Hierarchical Image Geolocalization Using ViT-based Encoders and Satellite Map Images

Yoon Shik Kim* Jae Won Jang* Cheong Su Kim
Seoul National University
{yoonshik, pert0407, cjd1125}@snu.ac.kr

Abstract

Image geolocalization, the task of determining the precise geographical coordinates of an image, remains a challenging problem in computer vision. Traditional approaches often struggle with generalization across diverse datasets and exhibit significant urban-rural disparities. In this study, we introduce a novel geolocalization model that leverages user-submitted photographs and satellite map images. Our model uses a ViT-based hierarchical prediction pipeline consisting of multiple zoom levels to progressively narrow down the location of an image. The photo encoder extracts feature embeddings from the input image, which are then compared with embeddings from satellite images processed by map encoders at various zoom levels. This iterative process aims to improve geolocation precision.

Our preliminary results indicate that while the model's performance does not surpass state-of-the-art methods in terms of accuracy, it provides valuable insights and a new approach to the geolocalization task. The research highlights the challenges in data collection and integration, and suggests directions for future improvements. Additionally, we propose future work incorporating GeoCLIP, a well-performing model, to enhance performance by leveraging its robust initial predictions and harmoniously integrating it with our model during the training phase.

1. Introduction

Image geolocalization, the task of determining the geographical location of a photograph, poses significant challenges in the realm of computer vision. This problem is exemplified by the popular online game GeoGuessr, where players must deduce locations from panoramic Street View images. The complexity of image geolocalization arises from the vast diversity and dynamic nature of Earth's landscapes, compounded by seasonal variations and the impacts of climate change.

Recent advances in the field have treated image geolocalization primarily as a classification problem, employing hierarchical models and state-of-the-art techniques like vision transformers and contrastive pretraining to enhance accuracy. Despite these advancements, a critical challenge remains: models often struggle to generalize across highly diverse and previously unseen datasets, highlighting the need for methodologies that go beyond training-test distribution alignments. Furthermore, the urban-rural disparity in image data complicates model training, with urban areas being overrepresented compared to less-documented rural or undeveloped regions.

To address these challenges, our study proposes a novel approach that leverages map tile images from the internet, consisting of both satellite imagery and overlaid text and symbols, alongside photograph to improve geolocalization accuracy. By incorporating map images with distinctive features such as sea, land, and city boundaries, our method enriches the data pool, allowing for a more nuanced interpretation of the geospatial elements present in user images.

Our model utilizes ViT-based image encoders in a 7-stage hierarchical prediction pipeline to predict the location of each photo. Specifically, the model consists of a "photo encoder," which is an unmodified CLIP image encoder for extracting embeddings from images, and seven "map encoders," which have been fine-tuned on satellite map tile images for each of the seven zoom levels in our hierarchical prediction pipeline. This design allows the model to iteratively narrow down the search area by comparing photo embeddings with map tile embeddings at progressively higher zoom levels.

In our prediction pipeline, the map tiles at each zoom level are divided into a 4x4 grid, and the top k similar patches are selected based on cosine similarity with the photo. These selected patches are then further divided and passed through the next zoom level's map encoder, iteratively refining the geolocation prediction. This process enhances the model's ability to accurately pinpoint locations by leveraging both visual cues from photos and detailed geographic information from map tiles.

The proposed approach demonstrates high performance in initial experiments, with the model showing significant

*Equal contribution

improvements over traditional methods. However, it is worth noting that our model exhibits lower performance compared to some state-of-the-art models on the given Im2GPS benchmark. Future work will focus on addressing these limitations and refining the model to better suit the task.

In summary, our research presents a promising direction for enhancing image geolocalization by integrating photo with satellite map tile images and employing a hierarchical prediction pipeline.

2. Related Work

2.1. Image Geolocalization Problem Setting

Image geolocalization, the complex challenge of deriving geographical coordinates from visual data, remains a critical focus in computer vision. Historically, traditional methods like IM2GPS utilized hand-crafted features and relied on extensive databases for nearest-neighbor retrieval, but these approaches struggled with scalability and were impractical for global application due to the vast data requirements [3]. These methods faced significant limitations in handling the diversity and volume of data required for effective global geolocalization, reflecting the early challenges in the field.

