Where We Are and What We’re Looking At

Brandon Clark, Alec Kerrigan, Parth Kulkarni, Vicente Cepeda, Mubarak Shah

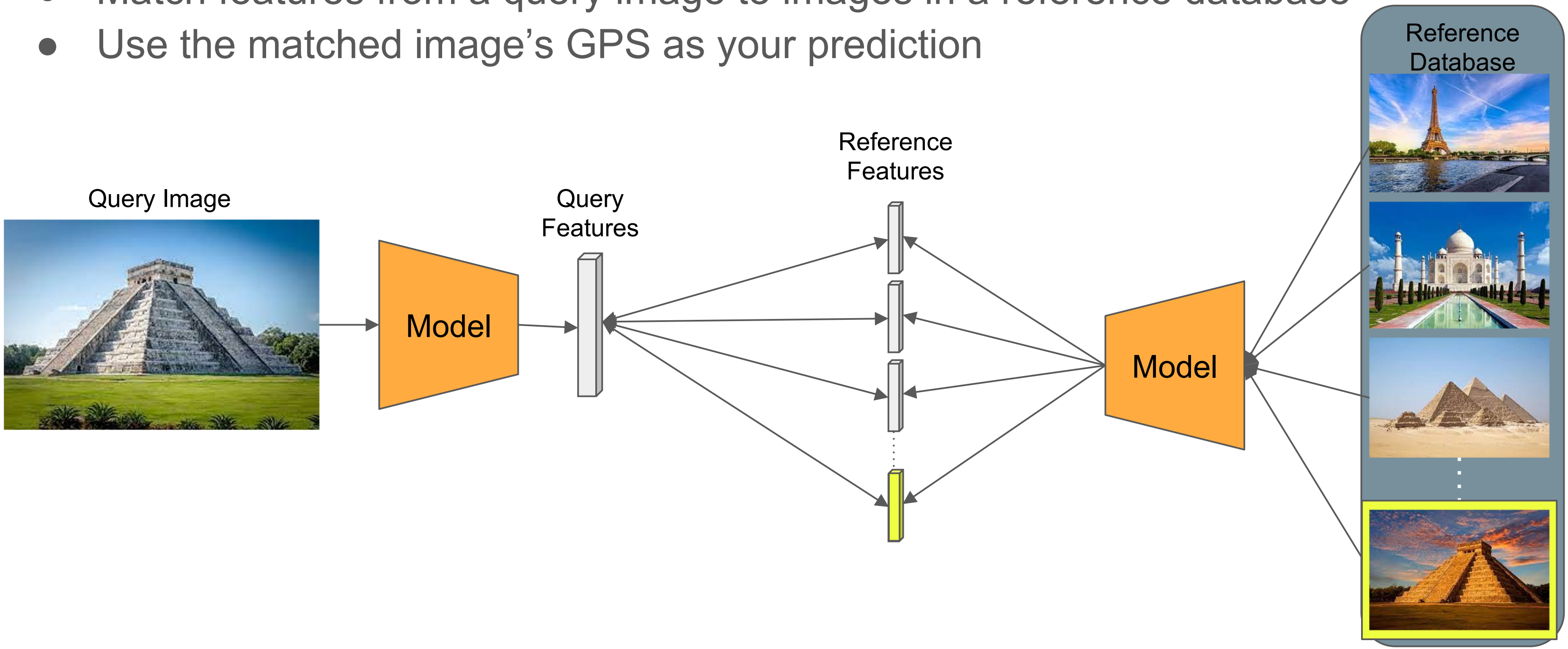
Image Geo-localization

- Geo-localization deals with predicting the GPS Coordinates of a query Image
- This task has been explored with two main techniques
    - Retrieval
    - Classification

???

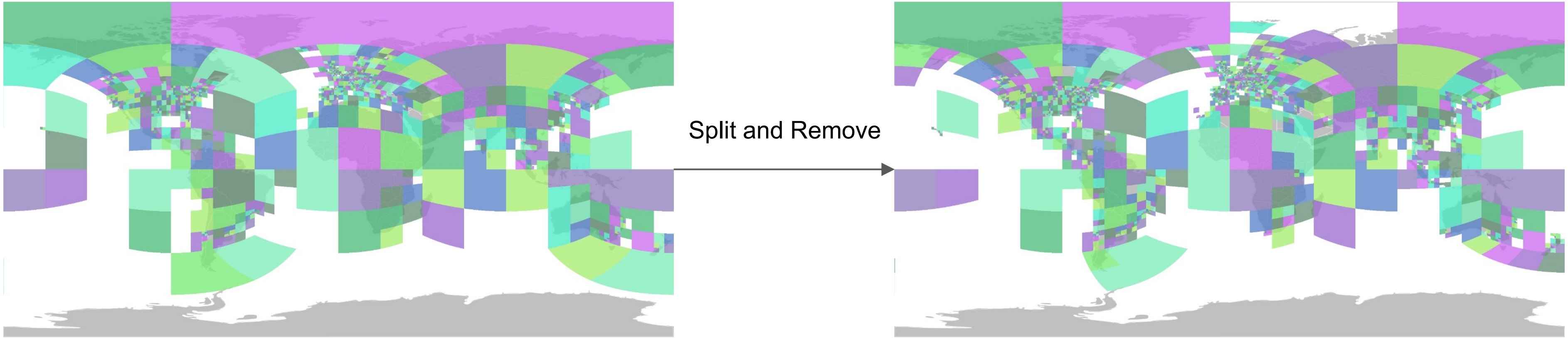
Retrieval

- Match features from a query image to images in a reference database
- Use the matched image's GPS as your prediction



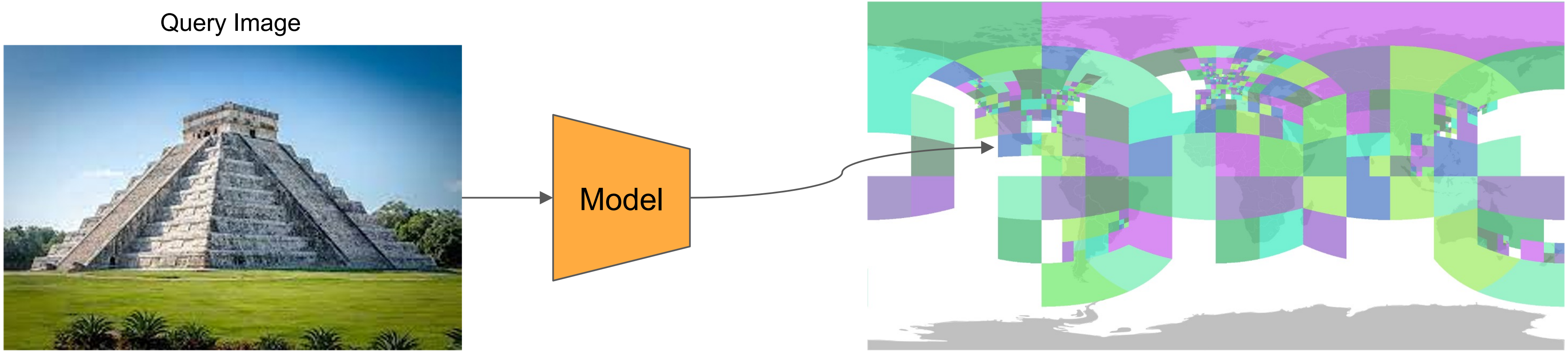
Classification

- Project the Earth onto a cube with each side as a class
- Split classes that are “too large” into 4 new smaller classes
    - “Too large” is defined by having more than t_{max} training images inside of it
- Remove classes that are “too small”
    - “Too small” is defined by having more than t_{min} training images inside of it



Classification

- Split Earth into geographic classes based on the training set
- Predict which class an image belongs to

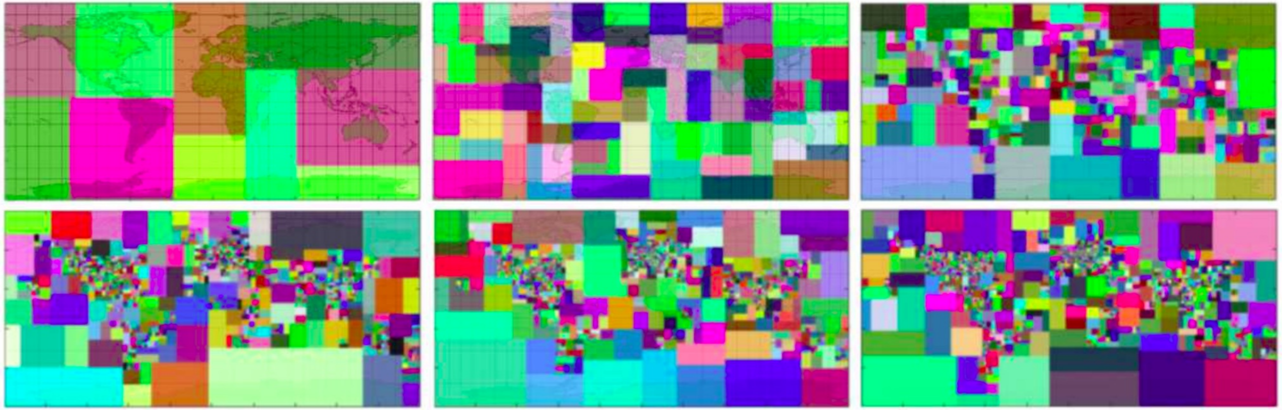


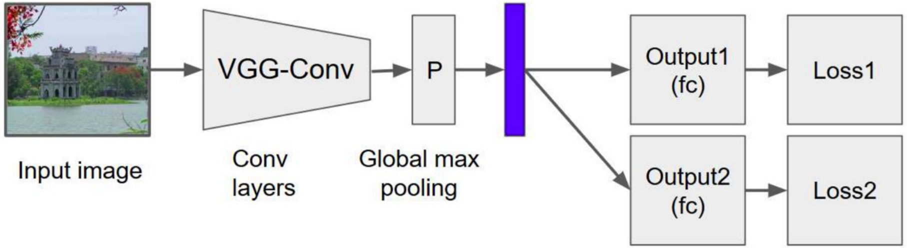
Advantages of Classification Approach

- Classification provides an immediate prediction with one forward pass
- Allows you to cover the entire Earth in cells
- Can use multiple hierarchies of cells to refine prediction

