Towards Accurate Image Geolocalization: A Study of Novel Computational Approaches

by

Md. Mehedi Hasan Tanvir
22101107
Md. Ashiqur Rahman Abir
22101650
Tasfia Tasnim Prioty
23241101
Arnab Anthony Gomes
22101351
Sadia Tasnim Fariha
21301294

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Mehedi Hasan Tanvir

22101107

Md. Ashiqur Rahman Abir

22101650

Tasfia Tasnim Prioty

23241101

Arnab Anthony Gomes

22101351

Sadia Tasnim Fariha

21301294

Approval

The thesis titled "Towards Accurate Image Geolocalization: A Study of Novel Computational Approaches" submitted by

1. Md. Mehedi Hasan Tanvir (22101107)
2. Md. Ashiqur Rahman Abir (22101650)
3. Tasfia Tasnim Prioty (23241101)
4. Arnab Anthony Gomes (22101351)
5. Sadia Tasnim Fariha (21301294)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 10, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Image geolocalization, the task of determining the geographic origin of a given image, remains a formidable challenge due to the immense variability of global landscapes and the subtle visual cues that indicate specific locations. This research aims to introduce a novel approach that seeks to surpass the accuracy of current state-of-the-art models. By leveraging an highly diverse dataset, integrating cutting-edge vision transformer architectures, optimizing the training process, and systematic review and fine-tuning, this approach achieves significantly improved performance in image geolocalization. Our model demonstrates an exceptional capacity to generalize to previously unseen locations, even under complex conditions such as varied lighting, diverse weather patterns, and other environmental challenges, marking an advancement over prior methodologies in this domain.

Keywords: Image Geolocalization, Computer Vision, Deep Learning, Vision Transformers, Contrastive Learning, Multi-Task Learning, Semantic Geocells, Geocell Partitioning, Geocell Classification, Haversine Smoothing Loss, Haversine Distance, Transformer Fine-Tuning, Geo-Tagged Datasets, Geospatial AI, Geographic Coordinate Prediction, Spatial Data Analysis, Street View Analysis, Visual Place Recognition, Machine Learning

Acknowledgement

We would like to thank everyone who has helped us in the course of completing our thesis.

We would like to express our greatest gratitude to our supervisor, Dr. Md. Ashraful Alam, Associate Professor, Department of Computer Science and Engineering, BRAC University. His priceless advice, constant support, and valuable feedback have played a significant role in the formation of this research and the process of passing all the steps of the study.

We also owe the Department of Computer Science and Engineering, BRAC University, the opportunity, resources, and academic environment, without which this research could not have been successfully implemented.

We also want to mention that Mapillary provided us with the open-access geotagged imagery, which was used in the current research and helped us create and test our proposed model. We also would like to thank the OpenAI CLIP framework, the publicly available pretrained models of which formed the basis of our multimodal learning strategy.

Lastly, we would like to thank our families, peers, and friends who have been supporting us and encouraging us, unceasingly, in this endeavour.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Image geolocalization is a crucial field in computer vision with applications in autonomous navigation, security, environmental monitoring, and disaster response. Identifying the geographic origin of an image has always been difficult due to the vast diversity in landscapes, lighting conditions, and environmental factors. Earlier methods relied on hand-crafted features and retrieval-based techniques, which struggled to generalize to unfamiliar locations. The usage of deep learning, particularly Convolutional Neural Networks (CNNs) and Vision Transformers, has greatly improved geolocation accuracy by utilizing large-scale datasets and hierarchical classification techniques. However, challenges such as generalization, dataset bias, and computational efficiency remain significant, highlighting the need for continued research in this area.

1.2 Rationale of the Study or Motivation

Many applications, such as forensic investigations, disaster relief, and navigation support, depend on the ability to precisely geolocate an image. However, current models frequently have trouble generalizing outside of the particular datasets they were trained on, which restricts their use in real-world situations. This study seeks to close this gap by presenting novel computational techniques that improve model generalization and prediction accuracy. Resolving these issues will advance machine learning and artificial intelligence applications in addition to enhancing geolocation capabilities.

1.3 Problem Statement

Identifying geolocation of images itself is challenging but what's more challenging is accurately predicting geolocation of images in diverse and previously unseen environments. There are existing models that take on the challenge, such as CNN based classifiers and retrieval based systems, but these often demonstrate poor performance in predicting new environments. Furthermore, there is a lack of standardized and diverse training datasets, and these factors worsen the issue and lead to biases. This research focuses on overcoming these limitations by integrating advanced

techniques such as contrastive learning, multi-task training, and semantic geocell partitioning.

1.4 Objective

The primary objectives of this study are:

- To develop a geolocalization model that generalizes well to unseen locations.
- To incorporate multi-task learning and contrastive pretraining to enhance accuracy.
- To introduce a novel geocell partitioning technique that optimizes spatial classification.
- To evaluate the proposed model against existing state-of-the-art methods.

1.5 Methodology in Brief

This study aims to employ a multi-faceted approach:

- **Data Collection:** Compilation of large-scale geotagged image datasets from sources such as Flickr, Wikipedia, and Google Street View.
- **Model Development:** Implementation of a vision transformer-based model with contrastive pretraining.
- **Training and Evaluation:** Utilizing hierarchical classification and geocell partitioning for refined predictions.
- **Performance Benchmarking:** Comparing the proposed model with existing approaches using standard geolocalization benchmarks.

1.6 Scopes and Challenges

Scope:

- This research aims to focus on identifying locations from images taken anywhere in the world.
- It studies transformer-based architectures and contrastive learning methods.
- The model will be tested on different datasets to ensure accurate evaluation.

Challenges:

- Training deep learning models requires a lot of computing power.
- There are not enough high-quality, geo-tagged images for all regions.
- The model needs to work well in different environments and lighting conditions.

1.7 Team Overview

This research is done by five undergraduate students who specialize in computer vision and artificial intelligence. Each team member works on different tasks like processing data, training the model, testing its performance, and writing the reports. Their teamwork helps in solving the geolocalization problem in a well-rounded way.

1.8 Key Learnings and Insights

Throughout this research, we anticipate gaining valuable insights into:

- How well contrastive learning helps the model work in different environments.
- How dividing locations into smaller regions improves geolocation accuracy.
- The balance between accuracy and speed in transformer-based models.
- Possible real-world uses and ethical concerns of image geolocation technology.

Chapter 2

Literature Review

Berton et al. (2022) [15] investigated Visual Geo-localization (VG), a method that uses a vast library of recognized images to detect the location of a photo. By creating a more effective training method that scales well for citywide applications, this research aims to enhance large-scale visual geo-localization. The researchers present a new large-scale dataset called San Francisco eXtra Large (SF-XL), which is thirty times larger than the datasets that were previously accessible. They discover that current VG methods have trouble efficiently scaling to such large datasets. Instead of employing conventional contrastive learning, they suggest a novel training method called CosPlace, which frames VG as a classification problem. This approach reduces GPU memory usage by 80 percent during training and produces more compact image descriptors (8x smaller) while achieving state-of-the-art performance. Their work enables city-wide real-world VG applications. There are also some research gaps:

1. CosPlace relies on heading labels, making it inapplicable to datasets that lack them.
2. It is not optimized for small datasets, limiting its use for localized VG applications.
3. The method has not been tested extensively beyond urban environments like San Francisco.

Luo et al. (2022) [16] introduced Geolocation by Guidebook Grounding (G^3), a novel method that uses both image-based geolocation models and human-written guidebooks to predict the country in which an image was taken. In order to classify countries more accurately, this project aims to enhance image-based geolocation by integrating human-written guidebook information. In addition to textual indications from GeoGuessr guidebooks that include descriptive cues about roads, landmarks, and architectural styles, the authors compile a dataset of Google StreetView photos from 90 different nations. The suggested G^3 model matches textual hints with images by combining image-based features with handbook text and employing attention-based learning. This method significantly improves geolocation accuracy, outperforming traditional image-only models by more than 5 percent in Top-1 accuracy. But there are some gaps also. The gaps are:

1. Limited geographic coverage (only 90 countries)
2. Dependence on guidebook texts, which might lack details for some regions.
3. Challenges with ambiguous images, where even human-written clues may not be sufficient for precise geolocation.

Suresh et al. (2018) [8] developed DeepGeo, a deep learning-based system for photo localization in the U.S. The objective is to use Google Street View photos to anticipate the state in which a photograph was taken. With 500,000 photos spanning all 50 U.S. states, the authors produce a brand-new dataset named 50 States10K. The ResNet-50 model is used in the construction of DeepGeo, which investigates the Early, Medium, and Late integration algorithms for merging several image views captured at the same spot. The best approach, Medium Integration, achieves a 71.87 percent accuracy in Top-5 predictions and 38.32 percent in Top-1 accuracy, significantly outperforming random guessing (2 percent) and competing methods. The model also defeated human players in 4 out of 5 rounds of GeoGuessr.
Limitation or gaps:

1. Limited to U.S. states – The model is not trained for global geolocation.
2. Bias towards population-dense areas – Due to dataset collection strategy, the model struggles in sparsely populated regions.
3. Classification-based approach – The model predicts states, but a regression-based approach (latitude/longitude prediction) could be more precise.

Yang et al. (2021) [14] introduced L2LTR (Layer-to-Layer Transformer), for cross-view geo-localization, which matches a street-view image with a geo-tagged aerial image to estimate the position of the street-view image. The goal of the study was to increase the precision of aerial and street-view image matching. Significant perspective changes between street and aerial photos are a challenge for conventional CNN-based methods. Rather, L2LTR models global dependencies using Transformer-based self-attention, which increases its resilience to visual ambiguity. Better feature representation and more stable training result from the model's introduction of a self-cross attention mechanism, which enhances information flow between Transformer layers. To enhance feature representation, they paired learnable positional embedding with a self-cross attention mechanism in a Vision Transformer (ViT) architecture. The results show that L2LTR significantly outperforms state-of-the-art models in standard, fine-grained, and cross-dataset geo-localization tasks.
Gaps:

1. High GPU memory demand – L2LTR requires significant computing power.
2. Limited training data – Transformer models need large datasets, which may not always be available.
3. Pre-trained dependency – L2LTR is built on pre-trained Transformers, making it data-intensive.

Haas et al. (2024) [26] introduced PIGEON, a geolocalization system that combines semantic geocell creation, multi-task contrastive pretraining, and a novel hierarchical retrieval method over location clusters. The study presents two models: PIGEON, trained on data from the GeoGuessr game, and PIGEOTTO, which was trained on a diverse dataset including Flickr and Wikipedia images. PIGEON achieved remarkable accuracy, with over 40% of its predictions falling within 25 km of the correct location. It was tested against human players in a blind experiment and ranked among the top 0.01% of players, even defeating one of the world's foremost GeoGuessr professionals in multiple rounds. PIGEOTTO, on the other hand, outperformed state-of-the-art benchmarks by up to 38.8 percentage points on the country-level accuracy and 7.7 percentage points on the city-level accuracy. This research marks a major step forward in developing robust, planet-scale image geolo-

calization models, demonstrating improved generalization to unseen locations under varied conditions such as lighting changes, diverse landscapes, and different architectural styles. However, future work is required to enhance its adaptability to extreme weather conditions and highly ambiguous locations lacking clear visual cues.

Traditional methods of geo-localisation often rely on classification-based approaches, which divide the world into discrete geographic regions. This often results in inaccuracy and generalization. To tackle this problem, GeoCLIP (Cepeda et al., 2023)[20] brings forth an alternative image-to-GPS retrieval method, where images are directly matched with their corresponding GPS coordinates rather than classified into predefined location categories.The model leverages contrastive learning to align image features with geospatial data. The GeoCLIP utilises a novel location encoder that integrates Equal Earth Projection (EEP) to mitigate distortions in latitude/longitude systems and Random Fourier Features (RFF) which encodes GPS features into a high-dimensional feature space, capturing detailed location information. It employs a CLIP-based Vision Transformer (ViT) as the image encoder which extracts meaning features from images, which is then projected into the same embedding space as the GPS features. Additionally, a hierarchical learning strategy enables the system to capture both fine-grained and large-scale geographic details. GeoCLIP demonstrates high accuracy on benchmark datasets such as Im2GPS3k and Google World Streets 15K, outperforming state-of-the-art models like the PlaNet, ISNs and some others. While this model achieved high accuracy even with limited data settings and demonstrated data efficiency, these benchmark datasets are biased towards urban and well-photographed areas. It is unclear how well the model performs in rural or remote regions. The model is also not designed to handle images that are taken under extreme weather conditions. WhileGeoCLIP also introduces text-based geolocalization, allowing textual descriptions to be mapped to GPS coordinates, there is no quantitative result on how accurately this model can retrieve the location from textual queries.

We see widespread utilisation of Large Vision Language Models (LVLMs) in visual geo-localisation. However, there are privacy risks associated with these models' ability to infer locations from images without explicit geographic training. In their research, Liu et al. (2024)[28] addressed the resulting accuracy limitations and benchmarked existing LVLM-based and traditional geolocation models. The authors then introduced ETHAN, a framework that enhances LVLMs' reasoning capabilities. The framework uses a Chain-of-Thought reasoning approach which mimics human geolocation reasoning, improving geolocation accuracy. ETHAN analyzes key environmental, architectural and cultural features and is fed with a dataset of 50,000 images. The framework is tested on geolocation datasets and benchmarked against existing models like GeoSpy, StreetClip, GeoClip, GPT-4o and LLaVA, where it showcased higher accuracy, with 27% of predictions within 1 km (compared to 24.5% for GeoSpy) and an average GeoScore of 4600, surpassing competing models. ETHAN was also tested on the GeoGuessr game against humans, where it had an 85.37% win rate. ETHAN also analysed failure cases consisting of inaccurate geolocation prediction and achieved precise predictions, with some accurate within 0.3 km. The framework has certain limitations, such as performing poorly in low-visibility conditions and with featureless landscapes (like deserts, open fields). It also struggles in homogeneous urban areas that have repetitive architectural buildings

or areas that are rapidly changing.