The fundamental complexity of image geolocalization arises from the need to accurately interpret varied and often ambiguous visual cues within diverse environments. Factors such as changes in lighting, weather conditions, and seasonal variations further complicate this task. Initially, the field relied on static databases that could not effectively adapt to the dynamic nature of global landscapes, often resulting in poor generalization beyond the specific regions represented in the training data.

Moreover, traditional image geolocalization techniques were constrained by their dependency on clear, distinguishable landmarks that are not universally present in all geographic locations. This reliance on distinct features meant that rural or undeveloped areas, which typically lack such landmarks, were notably challenging for early geolocalization systems. The initial problem setting in image geolocalization thus required a shift from reliance on extensive image libraries to more adaptive and scalable solutions capable of dealing with the inherent variability and complexity of global environments.

Recent advancements in image geolocalization have addressed many challenges by utilizing deep learning techniques and leveraging large-scale datasets. Hierarchical models, vision transformers, and contrastive pretraining methods have significantly improved accuracy. However, these models often struggle to generalize across various and unseen datasets, indicating a need for approaches that can effectively overcome training-test distribution discrepancies.

Our study aims to further enhance the accuracy and robustness of image geolocalization by introducing a sophisticated hierarchical prediction model. This model leverages both photograph and satellite map images to provide a comprehensive geolocation solution. The process begins with a Vision Transformer (ViT)-based photo encoder, which extracts detailed feature embeddings from the input images. These embeddings are subsequently compared with those generated by map encoders, fine-tuned on satellite map tiles at different zoom levels. Starting from a broad zoom level, the model iteratively refines the search area by selecting the top k tiles that exhibit the highest cosine similarity with the photo embeddings. This iterative narrowing down process continues through progressively finer zoom levels, enhancing the precision of geolocation predictions.

Incorporating satellite map images enriched with distinctive features such as textual indications of sea, land, and city boundaries allows our model to interpret complex geospatial elements with greater accuracy. This methodology provides a nuanced understanding of the geographical context in user images, addressing the shortcomings of models that depend solely on photograph. By integrating these diverse data sources, our approach enhances the model's ability to generalize across varied environments.

2.2. Vision Transformers for Geolocalization

The evolution of image geolocalization has been significantly influenced by the shift from traditional methodologies to deep learning-based approaches. Initially, methods depended heavily on manually crafted features and extensive image databases. However, the field has since transitioned to more holistic end-to-end learning strategies that leverage the power of deep neural networks, fundamentally transforming geolocalization practices (Masone Caputo, 2021)[9].

Introduced by Google in 2016, the PlaNet model marked a pivotal turn by utilizing convolutional neural networks (CNNs) to classify geographic locations into predefined 'geocells' (Weyand et al., 2016)[19]. This approach stemmed from the challenges associated with using regression models for direct geographic coordinate prediction, which struggled due to the intricate variations in geographic data and the complex relationships between coordinates.

As deep learning technology advanced, it facilitated a renewed examination of the IM2GPS system (Vo et al., 2017)[17], inspired the application of CNNs on expansive mobile image datasets (Howard et al., 2017)[6], and even integrated these models into competitive settings such as the GeoGuessr game, where AI competes against human players. The integration of classification and retrieval methods has also been refined, adopting a hierarchical retrieval framework that mirrors prototypical networks with fixed parameters (Kordopatis-Zilos et al., 2021)[7].

With the emergence of transformer architectures, originally developed for natural language processing (Vaswani et al., 2017)[16], a new avenue has opened in computer