Previous Findings

- *Revisiting IM2GPS in the Deep Learning Era*, Vo et. al. ICCV 2017
- Using multiple partitions of the Earth helps accuracy





Previous Findings

- *Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification*, Muller-Budack et. al. ECCV 2018
- Combine Hierarchical Predictions

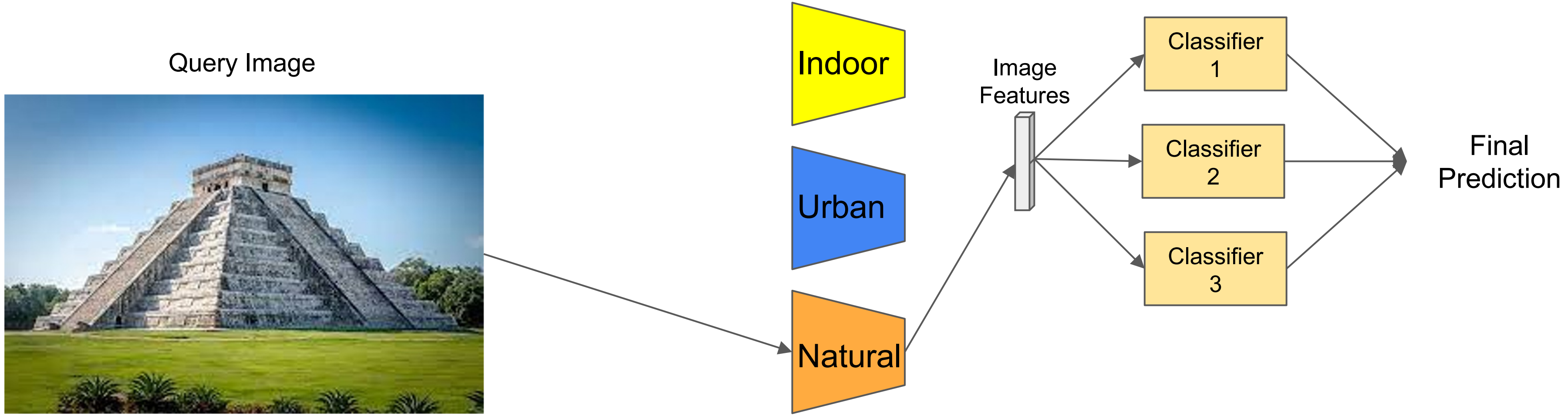
$$P'(\text{Eiffel Tower}) = P(\text{Eiffel Tower}) * P'(\text{Paris})$$

$$P'(\text{Paris}) = P(\text{Paris}) * P'(\text{France})$$

$$P'(\text{France}) = P(\text{France}) * P(\text{Europe})$$

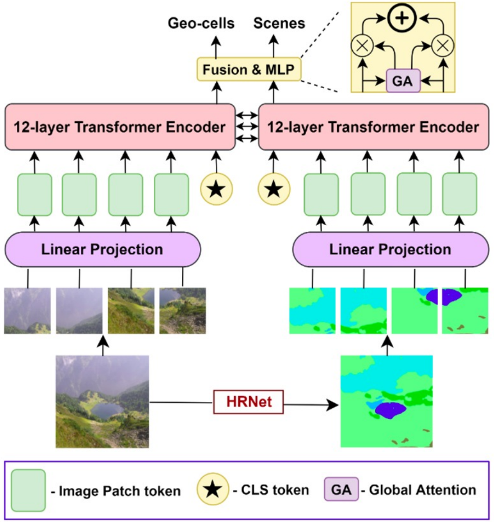
Previous Findings

- *Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification*, Muller-Budack et. al. ECCV 2018
- "Individual Scene Networks" (ISNs)
    - Training images are labelled as "Indoor", "Urban", or "Natural" by a trained model
    - Separate network for each scene label



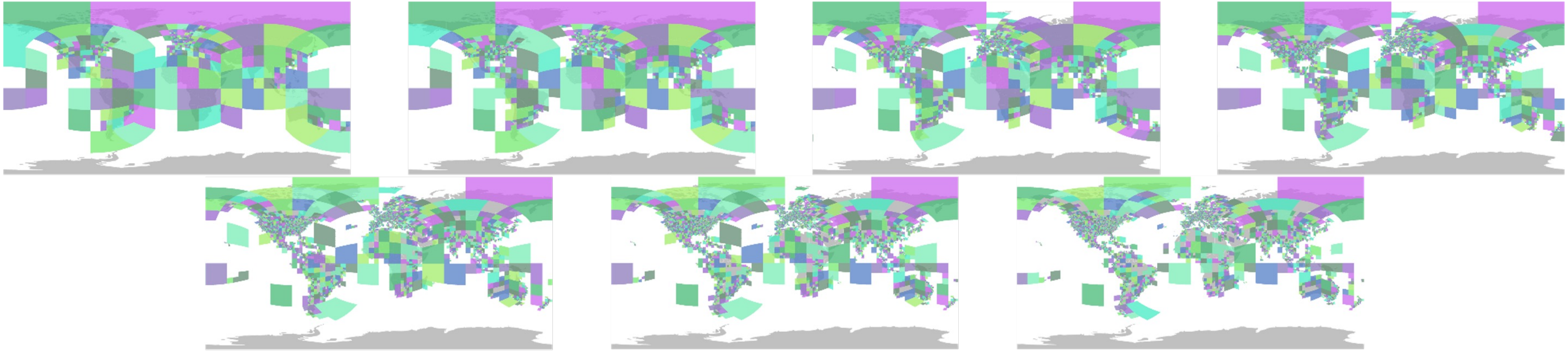
Previous Findings

- *Where in the World is this Image? Transformer-based Geo-localization in the Wild*, Pramanick et. al. ECCV 2022
- First to use Transformers for Geo-classification (Translocator)
- Used Semantic Segmentation to improve accuracy



Our Approach: Hierarchies

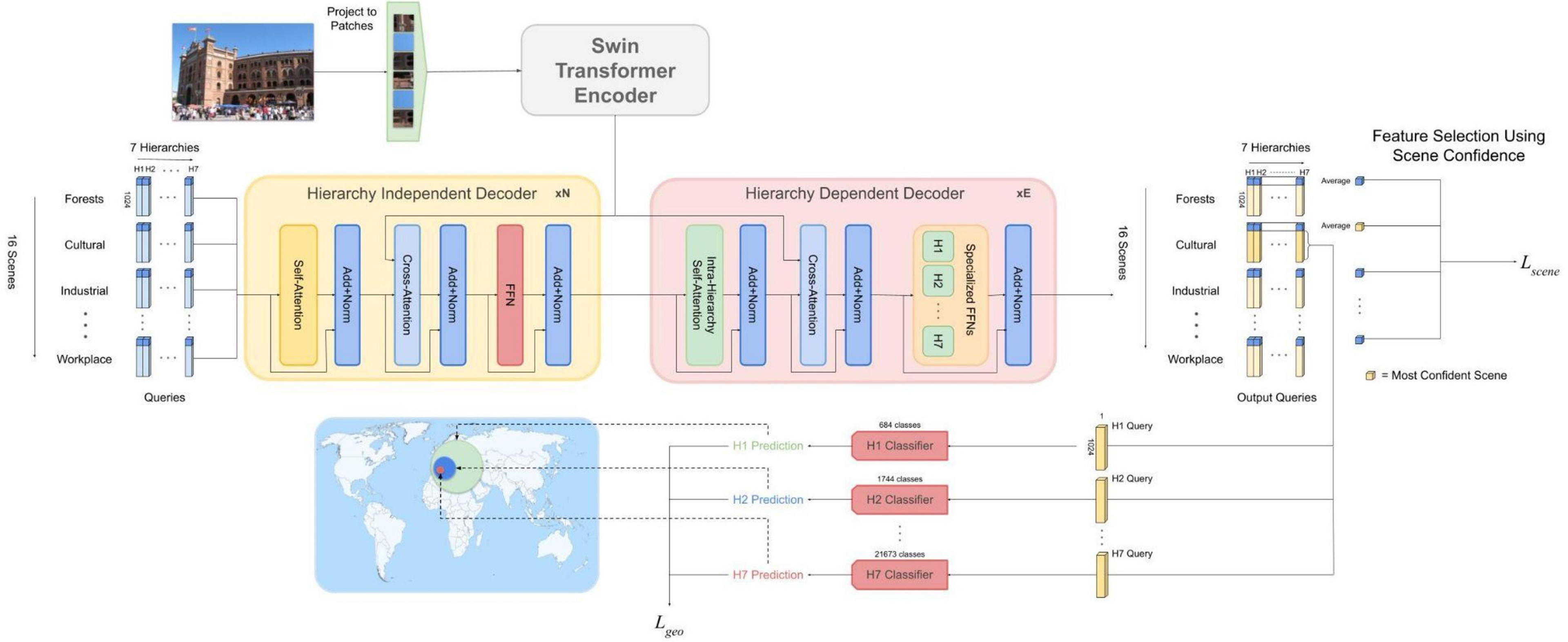
- Previous works only get one set of features for a query image
  - Different hierarchies might need to look at different features
- Our approach extracts features for every geographic hierarchy used
  - 7 hierarchies
    - Ablations on 1, 3, 5, and 7 hierarchies



Our Approach: Scenes

- Previous papers only use 3 scene labels (indoor, urban, natural)
  - Use the labels directly (ISNs) or predict the label (Translocator)
  - While these three labels are easily distinguishable, this can be taken to deeper levels
- We extract features for 16 different scene labels
  - Ablations on 0, 3, 16, 365