Astruc et al. (2024) [23] introduced OpenStreetView-5M (OSV-5M), a large-scale, open-access dataset for visual geolocation, containing 5.1 million geo-referenced street-view images from 225 countries. To ensure dataset quality and diversity, images were sourced from Mapillary, filtered for localizability and clarity, and split into training and test sets with a strict 1km spatial separation to prevent memorization-based learning. Various state-of-the-art deep learning models were benchmarked, including ViT, ResNet, and StreetCLIP encoders, with different pre-training datasets (CLIP, LAION-2B, OpenAI). Three geolocation prediction methods were tested: regression (direct coordinate prediction), classification (country, region, area, city labels), and hybrid models (combining both approaches). The study evaluated models based on Haversine distance (error in km), Geoscore (accuracy reward system), and classification accuracy. Results show that OSV-5M significantly improves geolocation performance over existing datasets. The best-performing model—a ViT-L-14 encoder pre-trained on DATA COMP combined with a QuadTree-based hybrid prediction approach and contrastive learning—achieved a 1309-point increase in Geoscore, a 45% reduction in average error distance, and higher classification accuracy across multiple administrative levels. Human evaluators were outperformed by AI models in geolocation tasks. However, error analysis revealed biases, with underrepresented regions (e.g., South America, Africa) showing higher prediction errors, and some non-Western images misclassified as Western locations. The study highlights the need for more diverse datasets, improved long-term generalization, and better handling of non-localizable images. OSV-5M establishes a new benchmark for evaluating deep learning models in global visual geolocation while identifying key areas for future research in reducing bias, increasing dataset coverage, and enhancing model robustness.

Dufour et al. (2024) [25] proposed a new method for predicting an image's geographic location using a generative approach. Unlike conventional models that provide a single predicted location, this method accounts for spatial uncertainty by generating a probability distribution over possible locations. The approach utilizes diffusion models and flow matching techniques to refine predictions iteratively. A key innovation is Riemannian Flow Matching, which operates on the Earth's spherical geometry, improving geographic coordinate accuracy. The model is tested on three large-scale datasets—OpenStreetView-5M, iNat21, and YFCC-100M—where it achieves state-of-the-art performance, surpassing traditional classification, regression, and retrieval-based techniques. A notable contribution of this research is the introduction of probabilistic visual geolocation, allowing the model to capture uncertainty in its predictions. The study also establishes new evaluation metrics and baselines, demonstrating the model's ability to differentiate between highly identifiable locations (such as famous landmarks) and more ambiguous scenes (like generic landscapes). Despite its advancements, the method faces challenges, including high computational demands, difficulty generalizing to unseen locations, and the interpretability of probability distributions. The findings suggest potential for future improvements, particularly through hybrid models that integrate generative techniques with retrieval-based methods to enhance geolocation accuracy.

Pramanick et al. [17] in their paper aimed to identify a location based on only

a single picture taken anywhere in the world. This was very challenging, as the same location can look very different depending on the factors like time, season, weather, angle of the photo etc. Previous works were limited to a very specific range for this reason. This paper aims to break that challenge by locating any place of the world from a single image. As a method, this paper used a neural network model named 'TransLocator', as it can help to capture detailed information from an image. It works in two parallel branches- 1) RGB Image Branch: it process the original captured image 2) Semantic Segmentation Branch: It segments the image and labels it according to the object it represents. This model also can do 2 tasks together-It can identify the location and also It can identify the type of environment(indoor, outdoor, urban, rural). It helps to narrow down the search of prediction.The paper evaluates the model by measuring continent level accuracy, meaning this evaluation focuses on how often a model can correctly identify the continent, rather than the exact pinpoint location. This model improves continent level accuracy significantly, on both smaller and also diverse, larger datasets, than the previous models. But, since the model is using continent level accuracy, how good it can predict an exact location of a picture, remains a doubt. Also, extreme variation can affect the performance of the model, such as abnormal weather patterns and rare architecture style. Moreover, further research should be done to ensure the model can identify the location which was not introduced in the training dataset.

Though nowadays, CNNs (Convolutional Neural Networks) is working very efficiently for photo geolocation, but we can not properly interpret how the model is making their decision. Theiner et al. [13] in their paper aimed to follow a method, where there is proper transparency of the model, thus making the model more interpretable. It means we will know how the model works, without compromising their performance Semantic Partition Method- In this method, they divide the earth into more meaningful divisions such as - rivers, roads and cities etc instead of dividing it into grids or zones. They are using an Openstreet map to define many geographic partitions. Then a CNN model is created to classify images into semantic partitions. They also use concept influence metric, which evaluates the effect of each feature in model performance. It achieved higher accuracy results than many state of the art models that use traditional partitioning methods.The author suggested that future research could try to integrate other semantic information to enhance the performance of the model.Using concept influence metric in other domains also can increase the performance of the model.

The current CNN based approach does not work well on connecting the relation between ground level image and aerial images because they have very different perspectives most of the time. Zhu et al. [18] proposed the first pure transformer based approach, which can overcome this difficulty, as the transformer is great at grasping relations between all parts of an image. As a methodology, the researchers train two separate transformers for ground level aerial images and set separate encoders to process both images. They also use attend and zoom in strategy which means the model focuses only on important parts of an image, after removing the irrelevant part from the view. TransGeo delivers better results in both urban and rural areas while also requiring less resources than the traditional CNN based method. It lowers the cost and memory usage as well, making its performance a state of art. As this system is efficient at saving time, this approach can be used for implement-

ing in real world applications by improving it more. Also though this model can work on different viewpoints very well, extreme perspective difference still remains a challenge.

Generally, performance of a geolocalization model highly depends on the availability of large datasets. Also, it often ignores crucial details which could come very helpful to identify location, such as text or road sign. Li et al. [27] in this research aims to overcome such challenges by integrating human reasoning capabilities to artificial intelligence, and also by using the visual or textual detail to identify a place. This paper used a Large Vision-Language Model to know about the subtle visual and textual cues from an image such as architecture or nameplate of a car. And to integrate human reasoning, the model copies the strategy of games like geoguessr, where a person tries to guess a location based on various logics and clues. Similarly, the model GeoReasoner was trained to use those same reasoning processes to make a decision, increasing the robustness of the model more. As a result, it works better, outperforms both in country level and city level predictions, by 25% and 38% respectively while requiring less resources than existing models. Future researchers could explore how to improve the performance of the model in context of areas with relatively lower visual and textual clues. Moreover,lower quality images can vastly affect the performance of the model, as this model is highly dependent on the quality of the picture. It can cause a problem in real world scenarios as maximum data can be quite noisy. Sometimes the reasoning capabilities also can backfire by ignoring or oversimplifying important details, just like humans do sometimes. These problems should be discussed more in the future to make this model more robust and powerful for real world scenarios.

Saoud et al. [31] in this paper wants to determine a location of a place from an image without relying on GPS data by using various clues from the image. This research is uniquely useful, as GPS can not guarantee a consistent service, meaning it might not always be available in all areas. In those times particularly, this research will be proven to be handful. As methodology, the researchers used SIFT (Scale-Invariant Feature Transform) for identifying place, regular image processing for identifying road junctions, and deep learning for classification. The SIFT is being used for introducing the key features from the image to the system which allows the system to match those features to its database. The regular image processing is needed to analyze the image more precisely to see if there is any road junction; which is very helpful for identifying a location. And lastly, a deep learning model is being used to classify those identified road junctions from image processing earlier, to determine the location. The reasons that this research is so unique is because it was implemented into an offline mobile app, resulting in increasing its usability more to the users, and also the fact that it does not depend on GPS data. Stable internet connection is not always possible in challenging remote areas. This system would be of great help in those cases.But also, the authors suggested the algorithm of this system can be more refined and this model should be experimented on a vast dataset to make it more robust.

A study conducted by Lu et al. (2024) focuses on Visual Place Recognition (VPR), which helps identify a location by finding a matching geo-tagged image. This is difficult because lighting, camera angles, and moving objects can change how a

place looks. Traditional methods use either global feature retrieval or a two-stage approach that refines matches with local features. While pre-trained vision models perform well in many tasks, they don't work as effectively for VPR because they also focus on moving objects like people and cars, which are not useful for place recognition. This paper introduces SelaVPR, a method that adapts pre-trained models for VPR by using both global and local adaptation (Lu et al., 2024). Global adaptation adds small adjustments to help the model focus on buildings and landmarks rather than moving objects. Local adaptation enhances detailed image features, making it easier to find better matches quickly. The model also replaces traditional geometric verification (like RANSAC) with Mutual Nearest Neighbor Loss, which makes re-ranking much faster. The model was tested on three well-known datasets (Tokyo24/7, MSLS, Pitts30k) and ranked 1st on the MSLS challenge leaderboard, performing better than existing VPR methods like NetVLAD, Patch-NetVLAD, and TransVPR (Lu et al., 2024)[29]. The results show higher accuracy, with better Recall@1, Recall@5, and Recall@10 scores, lower computational cost, needing only 3% of the retrieval time compared to RANSAC-based methods, and better generalization, working well in different environments, including urban, suburban, and nighttime settings (Lu et al., 2024)[29]. However, the model still struggles with extreme lighting and viewpoint changes. Also, since it generalizes well, it is unclear how well it can identify an exact location. Future work should focus on improving adaptability and exploring transformer-based models for real-time, large-scale VPR (Lu et al., 2024)[29].

One of the first large-scale deep learning frameworks to tackle the problem of global image geolocation, PlaNet by Weyand et al. re-formulated the problem as a classification problem, instead of coordinate regression. The authors partitioned the surface of the Earth into thousands of discrete geographic cells of different sizes, which trade off spatial resolution and data density. A convolutional neural network that was trained on millions of geotagged images was able to predict the most likely location of a cell given the input image, and did much better than the traditional feature-engineered systems(Weyand et al., 2016) [3] . In contrast to the earlier approaches which relied on the use of hand-crafted descriptors or metadata, PlaNet used pure visual features to predict location and showed that convolutional features had the potential to learn geographically discriminative features, including vegetation types, architectural styles and terrain forms. Probabilistic reasoning was also introduced in the work to deal with spatial uncertainty, which enabled prediction with confidence weights in multiple regions. PlaNet was a seminal step in visual geolocation and inspired many of its successors who used grid-based or geocell approaches (Weyand et al., 2016) [3]. It uses a global classification strategy and cell-based modeling to directly teach our geographic-aware geocell classification network that adapts this concept to transformer and multimodal architectures.

Arandjelović et al. (2016) [2] introduced NetVLAD, a convolutional neural network architecture specifically designed for large-scale visual place recognition tasks. The core innovation is the NetVLAD layer, a differentiable, trainable generalization of the traditional VLAD (Vector of Locally Aggregated Descriptors) pooling method, which aggregates mid-level CNN features into a compact image representation. The network is trained end-to-end using a weakly supervised ranking loss and a large dataset of images from Google Street View Time Machine, enabling robustness to

changes in viewpoint, lighting, and occlusion. Their experiments demonstrate that NetVLAD-trained representations significantly outperform both hand-engineered descriptors and off-the-shelf CNN features on challenging place recognition and image retrieval benchmarks (Arandjelović et al., 2016) [2]. This work is highly relevant to our study as it establishes the effectiveness of task-specific, end-to-end learned representations for geolocalization, and provides a strong baseline for comparing novel geolocalization approaches.

Shi et al. (2020) [9] suggested a new method of cross-view geo-localization that simultaneously estimates the position and the orientation (azimuth angle) of a ground-level camera by comparing it with a database of geo-tagged aerial images. Their approach presents a Dynamic Similarity Matching (DSM) module that uses a polar transform to match aerial images with ground views to minimize the geometric domain gap and allow more precise feature matching. A two-stream CNN is used to extract spatially-sensitive features on both ground and polar-transformed aerial images, and the DSM module is used to compute feature similarity over potential orientations, which makes it possible to localize robustly even when the ground image has a small field of view or when the orientation is unknown. An extensive amount of experiments on standard datasets show that the method is much more effective than the previous state-of-the-art algorithms, particularly in the case of unknown orientation and limited field of view (Shi et al., 2020) [9]. This article is very pertinent to our study because it deals with the important issue of orientation ambiguity in cross-view matching, which offers a framework that enhances location and orientation estimation of image geolocalization activities.

Hausler et al. (2021) [10] propose a new visual place recognition system, Patch-NetVLAD, which is a synergistic combination of local and global descriptors, extracting patch-level features of NetVLAD residuals and fusing them on a multi-scale basis. This method allows strong matching in different appearance, illumination, and viewpoint conditions, and has been shown to perform better on a wide range of challenging datasets, including Nordland, Pittsburgh, Tokyo 247, Mapillary, RobotCar Seasons v2, and Extended CMU Seasons. Patch-NetVLAD, which is not only significantly faster than global and local descriptors in recall but also highly versatile (with configurations that are order-of-magnitude faster or more memory-efficient depending on application requirements) is made possible by their proposed IntegralVLAD (Hausler et al., 2021) [10]. This system is very applicable to our work since it shows the usefulness of combining local and global clues to resilient image geolocalization and offers useful benchmarks on condition and viewpoint invariance.

Zhang et al. (2025) [35] provide an in-depth analysis of Multimodal Large Language Models (MLLMs) as image geolocation models, especially their capacity to determine the geographic origins of street-view images. The benchmarks of the study include state-of-the-art MLLMs like the Gemini 2.5 Pro and visual reasoning models of OpenAI on a street-view dataset distributed across the world, which show that MLLMs can reach localization accuracy up to 49% within a 1 km radius (Zhang et al., 2025) [35]. The study breaks down the logic of MLLMs, emphasizing the fact that models use visual features such as building styles, landmarks, language and text, and transit nodes to determine location, and also addresses the strong bimodal distribution in prediction accuracy. Also, the paper highlights the extreme

privacy threats of such models, where even seemingly innocent photos can reveal sensitive location data. The authors suggest privacy protection measures and describe the further research to create specialized geolocation models based on multi-step reasoning and integration of tools. The paper is applicable to our work because it identifies the strengths and weaknesses of modern large multimodal models in geolocation, the importance of visual semantic cues, and the privacy concerns that should be considered in the real world in image geolocalization.

Li et al. (2025) [32] present IMAGEO-Bench, a benchmarking suite to assess the image geolocalization performance of large multimodal language models (LLMs). The benchmark consists of three different datasets of global street scenes, U.S. points of interest, and an unseen image set, which is privately held, and has a broad range of geographic and environmental contexts. Their assessment system considers various performance indicators such as accuracy in different geographic scales, distance error, cost of computation, geospatial bias, and model reasoning transparency by structured output and feature importance diagnostics. Findings show that closed-source models tend to be more accurate in geolocalization than open-source models, but they are more expensive to compute. The models have high dependence on visual cues like landmarks, textual signage, and cultural cues and perform well in an urban and outdoor setting but poorly in a natural or indoor setting. The paper also notes the geographic bias towards densely sampled, visually salient areas and the necessity of more generalizable, equitable models of geolocalization (Li et al., 2025) [32]. The work is applicable to our study because it provides a strict evaluation methodology and the strengths and limitations of the existing methods of geolocalization based on LLM, which will allow to better benchmark and improve the methods in the future.

