Vision Foundation Models in Remote Sensing

A survey

SIQI LU, JUNLIN GUO, JAMES R. ZIMMER-DAUPHINEE, JORDAN M. NIEUSMA, XIAO WANG, PARKER VANVALKENBURGH, STEVEN A. WERNKE, AND YUANKAI HUO

Artificial intelligence (AI) technologies have profoundly transformed the field of remote sensing (RS), revolutionizing data collection, processing, and analysis. Traditionally reliant on manual interpretation and task-specific models, RS research has been significantly enhanced by the advent of foundation models (FMs)—large-scale pretrained AI models capable of performing a wide array of tasks with unprecedented accuracy and efficiency. This article provides a comprehensive survey of FMs in the RS domain. We categorize these models based on their architectures, pretraining datasets, and methodologies. Through detailed performance comparisons, we highlight emerging trends and the significant advancements achieved by those FMs. Additionally, we discuss technical challenges, practical implications, and future research directions, addressing the need for high-quality data, computational resources, and improved model generalization. Our research also finds that pretraining methods, particularly self-supervised learning (SSL) techniques like contrastive learning (CL) and masked autoencoders (MAEs), remarkably enhance the performance and robustness of FMs. This survey aims to serve as a resource for researchers and practitioners by providing a panorama of advances and promising pathways for the continued development and application of FMs in RS.

INTRODUCTION

AI technologies have profoundly transformed the field of RS, revolutionizing how data are collected, processed, and analyzed. Traditionally, RS projects relied heavily on manual interpretation and task-specific models that required

DISCLAIMER

Any subjective views or opinions that might be expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

Digital Object Identifier 10.1109/MGRS.2025.3541952

extensive labeled datasets and significant computational resources. However, the advent of AI and deep learning (DL) has ushered in a new era in which large-scale pretrained models, known as *FMs*, are capable of performing a wide array of tasks with unprecedented accuracy and efficiency. These advancements have not only enhanced the potential applications of RS but have also opened new avenues for its usage across various domains.

In recent years, numerous vision FMs have emerged, demonstrating remarkable performance in handling diverse RS tasks. These models have shown the potential to significantly improve performance on multiple downstream tasks such as scene classification, semantic segmentation, object detection, and more. By leveraging vast amounts of pretraining data and sophisticated architectures, these FMs have set new benchmarks in the field, making them indispensable tools for researchers and engineers alike.

This article aims to provide a comprehensive survey of vision FMs in the RS domain ad rem and is limited to FMs released between June 2021 and June 2024. This timeframe marks a surge in the development of modern FMs, including vision transformers (ViTs) and advanced SSL techniques. Although early models like Tile2Vec [47] and others laid the groundwork for representation learning in RS, they were typically limited in scale and generalization capabilities. Furthermore, numerous review articles and papers have already provided comprehensive overviews of these pre-2021 models. Our review, therefore, focuses on recent developments to highlight the unique contributions and innovations that have emerged in the past few years.

In Figure 1 [58], vision FMs are listed in chronological order. To facilitate navigation and enhance utility for researchers, we categorized existing models based on their perception levels (e.g., image level, region level, and pixel level). This organization helps clarify which models have been tested for general image-based challenges or specialized applications, such as environmental monitoring, land cover mapping, archaeological exploration, disaster management, and more. It is essential to distinguish between applications that models have been explicitly tested on and those for which they could potentially be effective. In this review, the fact that a model has not been tested on a particular application does not mean it won't perform well. FMs, especially convolutional neural network (CNN) backbones like residual networks [Residual Neural Networks (ResNets)] [36] and ViTs [25], may still be suitable for various downstream tasks, even if prior work has not yet demonstrated this (Figure 1).

Our contributions include the following:

1) We provide an exhaustive review of the current state of vision FMs proposed in the field of RS, starting from the background and methodologies of these models to specific applications across different domains and tasks in a hierarchical and structured manner.
2) We provide the categorization and analysis of the models based on their application in both image analysis (Table 1) and practical applications (Table 2). We discuss the architecture, pretraining datasets, pretraining methods, and performance of each model.
3) We provide a discussion of challenges and unresolved aspects related to FMs in RS. We pinpoint new trends, raise important questions, and propose future directions for further exploration.

BACKGROUND

REMOTE SENSING

RS refers to the process of acquiring information about objects or areas from a distance, typically using satellite or airborne sensors. These technologies and techniques

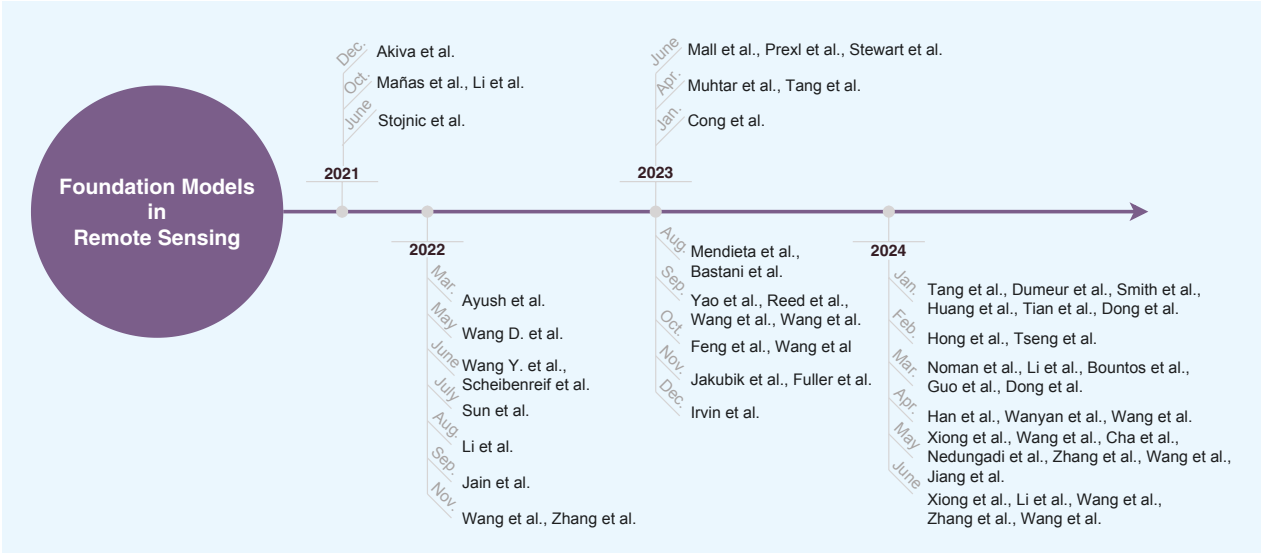

FIGURE 1. An overview of some well-known FMs for RS from June 2021 to June 2024. Detailed reference numbers are listed in Tables 2–4.

serve vital roles in diverse fields, enabling the collection of data over geographic areas without physical contact. Applications of RS include Earth observation, digital archaeology, urban planning and development, and disaster management. The field of RS has developed rapidly since the mid-20th century. Initially, RS predominately consisted of analog photographic techniques via aerial and satellite platforms, which provided limited spectral and spatial resolution. The launch of early Earth observation satellites, such as the Landsat program (which commenced in 1967 [112]), marked a significant advancement, enabling consistent and wide-ranging data collection for environmental monitoring.

Modern RS employs a variety of sensors suited for specific types of data collection, including optical, thermal, and radar. Optical sensors capture a wide variety of spectral bands, including visible and near-infrared light, allowing for the detailed imaging of land cover and vegetation health. Thermal sensors detect heat emitted or reflected from Earth's surface, which is useful for monitoring volcanic activity, forest fires, and climate change monitoring. Radar sensors can penetrate clouds and vegetation, providing critical information in all-weather conditions and for applications such as soil moisture estimation and urban infrastructure mapping [17], [71].

In recent years, RS has found applications in many fields. With regard to environmental monitoring, it is used to track deforestation, to monitor air and water quality, and to assess the impacts of climate change [30], [39]. In agriculture, RS helps in crop health monitoring, yield estimation, and efficient resource management [71]. Urban planning and development benefit from RS through the monitoring of urban sprawl, infrastructure development, and land use planning [17], [48]. Furthermore, in disaster management, RS is crucial for assessing the damage caused by natural disasters, aiding in the planning and execution of relief operations [1], [30].

The integration of RS data with Geographic Information Systems (GIS) has further enhanced its utility. GIS provides a framework for capturing, storing, analyzing, and visualizing spatial and geographic data. When combined with RS data, GIS can be used to create detailed and dynamic maps and models for various applications. This synergy is particularly valuable in resource management, urban planning, and disaster response, where accurate and timely information is critical [17], [30], [71].

FOUNDATION MODELS FOR REMOTE SENSING

FMs refer to large-scale pretrained models that provide a robust starting point for various downstream tasks across different domains [50]. These models leverage extensive datasets and advanced architectures, enabling them to capture complex patterns and features that can be fine-tuned for specific applications with minimal additional training. In RS, FMs are particularly valuable due to the diverse and complex nature of the data (Figure 2), including multispectral and multitemporal imagery. Techniques such as SSL [51] and transformers [93] have significantly enhanced the performance and efficiency of tasks such as image classification, object detection, and change detection, addressing the unique challenges posed by RS data [19].

A major strength of these models lies in their ability to utilize SSL to learn effective representations from largely unlabeled data, which is often abundant in RS scenarios [38]. By integrating advanced architectures like

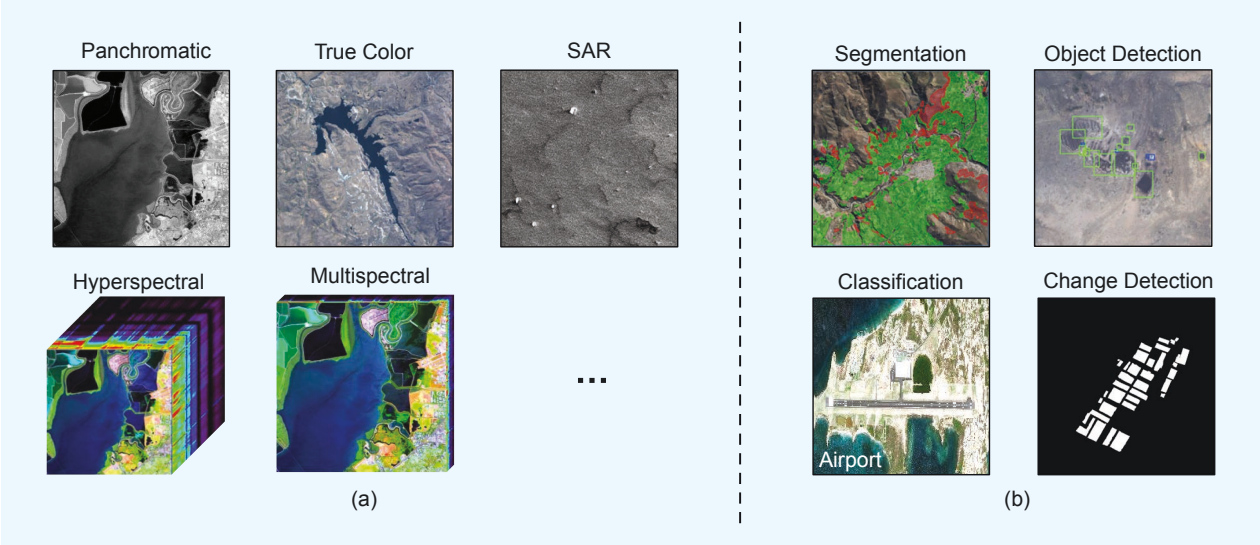

FIGURE 2. Examples of (a) data types used in those FMs and (b) downstream tasks that can be done by FMs. (a) Data: 1) Panchromatic [4], 2) True Color, 3) SAR [94], 4) Hyperspectral [4], and 5) Multispectral [4]. (b) Downstream tasks: 1) Segmentation, 2) Object Detection, 3) Classification [15], and 4) Change Detection [76]. (Source: True Color, Segmentation, and Object detection images copyright MAXAR 2024, provided through the NextView License Agreement.)

TABLE 1. A SUMMARY OF THE PRETRAINING METHODS UTILIZED AND IMAGE ANALYSIS TASKS EVALUATED ACROSS DIFFERENT MODELS. IMAGE LEVEL, PIXEL LEVEL, REGION LEVEL, AND SPATIAL-TEMPORAL CLASSIFY THE TASKS IN IMAGE ANALYSIS, WHILE CL AND PREDICTIVE CODING INDICATE THE DIFFERENT SELF-SUPERVISED PRETRAINING STRATEGIES THAT EACH STUDY USED.