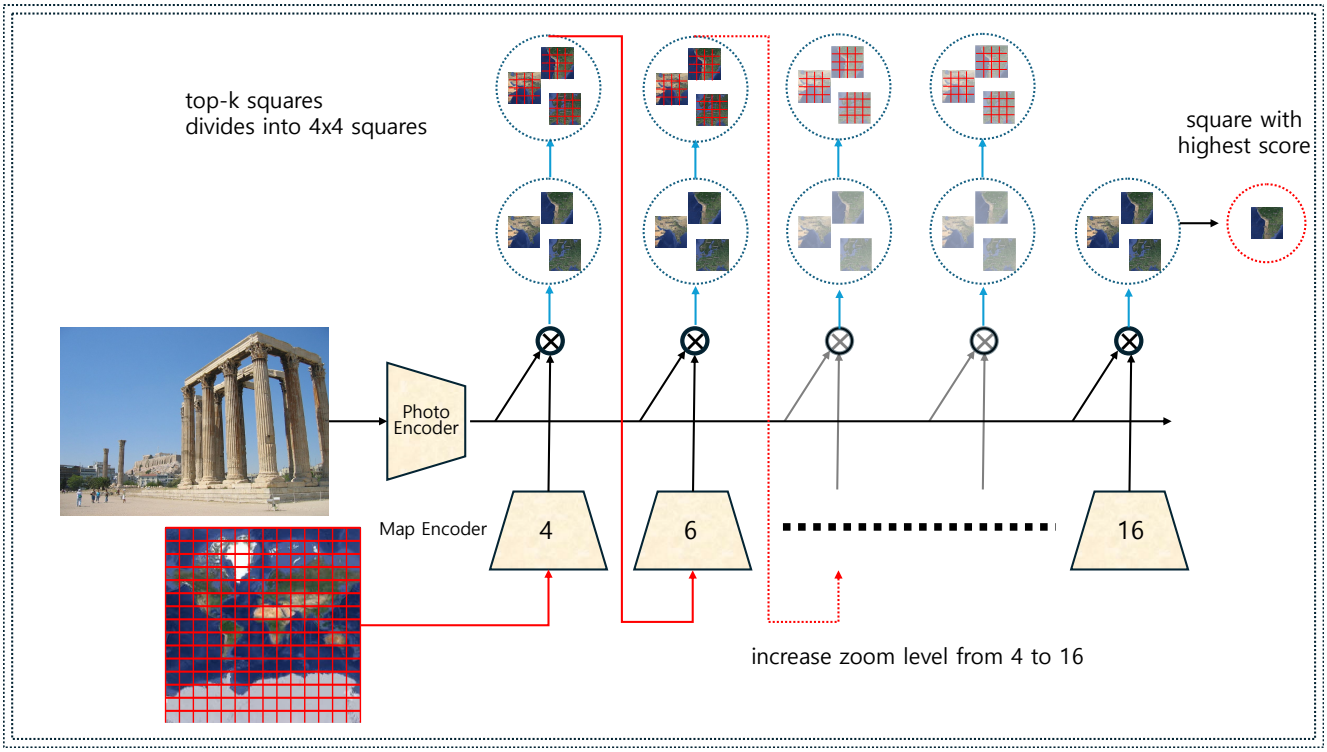

Figure 1. The pipeline of our geolocation model. The process begins with the photo encoder, which extracts feature embeddings from the input image. These embeddings are then compared with embeddings from satellite images processed by the map encoders at multiple zoom levels (4, 6, 8, 10, 12, 14, and 16). At each zoom level, the model selects the top-k satellite images with the highest cosine similarity scores compared to the photo embeddings. The selected satellite images are then divided into 4x4 grids and passed to the next zoom level's map encoder. This process is iteratively refined through higher zoom levels. The final prediction is made by selecting the satellite image with the highest similarity score at zoom level 16, progressively narrowing down the location of the input image with high accuracy.

vision. Vision transformers (ViT) and their multi-modal variants, including OpenAI's CLIP and GPT-4V, have been successfully adapted for image geolocalization, demonstrating substantial improvements in handling complex visual tasks (Kolesnikov et al., 2021; Radford et al., 2021; Agarwal et al., 2021; Pramanick et al., 2022; Luo et al., 2022; Zhu et al., 2022; OpenAI, 2023).[15][13]

To enhance the accuracy and robustness of image geolocalization, our study employs a Vision Transformer (ViT) architecture, specifically utilizing the CLIP image encoder with the ViT-B/32 model. Our approach features a multi-stage hierarchical system that processes photograph and satellite map images. The ViT-B/32-based photo encoder extracts detailed feature embeddings from input photograph, providing a robust visual representation.

These photo embeddings are compared against embeddings generated by a series of map encoders, fine-tuned on satellite map tiles at different zoom levels. Starting from a broad zoom level, the model narrows down the search area iteratively by selecting the top k tiles with the highest cosine similarity. This process refines geolocation predictions at progressively finer zoom levels, leveraging the hierarchical structure of map tiles.