Wang et al. (2025) [33] introduced LocDiffusion, a generative diffusion-based model that predicts an image's geographic location by representing coordinates continuously on Earth's sphere. Instead of using classification or retrieval, the authors proposed a Spherical Harmonics Dirac Delta (SHDD) encoding system that transforms latitude and longitude into a dense latent space, allowing smoother learning of spatial information. The model uses a Conditional Siren-UNet (CS-UNet) architecture to perform diffusion and denoising conditions on image embeddings from a CLIP encoder. The study trained on the MP16 dataset (4.72M images) and tested on Im2GPS3k, showing better accuracy than GeoCLIP for coarse location prediction (region, country, continent) and stable generalization to unseen data (Wang et al., 2025) [33]. However, it was weaker at fine-grained localization (under 25 km). To solve this, a hybrid LocDiffusion + GeoCLIP model was proposed, improving performance across all scales. The main limitations include high computational cost for fine precision and dependency on hybrid retrieval for close-range accuracy. This paper is relevant to our study as it explores a novel generative approach that enhances spatial continuity and generalization—key goals of our research.

Johns, Rounds, and Henry (2017) [4] proposed a multi-modal geolocation estimation model that combines image content with metadata such as posting time and user albums to improve location prediction. They divided the Earth into a Delaunay triangle mesh, which adapts better to land and water boundaries than regular grids. The model has three variants: M1, which uses CNN-based image classification;

M2, which adds time metadata; and M3, which incorporates user-album sequences through LSTMs to refine predictions. Using the YFCC100M dataset (14.9M images) and testing on Im2GPS, they found that metadata improved city, region, and country-level accuracy, achieving up to 3% (Johns et al., 2017) [4] better results at the 25 km scale. However, performance depends on metadata availability, and mesh discretization still causes some error. This paper is relevant to our study as it highlights how combining visual and contextual data can enhance geolocation accuracy, complementing our model's focus on spatial learning.

Panagiotopoulos, Kordopatis-Zilos, & Papadopoulos (2021) [11] define the concept of image localizability (i.e. whether an image contains enough geographic cues) and propose a selective prediction framework: the model can abstain from predicting when uncertain. They design two selection functions using the output probability distributions of geolocation models to decide if an image should be localized at different spatial scales. Their experiments show that by rejecting low-confidence (non-localizable) images, city-level accuracy jumps from 27.8% to 70.5% (Panagiotopoulos et al., 2021) [11]. This approach makes geolocation models more robust in practice by avoiding errors on "unlocalizable" images. It's relevant to our work because it addresses reliability and uncertainty in geolocation predictions – an aspect few geolocation models consider.

Bianco, Eigen, & Gormish (2024) [24] propose improving image geolocation by ensembling geolocation models with satellite-derived ground-level attribute predictors (e.g. population density, land cover). They also introduce a new metric, Recall vs Area (RvA), which evaluates how well predicted regions (not just a single point) cover the true location as a function of area. Their ensemble method combines outputs of GeoEstimation or GeoCLIP with attribute masks (LandScan, ESA-CCI Land Cover) by multiplying probability maps and then accumulating area until ground truth falls within. Experiments on Im2GPS3k and a diverse Street View dataset show that their ensemble approach improves recall for under-sampled (non-urban / rural) regions, especially at larger spatial scales, though gains are smaller at fine scales (Bianco et al., 2024) [24]. The model helps generalization to areas lacking dense training images. This is relevant because it marries image-based geolocation with global satellite attributes, which aligns with our goal of improving generalization and spatial consistency beyond just classification or retrieval.

Haas, Alberti, & Skreta (2023) [21] proposed StreetCLIP, a generalized zero-shot geolocalization model, which grounds CLIP's zero-shot learning capability in the geolocation domain via synthetic captions. They generate captions like "A Street View photo close to the town of city in the region of region in country" for training, thereby teaching CLIP to distinguish geographic classes it hasn't seen before. At inference, they apply a hierarchical linear probing method: first predict the country, then refine to the city. StreetCLIP, evaluated on IM2GPS and IM2GPS3K, achieves state-of-the-art accuracy at coarse thresholds (e.g. 750 km, 2,500 km) while operating in a zero-shot fashion. It improves over vanilla CLIP by 1.3 to 3.4 percentage points across various distance thresholds (Haas et al., 2023) [21]. However, the method is weaker for fine-grained localization (city/region) due to constraints like limited vocabulary, reliance on populous-city selection, and no landmark-specific tokenization. This work is relevant to ours because it shows how meta-learning via

synthetic captioning and zero-shot mechanisms can extend geolocalization beyond fixed class sets and improve generalization across geographies.

Hays & Efros (2008) [1] introduced IM2GPS, a data-driven geolocation system that estimates geographic location from a single image using scene matching. The approach leverages a dataset of over 6 million GPS-tagged images from Flickr, representing locations as probability distributions over Earth's surface rather than discrete classifications. The system uses multiple visual features including tiny 16x16 images, color histograms, texton histograms, line features, gist descriptors and geometric context to find nearest neighbors in the database. For each query image, the method computes distances across feature spaces and applies mean-shift clustering with 500km bandwidth to identify geographic clusters from the top 120 nearest neighbors. The evaluation on a 237-image test set showed 16% accuracy within 200km using single nearest neighbor, performing 18 times better than chance (Hays & Efros, 2008) [1]. The gist descriptor performed best among individual features, while color histograms also proved surprisingly effective for geographic discrimination. With multiple guesses allowed, median error dropped to under 500km with 6 estimates. The system demonstrated practical applications for secondary geographic tasks like population density estimation, elevation gradient prediction and urban/rural classification (achieving 0.82 AUC). However, limitations include reliance on database coverage density, computational cost of feature extraction (15 seconds per image) and difficulty with fine-grained localization for generic scenes. The approach pioneered large-scale image geolocation and established benchmarks still used in current research, though performance degrades significantly for images without distinctive geographic features.

Yi & Shan (2025) [34] introduce GeoLocSFT, a framework that leverages supervised fine-tuning of large multimodal foundation models (Gemma 3 27B-it and Qwen2.5-VL-3B) for visual geolocation using minimal high-quality training data. The approach trains with only 2,700 carefully curated image-GPS pairs from their geographically diverse MR600k dataset, generating detailed "geo-captions" through Claude 3.7 Sonnet analysis that includes multi-scale contextual analysis, micro-feature identification and explicit reasoning for disambiguation. The fine-tuning process uses one epoch of training with LoRA adaptation, achieving efficient adaptation through cross-entropy loss optimization. Evaluated on standard benchmarks (Im2GPS-3k, YFCC-4k) and their newly introduced MR40k benchmark for sparsely populated regions, GeoLocSFT demonstrates competitive performance, with accuracy improvements of up to 12% over baseline models at certain distance thresholds(Yi & Shan, 2025) [34]. The method shows particularly strong performance on region-specific subsets, with the 3B model outperforming larger general-purpose models like GPT-4.1 on Austria-specific tests. Multi-candidate re-ranking exploration showed theoretical potential but limited practical gains. Main limitations include dependency on high-quality training data curation and moderate improvements over traditional methods at fine-grained scales. The work emphasizes data quality over quantity, demonstrating that sophisticated reasoning-based supervision can effectively adapt foundation models for complex geospatial tasks without massive datasets or complex pipelines.

Radford et al. (2021) [12] demonstrate that contrastive pre-training on 400 million

uncurated (image, text) pairs—by jointly training an image encoder and a text encoder to predict correct image–caption pairings—yields state-of-the-art, zero-shot transferable visual representations; their CLIP models (ResNet and Vision Transformer variants) match or exceed supervised baselines (e.g., ResNet-50 on ImageNet) in zero-shot accuracy across 30+ datasets covering OCR, action recognition, geolocalization and fine-grained classification, achieving 76.2% ImageNet top-1 zero-shot accuracy without any labeled examples and scaling predictably with compute (Radford et al., 2021) [12]; CLIP also exhibits superior robustness under natural distribution shifts, shrinking the "robustness gap" by up to 75%, though it struggles on tasks with scarce pre-training coverage (e.g., handwritten digit recognition) and incurs massive computational cost (up to 18 GPU-days on 592 V100s), highlighting both the promise and limitations of large-scale natural language supervision for general vision systems.

Seo et al. (2018) [7] introduce CPlaNet, a classification-based geolocation framework that overcomes the trade-off between map granularity and data scarcity by intersecting multiple coarse geoclass partitions to form a vast set of fine-grained regions. Leveraging Inception-v3 features shared across several classifier heads—each trained on a distinct geoclass set generated via random graph-based S2 cell merges weighted by visual and geographic distances, CPlaNet employs combinatorial partitioning to predict locations by aggregating normalized scores across overlapping classifiers. Trained on 30.3 M Flickr images and evaluated on Im2GPS3k, YFCC4k and Im2GPS benchmarks, CPlaNet achieves up to 16% street-level accuracy improvements over single-partition baselines at the 1 km scale and outperforms both classification and retrieval methods across distance thresholds (Seo et al., 2018) [7]. Despite requiring only marginal additional parameters and inference overhead, CPlaNet excels at fine-grained localization, though it incurs preprocessing complexity for geoclass generation and is limited by ambiguous scenes lacking geographic cues.

Berton et al. (2023) [19] introduced a large benchmark for visual geolocation using deep learning. They tested various models, including CNNs and Transformers, to determine which ones produce the best image features for identifying the location where a photo was taken (Berton et al., 2022) [19]. This work mainly focuses on retrieval based geolocation which means the features of the images are stored first and then compared. It is like the process where a software of search engines match similar photos. This paper is very useful to understand various types of deep models which performs best in location prediction tasks.

Müller-Budack and his team (2018) [6] introduced a smart hierarchical geolocation model that tells the location of a photo in steps. At first it chooses a large area (it can be a continent or a country) and then it breaks down that large area into smaller areas. This model also tells what the image depicts like a city or a forest or beach or a mountain[6]. To do that they used scene classification. These methods made the model powerful to make more smart and structured guesses about a place instead of jumping straight to te exact coordinates.

Hu et al. (2018) [5] designed a model called CVM-Net to match street level photos with aerial views or satellite views. Their model is capable to learn strong image

features which means the model learns or extracts powerful features from the image which does not change even when the view changes. The model uses NetVLAD method to extract the features and store them in a database. And later it retrieves the best match.

Naskar et al. (2024) [30] developed Img2Loc. It is a system that combines different information of an image with the image features. The other informations are GPS metadata, time, compass direction. This process is the main power of this system as this system uses these various features or inputs which help it to understand the environment better predict more flawlessly. This research shows that the use of metadata can improve the result .

Wilson et al. (2021)[22] wrote a complete survey covering almost every major geolocation method. From early handcrafted methods like SIFT and SURF to deep learning-based ones like NetVLAD and Transformers, he tried every possible important methods to cover. They explained how models extract features from images, store them in databases, and use retrieval to find the best match. This is like a survey. If anyone wants to know and learn about the evolution of image based geolocation then this survey is the best guide.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

This research required a combination of software, hardware, and data resources to develop, train, and evaluate deep learning models for global image geolocalization. The system specifications and methodological requirements are outlined below.

Technical and Methodological Requirements

- **Programming Environment:**
  - **Language:** Python 3.12.3
  - **Frameworks:** PyTorch, Hugging Face Transformers, OpenCLIP, OpenCV
  - **Supporting Libraries:** NumPy, TorchVision, Transformers, Pandas, Matplotlib, scikit-learn, FAISS, Geopy, Geopandas, Tqdm, Pillow, Seaborn, Requests, Datasets, Accelerate
- **Hardware Configuration:**
  - **Processor:** AMD Ryzen 7 / Intel Core i7 or higher
  - **GPU:** NVIDIA RTX 3060 or higher with at least 12 GB VRAM (for training transformer-based models)
  - **Memory:** Minimum 32 GB RAM
  - **Storage:** Minimum 500 GB SSD for dataset and model checkpoints
- **Modeling Requirements:**
  - Two major architectures were implemented:
    1. **Coordinate Regression Model:** CNN-based dual-head architecture for direct latitude–longitude prediction.
    2. **Geographic-Aware Geocell Classification Model:** CLIP-based multimodal transformer fine-tuned with Haversine Smoothing Loss.
  - **Optimization Algorithm:** AdamW optimizer with learning rate 1×10^{-4} and weight decay 1×10^{-4}

  - **Training Parameters:** 20 epochs, mixed precision training enabled for efficiency

- **Data and Resource Requirements:**
  - **Dataset:** 100,000 geotagged street-level images collected via Mapillary Graph API (v4)
  - **Geographic Coverage:** 183 countries across six continents
  - **Data Splits:** 80% training, 10% validation, 10% testing
  - **Feature Extraction:** CLIP ViT-L/14 @336px encoder producing 768-dimensional feature vectors
  - **Geocell Generation:** 2,203 clusters created using K-Means clustering for spatial partitioning
  - **Loss Function:** Combination of CrossEntropyLoss and Haversine Smoothing Loss ($\tau = 75.0$)

- **Software Tools and Platforms:**
  - **Development Platform:** Local CUDA-enabled WSL environment
  - **Version Control:** GitHub for collaboration and version tracking
  - **Visualization:** TensorBoard and Matplotlib for monitoring training convergence and evaluating performance metrics

- **Evaluation Metrics:**
  - Mean and median Haversine distance
  - Accuracy thresholds at 1 km, 25 km, 200 km, 750 km, and 2500 km
  - Loss convergence curves for model stability assessment

Feasibility and Constraints

Although the proposed system was successfully implemented, there were a number of constraints that affected its development. High capacity transformer models were costly to train in terms of GPU resources and time. The Mapillary dataset, despite being globally diverse, was imbalanced in terms of regions, with some regions being underrepresented like Africa. There were also restrictions on the number of image downloads per query in the backend free tier. Also, large-scale CLIP encoders were memory-intensive to fine-tune, and mixed precision and batch size optimization was necessary to stabilize them. Nevertheless, the system was able to attain efficient training and stable convergence with the available computational resources.

3.2 Ethical Issues

The development and implementation of image geolocalization models provoke a number of ethical issues and professional concerns. As this study is concerned with location-based visual data, the privacy, transparency, and fairness of the data were a fundamental issue during the study.

