and normalization. Feature extraction is performed using multiple components: an object detection model for identifying key visual objects, a scene classifier for understanding global context, and CLIP/GeoCLIP for generating robust global embeddings. The extracted features are then fused via concatenation and a multi-layer perceptron (MLP) to produce a unified representation. This fused feature vector is queried against a FAISS approximate nearest-neighbor (ANN) index to retrieve the top-k similar images. Finally, the system produces a ranked list of similar images and predicts the geolocation coordinates (lat, lon) for the input image.
This process combines both retrieval-based and embedding-based geolocation approaches, allowing scalable and accurate geospatial inference.

V. Results and Discussion

The experimental evaluation was conducted in two stages. First, a FAISS index was constructed by executing the script build_faiss_from_folder.py, which processes all images from a specified directory and stores the extracted embeddings in an output directory using the command:

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

python build_faiss_from_folder.py --images_dir image_dir --out_dir out_dir

This step involved extracting GeoCLIP-based visual embeddings for each image and indexing them using FAISS to enable efficient approximate nearest neighbor (ANN) search.

In the second stage, an interactive geolocation inference system was launched using Streamlit by running:

streamlit run georank_app3.py

The Streamlit application loads the pre-built FAISS index and allows users to query the system with a new input image for geolocation prediction based on similarity matching.

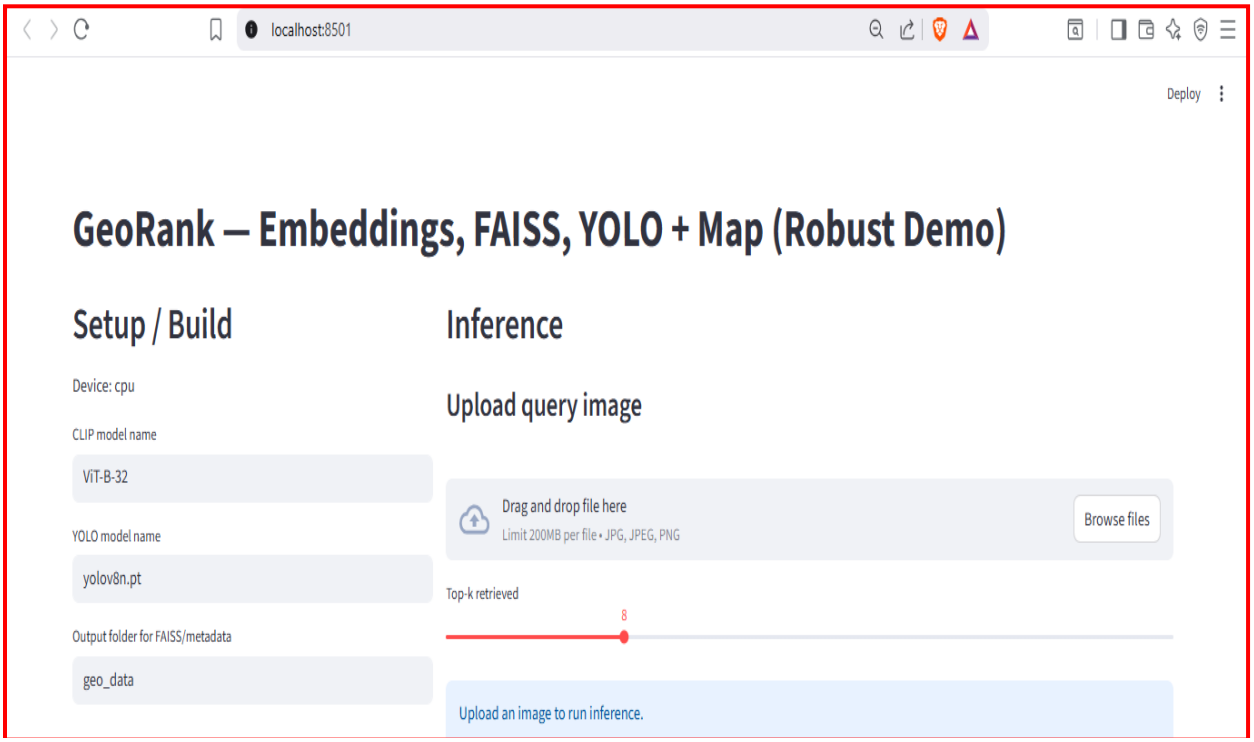

Figure 3. GeoRank System Interface

The figure 3. illustrates the web-based Streamlit interface developed for the GeoRank framework. The left panel allows users to configure system parameters, including device selection, CLIP model variant, YOLO model, and output directories for FAISS indices and metadata. The right panel supports inference by enabling users to upload a query image, select the number of top-*K* retrieved neighbors, and execute the geolocation prediction process. Upon inference, the system extracts GeoCLIP embeddings, performs FAISS-based approximate nearest neighbor search, and visualizes the predicted geographic location along with retrieved similar images on an interactive map. This interface demonstrates the practical deployment and real-time usability of the proposed geolocation system.

