Image Geolocation within China

Yiqi Dong
IIIS
Tsinghua University
dongyq24@mails.tsinghua.edu.cn

Langrui Qi
IIIS
Tsinghua University
qlr24@mails.tsinghua.edu.cn

Guotao Shi
IIIS
Tsinghua University
sgt24@mails.tsinghua.edu.cn

Abstract

Our project develops an Image2Geolocation system which combines multiple methods and capable of predicting the geographic location of an input image within China, which has strong potential for real-world applications in Open source intelligence, an important area of Intelligence studies.

1 Introduction

Image-to-geolocation (Image2Geo) is the task of predicting the geographic origin of a given image based solely on its visual content. This problem has garnered increasing attention in recent years due to its wide range of applications in areas such as open-source intelligence, disaster response, social media verification, and entertainment. While many previous works have focused on global-scale geolocation—attempting to localize images anywhere on Earth—there is growing interest in solving this task within specific countries or regions, where visual cues can be more subtle and the distribution of image content more imbalanced.

Most traditional approaches to Image2Geo rely on computer vision (CV) techniques such as CNN-based classification, image retrieval, or handcrafted feature matching. Systems like PlaNet or the popular online game GeoGuessr have demonstrated the feasibility of global image localization. However, these methods tend to struggle in country-scale settings like China, where regional architectural styles, landscapes, and signage often show only minor visual differences. Furthermore, conventional CV systems are generally limited to surface-level feature extraction, lacking the capacity for context-aware reasoning, such as interpreting landmarks or recognizing textual hints in images.

In contrast, recent advances in Large Vision-Language Models (LVLMs)—notably the multimodal capabilities of models like ChatGPT-o3—have shown that these models can perform high-level reasoning about images when guided through Chain-of-Thought (CoT) prompting. For example, when asked to infer a location from an image of a street scene, these models can analyze visual clues such as architecture, language on signs, vegetation, and road design to produce surprisingly accurate geolocation guesses. This emerging capability suggests that combining LVLMs with traditional CV pipelines could enhance localization performance, particularly in tasks that demand contextual understanding.

Another source of inspiration for our work comes from the increasing popularity of TuXun , a competitive online game that challenges players to guess the location of real-world street-view images within China. The game highlights both the feasibility and difficulty of country-level geolocation and provides a benchmark for evaluating human and AI performance on this task.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In this project, we propose a hybrid image geolocation system specifically designed for images within China. Our system integrates multiple modules: a fast, discriminative CV-based classification pipeline; an image-text recognition module for extracting location-relevant keywords; and a reasoning module powered by LVLMs with CoT prompting. This design enables both efficient image filtering and fine-grained inference, aiming to balance speed, interpretability, and accuracy.

2 Related Work

2.1 Traditional Image-to-Geolocation Methods

Early works in image geolocation, such as Im2GPS and Google's PlaNet, frame it as a classification or retrieval task using CNNs trained on large-scale image–location pairs. These methods demonstrated promise at global scale but typically falter when distinguishing between visually similar regions within a single country, where subtle visual differences (e.g., architecture, signage) play a decisive role.

2.2 Modern Image-to-Geolocation Methods

One of the earliest deep learning approaches to image geolocation was PlaNet, which formulated the task as a classification problem over geocells using convolutional neural networks. While effective at a global scale, its performance deteriorates in regions with subtle intra-country visual cues.

Recently, Haas et al. introduced PIGEON [1], a state-of-the-art system that leverages semantic geocell creation and contrastive pretraining to predict image locations at planetary scale. Unlike previous classification-based models, PIGEON incorporates image-text representations and learns geospatial relationships in a hierarchical manner. While their model performs remarkably well globally, it does not specifically address fine-grained localization within a single country like China, where visual ambiguity poses unique challenges. Nonetheless, their approach inspires our hybrid architecture by demonstrating the benefit of combining semantic priors with deep representations.

2.3 LVLMs for Geolocation with Chain-of-Thought Reasoning

Recent studies highlight the remarkable capabilities of Large Vision-Language Models (LVLMs) in geolocation when guided by Chain-of-Thought (CoT) prompting.

