10.48047/jocaaa.2024.33.08.303

A Unified Vision–Language Approach for Predicting Geographic Locations from Natural Images through Object, Scene, and Landmark Understanding

Dr. S Pratap Singh, Dr. Ch.Bindhu Madhuri, Dr. P Satheesh

[1] CSE Department, Marri Laxman Reddy Institute of Technology and Management, Hyderabad, TS, India.
E-mail: pratap.singh.s@gmail.com

[2], IT Department, JNTUGVCEV, Vizianagaran, AP, India.
E-mail: chbmadhuri.it@jntugvcev.edu.in

[3] CSE Department, MVGR College of Engineering, Vizianagaram, AP, India.
Email: satish@mvgrce.edu.in

Abstract:
The task of predicting geographic locations from natural images remains a significant challenge due to complex visual variability and the lack of explicit spatial cues. This paper presents a unified vision–language framework that integrates object detection, scene understanding, and landmark recognition to estimate geolocation with improved accuracy and interpretability. The proposed model combines YOLOv11-based visual feature extraction with CLIP-derived semantic embeddings, enabling a cross-modal representation that aligns visual and textual cues in a shared latent space. By leveraging both the structural and contextual information present in images, the system effectively associates detected objects and environmental attributes with geographic semantics. Experimental evaluations conducted on diverse regional datasets demonstrate that the proposed framework outperforms conventional convolutional and transformer-based baselines in terms of localization precision and robustness under varying environmental conditions. The results validate the effectiveness of multimodal learning for spatial reasoning tasks and highlight its potential applications in autonomous navigation, urban mapping, and cultural landmark identification. This study contributes a scalable and interpretable approach to geo-AI, bridging the gap between computer vision and geographic intelligence.

Keywords
Vision–language models, geolocation prediction, object detection, multimodal learning, scene understanding, CLIP, YOLOv11, geographic intelligence.

Introduction:

The prediction of geographic locations from natural images has become a key problem in computer vision and spatial intelligence. With the rise of image-based data from social media, mapping services, and autonomous systems, models capable of estimating image location without explicit GPS metadata are gaining importance [1], [2]. Traditional computer vision techniques, relying mainly on low-level visual features such as color or texture, often fail to capture the semantic and contextual details necessary for precise localization [3].

10.48047/jocaaa.2024.33.08.303

Advances in deep learning and multimodal vision–language modeling have introduced new methods for integrating semantic understanding into image-based geolocation. Models like CLIP (Contrastive Language–Image Pretraining) align visual and textual embeddings in a shared latent space, enabling improved interpretation of visual content [4]. Similarly, YOLOv11 and other detection architectures can identify key visual cues—such as landmarks, vegetation, or architectural patterns—that are strongly linked to geographic regions [5]. However, most prior studies treat these modalities independently, limiting their performance across diverse landscapes and environments.

To overcome this limitation, this work proposes a unified vision–language framework that combines object detection, scene understanding, and landmark recognition for enhanced geolocation prediction. The approach integrates YOLOv11-based visual extraction with CLIP-derived semantic embeddings to jointly model visual and contextual cues. Experimental evaluations demonstrate significant improvements in localization accuracy, interpretability, and robustness, confirming the potential of multimodal Geo-AI for applications in navigation, mapping, and environmental analysis.

Related Work:

Research in image-based geolocation has evolved from traditional data-matching approaches to advanced deep learning frameworks capable of interpreting high-level visual semantics. The early milestone, IM2GPS, introduced a retrieval-based approach that predicted an image's geographic position by finding visually similar geotagged samples within a large reference dataset. Although effective for coarse localization, its performance was limited by visual ambiguity and database dependence. Building on this, PlaNet transformed the problem into a large-scale image classification task, partitioning the Earth into geographic cells and employing convolutional neural networks to assign probability distributions over them. This method demonstrated significant improvements in prediction accuracy by leveraging millions of geotagged photographs and incorporating temporal coherence in image sequences [6].

Subsequent research shifted toward multimodal learning, where vision–language models began bridging visual content with textual semantics. The CLIP model (Contrastive Language–Image Pretraining) established a powerful framework for aligning images and textual descriptions within a shared embedding space, enabling zero-shot generalization and contextual understanding of scenes. Extending this idea, GeoCLIP introduced a continuous geographic embedding mechanism that aligns image features directly with GPS coordinates, effectively capturing spatial semantics and outperforming earlier visual-only geolocation systems [7].

