Edge AI for Earth Observation

Ning Ruan and Kun Li, *Henan Normal University, Xinxiang, 453007, China*
Qiyang Zhang, *Peking University, Beijing, 100871, China*
Lauri Lovén, *University of Oulu, 90014, Oulu, Finland*
Praveen Kumar Donta, *Stockholm University, 16425, Stockholm, Sweden*
Yi Jia, *National Natural Science Foundation of China, Beijing, 100006, China*
Schahram Dustdar, *TU Wien, 1040 Wien, Austria, and UPF ICREA, 08018, Barcelona, Spain*

Earth observation (EO), edge computing, and artificial intelligence (AI) are rapidly advancing technologies with diverse applications and benefits. Integrating edge computing and AI with EO enables the preprocessing and analysis of EO data near its source, supporting efficient decision-making and in-orbit information interpretation. In this context, this article provides a review of the current state of edge AI in EO applications, summarizes the key challenges, including data sample limitations, computing resource constraints, catastrophic forgetting, and difficulties with satellite–ground coordination. Also, we explore possible solutions and techniques such as including generalization under small sample conditions, lightweight model design and training (e.g., pruning, quantization, distillation), continuous learning for multiple tasks, and satellite–ground continuum systems (e.g., federated learning and resource-constrained inference). Finally, we outline possible future research directions to address the challenges using edge AI for EO scenarios.

Earth Observation (EO) aims to monitor Earth's surface through various remote sensing (RS) technologies, improving the ability to better understand RS data. EO research has been successfully applied across diverse domains. Today, an increasing number of satellites equipped with various sensors are launched to perform EO tasks. These advancements generate massive volumes of high-resolution, large-scale imagery every day, enabling a wide range of applications. However, effectively processing such vast data is challenging.

On another side, artificial intelligence (AI) has proven highly effective in supporting onboard applications, such as intelligent interpretation and real-time dynamic tracking. Deep learning (DL) techniques have been successfully employed to reduce transmission bandwidth, processing time, and resource consumption in EO applications, as shown in Table 1, including disaster monitoring, precision agriculture, and so on. For example, the integration of lightweight models on Huawei Ascend 310 processors reduced flood mapping latency from 3 h to only 8 min while maintaining accuracy. There is a growing interest in developing faster, more accurate, and more compact intelligent algorithms to meet evolving needs. One of the key strengths of DL lies in its ability to train models with robust feature self-learning and generalization capabilities. Numerous DL models for data interpretation have been proposed and widely adopted. However, their high complexity and reliance on large-scale data present significant challenges, especially considering the massive data volume and strict timeliness requirements inherent to EO analysis. Also, DL-based applications typically depend on the substantial computing power of hardware systems. Deploying DL models on resource-constrained satellites to achieve intelligent

Digital Object Identifier 10.1109/MIC.2025.3587325
Date of publication 10 July 2025; date of current version 27 August 2025.

TABLE 1. Representative use-cases of edge AI in Earth observation.

Use Case	Description	Edge AI Model Used	Benefits
Disaster monitoring	Real-time wildfire or flood detection from satellite imagery	Lightweight CNN, YOLOv5	Low-latency alerts, local decision-making
Precision agriculture	Crop health monitoring and yield estimation using multispectral data	MobileNet, Vision Transformer	Reduced data transfer, on-site analysis
Urban change detection	Tracking construction and land use changes over time	Siamese CNN, UNet	Efficient spatiotemporal analysis
Maritime surveillance	Detecting illegal ships or oil spills from satellite data	Object detection models	Near real-time monitoring over vast ocean areas
Environmental monitoring	Air/water quality estimation using satellite data + ground IoT integration	Hybrid models (MLP + SensorFusion).	Decentralized sensing and processing

operations remains a major challenge for advancing intelligent technology.

Therefore, addressing the dynamic nature of tasks and meeting stringent task requirements are critical. Developing AI techniques for EO scenarios can greatly facilitate efficient information acquisition and decision-making. Nevertheless, several key challenges remain:

