remote sensing

MDPI

Review

Advances on Multimodal Remote Sensing Foundation Models for Earth Observation Downstream Tasks: A Survey

Guoqing Zhou [1], Lihuang Qian [1,*] and Paolo Gamba [2]

[1] Guangxi Key Laboratory of Spatial Information and Geomatics, College of Geomatics and Geoinformation, Guilin University of Technology, Guilin 541004, China; gzhou@glut.edu.cn
[2] Department of Electrical, Biomedical, and Computer Engineering, University of Pavia, 27100 Pavia, Italy; paolo.gamba@unipv.it
* Correspondence: 1020221941@glut.edu.cn

Highlights

What are the main findings?

- This review develops an innovative taxonomy for vision–X (including vision, language, audio, and position) multimodal remote sensing foundation models (MM-RSFMs) according to their backbones, encompassing CNN, Transformer, Mamba, Diffusion, vision–language model (VLM), multimodal large language model (MLLM), and hybrid backbones.
- A thorough analysis of the problems and challenges confronting MM-RSFMs reveals a scarcity of high-quality multimodal datasets, limited capability for multimodal feature extraction, weak cross-task generalization, absence of unified evaluation criteria, and insufficient security measures.

What is the implication of the main finding?

- The taxonomy assists readers in developing a systematic understanding of the intrinsic characteristics and interrelationships between cross-modal alignment and multimodal fusion in MM-RSFMs from a technical perspective.
- By analyzing key issues and challenges, targeted improvements can be implemented to improve the generalization, interpretability, and security of MM-RSFMs, thereby advancing their research progress and innovative applications.

Abstract

Remote sensing foundation models (RSFMs) have demonstrated excellent feature extraction and reasoning capabilities under the self-supervised learning paradigm of "unlabeled datasets—model pre-training—downstream tasks". These models achieve superior accuracy and performance compared to existing models across numerous open benchmark datasets. However, when confronted with multimodal data, such as optical, LiDAR, SAR, text, video, and audio, the RSFMs exhibit limitations in cross-modal generalization and multi-task learning. Although several reviews have addressed the RSFMs, there is currently no comprehensive survey dedicated to vision–X (vision, language, audio, position) multimodal RSFMs (MM-RSFMs). To tackle this gap, this article provides a systematic review of MM-RSFMs from a novel perspective. Firstly, the key technologies underlying MM-RSFMs are reviewed and analyzed, and the available multimodal RS pre-training datasets are summarized. Then, recent advances in MM-RSFMs are classified according to the development of backbone networks and cross-modal interaction methods of vision–X, such as vision–vision, vision–language, vision–audio, vision–position, and vision–language–audio. Finally, potential challenges are analyzed, and perspectives for

check for updates

Academic Editor: Lefei Zhang

Received: 29 August 2025
Revised: 15 October 2025
Accepted: 22 October 2025
Published: 24 October 2025

Citation: Zhou, G.; Qian, L.; Gamba, P. Advances on Multimodal Remote Sensing Foundation Models for Earth Observation Downstream Tasks: A Survey. *Remote Sens.* **2025**, *17*, 3532. https://doi.org/10.3390/rs17213532

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

MM-RSFMs are outlined. This survey from this paper reveals that current MM-RSFMs face the following key challenges: (1) a scarcity of high-quality multimodal datasets, (2) limited capability for multimodal feature extraction, (3) weak cross-task generalization, (4) absence of unified evaluation criteria, and (5) insufficient security measures.

Keywords: remote sensing foundation model (RSFM); multimodal data; generative pre-trained Transformer (GPT); Earth observation; self-supervised learning

1. Introduction

In recent years, deep learning (DL) and vision–language models (VLMs) have been extensively applied to a wide range of Earth Observation (EO) tasks, including but not limited to change detection [1], semantic segmentation [2], scene classification [3], target detection [4], and disaster management [5]. However, traditional remote sensing (RS) models based on machine learning and neural networks are typically designed for specific tasks and single modalities. This greatly limits their versatility and adaptability when confronted with terabyte-scale multimodal data generated daily, such as optical, LiDAR, SAR, video, audio, and text. Moreover, these models often underutilize the abundant multi-source RS data and their inherent multimodal features, resulting in limited representation and generalization ability in practical multi-task applications. Additionally, many DL-based RS models rely on ImageNet for supervised pre-training. Since ImageNet consists of natural images, these models struggle to process unlabeled remote sensing images (RSIs) and complex RS scenes. Consequently, there is a growing need for a more generic and efficient pre-training framework tailored for RS interpretation tasks. The emergence of remote sensing foundation models (RSFMs) promises to tackle the above challenges.

As a generic artificial intelligence (AI) pre-training paradigm, the foundation model can be trained in a self-supervised manner on large-scale unlabeled datasets using various DL techniques, including but not limited to graph neural networks (GNNs) [6], convolutional neural networks (CNNs) [7], generative adversarial networks (GANs) [8], multilayer perceptrons (MLPs) [9], generative pre-trained Transformers (GPTs) [10], and variational autoencoders (VAEs) [11]. As a vertical application of foundation models in RS, the RSFM is a generic model that employs pre-training techniques to learn representations from large-scale unlabeled RS data in a self-supervised way [12–21]. It has demonstrated promising initial performance on specific modal datasets [22–26] and has been successfully adapted to various EO downstream tasks, such as semantic segmentation, scene classification, change detection, image captioning, and visual grounding, through fine-tuning, prompt tuning, or context learning.

As the first pre-training foundation model tailored for RS data in the world, RingMo has significantly advanced both the research and application of RSFMs since its release in July 2022 [12]. Accordingly, this article identifies 2022 as the inaugural year of RSFMs development. Furthermore, a detailed survey was conducted using the Web of Science (WOS) database to retrieve peer-reviewed papers and review articles on RSFMs published from 2022 to 2024, as summarized in Figure 1. During the retrieval process, this article employed important information such as subject keywords and publication time to ensure accuracy. The subject keywords were “remote sensing foundation model”, the publication time was “2022–2024”, and the publication types were “article” and “review article”. Subsequent screening based on titles, keywords, and abstracts yielded 590 relevant publications in the field of RSFMs. Figure 1a illustrates a year-on-year increase in the number of publications related to RSFM research. The growth rate was approximately

23.08% from 2022 to 2023 and surged to about 53.98% from 2023 to 2024, reflecting growing research interest in this field. Figure 1b presents the proportion of methodology keywords we selected based on the dominant models or methods identified in the publications. Among them, "Transformer" and "self-supervised learning" emerged as the most prominent techniques, accounting for 11.47% and 11.01%, respectively. Among Transformer-related publications, the focus lay primarily on Vision Transformer (ViT) and generative pre-trained Transformer (GPT). Publications on self-supervised learning mainly addressed contrastive learning and generative learning. Other notable methods included "Masked Image Modeling" (10.55%), "neural network" (10.09%), "Segment Anything Model" (9.63%), and "fine-tuning" (9.17%). The proportion of publications related to "multimodal model" reached 8.92%, underscoring the growing importance of multimodality in RSFM research. Figure 1c displays the distribution of publications across journals about RSFM research from 2022 to 2024. *IEEE Transactions on Geoscience and Remote Sensing* (TGRS) published the highest proportion at 31.43%, followed by *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (JSTAR) at 13.33%, *and ISPRS Journal of Photogrammetry and Remote Sensing* at 7.62%. Together, these three journals accounted for more than half of all publications in the field.

Although significant interest has been aroused within the RS and geoscience communities, current RSFMs remain largely focused on model pre-training and downstream tasks involving single-modal data and specific applications. Several key developmental milestones of RSFMs in recent years are illustrated in Figure 2. For RGB images, Zhang et al. employed CNN and Transformer techniques for target recognition in RGB-D images [27]. Cai et al. applied a multi-level Transformer model with self-attention mechanisms to implement per-pixel segmentation in RGB road scenes [28]. Hu et al. proposed a human-centric multimodal fusion network that fuses RGB images with depth and optical flow data [29]. Li et al. combined CNN with encoder–decoder networks to propose an RGB–thermal image segmentation method based on parameter sharing and attention fusion models [30]. Qin et al. introduced a semantic segmentation network using foreground attention to address the problem of sun-glint correction in high-resolution marine Unmanned Aerial Vehicle (UAV) RGB images [31]. For hyperspectral images (HSIs), Mei et al. proposed a group-aware hierarchical Transformer to mitigate feature dispersion issues using a multi-head self-attention (MHSA) mechanism [32]. Gong et al. employed deep feature embedding convolutional networks to improve the interpretability and classification accuracy of HSIs [33]. Zhang et al. proposed a language-aware domain generalization network leveraging contrastive learning and image–text dual encoders for cross-scene HSI classification [34]. For multispectral images (MSIs), Zhou et al. developed an inverse neural network to tackle the issue of pan-sharpening between high-resolution and low-resolution MSIs [35]. Zhou et al. proposed a self-organizing pixel entangled neural network for accurate unsupervised classification of MSIs [36]. Zheng et al. applied a spectral knowledge transfer framework with an autoencoder model to improve the spectral resolution in MSI change detection [37]. For 3D LiDAR images, Zheng et al. utilized density clustering and hierarchical clustering for noise point filtering and target recognition in LiDAR data [38]. Farmonov et al. integrated the 3D structural features from LiDAR data with the spectral features of hyperspectral data via attention mechanism and CNNs to improve the efficiency and accuracy of extracting crop shape and texture features [39]. Ma et al. developed a deep neural network based on ViTs to achieve 3D LiDAR moving-object segmentation [40]. For SAR images, Zhao et al. designed a domain-adaptation Transformer to improve object detection accuracy of the unlabeled multi-source satellite-borne SAR images [41]. Yasir et al. combined CNNs with nonlinear regression techniques to extract geometric features of ships from SAR data [42]. Wang

et al. employed hierarchical embedding and incremental evolutionary networks for SAR target recognition in open and few-shot scenes [43]. Undoubtedly, a large number of foundation models have been widely applied to specific EO tasks such as object detection and semantic segmentation, demonstrating superior performance and accuracy compared with traditional models. However, RSFMs designed for single-modal data struggle to accommodate diverse downstream tasks and multimodal RS data. This limitation has prompted a growing shift among researchers toward multimodal remote sensing foundation models (MM-RSFMs).

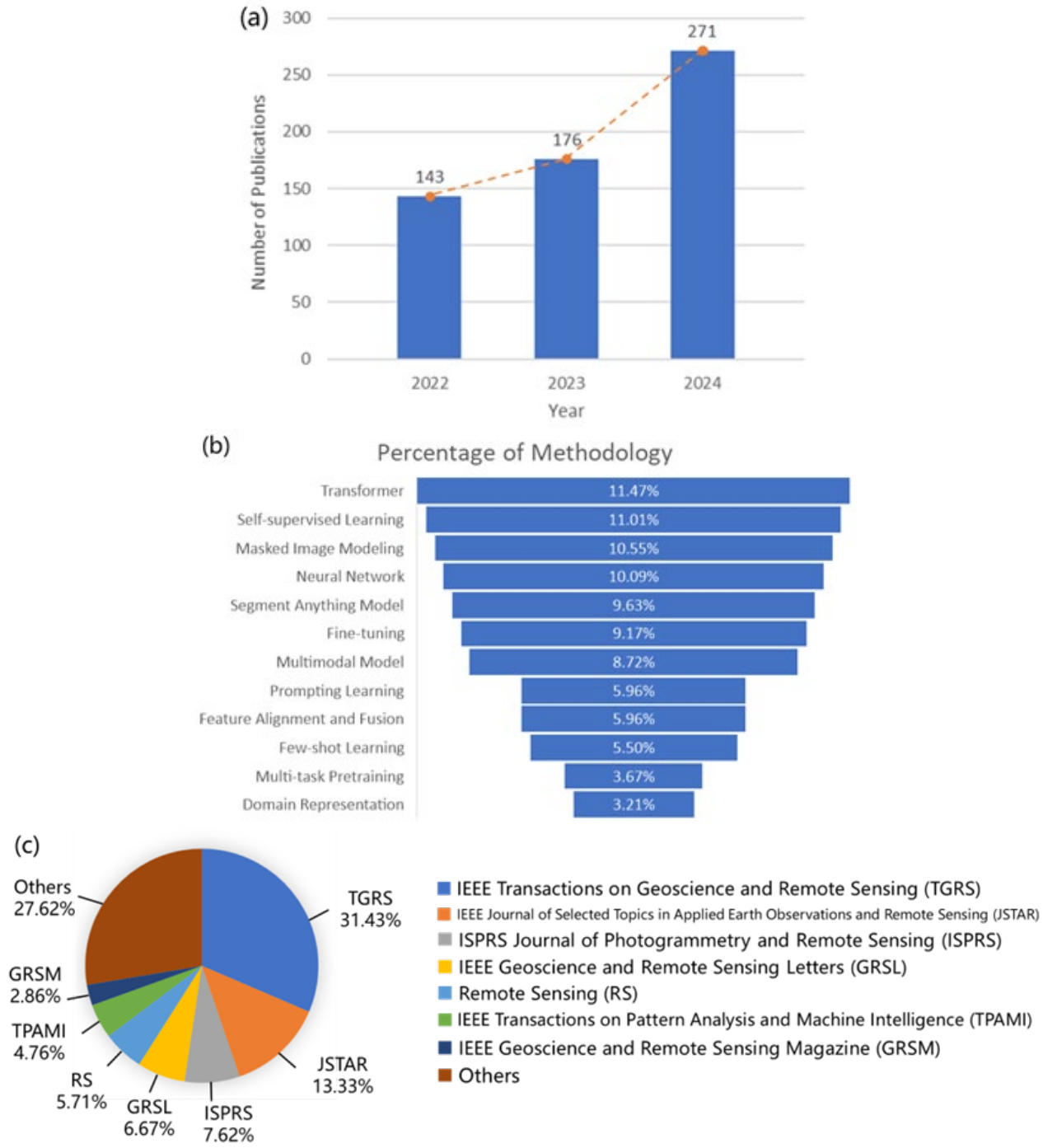


Figure 1. The number and proportion of publications of RSFMs on the basis of (**a**) publication year, (**b**) existing methodologies, and (**c**) various journals.

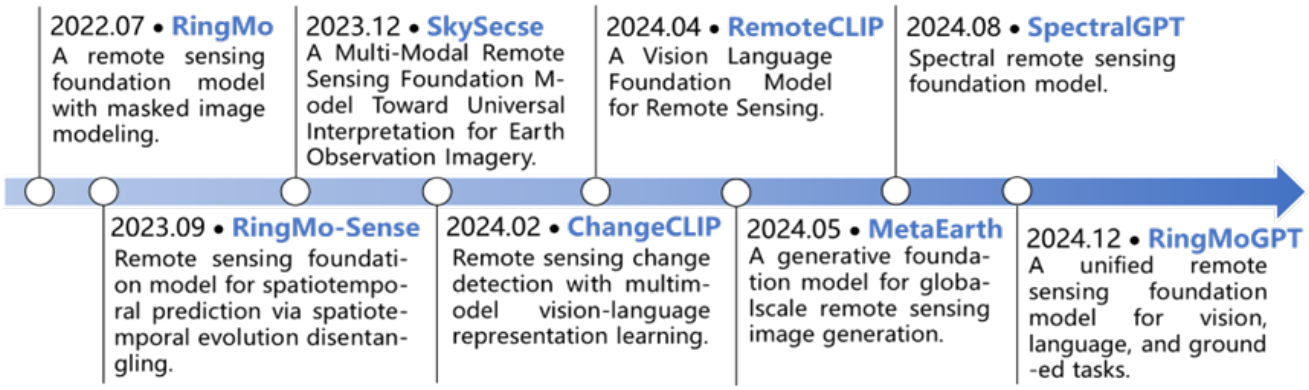


Figure 2. Major milestones in the development of RSFMs.

The multimodal data acquired from diverse RS sensors can comprehensively characterize object features across multiple dimensions, such as semantics, spectrum, space, time, wavelength, and reflectivity. However, these data also exhibit significant intermodal discrepancies and fusion complexities Therefore, MM-RSFMs must possess capabilities for cross-modal alignment, feature extraction, and multi-task learning [44,45]. In order to review the research advances of RSFMs, Huang et al. provided a comprehensive survey of RS vision and multimodal foundation models across four dimensions: architecture design, training methods, training datasets, and evaluation criteria [46]. Zhang et al. categorized the RSFMs from the perspective of mining and utilizing geoscience knowledge [47]. Fu et al. summarized RSFMs based on single-temporal and multi-temporal data, developing a framework for a new generation of generic predictive RSFMs [48]. Zhang et al. systematically reviewed the existing generic geoscience foundation models in terms of the key advantages, applications, technological progress, and challenges [49]. Yan et al. focused on the research progress of multimodal remote sensing large models, interpretable remote sensing large models, and human feedback reinforcement learning from three levels: data, model, and downstream tasks [50]. Li et al. systematically summarized the research progress and development trend of vision–language foundation models in RS from the task level of visual question answering, image captioning, and semantic segmentation [51]. Bao et al. reviewed the overall architecture and application pathways of the state space model represented by Vision Mamba in RS tasks [52]. Unlike existing surveys and reviews of RSFMs, this article provides a comprehensive analysis of the latest progress, key challenges, and future trends of MM-RSFMs.

The main contributions of this article are as follows:

(1) Comprehensive Survey: This article provides the first comprehensive survey of vision–X (including vision, language, audio, and position) MM-RSFMs specifically designed for EO downstream tasks. It systematically reviews research progress, technological innovation, model architecture, key issues, and development trends of MM-RSFMs across five dimensions: pre-training data, key technologies, backbones, cross-modal interactions, and problems and prospects.

(2) Innovative Taxonomy: This article develops an innovative taxonomy framework for MM-RSFMs and reviews the development of them according to multimodal backbones, such as CNN backbones, Transformer backbones, Mamba backbones, Diffusion backbones, vision–language model (VLM) backbones, multimodal large language model (MLLM) backbones, and hybrid backbones. This taxonomy aids readers in developing a systematic understanding of the intrinsic characteristics and interrelationships among cross-modal alignment and multimodal fusion in MM-RSFMs.

(3) Thorough Analysis: This work conducts a thorough analysis of the problems and challenges confronting MM-RSFMs and predicts future directions. It analyzes the key issues of MM-RSFMs from five aspects: scarcity of high-quality multimodal datasets, limited capability for multimodal feature extraction, weak cross-task generalization, absence of unified evaluation criteria, and insufficient security measures. These insights are intended to facilitate further research progress and innovative applications in the field.

This article is organized into seven sections, as illustrated in Figure 3: Introduction, Key Technologies, Backbones, Vision–X RSFMs, Problems and Challenges, and Conclusion and Outlook. Section 2 discusses multimodal RS pre-training data. Section 3 introduces the key technologies underpinning MM-RSFMs, covering four aspects: self-supervised learning, Vision Transformers, Masked Image Modeling, and model fine-tuning. Section 4 analyzes the backbones of MM-RSFMs, including CNN, Transformer, Mamba, Diffusion, VLM, MLLM, and hybrid backbones. In Section 5, the latest progress of MM-RSFMs is reviewed according to the method of cross-modal interaction: vision–X (including vision, language, audio, and position). Section 6 examines the challenges facing MM-RSFMs from five perspectives: high-quality datasets, multimodal feature extraction, generalization ability, evaluation criteria, and security measures. Finally, Section 7 concludes the article by summarizing the technical strengths of MM-RSFMs and suggesting promising directions for future research.

