
IMG2LOC: REVISITING IMAGE GEOLOCALIZATION USING MULTI-MODALITY FOUNDATION MODELS AND IMAGE-BASED RETRIEVAL-AUGMENTED GENERATION

Zhongliang Zhou*
University of Georgia
zz42551@uga.edu

Jielu Zhang*
University of Georgia
jz20582@uga.edu

Zihan Guan
University of Virginia
bxv6gs@virginia.edu

Mengxuan Hu
University of Virginia
qtq7su@virginia.edu

Ni Lao
University of Georgia
noon99@gmail.com

Lan Mu
University of Georgia
mulan@uga.edu

Sheng Li†
University of Virginia
shengli@virginia.edu

Gengchen Mai†
University of Georgia
gengchen.mai25@uga.edu

ABSTRACT

Geolocating precise locations from images presents a challenging problem in computer vision and information retrieval. Traditional methods typically employ either classification—dividing the Earth’s surface into grid cells and classifying images accordingly, or retrieval—identifying locations by matching images with a database of image-location pairs. However, classification-based approaches are limited by the cell size and cannot yield precise predictions, while retrieval-based systems usually suffer from poor search quality and inadequate coverage of the global landscape at varied scale and aggregation levels. To overcome these drawbacks, we present **Img2Loc**, a novel system that redefines image geolocation as a text generation task. This is achieved using cutting-edge large multi-modality models (LMMs) like GPT-4V or LLaVA with retrieval augmented generation. **Img2Loc** first employs CLIP-based representations to generate an image-based coordinate query database. It then uniquely combines query results with images itself, forming elaborate prompts customized for LMMs. When tested on benchmark datasets such as Im2GPS3k and YFCC4k, **Img2Loc** not only surpasses the performance of previous state-of-the-art models but does so without any model training.

Keywords Image Localization · Large Multi-modality Models · Vector Database

1 Introduction

The field of visual recognition has witnessed a marked improvement, with state-of-the-art models significantly advancing in areas such as object classification [1, 2, 3, 4], object detection [5, 6, 7], semantic segmentation [8, 9, 10, 11], scene parsing [12, 13], disaster response [14, 15], environmental monitoring [16] among others [17, 18, 19]. As progress moves forward, the information retrieval community is widening its focus to include the prediction of more detailed and intricate attributes of information. A key attribute in this expanded scope is image geolocation [20, 21, 22], which aims to determine the exact geographic coordinates given an image. The ability to accurately geolocalize images is crucial, as it provides possibilities for deducing a wide array of related attributes, such as temperature, elevation, crime rate, population density, and income level, providing a comprehensive insight into the context surrounding the image.

In our study, we delve into predicting the geographic coordinates of a photograph solely from the ground-view image. Predictions are considered accurate if they closely match the actual location (Figure 1). Prevailing research approaches fall under the categories of either retrieval-based or classification-based methods. Retrieval-based techniques compare query images against a geo-tagged image database [23, 24, 25, 26, 27, 28, 29, 30], using the location of the image that closest matches the query image to infer its location. Although straightforward, this method faces hurdles such as the

*These authors contributed equally to this work

†Corresponding authors

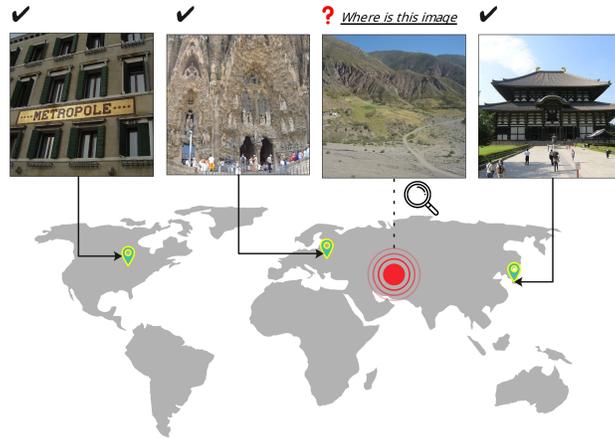


Figure 1: The image geolocation problem refers to predicting the coordinates of any given image.

complexity of feature extraction, the computational intensity of nearest neighbor search, and potential inaccuracies from over-reliance on database image locations. Alternatively, classification-based methods treat geolocation as a classification task [31, 32, 33, 34, 35, 36, 37], segmenting the Earth into discrete cells and training neural networks to classify images into these cells. However, this approach can yield significant errors, especially when the actual location of an image differs significantly from the center of its assigned cell. Moreover, the predefined cell structure introduces inherent limitations and biases, decreasing generalizability and accuracy across various global locations.