MONTH AND YEAR	ARCHITECTURE	MODEL NAME	IMAGE LEVEL	PIXEL LEVEL	REGION LEVEL	SPATIAL-TEMPORAL	CL	PREDICTIVE CODING
June 2021	ResNet-50	CMC-RSSR [84]	✓				✓	
Oct. 2021	ResNet-50	SeCo [66]	✓			✓		
Oct. 2021	ResNet-50	GeoKR [56]	✓	✓	✓			
Dec. 2021	ResNet-34	MATTER [2]	✓	✓		✓		✓
Mar. 2022	ResNet-50	GASSL [6]	✓	✓	✓		✓	
May 2022	ViTAEv2-S	RSP [96]	✓	✓	✓	✓		
June 2022	ViT-S/8	DINO-MM [105]	✓				✓	
June 2022	Swin Transformer	Scheibenreif et al. [79]	✓	✓			✓	
July 2022	ViT/Swin Transformer	RingMo [87]	✓	✓	✓	✓		✓
Aug. 2022	ResNet-50	GeCO [57]	✓	✓	✓			✓
Sep. 2022	BYOL	RS-BYOL [45]	✓	✓			✓	
Nov. 2022	ViT-B	CSPT [104]	✓		✓			✓
Nov. 2022	ViT	RVSA [100]	✓	✓	✓			✓
Jan. 2023	MAE-based Framework	SatMAE [16]	✓	✓				✓
Apr. 2023	TOV	TOV [89]	✓	✓	✓			✓
Apr. 2023	Teacher-Student Self-Distillation	CMID [70]	✓	✓	✓	✓		
June 2023	CACo	CACo [67]	✓	✓		✓	✓	
2023 Jun	ResNet-18	IaI-SimCLR [77]	✓				✓	
June 2023	ResNet	SSL4EO-L [83]		✓			✓	
Aug. 2023	Teacher-Student	GFM [69]	✓	✓		✓		✓
Aug. 2023	Swim Transformer	SatLasPretrain [7]	✓	✓				
Sep. 2023	Multi-Branch	RingMo-Sense [119]		✓				✓
Sep. 2023	ViT	Scale-MAE [78]	✓	✓				✓
Sep. 2023	CNN-Transformer	RingMo-lite [109]	✓	✓	✓	✓		✓
Sep. 2023	Multimodel SSL	DeCUR [102]	✓	✓				✓
Oct. 2023	MSFE+MMFH	Feng et al. [27]	✓	✓	✓	✓		✓
Oct. 2023	ViT	FG-MAE [108]	✓	✓				✓
Nov. 2023	ViT	Prithvi [46]		✓				✓
Nov. 2023	Multimodal Encoder	CROMA [28]	✓	✓			✓	✓
Dec. 2023	ViT	USat [44]	✓					✓
Jan. 2024	ViT-B	Cross-Scale MAE [88]	✓	✓				✓
Jan. 2024	Unet+Transformer	U-BARN [26]	✓	✓				
Jan. 2024	Autoregressive Transformer	EarthPT [82]	✓					✓
Jan. 2024	Teacher-Student Network	GeRSP [42]	✓	✓	✓		✓	✓
Jan. 2024	Dual-Branch	SwiMDiff [91]	✓					
Jan. 2024	Generative ConvNet	SMLFR [22]		✓	✓			✓
Feb. 2024	3D GPT	SpectralGPT [40]	✓	✓		✓		✓
Feb. 2024	MAE-based Framework	Presto [92]		✓			✓	✓
Mar. 2024	SatMAE	SatMAE++ [73]	✓					✓
Mar. 2024	Joint-Embedding Predictive Architecture	SAR-JEPA [58]	✓					✓
Mar. 2024	ViT	FoMo-Bench [8]	✓	✓	✓			✓
Mar. 2024	Factorized Multi-Modal Spatiotemporal Encoder	SkySense [32]	✓	✓	✓	✓		✓
Mar. 2024	Multi-Modules	UPetu [24]	✓	✓		✓		✓
Apr. 2024	Swim Transformer	msGFM [33]	✓	✓				✓
Apr. 2024	DINO	DINO-MC [111]	✓			✓	✓	
May 2024	OFA-Net	OFA-Net [118]	✓	✓				✓
May 2024	Shared Encoder, Task-Specific Decoders	MTP [99]	✓	✓	✓	✓		
May 2024	ViT	BFM [11]		✓	✓			✓
May 2024	MP-MAE	MMEarth [72]	✓	✓				✓

(*Continued*)

TABLE 1. A SUMMARY OF THE PRETRAINING METHODS UTILIZED AND IMAGE ANALYSIS TASKS EVALUATED ACROSS DIFFERENT MODELS. IMAGE LEVEL, PIXEL LEVEL, REGION LEVEL, AND SPATIAL-TEMPORAL CLASSIFY THE TASKS IN IMAGE ANALYSIS, WHILE CL AND PREDICTIVE CODING INDICATE THE DIFFERENT SELF-SUPERVISED PRETRAINING STRATEGIES THAT EACH STUDY USED. (*Continued*)

MONTH AND YEAR	ARCHITECTURE	MODEL NAME	IMAGE LEVEL	PIXEL LEVEL	REGION LEVEL	SPATIAL-TEMPORAL	CL	PREDICTIVE CODING
May 2024	ViT	CtxMIM [90]	✓	✓	✓			✓
May 2024	HiViT	SARATR-X [54]	✓		✓			✓
May 2024	Transformer	SoftCon [106]	✓	✓		✓	✓	
May 2024	ViT	LeMeViT [49]		✓	✓	✓		
June 2024	Masked Autoencoder	S2MAE [59]	✓			✓		✓
June 2024	CNN-Transformer	RS-DFM [110]		✓	✓			
June 2024	MAE-based	A2-MAE [122]	✓	✓		✓		
June 2024	ViT	HyperSIGMA [95]	✓	✓	✓	✓		✓
June 2024	Dynamic OFA	DOFA [117]	✓	✓				✓

transformers [93], FMs in RS can handle the unique characteristics of geospatial data, such as varying spatial resolutions and temporal dynamics, without requiring separate task-specific models.

The evolution of FMs has been driven by advancements in DL and the availability of large datasets. Initially, CNNs like ResNet [36] paved the way for improved image recognition and classification tasks [65]. The introduction of transformers, which use self-attention mechanisms to model long-range dependencies, has further advanced the capabilities of FMs in handling large-scale image data [16]. ViTs [25] extend the transformer architecture to process image data by treating image patches as sequences of tokens, enabling models to learn both local and global relationships. This capability makes transformers particularly effective for semantic segmentation and change detection tasks, where capturing long-range dependencies is crucial, especially in high-resolution satellite imagery.

Notable FMs in RS include SatMAE [16], which pretrains transformers for temporal and multispectral satellite imagery; Scale-MAE [78], a scale-aware MAE for multiscale geospatial representation learning; and DINO-MC [111], which extends global-local view alignment for SSL with RS imagery. These models have shown remarkable performance in various RS tasks, such as scene classification, object detection, and change detection.

Despite their success, FMs face several challenges, including the need for high-quality and diverse training data, significant computational resources, and effective domain adaptation to specific RS tasks [73]. Addressing these challenges will be crucial for the continued advancement of FMs in RS.

RELATED REVIEW ARTICLES AND PAPERS

AI in RS has been a growing area of research, with numerous review articles and papers providing insights into AI advancements and their applications. In this section, we summarize the most influential reviews on FMs in RS.

In 2016, Zhang et al., in their foundational review, "Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art" [121], introduced DL techniques to RS, focusing on CNNs for tasks such as image classification and object detection. This work highlighted both the promise and challenges of early AI integration in RS, setting the stage for subsequent advancements.

In 2017, Zhu et al.'s "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources" [37] delved into diverse AI applications, including hyperspectral analysis and synthetic aperture radar (SAR) interpretation. It also provided an extensive resource list, capturing the rapid adoption of DL in addressing complex RS challenges, paving the way for more advanced AI models in the following years.

More recent reviews have focused on advanced AI models and methods. Wang et al.'s 2022 review, "Self-Supervised Learning in Remote Sensing" [103], highlighted the ability of SSL methods to utilize large volumes of unlabeled data, significantly reducing dependence on labeled datasets while maintaining high performance in RS tasks. The review also identified key challenges and future directions, emphasizing SSL's potential to handle large-scale RS data complexities.

Zhang and Zhang (2022), in "Artificial Intelligence for Remote Sensing Data Analysis: A Review of Challenges and Opportunities" [120], offered a comprehensive overview of AI algorithms, synthesizing findings from more than 270 studies. It emphasized ongoing challenges such as explainability, security, and integrating AI with other computational techniques, serving as a road map for future innovation in AI-driven RS.

Aleissaee et al.'s 2023 survey, "Transformers in Remote Sensing" [3], explored the impact of transformer-based models across various RS tasks, comparing them with CNNs. It identified both strengths and limitations, along with unresolved challenges, providing a detailed road map for future research on transformers' role in RS.

Li et al.'s 2024 review, "Vision-Language Models in Remote Sensing" [60], examined the increasing significance

TABLE 2. THIS TABLE ILLUSTRATES VARIOUS TASKS IN DIFFERENT APPLICATIONS FOR RS. KEY AREAS INCLUDE ENVIRONMENTAL MONITORING, AGRICULTURE, URBAN PLANNING AND DEVELOPMENT, DISASTER MANAGEMENT, AND ARCHAEOLOGY. EACH DOMAIN COMPRISES SPECIFIC TASKS IN DIFFERENT IMAGE ANALYSIS LEVELS, LIKE IMAGE LEVEL, PIXEL LEVEL, REGION LEVEL, AND SPATIAL-TEMPORAL. THE RELATIONSHIPS BETWEEN THESE TASKS AND THEIR APPLICATIONS ARE DEPICTED THROUGH CHECKMARKS, EMPHASIZING THE INTERCONNECTED NATURE OF IMAGE ANALYSIS METHODS ACROSS DIFFERENT FIELDS.

		IMAGE ANALYSIS BY LEVELS				
TASKS		**IMAGE LEVEL**	**PIXEL LEVEL**	**REGION LEVEL**	**SPATIAL-TEMPORAL**	**RELATED WORK**
Environmental monitoring	Land cover change detection				✓	[2], [24], [32], [40], [49], [59], [66], [67], [69], [70], [87], [91], [95], [96], [99], [106], [109], [111], [122]
	Deforestation monitoring		✓			[2], [6], [7], [16], [22], [24], [26], [27], [28], [32], [33], [40], [42], [45], [49], [56], [57], [67], [69], [70], [78], [79], [87], [88], [89], [92], [95], [96], [99], [100], [102], [106], [108], [109], [110], [119], [117], [118], [122], [90]
	Water body analysis		✓	✓	✓	[27], [32], [49], [70], [87], [95], [96], [99]
	Forest cover mapping		✓		✓	[27], [40], [49], [59], [67], [69], [87], [95], [96], [106], [109], [122]
	Biomass estimation					[77]
	Weather/climate prediction	✓				[82], [101], [119]
	Cloud removal					[33]
	Moisture content measurement					[92]
Agriculture	Crop type mapping	✓	✓	✓	✓	[27], [32], [49], [70], [83], [87], [95], [96], [99]
	Weed detection			✓		[6], [8], [22], [27], [32], [49], [56], [57], [70], [87], [89], [95], [96], [99], [100], [104], [110], [90]
	Disease monitoring	✓	✓		✓	[6], [27], [22], [32], [49], [56], [57], [70], [87], [89], [95], [96], [99], [100], [110], [90]
	Forecasting					[82], [119]
	Soil parameter estimation					[117]
	Yield estimation	✓	✓			[2], [6], [7], [16], [24], [26], [27], [32], [33], [40], [42], [45], [56], [57], [67], [69], [70], [78], [79], [87], [88], [89], [95], [99], [96], [100], [102], [106], [108], [109], [118], [117], [122], [90]
	Agricultural pattern segmentation		✓			[66]
Archaeology	Artifact classification and recognition	✓		✓		[2], [6], [7], [8], [16], [26], [24], [27], [32], [33], [40], [42], [44], [45], [56], [57], [59], [58], [66], [67], [69], [70], [72], [77], [78], [82], [88], [91], [89], [79], [84], [87], [95], [96], [99], [105], [100], [102], [104], [106], [108], [109], [111], [118], [90], [117], [122]
	Detection of archaeological structures			✓		[6], [8], [22], [27], [32], [49], [56], [57], [70], [87], [89], [95], [99], [96], [100], [104], [110], [90]
	Semantic segmentation		✓			[2], [6], [7], [16], [22], [24], [26], [27], [28], [32], [33], [40], [49], [42], [45], [56], [57], [67], [69], [70], [78], [79], [87], [89], [96], [88], [92], [95], [99], [100], [102], [109], [119], [106], [108], [110], [117], [118], [122], [90]
	Texture/structural analysis					[2]
	Pattern recognition		✓	✓		[6], [22], [32], [49], [27], [56], [57], [70], [87], [89], [95], [96], [99], [100], [110], [90]
Urban planning and development	Traffic monitoring	✓		✓	✓	[27], [32], [49], [70], [87], [95], [96], [99]
	Land cover/use classification	✓				[2], [6], [7], [8], [16], [24], [26], [27], [32], [33], [40], [45], [57], [59], [42], [44], [58], [56], [66], [67], [69], [70], [72], [77], [78], [79], [83], [84], [87], [89], [82], [88], [91], [96], [99], [95], [102], [117], [105], [100], [109], [104], [108], [106], [111], [118], [122], [90]
	Road crack detection			✓		[6], [8], [27], [22], [32], [49], [56], [57], [70], [87], [89], [95], [96], [99], [100], [104], [110], [90]

(*Continued*)

TABLE 2. THIS TABLE ILLUSTRATES VARIOUS TASKS IN DIFFERENT APPLICATIONS FOR RS. KEY AREAS INCLUDE ENVIRONMENTAL MONITORING, AGRICULTURE, URBAN PLANNING AND DEVELOPMENT, DISASTER MANAGEMENT, AND ARCHAEOLOGY. EACH DOMAIN COMPRISES SPECIFIC TASKS IN DIFFERENT IMAGE ANALYSIS LEVELS, LIKE IMAGE LEVEL, PIXEL LEVEL, REGION LEVEL, AND SPATIAL-TEMPORAL. THE RELATIONSHIPS BETWEEN THESE TASKS AND THEIR APPLICATIONS ARE DEPICTED THROUGH CHECKMARKS, EMPHASIZING THE INTERCONNECTED NATURE OF IMAGE ANALYSIS METHODS ACROSS DIFFERENT FIELDS. (*Continued*)

		IMAGE ANALYSIS BY LEVELS				
TASKS		IMAGE LEVEL	PIXEL LEVEL	REGION LEVEL	SPATIAL-TEMPORAL	RELATED WORK
	Air quality monitoring	✓		✓		[6], [22], [27], [33], [32], [49], [95], [56], [57], [70], [87], [89], [96], [99], [100], [110], [90]
	Building extraction	✓	✓			[27]
	Object/video tracking			✓	✓	[119]
	Infrastructure monitoring					[44], [119]
Disaster management	Landslide risk monitoring	✓	✓	✓	✓	[27], [32], [49], [70], [87], [95], [96], [99]
	Disaster response					[117]
	Real-time detection and mapping		✓	✓	✓	[27], [32], [49], [70], [87], [95], [96], [99]
	Building damage assessment	✓	✓	✓		[27], [32], [49], [70], [87], [95], [96], [99]
	Critical infrastructure detection	✓		✓		[6], [8], [22], [32], [49], [27], [56], [57], [70], [87], [89], [96], [100], [95], [99], [104], [110], [90]
	Flood/fire mapping and prediction	✓	✓		✓	[27], [40], [67], [69], [87], [96], [49], [95], [106], [109], [122]
	Crowd and vehicle detection			✓	✓	[2], [24], [32], [49], [59], [40], [66], [67], [69], [70], [87], [91], [95], [96], [99], [109], [106], [111], [122]

of vision-language models (VLMs), which combine visual and textual data. It highlighted VLMs' potential in applications like image captioning and visual question answering, emphasizing a shift toward richer semantic understanding in RS tasks.

Additionally, the recent work, "On the Foundations of Earth and Climate Foundation Models" [97], provided a comprehensive review of existing FMs, proposing features like geolocation embedding and multisensory capability. It outlined key traits for future Earth and climate models, contributing to a broader discussion on foundational advancements in geospatial AI.