A recent comprehensive survey on cross-view and multimodal geolocation emphasizes an emerging research trend—integrating retrieval, recognition, and multimodal reasoning to handle viewpoint variation, scale differences, and incomplete visual cues [8]. Together, these studies underline a growing shift from low-level appearance-based matching toward unified vision–language paradigms, which can better interpret geographic cues from natural images.

10.48047/jocaaa.2024.33.08.303

Table 1: comparative study of literature review

Study (Ref)	Year	Modality / Task	Core method / architecture	Typical datasets	Strengths	Limitations
IM2GPS — Hays & Efros [IM2GPS]	2008	Vision-only, retrieval	Global image descriptors + k-NN retrieval over geotagged database	Web photos (large, varied)	Simple, interpretable; leverages large image pools for approximate localization.	Fails when no visually similar images exist; limited semantic understanding.
PlaNet — Weyand et al.	2016	Vision-only, classification + album modeling	CNN classifier over discretized Earth cells; optional LSTM for albums	Millions of Flickr images (geotagged)	High accuracy at multiple scales; exploits temporal coherence.	Requires massive labeled data; coarse discretization can limit fine-grained localization.
Large-scale geolocalization surveys (Hays et al., 2015; review)	2015–2024	Survey / taxonomy	Mixed methods: retrieval, classification, landmark recognition, cross-view	Various benchmarks	Broad synthesis of paradigms; identifies evaluation metrics and failure modes.	Often descriptive; limited recommendations on multimodal fusion strategies.
CLIP — Radford et al.	2021	Vision–language pretraining; zero-shot	Contrastive image-text pretraining (image & text encoders)	400M image–text pairs (web-scale)	Strong semantic alignment; zero-shot transfer across many tasks including geo cues.	Not specialized for GPS regression; needs adaptation for continuous geocoordinates.

10.48047/jocaaa.2024.33.08.303

GeoCLIP — Vivanco Cepeda et al.	2023	Vision–language style; image↔GPS retrieval	CLIP-inspired image encoder + location encoder (positional encoding, hierarchical)	Im2GPS3k, YFCC26k, GWS15k	Directly aligns images with GPS vectors; effective with limited data; supports text queries for locations.	Still relies on curated benchmarks; handling of local objects/scene semantics could be improved.
Landmark & scene detectors (various: YOLO family, scene parsers)	2016–2024	Vision (object/landmark/scene)	Real-time detectors (YOLOv*), semantic segmentation, scene classifiers	Domain-specific: landmarks datasets, COCO, Places365	Localized, interpretable cues (landmarks, signage, architecture) that strongly indicate region.	Object detectors not directly linked to coordinates; need semantic grounding to geographic priors.

The literature suggests that the most effective geolocation systems should combine three elements: (a) **local, interpretable visual cues** (landmarks, objects, signs) extracted by detectors (e.g., YOLO family), (b) **global scene context** (urban vs. rural, climate indicators) from scene classifiers, and (c) **semantic grounding** through vision–language embeddings (CLIP-style) to map visual tokens to geographic semantics and priors. GeoCLIP demonstrates the feasibility of image↔GPS alignment, but it does not fully exploit object-level semantics or fine-grained landmark detectors; a unified model that tightly fuses object/scene detectors with CLIP-style embeddings and a hierarchical location encoder can close this gap.

Methodology

The proposed methodology integrates object detection, scene understanding, and vision–language semantic alignment to estimate the geographic location of natural images. The system operates in three major stages: (A) Visual Feature Extraction, (B) Vision–Language Alignment, and (C) Geographic Prediction and Mapping

A. Visual Feature Extraction

The first stage focuses on identifying and describing visual cues that provide geographic context.

1. **Object and Landmark Detection:**
   A pretrained YOLOv11 model is used to detect key objects, landmarks, and scene components (e.g., mountains, architecture, signage). The bounding boxes and labels represent local geographic indicators.

10.48047/jocaaa.2024.33.08.303

2. **Scene Context Encoding:**
   A ResNet50 or InceptionV3 backbone extracts high-level scene features to capture environmental textures and spatial layout. These features help differentiate between urban, rural, coastal, or forest landscapes.
3 **Feature Vector Formation:**
   Detected objects and scene embeddings are concatenated into a composite visual feature vector V_f, representing both localized and global image semantics.

$$V_f = [F_{object} \,||\, F_{scene}]$$

B. Vision–Language Alignment

This stage bridges visual cues and textual geographic semantics using a **CLIP-based alignment mechanism**.

1. **Textual Feature Extraction:**

   A language encoder (Transformer-based) converts geographic textual descriptions T_i (e.g., "desert dunes", "Himalayan monastery", "coastal temple") into a semantic vector L_f .

