Geo-Foundation Models: Reality, Gaps and Opportunities (Vision Paper)

Yiqun Xie[1], Zhaonan Wang[2], Gengchen Mai[3], Yanhua Li[4], Xiaowei Jia[5], Song Gao[6], Shaowen Wang[2]
[1]University of Maryland, [2]University of Illinois Urbana-Champaign, [3]University of Georgia, [4]Worcester Polytechnic Institute, [5]University of Pittsburgh, [6]University of Wisconsin-Madison
xie@umd.edu,{znwang,shaowen}@illinois.edu,gengchen.mai25@uga.edu,yli15@wpi.edu,xiaowei@pitt.edu,song.gao@wisc.edu

ABSTRACT

With the recent rapid advances of revolutionary AI models such as ChatGPT, foundation models have become a main topic for the discussion of future AI. Despite the excitement, the success is still limited to specific types of tasks. Particularly, ChatGPT and similar foundation models have unique characteristics that are difficult to replicate for most geospatial tasks. This paper envisions several major challenges and opportunities in the creation of geospatial foundation (geo-foundation) models, as well as potential future adoption scenarios. We also expect that a major success story is necessary for geo-foundation models to take off in the long term.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Spatial-temporal systems**.

KEYWORDS

AI, GeoAI, foundation models, geospatial data

ACM Reference Format:
Yiqun Xie[1], Zhaonan Wang[2], Gengchen Mai[3], Yanhua Li[4], Xiaowei Jia[5], Song Gao[6], Shaowen Wang[2]. 2023. Geo-Foundation Models: Reality, Gaps and Opportunities (Vision Paper). In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625616

1 INTRODUCTION

While ChatGPT and similar foundation models for generative tasks have fueled huge excitement around AI and raised the expectation, the success is still limited to specific types of tasks. Recent discussions have considered various possibilities of adopting ChatGPT for language-related tasks in geospatial fields [14, 16]. However, through a closer look, we can find there are unique characteristics of these foundation models that make them hard to adapt to general geospatial tasks. In this paper, we envision several major challenges and opportunities in the development of geo-foundation models, as well as potential adoption scenarios. We start with a brief overview of the current status of foundation models and GeoAI as follows.

[*]Yiqun Xie and Zhaonan Wang contributed equally to this research.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3625616

1.1 Success Stories from Broad AI

1.1.1 Imitating Human Behaviors. The term *Artificial Intelligence* (AI), originally coined in 1950s, refers to a machine's ability to do tasks that require human intelligence. This field has undergone ups and downs, and recently become tangible to everyone's life because of the revolutionary milestone of ChatGPT. It is essentially powered by a Large Language Model (LLM), namely Generative Pre-trained Transformers (GPT), to simulate human-like conversations with users. LLMs such as ChatGPT and other generative models such as DALL-E 2 are major success stories that are able to imitate human behaviors while being substantially more efficient.

1.1.2 Surpassing Human Capabilities. There are also specific cases where AI algorithms reach beyond-human performance, such as AlphaGo on playing the board game Go and AlphaFold on predicting complex 3D protein structures. While these are brilliant successes, each effort tends to take a huge amount of resources and there are only very few successes to date for specific applications.

1.2 Foundations of "Geo"-AI are Still Needed

"Geo"-AI has been researched for many years. However, the necessity of the term "Geo" from the technical perspective still needs to be thoroughly developed to establish its foundations. This means that, especially in the long term, GeoAI should not be only applying general AI methods to geospatial data. Instead, it should represent a family of spatially-oriented AI methods that explicitly address the challenges posed by geospatial data (e.g., spatial heterogeneity, spatial data representation learning, etc.), which violate certain assumptions of general AI methods and cannot be well-addressed without GeoAI. These foundations are much needed for further research on "Geo" foundation models. Otherwise, it would just be typical foundation models being applied to geospatial datasets, which does not justify the "Geo" prefix - having a prefix for each of the numerous domains using AI is the same as having none.