- *Limited sample availability:* Unlike visual images, EO data often suffer from a scarcity of effective samples. Satellite scenes typically encompass complex backgrounds and diverse target categories; however, the number of labeled and usable samples remains limited. This issue is particularly severe for newly appeared or altered targets, leading to a significant class imbalance in data distribution.
- *Computing resource constraints*: EO data frequently cover vast areas, ranging from several kilometers to hundreds of kilometers, with individual datasets reaching sizes of several gigabytes. This creates substantial storage demands for satellite systems. Additionally, modern DL models, characterized by increasing complexity, demand substantial computational resources. Therefore, addressing the tradeoff between model complexity and power consumption is critical.
- *Catastrophic forgetting*: In EO data analysis, new tasks and categories continually emerge. Existing algorithms frequently demonstrate limited generalization capability when processing sequentially introduced data, leading to catastrophic forgetting of previously learned knowledge. Addressing this issue requires the development of continuous learning methods to maintain high accuracy and performance over time, remains an urgent and critical challenge.
- *Satellite–ground bottleneck*: The challenge lies in bridging the computational and communication gaps between satellite-based and ground-based systems. Satellites typically operate under constraints such as limited onboard computing resources, energy consumption, and bandwidth, making real-time data processing and transmission to ground stations difficult. In contrast, ground systems often face latency issues when receiving and processing data from space platforms, particularly when handling large-scale EO data.
- *Reliability challenges in spaceborne AI*: Despite AI's potential in satellite systems, its reliability remains critically challenged by radiation-induced faults, hardware degradation, and algorithmic vulnerabilities. Issues that are further aggravated by extreme environmental conditions and stringent operational demands.

To address these challenges, we explore how Edge AI can benefit EO. For instance, Edge AI has been shown to enhance the quality of low-light satellite images.[2] This article examines various advantages of incorporating Edge AI into EO, as summarized in Figure 1. The figure outlines the challenges, current AI research in EO, and typical applications, aiming to advance AI-driven EO development. From top to bottom, it illustrates the challenges, corresponding solutions, and potential applications. We provide a detailed analysis of the challenges and solutions, concluding with future research directions.

GENERALIZATION IN SMALL SAMPLE SCENARIOS

The performance improvement achieved by DL algorithms is based on large amounts of training data. However, DL algorithms often face the challenge of

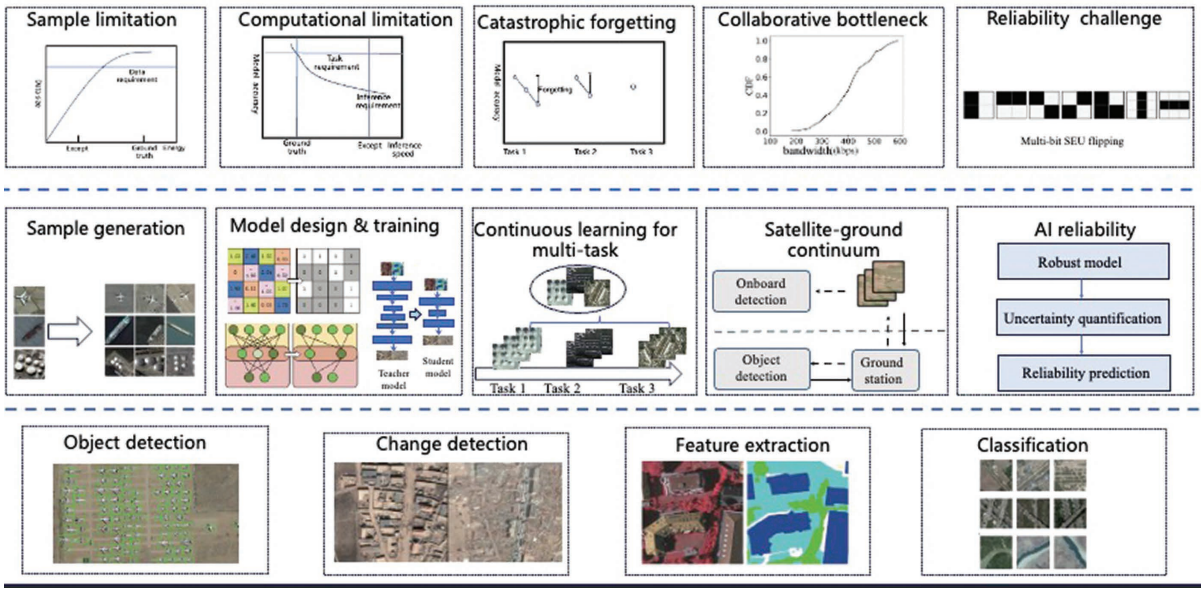

FIGURE 1. The architecture for EO satellites comprises three main components: challenges (top), solutions (middle), and supporting applications (bottom).

small sample sizes. Existing solutions can be categorized into two classes. The first exploits the inherent characteristics of images to address data volume requirements. The second involves knowledge reuse, which incorporates prior knowledge through strategies. These methods support the learning of new object categories, thus reducing the dependence on large data volumes and enabling intelligent interpretation. Table 2 summarizes the comparison of sample generalization methods.

