THE TRILLION-PIXEL CHALLENGE

Lexie Yang, ORNL
Dalton Lunga, ORNL
Jordan Lieberman, NGA
Timothy Doster, PNNL
Hannah Kerner, UMD
May Casterline, NVIDIA
Eric Shook, UMN
Edmon Begoli, ORNL
Rahul Ramachandran, NASA
Jitendra Kumar, ORNL
Fabio Pacifici, Maxar
Shawn Newsam, UC Merced
Steven Ward, ORNL
Budhu Bhaduri, ORNL

geoai.ornl.gov/trillion-pixel

DECEMBER 2021

ORNL/LTR-2021/2326

Geospatial Science and Human Security Division

2021 GeoAI TrillionPixel Workshop Report

December 2021

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831-6283
managed by
UT-Battelle LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

EXECUTIVE SUMMARY

The convergence of geospatial big data with advancements from artificial intelligence, cloud infrastructure, and high-performance computing continues to revolutionize mapping and analysis of EarthâĂŹs surface in unprecedented detail. Rapid innovations in sensing technologies will soon collect national and global geospatial data at even higher resolution resolutions and frequencies. These developments offer the potential for breakthroughs in science, policy, and national security via end-to-end GeoAI systems that can provide fresh insights into how humans occupy and alter their environment over time.

At the 2021 GeoAI Trillion Pixel workshop, international subject matter experts from government, academia, industry, and nonprofit organizations gathered virtually to discuss the Trillion Pixel GeoAI Challenge. The event focused on six major themes that are currently influencing scientific innovation and breakthroughs. A particular focus was paid to societal impacts. As an additional takeaway message, the gathering identified key application gaps and challenges that are in need of stronger community partnerships and collaborations.

The two-day workshop focused on the progress made from the first event held in 2019 * . The following emerging gaps in each of the the six themes were identified and featured through speaker presentations:

- **Trillion Pixels Grand Challenges:** Better human and machine teaming to foster deeper domain knowledge integration
- **Generalization and Transferability:** Consistent benchmarking standards and generalization toward global inferencing
- **Scalable Geospatial Processing Architectures:** Comprehensive yet application-specific optimization in processing architectures from step 0 (data acquisition)
- **Trustworthiness in GeoAI Systems:** Initiatives to set the requirements of building trustworthiness in GeoAI systems
- **GeoAI Beyond Pixels:** Significant research and development challenges in multimodality data integration
- **Collaborations and Community Engagements:** Identifying partnerships that can bring in major players from the private sector, academia, and international agencies to establish effective collaboration. Launching more initiatives to reach diverse communities

Six hundred sixty-seven participants representing various sectors including US federal agencies, private industry, academia, international agencies, and the US Department of Energy national laboratories (Table 1) joined this virtual event. And, twenty-nine panelists shared their critical perspectives on the six themes.

In summarizing the event, several key directions were identified that require a radical shift to address the Trillion Pixel challenge:

- The integration and fusion of multisource geospatial data offers broader impacts.
- Setting format and benchmark metrics standards or guidelines will facilitate knowledge exchange and advancement.

*Lunga, Dalton D., Alemohammad, Hamed, Liu, Yan, Newsam, Shawn, Pacifici, Fabio, Santos-Villalobos, Hector, Shook, Eric, Stewart, Robert N., Voisin, Sophie, Yang, Lexie, and Bhaduri, Budhu L. The Trillion Pixel GeoAI Challenge Workshop. United States: N. p., 2019. Web. doi:10.2172/1606744.

Table 1. Workshop participation by sector

Sector	No. participants
Academia	327
Federal agency	55
Industry	178
International agency	16
Laboratory	91

- Transfer learning approaches are critical for global generalization. Domain knowledge should be considered in this process.
- Seek both global support (resources) and local perspectives while looking for effective solutions from GeoAI systems.
- Build trustworthiness in GeoAI systems, promote model transparency and be aware of sample and algorithm bias.
- Close the human–machine teaming loop and bring domain knowledge into GeoAI system design.
- Design flexible building blocks in scalable processing chains and test optimizing solutions in different user cases.
- Develop initiatives to encourage open data/code sharing, mentoring programs, and increase community inclusion.

