Overview of Foundation Models in Remote Sensing

1st Teo Stojkovski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
teo.stojkovski@students.finki.ukim.mk

2nd Evrosina Stojkoska
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
evrosina.stojkoska@students.finki.ukim.mk

3rd Marko Petrov
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
marko.petrov@finki.ukim.mk

4th Ema Pandilova
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ema.pandilova@finki.ukim.mk

5th Vlatko Spasev
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
vlatko.spasev@finki.ukim.mk

6th Ivan Kitanovski
Faculty of Computer Science and Engineering
University Ss Cyril and Methodius
Skopje, N. Macedonia
ivan.kitanovski@finki.ukim.mk

***Abstract*—Recent advances in artificial intelligence, particularly the rise of foundation models, are reshaping the landscape of remote sensing. Traditionally reliant on handcrafted features and task-specific models, remote sensing has faced challenges in scalability, adaptability, and generalization across diverse geospatial data sources. Foundation models—large-scale, pre-trained architectures developed using self-supervised learning—offer a transformative solution by enabling cross-task generalization, improved performance in object detection and segmentation, and reduced reliance on annotated datasets. This paper provides a comprehensive review of foundation models in the context of remote sensing, highlighting their architectures, training strategies, and applications across environmental monitoring, agriculture, disaster response, and urban planning. We also address the technical and practical challenges of integrating these models into heterogeneous geospatial workflows, including variations in data resolution, sensor modalities, and real-time processing demands.**

***Index Terms*—Remote Sensing Images, Object Detection, Deep Learning, Earth Observation**

I. Introduction

The rapid advancements in artificial intelligence and deep learning have significantly transformed the field of remote sensing, enabling the extraction of meaningful insights from vast amounts of satellite and aerial imagery [1]. As an essential technology for Earth observation, remote sensing plays a crucial role in various domains, including environmental monitoring, agriculture, disaster management, and urban planning. However, traditional remote sensing methods often rely on handcrafted features and task-specific models, which can be limited in scalability, generalization, and adaptability to diverse geospatial data sources.

The emergence of foundation models [2] has introduced a paradigm shift in remote sensing by leveraging large-scale pre-trained architectures capable of generalizing across multiple downstream tasks. These models, trained on extensive datasets using self-supervised learning techniques, offer significant improvements in object detection, image segmentation, and change detection by effectively capturing spatial and temporal patterns in geospatial data. Unlike conventional deep learning approaches that require extensive labeled datasets for each specific application, foundation models enable knowledge transfer [3] and zero-shot [4] learning, reducing the dependency on annotated data while enhancing model robustness.

Despite these advantages, integrating foundation models into remote sensing presents unique challenges. Remote sensing data varies widely in resolution, color bands, and sensor types, making it challenging to process using a single approach. To address this, model architectures must be designed to handle different types of inputs effectively, ensuring accurate analysis across diverse geospatial datasets [5]. Additionally, computational efficiency and the need for domain-specific adaptations remain key concerns, particularly in high-resolution satellite imagery applications where real-time processing is crucial.

This paper aims to offer a thorough overview of foundation models in remote sensing, exploring their architecture, training methodologies, and performance across various geospatial tasks. We examine their impact on traditional remote sensing workflows, assess the latest advancements in the field, and discuss existing challenges and potential future developments.

The paper is organized as follows: In the second section we provide a background in the topic of remote sensing, the data types present, the basic concepts of processing the data and challenges. The third section provides an overview of the foundation models in the context of remote sensing, explaining their details and characteristics. In the fourth section, we com-

pare and discuss the different models and their capabilities. The concluding remarks are provided in the final section.

II. Background

A. Remote sensing data modalities

Remote sensing was limited and time-consuming in the past, mainly due to its reliance on data acquired through aerial photography and ground-based analysis. With the refinement of sensor technology and satellite imagery, modern remote sensing has revolutionized how we view and understand the surface of the Earth. Today, vast amounts of data are being collected through an array of advanced sensors, enabling precise monitoring of environmental and man-made changes.

