MMLandmarks: a Cross-View Instance-Level Benchmark for Geo-Spatial Understanding

Oskar Kristoffersen[1,2], Alba Reinders Sánchez[1], Morten Rieger Hannemose[1,2],
Anders Bjorholm Dahl[1,2], Dim P. Papadopoulos[1,2]

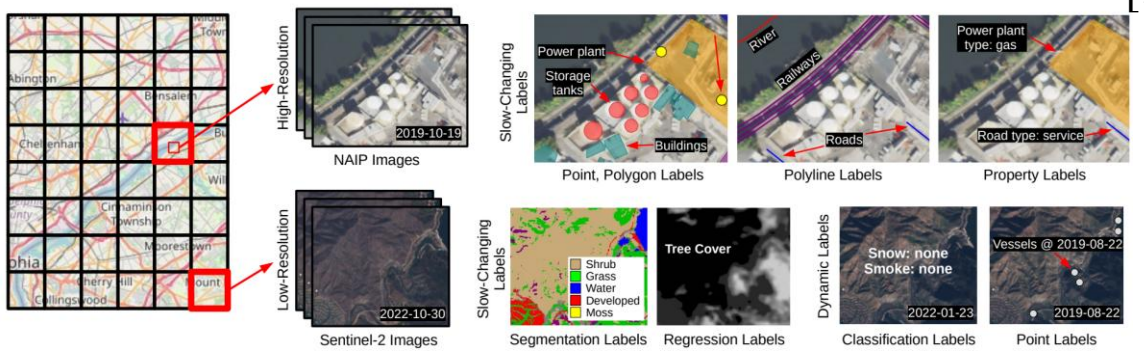
Technical University of Denmark[1]

Pioneer Center for Artificial Intelligence[2]

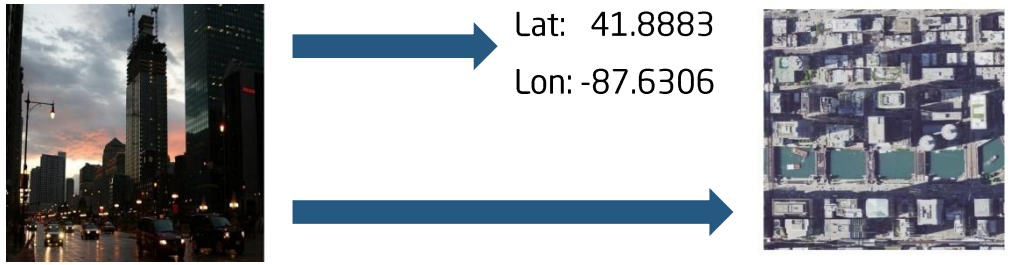
Motivation - Geospatial data

[1]

- **Multimodal data**
  - Strong embeddings
  - Foundation models for geospatial tasks



- **Aligning modalities** from the **same location**
  - Geolocalization
  - Cross-view localization