Building on these reviews, our study provides a comprehensive analysis of FMs developed from June 2021 to June 2024, focusing on advances in SSL and transformer-based architectures. Unlike previous reviews, which focused mainly on individual techniques, we explore their combined potential in RS tasks like semantic segmentation, multispectral analysis, and change detection. For instance, SatMAE [16] demonstrates the effective use of SSL for pretraining transformers, enabling improved segmentation in complex multispectral imagery, while Scale-MAE employs scale-aware MAEs for better handling of varied spatial resolutions in RS data.

Our study also highlights new models like DINO-MC [111], which integrates global-local view alignment for SSL, making it particularly effective for identifying changes in high-resolution satellite imagery. By systematically examining these innovations, we illustrate how recent models address persistent challenges like domain adaptation and computational efficiency. For example, efficient self-attention mechanisms in Scale-MAE [78] help reduce computation costs, while enhanced geolocation embeddings in models like SatMAE improve performance in geospatial feature extraction.

In contrast to earlier reviews, which often remained theoretical, we emphasize both the theoretical advancements and practical applications of recent models. For example, DINO-MC's [111] and ORBIT's [101] real-world applications in environmental monitoring and disaster response highlight their practical impact, demonstrating how new FMs can be effectively leveraged to address pressing challenges in geospatial analysis.

PRETRAINING METHODS

Pretraining serves as a critical step in developing FMs, enabling them to learn transferable and generalized representations from large-scale datasets. This process leverages self-supervised or supervised learning methods to extract domain-agnostic features that can be adapted to various downstream tasks. In this section, we explore the key pretraining methods utilized commonly in FMs for RS, explaining the mechanism of these methods and their roles in enhancing model performance and addressing challenges in this field.

SELF-SUPERVISED LEARNING

SSL has emerged as a cornerstone of pretraining FMs, offering a paradigm where models learn representations by predicting parts of the input data from other parts. This approach reduces reliance on expensive and time-consuming

labeled datasets, making it particularly advantageous in fields like RS, where labeled data are often scarce or challenging to obtain.

SSL allows models to exploit vast amounts of unlabeled data, learning rich and generalizable representations that transfer well to downstream tasks such as scene classification, semantic segmentation, object detection, and change detection. By uncovering underlying data structures and patterns, SSL not only enhances model robustness but also improves adaptability across diverse domains and resolutions of RS imagery [103]. Figure 3 illustrates the general pipeline of SSL. Two SSL methods commonly used in vision FMs for RS are predictive coding and CL, each offering unique mechanisms to harness information from unlabeled data.

PREDICTIVE CODING

Predictive coding leverages a generative approach, where the model learns to predict missing or occluded parts of an image based on visible portions. This strategy helps capture spatial and contextual relationships in RS imagery, which often contains diverse textures, complex scenes, and varying resolutions.

In RS, predictive coding can be applied to tasks such as gap filling in satellite imagery, where the model learns to infer missing data caused by sensor limitations or occlusions like cloud cover. Popular implementations of predictive coding frameworks include autoencoder-based architectures, masked image modeling (MIM) techniques like those used in MAEs [34], and autoregressive models. These methods are particularly effective in learning fine-grained details critical for high-resolution imagery and specialized tasks.

CONTRASTIVE LEARNING

CL is another powerful SSL technique that focuses on distinguishing between similar and dissimilar samples in the data. The key idea is to bring representations of similar (positive) samples closer together while pushing apart those of dissimilar (negative) samples. This encourages the model to learn discriminative and invariant features that are crucial for RS tasks.

CL frameworks such as SimCLR [13], MoCo [35], DINO [9], and BYOL [29] have shown promise in RS applications. They use augmentations like random cropping, rotations, and spectral band dropping to generate positive pairs, enabling the model to learn robust representations invariant to these transformations. For instance, in multispectral or hyperspectral imagery, CL can help models capture spectral signatures across varying conditions, improving performance in tasks like crop classification or land cover mapping [103]. CL is especially relevant in RS when labeled datasets are highly imbalanced as it enables models to learn from underrepresented classes or regions without explicit labels.

By combining approaches like predictive coding and CL, SSL has significantly advanced the development of

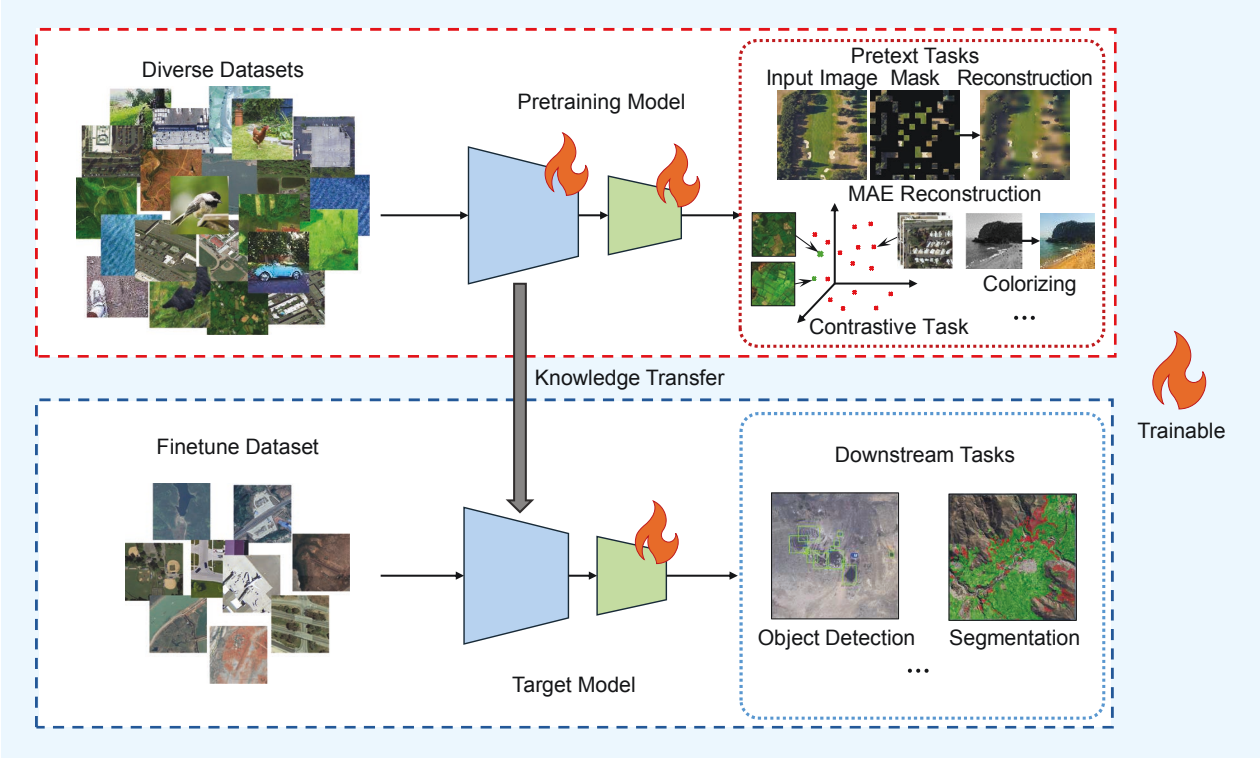

FIGURE 3. The general pipeline of SSL [51]. Diverse dataset images and pretext task images are acquired from ImageNet [18], BigEarthNet [85], and MillionAID [64]. The Finetune dataset includes images from DIOR [55]. (Source: Object Detection and Segmentation, copyright MAXAR 2024, provided through the NextView License Agreement.)

vision FMs in RS. These methods allow models to leverage vast unlabeled datasets while maintaining adaptability across diverse spatial resolutions, spectral bands, and application scenarios. On the other hand, it is important to note that there are many other SSL methods that can be employed for such tasks. Other innovative methods, such as teacher-student self-distillation frameworks, have also demonstrated potential in RS applications. For example, CMID [70] achieves promising performance by combining CL and MIM in a teacher-student self-distillation framework. This structure enables it to capture both global and local features, making it effective for diverse RS tasks. The diversity of SSL techniques highlights the versatility and evolving nature of SSL, underscoring its critical role in unlocking the full potential of RS imagery.

SUPERVISED PRETRAINING

Supervised pretraining is a fundamental approach in DL, where models are trained using labeled datasets to minimize prediction errors for specific tasks, such as image classification. This method allows models to learn direct mappings between input features and target labels, fostering the development of detailed and task-specific representations. For instance, models like ResNet [36] and Visual Geometry Group Network (VGGNet) [81] trained on large-scale datasets such as ImageNet [18] have demonstrated how supervised pretraining can capture robust feature hierarchies that are highly transferable to related tasks, including semantic segmentation and object detection.

In RS, supervised pretraining has shown promise for tasks such as land cover classification and object detection using high-resolution satellite imagery [96]. However, the dependency on large-scale labeled datasets presents a major limitation. Creating labeled datasets for RS tasks, particularly when involving multispectral or hyperspectral data, is resource intensive and often requires domain expertise for annotation. For example, labeling pixel-level data for land cover classification or delineating objects in complex urban environments can be prohibitively time consuming. Furthermore, labeled data in RS are often domain specific, limiting the generalizability of models trained on one dataset to other applications or regions [37].

These challenges highlight the need for innovative strategies to address the reliance on labeled data. Such limitations have motivated the development of alternative approaches, including self-supervised pretraining methods, which leverage the abundance of unlabeled data to learn general-purpose representations without manual annotation.

IMAGE ANALYSIS METHODS

IMAGE PERCEPTION AT DIFFERENT LEVELS

FMs in RS enable image analysis at three primary levels: the image level, region level, and pixel level. These levels address varying spatial, contextual, and application-specific needs, providing the foundation for a wide range of tasks, such as environmental monitoring, urban planning, disaster response, and more. The following sections outline the distinct objectives and applications at each level. A detailed summary of the models evaluated for these tasks is provided in Table 3. The following sections outline the distinct objectives and applications at each level.

IMAGE LEVEL

Image-level analyses focus on classification tasks, categorizing entire images or large image segments into predefined classes, such as urban, forest, water bodies, or agricultural areas. This approach provides broad high-level insights into geographic regions and is instrumental in large-scale applications like land use mapping, land cover classification, and resource management. By classifying entire scenes, this level of analysis enables the efficient monitoring of extensive areas, supporting decision making in environmental management and policy planning.

REGION LEVEL

Region-level analysis identifies and localizes specific objects within an image, such as buildings, vehicles, ships, or other structures. Unlike image-level analysis, which provides holistic classifications, region-level tasks focus on object detection, which is used to detect individual entities and their spatial locations. This analysis is critical for targeted applications like urban planning, where the detection of infrastructure is essential, as well as disaster response and security, where identifying damaged buildings or vulnerable areas can significantly aid in timely interventions.

PIXEL LEVEL

Pixel-level analysis offers the most granular form of image perception, assigning a label to every pixel within an image. This includes tasks such as semantic segmentation, where each pixel is classified into categories like vegetation, water, or buildings; it also includes change detection, which identifies temporal differences between images captured at different times. Pixel-level analysis is indispensable for creating highly detailed maps used in applications like precision agriculture, deforestation tracking, and disaster management. The ability to analyze fine-grained details enables more accurate assessments and actionable insights for these critical areas.

BACKBONE

CONVOLUTIONAL NEURAL NETWORKS

CNNs [74] are a fundamental architecture in DL, designed to extract hierarchical spatial features from images through the use of convolutional layers. Each convolutional layer applies filters to the input data, detecting patterns like edges, textures, and shapes at different levels of abstraction.

TABLE 3. OVERVIEW OF RECENT FMS IN RS, CATEGORIZED BY ARCHITECTURE, MODEL NAME, PRETRAINING DATASET, RESOLUTION, GEOGRAPHIC COVERAGE, IMAGE ANALYSIS LEVELS, VISUAL ENCODER, PRETRAINING METHODS, AND THE NUMBER OF PARAMETERS.

MODEL NAME	ARCHITECTURE	PRETRAINING DATASET	RESOLUTION (M)	GEOGRAPHIC COVERAGE	IMAGE ANALYSIS LEVELS	PRETRAIN METHODS	NO. OF PARAMS
CMC-RSSR [84]	ResNet-50	NWPU-DOTA [113], BigEarthNet [85], ImageNet [18]	0.2–60	Global	Image level	Contrastive multiview coding	23 million
SeCo [66]	ResNet-50	*Sentinel-2* imagery	10, 20, 60	200,000 locations worldwide	Image level, spatial-temporal	CL	23.5 million
GeoKR [56]	ResNet-50	Levir-KR [56]	0.8–16	Global	Image level, pixel level, region level	Geo-graphical knowledge supervision	23.5 million/ 138 million
MATTER [2]	ResNet-34	*Sentinel-2* Imagery	—	Rural and remote regions with few changes	Image level, pixel level	SSL	21.3 million
GASSL [6]	ResNet-50	fMoW [15], GeoImageNet [18]	—	Seven continents	Image level, pixel level region level	CL	23.5 million
RSP [96]	ViTAEv2-S	MillionAID [63], [64]	0.5–153	Global	Image level, pixel level, region level, spatial-temporal	Supervised learning	24.8 million/ 23.5 million/ 29 million
DINO-MM [105]	ViT-S/8	BigEarthNet-MM [86]	10	Global	Image level	SSL	22 million
Scheibenreif et al. [79]	Swin Transformer	SEN12MS [80]	10	Global	Image level, pixel level	CL	—
RingMo [87]	ViT/Swin Trans-former	2 million RS images	0.3–30	Six Continents	Image level, pixel level, region level, spatial-temporal	MIM	—
GeCO [57]	ResNet-50	Levir-KR [56]	0.8–16	Global	Image level, pixel level, region level	SSL	23.5 million
RS-BYOL [45]	BYOL	Sen12MS [80]	10–20	Global	Image level, pixel level	SSL	23.5 million
CSPT [104]	ViT-B	ImageNet-1K [18]	—	Global	Image level, region level	SSL	86 million
RVSA [100]	ViT	MillionAID [63], [64]	0.5–153	Global	Image level, pixel level, region level	MAE	100 million
SatMAE [16]	MAE-based Framework	fMoW *Sentinel-2* [15]	10, 20, 60	Global	Image level, pixel level	MAE	307 million
TOV [89]	TOV	TOV-NI, TOV-RS	—	Global	Image level, pixel level, region level	SSL	—
CMID [70]	Teacher-student Self-Distillation	MillionAID [63], [64]	Varied	Global	Image level, pixel level, region level, spatial-temporal	SSL	25.6 million/ 87.8 million
CACo [67]	ResNet-18/50	*Sentinel-2* imagery	10	Global	Image level, pixel level, spatial-temporal	SSL	11.7 million/ 23.5 million
IaI-SimCLR [77]	ResNet-18	SEN12MS	—	Global	Image level	CL	11.7 million
SSL4EO-L [83]	ResNet/ViT	ImageNet [18], MoCo [35], SimCLR [13]	30	Global	Pixel level	SSL	11.7 million/ 23.5 million/ 86 million
GFM [69]	Teacher-Student	GeoPile [69]	—	Global	Image level, pixel level	Continual pretraining	—
SatlasPretrain [7]	SatlasNet	GeoPile [69]	1, 10	Global	Image level, pixel level	Multitask learning	88 million
RingMo-Sense [119]	Multi-Branch	RS Spatiotemporal Dataset	—	Global	Pixel level	SSL	—
Scale-MAE [78]	ViT-Large	FMoW [15]	—	Global	Image level, pixel level	MAE	322.9 million
RingMo-lite [109]	CNN-Transformer	AID [115]	0.3–30	Global	Image level, pixel level, region level, spatial-temporal	FD-MIM	60% less than RingMo

