Image-Based Geolocation Estimation using CNNs

Authors:
Kathy Lo, Chia-Chun Hsiao, Kai-Po Chang, Sophie Huang, Yu-Han Huang

Abstract:

Geolocation estimation from images is a challenging task that can predict the geographical location of a photo using only its visual content. This is particularly difficult because photos often lack metadata, and their visual features can vary due to environmental conditions, lighting, and time of year. Applications for this technology include photo organization, disaster response, and environmental monitoring, where location information plays a critical role.

Our project introduces a new deep learning framework that combines adaptive spatial partitioning and scene classification to improve geolocation accuracy. To address the complexity of Earth's surface, we apply the $S2\ geometry\ library3$ to create an adaptive quadtree structure that divides the Earth into hierarchical geographical nodes. This method ensures a balanced distribution of image density across spatial partitions, allowing finer divisions in well-photographed areas and coarser divisions in sparsely captured regions.

To enhance location prediction, we use scene classification to pre-train $ResNet$ model trained on the $Places2$ dataset, which contains 365 scene categories. Each image is assigned scene labels at multiple levels of granularity: fine-grained (365 categories), mid-level (16 groups), and coarse-level (3 types: indoor, natural, and urban). This hierarchical classification bridges detailed scene understanding with broader contextual information, enabling our framework to leverage both fine and coarse cues for geolocation.

The framework operates through four key stages to enhance geolocation accuracy. First, adaptive region partitioning uses S2-based hierarchical subdivision to create spatial nodes with a balanced number of images. Second, we apply scene classification to add geolocation predictions by incorporating environmental semantics. Third, multi-scale learning refines predictions across coarse, medium, and fine spatial scales to improve precision. Lastly, hierarchical geolocation prediction determines precise GPS coordinates by calculating the mean location of training images within the predicted region.g

Our evaluation uses the $IM2GPS$ Test Set, which contains images from diverse global locations, including urban landmarks, natural landscapes, and architectural scenes. Experimental results demonstrate significant improvements in geolocation accuracy, particularly for visually distinct areas. Thus, the adaptive partitioning minimizes biases by searching areas with more photos.

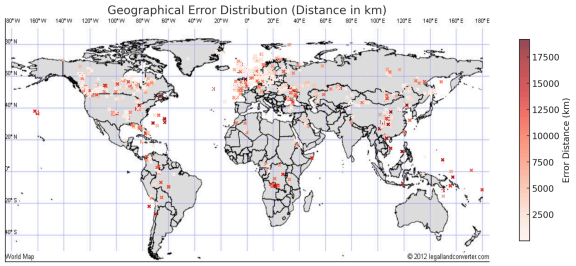

Figure 1. Geographical Error Distribution in im2gps3k dataset

Figure 1 illustrates the geographical error distribution for our geolocation estimation model. Most errors are concentrated in lighter shades, which indicates lower distance errors in most regions. Our results demonstrate that our trained model performs well in accurately localizing images across densely photographed areas, indicating that combining spatial partitioning, scene classification, multi-scale learning, and hierarchical spatial information can create a reliable and scalable method for image-based geolocation estimation.

1. Introduction

1.1 Problem Statement

Predicting where a photo was taken just from its visual features is no small feat. Without any clear metadata, an image can look completely different depending on the weather, time of day, or season conditions. Plus, our planet is huge and diverse—ranging from metropolises to remote natural landscapes—so identifying a location isn't straightforward without some extra context. If we can get a closed or even correct answer, it will be helpful in multi areas like photo organization, disaster response, environment monitoring, and etc.

1.2 Motivation

With the rapid growth of visual data on social media, photo-sharing platforms, and surveillance systems, there is an increasing need for reliable methods to determine where a photo was taken. While traditional methods rely on GPS metadata, such information is often unavailable or intentionally removed for privacy reasons. To deal with this, we use classical methods and deep learning models as powerful tools for extracting meaningful patterns from visual data to infer location. However, existing methods often overlook the importance of environmental context, such as whether an image depicts natural scenery, urban settings, or indoor spaces, which can provide valuable clues for geolocation. By integrating scene classification with hierarchical spatial partitioning, we believe our approach could bridge this gap and improve the precision of location predictions using only visual cues.

1.3 Related Work

We draw inspiration from the long history of predicting a photo's location using visual content, beginning with Im2GPS by Hays and Efros [4]. Im2GPS matched a query image to a database of geotagged images to estimate the location. Although this retrieval-based approach could work well for landmarks, it struggled with generic or blurry images that lacked distinctive features. Building on this, Weyand et al. introduced PlaNet [11], which turned geolocation into a classification problem. The idea of PlaNet is to divide the Earth into adaptive region nodes, ensuring balanced data distribution across regions. However, PlaNet still faced challenges when dealing with images from underrepresented or ambiguous regions.

