

Review

When Remote Sensing Meets Foundation Model: A Survey and Beyond

Chunlei Huo ^{1,2,*} , Keming Chen ³, Shuaihao Zhang ³, Zeyu Wang ³ , Heyu Yan ³, Jing Shen ^{4,5}, Yuyang Hong ², Geqi Qi ^{4,5}, Hongmei Fang ³ and Zihan Wang ³

¹ Information Engineering College, Capital Normal University, Beijing 100048, China

² University of Chinese Academy of Sciences, Beijing 101499, China; hongyuyang23@mails.ucas.ac.cn

³ Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100086, China; chenkm@aircas.ac.cn (K.C.); zhangshuaihao23@mails.ucas.ac.cn (S.Z.); wangzeyu231@mails.ucas.ac.cn (Z.W.); yanheyu23@mails.ucas.ac.cn (H.Y.); fanghongmei23@mails.ucas.ac.cn (H.F.); wangzihan232@mails.ucas.ac.cn (Z.W.)

⁴ Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China; shenjing2023@ia.ac.cn (J.S.); qigeqi2023@ia.ac.cn (G.Q.)

⁵ State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100086, China

* Correspondence: clhuo@cnu.edu.cn

Abstract: Most deep-learning-based vision tasks rely heavily on crowd-labeled data, and a deep neural network (DNN) is usually impacted by the laborious and time-consuming labeling paradigm. Recently, foundation models (FMs) have been presented to learn richer features from multi-modal data. Moreover, a single foundation model enables zero-shot predictions on various vision tasks. The above advantages make foundation models better suited for remote sensing images, where image annotations are more sparse. However, the inherent differences between natural images and remote sensing images hinder the applications of the foundation model. In this context, this paper provides a comprehensive review of common foundation models and domain-specific foundation models for remote sensing, and it summarizes the latest advances in vision foundation models, textually prompted foundation models, visually prompted foundation models, and heterogeneous foundation models. Despite the great potential of foundation models for vision tasks, open challenges concerning data, model, and task impact the performance of remote sensing images and make foundation models far from practical applications. To address open challenges and reduce the performance gap between natural images and remote sensing images, this paper discusses open challenges and suggests potential directions for future advancements.

Keywords: foundation model; remote sensing; pre-training; fine-tuning; adapter; segment anything model; state-space model; visual grounding



Academic Editor: Andrzej Stateczny

Received: 26 October 2024

Revised: 1 January 2025

Accepted: 3 January 2025

Published: 7 January 2025

Citation: Huo, C.; Chen, K.; Zhang, S.; Wang, Z.; Yan, H.; Shen, J.; Hong, Y.; Qi, G.; Fang, H.; Wang, Z. When Remote Sensing Meets Foundation Model: A Survey and Beyond. *Remote Sens.* **2025**, *17*, 179. <https://doi.org/10.3390/rs17020179>

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/>).

1. Introduction

We are currently in the era of remote sensing, big data, and artificial intelligence. With the continuous progress of imaging techniques and artificial intelligence advances, high-resolution remote sensing images have been widely used for various applications, such as resource exploration, environmental monitoring, precision agriculture, military reconnaissance, and so on [1,2]. Compared to traditional machine learning algorithms, DNNs have revolutionized the accuracy of remote sensing vision tasks with rich annotation data and powerful GPU computing power. However, DNNs are far from the remote

sensing requirements. On one hand, DNNs are usually impacted by the laborious and time-consuming labeling requirement and single-task DNN limitation (e.g., a DNN is generally trained for each visual recognition task). On the other hand, remote sensing image annotations are more sparse.

Recently, foundation models (FMs) have been a hot and interesting topic in many fields, and they have demonstrated strong versatility in visual and language understanding tasks, e.g., ChatGPT [3], Gemini [4], and visual ChatGPT [5]. FMs have shown promising signs in remote sensing; however, the essential differences between remote sensing images and natural images make the direct application of FMs limited for remote sensing images [6]. Therefore, it is necessary to develop domain-specific FMs peculiar to remote sensing by carefully considering the common technique shared between different domains and the inherent characteristics of remote sensing images. In this context, this paper aims to comprehensively review common foundation models and domain-specific foundation models for remote sensing. Compared to related surveys [7–9], this paper covers the newest FMs and emerging advances (e.g., prompt-related FMs, Mamba vision FMs), which are very important for developing remote-sensing-specific FMs.

The paper is organized as follows. First, the basic principles of foundation models and common techniques are reviewed, and the specific methods of foundation models in remote sensing are then surveyed. The challenges are summarized and the follow-up research directions are explored.

2. Foundation Model for General Vision Tasks

The flowchart of FMs for general vision tasks and remote sensing images is shown in Figure 1. A foundation model is essentially a backbone for feature learning, and fine-tuning or an adapter is usually used for different domains or downstream tasks. The following three characteristics make FMs very promising for vision tasks:

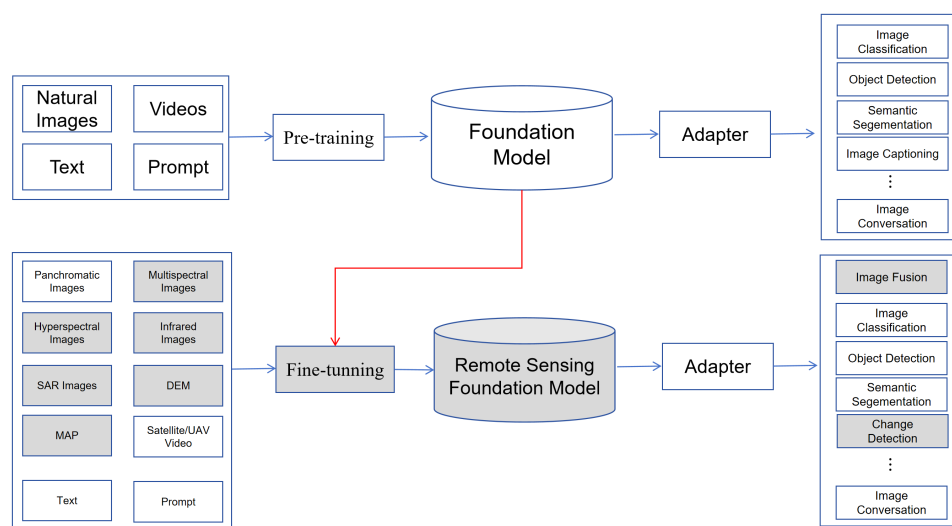


Figure 1. Flowchart of FMs for vision tasks on natural images and remote sensing images.

- (1) Model size. The massive model size enables FMS to capture more complex patterns hidden in the data, generating more representative features.
- (2) Learning strategy. By taking advantage of related data (including text, images, audio, and video), FMs use self-supervised or semi-supervised learning without human supervision or with minimal human interaction, which alleviates the time-consuming and labor-consuming labeling work.
- (3) Adaptation. FMs are trained on broad data and could be easily adapted to various downstream tasks [10].

Besides traditional vision foundation models and vision–language foundation models, several research efforts are devoted to developing visually prompted foundation models, e.g., Segment Anything Model (SAM) [11]. To provide a systematic review of recent foundation models, this paper focuses on the following foundation models: vision FMs, textually prompted FMs (i.e., vision–language FMs), visually prompted FMs, and heterogeneous FMs. Figure 2 lists state-of-the-art foundation models of different types. Below, we review the above foundation models from the perspective of pre-training and fine-tuning.

Vision foundation model	PyramidViT SwinTransformer Dino MoCov3 SegFormer DAT MIM ViT LeViT LocalViT	InternImage	LocalMamba EfficientVMamba VMamba Vim		
Textually prompted foundation model	ALIGN EVA MaskCLIP Florence OpenCLIP MaskCLIP GroundingDINO SLIP GLIP CRIS FILIP RegionCLIP CLIPA ViLD WenLan GroupViT CLIP	MetaLM KOSMOS SimVLM CapPa CoCa Frozen Flamingo	MaskVLM FLAVA IBLIP UniDetector FIBERBLIP-2 InstructBLIP X-Decoder PaLI	miniGPT4 VisualChatGPT LLaMA-Adapter LLaVA GPT4	Gemini 1.0 Pro Vision Mini-Gemini Gemini
Visually prompted foundation model	SegGPT CLIPSeg	CAT SEEM TAM SAM-DA SAM-PT SAM-Track AutoSAM DeSAM SAMM SAM RefSAM			
Heterogeneous foundation model	VisionLLM Prismer Painter	CLIP2Video AudioCLIP Valley COSA MACAW-LLM ImageBind	VIMA VOYAGER MineDojo COSA Palm-E		

Figure 2. State-of-the-art foundation models. Milestone models are marked in red.

2.1. Foundation Model Pre-Training

2.1.1. Vision Foundation Model

Vision foundation models aim to encode visual features, and they can be categorized into ViT-based, CNN-based, and Mamba-based types. By taking advantage of the attention mechanism, ViT defines a new paradigm for vision tasks. To enhance ViT, many variants (e.g., Swin Transformer [12], PyramidViT [13], LeViT [14], LocalViT [15], and MIM [16]) are proposed. It is worth noting that InternImage [17] is a large-scale CNN-based foundation model, which learns stronger and more robust patterns like ViT by reducing the inductive bias of CNNs. Vision Mamba models such as ViM [18] and VMamba [19] mitigate the modeling constraints of CNNs and offer Transformer-like modeling strengths by global receptive fields and dynamic weighting. Mamba-type foundation models perform excellently on long-sequence modeling tasks. Compared with other foundation models, vision foundation models are limited due to the uncertainty of images and the lack of powerful semantic representation taken by the language model.

2.1.2. Vision–Language Foundation Model

A vision–language foundation model is the combination of a vision model and a natural language model. The vision model captures visual features from the images, while the language model encodes information from the text. For instance, the newest GPT-4V [20] integrates image inputs into large language models (LLMs), transforming FMs from language-only systems to multi-modal powerhouses. Vision–language foundation models aimed at jointly understanding visual and textual modalities. To better connect vision models and language models, vision–language models are usually pre-trained by the following three objectives [21]: contrastive objective, generative objective, and alignment objective.

(1) Contrastive objective

The rationale of the contrastive objective is to learn discriminative image–text representations by pulling the embeddings of paired images and texts close while pushing others away. The milestone of the contrastive objective is CLIP [22]. CLIP jointly trained an image encoder and text encoder, and the contrastive loss aimed to maximize the cosine similarity between the paired image and text embeddings. To address the limitation of CLIP in computationally expensive pre-processing and cleaning, ALIGN [23] trained a dual encoder-based model on one billion noisy image–caption datasets, where cosine similarity was optimized via normalized softmax loss. To handle representations from space–time–modality space, Florence [24] expanded the images and textual representations from coarse to fine, from static to dynamic, and from RGB to multiple modalities. Finer-level information between modalities is hard to represent by CLIP-like methods, and FILIP [25] aimed to capture fine-grained semantic alignment by the token-wise cross-modal interaction. Motivated by the masked auto-encoder, FLIP [26] enhances CLIP by masking 50–75% of input pixels, and the computation complexities are reduced by 2–4 times. In contrast, MaskCLIP [27] learns local semantic features by the random masking of the input image and teacher-based self-distillation; EVA-CLIP [28] masks visual inputs to address the instability and optimization efficiency.

(2) Generative objective

The generative objective aims to learn semantic features and generate images or texts via masked modeling techniques. Masked image modeling is used to mask certain patches in an image, and the encoder is trained to learn image context information and reconstruct images conditioned on unmasked patches. For instance, a large portion of patches (i.e., 75%) were masked out in training KELIP [29] and SegCLIP [30]. Masked language modeling masks a fraction of tokens in each input text and predicts the masked tokens. In contrast, masked cross-modal modeling works by jointly masking and reconstructing text tokens and image patches. For instance, for FLAVA [31], the masked portions of image patches and text tokens are 40% and 15%, respectively. For an image, image-to-text generation aims to generate descriptive texts and capture fine-grained vision–language correlation. For example, COCA [32] produces intermediate embeddings by encoding an input image, and embeddings are then decoded into descriptive texts.

(3) Alignment objective

The alignment objective aligns the image–text pair in an embedding space via global matching between image and text or local matching between region and word. For instance, GLIP [33] and DetCLIP [34] model local fine-grained vision–language correlation by replacing object classification logits with region–word alignment scores.