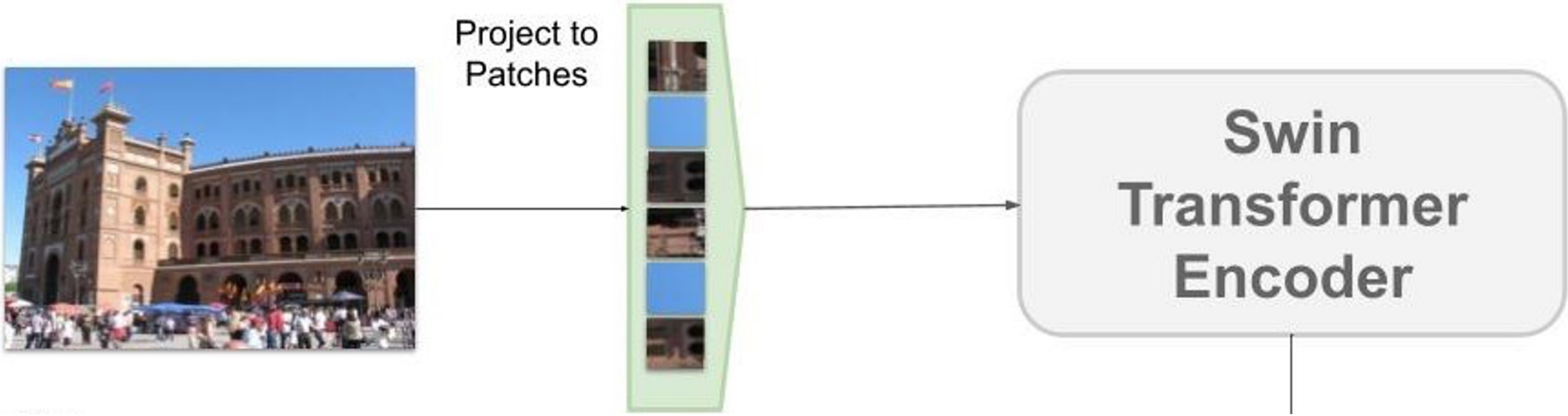
Model



Project to Patches
Swin Transformer Encoder
7 Hierarchies
H1 H2 . . . H7
Forests
Cultural
Industrial
Workplace
16 Scenes
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
Output Queries
Feature Selection Using Scene Confidence
Average
Average
L_{scene}
= Most Confident Scene
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}

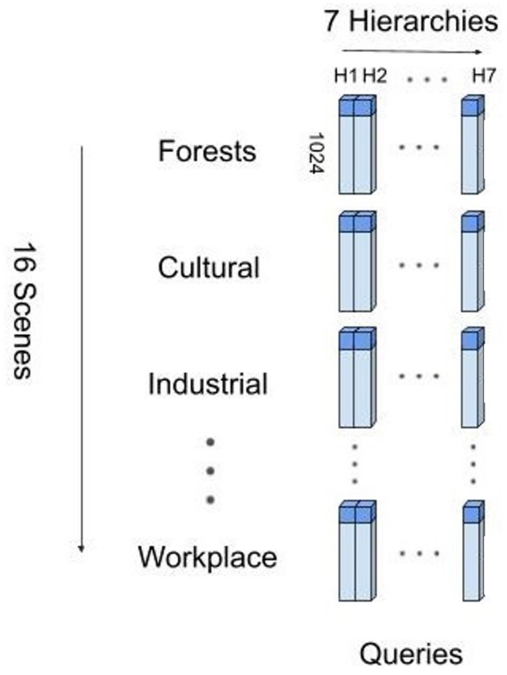
Encoder

- Swin Transformer
- Pre-Trained on ImageNet
- Outputs 7x7x1024 Tensor



Decoder Queries (Hierarchy Queries)

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

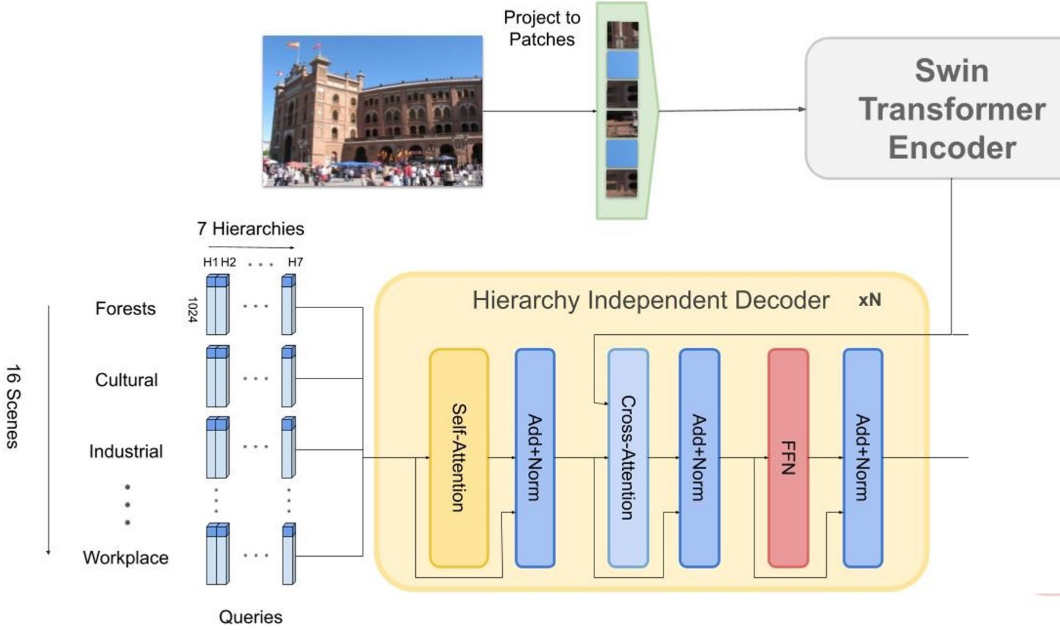


Hierarchy Independent Decoder

- Queries extract image features via Cross-Attention

$$
\begin{aligned}
y^{SA} &= MSA(LN(GQ^{k-1})) + GQ^{k-1}, \\
y^{CA} &= CA(LN(y^{SA}), \\
GQ^{k} &= FFN(LN(y^{CA})) + y^{CA}
\end{aligned}
$$

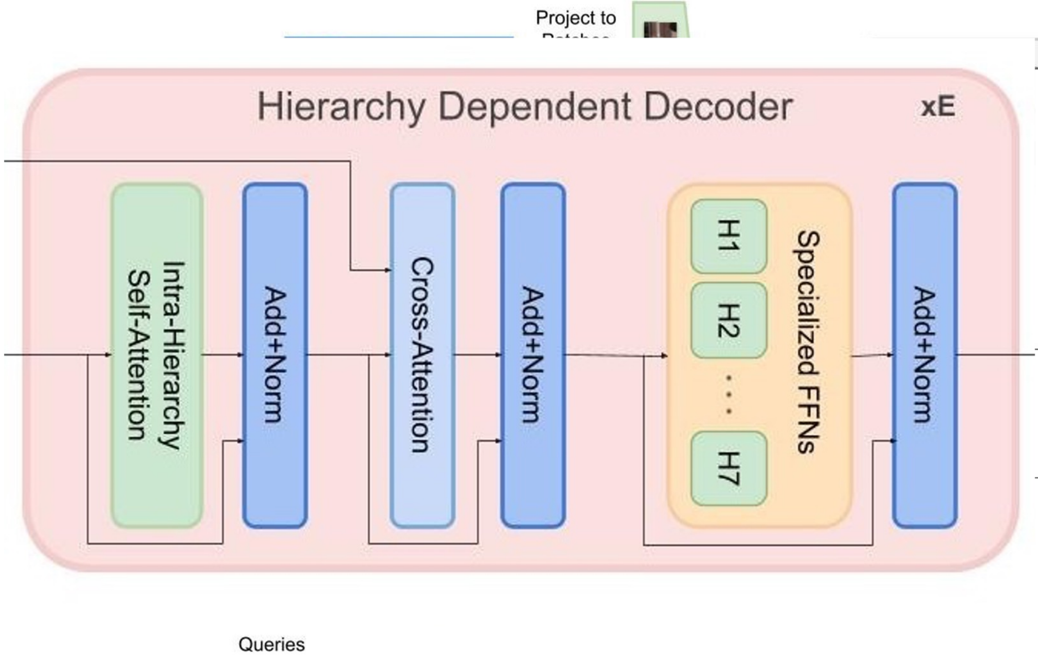
$$
\begin{aligned}
y^{SA} &= MSA(LN(GQ^{k-1})) + GQ^{k-1}. \\
y^{CA} &= CA(LN(y^{SA}, LN(X)) + y^{SA}, \\
GQ^{k} &= FFN(LN(y^{CA})) + y^{CA}.
\end{aligned}
$$



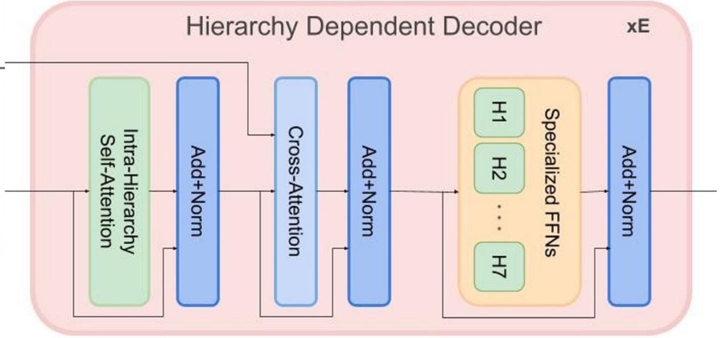
Hierarchy Dependent Decoder

$$y^{SA} = MSA(LN(GQ_h^{k-1})) + GQ_h^{k-1},$$
$$y^{CA} = CA(LN(y^{SA}), LN(X)) + y^{SA},$$
$$GQ_h^k = FFN_h(LN(y^{CA})) + y^{CA}.$$

