EDITORIAL

Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation

Danfeng Hong,[1,2,8] Chenyu Li,[1,3,8] Bing Zhang,[1,4,*] Naoto Yokoya,[5] Jon Atli Benediktsson,[6] and Jocelyn Chanussot[7]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Mathematics and Statistics, Southeast University, Nanjing 211189, China
[4]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[5]Graduate School of Frontier Sciences, The University of Tokyo, Chiba 277–8561, Japan
[6]Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik 102, Iceland
[7]University Grenoble Alpes, Grenoble 38000, France
[8]These authors contributed equally
*Correspondence: zb@radi.ac.cn

Received: January 30, 2024; Accepted: March 6, 2024; Published Online: March 14, 2024; https://doi.org/10.59717/j.xinn-geo.2024.100055

Citation: Hong D., Li C., Zhang B., et al., (2024). Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation. The Innovation Geoscience **2(1)**: 100055.

MULTIMODAL REMOTE SENSING BIG DATA

Earth observation (EO) techniques have undergone rapid development, facilitating comprehensive measurement and monitoring of the Earth's various facets, including land surface, subsurface, air, and water quality, as well as the well-being of humans, plants, and animals. Among these techniques, remote sensing (RS) emerges as a pivotal contact-free method for EO. RS enables the extraction of relevant information regarding the physical properties of Earth and its environmental systems from space. The abundance of diverse RS information introduces the concept of multimodality.[1] For simplicity, multimodal data refers to the description of the same object through various pieces of information or properties, such as image, text, sound, social media data, and video, which enhances our ability to gain a comprehensive understanding of the Earth[2] through the integration of multiple perspectives, including but not limited to agriculture, forestry, ecology, and the urban domains.