(*Continued*)

TABLE 3. OVERVIEW OF RECENT FMS IN RS, CATEGORIZED BY ARCHITECTURE, MODEL NAME, PRETRAINING DATASET, RESOLUTION, GEOGRAPHIC COVERAGE, IMAGE ANALYSIS LEVELS, VISUAL ENCODER, PRETRAINING METHODS, AND THE NUMBER OF PARAMETERS. (*Continued*)

MODEL NAME	ARCHITECTURE	PRETRAINING DATASET	RESOLUTION (M)	GEOGRAPHIC COVERAGE	IMAGE ANALYSIS LEVELS	PRETRAIN METHODS	NO. OF PARAMS
DeCUR [102]	Multimodel SSL	SSL4EO-S12 [107], RGB-DEM/depth	Varied	Global	Image level, pixel level	SSL	23.5 million
Feng et al. [27]	MSFE+MMFH	Multimodal Dataset	Varied	Global	Image level, pixel level, region level, spatial-temporal	SSL	—
FG-MAE [108]	ViT	SSL4EO-S12 [107]	10	Global	Image level, pixel level	MAE	—
Prithvi [46]	ViT	Harmonized Landsat *Sentinel-2*	30	Contiguous United States	Pixel level	MAE	100 million
CROMA [28]	Multimodal Encoder	SSL4EO [107]	10	Areas surrounding human settlements	Image level, pixel level	CL, MAE	86 million
USat [44]	ViT	Satlas [7]	Varied	Global	Pixel level	MAE	—
Cross-Scale MAE [88]	ViT-B	fMoW [15]	—	Global	Image level, pixel level	MAE	86 million
U-BARN [26]	Unet+Transformer	*Sentinel-2* imagery	Varied	France	Image level, pixel level	SSL	—
EarthPT [82]	Transformer	*Sentinel-2* Imagery	10	United Kingdom	Image level	Autoregressive SSL	700 million
GeRSP [42]	Teacher-Student Network	ImageNet [18], MillionAID [63], [64]	0.5–153	Global	Image level, pixel level, region level	SSL, SL	—
SwiMDiff [91]	Dual-Branch	Sen12MS [80]	Varied	Global	Image level, spatial-temporal	SSL	11.7 million
SMLFR [22]	Generative ConvNet	GeoSense [22]	0.05–150	Multiple continents	Pixel level, region level	SSL	88 million/ 197 million
SpectralGPT [40]	3D GPT	*Sentinel-2* imagery	Varied	Global	Image level, pixel level, spatial-temporal	MAE	100 million/ 300 million/ 600 million
Presto [92]	MAE-based framework	Presto-21.5M [92]	10	Global	Crop-type segmentation	MAE	402,000
SatMAE++ [73]	SatMAE	fMoW [15]	Varied	Global	Image level	Multiscale pretraining	—
SAR-JEPA [58]	Joint-Embedding Predictive Architecture	100,000 SAR Images	Varied	Global	Image level	SSL	—
FoMo-Bench [8]	ViT	Multiple	Varied	Global	Image level, pixel level, region level	MAE	101 million/ 110 million
SkySense [32]	Factorized Multi-Modal Spatiotemporal Encoder	Multiple	Varied	Global	Image level, pixel level, region level, spatial-temporal	CL	2.06B
UPetu [24]	Multi-Modules	GeoSense [22]	—	Global	Image level, pixel level, spatial-temporal	SSL	0.65 million
msGFM [33]	Swin Transformer	GeoPile-2 [69]	0.1–153	Global	Image level, pixel level	MIM	89 million
DINO-MC [111]	DINO	SeCo-100K [66]	10–60	Global	Image level, spatial-temporal	SSL	—
OFA-Net [118]	OFA-Net	Multimodal Dataset	Varied	Global	Image level, Pixel level	MIM	—
MTP [99]	Shared Encoder Task-Specific Decoders	SAMRS [98]	Varied	Global	Image level, pixel level, region level, spatial-temporal	Multitask pretraining	More than 300 million
BFM [11]	ViT	MillionAID [63], [64]	0.5–153	Global	Pixel level, region level	MAE	86 million/ 605.26 million/ 1.36 billion/ 2.42 billion
MMEarth [72]	MP-MAE	Multimodal, geospatial data	—	Global	Image level, pixel level	MP-MAE	3.7 million to 650 million

(*Continued*)

TABLE 3. OVERVIEW OF RECENT FMS IN RS, CATEGORIZED BY ARCHITECTURE, MODEL NAME, PRETRAINING DATASET, RESOLUTION, GEOGRAPHIC COVERAGE, IMAGE ANALYSIS LEVELS, VISUAL ENCODER, PRETRAINING METHODS, AND THE NUMBER OF PARAMETERS. (*Continued*)

MODEL NAME	ARCHITECTURE	PRETRAINING DATASET	RESOLUTION (M)	GEOGRAPHIC COVERAGE	IMAGE ANALYSIS LEVELS	PRETRAIN METHODS	NO. OF PARAMS
CtxMIM [90]	ViT	WorldView-3 imagery	Varied	Asia	Image level, pixel level, region level	MIM	88 million
SARATR-X [54]	HiViT	SAR datasets	0.1–3	Global	Image level, region level	MIM	66 million
SoftCon [106]	Siamese Network with ResNet and ViT Backbones	SSL4EO-S12-ML [107]	—	Global	Image level, pixel level, spatial-temporal	Multilabel soft CL	23 million, 23 million, 86 million
LeMeViT [49]	Hierarchical ViT	MillionAID [63], [64]	—	—	Image level, pixel level, region level, spatial-temporal	Dual cross-attention with learnable meta token adaptation	8.33 million to 52.61 million
S2MAE [59]	3D Transformer-based MAE	fMoW-Sentinel [15], BigEarthNet [85]	—	Global	Image level, spatial-temporal	3D MAE	—
RS-DFM [110]	Multiplatform Inference Framework	AirCo-MultiTasks [110]	—	—	3D region level, pixel level	Generalized feature mapping with relative depth estimation	—
A2-MAE [122]	ViT-Large	Spatial-Temporal-Spectral Structured Dataset (STSSD)	0.8–30 m	Global	Image level, pixel level, spatial-temporal	Anchor-aware masking strategy and geographic encoding module	304 million
HyperSIGMA [95]	ViT based	HyperGlobal-450K [95]	30 m	Global	Image level, region level, anomaly detection, spatial-temporal	MAE	More than 1 billion
DOFA [117]	Dynamic OFA	Multiple	1–30	Global	Image level, pixel level	MIM	111 million/ 337 million

FD-MIM: feature-distilled masked image modeling.

This makes CNNs well suited for handling complex visual tasks in RS, such as image classification, segmentation, and object detection.

ResNets [36], a type of CNN, address the degradation problem in deep neural networks by introducing residual connections, which allow gradients to bypass certain layers, facilitating the training of very deep networks. This capability is particularly beneficial in RS, where deep models are often required to capture the intricate details and variations in satellite images. ResNet, as an example, is characterized by its residual blocks, which include shortcut connections that bypass one or more layers. The residual block can be described by the following equation:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

where $\mathbf{y}$ is the output, $\mathcal{F}$ represents the residual mapping to be learned, $\mathbf{x}$ is the input, and $\{W_i\}$ are the layer weights [36].

ResNet has various architectures, like ResNet-50, ResNet-101, and ResNet-152, with the number indicating the total layers. These networks have shown remarkable performance in various vision tasks due to their ability to train deeper networks without degradation. In RS, ResNets are widely used for image classification, object detection, and change detection tasks [30]. For example, ResNet-based models can classify different land cover types [31], [114], detect objects like buildings and vehicles [30], and monitor changes [31], [75] in the landscape over time by comparing temporal sequences of satellite images.

TRANSFORMERS AND VISION TRANSFORMERS

Transformers, adapted for computer vision as ViTs, model long-range dependencies through self-attention, making them effective for complex geospatial data. Figure 4 illustrates the architecture of ViT. ViTs treat images as sequences of patches, capturing global and local patterns, which is

useful for segmentation and change detection. The self-attention mechanism computes the following:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q (query), K (key), and V (value) are the input matrices, and d_k is the dimension of the key vectors [93].

By incorporating these methodologies, FMs for RS can leverage vast amounts of data, handle complex structures, and achieve state-of-the-art performance across various applications. These methodologies enable models to effectively address the unique challenges of RS, such as large image sizes, diverse data sources, and the need for high accuracy in environmental monitoring and analysis. In the following sections, we will explore specific applications of these methodologies in different RS tasks, analyze their performance, and discuss the datasets used to train and evaluate these models.

DATA AND TASKS

DATA

Datasets play a crucial role in RS, providing the foundation for training and evaluating models. High-quality datasets enable models to learn accurate representations of Earth's surface, improving their performance on various RS tasks. In Figure 2, we showcase some examples of the data used for training FMs and their downstream tasks. In this section, we provide an overview of commonly used datasets in Table 4 for RS, discussing their characteristics, applications, and relevance to FMs. These datasets, with their varying resolutions, categories, and geographic coverage, provide a rich resource for advancing RS research and applications. They facilitate the development of robust models capable of addressing diverse challenges in understanding and interpreting Earth's surface through RS technologies.

Datasets used in RS vary significantly in size, from hundreds of thousands of samples, as seen in RSD46-WHU [62], [116], to more than a million, as seen in MillionAID [63], [64]. Generally, larger datasets contribute to model generalization by encompassing diverse geographic areas, seasonal variations, and environmental conditions. Dataset resolutions also range from high (submeter), suitable for tasks requiring detailed spatial analysis, to moderate (10–60 m), as with SEN12MS [80] and SSL4EO-S12 [107], which support broader pattern recognition applications.

These datasets leverage various sensor types, including red, green, blue (RGB), multispectral, hyperspectral, and SAR. For instance, SEN12MS [80] integrates both SAR and multispectral imagery, enabling models to learn from distinct data modalities. This diversity in sensor types is critical for robust model development as each sensor type captures unique surface characteristics, supporting tasks that benefit from cross-modal information.

FMs, in particular, benefit from such large-scale multimodal datasets, which support self-supervised and supervised training approaches across tasks such as scene classification, segmentation, and object detection. For further insight, "Commonly Used Pretrain Dataset for Remote Sensing" includes detailed descriptions of each dataset's structure, unique characteristics, and application roles, enhancing the understanding of their impact on RS advancements.

TASKS

Different applications in RS address particular real-world challenges by leveraging the capabilities of FMs. These tasks include environmental monitoring, archaeology, agriculture, urban planning and development, and disaster management. To highlight the versatility of FMs in RS, we present Table 3, which categorizes models based on their applicability to various applications as well as the different image analysis methods used. This table serves as a quick reference for researchers to identify suitable models for their specific needs.

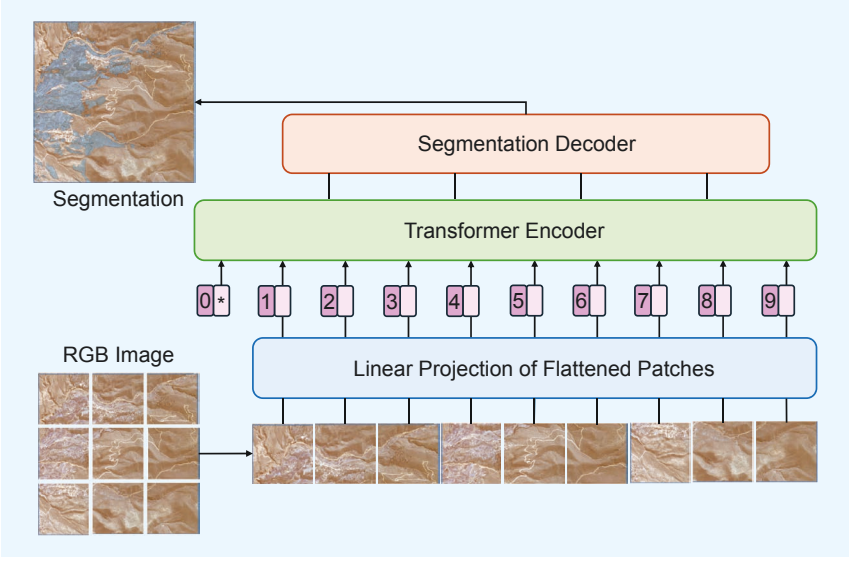

FIGURE 4. The Vision transformer architecture. (Source: RGB Image and Segmentation, copyright MAXAR 2024, provided through the NextView License Agreement.)

ENVIRONMENTAL MONITORING

According to Himeur et al. [39], environmental monitoring utilizes RS models to observe and track environmental changes, including deforestation, desertification, and pollution. These models play a crucial role in analyzing the effects of human activities and natural phenomena on the environment.

TABLE 4. THIS TABLE SUMMARIZES A SET OF COMMONLY USED PRETRAINED DATASETS FOR RS, INCLUDING DETAILS ON THE DATASET, SENSOR TYPE, GEOGRAPHIC COVERAGE, AND RELATED APPLICATIONS.