In order to capture the features of the Earth's surface, remote sensing uses a variety of data modalities (Figure 1), each having unique imaging methods, spectral ranges, and spatial resolution. These data types are essential for applications such as climate monitoring, disaster response, agriculture, and urban planning. In this section, we briefly analyze the following types of data modalities:

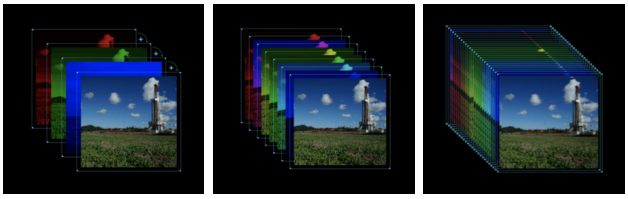
Optical RGB (**R**ed, **G**reen, and **B**lue) imagery [6] plays a fundamental role in remote sensing by capturing data in the visible spectrum and providing full-color images that closely resemble what the human eye sees. It is a useful tool for numerous applications, such as urban planning, disaster response, and object detection, due to its accessibility and easily interpreted output. RGB imagery is frequently obtained from multiple platforms, such as unmanned aerial vehicles (UAVs) and drones for regional observations, planes for large-scale object detection, and satellites for global remote sensing. It usually uses an 8-bit per channel system, where each channel records intensity values that define an image's color composition and brightness. These values range from 0 (black) to 255 (white). However, despite its technical simplicity, RGB imagery has limitations, particularly its sensitivity to lighting conditions and its restricted spectral range compared to multispectral and hyperspectral imaging.

Multispectral remote sensing [7] acquires data across the visible and infrared spectrum, with varying spatial and spectral resolutions. In comparison with hyperspectral imagery, multispectral images typically occur with lower spectral, but higher spatial resolution, typically ranging from 5 to 12 bands [8]. This makes multispectral data popular for long-term environmental research. These data are obtained through various platforms, including satellites such as Landsat, Sentinel-2, and MODIS.

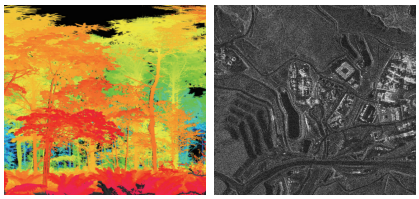
Hyperspectral remote sensing [8] captures data in a profoundly ranged specter, offering images with high spectral resolution. Compared to multispectral imagery, which provides data in a limited number of broad bands, hyperspectral data has much higher band number, ranging from hundreds to thousands. This spectral resolution allows the detection of information that is invisible to multispectral sensors, making hyperspectral data particularly valuable for complex data analysis in fields such as agriculture, forestry and environmental monitoring.

LiDAR (**L**ight **D**etection and **R**anging) technology [9] uses laser pulses for measuring distances, by capturing their reflection from objects in the environment. In remote sensing, this technology enables high-resolution 3D mapping and is widely used for capturing terrain data, water structures, man-made objects, and detecting deformations in structural engineering. LiDAR systems are deployed on terrestrial platforms, unmanned aerial vehicles (UAVs), aircraft, and satellites, supporting various applications such as urban planning, forestry, and object detection.

SAR (**S**ynthetic **A**perture **R**adar) [10] is a widely used data modality in remote sensing due to its ability to operate in all weather conditions and at any time of day. Unlike optical sensors, which rely on sunlight, SAR emits microwave signals and measures their returns, allowing it to penetrate certain surfaces. This makes SAR particularly valuable for applications such as environmental monitoring, disaster assessment, land cover classification, and infrastructure mapping.