2.3. Utilizing CLIP for Geolocalization

Our work builds on recent advances in computer vision by employing OpenAI's CLIP (Contrastive Language–Image Pre-training) model for the task of image geolocalization. CLIP integrates the power of vision transformers with natural language processing, making it highly effective for robust image recognition tasks crucial in geolocalization. This integration allows CLIP to understand and process both visual and textual data, which is particularly advantageous in the complex task of determining geographical locations from images. Notably, CLIP's zero-shot learning capabilities enable it to perform well across a variety of diverse and previously unseen datasets, addressing the challenge of geolocalization where environments can vary dramatically.

CLIP's utility is further enhanced by pretraining it with auxiliary geographic, demographic, and climate data within a multi-task learning framework. This approach significantly improves location accuracy by providing the model with a richer context for interpreting visual data. By broadening the model's exposure to various environmental conditions,

CLIP becomes better equipped to generalize across different geographies. For instance, training on demographic and climate data helps the model recognize patterns associated with specific regions, such as vegetation types in tropical climates or architectural styles in urban areas, thus enhancing its predictive accuracy [18].

Moreover, CLIP can interpret contextual clues that emerge from metadata or textual descriptions associated with geographic locations. This capability allows us to use descriptive location-based cues from data labels or annotations that accompany training images, effectively bridging the gap between visual data and geographical semantics. Terms like "mountainous" or "coastal" found in image descriptions can influence the model in refining its geolocation predictions. By aligning visual and textual data, CLIP can leverage additional contextual information to improve overall accuracy. For example, an image tagged with "desert" can help the model prioritize certain features such as sand dunes or sparse vegetation, which are characteristic of desert landscapes.

The inclusion of CLIP in geolocalization framework demonstrates a significant advancement in applying modern machine learning techniques to traditional computer vision problems. By combining the strengths of vision transformers and natural language processing, CLIP sets a new benchmark for how complex visual and textual data can be integrated to enhance geolocalization tasks. This hybrid approach not only improves the robustness and accuracy of location predictions but also opens new possibilities for further research and application in related fields. The ability to interpret and integrate multi-modal data sources makes CLIP a powerful tool for tackling the inherent challenges of geolocalization, paving the way for more accurate and reliable geospatial data interpretation.

2.4. GeoCLIP

Building upon the CLIP framework, GeoCLIP extends its capabilities by specifically focusing on geolocalization tasks.(2023)[1] GeoCLIP leverages contrastive learning on geolocated image pairs to provide robust initial geolocation predictions. This model is pretrained with auxiliary geographic, demographic, and climate data within a multi-task learning framework, which not only improves location accuracy but also broadens the model's exposure to various environmental conditions, enhancing its ability to generalize across different geographies.

GeoCLIP can interpret contextual clues that emerge from metadata or textual descriptions associated with geographic locations. This capability allows the model to use descriptive location-based cues from data labels or annotations that accompany training images, effectively bridging the gap between visual data and geographical semantics. Terms like "mountainous" or "coastal" found in image descriptions can influence the model in refining its geolocation predictions, aligning visual and textual data to improve overall accuracy.

While we did not integrate GeoCLIP directly into our current model, we recognize its potential for future enhancements. As future work, we propose incorporating GeoCLIP to assist in refining geolocation predictions at higher zoom levels. By leveraging GeoCLIP's initial predictions and integrating them into our hierarchical model, we aim to improve the overall accuracy and robustness of our geolocalization system. This future integration could enhance our model's performance, particularly in complex and diverse environments, further advancing the state of geolocation technologies.

2.5. Semantic Geocell Partitioning and Advanced Geolocalization Approaches

Our methodology advances beyond traditional image-to-image retrieval by employing semantic partitioning of the Earth into classes or "geocells," using hierarchical clustering and Voronoi tessellation to adapt dynamically based on the training dataset distribution. This approach addresses the imbalance in class sizes and enhances model performance across diverse geographical distributions [11]. The PIGEON model further expands on these techniques with its use of multi-task contrastive pretraining that incorporates geographical, demographic, and climate data to improve generalization across various and previously unseen locations, setting a new standard in the field [4].

