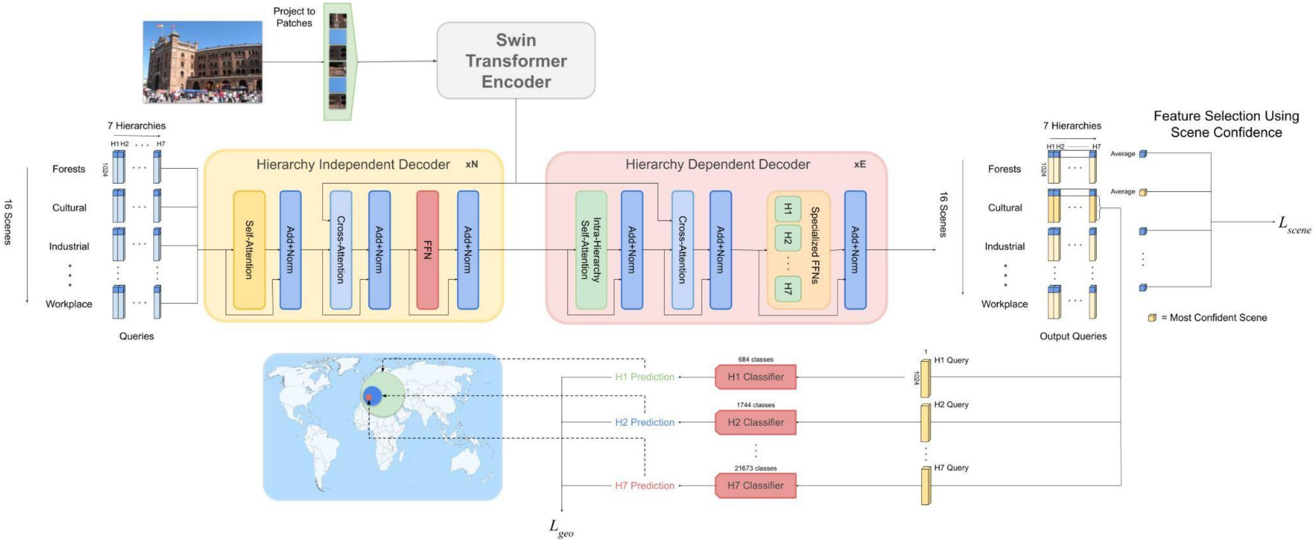
Where We Are and What We’re Looking At: Query Based Worldwide Image Geo-localization Using Hierarchies and Scenes

Brandon Clark, Alec Kerrigan, Parth Kulkarni, Vicente Vivanco Cepeda, Dr. Mubarak Shah

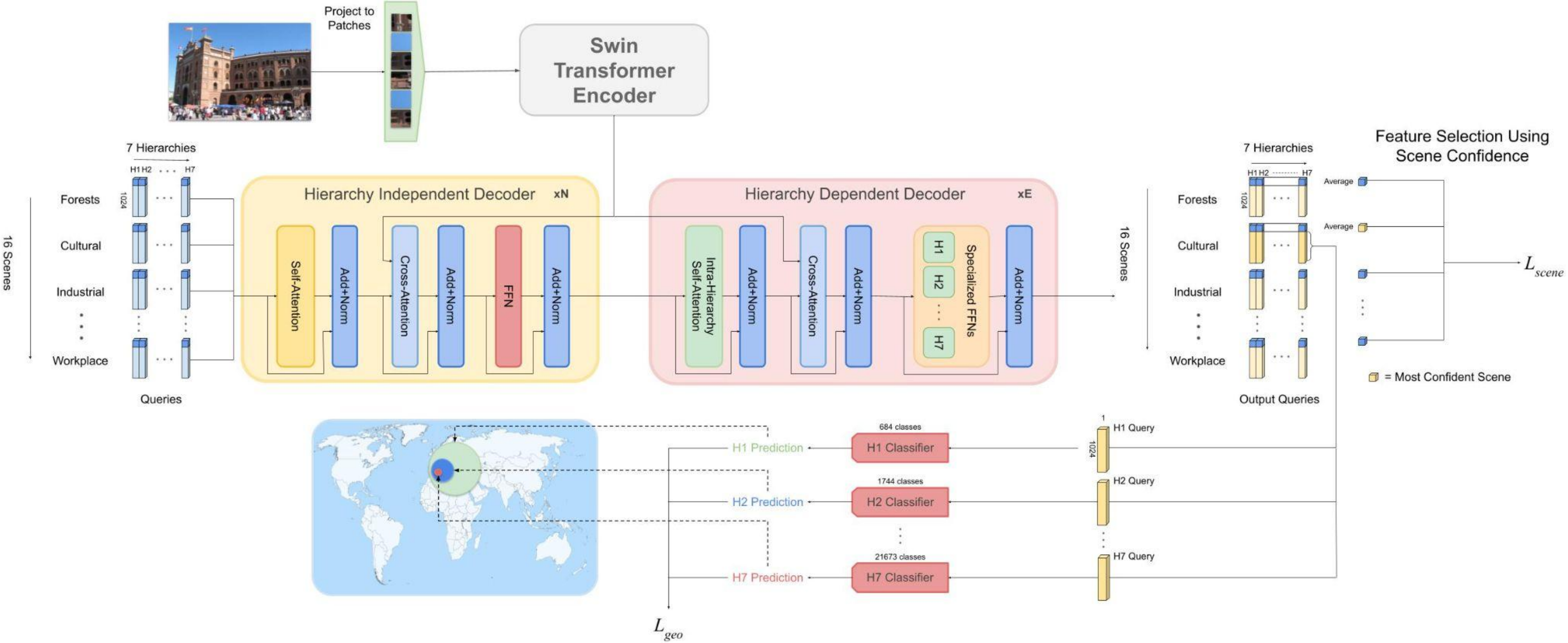
Paper Tag: THU-PM-246

Preview

- First Transformer Decoder based architecture
- Extract different features for each hierarchy and scene by using Geographic Queries
- Propose a new testing dataset with real world images and reduce geographic biases
- Improve location prediction on existing datasets by as much as 40%