1. INTRODUCTION

We envision focusing on several key questions will accelerate the efforts needed to tackle the grand trillion pixel challenge:

- What are the gaps and limitations of GeoAI addressing end user application challenges in 2021?
- What do we expect the next generation of GeoAI to look like for solving end use grand challenges, engaging with end users across disciplines, developing certain data privacy standards, and declaring GeoAI assurance?
- What are the current promising directions to promote model generalization? How should we approach model transferability of multimodal GeoAI data?
- Are current large-scale processing architectures meeting today's demands? Are they well designed to manage the demands of tomorrow?
- What should a GeoAI Trust framework look like? Is it application dependent? What are the suitable requirements and components to such a framework? If data, model, and infrastructure are key, is there an integral approach to infuse trust in the responsible systems?
- What are the early successes and challenges of GeoAI for multimodal/crossmodal Earth observation (EO) analysis?

- What would incentivize community engagement and collaboration?

We summarize the discussions and presentations that addressed key questions in each theme, including the trends and challenges, and provide a forward-looking analysis.

2. TRILLION PIXELS GRAND CHALLENGES

Availability of sensors and data and and open-source GeoAI algorithms has led to several successful use cases, including feature extraction, imagery classification tasks, and broadly influenced environmental science, national security, and Earth systems science domains. Successful cases often build upon effective public–private partnerships, which continue to be the critical part to promote democratization of machine learning and artificial intelligence.

In addition to the current "static" map and successes from various domain applications, communities are moving to extract insights about activities and integrate multiple sensors. One of the most recent successful stories might be understanding urban activities or social processes research during the COVID-19 pandemic. The availability of diverse imagery sensors motivates better integration of spatial, spectral, and temporal data into GeoAI pipelines to enable them to automatically detect, characterize, and monitor processes at a global scale.

The panelists identified several challenges:

- Reusable and applicable pretrained models are not understood well, which is a curse of a million frameworks, and adapting these frameworks to individual communities poses similar problems.
- More consistency is needed in quality measures for a wider and clear understanding of AI's potential.
- When the AI is deployed on the user end, more questions are raised about how we can move beyond algorithm development, focus on human-machine teaming, and build GeoAI trust.

Managing computational and data resources remains challenging for certain applications. Each domain and application has its own data requirements for accuracy, and subsequently, the computing demands. Knowing the problem, the data and the audience can continuously help us move toward solving this problem. Some of the available datasets are still underutilized (e.g., the 30 years of Landsat imagery archive). Even though the resolution might not be sufficient for certain applications, the rich information embedded in this archive can benefit environmental studies.

3. GENERALIZATION AND TRANSFERABILITY

The call for better generalization and transferability for GeoAI models is motivated from the unprecedented speed and volume of data, variety of data sources, authoritative and nonauthoritative data sources, and mixed quality of data sources with varying uncertainties. Highly complex structures with diverse data sources and types make generalizable machine learning for geospatial problems a challenge, requiring endeavors from machine learning experts as well as remote sensing and domain experts.

While there has been significant progress in addressing this challenge, thanks to many open datasets, mainly optical imagery, and vigorous research efforts, the set of sensors represented in these open data repositories is relatively small compared to the approximately 274 unique EO sensors in orbit. Moreover,

diverse representation of objects (e.g., rotating objects, different sizes of objects, lighting conditions, weather conditions) also poses a challenge to achieving greater model generalization and transferability. Several promising trends are to address these emerging challenges, including machine learning/deep learning architectures that can take advantage of the spectral and spatial relationships inherent in remote sensing data, such as hybrid 2D/3D convolutional neural networks, meta-learning, the use of generative adversarial networks for image-to-image translations, transformer architectures, weak supervision, few shot learning and out-of-distribution detection. The community has been also seeking a fair and accessible mechanism to benchmark different models and compare results from disparate techniques. We can foresee the problem will be more difficult when we combine different sensors.

Global inferencing is also a huge challenge and will be a critical metric to assess the generalization and transferability. So far it is still an open problem and is fundamentally hard, given the cost of the data, training, and pipeline-building process. More importantly, we need a way to validate the results, which can be equivalent to the cost of the data. Even though current literature and research activities in GeoAI are more focused on smaller scale datasets and experiments, they are really important as building blocks to solve this problem. One important aspect we should also pay attention in the future to is the role of domain users and experts when building generalizable and transferable models. From a decision-making perspective, for example, quantifying uncertainty from GeoAI models can give domain users and experts confidence scores to help them make better decisions about whether to reuse results that have relatively low uncertainty in the source distribution. Domain experts also pose the knowledge to guide model developments and innovations for achieving some forms of generalizations, such as seasonal or temporal invariance. Developers of GeoAI systems should also work with domain experts to define the resource and performance requirements of the system early on to ensure the system can meet the constraints of the end users.