(a) RGB (b) Multispectral (c) Hyperspectral



(d) LiDAR (e) SAR

Fig. 1. Remote sensing data modalities: (a) RGB, (b) Multispectral, (c) Hyperspectral, (d) LiDAR [11], (e) SAR. Image sourced from: https://www.critchlow.co.nz/resources/blog/next-generation-satellite-services-high-definition-imagery-daily-change-detection-multi-spectral-derived-data, https://www.cg.informatik.uni-siegen.de/de/exploration-synthetic-aperture-radar-sar-images

B. Remote sensing processing concepts

Data gathered through remote sensing must be processed to extract meaningful insights. Traditionally, this process relied on manual interpretation and task-specific models designed for specific applications. While traditional methods are straightforward, they often require extensive labeled datasets and significant data processing capabilities [12].

With the rise of deep learning, remote sensing has undergone a revolution, enabling the automation of data analysis with improved accuracy and efficiency.

Vision Transformers (**ViTs**) have proven to be a high-performance architectural framework for remote sensing image analysis and classification. ViTs process images by dividing them into small patches, allowing the model to focus on patterns and details within each section, leading to classification

accuracies of up to 98.49%. This architectural design makes ViTs highly effective at classifying complex remote sensing imagery [13].

C. Challenges in remote sensing

Due to the complexity of data in remote sensing, several challenges arise during image processing. Image classification relies on labeled samples, typically obtained through ground surveys or image photointerpretation. However, ground surveys offer accuracy with high cost, while photointerpretation is more affordable but has low precision. Variations in spectral signatures caused by environmental factors, seasonal changes, and sensor differences, complicate image segmentation. Although deep neural networks (DNNs) have enhanced segmentation, they are limited by the low volume of large labeled datasets. Object detection faces reduced detection accuracy due to the variation of object sizes, shapes, and spectral features, as well as the interference of environmental noise, clouds, and shadows. Traditional algorithms such as Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) often require voluminous labeled datasets, which challenges their ability to detect small or obscured objects [14].

III. Foundation models in Remote Sensing

A. What are Foundation models and why do we need them?

Foundation models (FMs) are large-scale, pre-trained models designed to serve as a foundation for various tasks across different field [15]. These models are trained on vast datasets using advanced architectures, allowing them to learn complex patterns and features. This pre-training makes them adaptable to specific applications with minimal additional training. Techniques like self-supervised learning (SSL) [16] and transformer-based [17] architectures have significantly improved their effectiveness in tasks such as image classification, object detection, and change detection.

One of the key advantages of FMs is their ability to learn from largely unlabeled data using SSL. This is especially beneficial in remote sensing, where acquiring labeled datasets is often challenging [18].

Several notable foundation models have been developed specifically for remote sensing, such as SatMAE which is a transformer-based model designed for temporal and multispectral satellite imagery [19].

Despite their advancements, FMs still face challenges, including the need for high-quality and diverse training data, significant computational resources, and effective adaptation to specific remote sensing applications [20]. Overcoming these limitations will be essential for the continued progress of foundation models in remote sensing.

B. Types of Models

Early models relied on convolutional neural networks (CNNs), such as ResNet [21], which improved image recognition and classification [22]. Vision transformers (ViTs) adapted this approach for image data by treating image patches as sequences of tokens, allowing models to recognize both local and global relationships [23]. This is particularly beneficial for tasks like semantic segmentation and change detection, which require understanding fine details in high-resolution satellite images.

Most foundation models [24] cannot perform zero-shot inference due to the lack of joint vision-language modeling. To address these challenges, researchers developed a vision-language foundation model for remote sensing, named RemoteCLIP. The goal was to learn strong visual features with rich semantics from satellite imagery while aligning these features with text embeddings [25].

Many of the current methods are inspired by self-supervised learning (SSL) techniques in computer vision, especially Masked Image Modeling (MIM) [26] [27] [28]. Recent models like SatMAE [19], Scale-MAE [29], ViTAE [30], Billion-scale MAE [31], RingMo [32], and GFM [33] have successfully leveraged MIM with large Vision Transformers (ViTs) trained on vast satellite imagery datasets, yielding promising outcomes.