Given the inherent limitations of both retrieval and classification approaches, we are transitioning to the more recent and increasingly dominant approach of foundation models. In this vein, we propose a generative approach that predicts the geographic coordinates of new query images using a reference gallery and multimodality language models, named **Img2Loc**. Initially, we transform all geo-tagged images into embeddings using the CLIP model [38], creating a vast embedding space. To navigate this space efficiently, we use a vector database with GPU-accelerated search algorithms, quickly pinpointing and retrieving gallery images similar to the query images. Next, we formulate elaborate prompts integrating the image and the geographical coordinates of these reference points and feed them into state-of-the-art multi-modality models like GPT-4V [39] or LLaVA [40], known for their adeptness in generating accurate outputs from combined image and text inputs. To further improve the accuracy, we introduce negative sampling by identifying and using the most dissimilar points in the database as a counterreference. This sharpens the model’s ability to distinguish between relevant and irrelevant data points. Our model, when evaluated on well-established datasets such as Im2GPS3k [32] and YFCC4k [32], demonstrates notable advances, significantly outperforming the prior state-of-the-art methods without any model fine-tuning. This highlights the efficacy of our generative approach, which synergizes the retrieval method’s strengths with the advanced understanding and generative prowess of contemporary language models.

In summary, our study makes significant strides for the task of image geolocation, marked by the following contributions:

- To the best of our knowledge, this study is the first successful demonstration of multi-modality foundation models in addressing the challenges of geolocation tasks.
- Our approach is training-free, avoiding the need for specialized model architectures and training paradigms and significantly reducing the computational overhead.
- Using a refined sampling process, our method not only identifies reference points closely associated with the query image but also effectively minimizes the likelihood of generating coordinates that are significantly inaccurate.
- We achieve outstanding performance on challenging benchmark datasets including Im2GPS3k and YFCC4k compared with other state-of-the-art approaches.

2 Related Work

Image Geolocation as a classification task. The predominant approach for the image geolocation problem involves first segmenting the planet’s surface into discrete grids, such as the Google S2 grid, and assigning a geographic

coordinate to each grid [31, 32, 33, 34, 35, 37, 36]. This methodology permits a model to directly predict a class, thereby simplifying the complex task of geolocation into a more manageable form of classification. To refine this approach and introduce granularity into the prediction, recent advances have involved partitioning the Earth’s surface into multiple levels, offering a hierarchical, multi-scale perspective of localization [33]. However, while this cell-based classification system simplifies the prediction process, it inherently introduces localization errors, particularly if the actual location of interest lies far from the center of the predicted cell. This discrepancy stems from the coarse nature of cell-based classification, where the precision of localization is inherently limited by the size and scale of the cells defined in the model.

Image Geolocation as a retrieval task. In another direction, image geolocation has significantly evolved from rudimentary methods to sophisticated retrieval-based systems over the years [23, 24, 25, 26, 27, 28, 29, 30]. Retrieval-based systems, recognized for their intuitiveness, leverage these extensive databases to find matches for query images based on feature similarity in a multi-dimensional space. However, creating planet-level reference datasets for these systems presents formidable challenges, not limited to scale but also encompassing data diversity, temporal changes, and the need for precise annotations. To address the intrinsic differences in ground and aerial perspectives, separate models are often adopted, with the integration of these models aiming to provide a more comprehensive understanding and robust localization system [41]. Nonetheless, this integration introduces the significant hurdle of misalignment between perspectives, potentially undermining accuracy. An innovative solution to this challenge is the implementation of non-uniform cropping [27], which selectively focuses on the most informative patches of aerial images. This method prioritizes features that offer distinctive geographical cues, enhancing precision by addressing the issue of non-uniform feature distribution across different views.