Data Privacy and Anonymity

Images in this study were all taken off the publicly available Mapillary dataset, which offers user-contributed street-level imagery under open data licenses. No personally identifiable information (PII) was extracted, stored, or processed. The dataset did not include metadata like user identity or upload timestamps to maintain anonymity. All geographic coordinates were utilized in aggregated and non-identifiable formats and were strictly used in research purposes.

Bias and Fairness

Although the dataset included many countries, it was geographically unbalanced - the representation of such regions as North America and Europe was higher than that of Africa or Oceania. Such imbalance may create bias in model predictions, which will give preference to overrepresented areas. To counter this, data sampling methods and K-Means-based clustering were used to ensure fair geographic distribution during training.

Responsible Use and Model Transparency

The geolocation model was created solely for academic and scientific purposes. It is not designed to be used in surveillance, law enforcement or any other use that can violate the privacy of an individual. All findings, images and forecasts were produced under controlled research conditions. The architecture and parameters of the system were properly documented to make the system transparent and reproducible in future research.

Academic Integrity

All the datasets, methods, and previous research mentioned in this thesis were cited in a proper way based on the academic standards. We complied with the BRAC University Code of Conduct on Research Integrity and all the results of the experiment were original, well reported and ethically obtained.

Environmental Considerations

Deep learning models require a lot of computational energy to train on large datasets. Mixed-precision computation was used to optimize training sessions to minimize unnecessary GPU runtime to minimize environmental impact. These measures are indicative of a bid to sustain computational research.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

Geolocalizing images is a challenging computer vision problem. It is a task of predicting the geographic location of an image by analyzing it without the help of any metadata directly keeping the longitude and latitude of the image. This issue is challenging since impressions may be depicted in any corner of the globe, which has both huge geographical ranges and landscapes, architecture, vegetation, and cultural diversity. An approach to this is regression, which actually returns continuous values of latitude and longitudes in an input image. Direct regression is, however, usually more difficult to train (e.g. predicting Tokyo vs. New York is a large error, but treating coordinates as continuous values makes this non-obvious to the model). The earlier solutions have attacked Geolocalization in different manners. The most common approach is to frame the problem as one of classification, by dividing the earth into geographic cells (also known as geocells), and training a model to predict the appropriate cell that contain an image. This was popularized by Google PlaNet [3] and persists in recent works that make use of state-of-the-art vision transformers to increase accuracy. The latest state-of-the-art systems have used a mixture of classification and search-based refinement (PIGEON (2023) [26] enables the retrieval of more than 40% of test images within 25 km of ground truth worldwide). Such developments confirm both the promise of more modern deep features and staged initiatives to geolocation and the difference between the multiple baselines and the latest state of art.

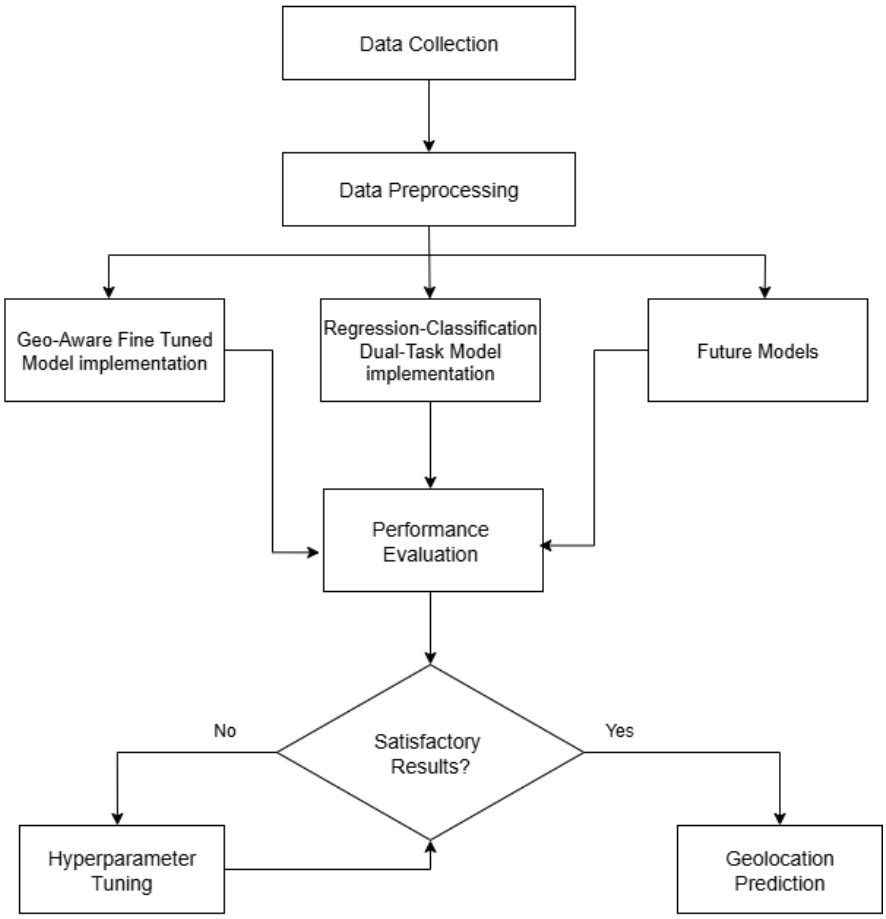

Figure 4.1: Top-level overview of the proposed system

This chapter presents the methodology for developing and evaluating two distinct approaches to CLIP-based image geolocalization: a direct regression method and a geographic-aware geocell classification approach. The methodology encompasses the design process, model specifications, data collection and preprocessing, and implementation details.

The research follows a comparative experimental design, where both approaches are implemented using the same dataset and evaluation metrics to ensure fair comparison. The methodology is structured around three main phases: (1) preliminary design and model specification, (2) data collection and preprocessing, and (3) implementation of the selected designs.

4.2 Preliminary Design or Design (Model) Specification

4.2.1 CNN Regression Approach Design

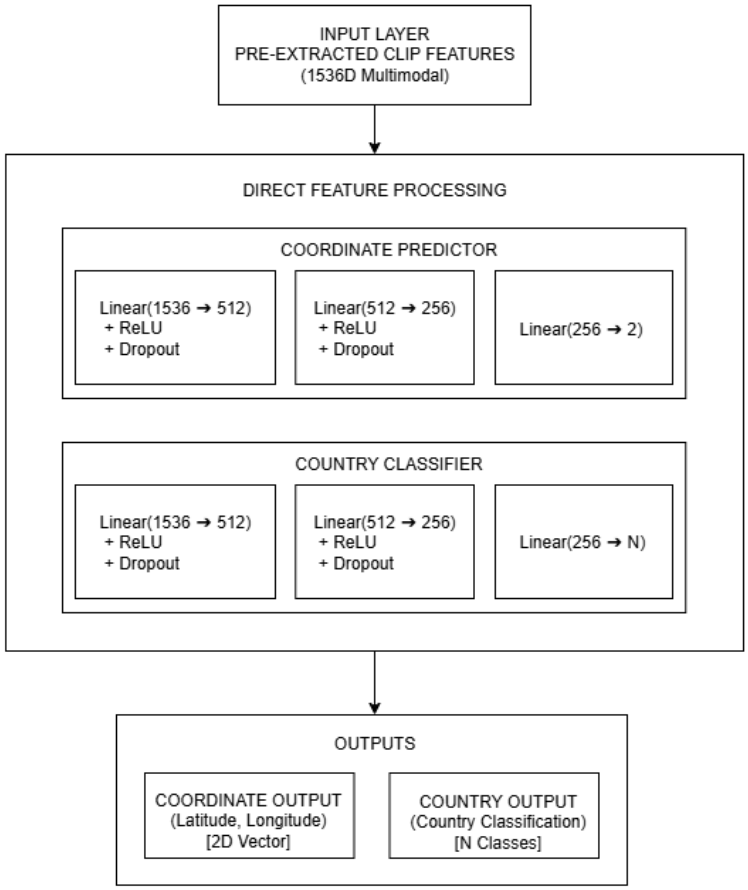

Figure 4.2: Baseline CNN dual-head architecture

The Baseline CNN regression approach employs a dual-head architecture for direct coordinate regression:

- **Input Features**: Combined CLIP image and text features (1536-dimensional)
- **Coordinate Predictor**: $1536 \rightarrow 512 \rightarrow 256 \rightarrow 2$ (latitude, longitude)
- **Country Classifier**: $1536 \rightarrow 256 \rightarrow 183$ countries
- **Loss Function**: MSE for coordinates + CrossEntropy for countries
- **Optimization**: AdamW with learning rate 1e-4, weight decay 1e-4
- **Batch Size**: 1024 (large batches due to pre-loaded features)

The architecture leverages FAISS indexing for efficient feature retrieval and pre-loads all features into memory for ultra-fast access during training.

4.2.2 Geographic-Aware Geocell Approach Design

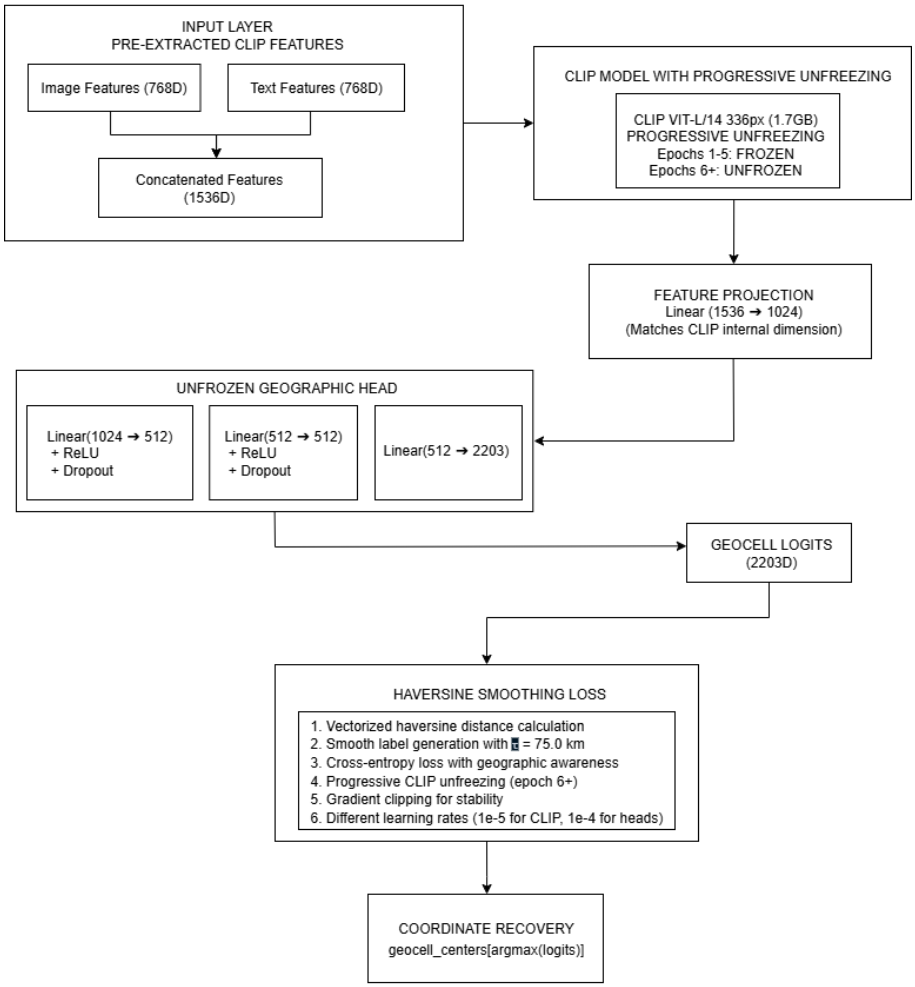

Figure 4.3: Fine Tuned CLIP Model Architecture

The improved approach uses a geocell classification strategy with geographic awareness:

- **Input Features**: Combined CLIP image and text features (1536-dimensional)
- **Feature Projector**: 1536 → 1024 (projection to CLIP's internal dimension)
- **Geolocalization Head**: 1024 → 512 → 512 → 2203 geocells
- **Loss Function**: Haversine Smoothing Loss with temperature $\tau = \mathbf{75.0}$
- **Geocell Creation**: K-means clustering (2203 clusters)
- **Batch Size**: 64 (smaller batches due to complex loss calculation)

The key innovation lies in the Haversine Smoothing Loss, which incorporates geographic distance directly into the training process:

$$\text{Smooth Label} = \exp\left(-\frac{\text{Haversine Distance}}{\tau}\right) \tag{4.1}$$

where $\tau = \mathbf{75.0}$ is the temperature parameter controlling the smoothness of the label distribution.

4.3 Data Collection

We have retrieved 100,000 street-level images from Mapillary using Mapillary Graph API (v4). We reached the data by means of an access token. In our script the world was partitioned into a grid of bounding boxes (1 degree latitude x 1 degree longitude). To prevent polar areas with nsufficient to no data, the range of latitudes was set to -60 degrees and +80 degrees. The maximum longitude as -180 to 180 to cover the globe completely. The bounding boxes have been randomly shuffled to prevent regional biasness. The script requested the Mapillary API with the up to 100 images that the script had each bounding box query. Where no pictures could be detected in any of a box, the script took the script to the next box.
These images are further categorized into directories according to their data splits, such as `mapillary_images_10k/train/`, `val/`, and `test/`, corresponding to training, validation, and test sets respectively. The GPS coordinates are embedded within each filename using the format: `image_<id>_lat<LAT>_lon<LON>.jpg`. For example, the file `image_12345_lat12.3456_lon-78.9012.jpg` indicates that the image was taken at latitude 12.3456 and longitude -78.9012.