MONTH AND YEAR	DATASET	TITLE	PATCH SIZE	SIZE	RESOLUTION (M)	SENSOR	CATEGORIES	GEOGRAPHIC COVERAGE	IMAGE TYPE	APPLICATION
2017	RSD46-WHU [62], [116]	—	256 × 256	117,000	0.5–2	Google Earth, Tianditu	46	Global	RGB	Scene classification
Apr. 2018	fMoW [15]	Functional Map of the World	—	1,047,691	—	Digital Globe	63	207 of 247 countries	Multispectral	Scene classification, object detection
May 2019	DOTA [113]	DOTA: A Large-scale Dataset for Object Detection in Aerial Images	800 × 800 to 20,000 × 20,000	11,268	Various	Google Earth, *GF-2* satellite, and aerial images	18	Global	RGB	Object detection
June 2019	SEN12MS [80]	SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral *Sentinel-1/2* Imagery for Deep Learning and Data Fusion	256 × 256	541,986	10	*Sentinel-1*, *Sentinel-2*, MODIS land cover	33	Globally distributed	SAR/multispectral	Land cover classification, change detection
June 2019	BigEarthNet [85]	BigEarthNet: A Large-Scale Benchmark Archive For Remote Sensing Image Understanding	20 × 20 to 120 × 120	590,326	Various	*Sentinel-2*	43	Europe	Multispectral	Scene classification, object detection
June 2019	SeCo [66]	Seasonal Contrast: Unsupervised Pre-Training from Uncurated Remote Sensing Data	264 × 264	~1 M	10–60	*Sentinel-2*	—	Global	Multispectral	Seasonal change detection, land cover classification over seasons
March 2021	MillionAID [63], [64]	Million-AID	110–31,672	1,000,848	Various	Google Earth	51	Global	RGB	Scene classification
July 2021	Levir-KR [56]	Geographical Knowledge-driven Representation Learning for Remote Sensing Images	—	1,431,950	Various	*Gaofen-1*, *Gaofen-2*, *Gaofen-6*	8	Global	Multispectral	Change detection, scene classification
Apr. 2022	TOV-RS-Balanced [89]	TOV: The Original Vision Model for Optical Remote Sensing Image Understanding via Self-supervised Learning	600 × 600	500,000	1–20	Google Earth	31	Global	RGB	Scene classification, object detection, semantic segmentation
July 2022	SeasoNet [53]	SeasoNet: A Seasonal Scene Classification, Segmentation and Retrieval dataset for satellite Imagery over Germany	up to 120 × 120	1,759,830	10–60	*Sentinel-2*	33	Germany	Multispectral	Scene classification, scene segmentation
Nov. 2022	SSL4EO-S12 [107]	SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation	264x 264	3,012,948	10–60	*Sentinel-1*, *Sentinel-2*	—	Global	SAR/Multispectral	SSL
Oct. 2023	SAMRS [98]	SAMRS: Scaling-up Remote Sensing Segmentation Dataset with Segment Anything Model	600 × 600 to 1,024 × 1,024	105,090	Various	HRSC2016, DOTA-V2.0, DIOR, FAIR1M-2.0	37	Global	High-resolution	Semantic segmentation, instance segmentation, object detection

(*Continued*)

TABLE 4. THIS TABLE SUMMARIZES A SET OF COMMONLY USED PRETRAINED DATASETS FOR RS, INCLUDING DETAILS ON THE DATASET, SENSOR TYPE, GEOGRAPHIC COVERAGE, AND RELATED APPLICATIONS. *(Continued)*

MONTH AND YEAR	DATASET	TITLE	PATCH SIZE	SIZE	RESOLUTION (M)	SENSOR	CATEGORIES	GEOGRAPHIC COVERAGE	IMAGE TYPE	APPLICATION
June 2023	CACo [67]	Change-Aware Sampling and Contrastive Learning for Satellite Images	Variable	–	10	*Sentinel-2*	–	Urban and rural areas	Multispectral	Semantic segmentation, change detection, SSL
Oct. 2023	SatlasPretrain [7]	SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding	512 × 512	856,000	1–10 (*Sentinel-2*), 0.5–2 (NAIP)	*Sentinel-1*, *Sentinel-2*, Landsat, and NAIP	137	Global	Multispectral, high resolution	Land cover classification, segmentation, change detection
Oct. 2023	SSL4EO-L [83]	SSL4EO-L: Datasets and Foundation Models for Landsat Imagery	264 × 264	5,000,000	30	*Landsat 4–5* TM, *Landsat 7* ETM+, *Landsat 8–9* OLI/TIRS	–	Global	Multispectral	Cloud detection, land cover classification, semantic segmentation
July 2024	MMEarth [72]	MMEarth: Exploring Multi-Modal Pretext Tasks For Geospatial Representation Learning	128 × 128	1,200,000	10	*Sentinel-2*, *Sentinel-1*, Aster DEM	46	Global	Multispectral, SAR, climate	Land cover classification, semantic segmentation

AGRICULTURE

In agriculture, RS models are used to monitor crop health, estimate yields, and manage agricultural practices. According to Kamilaris and Prenafeta-Boldú [52], these models help optimize resource use and improve agricultural productivity.

ARCHAEOLOGY

In archaeology, RS models have been used to identify and analyze archaeological features and sites. According to Argyrou and Agapiou [5], these models help detect features such as ruins, artifacts, and ancient structures from satellite imagery, leveraging technologies like CNNs and ViTs to process high-resolution images and capture fine details. Mantovan and Nanni [68] also highlight the effectiveness of AI models, particularly CNNs, in locating challenging terrestrial archaeological sites and processing multispectral data.

URBAN PLANNING AND DEVELOPMENT

In urban planning and development, RS models are used to monitor and analyze urban expansion, infrastructure development, and land use changes. According to Jha et al. [48], these models play a critical role in managing urban growth, planning new developments, and assessing the impact of urbanization by providing essential data for smart city planning and sustainable development.

DISASTER MANAGEMENT

RS models play a crucial role in disaster management by providing timely information on affected areas. According to Abid et al. [1], these models are used to detect and assess damage from natural disasters like earthquakes, hurricanes, and floods, enabling rapid response and recovery efforts.

DISCUSSION

The rapid advancement in FMs for RS underscores their transformative potential across various applications. As the field continues to evolve, it is crucial to synthesize the findings, address technical challenges, understand the practical implications, and identify future research directions. In this section, we make a comprehensive analysis of these aspects, aiming to offer insights and guidance for future development and application of RS FMs.

SYNTHESIS OF FINDINGS

In our survey of FMs for RS, we identified significant advancements and trends that highlight the evolving capabilities and applications of these models. The performance metrics of various models across different downstream tasks, such as scene classification, semantic segmentation, object detection, and change detection, reveal the following key findings.

MODEL PERFORMANCE

In this section, we present the performance metrics of recent FMs in RS based on results reported in the original articles and papers. *All performance numbers*

mentioned here are sourced directly from the original studies to ensure accuracy and consistency in evaluating these models. These metrics provide insights into the models' effectiveness across tasks like semantic segmentation, object detection, and change detection, highlighting their strengths and limitations under different experimental setups.

IMAGE LEVEL

The performance of FMs on the BigEarthNet dataset [85] for classification tasks shows variations in accuracy, as presented in Table 5. Overall, msGFM [33] has the top performance of 92.90% [mean average precision (mAP)], followed closely by SkySense [32] with a performance of 92.09%. Other notable performers include DeCUR [102], which achieved an mAP of 89.70%, and DINO-MC [111], with an mAP of 88.75%. SeCo [66] also demonstrated strong performance with an mAP of 87.81%, while DINO-MM [105] reached an mAP of 87.10%. On the other hand, models like CACo [67] and FoMo-Bench [8] have an mAP of 74.98% and F1 score of 68.33%, respectively, showing competitiveness but room for improvement.

The high mAP scores of msGFM [33] and SkySense [32] highlight their efficiency in classification tasks, making them suitable for applications requiring high accuracy. Other FMs, such as DINO-MM [105] and DeCUR [102], also provide strong performance with the potential for further optimization. The variety in performance metrics underscores the evolving capabilities and specialization of FMs in handling complex classification tasks within datasets like BigEarthNet [85].

The classification advancements observed in RS models stem from sophisticated pretraining techniques that capture both spatial and spectral complexity across vast datasets. SkySense, for example, shows an average improvement of 2.76% over recent models by implementing multigranularity CL on a diverse dataset of 21.5 million optical and SAR sequences [32]. This approach enables SkySense to

COMMONLY USED PRETRAIN DATASET FOR REMOTE SENSING

The RSD46-WHU [62], [116] dataset, introduced in 2017, is sourced from Google Earth and Tianditu. It contains 117,000 images with a patch size of 256 pixels and spatial resolutions ranging from 0.5 to 2 m per pixel. Covering 46 categories globally, this dataset is primarily used for scene classification. Similarly, the Functional Map of the World (fMoW) [15], released in April 2018, comprises more than 1 million images from Digital Globe. Spanning 63 categories across 207 countries, it includes multispectral images used for both scene classification and object detection.

In May 2019, the DOTA [113] dataset was proposed, known for its large-scale aerial image object detection capabilities. It includes 11,268 images of various resolutions from Google Earth, the *GF-2* satellite, and aerial sources, covering 18 categories globally. Another significant dataset, SEN12MS [80], released in June 2019, contains 541,986 images from *Sentinel-1*, *Sentinel-2*, and MODIS Land Cover. With a patch size of 256×256 pixels, it supports land cover classification and change detection tasks.

BigEarthNet [85], also from June 2019, consists of 590,326 images with varying sizes from 20×20 to 120×120 pixels, sourced from *Sentinel-2*. It covers 43 categories across Europe and is used for scene classification and object detection. The SeCo [66] dataset, another June 2019 release, contains approximately 1 million images with a resolution of 2.65×2.65 km from *Sentinel-2*. It is designed for seasonal change detection and land cover classification over seasons.

The MillionAID dataset [63], [64], introduced in March 2021, includes more than 1 million images of various sizes from Google Earth. Covering 51 categories globally, it is used for scene classification. Levir-KR, released in July 2021, contains 1,431,950 images from the *Gaofen-1*, *Gaofen-2*, and *Gaofen-6* satellites, supporting change detection and scene classification applications.

SoundingEarth [S1], introduced in August 2021, comprises 50,545 images of 1,024-pixel size from Google Earth, combining RGB and audio data for RS. The TOV-RS-Balanced dataset [89] from April 2022 includes 500,000 images with a 600-pixel size from Google Earth, covering 31 categories globally, and is used for scene classification, object detection, and semantic segmentation.

SeasoNet [53], released in July 2022, features 1,759,830 images from *Sentinel-2* with patch sizes from 20 to 120 pixels, supporting seasonal scene classification, segmentation, and retrieval over Germany. Lastly, the SSL4EO-S12 dataset [107] from November 2022 contains more than 3 million images from *Sentinel-1* and *Sentinel-2*, with a patch size of 264×264 pixels. Since this dataset does not contain any labels, it is commonly used for SSL.

In recent years, additional datasets have further enriched the resources available for RS research. The SAMRS dataset [98], released in October 2023, offers a high-resolution collection of images sourced from datasets like HRSC2016 and FAIR1M-2.0, tailored for advanced segmentation tasks. With more than 105,000 images and resolutions up to 1,024×1,024 pixels, SAMRS supports semantic and instance segmentation as well as object detection, contributing to the development of scalable segmentation models for RS.

Focusing on change-aware learning, CACo [67], launched in June 2023, provides a variable patch-size dataset sourced from *Sentinel-2*. This dataset is optimized for change detection and CL, specifically addressing urban and rural landscapes. By prioritizing contrastive and self-supervised tasks, CACo aids in developing models that can adapt to changes in satellite imagery across various environments.

The SatlasPretrain [7] dataset, introduced in October 2023, is a large-scale collection with more than 856,000 images combining *Sentinel-2* and NAIP high-resolution sources. With multispectral and high-resolution imagery, SatlasPretrain supports applications such as land cover classification, segmentation, and change detection, further advancing research in high-resolution satellite image analysis.

The SSL4EO-L [83] dataset, released in October 2023, represents a vast resource with more than 5 million images from Landsat, designed for SSL in cloud detection and land cover classification. By focusing on multiyear Landsat imagery, SSL4EO-L enables robust training for applications that benefit from long-term temporal coverage and cloud-resilient classification.

Finally, MMEarth [72], introduced in July 2024, combines data from *Sentinel-1*, *Sentinel-2*, and Aster DEM, providing more than 1.2 million images for multimodal applications. This dataset supports land cover classification and semantic segmentation, enabling researchers to leverage multiple sensor types and climate data for improved geospatial representation learning.

These datasets, with their varying resolutions, categories, and geographic coverage, provide a rich resource for advancing RS research and applications. They facilitate the development of robust models capable of addressing diverse challenges in understanding and interpreting Earth's surface through RS technologies.

Reference

[S1] K. Heidler et al., "Self-supervised audiovisual representation learning for remote sensing data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 116, Feb. 2023, Art. no. 103130, doi: 10.1016/j.jag.2022.103130.

learn nuanced spatial and temporal relationships across modalities, enhancing generalization in varied environmental conditions. Such multigranular representation proves crucial in RS, where scene classification often depends on subtle spectral differences that simpler models may overlook. Likewise, HyperSIGMA [95], pretrained on the expansive HyperGlobal-450K hyperspectral dataset [95], leverages its sparse sampling attention mechanism to optimize spectral-spatial feature extraction in high-dimensional hyperspectral data. By selectively focusing on critical spectral bands and reducing redundancy, HyperSIGMA achieves high classification accuracy across hyperspectral scenes, a marked improvement over previous models that struggled with hyperspectral data complexity. These models highlight the importance of designing pretraining strategies that capture multimodal features and effectively utilize dataset diversity as these elements directly impact the robustness and accuracy of classification in RS applications.

TABLE 5. THIS TABLE PROVIDES AN OVERVIEW OF THE PERFORMANCE METRICS FOR VARIOUS MODELS APPLIED TO THE BIGEARTHNET DATASET [85] FOR IMAGE-LEVEL TASKS. THE PERFORMANCE IS MEASURED USING MEAN AVERAGE PRECISION (MAP) AND F1 SCORE.

DATASET	MODEL	PERFORMANCE (%)	METRICS
BigEarthNet [85]	SeCo [66]	87.81	mAP
	CMC-RSSR [84]	82.9	mAP
	DINO-MM [105]	87.1	mAP
	CACo [67]	74.98	mAP
	GFM [69]	86.3	mAP
	DINO-MC [111]	88.75	mAP
	CROMA [28]	86.46	mAP
	DeCUR [102]	89.7	mAP
	CtxMIM [90]	86.88	mAP
	FG-MAE [108]	78	mAP
	USat [44]	85.82	mAP
	FoMo-Bench [8]	69.33	F1 score
	SwiMDiff [91]	81.1	mAP
	SpectralGPT [40]	88.22	mAP
	SatMAE++ [73]	85.11	mAP
	msGFM [33]	**92.9**	mAP
	SkySense [32]	92.09	mAP
	MMEarth [72]	78.6	mAP
	Shallow CNN* [85]	**70.98**	**F1 score**

*Performance for the shallow CNN model is sourced from the original BigEarthNet [85] paper. Bold values indicate the best-performing model for the corresponding metric.

