ICONIP 2024

31st International Conference on Neural Information Processing

December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Improving Image Geolocation with Multimodal Deep Learning

Bartosz Siński, **Adam Żychowski**, Jacek Mańdziuk

Faculty of Mathematics and Information Science,

Warsaw University of Technology, Warsaw, Poland

ICONIP 2024

31st International Conference on Neural Information Processing

December 2–6, 2024 · Auckland, New Zealand iconip2024.org

ICONIP 2024

31st International Conference on Neural Information Processing

December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Sponsors: g.tec · NEW ZEALAND · 100% PURE NEW ZEALAND · AUT KNOWLEDGE ENGINEERING & DISCOVERY RESEARCH INNOVATION · Springer

Supported by: AUT UNIVERSITY · UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND · NTU Nottingham Trent University · Asia Pacific Neural Network Society

Problem description

- **Image Geolocalization** - computer vision problem of **finding the location of the place that is represented by the image**
- various level of difficulty
- popular game GEOGUESSR
- real-world applications (e.g. detecting military objects location)

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Sponsors: g.tec NEW ZEALAND 100% PURE NEW ZEALAND AUT KNOWLEDGE ENGINEERING & DISCOVERY RESEARCH INNOVATION Springer

Supported by: AUT UNIVERSITY UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND NTU Nottingham Trent University AsiaPacificNeuralNetworkSociety

Image data

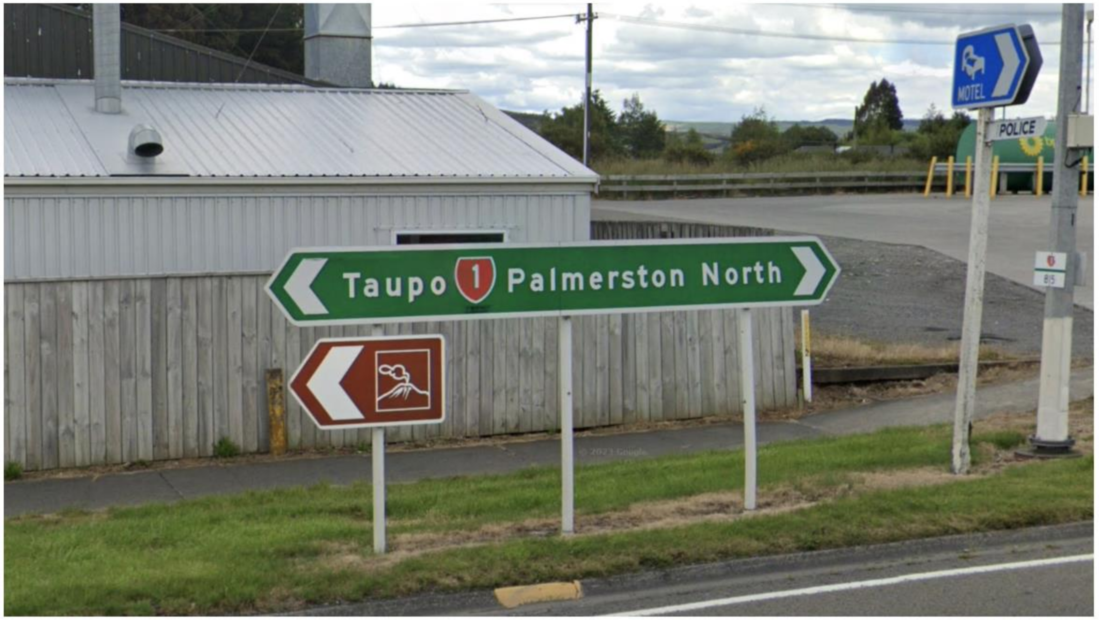
- 322,536 images from 90 countries
- panoramas downloaded from Google Street view

Luo, Grace, et al. "*G^3: Geolocation via Guidebook Grounding*" arXiv preprint arXiv:2211.15521 (2022).