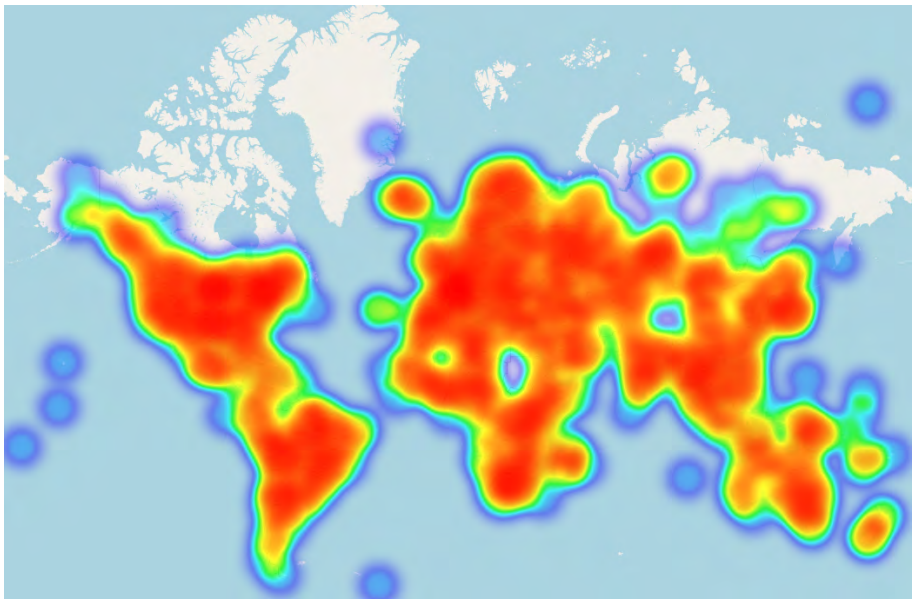

Figure 4.4: Geographic distribution of collected Mapillary images shown as a heatmap.

Table 4.1: Continent Overview of the Dataset

Continent	**Total Images**	**Unique Countries**
Europe	24,424	37
North America	21,077	14
Asia	17,607	38
Africa	13,243	41
South America	12,274	12
Oceania	7,035	7

Table 4.2: Top 20 Countries in the Dataset

Country	**Image Count**	**Percentage (%)**
USA	13,145	13.145
Russia	7,505	7.505
Brazil	6,334	6.334
Australia	5,645	5.645
Canada	4,471	4.471
India	2,547	2.547
South Africa	2,257	2.257
France	1,950	1.950
Turkey	1,635	1.635
Peru	1,527	1.527
Norway	1,510	1.510
Argentina	1,470	1.470
Finland	1,412	1.412
Mexico	1,303	1.303
Sweden	1,281	1.281
China	1,242	1.242
Spain	1,139	1.139
Japan	1,123	1.123
Kazakhstan	1,103	1.103
Indonesia	1,041	1.041
Total (Top 20)	61,910	61.91

4.3.1 Data Cleaning

The data cleaning process involved several steps:

1. **Coordinate Validation**: Verification of latitude/longitude values within valid ranges (-90° to 90° for latitude, -180° to 180° for longitude)

2. **Duplicate Removal**: Elimination of duplicate images based on file hash and coordinates

3. **Outlier Detection**: Removal of images with coordinates that are clearly erroneous (e.g., coordinates in oceans for street scenes)

4. **Country Mapping**: Standardization of country names and mapping to consistent identifiers

4.3.2 Data Transformation

Figure 4.5: Sample raw dataset images

The image filenames were gleaned with a preprocessing script to read the latitude and longitude and create CSV files based on the data split (e.g., train_labels.csv). The number of columns in every CSV were three; name of the image, latitude, and longitude.
The images were resized to 336x336 pixels. The CLIP image processor performs resizing, central cropping and normalization as per the pretrained backbone demands of CLIP, but for our convenience of handling the data well we pre-resized the images. We also performed augmentation to 30% of the training set images, adding random noise, random flip, random brightness adjustments.

Figure 4.6: Dataset images after preprocessing

4.3.3 Data Integration

All preprocessed data were systematically integrated into a structured database system. A relational schema was developed to store image metadata, extracted features, and geolocation information. CLIP features were aligned with their corresponding image entries. A consistent mapping between relative and absolute paths was maintained for cross-platform compatibility.

4.3.4 Data Reduction

Data reduction techniques applied:

1. **Feature Dimensionality**: Use of 768-dimensional CLIP features instead of raw images
2. **Batch Processing**: Efficient batch loading of features during training
3. **Memory Optimization**: Pre-loading of features into memory

4. **Geocell Clustering**: Reduction of prediction space from continuous coordinates to discrete geocells

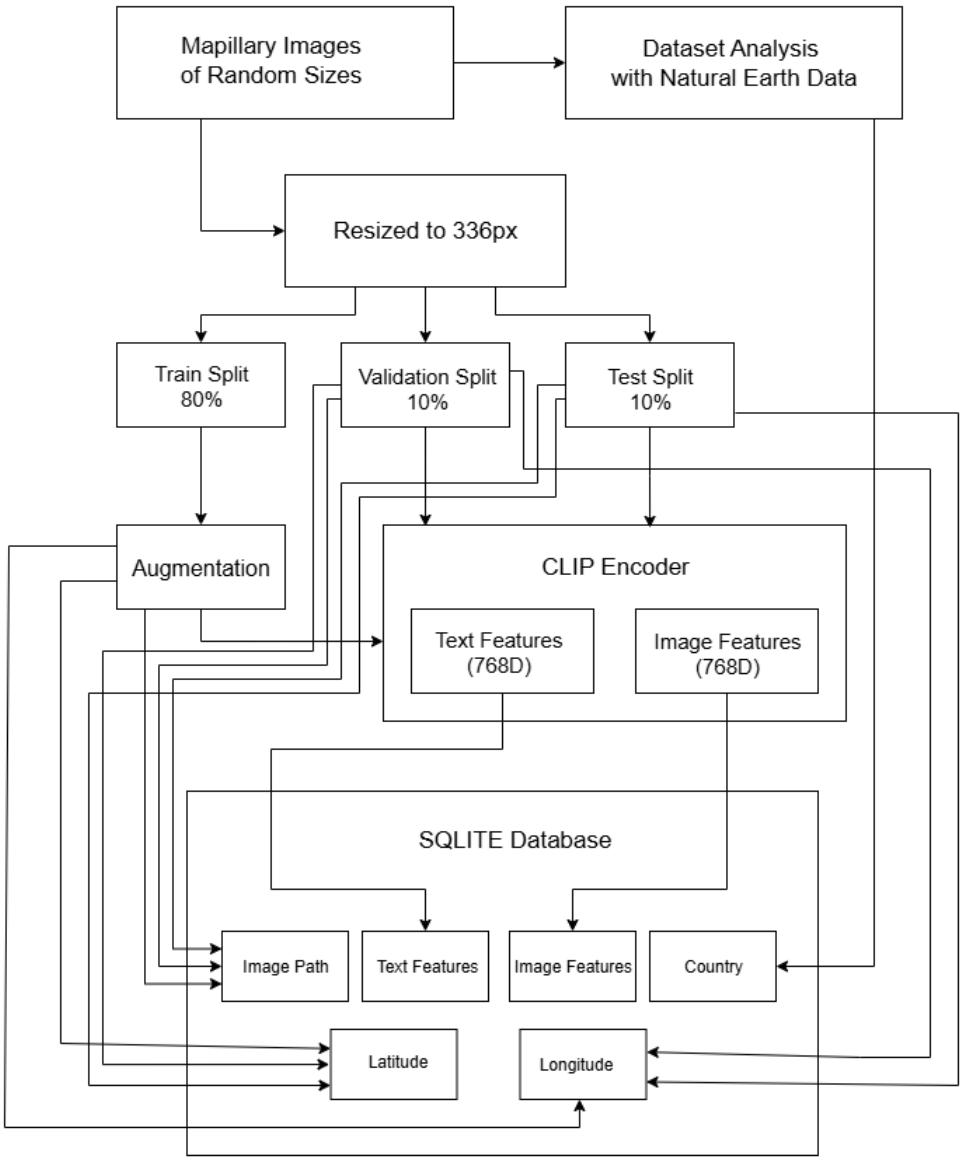

Figure 4.7: Data pre-processing workflow

4.3.5 Summary of Preprocessed Data

The dataset consists of 100,000+ geotagged images from the Mapillary 100k dataset with the following characteristics:

- **Total Images**: 100,000+ geotagged images
- **Geographic Coverage**: Global (183 countries)

- **Split Strategy**: 80% training, 10% validation, 10% test
- **Feature Extraction**: CLIP ViT-L/14 @336px (768-dimensional image features)
- **Quality Control**: Manual verification of coordinates

4.4 Implementation of Selected Design

This research implemented a systematic nine-stage progression, culminating in a geographically-aware multimodal geolocalization model. Each stage built upon the insights, provided by its predecessors. The architecture has developed out of simple coordinate regression, to a highly-sensitive geocell classification system.
The principles of our approach could be highlighted as follows: 1) setting **Baseline Models** (Stages 1-2); 2) launching **Geographic-Aware Innovations** (Stages 3-5); 3) **Multimodal Features** integration (Stage 6); and 4) developing the **Final Optimized Models** (Stages 7-9).

4.4.1 Baseline Model Implementations

The first stage was aimed at the creation of a performance baseline with the help of traditional. and CLIP-based approaches. The **Baseline CNN Model** (Stage 1) employed a used a two-task framework of coordinate regression and country labelling, based on 1536-dimensional multimodal CLIP features that had been pre-extracted. This model, despite its Multi-task design, never learnt to be an effective geolocalizer, with 0.00% accuracy in 1 km and 25 km distance errors, and a mean distance error of 3191.52 km. One of the weaknesses was that it was reliant on processing direct features without mapping to the internal representation space of CLIP.
The **Baseline CLIP Model** (Stage 2) tried to enhance the performance with the assistance of 768-dimensional CLIP image features with a typical three-layer Multilayer Perceptron (MLP) for direct coordinate regression, while the CLIP ViT-L/14@336px model was loaded and kept frozen. The mean distance was also large as a result of this approach, mean distance error (2,163.21 km) proving that simple MSE regression on raw CLIP features is not good enough to geolocalize.

4.4.2 Geographic-Aware Innovations

The poor results of the baseline models led to the adoption of geographic consciousness, transferring the problem out of pure regression, into a more solid classification paradigm.

Geographic Smoothing and Geocell Classification

The **Haversine Smoothing CLIP Model** (Stage 3) marked the first major innovation. It substituted the conventional MSE loss wtih a novel Haversine smoothing loss. This loss function is used to penalize the wrong predictions by the great-circle

distance d between the predicted and actual coordinates, which are calculated with the help of the haversine formula:

$$d = 2r \times \arcsin\left(\sqrt{\sin^2(\Delta\text{lat}/2) + \cos(\text{lat}_1)\cos(\text{lat}_2)\sin^2(\Delta\text{lon}/2)}\right)$$

The loss generates smooth probability labels using a temperature parameter $\tau = 75.0$ km. This method stabilized training instantly seeing 5.95% accuracy at 1km and getting the mean distance error down to 923.09 km.