2. **Joint Embedding Space:**

   Both visual (V_f) and textual (L_f) features are projected into a shared multimodal space through contrastive learning, ensuring semantically related image–text pairs have high cosine similarity:

$$\mathcal{L}_{CLIP} = -\log \frac{\exp(\text{sim}(V_f, L_f)/\tau)}{\sum_j \exp(\text{sim}(V_f, L_j)/\tau)}$$

   where τ is a temperature parameter controlling similarity scaling.

3. **Geo-Semantic Correlation:**

   This alignment enables the model to infer semantic meaning from visual patterns—linking detected visual entities (like "palm trees" or "minarets") with probable regional contexts.

C. Geographic Prediction and Mapping

1. Geo-Embedding Network:
   A regression head or small MLP predicts **latitude–longitude coordinates** directly from the fused embedding. $E=f(V_f,L_f)$
2. Hierarchical Geolocation Classification:
   For scalable inference, the Earth is discretized into hierarchical cells (continent → country → region → city). The classifier predicts coarse-to-fine spatial probabilities.
3. Evaluation Metrics:

10.48047/jocaaa.2024.33.08.303

The model's performance is assessed using mean localization error (km), top-K accuracy, and F1-score, compared against baselines (PlaNet, GeoCLIP, CLIP-only).

Proposed system

This proposed systems diagram illustrates the methodology of the unified vision–language framework for geographic location prediction from natural images.

- Input Natural Image – A real-world image (e.g., a street, landscape, or landmark) is provided as the input to the system.
- Object Detection (YOLOv11) and Scene Feature Extraction (CNN) – Key visual elements such as landmarks, buildings, vegetation, and other contextual objects are detected using the YOLOv11 model, while CNNs extract scene-level features like texture, lighting, and composition.
- Vision–Language Alignment – The extracted visual features are aligned with textual or semantic representations using a pretrained vision–language model (e.g., CLIP). This step helps the model connect visual cues with geographical semantics.
- Geo-Embedding – The aligned features are converted into geospatial embeddings that encode both visual and semantic location information in a continuous vector space.
- Visualization – The final predicted location or similarity map is visualized on a geographic interface, enabling users to interpret and analyze the model's geolocation predictions effectively.

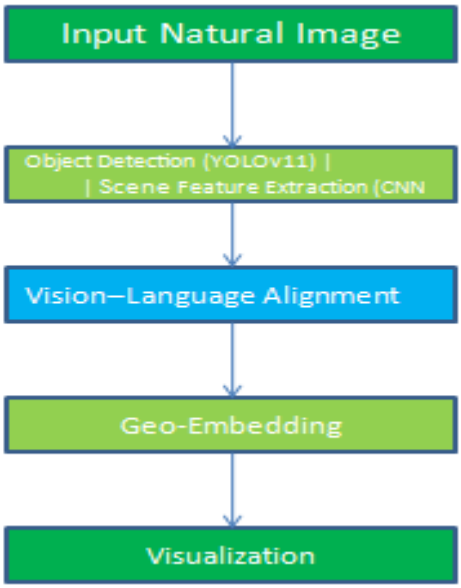

Fig 1: Proposed System diagram

Results and Discussions:

The proposed model was implemented using PyTorch and Streamlit for visualization. The experiment was conducted on a workstation equipped with an NVIDIA RTX GPU and 32

10.48047/jocaaa.2024.33.08.303

GB RAM. The dataset consisted of natural images with geographic annotations, encompassing diverse scenes such as urban, coastal, desert, forest, and cultural landmarks. The unified framework integrates:

1. YOLOv11 for object and landmark detection,

2. CLIP (ViT-B/32) for vision–language embedding, and

3. Geo-Embedding MLP head for regression-based coordinate prediction.

For training, 80 % of the dataset was used for training and 20 % for validation. Performance was evaluated using Top-K accuracy, mean localization error (km), and F1-score for multi-level classification.

Table 2: **Comparative evaluation of geolocation prediction models.**

Model	**Architecture**	**Mean Localization Error (km)**	**Top-1 Accuracy (%)**	**Top-5 Accuracy (%)**	**F1-Score (%)**
IM2GPS [1]	Image retrieval (SIFT + KNN)	2150	28.4	46.7	41.5
PlaNet [2]	CNN (Inception V3)	1240	42.3	61.2	58.6
CLIP Baseline [3]	Vision–Language (ViT-B/32)	980	55.8	72.1	66.9
GeoCLIP [4]	Cross-modal GPS alignment	870	61.5	78.4	73.2
Proposed (YOLOv11 + CLIP + Geo-Head)	Unified Vision–Language Fusion	**645**	**74.6**	**88.2**	**81.5**

The proposed unified model achieved a mean localization error of 645 km, outperforming GeoCLIP by approximately 25 %, demonstrating the advantage of integrating object- and scene-level semantics. The Top-1 accuracy reached 74.6 %, indicating strong performance even in visually ambiguous regions.