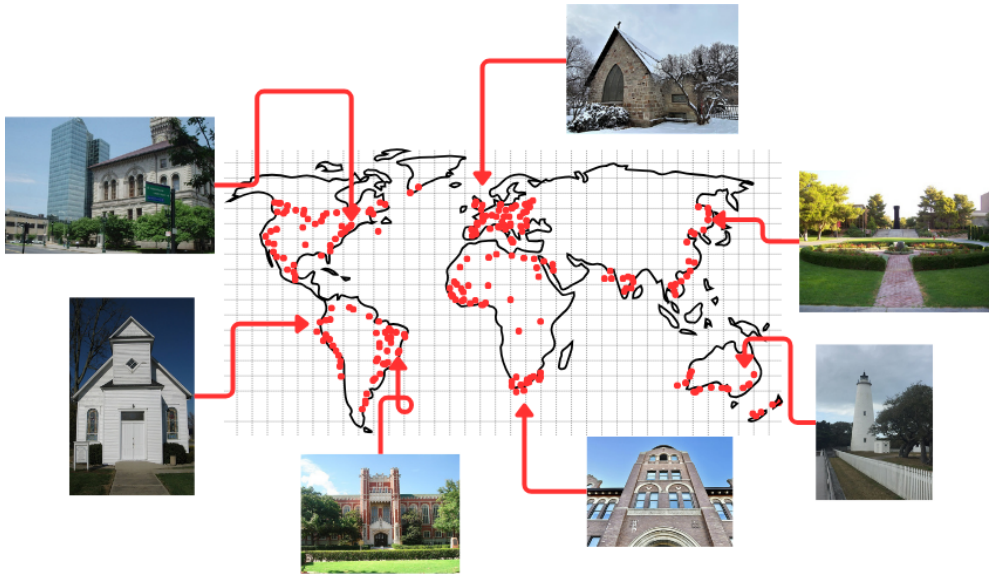


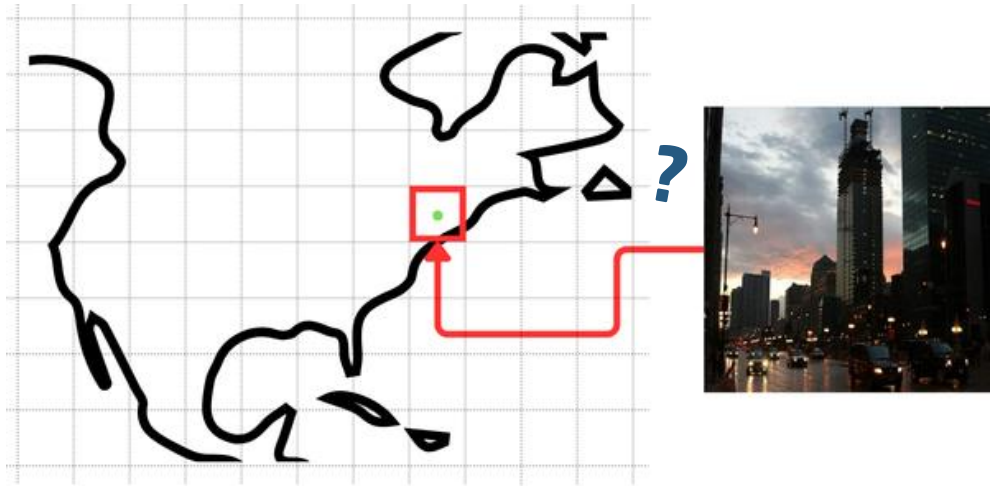
- Models solve **individual geospatial tasks** in **isolation**.
- No geospatial dataset supports **instance-level** correspondence across **multiple modalities**.

[1] Bastani et al. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding, ICCV'23

Motivation - Geolocalization

- Cell-based classification from a grid
  - Lacks fine-grained predictions (<10km)
  - Difficult to capture enough information from a single image