To address the limitations, Vo et al. revisited Im2GPS and introduced a multi-scale partitioning approach [10]. By combining information from coarse, medium, and fine geographic partitions, this method allowed models to refine their predictions using both global and local features. This idea strongly influenced our work, where we also adopt multi-resolution partitioning to improve accuracy.

Another key advancement is the use of scene context, which helps models understand the type of environment shown in an image. Zhou et al. introduced the Places2 Dataset [13], a large collection of images labeled into 365 scene categories, such as indoor, natural, and urban environments. These hierarchical labels bridge low-level image features and geographic information. Inspired by this, our framework uses scene classification to incorporate environmental context, making geolocation predictions more precise.

Our work is also influenced by hierarchical classification methods like YOLO9000 (Redmon and Farhadi, 2017) [5]. YOLO9000 successfully combined coarse and fine object categories to improve detection, and we apply a similar idea to geolocation by combining spatial probabilities from different scales. Building on this, we introduce Context-Specific Networks (CSNs)(introduced in section 2.3), which are tailored to predict geolocations based on specific scene types, such as indoor environments or natural landscapes.

1.4 Approach summary

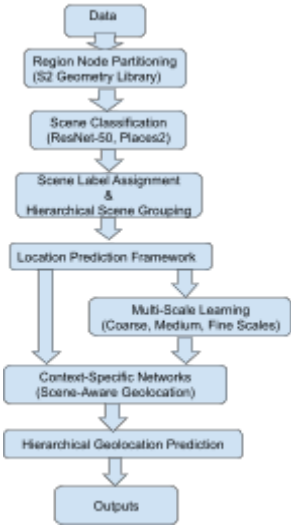

Fig2. Geolocation estimation workflow.

Our geolocation prediction framework makes GPS estimation by integrating hierarchical spatial partitioning, deep learning, and contextual insights. In the first step, we organize raw image data by splitting the Earth into a hierarchy of geographically balanced regions using S2 geometry and quadtree structures to balance image density across regions. Next, we use these regions to classify each image using deep neural networks to identify the type of scene and give them several tags. Finally, we adopt a multi-scale learning strategy. At larger scales, we capture broad patterns, while at finer scales, we zoom in for more detailed, context-specific insights. Together, the hierarchical partitioning, scene-aware classification, and multi-scale refinement create a flexible framework that

adapts to diverse environments. Figure 2 shows the workflow of our approach.

2. Details of the approach:

Our core idea is to divide the Earth's surface into adaptive spatial regions, ensuring that each region contains a balanced number of images. By using this adaptive partitioning, we aim to mitigate biases and improve the precision of location predictions. To enhance the framework, we use the visual content of images to extract information about their surroundings. Specifically, each image is assigned a scene label from a set of 365 categories derived from the $Places2$ dataset, which captures diverse environmental scenarios.

Building on this foundation, we incorporate strategies to integrate both the scene-based information and multiple spatial partitioning levels to improve accuracy. These strategies enable the framework to consider a variety of spatial resolutions and scene contexts, ensuring more reliable geolocation predictions. To determine the location of an image, our approach combines predictions across different levels of granularity, resulting in a hierarchical estimation of GPS coordinates.

2.1 Region Node Partitioning

To generate a set of distinct geographical region nodes C, we utilize the $S2\ geometry\ library3$. The process starts by projecting the Earth's surface onto a cube, where each face represents an initial node. Using the GPS coordinates of the images, the nodes are adaptively divided in a hierarchical approach. With a quadtree structure, subdivision begins at the root and continues until no node contains more images than the specified maximum threshold τ_{max}.

Hierarchical Subdivision:
The adaptive partitioning begins with coarse-level nodes, representing large geographical areas. Each node is subdivided into four smaller nodes (quadtree structure) until the number of images in every node is less than or equal to τ_{max}.

Filtering Sparse Nodes:
Nodes containing fewer than τ_{min} images are removed to exclude regions with limited distinguishing features, such as poles or oceans. To address edge cases, where images are located near the boundaries between dense and sparse nodes, we implement a series of additional checks to ensure proper allocation. Images at these boundaries are evaluated based on their geographical proximity and feature similarity to neighboring dense nodes. If an image is closer to or better matches the characteristics of a dense node, it is reassigned accordingly. This reassignment would ensure that edge images contribute to meaningful regions while preserving a balanced node representation.