Additionally, the introduction of GeoCLIP by Vivanco Cepeda et al. represents a significant breakthrough by employing the CLIP model to align image features directly with GPS coordinates, thereby transforming the geolocalization challenge into an image-to-GPS retrieval task. This method not only improves the accuracy of localizing images from diverse locations but also demonstrates a structured learning approach that iteratively refines geolocation estimates, achieving superior performance with reduced training data requirements.

In our study, we employ a simplified yet effective approach to geocell partitioning. Initially, the satellite map is divided into a 16x16 grid at a broad zoom level. As the model processes these tiles, it selects the top k tiles with the highest cosine similarity to the photo embeddings. These selected tiles are then further divided into smaller 4x4 grids, and the process is repeated at progressively finer zoom levels. This hierarchical method allows us to efficiently manage and process large datasets by focusing computational resources on the most relevant areas, thus improving geolocation accuracy. By starting with a broad overview and iteratively refining the search area, our model ensures that even minute details are considered in the final geolocation prediction.

2.6. Cross-view and Multimodal Approaches

In line with the latest research trends, cross-view and multimodal approaches significantly bolster the robustness of geolocalization systems. These techniques merge ground-level and aerial imagery to enrich feature sets, enhancing geolocalization accuracy, particularly in under-documented and rural areas where data may be sparse [20]. By integrating both perspectives, models can leverage a more comprehensive understanding of geographical contexts, which is critical for accurate geolocation in diverse environments. Ground-level images provide detailed, localized information, while aerial or satellite views offer broader contextual insights, making the combined approach more effective.

Recent studies have demonstrated that cross-view learning, where ground-level images are complemented by aerial or satellite views, can greatly enhance the detail and contextual information available to geolocalization models. This approach mitigates the limitations faced by models relying solely on ground-level imagery, which can be sparse or ambiguous, especially in rural or undeveloped regions. The additional perspective provided by aerial views helps to fill in the gaps left by ground-level images, offering a more holistic view of the environment. For instance, while a ground-level image might show a particular building or landscape feature, an aerial view can place that feature within a broader geographic context, making it easier to pinpoint the location accurately.

Multimodal techniques further enhance geolocalization performance by incorporating various data types, such as textual descriptions, climate data, and demographic information. These additional data sources provide valuable context that can refine geolocation predictions, making the models more resilient to variations in environmental conditions and more adaptable to different geographies. By integrating these diverse data sources, multimodal approaches create a richer and more informative dataset for geolocalization models. This not only improves the accuracy of predictions but also enhances the model's ability to generalize across different environments, demonstrating their robustness and adaptability.

3. Method

3.1. Model Architecture

We believe previous approaches to image geolocalization were limited by the lack of common-sense knowledge about the real world. To mitigate this, we propose a model that utilizes map tile images from the internet consisting of both satellite imagery and overlaid text and symbols that can provide crucial zero-shot insight into the cultural and geographical context of each location.

We propose a model that only uses ViT-based image encoders in a 7-stage hierarchical prediction pipeline to predict the location of each photo. Our model consists of 8 total image encoders, all derived from the pre-trained CLIP image encoder that uses the ViT-B/32 architecture. The first encoder, which we designate the "photo encoder", is an unmodified CLIP image encoder for extracting embeddings from images. The remaining 7 encoders, which we designate the "map encoders", have been fine-tuned on the satellite map tile images for each of the 7 zoom levels in our hierarchical prediction pipeline by unfreezing the last two layers of the transformer.

In order to effectively utilize existing map tiles, we set our geocells to correspond to available satellite map tiles from Google Maps, which consist of regions that have been divided from a Mercator projection by powers of two.

In the Google Maps API for map tiles, each tile is uniquely identified using three integer values x, y, and z. z indicates the zoom level, where each side of the square tile corresponds to the length of each side of the original square Mercator map divided by a factor of 2^z. Then, x and y corresponds to the zero-based index of each tile in the horizontal and vertical directions respectively, when the map has been divided into 2^z squares on each side. Hence, the center of the map tile identified by (x, y, z) corresponds to the following latitude ϕ and longitude λ (in degrees):

$$\phi = \frac{180}{\pi}\left[2\arctan\left\{e^{2\pi\left(0.5-\frac{y+0.5}{2^z}\right)}\right\} - \frac{\pi}{2}\right] \quad (1)$$

$$\lambda = 360 \cdot \frac{x+0.5}{2^z} - 180 \quad (2)$$

For the 7 stages of our prediction pipeline, we utilize map tiles of zoom levels 4, 6, 8, 10, 12, 14, and 16. This allows the model to easily identify the smaller tiles of the next level corresponding to each tile of the previous level by simply multiplying the x and y values by 4.