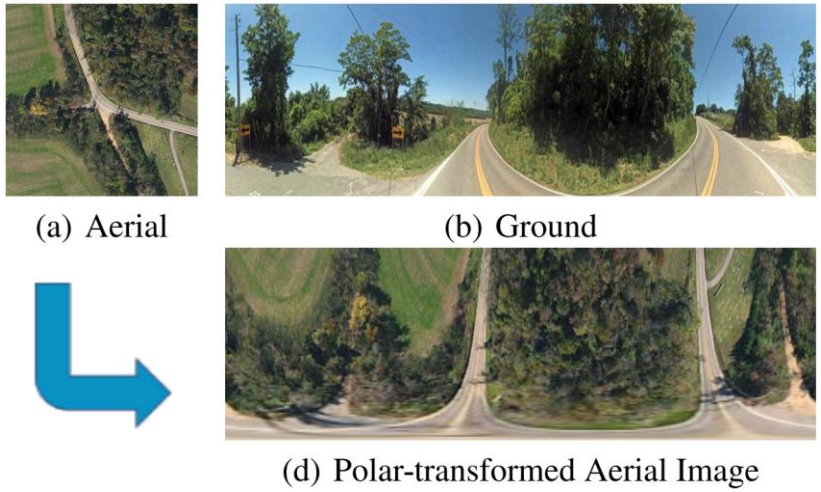


Motivation – Cross-view Localization

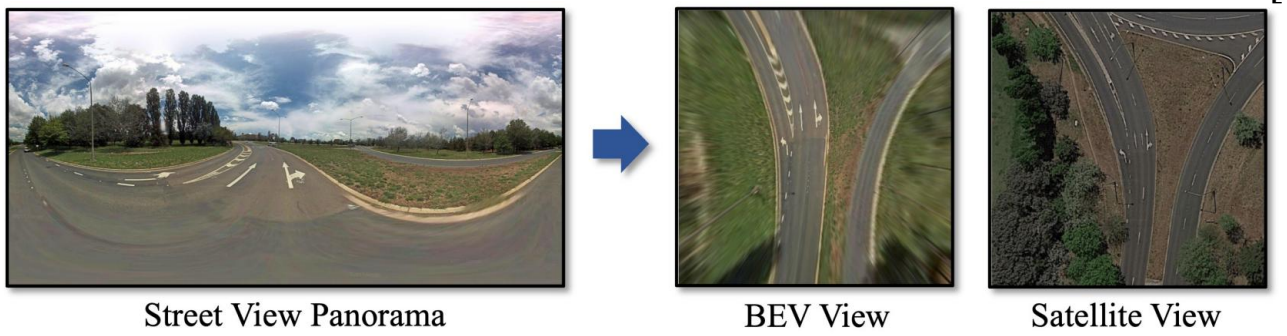
Current Cross-View datasets:

- 360-degree street-view images
- Have limited coverage & diversity: Streets and cities
- Build reliance on road structure

[2]

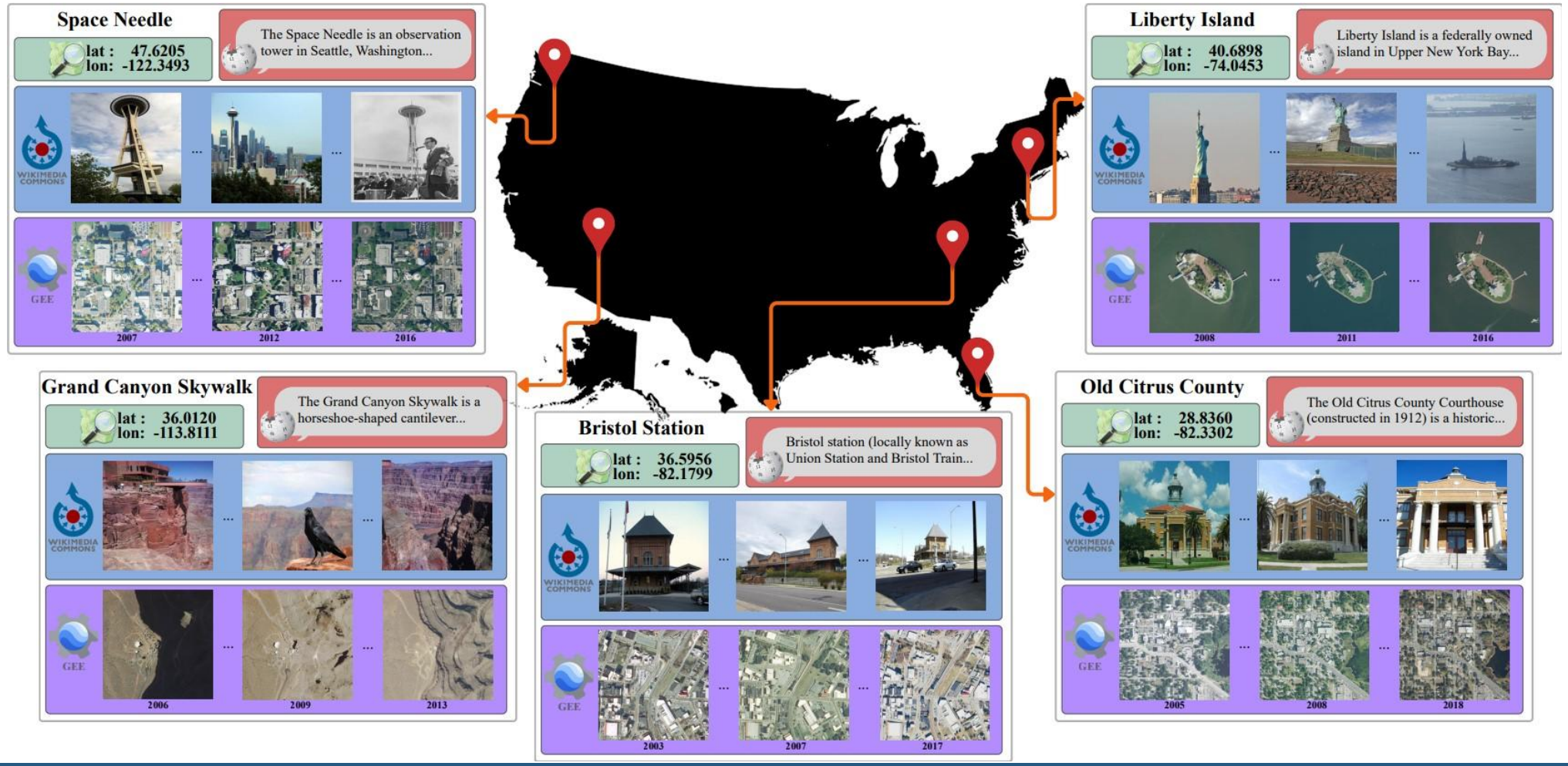


[3]