Geolocation Accuracy with different partitioning levels

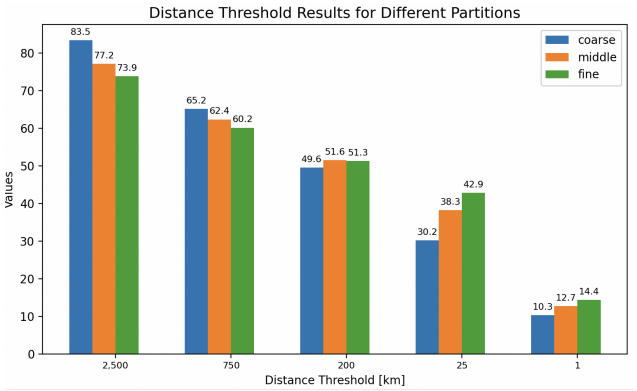

Figure 3. Distance Threshold Results for Different Partition Levels

Figure 3 presents a comparison of geolocation accuracy under various region node partitioning strategies across different distance thresholds. The results evaluate the base approach with and without additional subdivisions, using the coarse, middle, and fine partitioning levels. The base method serves as the reference and its accuracy is displayed at the center of the x-axis for each threshold. It can be observed that finer subdivisions improve accuracy at smaller distance thresholds, while coarser partitions tend to perform better at larger thresholds. This demonstrates the trade-off between partition granularity and geolocation precision.

Benefits of the Region Node Partitioning
The adaptive region partitioning method offers clear benefits over dividing the Earth's surface into equal areas. It avoids creating large regions in places with few photos and small regions in areas with many photos. We improve geolocation accuracy by focusing on areas with dense images, especially for landmarks and urban locations.

2.2 Hierarchical Scene Classification for Visual Context

ResNet Architecture:
We use a $ResNet$ model with 50 layers as the backbone for scene classification. The model is trained on the $Places2$ dataset [13]with 365 scene categories. Its ability to differentiate between diverse environments aligns with the objectives of our geolocation estimation framework.

Scene Label Assignment:
The model computes probabilities $P(c_i)$ for each of the 365 categories $c_i \in S_{365}$. Scene labels are assigned using the highest-probability category:

$$Label(x) = max_{c_i \in S_{365}} P(c_i)$$

Hierarchical Scene Grouping:
The provided scene hierarchy groups the 365 categories into 16 superordinate groups S_{16} and 3 broader categories S_3: indoor, natural (outdoor, natural), and urban (outdoor, man-made). Probabilities for each group are calculated by summing the probabilities of all classes within the group. The final label is selected based on the highest aggregated probability.

Handling Overlapping Categories:
Some categories belong to multiple superordinate groups, such as natural outdoor and man-made outdoor. To normalize, the probability of these categories is divided by the number of their assigned groups. The normalization enhances the model's interpretability by fairly distributing probabilities across overlapping groups. It also improves accuracy by preventing one category from overly influencing predictions.

We believe hierarchical classification can 1) enhance geolocation accuracy by extracting scene semantics at multiple levels, 2) link fine-grained scene features to broader environmental categories, and 3) provide richer contextual information.

2.3 Location Prediction Framework

We implemented a deep learning-based framework that integrates 1) hierarchical spatial partitioning from Section 2.2, 2) multi-scale learning, and 3) scene-aware context. By using convolutional neural networks(CNNs), we combine global and local spatial information while incorporating environmental context to improve prediction accuracy.

Basic Predictor
The foundation of our location prediction system is a single-resolution classifier that uses only the single region node partitioning as described in Section 2.1. We employed a ResNet-based CNN to extract hierarchical features from input images. The network includes convolutional layers, batch normalization, ReLU activations, and residual connections. Then we added a global average pooling layer, followed by a fully connected layer that outputs class probabilities for the geographical cells. The number of neurons in the layer corresponds to the total number of region node R.

To train the model, we optimized a cross-entropy loss function, minimizing the difference between predicted probabilities and the ground-truth labels:

$$L_{loc}^{single} = -\sum_{i=1}^{|Z|} R_i^{GT} \cdot log\left(\widehat{R_i}\right)$$

where $\widehat{R_i}$ is the predicted probability of region node i, and R_i^{GT} is the one-hot encoded ground-truth label.

Multi-Scale Learning Framework
To capture information at multiple spatial levels, we extended the basic predictor model with a multi-scale learning framework. This multi-scale learning framework can learn geolocation features at multiple spatial resolutions such as coarse, medium, and fine scales, which enables the network to integrate information at different levels of granularity.

We implemented additional fully connected layers for each spatial scale, where each layer outputs class probabilities for its corresponding region nodes. We calculated the multi-partitioning classification loss as the average of the cross-entropy losses across all scale:

$$L_{loc}^{multi} = \frac{1}{n}\sum_{j=1}^{n} L_{loc}^{single}\left(R_j\right)$$

where n is the number of spatial resolutions and R_j represents the region nodes for each partitioning. The model can now

simultaneously learn global and local spatial features and the effect is discussed later in Section 4.4.

Context-Specific Networks

To further enhance predictions, we integrate the environmental scene context into our multi-scale learning framework, using what we call Context-Specific Networks(CSNs). Each CSN specializes in a particular type of scene (e.g, either indoor, urban, or natural), thereby reducing the complexity of geolocation feature learning by narrowing the focus to specific scenarios.

