GIMI: A GEOGRAPHICAL GENERALIZABLE IMAGE-TO-IMAGE SEARCH ENGINE WITH LOCATION-EXPLICIT CONTRASTIVE EMBEDDING

Hao Li, Jiapan Wang, Balthasar Teuscher, Peng Luo, Martin Werner
Technical University of Munich, Munich, Germany
corresponding to: hao_bgd.li@tum.de

Gengchen Mai
University of Georgia, USA

Danfeng Hong
Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China

ABSTRACT

To query and localize objects of interest among massive and multi-modality big geospatial data (BGD) is fundamental in spatial data science and Earth system science (ESS). However, the effective and efficient searching among an extensive collection of geospatial data (e.g., global satellite imagery) for interesting patterns can be challenging, often requiring domain-specific prior knowledge (i.e., training labels) and intensive computational resources. Towards addressing this challenge, we introduce GIMI, a geographical generalizable image-to-image neural search engine that extends *the cluster hypothesis* from information retrieval theory - closely associated documents tend to be relevant to the same requests - to geospatial data. We explicitly integrate geo-location information into the contrastive learning of image embeddings via a general distance-penalized triplet loss. On this basis, GIMI is designed to support a wide range of search queries, including embedding-based similar search and spatial-constrained nearest neighborhood search. As a case study, we select the task of post-disaster damage building search to demonstrate the general idea behind GIMI and evaluate its model performance in a critical real-world searching scenario. Experiments show that GIMI achieves promising searching performance, w.r.t accuracy and efficiency, in selected areas affected by the 2023 Kahramanmaraş Earthquake in Turkey.

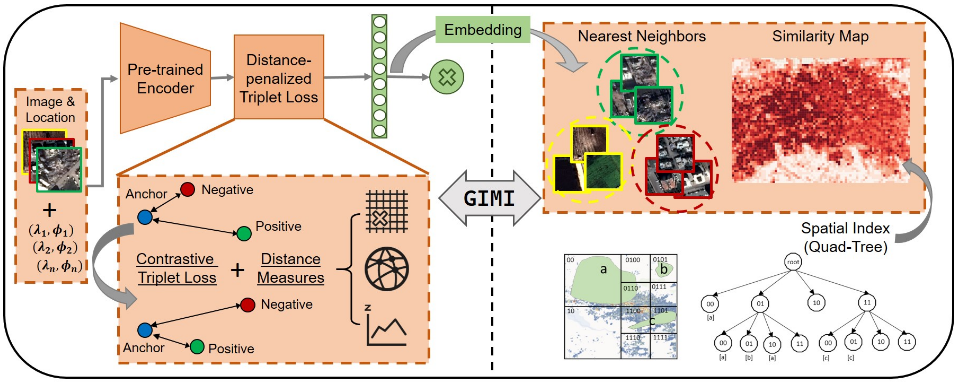
1 INTRODUCTION

Earth Observation (EO) via Remote sensing (RS) is one of the most fascinating and fast-growing techniques for collecting big geospatial data (BGD). More recently, it has become possible to gather multiple sensor modalities (e.g., high-resolution aerial images, multi- and hyperspectral images, Radar data, etc.) (Hong et al., 2023) to observe the Earth's surface from space at an unprecedented scale and frequency. These EO data form the backbone of state-of-the-art environmental and Earth System Science (ESS) research and the investigation of global challenges such as climate change, urbanization, and natural disasters.

However, the increasing resolution, quality, number of observations, coverage, and the amount of RS imagery that is being generated, all together posing pressing needs on how to search and mine large collections of EO data effectively and efficiently for interesting patterns. In this context, not only does one need to know where to look to find objects of interest, but also what model to use for different searching tasks. What if prior efforts had already created models on similar tasks but in another geographical study? Numerous downstream applications become possible if we can make large RS data collections searchable by content, metadata, and analytic tasks (Cavallaro et al., 2021; Li et al., 2023a).

