Science Bulletin 69 (2024) 3652–3656

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

Perspective

Foundation model for generalist remote sensing intelligence: Potentials and prospects

Mi Zhang [a,c], Bingnan Yang [a], Xiangyun Hu [a,c], Jianya Gong [a,b,c,*], Zuxun Zhang [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
[c] Hubei Luojia Laboratory, Wuhan 430079, China

With the advent of Earth observation satellites, the remote sensing (RS) dataset has experienced exponential growth, significantly enhancing scientific research and applications. By early 2024, the global Earth observation constellation comprises 1,379 satellites, with projections indicating an increase to 5,500 by 2033. On a daily basis, these satellites produce more than 20 TB of raw data, leading to an accumulation exceeding 500 PB [1]. The surge in data volume poses challenges in storage, analysis, and management within the remote sensing domain. Foundation models like ChatGPT, SAM, and CLIP [2], present novel approaches that improve efficiency and drive innovation in remote sensing data processing. Leveraging extensive training datasets, these models demonstrate promise across a range of remote sensing tasks [3–5].

Foundation models feature extensive parameters, ranging from tens of millions to hundreds of billions. These models adopt large-scale Transformer [6] networks in a self-supervised manner, demonstrating proficiency in language comprehension, vision-language interaction, and multi-modal interpretation. The most popular foundation models for remote sensing are summarized in Table S1 (online). The typical architecture of these models can be standardized as follows: (1) Modal-specific encoding and tokenization; (2) alignment and fusion of multi-modal representations; (3) incorporation of additional Transformer layers to facilitate cross-modal connectivity and integration; (4) implementation of task-specific decoders for pretraining or downstream tasks (Fig. S1 online). FMs are generally pretrained in a self-supervised manner using extensive datasets and subsequently fine-tuned in a supervised manner on domain-specific datasets for downstream tasks. While many approaches involve adapting models from pretrained weights, their effectiveness is primarily constrained to tasks focused on images. This limitation stems from the absence of timely feedback based on prompts, indicating a deficiency in context-aware learning capabilities. Hence, the swift rise of knowledge encoding and human-in-the-loop schema has demonstrated substantial potential in advancing intelligent interpretation within remote sensing applications.

In the realm of generalist remote sensing intelligence, foundation models have the potential to provide valuable assistance in environmental monitoring, offering crucial insights into shifts in habitats, as well as contributing to urban development and planning efforts. Through cross-domain transfer learning and generalization, foundation models have instigated a paradigm shift within remote sensing communities [7], garnering significant interest in exploring the realm of comprehensive remote sensing intelligence.

To date, a limited number of studies have explored the application of FMs in RS-related fields like earth, climate, and environmental science [8,9]. As shown in Fig. 1, our focus lies on the utilization of foundation models in common remote sensing tasks including scene classification, object detection, land-use classification, change detection and video-related interpretation. Several studies [10] have delved into these tasks, emphasizing the varied applications of vision-language models. We delve deeper into the challenges and future initiatives aimed at advancing comprehensive remote sensing intelligence.

Specifically, we characterize comprehensive remote sensing intelligence as a versatile foundation model capable of seamlessly integrating multi-modal geospatial data to perform a wide range of static and dynamic remote sensing tasks. This foundation model contains enormous parameters, is trained on extensive datasets, facilitates multi-modal collaboration, particularly in the realm of vision-language tasks, enables iterative improvement through human feedback, and has the potential to evolve into a self-operating artificial intelligence (AI) agent.