Sponsors: g.tec NEW ZEALAND 100% PURE NEW ZEALAND AUT KNOWLEDGE ENGINEERING & DISCOVERY RESEARCH INNOVATION Springer
Supported by: AUT UNIVERSITY UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND NTU Nottingham Trent University AsiaPacificNeuralNetworkSociety

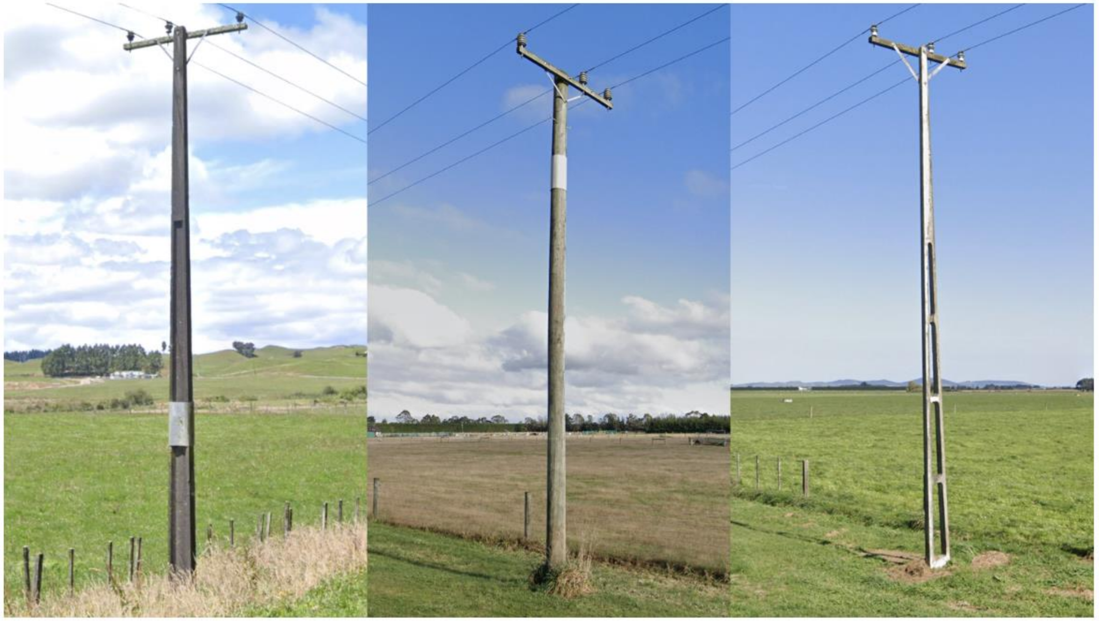
Text data

6041 clues



New Zealand uses **green directional signs.** If the sign is on a state highway, the highway number will always be shown in a **red crest**.

NOTE: Brown signs indicate the direction to landmarks, which can be useful when pinpointing.



The most common pole type found in New Zealand is made of concrete and has **one long indent** which runs most of the way up the pole. Most concrete poles have small **silver** possum guards. Circular wooden poles can also be found, but are less common. You can also see concrete holey poles in New Zealand.

https://www.plonkit.net/new-zealand

Proposed solution

- Image Encoder - StreetCLIP based model
- Text encoder - RoBERTa model
- self-attention mechanism