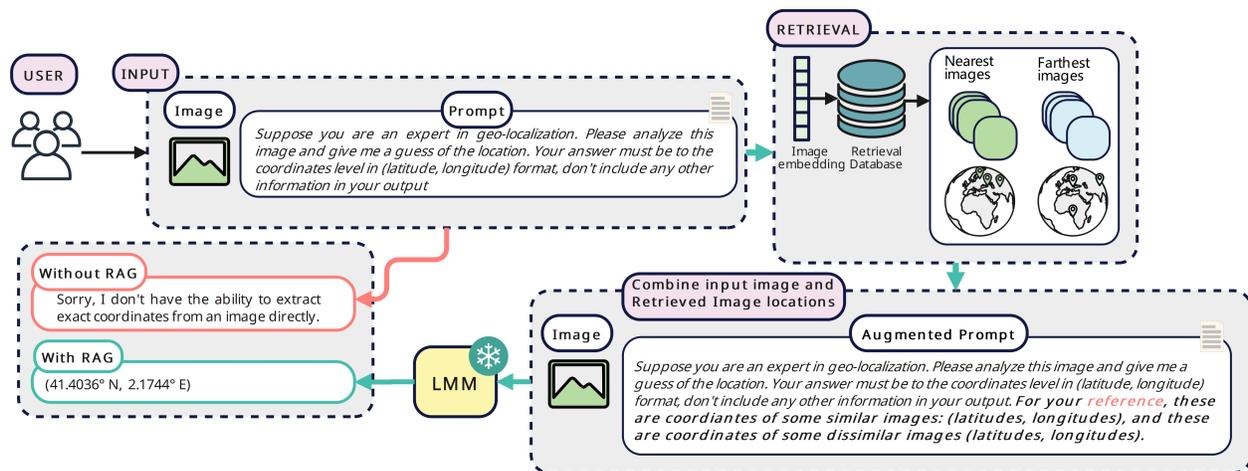


Figure 2: The architecture of the proposed framework.

Multi-modality foundation models and Retrieval-Augmented Generation. Large language models like GPT-4 [42] and LLaMA [43] have set new benchmarks in natural language processing, surpassing human performance in a variety of tasks as evidenced by their results on SuperGLUE [44] or BIG-bench [45]. Their exceptional zero-shot capabilities enable these generative models to be applicable across a wide range of research domains. Building on this success, multi-modality models such as GPT-4V [39] and LLaVA [40] have extended the prowess of large language models (LLMs) into the visual domain. However, alongside these remarkable achievements, certain challenges have become evident, most notably issues related to hallucination and reliance on outdated knowledge databases. These issues can compromise the reliability and trustworthiness of models’ outputs. To address these concerns, innovative methodologies such as the chain of thoughts (COT) [46] reasoning and Retrieval-Augmented Generation (RAG) [47] have been developed. These approaches significantly enhance the fidelity of the models’ responses. In particular, RAG represents a groundbreaking advancement by merging the powerful reasoning capabilities of foundation models (FM) with up-to-date external information. This is achieved by augmenting the input prompt with pertinent information retrieved from a comprehensive and up-to-date database. Such a process ensures that the model’s generations are not only creative and contextually aware but also grounded in solid, verifiable evidence. Consequently, this approach markedly improves the accuracy and relevance of the results, mitigating some of the earlier concerns associated with large language models. The integration of external databases into the generative process ensures that the output of these models remains both innovative and anchored in reality.

3 Method

Our method allows the user to input any image of interest for geolocation. Subsequently, the image is processed by the query and retrieval module, wherein the locations of the most similar and most dissimilar images are extracted. The image, along with these two sets of locations, is then fed into a multi-modality model. Finally, the geolocation result is displayed as an interactive map, which can be explored via a web interface. We will explain each step in the following sections.

3.1 Construction of the Image-Location Database

The core of the retrieval-based image localization system lies in how images are encoded into the database and how the nearest neighbor search is performed. Here, We utilize the CLIP model [38] for feature encoding and employ FAISS for the storage of resulting embeddings (Figure 3).

CLIP-Based Feature Encoding. The CLIP model, which is widely used as a fundamental representation model for various downstream tasks [48], is introduced to generate semantic embeddings of images. In particular, it accepts the query image and outputs the semantic embeddings of the outputs, which encapsulate rich information about the image in a condensed vector space. Utilizing the MediaEval Placing Tasks 2016 (MP-16) dataset [49], we have constructed a database encompassing over four million image embedding-location pairs, providing comprehensive coverage of the Earth’s surface.

Efficient Nearest Neighbor Search in the Vector Database. Once image embeddings are generated, it becomes crucial to store them in an efficient manner to facilitate effective search operations. To address the challenge, we use FAISS, a vector-based data storage system [50], which utilizes flat indexes and GPU parallel computation techniques to enhance efficiency. Then, to find the nearest neighbor for the query image, we propose using the inner product of the image embedding provided by the CLIP as the measurement. The underlying principle is straightforward: a higher inner product value signifies a greater level of similarity, and vice versa. This system allows us to generate an arbitrary number of nearest neighbors with ease.