TABLE 6. THIS TABLE PROVIDES AN OVERVIEW OF THE PERFORMANCE METRICS FOR VARIOUS MODELS APPLIED TO THE ISPRS POTSDAM [43] DATASET FOR PIXEL-LEVEL TASKS. THE PERFORMANCE IS MEASURED USING MEAN INTERSECTION OVER UNION (MIOU) AND OVERALL ACCURACY (OA).

DATASET	MODEL	PERFORMANCE (%)	METRICS
ISPRS Potsdam	GeoKR [56]	70.48	mIoU
	RSP [96]	65.3	mIoU
	RingMo [87]	91.74	OA
	RVSA [100]	91.22	OA
	TOV [89]	60.34	mIoU
	CMID [70]	**87.04**	mIoU
	RingMo-lite [109]	90.96	OA
	Cross-Scale MAE [88]	76.17	mIoU
	SMLFR [22]	**91.82**	OA
	SkySense [32]	**93.99**	mF1
	UPetu [24]	83.17	mIoU
	BFM [11]	92.58	OA
	R-SegNet* [23]	**91.37**	**OA**

*Non-FM. Bold values indicate the best-performing model for the corresponding metric.

PIXEL LEVEL

For the segmentation tasks, we compared 12 FMs that have been tested on the International Society for Photogrammetry and Remote Sensing (ISPRS) Potsdam dataset. As shown in Table 6, SkySense [32] has the better performance out of all 12 models, with an mF1 score of 93.99%. CMID [70] stands out with the highest mean intersection over union (mIoU) of 87.04%, demonstrating its superior capability in accurately segmenting different regions within the dataset. For overall accuracy (OA) performance, BFM [11] has the highest OA score of 91.82%. Cross-Scale MAE [88], UPetu [24], and RSP [96] have mIoU scores of 76.17%, 83.17%, and 65.30%, respectively, showing competitive segmentation capabilities. GeoKR [56] reaches an mIoU of 70.48%, indicating robust segmentation performance but with room for improvement compared to CMID [70]. TOV scores the lowest mIoU at 60.34%, suggesting that it may struggle with finer segmentation tasks compared to the other models.

The performance metrics for the models applied to the ISPRS Potsdam dataset reveal significant variations in their effectiveness in segmentation tasks. SkySense [32] and CMID [70] emerge as top performers in mF1 score and mIoU, respectively, while SMLFR [22], RingMo [87], and RingMo-lite [109] demonstrate strong OA. These insights can guide the selection and optimization of models for specific RS applications, ensuring the best possible performance for the task at hand.

For the change detection tasks, we compared the performance of FMs on the OSCD and LEVIR-CD datasets (Table 7). The models were evaluated based on their F1 scores, which provide a balanced measure of precision and recall. As shown in the table, the performance varies significantly across different models and datasets.

SkySense [32] achieves the highest F1 score of 60.06% on the OSCD dataset, demonstrating its superior ability to accurately detect changes. GFM [69] follows with an F1 score of 59.82%, indicating strong performance in change detection tasks. SpectralGPT [40] also performs well with an F1 score of 54.29%. Other notable models include DINO-MC [111] with an F1 score of 52.71% and CACo [67] with an F1 score of 52.11%. SeCo [66] records the lowest

F1 score at 46.94%, suggesting that it may require further optimization to enhance its change detection capabilities.

In contrast, the LEVIR-CD dataset reveals higher performance metrics across the models. MTP [99] achieves the highest F1 score of 92.67%, and SkySense [32] follows closely with an F1 score of 92.58%, demonstrating their robust performance. SWiMDiff reaches a lower F1 score of 80.90% compared to its peers but still indicates effective performance in the LEVIR-CD [12] dataset.

REGION LEVEL

In Table 8, the performance of FMs on the DOTA, DIOR, and DIOR-R datasets for object detection is evaluated based on their mAP and average precision at 50% (AP50). On the DOTA dataset, RVSA [100] achieves the highest mAP of 81.24% in accurately detecting objects, followed by SMLFR [22] and RSP [96] with mAPs of 79.33% and 77.72%. CMID [70], GeRSP [42], and BFM [11] also demonstrate moderate performances with mAPs of 72.12%, 67.40%, and 58.69%. For the DIOR and DIOR-R datasets, MTP [99] and SkySense [32] are the top performers with an AP50 of 78% and an mAP of 78.73%, respectively, showcasing their superior object detection capabilities. These insights can guide the selection and optimization of models to ensure the best possible performance for specific RS applications.

INFLUENCE OF PRETRAINING METHODS

Various pretraining methods have a substantial impact on the performance of FMs in RS. Models pretrained using SSL techniques, such as CL and MAE, consistently exhibit superior performance compared to those pretrained with traditional supervised learning. For instance, SkySense, which uses a multigranularity CL approach, outperforms other models by approximately 3.6% in scene classification and object detection tasks [32]. Similarly, Seco, based on seasonal contrast learning, yields superior performance for land cover classification, improving metrics by up to 7% over ImageNet-pretrained models [66]. In handling multitemporal and multispectral data, models like SatMAE [16] and Scale-MAE [78], using masked autoencoding, achieve improvements in change detection, with SatMAE showing up to a 14% performance gain in land cover classification [16] and Scale-MAE offering a 1.7% mIoU improvement for segmentation across varied resolutions [78]. These findings highlight the critical role of innovative pretraining methods in maximizing the effectiveness of FMs and suggest that the continued exploration and refinement of these techniques are essential for advancing RS capabilities.

FMs like SatMAE, RingMo, A2-MAE, and ORBIT each demonstrate strong performance, but practical tradeoffs are essential to consider, especially for application-specific constraints [16], [[87], [[101], [[122]. SatMAE, based on a transformer architecture, effectively leverages temporal and multispectral embeddings to capture complex spatial-temporal patterns in satellite imagery. This strength, however, comes at the cost of significant computational

TABLE 7. THIS TABLE PROVIDES AN OVERVIEW OF THE F1 SCORE FOR VARIOUS MODELS APPLIED TO THE ONERA SATELLITE CHANGE DETECTION (OSCD) DATASET [10] AND THE LEVIR-CD DATASET [12] FOR SPATIOTEMPORAL DOWNSTREAM TASKS.

DATASET	MODEL	F1 SCORE
OSCD [10]	SeCo [66]	46.94
	MATTER [2]	49.48
	CACo [67]	52.11
	GFM [69]	59.82
	SWiMDiff [91]	49.6
	SpectralGPT [40]	54.29
	SkySense [32]	**60.06**
	DINO-MC [111]	52.71
	HyperSIGMA [95]	59.28
	MTP [99]	53.36
	CNNs* [10]	**89.66 (OA)**
LEVIR-CD [12]	RSP [96]	90.93
	RingMo [87]	91.86
	RIngMo-lite [109]	91.56
	SwiMDiff [91]	80.9
	SkySense [32]	92.58
	UPetu [24]	88.5
	STANet* [12]	**85.4**

*Performance for the models is sourced from the original dataset articles and papers. STANet is the Spatial-Temporal Attention Network. Bold values indicate the best-performing model for the corresponding metric.

TABLE 8. THIS TABLE PROVIDES AN OVERVIEW OF THE PERFORMANCE METRICS FOR VARIOUS MODELS APPLIED TO THE DOTA [20], [21], [113] DATASET, DIOR [55], AND DIOR-R [14] DATASET FOR REGION-LEVEL TASKS. THE PERFORMANCE IS MAINLY MEASURED USING MAP.

DATASET	MODEL	PERFORMANCE (%)	METRICS
DOTA	RSP [96]	77.72	mAP
	RVSA [100]	**81.24**	mAP
	TOV [89]	26.1	mAP50
	CMID [70]	72.12	mAP
	GeRSP [42]	67.4	mAP
	SMLFR [22]	79.33	mAP
	BFM [11]	58.69	mAP
	YOLOv2-D* [21]	**60.51**	**AP**
DIOR	RingMo [87]	75.8	mAP
	CSPT [104]	69.8	mAP
	RingMo-lite [109]	73.4	mAP
	GeRSP [42]	72.2	mAP
	MTP [99]	**78**	AP50
	Faster R-CNN* [55]	**74.05**	**mAP**
DIOR-R	RVSA [100]	71.05	mAP
	SMLFR [22]	72.33	mAP
	SkySense [32]	**78.73**	mAP
	MTP [99]	74.54	mAP
	BFM [11]	73.62	mAP
	AOPG* [14]	**64.41**	**mAP**

*Model performance is acquired from original dataset articles and papers. AOPG: Anchor-free Oriented Proposal Generator. Bold values indicate the best-performing model for the corresponding metric.

requirements, which may not be feasible for real-time monitoring applications in resource-constrained environments.

In contrast, RingMo provides a more lightweight vision transformer architecture, offering efficient model inference and a balance between performance and computational demands. This makes RingMo particularly suitable for rapid-inference tasks like disaster response monitoring, where real-time processing is critical [87]. A2-MAE introduces an anchor-aware masking strategy, optimizing spatial-temporal-spectral representations and allowing the effective integration of multisource data. This design enhances its adaptability to varied data resolutions and modalities, yet the model's complex encoding techniques add to its computational load, suggesting a fit for applications that require high accuracy over efficiency [122].

Finally, ORBIT, designed with 113 billion parameters, is exceptionally scalable, achieving high-throughput performance for Earth system predictability tasks. While it excels in large-scale predictive tasks, the model's considerable resource requirements limit its deployment to specialized high-performance computing environments [101]. These tradeoffs highlight the importance of selecting a model that aligns with specific operational goals, whether for maximizing accuracy or minimizing computational overhead.

Furthermore, recent studies comparing SSL approaches highlight the distinct advantages of generative methods like MAEs over contrastive methods for time-series data, especially when labeled data are limited [61]. Unlike contrastive approaches that emphasize distinguishing between similar and dissimilar pairs, generative methods such as MAE reconstruct data from masked segments, allowing them to capture complex underlying structures and relationships within the data. This reconstruction-based learning proves particularly advantageous for time-series and multispectral applications in RS, where temporal and spectral dependencies are essential. Consequently, MAE-based models can achieve stronger representations under sparse labeling conditions, positioning them as powerful tools for RS tasks that require nuanced temporal analysis.

PRACTICAL IMPLICATIONS

FMs offer transformative capabilities in RS by building upon established applications like multispectral and time-series data analysis. While these applications have traditionally relied on machine learning and DL, FMs reduce the need for labeled data and enable rapid adaptation to new tasks, providing robust solutions in areas previously limited by data constraints and task-specific architectures. Consequently, the advancements in FMs have significant practical implications across various areas.

- *Environmental monitoring:* Models like GASSL [6] and SatMAE [16] offer detailed assessments of environmental changes, aiding in conservation efforts and policy-making. These models excel in monitoring deforestation, desertification, and pollution levels, providing actionable insights for environmental management. By integrating multispectral and temporal data, these models can track changes over time, allowing for the early detection of environmental degradation and the formulation of timely interventions. This capability is particularly important for the sustainable management of natural resources as well as reducing the impacts of climate change.
- *Agriculture and forestry:* FMs such as EarthPT [82] and GeCo [57] deliver valuable insights into crop health, yield predictions, and land use management, optimizing agricultural practices and resource allocation. For instance, RSP [96], leveraging multispectral data, enhances precision agriculture by accurately monitoring crop conditions and predicting yields. These models can detect the early signs of crop stress, diseases, and pest infestations, enabling farmers to take proactive measures. Additionally, they aid in forestry management by providing detailed maps of forest cover, estimating biomass, and monitoring deforestation activities, thereby supporting conservation efforts and sustainable forestry practices.
- *Archaeology:* The use of FMs in archaeology revolutionizes the way archaeological features and sites are discovered, mapped, and analyzed. Models such as GeoKR [56], RingMo [87], etc. can process high-resolution satellite imagery and multispectral data to enhance the detection and mapping of archaeological features that might be difficult to discern with the naked eye. Others, like MATTER [2], can accomplish texture and material analysis to help identify various surfaces. They enable large-scale surveys, allowing archaeologists to identify potential sites of interest over vast areas efficiently. Although thorough exploration still requires on-site visits and excavations or other terrestrial investigations, these significantly improve the initial identification and mapping process. Additionally, these models can track changes over time, helping archaeologists monitor environmental and human impacts and providing crucial information for preservation and restoration. This enhances the efficiency and accuracy of surveys and opens new possibilities for discovering unknown sites.
- *Urban planning and development:* RS models like CMID [70] and SkySense [32] are pivotal for monitoring urban expansion, infrastructure development, and land use changes. These models facilitate sustainable urban growth and development planning by providing high-resolution data analysis and trend forecasting. They enable city planners to assess the impact of urbanization on natural habitats, optimize land use, and plan infrastructure projects more effectively.
- *Disaster management:* Models such as OFA-Net [118], DOFA [117], and Prithvi [46] are instrumental in flood mapping as well as fire detection. These models provide critical real-time data that help in identifying affected areas quickly, enabling timely and effective response measures. This capability supports emergency responders in prioritizing resource allocation and implementing evacuation plans, thereby reducing the impact of natu-

ral disasters. Additionally, these models assist in post-disaster recovery by assessing damage and monitoring the recovery process over time. By integrating various data sources, they enhance the ability to make informed decisions, coordinate response efforts, and plan for future disaster mitigation strategies.

The improvements in accuracy across the models discussed have profound implications for real-world RS applications. In deforestation monitoring, for instance, models like GFM achieve high pixel-level accuracy in semantic segmentation, showing up to a 4.5% improvement over baseline models, which enhances the precision of mapping forest cover changes, supporting conservation efforts [101]. Similarly, HyperSIGMA achieves an impressive 6.2% accuracy boost in hyperspectral vegetation monitoring, providing invaluable data for assessing forest health and biodiversity [95].

In urban planning, models like UPetu excel in infrastructure mapping by integrating multimodal data, such as optical and radar imagery, achieving more than 5% higher accuracy compared to single-modality models, which allows urban planners to make more informed land use decisions [24]. Additionally, RingMo enhances object detection accuracy by 3.7% over traditional supervised models, effectively identifying dense urban features critical for disaster management and urban infrastructure assessment [87].

Finally, ORBIT demonstrates exceptional scalability, processing large climate datasets with a scaling efficiency of up to 85%, which supports applications in long-term environmental monitoring, such as climate change prediction and seasonal forecasting. This scalability not only advances traditional RS workflows but also enables complex multitemporal analyses and predictive modeling, which were previously challenging with conventional methods [101].