2 MAJOR CHALLENGES

2.1 Challenges in Replicating the Success of "ChatGPT" and Generative Models

While LLMs such as ChatGPT have shown revolutionary success and demonstrated strong capabilities in a variety of language-based and generative tasks, the success can be hard to translate to many-geospatial problems. The success of ChatGPT-type of foundational LLMs (and broad generative models) has two important ingredients:

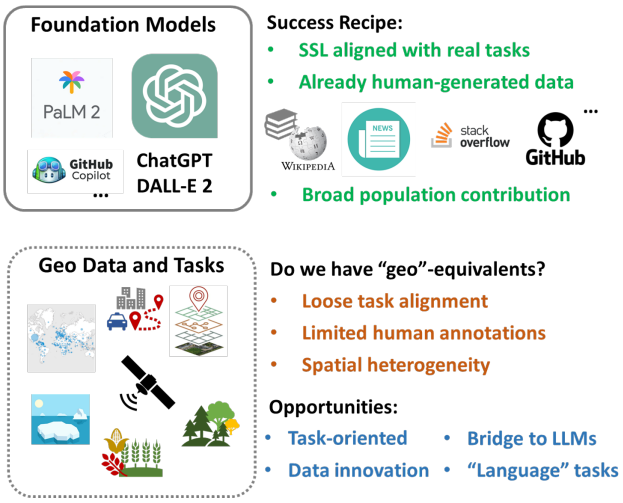

Figure 1: Challenges and opportunities for geo-foundation models.

(1) A very large model with a number of parameters at billions-to-trillions level; and (2) Self-supervised learning (SSL) based pretraining [11]. Specifically, there are several characteristics of the SSL component that are non-trivial to replicate:

- **Tight alignment between SSL and real-tasks:** SSL is mainly used for pretraining large models, which are otherwise very difficult to train due to their huge amount of parameters. The usefulness of SSL depends on the alignment between SSL (a.k.a. pretext) and downstream tasks [20]. In text-related tasks such as language generation or chatting, the commonly used SSL tasks (e.g., word masking and prediction) can be well-aligned with the actual downstream NLP tasks. This is similar to general generative tasks in computer vision. For example, the objective of GAN or diffusion models directly aligns with the task of realistic image generation [11]. However, many geospatial problems, except a few such as trajectory generation/classification (e.g., [7]), do not naturally have SSL tasks whose objectives are tightly aligned with the downstream applications. For example, in satellite-based flood mapping, the goal of classifying "flooded" vs "non-flooded" is not directly related to existing SSL tasks such as inpainting.
- **Imitating humans with human-generated "labels":** As AI aims to imitate human intelligence, most AI models learn from human-generated labels. This is nearly ideal for text-related language problems, as text – no matter from Wikipedia, books, news, online chats, forums, code repositories, or elsewhere – is mostly created by humans with careful thinking. Thus, in some sense, the vast amount of data being leveraged in SSL for text-related tasks is not just regular data, but "annotated" data containing human-generated labels. This is highly different from other tasks such as computer vision, where the photos/videos do not contain human-generated labels needed for most applications. For example, to train an object detector in self-driving cars, we still need additional manual effort to label these scenes, which significantly constrains the amount of labels. This is even more challenging for many geospatial tasks where field work or expensive in-situ sensors are commonly needed for label collection [19] (more details in Sec. 2.2).
- **Significant contribution from broad population:** While both texts in articles and object annotations in images (or other labels) are human-generated content, a key distinction between the two can make it very difficult to have a similarly large human-labeled dataset in broad applications beyond text-related language problems. The difference is that most of the text-based materials are generated as an essential part of the creators' daily life, such as work, recreation, or other communications. However, labels such as object annotations in images are not, and they are mostly created with the pure purpose of training or evaluating machine learning models, significantly confining the possible amount of contribution. Not only that, the text data are generated by a substantial proportion of the broad population. Such a level of human-generated content is off-the-chart and difficult to obtain for broad tasks including many geospatial problems.

Thus, SSL with super large models may be a perfect recipe for text-related language problems as well as general generative problems (e.g., image/video generation),[1] the success may not be easily achievable in many important geospatial problems.