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

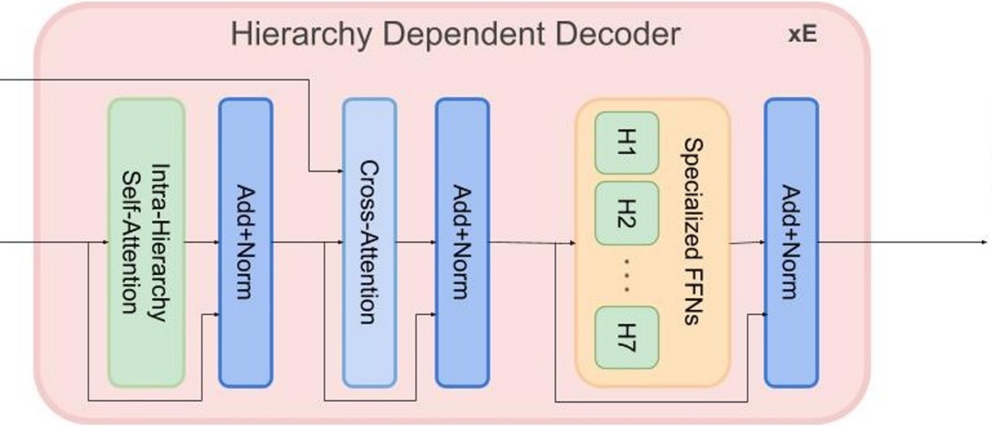


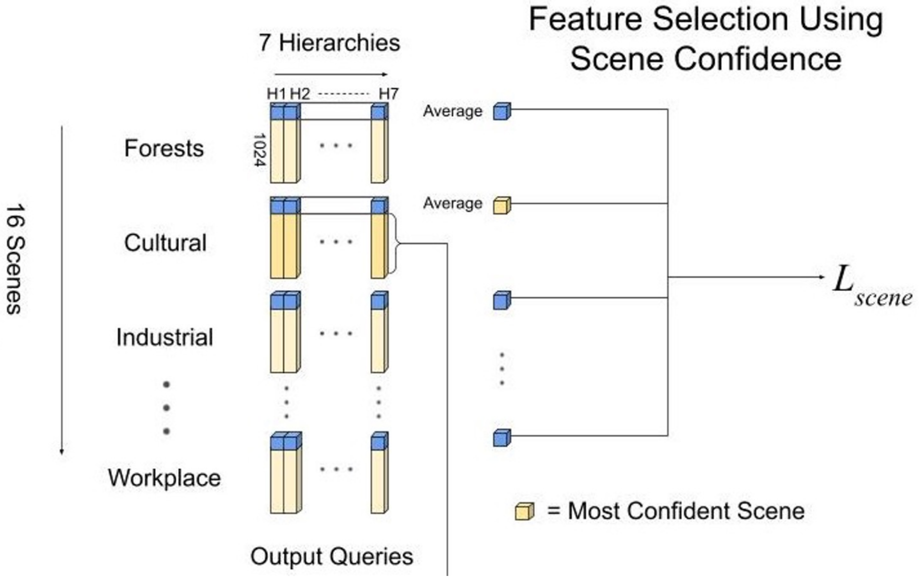
$$y^{SA} = MSA(LN(GQ_h^{k-1})) + GQ_h^{k-1},$$
$$y^{CA} = CA(LN(y^{SA}), LN(X)) + y^{SA},$$
$$GQ_h^k = FFN_h(LN(y^{CA})) + y^{CA}$$



Scene Selection

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

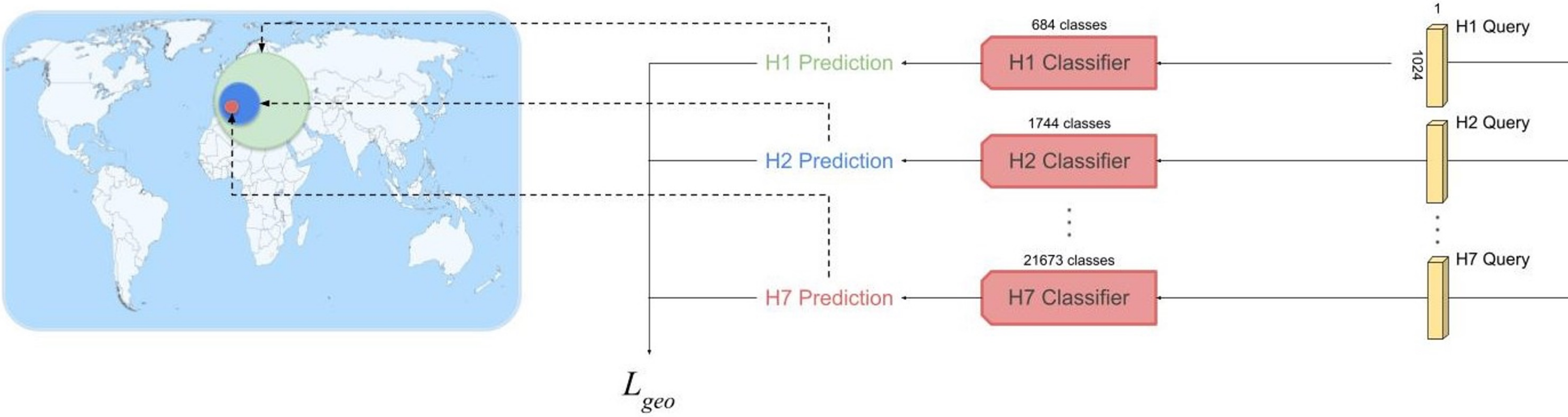




Classification

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

$$p(X|C_a^{H_7}) = p(X|C_a^{H_7}) * p(X|C_b^{H_6}) * \ldots * p(X|C_g^{H_1}),$$



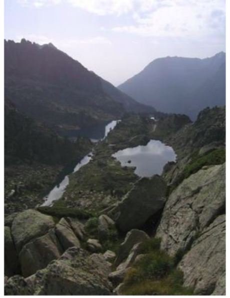
Model and Training Information

- 6 layers of Hierarchy Independent Decoder
- 2 layers of Hierarchy Dependent Decoder