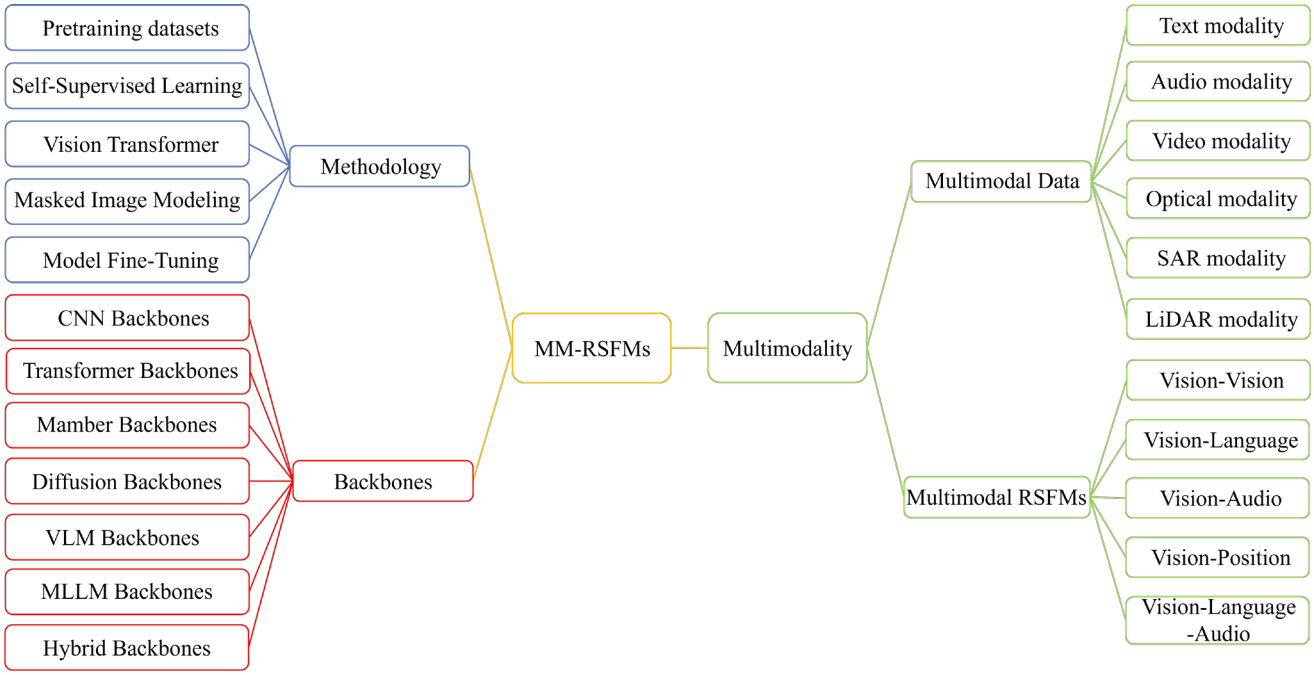


Figure 3. The overall structure of this article.

2. Multimodal RS Pre-Training Data

Multimodal RS pre-training data serves as a core resource enabling foundation models to learn jointly across multiple modalities, including optical, LiDAR, SAR, text, video, and audio. Compared with single-modal RS data, multimodal RS data requires higher robustness and generalization capability from the foundation model. By training with large-scale unlabeled multimodal RS data, RSFMs can obtain universal feature representations from various modalities and then efficiently adapt to various object scenes or downstream tasks through self-supervised learning. This section categorizes and summarizes four types of

multimodal RS pre-training datasets: vision + vision, vision + language, vision + audio, and vision + position.

Self-supervised learning is an essential feature of pre-training for RSFMs. Specifically, during the pre-training phase, self-supervised methods are applied to large-scale unlabeled remote sensing datasets to extract implicit labels reflective of ground object features. Dataset acquisition and processing are critical components of the model pipeline, operating in a relationship that is both independent and complementary. Consequently, the quality of pre-training datasets significantly influences the overall performance of RSFMs and their adaptability to downstream tasks. Especially in the context of MM-RSFMs, single-modal datasets can no longer meet the requirements for training generic foundation models or address the growing demand for multimodal RS data in EO tasks. Constructing large-scale, high-quality pre-training datasets that encompass diverse modalities and structures of RS data remain a persistent and critical technical challenge. As summarized in Table 1, multimodal RS pre-training datasets can be categorized into four types: vision + vision, vision + language, vision + audio, and vision + position.

Table 1. Multimodal remote sensing datasets.

Modality	Dataset	Year	Type	Size	Feature Description
Vision + Vision	Houston 2013 [22]	2021	HS, MS	15,029	The multimodal RS benchmark dataset obtained by reducing the spatial and spectral resolution of the original HSIs covers 20 types of land features such as roads, buildings, and vegetation, and is used for land-cover classification.
	Augsburg [22]	2021	HS, SAR, DSM	78,294	This multimodal RS benchmark dataset consists of spaceborne HSIs, Sentinel-1 dual-polarization (VV-VH) PolSAR images, and digital surface model (DSM) images, and is used for land-cover classification.
	DKDFN [26]	2022	MS, SAR, DEM	450	A multimodal dataset composed of Sentinel-1 bipolar SAR images, Sentinel-2 MSIs, and SRTM digital elevation model data is provided, with a resolution of 10 m, and is used for land-cover classification.
	Potsdam [53]	2021	DSM, IRRG, RGB, RGBIR	38	An RS benchmark dataset for urban semantic segmentation, in which Potsdam city, covered by airborne orthophoto images, features large building complexes, narrow streets, and dense settlement structures.
	Trento [54]	2020	HS, LiDAR	30,414	A multimodal dataset for RS land classification, derived from airborne HSIs and LiDAR data from the Trento rural area, with a spatial resolution of 1 m.
Vision + Language	RemoteCLIP [17]	2024	RSI, Text	828,725	An RS image–text dataset specifically established for the pre-training task of vision–language models, featuring a wide variety of scene types, image captions, semantics, and alignment features.
	iSAID [55]	2019	RSI, Text	2806	The first RS benchmark dataset composed of high-resolution aerial images and their annotations, featuring large-scale and multi-scale characteristics, is used for instance segmentation and object detection.
	RSICap [56]	2023	RSI, Text	2585	The RS image–text pairs containing image types, land feature attributes, and scene descriptions can be used for the pre-training of MM-RSFMs and for its downstream tasks.
	SkyScript [57]	2023	RSI, Text	2.6 M	A dataset of image–text pairs with a scale of millions, among which the RSIs are obtained from the Google Earth Engine (GEE) platform, and the corresponding semantic labels are obtained from the Open Street Map (OSM).
	LEVIR-CD [58]	2020	RSI, Text	637	The benchmark dataset for change detection, which includes a total of 31,333 individual buildings with changes, involves many variations such as sensor characteristics, atmospheric conditions, and lighting conditions.
	MMRS-1M [59]	2024	RSI, Text	1 M	A large-scale multi-sensor and multimodal RS instruction tracking dataset with image–text pairs, where the visual modality includes optical, SAR, and infrared, etc., and covers classification, image captions, and visual question answering, etc.

Table 1. *Cont.*

Modality	Dataset	Year	Type	Size	Feature Description
Vision + Audio	SoundingEarth [60]	2023	Image–Audio	50,545	An image–audio pair consisting of aerial images from 136 countries and corresponding crowdsourced audio from scenes, without the need for manual geographic annotation, can be used for geospatial perception and audio–visual learning tasks in RS.
	ADVANCE [61]	2020	Image–Audio	5075	A multimodal dataset for audio–visual aviation scene recognition tasks, consisting of audio data geotagged from FreeSound and high-resolution images from Google Earth.
	Ref-AVS [62]	2024	Video, Audio	40,020	A multimodal benchmark dataset rich in audio and visual descriptions, which can provide pixel-level annotations and multimodal cues for dynamic visual and auditory object segmentation tasks, covering a wide range of object categories.
Vision + Position	RingMoGPT [63]	2024	Image, Text	522,769	A unified RS basic model pre-training dataset for visual, language, and grounding tasks, containing over 500,000 pairs of high-quality image–text pairs, generated through a low-cost and efficient data generation paradigm.
	iNaturalist 2018 [64]	2018	Image, Text	859,000	The dataset used for object detection and classification consists of over 800,000 images from more than 5000 object categories, and each image has a ground-truth label.
	fMoW [65]	2018	Image, Metadata	1 M	The dataset used for identifying buildings and land-use tasks can infer object features such as location, time, solar angle, and physical size from satellite MSIs and the corresponding metadata of each image.
	MP-16 [66]	2017	Geotagged Images	4.72 M	This dataset consists of over 4 million geotagged images, enabling alignment between the images and their corresponding GPS locations. It can be used to develop global geographic positioning models for image-to-GPS retrieval methods.
	STAR [67]	2024	<Subject, Relationship, and Object>	400,000	The first benchmark dataset for satellite image scene-graph generation tasks, including complex scenarios such as airports, ports, and overpasses, covers 210,000 geographical entities and 400,000 target relation triples.

Million (M).

Currently, many multimodal RS datasets have been applied in downstream tasks such as target recognition, land-cover classification, semantic segmentation, and image captioning. This paper introduces some multimodal RS datasets from aspects such as data modality, release year, data volume, and characteristics, following the classification method of vision + vision, vision + language, vision + audio, and vision + position. Processing and learning of multimodal RS data with different features and structures in the same model requires cross-modal data fusion. Chen et al. proposed a Fourier-domain structural relationship analysis framework for unsupervised multimodal change detection, which utilizes modality-independent local and non-local structure relationships to solve the heterogeneity problem of modalities [68]. Zhu et al. proposed a feature-matching method based on repeatable feature detectors and rotation-invariant feature descriptors, which solved the problem of low feature-matching accuracy of multimodal RSIs due to differences in radiation and geometric structure [69]. To achieve cross-modal feature interaction learning between hyperspectral data and multispectral data, Li et al. proposed an X-shaped interactive autoencoder network. This method uses a dual autoencoder coupled network to capture the spatial–spectral information of cross-modal images [70].

3. Key Technologies for MM-RSFMs: From Self-Supervised Learning to Model Fine-Tuning

3.1. Self-Supervised Learning

As an important unsupervised learning technique in CV and natural language processing (NLP), self-supervised learning has proven to be an effective pre-training strategy for RS data. This technique enables the training of RSFMs without relying on a large amount

of labeled data. Instead of using manual annotations, self-supervised learning generates implicit labels from large volumes of unstructured and unlabeled source-domain datasets. It extracts and learns latent features of objects, along with general knowledge that benefits downstream tasks, ultimately facilitating transfer to target-domain datasets. Owing to its strong knowledge transfer capabilities and model generalizability, self-supervised learning techniques have significantly advanced the development of pre-trained RSFMs. Among them, contrastive learning and generative learning stand out as two prominent methods for model pre-training.

3.1.1. Contrastive Learning

Contrastive learning can be applied in model training for RSI processing and interpretation. This method aims to enable the model to learn and distinguish the data features of similar and different samples in the training data, so as to mine potential feature representations from a large number of unlabeled datasets through self-supervised methods. The core idea of contrastive learning is to generate positive and negative samples using data augmentation methods (such as CLSA [71], CGA2TC [72]) and to optimize the model using loss functions (such as InfoNCE [73], Triplet Loss [74]) to achieve the purpose of comparing and distinguishing the positive and negative samples in the feature space. The ultimate goal of contrastive learning is to make the feature representations of the dataset samples and the positive samples as close as possible, while being as different as possible from the representations of the negative samples, as illustrated in Figure 4. Typical contrastive learning methods include SimCLR [75] and MoCo [76].

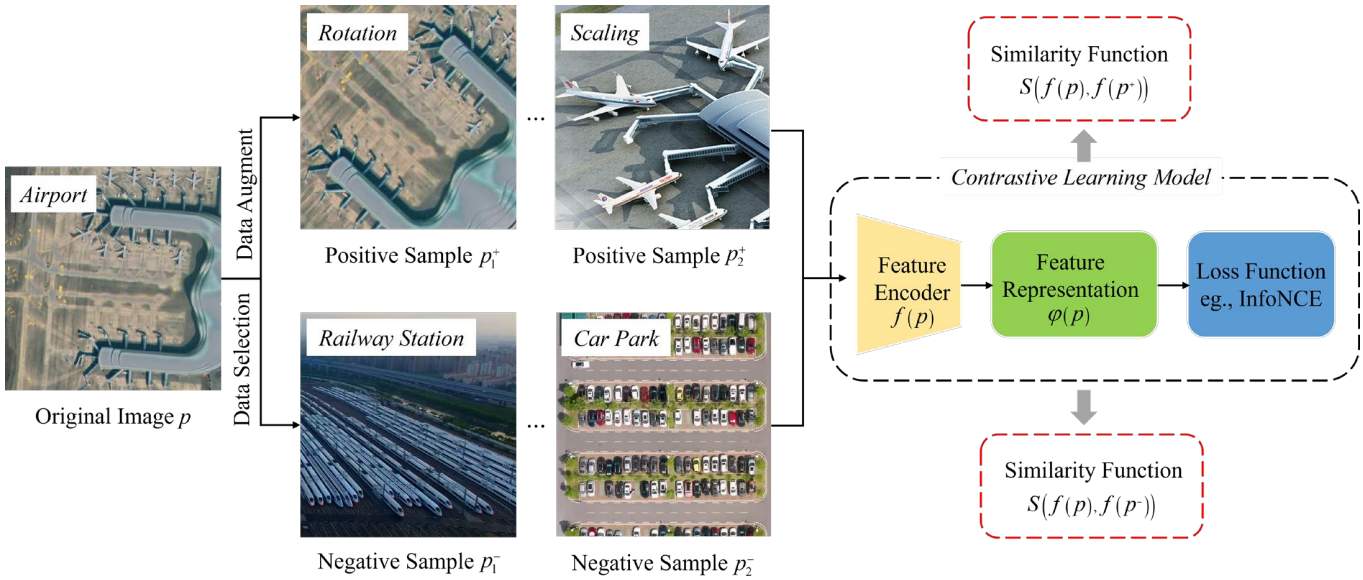


Figure 4. Contrastive learning for MM-RSFMs.

The contrastive learning method is used to pre-train the MM-RSFMs. Constructing positive and negative samples of image samples is a key point of the research. As shown in Figure 4, positive samples are first constructed by performing data augmentation such as translation, rotation, scaling, cropping, fusion, sharpening, grayscale conversion, and masking on the original multimodal RSIs. Then, negative samples are selected from the original dataset. Next, the similarity between the image samples and positive/negative samples is calculated using contrastive learning feature encoders and loss functions such as InfoNCE, after which the image feature representations are generated. For large-scale unlabeled RS data, pre-training with contrastive learning enables MM-RSFMs to learn and

master the features and spatial attributes of different structures and modalities of RS data, providing an unsupervised and generalizable model pre-training paradigm.

3.1.2. Generative Learning

Generative learning, another important branch of self-supervised learning, provides an ideal unsupervised learning paradigm for model pre-training in practical situations where it is difficult to obtain large-scale and high-quality labeled multimodal RS samples. In essence, generative learning refers to starting from the original data and training a model to learn the latent features and attribute information of the data, in order to generate new, similar data or restore missing data while maintaining the characteristics of the original data. Common generative learning methods include generative adversarial networks (GAN) [77], Masked Image Modeling (MIN) [78], variational autoencoders (VAE) [79], and Generative Pre-training Transformer (GPT) [80], etc. Figure 5 shows the differences between contrastive learning and generative learning methods in RSFMs pre-training.

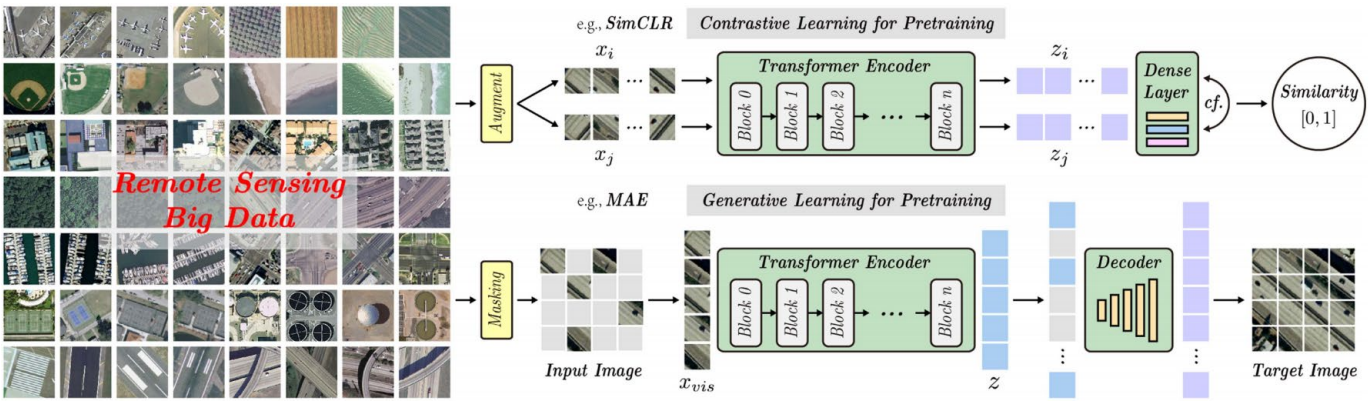


Figure 5. The differences between contrastive learning and generative learning for RSFMs pre-training [19].

The core objective of generative learning models is to generate new similar samples of the same category by learning the potential probability distribution of unlabeled dataset samples [81]. The process can be described as follows: for a random sample of images or a collection of image samples, the generative learning model needs to estimate the probability distribution of the image samples in the feature space, thereby generating samples of the same category based on this probability distribution. As a typical generative learning method widely used in the fields of computer vision and image processing in recent years, GANs mainly compose of a generator module and a discriminator module [82]. The generator module learns the image feature characteristics of a large number of unlabeled samples through deep neural networks to construct a parameterized probability distribution generation model, thereby generating sample images with similar real characteristics based on this model. The discriminator module uses neural networks to determine whether the sample images generated by the generator are real. The two modules compete with each other and mutually optimize during training.

3.2. Vision Transformer

The Vision Transformer (ViT) is a DL model that uses the Transformer network architecture to handle visual image tasks. Its core technologies include the multi-head self-attention mechanism (MHSA) and positional encoding [83–85]. The basic principle of the Vision Transformer is to divide the image into several small blocks containing various ground feature pixels, regard each small block as an image sequence unit, and then extract the spatial and temporal features of the image through the encoder–decoder of the Transformer

layer. The self-attention mechanism can perform weighted summation for each position in the input image sequence and extract the global dependence of the image sequence by modeling interactions among any pixels in the input image sequence [86]. The Vision Transformer adopts a method called "offset binary positional encoding" to capture the position information of each position in the input sequence. This method regards each position of the input sequence as a pixel point with a specific offset and then maps the coordinates of the pixel point to a binary vector, thereby achieving positional encoding [87]. As shown in Figure 6, the Transformer usually consists of an encoder and a decoder with the same network architecture, specifically including modules composed of multi-head attention mechanism, feedforward neural network, residual connection, and layer normalization. The encoder generates position encodings for the input sequence, while the decoder acquires all the encoded information and uses the context information it contains to generate the output sequence [87].

