Pixels to Places: Improving Zero-Shot Image Geolocalization using Prior Knowledge

Trent Borg[1] and Miriam Cha[2]
[1]DAF-MIT AI Accelerator
[2]MIT Lincoln Laboratory

***Abstract*—The ability to predict the geographic origin of a photo is critical for open-source investigation applications. However, image geolocalization is highly challenging due to the vast diversity of images captured worldwide. While vision transformer-based approaches have demonstrated success—even outperforming grandmasters in geolocation games like GeoGuessr—their performance does not generalize well to unseen locations. Prior methods rely solely on visual cues, neglecting broader contextual knowledge that image analysts typically employ. To bridge this gap, our research integrates the contextual understanding of geographic regions that imagery analysts possess into the geolocalization model. Specifically, we develop a variant of StreetCLIP, which embeds CLIP within geolocalization tasks and facilitates the incorporation of user-supplied prior knowledge such as continental or national boundaries. Our results on the IM2GPS3K benchmark dataset demonstrate a 10.66% improvement in regional prediction (within 200 km) and a 15.27% improvement in country-level prediction (within 750 km) over baseline models. Our results suggest that human-provided supervision can enhance image geolocalization accuracy, highlighting the potential of interactive systems where human expertise and AI work collaboratively to refine predictions.**

***Index Terms*—image geolocalization, CLIP, human-machine teaming, vision transformers**

I. Introduction

Predicting the location of an image anywhere in the world is a complex pattern recognition problem that has been attempted with various machine learning approaches. This task is challenging due to a variety of confounding factors such as variations in lighting, seasonal changes, construction, and imbalanced data between highly populated and rural areas. Early computer vision efforts relied on extracting image features such as color histograms, textures, line features, and geometric context to match images to a pre-existing database [1]. This created a reliance on extensive image databases which subsequent research circumvented by using deep learning models that could operate independently after initial training, eliminating the need for persistent large-scale storage. Later, Convolutional Neural Networks (CNNs) improved geolocation predictions by enabling feature extraction. Using feature extraction to break areas into regions with distinct characteristics allows models to sort the world into more than just cells by chunking them into categories (e.g. "forest" or "desert") or other regions with unique geographic patterns [2]. These CNNs have a proven track record in extracting effective low- and high-level representations such as edges or objects. However, CNNs can struggle with global context and long-range dependencies.

The emergence of Vision Transformers (ViT) [3] marked a significant advancement in the field. These transformer-based networks have shown superior performance over traditional CNNs by leveraging self-attention mechanisms, allowing them to capture long-range dependencies within images more effectively. In previous research, ViT-based approaches in image geolocalization have used techniques such as multitask learning [4] and hierarchical linear probing [2], and other techniques which culminate in systems that achieve greater precision than human experts.

However, state-of-the-art models have demonstrated predictive accuracy of only 50% within a 25 km radius [5], which remains insufficient for critical applications such as disaster response that require more accuracy. Human understanding of spatial relationships and contextual clues can be used to refine the model's outputs and improve trust and explainability, leading to more accurate localization in challenging scenarios. For that reason, human intuition and contextual knowledge play a valuable role in geolocation.

One example of state-of-the-art image geolocation pipelines is PIGEON/PIGEOTTO [5] which leverage geocells [6], multitask learning [4], hierarchical linear probes [2], and geographical feature data [2] to improve accuracy. Although the architecture for PIGEON and PIGEOTTO is available, the model and its weights are unpublished. Replicating the entire architecture would undoubtedly improve predictions, but evaluating the effectiveness of incorporating human feedback in a simplified environment is useful for understanding the effects. For this reason, StreetCLIP was selected as it is the most similar backbone model to PIGEON and PIGEOTTO, while retaining a more simple architecture. Traditionally, human-machine teaming has been limited to post-prediction validation, but our approach enables direct human interaction by allowing users to shape the natural language labels supplied to the model. This multimodal approach provides a means to address challenges such as imbalanced training data and the need for multiple models for different feature extraction tasks.