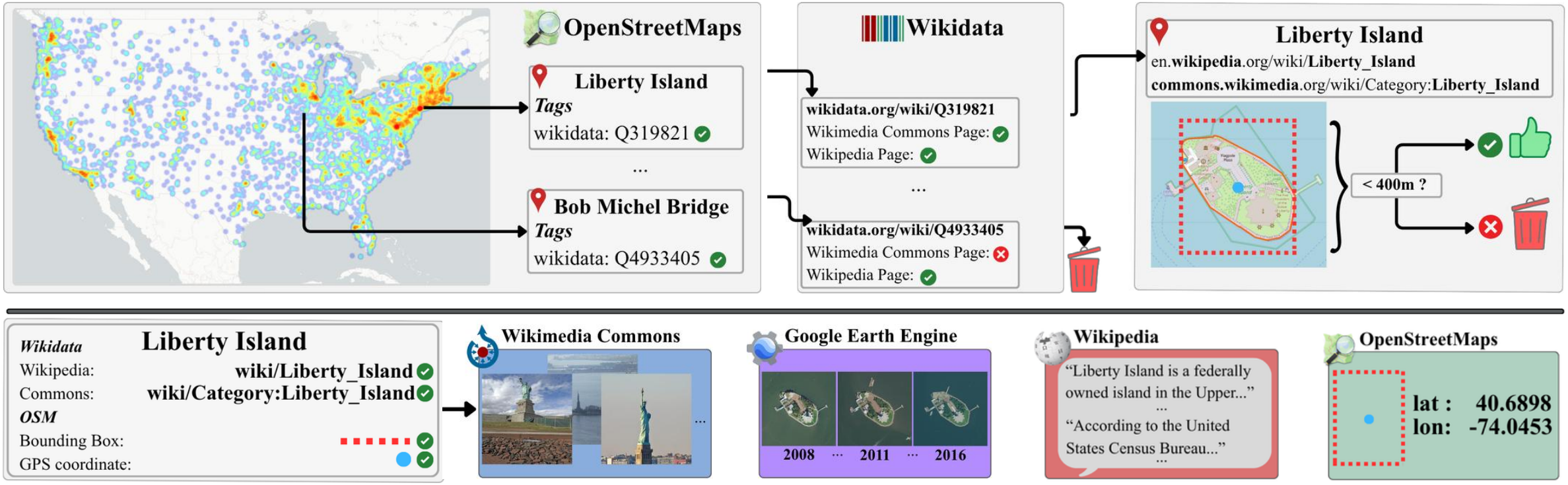


[2] Shi et al. Spatial-Aware Feature Aggregation for Cross-View Image based Geo-Localization, NeurIPS'19
[3] Ye et al. Cross-view image geo-localization with Panorama-BEV Co-Retrieval Network, ECCV'25