Neural search, as an emerging resaerch field in information retrieval, has significantly influences the development of modern searching systems via leveraging neural network (NN) models to extract non-trivial and cross-domain data representations, so-called **embeddings**, to support multimodal searching purposed (e.g., text, image, video, audio). Different from traditional search engines, which rely on keyword matching or other heuristics to retrieve relevant objects, neural search models, especially transformer-based architectures, showed revolutionary performance and speed by encoding the *"query"* and *"corpus"* representations in a high-dimensional vector embedding space, allowing for more nuanced and robust searching results based on their semantic similarity (Gordo et al., 2016; Karpukhin et al., 2020). More recently, the popularity of neural search has been boosted by the availability of pre-trained **Large Language Models (LLM)**, such as BERT and ChatGPT, with a living and inestimable impact on the entire philosophy behind multimodal data search.

However, adapting existing neural search engines to geographical data and geospatial analysis tasks is not trivial. One major challenge is the generalizability of pre-trained models, which can degrade fast when scaling up to different geographical areas, from regional to global scales (Li et al., 2022). In this context, this performance decrease phenomenon can be formulated as the **"Geographical Generalizability"** issue of AI models (Li et al., 2023b), which follows the terminology of "Replicability across Space" (Goodchild & Li, 2021). Fortunately, recent works seek to address this issue via explicitly feeding models with intuitions for spatial or spatio-temporal information, with so-called geographic priors (e.g., latitude and longitude, or geographical context) (Tang et al., 2015; Mac Aodha et al., 2019; Mai et al., 2020; Yang et al., 2022; Mai et al., 2023b), in order to improve the model's"Geographical Generalizability" and ensure more robust and consistent model performance. More recently, the potential of explicit location encoding has been intensively studied via self-supervised contrastive learning with diverse image analytic tasks (Mai et al., 2023a; Rußwurm et al., 2023; Klemmer et al., 2023; Hong et al., 2024; Cepeda et al., 2023). However, the implication of location-explicit embeddings for a geographically generalizable neural search of RS images remains a question mark. Inspired by early works in this direction, we aim to fill the research gap by integrating location information into an image-to-image neural search engine via a novel contrastive learning loss. Herein, we explicitly consider different distance measures and the implication of their map projection errors in a potential global location encoding scenario.



(a) Image & Position Embedding

(b) Image-to-Image Search

Figure 1: Overview of GIMI. (a) the image and location-explicit embedding module, which is featured with the proposed *Distance-penalized Triplet Loss* based on various geographical distance measures; (b) the image-to-image search module, which supports both *Embedding-based Nearest Neighbors Searching* and customized spatial index building.

In this work, we introduce GIMI, a geographical generalizable image-to-image neural search engine, which learns high-dimensional vector embeddings from geo-locations and image representations by explicitly integrating different geographical distance measures into a contrastive learning objective. Based on learned embeddings, GIMI allows for flexible similarity search with a predefined, customized index, such as a spatial index (e.g., Quada-tree or R*-tree), using K-Nearest Neighbors (KNN) algorithms. We evaluate the performance of GIMI in a real-world critical task of searching for damaged buildings after the 2023 Kahramanmaraş Earthquake in Turkey using Very High Resolution (VHR) satellite imagery. Experiment results confirm the effectiveness and efficiency of GIMI over traditional image classification approaches while significantly reducing the overhead of model retraining and boosting the inferencing speed.

2 METHODOLOGY

Given a list of geographical images as $\mathbb{X} = \{(\boldsymbol{x}_i, \mathbf{I}_i)|i = 1, \ldots, m\}$, where $\boldsymbol{x}_i = (\lambda_i, \phi_i)$ is the geographical location (with latitude and longitude) and $\mathbf{I}_i$ represent the image feature space. Furthermore, we define the pre-trained image encoder $f()$ as a high-dimensional nonlinear function $f(\mathbf{I}_i, \boldsymbol{\theta}) : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^D$, which is parameterized by $\boldsymbol{\theta}$ and would map the input image feature space (i.e., spatial-spectral dimension of $H \times W \times C$) into a vector embedding representation of D dimension. To inject location-specific knowledge into image embeddings, we fine-tune the image encoder with triplets of image tiles on a classification task, where each triplet $\mathbf{T}_i$ consists of an anchor image t_a, a positive image t_p that is believed similar to t_a, and a negative image t_n that is dissimilar.