II. Background

A. Contrastive Language-Image Pre-training (CLIP)

The key enabling technology for the state-of-the-art image geolocation pipeline is Contrastive Language-Image Pre-training (CLIP) [7], a model designed to associate images with the most relevant text labels. CLIP is trained by encoding both image and text pairs and then calculating the cosine similarity

between the text and image embeddings to ensure they are close in the shared vector space. The training objective uses a contrastive loss function, defined as cross-entropy loss:

$$L = \frac{1}{2}\left(L_{image \to text} + L_{text \to image}\right) \tag{1}$$

where

$$L_{image \to text} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\text{sim}(i_i, t_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(i_i, t_j)/\tau}} \tag{2}$$

$$L_{text \to image} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\text{sim}(t_i, i_i)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(t_i, i_j)/\tau}} \tag{3}$$

Here, $\text{sim}(\cdot, \cdot)$ represents cosine similarity, N is the batch size, and τ is the temperature parameter that controls the probability distribution. The variables i and t denote the image and text embeddings, respectively.

Once trained, CLIP can take an image and a set of potential labels, selecting the best match based on similarity score. CLIP is trained on YFCC100M [8], which is a dataset of photos from Flickr and Wikipedia. As the name suggests, the dataset consists of 100 million images ranging broadly in terms of content. The characteristics of CLIP that make it desirable for this use case is that it does well in a zero-shot environment and can be pre-trained further to improve out-of-domain performance.

B. StreetCLIP

A specialized adaption of CLIP, StreetCLIP [9], is trained on a dataset of 1.1 million Google Street View images from 275,000 locations. The training labels for these images are generated by reverse geocoding. When using StreetCLIP, in order to obtain an accurate result from the model, the correct label must be present in the data labels. In the original benchmark, as shown in Fig. 1, this is accomplished by supplying the country labels for nearly all the worlds countries and then supplying the city labels from the World Cities Database [10].

While other geolocation methods, such as Translocator [4], have shown improved performance compared to StreetCLIP in certain benchmarks—achieving 48.1% accuracy for predictions within 25 km—StreetCLIP remains an effective choice for zero-shot, out-of-distribution geolocation. Unlike models trained exclusively on curated datasets, StreetCLIP excels at processing any street-level image, making it highly adaptable.

III. Image Geolocalization with Prior Knowledge

The multimodal nature of CLIP and StreetCLIP allows for human-in-the-loop refinement by leveraging natural language inputs. As shown in Fig. 2, users can supply their own labels in lieu of using canned labels, such as country of city labels, to narrow down geolocation predictions. This interactive approach provides a more intuitive way to allow expert human input into the process by filtering the locations by country or city.

In general, geolocation tasks that use machine learning assume that the user knows nothing about the image they

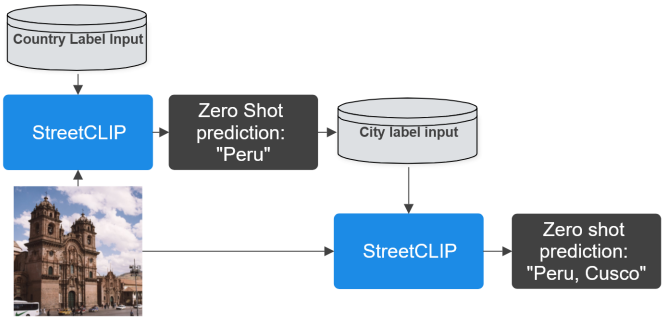

Fig. 1: **Prediction pipeline of StreetCLIP.** StreetCLIP was benchmarked by using a hierarchical linear probe, first predicting from a predefined list of 193 countries (country label input). Next, based on the predicted country, the top 30 city names (city label input) are selected from the World Cities database for city-level prediction.