2.1.3. Visually Prompted Foundation Model

A visually prompted foundation model can be prompted by various prompts, e.g., text, points, and bounding boxes. SAM [11] is a milestone of visually prompted models, and it is trained on 11 million images and 1.1 billion masks. The rationale of SAM involves encoding image and prompt embeddings, respectively, and predicting segmentation masks using a lightweight mask decoder. Two best-known training-free approaches based on SAM are Per-SAM [35] and its variant, PerSAM-F. For an image and a reference mask, PerSAM localizes the target concept before estimating the object’s position, while PerSAM-F reduces mask ambiguity by one-shot fine-tuning. SEED [36] takes more prompt types and exhibits strong generalization capabilities.

It is worth noting that the combination of SAM and vision–language foundation models is more powerful. For instance, Grounded-Segment-Anything [37] aims to detect

and segment anything with text inputs, where bounding boxes are obtained by Grounded-DINO [38] and text prompts, and target segmentations are achieved by SAM and bounding boxes. Caption-Anything [39] leverages SAM and large language models for zero-shot image captioning.

Other visually prompted models for image segmentation include CLIPSeg [40], SegGPT [41], and SEEM [42]. CLIPSeg achieved capabilities for zero-shot and one-shot segmentation, and the Transformer-based decoder is conditioned on text–visual CLIP embeddings. SegGPT aims to train a foundation model applicable to diverse segmentation tasks in an in-context learning paradigm.

2.2. Foundation Model Fine-Tuning

Fine-tuning is essential for foundation models adapting to specific domains or downstream tasks. Full fine-tuning needs training the entire model on a remote sensing dataset, allowing for complete adaptation to the specific domain. However, it requires significant computational resources and can lead to overfitting. In contrast, parameter-efficient fine-tuning (PEFT) updates only a small subset of parameters, such as the last few layers or specific attention heads [43]. PEFT methods like adapter, prompt tuning, and low-rank adaptation are commonly used to reduce computational costs and improve efficiency. An adapter is a small neural network inserted into the pre-trained model, facilitating task-specific adaptation without modifying the original model's parameters. CLIP-Adapter [44] learns new features by adding a bottleneck layer and performs residual-style feature blending with the pre-trained features. SAM-Adapter [45] adapts SAM to downstream tasks by injecting customized information into the network. MaskCLIP [27] adapts CLIP and masked self-distillation to learn local and global semantics. To further improve the segmentation performance, MaskCLIP+ [46] combines pseudo-labeling and self-training. Prompt tuning involves injecting task-specific information into the model's input and aims to guide the model to focus on relevant features. Low-rank adaptation reduces the parameter size by factorizing the model's weight matrices into low-rank matrices. In contrast, prompt-tuning methods involve finding the optimal prompts without fine-tuning the entire VLM. For example, CoOp [47] keeps the pre-trained models fixed and utilizes learnable vectors to model a prompt's context words. Visual prompt tuning enables pixel-level adaptation to downstream tasks. For example, DenseCLIP [48] adapts CLIP for dense prediction by converting the image–text matching problem in CLIP to a pixel–text matching problem.

By taking advantage of pre-trained models and the above fine-tuning techniques, FMs define many new tasks. For instance, CLIP defines concepts such as visual grounding, open-vocabulary object detection, referring image segmentation, and phrase grounding. Specifically, RegionCLIP [49] established a new SOTA for zero-shot and open-vocabulary object detection. CLIP-Driven CRIS [50] extended CLIP for the referring image segmentation task by introducing a visual–language decoder and a text-to-pixel contrastive loss. GLIP [33] reformulated the object detection task to phrase grounding, where all class names are combined in one sentence. Grounding–DINO [38] is an open-set object detector that combines the strengths of DINO with grounded pre-training, and it can detect arbitrary objects using human-provided text descriptions.

To tailor pre-trained foundation models to remote sensing tasks, fine-tuning and adapter techniques are usually employed. Specifically, fine-tuning aims to extend pre-trained backbones on natural images for the remote sensing domain by refining model parameters based on remote sensing images, and adapters are usually used for downstream tasks by adding additional layers to the refined backbone. For instance, SPT [51] is a novel fine-tuning technique for SAR images, which uses scattering information as a prompt and

integrates learnable parameters into the pre-trained model. MANet [52] adapts SAM for semantic segmentation and fuses information from different modalities to enhance the segmentation performance. Training-Efficient Adapting (TEA) [53] combines parameter-efficient fine-tuning with a top-down guidance mechanism to improve the efficiency and performance of foundation models in remote sensing. PETL [54] explores parameter-efficient techniques for adapting vision–language pre-trained models to remote sensing image–text retrieval tasks.

3. Remote Sensing Foundation Models

To systematically review recent advances in remote sensing foundation models, Figure 3 elaborated important efforts on datasets and foundation models.

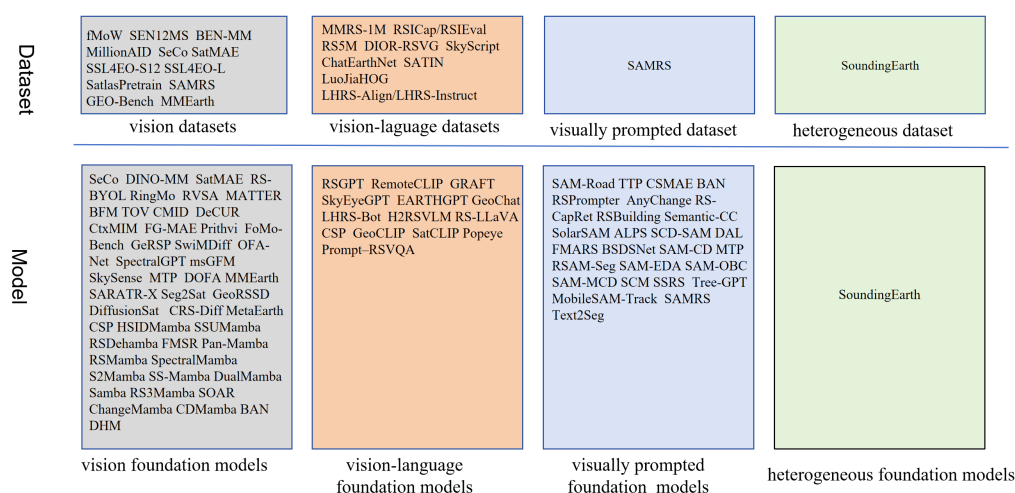


Figure 3. Overview of remote sensing datasets and foundation models.

3.1. Dataset

Some large-scale datasets are constructed for foundation model pre-training. Vision datasets are mainly for traditional tasks such as scene classification, semantic segmentation, and object detection. For instance, fMoW [55] contains 1 million images from 200 countries, and it aims to predict the functional purpose of buildings and land use. SEN12MS [56] consists of 180,662 triplets of dual-polarization SAR image patches, multi-spectral Sentinel-2 image patches and MODIS land cover maps. MillionAID [57] contains a million instances for scene classification. SatlasPretrain [58] consists of 302 M labels under 137 categories and seven label types: points, polygons, polylines, properties, segmentation labels, regression labels, and classification labels. Vision–language datasets have richer semantic information, and they are mainly constructed for downstream tasks such as image captioning and VQA. Table 1 lists some representative vision–language datasets. MMRS-1M [59] is the largest multi-modal multi-sensor instruction-following dataset, consisting of over 1M image–text pairs about optical, SAR, and infrared images. RSICap [60] contains rich and high-quality scene information (e.g., residential area, airport, or farmland) and object information (e.g., color, shape, quantity, absolute position, etc). These datasets help us pre-train or fine-tune a vision–language foundation model; however, differences between various datasets concerning captions and object information may hinder us from training a unified vision–language foundation model. Other types of datasets are rare. SAMRS [61] is a large-scale segmentation dataset that leverages SAM and existing object detection datasets. SoundingEarth [62] consists of about 50k pairs of overhead images and crowd-sourced audio samples, and it aims to train ResNet models and understand the key properties of a scene by mapping visual and auditory modalities into a common embedding space.

Table 1. Vision–language datasets.

Dataset	Description
MMRS-1M [59]	Tasks: Classification, detection, image captioning, VQA, visual grounding Contributions: MMRS-1M contains more than 1M image–text pairs and encompasses three visual modalities (e.g., optical, infrared and SAR). url(accessed on 19 December 2024): https://github.com/wivizhang/EarthGPT
RSICap [60]	Tasks: Image captioning, VQA Contributions: RSICap contains 2585 human-annotated captions with rich and high-quality scene information and object information. url(accessed on 19 December 2024): https://github.com/Lavender105/RSGPT?tab=readme-ov-file
RS5M [63]	Tasks: Image captioning Contributions: RS5M contains 5 million aligned image–text pairs. url(accessed on 19 December 2024): https://github.com/om-ai-lab/RS5M
DIOR-RSVG [64]	Tasks: VQA, image captioning, image–text retrieval Contributions: DIOR-RSVG is based on DIOR dataset, and it contains image–expression–box triplets for visual grounding tasks. url(accessed on 19 December 2024): https://github.com/zhanyang-nwpu/rsvg-pytorch
SkyScript [65]	Tasks: Zero-shot scene classification, zero-shot transfer learning for fine-grained object attribute classification and cross-modal retrieval Contributions: SkyScript contains 5.2 million image–text pairs and covers more than 29K distinct semantic tags. url(accessed on 19 December 2024): https://github.com/wangzhecheng/SkyScript
ChatEarthNet [66]	Tasks: Image captioning Contributions: ChatEarthNet is a new image–text dataset based on Sentinel-2 images. It consists of 173,488 image–text pairs with captions generated by ChatGPT-3.5 and a manual verification process. url(accessed on 19 December 2024): https://github.com/zhu-xlab/ChatEarthNet
SATIN [67]	Tasks: Zero- or low-shot classification Contributions: SATIN is curated from 27 existing remote sensing datasets, and it is organized into six distinct tasks and 250 distinct class labels. url(accessed on 19 December 2024): https://satinbenchmark.github.io/
LuoJiaHOG [68]	Tasks: Image–text retrieval Contributions: LuoJiaHOG is an image caption dataset containing 94856 images and 131 classes. url(accessed on 19 December 2024): https://huggingface.co/datasets/aleo1/LuojiaHOG
LHRS-Align and LHRS-Instruct [69]	Tasks: Scene classification, VQA, visual grounding Contributions: LHRS-Align comprises 1.15 million image–text pairs and captions are generated by a large language model. LHRS-Instruct is a multi-modal instruction-following dataset and it contains instruction data for visual reasoning, image description, and conversation. url(accessed on 19 December 2024): https://github.com/NJU-LHRS/LHRS-Bot
FIT-RS [70]	Tasks: Image caption, region caption, VQA, multi-label scene classification, object detection, multi-task conversation relation detection, relation reasoning, object reasoning Contributions: FIT-RS is a fine-grained instruction tuning dataset, and it contains 1,800,851 instructions samples for vision–language comprehension tasks, e.g., semantic relationships among objects in complex scenes. url(accessed on 19 December 2024): https://github.com/Luo-Z13/SkySenseGPT

3.2. Foundation Models

By effectively employing fine-tuning and adapter techniques [71], researchers can leverage the power of foundation models to tackle a wide range of remote sensing challenges, from object detection and segmentation to scene classification and change detection. Below, we review typical foundation models for remote sensing in depth.