Static remote sensing interpretation. Static interpretation of remote sensing data involves extensive aerial and satellite imagery that necessitates initial geometric preprocessing on the ground. This process is essential for producing static 4D products like Digital Orthophoto Maps (DOM), Digital Elevation Models (DEM), Digital Surface Models (DSM), and Digital Line Graphs (DL), culminating in static attribute interpretation. The interpretation process adheres to a coarse-to-fine approach, encompassing key tasks such as scene classification, object detection, land-use classification and change detection. Traditional approaches in interpretation mandate the development of deep learning models with limited parameters tailored to individual tasks, aiming to classify geospatial features accurately. By employing foundation models

* Corresponding author.
E-mail address: gongjy@whu.edu.cn (J. Gong).

https://doi.org/10.1016/j.scib.2024.09.017

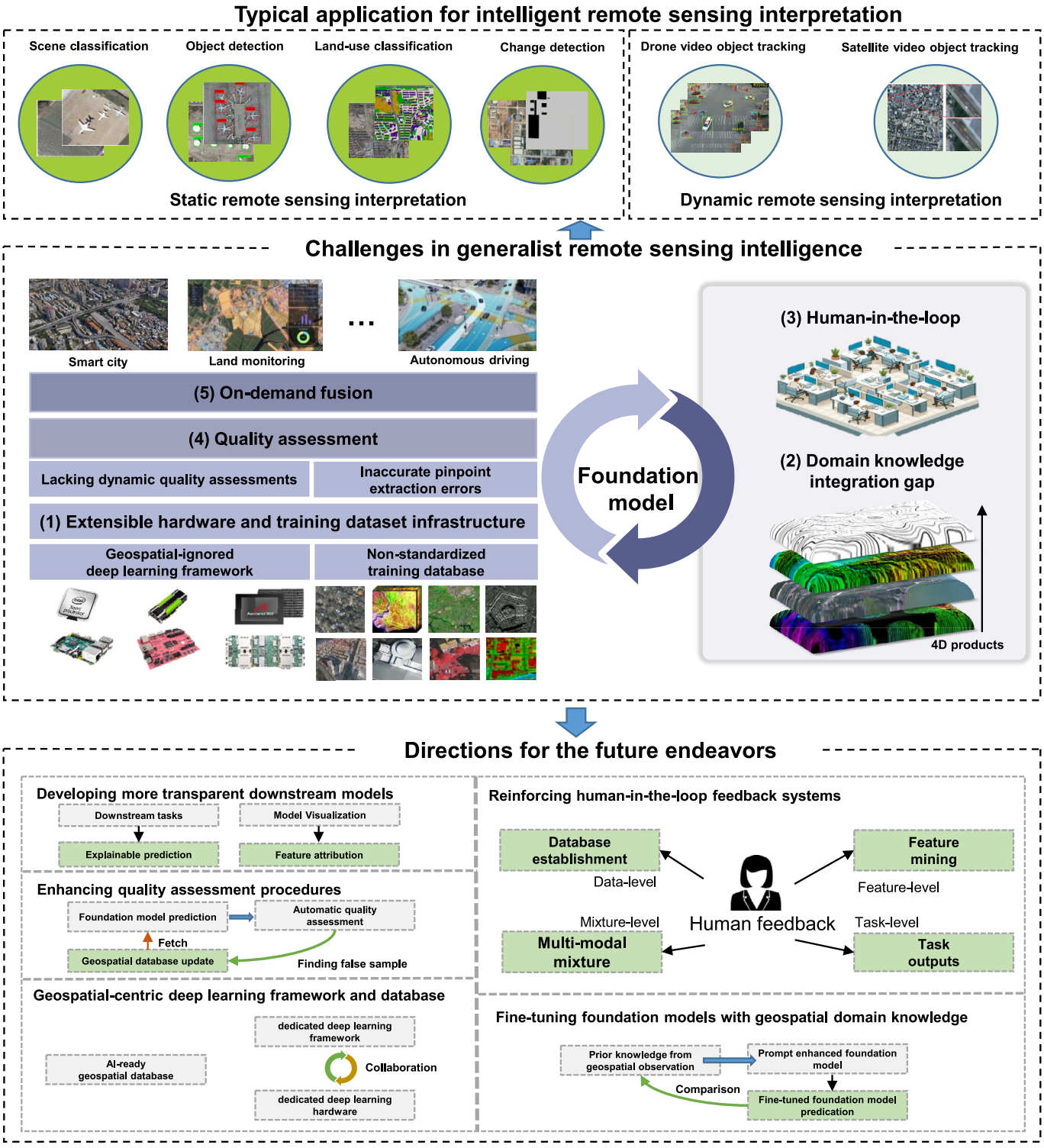

Fig. 1. Overview of potentials and prospects for large-scale remote sensing foundation model.