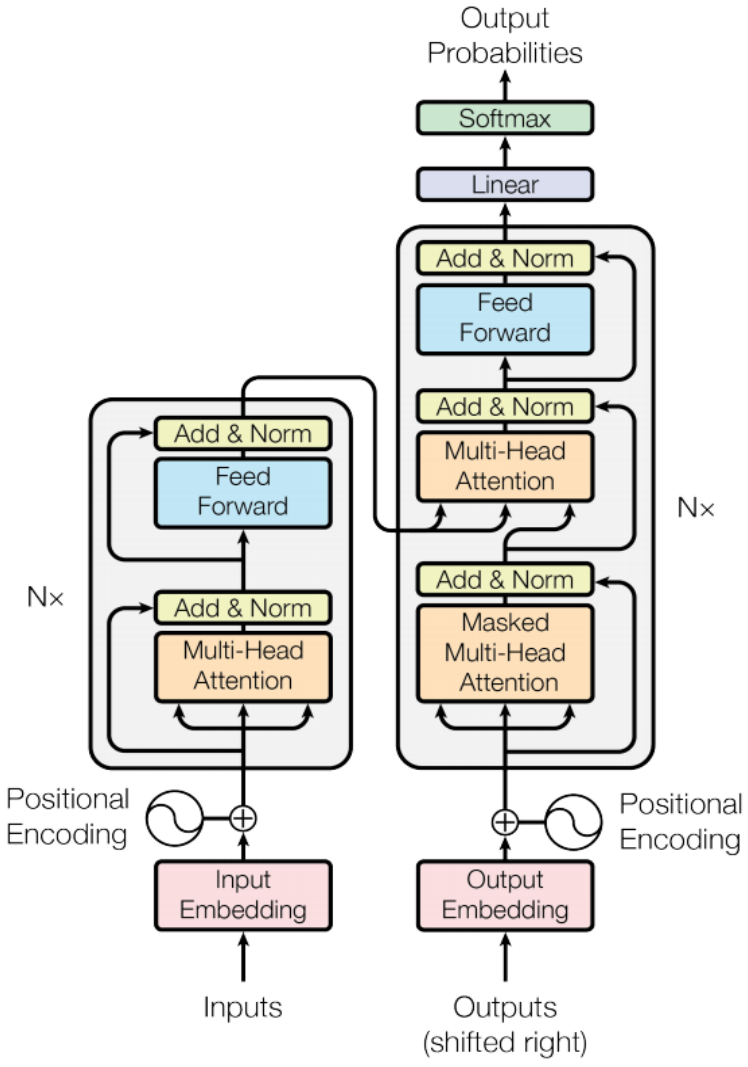


Figure 6. The architecture of the Transformer model [87].

With the significant success of Transformer-based generative learning methods in the foundation models of CV, more and more foundation models in the field of RS are using ViT and achieving good results. RingMo-Sense constructed an RS spatio-temporal prediction basic model by using the Swin Transformer encoder [13]. This model simultaneously achieves stable long-term predictions for RS video data and time-series images by using parameter sharing and progressive joint training strategies, demonstrating com-

petitive model performance in six downstream spatio-temporal tasks. For large-scale and diverse RS tasks, the RVSA model adopted a rotationally variable-sized window attention mechanism to replace the original full attention mechanism in the ViT [20]. This model can extract rich contextual information from different windows generated to significantly reduce computational costs and memory usage, while improving object representations. Compared with existing advanced methods, this model has better accuracy in downstream classification and segmentation tasks. For multi-scale EO tasks, the TTST model designed a residual token-selectable module and multi-scale feedforward neural network layers based on Transformer technology to achieve multi-scale feature mixing for object representation and global context information generation [88].

3.3. Masked Image Modeling

Masked Image Modeling (MIM) is a self-supervised learning method in the field of pre-training of RSFMs. This method achieves semantic encoding and feature learning by locally masking input images and using a deep network to extract the pixel and information of the unmasked parts to predict signals [89,90]. Compared with supervised pre-training and deep networks based on contrastive learning that only model long-range information, the MIM model can simultaneously model local and long-range information, though its semantic modeling capacity is somewhat lower [91,92]. As an important component of MM-RSFMs, the attention mechanism module in the MIM model can mine multi-scale geometric and spectral information in the RS pre-training dataset by assigning different attention weights to image mask blocks. When performing pre-training of RSFMs, the MIM model converts the important signals in different image mask blocks into token units, converts the tokens into a data structure represented by vectors through embedding, and finally inputs it to the encoder–decoder neural network module, such as the Masked Autoencoder Model (MAE). It can extract the spatio-temporal feature attributes of the image input sequence based on the prediction signals of the image mask module, thereby completing the pre-training of RSIs through self-supervised learning.

Currently, many DL-based RS models heavily rely on large labeled image datasets, which restricts their development in self-supervised learning capabilities for complex terrain scenes and large-scale unlabeled images. To address this shortcoming, Hou et al. proposed a language-assisted masked image pre-training model based on MIN (MILAN), which includes a prompt decoder and a semantic-aware mask sampling mechanism [93]. It also uses the Contrastive Language–Image Pre-training (CLIP) [94] image encoder results as the image reconstruction target and pre-trains and fine-tunes the Masked Autoencoder on the Image-1K dataset, ultimately achieving better accuracy than before in downstream semantic segmentation tasks.

Given the limited labeled data in RSIs and the limited representation capabilities of existing DL-based RS models, the AST model [95] combines the MIM pre-training model and the multi-scale Transformer architecture to propose an expandable and adaptive self-supervised image interpretation model, which can discover potential supervisory signals in a large amount of unlabeled RS data and learn multi-granularity semantic features. To alleviate the loss of image semantic information caused by the use of Token sparse technology, the AMAT model [96] proposes a self-adaptive mask auto-encoding mechanism and model training objective function based on MIM. This model is used for the pre-training and fine-tuning stages of the image classification model and can effectively reduce the model complexity and accuracy of the Transformer. As a new self-supervised pre-training method, the CMAE model [91] integrates the advantages of the MIM model and the contrastive learning method to learn image feature representations with strong instance distinguishability and local perceptibility, achieving excellent performance in

image classification, semantic segmentation, and object detection tasks. To address the domain gap between natural and remote sensing scenes and the poor generalization ability of RS models, the RingMo [12] model utilizes generative self-supervised learning and the MIM model to leverage the advantages of RSIs to establish a large-scale dataset and propose a training method for RSFMs focused on dense and small objects in complex RS scenes.

3.4. Model Fine-Tuning

RSFMs that use single-modal RS data as the pre-training object have certain feature extraction and object representation capabilities and have been successfully applied to specific RS interpretation tasks. However, such models often require a large amount of computing power and training costs, and it is difficult to ensure model accuracy when facing various RSI interpretation tasks such as scene classification, semantic segmentation, and change detection. Large-model fine-tuning technology can provide an effective and feasible solution to improve the generalization ability and task adaptability of RSFMs. Large-model fine-tuning refers to further training a pre-trained model using datasets specific to downstream tasks [97,98]. For RSFMs, fine-tuning technology enables the model to further train and optimize itself based on the specific EO downstream tasks and target datasets while learning the image features of the pre-training dataset, thus enabling the model to have model accuracy and generalization ability in various downstream tasks. The fine-tuning process is usually much faster than training a new RSFM from scratch, and often achieves better training results and model accuracy, especially in cases where RS data is limited. Common fine-tuning methods include, but are not limited to, full data fine-tuning, and parameter-efficient fine-tuning [99,100].

The main steps of RSFM fine-tuning include dataset preparation, model architecture optimization, hyperparameter setting, iteration convergence, and evaluation and application. The fine-tuning dataset preparation stage involves preparing a dataset that is conducive to improving the generalization ability and multi-task adaptability of the model, based on the differences between the pre-training dataset of the RSFMs and the target dataset of the specific task. This may involve data augmentation, data cleaning, and image annotation, etc. Model architecture optimization is adding additional fully connected layers or modifying the existing encoder structure based on the fine-tuning needs, in addition to the original remote sensing pre-training model architecture. The fine-tuning hyperparameter setting stage requires determining key training parameters such as learning rate, training epochs, and batch size based on the fine-tuning dataset and specific task characteristics [101,102]. The fine-tuning iteration stage involves executing fine-tuning with the selected dataset and optimized model architecture, followed by forward propagation, loss calculation, backpropagation, weight updates, and iterative convergence to achieve the best model training effect [103,104]. The key at this stage is to avoid overfitting of the RSFMs, which can be controlled through methods such as early stopping, regularization, etc. The evaluation and application stage requires evaluating the model accuracy and performance based on the results of fine-tuning iteration convergence in the previous stage, and continuously optimizing and adjusting the model, such as fine-tuning hyperparameters or improving the fine-tuning datasets [105,106]. Finally, the fine-tuned RSFMs will be applied to specific EO downstream tasks.

4. The Backbones of MM-RSFMs

In the early days, remote sensing models based on DL techniques such as CNN [107] and MLP [108] typically used supervised learning methods for training. However, these models were limited by high costs of manual labeling and low predictive capabilities, making them unable to effectively adapt to large-scale unlabeled RS data and complex

downstream tasks of EO. In recent years, with the rise in large language models and ChatGPT models, more and more scholars in the field of remote sensing have begun to study generic RSFMs. Recently, self-supervised learning techniques have become the mainstream method for pre-training of RSFMs, as they can utilize a large amount of unlabeled RSI data for model pre-training, enabling the model to apply the features learned from unlabeled RS data to various downstream tasks, such as scene classification, object detection, semantic segmentation, image captioning, and visual grounding [109]. Most RSFMs mainly handle single-modal RS data. However, RS has entered a new stage of multi-type sensors, including optical, LiDAR, SAR, and video data. Therefore, multimodality should become one of the important development trends of RSFMs.

Multimodal foundation models [110–112] use data from different modalities such as text, images, audio, and videos for learning and reasoning, utilizing technologies such as multimodal representation [113], modality fusion and alignment [114], cross-attention mechanism [115], multi-task learning [116], and large-model fine-tuning [117], to achieve feature learning and downstream tasks across modalities. It has become one of the popular research directions in general artificial intelligence. RS has multiple types of sensors and thus has multimodal properties. Moreover, RS data are essentially geographically integrated data with different time scales and spatial scales, so how to effectively extract multimodal features from RS data through RSFMs and applying them to different EO downstream tasks has become a problem that needs to be solved urgently. He et al. proposed an adaptive framework for multimodal RS data classification (FMA), which uses cross-spatial and cross-channel interaction modules to extract multimodal image features along spatial and channel dimensions, and explored various alignment methods between RSIs of different modalities, further improving the classification performance [117].

Currently, ChatGPT, based on large language models [118] and generative pre-training Transformer models, [119] has entered the era of multimodal AI large models for text and images. The resulting industry foundation models have been successfully applied in vertical fields such as industry [120], agriculture [121], meteorology [122], geographic information [123], finance [124], medicine [125], education [126], and law [127]. RSFMs are powerful, intelligent, and generic DL pre-training models that can learn and represent rich spectral, geometric, and semantic features from large-scale unlabeled RS data and apply them to downstream EO tasks. As an important extension, MM-RSFMs can explore the potential features of optical, LiDAR, SAR, text, video, audio, and other modal data through self-supervised learning and cross-modal alignment and fusion techniques to adapt to various downstream tasks of Earth Observation. Currently, research on RSFMs mainly focuses on two aspects: single-modal and multimodal RSFMs. Table 2 introduces MM-RSFMs published in the past three years across six aspects: publication year, modality, backbone model, parameter quantity, training volume, and downstream tasks. The architecture design of MM-RSFMs needs to consider two key factors: the fusion method and the fusion stage. The fusion method determines the interaction mode between different RS modalities, mainly including connection, element-wise multiplication, attention mechanism, etc. The fusion stage refers to the location where multimodal RS data interact, such as the input layer, intermediate layer, or output layer of the model. As illustrated in Figure 7, this article divides the backbones of the MM-RSFMs into seven categories: CNN backbones, Transformer backbones, Mamba backbones, Diffusion backbones, VLM backbones, MLLM backbones, and hybrid backbones.

Table 2. Several typical MM-RSFMs.

RSFM	Year	Data	Backbone	Parameter	Size	EO Downstream Tasks
RingMo-Sense [13]	TGRS 2023	videos, images	Transformer	-	1 M	VP, CF, REE, ODSV, MTSV, RSTIS
RingMo-SAM [14]	TGRS 2023	optical, SAR data	Transformer	-	1 M	SS, OD
SkySense [15]	CVPR 2023	optical, SAR data	Transformer	10 B	21.5 M	SS, OD, CD, SC
ChangeCLIP [16]	ISPRS 2024	image, text	VLM	-	0.06 M	CD
RemoteCLIP [17]	TGRS 2024	image, text	VLM	304 M	0.16 M	ITR, IP, KC, FSC, ZSIC, OCRSI
MetaEarth [18]	TPAMI 2024	optical image, GI	Diffusion	600 M	3.1 M	IG, RSIC
EarthGPT [59]	TGRS 2024	image–text pairs	MLLM	400 M	1 M	IC, RLC, VQA, SC, VG, OD
RingMoGPT [63]	TGRS 2024	image–text pairs	MLLM	-	0.52 M	SC, OD, VQA, IC, GIC, CC

Million (M), billion (B), video prediction (VP), cloud forecasting (CF), radar echo extrapolation (REE), object detection in satellite videos (ODAV), multi-object tracking in satellite videos (MTSV), RS time-series images' segmentation (RSTIS), semantic segmentation (SS), change detection (CD), scene classification (SC), image–text retrieval (ITR), linear probing (IP), k-NN classification (KC), few-shot classification (FSC), zero-shot image classification (ZSIC), object counting in RSIs (OCRSI), geographical information (GI), image generation (IG), RSI classification (RSIC), image captioning (IC), region-level captioning (RLC), visual question answering (VQA), visual grounding (VG), grounded image captioning (GIC), and change captioning (CC).

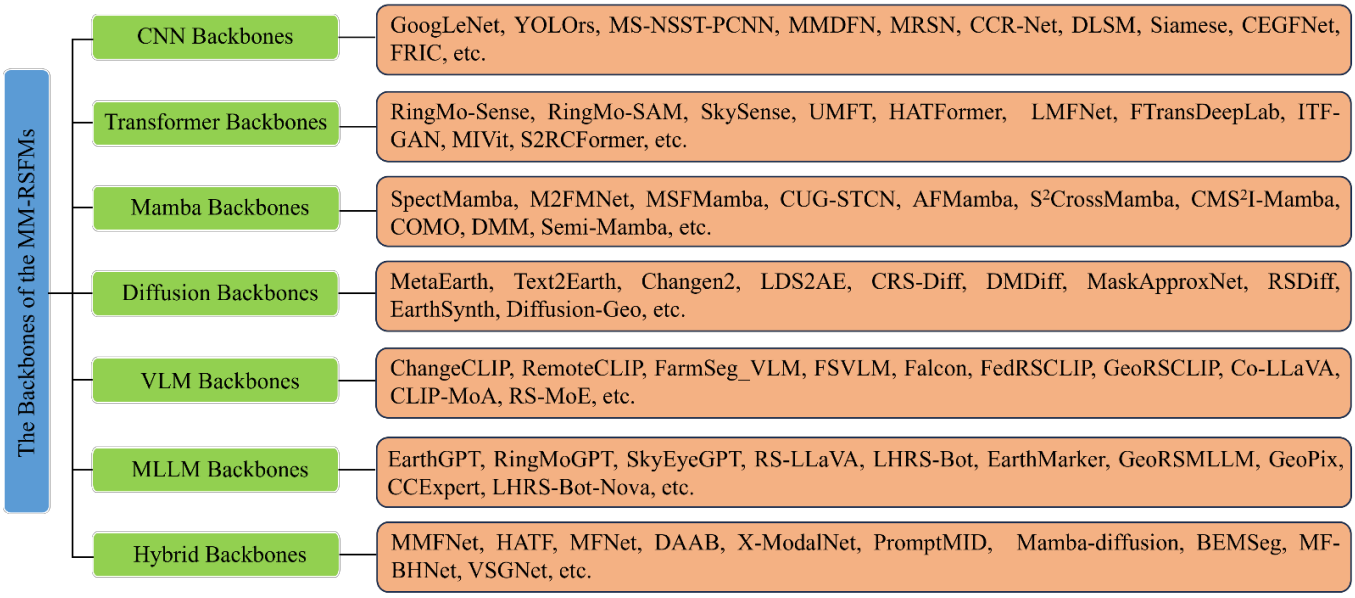


Figure 7. A chain of the backbones of the MM-RSFMs. GoogLeNet [128], YOLOrs [129], MS-NSST-PCNN [130], MMDFN [131], MRSN [132], CCR-Net [133], DLSM [134], Siamese [135], CEGFNet [136], FRIC [137], RingMo-Sense [13], RingMo-SAM [14], SkySense [15], UMFT [138], HATFormer [139], LMFNet [140], FTransDeepLab [141], ITF-GAN [142], MIVit [143], S2RCFormer [144], SpectMamba [145], M2FMNet [146], MSFMamba [147], CUG-STCN [148], AFMamba [149], S2CrossMamba [150], CMS2I-Mamba [151], COMO [152], DMM [153], Semi-Mamba [154], MetaEarth [18], Text2Earth [155], Changen2 [156], LDS2AE [157], CRS-Diff [158], DMDiff [159], MaskApproxNet [160], RSDiff [161], EarthSynth [162], Diffusion-Geo [163], ChangeCLIP [16], RemoteCLIP [17], FarmSeg_VLM [164], FSVLM [165], Falcon [166], FedRSCLIP [167], GeoRSCLIP [168], Co-LLaVA [169], CLIP-MoA [170], RS-MoE [171], EarthGPT [59], RingMoGPT [63], SkyEyeGPT [172], RS-LLaVA [173], LHRS-Bot [174], EarthMarker [175], GeoRSMLLM [176], GeoPix [177], CCExpert [178], LHRS-Bot-Nova [179], MMFNet [180], HATF [181], MFNet [182], DAAB [183], X-ModalNet [184], PromptMID [185], Mamba-diffusion [186], BEMSeg [187], MF-BHNet [188], VSGNet [189].

4.1. CNN Backbones

CNN is a typical DL model designed for processing grid data (such as RS images, videos, time-series signals, etc.), and its core advantages are reflected in local feature perception, convolutional kernel parameter sharing, and hierarchical feature extraction [128]. In particular, CNN has shown excellent performance in processing remote sensing images. It can effectively capture local features, such as edges and textures, while reducing the computational complexity of the model [129]. At the same time, it can extract multi-scale semantic features through multi-layer convolution and pooling, which has been gradually applied to the analysis of multimodal remote sensing data [130]. One of the key challenges in building an MM-RSFM based on CNN backbone is how to effectively integrate the information from different modal data. Common fusion strategies include input-layer fusion, middle-layer fusion, output-layer fusion, or close-location fusion [131,132]. The core value of CNN in the field of remote sensing is to efficiently extract spatial features and adapt to multi-source data. As is illustrated in Figure 8, the workflow of MM-RSFMs based on CNN backbones mainly includes data input, multi-scale feature extraction, multimodal feature fusion and representation, and task output. By combining geographical coordinates, time series, and other knowledge, it can further improve the comprehensive performance of MM-RSFMs.