3.2.1. Vision Foundation Models

Due to the complexity of remote sensing images, the foremost important extension of vision foundation models is to learn domain-specific representation and add domain-specific knowledge. For instance, DINO-MM [72] aimed to learn joint representation and the invariant representation of optical images and SAR images. Seg2Sat [73], map-sat [74], DiffusionSat [75], CRS-Diff [76], and MetaEarth [77] aimed at generating realistic or higher-resolution images by diffusion-based generative foundation models. CROMA [78] learns rich unimodal and multi-modal representations by combining contrastive and reconstruction objectives. MATTER [79] leverages multi-temporal satellite images to learn features that are robust to changes in illumination and viewing angle. Considering the differences between natural images and remote sensing images, GeoKR [80] and GASSL [81] enhance the foundation model through geographical knowledge and geography awareness. SAR-JEPA [82] learns multi-scale SAR target representations by knowledge-guided predictive architecture. RSBuilding [83] unifies building extraction and building change detection tasks by a task-prompted cross-attention decoder. CSPT [84] gradually bridged the domain gap and transferred large-scale data knowledge to the remote sensing domain. GFM [85] investigated the potential of continual pre-training from the large-scale ImageNet-22k model to the geospatial foundation model. To handle objects of arbitrary orientations in remote sensing images, RVSA [86] utilized a new rotated varied-size window attention to learn better object representation and handle objects of arbitrary orientations. EarthPT [87] accurately predicts pixel-level surface reflectance across the 400–2300 nm range. Prithvi [88] is a Transformer-based foundation model, and it could be fine-tuned for various tasks such as multi-temporal crop segmentation and flood mapping. To reduce the model size, SiamIntCD [89] adapts a CNN-based vision foundation model, InternImage [17], for change detection. It is the first CNN-based change detection vision foundation model and combines the advantages of Siamese architecture and InternImage.

By taking advantage of state-space models (SSMs) with selection mechanisms and hardware-aware architectures, Mamba is powerful in long-sequence modeling. Due to the quadratic complexity of the self-attention mechanism, the Mamba foundation model has recently been adapted for computer vision tasks, i.e., Vision Mamba (ViM). ViM's efficiency in processing large-size images makes it well-suited for high-resolution remote sensing data, and the bidirectional SSM captures detailed spatial information, enabling accurate feature extraction from complex remote sensing scenes. Mamba vision foundation models have been widely used for various tasks, e.g., denoising (HSIDMamba [90], SSUMamba [91]), dehazing (RSDehamba [92]), super-resolution (FMSR [93]), pan-sharpening (e.g., Pan-Mamba [94]), image classification (RSMamba [95], SpectralMamba [96], S^2 Mamba [97], SS-Mamba [98], DualMamba [99]), semantic segmentation (Samba [100], RS3Mamba [101]), object detection (SOAR [102]), change detection (ChangeMamba [103], CDMamba [104], BAN [105]), and compressive imaging (DHM [106]). HSIDMamba [90] outperforms the latest Transformer architectures by 30% efficiency by incorporating bi-directional continuous scanning, which strengthens spatial-spectral interactions from eight directions. RSDehamba [92] is an extremely lightweight Mamba-based model with only 1.80 M and outperforms baseline methods by more than 0.72 dB. Pan-Mamba [94] utilizes channel-swapping Mamba to initiate cross-modal interaction between partial panchromatic and

multi-spectral channels and cross-modal Mamba to exploit cross-modal relationships for improving the information representation capability. S^2 Mamba [97] utilizes bi-directional spectral scanning to explore semantic information from continuous spectral bands and the spatial–spectral mixture gate to adaptively incorporate representations from different dimensions. Samba [100] is the first Mamba method for the semantic segmentation of remote sensing images, where Samba blocks serve as the encoder and UperNet as the decoder. ChangeMamba [103] is the first Mamba method for change detection tasks and it adopts ViM as the encoder to learn global spatial contextual information. From the above analysis, it could be inferred that the common innovations of the above methods are learning the representative spatial, temporal, and spectral context and reducing computation complexity. While ViM foundation models have shown promising performances in remote sensing images, the inherent 1D nature of selective scanning technique presents challenges for 2D or higher-dimensional visual data, and critical spatial information may be lost. This problem requires further investigation in the future.

Details of state-of-the-arts vision foundation models can be found in Table 2.

Table 2. Vision foundation models.

Model Name	Description
DINO-MM [72]	Tasks: Representation learning Contributions: DINO-MM aims to jointly learn representations of SAR-optical data by concatenating all channels to a unified input, and it is flexible to learn representations from any combination of available modalities. url(accessed on 19 December 2024): https://github.com/zhu-xlab/DINO-MM
Seg2Sat [73]	Tasks: Generative foundation model Contributions: Seg2Sat explores diffusion algorithms and the terrain segmentation data to generate aerial images. url(accessed on 19 December 2024): https://github.com/RubenGres/Seg2Sat
map-sat [74]	Tasks: Realistic satellite image generation Contributions: map-sat aims to generate realistic satellite images from maps, and is a pre-trained diffusion model, and the pre-trained image diffusion model is augmented by ControlNet. url(accessed on 19 December 2024): https://github.com/miquel-espinoza/map-sat
DiffusionSat [75]	Tasks: Temporal generation, super-resolution Contributions: DiffusionSat is the first generative foundation model for remote sensing data based on the latent-diffusion model. DiffusionSat can generate high-resolution satellite images from numerical metadata and text, and a novel 3D-conditioning extension enables DiffusionSat for super-resolution, temporal generation, and in-painting. url(accessed on 19 December 2024): https://github.com/samar-khanna/DiffusionSat
CRS-Diff [76]	Tasks: Single and compositional image generation Contributions: CRS-Diff leverages the inherent advantages of diffusion models to generate visually clear images and enrich geographic and temporal information. url(accessed on 19 December 2024): https://github.com/Sonettoo/CRS-Diff?tab=readme-ov-file
MetaEarth [77]	Tasks: Unbounded and arbitrary-sized image generation Contributions: MetaEarth is a resolution-guided self-cascading generative framework, and it can generate images at any region with different geographical resolutions. url(accessed on 19 December 2024): https://jiupinjia.github.io/metaearth/

Table 2. Cont.

Model Name	Description
MATTER [79]	Tasks: Change detection, land cover classification and semantic segmentation Contributions: MATTER learns invariance to illumination and view angle, and it achieves material consistency and texture representation by leveraging multi-temporal images over unchanged regions. url(accessed on 19 December 2024): https://github.com/periakiva/MATTER
Prithvi [88]	Tasks: Flood mapping, wildfire scar segmentation, multi-temporal cloud gap imputation and crop segmentation. Contributions: Prithvi is a temporal vision transformer pre-trained on contiguous US Harmonised Landsat Sentinel 2 (HLS) data. The model accepts remote sensing data in a video format, and it captures spatial attention across multiple patches and temporal attention for each patch. url(accessed on 19 December 2024): https://github.com/NASA-IMPACT/Prithvi-EO-2.0
SiamIntCD [89]	Tasks: Change detection. Contributions: SiamIntCD adapts a CNN-based vision foundation model named InternImage [17] for model size reduction and change detection. It is the first CNN-based change detection vision foundation model and combines the advantages of Siamese architecture and InternImage. url(accessed on 19 December 2024): https://github.com/jingsl5/SiamIntCD
HSIDMamba [90]	Tasks: Hyperspectral denoising Contributions: By hyperspectral continuous scan blocks, HSIDMamba enhances long-range and local spatial-spectral information and strengthens spatial-spectral interactions. url(accessed on 19 December 2024):—
SSUMamba [91]	Tasks: Hyperspectral image denoising Contributions: SSUMamba is based on the spatial-spectral continuous scan Mamba block, which aims to improve denoising performance by capturing local spatial-spectral correlations, long-range spatial-spectral dependencies, and continuity in various directions. url(accessed on 19 December 2024): https://github.com/Ironkitty/SSUMamba
RSDehamba [92]	Tasks: Satellite image dehazing Contributions: RSDehamba integrates the state space model to enrich the global context of U-Net for image restoration. To transfer effective information between different regions, a direction-aware scan module is presented to sense the spatially varying haze distribution. url(accessed on 19 December 2024): —
FMSR [93]	Tasks: Image super-resolution Contributions: FMSR is a multi-level spatial-frequency fusion model. Multi-level features are recalibrated for accurate feature fusion via learnable scaling adaptors. url(accessed on 19 December 2024): https://github.com/XY-boy/FreMamba
Pan-Mamba [94]	Tasks: Pan-sharpening Contributions: Pan-Mamba is a novel pan-sharpening network, which leverages channel-swapping Mamba for initiating a lightweight cross-modal interaction and cross-modal Mamba for enhancing inherent cross-modal relationships. url(accessed on 19 December 2024): https://github.com/alexhe101/Pan-Mamba

Table 2. Cont.

Model Name	Description
RSMamba [95]	Tasks: Image classification Contributions: To alleviate the limitations of the vanilla Mamba in unidirectional modeling and position insensitivity, RSMamba augments Mamba’s capacity to model non-causal data by incorporating a position-sensitive dynamic multi-path activation mechanism. url(accessed on 19 December 2024): https://github.com/KyanChen/RSMamba
SpectralMamba [96]	Tasks: Hyperspectral image classification Contributions: SpectralMamba models hyperspectral data dynamics at two levels, i.e., the first level is the joint encoding of spatial regularity and spectral peculiarity in the spatial–spectral space by piece-wise sequential scanning, and the second level is the gated spatial–spectral merging in the hidden state space. url(accessed on 19 December 2024): https://github.com/danfenghong/SpectralMamba
S ² Mamba [97]	Tasks: Hyperspectral image classification Contributions: S2Mamba learns spatial contextual features by a patch cross-scanning module and explores spectral semantic information by a bidirectional spectral scanning module. The spatial–spectral mixture gate is performed to adaptively incorporate representative features across different dimensions. url(accessed on 19 December 2024): https://github.com/PURE-melo/S2Mamba
SS-Mamba [98]	Tasks: Hyperspectral image classification Contributions: SS-Mamba is a Mamba-based spectral–spatial learning framework, which aims to exploit the spectral–spatial information by converting HSI cube to spatial and spectral tokens as sequences and enhance spectral–spatial tokens using the HSI sample’s center region information. url(accessed on 19 December 2024): https://github.com/XY-boy/FreMamba
DualMamba [99]	Tasks: Hyperspectral image classification Contributions: DualMamba is a dual-stream Mamba-convolution network, and adaptive global-local fusion is utilized to dynamically adjust the weights of global and local spectral–spatial features. url(accessed on 19 December 2024): —
Samba [100]	Tasks: Semantic segmentation Contributions: Samba adopts Samba blocks to efficiently learn multi-level semantic information from high-resolution remote sensing images. It surpassed the mIoU metric of Segformer and ConvNeXt by 3.95% and 10.3% on the LoveD dataset. url(accessed on 19 December 2024): https://github.com/zhuqinfeng1999/Samba
RS3Mamba [101]	Tasks: Semantic segmentation Contributions: RS3Mamba constructs an auxiliary branch based on visual state space blocks and provides global information to the convolution-based branch, and features from the dual encoder are merged and enhanced by a collaborative completion module. url(accessed on 19 December 2024): https://github.com/sstary/SSRS
SOAR [102]	Tasks: Small body object detection for aerial imagery Contributions: SOAR employs Vision Mamba for precise location-aware visual understanding and a bidirectional state space model for effective visual context modeling. url(accessed on 19 December 2024): https://github.com/yash2629/S.O.A.R

Table 2. Cont.

Model Name	Description
ChangeMamba [103]	Tasks: Change detection Contributions: ChangeMamba adopts the cutting-edge Visual Mamba architecture to encode global spatial contextual information with the linear complexity, three decoders are presented to modeling spatial-temporal relationships for binary change detection, semantic change detection and building damage assessment tasks. url(accessed on 19 December 2024): https://github.com/ChenHongruixuan/MambaCD
CDMamba [104]	Tasks: Change detection Contributions: CDMamba utilizes Mamba to extract global features and convolution to enhance the local details. An adaptive global-local guided fusion block is proposed to facilitate the bitemporal interaction. url(accessed on 19 December 2024): https://github.com/zmoka-zht/CDMamba
BAN [105]	Tasks: Change detection Contributions: BAN is a bi-temporal adapter network, and it adapts the knowledge from pre-trained foundation models to the change detection task by the bi-temporal adapter branch and bridging modules. url(accessed on 19 December 2024): A new learning paradigm for foundation model-based remote-sensing change detection
DHM [106]	Tasks: Compressive Imaging Contributions: DHM adopts a dual Mamba strategy for efficient spectral compressive imaging, one is to model long-range dependencies across the entire HSIs, and the other is to capture local contexts by structured state-space sequences. url(accessed on 19 December 2024): https://github.com/JiahuaDong/DHM
SatMAE [107]	Tasks: Land cover classification, multi-label classification and building segmentation Contributions: SatMAE learns useful representations from temporal and multi-spectral satellite images by masked autoencoders and the joint encoding of positional, temporal and spectral information. However, computational resources can be strained by the increased length of token sequences. url(accessed on 19 December 2024): https://github.com/sustainlab-group/SatMAE
GeoRSCLIP [63]	Tasks: Zero-shot classification, cross-modal text–image retrieval, semantic localization Contributions: GeoRSCLIP proposed the concept of the domain vision–language model and fine-tuned the CLIP model for utilizing domain-specific data, RS5M, which is about 1000 times larger than existing largest image–text paired remote sensing datasets. url(accessed on 19 December 2024): https://github.com/om-ai-lab/RS5M?tab=readme-ov-file
RingMo [108]	Tasks: Classification, segmentation, detection Contributions: RingMo is a foundation model with masked image modeling and is designed for dense and small objects in complicated scenes. url(accessed on 19 December 2024): https://github.com/comeony/RingMo
RVSA [86]	Tasks: Classification and segmentation Contributions: By utilizing a new rotated varied-size window attention, RVSA learns better object representation and reduces the computational cost. url(accessed on 19 December 2024): https://github.com/ViTAE-Transformer/Remote-Sensing-RVSA

Table 2. Cont.

