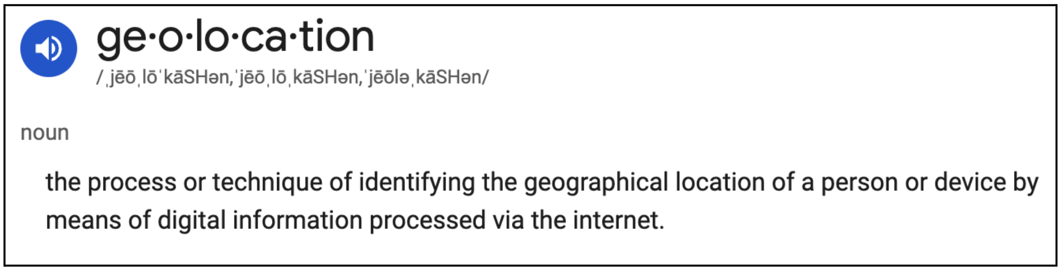
PROJECT VOILÀ

Richard Cheng

Project Overview



ge·o·lo·ca·tion

/ˌjēōˌlōˈkāSHən,ˈjēōˌlōˌkāSHən,ˈjēōləˌkāSHən/

noun

the process or technique of identifying the geographical location of a person or device by means of digital information processed via the internet.

Problem Statement:

The application of machine learning in geolocation isn't new, and existing AI-powered tools work well in urban environments.

However, existing approaches struggle with pictures taken in rural and wilderness areas, where scenes are visually more generic and less distinguishable. Naturally, more photos are taken by humans in places where humans visit more. This results in dataset imbalance and ambiguity, making direct "images → coordinates" prediction difficult.

Thus, I developed Project Voilà to solve this problem.

Project structure:

Voilà is a two-component system:

1. A convolutional neural network that takes in photos taken at ground level (ground CNN) learns a geographic embedding and uses a GeoIndex to retrieve a shortlist of plausible 5° regions.
2. A digital elevation model (DEM)-based reasoner then verifies which candidate region is most consistent with terrain patterns from ASTER GDEM.

The final output is the best-matching 5° cell (centroid) with the option to refine within-cell.

Central Idea:

Instead of forcing a model to guess globally in one shot, Voilà solves geolocation by retrieving plausible regions then verifying the candidates with terrain.

Component 1: Ground CNN + GeoIndex Retrieval

Goal of Component 1:

Given a ground image, produce a geographically meaningful embedding and retrieve a shortlist of candidate regions.

Design rationale:

- Classification/regression gives structured geo-supervision.
- Metric learning shapes the embedding space for nearest-neighbor search.

Ground model outputs:

- 256-D L2-normalized embedding
- grid classification over global cells (5° bins)
- optional 2D offset inside the predicted cell

GeoIndex design:

- embeddings (N×256)
- per-sample cell_id (5° cell)
- per-cell metadata (centroid lat/lon, counts, ranges)

Retrieval behavior:

- FAISS retrieves nearest samples
- results are deduped by cell to form Top-K candidate cells

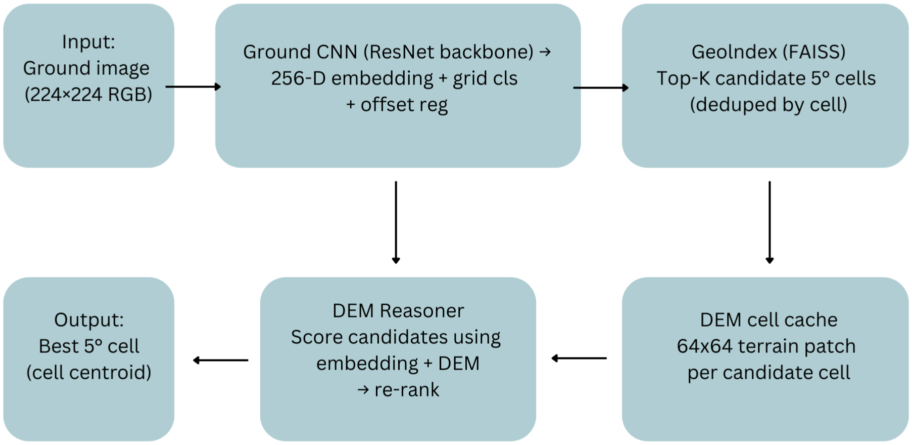


Figure 1: Model Architecture

Component 2: DEM Cache + Geospatial Reasoner

Goal of Component 2:

Given the Component 1 shortlist, pick the correct region using terrain evidence.

Design rationale:

- Terrain patterns are stable and informative even when RGB appearance is generic (flat coastal plains vs mountain valleys, etc.).

