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
{jia.pengyue,xiaopli2-c,yhwang25-c}@my.cityu.edu.hk,
{liuyiding.tanh,hahahenha,shiqiang.wang}@gmail.com,
duyantong94@hrbeu.edu.cn,weixt@sustech.edu.cn,
yindawei@acm.org,xianzhao@cityu.edu.hk

Abstract

Worldwide geolocalization aims to locate the precise location at the coordinate level of photos taken anywhere on the Earth. It is very challenging due to 1) the difficulty of capturing subtle location-aware visual semantics, and 2) the heterogeneous geographical distribution of image data. As a result, existing studies have clear limitations when scaled to a worldwide context. They may easily confuse distant images with similar visual contents, or cannot adapt to various locations worldwide with different amounts of relevant data. To resolve these limitations, we propose **G3**, a novel framework based on Retrieval-Augmented Generation (RAG). In particular, G3 consists of three steps, i.e., **G**eo-alignment, **G**eo-diversification, and **G**eo-verification to optimize both retrieval and generation phases of worldwide geolocalization. During Geo-alignment, our solution jointly learns expressive multi-modal representations for images, GPS and textual descriptions, which allows us to capture location-aware semantics for retrieving nearby images for a given query. During Geo-diversification, we leverage a prompt ensembling method that is robust to inconsistent retrieval performance for different image queries. Finally, we combine both retrieved and generated GPS candidates in Geo-verification for location prediction. Experiments on two well-established datasets IM2GPS3k and YFCC4k verify the superiority of G3 compared to other state-of-the-art methods. Our code[2] and data[3] are available online for reproduction.

1 Introduction

Worldwide image geolocalization [30] aims to pinpoint the exact shooting location for any given photo taken anywhere on Earth, as illustrated in Figure 1(a). Unlike geolocalization within specific regions (e.g., at city level) [19, 2, 25, 12, 23], worldwide geolocalization [29, 30, 43] greatly unleashes the potential of geolocalization, which is useful for various real-world applications, such as crime tracking and navigation. However, worldwide image geolocalization is extremely challenging, as images collected from around the world are featured with a myriad of elements, including varying landscapes, weather conditions, architectural styles, etc.

*Corresponding author.
[2]https://github.com/Applied-Machine-Learning-Lab/G3
[3]https://huggingface.co/datasets/Jia-py/MP16-Pro

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-1685

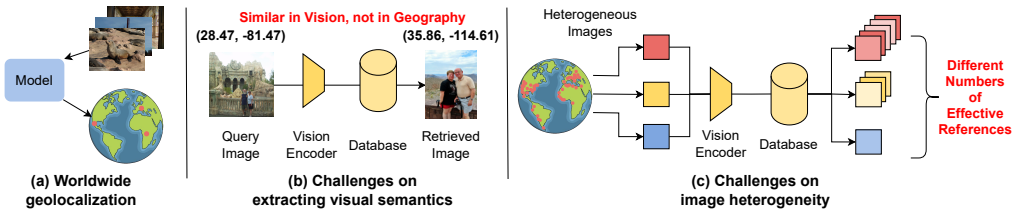


Figure 1: Illustration of limitations of prevailing methods in worldwide geolocalization.

Extensive research efforts have been devoted to the task, which can be broadly categorized into 1) classification-based, 2) retrieval-based, and 3) retrieval-augmented generation (RAG) methods. In particular, classification-based methods [32, 22, 20, 3] divide the entire geographical space into fixed grids, and classify each image into a particular grid. Retrieval-based methods convert image localization to either a image-to-image [2, 12, 23] or an image-to-GPS [29] retrieval problem, where the final prediction is a top-retrieved GPS exists in a given candidate database. Generation-based methods [43] recently achieved state-of-the-art performance on this task, via applying a retrieval-augmented generation (RAG) pipeline that leverages the strong reasoning and generalization ability of large multi-modality models (LMMs). They usually integrate the retrieved GPS coordinates in the input prompts of LMMs as references, to generate more accurate predictions.

Despite their initial success, existing studies still have clear limitations when scaled to a worldwide context, mainly due to two challenges as shown in Figure 1. **First**, it is very challenging to extract visual semantics that accurately indicate an image's geolocation, as two distant places could possibly have similar visual features. Conventional visual representations are ineffective in implying subtle location-aware semantics. **Second**, image data usually exhibits significant heterogeneity in its geographical distribution, which existing methods can hardly handle. For retrieval-based methods, it only performs well for image queries with many nearby images stored in the database, while many images at unpopular locations may have very few or even no similar data to be compared with, leading to large prediction errors. Worst still, such inconsistent retrieval performance significantly affects existing RAG-based methods that use a fixed number of references. As the retrieval performs inconsistently for different image queries, their generation process lacks the robustness to adapt to various image queries at different locations worldwide.

To address the aforementioned challenges, we propose G3, a novel RAG-based solution with expressive retrieval and robust generation for worldwide image geolocalization.

For the retrieval phase, we train multi-modality encoders to effectively capture location-aware visual similarity between images. Unlike existing RAG-based methods that only leverage conventional visual similarity for retrieval, we propose a multi-modality alignment process, namely **Geo-alignment**, which learns the representations of GPS coordinates, images, and textual descriptions in a joint manner. By doing this, the visual representations are able to capture fine-grained location-aware semantics for retrieving other images close by. In addition to the numerical GPS coordinates, the textual descriptions (e.g., city/country names) can largely enrich the location information to be aligned with the visual representations. Moreover, to train the multi-modality representations, we establish a new dataset, namely **MP16-Pro**, by including textual geographical descriptions to the original MP16 dataset [11]. We anticipate the dataset will benefit more future work for location-aware visual representation learning.

For the generation phase, we leverage a prompt ensembling method, namely **Geo-diversification**, which improves the robustness of prediction generation for different types of images. More specifically, it generates a diverse set of predictions via multiple retrieval-augmented prompts, each of which might be more useful for a certain type of query images. As such, the generated GPS candidates are more likely to contain the ground truth coordinates. Subsequently, we conduct **Geo-verification**, which combines both the retrieved and the generated GPS candidates, and compares their similarities with the query image using the learned multi-modality representations. The most similar GPS is returned as the final prediction. Extensive experiments are conducted on well-established datasets IM2GPS3k [30] and YFCC4K [26], and the results show the effectiveness of G3 compared to the other state-of-the-art baseline methods. We summarize the key contributions of our work as follows:

https://doi.org/10.52202/079017-1685

- We present G3, a novel solution for the worldwide geolocalization task. Our proposed method leverages 1) Geo-alignment to learn expressive location-aware representations of images, 2) Geo-diversification to improve the robustness of GPS candidate generation, and 3) Geo-verification to ensemble both retrieved and generated candidates for final prediction.
- 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. G3 demonstrates superior performance compared to other state-of-the-art baseline methods.