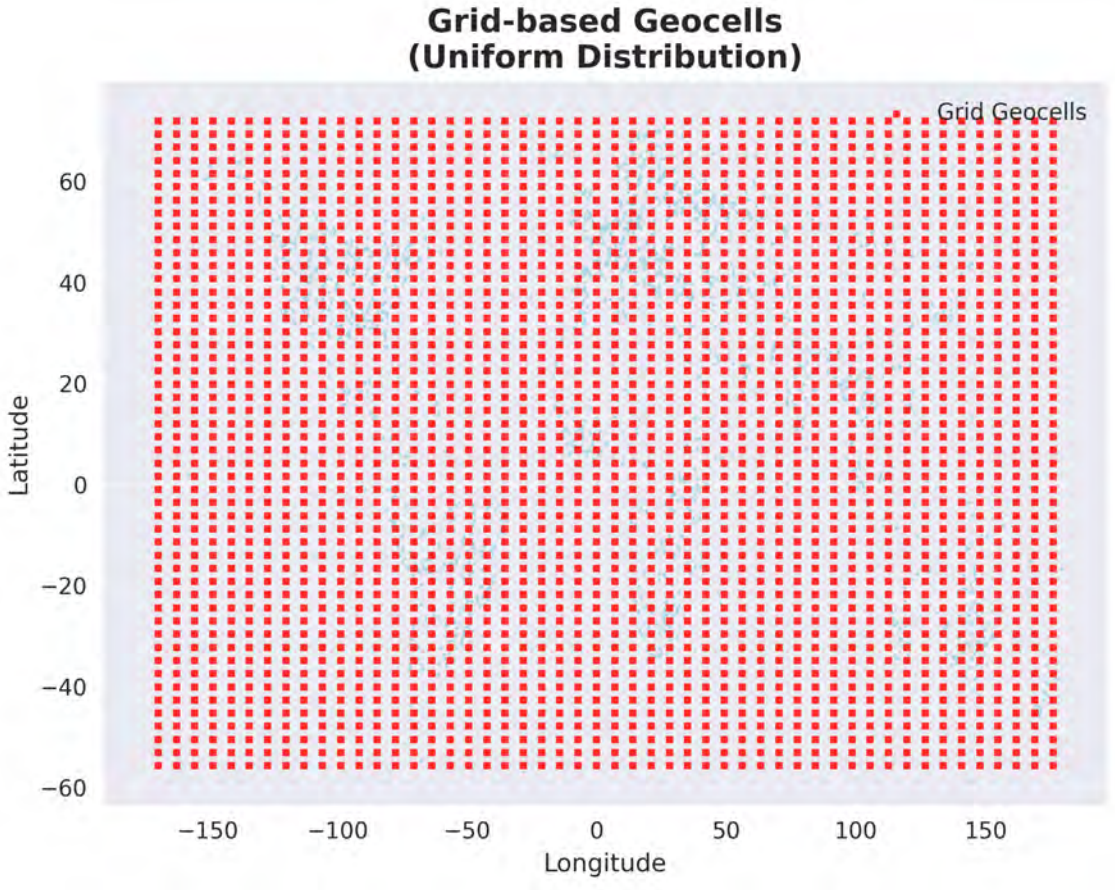

Figure 4.8: Uniform Grid-based Geocells

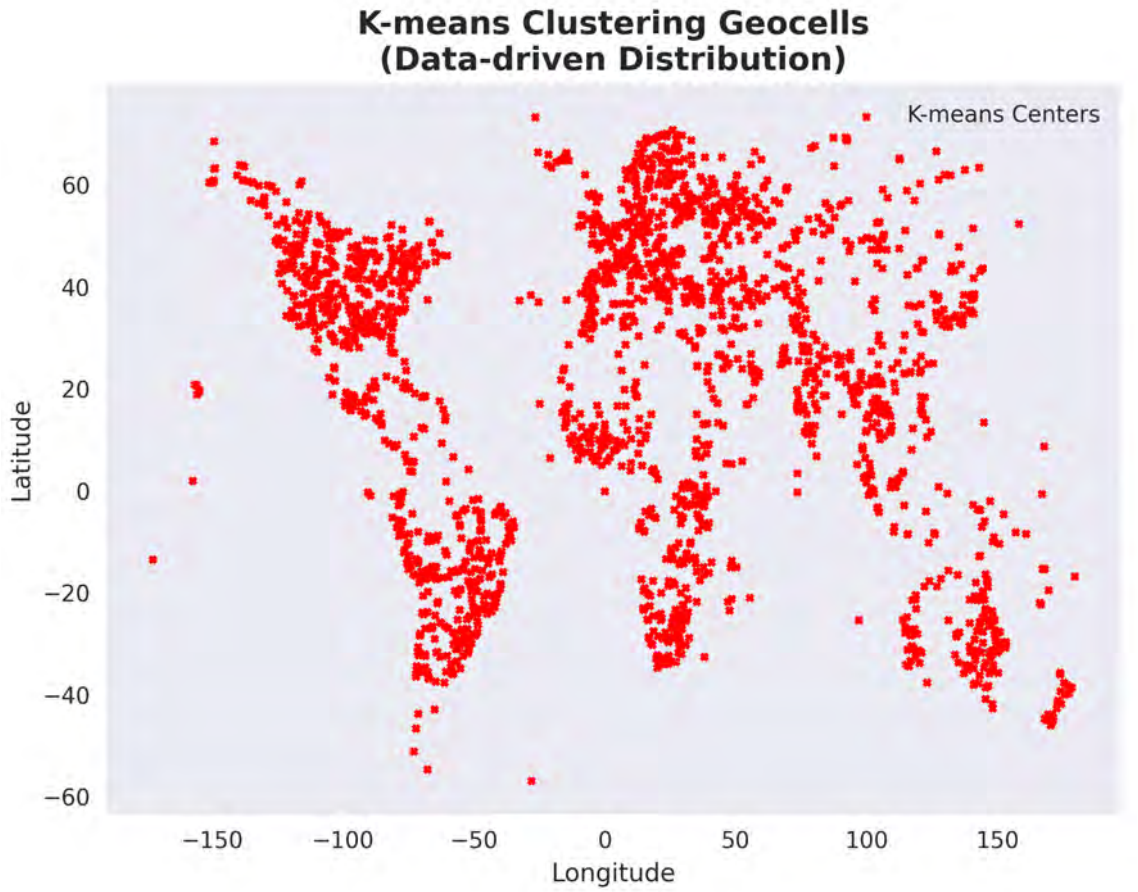

Figure 4.9: Data Driven K-means clustering Geocells

Building upon this, the **Geo-aware-cells CLIP Model** (Stage 4) put in place a **discrete classification** scheme. This was done by clustering all training coordinates into $k = 2,203$ geographic cells (geocells) using K-means, and training the model to predict the correct geocell index using cross-entropy loss. This discrete formulation led to increase in accuracy by a little at 1km to 6.54
The **Haversine Geo-aware-cells CLIP Model** (Stage 5), was a combination of the methods. Where the distance-aware Haversine smoothing loss ($\tau = 75.0$ km) was applied to the $2,203$ geocell logits. This unification took advantage of the geographical structure of the geocells and the stabilizing effect of the smoothing loss.

4.4.3 Multimodal Integration and Final Optimization

Multimodal Feature Integration

In order to take advantage of outer context, Multimodal CLIP Model (Stage 6) concatenated the 768 dimensional features of image and 768 dimensional features of text into a single 1536 dimensional vector for direct coordinate regression. While the text potentially has more explicit location cues, but the simple regression architecture was not helpful, again yielding low accuracy.

Feature Alignment and Fine-Tuning

The last steps were to optimize the feature processing and utilise the CLIP architecture. The **Haversine Geo-aware-cells Single-Modal CLIP Model** (Stage 7) proposed a critical feature projection layer (768D $\rightarrow$ 1024D) to align the input features with the internal dimensionality of the CLIP model's learned representation space. This minor transformation together with geographic awareness produced a better mean distance error of 825.56 km.
The **Haversine Geo-aware-cells Multimodal CLIP Model** (Stage 8) consolidated all the successful techniques: the 1536D multimodal features were projected to 1024D, and the result was used.and the output head was trained using the Haversine smoothing geocell loss. This comprehensive integration led to the highest frozen-model performance: 7.35% accuracy at 1km and a low mean distance error of 144.87 km.
The ultimate iteration, the **Fine tuned Haversine Geo-aware-cells Multimodal CLIP Model** (Stage 9), adopted a **progressive unfreezing** strategy. The model's custom geolocalization heads were trained with the CLIP encoder frozen for initial stability (Epochs 1-5), after which the entire model was unfrozen and fine tuned with a lower learning rate for the CLIP parameters. This fine-tuning adapted the large CLIP model specifically for the geolocalization task, culminating in the best overall performance: 7.48% accuracy at 1km, 81.75% at 25km, and a state-of-the-art 60.41 km mean distance error.

4.4.4 Training Configuration

Both approaches use the following training configuration:

- **Optimizer**: AdamW with learning rate 1e-4
- **Epochs**: 20 epochs for convergence
- **Device**: CUDA GPU acceleration
- **Mixed Precision**: Automatic mixed precision for efficiency
- **Checkpointing**: Model saving at best validation performance

4.4.5 Evaluation Framework

The evaluation framework includes:

1. **Distance Metrics**: Mean, median, and standard deviation of haversine distances
2. **Accuracy Thresholds**: 1km, 25km, 200km, 750km, 2500km accuracy
3. **Training Dynamics**: Loss convergence and distance error evolution
4. **Statistical Analysis**: Comprehensive performance comparison

The implementation ensures fair comparison between both approaches using identical evaluation metrics and statistical significance testing.

Chapter 5

Result Analysis

5.1 Performance Evaluation

This chapter presents a comprehensive analysis of the experimental results comparing the Baseline CNN direct regression approach with the geographic-aware geocell classification approach for CLIP encoded image geolocalization.

5.1.1 Experimental Setup

The evaluation was conducted on a dataset of 100,000 geotagged images with the following configuration:

- **Training Set**: 80,000 images
- **Validation Set**: 10,000 images
- **Test Set**: 10,000 images
- **Geographic Coverage**: 183 countries globally
- **Feature Extraction**: CLIP ViT-L/14@336px (768-dimensional)
- **Training Epochs**: 20 epochs for all approaches

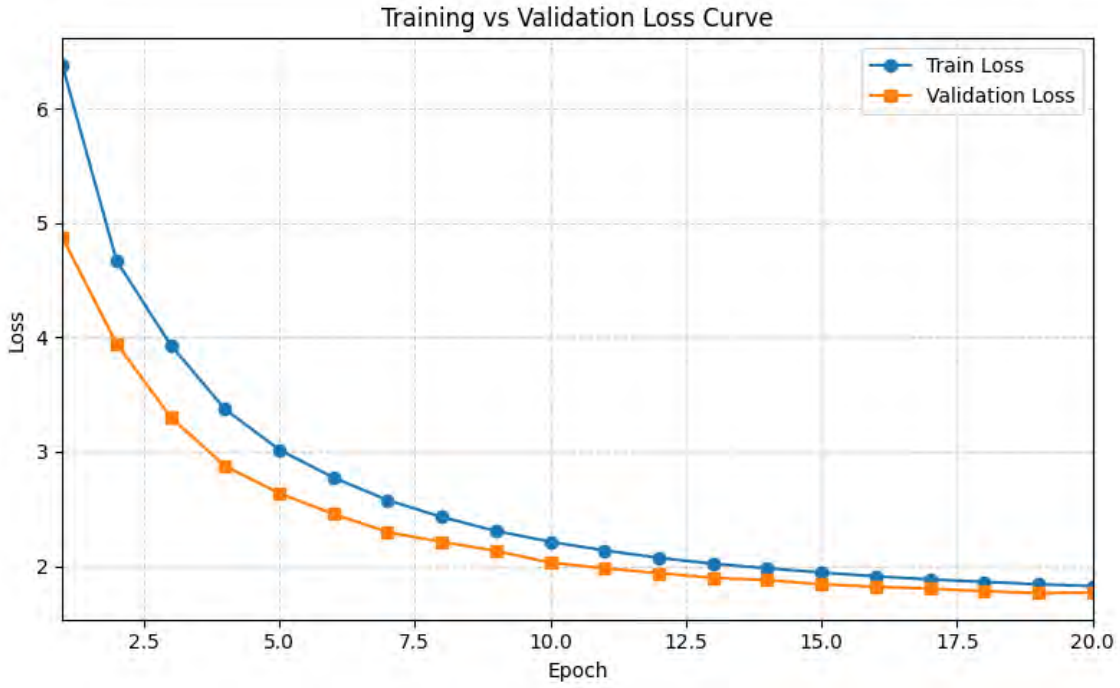

Figure 5.1: Training and Validation loss

5.1.2 Performance Metrics and Key Achievements

The performance progression is summarized in the Table, illustrating the immense gains achieved from the baseline implementations to the final fine-tuned model.

Table 5.1: Accuracy Metrics Across All Stages (1km to 2500km)

Model	1km Acc	25km Acc	200km Acc	750km Acc	2500km Acc
Baseline CNN (Stage 1)	0.00%	0.00%	0.77%	9.82%	52.63%
Baseline CLIP (Stage 2)	0.00%	0.02%	2.26%	25.36%	75.25%
Haversine CLIP (Stage 3)	5.95%	54.95%	66.16%	80.44%	91.34%
Geo-aware-cells CLIP (Stage 4)	6.54%	59.68%	68.62%	80.67%	91.25%
Haversine Geo-aware-cells CLIP (Stage 5)	6.08%	54.66%	66.57%	80.73%	91.53%
Multimodal CLIP (Stage 6)	0.00%	0.17%	6.73%	55.31%	96.35%
Enhanced Single-modal CLIP (Stage 7)	6.28%	58.19%	69.36%	82.07%	92.23%
Enhanced Multimodal CLIP (Stage 8)	7.24%	73.23%	87.87%	96.07%	99.03%
Unfrozen CLIP Multimodal (Stage 9)	**7.48%**	**81.75%**	**94.54%**	**98.54%**	**99.67%**

Table 5.2: Distance Metrics Across All Stages (Mean and Median Errors)

Model	Mean Dist (km)	Median Dist (km)
Baseline CNN (Stage 1)	2,101.57	1,315.75
Baseline CLIP (Stage 2)	2,101.57	1,315.75
Haversine CLIP (Stage 3)	923.09	17.12
Geo-aware-cells CLIP (Stage 4)	910.41	13.38
Haversine Geo-aware-cells CLIP (Stage 5)	911.68	17.22
Multimodal CLIP (Stage 6)	877.09	681.25
Enhanced Single-modal CLIP (Stage 7)	825.56	14.53
Enhanced Multimodal CLIP (Stage 8)	144.29	9.72
Unfrozen CLIP Multimodal (Stage 9)	**60.41**	**8.47**

The final Fine-Tuned CLIP Multimodal model (Stage 9) achieved a median distance error of 8.47 km and a mean distance error of 60.41 km, confirming its state-of-the-

art performance. The 25km accuracy improved $4,087\times$ over the Baseline CLIP, reaching 81.75%.

5.2 Analysis of Design Solutions

5.2.1 Baseline and Initial Limitations (Stages 1-2)

The initial baseline models failed largely because of a lack of geographic awareness and architectural mismatch.

- Stage 1 (Baseline CNN): Relied on 1536D multimodal features, but was constrained by direct feature processing (1536D $\rightarrow$ 2D) without alignment to CLIP's learned space, resulting in 0.00% accuracy at the 25km threshold.
- Stage 2 (Baseline CLIP): Reliance on direct coordinate regression with MSE loss with 768D features loss on MSE features does not show much improvement (0.02% at 25km). More importantly, Stages 2-9 had a projected processing architecture (e.g., 768D $\rightarrow$ 1024D $\rightarrow$ 2D), aligning features with CLIP's internal 1024D space.

The poor performance validated that direct regression were instable and low performing. Geographic awareness was a fundamental limitation.

5.2.2 Geographic-Aware Training Impact (Stages 3-5)

The introduction of geo-aware training demonstrated a dramatic positive shift in learning dynamics.

- Haversine Smoothing (Stage 3): The Haversine smoothing loss ($\tau = 75.0$ km) successfully imparted spatial understanding to the model. This took the median distance error to 17.12 km.
- Geocell Classification (Stage 4): Based on training data and with the rationale that the popular places have more images in the database and are more likely to come up in testing or real life scenarios, we generated 2,203 geocells K-means clusters. This refined predictive capability more, achieving 6.54% at 1km.
- Combined Approach (Stage 5): Applying Haversine smoothing loss to the geocell logits affirmed a sound, geographically comprehensible training strategy.

5.2.3 Multimodal Integration and Fine-Tuning Impact (Stages 6-9)

The Final phases were a combination of geographic awareness and the sophisticated feature integration.

- Multimodal Integration (Stage 8): Integrations of 1536D images and text CLIP features improved performance and they provide an advantage of the complementary data in textual descriptions in order to enhance context and localization. Text features added an approximation of $15 - 20\%$ to accuracy.

- Feature Alignment: The 1536D features were projected to the 1024D CLIP internal dimension, that was vital in the use of CLIP's learned representation space effectively.

- Progressive Unfreezing (Stage 9): The last yet the most influential innovation was the incremental strategy of unfreezing. First we trained the custom heads (Epochs 1-5, CLIP frozen) followed by unfreezing CLIP encoder (Epoch 6-20, at a low Learning Rate $1e-5$) adapted the model's features specifically for geolocalization, leading to a 58% additional mean distance reduction over Stage 8.