DEM cache design:

- For each geo-cell, precompute a 64×64 terrain patch extracted from ASTER GDEM tiles.
- ASTER GDEM provides near-global elevation coverage (83°N to 83°S) via tiled DEM products.

Reasoner model design:

- Input: (image embedding + K candidate DEM patches)
- Output: score per candidate → select best cell
- Training: cross-entropy over candidates

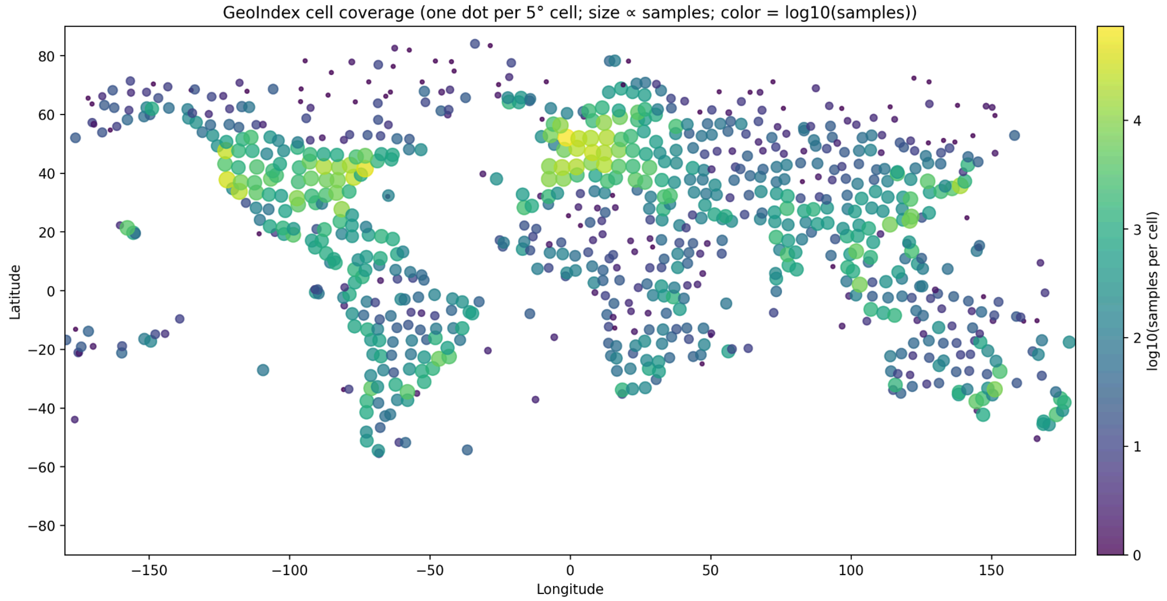


Figure 2: Global GeoIndex Cell Coverage

Development timeline

- Months 1-3:
  - literature review
  - system design
  - data planning
- Months 4-10:
  - implementation of GViT approach
  - training
  - evaluation and failure
- Months 11-15:
  - literature review
  - system redesign
  - implementation of GCNN + GeoIndex + DEM cache approach
  - training
  - evaluation and improvement

Initial failed GViT (single visual transformer) approach:

- oversimplified approach
- forced to guess globally in one step
- insufficient data to support the inherently data-hungry visual transformer

Results & Reflection

Table: Localization Error (km), Final Approach vs. Initial Approach

Highest across all test configurations	**GViT accuracy (km)**	**GCNN + GeoIndex + DEM cache accuracy (km)**
Mean	5762.53	408.09
Median	7198.43	63.94
Top1* mean	N/A	89.04
Top1* median	N/A	10.66

*Excluding self-matches (when image is already present in GeoIndex; corrupts the data)

Interpretation of results:

Compared to the initial GViT, the final pipeline (GCNN + GeoIndex + DEM cache) massively improves accuracy. It narrows the correct location to a city/region scale (tens to hundreds of kilometers) rather than continent scale (thousands of kilometers).

Reflection:

Voilà demonstrates that modular structure converts a global inverse problem into a tractable shortlist plus verification problem. Terrain acts as an independent source of evidence that improves robustness under RGB ambiguity, allowing improved accuracy for geolocating photos in rural and wilderness areas.

Even with the huge improvements after switching to the new GCNN + GeoIndex + DEM cache approach, the results that Project Voilà has achieved are not good enough for my intended application in investigative journalism and helping me find locations for photography. However, the past 15 months have taught me that the right response to failure is iteration and redesign. After the first 10 months of hard work produced no results, I pivoted to a new approach and achieved much-improved results. Thus, I intend to keep Project Voilà an ongoing project and continue exploring new approaches that may result in better model accuracy.

Code

https://github.com/jochueh/Project-Voila

Datasets & References

https://github.com/jochueh/Project-Voila/blob/main/README.md