Sample Generalization

Sample generation focuses on the automatic generation of EO data and utilizes the generated dataset to meet the training requirements. Traditional methods typically involve transformations such as translation, rotation, and filtering. However, these methods often rely on manual design and expert selection. This dependence limits their ability to generate new semantic information about ground objects, resulting in insufficient diversity in generated data. To address these limitations, two advanced approaches have gained attention: simulation-based and DL-based sample generation.

Simulation-Based Generation

This approach focuses on simulating and constructing specific ground objects for EO tasks, by leveraging simulation modeling platforms to generate synthetic data. Simulation-based generation eliminates the acquisition costs with traditional EO image data and allows flexible adjustment of ground object and imaging parameters (e.g., illumination, height, field of view) to meet specific task requirements.

Simulation-based sample generation primarily addresses two key data interpretation tasks: synthetic aperture radar (SAR) images and visible light images. SAR-based data simulation framework integrates computer-aided design modeling with simulation imaging. Specifically, this simulation incorporates noise and terrain information into the imaging process, effectively generating realistic simulated data. For visible light scenarios, where EO images often feature high resolution and complex background, a more practical approach to reducing modeling costs is to simulate detection targets at the instance level. This instance-based modeling strategy efficiently captures target-specific characteristics while maintaining scalability.

DL-Based Generation

DL-based generation leverages deep neural networks to approximate the distribution characteristics of EO data. This approach involves modeling and learning of the joint statistical distribution of EO data, enabling the trained model to generate new samples at either the feature level or image level. Most DL-based generation algorithms utilize generative adversarial networks.

TABLE 2. Comparative Study of different methods via sample generalization, lightweight model design, and multitask continuous learning methods.

Principle	Method	Advantages	Disadvantages
Comparison of different sample generalization methods			
Sample generalization	Traditional method	High efficiency	Select the generation method in combination with expert experience; limited generation of sample information abundance
	Simulation-based	Customize the image content	Professional knowledge is needed.
	DL-based	Automated generation	Model training is unstable.
Knowledge reuse	Transfer learning	High precision	Requirement of dataset is strict.
	Metric learning	High universality	Dual-source input has strict requirements for computing and storage.
	Meta learning	High speed and universality	Generalization performance relies on an auxiliary dataset rich in category information.
Comparison of lightweight model design and optimization methods			
Model design	Lightweight model	Balance model size and accuracy effectively	Based on human priors or massive machine search
Model optimization	Model pruning	Enable hardware-agnostic and fine-grained reduction of model space complexity	Due to the complexity of training, aggressive pruning may impair the model's feature representation capacity.
	Model quantization	Offer broad model applicability and enable substantial compression	Rely on specific underlying hardware support
Comparison of multitask continuous learning methods			
Sample reproduction	Storage of old samples	Balance training data	Increasing the sample storage space and limiting the application scenarios
	Pseudo-sample generation	Liberate data from constraints	Training is complicated, and the generative model tends to converge poorly.
Model structure expansion	Optimization constraint	Simple to implement	Task relevance is required, and the network capacity is easily saturated.
	Parameter isolation	Task adaptation	Additional parameter storage

Knowledge Reuse

Knowledge reuse involves the automatic extraction and modeling of relevant knowledge from an existing data domain which is then generalized to a target data domain. This approach aims to enhance the model's generalization accuracy in the target domain, even with limited sample availability. Based on the relationship between the two data domains, this method is categorized into three classes: transfer learning, metric learning, and meta learning.

Transfer Learning

Transfer learning exploits the similarities between data, tasks, and models to transfer knowledge learned from a source domain to a target domain. This approach aims to reduce the data dependency of models in the target domain. Based on the employed reuse strategy, transfer learning is classified into model reuse-based and feature mapping-based transfer learning. Model reuse-based method focuses on reusing pretrained models from the source domain. By selectively leveraging the structure and parameters of these models, it seeks to optimal solutions within the hypothesis space while reducing the requirement for training. Fine-tuning is the most typical method used in this approach. Furthermore, this method focuses on designing an optimal mapping space between source and target domains. The objective is to ensure that the features of each ground object type in the two domains exhibit high similarity within the mapping space while maintaining strong discriminability for learning tasks.