The prediction algorithm consists of the following steps:

Hierarchical Prediction

function PREDICT(X) ▷ X is the target photograph
  $k \leftarrow 64$
  $C \leftarrow \{$all tiles where $z = 4\}$
  $p \leftarrow$ PhotoEncoder(X)
  for $z = 4, 6, 8, 10, 12, 14, 16$ **do**
    $M \leftarrow \{\text{MapEncoder}_z(c) | c \in C\}$
    $S \leftarrow \{\text{CosineSimilarity}(p, m) | m \in M\}$
    if $z = 16$ **then**
      $i \leftarrow \text{argmax}(S)$
      return GetCenterLatLong(C_i)
    $I \leftarrow \{i \mid |\{s | s \in S, s \geq S_i\}| \leq k\}$
    $C \leftarrow \{C_i | i \in I\}$ ▷ narrow C down to top k
    $C \leftarrow \bigcup\{\text{GetNextLevelTiles}(c) | c \in C\}$

	Distance (% @ km)					Inference Time
	Street 1 km	**City 25 km**	**Region 200 km**	**Country 750 km**	**Continent 2,500 km**	
k=16	5.91	11.81	22.36	42.62	64.14	8'38"
k=32	6.35	11.11	15.87	28.57	55.56	20'48"
k=64	5.48	12.33	16.44	32.88	58.90	30'33"

Table 1. Inference Time and Performance Metrics for Different k Values

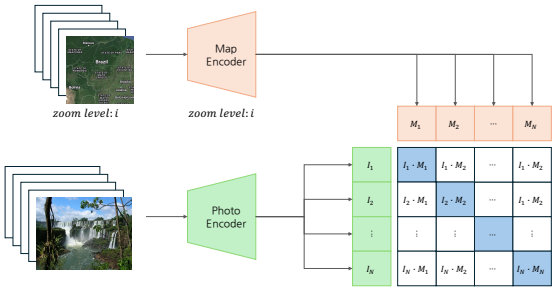

Figure 2. In the training process, we employed the contrastive learning method of the existing open model CLIP. The user's image and the corresponding guided patch were paired, and training was conducted to increase the cosine similarity of the two embeddings. As a result, the last two layers of the map encoder were trained.

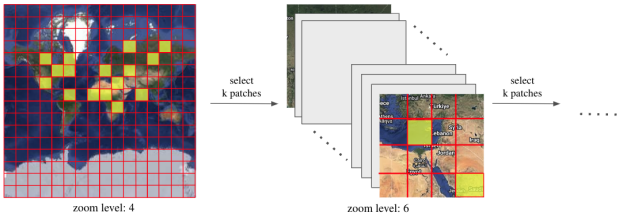

Figure 3. It describes the initial part of the model's inference process. Map patches at zoom level 4, divided from the original map into 16x16, pass through our model, leaving only the top 16 maps with high cosine similarity. This process continues iteratively.

3.2. Model Training

For each zoom level, the corresponding map encoder is trained by framing the task of matching each photo with the correct map tile as a classification problem. For each photo in a batch, the corresponding map tile of that zoom level is treated as the correct class, and the corresponding map tiles for the other photos in the batch are treated as incorrect classes. The cosine similarities of each photo embedding with the embeddings of the map tiles are put into a softmax function to output probabilities, and these are optimized using a cross-entropy loss.

Additionally, the fact that the map tiles of the coarser zoom levels are not very diverse and contain extremely coarse-grained information necessitates that we augment the training process using information from the lower zoom levels. As such, the map encoders are trained from the finest to coarsest zoom levels sequentially, where the encoder for zoom level 16 is fine-tuned from the original pre-trained encoder and each training process of the next levels is a fine-tuning of the trained model from the previous level. We suggest that this may improve the performance of retrieval in the coarser zoom levels by utilizing knowledge from the finer zoom levels.