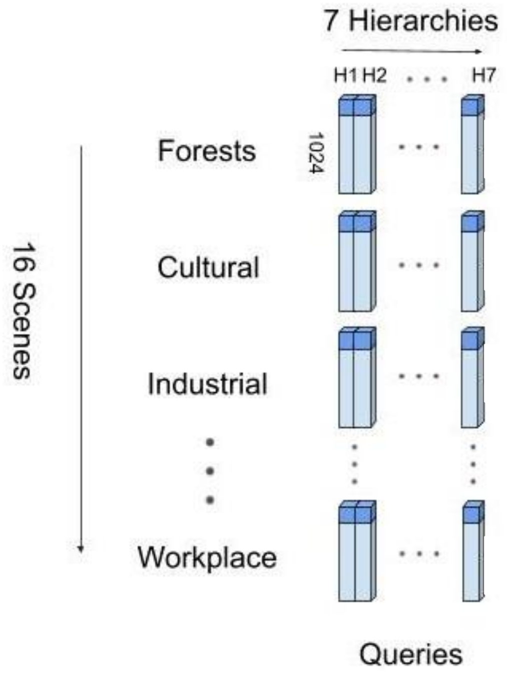


Model


Project to Patches
Swin Transformer Encoder
7 Hierarchies
H1 H2 . . . H7
16 Scenes
Forests
Cultural
Industrial
Workplace
1024
Queries
Hierarchy Independent Decoder
xN
Self-Attention
Add+Norm
Cross-Attention
Add+Norm
FFN
Add+Norm
Hierarchy Dependent Decoder
xE
Intra-Hierarchy Self-Attention
Add+Norm
Cross-Attention
Add+Norm
Specialized FFNs
H1
H2
H7
Add+Norm
16 Scenes
7 Hierarchies
H1 H2 H7
Forests
Cultural
Industrial
Workplace
1024
Output Queries
Feature Selection Using Scene Confidence
Average
Average
L_{scene}
= Most Confident Scene
1
1024
H1 Query
H2 Query
H7 Query
684 classes
H1 Classifier
1744 classes
H2 Classifier
21673 classes
H7 Classifier
H1 Prediction
H2 Prediction
H7 Prediction
L_{geo}

Decoder Queries (Hierarchy Queries)

- Each query is tasked to extract specific features
  - 7 Hierarchies * 16 Scenes = 112 Queries
- Dimension 1024
- Randomly initialized
- 0th channel is trained to be scene confidence