Image-Based Geolocation Using LVLMs demonstrates that an LVLM using CoT to analyze visual cues—such as vehicle styles, architecture, landscape, and cultural elements—can outperform traditional models and even humans in TuXun.

3 Methodology

3.1 Datasets

As previous works mainly focus on global level image Geolocation recognition and use abroad sources like google map streetview as their dataset, very few of the data is located at China Mainland, so we must construct our own dataset. To construct a large-scale China-specific geolocation dataset, we collected over 10,000 images based on the Baidu Maps and Tencent Maps APIs. Each image is associated with ground-truth location coordinates (latitude and longitude), which are explicitly provided by the respective map service APIs. We randomly split the dataset into 80% for training and 20% for validation.

3.2 System Architecture Overview

Our Image2Location system processes an input image through several interconnected modules to determine its geographical origin, it combines traditional CV approaches and modern LVLM with CoT. Also an Image text recognition block is used considering significant high frequency of word appearance in China Image2Geolocation.

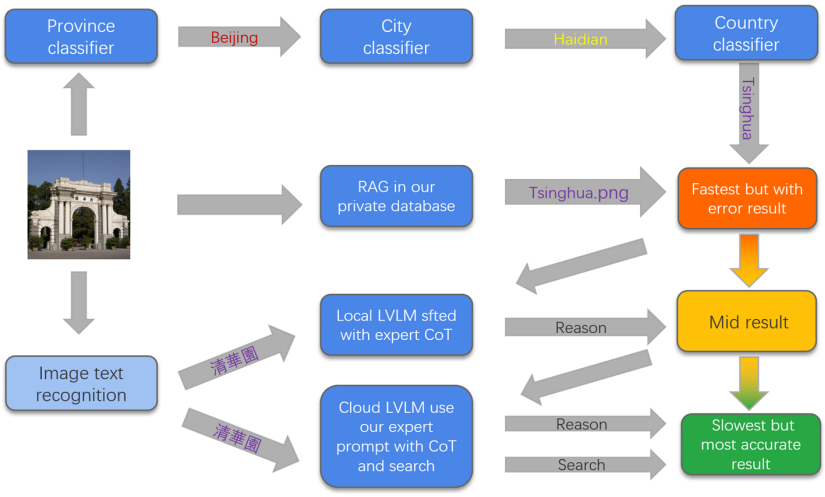

Figure 1: Our system Architecture

3.3 Multi-Stage Classification Pipeline

Given an input image, our system first passes it through a hierarchical classification pipeline consisting of three stages:

- Province Classifier: A lightweight convolutional neural network (CNN) predicts the province-level location.
- City Classifier: Given the province, a city-level classifier narrows the candidate regions.
- County/District Classifier: A final classifier predicts the most likely district.

This fast classification track provides an initial estimate, but may be error-prone, especially for ambiguous images with few distinguishing features.

3.4 Text-based Recognition

We noticed that in the Chinese dataset, due to the high population density in China, the text in the image is very helpful for image search. Therefore, we used an OCR model to extract the text in the image as the prompt of LVLM.

To extract semantic cues from the image, we incorporate a text recognition module based on PaddleOCR, a high-performance open-source OCR system developed by Baidu. PaddleOCR is capable of detecting and recognizing multi-language text with high accuracy, and supports complex scene text with various fonts and orientations.

In our pipeline, the OCR module identifies any visible textual content in the image—such as signs, landmarks, or address markers—and outputs the recognized text as a list of keywords.

3.5 LVLM-Based Inferential Reasoning

To further improve accuracy, we incorporate both local and cloud-based LVLMs:

- The local LVLM, fine-tuned with expert-designed Chain-of-Thought (CoT) prompts, reasons over the image content and text to provide a more informed estimate.
- The cloud LVLM (e.g., gemini) uses a more powerful reasoning prompt along with external search to produce the most accurate but computationally expensive result.

This multi-path reasoning mechanism allows us to balance speed and accuracy according to the application need.