Model Name	Description
BFM [109]	Tasks: Rotated object detection and semantic segmentation Contributions: BFM is the first billion-scale remote sensing foundation model, and it is achieved by pretraining a vision Transformer on extensive remote sensing images coupled with the parallelism strategy. url(accessed on 19 December 2024): —
CMID [110]	Tasks: Scene classification, semantic segmentation, object detection, and change detection Contributions: CMID adopts a teacher–student self-distillation architecture for combining contrastive learning and masked image modeling, and it aims to learn representations with both global semantic separability and local spatial perceptibility. url(accessed on 19 December 2024): https://github.com/NJU-LHRS/official-CMID
SARATR-X [111]	Tasks: Target recognition Contributions: SARATR-X ensures the diversity and scalability of foundation model by multi-scale gradient features and two-stage training. url(accessed on 19 December 2024): https://github.com/waterdisappear/SARatrX
MTP [112]	Tasks: Semantic segmentation, instance segmentation, rotated object detection, and change detection Contributions: Based on a shared encoder and task-specific decoder architecture, MTP explores a multi-task pretraining framework to address the discrepancy between upstream pre-training and downstream fine-tuning tasks. url(accessed on 19 December 2024): https://github.com/ViTAE-Transformer/MTP
DOFA [113]	Tasks: Classification, segmentation Contributions: Inspired by neuroplasticity for adjusting to new experiences or environmental shifts, DOFA trains a unified multi-modal foundation model for different data modalities across distinct tasks. url(accessed on 19 December 2024): https://github.com/zhu-xlab/DOFA
SkySense [114]	Single-modal and semantic segmentation, horizontal and oriented object detection, change detection, scene classification Contributions: SkySense is a generic billion-scale model pre-trained on 21.5 million temporal sequences of optical and SAR data. By multi-granularity contrastive learning, SkySense learns representations across different modalities and spatial granularities, and it surpasses 18 recent FMs on 16 datasets over seven tasks. url(accessed on 19 December 2024): —
msGFM [115]	Tasks: Scene classification, segmentation, cloud removal, and pan-sharpening Contributions: msGFM employs an innovative cross-sensor pre-training approach in masked image modeling, enabling the synthesis of joint representations from diverse sensors (e.g., SAR, Sentinel 2, RGB, DSM.) url(accessed on 19 December 2024): https://github.com/boranhhan/Geospatial_Foundation_Models

Table 2. Cont.

Model Name	Description
SpectralGPT [116]	Tasks: Single-label and multi-label scene classification, semantic segmentation, change detection Contributions: SpectralGPT is the first spectral RS foundation model with over 600 million parameters, and it is trained by a 3D generative Transformer on one million spectral images. By leveraging 3D tokens for spatial–spectral coupling, SpectralGPT captures spectrally sequential patterns via multi-target reconstruction. url(accessed on 19 December 2024): https://github.com/danfenghong/IEEE_TPAMI_SpectralGPT
SLFFM [117]	Tasks: Object detection, semantic segmentation, and change detection Contributions: A sparse and low-frequency filtered masked modeling framework is designed for representation learning of ConvNet foundation model. url(accessed on 19 December 2024): https://github.com/HIT-SIRS/SMLFR
SwiMDiff [118]	Tasks: Change detection, land-cover classification Contributions: SwiMDiff seamlessly integrates contrastive learning with a diffusion model. By pixel-level diffusion constraints, the encoder comprehensively captures fine-grained features and global semantic information. url(accessed on 19 December 2024): —
GeRSP [119]	Tasks: Object detection, semantic segmentation, and scene classification Contributions: GeRSP combines two pre-training branches and learns representations with the general and special knowledge using a teacher–student architecture. url(accessed on 19 December 2024): https://github.com/floatingstarZ/GeRSP
FG-MAE [120]	Tasks: Reconstructing HOG and NDI features Contributions: FG-MAE is a feature-guided masked autoencoder, and it aims to reconstruct various image features for medium-resolution SAR and multispectral images. url(accessed on 19 December 2024): https://github.com/zhu-xlab/FGMAE
CtxMIM [121]	Tasks: Land cover classification, semantic segmentation, object detection, and instance segmentation Contributions: CtxMIM formulates the original image patches as the reconstructive template, and it learns robust and transferable representations by a context-enhanced generative branch. url(accessed on 19 December 2024): —
DeCUR [122]	Tasks: Scene classification and semantic segmentation Contributions: By distinguishing inter- and intra-modal embeddings, DeCUR aims to integrate complementary information from different modalities, i.e., radar-optical, RGB elevation and RGB depth. url(accessed on 19 December 2024): https://github.com/zhu-xlab/DeCUR
CSP [123]	Tasks: Location representation Contributions: CSP encodes the images and corresponding geo-locations using a dual-encoder and learns effective location representations by a contrastive objective. url(accessed on 19 December 2024): https://gengchenmai.github.io/csp-website/

3.2.2. Textually Prompted Foundation Models

Compared to pure vision-based models, textually prompted foundation models define new vision tasks. Specifically, traditional label-featured image classification, object detection and semantic segmentation tasks are transformed to open-vocabulary image classification, open-vocabulary object detection, and open-vocabulary semantic segmentation, e.g., GRAFT [124] and YOLO-World [125].

The other advantage taken by textually prompted foundation models is the utilization of language model for image captioning and VQA, e.g., RSGPT [60] and RS-LLaVA [126]. RSGPT [60] created vision–language chat agents for remote sensing images. Interestingly, the addition of a language model makes the vision–language foundation model easily adapted for image captioning and visual grounding (e.g., SkyEyeGPT [127], EARTH-GPT [59]) and image–text retrieval and object counting (RemoteCLIP [128]), where visual grounding aims to locate relevant objects or regions based on a natural language query. Moreover, GeoChat [129] offers multitask conversational capabilities, i.e., image-level queries, region-specific dialogue, visually grounded objects, etc. Table 3 lists details of state-of-the-arts vision–language foundation models.

Table 3. Remote sensing vision–language foundation models.

Model Name	Description
EARTHGPT [59]	Tasks: Classification, detection, image captioning, VQA and visual grounding Contributions: EarthGPT is a pioneering multi-modal large language model designed for remote sensing image comprehension. It leverages a multi-modal mutual learning framework to seamlessly integrate various tasks and multi-sensor image interpretation in multi-turn dialogues. url(accessed on 19 December 2024): —
RSGPT [60]	Tasks: Image captioning, VQA Contributions: RSGPT is based on fine-tuning InstructBLIP on RSICap. For remote sensing domain adaptation, the Q-Former and the linear layer are trained, and the image encoder and LLM are frozen. url(accessed on 19 December 2024): https://github.com/Lavender105/RSGPT
LHRS-Bot [69]	Tasks: Scene classification, VQA, visual grounding Contributions: LHRS-Bot performs image understanding through multi-level vision–language alignment and curriculum learning. url(accessed on 19 December 2024): https://github.com/NJU-LHRS/LHRS-Bot
RS-LLaVA [126]	Tasks: Image captioning and VQA Contributions: RS-LLaVA is a large vision–language model for joint captioning and question answering in remote sensing images, and it is an enhanced version of LLaVA adapted by the LoRA method. url(accessed on 19 December 2024): https://github.com/BigData-KSU/RS-LLaVA?tab=readme-ov-file
SkyEyeGPT [127]	Tasks: Image captioning, visual grounding, VQA Contributions: SkyEyeGPT is a unified multi-modal, large language model designed for remote sensing (RS) vision–language understanding. It uses a two-stage tuning method to enhance instruction-following and multi-turn dialogue ability at different granularities, and it exhibits encouraging performances compared to GPT-4V in certain tests. url(accessed on 19 December 2024): https://github.com/ZhanYang-nwpu/SkyEyeGPT

Table 3. Cont.

Model Name	Description
RemoteCLIP [128]	Tasks: Zero-shot image classification, few-shot classification, image–text retrieval, object counting Contributions: RemoteCLIP is the first vision–language foundation model for remote sensing, and it adapts CLIP for seamless downstream applications by aligning vision–language representations and learning robust visual features with rich semantics. url(accessed on 19 December 2024): https://github.com/ChenDelong1999/RemoteCLIP
GeoChat [129]	Tasks: Image and region captioning, VQA, scene classification, visually grounded conversations and referring detection Contributions: GeoChat is the first grounded large vision language model for remote sensing scenarios. It finetunes LLaVA-1.5 by LoRA, which avoids forgetting original knowledge embedded in the large language model and extends its domain knowledge to remote sensing tasks. url(accessed on 19 December 2024): https://github.com/mbzuai-oryx/GeoChat
GRAFT [124]	Tasks: Image classification, retrieval, segmentation and VQA Contributions: GRAFT trains an image encoder without textual annotations; the key insight involves aligning the image encoder of CLIP using paired internet and satellite images. url(accessed on 19 December 2024): https://colab.research.google.com/drive/1C-7H4uptjgOBnmiQl6gcNAJKYvilsBxS?usp=sharing
H ² RSVLM [130]	Tasks: Scene classification, VQA, visual grounding Contributions: H²RSVLM is a helpful and honest remote sensing vision language model, and it can mitigate the incorrect generations by recognizing and rejecting unanswerable questions. url(accessed on 19 December 2024): https://github.com/opendatalab/H2RSVLM
GeoCLIP [131]	Tasks: Geo-localization Contributions: GeoCLIP employs GPS encoding by CLIP-inspired alignment between locations and images, and a continuous function is used to model the Earth and obtain semantic features suitable for geolocation. url(accessed on 19 December 2024): https://vicentevivan.github.io/GeoCLIP/
SatCLIP [132]	Tasks: Geo-localization Contributions: SatCLIP exploits a contrastive location–image pretraining objective, and it matches satellite images and the respective coordinates to learn an implicit neural representation of visual patterns on the globe. url(accessed on 19 December 2024): https://github.com/microsoft/satclip

3.2.3. Visually Prompted Foundation Models

Recently, the remote sensing community has shown strong interest in adapting SAM for segmentation-related applications, especially the combination of SAM and CLIP [133,134]. The fact that SAM provides accurate region segmentation and CLIP provides reliable semantics makes foundation models attractive for practical applications. Figure 4 shows two strategies of combining SAM and CLIP, i.e., pre-SAM and post-SAM. Specifically, for the pre-SAM strategy, a text prompt is used as input for Grounding DINO or CLIP Surgery, which generates bounding boxes or point prompts for SAM to produce a semantic segmentation map. In contrast, post-SAM employs SAM to generate all available segmentation masks and then utilizes CLIP to identify masks of interest by comparing semantic similarities with the text prompt. Awesome applications vary from different data

(optical images [135,136], multi-view remote sensing image [137], SAR images [138], aerial images [139], point cloud [140], LiDAR data [141]), different tasks (semantic segmentation [83,142], instance segmentation [143], shadow detection [144], change detection [145], image understanding [146]), domain adaptation [147], and zero-shot and one-shot learning [134].



Figure 4. Two strategies of combining SAM and CLIP. The figure is from [148].