with an extensive parameter count derived from masked pretraining, and employing cross-modal alignment to integrate the geometric and attribute information of 4D products, it becomes feasible to attain a 'one model, multiple uses' objective. This approach helps streamline the network design complexity in downstream tasks [4,5]. Furthermore, foundation models offer cross-modal perception capabilities, effectively capturing the complex semantic categories and spatial contextual relationships of geospatial features through language models [4,8]. For example, leveraging large language model (LLM) instruction tuning can harmonize diverse RS visual tasks into a Visual Question Answer (VQA) format [4].

Dynamic remote sensing interpretation. Dynamic interpretation in remote sensing has evolved with the introduction of portable drones and numerous video satellites, enabling real-time monitoring of specific targets through continuous video frames, a crucial method in earth observation. Conventional deep learning methods with a small number of parameters, such as YOLO [11], SSD [12],

etc., necessitate specialized architectures to address challenges like target occlusions, varying illumination conditions, and high-dynamic scenes. The limited parameter capacity of these models also restricts their generalization capabilities. In such contexts, expansive foundation models, particularly those utilizing Transformer architectures, can exploit attention mechanisms and multi-task learning strategies to adapt efficiently with minimal labeled data. This approach enhances the model's ability to capture nuanced variations and intricate patterns within video frames, consequently improving the generalization capacity and accuracy in interpreting dynamic video data [5]. Unlike static scenario, dynamic videos are more adept at capturing movement information and behavioral patterns of targets, consequently enhancing the efficacy of object detection and tracking.

In these context, large foundation models play a rule for intelligent remote sensing interpretation in the following ways:

(1) Scene classification. Foundation models can be utilized to automatically classify intricate and varied scenes into predetermined categories, thereby improving the comprehension of extensive geographical regions. By harnessing the extensive learning abilities of these models, scene classification (Table S2 online) can be accomplished with increased accuracy and efficiency, even in demanding scenarios such as overlapping classes or ambiguous landscapes. For instance, the straightforward adaptation of two prominent vision-language models (VLM) CLIP [2] and Blip [13] can surpass single- modal model in the few-shot RS image scene classification task.

(2) Object Detection. Foundation models excel at accurately identifying and locating specific objects within aerial and satellite imagery. Certain foundation models, such as RingMo and SkySense, develop and train task-specific decoders for object detection, whereas others, like EarthGPT and GeoChat, perform these tasks through VQA with instruction-tuned LLM decoders. With a wealth of parameters at their disposal, these models are adept at handling the challenges associated with detecting small or partially obscured objects, leading to notable enhancements in object detection rates across diverse environments (Table S3 online)..

(3) Land-use classification. Foundation models could be utilized to distinguish between various types of land use, including urban areas, agricultural lands, forests, and water bodies, with enhanced accuracy. By conducting extensive multi-modal analysis, these models can interpret intricate patterns and variations in imagery, resulting in more precise and dynamic land-use mapping (Table S4 online). In remote sensing FMs, this task is typically implemented as semantic segmentation.

(4) Change detection. Foundation models could be utilized to track changes over time, identifying modifications (Table S5 online) in landscapes, urban development, or environmental degradation. Their ability to process and analyze temporal data allows for the detection of subtle changes that may go unnoticed by conventional approaches, offering valuable insights for urban planning, environmental monitoring, and disaster management. Most RS foundation models approach the change detection task by predicting change segmentation masks, while some utilize the vision-language model framework for change detection captioning.