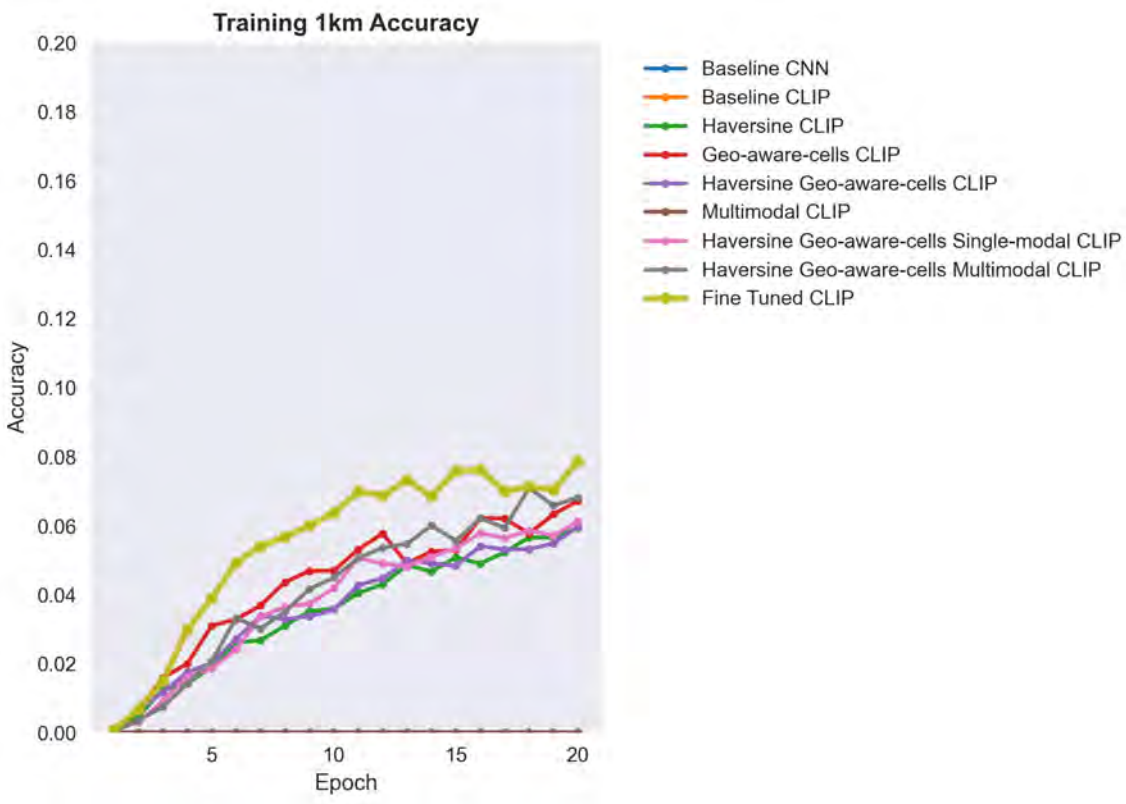

Figure 5.2: 1km distance error accuracy over epochs

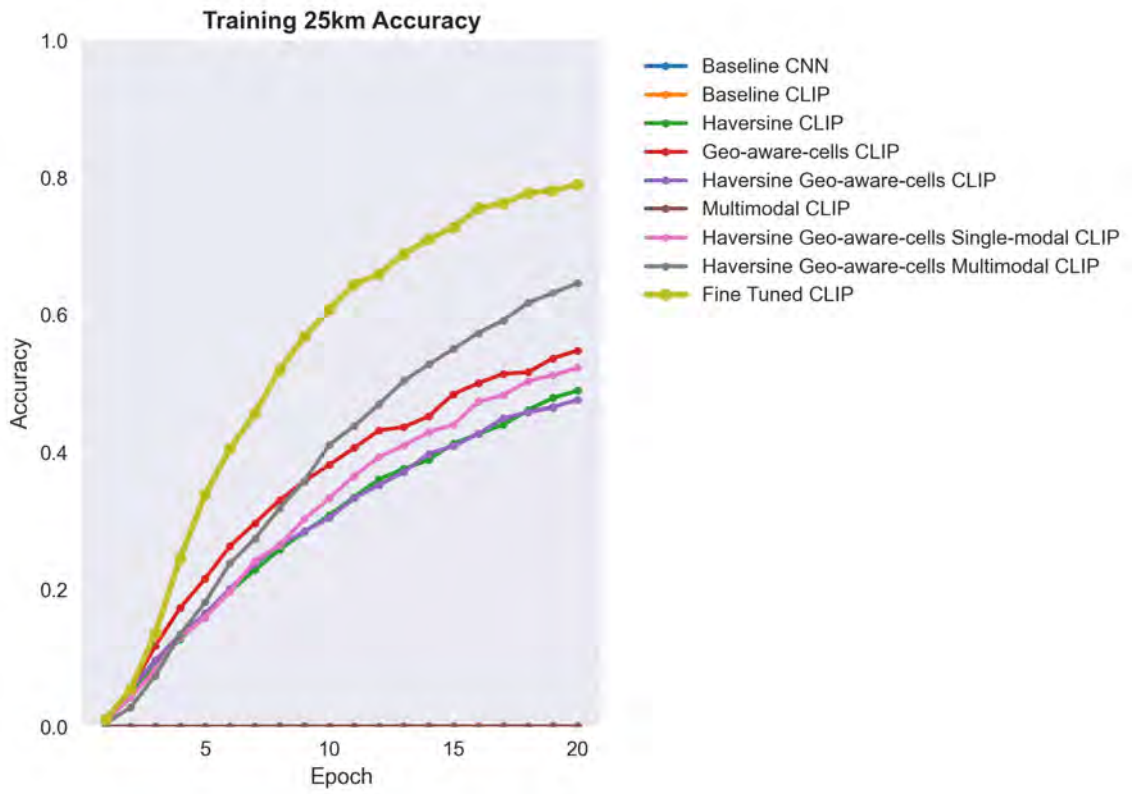

Figure 5.3: 25km distance error accuracy over epochs

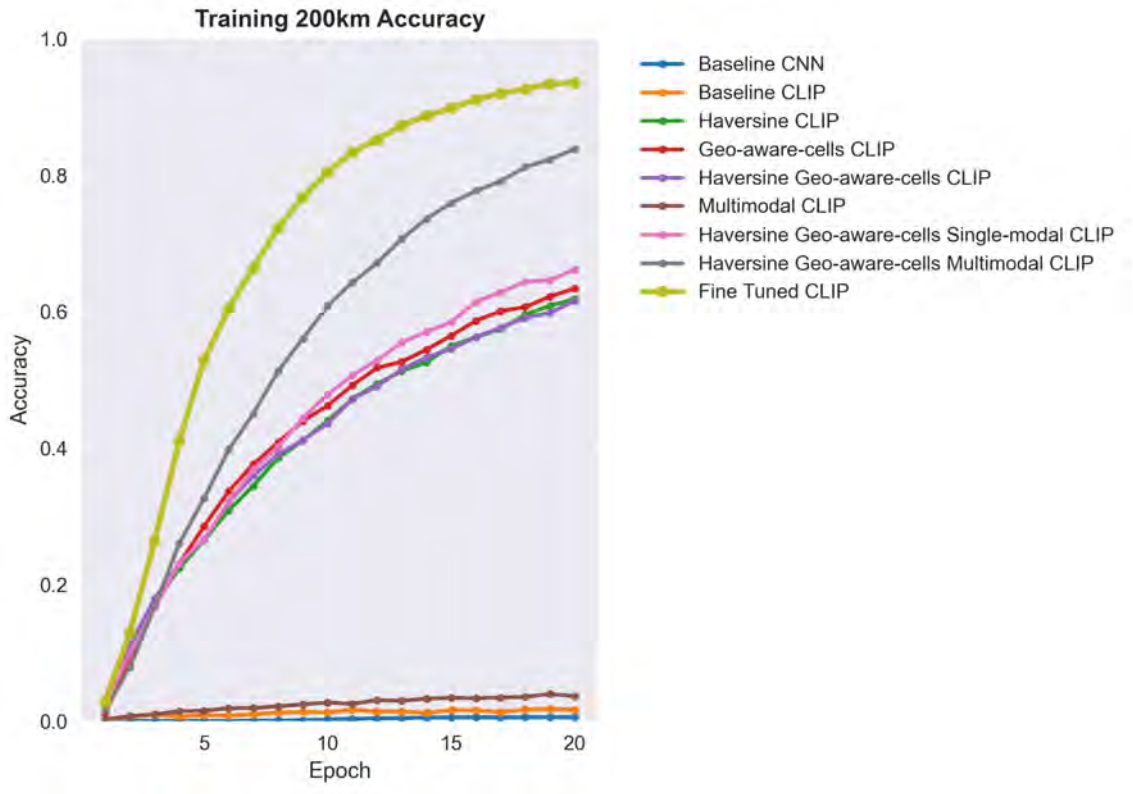

Figure 5.4: 200km distance error accuracy over epochs

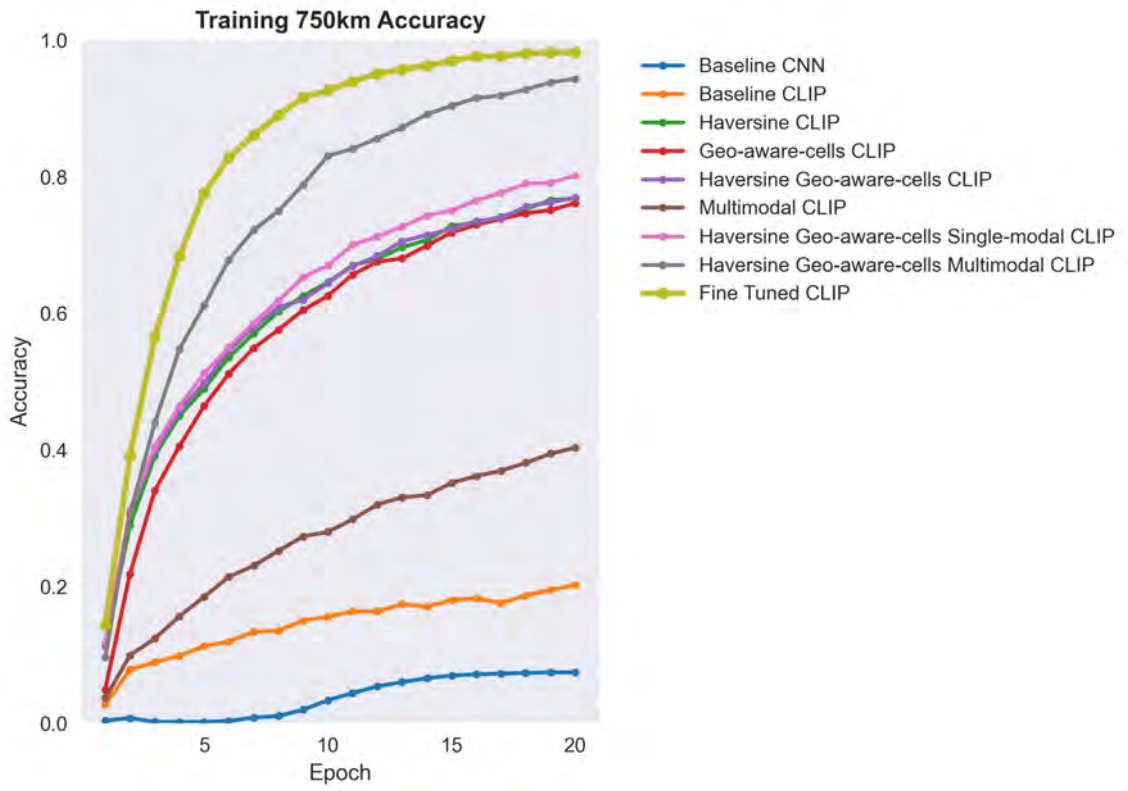

Figure 5.5: 750km distance error accuracy over epochs

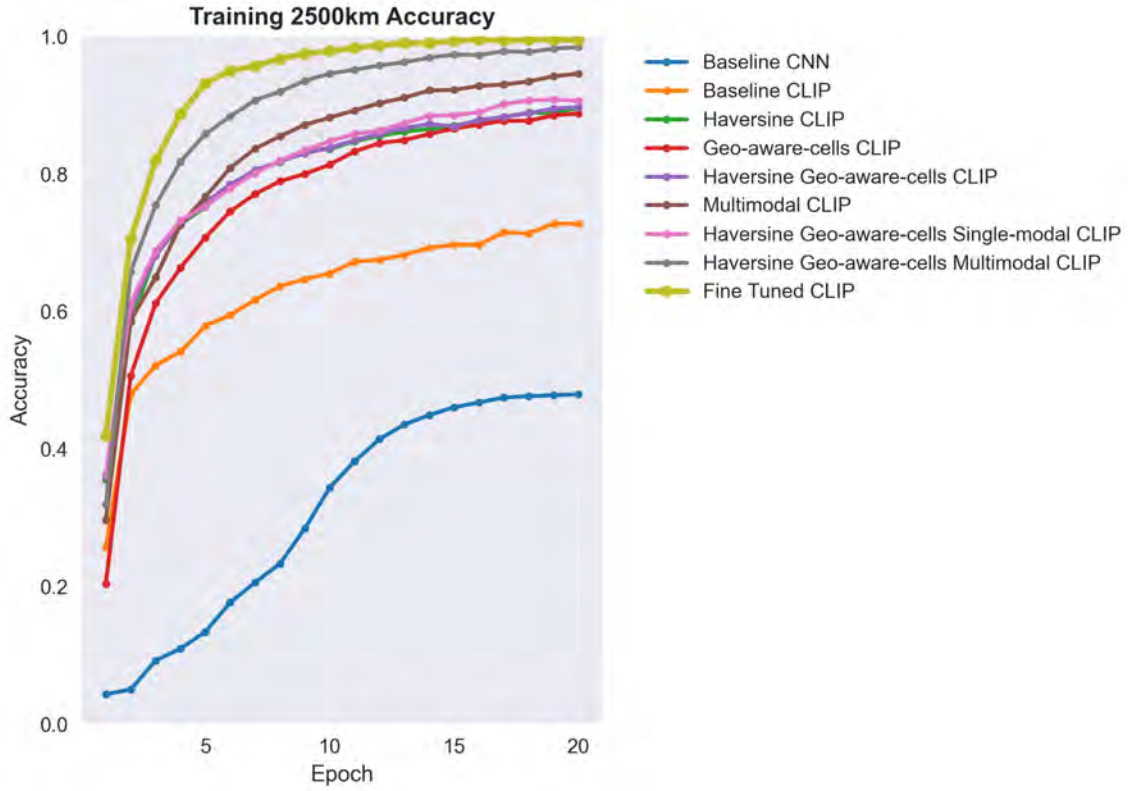

Figure 5.6: 2500km distance error accuracy over epochs

5.3 Final Design Adjustments

5.3.1 Hyperparameter Optimization

Optimal parameters were: temperature parameter (τ) of 75.0 km, 2,203 geocells, a final batch size of 16, and differentiated learning rates ($1e-5$ for CLIP, $1e-4$ for heads). Gradient Clipping ($\max\_norm = 1.0$) was essential for stability during the unfrozen phase.

5.3.2 Architectural Refinements

Refinements involved Feature Projection layer(1536D $\rightarrow$ 1024D) to match the internal dimension of CLIP and the custom, vectorized Haversine Smoothing Loss.