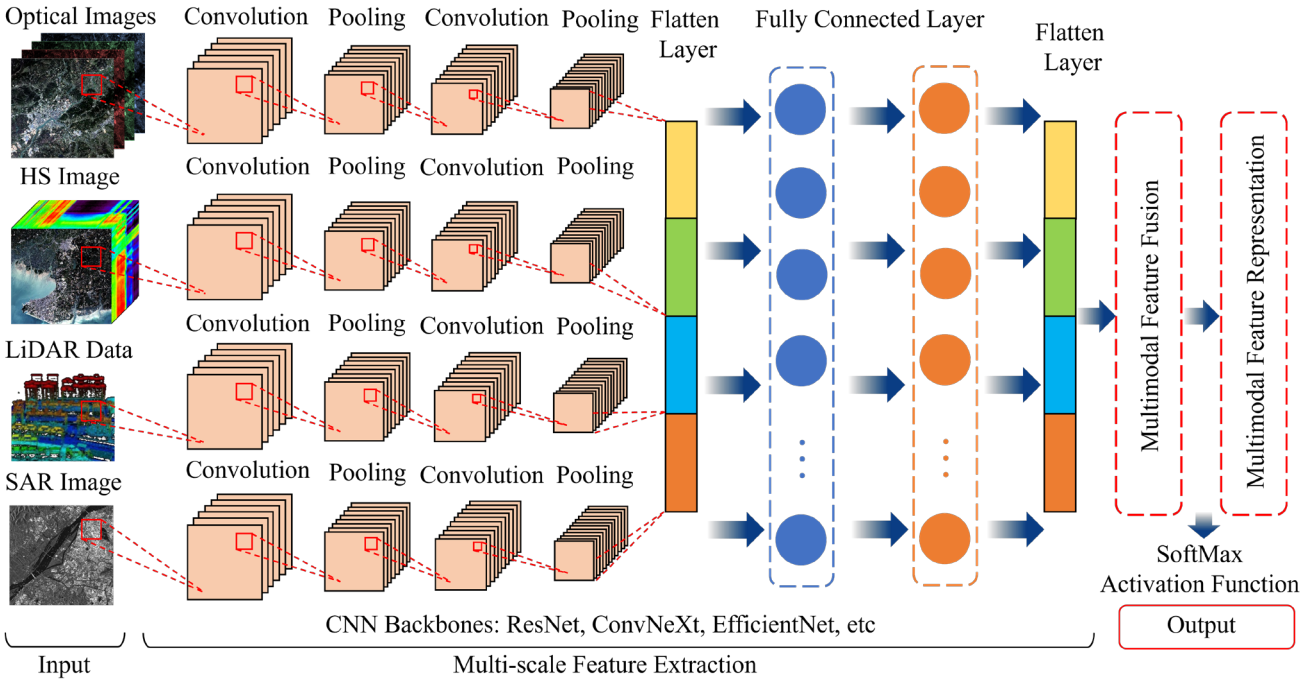


Figure 8. The framework of the CNN backbones for MM-RSFMs.

In recent years, some CNNs designed specifically for MM-RSFMs have emerged, achieving significant cross-modal fusion and feature extraction capabilities. Wu et al. proposed CCR-Net for classification of multimodal remote sensing data including hyperspectral, LiDAR, and SAR data, which learned the fusion representations through cross-modal reconstruction strategy [133]. In order to accomplish efficient discrimination and localization for multimodal RSIs, Uss et al. designed the DLSM model using a two-channel patch matching network [134]. Regarding the registration of multimodal RSIs, Zhang et al. proposed a Siamese network to learn descriptors for multimodal image patch matching [135]. With the aim of effectively integrating complementary information for scene parsing of multimodal RSIs, Zhou et al. introduced CEGFNet, which captures both high-level semantic features and low-level spatial details [136]. To address the few-shot issue in RS image captioning,

Zhou et al. developed a few-shot image captioning framework, which can train a simple foundation model with limited caption-labeled samples [137].

4.2. Transformer Backbones

As an attention mechanism model tailored for processing sequential data, the Transformer, with its powerful sequence modeling and global dependency modeling capabilities, has significant advantages in processing RS data that can be viewed as pixel sequences or image patch sequences [138]. Unlike Section 2, which introduces the working principles and attention mechanisms of the Vision Transformer, this subsection focuses on the characteristics and progress of the Transformer as a backbone network. Currently, MM-RSFMs based on Transformer backbones are one of the most cutting-edge and promising research directions in the field of RS, combining self-supervised learning and cross-modal alignment techniques with the multimodal and large-scale characteristics of RS data. Due to the fact that the Transformer does not limit the type of input data, different modalities (image patches, radar signals, text descriptions, time series, and even geographic coordinates) can be mapped to a unified vector space through appropriate embedding layers to achieve multimodal fusion [139]. As shown in Figure 9, the MM-RSFM based on mainstream Transformer backbones involve technologies such as data preprocessing and representation, Masked Image Modeling, encoder and decoder, self-supervised pre-training, multi-head self-attention mechanism, fine-tuning, knowledge transfer, and multi-task collaboration.

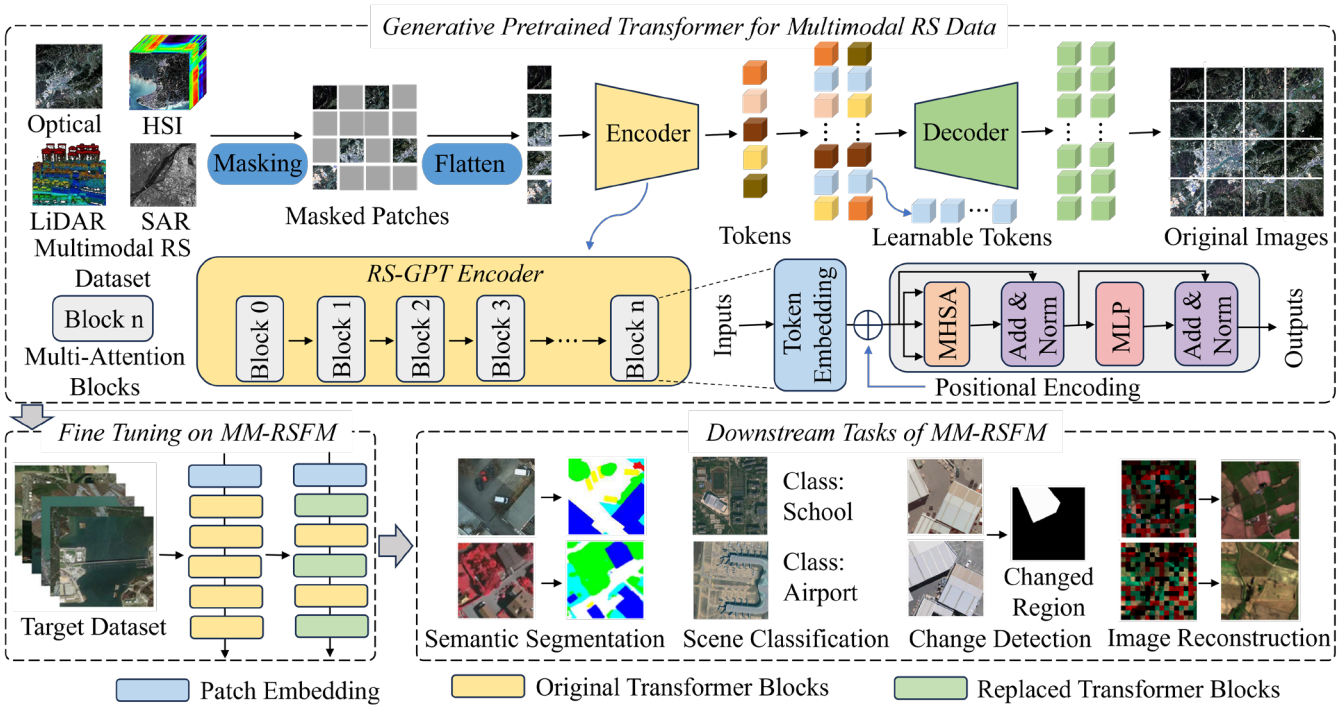


Figure 9. The overall workflow of the MM-RSFM based on the Transformer backbones.

Before the emergence of Transformer backbones, recurrent neural networks (RNNs) and their variants were the mainstream models for processing sequential data. However, RNNs suffer from gradient vanishing and difficulties in parallel computation. The attention mechanism in Transformers, especially self-attention and multi-head attention, effectively addresses these two issues, offering advantages such as dynamic weight allocation, modeling of global dependency relationships, and support for parallel computing. Despite the rapid development of the RSFMs based on Transformer backbones, fusing multimodal RS data to perform downstream tasks remains a challenge. To address the problem, Wang et al. [140] introduced a lightweight multimodal data fusion network for

digital surface model (DSM), RGB, and Near-Infrared (NIR) data, employing a multi-branch ViT to reconstruct and fuse multimodal features. Addressing the issue of achieving precise semantic segmentation of high-resolution RSIs, Feng et al. [141] created a multimodal fusion Transformer-based DeepLabv3+ model to acquire multimodal and multi-scale features from two modalities. For the sake of enhancing the fidelity of spatial details when fusing HSIs with MSIs, Zhu et al. [142] designed an implicit Transformer fusion GAN that integrated the advantages of the continuity perception mechanism and the self-attention mechanism. To tackle the problem of data redundancy in multimodal land-cover classification, Zhang et al. [143] proposed the Multimodal Informative ViT, which significantly reduced redundancy in the empirical distribution of separate and fused features in each modality. In order to capture the correlations and complementarity between different modalities in land-cover classification tasks, Xu et al. [144] introduced the spatial–spectral residual cross-attention Transformer, which makes full use of the multimodal features.

4.3. Mamba Backbones

Mamba is a sequence modeling architecture based on the selective state space model (SSM), which has long-range modeling capability and linear computational complexity for multimodal RS sequence data [145,146]. Undoubtedly, it is conducive to solving the bottleneck of computational efficiency and multi-scale feature fusion faced by traditional Transformers in high-resolution and long-sequence RS data. It is feasible to use parallel modules of spectral Mamba and spatial Mamba to achieve cross-modal feature representation and fusion [147]. Specifically, the MM-RSFM based on the Mamba backbone can employ various cross-modal interaction and self-adaptative scale modules to execute downstream tasks [148,149]. However, there is an urgent need for lightweight and unified multimodal framework. As is illustrated in Figure 10, the workflow of the MM-RSFMs based on Mamba backbones mainly includes data input, multimodal feature extraction, multimodal feature fusion, model pre-training, and downstream tasks.

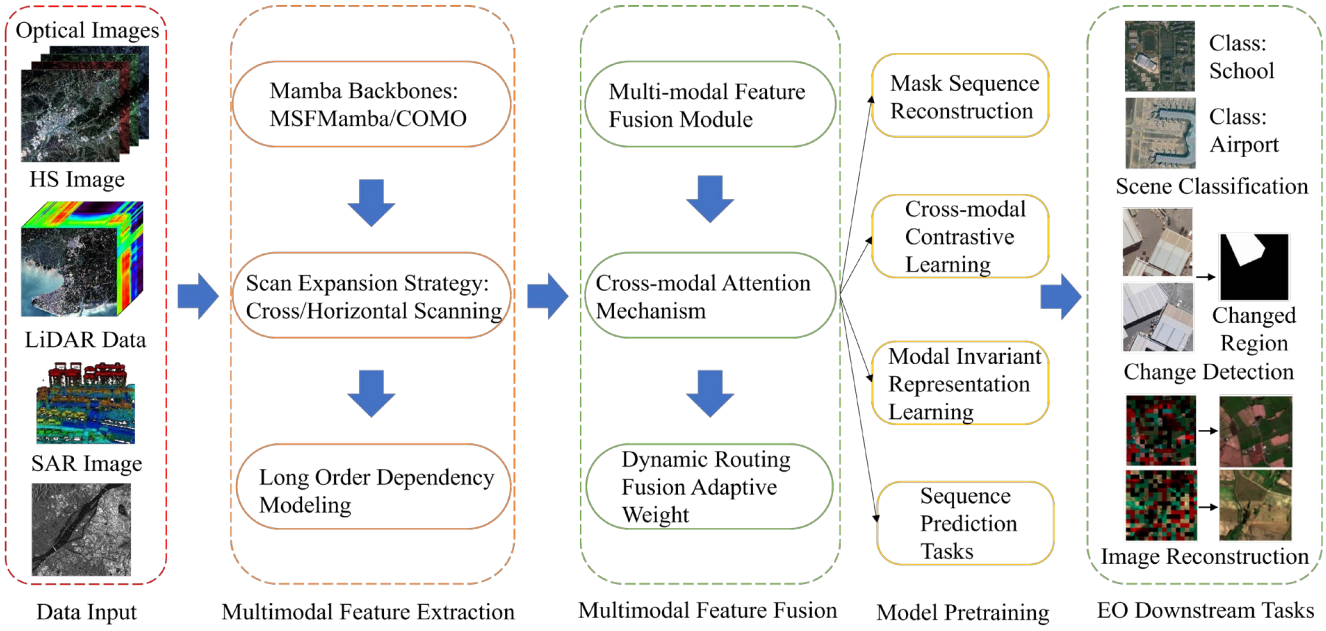


Figure 10. The overall workflow of the MM-RSFM based on the Mamba backbones.

Although the MM-RSFMs based on CNN and Transformer backbones have been successfully applied to a series of multimodal RS tasks, their practical performance is constrained by model complexity and computational efficiency. To address the above

issues, Zhang et al. introduced a multimodal fusion network based on the SSM for multimodal RSI classification, called S2CrossMamba, which is composed of the dual-branch spectral Mamba and spatial Mamba [150]. In order to achieve high-quality and efficient RSI fusion classification, Ma et al. proposed a cross-modal spatial–spectral interaction Mamba to capture global long-range dependencies and spatial–spectral features [151]. Regarding the problem of multi-sensor image matching in multimodal change detection, Liu et al. designed a cross-Mamba interaction and offset-guided fusion framework to capture multimodal features and reduce computational over-head [152]. For the challenges of multispectral-oriented object detection, Zhou et al. developed a disparity-guided multispectral Mamba that can adaptively merge cross-modal features and enhance feature representation [153]. To make full use of the representational complementarity and consistency between different modalities, Li et al. introduced a semi-supervised framework tailored for high-dimensional multimodal data fusion, achieving cross-modal learning and pixel-level annotations [154].

4.4. Diffusion Backbones

As a type of generative model based on probability diffusion process, Diffusion models learn data distribution through an iterative process of first adding noise and then removing noise [155]. Multimodal RS data, including text, images, videos, audio, and position, all contain noise, which enables the MM-RSFMs based on Diffusion backbones, with gradual denoising capabilities, to uniformly process these heterogeneous RS data and find a unified representation in the latent space, as illustrated in Figure 11. Consequently, it is beneficial for the MM-RSFMs to generate high-quality samples and perform model pre-training by employing Diffusion backbones [156,157]. The CRS-Diff is a typical two-stage RS generative model that simultaneously supports the control inputs of the text condition, metadata condition, and image condition [158]. However, the high computational complexity and difficult model optimization are major challenges in processing multimodal RS big data.

Although the MM-RSFMs based on Diffusion backbones have shown great potential and superiority in generating images, text, and other RS data modalities in recent years, there are still many problems to be solved in multimodal RS tasks. In order to tackle the problems of traditional RS cloud removal methods in maintaining image texture details and visual authenticity, Zhang et al. proposed a dual-branch multimodal conditionally guided Diffusion model to extract features adaptively based on the characteristics and differences of multimodal RS data [159]. For the limited generalizability and adaptability of previous models, Sun et al. introduced a multimodal RS change description method based on the Diffusion model, which consists of multi-scale change detection and frequency-guided complex filter modules [160]. To generate high-resolution satellite images from textual prompts, Sebaq et al. designed a two-stage Diffusion model to merge text and image embeddings within a shared latent space [161]. Regarding the scarcity of RS-labeled data, Pan et al. developed a generative foundation model based on the Diffusion backbone that can synthesize multi-category and cross-satellite labeled data for RSI interpretation tasks [162]. To address the issue of generated images lacking spatial and semantic constraints, Cai et al. proposed a controllable text-to-image generative model using the Diffusion backbone, which involves two stages: creating an image–text generation dataset and efficient tuning [163].

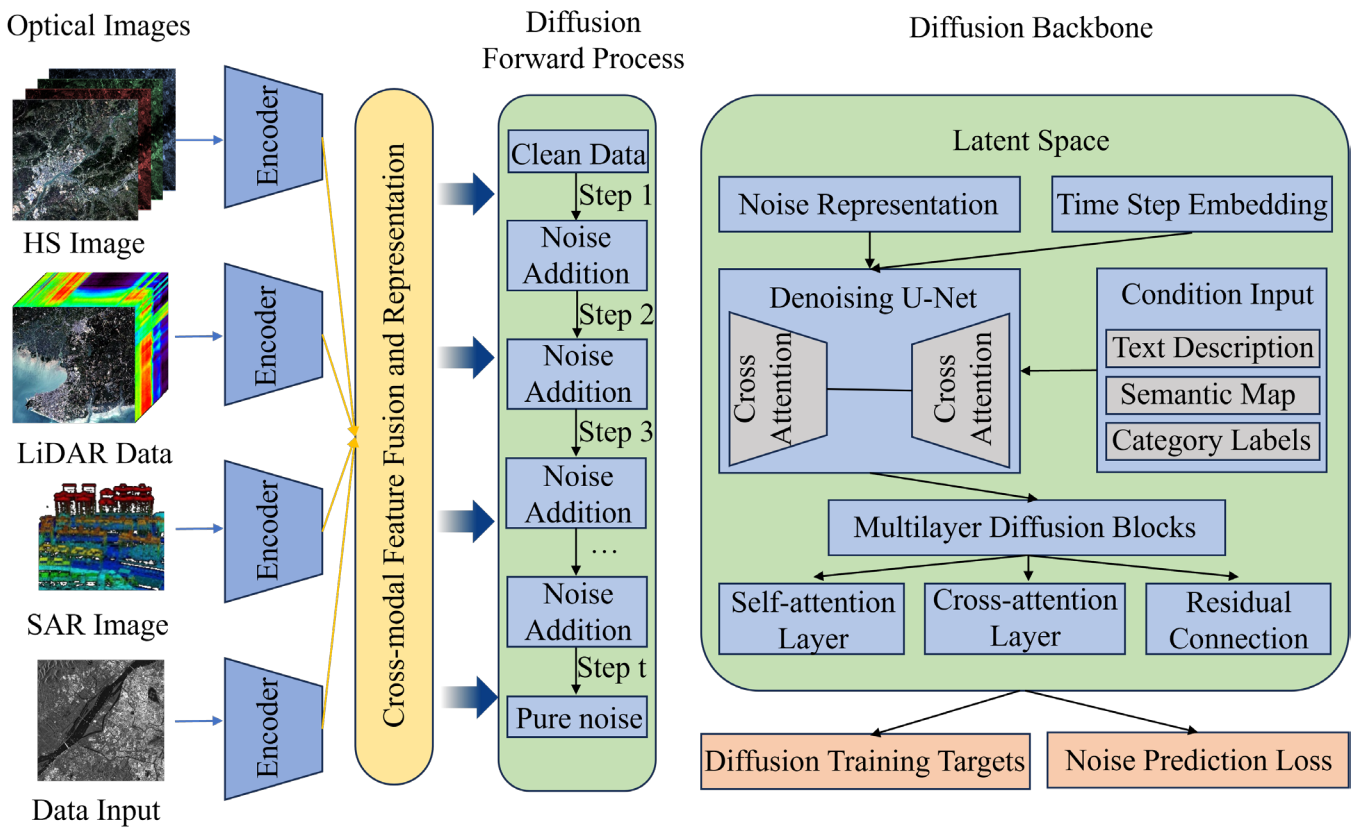


Figure 11. The architecture of the Diffusion backbone for MM-RSFMs.