SAM relies on manual guidance due to its interactive nature, and it is limited for automated remote sensing image segmentation. RSPrompter [143] learns to produce task-specific prompts, which help SAM generate semantically discernible segmentation results. AnyChange [149] is a zero-shot change detection approach, which adopts SAM to reveal intra-image and inter-image semantic similarities in the latent space. With zero-shot change detection capability. SAM-Road [142] adapts SAM for efficiently and accurately predicting road segments and intersections from satellite images. RingMo-SAM [136] can segment anything and identify object categories in optical and SAR images, and it improves the multi-object segmentation performances of complicated scenes by multi-box prompts. SAMCD [150] adapts FastSAM for change detection, where FastSAM focuses on some specific ground objects and models the semantic latent in bitemporal images. Some typical applications are listed in Table 4.

Table 4. Visually prompted foundation models for remote sensing.

Model Name	Description
RSBuilding [83]	Tasks: Building extraction, change detection Contributions: RSBuilding enhances cross-scene generalization and task universality by extracting image features based on the prior knowledge of FMs and unifying task representation by a cross-attention decoder with task prompts. url(accessed on 19 December 2024): https://github.com/Meize0729/RSBuilding
Text2Seg [133]	Tasks: Semantic segmentation Contributions: By employing visual foundation models for automatic prompt generation, Text2Seg overcomes the dependency of SAM on extensive annotations and enhances the model's ability to generalize across diverse datasets. Compared to the vanilla SAM model, relative improvements of Text2Seg over zero-shot prediction performance range from 31% to 225%. url(accessed on 19 December 2024): https://github.com/Douglas2Code/Text2Seg
BSDSNet [138]	Tasks: Land cover classification Contributions: BSDSNet is a SAM-based dual-stream feature extraction network, where the ConvNext encoder and ViT encoder are used to extract local and global information. By integrating local–global information, the SA-Gate module makes BSDSNet well-suited for classifying PolSAR images across different resolutions. url(accessed on 19 December 2024): —

Table 4. Cont.

Model Name	Description
SAM-Road [142]	Tasks: Road network graph extraction Contributions: SAM-Road aims to extract a road network graph by leveraging the inherent strengths of SAM. Probability masks of roads and intersections are produced by fine-tuning the image encoder of SAM, and graph topology is predicted by a lightweight Transformer-based graph neural network. url(accessed on 19 December 2024): https://github.com/htrc/sam_road
RSPrompter [143]	Tasks: Instance segmentation Contributions: RSPrompter learns to produce task-specific prompts, which help SAM generate semantically discernible segmentation results. url(accessed on 19 December 2024): https://github.com/KyanChen/RSPrompter
SAM-CD [145]	Tasks: Change detection Contributions: SAM-CD adapts FastSAM and a convolutional adaptor to aggregate the task-oriented change information, and a task-agnostic semantic learning branch is introduced to utilize semantic representations inherent to SAM features and model the semantic latent in bi-temporal images. url(accessed on 19 December 2024): https://github.com/DingLei14/SAM-CD
Tree-GPT [146]	Tasks: Image understanding Contributions: Tree-GPT is a modular LLM expert system for the forestry remote sensing data. Structured information is extracted from forest remote sensing images by the automatic or interactive generation of prompts, and it is used to help SAM generate the optimal tree segmentation results. url(accessed on 19 December 2024):—
SAM-EDA [147]	Tasks: Segmentation and domain adaptation Contributions: SAM-EDA integrates SAM into an unsupervised domain adaptation segmentation framework. The teacher–assistant model provides semantic pseudo-labels and guides the student model for learning. url(accessed on 19 December 2024):—
TTP [151]	Tasks: Change detection Contributions: TTP adapts SAM for change detection, where low-rank fine-tuning is presented to address the domain shift problem, a time-travel activation gate is utilized to discern inter-image correlation and a multi-level change prediction head to decode dense semantic information. url(accessed on 19 December 2024): https://github.com/KyanChen/TTP
Semantic-CC [152]	Tasks: Image captioning and VQA Contributions: Semantic-CC utilizes the latent knowledge of foundation models to alleviate extensive annotations and harnesses pixel-level semantics from change detection to guide comprehensive and accurate change descriptions. url(accessed on 19 December 2024):—
ALPS [153]	Tasks: Segmentation and annotation Contributions: ALPS is an innovative auto-labeling framework, and it leverages SAM to predict precise pseudo-labels without prior annotations or additional prompts. url(accessed on 19 December 2024): https://github.com/StriveZs/ALPS

Table 4. Cont.

Model Name	Description
SCD-SAM [154]	Tasks: Change detection Contributions: SCD-SAM combines MobileSAM and CNN to extract progressive semantic change features in parallel, and a dual decoder aggregates binary change features and semantic change features to alleviate the semantic gap across different scales. url(accessed on 19 December 2024): https://github.com/yzygit1230/SCD-SAM
RSAM-Seg [155]	Tasks: Semantic segmentation Contributions: RSAM-Seg is a tailored modification of SAM for remote sensing, where the Adapter-Scale and Adapter-Feature modules aim to generate image-informed prompts and eliminate manual intervention to provide prompts for SAM by incorporating high-frequency image information and image embedding features. url(accessed on 19 December 2024): https://chief-byte.github.io/RSAM-Seg-Site
SSRS [156]	Tasks: Semantic segmentation Contributions: SSRS exploits two novel concepts called SAM-Generated Object (SGO) and SAM-Generated Boundary (SGB) to enhance semantic segmentation performance, where the object consistency aims to leverage segmented regions lacking semantic information, and the boundary preservation loss directs the model's attention to the object boundary. url(accessed on 19 December 2024): https://github.com/sstary/SSRS

Despite the novelties of SAM, the following limitations hinder the application and need to be addressed in the future [157]: the requirement of strong prior knowledge, less effectiveness in low-contrast applications, a limited understanding of the professional data, and smaller and irregular object segmentation. Ren [158] found that SAM is sensitive to image resolution, and it produces poor performances on some object classes such as roads.

3.2.4. Other Foundation Models

Heterogeneous foundation models for remote sensing are rare. SoundingEarth [62] exploits the correspondence between co-located images and audio recordings, and it understands a scene by visual and auditory properties.

It is worth noting that the model size is an important factor in balancing the model performance and computation efficiency. On the one hand, a larger model will help the foundation model learn complex and representative features. On the other hand, a larger model will need more computational resources for training and inference. Generally, the size of foundation models for remote sensing images varies significantly depending on the architecture, training data, and specific tasks. Specifically, small-scale models contain around 20–30 million parameters (e.g., ResNet50-based models), medium-scale models about 80–100 million parameters (e.g., ViT-Base), and large-scale models over 1 billion parameters (e.g., ViT-Huge, Swin-Large). For instance, Prithvi [88] is based on ViT-Base architecture and contains around 100 million parameters; EarthPT [87] is based on ViT-Huge architecture and contains approximately 700 million parameters; RingMo 3.0 [159] is one of the largest remote sensing foundation models, and it contains over 10 billion parameters.

4. Discussion

4.1. Open Challenges

Theoretically, foundation models for natural images could be directly applied to remote sensing images due to their powerful capability of zero-shot or few-shot learning.

However, due to the significant differences between natural images and remote sensing images [6], foundation models for remote sensing images face several open challenges, and there exists a considerable performance gap between remote sensing images and natural images. Specifically, open challenges lie in the following three aspects:

(1) Data challenges

Data heterogeneity: Due to differences in imaging factors (e.g., imaging platform, imaging mechanism, and imaging condition), remote sensing data has high variations concerning modality (i.e., optical images, infrared images, hyperspectral images, and SAR images), spatial resolution, viewpoint, and so on. The variations hinder the direct application of foundation models to remote sensing images. For instance, as illustrated by Figure 5a, kySense performs differently on LEVIR-CD and OS-CD, i.e., 92.58% vs. 60.06%. The performances are significantly different on homogeneous optical images, let alone heterogeneous images between optical images and SAR images or hyperspectral images.

Annotation data scarcity: Multiple data modalities (e.g., images, text, and geographic information) are required to train foundation models. For instance, the availability of large-scale image–text pairs is crucial for developing a vision–language foundation model. However, datasets published in the remote sensing domain are less than natural images and far from practical applications, and obtaining sufficient amounts of high-quality multi-modal data is challenging due to the complex differences between multi-modal data and reliable alignment in terms of spatial–temporal dimension and semantical dimension.

Dataset evaluation complexity: To train powerful foundation models, high-quantity and high-quality requirements are simultaneously important for constructing datasets. However, it is challenging to evaluate and compare the qualities of different datasets, and captions about deep understanding and spatial reasoning are more difficult to assess. Dataset quality is impacted by complex factors, e.g., image quality, imaging condition, task definition, annotation tool, and annotation experience. The lack of dataset quality evaluation makes reliable model benchmarks difficult.

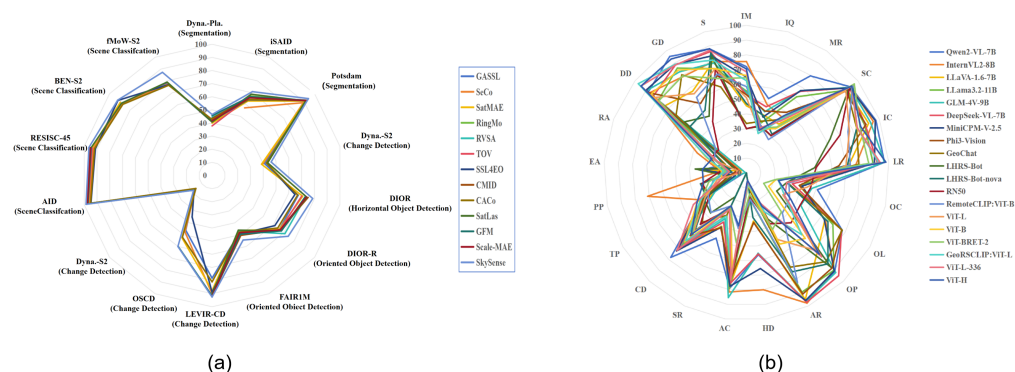


Figure 5. A performance comparison of state-of-the-art foundation models on various datasets and tasks. (a): Vision foundation models performance comparison; Data from [114]. (b): Vision–language foundation models performance comparison; Data from [160]. Image-level comprehension contains five leaf tasks: scene classification (SC), map recognition (MR), image caption (IC), image quality (IQ), and image modality (IM). Single-instance Identification comprises seven tasks: landmark recognition (LR), object presence (OP), object localization (OL), object counting (OC), attribute recognition (AR), visual grounding (VG) and hallucination detection (HD). Cross-instance discernment (CID) consists of three leaf tasks: attribute comparison (AC), spatial relationship (SR) and change detection (CD). Attribute reasoning (AttR) includes two key aspects: time property (TP) and physical property (PP). Assessment reasoning (AssR) contains resource assessment (RA) and environmental assessment (EA). Common sense reasoning (CSR) consists of three leaf tasks: disaster discrimination (DD), geospatial determination (GD) and situation inference (SI).

(2) Model challenges

Model architecture: Due to the significant difference between natural images and remote sensing images, naive model architectures for natural images are limited for remote sensing images. In detail, the desired model architecture for remote sensing requires considering the imaging mechanism, data heterogeneity, physical constraints, and domain-specific knowledge. Despite the great efforts in developing domain-specific model architecture, existing foundation models are far from the above expectations, which could be illustrated by the significant performance gap between natural images and remote sensing images.

Model performance: Foundation models achieved remarkable zero-shot prediction performances on natural images and image-level comprehension tasks. However, the performances on remote sensing images and semantic reasoning tasks are far from practical applications. As illustrated by Figure 5b, foundation models face challenges in addressing the attribute and assessment reasoning task. Specifically, all models scored below the random guess level, 50%, and many state-of-the-art models performed below 20%.

Model complexity: Larger model size helps foundation models generate more representative features and capture more complex patterns hidden in the data. However, prohibitive model size requires significant computational resources for training and inference, and it hinders foundation models for emergent applications (e.g., disaster monitoring, emergency search and rescue) and resource-limited tasks (e.g., onboard object detection). In addition, prohibitive model size may be prone to concealing the interpretability of the model.