While RS has long benefited from multispectral and temporal data, the adaptability, scalability, and efficiency of FMs unlock a new level of precision and accessibility in these applications. This advancement opens up opportunities to tackle complex and evolving challenges across domains—from environmental conservation to urban planning—that traditional models have struggled to address at scale.

FUTURE DIRECTIONS

Future research should prioritize several key areas as follows:

- *Efficient model development:* Exploring techniques such as model distillation, pruning, and quantization to reduce computational requirements without compromising performance is crucial. Additionally, developing scalable architectures that efficiently handle ultra-high-resolution images is essential. For instance, applying pruning techniques to models like SatMAE [16] could maintain performance while reducing computational load. Model adaptation techniques such as Low-Rank Adaptation (LoRA) [41] have emerged as effective methods for fine-tuning large-scale models with minimal computational overhead. By decomposing weight updates into low-rank matrices, LoRA [41] enables efficient adaptation without the need to modify the entire set of model parameters, making it suitable for resource-constrained environments or when frequent retraining is required. Incorporating methods like LoRA [41] can further enhance the applicability of FMs across diverse tasks and domains.
- *Multimodal data integration:* Enhancing methods for integrating and processing multimodal data (e.g., combining optical and radar imagery) will provide more comprehensive insights. Research on advanced SSL techniques capable of leveraging multimodal data is necessary. The OFA-Net [118] framework, which integrates multimodel data, serves as a promising direction for future models to emulate and improve upon.
- *Interdisciplinary collaboration:* Promoting collaboration among RS experts, AI researchers, and domain specialists can address complex challenges and drive innovation. For example, partnerships between AI researchers and environmental scientists can refine models like GASSL [6] for better environmental monitoring and conservation efforts.

Looking ahead, the consistent success of SSL methods in FMs marks an exciting frontier for future research. These models' ability to learn from unlabeled data and adapt to diverse RS tasks with minimal fine-tuning suggests that advancements in unsupervised learning techniques could greatly reduce reliance on large labeled datasets, which remain a significant bottleneck in many RS applications. However, as these models grow in size and complexity, balancing computational demands with the need for efficiency will become increasingly crucial. Future work may focus on developing more resource-efficient versions of FMs that maintain high performance, particularly for deployment in real-time monitoring systems or environments with limited computational resources.

LIMITATIONS

This survey has several limitations as follows:

- *Scope and coverage:* The review focuses on FMs released between June 2021 and June 2024. While the scope of this review is extensive and covers many significant developments, it is not exhaustive. Some recent advancements and innovations in the field may not be included due to their release timing or the lack of sufficient evaluation metrics at the time of writing. Consequently, certain cutting-edge models that have emerged in the latter part of this period or that have not yet been thoroughly evaluated might be omitted. This limitation underscores the need for readers to seek out the most current research and updates beyond the scope of this survey. Additionally, while FMs have been empirically tested on a specific set of downstream applications, their robust architectures and general-purpose training paradigms, such as

convolutional networks (e.g., ResNet) and ViTs, indicate their potential to perform well across a much broader range of tasks. The limited testing observed in current literature should not be seen as a constraint on their applicability but rather as an indication of the focus of existing research efforts. Given their design, these models are expected to generalize effectively to a wide variety of RS tasks, even beyond those explicitly tested. Future work should aim to explore and validate their performance across more diverse applications to unlock their full potential.

- *Evolving field:* The field of AI and RS is rapidly evolving, with continuous advancements and breakthroughs occurring at a fast pace. This dynamic nature necessitates ongoing reviews and updates to ensure the relevance and comprehensiveness of the survey. New techniques, methodologies, and models are constantly being developed, which can significantly impact the state of the art. Therefore, it is essential to recognize that this survey represents a snapshot in time and that continuous monitoring of the literature is required to capture the latest advancements and emerging trends. This approach will help maintain an up-to-date understanding of the field and incorporate new findings as they become available.

CONCLUSION

In this comprehensive survey, we have reviewed the recent advancements in FMs for RS. We categorized these models based on their pretraining methods, image analysis techniques, and applications across different areas, highlighting their unique methodologies and capabilities.

Our analysis covered various advanced techniques, including SSL, ViTs, and ResNets. These models have significantly improved performance on different image perception levels, like the region level, pixel level, and image level, as well as in applications like environmental monitoring, digital archaeology, agriculture, urban planning, and disaster management.

While significant progress has been made, several challenges persist, such as the need for more diverse and high-quality datasets, high computational requirements, and difficulties for different applications. Addressing these challenges will require further research and collaboration across disciplines.

In summary, this survey provides a detailed overview of the current state of FMs in RS, offering valuable insights and identifying future research directions. We recommend continued efforts in developing efficient model architectures, enhancing multimodal data integration, and expanding dataset diversity to fully realize the potential of these models in RS.

ACKNOWLEDGMENT

This work was supported by NSF 2419793, HAA-293452-23, Vanderbilt Seeding Success Grant, Vanderbilt Discovery Grant, and VISR Seed Grant. We extend gratitude to NVIDIA for their support by means of the NVIDIA hardware grant. This work was also supported by NSF NAIRR Pilot Award NAIRR240055. This manuscript has been co-authored by ORNL, operated by UT-Battelle, LLC under Contract DE-AC05-00OR22725 with the U.S. Department of Energy.

AUTHOR INFORMATION

Siqi Lu (vickie0647@gmail.com) received her B.S. degree in electrical engineering from the University of Illinois, Urbana-Champaign in 2023. She is currently a second-year master's student in the Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN 37235 USA, supervised by Dr. Yuankai Huo and Dr. Mitchell M. Wilkes. Her research interests include deep learning, medical image analysis, and software engineering. She is a Student Member of IEEE.

Junlin Guo (junlin.guo@vanderbilt.edu) received his B.S. degree in telecommunication engineering from Northeastern University in 2017 and his M.S. degree from the Department of Electrical and Computer Engineering, Vanderbilt University in 2020, focusing on functional MRI brain activation study. He is currently working toward his Ph.D. degree in the Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN 37235 USA. His research interests include medical image analysis, deep learning, computer vision, and brain study.

James R. Zimmer-Dauphinee (james.r.zimmer-dauphinee@vanderbilt.edu) received his B.A. degree in anthropology and his B.S. degree in mathematics from Georgia Southern University in 2011, his M.A. degree in anthropology from the University of Arkansas in 2014, and his Ph.D. degree in anthropology from Vanderbilt University in 2023. He is currently a postdoctoral fellow in the Spatial Analysis Research Laboratory, Vanderbilt University, Nashville, TN 37235 USA, funded by the GeoPACHA 2.0 Grant from the National Endowment for the Humanities. His research interests include developing deep learning models for large-scale autonomous archaeological satellite imagery surveys, geophysical methods, and spatial modeling to understand the impact of colonization on indigenous peoples.

Jordan M. Nieusma (jordan.m.nieusma@vanderbilt.edu) received her B.A. degree in English with a French minor from Haverford College and her M.S. degree in data science at Vanderbilt University in 2024. She is currently a research assistant in the Spatial Analysis Research Laboratory, Data Science Institute, Vanderbilt University, Nashville, TN 37235 USA.

Xiao Wang (wangx2@ornl.gov) received his B.S. degrees in mathematics and computer science from Saint John's University, MN, in 2012, his M.S. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, in 2016, and his Ph.D. degree in electrical and computer engineering from Purdue University in 2017. He pursued postdoctoral research at Harvard Medical School and Boston Children's Hospital until 2021. He is currently

a research staff scientist at Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA. His research interests include applying machine learning, medical physics, image processing, and high-performance computing to various imaging problems, including computerized tomography (CT) reconstruction, electron tomography imaging, and MRI. He was the 2022 AAPM Truth CT Reconstruction Challenge winner and a 2017 ACM Gordon Bell Prize finalist. He is a Senior Member of IEEE.

Parker VanValkenburgh (parker_vanvalkenburgh@brown.edu) received his Ph.D. from Harvard University. He is currently an associate professor of anthropology and interim director of Latin American and Caribbean studies at the Department of Anthropology, Brown University, Providence, RI 02912 USA. His research focuses on the impacts of colonialism and imperialism on Indigenous people and environments in the Peruvian Andes. He utilizes diverse materials and digital methodologies, including geographic information systems, to understand the transformation of relationships in imperial histories. He codirects the Paisajes Arqueológicos de Chachapoyas (PACha) project and the Geospatial Platform for Andean Culture, History, and Archaeology (GeoPACHA).

Steven A. Wernke (s.wernke@vanderbilt.edu) is an associate professor and chair of anthropology at the Department of Anthropology, Vanderbilt University, Nashville, TN 37235 USA, director of the Spatial Analysis Research Laboratory, and director of the Vanderbilt Institute for Spatial Research. He is an archaeologist and historical anthropologist of the Andean region of South America. His research takes place at the intersection of several disciplines: archaeology and history, pre-Hispanic and colonial studies, anthropology, and cultural geography. His research interests include center on the lived experiences of indigenous communities across the Spanish invasion of the Andes—especially how new kinds of communities, landscapes, and religious practices emerged out of successive attempts by the Inkas and the Spanish to subordinate and remake Andean societies in their own self-image. Methodologically, his work brings together analyses of archaeological and documentary datasets in geospatial frameworks.

Yuankai Huo (yuankai.huo@vanderbilt.edu) received his B.S. degree in electrical engineering from the Nanjing University of Posts and Telecommunications (NJUPT) in 2008 and his master's degree in electrical engineering from Southeast University in 2011. After graduation, he worked at Columbia University and the New York State Psychiatric Institute as a staff engineer and research officer from 2011 to 2014. He received his master's degree in computer science from Columbia University in 2014 and his Ph.D. degree in electrical engineering from Vanderbilt University in 2018. Then, he worked as a research assistant professor at Vanderbilt University and later, as a senior research scientist at PAII Labs. Since 2020, he has been a faculty member at the Department of Electrical Engineering and Computer Science and Data Science Institute, Vanderbilt University. He is currently an assistant professor in the Department of Computer Science, Vanderbilt University, Nashville, TN 37235 USA. He is a Senior Member of IEEE.

REFERENCES

[1] S. K. Abid et al., "Toward an integrated disaster management approach: How artificial intelligence can boost disaster management," *Sustainability*, vol. 13, no. 22, 2021, Art. no. 12560, doi: 10.3390/su132212560.

[2] P. Akiva, M. Purri, and M. Leotta, "Self-supervised material and texture representation learning for remote sensing tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8193–8205, doi: 10.1109/CVPR52688.2022.00803.

[3] A. A. Aleissaee et al., "Transformers in remote sensing: A survey,"*Remote Sens.*, vol. 15, no. 7, 2022, Art. no. 1860, doi: 10.3390/rs15071860.

[4] "English: Mono, Multi and hyperspectral cube and corresponding spectral signatures," Arbeck Systems, Sandton, South Africa, Mar. 2013. [Online]. Available: https://commons.wikimedia.org/wiki/File:Mono,_Multi_and_Hyperspectral_Cube_and_corresponding_Spectral_Signatures.svg

[5] A. Argyrou and A. Agapiou, "A review of artificial intelligence and remote sensing for archaeological research," *Remote Sens.*, vol. 14, no. 23, 2022, Art. no. 6000, doi: 10.3390/rs14236000.

[6] K. Ayush et al., "Geography-aware self-supervised learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10,161–10,170, doi: 10.1109/ICCV48922.2021.01002.

[7] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "SatlasPretrain: A large-scale dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 16,726–16,736, doi: 10.1109/ICCV51070.2023.01538.

[8] N. Ioannis Bountos, A. Ouaknine, and D. Rolnick, "FoMo-Bench: A multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models," 2024, *arXiv:2312.10114*.

[9] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9630–9640, doi: 10.1109/ICCV48922.2021.00951.

[10] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "OSCD - Onera satellite change detection," IEEE DataPort, Piscataway, NJ, USA, 2019. [Online]. Available: https://ieee-dataport.org/open-access/oscd-onera-satellite-change-detection

[11] K. Cha, J. Seo, and T. Lee, "A billion-scale foundation model for remote sensing images," 2024, *arXiv:2304.05215*.

[12] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662, doi: 10.3390/rs12101662.

[13] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations,"in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607, doi: 10.5555/3524938.3525087.

[14] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, Jun. 2022, doi: 10.1109/TGRS.2022.3183022.

[15] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6172–6180.

[16] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. Glasgow, Scotland: Curran Associates, 2022, vol. 35, pp. 197–211, doi: 10.5555/3600270.3600285.

[17] R. G. Congalton, "Remote sensing: An overview," *GISci. Remote Sens.*, vol. 47, no. 4, pp. 443–459, 2010, doi: 10.2747/1548-1603.47.4.443.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[19] P. Dias et al., "An agenda for multimodal foundation models for earth observation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 1237–1240, doi: 10.1109/IGARSS52108.2023.10282966.

[20] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, doi: 10.1109/CVPR.2019.00296.

[21] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022, doi: 10.1109/TPAMI.2021.3117983.

[22] Z. Dong, Y. Gu, and T. Liu, "Generative convNet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, Jan. 2024, doi: 10.1109/TGRS.2023.3348479.

[23] H. Zhu et al., "Deep convolutional encoder–decoder networks based on ensemble learning for semantic segmentation of high-resolution aerial imagery," *CCF Trans. High Perform. Comput.*, vol. 6, no. 4, pp. 408–424, Aug. 2024, doi: 10.1007/s42514-024-00184-0.

[24] Z. Dong, Y. Gu, and T. Liu, "UPetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, Mar. 2024, doi: 10.1109/TGRS.2024.3382734.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2021, *arXiv: 2010.11929*.

[26] I. Dumeur, S. Valero, and J. Inglada, "Self-supervised spatio-temporal representation learning of satellite image time series," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 4350–4367, Jan. 2024, doi: 10.1109/JSTARS.2024.3358066.

[27] Y. Feng et al., "A self-supervised cross-modal remote sensing foundation model with multi-domain representation and cross-domain fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2023, pp. 2239–2242, doi: 10.1109/IGARSS52108.2023.10282433.

[28] A. Fuller, K. Millard, and J. R. Green, "CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders," in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 5506–5538, doi: 10.5555/3666122.3666363.

[29] J.-B. Grill et al., "Bootstrap your own latent a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21,271–21,284, doi: 10.5555/3495724.3497510.

[30] S. Gui, S. Song, R. Qin, and Y. Tang, "Remote sensing object detection in the deep learning era—A review," *Remote Sens.*, vol. 16, no. 2, 2024, Art. no. 327, doi: 10.3390/rs16020327.

