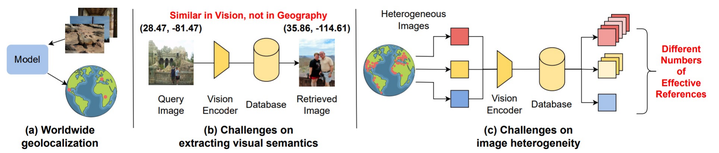
G3: An Effective and Adaptive Framework for Worldwide Geolocalization Using Large Multi-Modality Models

Pengyue Jia[1], Yiding Liu[2], Xiaopeng Li[1], Yuhao Wang[1], Yantong Du[1], Xiao Han[1], Xuetao Wei[3], Shuaiqiang Wang[2], Dawei Yin[2], Xiangyu Zhao[1*]

[1]City University of Hong Kong, [2]Baidu Inc., [3]Southern University of Science and Technology

Background & Motivation



Background

Worldwide image geolocalization aims to pinpoint the exact shooting location for any given photo taken anywhere on Earth. Unlike geolocalization within specific regions (e.g., at city level), worldwide geolocalization greatly unleashes the potential of geolocalization, which is useful for various real-world applications, such as crime tracking and navigation.

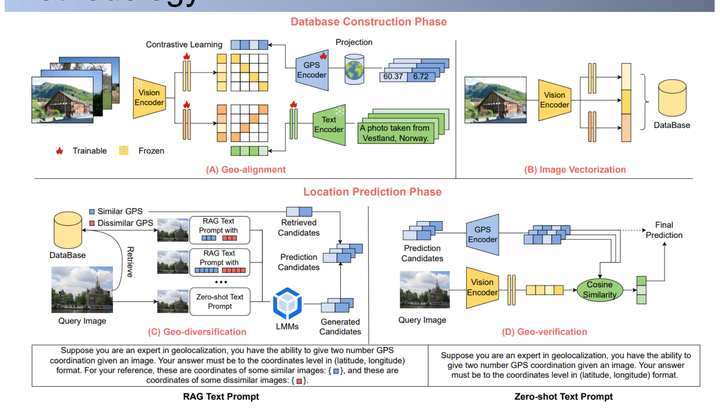
Motivation

- It is very challenging to extract visual semantics that accurately indicate an images' geolocation, as two distant places could possibly have similar visual features.
- Image data usually exhibits significant heterogeneity in its geographical distribution, which existing methods can hardly handle.

Contributions

- We present G3, a novel solution for the worldwide geolocalization task.
- We release a new dataset MP16-Pro, adding textual localization descriptions to each sample based on the original dataset MP16 to facilitate future research in the field.
- We extensively experiment with two well-established datasets IM2GPS3k and YFCC4K.

Methodology



Geo-alignment

- Image encoding $\mathbf{e}^{\text{image}}_{i,\text{text}} = f_{\text{text}}(\mathcal{V}(\mathbf{I}_i)),\ \mathbf{e}^{\text{image}}_{i,\text{gps}} = f_{\text{gps}}(\mathcal{V}(\mathbf{I}_i))$
- GPS encoding $\begin{cases}\mathbf{x} = \mathbf{R}\cdot(\lambda-\lambda_0)\\ \mathbf{y} = \mathbf{R}\cdot ln[\tan(\frac{\pi}{4}+\frac{\phi}{2})]\end{cases}$ $\mathbf{e}^{\text{gps}}_i = \sum_{k=1}^{K} f_k(\gamma(\mathbf{G}_i,\ \sigma_k))$
- Text encoding $\mathbf{e}^{\text{text}}_i = f(\mathcal{T}(\mathbf{T}_i))$

Optimization

$$\mathcal{L}_{a,b} = -\sum_{i=1}^{n} log(\frac{\exp(\text{logits}_{ii})}{\sum_{j=1}^{n}\exp(\text{logits}_{ij})}),\ \text{logits} = (\frac{\mathbf{e}^a}{\|\mathbf{e}^a\|_2})(\frac{\mathbf{e}^b}{\|\mathbf{e}^b\|_2})^{\mathrm{T}}\cdot\exp^{t_{a,b}}$$

Geo-diversification

We construct K RAG prompts with different numbers of reference coordinates, and each prompt will generate N results. S coordinate candidates of retrieved similar images will also be considered.

$$\{c_1^k, c_2^k, \cdots, c_n^k\} = \text{RAG}(p^k)$$
$$\{c_1, c_2, \cdots, c_m\}, \text{where } m = K\times N + S$$