4. SCALABLE GEOSPATIAL PROCESSING ARCHITECTURES

The majority of "big data" analytical systems are designed to operate on lightweight data types that scale by quantity of record, not density of record. The complexities encountered with geospatial imagery do not inherently fit this model, and the architecture to feed any type of processing engine becomes challenging at large scales. Because of data heterogeneity (vector, raster, time-series, 3D data, spectral data) and data gravity (data is big and difficult to move, and big data attracts compute), these unique challenges in GeoAI domain require us to think of data at all parts of the storage and system hierarchy: How do we assemble data? How do we drive it through the systems? How can we efficiently index geospatial data? And, how can we create a software and middleware layer that allows us to take advantage of these nodes, local storage and memory?

Currently there is a fairly rich ecosystem of compute libraries for handling big geospatial data, which potentially accelerates optimization of hardware and software that is essential to a scalable geospatial processing architectures. Recent advances and promising solutions include context aware networking (NVIDIA BlueField), unified analytics-centric memory architectures (Apache Arrow), file formats that support random access reads over HTTP (Cloud Optimized GeoTIFFs), deployment of spatiotemporal indexing techniques (S2 Geometry, Uber's H3), and RasterFrames' spatial context bookkeeping. In terms of the end-to-end workflow, one of the important directions is to alleviate this challenge from step 0 (data acquisition) and have a better automatic data curation and integration for both vector and raster, direct data transferring from all the sources/sensors in hierarchical and distributed ways to supercomputing and cloud computing for machine learning. We also need to create persistent data cubes for geospatial data so we can

extract features, iterate, test our processes, and run intelligent queries at a reasonable speed. Even though in some scenarios the scale of datasets and the cost of computing makes the use of AI or GeoAI accessible to only those who have those necessary resources to run GeoAI models, the research for sparser and sample efficient models and faster feature generation is still needed to bring more flexibility and faster response time. It could be the case that these deep models are already over parameterized, so maybe there is a way to do this in a more efficient way. The processing architectures we are looking for are likely dependent on the use cases, which is not one-size-fits-all.

Edge computing is also emerging in GeoAI, given that sensor perspective is critical and probably the fastest growing part of GeoAI - the number of nano satellites is still in thousands, but the number of smartphone cameras is in the billions. The amount of diagnostic data collected is significant, but there is not enough bandwidth to send all the data back. Those are parts of the untapped geospatial data in GeoAI. Bringing computation to the edge and designing edge hardware in a balanced way all the way to the data center will be a revolutionary change. However, the protocols that sustain the geospatial processing workflows from the edge to extreme scales have yet to be defined.

5. ETHICAL AND TRUSTWORTHY GEOAI SYSTEMS

Ethics and trustworthiness are increasingly important when designing, deploying, and monitoring of GeoAI tools and are a fundamental and integral part of advancing breakthroughs for the benefit of society. Establishing trust in AI is a multifaceted problem, involving factors such as ethics, security, safety, system integration, open and clear testing/verification, and legal/policy issues.

To establish trust in GeoAI applications and with policy makers, providing a full picture of the information provided by GeoAI systems uncertainties is useful, as well as explaining the assumptions that are behind the application of certain models, such that users can know where and when to apply certain models. For testing and verification, reporting quality of data and labels, guidelines for evaluation metrics, and the standards can facilitate the reproduction and validation of results by other scientists. Creating benchmarks would certainly help, but it is much harder than in computer vision or machine learning because problems are much broader. Moving forward, the GeoAI community could start to set standards to classify use cases by the risk of potential harm, which would enable calibrated policy responses, create tests and checklists for self-auditing GeoAI applications for industry, and measure GeoAI performance.

Open data have been an essential part of GeoAI advancements and are also helpful for reproducibility, which is extremely important for the scientific community. However, an inherent tension exists between increasing transparency of data and models to build trust and data security. The process of building trustworthiness needs to have a human, psychological component, where putting transparency and better communication is crucial. Transparency foments trust! However, the open challenge is to maintain the balance between transparency and privacy when we exploit the data.

6. GEOAI BEYOND PIXELS

Building models of global GeoAI systems means capturing data and information about many specific regions in situ or through airborne and satellite-based remote sensing. For example, fusion of lidar with radar could be a major advancement for biomass estimation, so is leveraging Synthetic Aperture Radar (SAR), LiDAR (Light Detection and Ranging), and optical data for deforestation studies. In many other applications, various forms of data are also useful, including points, vectors, raster data, field monitoring logs, unstructured data on websites, and crowdsourced data (e.g., Twitter feeds).