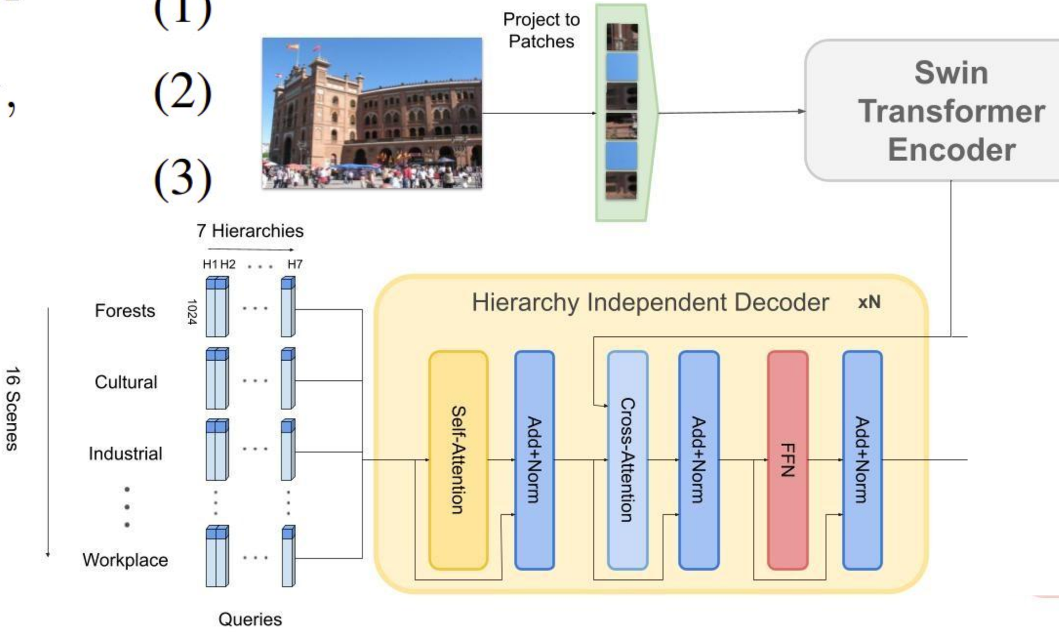
Hierarchy Independent Decoder

- Queries extract image features via Cross-Attention

$$y^{SA} = MSA(LN(GQ^{k-1})) + GQ^{k-1} \qquad (1)$$

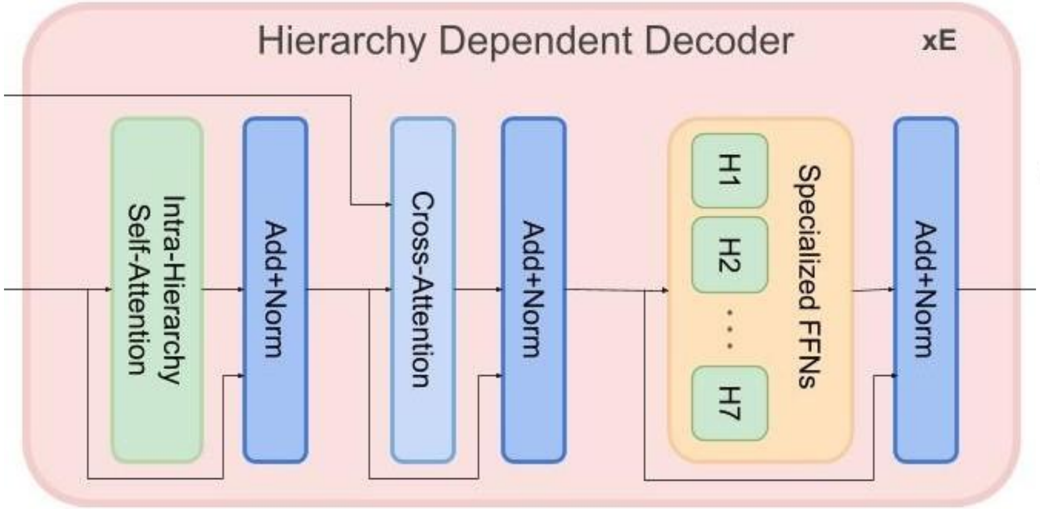
$$y^{CA} = CA(LN(y^{SA}, LN(X)) + y^{SA}, \qquad (2)$$

$$GQ^{k} = FFN(LN(y^{CA})) + y^{CA} \qquad (3)$$



Hierarchy Dependent Decoder

- Allows queries to specify which hierarchy they represent
- Self-Attention and FFNs are specific to each hierarchy



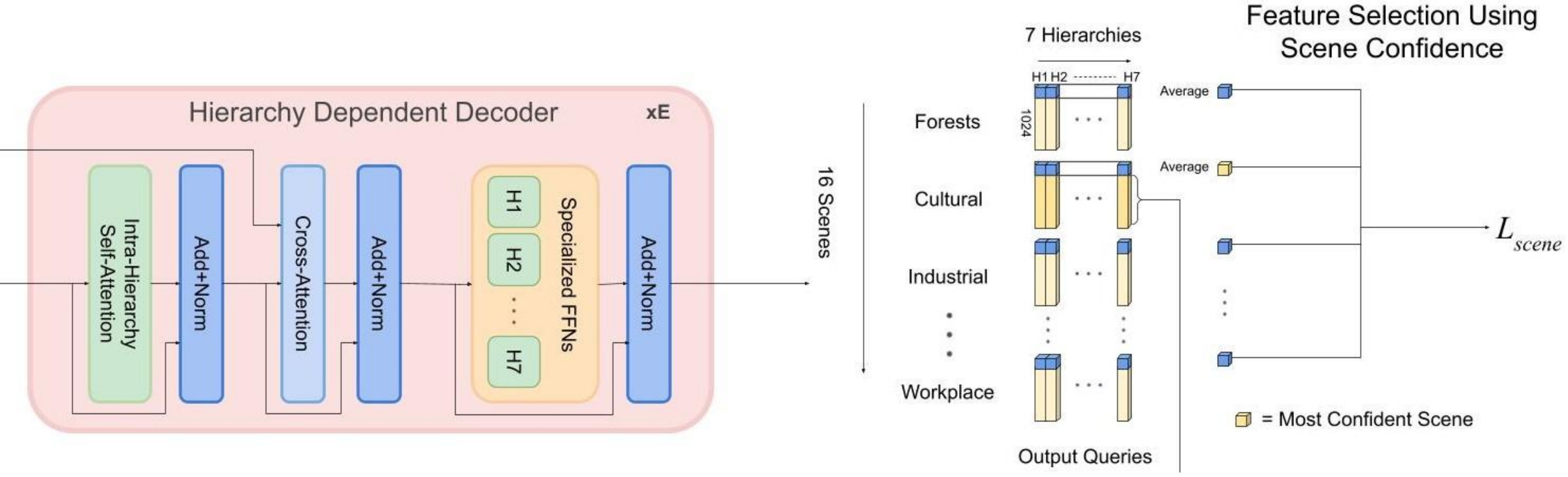
$$y^{SA} = MSA(LN(GQ_h^{k-1})) + GQ_h^{k-1}, \tag{4}$$