2.2 Challenges in Building the Datasets

Creating large and representative datasets for many geospatial problems can be challenging due to the following aspects: **(1) High cost:** Geospatial problems not only occur in highly developed countries and urban areas. Many of the most pressing issues facing our society – such as food security, wildfires and climate change – are closely related to rural or less developed regions, where the cost of ground-truth collection is very expensive. For example, field surveys are still one of the most common ways in large-scale crop monitoring for label collection [19]. **(2) Expert knowledge:** While citizen science platforms can be leveraged to increase participation from general population or internet citizens, labeling in many geospatial problems require certain expertise from training, such as recognizing different species of trees (over hundreds) and crops, identifying cropland with nutrient deficiency, classifying types of sea-ices, etc. **(3) Distribution shifts in space:** Geospatial problems often concern decisions and phenomena at large scales over time. Considering fundamental properties such as spatial heterogeneity [4, 21], this requires labels to be representative over space to cover the distribution shifts and localized efforts often do not generalize.

2.3 Challenges in Generalization

A motivation for building foundation models is to improve generalizability. In geospatial problems, the term "generalizability" should consider at least two different dimensions. First, it covers different types of tasks, which is similar to LLMs and so on. Second, it also needs to consider spatial generalizability within each task, which means the model should be able to adapt to different geographical regions and locations. This spatial generalization problem is commonly faced by practitioners in the geospatial domain. While the concept of a foundation model is attractive, it can be challenging to

[1] SSL tasks in the image-generation type of problems can be tightly aligned with the target real task and only require images/videos as input without major manual labeling efforts.

build a single model that fits all due to the fundamental property of spatial heterogeneity or variability. Given the physical and social complexity of real-world scenarios, our observations (i.e., features **X**) tend to be partial about the entire phenomena or events. The unobserved features (e.g., certain environmental conditions), however, are most likely not fixed constants over space [4, 21]. This potentially introduces conflicts between different locations, where similar observed features can correspond to different labels.

3 WHAT'S NEXT?

3.1 Success Modes and Opportunities

3.1.1 Task-oriented Foundation Models. Due to the spatial generalization challenge, forcing multiple tasks into one foundation model may introduce conflicts in training, considering that spatial heterogeneity or variability patterns can be different for different tasks. Thus, at the early stage, scoping down and building task-specific foundation models may help smooth the path. Such foundation models can focus on spatial generalization within individual tasks, which is already a major "geospatial" feat if successful. We envision several opportunities: **(1) Task-aligned self-supervised learning:** The alignment between SSL tasks and downstream tasks can directly impact the quality of pre-training [20]. In the ideal case, if an SSL task $\mathcal{T}_{SSL}$ is a dual (equivalent) problem of the original task $\mathcal{T}_{real}$, then the SSL task $\mathcal{T}_{SSL}$ can replace the need of human-labeled data in the supervised setting. Thus, the key is to design a task $\mathcal{T}_{SSL}$ to resemble $\mathcal{T}_{real}$ as close as possible. As a concrete example, Auto-CM defines an SSL task for the problem of cloud masking in satellite imagery, where the decrease of the loss on the SSL task directly depends on the model's ability to mask out clouds [23]. This dependency enforces the network to develop the ability to solve the original problem. **(2) Heterogeneity-aware spatial subtasking:** The ability to handle spatial heterogeneity is necessary for building a spatially generalizable foundation model. To achieve this, heterogeneity-aware learning frameworks are needed to recognize different data distributions and create sub-tasks with private parameters to resolve potential conflicts [21]. Sub-tasking alone does not address data imbalance issues over space, and advances in spatial adaptation and finetuning are needed for generalization to data-sparse regions. Physics- or process-based models built upon scientific theories may also be leveraged to enhance generalization in data-limited scenarios [10].

3.1.2 Bridges to Data-Rich LLMs. With advances in multi-modal learning, it is possible for non-text tasks to leverage latent features in LLMs that can be trained with huge datasets (e.g., vision-language pre-training[3]). Geospatial tasks can leverage this multi-modal potential, and the success will depend on the bridge-building ability between data-sparse and data-rich tasks. Recent examples of multi-modal learning between geographical information and text include geocoding [5] and scene classification for satellite images [17], but the level of success is still rather limited compared to the capabilities brought by LLMs. Spatiotemporal information may also be explored to combine with other modalities to improve prediction performance [7, 8, 15, 23].