Model evaluation: Evaluating different foundation models on a fair and uniform benchmark is very important. However, a fair evaluation is difficult. On the one hand, the evaluation is usually impacted by many other factors. For instance, by following the image resizing and normalization trick in pre-training, significant performance improvements (e.g., 32.28% on the So2Sat random split dataset and 11.16% on the EuroSAT dataset) could be achieved on downstream tasks [161]. In this context, reliable cross-domain evaluation is difficult to perform. On the other hand, a particular image can have several ground-truth descriptions, but the common evaluation strategies only compare a candidate sentence with reference sentences. Fair evaluation becomes difficult when consensus is low for reliable predictions.

(3) Application-specific challenges

Task difficulty: In contrast to object detection on natural images where objects are well postured, domain-specific task difficulties disturbed remote sensing, such as large object-size variation (e.g., airplane vs. airport), significant appearance differences caused by view-angle variation, and the sensor-related features (e.g., radiation feature, scattering feature, spectral feature, temperature feature). Foundation models will perform poorly if the above task difficulty is not carefully considered.

Task isolation: In the literature, most foundation models are performed on individual tasks independently, without considering their potential interactions and dependencies. In the remote sensing domain, it may lead to suboptimal performance and limited insights into the model's capabilities.

Application imbalance: Application imbalance refers to the uneven distribution of training data and evaluation metrics between different remote sensing applications. For instance, foundation models mainly focus on certain image types (e.g., optical images), certain object types (e.g., buildings, roads, airplanes), and other image types (e.g., SAR images, infrared images, hyperspectral images) and certain object types (e.g., fine-grained crop and forest classification [59,146,162–164]) are usually ignored. In fact, fine-grained

crop and forest classification face inherent challenges of feature learning (e.g., low interclass discrimination and high intraclass difference), and it is worth more in-depth research.

4.2. Future Directions

To address the above challenges, the following directions should be emphasized in the future.

(1) Multi-modal annotation augmentation

To solve data challenges about annotation quantity, annotation quality, modal consistency and application balance, the multi-modal annotation should be augmented. Considering the scarcity of high-quality image–text data and the prohibitive cost of manual captioning, AI-assisted annotations may be interesting. For instance, AIGC helps us generate variant images from the given captions, and Geochat helps generate captions from images. Of course, careful validation and correction are mandatory to improve the training data quality. Considering the semantic consistency between multi-sensor remote sensing images, annotations can be facilitated by utilizing annotations on optical images and image registration techniques. In this context, training data construction is essentially the fast annotation propagation over unchanged regions. In addition, we can take advantage of other auxiliary data to reduce the annotation labor. For example, land cover products, geographical location and OpenStreetMap tags could be used for geographical knowledge to provide supervision or annotation. For reasoning annotation, it is necessary to invite cognitive scientists and psychologists to reduce the semantic bias.

(2) Domain-specific model architecture

Different from nature images, remote sensing images implicitly hide some geographical priors, physical constraints and domain knowledge, and it is difficult for the naive foundation model to capture the latent information. In consequence, it is mandatory to develop foundation models specific to remote sensing domains. In our opinion, physics-constrained foundation models may be a potential solution. A physics-constrained foundation model is essentially a neural network architecture that incorporates physical laws (e.g., atmospheric scattering, radiative transfer, or surface reflectance models) and imaging principles into the backbone, the loss function, and the training process. This integration allows the foundation model to learn more accurate and reliable representations of the real world, where physical processes govern remote sensing data generation. For instance, although end-to-end networks are powerful in learning complex mapping, however, it is not easy for them to learn physics-related features from multispectral images, e.g., the normalized difference vegetation index (NDVI), the normalized difference water index (NDWI), the normalized difference soil index (NDSI), and non-homogeneous feature difference (NHFD). In other words, adding a remote-sensing-specific imaging model or prior knowledge to the foundation model will help improve the feature discrimination and model interpretation.

(3) Cognition-informed model enhancement

Cognition-informed foundation models for remote sensing involve a multi-faceted approach, combining foundation models with principles of human cognition (e.g., hierarchical learning, attention mechanisms, contextual understanding) and symbolic reasoning (e.g., knowledge graphs, causal inference). One important future task is to enhance the vision encoder, since the poor performance of foundation models mainly stems from visual incapability [165]. i.e., the gap between CLIP-based visual embeddings and vision-only self-supervised learning. In fact, the vision encoder is more challenging when considering the complexities of remote sensing data, and a cognition-driven vision encoder is expected to address the above limitations.

(4) Domain adaptation

Domain adaptation aims to improve the model's versatility by adapting foundation models to different remote sensing domains, such as multi-sensor, multi-resolution, and multi-temporal images. This involves transferring knowledge from a source domain (e.g., low-resolution satellite images) to a target domain (e.g., high-resolution UAV images). However, the following factors make domain adaptation challenging, the significant domain gap between different modalities, the scarcity of labeled data in many domains, and the expensive computational cost for large-scale datasets. To address these challenges, future research directions include unsupervised domain adaptation (i.e., developing techniques that can adapt models without requiring labeled target domain data), self-supervised learning (leveraging self-supervised learning to learn domain-invariant representations), meta-learning (training models that can quickly adapt to new domains with limited data), hybrid approaches (combining multiple domain adaptation techniques to achieve optimal performance), and continual learning (enabling models to learn continuously from data streams without forgetting previously acquired knowledge).

(5) Multi-task cooperative learning

Multi-task cooperative learning is a machine learning paradigm where multiple tasks are trained simultaneously, sharing information and improving each other's performance. For foundation models and remote sensing, multi-task cooperative learning can significantly enhance the model's ability to generalize to different tasks and data distributions. However, it is difficult to identify the appropriate relationships between different tasks, potentially resulting in bias in the model's learning process caused by data imbalance and expensive computational cost for training large-scale foundation models on multiple tasks. In the future, more efforts should be made to solve the above challenges by shared representation learning (shared encoder and task-specific decoders) and hierarchical multi-task learning (hierarchical task structure or progressive learning).

(6) Adaptive prompt learning

Adaptive prompt learning involves dynamically adjusting prompts to guide the behavior of foundation models. In remote sensing, it is particularly useful for tailoring the model's output to specific tasks and data characteristics. For instance, adaptive prompt learning can be improved by data-driven adaptation (i.e., adjusting the prompt based on the characteristics of the input data, such as image resolution, spectral bands, and noise levels.), task-driven adaptation (tailoring the prompt to the specific task, such as classification, segmentation, or object detection), and user-driven adaptation (i.e., allowing users to provide feedback to the model, which can be used to refine the prompt).

(7) Model evaluation

Evaluating the performance of foundation models on remote sensing images presents unique challenges due to the complex nature of the data and the diverse range of tasks involved. These challenges can be solved by constructing robust benchmark datasets that cover various remote sensing scenarios and developing advanced evaluation techniques. For instance, human-in-the-loop evaluation is expected to assess the quality and relevance of model predictions by incorporating human judgment.

(8) Model compression and optimization

Several methods have been proposed to reduce the size of foundation models, e.g., model architecture optimization (model pruning, model quantization, knowledge distillation) [166,167], hardware-optimized implementations by GPUs and TPUs [168], model compression techniques (e.g., parameter sharing, low-rank decomposition, sparse neural

networks) [169]. However, for some emergent applications such as onboard object detection, storage resources and computation resources are extremely limited, and lightweight foundation models must be studied, e.g., CNN-based foundation models [89], binary foundation models [170].

In short, researchers from different domains should enhance collaborations to solve the above challenges in the AI domain, the remote sensing domain, and the cognitive psychology domain.

5. Conclusions

Foundation models have become a hot topic in the remote sensing field, and they have initially shown promising signs in various remote sensing tasks. However, the essential differences between remote sensing images and natural images make the general foundation model limited for remote sensing images. This paper starts with a review of the common techniques of foundation models, and then a systematic review of remote sensing foundation models is conducted. The limitations of current techniques, as well as future directions, are discussed for future developments.

Author Contributions: Conceptualization, C.H. and K.C.; methodology, C.H. and S.Z.; investigation, C.H., Z.W. (Zeyu Wang), Z.W. (Zihan Wang), H.Y., J.S., Y.H. and G.Q.; visualization, J.S.; writing—original draft preparation, C.H., K.C., S.Z., H.F., Z.W. (Zeyu Wang) and Z.W. (Zihan Wang); data curation, J.S. and C.H.; writing—review and editing, C.H. and K.C.; supervision, C.H. and K.C.; funding acquisition, C.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported partially by National Natural Science Foundations of China (Grants No. 62071466).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

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

References

1. Zhang, L.; Zhang, L. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [\[CrossRef\]](#)
2. Sun, Z.; Sandoval, L.; Crystal-Ornelas, R.; Mousavi, S.M.; Wang, J.; Lin, C.; Cristea, N.C.; Tong, D.Q.; Carande, W.H.; Ma, X.; et al. A review of earth artificial intelligence. *Comput. Geosci.* **2022**, *159*, 105034. [\[CrossRef\]](#)
3. Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT. *arXiv* **2023**, arXiv:2302.09419. [\[CrossRef\]](#)
4. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
5. Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; Duan, N. Visual ChatGPT: Talking, drawing and editing with visual foundation models. *arXiv* **2023**, arXiv:2303.04671.
6. Rolf, E.; Klemmer, K.; Robinson, C.; Kerner, H. Mission critical—Satellite data is a distinct modality in machine learning. *arXiv* **2024**, arXiv:2402.01444.
7. Jiao, L.; Huang, Z.; Lu, X.; Liu, X.; Yang, Y.; Zhao, J.; Zhang, J.; Hou, B.; Yang, S.; Liu, F.; et al. Brain-inspired remote sensing foundation models and open problems: A comprehensive survey. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 10084–10120. [\[CrossRef\]](#)
8. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-Language models in remote sensing: Current progress and future trends. *arXiv* **2023**, arXiv:2305.05726. [\[CrossRef\]](#)
9. Zhou, Y.; Feng, L.; Ke, Y.; Jiang, X.; Yan, J.; Yang, X.; Zhang, W. Towards Vision-Language Geo-Foundation Model: A Survey. *arXiv* **2024**, arXiv:2406.09385.
10. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.

11. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Roll, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo W.Y.; et al. Segment anything. *arXiv* **2023**, arXiv:2304.02643.
12. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision Transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
13. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
14. Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; Douze, M. LeViT: A vision Transformer in Convnet’s clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
15. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; Van Gool, L. LocalViT: Bringing locality to vision Transformers. *arXiv* **2021**, arXiv:2104.05707.
16. Zhang, X.; Zeng, Y.; Zhang, J.; Li, H. Toward building general foundation models for language, vision, and vision-language understanding tasks. *arXiv* **2023**, arXiv:2301.05065.
17. Wang, W.; Dai, J.; Chen, Z.; Huang, Z.; Li, Z.; Zhu, X.; Hu, X.; Lu, T.; Lu, L.; Li, H.; et al. InternImage: Exploring large-scale vision foundation models with deformable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
18. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv* **2024**, arXiv:2401.09417.
19. Liu, X.; Zhang, C.; Zhang, L. Vision Mamba: A comprehensive survey and taxonomy. *arXiv* **2024**, arXiv:2405.04404.
20. Available online: <https://platform.openai.com/docs/guides> (accessed on 1 April 2024).
21. Awais, M.; Naseer, M.; Khan, S.; Anwer, R.M.; Cholakkal, H.; Shah, M.; Yang, M.; Khan, F.S. Foundational models defining a new era in vision: A survey and outlook. *arXiv* **2023**, arXiv:2307.13721.
22. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021.
23. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q.V.; Sung, Y.; Li, Z.; Duerig, T.; et al. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021.
24. Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.C.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A new foundation model for computer vision. *arXiv* **2021**, arXiv:2111.11432.
25. Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; Xu, C.; et al. FILIP: Fine-grained interactive language-image pre-training. *arXiv* **2021**, arXiv:2111.07783.
26. Li, Y.; Fan, H.; Hu, R.; Feichtenhofer, C.; He, K. Scaling language-image pre-training via masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 7–24 June 2023.
27. Dong, X.; Zheng, Y.; Bao, J.; Zhang, T.; Chen, D.; Yang, H.; Zeng, M.; Zhang, W.; Yuan, L.; Chen, D.; et al. MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
28. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv* **2023**, arXiv:2303.15389.
29. Ko, B.; Gu, G. Large-scale bilingual language-image contrastive learning. *arXiv* **2022**, arXiv:2203.14463.
30. Luo, H.; Bao, J.; Wu, Y.; He, X.; Li, T. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
31. Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; Kiela, D. FLAVA: A foundational language and vision alignment model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
32. Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; Wu, Y. CoCa: Contrastive captioners are image-text foundation models. *arXiv* **2022**, arXiv:2205.01917.
33. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J. Grounded language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
34. Yao, L.; Han, J.; Wen, Y.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; Xu, C.; Xu, H. DetCLIP: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv* **2023**, arXiv:2209.09407.
35. Zhang, R.; Jiang, Z.; Guo, Z.; Yan, S.; Pan, J.; Dong, H.; Gao, P.; Li, H. Personalize segment anything model with one shot. *arXiv* **2023**, arXiv:2305.03048.
36. Ge, Y.; Ge, Y.; Zeng, Z.; Wang, X.; Shan, Y. Planting a SEED of vision in large language model. *arXiv* **2023**, arXiv:2307.08041.