3.3. Inference Method

The inference process proceeds as follows. First, the Map Encoder fine-tuned through training with satellite map images at zoom level i, and the publicly available CLIP image encoder ViT are prepared. Here, i ranges from 4 to 16 with a spacing of 2. Using the i Map Encoder, the cosine similarity between the satellite map image patches at zoom level i and the user's photo is calculated, and the top k similar patches are selected. This process is illustrated in Figure 3. Then, the selected k patches are further divided into a 4 by 4 grid. These $k \times 16$ patches are passed through the $i + 2$ Map Encoder to obtain embeddings for each, which are then compared again with the user image to calculate similarity, yielding the top k high probability map patches. This process is iterated from $i = 4$ to $i = 14$, ultimately obtaining the highest probability map patch at zoom level 16. Through this process, the inference of the location where the user photo was taken is completed.

Since the process of obtaining k patches with high cosine similarity at each zoom level affects performance, the value of k is expected to influence the results. Therefore, we conducted inference while varying k values to 16, 32, and 64. This proved to be a significant factor affecting inference time. It is provided at Table 1

Benchmark	Method	Distance (% @ km)				
		Street 1 km	**City 25 km**	**Region 200 km**	**Country 750 km**	**Continent 2,500 km**
IM2GPS[5]	ISNs [10]	16.9	43.0	51.9	66.7	80.2
	Translocator [12]	19.9	48.1	64.6	75.6	86.7
	GeoDecoder [2]	22.1	50.2	69.0	80.0	89.1
	PIGEOTTO [4]	14.8	40.9	63.3	82.3	91.1
	Ours	6.8	11.3	18.1	34.0	65.9

Table 2. Compare with other models in IM2GPS Benchmark.
The metrics other than the performance of **Ours** were obtained from [4].

4. Experiments

4.1. Data Preparation

We conducted training using a portion of the Yahoo Flickr Creative Commons 100 Million dataset (YFCC100M) [14], which had been previously utilized in geoEstimation tasks [11]. This subset, structured according to the data introduced in the MediaEval Placing Task 2016 (MP-16) [8], consists of over 5 million geo-tagged images. The dataset images encompass a variety of scenes, including outdoor and indoor settings, as well as photos of food, people, and other scenes where inferring location information directly may be challenging. The photos were crawled from the Flickr web platform, and the data includes information such as the photo itself, the uploader's id, and images' latitude and longitude. Due to time constraints in our research, we utilized approximately 190,000 data points for training.

The MP-16 dataset provides latitude and longitude data for the locations where user images were taken. Therefore, to utilize 2D map images in our model, it is necessary to first convert 3D coordinates to 2D coordinates. For the Google images we used, which have a 2D map started at latitude 85 degrees and longitude -180 degrees, the following formula can be used to transform coordinates. Additionally, the formula to obtain the (x,y)-th map at zoom level k from the converted coordinates is as follows.

$$\text{East Distance (E)} = \text{false\_easting} + (\lambda - \lambda_o) \tag{3}$$

$$\text{North Distance (N)} = \text{false\_northing} + \ln\left(\tan\left(\frac{\pi}{4} + \frac{\phi}{2}\right)\right) \tag{4}$$

$$x = \left\lfloor \frac{E + \pi}{2\pi/2^k} \right\rfloor \tag{5}$$

$$y = \left\lfloor 2^k - \frac{N + \pi}{2\pi/2^k} \right\rfloor \tag{6}$$

Here, ϕ represents latitude in radians, λ represents longitude in radians, λ_o represents the central meridian in radians. 'false easting' and 'false northing' are values to correct for the easting and northing distances during the process of converting 3D coordinates to 2D rectangular coordinates. This time, both values were set to 0 before proceeding.

4.2. Hyperparameter Setting

The crawled BP-16 dataset was divided into train, valid, and test sets with a ratio of 10:1:1 for experimentation. During training, a batch size of 64 was used, while a batch size of 128 was used for validation. The optimizer employed was Adam with a learning rate of 0.0001, and the training proceeded for 5 epochs.

4.3. Result

Our model, trained on a large dataset of 190,000 user photo-map image pairs, exhibited performance comparable to Table 1 across various k values on five different length scales. This evaluation was conducted on the publicly available dataset, IM2GPS. However, due to the lengthy inference time per image as shown in the table, it was impractical to infer the entire test set within the limited time. For $k = 16, 32, 62$, 100%, 27%, and 30.6% of the test set were processed, respectively.