Hyperparameter	Value

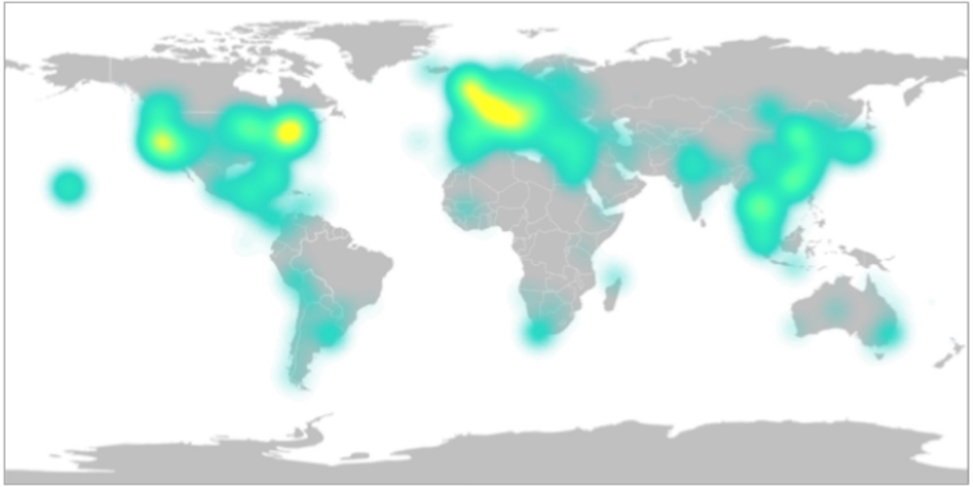
Training Dataset

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



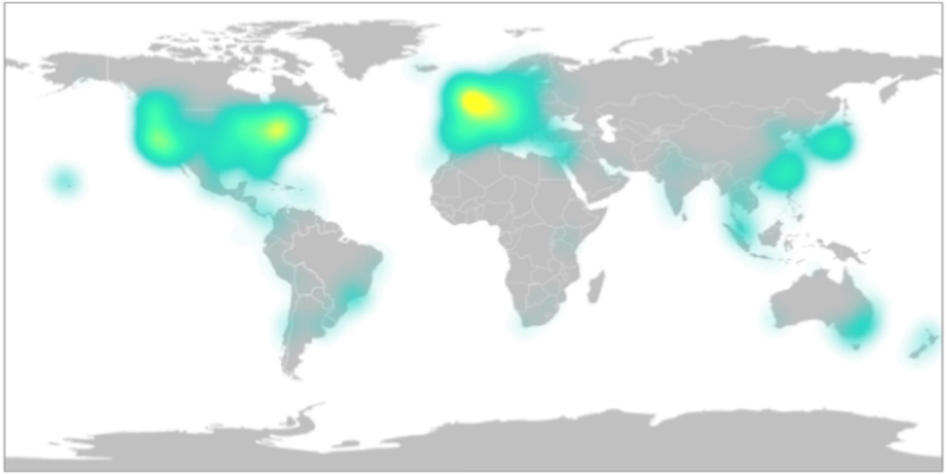
Testing Datasets

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



Testing Datasets

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

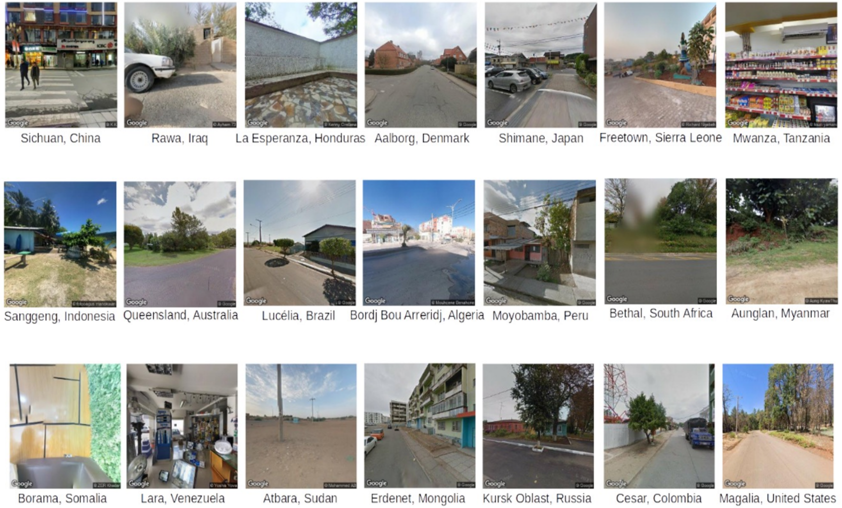


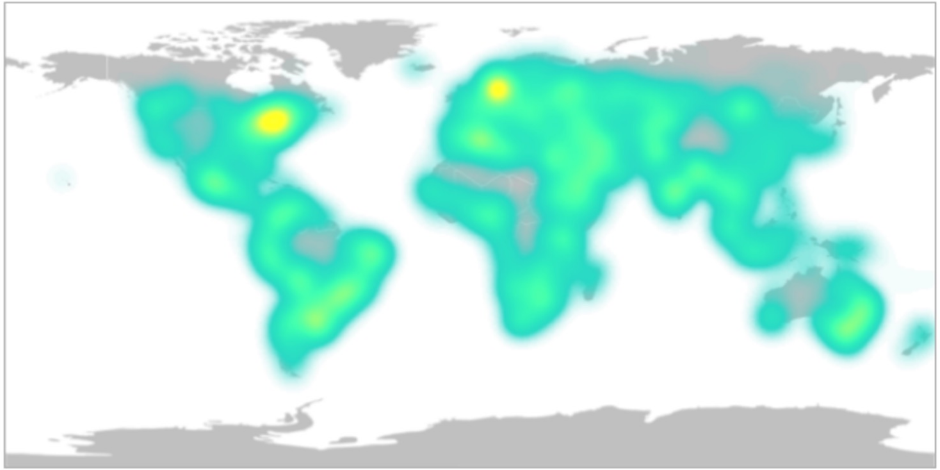
New Dataset

- There’s a problem with existing test sets
  - Landmarks or iconic places are simply a memory task
  - Flickr photos are from random social media users, so many images don’t even have geo-localizable information
  - Not evenly distributed across the Earth
- How do we fix this
  - Collect random images from Google Street View
  - Ensures the image is geo-localizable
  - Can evenly distribute the images over the Earth

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

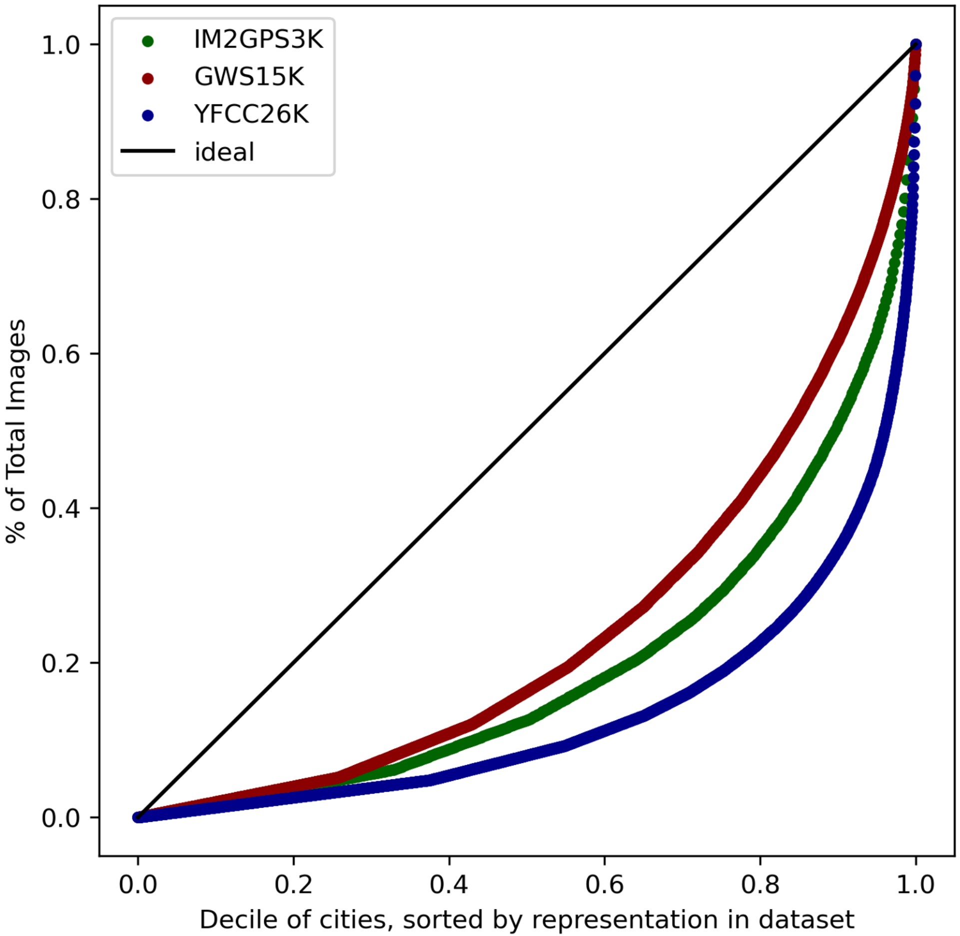




Lorenz Curve

- Helps show fairness of datasets

1. Sort Cities based on # of images
2. Take bottom 10% of Cities
3. Plot the % of Total images on the y-axis



Results

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, $\sigma = 4$ [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 1 **km**	City 25 **km**	Region 200 **km**	Country 750 **km**	Continent 2500 **km**
YFCC $4k$ [21]	[L]kNN, $\sigma = 4$ [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**

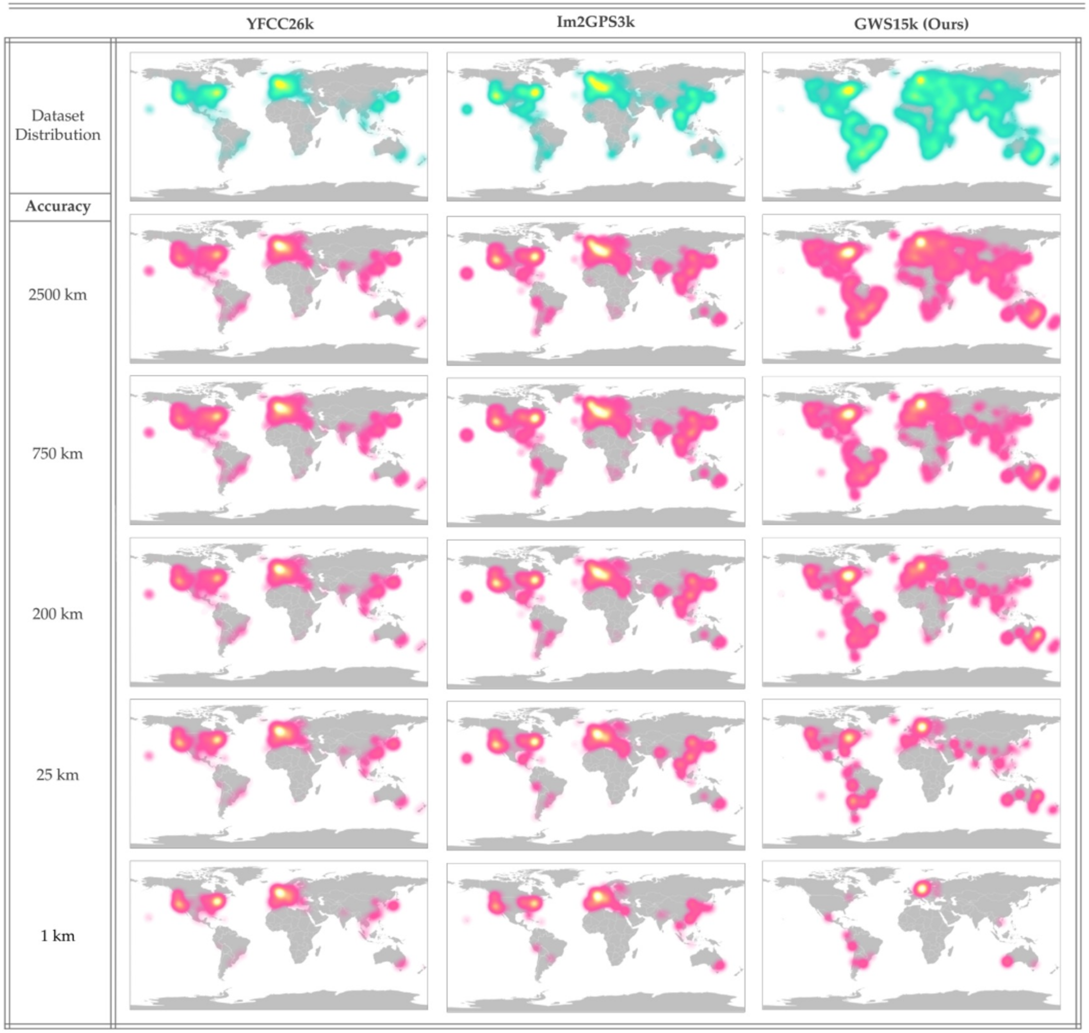
Results on YFCC26k

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 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 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**

Accuracy Distribution



Ablation Study on GeoDecoder Depth

Dataset	Depth	Distance (a_r [%] @ km)				
		Street 1 **km**	City 25 **km**	Region 200 **km**	Country 750 **km**	Continent 2500 **km**
Im2GPS3k [21]	3	11.9	32.9	45.0	59.5	75.4
	5	12.5	33.3	45.2	60.1	75.9
	8	**12.8**	**33.5**	**45.9**	**61.0**	76.1
	10	12.5	33.2	45.2	60.1	**76.2**
YFCC26k [18]	3	9.7	23.5	33.4	49.3	68.3
	5	9.9	23.6	33.8	49.6	68.5
	8	**10.1**	**23.9**	**34.1**	49.6	69.0
	10	10.0	23.7	33.6	**50.1**	**69.2**