Moreover, we posit that identifying images most dissimilar to the query image (positive neighbors) can also contribute to ruling out implausible locations, as they usually represent scenes that are geographically distant from the query image. This “negative neighbors search” is executed by finding the farthest neighbors for the negative query embedding. Upon completion of this search, the locations of both positive and negative neighbors are integrated into the subsequent step of our process.

3.2 Generate locations with augmented prompt

Current multi-modality foundation models, such as GPT-4V and LLaMA, accept input from both images and text to generate responses. In our approach, we conceptualize the task of image geolocation as a text generation task. Specifically, we prompt these foundation models to provide the precise latitude and longitude corresponding to a given image. We enhance the input prompt with additional information derived from our retrieval of similar and dissimilar locations (Figure 2). The similar images’ coordinates and dissimilar images’ coordinates will be appended to the text prompt as anchor information.

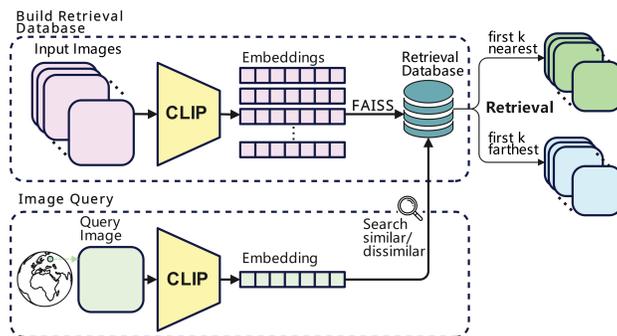


Figure 3: The retrieval and neighbor search pipeline.

4 Experiments

4.1 Datasets and Evaluation Details

We build our search database using the MediaEval Placing Tasks 2016 (MP-16) dataset [49], encompassing 4.72 million geotagged images from Flickr[†]. The performance of our model is evaluated using Im2GPS3k [32] and YFCC4k [32] datasets. We compute the geodesic distance between the predicted and actual geographical coordinates for each test image and quantify the proportion of these predictions that align with set distance thresholds (1km, 25km, 200km, 750km, and 2500km). In terms of multi-modality models, our focus is on GPT-4V and LLaVA, selected for their availability and superior performance. It’s noteworthy that our framework is designed for flexibility, allowing for seamless integration of the latest model releases as they become available.

4.2 Results

Dataset	Method	Distance (a_r [%] @ km)				
		Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
Im2GPS 3k [32]	[L]kNN, $\sigma = 4$ [32]	7.2	19.4	26.9	38.9	55.9
	PlaNet [35]	8.5	24.8	34.3	48.4	64.6
	CPlaNet [36]	10.2	26.5	34.6	48.6	64.6
	ISNs (M, f, S_3) [33]	10.1	27.2	36.2	49.3	65.6
	Translocator [31]	11.8	31.1	46.7	58.9	80.1
	GeoGuessNet [34]	12.8	33.5	45.9	61.0	76.1
	GeoCLIP [22]	14.11	34.47	50.65	69.67	83.82
	Img2Loc(LLaVA)	7.98	23.37	29.94	40.11	51.12
	Img2Loc(GPT4V)	17.10	45.14	57.87	72.91	84.68
YFCC 4k [32]	[L]kNN, $\sigma = 4$ [32]	2.3	5.7	11.0	23.5	42.0
	PlaNet [35]	5.6	14.3	22.2	36.4	55.8
	CPlaNet [36]	7.9	14.8	21.9	36.4	55.5
	ISNs (M, f, S_3) [33]	6.5	16.2	23.8	37.4	55.0
	Translocator [31]	8.4	18.6	27.0	41.1	60.4
	GeoGuessNet[34]	10.3	24.4	33.9	50.0	68.7
	GeoCLIP[22]	9.59	19.31	32.63	55.0	74.69
	Img2Loc(LLaVA)	7.93	14.20	19.51	29.98	39.72
	Img2Loc(GPT4V)	14.11	29.57	41.40	59.27	76.88

Table 1: Geo-localization accuracy of the proposed method compared to previous methods, across two baseline datasets.

The data presented in Table 1 demonstrates that our methods outperform previous classification and retrieval methods across all granularity levels on both tested datasets. On the Im2GPS3k dataset, we have achieved significant improvements over the prior top-performing method, GeoCLIP, without ever training any of the models on geo-tagged data (MP-16 dataset [49]) The improvements are +2.89%, +10.67%, +7.22%, +3.24%, and +0.86% at the 1km, 25km, 200km, 750km, and 2500km thresholds, respectively. Furthermore, on the YFCC4k dataset, our method surpasses the previous best model, GeoGuessNet, by margins of +3.81%, +5.17%, +7.5%, +9.27%, and +8.18% for the same respective distance thresholds.

