PGeoCLIP: Acceleration on Image geo-localization Using Precomputed Features

Chengwuzhou Wu

School of Big Data & Software Engineering, Chongqing University, Chongqing, China
20231458.stu.cqu.edu.cn

Abstract. Image-based Geo-localization refers to predicting the geographic location from an image. A noble image-to-GPS retrieval approach GeoCLIP, demonstrated outstanding performance below distance threshold metrics, while the substantial training time and computational overhead present considerable challenges. Acceleration in image geo-localization refers to the enhancement of both training and inference efficiency for image geo-localization models. GeoCLIP's time bottleneck lies in the feature extraction stage of its image encoder. In this paper, the author evaluated GeoCLIP's performance using precomputed features on the test dataset, and proposed an improved model PGeoCLIP based on GeoCLIP, to mitigate the performance degradation caused by using precomputed features. PGeoCLIP removes random data augmentation in the dataloader and introduces several selective improvements: a gated MLP mechanism in the location encoder, as well as a supervised constraint in the image encoder. These modifications significantly reduce training time (e.g., -12h for training one epoch) and inference time (e.g., -8.70s for one query), while maintaining competitive accuracy on the test dataset.

Keywords: image geo-localization, precomputed features, training acceleration.

1 Introduction

Ground-level image geo-localization is the task of determining the geographic location of a photograph captured from a ground-level perspective, enabling applications in mapping, navigation, and location-based services. Traditional manual approaches to image geo-localization can achieve considerable accuracy under ideal conditions. However, manual approaches are time-consuming, unstable, and often yield inconsistent results, as the performance largely depends on the varying levels of prior geographic knowledge among individuals.

Over the years, researchers have developed a substantial body of work in computer-aided geo-localization, building a rich foundation for further advancements in the field. Early works adopted image retrieval-based approaches, where a query image is matched against a large database of Geo-tagged reference images using handcrafted features and similarity metrics [1]. With the rise of deep learning, learned global descriptors became the dominant paradigm, significantly improving robustness and scalability in real-world environments [2]. More recent trends include transformer-

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_27

based architectures and multi-modal fusion methods, which leverage global context modeling and additional cues such as textual metadata or GPS priors to improve robustness under challenging conditions [3][4].

GeoCLIP [5] has recently achieved state-of-the-art performance by leveraging large-scale vision-language pretraining. While highly accurate on Accuracy@Distance, it suffers from substantial computation and training-time overheads, particularly in its image feature extraction stage [1][6]. These limitations of GeoCLIP [5] motivate the development of more efficient solutions without compromising accuracy.

To explore the potential of leveraging precomputed features for faster training, this work proposes PGeoCLIP. PGeoCLIP builds upon the GeoCLIP model, cutting down its training-time overhead by removing random data augmentations in the dataloader, so that image features extracted in frozen CLIP can be reused in every round of training. In the location encoder, PGeoCLIP introduced one training strategy and two optional functions to strengthen the feature embedding of 2D GPS coordinates: progressive hierarchical training strategy, gated MLP and channel attention mechanism. A supervised loss is incorporated to encourage the extracted features to be consistent with coarse-grained location categories.

In summary, this paper makes two primary contributions.

(1) This study evaluates the performance of GeoCLIP when trained with precomputed features on the test dataset, revealing the performance drop introduced by this acceleration strategy.

(2) This study proposes PGeoCLIP, an improved variant of GeoCLIP that incorporates architectural and training refinements to mitigate this degradation and achieve competitive accuracy with significantly reduced training time.

2 PGeoCLIP

2.1 Approach Overview

Image Geo-localization training datasets typically offer image-GPS pairs. Some methods handled the image and GPS relatively well. PGeoCLIP proposed a method tuned from GeoClip, which offers a location encoder and an image encoder. Location encoder encodes GPS information, represented by latitude-longitude pairs, into a high-dimensional embedding. Image encoder utilizes a frozen CLIP backbone to extract features from input images, while also supporting pre-computed features. The model performs location prediction based on the cosine similarity between the GPS embeddings and the image features. Figure 1 shows the PGeoCLIP overall structure.