2 Related Work

Image Geolocalization. Image Geolocalization is an important task in computer vision [48, 47, 49], spatial data mining [41, 42, 36, 6], and GeoAI [40, 39, 35, 37]. Previous work in image geolocalization can be divided into three main categories: classification-based methods, retrieval-based methods, and generation-based methods. (1) Classification-based methods [22, 30, 18, 32, 20, 3] divide the entire earth into multiple grid cells and assign the center coordinates as predicted values. Models are then trained to classify the input image into the correct cell. However, if the actual location of the image is far from the center of the predicted cell, there can still be significant errors, even if the cell prediction is correct. (2) Retrieval-based methods treat the image geolocalization task as a retrieval problem, typically maintaining a database of images [33, 17, 46, 34, 44, 27, 24, 45] or a gallery of GPS coordinates [29]. They take the most similar images and GPS coordinates to the query image as the predicted values. However, maintaining a global-level image database or GPS gallery is infeasible. (3) Generation-based methods employ large multi-modality models to generate the predicted coordinates for images [14]. Zhou *et al.* [43] introduced retrieval-augmented generation into the geolocalization task and took the retrieved similar images' coordinates as references to help generate predictions. However, they can not accurately extract visual semantics to indicate an image's location because they simply use visual similarity to retrieve references and suffer inaccurate prediction when facing heterogeneous query images. In this work, G3 introduces Geo-alignment to incorporate geographical information into image representations to help retrieve similar images in geography and proposes Geo-diversification and Geo-verification to enhance prediction performance and robustness.

Large Multi-modality Models. Inspired by the success of large models [9, 10, 15] in single domains like computer vision [31] and natural language processing [38], there has been increasing attention on large multi-modality models. CLIP [21] aligns image and text representations through contrastive learning and achieves remarkable model generalization with simple optimization objectives. The Large Language-and-Vision Assistant (LLAVA) [16] effectively combines CLIP's visual encoder with the powerful language model Vicuna, enhancing the model's general understanding of both visual and textual information through two-stage instruction tuning. GPT4V [1] is a large multimodal model released by OpenAI in 2023, allowing users to input text and images to obtain answers.

Retrieval-Augmented Generation. To mitigate the hallucination issue in LLMs, retrieval-augmented generation (RAG) [13] has emerged as a popular and effective technique. It enhances the reliability of content generated by LLMs by incorporating facts fetched from external sources. Specifically, certain factual knowledge is retrieved by a retriever from external sources based on a query. LLMs can access these retrieval results during the generation process to generate accurate outcomes. RAG preserves the generalization capabilities of LLMs while also introducing external information to enhance the reliability of generated content, efficiently alleviating the hallucination problem.

3 Methodology

Figure 2 illustrates the comprehensive architecture of G3, which consists of Geo-alignment, diversification, and verification, with two phases: database construction and location prediction. During the database construction phase, introduced in Section 3.1, Geo-alignment aligns image representations with textual image descriptions and GPS information to incorporate geographical information into representations. In the location prediction phase, introduced in Section 3.2, similar images will be first retrieved based on the nearest neighbor search from the database; Geo-diversification will then

 https://doi.org/10.52202/079017-1685

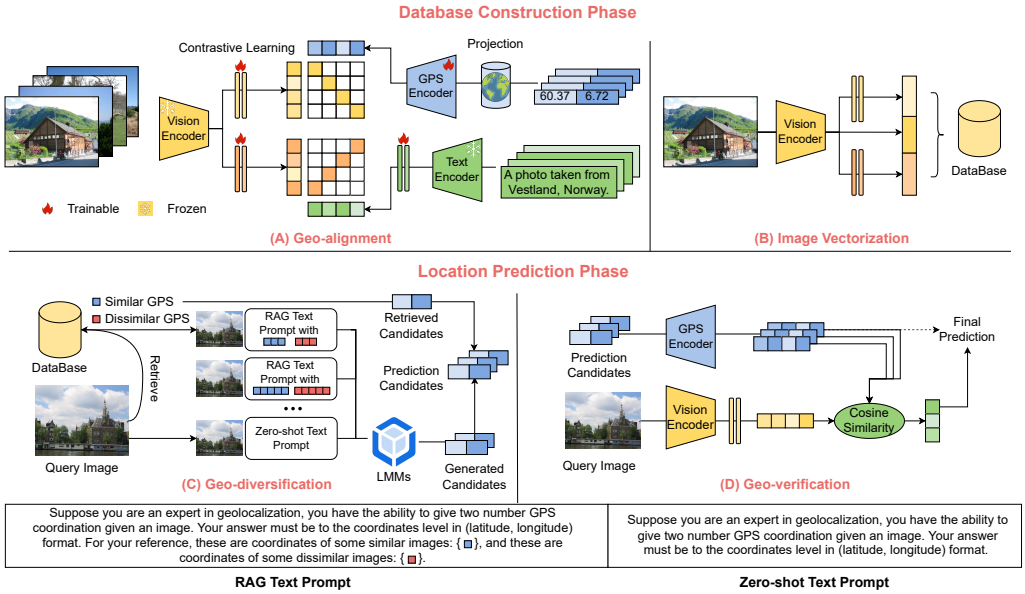


Figure 2: Overview of the framework of G3.

combine their coordinates in RAG prompts to generate diverse candidates, and Geo-verification finally selects the final predicted coordinates in a multi-modality space.

3.1 Database Construction

G3 requires an image database to preserve image representations. Existing work directly uses visual encoders (e.g., CLIP's ViT encoder or ResNet) to encode images. However, visual similarity cannot completely represent geographical proximity. To overcome this issue, we propose Geo-alignment, which incorporates geographical information into image representations by multi-modality alignment.

Geo-alignment. Geographical features can be divided into continuous and discrete types, which are essential in geolocalization. On the one hand, according to the first law of geography [28], "everything is related to everything else, but near things are more related to each other." Climate, terrain, and vegetation are continuous features that gradually change along latitude or longitude. On the other hand, discrete features (e.g., city/country names) are also conducive to determining geographical location. These features usually change abruptly at national borders. To encode images with representations tailored for geolocalization, we propose a multi-modality alignment method, Geo-alignment, as shown in Figure 2(A).