Metric Learning

This method relies on distance metrics to define similarity and dissimilarity relationships between objects. Metric learning utilizes convolutional neural networks to learn mapping functions that ensure high intra-class similarity and low inter-class similarity within the mapped feature space. During testing, the similarity between test data and labeled samples is evaluated in the mapped feature space, generating classification confidence scores for various ground objects. The primary advantage of this method lies in the strong generalization capability. This method adapts to interpretation tasks involving diverse ground objects without requiring retraining, making them highly versatile.

Meta Learning

Meta learning enhances model learning efficiency through experience accumulation by training models across multiple few-shot tasks and focusing on the acquisition of transferable knowledge. This approach endows models with human-like analogical reasoning capabilities, enabling rapid adaptation to novel land-cover categories with limited training samples. By leveraging training across a wide range of categories, meta learning produces a meta model with strong generalization capabilities, enabling quick adaptation to the small sample interpretation needs. However, the training depends datasets often lack annotated ground object categories, limiting the generalization performance. Additionally, the computational efficiency remains low, further hindering the practical applications.

LIGHTWEIGHT MODEL DESIGN AND OPTIMIZATION

EO images exhibit unique characteristics, including extensive spatial coverage, complex target elements, and diverse image modalities. These attributes require more advanced DL models to effectively capture image features. For instance, target elements occupy larger pixel area. This requires processing larger images to encompass these targets and employing deeper models to extract relevant classification information. Furthermore, EO image processing typically involves multimodal data, including SAR images, hyperspectral images, and digital surface models. Managing large-scale and complex scenarios typically requires the integration of multimodal data. This approach inherently increases the number of model parameters due to multichannel inputs and often involves coordinating multiple models. This section focuses primarily on lightweight model design and training.

Unlike satellite flight controllers, which are considered critical components, commercial off-the-shelf hardware typically offers lower reliability and computational capacity compared to ground-based systems. For example, according to BUPT-1 satellite, a Raspberry Pi provides approximately 7 W of computational power, whereas the Atlas chip achieves around 13 W. Moreover, depending on the solar panel area and sunlight availability, a CubeSat can harvest between 1 W and 150 W of energy at most. Consequently, the onboard AI models must be carefully designed to minimize the number of model parameters and maximize operational speed, while maintaining high performance as much as possible. Model design methods can generally be categorized into two approaches: model pruning and model quantization. Table 2 summarizes the comparison of lightweight model design and optimization methods.

Model Pruning

Model pruning addresses onboard resource constraints by reducing spatial complexity and enabling deployment on low power devices. Usually, as the pruning granularity increases, the degree of model lightweighting also increases. Large-granularity pruning algorithm is particularly well-suited for deploying models on satellite-borne systems. For instance, Zhu et al.[3] proposed a two-stage target detection method utilizing convolution kernel pruning. The method employs the sum of the absolute values of convolution kernel weights as the pruning criterion, ranking them in ascending order before conducting pruning. By streamlining the network structure, this approach enhances inference speed.

Model Quantization

Model quantization is a lightweight model design technique that compresses neural networks by reducing the bit width used to represent weights. Model quantization assumes that such high precision is unnecessary and instead replaces 32-b FLOPs with low-precision alternatives. Model quantization is typically classified into binary quantization, ternary quantization, and multivalue quantization, based on the number of bits used to store weights after quantization: 1) Binary quantization reduces 32-b FLOPs weights to two values (e.g., 0/1 or –1/1), compressing the model size to 1/32 of the original. 2) Ternary quantization introduces a third value (–1, 0, 1), minimizing performance loss without increasing computational complexity. 3) Multivalue quantization represents weights with higher precision than binary or ternary

quantization (e.g., 8-bit or 16-bit quantization), achieving performance closer to that of the original network. Model quantization reduces model size, memory usage, and hardware power consumption by lowering the model precision. However, quantization often leads to accuracy degradation. Consequently, current research focuses on achieving substantial model compression while maintaining accuracy.

EO applications primarily enhance the training process by leveraging the characteristics of multiscale targets and multimodal data. For EO images with multiscale features, feature learning is achieved by fusing feature maps from different stages. The output of each feature is typically supervised by an auxiliary objective function, which does not increase the computational cost during the inference but significantly improves performance. For EO images with multiple modalities, the training process for each modality is typically constrained by incorporating a regularization penalty term, which balances the speed of feature learning and prevents both under-fitting and over-fitting.

Knowledge distillation involves two network models: the student and the teacher network. The student network, typically a lightweight model, is designed to have a smaller size and limited computational capacity. In contrast, the teacher network is a pretrained, high-precision model. The goal of knowledge distillation is to improve the performance of the student network by transferring knowledge from the teacher network. During the training of student network, in addition to using labeled training data as supervision (as in traditional methods), the extracted knowledge serves as an auxiliary supervisory signal to guide the training process.