The process begins with scene classification as outlined in Section 2.2. A pre-trained scene classifier assigns probabilities S for each scene type. Images with a scene probability greater than a threshold τ_S are used to train the corresponding CSN.

To reduce the diversity in the data space, we fine tuned models initially trained without scene restrictions using images specific to a given environmental category S_k ($k \in \{indoor, natural, urban\}$). For example, an urban CSN learns features relevant to cityscapes, and a natural CSN focuses on landscapes like forests and mountains.

For query images, the scene classifier first predicts the most likely scene type. Based on this prediction, the corresponding CSN is selected to perform geolocation. Each CSN produces region probabilities across all spatial scales, which are further refined hierarchically(Section 2.4) to improve the final prediction.

2.4 Hierarchical Geolocation Prediction with Multi-Scale Spatial Representation

Built upon the framework introduced in Section 2.3, we enhance our prediction stability by applying the trained model on three evenly sampled crops of a query image. The technique reduces noise by managing differences in image orientation and content and averaging class probabilities from these crops. We believe this can balance stability and efficiency and provide enough coverage to ensure reliable predictions without increasing too much processing time.

Using the multi-partitioning described in Section 3.3, calculated class probabilities at multiple spatial resolutions. For each query image, the class with the highest probability is selected, and the corresponding geographical region is assigned. Since it includes both detailed and broad spatial information, we thought combining probabilities from all levels can help us make predictions more accurate.

Hierarchical Spatial Framework

In order to make every geographical region in the finest partition hierarchically linked to a larger parent area, we use the adaptive subdivision threshold introduced in Section 2.1. These methods create a geographical hierarchy spanning multiple spatial resolutions. Inspired by YOLO9000's hierarchical classification approach [5], we multiply the probabilities at each level of the hierarchy. As a result, we believe this can make coarser level predictions to refine the results for finer subdivisions.

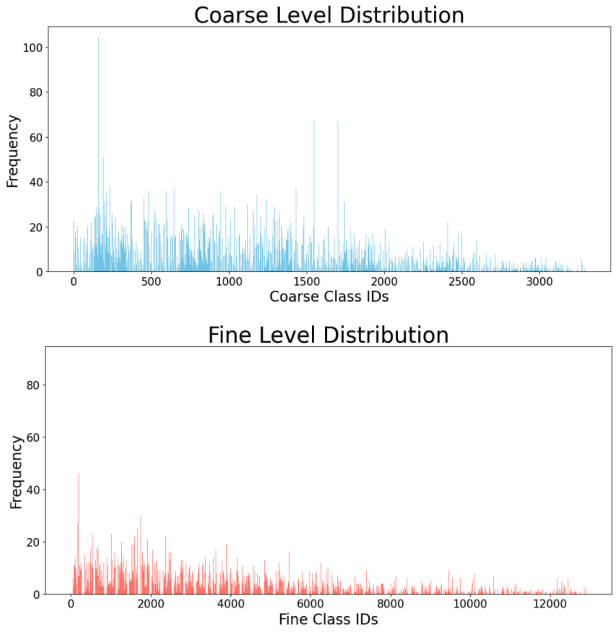

Fig4. Coarse-level distribution(top) and fine-level distribution(bottom)

In figure 4, two charts represent the distribution of classes at two different levels of granularity in our hierarchical geolocation model: coarse and fine. This highlights the hierarchical structure of our model, with coarse level nodes capturing border regions and fine level nodes providing detailed granularity.

Class-to-GPS Mapping

Based on the predicted class of the given query image, we determine the GPS coordinates for the query image. We calculate the mean coordinates of all training images within the predicted region rather than relying on the geographic

center. For instance, consider a region that contains both an ocean and a city near its boundary, if were were to use the geographic center as the predicted location, it would likely fall somewhere in the ocean, far from the actual area where most photos were taken.

3. Results

3.1 Experimental protocols

For hierarchical spatial partitioning, the Earth's surface was divided into region nodes at multiple spatial resolutions (coarse, medium, fine) using the S2 geometry library. Sparse region nodes (those with fewer than a minimum threshold of images) were excluded to enhance classification precision.

Then we used a pre-trained ResNet model (trained on the Places2 dataset) for multi-partitioning learning to assign hierarchical scene labels(indoor, natural, urban) and set stochastic gradient descent with an initial learning rate of 0.01, momentum of 0.9, and weight decay of 0.0001. For augmentation, we performed random cropping and flipping of images to 224x224 pixels. Finally, we used a subset of 25,600 images from the YFCC100M dataset for validation during training.

3.2 Training and test data

For training data, we used a subset of the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M) [13], containing approximately 4.72 million geo-tagged images.

To evaluate the effectiveness of the proposed framework, for test data, we use the *IM2GPS* Test Set, which consists of 237 images from diverse locations worldwide. This dataset is a mix of recognizable landmarks and generic scenes. To further evaluate, we used a larger Im2GPS3k testset with 3,000 geo-tagged images(more diverse and challenging benchmark).