Combining all modalities into one dataset

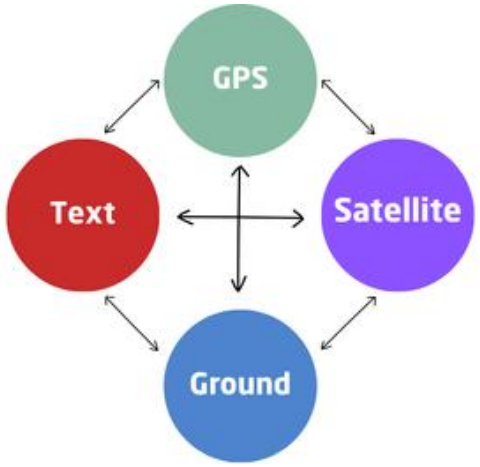


Our pipeline for data collection



Final dataset overview

- 18557 Landmarks across the United States of America



329k Ground Images

197k Satellite Images

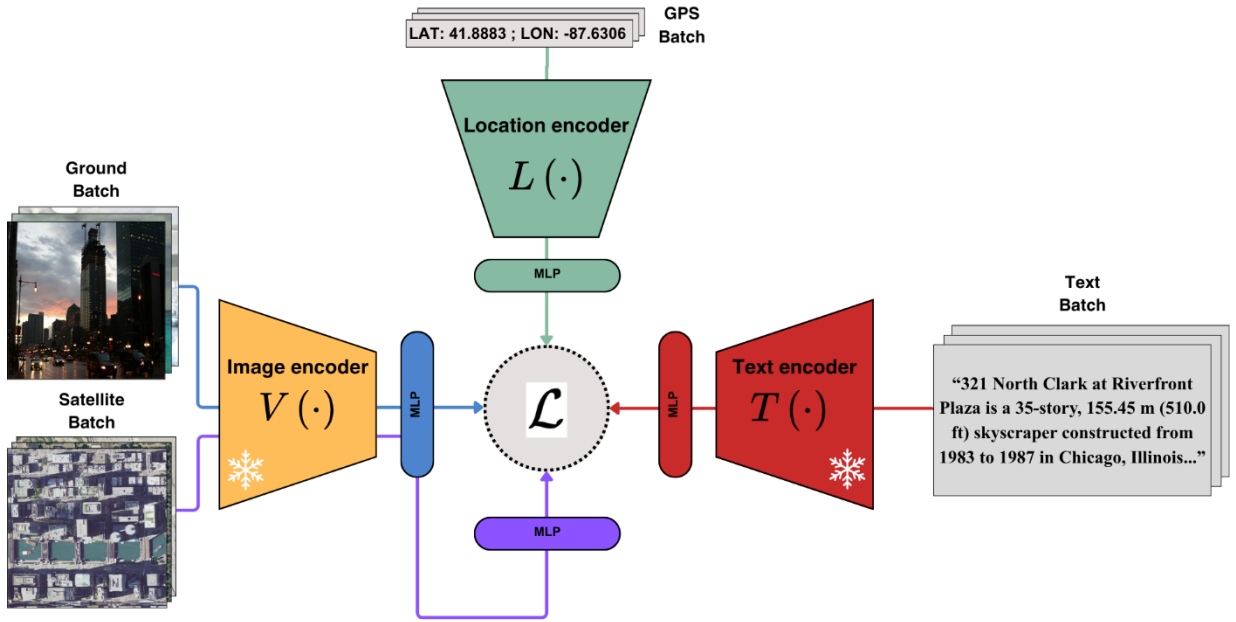
18,557 Coordinates

18,557 Text Information

A simple baseline

- Frozen CLIP Image and Text encoders.
- Initialized GeoCLIP [1] Location encoder
- Trainable MLP Projection heads
- Complete Contrastive Loss

$$\mathcal{L} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1 \; j \neq i}^{K} \mathcal{L}_{i,j}$$



[4] Cepeda et al. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization, NeurIPS'23

Some Results

	model	Satellite → Ground			Ground → Satellite		
		medR ↓	R@1 ↑	R@5 ↑	medR ↓	R@1 ↑	R@5 ↑
Zero-Shot	DiNOv2	28342	1.9	3.0	9335	3.3	8.8
	DiNOv3	14623	3.7	7.4	2241	6.1	15.5
	SigLIP	1546	7.2	15.25	679	10.7	23.8
	OAI-CLIP	519	10.4	21.2	620	11.9	26.4
	SigLIP2	682	8.6	18.6	140	14.6	31.6
	Uni-1652	62606	0.4	0.8	39961	0.2	0.9
	TransGeo-90°	40973	0.7	2.0	13425	0.4	1.7
	Sample4Geo-UNI	34988	3.0	6.2	40056	0.85	2.8
	MMCLIP	**23**	**30.4**	**52.4**	**48**	**20.5**	**46.5**