Image encoding. We use pretrained vision encoder and two trainable transformation layers to encode images: $\mathbf{e}^{\text{image}}_{i,\text{text}} = f_{\text{text}}(\mathcal{V}(\mathbf{I}_i))$, $\mathbf{e}^{\text{image}}_{i,\text{gps}} = f_{\text{gps}}(\mathcal{V}(\mathbf{I}_i))$, where $\mathbf{e}^{\text{image}}_{i,\text{text}}$ and $\mathbf{e}^{\text{image}}_{i,\text{gps}}$ are the i-th image representations in the batch that need to be aligned with textual geographical descriptions and GPS data. f_{text} and f_{gps} are the corresponding feed forward functions, $\mathcal{V}$ represents the fixed vision encoder, and $\mathbf{I}_i$ is the i-th image in a batch.

GPS coordinate encoding. To encode GPS coordinates, an appropriate projection is needed to transform latitude and longitude into a Cartesian coordinate system. We choose not to adopt the equal earth projection (EEP) used in previous work [29] because EEP primarily focuses on area accuracy while overlooking angular distortions, which is significant in modeling the trends of geographical features along latitude and longitude. As a result, we utilize Mercator projection for its conformal property. The formula of Mercator projection is shown below:

$$\begin{cases} \mathbf{x} = \mathbf{R} \cdot (\lambda - \lambda_0) \\ \mathbf{y} = \mathbf{R} \cdot ln[\tan(\frac{\pi}{4} + \frac{\phi}{2})] \end{cases} \tag{1}$$

where λ and ϕ are radians of longitude and latitude, λ_0 denotes the central meridian longitude. $\mathbf{R}$ is a proportional constant of Earth radius. The output $\mathbf{x}$ and $\mathbf{y}$ denote the transformed plane coordinates.

After projection, we follow previous work [29] to capture high-frequency patterns and hierarchical representations using random fourier features (RFF) with various frequencies. RFF function γ will transform the projected coordinate $\mathbf{G}_i = (\mathbf{x}_i,\ \mathbf{y}_i)$ first: $\gamma(\mathbf{G}_i) = [\cos(2\pi\mathbf{M}\mathbf{G}_i),\ \sin(2\pi\mathbf{S}\mathbf{G}_i)]^{\mathrm{T}}$. $\mathbf{M}$ denotes a matrix sampled from a Gaussian distribution $\mathbf{M} \sim \mathcal{N}(0, \sigma)$ to limit the frequencies. To capture hierarchical representations, we sum up the outputs with different σ: $\mathbf{e}_i^{\text{gps}} = \sum_{k=1}^{K} f_k(\gamma(\mathbf{G}_i,\ \sigma_k))$ where $\mathbf{e}_i^{\text{gps}}$ is the encoded gps representations for i-th sample in a batch, K denotes the number of hierarchical patterns, f_k is the feed forward function for k-th hierarchical layer. σ_k controls the frequency for k-th layer and $\sigma_k = 2^{log_2(\sigma_{min}+(k-1)(log_2(\sigma_{max})-log_2(\sigma_{min}))/(N-1)},\ \forall k \in \{1, \ldots, N\}$. σ_{min} and σ_{max} controls the range of σ_k.

Text encoding. We initially employ geographical reverse encoding to obtain textual descriptions of GPS coordinates. For instance, as illustrated in Figure 2(A), the GPS coordinates (60.37, 6.72) can be converted into the textual description "A photo taken from Vestland, Norway". These textual descriptions are inputs to a pre-trained text encoder, followed by feedforward networks for vector transformation. $\mathbf{e}_i^{\text{text}} = f(\mathcal{T}(\mathbf{T}_i))$ where $\mathbf{e}_i^{\text{text}}$ denotes the encoded textual descriptions for i-th sample in a batch, f is the feed forward transformation layer, $\mathcal{T}$ is the text encoder function, and $\mathbf{T}_i$ is the textual descriptions for i-th sample in a batch.

Optimization. Geo-alignment is optimized with the following objective to align image representations with textual descriptions and GPS information:

$$\mathcal{L}_{a,b} = -\sum_{i=1}^{n} log(\frac{\exp(\text{logits}_{ii})}{\sum_{j=1}^{n}\exp(\text{logits}_{ij})}),\ \text{logits} = (\frac{\mathbf{e}^a}{\|\mathbf{e}^a\|_2})(\frac{\mathbf{e}^b}{\|\mathbf{e}^b\|_2})^{\mathrm{T}} \cdot \exp^{t_{a,b}} \tag{2}$$

where $\mathcal{L}_{a,b}$ denotes the loss function of modality a to modality b, $\mathbf{e}$ is the encoded representations, and t is the temperature. G3 needs to align image representations with both textual descriptions and GPS data, so the final optimization objective is shown below: $\mathcal{L} = (\mathcal{L}_{\text{image,text}} + \mathcal{L}_{\text{image,gps}} + \mathcal{L}_{\text{text,image}} + \mathcal{L}_{\text{gps,image}})/2$.

Image vectorization. As illustrated in Figure 2(B), after Geo-alignment, we will vectorize the images in the dataset and store them in a database. To maintain image representations tailored for geolocation tasks, we concatenate the original visual representations with representations aligned with geographical information: $\mathbf{e}' = \text{concat}(\mathbf{e}^{\text{image}}, \mathbf{e}_{\text{text}}^{\text{image}}, \mathbf{e}_{\text{gps}}^{\text{image}})$. $\mathbf{e}'$ denotes the final representation, $\mathbf{e}^{\text{image}}$ represents the vector obtained directly through the pretrained vision encoder. $\mathbf{e}_{\text{text}}^{\text{image}}$ and $\mathbf{e}_{\text{gps}}^{\text{image}}$ are the image representations aligned with textual geographical descriptions and GPS information.

3.2 Location Prediction

Figure 2(C) and (D) illustrate the overview of the location prediction phase. Previous work [43] directly incorporates the GPS coordinates of similar retrieved images as references into a single RAG prompt to generate predictions. However, due to the heterogeneity of query images, the number of reference GPS coordinates varies when each sample achieves optimal prediction performance. To address this issue, Geo-diversification expands the candidate pool with prompts containing different numbers of reference coordinates, including zero (i.e., in a zero-shot manner), as shown in Figure 2(C). Illustrated in Figure 2(D), Geo-verification selects the best prediction coordinate for each sample using the well-trained Image-to-GPS encoders in Geo-alignment. In the location prediction phase, Geo-diversification and Geo-verification are introduced to enrich the diversity of generated predictions and select the predictions with the highest confidence.