3.6 Retrieval-Augmented Generation with LVLM

To enhance the reasoning capability of the vision-language model (LVLM), we integrate a Retrieval-Augmented Generation (RAG) framework into our pipeline. RAG augments the LVLM by retrieving relevant external knowledge based on the visual content and/or user queries, allowing the model to access information beyond its parametric memory.

3.7 Fine-Tuning with CoT Supervision

To enhance the reasoning ability of our local vision-language model, we perform supervised fine-tuning (SFT) on Qwen-VL, a powerful open-source Chinese vision-language model. The goal is to make the model more capable of deducing location-relevant clues from image content, even in the absence of explicit text.

We construct the fine-tuning dataset using Chain-of-Thought (CoT) explanations generated by Gemini (a proprietary LVLM known for its high-quality reasoning). For a large number of diverse geolocation images, we prompt Gemini with expert-designed CoT templates,not that the accurate answer is included in the prompt. The resulting explanations—often referencing features like architecture, terrain, license plates, or language signs—are paired with the original images as training data.

We then fine-tune Qwen-VL on this dataset to imitate Gemini's reasoning process. The fine-tuned model can now generate interpretable and accurate location guesses, even on previously unseen Chinese regions.

This CoT-based SFT strategy allows us to distill expensive reasoning capabilities from proprietary models into a lightweight, local model suitable for on-device inference.The code of this part is based on [2]

4 Experiments

To evaluate the effectiveness of our Image2Geolocation system, we use a held-out test set consisting of several thousand real-world images collected from the TuXun platform. Each image is annotated with its ground-truth GPS coordinates (latitude and longitude) based on the in-game metadata.

We compute the geodesic distance error between the predicted and ground-truth coordinates using the haversine formula, which calculates the shortest path over the Earth's surface.

4.1 Metrics

We report the following evaluation metrics:

- Mean Geolocation Error (in kilometers)
- Percentage of predictions within X km radius, for X = 200 km and 500 km

4.2 Results

Metric	Value
Arithmetic Mean Error Distance (km)	409.2
Geometric Mean Error (km)	165.5
% of images with error < 200 km	42.1%
% of images with error < 750 km	73.7%

Table 1: Geolocation performance on TuXun test dataset

Notably, the geometric mean error is significantly lower (165 km), indicating that the majority of predictions are relatively close to the ground truth, but a few large-error outliers inflate the arithmetic mean.

These results indicate that our system is able to localize a substantial proportion of images within a few hundred kilometers, despite the inherent difficulty of the task.

Metric	Value
Arithmetic Mean Error Distance (km)	975.74
Geometric Mean Error (km)	219.16
% of images with error < 200 km	34.6%
% of images with error < 750 km	65.4%

Table 2: gemini-2.5-pro performance on TuXun test dataset

These levels of precision are encouraging given the size and geographic diversity of China, and demonstrate the promise of combining hierarchical classification, OCR-based retrieval, and LVLM-based reasoning.

5 Conclusion

In this project, we present a hybrid system for image-based geolocation within China, combining hierarchical classifiers, OCR-based retrieval, and vision-language reasoning. Inspired by the growing popularity of TuXun and recent advances in large vision-language models (LVLMs), we aim to strike a balance between speed, interpretability, and localization accuracy.

Our method demonstrates promising performance on a real-world dataset collected from the TuXun platform, achieving a mean error of approximately 409 km, with about 74% of predictions falling within 750 km of the ground truth. These results validate the effectiveness of integrating both discriminative and generative components, especially in a linguistically and visually diverse environment like China.

Furthermore, we show that fine-tuning an open-source LVLM (Qwen-VL) using CoT-style supervision distilled from Gemini significantly enhances local inference capability, enabling interpretable, step-by-step reasoning about visual clues.

Future directions include incorporating map priors or satellite imagery, and reducing dependency on cloud-based models for cost-efficient deployment.

References

[1] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. *arXiv*, 2307.05845, 2023. Accepted at CVPR 2024.

[2] Gelei Deng Yuekang Li Tianwei Zhang Weisong Sun Yaowen Zheng Jingquan Ge Yang Liu Yi Liu, Junchen Ding. Image-based geolocation using large vision-language models. 2024.