5 Conclusion

In our study, we present Img2Loc, a cutting-edge system that harnesses the power of multi-modality foundation models and integrates advanced image-based information retrieval techniques for image geolocation. Our approach has demonstrated evidently-improved performance when compared to existing methods. We envision Img2Loc as a compelling example of leveraging modern foundation models to address complex problems in a streamlined and effective manner.

[†]<https://www.flickr.com/>

References

- [1] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.
- [2] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.
- [3] 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, 2023.
- [4] Gengchen Mai, Ni Lao, Yutong He, Jiaming Song, and Stefano Ermon. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *the Fortieth International Conference on Machine Learning (ICML 2023)*, 2023.
- [5] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.
- [7] 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*, pages 1–9, 2023.
- [8] Ziyang Wang and Chen Yang. Mixsegnet: Fusing multiple mixed-supervisory signals with multiple views of networks for mixed-supervised medical image segmentation. *Engineering Applications of Artificial Intelligence*, 133:108059, 2024.
- [9] Ziyang Wang, Meiwen Su, Jian-Qing Zheng, and Yang Liu. Densely connected swin-unet for multiscale information aggregation in medical image segmentation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 940–944. IEEE, 2023.
- [10] Ziyang Wang and Congying Ma. Dual-contrastive dual-consistency dual-transformer: A semi-supervised approach to medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 870–879, 2023.
- [11] Jielu Zhang, Zhongliang Zhou, Gengchen Mai, Lan Mu, Mengxuan Hu, and Sheng Li. Text2seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv preprint arXiv:2304.10597*, 2023.
- [12] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4003–4012, 2020.
- [13] Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 775–793. Springer, 2020.
- [14] Wenchong He, Arpan Man Sainju, Zhe Jiang, Da Yan, and Yang Zhou. Earth imagery segmentation on terrain surface with limited training labels: A semi-supervised approach based on physics-guided graph co-training. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–22, 2022.
- [15] Wenchong He, Arpan Man Sainju, Zhe Jiang, and Da Yan. Deep neural network for 3d surface segmentation based on contour tree hierarchy. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 253–261. SIAM, 2021.
- [16] Wenchong He, Zhe Jiang, Marcus Kriby, Yiqun Xie, Xiaowei Jia, Da Yan, and Yang Zhou. Quantifying and reducing registration uncertainty of spatial vector labels on earth imagery. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 554–564, 2022.
- [17] Jingdi Chen, Lei Zhang, Joseph Riem, Gina Adam, Nathaniel D Bastian, and Tian Lan. Ride: Real-time intrusion detection via explainable machine learning implemented in a memristor hardware architecture. In *2023 IEEE Conference on Dependable and Secure Computing (DSC)*, pages 1–8. IEEE, 2023.

- [18] Jingdi Chen, Tian Lan, and Nakjung Choi. Distributional-utility actor-critic for network slice performance guarantee. In *Proceedings of the Twenty-fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 161–170, 2023.
- [19] Jingdi Chen, Hanhan Zhou, Yongsheng Mei, Gina Adam, Nathaniel D Bastian, and Tian Lan. Real-time network intrusion detection via decision transformers. *arXiv preprint arXiv:2312.07696*, 2023.
- [20] 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.
- [21] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. *arXiv preprint arXiv:2006.12567*, 2020.
- [22] 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.
- [23] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocation with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. *arXiv preprint arXiv:2204.00097*, 2022.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] 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. *arXiv preprint arXiv:2204.13861*, 2022.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] Tobias Weyand, Ilya Kostrikov, and James Philbin. Planet-photo geolocation with convolutional neural networks. In *European Conference on Computer Vision*, pages 37–55. Springer, 2016.
- [36] Paul Hongsuck Seo, Tobias Weyand, Jack Sim, and Bohyung Han. Cplanet: Enhancing image geolocation by combinatorial partitioning of maps. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018.
- [37] Mike Izbicki, Evangelos E Papalexakis, and Vassilis J Tsotras. Exploiting the earth’s spherical geometry to geolocate images. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 3–19. Springer, 2020.

- [38] 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.
- [39] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [41] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019.
- [42] 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.
- [43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [44] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.
- [45] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.
- [46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [47] Deng Cai, Yan Wang, Lema Liu, and Shuming Shi. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3417–3419, 2022.
- [48] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.
- [49] 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.
- [50] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.