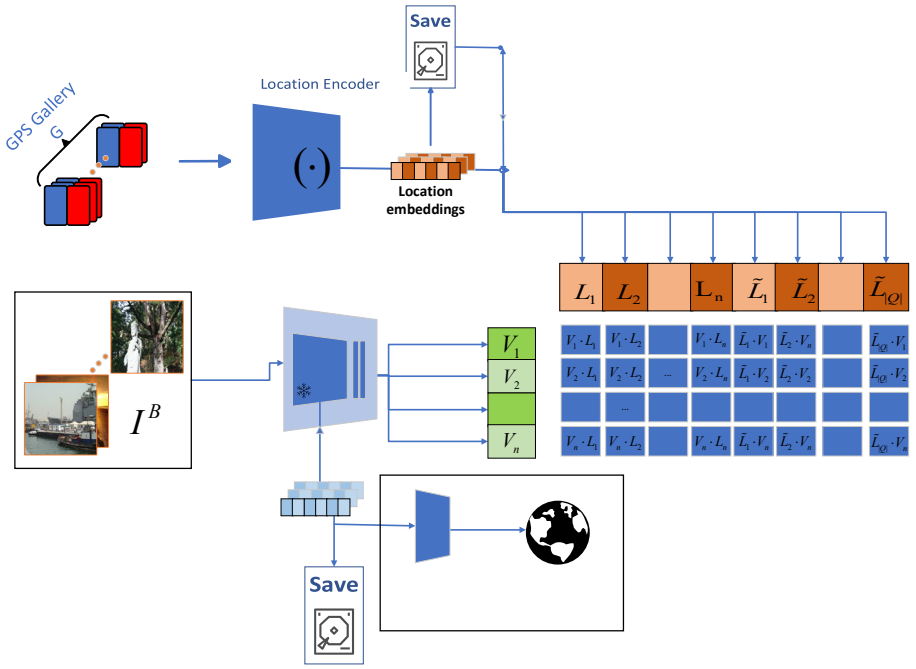

Fig. 1. Overall structure (Picture credit: Original)

2.2 Image Encoder

Following the findings reported in the supplementary material of GeoCLIP(Section 6: Motivation for using Pretrained CLIP as Image Encoder Backbone), PGeoCLIP adopts a pretrained and frozen CLIP as its image encoder, as it has shown strong generalization performance across different downstream tasks. This characteristic provides a feasible basis for using precomputed image features. In Image Encoder, the input consists of raw images processed by the dataloader, which are then passed through CLIP to extract features, followed by an MLP layer (M(·))to obtain the final image feature representation. The CLIP backbone (CLIP(·)) uses the pretrained CLIP ViT-L/14 weights and remains frozen during training. The subsequent MLP consists of two linear layers. In this paper, the author modifies the input of the image encoder to accept features F_i extracted by CLIP(F_i=CLIP(I_i),∀I_i∈D_train), thereby eliminating the need for repeated, time-consuming feature extraction during multiple training iterations. Instead, precomputed features are directly used in subsequent training. The modified Image Encoder V(·) and image feature V_i represent as follow.

$$V_i = \mathcal{V}(F_i) = \mathcal{M}\big(CLIP(I_i)\big), \forall i \in [1 \dots N] and\ I_i \in D_{train} \quad (1)$$

The use of pre-computed features ensures that images fed into the image encoder are processed only once in the dataloader, instead of being processed in every epoch as

before. The impact of this step on model performance resulting from this modification will be presented and analyzed in the Experiments section.

Supervised Constraint. CLIP features mainly encode semantic concepts (e.g., objects and scenes) rather than geographic information, which may limit their ability to distinguish between locations. To address this, PGeoCLIP add a supervised constraint in the image encoder to align image features with coarse-grained location. For defining the supervised categories, PGeoCLIP follows the idea used in PlaNet, which partitions the Earth into a fixed number of geographic cells. Each image is assigned to the cell corresponding to its GPS coordinates, and the image encoder is trained to predict the cell index. For each (gps, image) pair in the training dataset, this work assign a categorical label Label_i according to predefined geography divisions by mapping the GPS coordinates onto the map, thus forming triplets (gps, image, label). This work then trains a discriminator (D(·)) to predict the region label from the CLIP-extracted features, and computes a classification loss(Cross-Entropy, CE) by comparing the predicted labels with ground-truth.