Cross-View Localization

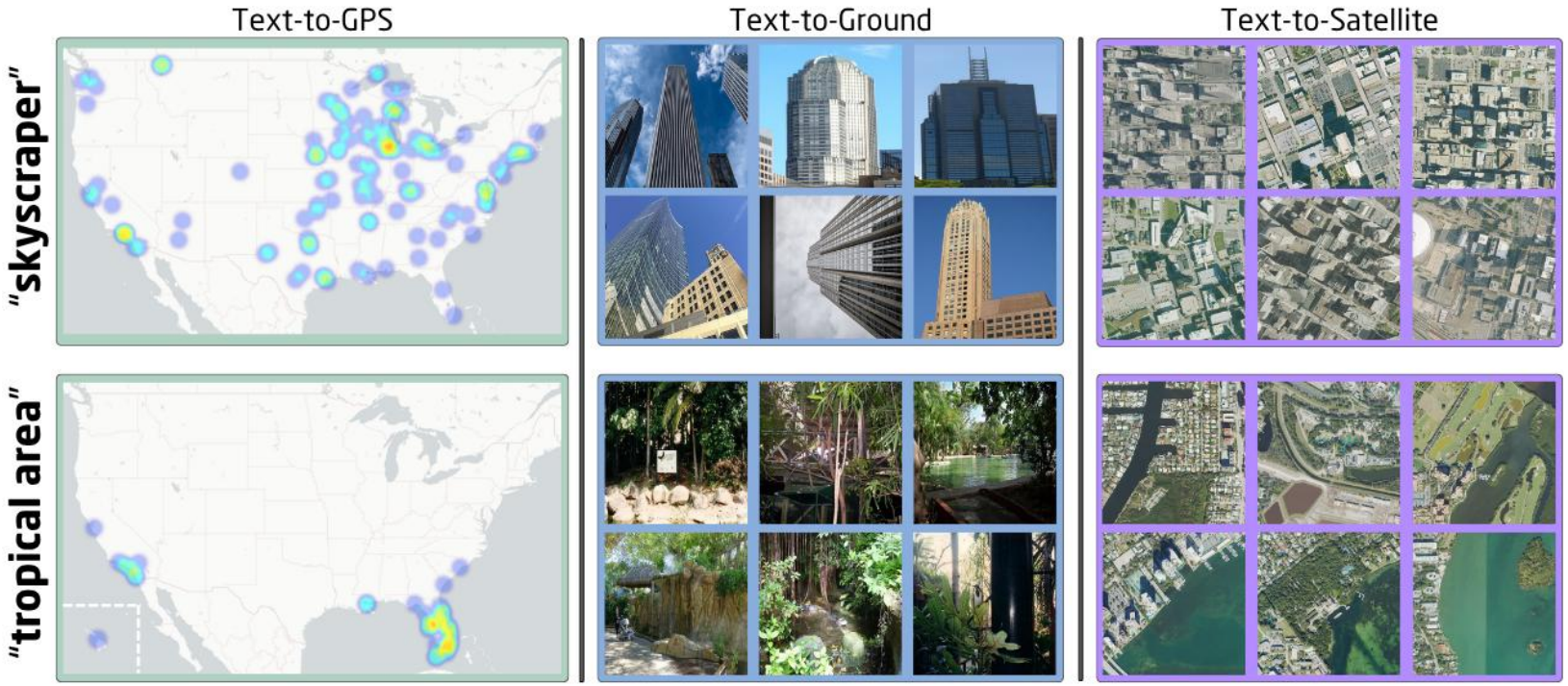
	Method	Distance (% @ km) ↑					
		Street 1 km		Region 200 km	Continent 2500 km		
		Ground → GPS			Satellite → GPS		
Zero-Shot	GeoReasoner	0.01	8.03	80.28	0.6	6.2	82.9
	StreetCLIP	0.53	48.89	88.48	0.6	49.6	93.9
	osv5m	1.28	19.19	57.25	0.7	21.2	70.8
	G3	12.86	45.52	**91.83**	8.8	47.9	97.0
	GeoCLIP	**21.37**	48.57	91.5	12.3	48.8	97.4
	MMCLIP	16.83	**51.78**	91.52	**36.9**	**81.1**	**99.7**

Geo-localization

- Comparison with off the shelf foundation and specialized models
- Goal: Illustrate flexibility, not superiority

- MMCLIP, is versatile when trained on Mmlandmarks.

Text-to-X modality alignement



Final Thoughts - MMLandmarks

- First large-scale, instance-level, multimodal dataset for geospatial understanding
- 18557 landmarks across the United States
- Landmark-level one to one correspondence: Ground images, Aerial images, Text and GPS
- Open Domain – permissive licenses

Dataset →

Paper →

Codebase →

Project page - scan me!

mmlandmarks.compute.dtu.dk


DTU