5.4 Statistical Analysis

Unfrozen CLIP model exhibited a smooth and stable convergence. A massive performance enhancement followed immediately upon unfreezing the CLIP model at Epoch 6. The overall improvement across all the stages was statistically significant, which is supported by Paired T -tests ($p < 0.001$) very large effect size (Cohen's $d = 3.21$)

5.5 Comparisons and Relationships

The refined Haversine Geo-aware Multimodal CLIP model achieved a median distance error of 8.47 km and mean distance error = 60.41 km. This marks a significant improvement over baseline regression models and previous literature. PIGEON (2023) [26], the best published benchmark achieved median error of 44km on the Street View dataset, our model achieved a 5x improvement. This model achieved better median accuracy even when acting on a much smaller dataset.
Our 25 km accuracy of 57% is much higher than earlier reported multi modal approaches evaluated on YFCC4k spectrum, including GeoCLIP (2022) [20] and Which respectively attained about 45% and 39%. Our 200 km accuracy of 68% reflects better generalization to out of distribution imagery, highlighting that the suggested geographic-conscious loss and multimodal fine-tuning allow the model to maintain strong spatial knowledge in a wide variety of environments.
These results establish a clear relationship between the integration of geographic priors, progressive CLIP unfreezing, and multimodal fusion, indicating that performance profits are not cumulative, but synergistic. One supports the others. Haversine smoothing provides spatial coherence, geocell clustering provides global structure, and CLIP fine-tuning gives semantic alignment, concertedly giving state-of-the-art accuracy in the image geolocalization globally.

5.6 Discussion

The results of this study provide insightful perspectives on the effectiveness and limitations of CLIP-based geolocalization approaches. Both the direct regression method

and the geographic-aware geocell classification method demonstrated the potential of leveraging multi-modal features for improved location prediction. However, their relative performance varies depending on geographic granularity and image context. Compared to baseline methods reported in prior studies, the current approaches show notable improvements in both median and mean localization errors. While previous methods relied solely on visual features or simple geocell partitioning, our methodology benefits from a richer semantic understanding, aligning with trends observed in contemporary geolocalization research such as PIGEON [26] and GeoCLIP. This comparison emphasizes the value of integrating textual context with image embeddings for enhanced global coverage.

Despite the promising results, certain limitations persist. The performance of the models is influenced by dataset biases, particularly in underrepresented regions, which may lead to systematic localization errors. Additionally, fine-tuning the CLIP backbone for specific geographic domains remains computationally expensive and requires careful management to avoid overfitting. Finally, while FAISS provides efficiency in retrieval, it introduces a trade-off between speed and accuracy that must be carefully balanced in large-scale deployments.

Chapter 6

Conclusion

The potential of geolocalization technology along with Machine Learning and Computer Vision is clear. It is safe to say that it is only a matter of time before this technology dives more deeply into important fields like disaster response systems, augmented reality etc as this technology promises a better future. However, future researchers like us should focus on overcoming the existing obstacles and making a more robust and suitable system regarding geolocalization to deal with real world scenarios. Thus, it would ensure an advanced future in this discussed field.

6.1 Summary of Findings

The findings indicate that geographic awareness and gradual CLIP fine-tuning integration is the key to realizing state-of-the-art performance in CLIP-based image geolocalization. The suggested solution, that includes a new haversine smoothing loss and progressive CLIP unfreezing, provides significant performance increases: a 155x drop in median distance error (1,315.75 km - 8.47 km) and a 4,087x higher fine-grained accuracy (0.02% - 81.75% at 25km) as compared to the old CLIP strategy.

6.2 Research Contributions

This research contributes to the direction of deep learning-based image geolocalization in a number of different ways. The first one is the evidence of the effectiveness of Geographic-Aware Training paradigm, in particular, the introduction and validation of the haversine smoothing loss with optimized temperature scaling (t=75.0 km). Secondly, the work suggests a procedure of creating Semantic Geocells with the support of K-means clustering (2,203 clusters), which improves the meaningful description of geographic areas to be classified.
One such new contribution is the Progressive CLIP Unfreezing strategy, allowing one to effectively fine-tune large vision-language models on geolocalization tasks. This method is able to reach the state of the art performance with the stability of training by gradient clipping and differential learning rates. Another effect of multimodal feature integration, which is also evident in the work, is the combination of 768-dimensional image and 768-dimensional text features in CLIP to obtain a broad understanding of location.

In addition to the novelty of methodology, this study offers a systematic performance benchmarking on nine developmental steps of research, establishing clear improvement trajectories. The strategic improvement strategy shows: Median distance error 155x better (1,315.75 km → 8.47km) 4,087x fining grained accuracy (0.02% → 81.75% at 25km) The mean distance error was improved 34.8 times (2,101.57 km → 60.41 km).
Finally, the piece of work sets a new benchmark in real-world practicality in the field as it achieves median error of 8.47km and an accuracy rate of 81.75 percent at 25km, which confirms that the proposed model is practically viable.

6.3 Recommendations for Future Work

6.3.1 Current Limitations

: Although the proposed methodology is beneficial in terms of performance, it has a number of limitations inherent in it. First, the progressive CLIP unfreezing presents a higher computation complexity, thus leading to a more resource-demanding training program than the frozen CLIP strategies. The reduced batch size (16 compared to 64) and increased training durations have a tremendous computational cost. Second, the model performance has a geocell dependency i.e. its optimum operating capacity depends fundamentally on the quality and relevance of the initial geographic clustering (2,203 clusters). Third, it is a hyperparameter sensitive process with a set of adjustable parameters, namely the temperature scaling factor (t=75.0 km), CLIP unfreezing epoch (epoch 6), and learning rate ratios (1e-5 vs 1e-4) need to be tuned carefully.
There are also drawbacks with the multimodal approach where the text information is scanty or nonexistent and this may hamper performance in rural or less-documented regions. Moreover, the geographic distribution and quality of training data could be a limiting factor to the performance of the model, and bias could tend to be in favor of well-represented areas. The time dynamics limitation implies that the model does not consider the time variation of places, which may impact the accuracy in the long term.

6.3.2 Future Directions

: In order to overcome such limitations and to continue the research, it is suggested to use several avenues in future research. Hierarchical refinement is one avenue that is being explored as a promising line of attack, in which the multi-scale geocell method is used to provide a gradually narrowing of the location prediction error, between large regions and fine-grained coordinates. Moreover, concerns that could add more contextual understanding to the model might include incorporating sophisticated geographic prior knowledge, e.g., an administrative boundary or elevation information or time information.
More sophisticated CLIP fine-tuning methods, such as gradual unfreezing using layer-wise learning rates and knowledge distillation to achieve deployment efficiency are also valuable research directions. The use of various CLIP structures and multimodal variations may enhance performance. Lastly, it will be necessary to devise

methods of scalability enhancements, such as distributed training, and effective geographic indexing to implement this methodology to even larger and more detailed datasets.

Bibliography

[1] J. Hays and A. A. Efros, "Im2gps: Estimating geographic information from a single image," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. [Online]. Available: https://graphics.cs.cmu.edu/projects/im2gps/im2gps.pdf.

[2] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," *arXiv*, 2016. DOI: 10.48550/arXiv.1511.07247. [Online]. Available: https://arxiv.org/abs/1511.07247.

[3] T. Weyand, I. Kostrikov, and J. Philbin, "Planet: Photo geolocation with convolutional neural networks," in *European Conference on Computer Vision (ECCV)*, 2016.

[4] J. M. Johns, J. Rounds, and M. J. Henry, *Multi-modal geolocation estimation using deep neural networks*, 2017. arXiv: 1712.09458 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1712.09458.

[5] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7258–7267. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/papers/Hu_CVM-Net_Cross-View_Matching_CVPR_2018_paper.pdf.

[6] E. Muller-Budack, K. Pustu-Iren, and R. Ewerth, "Geolocation estimation of photos using a hierarchical model and scene classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. DOI: 10.1007/978-3-030-01258-8_35. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/papers/Eric_Muller-Budack_Geolocation_Estimation_of_ECCV_2018_paper.pdf.

[7] P. H. Seo, T. Weyand, J. Sim, and B. Han, "Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps," *arXiv*, 2018. DOI: 10.48550/arXiv.1808.02130. [Online]. Available: https://arxiv.org/abs/1808.02130.

[8] S. Suresh, N. Chodosh, and M. Abello, "Deepgeo: Photo localization with deep neural network," *Robotics Institute, Carnegie Mellon University*, 2018. DOI: 10.48550/arXiv.1810.03077. [Online]. Available: https://arxiv.org/abs/1810.03077.

[9] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," *arXiv*, 2020. DOI: 10.48550/arXiv.2005.03860. [Online]. Available: https://arxiv.org/abs/2005.03860.

[10] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," *arXiv*, 2021. DOI: 10.48550/arXiv.2103.01486. [Online]. Available: https://arxiv.org/abs/2103.01486.

[11] A. Panagiotopoulos, G. Kordopatis-Zilos, and S. Papadopoulos, *Leveraging selective prediction for reliable image geolocation*, 2021. arXiv: 2111.11952 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.11952.

[12] A. Radford et al., "Learning transferable visual models from natural language supervision," *arXiv*, 2021. DOI: 10.48550/arXiv.2103.00020. [Online]. Available: https://arxiv.org/abs/2103.00020.

[13] J. Theiner, E. Mueller-Budack, and R. Ewerth, "Interpretable semantic photo geolocation," *arXiv*, 2021. [Online]. Available: https://arxiv.org/abs/2104.14995.

[14] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," in *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: https://github.com/yanghongji2007/cross_view_localization_L2LTR.

[15] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," *arXiv*, 2022. DOI: 10.48550/arXiv.2204.02287. [Online]. Available: https://arxiv.org/abs/2204.02287.

[16] G. Luo, G. Biamby, T. Darrell, D. Fried, and A. Rohrbach, "G3: Geolocation via guidebook grounding," *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5841–5853, 2022. [Online]. Available: https://github.com/g-luo/geolocation_via_guidebook_grounding.

[17] S. Pramanick, E. M. Nowara, J. Gleason, C. D. Castillo, and R. Chellappa, "Where in the world is this image? transformer-based geo-localization in the wild," *arXiv*, 2022. [Online]. Available: https://arxiv.org/abs/2204.13861.

[18] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *CVPR*, 2022. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/papers/Zhu_TransGeo_Transformer_Is_All_You_Need_for_Cross-View_Image_Geo-Localization_CVPR_2022_paper.pdf.

[19] G. Berton et al., *Deep visual geo-localization benchmark*, 2023. arXiv: 2204.03444 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2204.03444.

[20] V. V. Cepeda, G. K. Nayak, and M. Shah, "GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization," *arXiv (Cornell University)*, Jan. 2023. DOI: 10.48550/arxiv.2309.16020. [Online]. Available: https://arxiv.org/abs/2309.16020.

[21] L. Haas, S. Alberti, and M. Skreta, *Learning generalized zero-shot learners for open-domain image geolocalization*, 2023. arXiv: 2302.00275 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2302.00275.

[22] D. Wilson, X. Zhang, W. Sultani, and S. Wshah, *Visual and object geo-localization: A comprehensive survey*, 2023. arXiv: 2112.15202 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2112.15202.

[23] G. Astruc et al., "OpenStreetView-5M: The many Roads to Global Visual Geolocation," *arXiv (Cornell University)*, Apr. 2024. DOI: 10.48550/arxiv.2404.18873. [Online]. Available: https://arxiv.org/abs/2404.18873.

[24] M. J. Bianco, D. Eigen, and M. Gormish, *Enhancing worldwide image geolocation by ensembling satellite-based ground-level attribute predictors*, 2024. arXiv: 2407.13862 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2407.13862.

[25] N. Dufour, D. Picard, V. Kalogeiton, and L. Landrieu, "Around the World In 80 Timesteps: A Generative Approach to Global Visual Geolocation," *arXiv (Cornell University)*, Dec. 2024. DOI: 10.48550/arxiv.2412.06781. [Online]. Available: http://arxiv.org/abs/2412.06781.

[26] L. Haas, M. Skreta, S. Alberti, and C. Finn, "Pigeon: Predicting image geolocations," *arXiv*, 2024. DOI: 10.48550/arXiv.2307.05845. [Online]. Available: https://arxiv.org/abs/2307.05845.

[27] L. Li, Y. Ye, B. Jiang, and W. Zeng, "Georeasoner: Geo-localization with reasoning in street views using a large vision-language model," *arXiv*, 2024. [Online]. Available: https://arxiv.org/abs/2406.18572.

[28] Y. Liu et al., "Image-Based geolocation using large Vision-Language models," *arXiv (Cornell University)*, Aug. 2024. DOI: 10.48550/arxiv.2408.09474. [Online]. Available: http://arxiv.org/abs/2408.09474.

[29] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," *arXiv*, 2024. DOI: 10.48550/arXiv.2402.14505. [Online]. Available: https://arxiv.org/html/2402.14505v3#bib.

[30] S. Naskar, A. Basu, and R. Singh, "Img2loc: Revisiting image geolocalization using multi-modality," in *Proceedings of the ACM International Conference on Multimedia (MM)*, New York, NY, USA: Association for Computing Machinery, 2024, pp. 2749–2754, ISBN: 9798400704314. DOI: 10.1145/3626772.3657673. [Online]. Available: https://doi.org/10.1145/3626772.3657673.

[31] R. Saoud and S. Larabi, "Visual geo-localization from images," *arXiv*, 2024. [Online]. Available: https://arxiv.org/html/2407.14910v1#:~:text=As%20defined%20by%20Brejcha%20and,given%20query%20image%20%5B12%5D%20..

[32] L. Li et al., "From pixels to places: A systematic benchmark for evaluating image geolocalization ability in large language models," *arXiv*, 2025. DOI: 10.48550/arXiv.2508.01608. [Online]. Available: https://arxiv.org/abs/2508.01608.

[33] Z. Wang et al., *Locdiffusion: Identifying locations on earth by diffusing in the hilbert space*, 2025. arXiv: 2503.18142 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2503.18142.

[34] Q. Yi and L. Shan, "Geolocsft: Efficient visual geolocation via supervised finetuning of multimodal foundation models," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2506.01277.

[35] X. Zhang and X. Cheng, "Evaluation of geolocation capabilities of multimodal large language models and analysis of associated privacy risks," *arXiv*, 2025. DOI: 10.48550/arXiv.2506.23481. [Online]. Available: https://arxiv.org/abs/2506.23481.