5. Conclusion

Our model demonstrates high performance compared to idle inference. While we anticipated significant performance differences across various k values during the inference process, no clear trends were observed in reality. However, there were significant differences in inference times, indicating the importance of selecting a model based on this criterion. Nonetheless, as evident from Table 2, our model exhibits lower performance compared to other models on the given Im2GPS benchmark.

The geolocalization task, unlike many other tasks in computer vision, often lacks publicly available code or does not disclose crucial elements used in training, even if some parts of the code are shared. [4] mentioned that this choice was

Ablation	Distance (% @ km)				
	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
GeoCLIP[1]	17.30	41.77	60.76	77.22	89.87
Simply attach GeoCLIP to Ours only in inference	14.29	21.43	28.57	45.24	69.05

Table 3. Performance comparison of different methods on various distances.

made because widely publicizing geolocalization tasks could raise ethical concerns due to the potential risk of identifying someone's location based on others' photos. Therefore, due to such reasons, we faced challenges in utilizing our model to surpass previous research within a limited time, one of the methods being 'leveraging previous research"'. Additionally, the bottleneck of manually crawling datasets from the internet posed another challenge. It was difficult to utilize all training data used by existing models within the constraints of limited internet speed. Future research should address these deficiencies and further refine models to better suit the task.

5.1. Future works

Fortunately, the previous research, GEOCLIP [1], had its code publicly available and distributed in a form that could be imported and utilized. It is deemed that incorporating GEOCLIP into our current research could potentially improve performance. In fact, we devised experiments by simply connecting our model with GEOCLIP. After obtaining k satellite map patches, we leveraged GEOCLIP by including the map patch selected by the GEOCLIP model at the chosen coordinate position along with the $k+1$ patches for the next stage of inference. However, since this approach did not involve training GEOCLIP and our model harmoniously from the training phase, we did not achieve dramatic performance improvements. Instead, we observed a slight decrease in performance compared to when only GEOCLIP was used for inference, suggesting possible confusion with the GEOCLIP model. Comparison between two can be observed at Table 3 Nevertheless, it could be considered as future work to potentially enhance performance by leveraging GEOCLIP, a well-performing model, from the training phase, enabling it to receive map images as input, thus obtaining additional information about terrain or text on the map.

References

[1] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv preprint arXiv:2309.16020*, 2023. 4, 8

[2] Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, and Mubarak Shah. Where we are and what we're looking at: Query based worldwide image geo-localization using hierarchies and scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23182–23190, 2023. 7

[3] David J Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770, 2009. 2

[4] Lukas Haas, Silas Alberti, and Michal Skreta. Pigeon: Predicting image geolocations. *arXiv preprint arXiv:2307.05845*, 2023. 4, 7

[5] James Hays and Alexei A Efros. Im2gps: estimating geographic information from a single image. In *2008 ieee conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 7

[6] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 2

[7] Giorgos Kordopatis-Zilos, Panagiotis Galopoulos, Symeon Papadopoulos, and Ioannis Kompatsiaris. Leveraging efficientnet and contrastive learning for accurate global-scale location estimation. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 155–163, 2021. 2

[8] Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth JF Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. 7

[9] Carlo Masone and Barbara Caputo. A survey on deep visual place recognition. *IEEE Access*, 9:19516–19547, 2021. 2

[10] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 7

[11] Eric Muller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–579, 2018. 4, 7

[12] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. In *European Conference on Computer Vision*, pages 196–215. Springer, 2022. 7

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[14] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 7

[15] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021. 3

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[17] Nam Vo, Nathan Jacobs, and James Hays. Revisiting im2gps in the deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 2621–2630, 2017. 2

[18] Cheng-Xiang Wang, Xiaohu You, Xiqi Gao, Xiuming Zhu, Zixin Li, Chuan Zhang, Haiming Wang, Yongming Huang, Yunfei Chen, Harald Haas, et al. On the road to 6g: Visions, requirements, key technologies and testbeds. *IEEE Communications Surveys & Tutorials*, 2023. 4

[19] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 37–55. Springer, 2016. 2

[20] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 5