C. Foundation Models in Remote Sensing

Remote sensing (RS) is the process of gathering information about objects or areas from a distance, usually through satellite or airborne sensors [34].

Large Foundation Models (FMs) have demonstrated remarkable generalization and zero-shot learning abilities in multiple domains [35]. These models learn generalized representations of Earth's surface features and can be fine-tuned for a wide range of downstream tasks, such as land cover classification, object detection, change detection, and environmental monitoring. By leveraging transfer learning, foundation models reduce the need for extensive labeled data in domain-specific applications and significantly accelerate development cycles. Their ability to generalize across sensors, regions, and modalities (e.g., optical, SAR, hyperspectral) makes them especially powerful for global-scale analysis and applications in agriculture, disaster response, climate monitoring, and more.

Several notable foundation models have been developed specifically for remote sensing [25], such as SatMAE, Scale-MAE, ViTAE, Billion-scale MAE, RingMo, GFM, RemoteCLIP, SimCLR, MAE, SAM, and BERT. These models are often trained on large-scale, multi-sensor datasets, enabling them to extract rich, transferable representations from diverse imagery sources. Some, like RemoteCLIP and SimCLR, focus on contrastive learning approaches, while others, like MAE and SatMAE, use masked autoencoding to learn robust visual features. Their architectures are increasingly incorporating modality-specific adaptations to better handle unique remote sensing challenges, such as varying spatial resolutions, spectral bands, and geospatial context

IV. Discussion

The emergence of foundation models has led to significant advancements in remote sensing. Using deep learning architectures, these models have demonstrated remarkable improvements in the field, enhancing the accuracy and efficiency of

TABLE I
COMPARISON OF REMOTE SENSING MODELS BASED ON REFERENCES, PARAMETERS, PRETRAINING DATASET SIZE, MODALITIES, AND AVAILABILITY OF WEIGHTS AND ARCHITECTURE

Model	References	Parameters	Pretraining Dataset Size	Modalities	Weights Available	Architecture Available
SeCo	313	23.5M	1,000,000	Multispectral	✓	✓
GeoKR	275	138M	1,431,950	RGB, Multispectral	×	✓
RVSA	269	100M	1,000,848	RGB, Multispectral, SAR	✓	✓
SatMAE	265	307M	1,047,691	RGB, Multispectral, Temporal	✓	✓
RingMo	244	10B	2,000,000	RGB, SAR	✓	✓
RSP	219	25.8M	1,000,848	Multispectral, Hyperspectral	✓	✓
ScaleMAE	172	322.9M	1,047,691	RGB, Multispectral	×	✓
SkySense	112	2.06B	21,500,000	RGB, Multispectral, SAR	✓	×
Prithvi	98	100M	4,200,000	Multispectral	✓	✓
GFM	78	100M	600,000	RGB, Multispectral, SAR	✓	✓

tasks such as image classification, segmentation, and object detection. This section will examine the findings of recent studies, compare performance evaluations, and analyze their overall impact on remote sensing. Since, there are numerous foundation models, we decided to focus on the top 10 models by the number of references of their respective publications.The number of references was determined based on citation counts retrieved from Google Scholar, ensuring a consistent and objective comparison across models. Table I provides a summary of the key attributes of these foundation models, including references, parameters, dataset sizes, and modalities, highlighting the diversity of models used across various remote sensing tasks.

A. Performance of Foundation Models Across Tasks

The performance metrics presented here are directly sourced from original studies, ensuring accuracy and consistency in evaluating these foundation models on benchmark datasets. This provides key insights into their effectiveness and limitations in various tasks.

In **image classification**, foundation models have demonstrated strong performance on the BigEarthNet dataset [36]. Models like msGFM [37] and Skysense [38] achieved the highest mAP of 92.90% and 92.09%, respectively. Their superior performance makes them well suited for high-precision applications. Other models, such as DeCUR [39] (89.70%) and DINO-MC [40] (88.75%), have also shown strong classification capabilities. The variations in performance highlight the importance of model specialization, with some models excelling in specific environmental conditions [41].