(5) Video object tracking. Foundation models play a crucial role in tracking the movement of objects across a series of satellite or drone video frames, enhancing both accuracy and speed. These models effectively address the complexities of dynamic scenes, including swift movements, occlusions, and changing lighting conditions, making them well-suited for applications in surveillance, wildlife monitoring, and traffic management.

(6) Geoscience applications. Foundation models have also revolutionized various tasks within other geoscience disciplines. Their robust generalization, scalability, and multi-modal capabilities facilitate the precise capturing, simulation, and prediction of functions, interconnections, and changes in geospatial components within the earth system from static and dynamic viewpoints. Typical applications include climate and weather forecasting, smart architecture, geospatial localization, hydrology, etc.

Prior AI models used in remote sensing interpretation provided benefits such as high object detection precision, effective processing of extensive datasets, capability for multi-temporal analysis, and resilience in diverse environmental conditions. Nevertheless, these models encountered challenges related to language understanding, adaptability to unseen scenarios, capturing complex spatial relationships, and integrating diverse data sources effectively. Modern remote sensing foundation models utilize language modality to harmonize the design and execution of various tasks, fostering adaptability to unseen scenarios with an open-vocabulary approach. Additionally, they embrace a cross-modal paradigm to seamlessly integrate globally distributed multi-source geospatial data and elucidate their interconnections from the perspectives of time, space, and spectral views. The shift from earlier AI models to expansive foundation models in remote sensing interpretation signifies a substantial advancement in the capabilities of comprehensive analysis and comprehension. Nonetheless, the current utilization of foundation models in remote sensing, particularly in dynamic environmental monitoring and multi-modal data fusion, continues to encounter several challenges.

(1) Extensible hardware and training dataset infrastructure. The training of most foundation models heavily depends on NVIDIA GPUs, the PyTorch framework, and non-standardized remote sensing datasets, creating dependencies and constraints in the training process. The significant reliance on particular hardware and software may lead to the monopolization of the AI infrastructure market by selecting manufacturers. Regarding data infrastructures, the majority of current remote sensing deep learning datasets are tailored for specific tasks, resulting in insufficient volume size and storage formats for the requirements of remote sensing foundation models. This underscores the need to establish infrastructure for effectively converting the rapidly expanding volume of unprocessed remote sensing data into a state ready for AI utilization, especially for downstream applications [14].

(2) Domain knowledge integration gap. Current models primarily rely on the visual features extracted from remote sensing imagery and may struggle to provide accurate assessments of terrestrial targets. For instance, differentiating a river that dries up in autumn from bare land, roads, and other land features based solely on image characteristics can be challenging. This underscores the challenge that current remote sensing image interpretation models encounter in distinguishing between closely related or visually ambiguous land categories. These models often overlook the seasonal variations in land features and do not fully leverage available geographic information and expert physical knowledge, resulting in challenges in accurately identifying the categories.

(3) Human-in-the-loop. Current remote sensing interpretation models operate unidirectionally and disregard human feedback, potentially causing discrepancies between interpretation outcomes and user expectations. By neglecting human input, these models face challenges in tailoring to specific requirements and preferences, constraining their accuracy and adaptability in intricate scenarios. The absence of iterative learning hinders the optimization and improvement of foundation models for specific interpretation tasks.

(4) Quality assessment. In addressing the intricate requirements of 'scene- target-pixel' multi-level remote sensing image interpretation tasks, current models exhibit a deficiency in dynamic quality assessments and precise error localization. The reliability is commonly evaluated through manual selection of accuracy metrics and their empirical combination. Alternative post-processing

strategies statically rectify errors and enhance low-quality elements but do not dynamically expand the initial image interpretation samples or update model parameters. As a result, this limitation undermines the reliability of the models' predictive outputs.

(5) On-demand fusion. Current models frequently focus on a single image source or are customized for particular tasks, thereby restricting their adaptability and overlooking the extensive possibilities offered by multi-modal data. This oversight disregards the intricate needs of downstream tasks, ultimately constraining the models' utility and their capacity for generalization. Hence, to enhance adaptability and accuracy, it is crucial to employ flexible fusion strategies for integrating diverse data modalities-such as images, text, and sound—in alignment with task specifications. This approach not only refines model performance but also enhances its reliability in complex application settings.