Following *the cluster hypothesis* that ”closely associated documents tend to be relevant to the same requests” (Voorhees, 1985), the most common approach of contrastive learning is to simultaneously minimize the Euclidean distance between the embeddings of the anchor t_a and the positive image t_p while maximizing the distance to the negative t_n. However, this common approach does not consider the geographical location (i.e., $\boldsymbol{x}_i$ for triplet $\mathbf{T}_i$) into the embedding learning process. Herein, we assume that *the influence of different positive and negative samples may differ when spatially clustered or co-located*. To this end, we proposed a general **Distance-Penalized Triplet (DPT)** loss function by extending the origin idea with a customized geographical distance term. So to say, for each triple $\mathbf{T}_i = \{t_a, t_p, t_n\}$ and the geographical location (with latitude and longitude) $\boldsymbol{x}_i$, the distance-penalized triplet loss can be calculated as follows:

$$L_{DPT} = [||f(t_a) - f(t_p)||_2 - ||f(t_a) - f(t_n)||_2 + P(\boldsymbol{x}_i) + a]_+ \tag{1}$$

Where $f()$ is an off-the-shelf pre-trained image encoder (e.g., ResNet 50 or ViT base) whose parameter $\boldsymbol{\theta}$ will be fine-tuned. To prevent the encoder from pushing the negative image without limitation, a rectifier term with margin m is introduced to keep the maximum distance between the anchor and negative smaller than m. Moreover, we define $P()$ as a customized distance function based on the geographical locations $\boldsymbol{x}_i = \{x_a, x_p, x_n\}$ of the triplet $\{t_a, t_p, t_n\}$, which are normalized among all triplets. This will lead to an additional penalization term by explicitly considering the real-world geographical distance among anchors, positives, and negatives in the embedding space. Specifically, the distance penalization term is defined as:

$$P(x_a, x_p, x_n) = q(x_a, x_p) + q(x_a, x_n) - q(x_p, x_n) \tag{2}$$

Where q() can be any geographical distance measures, such as projected (e.g., Euclidean and Manhattan) or geodetic (e.g., “Great-circle distance”) distances. The key idea is to penalize triplets that are geographically close to each other and pay more attention to those geographically distinct training samples (both positive and negative). Therefore, the DPT loss can embed triplet locations as static before the training, which is thus free from the gradient vanishing problem. The map projection theory applies here to support a flexible and sophisticated distance modeling on the surface of the Earth (Grafarend & Krumm, 2014). More specifically, one needs to account for the planar approximation distance with distinct **Tissot’s indicatrix** (Laskowski, 1989) across the global, similar to Mai et al. (2023b); Rußwurm et al. (2023). Based on exact map projection and geodetic datum, one can consider the following distance penalization with GIMI:

- **Spherical Earth Distance**: $q(x_i, x_j) = R\sqrt{(\triangle\phi)^2 + (cos(\phi_m)\triangle\lambda)^2}$, where $\triangle\phi = \phi_i - \phi_j$ and $\triangle\lambda = \lambda_i - \lambda_j$ are in radians. R and ϕ_m refer to the radius of the EARTH and mean latitude, respectively
- **Ellipsoidal Earth Distance**: $q(x_i, x_j) = \sqrt{(M(\phi_m)\triangle\phi)^2 + (N(\phi_m)cos(\phi_m)\triangle\lambda)^2}$, where M and N are the meridional and its perpendicular (”normal”). See more details about Ellipsoidal projection in Grafarend & Krumm (2014)

For the long term, we anticipate a significant impact of different distance measures on the DPT loss, especially when considering global scale pre-training and fine-tuning, where the effect of **Tissot's indicatrix** will play a substantial role. One intuitive example is that one meter (e.g., in the same projected coordinate system) in Greenland and Singapore will mean a completely different distance, which can easily destroy a global pretraining objective.