4.5. VLM Backbones

VLMs are advanced multimodal AI models that combine the capabilities of computer vision and natural language to simultaneously process image and text data, while learning how to connect them logically [164]. The MM-RSFMs driven by VLM backbones are reshaping the paradigm of RS image interpretation and geographic information acquisition. Their core breakthrough lies in injecting human prior knowledge into multimodal AI systems through natural language and RS vision [165]. Due to the significant differences between RS images and natural images, applying VLM such as CLIP directly yields poor results. Therefore, how to modify VLMs to adapt to the MM-RSFMs is a key challenge, and at the same time, the problem of RS data lacking text descriptions urgently needs to be solved. As is illustrated in Figure 12, the workflow of the MM-RSFMs based on VLM backbones mainly includes data input, vision and text encoding, cross-modal feature interaction, model pre-training, and EO downstream tasks. The Falcon is an RS vision–language foundation model that possess powerful understanding and reasoning abilities at the image, region, and pixel levels [166].

The research on VLMs in the field of multimodal RS is still in its early stages, introducing natural language as a unified semantic interface to achieve cross-modal understanding and interactive analysis of RS data. To address the challenges of data heterogeneity and large-scale model transmission in traditional federated learning approaches, Lin et al. introduced a federated learning framework based on a VLM for RS scene classification [167]. In order to utilize existing large-scale pre-trained VLMs to perform RS downstream tasks, Zhang et al. proposed a large-scale image–text paired dataset and a large VLM method [168]. To mitigate hallucination phenomena and reduce computational burden, Liu et al. presented a VLM method for RS visual question answering tasks [169]. Given the remarkable domain gap between the pre-training data of CLIP and RS, Fu et al. designed an RS contrastive language–image pre-training (CLIP) framework with a learnable mixture of

adapters (MoA) module [170]. Regarding RS image captioning tasks, Lin et al. introduced a VLM method based on the mixture of experts (MoE), which consists of a novel instruction router and multiple lightweight large language models (LLMs) [171].

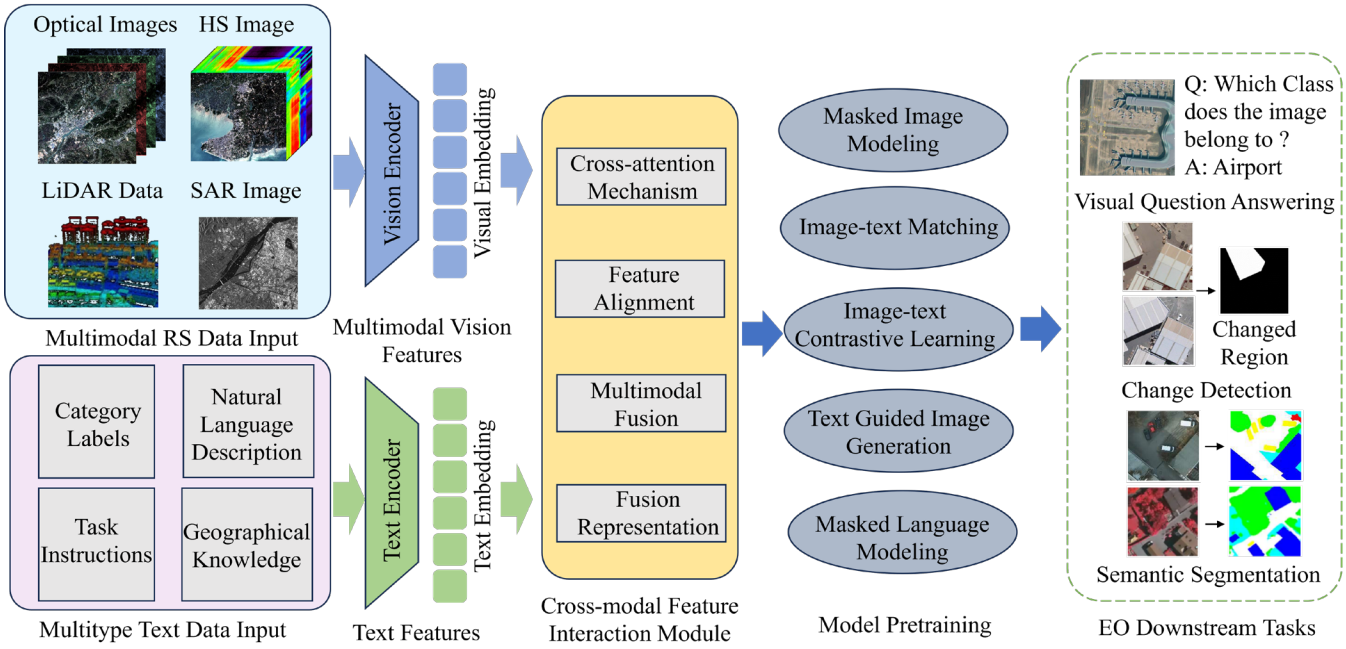


Figure 12. The overall workflow of the MM-RSFM based on the VLM backbones.

4.6. MLLM Backbones

MLLMs further enhance the multimodal information processing capabilities on the basis of traditional LLMs, which exhibit stronger generalization and complex reasoning abilities compared to previous multimodal methods such as CNN, Transformer, Mamba, Diffusion, and VLM [172]. Introducing MLLMs into the field of MM-RSFM can achieve deep understanding, reasoning, and decision-making of multimodal RS data [173,174]. The core architecture of MM-RSFM based on MLLM backbones mainly consists of four parts: multimodal data input, multimodal encoder, LLM, and task executor, among which the LLM is the cognitive core, as illustrated in Figure 13. The LLM of EarthMarker, an RS MLLM, is composed of a sharing visual encoding module and a text tokenizer module [175].

The MM-RSFMs based on MLLM backbones revolutionarily upgrades RS data interpretation from model optimization to human–machine dialog and autonomous agents. However, there are still many key issues that urgently need to be addressed. In order to effectively apply visual prompting to complicated RS scenes, Zhang et al. developed the first visual prompting-based RS MLLM named EarthMarker, that is capable of interpreting RS imagery at the image, region, and point levels [175]. For the difficulties of VLMs in complex instruction or pixel-level tasks, Zhang et al. proposed an RS vision–language task set and unified data representation [176]. To endow RS MLLMs with pixel-level dialog capabilities, Ou et al. designed a novel RS MLLMs called GeoPix that extends image understanding capabilities to the pixel level [177]. Regarding RS image change captioning tasks, Wang et al. introduced a new RS MLLM framework to capture multi-scale differences between bi-temporal images [178]. In order to efficiently cope with RS vision–language interpretation tasks, Li et al. proposed the LHRS-Bot-Nova model based on MLLM to expertly perform various RS understanding tasks aligned with human instructions [179].

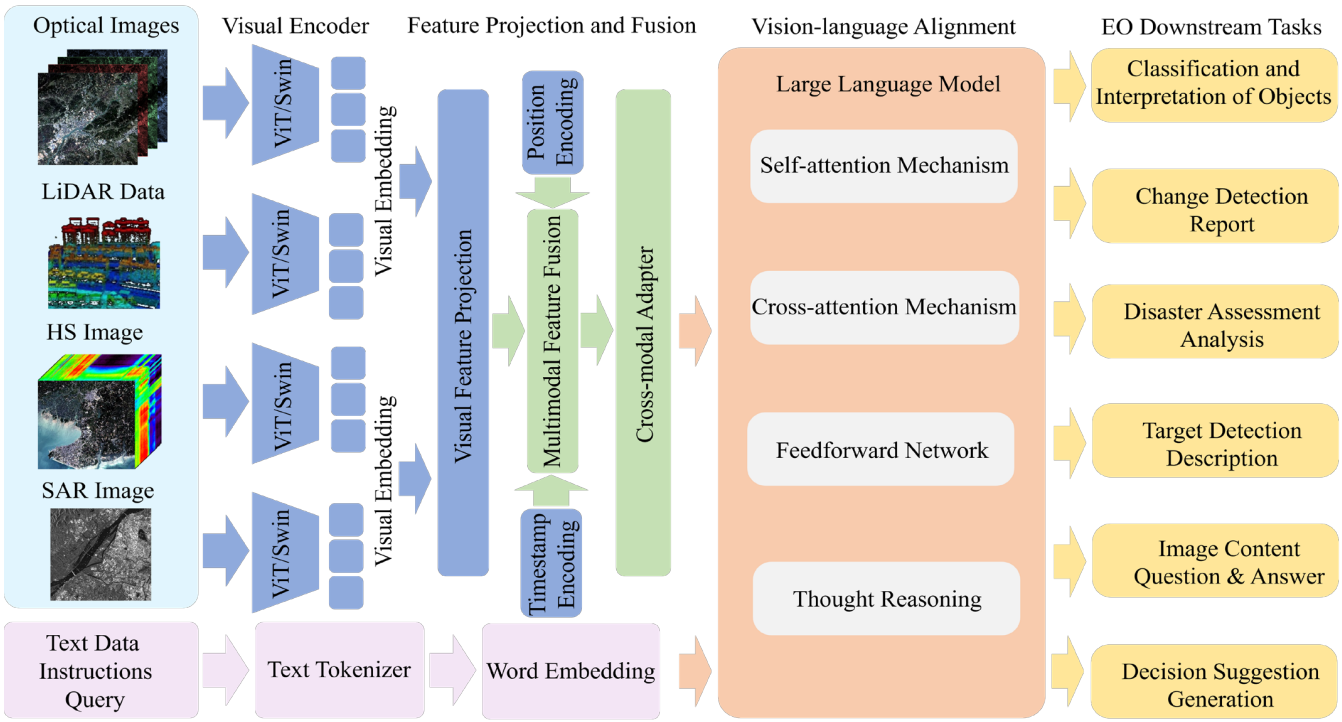


Figure 13. The architecture of the MM-RSFM based on the MLLM backbones.

4.7. Hybrid Backbones

The backbones of MM-RSEMs include, but are not limited to the following categories: CNN, Transformer, Mamba, Diffusion, VLM, and MLLM, each of which has its own advantages and disadvantages. In order to fully leverage the advantages of multiple backbones, many scholars usually employ hybrid backbones to develop and train a unified and generic MM-RSFM [180–184]. As shown in Figure 14, the architecture of the MM-RSFM based on hybrid backbones mainly consists of multimodal RS data input, cross-modal feature extraction and fusion using hybrid backbones, and EO downstream tasks. The PromptMID integrates Diffusion and Transformer models to achieve optical and SAR image matching [185].

As is well known, maintaining geometric consistency in multimodal RS semantic segmentation faces challenges. To tackle this issue, Du et al. designed a Mamba–Diffusion hybrid framework to preserve geometric consistency in segmentation masks [186]. To maintain semantic consistency in multimodal RS semantic segmentation, Liu et al. developed a bidirectional feature fusion and enhanced alignment-based multimodal semantic segmentation model for fusion and alignment of image and text information [187]. To effectively fuse the complementary features of SAR and optical images, Wang et al. introduced a hybrid multimodal fusion network for building height estimation to mine cross-modal feature and intermodal correlation [188]. Regarding the differences in image acquisition mechanisms between different sensors, Pan et al. proposed a multimodal fusion framework based on CNN and ViT for multimodal RS semantic segmentation [189].

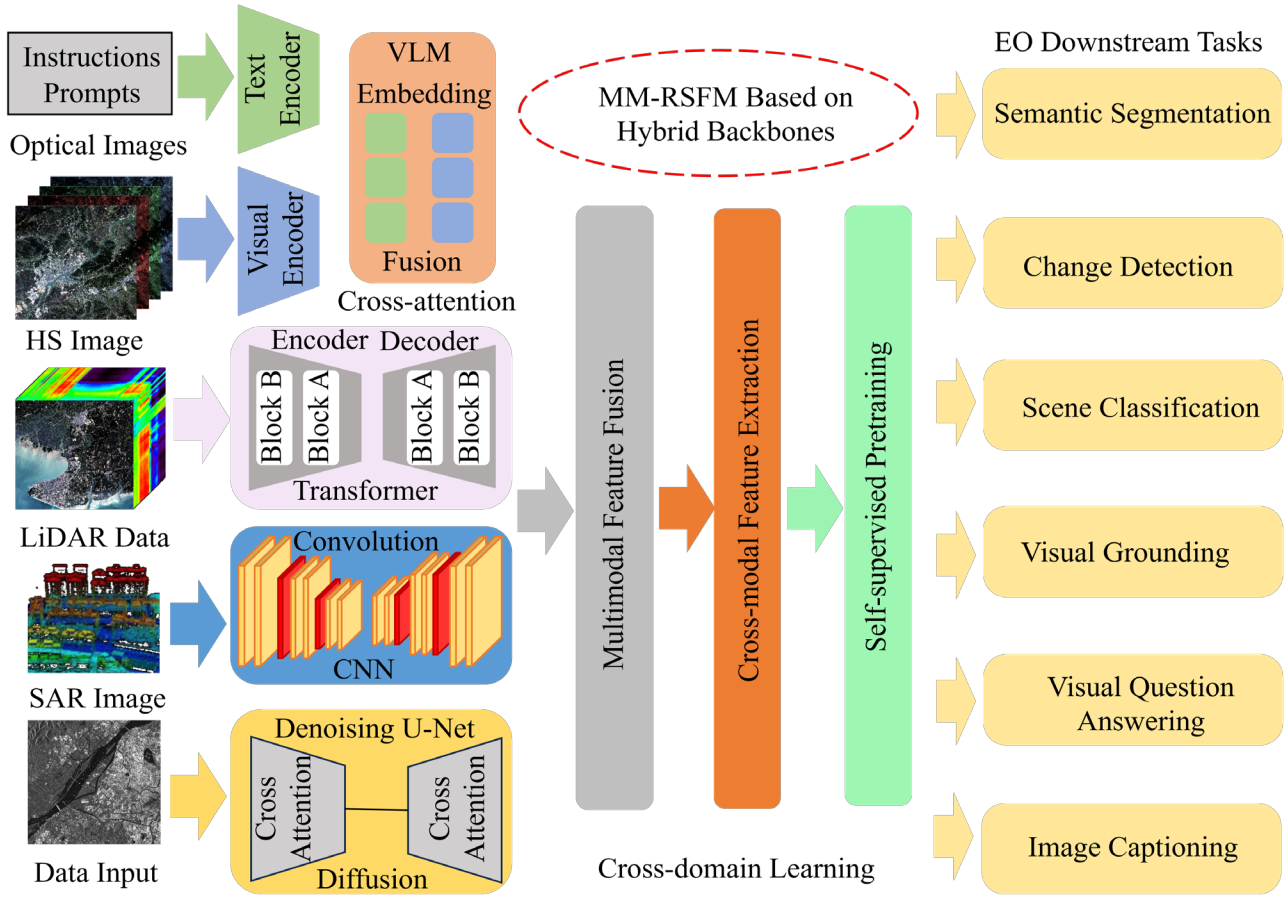


Figure 14. The architecture of the MM-RSFM based on the hybrid backbones.

5. Vision–X: MM-RSFMs

The MM-RSFMs take data from different modalities such as text, audio, video, optical images, SAR, LiDAR, etc., as the objects for learning and reasoning, and utilize technologies such as multimodal representation learning, modality fusion and alignment, cross-attention mechanism, multi-task pre-training, and large-model fine-tuning to achieve feature learning for cross-modal data and EO downstream tasks [190,191], including but not limited to semantic segmentation, change detection, scene classification, and visual positioning. Due to the significant differences in imaging sensors, spatial scale, target types, and background noise among different modalities of remote sensing data, compared with large language models like ChatGPT, building a generic MM-RSFM has encountered many difficulties. For example, embedding the complementary information of multimodal data into a unified feature representation and integrating various visual analysis tasks into a universal model framework. As is illustrated in Figure 15, this article classifies the MM-RSFMs according to the cross-modal interaction method of vision–x, such as vision–vision, vision–language, vision–audio, vision–position, and vision–language–audio.

5.1. Vision–Vision RSFMs

5.1.1. Optical–SAR RSFMs

Optical data contains rich spectral information, while SAR data has abundant geometric information. Both visual modal data have been used by multiple cross-modal RSFMs for pre-training and multi-scale feature fusion. RingMo-SAM [14] is a multimodal RSI segmentation foundation model based on the SAM model. This model achieves precise segmentation of optical and SAR RSIs through a category-decoupled mask decoder and

a multimodal feature hint encoder and can identify the categories of segmented objects. This model has good generalization ability and segmentation accuracy in complex RS scenes and multi-object segmentation tasks and also supports SAR feature hints to improve the segmentation performance of SAR images. SkySense [15] is an MM-RSFM with a billion-parameter scale that can handle optical and SAR modal data simultaneously. It is pre-trained using 2.15 million time-series RS data. The model inputs time-series optical and SAR RSIs into a multimodal spatio-temporal encoder, uses multi-granularity contrastive learning technology to learn cross-modal feature representations, and uses geo-text data to enhance the representation ability of the MM-RSFMS. Its model accuracy and performance outperform the existing 16 RSFMs in seven downstream tasks, and it can also flexibly adjust the module architecture according to diverse downstream tasks. The S2FL model [22] solves the problem of feature learning for heterogeneous data by decomposing optical and SAR multimodal RS data into modality-shared and modality-specific components, enabling effective complementation of multimodal data features. This model performs well in the land-cover classification task of three multimodal RS benchmark datasets.

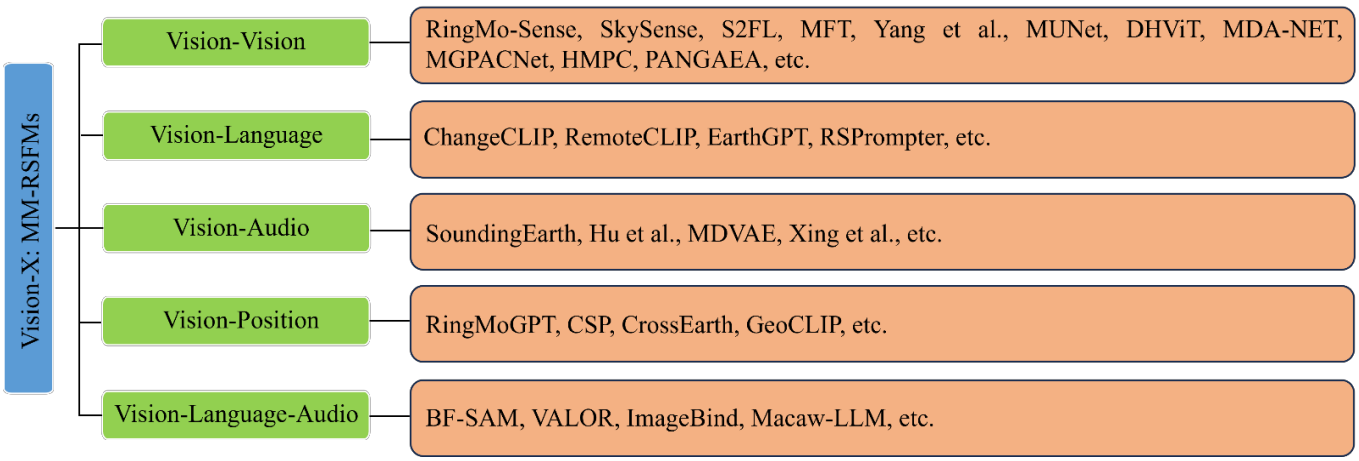


Figure 15. A chain of the types of MM-RSFMs. RingMo-Sense [13], SkySense [15], S2FL [22], MFT [44], Yang et al. [192], MUNet [193], DHViT [194], MDA-NET [195], MGPACNet [196], HMPC [197], PANGAEA [198], ChangeCLIP [16], RemoteCLIP [17], EarthGPT [59], RSPrompter [199], SoundingEarth [60], Hu et al. [200], MDVAE [201], Xing et al. [202], RingMoGPT [63], CSP [203], CrossEarth [204], GeoCLIP [205], BF-SAM [206], VALOR [207], ImageBind [208], Macaw-LLM [209].