3.1.3 Major Data Collection Efforts. To achieve generalizability via geo-foundation models, representative data collection is likely an essential step towards real success stories. While the data volume may not reach the level of LLMs' any time soon (Sec. 2.1, we anticipate that major data collection efforts will be carried out as they start to become the actual bottlenecks. Major investments or innovations may be required to build such datasets, and we envision the following opportunities: **(1) High-incentive spatial schemes:** The ground-truthing work may require strong incentives from the broad public to contribute. As many tasks require in-field observation beyond simple picture-looking (e.g., Amazon Mechnical Turk), new types of platforms may be needed with new payment and validation methods that explicitly consider spatial characteristics. **(2) AI-assisted labeling:** As general foundation models continue to mature, some can potentially be leveraged or customized to accelerate labelling for certain tasks. For example, CVAT has included recent models such as the Segment-Anything-Model [12]. While these models are still far from mature for direct automation in related geospatial tasks, they can be easily used as assistants. **(3) Low-cost sensors:** We also envision low-cost sensors to be more broadly deployed in combination with sparser installation of high-precision sensors as a way to collect in-situ labels (e.g., carbon emission, water level) at large scales to enable spatial generalization. **(4) Collaboration:** Academia-industry-government collaboration may be needed to enable geo-foundation models, as the data collection requires resources at scale (both monetary and human) beyond what are currently available to most academics.

3.1.4 Natural Language and Coding Tasks. The discussion above refers to more general geospatial tasks. For certain specific tasks, we expect the LLM-type of success to be more reproducible thanks to task similarities to LLM tasks. In principle, tasks that can be converted into standard coding should be lower-hanging fruits by prompting with LLMs. For example, spatial data processing, querying, and analysis workflows that can be formulated using Python code snippets with well-documented libraries can be generated by natural language commands. Geoparsing, as another instance, could be performed by zero-shot inference [14]. While these tasks are more direct applications of existing LLMs, they could substantially reduce manual efforts on such common procedures. Problem reformulation is another short-term possibility, where the similarity between the sequential structure of language and certain geospatial tasks can be leveraged to enhance the performance, such as POI recommendation [9, 13], activity detection [7], etc.

3.2 Adoption Scenarios

3.2.1 A Few Foundation Models. We envision that a more conservative scenario will be a few widely-adopted geo-foundation models being built or customized for geospatial problems. We expect to see task-oriented geo-foundation models (Sec. 3.1.1) appear first before more general versions can be developed. Such task-oriented models may also evolve into domain-oriented models that cover many tasks from the same application domain such as transportation and agriculture. To build more general geo-foundation models, major changes or advances in geospatial data generation (labels) may need to happen (Sec. 3.1.3).

3.2.2 Foundation Models for Everyone. A more radical or non-conservative scenario is that geo-foundation models may become a

replacement for our current deep learning models, just being bigger in size. A decade ago, training a deep neural network can be a luxury practice that can only be done by a few; they were also considered super-size models by the old standard. With advancements in software and hardware, deep learning models have become something that most researchers can easily train and use. Thus, it is not impossible that in the longer term, most of us will start to build and use these now-super-size models in our regular research activities. However, this may make the original meaning of "foundation models" disappear. In either scenario, efforts are needed to establish building blocks of "geospatial" AI for foundation models.

4 WHAT'S A GAME CHANGER?

To establish geo-foundation models, we envision that a game changer will be the release of a spatially-generalizable model, which robustly surpasses the performance of existing models at large scales and is broadly adopted by important real-world applications such as agriculture or transportation. This is needed to demonstrate the feasibility and practical value of geo-foundation models and establish the confidence of end-users and research communities. It is important to recognize that for geospatial problems in many critical sectors, the most common machine learning models being adopted in real products are still the very traditional methods, such as decision trees and random forests, even for tasks where deep learning or foundation models are thought to be good at. For example, USDA's Cropland Data Layer, as the most commonly used crop map for the US, still uses decision trees to generate a base map for manual refinement [2]. Thus, it is important to establish a real success story, even with a limited scope, before geo-foundation models can take off. In addition, a successful geo-foundation model needs to demonstrate unique capabilities to handle geospatial challenges – other than being one of the numerous customizations – to have long-term impacts.

5 OTHER ASPECTS: ETHICS AND MORE

This paper's discussion focuses on the aspects of technical feasibility related to the model performance. It is important to note that there are other aspects such as responsibility, fairness, ethics and sustainability that have been broadly discussed around foundation models [1], where most apply to geo-foundation models. Here we briefly discuss two unique aspects to consider in geo-foundation models: **(1) Locational fairness:** The large degree of freedom in deep neural networks has been shown to be easily biased over locations, where performances in certain regions can be largely compromised to chase a global score. Locational fairness metrics and frameworks (e.g., [6, 22]) need to be explicitly incorporated into geo-foundation models. **(2) Environmental justice:** The training of geo-foundation models can cause excessive carbon emissions and environmental problems. However, the models are likely not equally useful for people from different backgrounds and geographic regions [18]. Advances in AI training or new policies may be needed to regulate multi-scale environmental impact.