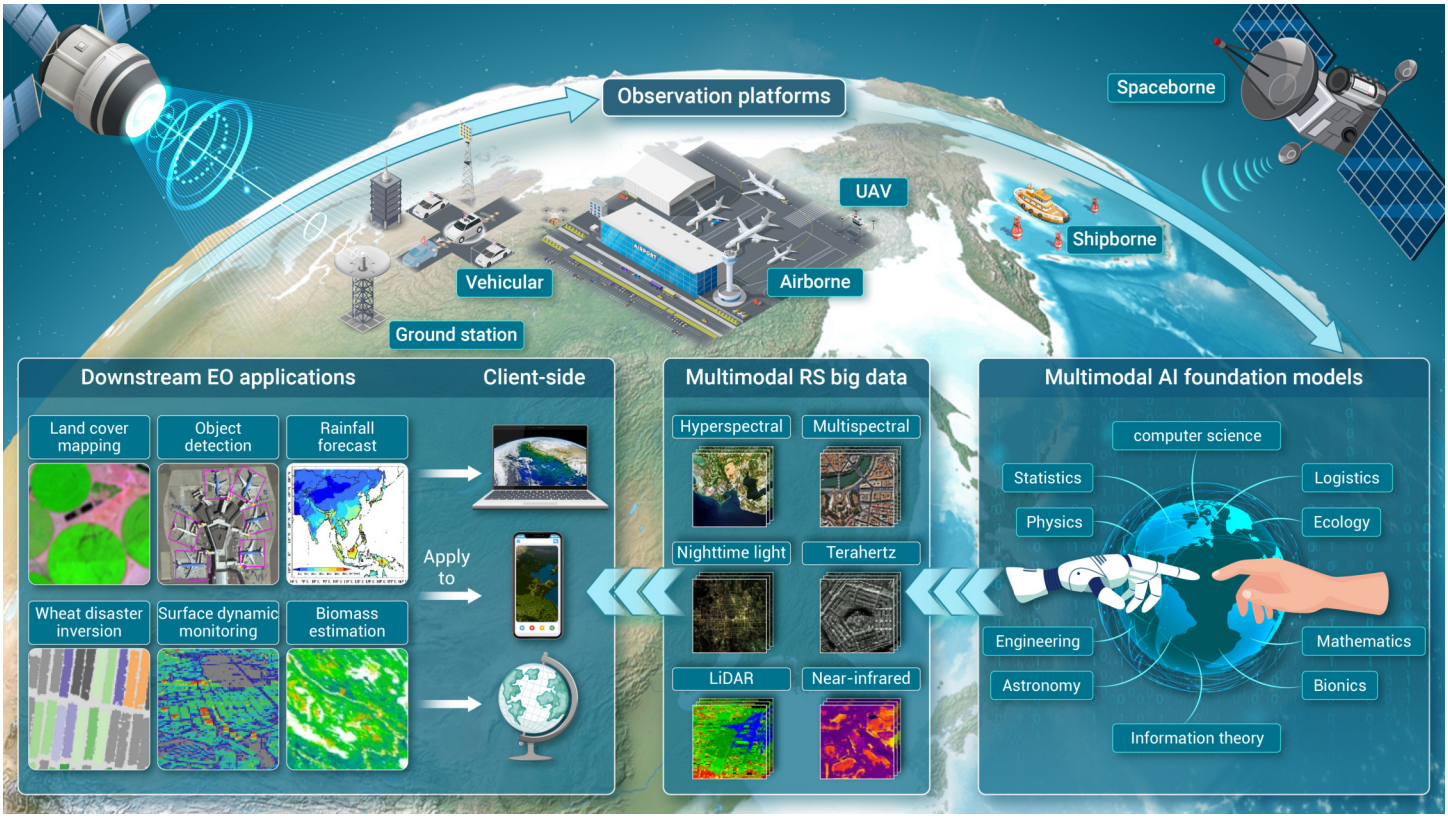

Figure 1. A cycle-chain RS intelligent interpretation system enabled by multimodal AI foundation models for RS big data in EO Start from different observation platforms, acquire the multimodal RS big data, train well-designed multimodal AI foundation models, act on downstream EO applications, apply to clients in practice, and finally feedback to the validation and design of payloads and platforms.

However, the escalating volume and diversity of RS data from various observation platforms, including spaceborne, airborne, and ground sources, underscores a pressing need to advance the multimodal processing and analysis capabilities of RS big data using artificial intelligence (AI) techniques.[3] This rapid expansion unavoidably introduces challenging difficulties, outlined as follows.

- Existing models significantly fall short in terms of their capacity for information extraction and analysis.
- Effectively harnessing and fully utilizing multimodal RS big data poses a significant bottleneck.
- There is a notable deficiency in deep information mining and homogenization of applications.

MULTIMODAL AI FOUNDATION MODELS

To overcome the difficulties mentioned above, researchers have dedicated their efforts to developing a high-precision RS intelligent interpretation system.[4] The system, as shown in Figure 1, encapsulates the cycle-chain process: utilizing the observation platforms, acquiring the multimodal RS big data, developing multimodal AI foundation models, applying for practical client applications, and finally feed-backing the validation and design of payloads and platforms. The establishment of this system hinges on several pivotal elements, i.e., the incorporation of massive multimodal RS big data, the utilization of high-performance computing power, and the integration of RS foundation models. In the current scenarios, satisfying the requirements for the first two elements is comparatively achievable to a great extent. However, a significant obstacle lies in the absence of customized multimodal AI foundation models that can effectively bridge the gap between RS big data and high-performance computing power. The foundation models are capable of thoroughly mining and extracting information from RS big data -- aiming to squeeze out every bit of information. This marks a transition into the era of foundation models, following the progression through statistical, physical, and big data models.

Very recently, there has been a significant upsurge in pretraining techniques centered around RS foundation models, especially in the context of utilizing spectral RS data. This surge has substantially broadened the capabilities of models across various EO-related applications. SpectralGPT[5] proposed by Hong et al. marks the first instance of a spectral RS foundation model specifically designed for spectral RS data. SpectralGPT undergoes training on an extensive dataset, encompassing over one million multimodal spectral RS images with variations in sizes, resolutions, time series, and regions. The model parameters of SpectralGPT exceed 600 million, making it currently the largest spectral foundation model in RS. Additionally, SpectralGPT has demonstrated significant potential in advancing multimodal RS big data applications within the field of geoscience, particularly across four downstream tasks: single-label scene classification, multi-label scene classification, semantic segmentation, and change detection.

REMAINING CHALLENGES AND FUTURE TRENDS

The rapid advancements in pretraining techniques, particularly those based on self-supervised learning, have spurred a growing focus on foundation models in the realms of computer vision and natural language processing. The tasks assigned to these pretraining agents are typically categorized into contrastive learning and generative learning. Despite the proliferation of numerous well-known foundation frameworks around the two mainstream pretraining techniques, their exploration in RS has been somewhat restrained. Particularly for multimodal RS data, a universal and consistent solution is yet to emerge in the development process of RS foundation models. Consequently, the development and promotion of multimodal AI foundation models in RS pose significant challenges, mainly stemming from the following three aspects.