[31] H. Zhu et al., "A spatial-channel progressive fusion ResNet for remote sensing classification," *Inform. Fusion*, vol. 70, pp. 72–87, Jun. 2021, doi: 10.1016/j.inffus.2020.12.008.

[32] X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27,662–27,673, doi: 10.1109/CVPR52733.2024.02613.

[33] B. Han, S. Zhang, X. Shi, and M. Reichstein, "Bridging remote sensors with multisensor geospatial foundation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27,852–27,862, doi: 10.1109/CVPR52733.2024.02631.

[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 15,979–15,988, doi: 10.1109/CVPR52688.2022.01553.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9726–9735, doi: 10.1109/CVPR42600.2020.00975.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[37] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017, doi: 10.1109/MGRS.2017.2762307.

[38] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," *Int. J. Mach. Learn. Cybern.*, early access, Nov. 2024, doi: 10.1007/s13042-024-02443-6.

[39] Y. Himeur, B. Rimal, T. Abhishek, and A. Abbes, "Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives," *Inf. Fusion*, vols. 86–87, pp. 44–75, Oct. 2022, doi: 10.1016/j.inffus.2022.06.003.

[40] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 1–18, Aug. 2024, doi: 10.1109/TPAMI.2024.3362475.

[41] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[42] Z. Huang, M. Zhang, Y. Gong, Q. Liu, and Y. Wang, "Generic knowledge boosted pretraining for remote sensing images,"*IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, Jan. 2024, doi: 10.1109/TGRS.2024.3354031.

[43] "2D semantic labeling contest – Potsdam," International Society for Photogrammetry and Remote Sensing (ISPRS), Baton Rouge, LA, USA, 2024. Accessed: Jul. 8, 2024. [Online]. Available: https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

[44] J. Irvin et al., "USat: A Unified Self-Supervised Encoder for Multi-Sensor Satellite Imagery," 2023, *arXiv:2312.02199*.

[45] P. Jain, B. Schoen-Phelan, and R. Ross, "Self-supervised learning for invariant representations from multi-spectral and SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7797–7808, Sep. 2022, doi: 10.1109/JSTARS.2022.3204888.

[46] J. Jakubik et al., "Foundation models for generalist geospatial artificial intelligence," 2023, *arXiv:2310.18660*.

[47] N. Jean, S. Wang, A. Samar, G. Azzari, D. Lobell, and S. Ermon, "Tile2Vec: Unsupervised representation learning for spatially distributed data," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 3967–3974.

[48] A. K. Jha, A. Ghimire, S. Thapa, A. M. Jha, and R. Raj, "A review of AI for urban planning: Towards building sustainable smart cities," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, 2021, pp. 937–944, doi: 10.1109/ICICT50816.2021.9358548.

[49] W. Jiang, J. Zhang, D. Wang, Q. Zhang, Z. Wang, and B. Du, "LeMeViT: Efficient vision transformer with learnable meta tokens for remote sensing image interpretation," 2024, *arXiv:2405.09789*.

[50] L. Jiao et al., "Brain-inspired remote sensing foundation models and open problems: A comprehensive survey," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10,084–10,120, Sep. 2023, doi: 10.1109/JSTARS.2023.3316302.

[51] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2021, doi: 10.1109/TPAMI.2020.2992393.

[52] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agriculture*, vol. 147, pp. 70–90, Apr. 2018, doi: 10.1016/j.compag.2018.02.016.

[53] D. Koßmann, V. Brack, and T. Wilhelm, "SeasoNet: A seasonal scene classification, segmentation and retrieval dataset for satellite imagery over Germany," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 243–246, doi: 10.1109/IGARSS46834.2022.9884079.

[54] Weijie, L. W. Yang, Y. Hou, Li Liu, Y. Liu, and X. Li, "SARATR-X: A foundation model for synthetic aperture radar images target recognition," 2024, *arXiv:2405.09365*.

[55] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020, doi: 10.1016/j.isprsjprs.2019.11.023.

[56] W. Li, K. Chen, H. Chen, and Z. Shi, "Geographical knowledge-driven representation learning for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022, doi: 10.1109/TGRS.2021.3115569.

[57] W. Li, K. Chen, and Z. Shi, "Geographical supervision correction for remote sensing representation learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–20, Aug. 2022, doi: 10.1109/TGRS.2022.3202499.

[58] W. Li et al., "Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture," *ISPRS J. Photogrammetry Remote Sens.*., vol. 218, pp. 326–338, Dec. 2024, doi: 10.1016/j.isprsjprs.2024.09.013.

[59] X. Li, D. Hong, and J. Chanussot, "S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27,696–27,705, doi: 10.1109/CVPR52733.2024.02616.

[60] Xiang Li, C. Wen, Yuan Hu, Z. Yuan, and X. Xiang Zhu, "Vision-language models in remote sensing: Current progress and future trends," *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 32–66, Jun. 2024, doi: 10.1109/MGRS.2024.3383473.

[61] Z. Liu, A. Alavi, M. Li, and X. Zhang, "Self-supervised learning for time series: Contrastive or generative?" 2024, *arXiv:2403.09809*.

[62] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017, doi: 10.1109/TGRS.2016.2645610.

[63] Yang Long, G.-S. Xia et al., "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, Apr. 2021, doi: 10.1109/JSTARS.2021.3070368.

[64] Y. Long, G.-S. Xia, L. Zhang, G. Cheng, and D. Li, "Aerial scene parsing: From tile-level scene classification to pixel-wise semantic labeling," 2022, *arXiv:2201.01953*.

[65] Y. Ma, S. Chen, S. Ermon, and D. B. Lobell, "Transfer learning in environmental remote sensing," *Remote Sens. Environ.*, vol. 301, Feb. 2024, Art. no. 113924, doi: 10.1016/j.rse.2023.113924.

[66] O. Mañas, A. Lacoste, X. Giró-I Nieto, D. Vazquez, and P. Rodríguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9394–9403, doi: 10.1109/ICCV48922.2021.00928.

[67] U. Mall, Bharath Hariharan and K. Bala, "Change-aware sampling and contrastive learning for satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5261–5270, doi: 10.1109/CVPR52729.2023.00509.

[68] L. Mantovan and L. Nanni, "The computerization of archaeology: Survey on artificial intelligence techniques," *SN Comput. Sci.*, vol. 1, no. 5, Aug. 2020, Art. no. 267, doi: 10.1007/s42979-020-00286-w.

[69] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 16,760–16,770, doi: 10.1109/ICCV51070.2023.01541.

[70] D. Muhtar, X. Zhang, P. Xiao, Z. Li, and F. Gu, "CMID: A unified self-supervised learning framework for remote sensing image understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, Apr. 2023, doi: 10.1109/TGRS.2023.3268232.

[71] R. Navalgund, V. Jayaraman, and P. Roy, "Remote sensing applications: An overview," *Current Sci.*, vol. 93, no. 12, pp. 1747–1766, Dec. 2007.

[72] V. Nedungadi, A. Kariryaa, S. Oehmcke, S. Belongie, C. Igel, and N. Lang, "*MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning,*" in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 164–182, doi: 10.1007/978-3-031-73039-9_10.

[73] M. Noman, M. Naseer, H. Cholakkal, R. M. Anwar, S. Khan, and F. S. Khan, "*Rethinking transformers pre-training for multi-spectral satellite imagery,*"in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27,811–27,819, doi: 10.1109/CVPR52733.2024.02627.

[74] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.

[75] E. Jonasova Parelius, "A review of deep-learning methods for change detection in multispectral remote sensing images," *Remote Sens.*, vol. 15, no. 8, 2023, Art. no. 2092, doi: 10.3390/rs15082092.

[76] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021, doi: 10.1109/TGRS.2020.3011913.

[77] J. Prexl and M. Schmitt, "Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2023, pp. 2136–2144, doi: 10.1109/CVPRW59228.2023.00207.

[78] C. J. Reed et al., "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4065–4076, doi: 10.1109/ICCV51070.2023.00378.

[79] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2022, pp. 1421–1430, doi: 10.1109/CVPRW56347.2022.00148.

[80] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS – A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," 2019, *arXiv:1906.07789*.

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[82] M. J. Smith, L. Fleming, and J. E. Geach, "EarthPT: A Time series foundation model for earth observation," 2024. [Online]. Avaialble: https://arxiv.org/abs/2309.07207

[83] A. J. Stewart et al., "SSL4EO-L: Datasets and foundation models for landsat imagery," 2023, *arXiv:2306.09424*.

[84] V. Stojnic and V. Risojevic, "Self-supervised learning of remote sensing scene representations using contrastive multiview coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2021, pp. 1182–1191, doi: 10.1109/CVPRW53098.2021.00129.

[85] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Piscataway, NJ, USA: IEEE, Jul. 2019, pp. 5901–5904, doi: 10.1109/IGARSS.2019.8900532.

[86] G. Sumbul et al., "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021, doi: 10.1109/MGRS.2021.3089174.

[87] X. Sun et al., "RingMO: A remote sensing foundation model with masked image modeling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–22, 2023, doi: 10.1109/TGRS.2022.3194732.

[88] M. Tang, A. Cozma, K. Georgiou, and H. Qi, "Cross-scale MAE: A tale of multi-scale exploitation in remote sensing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 1–13.

[89] C. Tao, J. Qi, G. Zhang, Q. Zhu, W. Lu, and H. Li, "TOV: The original vision model for optical remote sensing image understanding via self-supervised learning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4916–4930, Apr. 2023, doi: 10.1109/JSTARS.2023.3271312.

[90] M. Zhang, Q. Liu, and Y. Wang, "CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding," 2024, *arXiv:2310.00022*.

[91] J. Tian, J. Lei, J. Zhang, W. Xie, and Y. Li, "SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image,"*IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, Feb. 2024, doi: 10.1109/TGRS.2024.3371481.

[92] G. Tseng, R. Cartuyvels, I. Zvonkov, M. Purohit, D. Rolnick, and H. Kerner, "Lightweight, pre-trained transformers for remote sensing timeseries," 2024, *arXiv:2304.14065*.

[93] A. Vaswani et al., "Attention is all you need," 2023, *arXiv:1706.03762*.

[94] C. Wang et al., "A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode," *Geosci. Data J.*, vol. 6, no. 2, pp. 105–115, 2019, doi: 10.1002/gdj3.73.

[95] D. Wang et al., "HyperSIGMA: Hyperspectral intelligence comprehension foundation model," 2024, *arXiv:2406.11519*.

[96] D. Wang, J. Zhang, B. Du, G.-S. Xia, and D. Tao, "An empirical study of remote sensing pretraining," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–20, 2023, doi: 10.1109/TGRS.2022.3176603.

[97] X. X. Zhu et al., "On the foundations of earth and climate foundation models," 2024, *arXiv:2405.04285*.

[98] D. Wang et al., "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model,"in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 8815–8827, doi: 10.5555/3666122.3666507

[99] D. Wang et al., "MTP: Advancing remote sensing foundation model via multi-task pretraining,"*IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 11,632-11,654, Jun. 2024, doi: 10.1109/JSTARS.2024.3408154.

[100] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023, doi: 10.1109/TGRS.2022.3222818.

[101] X. Wang et al., "ORBIT: Oak ridge base foundation model for earth system predictability," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, 2024, pp. 1–11, doi: 10.1109/SC41406.2024.00007.

[102] Y. Wang, C. M. Albrecht, N. A. A. Braham, C. Liu, Z. Xiong, and X. X. Zhu, DeCUR: Decoupling common & unique representations for multimodal self-supervision, 2023, *arXiv:2309.05300*.

[103] Y. Wang, C. M. Albrecht, N. A. Ali Braham, L. Mou, and X. Xiang Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022, doi: 10.1109/MGRS.2022.3198244.

[104] T. Zhang et al., "Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain," *Remote Sens.*, vol. 14, no. 22, 2022, Art. no. 5675, doi: 10.3390/rs14225675.

[105] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Self-supervised vision transformers for joint SAR-optical representation learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2022, pp. 139–142, doi: 10.1109/IGARSS46834.2022.9883983.

[106] Y. Wang, C. M. Albrecht, and X. X. Zhu, "Multi-label guided soft contrastive learning for efficient earth observation pretraining," 2024, *arXiv:2405.20462*.

[107] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "SSL4eo-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation,"*IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 3, pp. 98–106, Sep. 2023, doi: 10.1109/MGRS.2023.3281651.

[108] Y. Wang, H. H. Hernández, C. M. Albrecht, and X. X. Zhu, "Feature guided masked autoencoder for self-supervised learning in remote sensing,"*IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 321–336, 2025, doi: 10.1109/JSTARS.2024.3493237.

[109] Y. Wang et al., "RingMO-Lite: A remote sensing multi-task lightweight network with CNN-transformer hybrid framework," 2023, *arXiv:2309.09003*.

[110] Z. Wang et al., "RS-DFM: A remote sensing distributed foundation model for diverse downstream tasks," 2024, *arXiv:2406.07032*.

[111] X. Wanyan, S. Seneviratne, S. Shen, and M. Kirley, "Extending global-local view alignment for self-supervised learning with remote sensing imagery," 2024, *arXiv:2303.06670*.

[112] M. Wulder et al., "Current status of Landsat program, science, and applications," *Remote Sens. Environ.*, vol. 225, pp. 127–147, Mar. 2019, doi: 10.1016/j.rse.2019.02.015.

[113] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 3974–3983.

[114] 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 Sens.*, vol. 14, no. 19, 2022, Art. no. 4883, doi: 10.3390/rs14194883.

[115] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017, doi: 10.1109/TGRS.2017.2685945.

[116] Z. Xiao, Y. Long, D. Li, C. Wei, G. Tang, and J. Liu, "High-resolution remote sensing image retrieval based on CNNs from a dimensional perspective," *Remote Sens.*, vol. 9, no. 7, 2017, Art. no. 725, doi: 10.3390/rs9070725.

[117] Z. Xiong et al., "Neural plasticity-inspired multimodal foundation model for earth observation," 2024, *arXiv:2403.15356*.

[118] Z. Xiong, Y. Wang, F. Zhang, and X. X. Zhu, "One for all: Toward unified foundation models for earth vision," 2024, *arXiv:2401.07527*.

[119] F. Yao et al., "RingMO-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–21, Sep. 2023, doi: 10.1109/TGRS.2023.3316166.

[120] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 270–294, Jun. 2022, doi: 10.1109/MGRS.2022.3145854.

[121] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016, doi: 10.1109/MGRS.2016.2540798.

[122] L. Zhang et al., "A^2-MAE: A spatial-temporal-spectral unified remote sensing pre-training method based on anchor-aware masked autoencoder," 2024, *arXiv:2406.08079*.

GRS