5.1.2. Optical–LiDAR RSFMs

In recent years, the fusion of high-spectral images and LiDAR has been widely used to improve the classification accuracy of RSIs. The modality fusion vision Transformer model [192] consists of a multimodal cross-attention module and a spectral self-attention module, which can partially solve the heterogeneous feature fusion and category representation problems of HSI and LiDAR and fully extract the spectral and spatial features of optical and LiDAR data. MUNet [193] is a brand-new high-spectral image multimodal unmixing network model. This model considers the height difference in LiDAR data through the attention mechanism and integrates the spatial information of LiDAR to assist unmixing, improving the hyperspectral unmixing accuracy in complex scenes. DHViT [194] develops a multimodal deep hierarchical visual Transformer architecture to achieve joint classification of high-spectral and LiDAR data. The model first processes hyperspectral images using a spectral sequence Transformer and extracts hierarchical spatial features from high-spectral and LiDAR data using a spatial hierarchical Transformer. Additionally, the cross-attention feature fusion mode in this model adaptively and dynamically fuses heterogeneous features from multimodal data, improving the joint classification accuracy

of multimodal RS data. MDA-NET [195] is a multi-level domain-adaptation network for promoting information collaboration and cross-scenario classification. This network builds adapters related to HSI and LiDAR modalities and uses mutual classifiers to aggregate feature information captured from different modal data to improve cross-scenario classification performance.

5.1.3. Optical–SAR–LiDAR RSFMs

CNN and SA are efficient techniques for multimodal RS data fusion. The MFT model [44] provides a multimodal RS data fusion scheme for land-cover classification problems. This model uses multi-head cross-patch attention mechanism and visual Transformer to enhance the fusion and classification capabilities of HSI, MSI, SAR, and LiDAR multimodal RS data and employs a novel multimodal feature merging token. The model demonstrates effectiveness and high accuracy in four public multimodal RS datasets. MG-PACNet [196] is a multi-scale geometric prior perception cross-modal network for RSI fusion classification. This network creates a geometric prior feature-enhanced residual module to extract boundary features from HS, SAR, and LiDAR multimodal RS data, and through a multi-scale global–local spatial–spectral feature extraction module, it extracts rich semantic information. Finally, a dual-attention fusion module is designed to enhance the complementarity of heterogeneous data. HMPC [197] provides a solution for multimodal RS data matching to address noise and nonlinear intensity differences, namely, the new feature descriptor of the Maximum Phase Consistency Histogram, which is based on the structural characteristics of the image, using a similarity calculation formula, a precise bilateral matching principle, and a consistency-checking algorithm for matching, demonstrating robustness in matching complex texture structures and noisy images. PANGAEA [198] establishes a highly scalable evaluation benchmark framework of the geographic spatial foundation model, which covers diverse datasets, tasks, resolutions, sensor modes, and temporal characteristics, achieving geographical globality and task diversity of existing benchmarks.

5.2. Vision–Language RSFMs

To enhance the advanced feature learning and zero-shot application capabilities of generic RSFMs, text data other than image modalities can be added, which reduces the amount of annotation data during fine-tuning and enables the model to acquire better language understanding and reasoning capabilities. Vision–language foundation models (such as CLIP) provide a new paradigm for MM-RSFMs. ChangeCLIP [16] is a vision–audio foundation model specifically tailored for RS change detection. This model re-creates the original CLIP to extract dual-temporal features of images and text and through a differential feature compensation module to capture detailed semantic changes between images and text, thereby achieving a comprehensive application of image–text semantic features and visual features and obtaining excellent accuracy in six datasets. RemoteCLIP [17] is the first vision–language foundation model for RS tasks which can simultaneously learn rich semantic and visual features. This model uses data scaling technologies to unify heterogeneous annotations into an image caption data format and further integrates drone images, solving the problem of scarce pre-training data and performing well in downstream tasks. EarthGPT [59] provides a general multimodal large language model for multi-sensor RS interpretation tasks. This model constructs a visual-enhanced perception mechanism, proposes cross-modal mutual understanding methods and instruction adjustment methods for multi-sensor, multi-task processing, enhances the interaction between visual perception and language understanding, and demonstrates superior performance in scene classification, image captioning, visual question answering (VQA), and other tasks. RSPrompter [199]

is an RSI automatic instance segmentation method based on SAM foundation model and prompt learning. This method combines the visual features of the image and semantic category information of images and learns to generate appropriate prompts for RSIs to generate semantically distinguishable segmentation results.

5.3. Vision–Audio RSFMs

Many DL-based RSI interpretation models rely on manually annotated image datasets for the supervised pre-training of the backbone network. The cross-modal collaboration of image modalities and audio modalities provides a new paradigm for the self-supervised pre-training of RS models. SoundingEarth [60] is a multimodal unlabeled dataset containing images and audio, consisting of co-located aerial images from around the world and crowd-sourced audio samples. Pre-training the ResNet model using this dataset enables mapping of visual and audio samples to a common embedding space through self-supervised methods, enhancing the multimodal feature representation and reasoning capabilities. Hu et al. proposed a two-stage progressive learning framework that only uses the correspondence between audio and vision to locate and identify sound objects in complex audio–visual scenes, achieving good performance in both localization and recognition tasks, as well as unsupervised object detection [200]. MDVAE [201] is a multimodal dynamic VAE applied to unsupervised audio–visual speech representation learning. This model introduces static latent variables to encode the time-invariant information in audio–visual speech sequences and effectively combines audio and visual information in an unsupervised manner with excellent model accuracy. In addition to training the base model from scratch, cross-modal representations of vision and audio can be achieved through optimization and fine-tuning of existing models. Xing et al. proposed a multimodal model based on an optimized cross-visual–audio and joint vision–audio generation framework, which consists of a multimodal latent aligner with a pre-trained ImageBind model, achieving outstanding performance on tasks such as joint video–audio generation, vision-guided audio generation, and audio-guided vision generation [202].

5.4. Vision–Position RSFMs

Images with geographic information labels are a type of important visual data, but the large amount of geographic tagging work is time-consuming and labor-intensive. Developing cross-modal foundation models for vision and position data can effectively solve this problem. RingMoGPT [63] is a unified MM-RSFM for vision, language, and position. This model is based on the idea of domain adaptation, specifically through location- and instruction-aware Q-Former and a change detection module for object detection and change captions. Its pre-training dataset contains over 500,000 high-quality image–text pairs, and the instruction-tuning dataset contains over 1.6 million question–answer pairs. The model has good performance and generalization ability in six downstream tasks, including scene classification, object detection, visual question answering, image captioning, basic image captioning, and change captioning. CSP [203] is a contrastive space pre-training model for geotagged images. This model uses self-supervised methods to directly learn rich geographic spatial information during pre-training, fine-tuning, and inference stages. Specifically, it uses dual encoders to separately encode images and their corresponding geographical location and uses a contrastive target to learn effective position representations from images, demonstrating significant model performance. CrossEarth [204] is a visual foundation model with strong cross-domain generalization ability. This model performs visual tasks through a specially designed data-level Earth-style injection pipeline and a model-level multi-task training pipeline. Moreover, for semantic segmentation tasks, the model outperforms existing state-of-the-art methods on a comprehensive benchmark across

different regions, spectral bands, platforms, and climates. GeoCLIP [205] is an image-GPS retrieval method based on the CLIP backbone. This method enforced alignment between images and their corresponding GPS locations by using random Fourier features for position encoding and constructs a hierarchical representation that captures information at different resolutions. This model is the first to use GPS encoding for geographic positioning.

5.5. Vision–Language–Audio RSFMs

This model mainly uses vision modality data and combines non-vision modality data for multimodal learning and representation, thereby enhancing the adaptability and generalization of RSFMs in different application scenes [206]. For example, combining vision data with semantically consistent text and audio data. VALOR [207] is a vision–audio–language perception pre-training model for multimodal understanding and generation. This model jointly models the three modalities in an end-to-end manner, specifically pre-training through multimodal grouping alignment (MGA) and multimodal grouping caption (MGC) tasks. This model has achieved state-of-the-art performance in a series of public cross-modal benchmark tests and has been extended to various downstream tasks such as retrieval, captioning, and question answering. ImageBind [208] is a method for learning joint embeddings across six modalities (image, text, audio, depth, heat, and IMU data). Training such joint embeddings does not require the combination of all paired data; only image-paired data is sufficient to bind the modalities together. Macaw-LLM [209] is a multimodal large language model (LLM) that can seamlessly integrate visual, audio, and text modalities. This model consists of three parts: the modality module, the cognitive module, and the alignment module. The modality module encodes multimodal data, the cognitive module performs multimodal pre-training for the model, and the alignment module coordinates different representations to seamlessly connect multimodal features to text features.

6. Challenges and Perspectives

With the survey conducted above, a few challenges can be outlined as follows.

6.1. A Scarcity of High-Quality Multimodal RS Datasets

Multimodal RS datasets consist of data acquired by diverse types of sensors operating across various electromagnetic spectral bands, such as visible light, infrared, and microwave, or employing different technical mechanisms like LiDAR, SAR, text, RS videos, and image captions. These datasets are essential for capturing three-dimensional information and spatio-temporal patterns of changes on the Earth's surface. MM-RSFMs rely on large-scale, high-quality multimodal RS data for pre-training. Although numerous public RS data resources and pre-training datasets are available, multimodal RS data in practice exhibit significant diversity in sources, types, and formats. High acquisition and annotation costs, coupled with limited data sharing and access restrictions, further hinder the assembly of such datasets. Moreover, the absence of unified standards and conventions for multimodal RS data complicates the construction of consistent and high-quality pre-training corpora. Variations in modalities, spatial scales, and noise levels among datasets exacerbate these challenges, increasing the difficulty of data preprocessing and cross-modal fusion and alignment. As a result, the effectiveness and generalization capability of MM-RSFMs are considerably impacted.

6.2. Limited Multimodal Feature Extraction Capability

Due to the substantial differences in imaging mechanisms, target types, spatial scales, and background noise among various types of RS data, effectively fusing these modalities and extracting representative cross-modal features through MM-RSFMs remains a major challenge. However, existing public MM-RSFMs still exhibit limitations in cap-

turing semantic, spatial, spectral, and relational features from multimodal pairs such as image–text and image–audio. To enhance the perceptual and reasoning capabilities of MM-RSFMs, particularly in tasks such as change detection and trend analysis, it is essential to perform comprehensive feature extraction from multimodal RS data across multiple scenes within the same geographic area. This process should integrate annotated text, geolocation information, and audio descriptors to establish richer contextual understanding. Furthermore, if MM-RSFMs fail to extract sufficient relevant information or discriminative features from the input data, especially when confronted with complex and dynamic scenarios, their adaptability to new data, unseen environments, and novel tasks will be significantly constrained.

6.3. Weak Cross-Task Generalization Ability

Even with access to large-scale pre-training datasets, MM-RSFMs may still experience overfitting and struggle to transfer and generalize multimodal features to new downstream tasks if the data sources are too restricted or the range of covered scenarios is insufficient. Currently, most RSFMs are designed for single-modality data and specialized tasks, lacking the ability to perform multi-task learning and generalize across diverse scenarios in complex EO applications. Furthermore, while many MM-RSFMs achieve strong results on specific benchmark datasets, their performance often declines significantly when applied to new datasets, and their cross-modal and cross-domain generalization capabilities remain inadequately validated. In practical applications, few-shot and weakly supervised learning approaches are often better suited to meet the demands of handling multiple downstream tasks. At the same time, establishing a unified and accurate evaluation benchmark is essential for comprehensively assessing the generalization capacity of MM-RSFMs.

6.4. The Absence of Unified Evaluation Criteria

With the continuous emergence of various MM-RSFMs, there is growing heterogeneity in their pre-training datasets, backbone architectures, parameter sizes, fine-tuning strategies, and downstream task designs. However, the absence of a unified evaluation benchmark makes it difficult to comprehensively assess their performance and accuracy. Early evaluations of RSFMs employed diverse datasets, tasks, and metrics, resulting in a lack of systematic and consistent evaluation standards. Although initiatives such as SkySense [15] have attempted to establish a unified benchmark across multiple datasets, covering tasks such as single-modal image-level classification, object-level detection, pixel-level segmentation, and multimodal time-series classification, this framework remains limited in handling diverse task types and multimodal data scenarios. Moreover, there is an increasing need to evaluate the generalization capability of pre-trained models under weak supervision settings, which better reflects real-world application requirements. For instance, it is essential to assess model robustness in downstream tasks involving few-shot learning and large volumes of noisy labels.

6.5. Insufficient Security Measures of Foundation Models

The absence of security measures constitutes a significant concern, as it directly undermines the credibility of MM-RSFMs and their feasibility for real-world deployment. Current evaluation criteria primarily focus on performance metrics such as accuracy and efficiency, while lacking systematic standards for assessing robustness, resistance to attacks, and privacy preservation. Major security risks facing MM-RSFMs include, but are not limited to, training data contamination, adversarial attacks on input data, leakage of sensitive information, model theft, and cross-modal security propagation. As MM-RSFMs become more widely adopted, targeted malicious techniques and risk scenarios are expected to proliferate. Consequently, security protection measures must evolve in parallel with ad-

vances in multimodal AI technology. It is therefore imperative to establish a comprehensive security evaluation benchmark for MM-RSFMs, incorporating standardized test sets for data pollution, adversarial examples, privacy breaches, and other vulnerabilities. At the technical level, a multi-layered protection framework spanning terminals, tasks, datasets, and models should be implemented to ensure holistic security.

7. Conclusions and Outlook

MM-RSFMs offer new opportunities for the intelligent interpretation of RSIs and EO downstream tasks. For data acquired from diverse types of sensors, MM-RSFMs are designed to extract comprehensive and representative multimodal spatio-temporal features and semantic information, integrating them effectively to tackle complex and variable EO downstream applications. This article provides a thorough and detailed review of the fundamental principles, key technologies, model backbones, multimodal paradigms, downstream tasks, current challenges, and future research directions pertaining to MM-RSFMs. Our aim is to furnish readers with a profound understanding of the state-of-the-art advances in this rapidly evolving field.

Although recent MM-RSFMs have demonstrated impressive representation and reasoning capabilities, their practical deployment remains constrained by several limitations: the scarcity of high-quality multimodal remote sensing datasets, limited multimodal feature extraction capacity, weak cross-task generalization, the lack of unified evaluation standards, and insufficient security measures. Nevertheless, by integrating diverse information sources, such as visual, linguistic, and positional data, and leveraging self-supervised learning to distill knowledge from massive datasets, MM-RSFMs are advancing remote sensing analysis from task-specific approaches toward general-purpose intelligence. This evolution holds the potential to fundamentally transform how we observe and understand the Earth, providing powerful decision-support capabilities for critical domains including climate change monitoring, sustainable development, and smart city planning. Looking ahead, due to their simplicity and effectiveness, prompt learning and feature adaptation methods represent promising research directions. Their application in RS is expected to further enhance the generalization power and interpretability of MM-RSFMs.

Author Contributions: Conceptualization, G.Z. and L.Q.; methodology, G.Z. and L.Q.; investigation, G.Z., L.Q. and P.G.; resources, G.Z.; writing—original draft preparation, L.Q.; writing—review and editing, G.Z., L.Q. and P.G.; visualization, G.Z., L.Q. and P.G.; supervision, G.Z.; project administration, G.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science of China, grant number 42461050; the Guangxi Surveying and Mapping LiDAR Intelligent Equipment Technology Mid-Test Base, grant number Guike AD23023012; the Guangxi Science and Technology Talent Grand Project, grant number Guike AD19254002; and the Innovation Project of Guangxi Graduate Education, grant number YCSW2025385.

Acknowledgments: The authors have reviewed and edited the manuscript and take full responsibility for the content of this publication. We would like to thank the editors and reviewers for their reviews, which improved the content of this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, G.; Wang, X.; Liu, S.; Wang, Y.; Gao, E.; Wu, J.; Lu, Y.; Yu, L.; Wang, W.; Li, K. MSS-Net: A lightweight network incorporating shifted large kernel and multi-path attention for ship detection in remote sensing images. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *143*, 104805. [CrossRef]

2. Hong, D.; Zhang, B.; Li, H.; Li, Y.; Yao, J.; Li, C.; Werner, M.; Chanussot, J.; Zipf, A.; Zhu, X.X. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sens. Environ.* **2023**, *299*, 113856. [CrossRef]
3. Li, X.; Sun, Y.; Peng, X. TA-MSA: A fine-tuning framework for few-shot remote sensing scene classification. *Remote Sens.* **2025**, *17*, 1395. [CrossRef]
4. Zhou, G.; Zhang, Z.; Wang, F.; Zhu, Q.; Wang, Y.; Gao, E.; Cai, Y.; Zhou, X.; Li, C. A multi-scale enhanced feature fusion model for aircraft detection from SAR images. *Int. J. Digit. Earth* **2025**, *18*, 2507842. [CrossRef]
5. Wu, K.; Zhang, Y.; Ru, L.; Dang, B.; Lao, J.; Yu, L.; Luo, J.; Zhu, Z.; Sun, Y.; Zhang, J.; et al. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. *Nat. Mach. Intell.* **2025**, *7*, 1235–1249. [CrossRef]
6. Yao, D.; Zhi-Li, Z.; Xiao-Feng, Z.; Wei, C.; Fang, H.; Yao-Ming, C.; Cai, W.-W. Deep hybrid: Multi-graph neural network collaboration for hyperspectral image classification. *Def. Technol.* **2023**, *23*, 164–176. [CrossRef]
7. Zhou, G.; Zhi, H.; Gao, E.; Lu, Y.; Chen, J.; Bai, Y.; Zhou, X. DeepU-Net: A parallel dual-branch model for deeply fusing multi-scale features for road extraction from high-resolution remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9448–9463. [CrossRef]
8. Wang, J.; Guo, S.; Huang, R.; Li, L.; Zhang, X.; Jiao, L. Dual-channel capsule generation adversarial network for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5501016. [CrossRef]
9. Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5314–5321. [CrossRef]
10. Lu, H.; Wei, Z.; Wang, X.; Zhang, K.; Liu, H. GraphGPT: A graph enhanced generative pretrained transformer for conditioned molecular generation. *Int. J. Mol. Sci.* **2023**, *24*, 16761. [CrossRef]
11. Duan, Z.; Lu, M.; Ma, J.; Huang, Y.; Ma, Z.; Zhu, F. QARV: Quantization-aware ResNet VAE for lossy image compression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 436–450. [CrossRef]
12. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5608420. [CrossRef]
13. Yao, F.; Lu, W.; Yang, H.; Xu, L.; Liu, C.; Hu, L.; Yu, H.; Liu, N.; Deng, C.; Tang, D.; et al. RingMo-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5620821. [CrossRef]
14. Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A foundation model for segment anything in multimodal remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5625716. [CrossRef]
15. Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024, Seattle, WA, USA, 16–22 June 2024; pp. 27672–27683.
16. Dong, S.; Wang, L.; Du, B.; Meng, X. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 53–69. [CrossRef]
17. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [CrossRef]
18. Yu, Z.; Liu, C.; Liu, L.; Shi, Z.; Zou, Z. MetaEarth: A generative foundation model for global-scale remote sensing image generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1764–1781. [CrossRef]
19. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef]
20. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5607315. [CrossRef]
21. Cha, K.; Seo, J.; Lee, T. A billion-scale foundation model for remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, 1–17. [CrossRef]
22. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, X.X. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 68–80. [CrossRef] [PubMed]
23. Zhou, G.; Qi, H.; Shi, S.; Sifu, B.; Tang, X.; Gong, W. Spatial-spectral feature fusion and spectral reconstruction of multispectral LiDAR point clouds by attention mechanism. *Remote Sens.* **2025**, *17*, 2411. [CrossRef]
24. Sun, X.; Tian, Y.; Lu, W.; Wang, P.; Niu, R.; Yu, H.; Fu, K. From single- to multi-modal remote sensing imagery interpretation: A survey and taxonomy. *Sci. China Inf. Sci.* **2023**, *66*, 140301. [CrossRef]
25. Li, J.; Hong, D.; Gao, L.; Yao, J.; Zheng, K.; Zhang, B.; Chanussot, J. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102926. [CrossRef]