37. Ren, T.; Liu, S.; Zeng, A. Grounded SAM: Assembling open-world models for diverse visual tasks. *arXiv* **2024**, arXiv:2401.14159.
38. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. *arXiv* **2023**, arXiv:2303.05499.
39. Wang, T.; Zhang, J.; Fei, J.; Ge, Y.; Zheng, H.; Tang, Y.; Li, Z.; Gao, M.; Zhao, S.; Shan, Y.; et al. Caption anything: Interactive image description with diverse multimodal controls. *arXiv* **2023**, arXiv:2305.02677.
40. Lüddecke, T.; Ecker, A. Image segmentation using text and image prompts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
41. Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. SegGPT: Segmenting everything in context. *arXiv* **2023**, arXiv:2304.03284.
42. Zou, X.; Yang, J.; Zhang, H.; Li, L.; Gao, J.; Lee, Y.J. Segment everything everywhere all at once. *arXiv* **2023**, arXiv:2304.06718.
43. Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S.Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv* **2024**, arXiv:2403.14608.
44. Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; Qiao, Y.J. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.* **2024**, *132*, 581–595. [[CrossRef](#)]
45. Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Zhang, S.; Wang, Y.; Li, Z.; Sun, L.; Mao, P.; Zang, Y. SAM fails to segment anything? SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more. *arXiv* **2023**, arXiv:2304.09148.
46. Zhou, C.; Loy, C.C.; Dai, B. Extract free dense labels from CLIP. In *European Conference on Computer Vision*; Springer Nature Switzerland: Cham, Switzerland, 2022.
47. 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](#)]
48. Rao, Y.; Zhao, W.; Chen, G.; Tang, Y.; Zhu, Z.; Huang, G.; Zhou, J.; Lu, J. DenseCLIP: Language-guided dense prediction with context-aware prompting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
49. Zhong, Y.; Yang, J.; Zhang, P. RegionCLIP: Region-based language-image pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
50. Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. CRIS: CLIP-driven referring image segmentation. *arXiv* **2021**, arXiv:2111.15174.
51. Guo, W.; Li, S.; Yang, J. Scattering prompt tuning: A fine-tuned foundation model for SAR object recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024.
52. Ma, X.; Zhang, X.; Pun, M.; Huang, B. MANet: Fine-Tuning Segment Anything Model for Multimodal Remote Sensing Semantic Segmentation. *arXiv* **2024**, arXiv:2410.11160.
53. Hu, L.; Lu, W.; Yu, H.; Yin, D.; Sun, X.; Fu, K. TEA: A training-efficient adapting framework for tuning foundation models in remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5648118. [[CrossRef](#)]
54. Yuan, Y.; Zhan, Y.; Xiong, Z. Parameter-efficient transfer learning for remote sensing image–text retrieval. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [[CrossRef](#)]
55. Christie, G.A.; Fendley, N.; Wilson, J.; Mukherjee, R. Functional map of the world. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
56. Schmitt, M.; Hughes, L.H.; Qiu, C.; Zhu, X. SEN12MS—A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *arXiv* **2019**, arXiv:1906.07789. [[CrossRef](#)]
57. Long, Y.; Xia, G.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.; Zhang, L.; Li, D. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 4205–4230. [[CrossRef](#)]
58. Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
59. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv* **2024**, arXiv:2401.16822.
60. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Li, X. RSGPT: A remote sensing vision language model and benchmark. *arXiv* **2023**, arXiv:2307.15266.
61. Wang, D.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Scaling-up remote sensing segmentation dataset with segment anything model. *arXiv* **2023**, arXiv:2305.02034.
62. Heidler, K.; Mou, L.; Hu, D.; Jin, P.; Li, G.; Gan, C.; Wen, J.; Zhu, X.X. Self-supervised audiovisual representation learning for remote sensing data. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103130. [[CrossRef](#)]
63. 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. *arXiv* **2023**, arXiv:2306.11300. [[CrossRef](#)]

64. Zhan, Y.; Xiong, Z.; Yuan, Y. RSVG: Exploring data and models for visual grounding on remote sensing data. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5604513. [\[CrossRef\]](#)
65. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. SkyScript: A large and semantically diverse vision-language dataset for remote sensing. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
66. Yuan, Z.; Xiong, Z.; Mou, L.; Zhu, X.X. ChatEarthNet: A global-scale image-text dataset empowering vision-language geo-foundation models. *arXiv* **2024**, arXiv:2402.11325.
67. Roberts, J.; Han, K.; Albanie, S. SATIN: A multi-task metadataset for classifying satellite imagery using vision-language models. *arXiv* **2023**, arXiv:2304.11619.
68. Zhao, Y.; Zhang, M.; Yang, B.; Zhang, Z.; Kang, J.; Gong, J. LuoJiaHOG: A hierarchy oriented geo-aware image caption dataset for remote sensing image-text retrieval. *arXiv* **2024**, arXiv:2403.10887.
69. Muhtar, D.; Li, Z.; Gu, F.; Zhang, X.; Xiao, P. LHRS-Bot: Empowering remote sensing with VGI-enhanced large multimodal language model. *arXiv* **2024**, arXiv:2402.02544.
70. Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; et al. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv* **2024**, arXiv:2406.10100.
71. Xing, J.; Liu, J.; Wang, J.; Sun, L.; Chen, X.; Gu, X.; Wang, Y. A survey of efficient fine-tuning methods for vision-language models-prompt and adapter. *Comput. Graph.* **2024**, *119*, 103885. [\[CrossRef\]](#)
72. Wang, Y.; Albrecht, C.M.; Zhu, X.X. Self-supervised vision transformers for joint sar-optical representation learning. In Proceedings of the IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Kuala Lumpur, Malaysia, 17–22 July 2022.
73. Available online: <https://github.com/RubenGres/Seg2Sat> (accessed on 12 October 2023).
74. Espinosa, M.; Crowley, E.J. Generate your own Scotland: Satellite image generation conditioned on maps. *arXiv* **2023**, arXiv:2308.16648.
75. Khanna, S.; Liu, P.; Zhou, L.; Meng, C.; Rombach, R.; Burke, M.; Lobell, D.B.; Ermon, S. DiffusionSat: A generative foundation model for satellite imagery. *arXiv* **2023**, arXiv:2312.03606.
76. Tang, D.; Cao, X.; Hou, X.; Jiang, Z.; Liu, J.; Meng, D. CRS-Diff: Controllable generative remote sensing foundation model. *arXiv* **2024**, arXiv:2403.11614.
77. Yu, Z.; Liu, C.; Liu, L.; Shi, Z.X.; Zou, Z. MetaEarth: A generative foundation model for global-scale remote sensing image generation. *arXiv* **2024**, arXiv:2405.13570. [\[CrossRef\]](#)
78. Fuller, A.; Millard, K.; Green, J.R. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *arXiv* **2023**, arXiv:2311.00566.
79. Akiva, P.; Purri, M.; Leotta, M. Self-supervised material and texture representation learning for remote sensing tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
80. Li, W.; Chen, K.; Chen, H.; Shi, Z. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5405516. [\[CrossRef\]](#)
81. Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-aware self-supervised learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
82. Li, W.; Wei, Y.; Liu, T.; Hou, Y.; Li, Y.; Liu, Z.; Liu, Y.; Liu, L. Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture. *arXiv* **2023**, arXiv:2311.15153. [\[CrossRef\]](#)
83. Wang, M.; Chen, K.; Su, L.; Yan, C.; Xu, S.; Zhang, H.; Yuan, P.; Jiang, X.; Zhang, B. RSBuilding: Towards general remote sensing image building extraction and change detection with foundation model. *arXiv* **2024**, arXiv:2403.07564. [\[CrossRef\]](#)
84. Zhang, T.; Gao, P.; Dong, H.; Yin, Z.; Wang, G.; Zhang, W.; Chen, H. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sens.* **2022**, *14*, 5675. [\[CrossRef\]](#)
85. Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; Chen, C.; Li, M. Towards geospatial foundation models via continual pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
86. 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.* **2023**, *61*, 5612822. [\[CrossRef\]](#)
87. Smith, M.J.; Fleming, L.; Geach, J.E. EarthPT: A time series foundation model for earth observation. *arXiv* **2023**, arXiv:2309.07207.
88. Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. Foundation models for generalist geospatial artificial intelligence. *arXiv* **2023**, arXiv:2310.18660.
89. Shen, J.; Huo, C.; Xiang, S. Siamese InternImage for Change Detection. *Remote Sens.* **2024**, *16*, 3642. [\[CrossRef\]](#)
90. Liu, Y.; Xiao, J.; Guo, Y.; Jiang, P.; Yang, H.; Wang, F. HSIDMamba: Exploring bidirectional state-space models for hyperspectral denoising. *arXiv* **2024**, arXiv:2404.09697.
91. Fu, G.; Xiong, F.; Lu, J.; Zhou, J. SSUMamba: Spatial-spectral selective state space model for hyperspectral image denoising. *arXiv* **2024**, arXiv:2405.01726. [\[CrossRef\]](#)