3.3 Example output

The Figure 5 table provided represents a sample of the output generated by our model, where we predict geographical coordinates (latitude and longitude) for given images. Each row includes the following:

img_id: The unique identifier for an image, which combines its Flickr ID and user information.
pred_lat: The latitude predicted by our model.
pred_lng: The longitude predicted by our model.

For instance, the first row corresponds to an image with the ID 104123223_7410c654ba_19_19355699@N00. For this image, our model predicts a latitude of 32.7313 and a longitude of -117.154, which places it near San Diego, California. Similarly, the predictions for the other rows align with their respective IDs.

By analyzing this output, we can validate the performance of our model in estimating geolocations based on visual features extracted from images. This table provides clear evidence of how our approach operates on diverse data and showcases the precision of our predictions.

img_id	pred_lat	pred_lng
104123223_7410c654ba_19_19355699@N00	32.7313	-117.154
1095548455_f636d22cbb_1277_8576809@N08	32.79269	-90.87
1185597181_0158ab4213_1311_43616936@N00	42.34779	-71.086
1199004207_0ce4e7a456_1285_16418049@N00	36.0611	27.18043
1257001714_3453f5fc4b_1405_11490799@N08	45.4119	-75.7143

Figure 5. Sample of the output generated by our model

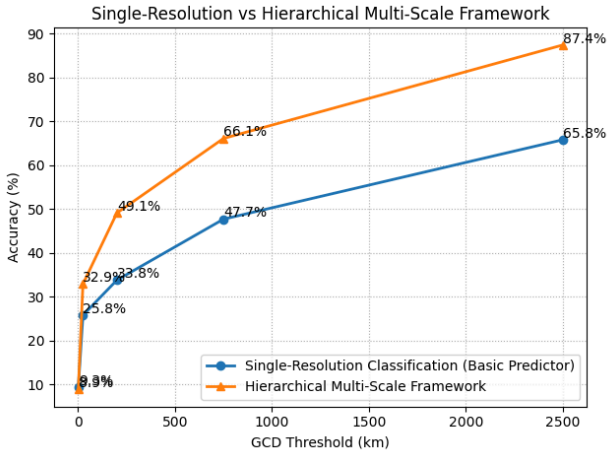

Figure 6. Accuracy of single resolution(basic predictor) versus multi-scale learning framework(with scene-aware context integration) across varying GCD thresholds (1 km to 2,500 km).

3.4 Evaluations

As part of the evaluation for the location prediction framework, we conducted a performance comparison between the Single-Resolution Classification (Basic Predictor) and the Hierarchical Multi-Scale Framework. The results of this analysis are presented in Figure 6.

The basic predictor performs slightly better at smaller thresholds (e.g., 1 km: 9.3% vs. 8.9%), likely because its single-resolution design focuses on localized features without interference from broader-scale integration when there are conflicts between predictions across scales (e.g., when coarse and fine scales produce divergent results). However, as the thresholds increase (200 km and beyond), the hierarchical multi-scale framework outperforms the basic predictor by a huge amount, achieving 87.4% accuracy at 2500 km compared to 65.8%. This demonstrates that the hierarchical framework is able to integrate features from multiple spatial scales, enabling better generalization and capturing both fine-grained and global patterns. On the other hand, the basic predictor struggles at larger scales due to its limited focus on a single resolution, while the hierarchical approach excels in diverse and complex geolocation scenarios.

We also analyzed the geographical error distribution of our model using the im2gps and im2gps3k datasets. The im2gps3k dataset is more diverse, covering a broader range of geographical regions compared to the im2gps dataset, which primarily focuses on urban areas. For example, while im2gps includes approximately 1,000 images predominantly from North America and Europe, im2gps3k extends to 3,000 samples, adding significant representation from Asia, Africa, and South America. This increased diversity provides a more challenging and comprehensive test for our model.

The results are visualized in the maps below(Figure 7.& Figure 8), where the color intensity and size of the markers correspond to the error distance in kilometers. Redder and larger markers indicate higher prediction errors. From the im2gps dataset shown as Figure 7. We observed that the majority of the predictions had relatively small error distances, concentrated in regions with high-density data points such as North America and Europe. However, there are some significant outliers, particularly in areas where the ground-truth locations are underrepresented in the training data. For example, certain regions in Africa and South America show higher errors due to the lack of similar visual features in the dataset.

When evaluating on the larger im2gps3k dataset shown as Figure 8, we noticed a similar pattern in error distribution. Regions with higher dataset density, such as urban areas and tourist hotspots, tend to have lower error distances. In contrast, sparsely sampled regions exhibit larger errors. The extended dataset, however, allows the model to generalize better in some underrepresented areas, reducing error magnitudes compared to the smaller dataset.