$$y^{CA} = CA(LN(y^{SA}), LN(X)) + y^{SA}, \tag{5}$$

$$GQ_h^k = FFN_h(LN(y^{CA})) + y^{CA} \tag{6}$$

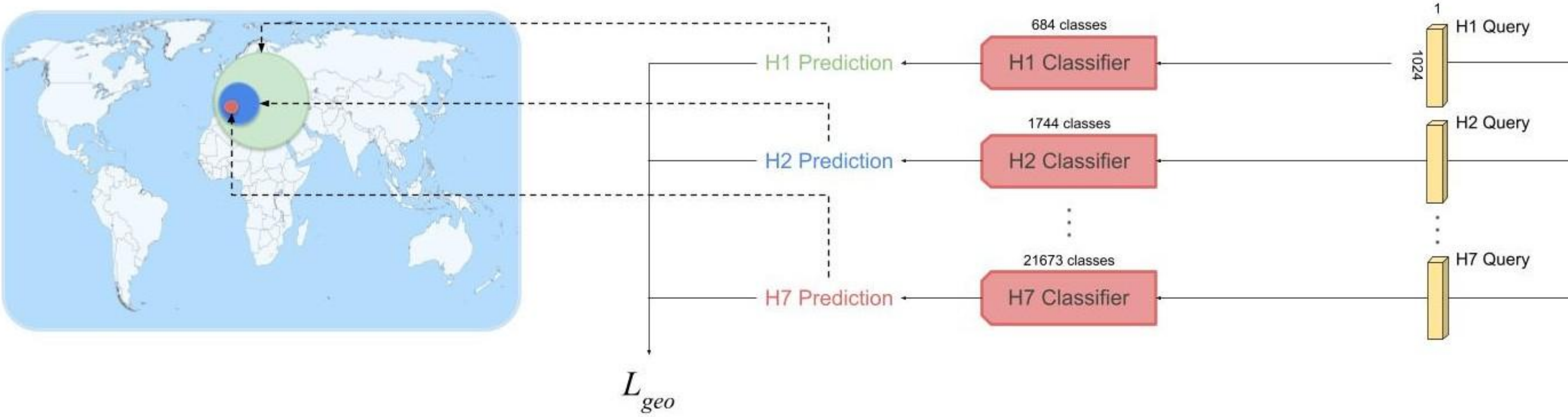
Scene Selection

- Average 0th Channel for each scene
- Highest value is the selected scene



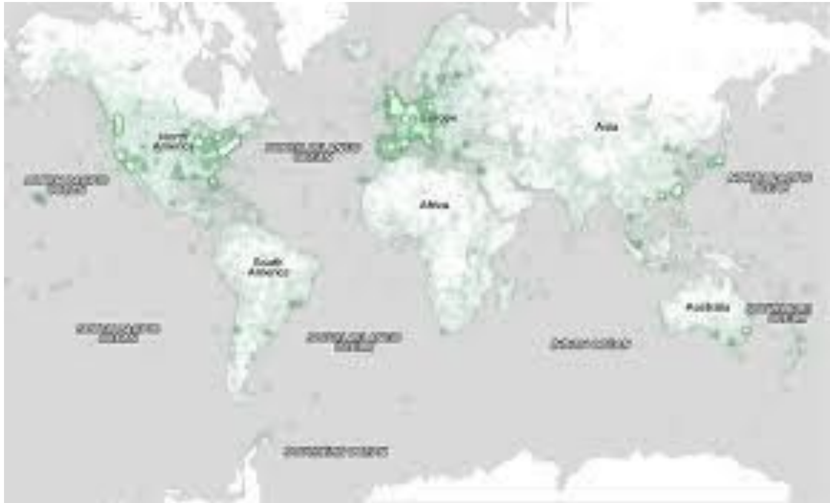
Classification

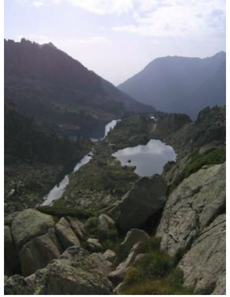
- Selected queries go to their specified classification layers
- Predictions from each hierarchy are used to make a final prediction



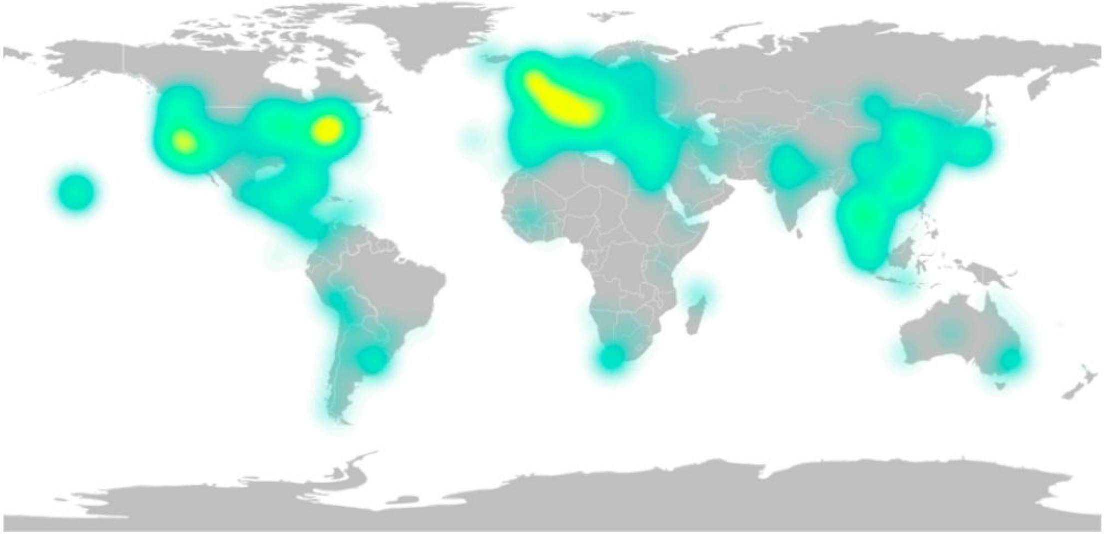
Training Dataset

- MediaEval Places 2016 (MP16)
  - 4.7M Images with GPS from Yahoo and Flickr
  - Subset of YFCC100M
  - Uncurated dataset