Ablation Study on Scene Prediction Method

Dataset	Method	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS3k [21]	No Scene Prediction	11.7	31.5	42.3	57.0	72.3
	Scene Prediction [12]	12.2	32.8	44.3	59.5	75.8
	Ours	**12.8**	**33.5**	**45.9**	**61.0**	**76.1**
YFCC26k [18]	No Scene Prediction	9.4	22.9	32.6	48.0	65.4
	Scene Prediction [12]	9.7	23.2	33.0	48.8	67.0
	Ours	**10.1**	**23.9**	**34.1**	**49.6**	**69.0**

Ablation Study on Number of Scenes

Dataset	# of Scenes	Distance (a_r [%] @ km)				
		Street 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 2500 **km**
Im2GPS3k [10]	0	11.8	30.4	46.2	58.3	77.6
	3	12.0	31.7	47.0	59.8	78.4
	16	**12.2**	**32.0**	**47.9**	**60.5**	**79.8**
	365	11.9	31.8	47.2	58.5	78.6
YFCC26k [9]	0	8.0	19.8	30.1	44.6	62.2
	3	8.4	20.5	31.0	46.0	64.8
	16	**8.7**	21.4	**31.6**	**47.8**	**66.2**
	365	8.5	**21.6**	30.2	46.4	64.9

Ablation Study on Hierarchy Dependent Decoder

Dataset	Layers	Distance (a_r [%] @ km) Street 1 **km**	City 25 **km**	Region 200 **km**	Country 750 **km**	Continent 2500 **km**
Im2GPS 3k [21]	0	12.2	33.2	45.5	60.3	75.8
	2	**12.8**	**33.5**	**45.9**	**61.0**	**76.1**
	4	12.8	33.4	45.0	60.7	75.6
	6	12.6	33.2	44.5	59.9	75.3
YFCC26k [18]	0	9.7	23.5	33.8	49.2	68.7
	2	**10.1**	**23.9**	**34.1**	**49.6**	**69.0**
	4	9.9	23.4	33.6	49.0	68.3
	6	8.7	22.6	33.0	48.6	67.6

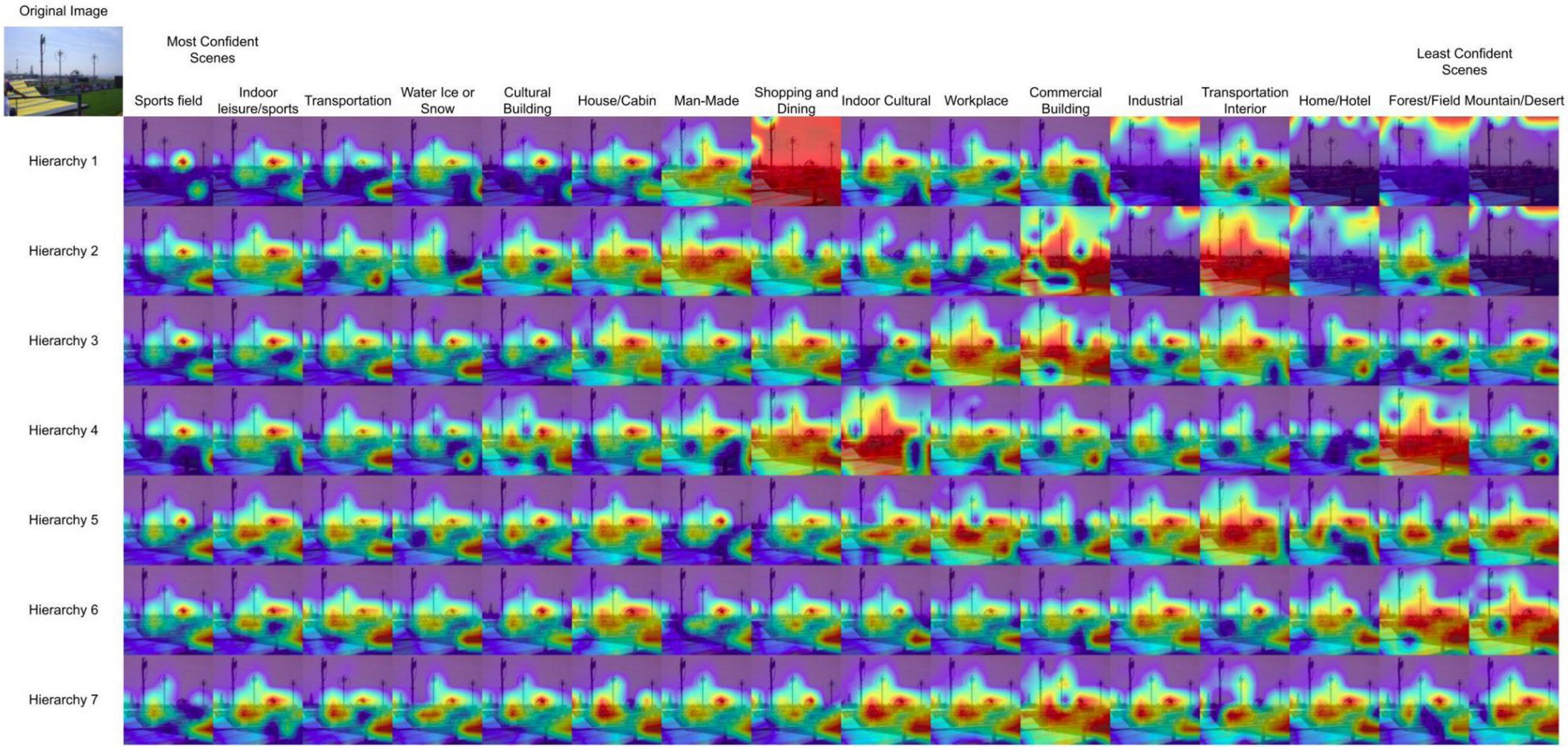
Ablation Study on Encoder Type

Dataset	Model	Distance (a_r [%] @ km)				
		Street 1 **km**	**City** 25 **km**	**Region** 200 **km**	**Country** 750 **km**	**Continent** 2500 **km**
YFCC26k [18]	ViT	6.9	17.3	27.5	40.5	59.5
	Swin	9.6	22.3	33.6	48.0	67.5
	Ours (ViT)	8.7	21.4	31.6	47.8	66.2
	Ours (Swin)	**10.1**	**23.9**	**34.1**	**49.6**	**69.0**

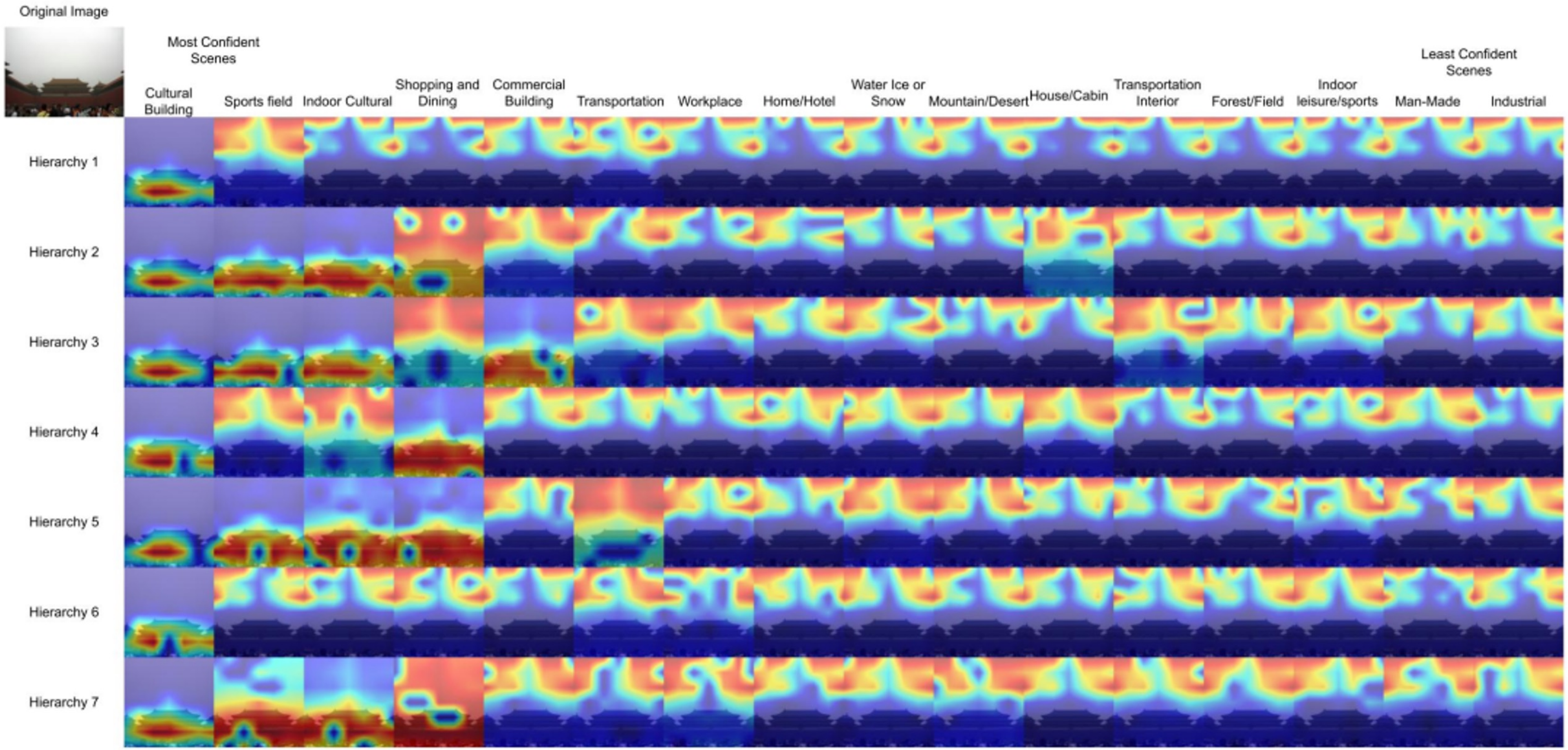
Ablation Study on Number of Hierarchies

Dataset	# of hierarchies	Distance (a_r [%] @ km) Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS3k [10]	1	9.8	29.6	41.1	56.4	73.5
	3	12.8	34.5	**46.1**	**61.5**	**76.7**
	5	13.4	34.4	45.4	61.1	76.1
	7	**14.3**	**34.8**	45.7	61.3	76.0
YFCC26k [9]	1	6.7	18.2	29.0	45.2	64.0
	3	10.1	**24.3**	34.7	**50.1**	**67.8**
	5	10.2	24.1	**34.8**	50.0	67.7
	7	**10.8**	23.5	34.0	49.3	67.4
GWS15k	1	0.0	0.9	5.7	21.8	44.0
	3	0.2	1.3	7.9	**25.4**	**49.4**
	5	**0.6**	**1.7**	**8.1**	24.3	48.0
	7	0.2	1.0	6.9	22.7	46.2