Semantic segmentation. Evaluations on ISPRS Postdam dataset show that SkySense [38] achieved the highest mF1 score of 93.99%, while CMID [42] attained the highest mIoU score of 87.04%, making them highly effective in capturing detailed spatial patterns. Models like BFM, Cross-scale MAE, and UPetu, also demonstrate competitive segmentation abilities, highlighting the significant variations in model performance across segmentation tasks [41].

For **object detection**, foundation models were evaluated on the DOTA dataset, where RVSA [30] achieved the highest mAP of 81.24%, followed by SMLFR [43] (79.33%) and RSP [44] (77.72%). On the DIOR and DIOR-R datasets, MTP [45] and SkySense [38] demonstrated strong performance, with an AP50 of 78% and a mAP of 78.73%, respectively. These results highlight the ability of these models to accurately detect objects across various remote sensing applications [41].

In **change detection** tasks, models such as SkySense [38] and GFM [33] achieved the highest F1 scores of 60.06% and 59.82% on the OSCD dataset. Meanwhile, performance was significantly higher on the LEVIR-CD dataset, with MTP [45] attaining an F1 score of 92.67%, closely followed by SkySense [38] (92.58%) [41].

B. Influence of pre-training methods

The rapid advancement of pre-training techniques has greatly expanded the capabilities of foundation models in remote sensing, allowing them to extract complex features from vast amounts of data. By leveraging self-supervised learning techniques, such as contrastive learning (CL) and masked autoencoders (MAE), foundation models have outperformed traditional learning methods in every aspect.

Skysense, which implements a multi-granularity contrastive learning, achieves approximately 3.6% higher accuracy in scene classification and object detection [38] compared to other conventional models. Similarly, Seco, utilizing seasonal contrast learning, improves land-cover classification by up to 7% over ImageNet pre-trained models [46]. Models such as SatMAE and ScaleMae, which utilize masked autoencoding, have demonstrated advancements in multi-temporal and multispectral data processing. SatMae [19] achieved an improvement of up to 14% in land cover classification, while ScaleMae [29] showed an increase in mIoU by 1.7% in segmentation precision.

Among these foundation models, RingMo stands out as the largest and most complex, with 10 billion parameters. Meanwhile, SkySense was trained on the largest dataset, or a combination of multiple datasets, consisting of over 21.5 million images, making it highly effective for multi-modal remote sensing tasks. Overall, these results underscore the importance of pre-training methods in maximizing the performance of foundation models, making them more reliable and efficient in handling complex data [41].

C. Practical Implications and Applications

Foundation models transform the field of remote sensing by enhancing traditional applications like multi-spectral and time-

series data analysis. These models offer higher performance with less reliance on huge labeled data. In this section, we will briefly discuss some of the practical implications of foundation models across different areas:

- **Environmental Monitoring**. Foundation models play a fundamental role in monitoring deforestation, desertification, and pollution. Models like GASSL [47] and SatMAE [48] use multi-spectral and temporal data to detect environmental degradation early, reducing the response time and minimizing the risk of larger disasters.
- **Agriculture and Forestry**. Models like EarthPT [49] and GeCo [50] provide valuable information on crop health, yield prediction, and land use management. RSP [44], for instance, enables precise monitoring of crop conditions, allowing for the early detection of crop stress signs. These models are also highly effective in forestry management, providing detailed maps of forest cover, biomass estimation, and deforestation monitoring.
- **Archaeology**. Models such as GeoKR [51] and RingMo [32] employ high-resolution satellite images and multispectral data to detect archaeological features that are difficult to identify with the naked eye. Meanwhile, models like MATTER [52] analyze texture and material compositions, offering insights into various surface characteristics.
- **Urban Planning and Development**. Foundation models are essential for monitoring urban expansion, infrastructure growth, and land use changes. CMID [42] and SkySense [38] are models that allow advanced urban planning, providing insights on the impact of urbanization on natural habitats and help city planners optimize land use decisions.
- **Disaster Management**. Remote sensing models such as OFA-Net [53], DOFA [54] and Prithvi [55] provide crucial real-time data for detecting affected areas, minimizing response time, and supporting emergency responders in resource allocation and evacuation planning. Furthermore, these models also play a vital role in post-disaster recovery by assessing damages and monitoring environmental changes.