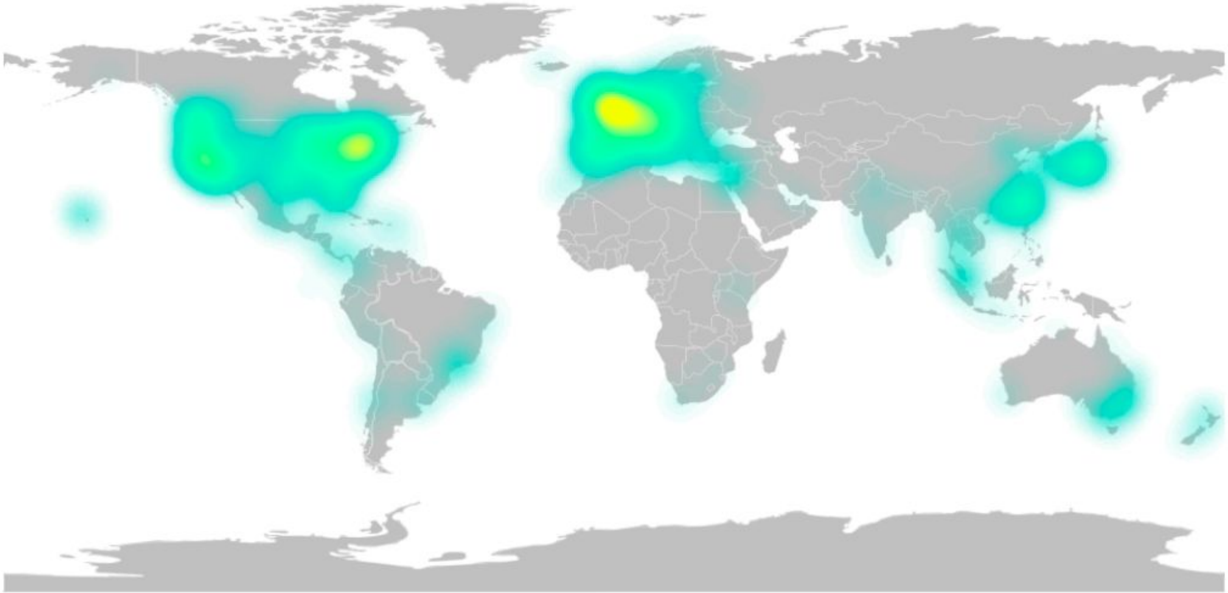
Testing Datasets



- Im2GPS
  - ~300 Images
- Im2GPS3k
  - ~3k Images
- Curated sets of landmarks

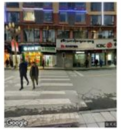
Testing Datasets

- YFCC4k
  - ~4k Images
- YFCC26k
  - ~26k Images
- Uncurated
- Subset of YFCC100M



Google World Streets 15k (GWS15k)

1. Pick a Country with probability based on surface area
2. Pick a town or city in that country
3. Pick a random coordinate within 5Km of the town/city