3 DATA AND EXPERIMENT

As a case study, we select the city of Adiyaman in Turkey (of 29.9 km^2), which has been severely affected by the 2023 Kahramanmaraş Earthquake in Turkey. We obtained VHR satellite imagery before and after the earthquake from the Pléiades 1A and 1B satellites as part of the open data supported by ITU-CSCRS [1], Turkey. In addition, 1456 geo-tagged images in three categories (i.e., 594 -*collapsed building*, 594 -*healthy building* and 268 -*non-building*) have been manually labeled as our search base following a train and test ratio of 0.2.

To evaluate the effectiveness of **GIMI**, we adopt two distinct pretrained encoder architecture, namely **ResNet-18** (He et al., 2015) and **ViT-B-16** (Dosovitskiy et al., 2021) pretrained on ImageNet1K_V1 (Russakovsky et al., 2015)). Then, we examine three distinct embeddings with the plain embedding from the pretrained base encoder: 1) **Softmax** embeddings using the cross-entropy loss; 2) Contrastive embeddings using a normal **Triplet loss**; 3) Contrastive embeddings using the **DPT loss**. Next, we pass query images through the image and location embedding models to get a fixed-length vector representation of each image. Then, we calculate **K-Nearest-Neighbors** in the embedding feature space and sort them according to their cosine similarity. Lastly, the search results are presented by a ranking list (or heatmap) together with similarity scores of all searching candidates.

Table 1: Comparison of GIMI's embeddings for damage building search.

Encoder	Method	Fine-tuned	Top 5% (%)	NDCG at 5%	Top 10% (%)	NDCG at 10%
ResNet	Base	✗	68.97 ± 19.87	0.908 ± 0.081	66.30 ± 17.29	0.912 ± 0.072
	Softmax	✓	92.22 ± 23.19	**0.985** ± 0.055	91.68 ± 23.44	0.974 ± 0.084
	Triplet Loss	✓	**94.06** ± 17.24	0.982 ± 0.060	94.11 ± 14.65	0.985 ± 0.048
	DPT Loss	✓	94.03 ± 17.61	0.983 ± 0.062	**95.32** ± 13.24	**0.987** ± 0.048
ViT	Base	✗	67.39 ± 13.83	0.921 ± 0.059	61.21 ± 10.99	0.907 ± 0.053
	Softmax	✓	93.46 ± 15.49	0.982 ± 0.055	91.50 ± 17.95	0.977 ± 0.069
	Triplet Loss	✓	96.09 ± 12.00	0.990 ± 0.035	93.97 ± 13.81	0.988 ± 0.040
	DPT Loss	✓	**98.04** ± 5.63	**0.995** ± 0.015	**96.96** ± 7.07	**0.993** ± 0.024

Table 1 reports the preliminary results from our case study, where we compare different settings of GIMI, each using 500 bootstrapped query images, and calculate the average accuracy of the Top 5% and Top 10% searching results. In addition, the average Normalized Discounted Cumulative Gain (NDCG) with a binary relevance score is calculated at Top 5% and Top 10% positions (Wang et al., 2013). Two findings are important here: first, contrastive embeddings, especially using the DPT Loss, outperforms classic deep features (both Softmax and Triplet Loss) and leads to superior accuracy and NDCG; second, GIMI with ViT yields the best performance which allows further extension via global-scale representative learning approaches in e.g.,Mai et al. (2023a); Rußwurm et al. (2023).