Geo-diversification. Due to the heterogeneity of query images, the number of reference coordinates introduced in the RAG process varies when each sample achieves optimal prediction performance. To solve this issue, we introduce Geo-diversification. Specifically, we first construct K RAG prompts with different numbers of reference coordinates (0 reference coordinates equals zero-shot generation), and each prompt will generate N results. This process can be represented as follows: $\{c_1^k, c_2^k, \cdots, c_n^k\} = \text{RAG}(p^k)$ where c^k denotes the candidate coordinate generated by the k-th RAG prompt, and p^k is the k-th RAG prompt. The final candidate pool contains the top S coordinate candidates of retrieved similar images and the generated coordinate candidates as shown in Figure 2. The final candidate pool is denoted as $\{c_1, c_2, \cdots, c_m\}$, where $m = K \times N + S$.

https://doi.org/10.52202/079017-1685

Geo-verification. Given the coordinate candidates set $\{c_1, c_2, \cdots, c_m\}$, selecting the best guess is essential and challenging. We reinvent the well-trained Image-to-GPS model in Geo-alignment to achieve this target, as shown in Figure 2(D). The similarity between image representations $\mathbf{e}_{\text{gps}}^{\text{image}}$ and GPS representations $\mathbf{e}^{gps}$ are derived by $\text{sim} = \mathbf{e}_{\text{gps}}^{\text{image}}(\mathbf{e}^{gps})^{\text{T}}$, and the coordinate c_j with the highest similarity is selected as the final prediction by $j = \text{argmax}(\text{sim}_j),\ j \in \{1, 2, \cdots, m\}$.

4 MP16-Pro Dataset

To facilitate subsequent research, we propose the MP16-Pro dataset by adding textual geographical descriptions to each sample from the MediaEval Placing Tasks 2016 (MP-16) dataset [11]. Specifically, we utilize the open-source geocoding tool Nominatim to obtain multi-level geographical textual descriptions for each sample's GPS location (**total 4.72 million locations**). There are eight geographical unit levels: neighborhood, city, county, state, region, country, country code, and continent. Some examples are given in Appendix A.1 for reference. Geographical text descriptions provide additional information for geolocalization tasks and enable models to transcend the original paradigm solely supporting image and GPS alignment, facilitating more diverse modeling approaches.

5 Experiments

Datasets and evaluation metrics: For database construction and model training, we use the MP16-Pro dataset we released. It contains 4.72 million geotagged images from Flickr [4]. However, given that the dataset was released in 2016, currently, 4.12 million images within the dataset remain accessible. Following previous work [29, 43], we evaluate G3 with public datasets (IM2GPS3k [7] and YFCC4K [26]) and a threshold metric. Given the predicted coordinates and the ground truths, this metric quantifies the percentage of predictions where the distance to the ground truth falls within specified thresholds (1km, 25km, 200km, 750km, and 2500km).

Implementation details: We use faiss [4] to deploy the image database. The vision and text encoders are pretrained ViT-L/14 and a masked self-attention transformer from CLIP [21]. The dimensions for two trainable layers of f_{text}, f_{gps}, f are 768 and 768. The input dimension of GPS encoder is 512, and the dimensions for four hidden layers of f_k in Equation 3.1 are 1024, the output dimension is 512. For the Earth radius, we set it as 6378137.0. For RFF, we use three hierarchies with σ_{min} as 2^0 and σ_{max} as 2^8. GPT4V [5] is selected as the LMMs in this paper. Its temperature is set to 1.2. The number of RAG prompts K is set to 4, and the number of candidates for each RAG N is set to 5 for IM2GPS3K and 1 for YFCC4K. The number of similar image coordinates taken into account in candidates is 0 for IM2GPS3K and 1 for YFCC4K. G3 is trained using AdamW optimizer with learning rate 3e-5 and weight decay 1e-6. A step linear scheduler is employed with gamma 0.87, and the training epoch is set to 10. Training batch size is set to 256 and temperature t in Equation 2 is initialized as 3.99. All experiments are conducted with Pytorch and one NVIDIA H800 GPU. Please refer to Appendix A.2 for more details on the training environment, training time, and API cost. We also mention the limitations of G3 in Appendix A.4.

Baselines: To evaluate G3 in geolocalization, we follow previous work [29, 43] and select the following baselines for comparison: [L]kNN,σ=4 [30], PlaNet [32], CPlaNet [22], ISNs [18], Translocator [20], GeoDecoder [3], GeoCLIP [29], Img2Loc [43], PIGEON [5]. The detailed descriptions of baselines are in Appendix A.5. Due to the lack of available implementations for Img2Loc, we reproduce it based on its paper and release it in our repository for future research.

5.1 Comparison with State-of-the-art Methods

To verify the effectiveness of G3, we conduct comparative experiments on IM2GPS3K and YFCC4K with other state-of-the-art methods. The results are shown in Table 1. (1) G3 is superior to all the other baselines on almost all metrics. In addition, compared to the second best methods, G3 achieves **8.5%**, **2.8%**, **3.3%** improvements on IM2GPS3K in the 1km, 25km, 200km thresholds and **21.3%**, **16.9%**, **13.5%**, **3.3%**, **0.6%** improvements on YFCC4K in the 1km, 25km, 200km, 750km, 2500km thresholds. (2) G3, Img2Loc, GeoCLIP, and PIGEON achieve leading results, which

[4] https://www.flickr.com/

[5] https://openai.com/

Table 1: Overall experimental results on IM2GPS3K and YFCC4K.

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**

Table 2: Ablation study on IM2GPS3K and YFCC4K.

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**

can be attributed to the other methods taking the worldwide geolocalization task as a classification problem, introducing inevitable systemic biases. (3) G3 demonstrates significant improvements over GeoCLIP because GeoCLIP is constrained by the settings of the GPS gallery, which can not cover the entire globe. (4) Compared to Img2Loc, G3, through Geo-alignment that aligns images with discrete and continuous geographical features, achieves more precise retrieval of reference coordinates for subsequent RAG processes. Additionally, Geo-diversification and Geo-verification effectively expand the candidate pool and filter out the confident prediction results, further enhancing geolocalization performance. Overall, G3 achieves the best performance on all datasets across almost all metrics, which verifies the effectiveness of G3.

5.2 Ablation Study

To understand the specific effects of each module in G3, we design the following variants:

- **w/o Geo-A:** G3 without Geo-alignment. Directly using ViT in CLIP for database construction.
- **w/o Geo-D:** G3 without Geo-diversification. Generating prediction with one RAG prompt with 10 positive samples and 10 negative samples (the parameter has been tuned).
- **w/o Geo-V:** G3 without Geo-verification. Instead of using the well-trained Image-GPS model in Geo-alignment, this variant randomly selects the final prediction from candidates.