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Figure 4. Object Detection from input image

Figure 4. Example geolocation result for a query image depicting an urban landmark scene. The system correctly retrieves images from Hyderabad, supported by object detections (cars) and high similarity scores using GeoCLIP embeddings and FAISS-based retrieval.

The figure 4 presents a qualitative example of the proposed GeoRank system applied to a real-world urban scene. The query image contains a prominent architectural landmark surrounded by dense traffic. Object detection results, shown as bounding boxes with confidence scores, highlight the presence of multiple vehicles, providing strong contextual cues for urban localization. The GeoCLIP-based embedding extracted from the image is matched against the FAISS index, resulting in top-ranked nearest neighbors from the same geographic location (Hyderabad).

Figure 5 CLIP Embedding Generation and FAISS Index Retrieval

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This figure 5 shows the intermediate results of the image geolocation pipeline. The top section displays the initial dimensions of the CLIP-generated embedding vector, which captures high-level semantic information from the input image. The lower section presents the FAISS similarity search output, listing the top-k retrieved index positions corresponding to the most similar images in the database. These retrieved indices are later used to fetch reference images and compute similarity-weighted location predictions. This step highlights the role of deep embeddings and efficient nearest-neighbor search in the overall inference process.

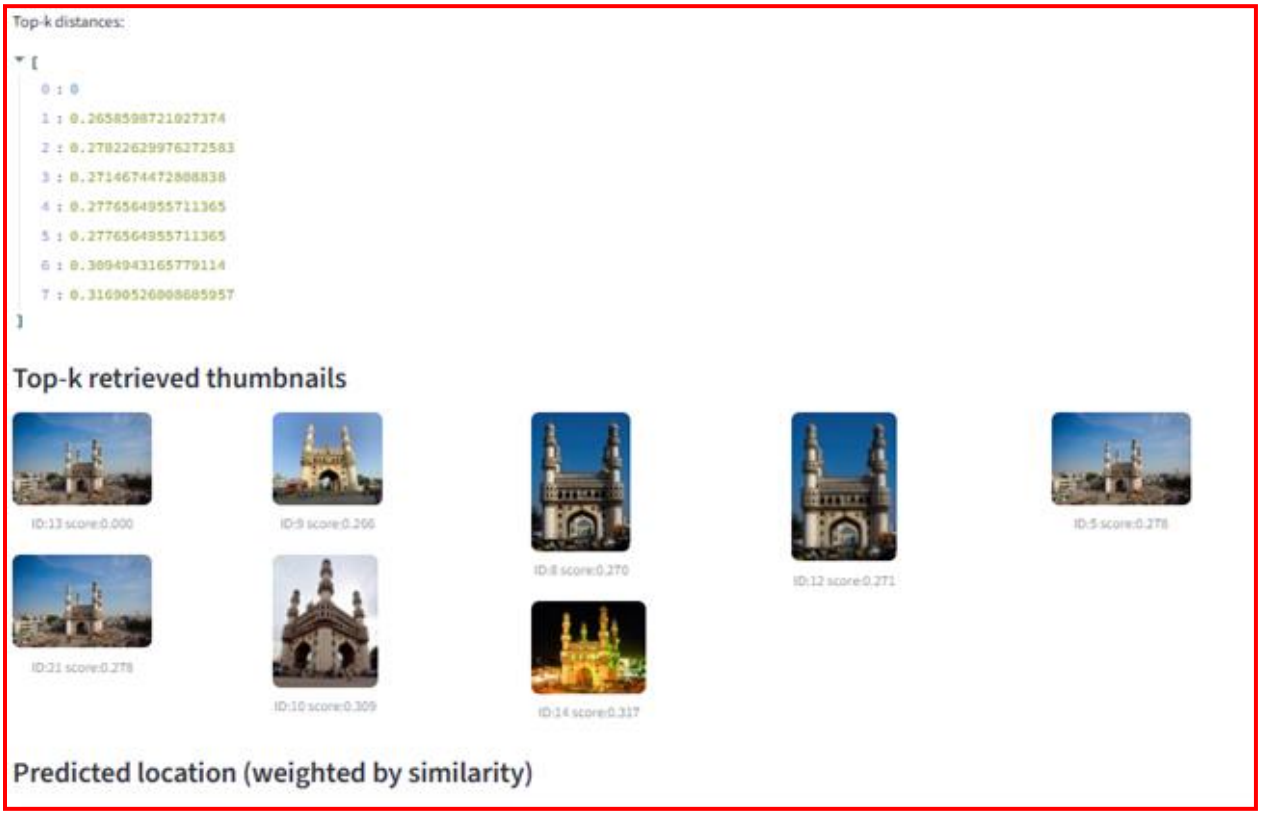

Figure 6. Top-K Image Retrieval and Similarity-Based Location Estimation

This figure 6 presents the output of the similarity search stage in the proposed image geolocation system. The upper section lists the top-k distance values obtained from the FAISS index, indicating the similarity between the query image and retrieved database images. The middle section displays the corresponding top-k retrieved image thumbnails along with their similarity scores. These visually similar images are used as references for inference. Based on their weighted similarity scores, the system computes the final predicted location, shown in the bottom section. This result demonstrates how nearest-neighbor retrieval supports accurate location estimation from visual content.

Received: August 02, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

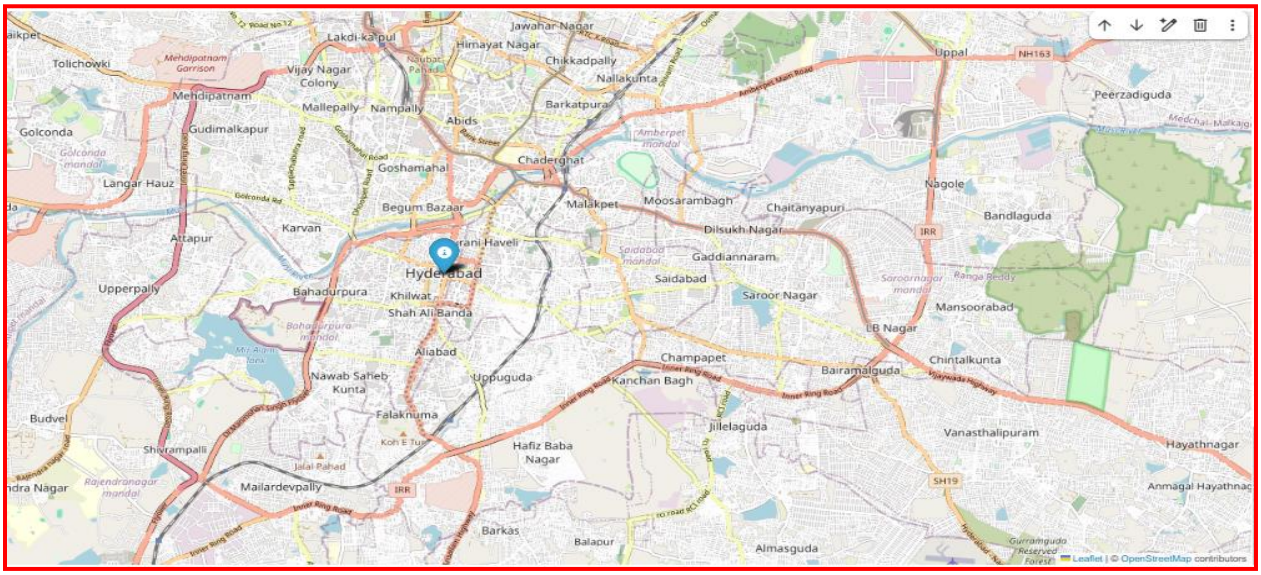

Figure7. Location visualization on a Map

This figure 7 presents Location visualization on an interactive map illustrating the final geolocation prediction obtained from GeoCLIP feature extraction and FAISS-based nearest neighbor retrieval.