This encourages the extracted features to be more geographically discriminative, making them better aligned with the fine-grained GPS embeddings. The constraint loss can be represented as follow.

$$L_{cons} = CE(D(f_{CLIP}(I_i), Label_i) \tag{2}$$

2.3 Location Encoder

In the Location encoder, the model applies the Equal Earth Projection (EEP) to transform raw GPS coordinates (i.e., G_i ∈ R^2, ∀i ∈ {1…N}), and incorporates Random Fourier Features (RFF) along with a Hierarchical Representation to address the spectral bias problem encountered by standard MLPs in GPS encoding. A gated MLP module is employed to provide weights for the summation of hierarchically extracted multi-level embeddings.

Random Fourier Features with gated MLP. The Random Fourier Features (RFF) module employs multiple σ values to capture frequency-specific components from the Fourier representations of GPS coordinates. As the frequency range in RFF depends on the σ value, the work proposes an exponential assignment strategy for choosing the σ values in the Location encoder. Specifically, for a given range of σ values (i.e., [σ_min,σ_max], in this paper, σ_min=2^0, σ_max=2^12, M=4), it obtains the σ values for M level hierarchy using the following:

$$\sigma_i = 2^{log_2(\sigma_{min})+(i-1)(log_2(\sigma_{max}))/(M-1)}, \forall i \in \{1 \dots M\} \tag{3}$$

The features corresponding to each frequency are processed by dedicated MLP layers f_i,i∈ {1, 2, ⋯, M}, and their outputs are subsequently aggregated via linear summation to obtain the final encoding. In this work, the author adds a gated MLP (GM(.)) between EEP and feature sum. Given GPS coordinates processed by the EEP, the gated MLP outputs weights for the capsules corresponding to each frequency. The final representation is obtained by performing a weighted combination of features across all frequencies during feature aggregation.

$$w = GM(G_i) = [w_1, w_2, \ldots, w_M] \tag{4}$$

Below the author summarizes the overall operations performed by the location encoder L(.)for obtaining rich high-dimensional encoded features (L_i) of a GPS 2D coordinate G_i:

$$L_i = \mathcal{L}(G_i) = \sum_{i=1}^{M} w_i f_i \left(\gamma(EEP(G_i), \sigma_i)\right) \tag{5}$$

Similar to the precomputed features in image encoder, use precomputed location features during inference can significantly improve the model's inference speed, especially when running on a CPU. The inference time is reported in the experimental section.

2.4 Loss

This work adapts a dynamic queue Q of GPS locations with a fixed length S to every batch B of training data D_train. During training, the model generates embeddings for these GPS coordinates ($\widetilde{L}$). A regulating parameter α is introduced to balance the contribution of the constraint loss on the total loss.

The overall training objective is to minimize the following loss:

$$loss = -log \frac{exp(V_i \cdot L_i)}{\sum_{i=0}^{|B|} exp(V_i \cdot L_i) + \sum_{i=0}^{S} exp(V_i \cdot \widetilde{L_\iota})} + \alpha \frac{1}{S} \sum_{i=1}^{S} CE(D(f_{CLIP}(I_i), Label_i) \tag{6}$$

3 Experiment

3.1 Datasets

This work performs experiments on publicly available benchmark datasets to have a fair comparison with existing works. For training, this work uses the MediaEval Placing Tasks 2016(MP-16) dataset [7], which consists of 4.72 million geotagged images from Flickr [8]. For testing, this work tests the trained model on Im2GPS3k [1] and YFCC26k [9].