Table 2 shows the experimental results. We can draw the following conclusions: (1) All three modules significantly contribute to the final performance. (2) G3 achieves better performance than w/o Geo-A for Geo-alignment incorporates geographical information into image representations. As a result, the retrieved images are geographically similar to the query image, enhancing the effectiveness of references in the RAG process. (3) G3 is superior to w/o Geo-D, for the number of reference coordinates varies when each sample achieves the optimal prediction performance facing heterogeneous query images. The absence of Geo-diversification leads to suboptimal candidates. (4) Comparing G3 with w/o Geo-V, we observe a significant performance drop in w/o Geo-V, indicating the necessity of Geo-verification.

5.3 Hyperparameter Analysis

In the generation process within G3, two hyperparameters directly impact the results: the number of RAG prompts and the number of candidate coordinates generated by each single prompt.

https://doi.org/10.52202/079017-1685

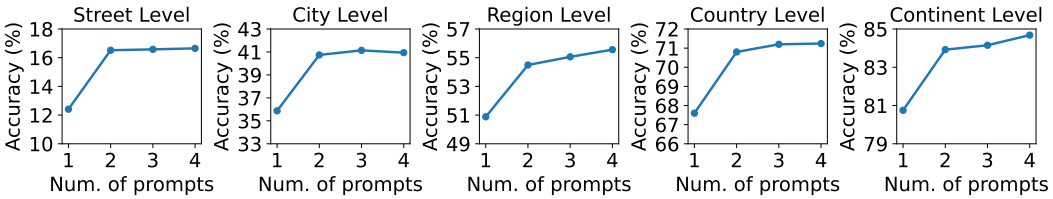


Figure 3: Varying the number of RAG prompts on IM2GPS3K.

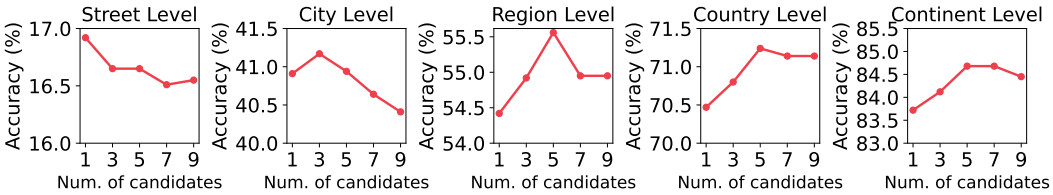


Figure 4: Varying the number of candidates for each RAG prompt on IM2GPS3K.

Number of RAG prompts. To investigate the impact of varying numbers of RAG prompts, we design the following experiment: We employ four sets of RAG prompts with different reference coordinates: 0 positive, 0 negative; 5 positive, 5 negative; 10 positive, 10 negative; 15 positive, 15 negative. Starting with the first prompt, subsequent prompts will be sequentially added to change the number of RAG prompts. The number of candidates generated by each prompt is fixed to 5. As illustrated in Figure 3, the influence of RAG prompt counts on prediction performance is consistent across different metric thresholds. A significant enhancement is observed when the number increases from 1 to 2. The reason is that the zero-shot prompt (RAG prompt with 0 positive and 0 negative reference coordinates) fails to provide high-quality predictions for global images with insufficient information. The model's performance gradually improves as the number increases from 2 to 4 because having more candidates will increase the possibility of containing the ground truth coordinates.

Number of candidates. Figure 4 shows the results varying the number of candidates for each prompt. We fix the number of prompts to 4 in this experiment. We observe that the turning points where performance begins to decline exhibit an increasing trend at different levels. Specifically, at the street level, performance declines after just one candidate, at the city level after three, at the region and country levels after five, and at the continent level after seven. Three key points merit attention: (1) The initial upward trend occurs because the generation of LMMs involves randomness. Introducing more candidates can alleviate the randomness. (2) As the number of candidates increases, performance ultimately drops, likely due to the introduction of more noise in the predictions from additional generations. (3) The turning points of decline differ by level because broader levels demonstrate greater tolerance to predictive bias when more noise candidates are included.

5.4 Effectiveness of Geo-alignment and Mercator Projection

To assess the effectiveness of Geo-alignment and Mercator projection, we conduct the following experiments focusing on the reference retrieval phase: We build image databases using different embedding techniques and then retrieve the Top-N images closest to the query image. The geodesic distances will be calculated between their coordinates and the query image. The embedding variants are illustrated as follows:

- **CLIP ViT:** Directly using the visual encoder ViT in CLIP for image embedding.
- **G3+EEP:** Geo-alignment with Equal Earth Projection (EEP).
- **G3+Mercator:** Geo-alignment with Mercator Projection.

Table 3 shows the statistics of the geodesic distances of retrieval reference images with different embedding methods. We can draw the following conclusions: (1) G3+EEP outperforms CLIP ViT as the latter only considers visual similarity, while image representations in G3+EEP encompass both

Table 3: Distance statistics of retrieval reference images with different embedding methods. Avg., Md., Max., and Min. are the average, median, maximum, and minimum distances to the query image.

Methods	Top-5 Candidates				Top-10 Candidates				Top-15 Candidates			
	Avg.	Md.	Max.	Min.	Avg.	Md.	Max.	Min.	Avg.	Md.	Max.	Min.
CLIP ViT	2554.7	2244.6	5048.4	800.7	2645.9	2269.9	6376.9	513.3	2704.1	2307.7	7142.2	404.8
G3+EEP	2361.5	2089.0	4618.2	735.3	2434.5	2102.2	5808.9	**479.3**	2464.3	2104.3	6529.2	**369.6**
G3+Mercator	**2299.2**	**2054.9**	**4474.7**	**699.0**	**2362.5**	**2035.4**	**5569.6**	482.2	**2405.0**	**2046.9**	**6341.0**	373.0

Table 4: 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**

visual and geographical similarity, which is essential for geolocalization tasks. (2) G3+Mercator performs better than G3+EEP, as the EEP projection method emphasizes area projection accuracy while overlooking angular distortions, which increases the training complexity and limits the performance.

5.5 Impact of LMMs on G3

To explore the impact of LMMs on G3, we conduct the experiments of G3 and Img2Loc with LLaVA (LLaVA-Next-LLaMA3-8b) on IM2GPS3K. From Table 4 we can find that: (1) After switching LMMs from GPT4V to LLaVA, the performance of G3 shows some decline across various metrics but remains competitive. (2) Additionally, compared to Img2Loc (LLaVA), G3 (LLaVA) significantly outperforms Img2Loc (LLaVA), demonstrating the effectiveness of the proposed modules. (3) Finally, by comparing the performance of G3 equipped with LLaVA and GPT4V to Img2Loc equipped with LLaVA and GPT4V, we can observe that G3 shows more stable performance across different LMMs.