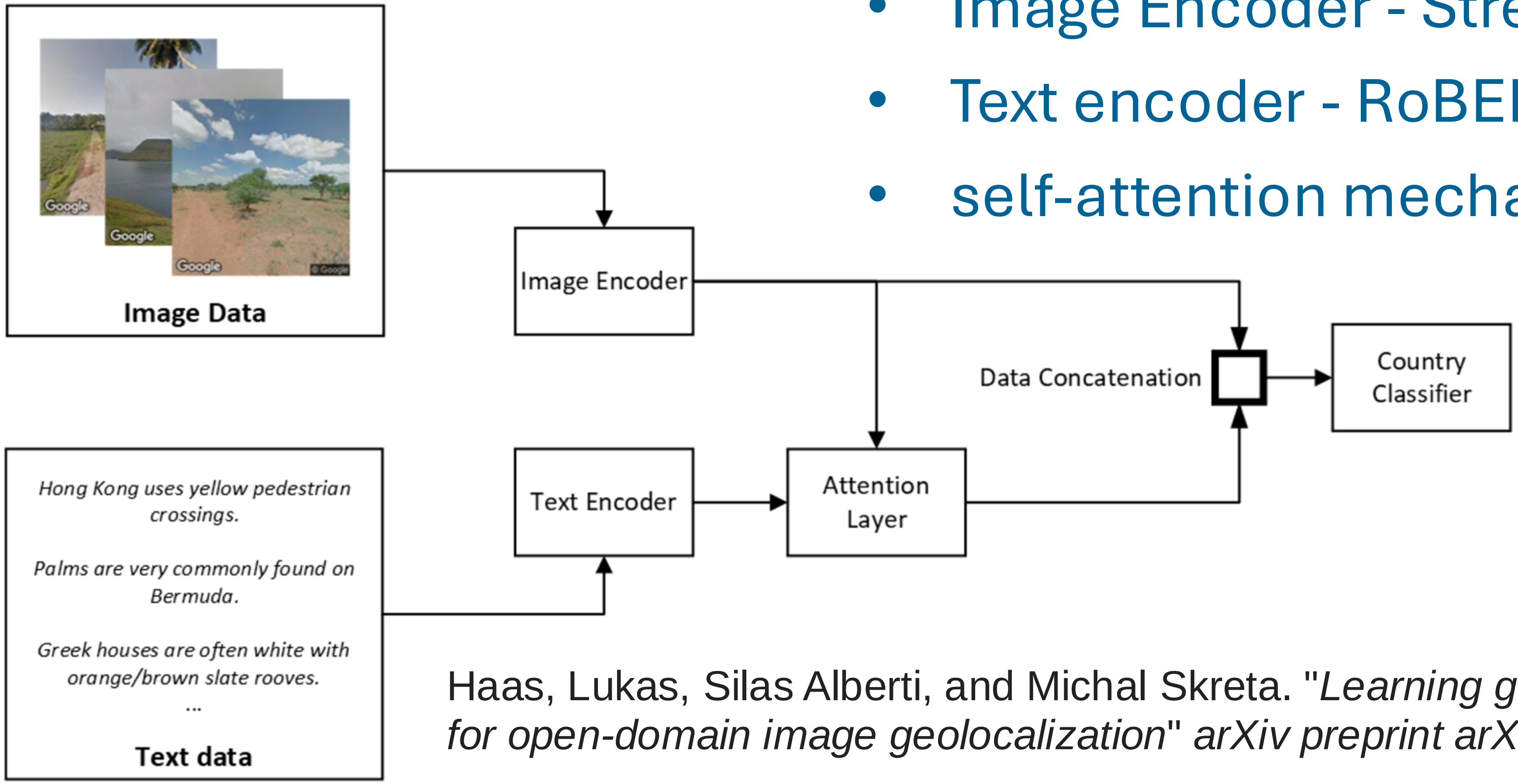


Image Data

Image Encoder

Data Concatenation

Country Classifier

Attention Layer

Text Encoder

Hong Kong uses yellow pedestrian crossings.

Palms are very commonly found on Bermuda.

Greek houses are often white with orange/brown slate rooves.

...

Text data

Haas, Lukas, Silas Alberti, and Michal Skreta. "*Learning generalized zero-shot learners for open-domain image geolocalization*" *arXiv preprint arXiv:2302.00275* (2023).

Liu, Yinhan. "*RoBERTa: A robustly optimized BERT pretraining approach*" *arXiv preprint arXiv:1907.11692* 364 (2019).