Knowledge distillation must account for the large-scale variations and irregular shapes of target objects. Effective distillation architectures address these challenges by extracting multiscale features from different stages of the teacher network and capturing shape-related features through its input representations. Additionally, self-distillation techniques have emerged as promising approaches, where the student network leverages its internal layers for guidance. Shallow layers can learn contextual attention information from deeper layers, while deeper layers can capture semantic attention information from shallower layers. These techniques are particularly beneficial for improving student networks in EO applications.

CONTINUOUS LEARNING FOR MULTIPLE TASKS

Current deep neural network models are typically static, which are designed for specific tasks and cannot expand over time. When new data becomes available, these models cannot update without compromising its performance on the original task, leading to catastrophic forgetting. For instance, in detection and recognition tasks, it is common to train separate models for different targets, or even create distinct models for various subtypes. This approach is not only complex but also inefficient. As EO images can be updated daily, static models struggle to incorporate new data timely, thereby limiting their adaptability. To illustrate the challenge of catastrophic forgetting, we refer to empirical results reported by Rebuffi et al.,[4] where sequential training on the CIFAR-100 dataset caused the model's accuracy on earlier tasks to drop from 70% to below 30% without any continual learning strategies. This demonstrates the severity of performance degradation in continuous learning settings, which similarly impacts real-world satellite learning scenarios.

Multitask continuous learning addresses the challenge from an infinite stream of data, aiming at extending acquired knowledge to future tasks. The samples from different stages of the data stream correspond to distinct tasks. During the training stage, the dataset includes only samples from the current task, with previous task samples being unavailable. During the inference stage, the model must maintain high prediction accuracy for both the previous and current tasks. Existing methods are primarily categorized into two approaches: sample reproduction and model structure expansion. Table 2 summarizes the comparison of multitask continual learning methods.

Sample Reproduction

Sample reproduction mitigates catastrophic forgetting by storing samples from previous tasks or generating pseudosamples using generative models. During the training of new tasks, these stored or generated samples are replayed to balance the training data across tasks, thereby alleviating the forgetting issue.

Extensive research has been conducted on sample reproduction methods to address catastrophic forgetting. Rebuffi et al.[5] proposed to select samples that best approximate the class-wise mean in the learned feature space, which are then stored in a memory pool for subsequent training. To prevent the storage pool from overflowing as the number of tasks grows, samples from previous tasks are reselected based on the same criteria after completing the current task's training. However, this method results in prediction bias, due to the significant imbalance between new and old task samples, where the model is more likely to predict inputs as belonging to new tasks. To mitigate this, Li et al.6 adopted knowledge distillation by using the previous model's output on current task samples as a soft

label for the prior tasks, thereby reducing forgetting and improving knowledge transfer. However, maintaining a memory pool of representative samples significantly increases memory overhead.

Model Structure Expansion

To address the challenge of unavailable historical data from prior tasks, a model structure extension approach has been proposed. This method involves constraining parameter update strategies or isolating model parameters, effectively partitioning the model into subsets dedicated to specific tasks. This method enhances task adaptability and mitigates catastrophic forgetting without relying on historical data.

The earliest approaches relied on the distribution of parameters from previous models as prior knowledge. However, due to the vast number of model parameters, these approaches were highly complex. To address this, Rusu et al.[7] proposed progressive neural networks, which utilize lateral connections to reuse resources. The new parameters are added for learning the current task while preserving the weights of parameters associated with previous tasks. Within a fixed network architecture, parameter isolation can be achieved by identifying parameters used for previous tasks and masking them during training for the current task. Currently, model structure expansion remains in the early exploratory stages for EO applications.

SATELLITE–GROUND CONTINUUM

Figure 2 illustrates the satellite–ground continuum system supporting inference and federated learning.