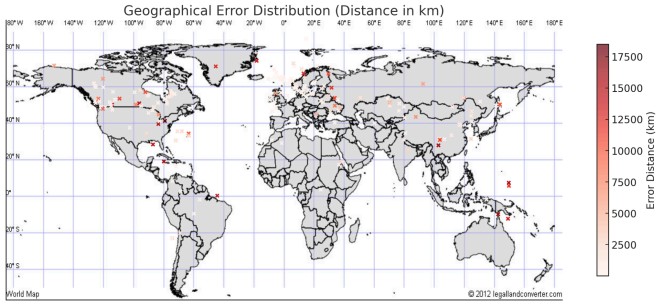

Figure 7. Geographical Error Distribution in im2gps dataset

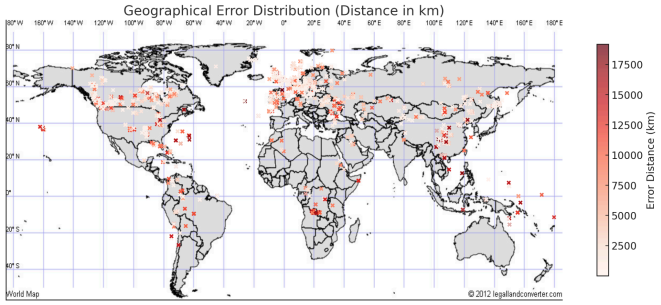

Figure 8. Geographical Error Distribution in im2gps3k dataset

To further evaluate our model's performance, we analyzed the best and worst prediction results in both the im2gps and im2gps3k datasets. Below are the observations based on distance error calculations for specific examples.

In both datasets, the top five best predictions demonstrate that our model can achieve highly accurate geolocation estimates, with errors as low as a few meters. In the im2gps dataset(Figure 9), the prediction for an image in Paris (Paris_00131) resulted in a negligible distance error of 0.0358 km, showcasing the model's capability to recognize prominent landmarks.Similarly, in the im2gps3k dataset(Figure 11), the best predictions have minimal errors, often under 1 km, which aligns well with urban environments and well-documented regions.

The worst predictions highlight the limitations of our approach.For the im2gps dataset(Figure 10), images from Ecuador (Ecuador_00016) and other underrepresented regions exhibited significant errors exceeding 15,000 km, suggesting difficulty in identifying less-distinctive visual features.In the im2gps3k dataset(Figure 12), the largest errors were similarly pronounced, with outliers such as 322234321 showing a distance error of over 19,000 km. These failures often occurred in regions with ambiguous visual cues or low representation in the training set.

img_id	Real Lat	Real Lon	Pred Lat	Pred Lon	Distance error(Km)
Paris_00131	48.86313	2.336654	48.8633	2.336193	0.038524344
Spain_00083	37.38902	-5.99407	37.38831	-5.99432	0.082086451
Paris_00005	48.85372	2.347738	48.85245	2.348827	0.162082434
429777514	41.40349	2.174423	41.40487	2.173752	0.163132996
97344248	41.39535	2.169113	41.39423	2.167114	0.208107603

Figure 9. Best 5 result as distance in im2gps dataset

img_id	Real Lat	Real Lon	Pred Lat	Pred Lon	Distance error(Km)
779600060	-34.9403	138.5856	42.35726	-71.0666	17327.06
Ecuador_00016	-0.96744	-77.764	-13.1698	130.7576	16499.29
Bangkok_00011	9.376241	99.94091	-23.1482	-44.5663	15943.9
453214262	35.52143	-111.373	-26.2885	115.5978	15461.5
522546192	-14.6667	145.4485	28.09119	-77.606	15342.06

Figure 10. Worst 5 result as distance in im2gps dataset

img_id	Real Lat	Real Lon	Pred Lat	Pred Lon	Distance error(Km)
230313836	40.74872	-73.9861	40.74852	-73.986	0.022476
285425292	45.43966	12.32597	45.43991	12.3261	0.0295
872673086	40.75783	-73.9857	40.75753	-73.986	0.041704
297480819	51.50747	-0.12795	51.50728	-0.12726	0.05277
266531088	37.80855	-122.41	37.80907	-122.41	0.058039

Figure 11. Best 5 result as distance in im2gps3k dataset

img_id	Real Lat	Real Lon	Pred Lat	Pred Lon	Distance error(Km)
322343421	37.76312	-25.3413	-42.3933	147.0788	19190.44
447024764	-12.6119	-69.0372	5.409659	116.7332	18994.11
304304112	51.47061	-0.2123	-43.408	172.7573	18975.37
123607970	1.286665	103.8387	8.990037	-79.7476	18805.26
377436986	13.44364	99.98322	-17.4901	-69.4843	18800.34

Figure 12. Worst 5 result as distance in im2gps3k dataset

The error distribution in the im2gps dataset(Figure 13) shows a median error below 1,500 km, but the mean error (2,122.52 km) indicates the presence of significant outliers.