3.2 Evaluation Metrics

During testing, this work conducts image-to-GPS retrieval against a GPS gallery of 100K coordinates. This paper evaluates the results using a threshold metric, similar to the protocol followed in previous works. For each image in the test dataset pair (image, GPS), the trained model predicts the most possible GPS coordinate, and computes the Geodesic distance between the predicted value and the GPS ground truth in the same pair. This work then performs stratified static on the distance errors and reports the percentage of samples failing within predefined distance thresholds: 2500km, 750km, 200km, 25km and 1km.

3.3 Train details

The proposed PGeoCLIP model was trained for 10 epochs with a batch size of 512 using the Adam optimizer. This paper used two parameter groups with different learning rates: the image encoder parameters were updated with a learning rate of 3*10-5, while the entire location encoder was trained with a smaller learning rate of 3*10-5 and a weight decay of 1*10-6. To ensure stable convergence, this study applied a step-wise learning rate decay using StepLR with a decay factor of 0.87 after every epoch.

When using precomputed CLIP features, the input to the image encoder was replaced with these pre-extracted embeddings, thus eliminating the need for repeated feature extraction during each epoch and reducing computational overhead. Model checkpoints were saved every three epochs.

3.4 Training and Inference Speedup

Training time. This work measures the training time cost of the model by the time required for one epoch, as shown in Table 1. In GeoCLIP, for each image in the training dataset, the dataloader applied a series of data augmentation operations, such as: Random-Resized-Crop, Random-Horizontal-Flip and Random-Grayscale. Such augmentations, however, required recomputing image features in every training iteration, resulting in a considerable time overhead. PGeoCLIP removed the random data augmentations applied before training, computing each feature only once and storing it for later use during training. This greatly reduces the training time but inevitably results in a loss of model accuracy.

Table 1. Train time comparison between GeoCLIP and PGeoCLIP.

Train Time for 1 Epoch	Data Preparing	Training	Average
GeoCLIP	12 h	2m40s	12h+2m40s
PGeoCLIP	12h	3m	(12h/num_epoch)+3m

Inference time. Once the model is trained, the parameters of the location encoder remain unchanged during inference. Using precomputed features for the GPS gallery can accelerate the inference process, especially when computational resources are limited. Table 2 shows that the use of precomputed features reduces the inference time from tens of seconds to the second level.

Table 2. Inference time on CPU for given n pictures.

Model	avg 1 picture(s)	5 pictures(s)
PGeoCLIP with precomputed GPS gallery	9.36	46.81
PGeoCLIP	0.66	3.31

3.5 Comparison

The use of precomputed features leads to a slight performance drop across all scales. PGeoCLIP mitigates this performance degradation while reducing the training time. In this work, the author reproduces the results, using the training code released by

GeoCLIP, while replacing the input to the image encoder with precomputed features. The performance of the resulting model is reported in the rows labeled Precomputed in the following tables. As shown in Table 3 and Table 4, PGeoCLIP outperforms both GeoCLIP with precomputed features and state-of-art methods prior to GeoCLIP on the test dataset.

Table 3. Accuracy comparison of PGeoCLIP and other models on the Im2GPS3k dataset on acc@distance metrics.

IM2GPS3k [1]	2500km	750km	200km	25km	1km
GeoCLIP [5]	83.82	69.67	50.65	34.47	14.11
Precomputed	80.93	65.65	48.23	32.68	12.14
PGeoCLIP	82.25	66.60	47.15	30.96	11.97
Translocator [10]	80.1	58.9	46.7	31.1	11.8
GeoDecoder [11]	76.1	61.0	45.9	33.5	12.8

Table 4. Accuracy comparison of PGeoCLIP and other models on the YFCC26k dataset on acc@distance metrics.

YFCC26k [9]	2500km	750km	200km	25km	1km
GeoCLIP [5]	76.02	57.47	36.69	22.19	11.61
Precomputed	74.97	56.66	35.71	21.79	10.15
PGeoCLIP	75.22	57.09	36.08	22.16	10.58
Translocator [10]	60.6	41.3	28.0	17.8	7.2
GeoDecoder [11]	69.0	49.6	34.1	23.9	10.1

3.6 Ablations

To assess the contribution of each component in PGeoCLIP, this work conducts ablation studies on the IM2GPS3k benchmark (Table 5). Several observations can be made:

Effect of layer number. Increasing the number of location encoder capsules from 3 to 4 slightly improves performance at all thresholds. This indicates that 4 capsules have a positive effect on the model.

