Exploring Planet-Scale Image Geolocation with PIGEON

[Hoffmann Szabocs]

January 8, 2026

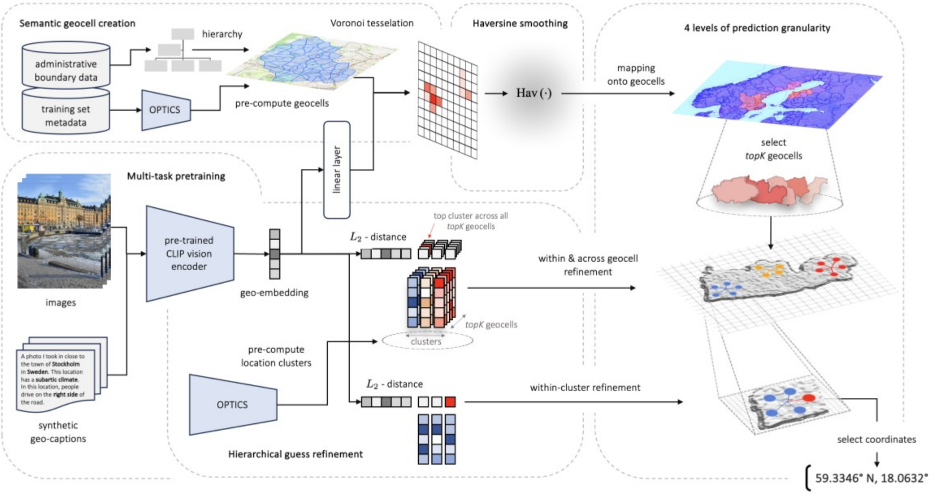
Aims of the Project

- Understand how modern AI systems infer geographic location from a single image.
- Study PIGEON / PIGEOTTO as a concrete, state-of-the-art example.
- Analyse how CLIP, semantic geocells and refinement in embedding space interact.
- Develop and articulate my own refinement idea based on weighted barycentric coordinates.
- Sketch future directions that combine PIGEON with a text-aware, ChatGPT-like model.

PIGEON

- Reframes geolocation as **classification** over geocells instead of direct coordinate regression.
- Uses a **CLIP-based vision encoder** to map images into a shared embedding space.
- Trains a geocell head with **distance-aware label smoothing** based on haversine distance.
- Adds a **non-parametric refinement layer** that retrieves training locations in embedding space.
- Achieves high accuracy by combining global classification with local nearest-neighbour refinement.

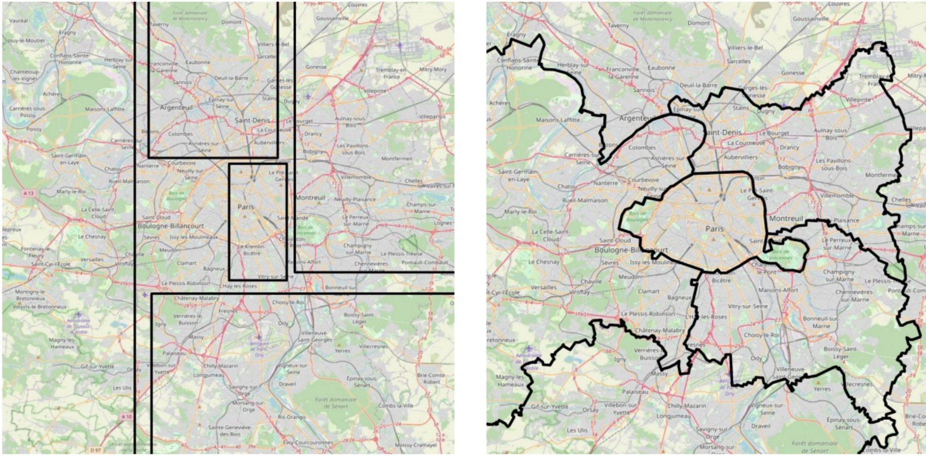
PIGEON: Predicting Image Geolocations



A photo I took in close to the town of **Stockholm** in **Sweden**. This location has a **subartic climate**. In this location, people drive on the **right side** of the road.

Semantic Geocells

- Earlier systems relied on artificial grids with arbitrary boundaries.
- PIGEON builds **semantic geocells**:
  - starts from real administrative polygons (countries, regions),
  - merges small regions until each cell has sufficient training data,
  - further splits dense urban areas by clustering locations (OPTICS) and using Voronoi cells.
- The label space respects political borders where possible while providing higher resolution in data-rich, visually complex regions.