The first quartile (Q1) error of 9.44 km and third quartile (Q3) error of 1,484.94 km demonstrate that most predictions are within reasonable bounds, with the majority of errors clustering below 2,500 km.

The mean error of 3,117.00 km of the error distribution in the im2gps3k dataset(Figure 14) is higher than that of the im2gps dataset, reflecting the increased diversity and complexity of the larger dataset.

The Q1 and Q3 range (28.25 km to 3,542.57 km) indicates a broader spread in error distances, with the model struggling more in sparsely populated or visually ambiguous regions. This indicates that our model performs well in the majority dataset; however, some extreme test data become a burden in the average results. We will dive into data that don't perform well in the next part's discussion.

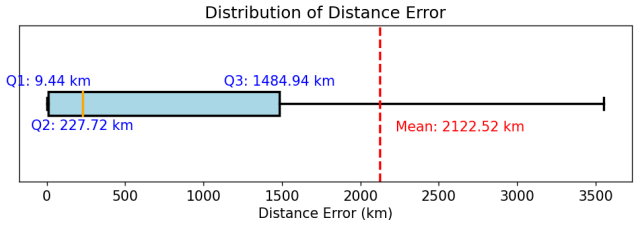

Figure 13. Distribution of Distance Error in im2gps dataset

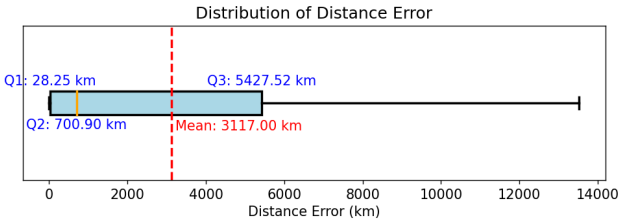

Figure 14. Distribution of Distance Error in im2gps3k dataset

4. Discussion and conclusions:

4.1 Main Insights and Observations

Accurate Predictions in Urban and Iconic Regions: Urban landmarks or areas with well-recognized features(e.g., North America and Western Europe) perform better, while regions with less frequent representation in the dataset (e.g., Africa, South America, and the Pacific Islands) consistently produce larger errors. This reflects a bias in the training data distribution and highlights the need for more balanced global datasets.

Extreme Mispredictions:

The largest errors exceeded 15,000 km, where predictions landed in completely opposite areas. For example: an image in South America was mistaken for a location in Australia. These extreme mispredictions suggest that certain images lack sufficient identifiable geographic features, leading the model to perform poorly.

Patterns in the Errors:
We observed that errors are often influenced by longitude mismatches, with predictions occasionally jumping across continents or oceans. Generic natural landscapes like forests, beaches, and mountains are more challenging, since they may look similar to many other parts of the world. Similar weather, plants, and landscape across different countries can also mislead the model; for instance, certain plants in Alaska and siberia in Russia can appear quite alike, causing confusion. For example, in Fig15, our model can specify the image is in high latitude due to the snow and guess the latitude is 41.142337, with 44.16537094 as the correct answer. However, our model is struggling in guessing which guess is -73.236408, while the real location is -110.9542847. We can find out that most of the error comes from the mismatch of longitude.

Fig15. image 425947697_ecb480e925_157_15376845@N00

Impact of Dataset Scale and Coverage:
The im2gps3k dataset demonstrates a higher mean error (3,117 km) compared to the im2gps dataset (2,122 km). This indicates that as datasets scale and include more diverse and challenging examples, the model struggles to generalize effectively.

Actual Accuracy:
Just as quartile information shown in Fig13, we got only 200 km distance error in median while 2122 km mean distance error. We can say that the mean result was encumbered by the extreme misprediction which seldom contained information in the picture as shown in Figure 16.

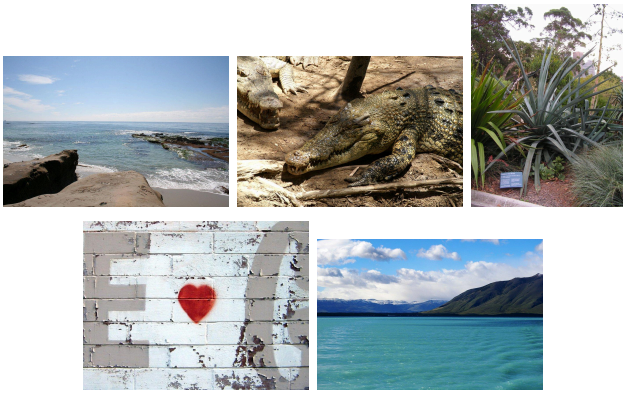

Figure 16 . The worst 5 results as distance.