V. Conclusion

In this paper, we explored various foundation models, highlighting their role in remote sensing applications. Most of these models utilize RGB and multispectral data to address tasks such as object detection, land cover classification, and change detection. Their ability to process diverse data modalities makes them valuable for applications in agriculture, urban planning, environmental monitoring, and disaster management. Additionally, this paper emphasizes the importance of pretraining these models on large-scale datasets to enhance their generalization and performance across different remote sensing tasks. By understanding their capabilities and limitations, we can effectively use these models to enhance remote sensing analysis and decision-making.

Acknowledgement

The authors thank the Faculty of computer science and engineering at the Ss. Cyril and Methodius University in Skopje for the provided finacial support under the SatTime ("Analysis of satellite image time-series") project.

References

[1] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE geoscience and remote sensing magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[2] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundation models defining a new era in vision: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[4] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[5] H. Wang, E. Skau, H. Krim, and G. Cervone, "Fusing heterogeneous data: A case for remote sensing and social media," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 6956–6968, 2018.

[6] A. Kior, L. Yudina, Y. Zolin, V. Sukhov, and E. Sukhova, "Rgb imaging as a tool for remote sensing of characteristics of terrestrial plants: A review," *Plants*, vol. 13, no. 9, p. 1262, 2024.

[7] Y. Zhao, X. Zhang, W. Feng, and J. Xu, "Deep learning classification by resnet-18 based on the real spectral dataset from multispectral remote sensing images," *Remote Sensing*, vol. 14, no. 19, p. 4883, 2022.

[8] T. Adão, J. Hruška, L. Pádua, J. Bessa, E. Peres, R. Morais, and J. J. Sousa, "Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry," *Remote sensing*, vol. 9, no. 11, p. 1110, 2017.

[9] Z. Wang and M. Menenti, "Challenges and opportunities in lidar remote sensing," *Frontiers in Remote Sensing*, vol. 2, p. 641723, 2021.

[10] C.-a. Liu, Z.-X. Chen, S. Yun, J.-s. Chen, T. Hasi, and H.-z. Pan, "Research advances of sar remote sensing for agriculture applications: A review," *Journal of integrative agriculture*, vol. 18, no. 3, pp. 506–525, 2019.

[11] K. Omasa, F. Hosoi, and A. Konishi, "3d lidar imaging for detecting and understanding plant responses and canopy structure," *Journal of experimental botany*, vol. 58, no. 4, pp. 881–898, 2007.

[12] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "Ai foundation models in remote sensing: A survey," *arXiv preprint arXiv:2408.03464*, 2024.

[13] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, no. 3, p. 516, 2021.

[14] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big data for remote sensing: Challenges and opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, 2016.

[15] L. Jiao, Z. Huang, X. Lu, X. Liu, Y. Yang, J. Zhao, J. Zhang, B. Hou, S. Yang, F. Liu *et al.*, "Brain-inspired remote sensing foundation models and open problems: A comprehensive survey," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 10 084–10 120, 2023.

[16] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He *et al.*, "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *International Journal of Machine Learning and Cybernetics*, pp. 1–65, 2024.

[19] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. Lobell, and S. Ermon, "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197–211, 2022.

[20] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwer, S. Khan, and F. S. Khan, "Rethinking transformers pre-training for multi-spectral satellite imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 811–27 819.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sensing of Environment*, vol. 301, p. 113924, 2024.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[25] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, "Remoteclip: A vision language foundation model for remote sensing," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[27] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[28] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[29] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell, "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088–4099.