Sichuan, China

Rawa, Iraq

La Esperanza, Honduras

Aalborg, Denmark

Shimane, Japan

Freetown, Sierra Leone

Mwanza, Tanzania

Sanggeng, Indonesia

Queensland, Australia

Lucélia, Brazil

Bordj Bou Arreridj, Algeria

Moyobamba, Peru

Bethal, South Africa

Aunglan, Myanmar

Borama, Somalia

Lara, Venezuela

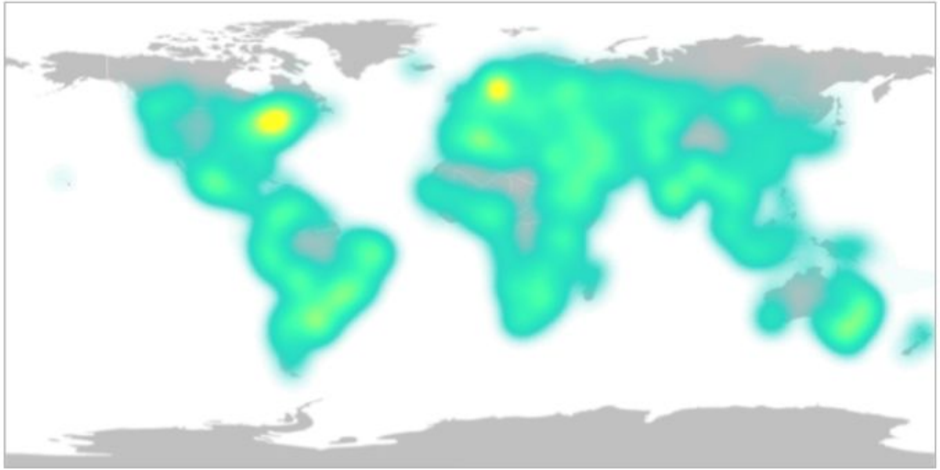
Atbara, Sudan

Erdenet, Mongolia

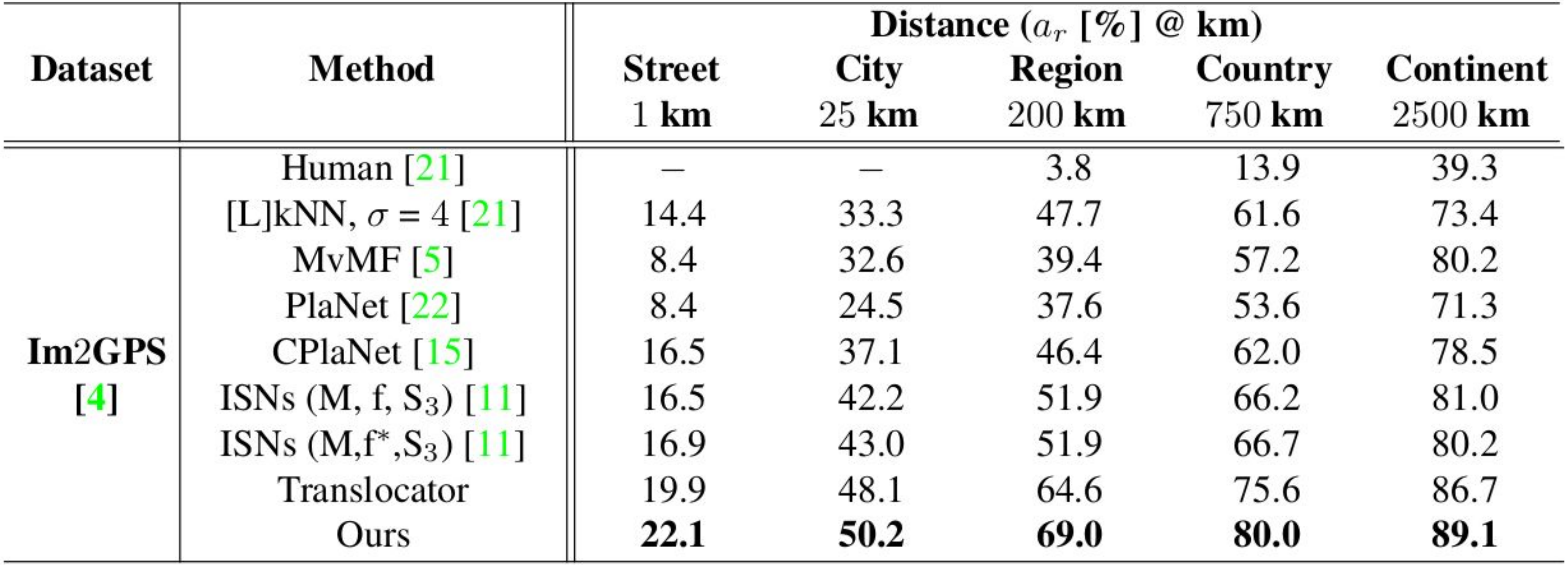
Kursk Oblast, Russia

Cesar, Colombia

Magalia, United States

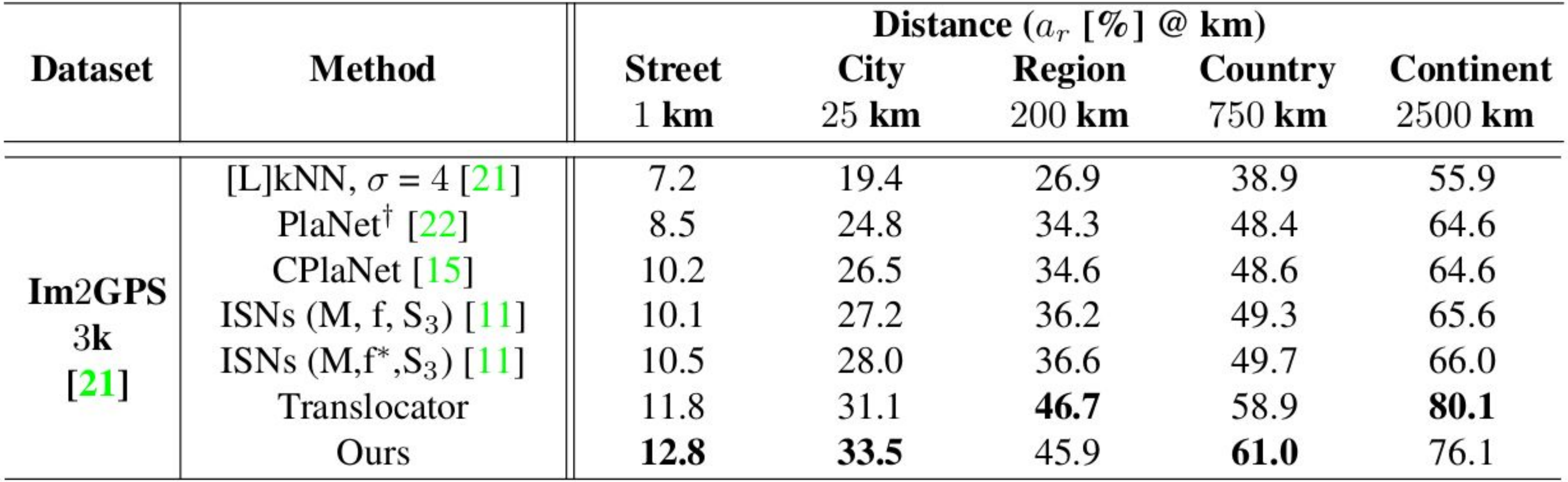


Results on Im2GPS



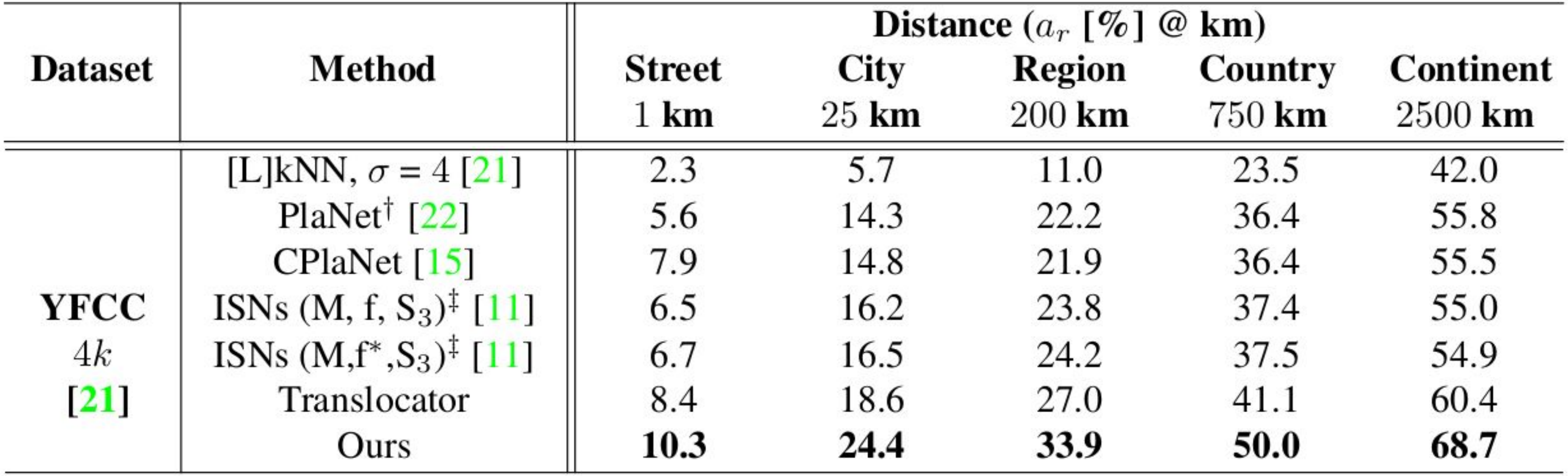
Dataset	Method	Distance (a_r [%] @ km)				
		Street 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 2500 **km**
Im2GPS [4]	Human [21]	–	–	3.8	13.9	39.3
	[L]kNN, σ = 4 [21]	14.4	33.3	47.7	61.6	73.4
	MvMF [5]	8.4	32.6	39.4	57.2	80.2
	PlaNet [22]	8.4	24.5	37.6	53.6	71.3
	CPlaNet [15]	16.5	37.1	46.4	62.0	78.5
	ISNs (M, f, S_3) [11]	16.5	42.2	51.9	66.2	81.0
	ISNs (M,f*,S_3) [11]	16.9	43.0	51.9	66.7	80.2
	Translocator	19.9	48.1	64.6	75.6	86.7
	Ours	**22.1**	**50.2**	**69.0**	**80.0**	**89.1**