Geo-verification

We reinvent the well-trained Image-to-GPS model in Geo-alignment to verify the prediction.

$$\text{sim} = \mathbf{e}^{\text{image}}_{\text{gps}}(\mathbf{e}^{gps})^{\mathrm{T}}$$
$$j = \text{argmax}(\text{sim}_j),\ j\in\{1,2,\cdots,m\}$$

Experiments

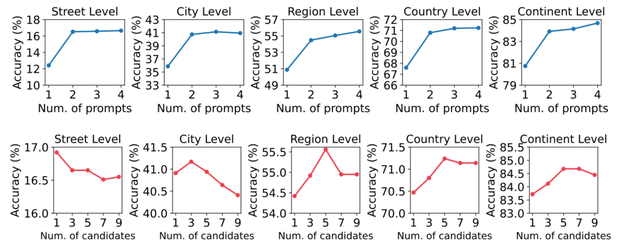
Overall Experimental Results

Methods	IM2GPS3K					YFCC4K				
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
[L]kNN, sigma=4 [30]	7.2	19.4	26.9	38.9	55.9	2.3	5.7	11	23.5	42
PlaNet [22]	8.5	24.8	34.3	48.4	64.6	5.6	14.3	22.2	36.4	55.8
CPlaNet [22]	10.2	26.5	34.6	48.6	64.6	7.9	14.8	21.9	36.4	55.5
ISNs [18]	10.5	28	36.6	49.7	66	6.5	16.2	23.8	37.4	55
Translocator [20]	11.8	31.1	46.7	58.9	80.1	8.4	18.6	27	41.1	60.4
GeoDecoder [3]	12.8	33.5	45.9	61	76.1	10.3	24.4	33.9	50	68.7
GeoCLIP [29]	14.11	34.47	50.65	69.67	83.82	9.59	19.31	32.63	55	74.69
Img2Loc [43]	15.34	39.83	53.59	69.7	82.78	19.78	30.71	41.4	58.11	74.07
PIGEON [5]	11.3	36.7	53.8	**72.4**	**85.3**	10.4	23.7	40.6	62.2	77.7
Ours	**16.65**	**40.94**	**55.56**	71.24	84.68	**23.99**	**35.89**	**46.98**	**64.26**	**78.15**

Ablation Study

Methods	IM2GPS3K					YFCC4K				
	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
w/o Geo-A	15.71	40.64	54.85	70.8	84.05	20.8	32.72	44.25	61.83	76.64
w/o Geo-D	16.35	40.51	53.89	69.2	83.11	20.28	31.87	43.67	60.84	76.25
w/o Geo-V	14.98	38.27	51.25	67.6	81.18	19.03	30.24	40.93	57.93	72.83
Ours	**16.65**	**40.94**	**55.56**	**71.24**	**84.68**	**23.99**	**35.89**	**46.98**	**64.26**	**78.15**

Hyperparameter Analysis



Impact of LMMs on G3

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
Img2Loc (LLaVA)	10.21	29.06	39.51	56.36	71.07
Img2Loc (GPT4V)	15.34	39.83	53.59	69.70	82.78
G3 (LLaVA)	14.31	35.87	49.42	66.93	81.78
G3 (GPT4V)	**16.65**	**40.94**	**55.56**	**71.24**	**84.68**

Necessity Analysis of Three Representations Alignment

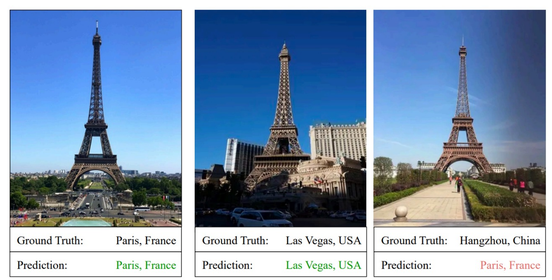
Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
IMG	15.71	40.64	54.85	70.8	84.05
IMG+GPS	**16.91**	**41.41**	55.02	70.94	84.18
IMG+GPS+TEXT(G3)	16.65	40.94	**55.56**	**71.24**	**84.68**

Case Study on Reference Image Retrieval

Experiments on Textual Description Granularity

Methods	Street 1km	City 25km	Region 200km	Country 750km	Continent 2500km
G3-N	16.44	40.64	54.35	70.57	83.98
G3	**16.65**	**40.94**	**55.56**	**71.24**	**84.68**

Failure Analysis



Code Dataset AML Lab HomePage