5.6 Necessity Analysis of Three Representations Alignment in Geo-alignment

To verify the necessity of aligning the three representations in Geo-alignment, we conduct experiments of the following variants on IM2GPS3K:

- **IMG:** Directly using pre-trained CLIP vision encoder as the encoder.
- **IMG+GPS:** Aligning Image representations with GPS representations in Geo-alignment, the textual descriptions are not used.
- **IMG+GPS+TEXT(G3):** Aligning three modalities simultaneously in Geo-alignment.

Table 5 shows that: (1) By comparing IMG+GPS+TEXT, IMG+GPS, and IMG, we find that adding GPS and text information can both enhance the feature representation compared to using the original image information alone. (2) By comparing IMG+GPS+TEXT with IMG+GPS, we find that IMG+GPS performs better at smaller scales, while IMG+GPS+TEXT performs better at larger scales. This might be because GPS is suitable for modeling variations at smaller scales, whereas text descriptions do not vary significantly at small scales and may even remain the same.

Table 5: Results of necessity analysis.

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**

5.7 Case Study

Case study on reference image retrieval. Figure 5 visually demonstrates the superiority of G3 in reference image retrieval. It is evident that if CLIP's ViT is used as the image encoder, the model

https://doi.org/10.52202/079017-1685

Figure 5: Reference image retrieval with CLIP ViT and G3.

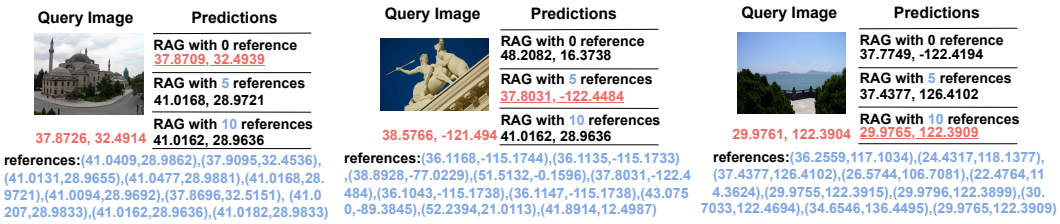


Figure 6: Predictions given with different numbers of references facing heterogeneous images.

primarily focuses on the human figures in the image (i.e., 'two people posing together in the center of the photo') while neglecting background elements beneficial for geolocalization. In Geo-alignment, G3 incorporates geographical information into the image representations. As a result, retrieved images are more focused on geographical proximity (three reference images within 1 km of the actual shooting location are retrieved in the top-5 candidate images). These valuable reference images further assist the RAG process to enhance final prediction performance.

Case study on heterogeneous query image in RAG process. Figure 6 provides three examples illustrating the best prediction occurs when using RAG prompts with different numbers of references facing heterogeneous query images. (1) RAG with 0 references achieves the best performance for the first query image. This is because, on the one hand, the references are filled with biased coordinates, and on the other hand, the building in the figure is a famous landmark named Selimiye Camii mosque. The pre-trained LMMs effectively provide the longitude and latitude of this landmark based on its world knowledge. (2) For the second query image, RAG with 5 references performs best because the optimal reference appeared in the fifth position. More references do not add extra helpful information but instead introduce more noise, causing the performance of RAG with 10 references to decline; RAG with 0 references produces incorrect predictions due to the absence of clear landmark indicators in this image. (3) For the third query image, RAG with 10 references yields the best accuracy, as the references from 6 to 10 provide substantial helpful information, whereas the first five reference coordinates are far from the ground truth. Overall, from these examples, we can discern some common patterns: for images with prominent landmark features, RAG with 0 references often yields good results; for images with less informative content (such as oceans, skies, or indoor scenes), RAG with 10 references makes more comprehensive judgments based on a greater number of references; and for images with distinct regional features (images between the first two settings), RAG with 5 references will achieve satisfactory prediction accuracy.

6 Conclusion

In this paper, we propose a novel worldwide geolocalization framework named G3. First, we introduce Geo-alignment to capture location-aware semantics in images by aligning images with textual geographical descriptions and GPS information. Second, Geo-diversification is proposed to improve the robustness of prediction generation via a prompt ensemble technique. Finally, Geo-verification selects the final coordinate prediction using the learned multi-modality representations. G3 is evaluated on two well-established datasets, IM2GPS3K and YFCC4K, and achieves state-of-the-art performance. In addition, 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. All the code and data used in this work have been released public.

Acknowledgements

This research was partially supported by Collaborative Research Fund (No.C1043-24GF), Research Impact Fund (No.R1015-23), APRC - CityU New Research Initiatives (No.9610565, Start-up Grant for New Faculty of CityU), CityU - HKIDS Early Career Research Grant (No.9360163), Hong Kong ITC Innovation and Technology Fund Midstream Research Programme for Universities Project (No.ITS/034/22MS), Hong Kong Environmental and Conservation Fund (No. 88/2022), and SIRG - CityU Strategic Interdisciplinary Research Grant (No.7020046), Huawei (Huawei Innovation Research Program), Tencent (CCF-Tencent Open Fund, Tencent Rhino-Bird Focused Research Program), Ant Group (CCF-Ant Research Fund, Ant Group Research Fund), Alibaba (CCF-Alimama Tech Kangaroo Fund (No. 2024002)), CCF-BaiChuan-Ebtech Foundation Model Fund, and Kuaishou.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 726–743. Springer, 2020.

[3] 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.

[4] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[5] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12893–12902, 2024.

[6] Xiao Han, Xiangyu Zhao, Liang Zhang, and Wanyu Wang. Mitigating action hysteresis in traffic signal control with traffic predictive reinforcement learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 673–684, 2023.

[7] 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.

[8] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

[9] Pengyue Jia, Jingtong Gao, Xiaopeng Li, Zixuan Wang, Yiyao Jin, and Xiangyu Zhao. Second place overall solution for amazon kdd cup 2024. In *Amazon KDD Cup 2024 Workshop*, 2018.

[10] Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. Mill: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2498–2518, 2024.

[11] 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.

[12] Seongwon Lee, Hongje Seong, Suhyeon Lee, and Euntai Kim. Correlation verification for image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5384, 2022.

 https://doi.org/10.52202/079017-1685

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[14] Ling Li, Yu Ye, Bingchuan Jiang, and Wei Zeng. Georeasoner: Geo-localization with reasoning in street views using a large vision-language model. In *Forty-first International Conference on Machine Learning*.

[15] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. *arXiv preprint arXiv:2312.15450*, 2023.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[17] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019.

[18] 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.

[19] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

[20] 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.

[21] 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.