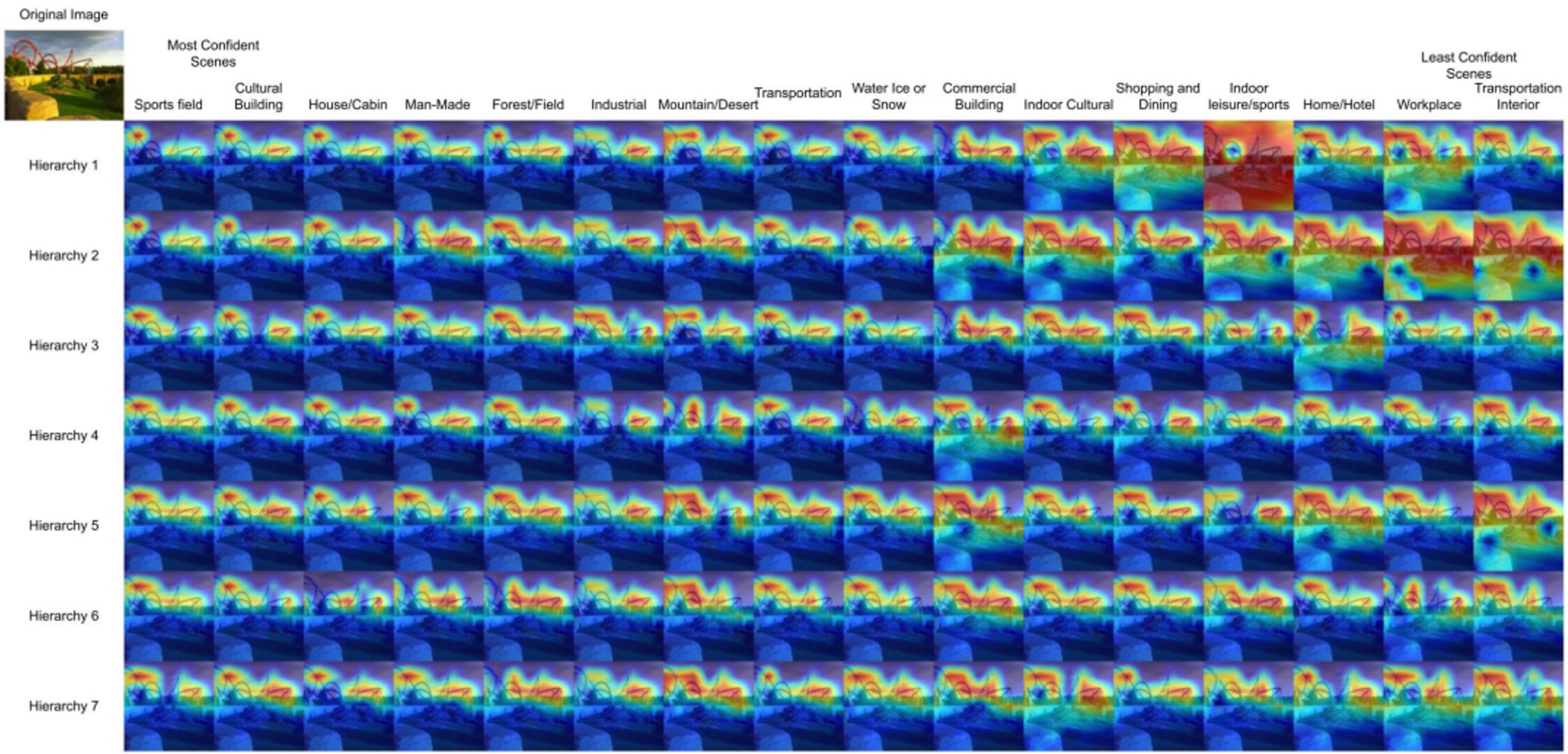
Qualitative Results Im2GPS3k



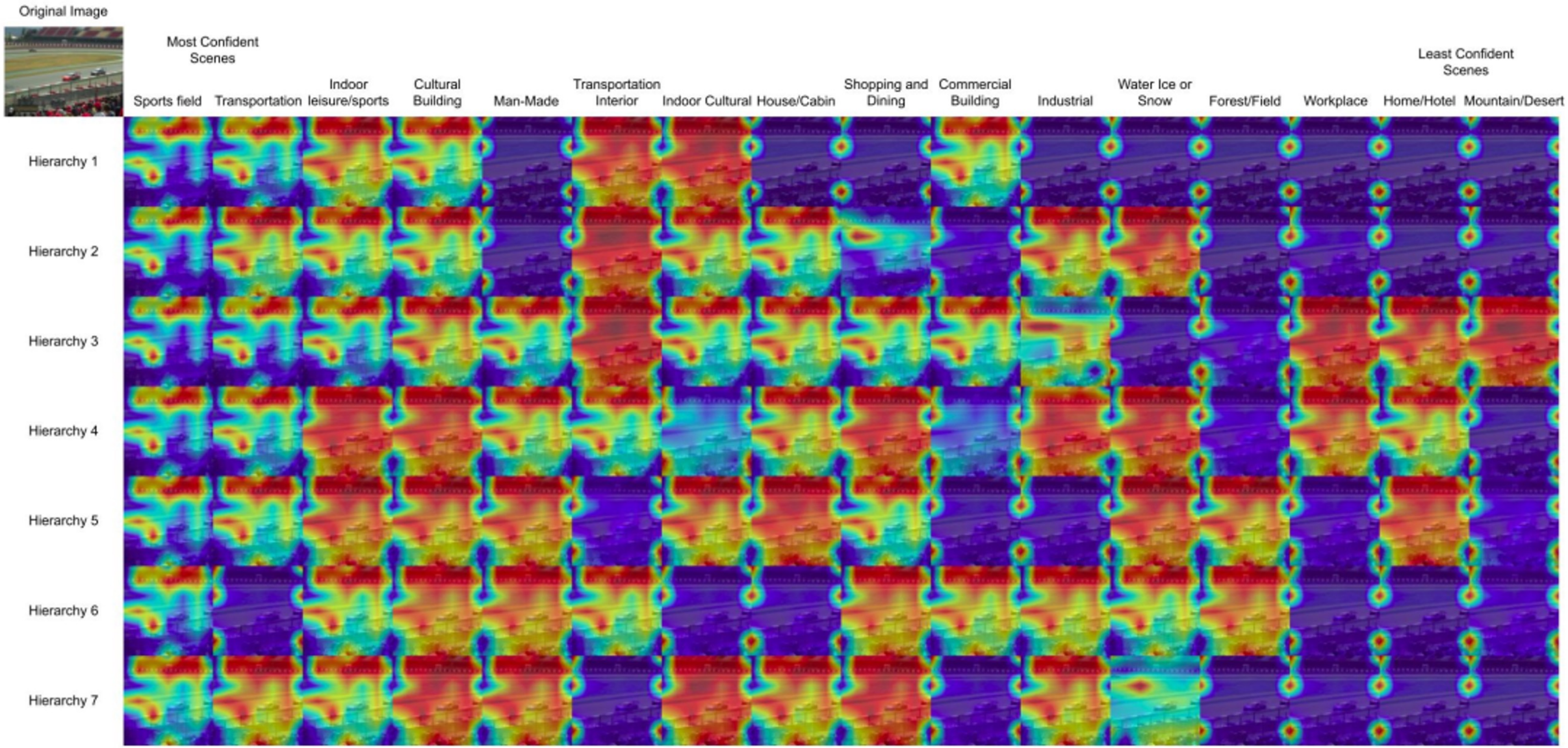
Qualitative Results Im2GPS3k



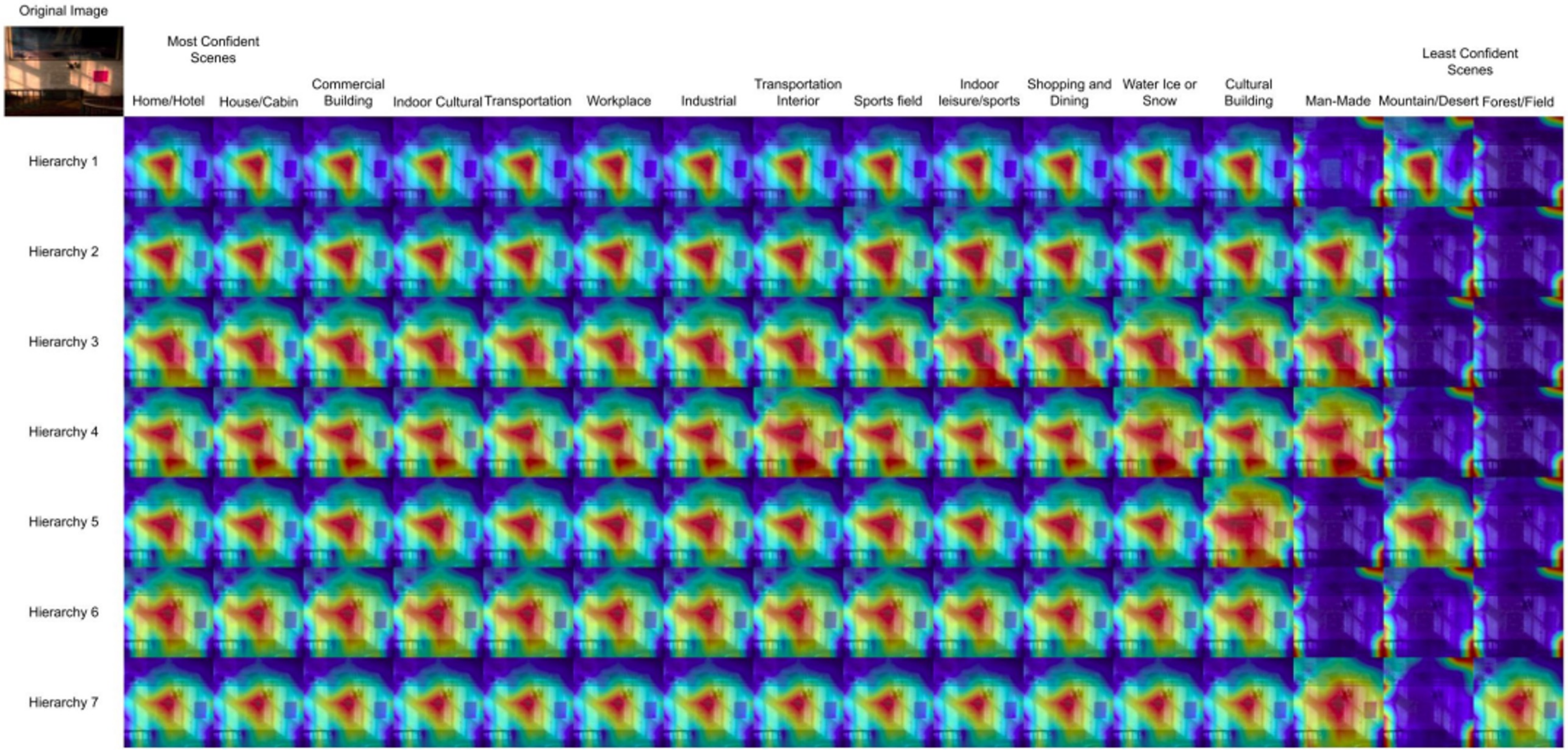
Qualitative Results Im2GPS3k



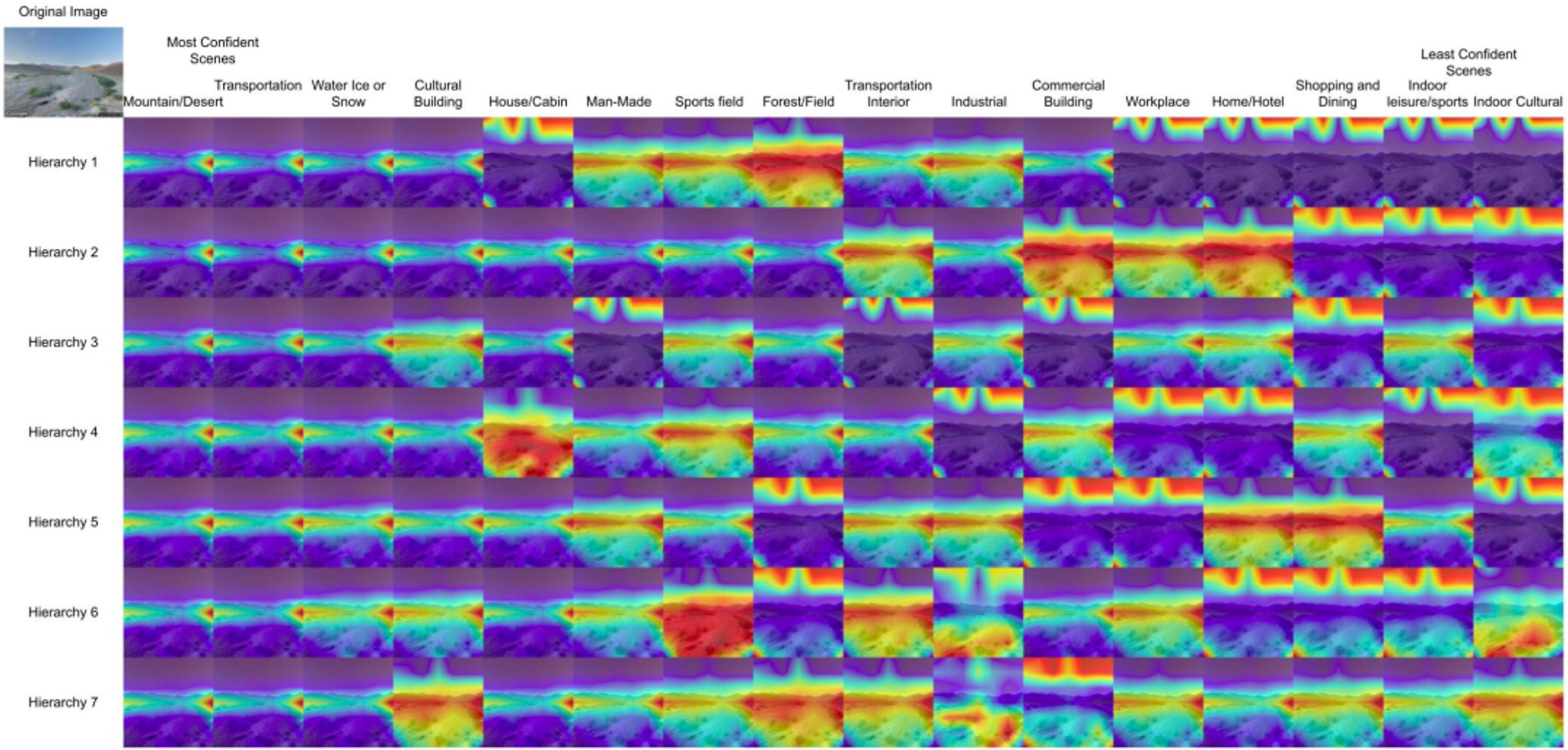
Qualitative Results YFCC26k