These limitations present opportunities for progress and steer future initiatives. Leveraging advanced AI algorithms and technologies, the utilization of large-scale foundation models in comprehensive remote sensing intelligence shows potential for enhancing hardware infrastructure, enhancing specialized frameworks, and refining application models. The key initiatives involve constructing an extensible geospatial database and a resilient deep learning framework, bridging the domain knowledge integration gap, strengthening human-in-the-loop feedback systems, enhancing quality assessment procedures, and bolstering the models' reliability and adaptability across diverse contexts.

(1) Elevating database, framework and hardware infrastructures for geospatial-centric foundation models. The training of current foundation models is restricted by the limited coverage of categories and sensor diversity in image samples, impacting their ability to generalize across extensive spatiotemporal domains. Solutions entail the automatic identification and expansion of new categories, auto-annotation of geospatial samples, and enhancing sample precision. An actionable solution is to establish an automatic or semi-automatic annotation workflow, similar to SAM, to effectively generate diverse labels for extensive raw data. Furthermore, data from diverse sources must be meticulously collected, accurately matched, and systematically organized into an AI-ready state that empowers scientists to access, collaborate on, and analyze multi-source data as needed [15]. Moreover, creating domain-specific deep learning frameworks that encompass the unique 'time–space-spectrum-angle' attributes of remote sensing could serve as a viable solution [3]. Additionally, the close collaboration between the deep learning framework and computing hardware warrants further exploration.

(2) Fine-tuning foundation models with geospatial domain knowledge. Leveraging geospatial domain-specific knowledge during foundation model training can greatly improve prediction accuracy and relevance. Integrating expert insights and contextual information regarding geographic features, seasonal variations, and environmental factors assist models in gaining a deeper understanding of earth observation data. Practically, potential methods include creating large-scale datasets with detailed expert knowledge or using LLMs to fuse multi-modal representations. Besides, integrating knowledge graphs through restructuring into linguistic prompts or extracting cross-modality knowledge correlations is a viable yet underexplored solution.

(3) Reinforcing human-in-the-loop feedback systems. Incorporating human feedback into the foundation model learning loop facilitates ongoing refinement and adaptation. Expert review, correction, and annotation of model outputs contribute to performance enhancement over time and establish user confidence in automated interpretations. Human feedback can be utilized at the data, feature, mixture and task levels to rectify database errors, facilitate feature extraction, guide multi-modal interactions, and enhance task accuracy outputs, respectively (Fig. S1 online). A prevalent approach for implementation involves utilizing reinforcement learning to modify the behavior of foundation models.

(4) Enhancing quality assessment procedures. Creating advanced metrics and evaluation frameworks to precisely evaluate the quality and reliability of model outputs is essential. Thorough quality control guarantees that foundation model interpretations adhere to stringent standards required for critical applications such as environmental monitoring and disaster management. Possible solutions involve implementing mathematical uncertainty quantification, developing a judgmental LLM, and establishing an LLM agent-based automated assessment framework.

(5) Developing more transparent downstream models. Transparency in AI models is essential for their acceptance and utility insensitive remote sensing downstream tasks. Developing models that provide interpretable predictions and decisions allow users to comprehend the reasoning behind interpretations, enhancing trust in the technology. This can be accomplished through methods such as model visualization, feature attribution, and providing clear explanations of model behavior.

The development of a large-scale foundation model for generalist remote sensing intelligence, incorporating extensible infrastructure, enhancing existing open-source models, and integrating geospatial knowledge into the model, requires sustained endeavors. Upon creating a model at the scale of hundreds of billions, ongoing adaptation will be necessary to address evolving downstream needs. Presently, there is a deficiency in comprehensive evaluation frameworks for foundation models. The existing standards for quality and performance assessment have limitations in their relevance to remote sensing, underscoring the need for additional research and development to establish robust, scalable solutions.