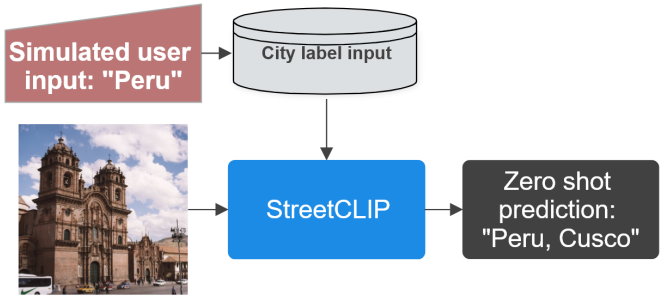

Fig. 2: **Prediction pipeline with user feedback.** The user provides the country name based on prior knowledge, which is used to retrieve the top 30 city labels from the World Cities database. For benchmarking, the user input is simulated by supplying the true country name assuming the user can infer it from context clues.

are attempting to make a prediction about. However, for this experiment, the opposite is assumed to be true. It is reasonable to believe that the user may have knowledge of the continent, country, or even city where the image was taken. Although machines are useful in participating in visual recognition tasks, humans are uniquely suited to identifying objects that can serve as clues about where the image was taken. Clues such as identifying text or a flag in an image which can vastly narrow down an image location is a challenge for machine learning that requires an additional model, training or technique to account for. In contrast, identifying a flag or sign is a simple and small part of image analysis for a human that greatly narrows down the possible location. In working with the model, this research will demonstrate that human-machine teaming can have a positive impact on the accuracy of the StreetCLIP models' predictions.

IV. Datasets and Metric

IM2GPS [1], a test set of 237 Flickr images covering diverse global locations, was used as a benchmark for evaluation. This

TABLE I: Comparison of geolocation methods across different geographic scales.

Benchmark	Method	Distance (% @ km)			
		City 25 km	Region 200 km	Country 750 km	Continent 2,500 km
IM2GPS	StreetCLIP	28.3	45.1	74.7	88.2
	StreetCLIP with priors	**30.34**	**53.41**	**79.05**	**90.17**
IM2GPS3K	StreetCLIP	22.4	37.4	61.3	80.4
	StreetCLIP with priors	**26.85**	**48.06**	**76.57**	**91.69**

dataset does not overlap with the CLIP or StreetCLIP training corpus and includes images ranging from easily geolocatable landmarks to nearly impossible to locate scenes. Additionally, IM2GPS3K [11], a larger set of 3,000 test images from Flickr, was used.

To measure the accuracy of predictions, the Haversine distance formula is used to calculate the distance between the predicted and actual locations:

$$d = 2R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{4}$$

where R is the radius of the earth (set to 6371 km). The variables ϕ_1, λ_1 and ϕ_2, λ_2 correspond to the latitude and longitudes coordinates of the two locations. Ground-truth coordinates from the IM2GPS and IM2GPS3K datasets serve as reference locations. The model's predicted location is compared against the truth location using the Haversine distance. The model does not have access to the metadata for any of the test images to assist in the prediction process.

V. Experiments

An image is provided to the StreetCLIP model along with a list of possible location names derived from the World Cities database [10]. To simulate a human's prior knowledge of the area, the correct country is given rather than requiring the model to infer it using the linear probe methodology. The country label is used to filter the top 30 cities within the country, and then a label template is applied during inference. The template applied during inference is modified based on the user input as: *"A street view photo of* {*Country*}*,* {*City*}*."*

As shown in Table I, this simple yet effective modification resulted in measurable performance improvements. Specifically, the model's accuracy increased by 4.45% for predictions within 25 km, 10.66% within 200 km, 15.27% within 750 km, and 11.29% within 2500km for IM2GPS3K. These results are encouraging indications that human input can provide positive downstream effects. The StreetCLIP model used in these experiments is based on the *clip-vit-large-patch14-336* architecture from OpenAI [7]. The corresponding hyperparameters for this ViT model are detailed in Table II.