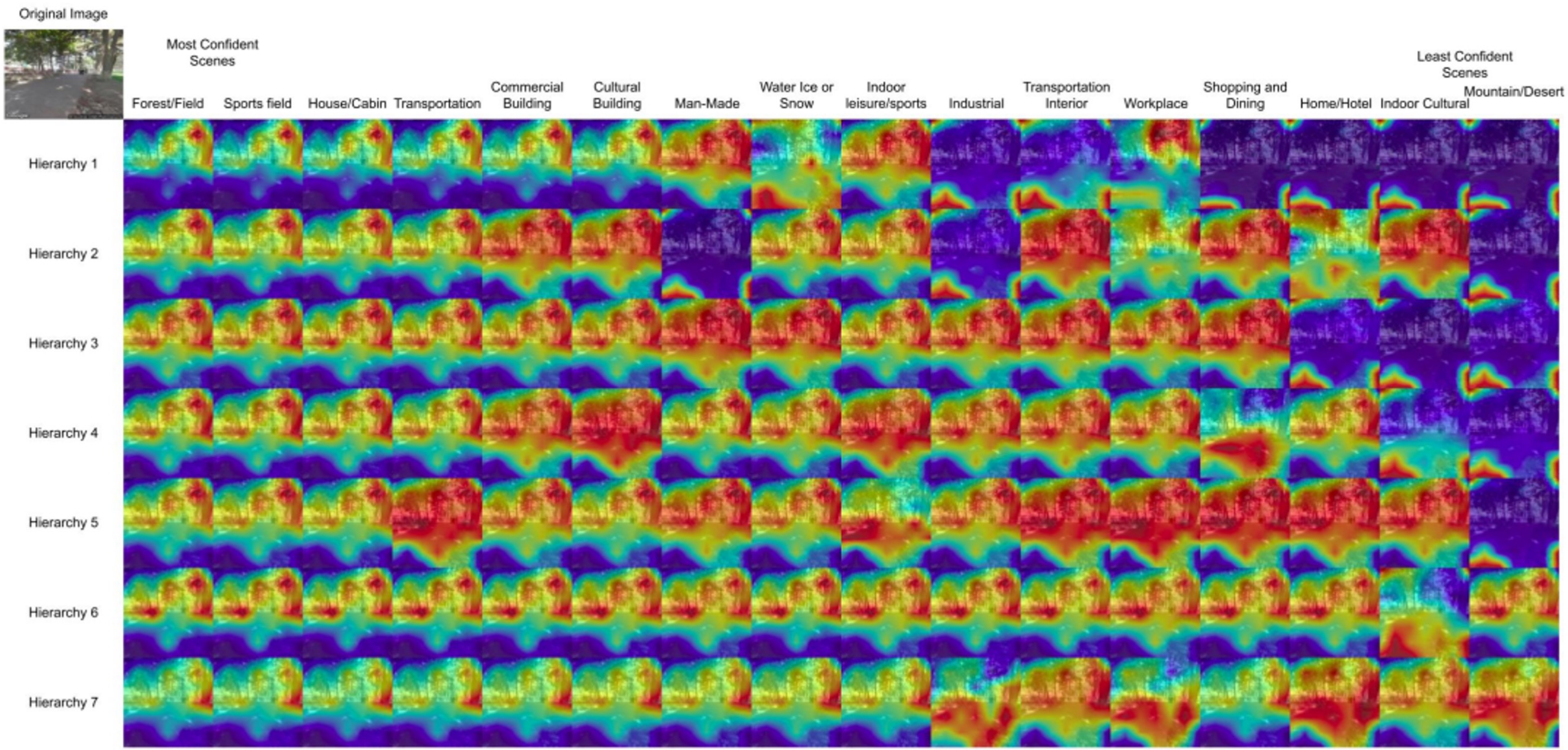
Qualitative Results YFCC26k



Qualitative Results GWS15k



Qualitative Results GWS15k



GWS15k Images Predicted <1Km

GWS15k Images Predicted <25Km

GWS15k Images Predicted <200Km

GWS15k Images Predicted <750Km

GWS15k Images Predicted <2500Km

GWS15k Images Predicted >3000Km