26. Li, Y.; Zhou, Y.; Zhang, Y.; Zhong, L.; Wang, J.; Chen, J. DKDFN: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification. *ISPRS J. Photogramm. Remote Sens.* **2022**, *186*, 170–189. [CrossRef]
27. Zhang, Y.; Yin, M.; Wang, H.; Hua, C. Cross-level multi-modal features learning with transformer for RGB-D object recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7121–7130. [CrossRef]
28. Cai, S.; Wakaki, R.; Nobuhara, S.; Nishino, K. RGB road scene material segmentation. *Image Vis. Comput.* **2024**, *145*, 104970. [CrossRef]
29. Hu, Z.; Xiao, J.; Li, L.; Liu, C.; Ji, G. Human-centric multimodal fusion network for robust action recognition. *Expert Syst. Appl.* **2023**, *239*, 122314. [CrossRef]
30. Li, G.; Lin, Y.; Ouyang, D.; Li, S.; Luo, X.; Qu, X.; Pi, D.; Li, S.E. A RGB-thermal image segmentation method based on parameter sharing and attention fusion for safe autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2023**, *25*, 5122–5137. [CrossRef]
31. Qin, J.; Li, M.; Zhao, J.; Li, D.; Zhang, H.; Zhong, J. Advancing sun glint correction in high-resolution marine UAV RGB imagery for coral reef monitoring. *ISPRS J. Photogramm. Remote Sens.* **2024**, *207*, 298–311. [CrossRef]
32. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral image classification using group-aware hierarchical transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
33. Gong, Z.; Zhou, X.; Yao, W.; Zheng, X.; Zhong, P. HyperDID: Hyperspectral intrinsic image decomposition with deep feature embedding. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5506714. [CrossRef]
34. Zhang, Y.; Zhang, M.; Li, W.; Wang, S.; Tao, R. Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5501312. [CrossRef]
35. Zhou, M.; Huang, J.; Hong, D.; Zhao, F.; Li, C.; Chanussot, J. Rethinking pan-sharpening in closed-loop regularization. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 14544–14558. [CrossRef] [PubMed]
36. Zhou, G.; Yang, F.; Xiao, J. Study on pixel entanglement theory for imagery classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5409518. [CrossRef]
37. Zheng, H.; Li, D.; Zhang, M.; Gong, M.; Qin, A.K.; Liu, T.; Jiang, F. Spectral knowledge transfer for remote sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 4501316. [CrossRef]
38. Zheng, J.; Yang, S.; Wang, X.; Xiao, Y.; Li, T. Background noise filtering and clustering with 3d lidar deployed in roadside of urban environments. *IEEE Sens. J.* **2021**, *21*, 20629–20639. [CrossRef]
39. Farmonov, N.; Esmaeili, M.; Abbasi-Moghadam, D.; Sharifi, A.; Amankulova, K.; Mucsi, L. HypsLiDNet: 3-D–2-D CNN model and spatial–spectral morphological attention for crop classification with DESIS and LIDAR data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11969–11996. [CrossRef]
40. Ma, C.; Shi, X.; Wang, Y.; Song, S.; Pan, Z.; Hu, J. MosViT: Towards vision transformers for moving object segmentation based on Lidar point cloud. *Meas. Sci. Technol.* **2024**, *35*, 116302. [CrossRef]
41. Zhao, S.; Luo, Y.; Zhang, T.; Guo, W.; Zhang, Z. A domain specific knowledge extraction transformer method for multisource satellite-borne SAR images ship detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 16–29. [CrossRef]
42. Yasir, M.; Liu, S.; Mingming, X.; Wan, J.; Pirasteh, S.; Dang, K.B. ShipGeoNet: SAR image-based geometric feature extraction of ships using convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5202613. [CrossRef]
43. Wang, L.; Yang, X.; Tan, H.; Bai, X.; Zhou, F. Few-shot class-incremental sar target recognition based on hierarchical embedding and incremental evolutionary network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5204111. [CrossRef]
44. Roy, S.K.; Deria, A.; Hong, D.; Rasti, B.; Plaza, A.; Chanussot, J. Multimodal fusion transformer for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5515620. [CrossRef]
45. Sun, Y.; Lei, L.; Guan, D.; Kuang, G.; Li, Z.; Liu, L. Locality preservation for unsupervised multimodal change detection in remote sensing imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 6955–6969. [CrossRef] [PubMed]
46. Huang, Z.; Yan, H.; Zhan, Q.; Yang, S.; Zhang, M.; Zhang, C.; Lei, Y.; Liu, Z.; Liu, Q.; Wang, Y. A Survey on remote sensing foundation models: From vision to multimodality. *arXiv* **2025**, arXiv:2503.22081. [CrossRef]
47. Zhang, Y.; Li, Y.; Dang, B.; Wu, K.; Guo, X.; Wang, J.; Chen, J.; Yang, M. Multi-modal remote sensing large foundation models: Current research status and future prospect. *Acta Geod. Cartogr. Sin.* **2024**, *53*, 1942–1954. [CrossRef]
48. Fu, K.; Lu, W.; Liu, X.; Deng, C.; Yu, H.; Sun, X. A comprehensive survey and assumption of remote sensing foundation modal. *J. Remote Sens.* **2023**, *28*, 1667–1680. [CrossRef]
49. Zhang, H.; Xu, J.-J.; Cui, H.-W.; Li, L.; Yang, Y.; Tang, C.-S.; Boers, N. When Geoscience meets foundation models: Toward a general geoscience artificial intelligence system. *IEEE Geosci. Remote Sens. Mag.* **2024**, 2–41. [CrossRef]
50. Yan, Q.; Gu, H.; Yang, Y.; Li, H.; Shen, H.; Liu, S. Research progress and trend of intelligent remote sensing large model. *Acta Geod. Cartogr. Sin.* **2024**, *53*, 1967–1980. [CrossRef]
51. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-language models in remote sensing: Current progress and future trends. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 32–66. [CrossRef]

52. Bao, M.; Lyu, S.; Xu, Z.; Zhou, H.; Ren, J.; Xiang, S.; Li, X.; Cheng, G. Vision Mamba in remote sensing: A comprehensive survey of techniques, applications and outlook. *arXiv* **2025**, arXiv:2505.00630. [CrossRef]
53. Yang, X.; Li, S.; Chen, Z.; Chanussot, J.; Jia, X.; Zhang, B.; Li, B.; Chen, P. An attention-fused network for semantic segmentation of very-high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 238–262. [CrossRef]
54. Hang, R.; Li, Z.; Ghamisi, P.; Hong, D.; Xia, G.; Liu, Q. Classification of hyperspectral and lidar data using coupled CNNs. *IEEE Trans. Geosci. Sens.* **2020**, *58*, 4939–4950. [CrossRef]
55. Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.-S.; Bai, X. iSAID: A large-scale dataset for instance segmentation in aerial images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 28–37.
56. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; Li, X. RSGPT: A remote sensing vision language model and benchmark. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 272–286. [CrossRef]
57. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 5805–5813. [CrossRef]
58. Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
59. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5917820. [CrossRef]
60. Heidler, K.; Mou, L.; Hu, D.; Jin, P.; Li, G.; Gan, C.; Wen, J.-R.; Zhu, X.X. Self-supervised audiovisual representation learning for remote sensing data. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *116*, 103130. [CrossRef]
61. Hu, D.; Li, X.; Mou, L.; Jin, P.; Chen, D.; Jing, L.; Zhu, X.; Dou, D. Cross-task transfer for geotagged audiovisual aerial scene recognition. *arXiv* **2020**, arXiv:2005.08449. [CrossRef]
62. Wang, Y.; Sun, P.; Zhou, D.; Li, G.; Zhang, H.; Hu, D. Ref-Avs: Refer and segment objects in audio-visual scenes. In Proceedings of the IEEE European Conference on Computer Vision (ECCV) 2024, Milan, Italy, 29 September–4 October 2024.
63. Wang, P.; Hu, H.; Tong, B.; Zhang, Z.; Yao, F.; Feng, Y.; Zhu, Z.; Chang, H.; Diao, W.; Ye, Q.; et al. RingMoGPT: A unified remote sensing foundation model for vision, language, and grounded tasks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5611320. [CrossRef]
64. Horn, G.; Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A. The iNaturalist species classification and detection dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8769–8778.
65. Christie, G.; Fendley, N.; Wilson, J.; Mukherjee, R. Functional Map of the World. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6172–6180.
66. Larson, M.; Soleymani, M.; Gravier, G.; Ionescu, B.; Jones, G.J. The benchmarking initiative for multimedia evaluation: MediaEval 2016. *IEEE Multimed.* **2017**, *24*, 93–96. [CrossRef]
67. Li, Y.; Wang, L.; Wang, T.; Yang, X.; Luo, J.; Wang, Q.; Deng, Y.; Wang, W.; Sun, X.; Li, H.; et al. STAR: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *arXiv* **2024**, arXiv:2406.09410. [CrossRef] [PubMed]
68. Chen, H.; Yokoya, N.; Chini, M. Fourier domain structural relationship analysis for unsupervised multimodal change detection. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 99–114. [CrossRef]
69. Zhu, B.; Yang, C.; Dai, J.; Fan, J.; Qin, Y.; Ye, Y. R_2FD_2: Fast and Robust matching of multimodal remote sensing images via repeatable feature detector and rotation-invariant feature descriptor. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5606115. [CrossRef]
70. Li, J.; Zheng, K.; Li, Z.; Gao, L.; Jia, X. X-shaped interactive autoencoders with cross-modality mutual learning for unsupervised hyperspectral image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5518317. [CrossRef]
71. Wang, X.; Qi, G.-J. Contrastive learning with stronger augmentations. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5549–5560. [CrossRef] [PubMed]
72. Yang, Y.; Miao, R.; Wang, Y.; Wang, X. Contrastive graph convolutional networks with adaptive augmentation for text classification. *Inf. Process. Manag.* **2022**, *59*, 102946. [CrossRef]
73. Ding, L.; Liu, L.; Huang, Y.; Li, C.; Zhang, C.; Wang, W.; Wang, L. Text-to-image vehicle re-identification: Multi-scale multi-view cross-modal alignment network and a unified benchmark. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 7673–7686. [CrossRef]
74. Ma, H.; Lin, X.; Yu, Y. I2F: A unified image-to-feature approach for domain adaptive semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *46*, 1695–1710. [CrossRef]
75. Li, S.; Liu, Z.; Zang, Z.; Wu, D.; Chen, Z.; Li, S.Z. GenURL: A general framework for unsupervised representation learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 286–298. [CrossRef]
76. Han, D.; Cheng, X.; Guo, N.; Ye, X.; Rainer, B.; Priller, P. Momentum cross-modal contrastive learning for video moment retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 5977–5994. [CrossRef]
77. Zhang, W.; Li, Z.; Li, G.; Zhuang, P.; Hou, G.; Zhang, Q.; Li, C. GACNet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5503314. [CrossRef]

78. Ma, X.; Liu, C.; Xie, C.; Ye, L.; Deng, Y.; Ji, X. Disjoint masking with joint distillation for efficient masked image modeling. *IEEE Trans. Multimed.* **2023**, *26*, 3077–3087. [CrossRef]
79. Mantripragada, K.; Qureshi, F.Z. Hyperspectral pixel unmixing with latent dirichlet variational autoencoder. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5507112. [CrossRef]
80. De Santis, E.; Martino, A.; Rizzi, A. Human versus machine intelligence: Assessing natural language generation models through complex systems theory. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 4812–4829. [CrossRef]
81. Si, L.; Dong, H.; Qiang, W.; Song, Z.; Du, B.; Yu, J.; Sun, F. A Trusted generative-discriminative joint feature learning framework for remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5601814. [CrossRef]
82. Huang, Y.; Zheng, H.; Li, Y.; Zheng, F.; Zhen, X.; Qi, G.; Shao, L.; Zheng, Y. Multi-constraint transferable generative adversarial networks for cross-modal brain image synthesis. *Int. J. Comput. Vis.* **2024**, *132*, 4937–4953. [CrossRef]
83. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
84. Jiang, M.; Su, Y.; Gao, L.; Plaza, A.; Zhao, X.-L.; Sun, X.; Liu, G. GraphGST: Graph generative structure-aware transformer for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5504016. [CrossRef]
85. Zhang, Q.; Xu, Y.; Zhang, J.; Tao, D. ViTAEv2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *Int. J. Comput. Vis.* **2023**, *131*, 1141–1162. [CrossRef]
86. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M. Transformers in vision: A survey. *ACM Comput. Surv.* **2022**, *54*, 200. [CrossRef]
87. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. UniFormer: Unifying convolution and self-attention for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12581–12600. [CrossRef] [PubMed]
88. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010. [CrossRef]
89. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C.-W.; Zhang, L. TTST: A top-*k* token selective transformer for remote sensing image super-resolution. *IEEE Trans. Image Process.* **2024**, *33*, 738–752. [CrossRef] [PubMed]
90. Tu, L.; Li, J.; Huang, X.; Gong, J.; Xie, X.; Wang, L. S^2HM^2: A spectral–spatial hierarchical masked modeling framework for self-supervised feature learning and classification of large-scale hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5517019. [CrossRef]
91. Huang, Z.; Jin, X.; Lu, C.; Hou, Q.; Cheng, M.-M.; Fu, D.; Shen, X.; Feng, J. Contrastive masked autoencoders are stronger vision learners. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2506–2517. [CrossRef]
92. Qian, Y.; Wang, Y.; Zou, J.; Lin, J.; Pan, Y.; Yao, T.; Sun, Q.; Mei, T. Kernel masked image modeling through the lens of theoretical understanding. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 13512–13526. [CrossRef]
93. Hou, Z.; Sun, F.; Chen, Y.; Xie, Y.; Kung, S.-Y. MILAN: Masked image pretraining on language assisted representation. *arXiv* **2022**, arXiv:2208.06049. [CrossRef]
94. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
95. He, Q.; Sun, X.; Yan, Z.; Wang, B.; Zhu, Z.; Diao, W.; Yang, M.Y. AST: Adaptive self-supervised transformer for optical remote sensing representation. *ISPRS J. Photogramm. Remote Sens.* **2023**, *200*, 41–54. [CrossRef]
96. Chen, X.; Liu, C.; Hu, P.; Lin, J.; Gong, Y.; Chen, Y.; Peng, D.; Geng, X. Adaptive masked autoencoder transformer for image classification. *Appl. Soft Comput.* **2024**, *164*, 111958. [CrossRef]
97. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C.-M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **2023**, *5*, 220–235. [CrossRef]
98. Radenovic, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1655–1668. [CrossRef]
99. Dong, Z.; Gu, Y.; Liu, T. UPetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5616613. [CrossRef]
100. Song, B.; Yang, H.; Wu, Y.; Zhang, P.; Wang, B.; Han, G. A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4408818. [CrossRef]
101. Wang, H.; Li, J.; Wu, H.; Hovy, E.; Sun, Y. Pretrained language models and their applications. *Engineering* **2023**, *25*, 51–65. [CrossRef]
102. Zhu, J.; Li, Y.; Yang, K.; Guan, N.; Fan, Z.; Qiu, C.; Yi, X. MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5610413. [CrossRef]
103. Guo, M.-H.; Zhang, Y.; Mu, T.-J.; Huang, S.X.; Hu, S.-M. Tuning vision-language models with multiple prototypes clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 11186–11199. [CrossRef] [PubMed]

104. Zhou, G.; Qian, L.; Gamba, P. A novel iterative self-organizing pixel matrix entanglement classifier for remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5407121. [CrossRef]
105. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [CrossRef] [PubMed]
106. Zhou, G.; Liu, W.; Zhu, Q.; Lu, Y.; Liu, Y. ECA-MobileNetV3(Large)+SegNet model for binary sugarcane classification of remotely sensed images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 4414915. [CrossRef]
107. Zhang, C.; Pan, X.; Li, H.; Gardiner, A.; Sargent, I.; Hare, J.; Atkinson, P.M. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *140*, 133–144. [CrossRef]
108. An, X.; He, W.; Zou, J.; Yang, G.; Zhang, H. Pretrain a remote sensing foundation model by promoting intra-instance similarity. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5643015. [CrossRef]
109. Baltrusaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [CrossRef] [PubMed]
110. Rajpurkar, P.; Lungren, M.P. The Current and future state of ai interpretation of medical images. *N. Engl. J. Med.* **2023**, *388*, 1981–1990. [CrossRef] [PubMed]
111. Jin, L.; Li, Z.; Tang, J. Deep semantic multimodal hashing network for scalable image-text and video-text retrievals. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 1838–1851. [CrossRef] [PubMed]
112. Lin, R.; Hu, H. Adapt and explore: Multimodal mixup for representation learning. *Inf. Fusion* **2023**, *105*, 102216. [CrossRef]
113. Li, D.; Xie, W.; Li, Y.; Fang, L. FedFusion: Manifold-driven federated learning for multi-satellite and multi-modality fusion. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5500813. [CrossRef]
114. Khan, M.; Gueaieb, W.; El Saddik, A.; Kwon, S. MSER: Multimodal speech emotion recognition using cross-attention with deep fusion. *Expert Syst. Appl.* **2024**, *245*, 122946. [CrossRef]
115. Wu, Y.; Liu, J.; Gong, M.; Gong, P.; Fan, X.; Qin, A.K.; Miao, Q.; Ma, W. Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding. *IEEE Trans. Multimed.* **2023**, *26*, 1626–1638. [CrossRef]
116. Krishna, R.; Wang, J.; Ahern, W.; Sturmfels, P.; Venkatesh, P.; Kalvet, I.; Lee, G.R.; Morey-Burrows, F.S.; Anishchenko, I.; Humphreys, I.R.; et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **2024**, *384*, eadl2528. [CrossRef]
117. He, X.; Chen, Y.; Huang, L.; Hong, D.; Du, Q. Foundation model-based multimodal remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5502117. [CrossRef]
118. Min, B.; Ross, H.; Sulem, E.; Ben Veyseh, A.P.; Nguyen, T.H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.* **2023**, *56*, 30. [CrossRef]
119. Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; Wang, B. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **2024**, *21*, 1470–1480. [CrossRef]
120. Saka, A.; Taiwo, R.; Saka, N.; Salami, B.A.; Ajayi, S.; Akande, K.; Kazemi, H. GPT models in construction industry: Opportunities, limitations, and a use case validation. *Dev. Built Environ.* **2024**, *17*, 100300. [CrossRef]
121. Zhao, B.; Jin, W.; Del Ser, J.; Yang, G. ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. *Neurocomputing* **2023**, *557*, 126708. [CrossRef]
122. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef]
123. Zhang, Y.; Wei, C.; He, Z.; Yu, W. GeoGPT: An assistant for understanding and processing geospatial tasks. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103976. [CrossRef]
124. Huang, A.H.; Wang, H.; Yang, Y. FinBERT: A large language model for extracting information from financial text. *Contemp. Account. Res.* **2023**, *40*, 806–841. [CrossRef]
125. Omiye, J.; Gui, H.; Rezaei, S.; Zou, J.; Daneshjou, R. Large language models in medicine: The potentials and pitfalls: A narrative review. *Ann. Intern. Med.* **2024**, *177*, 210–220. [CrossRef] [PubMed]
126. Mann, S.; Earp, B.; Moller, N.; Vynn, S.; Savulescu, J. AUTOGEN: A personalized large language model for academic enhancement-ethics and proof of principle. *Am. J. Bioeth.* **2023**, *23*, 28–41. [CrossRef] [PubMed]
127. Gong, S.; Luo, X. DGGCCM: A hybrid neural model for legal event detection. *Artif. Intell. Law* **2024**, 1–41. [CrossRef]
128. Zhang, Y.; Fu, K.; Sun, H.; Sun, X.; Zheng, X.; Wang, H. A multi-model ensemble method based on convolutional neural networks for aircraft detection in large remote sensing images. *Remote Sens. Lett.* **2018**, *9*, 11–20. [CrossRef]
129. Sharma, M.; Dhanaraj, M.; Karnam, S.; Chachlakis, D.G.; Ptucha, R.; Markopoulos, P.P. YOLOrs: Object detection in multimodal remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *14*, 1497–1508. [CrossRef]
130. Li, W.; Wu, J.; Liu, Q.; Zhang, Y.; Cui, B.; Jia, Y. An effective multimodel fusion method for SAR and optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5881–5892. [CrossRef]
131. Ma, W.; Guo, Q.; Wu, Y.; Zhao, W.; Zhang, X.; Jiao, L. A novel multi-model decision fusion network for object detection in remote sensing images. *Remote Sens.* **2019**, *11*, 737. [CrossRef]