Results on Im2GPS3k



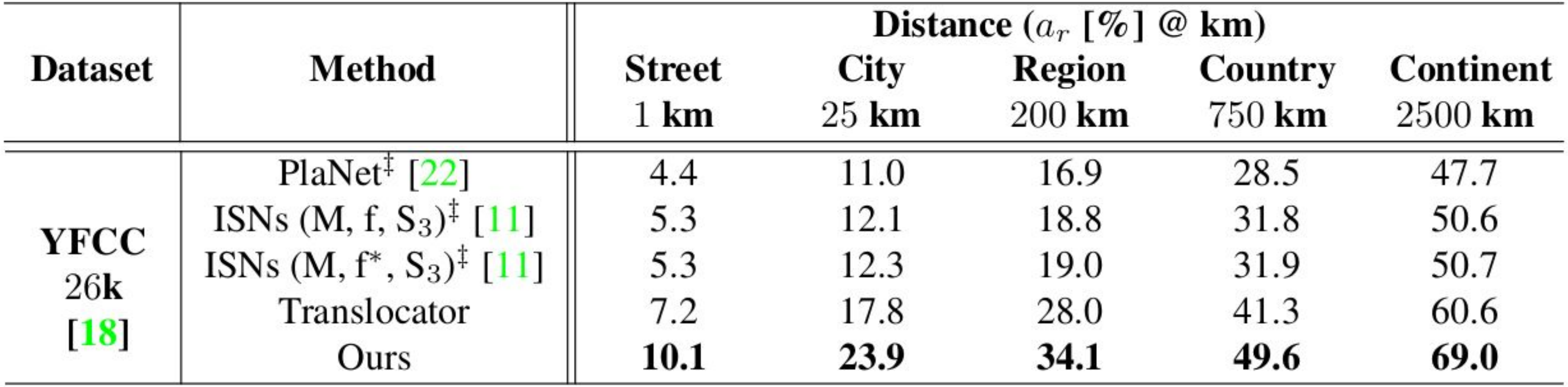
Dataset	**Method**	**Distance** (a_r **[%] @ km)**				
		Street 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 2500 **km**
Im2GPS 3k [21]	[L]kNN, $\sigma = 4$ [21]	7.2	19.4	26.9	38.9	55.9
	PlaNet† [22]	8.5	24.8	34.3	48.4	64.6
	CPlaNet [15]	10.2	26.5	34.6	48.6	64.6
	ISNs (M, f, S_3) [11]	10.1	27.2	36.2	49.3	65.6
	ISNs (M,f*,S_3) [11]	10.5	28.0	36.6	49.7	66.0
	Translocator	11.8	31.1	**46.7**	58.9	**80.1**
	Ours	**12.8**	**33.5**	45.9	**61.0**	76.1

Results on YFCC4k



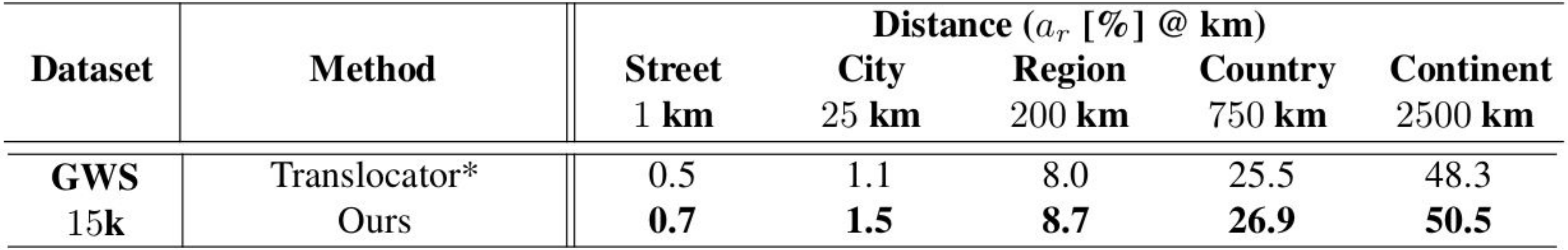
Dataset	**Method**	**Distance** (a_r **[%] @ km**)				
		Street	**City**	**Region**	**Country**	**Continent**
		1 **km**	25 **km**	200 **km**	750 **km**	2500 **km**
YFCC $4k$ [21]	[L]kNN, σ = 4 [21]	2.3	5.7	11.0	23.5	42.0
	PlaNet† [22]	5.6	14.3	22.2	36.4	55.8
	CPlaNet [15]	7.9	14.8	21.9	36.4	55.5
	ISNs (M, f, S_3)‡ [11]	6.5	16.2	23.8	37.4	55.0
	ISNs (M,f*,S_3)‡ [11]	6.7	16.5	24.2	37.5	54.9
	Translocator	8.4	18.6	27.0	41.1	60.4
	Ours	**10.3**	**24.4**	**33.9**	**50.0**	**68.7**