Although experts know the value of domain-specific data, the data volume and heterogeneity are the main barrier to fully unlocking AI's potential in this space. It is imperative that we prioritize developing data standards (such as STAC (SpatioTemporal Asset Catalog), platforms, and algorithms to mitigate the friction of combining datasets and address the associated technical and engineering barriers. While navigating the best practices for integrating various forms of data, domain scientists are seeking interpretability and explainability from AI model predictions to gain scientific insights. One critical aspect that can be helpful to defy the "black-box" nature of AI algorithms is to be able to quantify uncertainty. From a scientific integrity perspective, this capability will allow domain scientists to know when and when not to combine datasets or reason about AI model selection. Further, uncertainties can provide meaningful interpretation when probabilities are given based on the domain-specific knowledge. Domain scientists need to be included in the development of GeoAI systems that ingest more than pixels, which mirrors the need to have human-in-the-loop when curating data and explaining GeoAI model outcomes.

Another challenge is that even though open science is an emerging theme at NASA, the European Space Agency (ESA), and other major sponsors of satellite-based research, a culture of competition can still discourage code sharing. Compared to many optical remote sensing datasets that are open-sourced, other types of data are still in a great demand to push the developments of GeoAI beyond pixels. We should move forward with an expectation that products that lack an open implementation are not as useful. Funding agencies should consider enacting policies that will require open data/algorithms to promote and incentivize an open science culture.

7. COLLABORATIONS AND COMMUNITY ENGAGEMENTS

On January 15, 2021, then President-Elect Biden sent a letter to Dr. Eric S. Lander, his appointee as the president's science advisor and nominee as director of the Office of Science and Technology Policy, tasking him to refresh and reinvigorate the United States' national science and technology strategy. The letter poses five questions. The first two questions are : what can we learn from the pandemic about what is possible, or what ought to be possible, to address the widest range of needs related to our public health? How can breakthroughs in science and technology create powerful new solutions to address climate change, propelling market-driven change, jump starting economic growth, improving health, and growing jobs, especially in communities that have been left behind? Active collaborations and community engagements will be key to GeoAI playing a role in addressing these and other societal grand challenges.

Enabling an AI model design to solve GeoAI grand challenges requires more interdisciplinary collaboration because of the scale of resources and complexities of policy-making needed to sustain the ecosystem. Three key steps include data transformation, which creates data interfaces based on user needs; knowledge integration, which combines physical, social, economic and other types of data; and decision support, which provides recommendations for national leaders' action. A search for compelling partnerships, providing visible and tangible benefits for all parties involved, is the key to facilitate such interdisciplinary collaboration, based on the experience of the ESA Θ-lab, which promotes particularly high-risk, high-reward research in EO across Europe. As one of the greatest examples of collaboration and partnership, the SpaceNet series of challenges/datasets involves CosmiQ works, Maxar, AWS, NGA, IEEE GRSS, Topcoder, Planet, Capella Space, and Intel AI, among others, and plays an important role to move us forward in solving the challenges discussed in different sessions, including benchmarking, data sharing, and computing resources.

Collaboration will also facilitate the infusion of local and global perspectives in solving the grand

challenges. Societal challenges such as sustainability, addressing climate change, and achieving carbon neutrality are often global in scale; however, the problems and their solutions vary spatially or locally. For example, collaboration with local partners is the easiest way to ensure that local views, local insights, and historical background are represented and considered, but the context of useful data for a GeoAI framework that is globally available or the decision- and policy-making should be considered at a different scale. Although the current leaders in the GeoAI community are creating and forming such collaborations with different initiatives, young talent is equally crucial to reinvigorate a more inclusive and tactful GeoAI community. Bringing in creativity and developing thought leadership in education can be the first step for them to raise their sense of duty in fostering collaborations and community engagements and maximize their relevance to society. For example, in the course "Space Technology for the Development Leader" offered by Professor Danielle Wood at MIT, students are encouraged to envision different kinds of collaborations in the area of EO communication and to consider what strategies could make those efforts meaningful. Aspects to be considered include the needs and viewpoints of national leaders, people at the city and regional scales, companies, and a varied group of organizations. Opportunities in mentorship and internship programs are also important to entice young talent from different areas to work in the intersection of computer vision, machine learning, and geospatial science. Those programs expose the next generation of researchers to the field and impacts of geospatial data and highlight the relevance of their diverse backgrounds and unique perspectives to grow the GeoAI community collectively.

However, we also observe that similar to other disciplines, the benefits of science and technology remain unevenly distributed across racial, gender, economic, and geographic lines in the GeoAI domain. Private sector firms, research institutions, and government agencies are undertaking a number of initiatives and recruitment priorities to focus on inclusion, to increase the representation of minorities, and to reach out to specific communities.