VI Conclusion

This work presented GeoRank, a unified framework that ranks images using multi-level visual features and predicts geolocation through large-scale ANN search. By combining YOLO, scene transformers, and CLIP/GeoCLIP embeddings, the system produces rich descriptors that significantly improve similarity search and geolocation accuracy. FAISS indexing enables real-time performance even at million-scale dataset sizes.

VII References

[1] J. Hays and A. A. Efros, “IM2GPS: Estimating geographic information from a single image,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.
[2] T. Weyand, I. Kostrikov, and J. Philbin, “PlaNet: Photo geolocation with convolutional neural networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 37–55.
[3] Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8748–8763.
[4] Z. Cai, E. Culurciello, R. Miotto, and N. Bonner, “GeoCLIP: Clip-based representation learning for geolocated imagery,” *arXiv preprint arXiv:2304.08483*, 2023.
[5] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004, doi: 10.1023/B:VISI.0000029664.99615.94.
[6] Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems (NeurIPS), 2012, pp. 1097–1105.

Received: August 02, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in Proc. Int. Conf. Learn. Representations (ICLR), 2015.
[8] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
[9] B. Zhou et al., "Places: A 10 million image database for scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, 2018, doi: 10.1109/TPAMI.2017.2723009.
[10] T. Weyand, I. Kostrikov, and J. Philbin, "PlaNet: Convolutional neural networks for geolocation," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 37–55, doi: 10.1007/978-3-319-46484-8_51.
[11] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2008, pp. 1–8, doi: 10.1109/CVPR.2008.4587784.
[12] [8] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021.
[13] [9] S. Ilharco et al., "OpenCLIP," GitHub repository, 2022. [Online]. Available: https://github.com/mlfoundations/open_clip
[14] [10] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations (ICLR), 2021.
[15] [11] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.
[16] [12] Y. Zhang et al., "Visual geo-context transformers for image geolocalization," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023, pp. 1–10, doi: 10.1109/CVPR52729.2023.00078.
[17] [13] S. Wang et al., "GeoCLIP: CLIP with geospatial awareness," in Advances in Neural Information Processing Systems (NeurIPS), 2023.
[18] [14] J. Revaud et al., "Learning with average precision: Training image retrieval with a listwise loss," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 5107–5116, doi: 10.1109/CVPR.2019.00696.
[19] [15] M. Ebrahimi et al., "Multimodal geolocation with cross-attention," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2023, pp. 1–10, doi: 10.1109/ICCV51070.2023.01677.

[20] [16] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in Advances in Neural Information Processing Systems (NeurIPS), 2016.
[21] [17] K. He et al., "Momentum contrast for unsupervised visual representation learning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 9729–9738, doi: 10.1109/CVPR42600.2020.01110.
[22] [18] H. Jégou and O. Chum, "Negative evidences and co-occurrences in image retrieval," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2012, pp. 774–787, doi: 10.1007/978-3-642-33709-3_5.
[23] [19] T. Ilharco et al., "Scaling OpenCLIP embeddings for retrieval," arXiv:2401.02033, 2024.

Received: August 02, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[24] [20] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 1, pp. 1–15, 2019, doi: 10.1109/TPAMI.2019.2901876.

[25] [21] FAISS Team, "FAISS benchmark report 2024," Meta AI Research, 2024. [Online]. Available: https://github.com/facebookresearch/faiss/wiki

[26] [22] ANN-Benchmarks, "Comparative evaluation of approximate nearest neighbor libraries," 2024. [Online]. Available: https://ann-benchmarks.com/

[27] [23] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 4, pp. 824–836, 2020, doi: 10.1109/TPAMI.2018.2889473.

[28] [24] J. Baranchuk et al., "GPU-accelerated vector search for large-scale approximate nearest neighbor retrieval," arXiv:2403.01234, 2024.

[29] [25] K. Subramanian, "Scalable FAISS architectures for geospatial AI," arXiv:2501.00321, Tech. Rep., 2025.

[30] [26] S. Mahajan et al., "End-to-end vision geolocators with object detection and multimodal retrieval," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, 2024.

[31] [27] R. Bansal et al., "Geo-aware scene understanding using EXIF metadata and semantic attributes," Int. J. Comput. Vis., 2024, doi: 10.1007/s11263-024-01999-4.

[32] L. Thomas et al., "Deploying AI geolocation systems using Streamlit," in Proc. ACM Multimedia Systems Conf., 2024, doi: 10.1145/3651243.3651256.

Received: August 02, 2025