[22] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.

[23] Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, and Bingyi Cao. Global features are all you need for image retrieval and reranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11036–11046, 2023.

[24] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[25] Fuwen Tan, Jiangbo Yuan, and Vicente Ordonez. Instance-level image retrieval using reranking transformers. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115, 2021.

[26] 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.

[27] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017.

[28] Waldo R Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[29] Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] 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.

[31] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[32] 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.

[33] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[34] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

[35] Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S Joe Qin, and Hongwei Zhao. Mlpst: Mlp is all you need for spatio-temporal prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3381–3390, 2023.

[36] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3195–3205, 2023.

[37] Zijian Zhang, Xiangyu Zhao, Hao Miao, Chunxu Zhang, Hongwei Zhao, and Junbo Zhang. Autostl: Automated spatio-temporal multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4902–4910, 2023.

[38] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[39] Xiangyu Zhao, Wenqi Fan, Hui Liu, and Jiliang Tang. Multi-type urban crime prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4388–4396, 2022.

[40] Xiangyu Zhao and Jiliang Tang. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 497–506, 2017.

[41] Xiangyu Zhao, Tong Xu, Yanjie Fu, Enhong Chen, and Hao Guo. Incorporating spatio-temporal smoothness for air quality inference. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 1177–1182. IEEE, 2017.

[42] Xiangyu Zhao, Tong Xu, Qi Liu, and Hao Guo. Exploring the choice under conflict for social event participation. In *Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I 21*, pages 396–411. Springer, 2016.

[43] Zhongliang Zhou, Jielu Zhang, Zihan Guan, Mengxuan Hu, Ni Lao, Lan Mu, Sheng Li, and Gengchen Mai. Img2loc: Revisiting image geolocalization using multi-modality foundation models and image-based retrieval-augmented generation. *arXiv preprint arXiv:2403.19584*, 2024.

 https://doi.org/10.52202/079017-1685

[44] 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.

[45] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023.

[46] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021.

[47] Yuanshao Zhu, Yongchao Ye, Ying Wu, Xiangyu Zhao, and James Yu. Synmob: Creating high-fidelity synthetic gps trajectory dataset for urban mobility analysis. *Advances in Neural Information Processing Systems*, 36:22961–22977, 2023.

[48] Yuanshao Zhu, Yongchao Ye, Shiyao Zhang, Xiangyu Zhao, and James Yu. Difftraj: Generating gps trajectory with diffusion probabilistic model. *Advances in Neural Information Processing Systems*, 36:65168–65188, 2023.

[49] Yuanshao Zhu, James Jianqiao Yu, Xiangyu Zhao, Qidong Liu, Yongchao Ye, Wei Chen, Zijian Zhang, Xuetao Wei, and Yuxuan Liang. Controltraj: Controllable trajectory generation with topology-constrained diffusion model. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4676–4687, 2024.

4f/a0/3963216890.jpg

eb/a7/193938478.jpg

4b/5c/8178901047.jpg

Figure 7: Image data in MP16-Pro Dataset.

Table 6: More details on training and inference parameters.

Parameter	G3
GPU	H800 80G * 1
Training Time	7 hours / epoch * 10 epoch
Total params	441,266,179
Trainable params	13,648,131 (3.09%)
GFLOPS	304.38
Dataset Samples	4.12M
Batch Size	256
GPU Memory	24G
Text Processor	Huggingface CLIP default text processor
Vision Processor	Huggingface CLIP default vision processor
Token per RAG prompt	Input: 200+30 $\times$ k Output: 18 $\times$ n for each RAG prompt
Token for IM2GPS3K	Input: 5.1M($51) Output: 2.16M($64.8)
Token for YFCC4K	Input: 7.2M($72) Output: 3.06M($91.8)

A Appendix

A.1 MP16-Pro Dataset Samples

To facilitate understanding of the MP16-Pro dataset, we provide some samples from the dataset in this section. The dataset contains two parts: the image data and the metadata for images.

Figure 7 gives three examples from the MP16-Pro dataset. Take these three images as examples, MP16-Pro adds extra textual descriptions based on their coordinates:

- For image 4f/a0/3963216890.jpg, LAT: 47.217578, LON: 7.542092, neighbourhood: Wengistein, city: Solothurn, county: Amtei Solothurn-Lebern, state: Solothurn, region: NA, country: Switzerland, country_code: ch, continent: NA.
- For image eb/a7/193938478.jpg, LAT: 39.950477, LON: -75.157535, neighbourhood: Center City, city: Philadelphia, county: Philadelphia County, state: Pennsylvania, region: NA, country: United States, country_code: us, continent: NA.
- For image 4b/5c/8178901047.jpg, LAT: -34.580365, LON: -58.425464, neighbourhood: Palermo, city: Buenos Aires, county: NA, state: Autonomous City of Buenos Aires, region: NA, country: Argentina, country_code: ar, continent: NA.

All the data has been released online[6].

A.2 More Details on Training and Inference

In this section, we detail the training and inference process by providing information about the training environment, text and vision processors, and token costs. The Input token cost and output token cost for each RAG prompt are $200 + 30 \times k$ and $18 \times n$, where 200 is the fixed token cost for prompts and images with resolution parameter "low", and the 30 is related to the reference

[6]https://huggingface.co/datasets/Jia-py/MP16-Pro

https://doi.org/10.52202/079017-1685

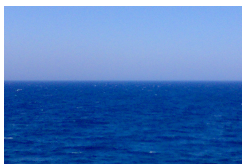
Error $\leq$	Images		
1KM			
25KM			
200KM			
750KM			
2500KM			

Figure 8: Example query images from IM2GPS3K that G3 localization error falls in 1km, 25km, 200km, 750km, 2500km thresholds.

coordinates, k denotes the number of references in the prompt. For the output, 18 is the token cost for one-time generation, and n is the times of generations.

A.3 G3 with Query Image

Figure 8 showcases some example query images from the IM2GPS3K dataset, with different rows representing the varying localization errors of G3 on these images. We can observe patterns in query images with errors ranging from 1km to 2500km. Images with lower errors often feature prominent landmarks or regional characteristics, such as buildings, decorations, and symbols. On one hand, these images are more likely to have more similar images in the database. On the other hand, LMMs are more sensitive to this type of information and can provide more accurate predictions. In contrast, query images with higher errors contain less effective information; large expanses of ocean, water bodies, and snow scenes offer limited assistance for geolocalization.

Table 7: Experimental results on textual description granularity on IM2GPS3K.