Satellite–Ground Inference

Recently, orbital edge computing architecture has been proposed to overcome the limitations of the "bent-pipe" architecture, where EO data must be transmitted to the ground for processing. The core concept is to distribute computing tasks across satellites in low-Earth orbit (LEO) constellation, effectively reconstructing the computing pipeline. Kodan[8] focused on computations performed on a single satellite. By discarding irrelevant satellite images early, Kodan reduces the data volume transmitted to Earth. Serval[9] concentrated on the management of computing tasks between satellites and ground stations using fixed DL models. By leveraging the predictability of satellite orbits, Serval separates computing tasks between satellites and ground stations to minimize the end-to-end request latency. Earth+[10] leveraged images across an entire satellite constellation to enhance imagery compression, enabling more images, especially those

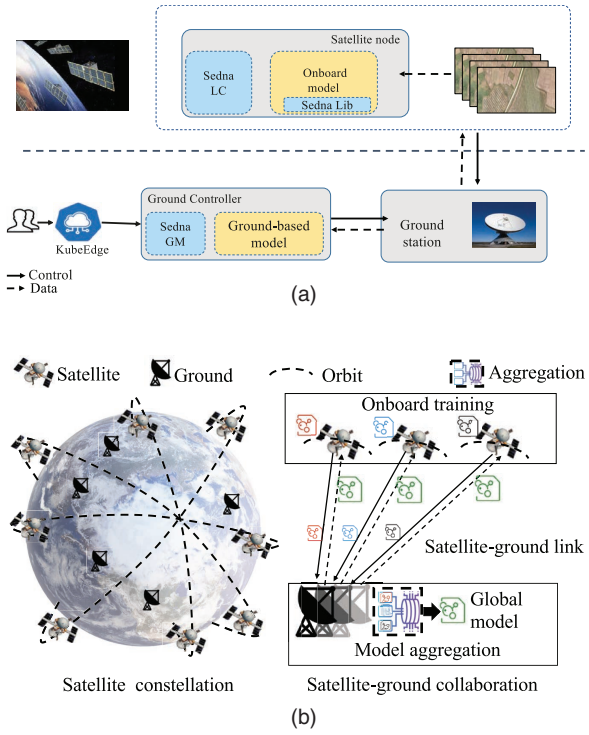

FIGURE 2. Satellite–ground continuum. (a) Inference. (b) Satellite–ground federated learning.

of the same area, to be downloaded to the ground efficiently. However, existing studies often neglect the reliability of key system variables such as computational capabilities and energy constraints. These resource limitations are key contributors to computational bottlenecks in EO systems. Furthermore, the stringent size and weight restrictions of satellites impose additional limitations on processing speed and operational efficiency. To address these dual challenges of energy and computational constraints, TargetFuse[11] proposed an end-to-end satellite–ground collaborative inference framework. This approach not only enhances inference accuracy but also explicitly accounts for the inherent limitations of satellite-based systems.

Federated Learning

Within a distributed learning architecture, a satellite–ground federated learning system is proposed, where LEO satellites perform local model training while collectively contributing to a global model through periodic parameter aggregation. In this architecture, LEO satellites function as edge devices that communicate with ground stations acting as parameter servers. This approach enables interconnection across large-scale constellations, reduces the high communication costs

of satellite networks, and effectively protects satellite data privacy.

Previous works have primarily focused on addressing challenges arising from satellite heterogeneity, particularly the latency and staleness issues inherent in synchronous or asynchronous federated learning algorithms. For instance, So et al.[12] proposed an adaptive aggregation buffer to maintain parameter freshness, improving system accuracy. To mitigate inefficiencies in collaborative training and slow model convergence, Zheng et al.[13] introduced a substructure scheme that enables heterogeneous local model training, considering different computing, memory, and communication constraints of LEO satellites. Shi et al.[14] developed a framework tailored for LEO mega-constellation networks, significantly reducing the reliance on low data-rate and intermittent satellite–ground links. Moreover, neural quantization has been explored to reduce communication costs by eliminating redundant gradient information during training. Yang et al.[15] proposed a precision-aware federated quantization training algorithm that supports in-orbit satellite training under dynamic satellite–ground connections.

AI RELIABILITY

AI has demonstrated great potential in resource-constrained and dynamic satellite environments. Nevertheless, the reliability of AI models deployed on satellite presents significant challenges. Factors such as transient faults caused by radiation or electromagnetic interference, hardware degradation due to harsh environmental conditions, limited computational resources, and the inherent sensitivity of DL models to minor perturbations can severely affect system performance,[16, 17] These issues are further exacerbated by the massive data volumes and stringent timeliness requirements typical of EO applications. Therefore, ensuring the reliability of AI systems in orbit is essential in achieving consistent and trustworthy operations.

To address these challenges, several mitigation strategies have been explored. Redundancy techniques, such as Triple Modular Redundancy,[18] allow critical operations to be duplicated to enhance fault tolerance. Developing fault-tolerant models through robust training with adversarial or corrupted data, as well as employing ensemble methods, improves resilience against errors.[19] Furthermore, lightweight error detection techniques, including checksums, hashing, and anomaly detection modules, are favored due to their minimal computational overhead, making them suitable for onboard deployment.[20] These approaches aim to enhance the robustness of AI systems operating in constrained and hostile environments.