TABLE II: CLIP ViT Hyperparameters

Hyperparameter	Value
Batch size	32768
Vocabulary size	49408
Training epochs	32
Maximum temperature	100.0
Weight decay	0.2
Warm-up iterations	2000
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	10^{-6}

VI. Discussion and Limitations

The hierarchical linear probe methodology shares limitations observed in prior StreetCLIP evaluations [9]. Since the model predicts only location names, the assigned coordinates are typically geocoded center points or predefined points from a database. The model is capable of predicting street names, but global databases of street names would be not only massive but also frequently updated. Furthermore, a large portion of the globe, especially rural areas, do not have street addresses. Therefore, without other techniques and methods to remedy this issue, the model is best suited to predict at the city-level scale.

VII. Conclusions and Future Work

Machine learning models have demonstrated the ability to outperform humans in geolocation tasks [12], yet human intuition remains valuable in addressing technical limitations. Challenges such as limited imagery in rural regions, rapidly changing environmental conditions (e.g. natural disasters, seasonal variations, and day/night differences), and geopolitical restrictions (e.g. countries with restricted mapping) can create gaps where human input can be beneficial. Our experiments show that incorporating user's prior knowledge significantly improves geolocation accuracy at city, regional, country, and continent levels.

Future work will explore further enhancements to human-machine teaming collaboration. Specifically, the research will incorporate a capability to allow users to input textual descriptions of terrain or environmental features to refine predictions, pushing the model towards predicting areas that match the users' description. Another promising direction may be leveraging geocells with associated metadata. Filtering predictions based on user-described attributes such as coastlines, mountains, rivers, or urban features could offer an alternative strategy when country or city names are unknown.

Acknowledgment

We would like to thank the MIT Lincoln Laboratory and the Air Force AI Accelerator for their support and guidance on this research. We would also like to thank the peers and mentors that offered their time and guidance to help make this study successful. Research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted

as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] J. Hays and A. A. Efros, "Im2gps: estimating geographic information from a single image," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. [Online]. Available: http://graphics.cs.cmu.edu/projects/im2gps/im2gps.pdf

[2] E. Müller-Budack, K. Pustu-Iren, and R. Ewerth, "Geolocation estimation of photos using a hierarchical model and scene classification," in *European Conference on Computer Vision (ECCV)*, 2018. [Online]. Available: https://openaccess.thecvf.com/content_ECCV_2018/papers/Eric_Muller-Budack_Geolocation_Estimation_of_ECCV_2018_paper.pdf

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] 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 preprint arXiv:2204.13861*, 2022. [Online]. Available: https://arxiv.org/abs/2204.13861

[5] L. Haas, M. Skreta, S. Alberti, and C. Finn, "Pigeon: Predicting image geolocations," *arXiv preprint arXiv:2307.05845*, 2023. [Online]. Available: https://arxiv.org/pdf/2307.05845

[6] T. Weyand, I. Kostrikov, and J. Philbin, "Planet - photo geolocation with convolutional neural networks," in *Computer Vision – ECCV 2016*, vol. 9912, 2016, pp. 37–55.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. [Online]. Available: https://arxiv.org/pdf/2103.00020

[8] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, pp. 64–73, 2016.

[9] L. Haas, S. Alberti, and M. Skreta, "Learning generalized zero-shot learners for open-domain image geolocalization," *arXiv preprint*, 2023.

[10] Simplemaps. (2019) World cities database. [Online]. Available: https://simplemaps.com/data/world-cities

[11] N. Vo, N. Jacobs, and J. Hays, "Revisiting im2gps in the deep learning era," *arXiv preprint arXiv:1705.04838*, 2017. [Online]. Available: https://arxiv.org/abs/1705.04838

[12] L. Haas, M. Skreta, S. Alberti, and C. Finn, "Pigeon: Predicting image geolocations," *arXiv preprint arXiv:2307.05845*, 2023. [Online]. Available: https://arxiv.org/pdf/2307.05845