132. Zhao, H.; Chen, C.; Xia, C. Multimodal remote sensing network. In Proceedings of the 2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Athens, Greece, 31 October–2 November 2023; pp. 1–4.
133. Wu, X.; Hong, D.; Chanussot, J. Convolutional neural networks for multimodal remote sensing data classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5517010. [CrossRef]
134. Uss, M.; Vozel, B.; Lukin, V.; Chehdi, K. Efficient Discrimination and localization of multimodal remote sensing images using CNN-based prediction of localization uncertainty. *Remote Sens.* **2020**, *12*, 703. [CrossRef]
135. Zhang, H.; Ni, W.; Yan, W.; Xiang, D.; Wu, J.; Yang, X.; Bian, H. Registration of multimodal remote sensing image based on deep fully convolutional neural network. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3028–3042. [CrossRef]
136. Zhou, W.; Jin, J.; Lei, J.; Hwang, J.-N. CEGFNet: Common extraction and gate fusion network for scene parsing of remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5405110. [CrossRef]
137. Zhou, H.; Xia, L.; Du, X.; Li, S. FRIC: A framework for few-shot remote sensing image captioning. *Int. J. Digit. Earth* **2024**, *17*, 2337240. [CrossRef]
138. Azeem, A.; Li, Z.; Siddique, A.; Zhang, Y.; Zhou, S. Unified multimodal fusion transformer for few shot object detection for remote sensing images. *Inf. Fusion* **2024**, *111*, 102508. [CrossRef]
139. Liu, B.; Huang, Z.; Li, Y.; Gao, R.; Chen, H.-X.; Xiang, T.-Z. HATFormer: Height-aware transformer for multimodal 3D change detection. *ISPRS J. Photogramm. Remote Sens.* **2025**, *228*, 340–355. [CrossRef]
140. Wang, T.; Chen, G.; Zhang, X.; Liu, C.; Wang, J.; Tan, X.; Zhou, W.; He, C. LMFNet: Lightweight multimodal fusion network for high-resolution remote sensing image segmentation. *Pattern Recognit.* **2025**, *164*, 111579. [CrossRef]
141. Feng, H.; Hu, Q.; Zhao, P.; Wang, S.; Ai, M.; Zheng, D.; Liu, T. FTransDeepLab: Multimodal fusion Transformer-based deeplabv3+ for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4406618. [CrossRef]
142. Zhu, C.; Zhang, T.; Wu, Q.; Li, Y.; Zhong, Q. An implicit Transformer-based fusion method for hyperspectral and multispectral remote sensing image. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *131*, 103955. [CrossRef]
143. Zhang, J.; Lei, J.; Xie, W.; Yang, G.; Li, D.; Li, Y. Multimodal informative ViT: Information aggregation and distribution for hyperspectral and LiDAR classification. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7643–7656. [CrossRef]
144. Xu, Y.; Cao, L.; Li, J.; Li, W.; Li, Y.; Zong, Y.; Wang, A.; Rao, Y.; Deng, S. S^2RCFormer: Spatial-spectral residual cross-attention Transformer for multimodal remote sensing data classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 16176–16193. [CrossRef]
145. Dong, Z.; Cheng, D.; Li, J. SpectMamba: Remote sensing change detection network integrating frequency and visual state space model. *Expert Syst. Appl.* **2025**, *287*, 127902. [CrossRef]
146. Pan, H.; Zhao, R.; Ge, H.; Liu, M.; Zhang, Q. Multimodal fusion Mamba network for joint land cover classification using hyperspectral and Lidar data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 17328–17345. [CrossRef]
147. Gao, F.; Jin, X.; Zhou, X.; Dong, J.; Du, Q. MSFMamba: Multiscale feature fusion State Space Model for multisource remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5504116. [CrossRef]
148. Wang, H.; Chen, W.; Li, X.; Liang, Q.; Qin, X.; Li, J. CUG-STCN: A seabed topography classification framework based on knowledge graph-guided vision mamba network. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *136*, 104383. [CrossRef]
149. Weng, Q.; Chen, G.; Pan, Z.; Lin, J.; Zheng, X. AFMamba: Adaptive fusion network for hyperspectral and LiDAR data collaborative classification base on mamba. *J. Remote Sens.* **2025**, *135*, 1–15. [CrossRef]
150. Zhang, G.; Zhang, Z.; Deng, J.; Bian, L.; Yang, C. S^2CrossMamba: Spatial–spectral cross-Mamba for multimodal remote sensing image classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5510705. [CrossRef]
151. Ma, M.; Zhao, J.; Ma, W.; Jiao, L.; Li, L.; Liu, X.; Liu, F.; Yang, S. A Mamba-aware spatial–spectral cross-modal network for remote sensing classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4402515. [CrossRef]
152. Liu, C.; Ma, X.; Yang, X.; Zhang, Y.; Dong, Y. COMO: Cross-mamba interaction and offset-guided fusion for multimodal object detection. *Inf. Fusion* **2025**, *125*, 103414. [CrossRef]
153. Zhou, M.; Li, T.; Qiao, C.; Xie, D.; Wang, G.; Ruan, N.; Mei, L.; Yang, Y.; Shen, H.T. DMM: Disparity-guided multispectral Mamba for oriented object detection in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5404913. [CrossRef]
154. Li, Y.; Li, D.; Xie, W.; Ma, J.; He, S.; Fang, L. Semi-mamba: Mamba-driven semi-supervised multimodal remote sensing feature classification. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 9837–9849. [CrossRef]
155. Liu, C.; Chen, K.; Zhao, R.; Zou, Z.; Shi, Z. Text2Earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 238–259. [CrossRef]
156. Zheng, Z.; Ermon, S.; Kim, D.; Zhang, L.; Zhong, Y. Changen2: Multi-temporal remote sensing generative change foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 725–741. [CrossRef]
157. Qu, J.; Yang, Y.; Dong, W.; Yang, Y. LDS2AE: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities. In Proceedings of the 38th AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38. [CrossRef]

158. Tang, D.; Cao, X.; Hou, X.; Jiang, Z.; Liu, J.; Meng, D. CRS-Diff: Controllable remote sensing image generation with Diffusion model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5638714. [CrossRef]
159. Zhang, W.; Mei, J.; Wang, Y. DMDiff: A dual-branch multimodal conditional guided Diffusion model for cloud removal through sar-optical data fusion. *Remote Sens.* **2025**, *17*, 965. [CrossRef]
160. Sun, D.; Yao, J.; Xue, W.; Zhou, C.; Ghamisi, P.; Cao, X. Mask approximation net: A novel Diffusion model approach for remote sensing change captioning. *IEEE Trans. Geosci. Remote Sens.* **2025**, 1. [CrossRef]
161. Sebaq, A.; ElHelw, M. RSDiff: Remote sensing image generation from text using Diffusion model. *Neural Comput. Appl.* **2024**, *36*, 23103–23111. [CrossRef]
162. Pan, J.; Lei, S.; Fu, Y.; Li, J.; Liu, Y.; Sun, Y.; He, X.; Peng, L.; Huang, X.; Zhao, B. EarthSynth: Generating informative earth observation with Diffusion models. *arXiv* **2025**, arXiv:2505.12108. [CrossRef]
163. Cai, M.; Zhang, W.; Zhang, T.; Zhuang, Y.; Chen, H.; Chen, L.; Li, C. Diffusion-Geo: A two-stage controllable text-to-image generative model for remote sensing scenarios. In Proceedings of the IGARSS 2024—2024 IEEE International Geoscience and Remote Sensing Symposium, Athens, Greece, 7–12 July 2024; pp. 7003–7006.
164. Wu, H.; Mu, W.; Zhong, D.; Du, Z.; Li, H.; Tao, C. FarmSeg_VLM: A farmland remote sensing image segmentation method considering vision-language alignment. *ISPRS J. Photogramm. Remote Sens.* **2025**, *225*, 423–439. [CrossRef]
165. Wu, H.; Du, Z.; Zhong, D.; Wang, Y.; Tao, C. FSVLM: A vision-language model for remote sensing farmland segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4402813. [CrossRef]
166. Yao, K.; Xu, N.; Yang, R.; Xu, Y.; Gao, Z.; Kitrungrotsakul, T.; Ren, Y.; Zhang, P.; Wang, J.; Wei, N.; et al. Falcon: A remote sensing vision-language foundation model. *arXiv* **2025**, arXiv:2503.11070. [CrossRef]
167. Lin, H.; Zhang, C.; Hong, D.; Dong, K.; Wen, C. FedRSCLIP: Federated learning for remote sensing scene classification using vision-language models. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 260–275. [CrossRef]
168. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. RS5M and GeoRSCLIP: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5642123. [CrossRef]
169. Liu, F.; Dai, W.; Zhang, C.; Zhu, J.; Yao, L.; Li, X. Co-LLaVA: Efficient remote sensing visual question answering via model collaboration. *Remote Sens.* **2025**, *17*, 466. [CrossRef]
170. Fu, Z.; Yan, H.; Ding, K. CLIP-MoA: Visual-language models with mixture of adapters for multitask remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4703817. [CrossRef]
171. Lin, H.; Hong, D.; Ge, S.; Luo, C.; Jiang, K.; Jin, H.; Wen, C. RS-MoE: A vision–language model with mixture of experts for remote sensing image captioning and visual question answering. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5614918. [CrossRef]
172. Zhan, Y.; Xiong, Z.; Yuan, Y. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS J. Photogramm. Remote Sens.* **2025**, *221*, 64–77. [CrossRef]
173. Bazi, Y.; Bashmal, L.; Al Rahhal, M.M.; Ricci, R.; Melgani, F. RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sens.* **2024**, *16*, 1477. [CrossRef]
174. Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; Xiao, P. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. *Comput. Vis. ECCV* **2024**, *15132*, 440–457. [CrossRef]
175. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Li, J.; Mao, X. EarthMarker: A visual prompting multimodal large language model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *63*, 5604219. [CrossRef]
176. Zhang, Z.; Shen, H.; Zhao, T.; Chen, B.; Guan, Z.; Wang, Y.; Jia, X.; Cai, Y.; Shang, Y.; Yin, J. GeoRSMLLM: A multimodal large language model for vision-language tasks in geoscience and remote sensing. *arXiv* **2025**, arXiv:2503.12490.
177. Ou, R.; Hu, Y.; Zhang, F.; Chen, J.; Liu, Y. GeoPix: A multimodal large language model for pixel-level image understanding in remote sensing. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 324–337. [CrossRef]
178. Wang, Z.; Wang, M.; Xu, S.; Li, Y.; Zhang, B. CCExpert: Advancing MLLM capability in remote sensing change captioning with difference-aware integration and a foundational dataset. *arXiv* **2024**, arXiv:2411.11360. [CrossRef]
179. Li, Z.; Muhtar, D.; Gu, F.; He, Y.; Zhang, X.; Xiao, P.; He, G.; Zhu, X. LHRS-Bot-Nova: Improved multimodal large language model for remote sensing vision-language interpretation. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 539–550. [CrossRef]
180. Liu, P. A multimodal fusion framework for semantic segmentation of remote sensing based on multilevel feature fusion learning. *Neurocomputing* **2025**, *653*, 131233. [CrossRef]
181. Liu, X.; Wang, Z.; Gao, H.; Li, X.; Wang, L.; Miao, Q. HATF: Multi-modal feature learning for infrared and visible image fusion via hybrid attention Transformer. *Remote Sens.* **2024**, *16*, 803. [CrossRef]
182. Ma, X.; Zhang, X.; Pun, M.-O.; Huang, B. A unified framework with multimodal fine-tuning for remote sensing semantic segmentation. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5405015. [CrossRef]
183. Wang, J.; Su, N.; Zhao, C.; Yan, Y.; Feng, S. Multi-modal object detection method based on dual-branch asymmetric attention backbone and feature fusion pyramid network. *Remote Sens.* **2024**, *16*, 3904. [CrossRef]
184. Hong, D.; Yokoya, N.; Xia, G.-S.; Chanussot, J.; Zhu, X.X. X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *167*, 12–23. [CrossRef]

185. Nie, H.; Luo, B.; Liu, J.; Fu, Z.; Zhou, H.; Zhang, S.; Liu, W. PromptMID: Modal invariant descriptors based on Diffusion and vision foundation models for optical-SAR image matching. *arXiv* **2025**, arXiv:2502.18104. [CrossRef]
186. Du, W.-L.; Gu, Y.; Zhao, J.; Zhu, H.; Yao, R.; Zhou, Y. A Mamba-Diffusion framework for multimodal remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6016905. [CrossRef]
187. Liu, Q.; Wang, X. Bidirectional feature fusion and enhanced alignment based multimodal semantic segmentation for remote sensing images. *Remote Sens.* **2024**, *16*, 2289. [CrossRef]
188. Wang, S.; Cai, B.; Hou, D.; Ding, Q.; Wang, J.; Shao, Z. MF-BHNet: A hybrid multimodal fusion network for building height estimation using sentinel-1 and sentinel-2 imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4512419. [CrossRef]
189. Pan, C.; Fan, X.; Tjahjadi, T.; Guan, H.; Fu, L.; Ye, Q.; Wang, R. Vision foundation model guided multimodal fusion network for remote sensing semantic segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 9409–9431. [CrossRef]
190. Liu, F.; Zhang, T.; Dai, W.; Zhang, C.; Cai, W.; Zhou, X.; Chen, D. Few-shot adaptation of multi-modal foundation models: A survey. *Artif. Intell. Rev.* **2024**, *57*, 268. [CrossRef]
191. Zhang, M.; Yang, B.; Hu, X.; Gong, J.; Zhang, Z. Foundation model for generalist remote sensing intelligence: Potentials and prospects. *Sci. Bull.* **2024**, *69*, 3652–3656. [CrossRef] [PubMed]
192. Yang, B.; Wang, X.; Xing, Y.; Cheng, C.; Jiang, W.; Feng, Q. Modality fusion vision transformer for hyperspectral and LiDAR data collaborative classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 17052–17065. [CrossRef]
193. Han, Z.; Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Chanussot, J. Multimodal hyperspectral unmixing: Insights from attention networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5524913. [CrossRef]
194. Xue, Z.; Tan, X.; Yu, X.; Liu, B.; Yu, A.; Zhang, P. Deep hierarchical vision transformer for hyperspectral and LIDAR data classification. *IEEE Trans. Image Process.* **2022**, *31*, 3095–3110. [CrossRef]
195. Zhang, M.; Zhao, X.; Li, W.; Zhang, Y.; Tao, R.; Du, Q. Cross-scene joint classification of multisource data with multilevel domain adaption network. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 11514–11526. [CrossRef]
196. Song, X.; Jiao, L.; Li, L.; Liu, F.; Liu, X.; Yang, S.; Hou, B. MGPACNet: A multiscale geometric prior aware cross-modal network for images fusion classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4412815. [CrossRef]
197. Wu, Q.; Li, Z.; Zhu, S.; Xu, P.P.; Yan, T.T.; Wang, J. Nonlinear intensity measurement for multi-source images based on structural similarity. *Measurement* **2021**, *179*, 109474. [CrossRef]
198. Marsocci, V.; Jia, Y.; Bellier, G.; Kerekes, D.; Zeng, L.; Hafner, S.; Gerard, S.; Brune, E.; Yadav, R.; Shibli, A.; et al. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv* **2024**, arXiv:2412.04204. [CrossRef]
199. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4701117. [CrossRef]
200. Hu, D.; Wei, Y.; Qian, R.; Lin, W.; Song, R.; Wen, J.-R. Class-aware sounding objects Localization via audiovisual correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9844–9859. [CrossRef] [PubMed]
201. Sadok, S.; Leglaive, S.; Girin, L.; Alameda-Pineda, X.; Séguier, R. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. *Neural Netw.* **2024**, *172*, 106120. [CrossRef] [PubMed]
202. Xing, Y.; He, Y.; Tian, Z.; Wang, X.; Chen, Q. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. *arXiv* **2024**, arXiv:2402.17723.
203. Mai, G.; Lao, N.; He, Y.; Song, J.; Ermon, S. CSP: Self-supervised contrastive spatial pre-training for geospatial-visual representations. *arXiv* **2023**, arXiv:2305.01118.
204. Gong, Z.; Wei, Z.; Wang, D.; Hu, X.; Ma, X.; Chen, H.; Jia, Y.; Deng, Y.; Ji, Z.; Zhu, X.; et al. CrossEarth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv* **2024**, arXiv:2410.22629. [CrossRef]
205. Cepeda, V.; Nayak, G.; Shah, M. GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv* **2023**, arXiv:2309.16020.
206. Gong, Z.; Li, B.; Wang, C.; Chen, J.; Zhao, P. BF-SAM: Enhancing SAM through multi-modal fusion for fine-grained building function identification. *Int. J. Geogr. Inf. Sci.* **2024**, *39*, 2069–2095. [CrossRef]
207. Liu, J.; Chen, S.; He, X.; Guo, L.; Zhu, X.; Wang, W.; Tang, J. VALOR: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 708–724. [CrossRef]
208. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. ImageBind One Embedding Space to Bind Them All. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 15180–15190.
209. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-LLM: Multi-modal language modeling with image, audio, video, and text integration. *arXiv* **2023**, arXiv:2306.09093.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.