[30] D. Wang, Q. Zhang, Y. Xu, J. Zhang, B. Du, D. Tao, and L. Zhang, "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2022.

[31] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," *arXiv preprint arXiv:2304.05215*, 2023.

[32] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2022.

[33] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 806–16 816.

[34] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, S. A. Wernke, Y. Huo *et al.*, "Vision foundation models in remote sensing: A survey," *IEEE Geoscience and Remote Sensing Magazine*, 2025.

[35] A. Xiao, W. Xuan, J. Wang, J. Huang, D. Tao, S. Lu, and N. Yokoya, "Foundation models for remote sensing and earth observation: A survey," *arXiv preprint arXiv:2410.16602*, 2024.

[36] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*. IEEE, 2019, pp. 5901–5904.

[37] B. Han, S. Zhang, X. Shi, and M. Reichstein, "Bridging remote sensors with multisensor geospatial foundation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 852–27 862.

[38] X. Guo, J. Lao, B. Dang, Y. Zhang, L. Yu, L. Ru, L. Zhong, Z. Huang, K. Wu, D. Hu *et al.*, "Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 672–27 683.

[39] Y. Wang, C. M. Albrecht, N. A. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, "Decur: decoupling common & unique representations for multimodal self-supervision," 2023.

[40] X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley, "Extending global-local view alignment for self-supervised learning with remote sensing imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2443–2453.

[41] S. Lu, J. Guo, J. R. Zimmer-Dauphinee, J. M. Nieusma, X. Wang, P. VanValkenburgh, S. A. Wernke, and Y. Huo, "Vision foundation models in remote sensing: A survey," *arXiv preprint arXiv:2408.03464*, 2024.

[42] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "Cmid: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–17, 2023.

[43] Z. Dong, Y. Gu, and T. Liu, "Generative convnet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[44] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2022.

[45] D. Wang, J. Zhang, M. Xu, L. Liu, D. Wang, E. Gao, C. Han, H. Guo, B. Du, D. Tao *et al.*, "Mtp: Advancing remote sensing foundation model via multi-task pretraining. arxiv," *arXiv preprint arXiv:2403.13430*, 2024.

[46] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9414–9423.

[47] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 181–10 190.

[48] G. Cheng, J. Wang, K. Li, X. Xie, C. Lang, Y. Yao, and J. Han, "Anchor-free oriented proposal generator for object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[49] A. Stewart, N. Lehmann, I. Corley, Y. Wang, Y.-C. Chang, N. A. Ait Ali Braham, S. Sehgal, C. Robinson, and A. Banerjee, "Ssl4eo-l: Datasets and foundation models for landsat imagery," *Advances in Neural Information Processing Systems*, vol. 36, pp. 59 787–59 807, 2023.

[50] W. Li, K. Chen, and Z. Shi, "Geographical supervision correction for remote sensing representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–20, 2022.

[51] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[52] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8203–8215.

[53] Z. Xiong, Y. Wang, F. Zhang, and X. X. Zhu, "One for all: Toward unified foundation models for earth vision," in *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2024, pp. 2734–2738.

[54] Z. Xiong, Y. Wang, F. Zhang, A. J. Stewart, J. Hanna, D. Borth, I. Papoutsis, B. L. Saux, G. Camps-Valls, and X. X. Zhu, "Neural plasticity-inspired multimodal foundation model for earth observation," *arXiv preprint arXiv:2403.15356*, 2024.

[55] J. Jakubik, S. Roy, C. Phillips, P. Fraccaro, D. Godwin, B. Zadrozny, D. Szwarcman, C. Gomes, G. Nyirjesy, B. Edwards *et al.*, "Foundation models for generalist geospatial artificial intelligence," *arXiv preprint arXiv:2310.18660*, 2023.