In conclusion, large-scale foundation models have the potential to transform remote sensing by enhancing interpretation accuracy across various applications. Nevertheless, challenges such as needing extensible hardware and datasets, integrating geospatial knowledge, ensuring robust human-in-the-loop feedback, and developing comprehensive quality assessment frameworks must be addressed. Continuous innovation and refinement are crucial for unlocking the complete potential of foundation models to address the dynamic and complex requirements of intelligent remote sensing interpretation.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the Key Research and Development Program of Hubei Province (2023BAB173), the State Key Laboratory of Geo-Information Engineering (SKLGIE2021-M-3-1), the National Natural Science Foundation of China (41901265), Major Program of the National Natural Science Foundation of China (92038301), and supported in part by the Special Fund of Hubei Luojia Laboratory (220100028). We extend our sincere gratitude to Ph.D student Yuanxin Zhao for his diligent efforts in meticulously refining the figures.

Appendix A. Supplementary materials

Supplementary materials to this perspective can be found online at https://doi.org/10.1016/j.scib.2024.09.017.

References

[1] Li J. Advances in high-resolution earth observation satellite remote sensing technologies in China. Sci Technol Foresight 2022;1:112–25.
[2] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Int Conf Mach Learn, PMLR, 2021, pp. 8748–8763.
[3] Zhang Z, Zhang M, Gong J, et al. Luojiaai: A cloud-based artificial intelligence platform for remote sensing image interpretation. Geo-spat Inf Sci 2023;26:218–41.
[4] Zhang W, Cai M, Zhang T, et al. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. IEEE Trans Geosci Remote Sens 2024;62:5917820.
[5] Guo X, Lao J, Dang B, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: IEEE Conf Comput Vis Pattern Recognit, 2024, pp. 27672–27683.
[6] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30:5998–6008.
[7] Hong D, Li C, Zhang B, et al. Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation. Innovation Geosci 2024;2:100055.
[8] Mai G, Huang W, Sun J, et al. On the opportunities and challenges of foundation models for Geoai (vision paper). ACM Trans Spat Algor Syst 2024;10:1–46.
[9] Ma Y, Chen S, Ermon S, et al. Transfer learning in environmental remote sensing. Remote Sens Environ 2024;301:113924.
[10] Li X, Wen C, Hu Y, et al. Vision-language models in remote sensing: Current progress and future trends. IEEE Geosci Remote Sens Mag 2024;2:32–66.
[11] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection. IEEE Conf Comput Vis Pattern Recognit 2016:779–88.
[12] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector. In: Eur Conf Comput Vis, Springer, 2016, pp. 21–37.
[13] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Int Conf Mach Learn, PMLR, 2022, pp. 12888–12900.
[14] Scheffler M, Aeschlimann M, Albrecht M, et al. Fair data enabling new horizons for materials research. Nature 2022;604:635–42.
[15] Li X, Feng M, Ran Y, et al. Big data in earth system science and progress towards a digital twin. Nat Rev Earth Environ 2023;4:319–32.

Mi Zhang is an associate researcher at the School of Remote Sensing and Information Engineering, Wuhan University. He serves as the chief artificial intelligence scientist in Handleray Corporation and technical director in WHU-LuoJiaAI Group. His research interest mainly includes computer vision, machine learning, with particular interest in semantic object segmentation and the construction of deep learning framework.

Jianya Gong received the Ph.D. degree from the Wuhan Technical University of Surveying and Mapping, China, in 1992. He is currently an Academician with the Chinese Academy of Sciences and a Professor and the Dean of the School of Remote Sensing and Information Engineering, Wuhan University. He is also the President of Commission VI of the International Society for Photogrammetry and Remote Sensing. His current research interest includes remote sensing image processing, spatial data infrastructure, geospatial data interoperability, and artificial intelligence.