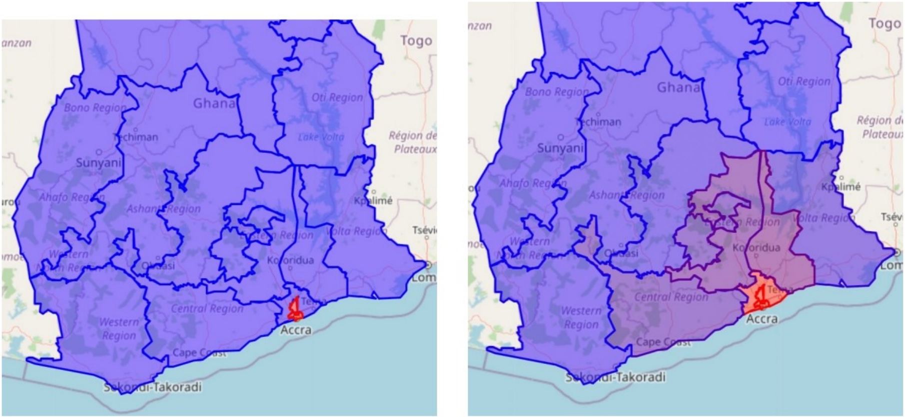


(a) With naive, rectangular geocells.

(b) With our semantic geocells.

Distance-Aware Label Smoothing

- Instead of one-hot geocell labels, PIGEON uses a **distance-aware** target distribution.
- For each image, the true coordinates are compared to every cell centroid via the **haversine distance**.
- Probabilities are assigned so that nearby cells receive non-zero mass; distant cells are strongly down-weighted.
- This formalises the idea that a nearby error is less severe than a distant one and cuts issues at geocell boundaries.



(a) Without haversine smoothing.

(b) With haversine smoothing.

Figure 3: Impact of applying haversine smoothing over neighboring geocells for a location in Accra, Ghana.

Refinement in Embedding Space

- After training the geocell classifier, PIGEON adds a refinement stage in CLIP embedding space.
- Within each geocell, training locations are **clustered** and represented by prototype embeddings.
- At inference:
  - the model selects the top–K most probable geocells,
  - chooses the best cluster within this subset,
  - outputs the coordinates of the **nearest training location** in that cluster.
- Parametric network narrows down the region; refinement behaves like a nearest-neighbour retrieval mechanism.

My Refinement Idea (1): Motivation

- The original refinement picks a **single** nearest training location.
- This can be sensitive to:
  - outliers,
  - mislabelled images,
  - local irregularities in the training set.
- I propose to replace this with a **weighted barycentric** prediction over several neighbours.
- Intuition: use the structure of the top–K geocells and nearby examples more smoothly.

My Refinement Idea (2): Weighted Barycentric Scheme

- Step 1: obtain the top–K geocells from the classifier and restrict to training points inside them.
- Step 2: in CLIP space, find the M nearest training images to the query.
- Step 3: for neighbours with coordinates (λ_i, ϕ_i) and distances d_i, define

$$w_i = \frac{\exp(-\alpha d_i)}{\sum_{j=1}^{M} \exp(-\alpha d_j)}, \qquad \alpha > 0.$$

- Step 4: predict

$$\hat{\lambda} = \sum_{i=1}^{M} w_i \lambda_i, \qquad \hat{\phi} = \sum_{i=1}^{M} w_i \phi_i.$$

Barycentric Refinement: Intuition

- Top–K geocells still provide a **spatial prior**, limiting the search region.
- The barycentric combination:
  - smooths piecewise-constant nearest-neighbour behaviour,
  - may reduce sensitivity to single atypical or mislabelled points,
  - can be seen as a “soft” Voronoi assignment or kernel regression in embedding space.
- The method is compatible with the existing PIGEON pipeline and could be tested with the same evaluation metrics.

Future Direction: Text-Aware Geo-AI

- PIGEON excels at using visual cues, but it does not explicitly **read text** in the scene.
- A multimodal ChatGPT-like model could:
  - perform OCR on street signs, shop names, number plates,
  - detect language and script,
  - apply world knowledge about city names, brands, and road systems.
- These textual observations could be turned into soft constraints on the likely region, country, or even a particular city.
- Such constraints could be used to reweight geocell probabilities or neighbour weights in the barycentric scheme.

Conclusion and Next Steps

- I developed a structured understanding of PIGEON's architecture:
  - semantic geocells,
  - distance-aware label smoothing,
  - refinement in embedding space.
- I proposed a refinement method based on weighted barycentric coordinates inside the top–K geocells.
- I outlined how a text-aware, ChatGPT-like model could complement PIGEON by exploiting textual cues.
- Next steps:
  - implement toy experiments with barycentric refinement,
  - design simple fusion schemes between visual and textual constraints,
  - move towards a concrete research plan and prototype system.

AI Usage

- The LaTeX code for this slideshow
- Some assistance translating and understanding larger concepts in the original article
- Testing ChatGPT as an OCR model