4.2 Potential Solutions

Improving Model Generalization:
Train the model on a more geographically balanced dataset to address regional biases. Include more images from underrepresented areas like Africa, South America, and remote Pacific Islands. Apply ensemble models that combine predictions from multiple networks trained on different scales or features, improving accuracy and consistency in challenging cases.

Handling Ambiguous Images:
We can enhance geographic predictions for natural landscapes by integrating climate-based features such as vegetation indices (NDVI, EVI), elevation data, and climate zones into our model. These features can be extracted from datasets like MODIS and SRTM and represented as additional vectors. By combining them with visual features from a CNN in a multi-input neural network, we can create a more context-aware framework. Additionally, we can develop specialized modules for unique patterns like coastlines or deserts. This approach can improve predictions in visually similar or underrepresented regions, making the model more accurate and scalable across diverse landscapes.

Addressing Extreme Errors:
Mark out predictions that are geographically unlikely. For example, large jumps across hemispheres could trigger extra validation steps. Implement confidence scoring: Have the model output a confidence score for each prediction, allowing uncertain predictions to be handled differently.

5. Connection to course material:

Our projects builds on several foundational concepts covered in the course, which directly ties to the lectures and techniques as:

Preprocessing:
At the beginning ,we need to process it in a way that makes its features easier to analyze. This involves steps like sampling, filtering, and edge detection. Which directly relates to the lectures on September 4 and 6 (Image Filtering), and September 18 (Edge Detection) and helps us organize it into a hierarchical geo-classification structure which we've learned in this lecture. Moreover, skills like Fourier transforms(Fourier analysis in Sept. 6 & 11) and interpolation (Image processing in Aug. 30) help us ensure images with different sizes and resolutions can be compared correctly.

Feature Extraction:
After preprocessing, we look for unique features that can guide us toward an image's geographical location. This is where methods like corner detection(Corner detection Sept. 20) and SIFT(SIFT keypoint detection in Sept. 25 & SIFT in Assignment 3) which were taught in class can help. SIFT allows us to find and match distinctive landmarks across images and solve problems of multi-resolution inputs and inputs of different scales.

Spatial Relationships:
Once we've extracted the features, we need to understand how they relate to one another in space. Accurately placing a new image into a set of known geographical regions involves applying spatial transformations. Techniques like affine transformations(Alignment Oct. 9) and image warping help us align images correctly and refine the overall hierarchy of classifications..

6. Statement of individual contribution:

Sophie Huang: contributed to dataset collection, implemented adaptive region partitioning, integrated scene-aware context classification, and collaborated on developing Context-Specific Networks.

Kathy Lo: worked on dataset collection and preprocessing, implemented the multi-scale learning framework, collaborated on Context-Specific Networks, and assisted with debugging

Chia-Chun Hsiao: focused on hierarchical geolocation prediction and evaluation. She contributed to scene classification integration and provided significant input for the report's methodology sections.

Yu-Han Huang: managed experimental protocols, including testing, performance evaluation, and result visualization. He also contributed to integrating scene-aware context classification.

Kai-Po Chang: supported model testing and evaluation, analyzed quantitative results, and assisted with debugging. He also contributed to the multi-scale learning framework and the final hierarchical geolocation prediction step.

7. References

[1] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. [Online]. Available: https://arxiv.org/abs/1704.04861

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[3] *GeoSpottr: Image-Based Geolocalization*. GitHub Repository. [Online]. Available: https://github.com/pkardjian/GeoSpottr

[4] Hays, J., & Efros, A. A. *"IM2GPS: Estimating geographic information from a single image."* Carnegie Mellon Graphics, 2008. [Online]. Available: http://graphics.cs.cmu.edu/projects/im2gps/

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517-6525. [Online]. Available: https://ieeexplore.ieee.org/document/8100173

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[8] *Outperforming Humans in GeoGuessr with Deep Learning - DLCV Project*. GitHub Repository. [Online]. Available: https://github.com/valdrin-uni/DLCV_Project_GeoGuessr_AI/tree/main

[9] Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., & Li, L.-J. "The New Data and the New Challenges for Multimedia Research." *Proceedings of ACM Multimedia*, 2016. [Online]. Available: https://arxiv.org/abs/1503.01817

[10] Vo, N., Jacobs, N., & Hays, J. *"Revisiting IM2GPS in the Deep Learning Era."* arXiv preprint arXiv:1705.04838, 2017. [Online]. Available: https://arxiv.org/abs/1705.04838

[11] Weyand, T., Kostrikov, I., & Philbin, J. *"PlaNet – Photo Geolocation with Convolutional Neural Networks."* arXiv preprint arXiv:1602.05314, 2016. [Online]. Available: https://arxiv.org/abs/1602.05314

[12] Zhang, W., Kosecka, J.: Image based localization in urban environments. In: 3DPVT 2006: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT 2006), pp. 33–40 (2006)

[13] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. "Places: A 10 Million Image Database for Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. [Online]. Available: http://places2.csail.mit.edu/PAMI_places.pdf