4 DISCUSSION AND CONCLUSION

Fast and accurate retrieval of satellite images (i.e., **Image-to-Image search**) from massive EO data archive emerges as a substantial task, especially under a **disaster mapping** scenario, where speed and accuracy are counted by human lives. Multiple challenges and needs are entangled herein, thus requiring an integrated solution to simultaneously ensure the model's accuracy, speed, and geographical generalizability. In this context, GIMI is the first kind of such image-to-image search engine designed especially for geographical applications, which contribute to existing methods for image classification and retrieval of EO data from a novel perspective. Our future work will focus on 1) extending the **Distance-Penalized Triplet (DPT)** loss into a self-supervised approach at a

[1] `https://web.cscrs.itu.edu.tr/kahramanmaras-earthquakes/`

global scale; 2) benchmarking scalable searching methods, such as Faiss (Douze et al., 2024) and ScaNN (Guo et al., 2020), with GIMI embeddings; (3) validating the geographical generalizable GIMI. We are looking forward to a broad application of GIMI and its variations in real-world and critical mapping scenarios.

5 REFERENCE

REFERENCES

Gabriele Cavallaro, Dora B Heras, Dalton Lunga, Martin Werner, and Andreas Züfle. *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data*. ACM, 2021.

Vicente Vivanco Cepeda, Gaurav Kumar Nayak, and Mubarak Shah. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

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

Michael F Goodchild and Wenwen Li. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118(35): e2015759118, 2021.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pp. 241–257. Springer, 2016.

Erik W Grafarend and Friedrich W Krumm. *Map projections*. Springer, 2014.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 2020. URL `https://arxiv.org/abs/1908.10396`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Danfeng Hong, Bing Zhang, Hao Li, Yuxuan Li, Jing Yao, Chenyu Li, Martin Werner, Jocelyn Chanussot, Alexander Zipf, and Xiao Xiang Zhu. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299:113856, 2023.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. DOI:10.1109/TPAMI.2024.3362475.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pp. 6769–6781. Association for Computational Linguistics (ACL), 2020.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

Piotr H. Laskowski. The traditional and modern look at tissot's indicatrix. *The American Cartographer*, 16(2):123–133, 1989. doi: 10.1559/152304089783875497.

Hao Li, Benjamin Herfort, Sven Lautenbach, Jiaoyan Chen, and Alexander Zipf. Improving openstreetmap missing building detection using few-shot transfer learning in sub-saharan africa. *Transactions in GIS*, 26(8):3125–3146, 2022.

Hao Li, Gabriele Cavallaro, Dora B. Heras, Dalton Lunga, Martin Werner, and Andreas Züfle. Geosearch '23: Proceedings of the 2nd acm sigspatial international workshop on searching and mining large collections of geospatial data. New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9798400703522.

Hao Li, Jiapan Wang, Johann Maximilian Zollner, Gengchen Mai, Ni Lao, and Martin Werner. Rethink geographical generalizability with unsupervised self-attention model ensemble: A case study of openstreetmap missing building detection in africa. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–9, 2023b.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019.

Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, and Ni Lao. Multi-scale representation learning for spatial feature distributions using grid cells. In *International Conference on Learning Representations*, 2020.

Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. *arXiv preprint arXiv:2305.01118*, 2023a.

Gengchen Mai, Yao Xuan, Wenyun Zuo, Yutong He, Jiaming Song, Stefano Ermon, Krzysztof Janowicz, and Ni Lao. Sphere2vec: A general-purpose location representation learning over a spherical surface for large-scale geospatial predictions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 202:439–462, 2023b.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015.

Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *arXiv preprint arXiv:2310.06743*, 2023.

Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pp. 1008–1016, 2015.

Ellen M. Voorhees. The cluster hypothesis revisited. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '85, pp. 188–196, New York, NY, USA, 1985. Association for Computing Machinery. ISBN 0897911598. doi: 10.1145/253495.253524. URL `https://doi.org/10.1145/253495.253524`.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Tie-Yan Liu, and Wei Chen. A theoretical analysis of ndcg type ranking measures, 2013.

Lingfeng Yang, Xiang Li, Renjie Song, Borui Zhao, Juntian Tao, Shihao Zhou, Jiajun Liang, and Jian Yang. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10945–10954, 2022.