PROSPECTS AND CHALLENGES

Edge AI has advanced significantly in the past decade, including in EO applications; however, it still faces challenges in real-world scenarios.

Small Sample Learning

Current small sample learning research typically assumes limited, fully labeled data supplemented by ample unlabeled samples. Many methods, such as those based on feature mapping or deep generative transfer learning, often rely on such unlabeled data. Yet in practice, especially for new target categories in EO applications, gathering even unlabeled data can be challenging. Furthermore, data acquisition and labeling remain nontrivial. Considering the complexities of satellite data collection and the need for rapid decision-making in edge AI for RS tasks, investigating single-sample or even zero-sample approaches offers a promising research avenue.

Lightweight Multitask Models

Current approaches to designing and training lightweight EO models typically focus on one specialized task or a single functional component. However, as the volume of EO data increases and the demand for diverse applications grows, multitask processing for remote sensing image interpretation has emerged as a critical research frontier. Consequently, there is an urgent need to develop lightweight models that extend from single to multiple functionalities and from single to multiple tasks.

Introducing Domain Expert Knowledge

The diverse imaging modes in EO, combined with imbalanced foregrounds and backgrounds and densely distributed ground targets, pose major challenges for multitask continuous learning, often leading to catastrophic forgetting. Incorporating domain expert knowledge helps the model differentiate critical parameters more effectively. By leveraging such expertise, the importance of each task can be prioritized, enhancing the penalty for altering key parameters. This approach both constrains and guides the continuous learning process, ultimately mitigating or even preventing catastrophic forgetting.

In-Orbit Fusion of Multisource Data

The integration of complementary information from multisource data, including visible light, hyperspectral, and SAR sensor data, has emerged as a key trend in extracting more detailed and accurate information

about terrestrial targets and objects. In EO applications, different sensor types provide EO data with distinct characteristics, offering varying strengths and limitations in terms of resolution, spectral information, and temporal coverage. Research in multisource information fusion and multimodal DL algorithms focuses on leveraging these diverse data sources to enhance information extraction. By combining advanced fusion algorithms with DL techniques, it is possible to achieve more precise data interpretation and intelligent decision-making for EO tasks.

Enhancing AI Reliability

While existing mitigation strategies offer promising solutions, further research is needed to develop adaptive reliability techniques that dynamically respond to changing environmental and operational conditions. Co-designing AI models with fault-aware hardware architectures represents a promising direction for achieving system-wide resilience. Moreover, incorporating self-healing mechanisms and online learning capabilities could enable AI systems to autonomously detect, adapt to, and recover from faults during deployment. Continued advancements in these areas will be crucial for the future deployment of robust, reliable, and intelligent systems.

CONCLUSION

Integration of EO applications with edge AI is essential for advancing technology and expanding its application scenarios. This article reviews the theoretical foundations and recent advancements in intelligent EO, focusing on technical frameworks and research outcomes. These efforts aim to address key challenges in the field. Moreover, the development and integration of these technologies will play a critical role in the future EO research.