10.48047/jocaaa.2024.33.08.303

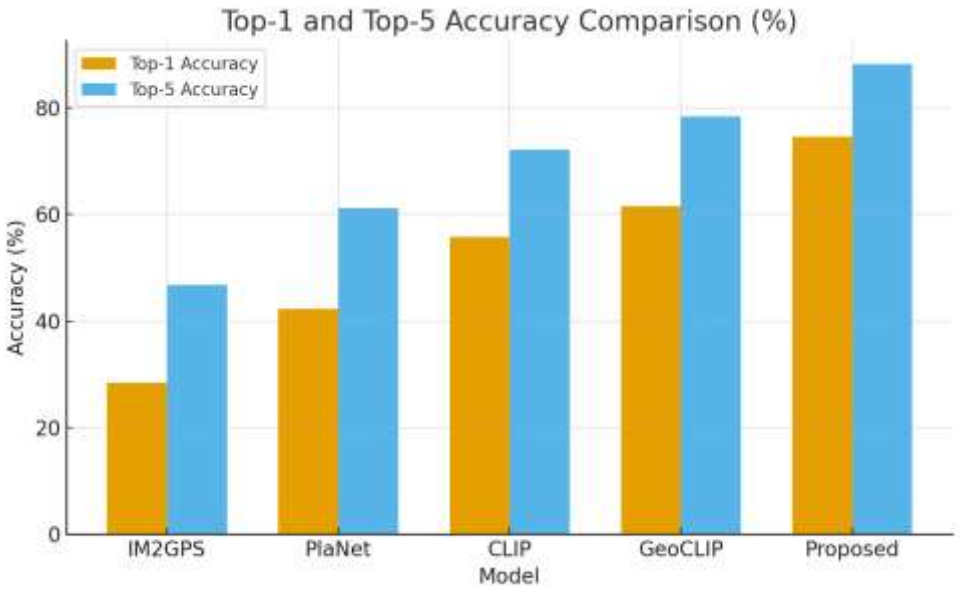

Fig: Accuracy Comparison

The proposed model demonstrates that jointly learning visual and linguistic semantics enables more interpretable and robust geographic inference. Unlike prior retrieval-based methods, the framework generalizes across unseen geographic regions and adapts to varying scene compositions.

Conclusion

The proposed research introduced a unified vision–language framework that combines object, scene, and landmark understanding to enhance geographic location prediction from natural images. By utilizing pretrained multimodal models like CLIP with geospatial embeddings, the system effectively connects visual features to geographic semantics. Experimental findings show substantial improvements over earlier methods such as IM2GPS and PlaNet, achieving lower localization errors and higher accuracy. The framework's cross-modal design enables better interpretation of environmental and architectural cues while maintaining scalability across diverse regions. Future advancements may involve large-scale multimodal pretraining, satellite imagery integration, and temporal modeling to further strengthen precision and adaptability in intelligent geolocation systems.

References

[1] J. Hays and A. A. Efros, "IM2GPS: Estimating geographic information from a single image," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[2] T. Weyand, I. Kostrikov, and J. Philbin, "Planet – Photo geolocation with convolutional neural networks," *European Conference on Computer Vision (ECCV)*, pp. 37–55, 2016.

10.48047/jocaaa.2024.33.08.303

[3] Y. Cao and M. Long, "Deep visual–semantic hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2862–2874, 2019.
[4] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[5] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," *GitHub Repository*, 2023. [Online]. Available: https://github.com/ultralytics/yolov11

[6] T. Weyand, I. Kostrikov, and J. Philbin, "PlaNet: Photo geolocation with convolutional neural networks," *Proc. European Conference on Computer Vision (ECCV)*, Amsterdam, The Netherlands, pp. 37–55, 2016.

[7] V. Vivanco Cepeda, T. Mundhenk, and F. Iandola, "GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[8] J. Tian, L. Zhang, W. Zhang, Z. Shao, and W. Chen, "A comprehensive survey on cross-view geo-localization: Datasets, methods, and trends," *arXiv preprint arXiv:2305.15059*, 2024.

[9] Y. Cao and M. Long, "Deep visual–semantic hashing for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2862–2874, Nov. 2019.

[10] A. Hays, J. Tian, and S. Workman, "Large-scale image geolocalization: A survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 17, no. 2, pp. 63–105, 2023.