Results on YFCC26k



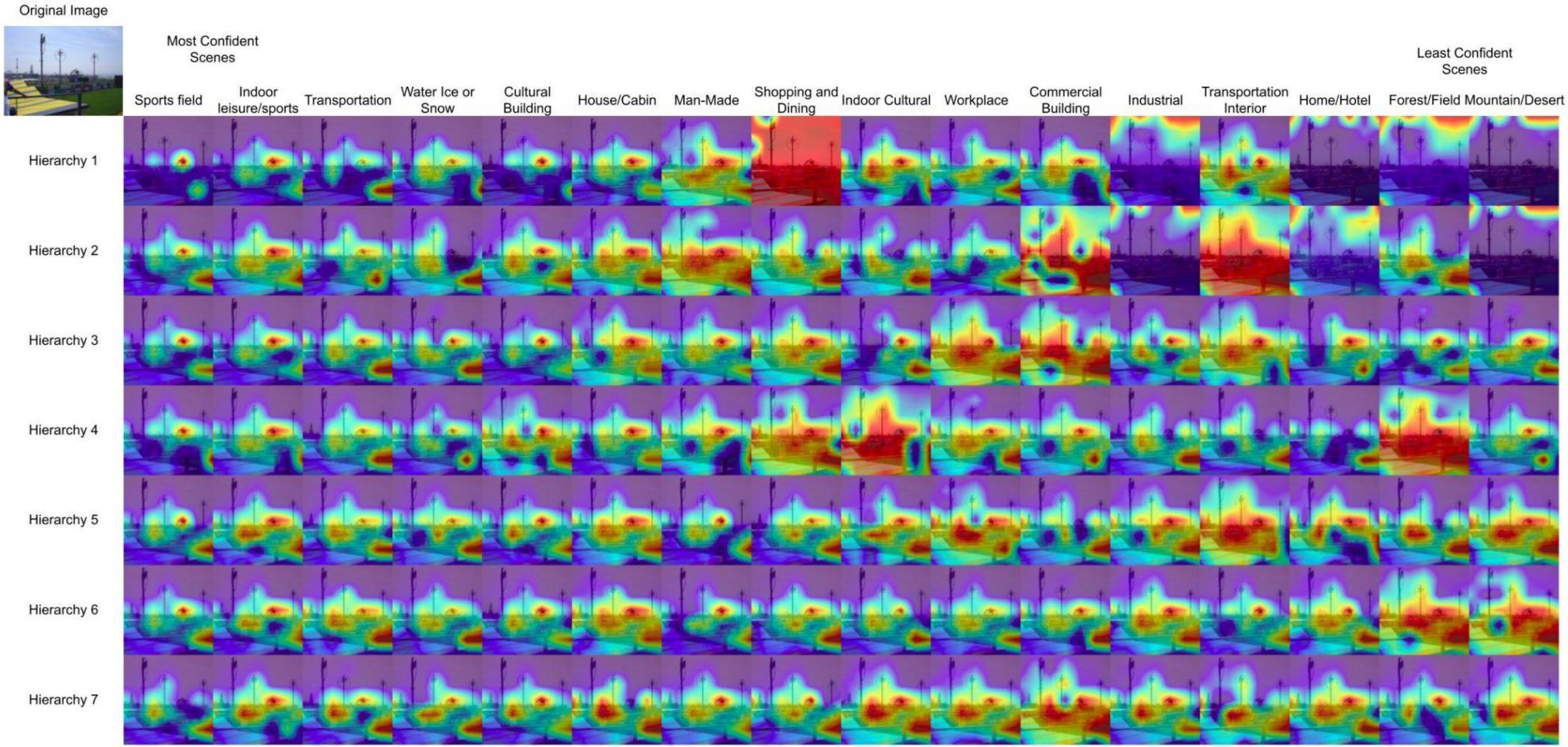
Dataset	Method	Distance (a_r [%] @ km)				
		Street 1 **km**	City 25 **km**	Region 200 **km**	Country 750 **km**	Continent 2500 **km**
YFCC 26**k** [18]	PlaNet‡ [22]	4.4	11.0	16.9	28.5	47.7
	ISNs (M, f, S_3)‡ [11]	5.3	12.1	18.8	31.8	50.6
	ISNs (M, f*, S_3)‡ [11]	5.3	12.3	19.0	31.9	50.7
	Translocator	7.2	17.8	28.0	41.3	60.6
	Ours	**10.1**	**23.9**	**34.1**	**49.6**	**69.0**

Results on GWS15k



Dataset	**Method**	**Distance** (a_r **[%] @ km)**				
		Street	**City**	**Region**	**Country**	**Continent**
		1 **km**	25 **km**	200 **km**	750 **km**	2500 **km**
GWS 15**k**	Translocator*	0.5	1.1	8.0	25.5	48.3
	Ours	**0.7**	**1.5**	**8.7**	**26.9**	**50.5**

Qualitative Results Im2GPS3k



Conclusion

- Extracting features unique to each geographic hierarchy is important for geo-localization
- Our model shows significant quantitative improvements and our qualitative results show the model works as intended
- Our new testing dataset solves the limitations of other test sets, while also showing the limitations of the problem
- Code is available at https://github.com/AHKerrigan/GeoGuessNet

References

- Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. arXiv preprint arXiv:2204.13861, 2022. 1, 2, 3, 7, 8
- Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In Proceedings of the IEEE international conference on computer vision, pages 2621–2630, 2017. 1, 2, 3, 6, 7, 8
- Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 563–579, 2018. 1, 2, 3, 5, 6, 7