Firstly, *diversity in the types of multimodal RS data*. RS data obtained from various platforms exhibit notable differences in image quality. For instance, unmanned aerial vehicle images demonstrate higher clarity and finer details compared to those acquired from satellites and airborne platforms. Heterogeneous RS data from different sensors, such as optical and Synthetic Aperture Radar, encompass distinct data types originating from diverse imaging mechanisms. Beyond gridded RS image data, point cloud data and social media data fall under structured data, necessitating particular attention to data processing and considerations for model inputs. Moreover, multi-band RS data, such as RGB, multispectral, and hyperspectral, exhibit variations in the number of spectral bands. Even within multispectral RS data, there is notable inconsistency in the band count across different sensors.

Secondly, *fragmentation in a variety of existing multimodal RS models*. Currently, numerous multimodal RS models have been proposed for different purposes or applications, such as information extraction, modality fusion, land cover classification, change detection, scene recognition, quantitative inversion, and environmental monitoring. While these models exhibit strong professionalism and are designed for customization for specific tasks, their inflexibility becomes apparent when confronted with new tasks. The installation and adaptation of models can be cumbersome and time-consuming in such scenarios. As the model library accumulates to a large extent, the task of selecting a suitable model also becomes increasingly challenging. Such a usage strategy of multimodal RS models does not align with the requirements of practical applications.

Thirdly, *insufficient, expensive, and time-consuming data annotation*. Despite the explosive growth of multimodal RS data in terms of types and quantities, the process of data annotation remains a challenge, posing difficulties in achieving high accuracy and efficiency in EO. Indeed, it is recognized that the speed of RS data annotation is considerably slower than that of data acquisition. On the other hand, addressing various EO tasks requires annotating different types of samples, amplifying the cost and complexity of sample annotation. Furthermore, the diversity of multimodal RS data further exacerbates the difficulties in the annotation process, particularly in expert visual interpretation.

Finally, *lack of modeling for specific attributes or knowledge of RS data*. Current multimodal AI foundation models primarily stem from the fields of computer vision and natural language processing. This tendency also results in a common phenomenon where nearly all visual foundation models are constructed based on RGB images. The corresponding negative effects primarily manifest in issues related to transferability and applicability. There is a noticeable gap between natural RGB images and spectral or SAR RS data. This gap inevitably results in performance degradation in EO tasks when using trained multimodal RS foundation models. The degradation occurs due to the lack of embedding specific knowledge (e.g., phenological characteristics, weather factors, geographic location) or information on RS data in the training process.

Considering the above remaining challenges, several possible future trends or solutions can be proposed accordingly, i.e., 1) establishing a unified multimodal RS foundation model to enhance universality for integrating RS data diversity and model fragmentation; 2) leveraging unsupervised or self-supervised learning strategies to reduce training costs in annotation; 3) embedding specific attributes and knowledge of RS data into models to extend beyond the focus on RGB data.

CONCLUSIONS

Multimodal AI foundation models represent the future of RS big data analysis, ready to unleash the potential inherent in multimodal RS data for diverse EO tasks. These models have the capability to harness the richness of multimodal RS big data, providing a robust framework for addressing the complexities of EO applications. By unifying different data types and modalities, these models enhance our overall understanding and analysis of the Earth's surface and environment. The shift towards multimodal foundation models signifies a promising advancement in optimizing the utilization of RS big data for a myriad of EO objectives, marking a transformative era in the field.

REFERENCES

1. Dalla Mura, M., Prasad, S., Pacifici, F., et al. (2015). Challenges and opportunities of multimodality and data fusion in remote sensing. Proceedings of the IEEE, **103**(9): 1585–1601. DOI: 10.1109/JPROC.2015.2462751.
2. Hong, D., Zhang, B., Li, H., et al. (2023). 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. DOI:10.1016/j.rse.2023.113856.
3. Xu, Y., Liu, X., Cao, X., et al. (2021). Artificial intelligence: A powerful paradigm for scientific research. The Innovation **2**(4): 100179. DOI: 10.1016/j.xinn.2021.100179.
4. Zhang, B., Wu, Y., Zhao, B., et al. (2022). Progress and challenges in intelligent remote sensing satellite systems. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **15**: 1814–1822. DOI: 10.1109/JSTARS.2022.3148139.
5. Hong, D., Zhang, B., Li, X., et al. (2024). SpectralGPT: Spectral remote sensing foundation model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024. DOI: 10.1109/TPAMI.2024.3362475.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2022YFB3903401), the National Natural Science Foundation of China (42241109, 42271350), and the MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

DECLARATION OF INTERESTS

The authors declare no competing interests.