Effect of gated MLP. Introducing the gated MLP without the supervised constraint brings marginal improvements (e.g., +0.10 at 2500km, +0.02 at 200km), suggesting that gating helps the model better regulate frequency-specific representations. However, the gains remain limited when used alone.

Effect of supervised constraint. Adding the supervised constraint alone yields inconsistent performance. While it preserves discriminative power at mid threshold(200km), it does not consistently improve fine-grained accuracy. Likely because the author divided the globe into eight continents in a simplistic manner, which prevented the semantic CLIP features from being properly aligned with geographic regions.

Combined effect (PGeoCLIP). When both gated MLP and supervised constraint are integrated, the model achieves the best overall performance, surpassing all other settings across the thresholds. This indicates that the two components are

complementary: the gated MLP enhances representation control, while the supervised constraint introduces geographic discriminability.

Table 5. Ablation Results of PGeoCLIP on the IM2GPS3k Dataset

IM2GPS3k [1]	2500km	750km	200km	25km	1km
Precomputed no gated no constraint 3 layers	80.93	65.65	48.23	32.68	12.14
no gated no constraint 4 layers	81.41	65.84	47.78	32.92	12.95
Gated no constraint 4 layers	81.58	66.01	48.0	33.50	13.06
no gated constraint 4 layers	80.78	65.26	48.04	32.36	12.51
PGeoCLIP gated constraint 4 layers	81.68	66.57	48.61	33.30	13.04

4 Conclusion

This work evaluates the performance of GeoCLIP trained with precomputed image features and proposes a modified model PGeoCLIP based on GeoCLIP that eliminates random data augmentation during the training stage, adds gated MLP and supervised constraint mechanisms, and enables precomputed inputs. PGeoCLIP computes image features once and saves them to local storage, enabling faster subsequent training by reusing them. This approach significantly reduces training time, as measured by the duration of a single training epoch, but inevitably introduces a certain degree of accuracy loss. Furthermore, this work explored approaches to mitigate performance degradation. The method is designed to alleviate this degradation while maintaining the advantage of reduced training time: The Gated MLP is introduced to improve the hierarchical encoding capability of the location encoder for GPS coordinates. The supervised constraints try to associate semantic CLIP features with geographic regions. Experimental results demonstrate that, although some accuracy is sacrificed, PGeoCLIP achieves a favorable trade-off between computational efficiency and predictive performance.

References

1. Hays, J., Efros, A.A.: Im2GPS: Estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)
3. Wilson, D., Zhang, X., Sultani, W., Wshah, S.: Visual and object geo-localization: A comprehensive survey. In: arXiv:2112.15202 (2021)
4. Zhu, S., Shah, M., Chen, C.: 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, pp. 1162–1171 (2022)

5. Vivanco Cepeda, V., Nayak, G.K., Shah, M.: GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. Advances in Neural Information Processing Systems 36, 8690–8701 (2023)
6. Vo, N., Jacobs, N., Hays, J.: Revisiting Im2GPS in the deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2621–2630 (2017)
7. Larson, M., Soleymani, M., Gravier, G., Ionescu, B., Jones, G.J.: The benchmarking initiative for multimedia evaluation: MediaEval 2016. IEEE MultiMedia 24(1), 93–96 (2017)
8. Flickr: Upgrade everything you do with Flickr Pro. https://www.flickr.com, (2025/09/20)
9. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Li, L.J., et al.: YFCC100M: The new data in multimedia research. Communications of the ACM 59(2), 64–73 (2016)
10. Pramanick, S., Nowara, E.M., Gleason, J., Castillo, C.D., Chellappa, R.: Where in the world is this image? Transformer-based geo-localization in the wild. In: European Conference on Computer Vision, pp. 196–215. Springer, Cham (2022)
11. Clark, B., Kerrigan, A., Kulkarni, P.P., Cepeda, V.V., Shah, M.: 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, pp. 23182–23190 (2023)

.