Experimental setup

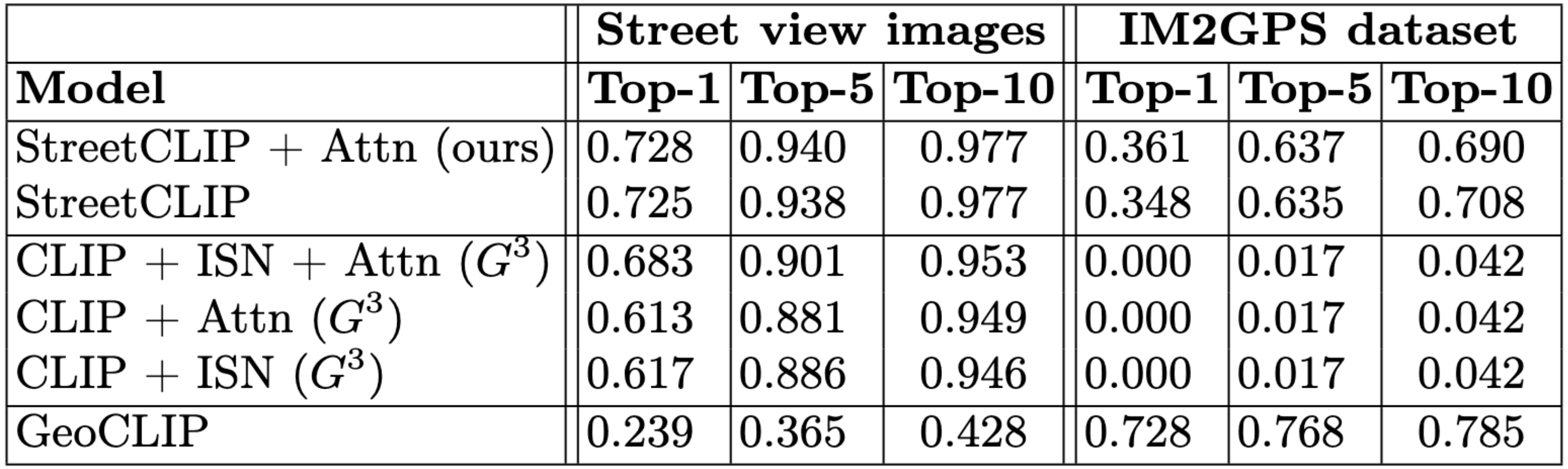
Street view images

IM2GPS dataset

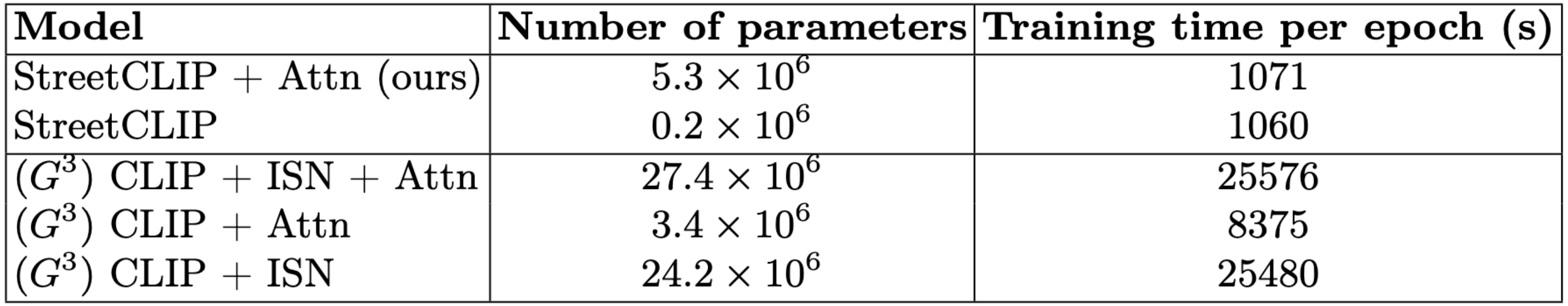
Hays, James, and Alexei A. Efros. "*IM2GPS: estimating geographic information from a single image*" *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 2008.

Sponsors: g·tec 100% PURE NEW ZEALAND AUT KNOWLEDGE ENGINEERING & DISCOVERY RESEARCH INNOVATION Springer
Supported by: AUT UNIVERSITY UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND NTU Nottingham Trent University Asia Pacific Neural Network Society

Results



	Street view images			IM2GPS dataset		
Model	**Top-1**	**Top-5**	**Top-10**	**Top-1**	**Top-5**	**Top-10**
StreetCLIP + Attn (ours)	0.728	0.940	0.977	0.361	0.637	0.690
StreetCLIP	0.725	0.938	0.977	0.348	0.635	0.708
CLIP + ISN + Attn (G^3)	0.683	0.901	0.953	0.000	0.017	0.042
CLIP + Attn (G^3)	0.613	0.881	0.949	0.000	0.017	0.042
CLIP + ISN (G^3)	0.617	0.886	0.946	0.000	0.017	0.042
GeoCLIP	0.239	0.365	0.428	0.728	0.768	0.785



Model	**Number of parameters**	**Training time per epoch (s)**
StreetCLIP + Attn (ours)	5.3×10^6	1071
StreetCLIP	0.2×10^6	1060
(G^3) CLIP + ISN + Attn	27.4×10^6	25576
(G^3) CLIP + Attn	3.4×10^6	8375
(G^3) CLIP + ISN	24.2×10^6	25480

Results

Mislabelled countries

1. Lithuania & Latvia: 14
2. Palestinian Territory & Israel: 11
3. Palestinian Territory & Jordan: 9
4. Puerto Rico & Dominican Republic: 9
5. Lithuania & Estonia: 9
6. United States & Canada: 8
7. Ukraine & Lithuania: 8
8. Guatemala & Ecuador: 7
9. Sweden & Norway: 7
10. Serbia & Croatia: 6

Continents

Continent	**Top-1**	**Top-5**	**Top-10**
Europe	0.429	0.714	0.805
Asia	0.570	0.835	0.923
Africa	0.603	0.910	0.993
South & North America	0.463	0.698	0.927