REFERENCES

1. D. Tuia et al., "Artificial intelligence to advance earth observation: A review of models, recent trends, and pathways forward," *IEEE Geosci. Remote Sens. Mag.*, early access, Sep. 9, 2024, doi: 10.1109/MGRS.2024.3425961.
2. T.-A. Bui, P.-J. Lee, C.-S. Liang, P.-H. Hsu, S.-H. Shiu, and C.-K. Tsai, "Edge-computing-enabled deep learning approach for low-light satellite image enhancement," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4071–4083, 2024, doi: 10.1109/JSTARS.2024.3357093.
3. J. Zhu, Y. Zhao, and J. Pei, "Progressive kernel pruning based on the information mapping sparse index for CNN compression," *IEEE Access*, vol. 9, pp. 10,974–10,987, 2021, doi: 10.1109/ACCESS.2021.3051504.
4. R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan, "Measuring catastrophic forgetting in neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 3390–3398, doi: 10.1609/aaai.v32i1.11651.
5. S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5533–5542, doi: 10.1109/CVPR.2017.587.
6. Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018, doi: 10.1109/TPAMI.2017.2773081.
7. A. A. Rusu et al., "Progressive neural networks," 2016, *arXiv:1606.04671*.
8. B. Denby, K. Chintalapudi, R. Chandra, B. Lucia, and S. Noghabi, "Kodan: Addressing the computational bottleneck in space," in *Proc. 28th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2023, vol. 3, pp. 392–403.
9. B. Tao, O. Chabra, I. Janveja, I. Gupta, and D. Vasisht, "Known knowns and unknowns: Near-realtime earth observation via query bifurcation in serval," in *Proc. 21st USENIX Symp. Netw. Syst. Des. Implementation*, 2024, pp. 809–824.
10. K. Du, Y. Cheng, P. Olsen, S. Noghabi, and J. Jiang, "Earth+: On-board satellite imagery compression leveraging historical earth observations," in *Proc. 30th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2025, vol. 1, pp. 361–376.
11. Q. Zhang et al., "Resource-efficient in-orbit detection of earth objects," in *Proc. IEEE Conf. Comput. Commun.*, 2024, pp. 551–560, doi: 10.1109/INFOCOM52122.2024.10621328.
12. J. So, K. Hsieh, B. Arzani, S. Noghabi, S. Avestimehr, and R. Chandra, "FedSpace: An efficient federated learning framework at satellites and ground stations," 2022, *arXiv:2202.01267*.
13. Z. Lin, Z. Chen, Z. Fang, X. Chen, X. Wang, and Y. Gao, "FedSN: A federated learning framework over heterogeneous LEO satellite networks," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 1293-1307, Mar. 2025, doi: 10.1109/TMC.2024.3481275.
14. Y. Shi, L. Zeng, J. Zhu, Y. Zhou, C. Jiang, and K. B. Letaief, "Satellite federated edge learning: Architecture design and convergence analysis," *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, pp. 15,212–15,229, Oct. 2024, doi: 10.1109/TWC.2024.3427377.
15. C. Yang et al., "Communication-efficient satellite-ground federated learning through progressive weight quantization," *IEEE Trans. Mobile Comput.*, vol. 23, no. 9, pp. 8999–9011, Sep. 2024, doi: 10.1109/TMC.2024.3358804.

16. R. C. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 3, pp. 305–316, Sep. 2005, doi: 10.1109/TDMR.2005.853449.
17. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.
18. M. Nicolaidis, "Design for soft error mitigation," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 3, pp. 405–418, Sep. 2005, doi: 10.1109/TDMR.2005.855790.
19. X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2805–2824, Sep. 2019, doi: 10.1109/TNNLS.2018.2886017.
20. Z. Wang, J. Tian, H. Fang, L. Chen, and J. Qin, "LightLog: A lightweight temporal convolutional network for log anomaly detection on the edge," *Comput. Net.*, vol. 203, Feb. 2022, Art. no. 108616, doi: 10.1016/j.comnet.2021.108616.

NING RUAN is a lecturer at the School of Computer and Information Engineering, Henan Normal University, Xinxiang, 453007, China. His research interests include deep learning and data mining. Contact him at 2017030@htu.edu.cn.

KUN LI is a undergraduate student at the School of Computer and Information Engineering, Henan Normal University, Xinxiang, 453007, China. His research interests include satellite computing and deep learning. Contact him at 2046271174@qq.com.

QIYANG ZHANG is a postdoctoral researcher at the School of Computer Science, Peking University, Beijing, 100871, China. His research interests include training/inference in distributed continuum systems. He is the corresponding author of this article. He is a Member of IEEE. Contact him at qiyangzhang@pku.edu.cn.

LAURI LOVÉN is a postdoctoral researcher and the coordinator of the distributed intelligence strategic research area in the 6G Flagship research program, University of Oulu, 90014, Oulu, Finland. His research interests include learning in distributed continuum systems. He is a Senior Member of IEEE. Contact him at lauri.loven@oulu.fi.

PRAVEEN KUMAR DONTA is an associate professor at the Department of Computer and Systems Sciences, Stockholm University, 10691, Stockholm, Sweden. His research interests include learning in distributed continuum systems. He is a Senior Member of IEEE. Contact him at praveen@dsv.su.se.

YI JIA is a research associate at the High Tech Research and Development Center, National Natural Science Foundation of China, Beijing, 100006, China. His research areas include science and technology management and edge intelligence. Contact him at jiayi@nsfc.gov.cn.

SCHAHRAM DUSTDAR is a full professor of computer science and head of the Research Division of Distributed Systems at TU Wien, 1040 Wien, Austria, and UPF ICREA, 08018, Barcelona, Spain. His research interests include the investigation of all aspects related to edge/fog/cloud computing. He is a Fellow of IEEE. Contact him at dustdar@dsg.tuwien.ac.at.