6 CONCLUSIONS

We envisioned the challenges and opportunities for geo-foundation models given the recent advancement and excitement around the ChatGPT-type of success. We found that there are major differences between geospatial tasks and the success-recipe behind typical generative models. Narrower-scope formulations, spatially-oriented models, multi-modal bridges and major data collection efforts are needed. More importantly, a success story focused on solving challenging real-world problems is still missing, but critical to demonstrate the feasibility and practical value of geo-foundation models.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2105133, 2126474, 2147195 and 2118329; NASA under Grant No. 80NSSC22K1164 and 80NSSC21K0314; USGS under Grant No. G21AC10207; Google's AI for Social Good Impact Scholars program; the DRI award and the Zaratan supercomputing cluster at the University of Maryland; and Pitt Momentum Funds award and CRC at the University of Pittsburgh.

REFERENCES

[1] Rishi Bommasani et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).
[2] CDL 2023. USDA Cropland Data Layer. https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php.
[3] Zhe Gan et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends Comput. Graph.* 14, 3–4 (2022), 163–352.
[4] Michael F Goodchild and Wenwen Li. 2021. Replication across space and time must be weak in the social and environmental sciences. *PNAS* 118, 35 (2021).
[5] Milan Gritta et al. 2018. Which melbourne? augmenting geocoding with maps. *ACL*.
[6] Erhu He et al. 2023. Physics Guided Neural Networks for Time-aware Fairness: An Application in Crop Yield Prediction. In *AAAI*.
[7] Mingzhi Hu et al. 2023. Self-supervised Pre-training for Robust and Generic Spatial-Temporal Representations. In *ICDM*.
[8] Mingzhi Hu, Xin Zhang, Yanhua Li, Xun Zhou, and Jun Luo. 2023. ST-iFGSM: Enhancing Robustness of Human Mobility Signature Identification Model via Spatial-Temporal Iterative FGSM. In *KDD*.
[9] Jizhou Huang et al. 2022. ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. In *KDD*. 3029–3039.
[10] Xiaowei Jia et al. 2021. Physics-guided machine learning from simulation data: An application in modeling lake and river systems. In *ICDM*. 270–279.
[11] Longlong Jing and Yingli Tian. 2020. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 43, 11 (2020), 4037–4058.
[12] Alexander Kirillov et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
[13] Zekun Li et al. 2022. SpaBERT: A Pretrained Language Model from Geographic Data for Geo-Entity Representation. *EMNLP* (2022).
[14] Gengchen Mai et al. 2022. Towards a foundation model for geospatial artificial intelligence (vision paper). In *SIGSPATIAL*. 1–4.
[15] Gengchen Mai et al. 2023. CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. *ICML* (2023).
[16] Gengchen Mai et al. 2023. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798* (2023).
[17] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.
[18] Meilin Shi et al. 2023. Thinking Geographically about AI Sustainability. *AGILE: GIScience Series* 4 (2023).
[19] Xiao-Peng Song, Matthew C Hansen, Peter Potapov, Bernard Adusei, Jeffrey Pickering, Marcos Adami, Andre Lima, Viviana Zalles, Stephen V Stehman, Carlos M Di Bella, et al. 2021. Massive soybean expansion in South America since 2000 and implications for conservation. *Nature sustainability* 4, 9 (2021), 784–792.
[20] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. 2021. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems* 34 (2021), 22682–22694.
[21] Yiqun Xie et al. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *ICDM*. 767–776.
[22] Yiqun Xie et al. 2022. Fairness by "Where": A Statistically-Robust and Model-Agnostic Bi-level Learning Framework. *AAAI* 36, 11 (Jun. 2022), 12208–12216.
[23] Yiqun Xie et al. 2023. Auto-CM: Unsupervised Deep Learning for Satellite Imagery Composition and Cloud Masking Using Spatio-Temporal Dynamics. In *AAAI*.