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**

A.4 Limitations

The limitation of G3 is mainly about its efficiency. G3 relies on a large number of candidates generated through Geo-diversification. However, the noise candidates generated by Geo-diversification lead to low efficiency in the use of computational resources. A potential solution is to expand the database or enhance the geolocalization ability of LMMs by fine-tuning to help create prediction candidates more efficiently and effectively. In another aspect, the image vector length in Geo-alignment comprises three concatenated vectors, which increases storage demand and retrieval latency. A potential solution is to use quantizers, such as ProductQuantizer [8], to accelerate vector retrieval speed and reduce storage pressure.

A.5 Details on Baselines

- **[L]kNN,$\sigma = 4$ [30].** KNN makes use of the top k NN images and aggregates their coordinates to a prediction point. As k decreases, σ decreases, this method transforms to NN.
- **PlaNet [22].** PlaNet is the first work posing worldwide geolocalization as a classification task by dividing the globe into thousands of multi-scale geographical cells. It can combine the complex clues in images to help pinpoint the shooting location of images.
- **CPlaNet [22].** CPlaNet tries to solve the trade-off of cell granularity. It introduces combinatorial partitioning, which creates detailed output classes by intersecting broad earth partitions, with each classifier voting for overlapping classes.
- **ISNs [18].** ISNs combines the hierarchical information that existed in the partitionings and the photo's scene contextual information (e.g., indoor, natural, or urban, etc.) to give the prediction.
- **Translocator [20].** Translocator is a dual-branch transformer network that focuses on the detailed clues in images and generates robust feature representations. The semantic segmentation map and the entire image will be the input to translator.
- **GeoDecoder [3].** GeoDecoder argues that previous work fails to exploit the detailed cues in different hierarchical levels. It proposes a cross-attention network to capture the complex relationships between different hierarchical features.
- **GeoCLIP [29].** GeoCLIP is based on the CLIP backbone model and first introduces a GPS encoder to transform coordinates into embeddings in worldwide geolocalization tasks.
- **Img2Loc [43].** Img2Loc combines the RAG paradigm into worldwide geolocalization and is the latest work in this field. It first retrieves similar images via visual similarity and puts the coordinates of these images into RAG prompt to help generate predictions.
- **PIGEON [5].** PIGEON introduces an innovative approach combining semantic geocell creation, multi-task contrastive pretraining, and a novel loss function, and uniquely enhances guess accuracy through retrieval over location clusters.

A.6 More Experimental Results on Geo-alignment.

In this section, we conduct experiments on incorporating more fine-grained textual descriptions in Geo-alignment. Specifically, in addition to including the city, county, and country information in the textual descriptions of coordinates, we also introduce neighborhood information, which is the most fine-grained data that can be obtained from Nominatim. We use G3-N to denote this variant and keep the other hyperparameters the same as G3. The experimental results on IM2GPS3K are presented in Table 7.

From the results, we can see that G3 outperforms G3-N across all metrics. This may be because the text encoder's pre-training corpus contains very few instances of neighborhood-level information,

https://doi.org/10.52202/079017-1685

resulting in weaker modeling capabilities for neighborhood names. Therefore, introducing neighborhood information into the textual descriptions of coordinates actually adds noise, which negatively impacts the effectiveness of Geo-alignment and subsequently reduces the model's prediction accuracy.

A.7 Hard Sample and Failure Analysis

Hard Sample. In this section, we give a hard sample to illustrate the effectiveness of G3. Figure 9 shows a man holding an American flag in France, the text on the left says "United States of America" in French. G3 can accurately give the prediction coordinate latitude: 48.8529 and longitude: 2.3632 located in Paris, France. This demonstrates that G3 can effectively avoid the influence of text in images, thereby focusing on the location where the image was taken, showing strong stability.

Figure 9: Hard sample.

Failure Analysis. Figure 10 presents the results of G3 predicting the coordinates of the Eiffel Tower and its replicas. G3 is able to locate the Eiffel Tower in France and its replica in the USA, but fails to distinguish the replica in China. The prediction for the replica in the USA is correct, while the prediction for the replica in China is incorrect. This may be because there are more reference objects around the tower in the USA, which reduces the difficulty of the model's prediction. In contrast, the lack of reference objects around the tower in China confuses G3's judgment.

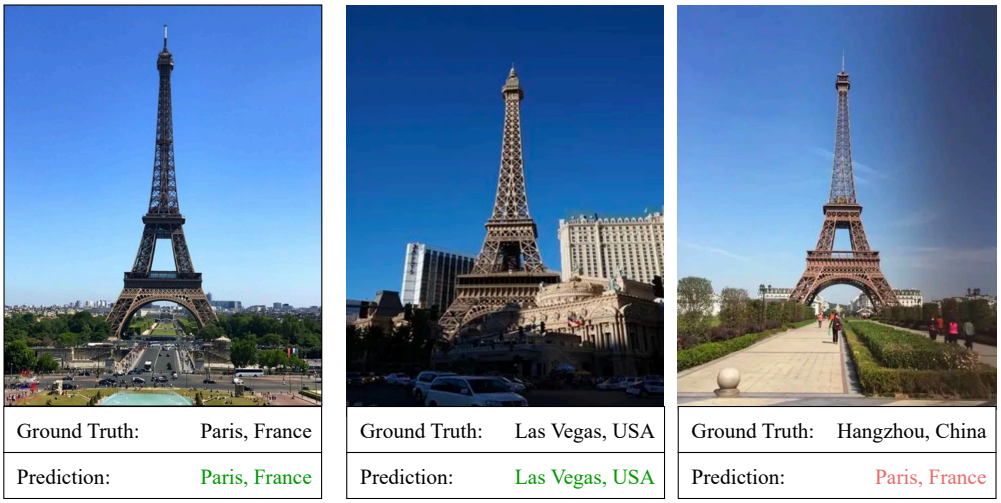


Figure 10: Experimental results of G3 on predicting coordinates of Eiffel Tower and its replica.

https://doi.org/10.52202/079017-1685

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize the contributions and scope in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have mentioned the limitations of our work in Appendix A.4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-1685

Justification: We have contained the formula derivation process and the experimental results to prove the assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have introduced all the details of G3 in our paper. In addition, our submission also contains the materials to reproduce the main experimental results, including code, data, hyperparameters settings, etc.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the link to the GitHub repository in our paper to provide open access to the data and code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Section 5, Appendix A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted repeated experiments and calculated the average to mitigate the variability of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-1685

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have included the cost information in Appendix A.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention the societal impact on the introduction section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our released dataset is based on public dataset MP16 and will also follow the license of MP16.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-1685

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we have included the descriptions in this paper and in the supplementary document.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There are no crowdsourcing experiments or research involving human subjects in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No potential risks are found in this work.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.