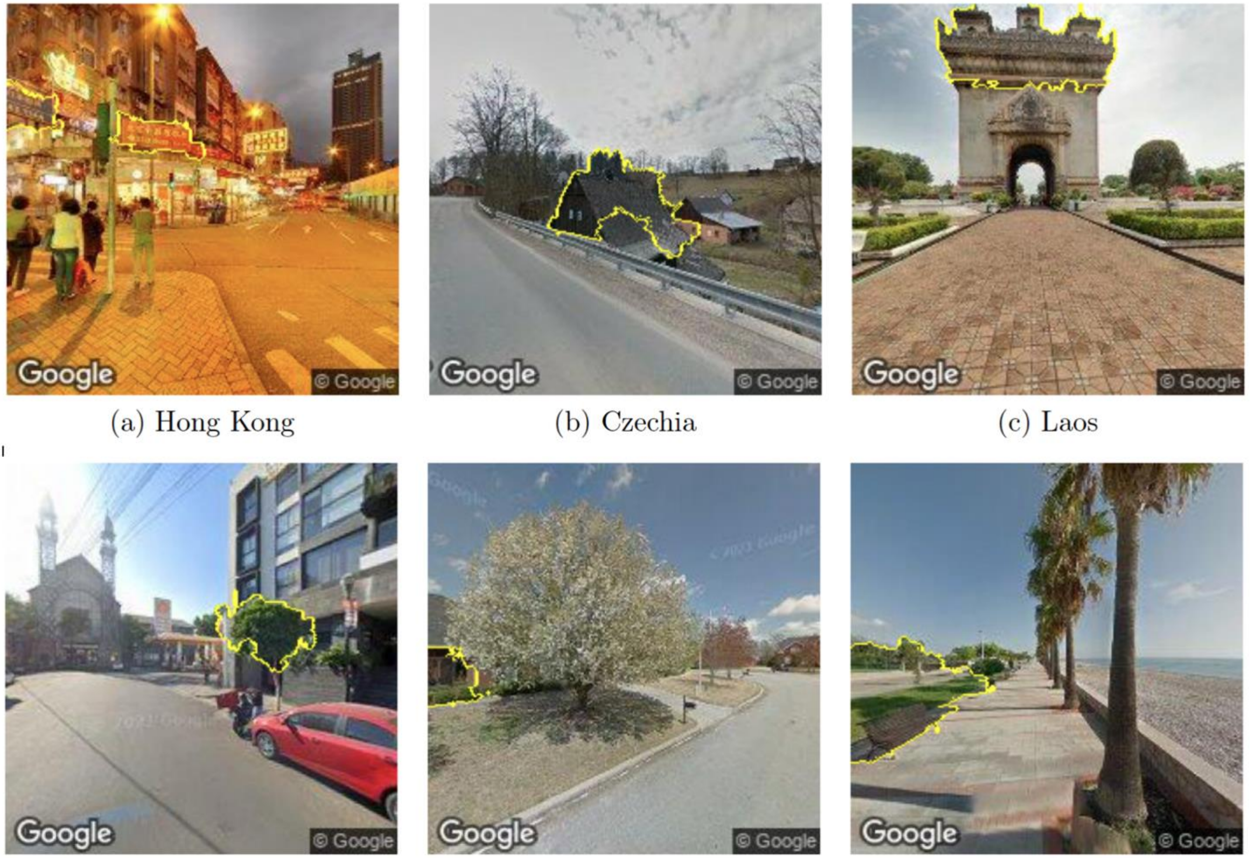
Highest recall

Greenland:	1.000
Faroe Islands:	1.000
Ireland:	1.000
Guam:	0.976
Iceland:	0.974.

Lowest recall

Slovakia:	0.200
Lithuania:	0.357
Austria:	0.414
Spain:	0.418
Belgium:	0.455

Explanations


(a) Hong Kong (b) Czechia (c) Laos

(d) Mexico (e) United States (f) Spain

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "*Why should i trust you? Explaining the predictions of any classifier*" *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

Summary

- Proposed geolocation model successfully classifies images to the countries that they represent in multimodal manner.
- Images come from street view panoramas.
- Text data gathered from Geoguessr community tutorial websites and forums.
- Proposed model surpassed the state-of-the-art G^3 model in accuracy on both the street view images test set and the IM2GPS benchmark dataset.
- Training time and number of trainable parameter significantly reduced compared to the G^3 model.

Thank you