92. Zhou, H.; Wu, X.; Chen, H.; Chen, X.; He, X. RSDehamba: Lightweight vision mamba for remote sensing satellite image dehazing. *arXiv* **2024**, arXiv:2405.10030.
93. Xiao, Y.; Yuan, Q.; Jiang, K.; Chen, Y.; Zhang, Q.; Lin, C. Frequency-assisted Mamba for remote sensing image super-resolution. *arXiv* **2024**, arXiv:2405.04964. [[CrossRef](#)]
94. He, X.; Cao, K.; Yan, K.; Li, R.; Xie, C.; Zhang, J.; Zhou, M. Pan-Mamba: Effective pan-sharpening with state space model. *arXiv* **2024**, arXiv:2402.12192. [[CrossRef](#)]
95. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z.X. RSMamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [[CrossRef](#)]
96. Yao, J.; Hong, D.; Li, C.; Chanussot, J. SpectralMamba: Efficient mamba for hyperspectral image classification. *arXiv* **2024**, arXiv:2404.08489.
97. Wang, G.; Zhang, X.; Peng, Z.; Zhang, T.; Jia, X.; Jiao, L. S2Mamba: A spatial-spectral state space model for hyperspectral image classification. *arXiv* **2024**, arXiv:2404.18213.
98. Huang, L.; Chen, Y.; He, X. Spectral-spatial Mamba for hyperspectral image classification. *arXiv* **2024**, arXiv:2404.18401.
99. Sheng, J.; Zhou, J.; Wang, J.; Ye, P.; Fan, J. DualMamba: A lightweight spectral-spatial Mamba-convolution network for hyperspectral image classification. *arXiv* **2024**, arXiv:2406.07050. [[CrossRef](#)]
100. Zhu, Q.; Cai, Y.; Fang, Y.; Yang, Y.; Chen, C.; Fan, L.; Nguyen, A. Samba: Semantic segmentation of remotely sensed images with state space model. *arXiv* **2024**, arXiv:2404.01705. [[CrossRef](#)]
101. Ma, X.; Zhang, X.; Pun, M.O. RS3Mamba: Visual state space model for remote sensing image semantic segmentation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6011405. [[CrossRef](#)]
102. Verma, T.; Singh, J.; Bhartari, Y.; Jarwal, R.; Singh, S.; Singh, S. SOAR: Advancements in small body object detection for aerial imagery using state space models and programmable gradients. *arXiv* **2024**, arXiv:2405.01699.
103. Chen, H.; Song, J.; Han, C.; Xia, J.; Yokoya, N. ChangeMamba: Remote sensing change detection with spatio-temporal state space model. *arXiv* **2024**, arXiv:2404.03425.
104. Zhang, H.; Chen, K.; Liu, C.; Chen, H.; Zou, Z.; Shi, Z.X. CDMamba: Remote sensing image change detection with Mamba. *arXiv* **2024**, arXiv:2406.04207.
105. Li, K.; Cao, X.; Meng, D. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5610112. [[CrossRef](#)]
106. Dong, J.; Yin, H.; Li, H.; Li, W.; Zhang, Y.; Khan, S.; Khan, F.S. Dual hyperspectral Mamba for efficient spectral compressive imaging. *arXiv* **2024**, arXiv:2406.00449.
107. Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; Ermon, S. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 197–211.
108. 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.* **2023**, *61*, 5612822. [[CrossRef](#)]
109. Cha, K.; Seo, J.; Lee, T. A billion-scale foundation model for remote sensing images. *arXiv* **2023**, arXiv:2304.05215. [[CrossRef](#)]
110. Muhtar, D.; Zhang, X.; Xiao, P.; Li, Z.; Gu, F. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5607817. [[CrossRef](#)]
111. L, W.; Yang, W.; Hou, Y.; Liu, L.; Liu, Y.; Li, X. SARATR-X: A foundation model for synthetic aperture radar images target recognition. *arXiv* **2024**, arXiv:2405.09365.
112. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing remote sensing foundation model via multi-task pre-training. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654. [[CrossRef](#)]
113. Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A.J.; Hanna, J.; Borth, D.; Papoutsis, I.; Saux, B.L.; Camps-Valls, G.; Zhu, X.X. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv* **2024**, arXiv:2403.15356.
114. 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, Seattle, WA, USA, 17–18 June 2024.
115. Han, B.; Zhang, S.; Shi, X.; Reichstein, M. Bridging remote sensors with multisensor geospatial foundation models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024.
116. Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *arXiv* **2023**, arXiv:2311.07113. [[CrossRef](#)] [[PubMed](#)]
117. Dong, Z.; Gu, Y.; Liu, T. Generative ConvNet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5603816. [[CrossRef](#)]
118. Tian, J.; Lei, J.; Zhang, J.; Xie, W.; Li, Y. SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5613213. [[CrossRef](#)]
119. Huang, Z.; Zhang, M.; Gong, Y.; Liu, Q.; Wang, Y. Generic knowledge boosted pre-training for remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5605913.

120. Wang, Y.; Hern'andez, H.H.; Albrecht, C.M.; Zhu, X.X. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv* **2023**, arXiv:2310.18653. [[CrossRef](#)]
121. Zhang, M.; Liu, Q.; Wang, Y. CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding. *arXiv* **2023**, arXiv:2310.00022.
122. Wang, Y.; Albrecht, C.M.; Braham, N.A.A.; Liu, C.; Xiong, Z.; Zhu, X.X. DeCUR: Decoupling common and unique representations for multimodal self-supervision. *arXiv* **2023**, arXiv:2309.05300.
123. Nayak, N.V.; Yu, P.; Bach, S.H. Learning to compose soft prompts for compositional zero-shot learning. *arXiv* **2022**, arXiv:2204.03574.
124. Mall, U.; Phoo, C.P.; Liu, M.K.; Vondrick, C.; Hariharan, B.; Bala, K. Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv* **2023**, arXiv:2312.06960.
125. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-time open-vocabulary object detection. *arXiv* **2024**, arXiv:2401.17270.
126. 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](#)]
127. Zhan, Y.; Xiong, Z.; Yuan, Y. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *arXiv* **2024**, arXiv:2401.09712.
128. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Zhou, J. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [[CrossRef](#)]
129. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.H.; Khan, F.S. GeoChat: Grounded large vision-language model for remote sensing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024.
130. Pang, C.; Wu, J.; Li, J.; Li, J.; Liu, Y.; Sun, J.; Li, W.; Wang, S.; Feng, L.; Xia, G.-S.; et al. H2RSVLM: Towards helpful and honest remote sensing large vision language model. *arXiv* **2024**, arXiv:2403.20213.
131. Vivanco Cepeda, V.; Nayak, G.K.; Shah, M. GeoCLIP: CLIP-inspired alignment between locations and images for effective worldwide geo-localization. *arXiv* **2023**, arXiv:2309.16020.
132. Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; Rußwurm, M. SatCLIP: Global, general-purpose location embeddings with satellite imagery. *arXiv* **2023**, arXiv:2311.17179.
133. Zhang, J.; Zhou, Z.; Mai, G.; Mu, L.; Hu, M.; Li, S. Text2Seg: Remote sensing image semantic segmentation via text-guided visual foundation models. *arXiv* **2023**, arXiv:2304.10597.
134. Osco, L.P.; Wu, Q.; Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.; Li, J.; Junior, J.M. The segment anything model (SAM) for remote sensing applications: From zero to one shot. *arXiv* **2023**, arXiv:2306.16623. [[CrossRef](#)]
135. Zhou, X.; Liang, F.; Chen, L.; Liu, H.; Song, Q.; Vivone, G.; Chanussot, J. MeSAM: Multiscale enhanced segment anything model for optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5623515. [[CrossRef](#)]
136. 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](#)]
137. Qi, Z.; Liu, C.; Liu, Z.; Chen, H.; Wu, Y.; Zou, Z.; Sh, Z. Multi-view remote sensing image segmentation with SAM priors. *arXiv* **2024**, arXiv:2405.14171.
138. Wang, Y.; Zhang, W.; Chen, W.; Chen, C. BSDSNet: Dual-stream feature extraction network based on segment anything model for synthetic aperture radar land cover classification. *Remote Sens.* **2024**, *16*, 1150. [[CrossRef](#)]
139. Xue, B.; Cheng, H.; Yang, Q.; Wang, Y.; He, X. Adapting segment anything model to aerial land cover classification with low-rank adaptation. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 2502605. [[CrossRef](#)]
140. Kang, J.; Chen, N.; Li, M.; Mao, S.; Zhang, H.; Fan, Y.; Liu, H. A point cloud segmentation method for dim and cluttered underground tunnel scenes based on the segment anything model. *Remote Sens.* **2023**, *16*, 97. [[CrossRef](#)]
141. Ji, Y.; Wu, W.; Wan, G.; Zhao, Y.; Wang, W.; Yin, H.; Tian, Z.; Liu, S. Segment anything model-based building footprint extraction for residential complex spatial assessment using LiDAR data and very high-resolution imagery. *Remote Sens.* **2024**, *16*, 2661. [[CrossRef](#)]
142. Hetang, C.; Xue, H.; Le, C.; Yue, T.; Wang, W.; He, Y. Segment anything model for road network graph extraction. *arXiv* **2024**, arXiv:2403.16051.
143. Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z.X. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 4701117. [[CrossRef](#)]
144. Chen, X.; Wu, W.; Yang, W.; Qin, H.; Wu, X.; Mao, X. Make segment anything model perfect on shadow detection. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4410713. [[CrossRef](#)]
145. Sun, Z.; Song, H.; Zhang, K.; Dong, G.; Liang, L.; Zhao, Y. Segment anything model guided semantic knowledge learning for remote sensing change detection. In Proceedings of the ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024.

146. Du, S.; Tang, S.; Wang, W.; Li, X.; Guo, R. Tree-GPT: Modular large language model expert system for forest remote sensing image understanding and interactive analysis. *arXiv* **2023**, arXiv:2310.04698. [\[CrossRef\]](#)
147. Wang, Z.; Zhang, Y.; Zhang, Z.; Jiang, Z.; Yu, Y.; Li, L.; Li, L. Exploring semantic prompts in the segment anything model for domain adaptation. *Remote Sens.* **2024**, *16*, 758. [\[CrossRef\]](#)
148. Available online: <https://github.com/Douglas2Code/Text2Seg> (accessed on 7 June 2023).
149. Zheng, Z.; Zhong, Y.; Zhang, L.; Ermon, S. Segment any change. *arXiv* **2024**, arXiv:2402.01188.
150. Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; Bruzzone, L. Adapting segment anything model for change detection in VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611711. [\[CrossRef\]](#)
151. Chen, K.; Liu, C.; Li, W.; Liu, Z.; Chen, H.; Zhang, H.; Zou, Z.; Shi, Z.X. Time travelling pixels: Bitemporal features integration with foundation model for remote sensing image change detection. *arXiv* **2023**, arXiv:2312.16202.
152. Zhu, Y.; Li, L.; Chen, K.; Liu, C.; Zhou, F.; Shi, Z.X. Semantic-CC: Boosting remote sensing image change captioning via foundational knowledge and semantic guidance. *arXiv* **2024**, arXiv:2407.14032. [\[CrossRef\]](#)
153. Zhang, S.; Wang, Q.; Liu, J.; Xiong, H. ALPS: An auto-labeling and pre-training scheme for remote sensing segmentation with segment anything model. *arXiv* **2024**, arXiv:2406.10855.
154. Mei, L.; Ye, Z.; Xu, C.; Wang, H.; Wang, Y.; Lei, C.; Yang, W.; Li, Y. SCD-SAM: Adapting segment anything model for semantic change detection in remote sensing imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5626713. [\[CrossRef\]](#)
155. Zhang, J.; Yang, X.; Jiang, R.; Shao, W.; Zhang, L. RSAM-Seg: A SAM-based approach with prior knowledge integration for remote sensing image semantic segmentation. *arXiv* **2024**, arXiv:2402.19004.
156. Ma, X.; Wu, Q.; Zhao, X.; Zhang, X.; Pun, M.; Huang, B. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *arXiv* **2023**, arXiv:2312.02464. [\[CrossRef\]](#)
157. Ji, W.; Li, J.; Bi, Q.; Li, W.; Cheng, L. Segment anything is not always perfect: An investigation of SAM on different real-world applications. *arXiv* **2023**, arXiv:2304.05750. [\[CrossRef\]](#)
158. Ren, S.; Luzi, F.; Lahrichi, S.; Kassaw, K.; Collins, L.M.; Bradbury, K.; Malof, J.M. Segment anything, from space? In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024.
159. Available online: <https://github.com/comeony/RingMo> (accessed on 17 February 2023).
160. An, X.; Sun, J.; Gui, Z.; He, W. COREval: A comprehensive and objective benchmark for evaluating the remote sensing capabilities of large vision-language models. *arXiv* **2024**, arXiv:2411.18145.
161. Corley, I.A.; Robinson, C.; Dodhia, R.; Ferres, J.M.; Najafirad, P. Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters. *arXiv* **2023**, arXiv:2305.13456.
162. Chang, Y.; Stewart, A.J.; Bastani, F.; Wolters, P.; Kannan, S.; Huber, G.R.; Wang, J.; Banerjee, A. On the generalizability of foundation models for crop type mapping. *arXiv* **2024**, arXiv:2409.09451.
163. Li, J.; Xu, M.; Xiang, L.; Chen, D.; Zhuang, W.; Yin, X.; Li, Z. Foundation models in smart agriculture: Basics, opportunities, and challenges. *Comput. Electron. Agric.* **2024**, *222*, 109032. [\[CrossRef\]](#)
164. Tan, C.; Cao, Q.; Li, Y.; Zhang, J.; Yang, X.; Zhao, H.; Wu, Z.; Liu, Z.; Yang, H.; Wu, N.; et al. On the promises and challenges of multimodal foundation models for geographical, environmental, agricultural, and urban planning applications. *arXiv* **2023**, arXiv:2312.17016.
165. Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; Xie, S. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. *arXiv* **2024**, arXiv:2401.06209.
166. Xu, M.; Yin, W.; Cai, D.; Yi, R.; Xu, D.; Wang, Q.; Wu, B.; Zhao, Y.; Yang, C.; Wang, S.; et al. A survey of resource-efficient LLM and multimodal foundation models. *arXiv* **2024**, arXiv:2401.08092.
167. Frantar, E.; Alistarh, D. SparseGPT: Massive language models can be accurately pruned in one-shot. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
168. Pietro'n, M.; Żurek, D. Speedup deep learning models on GPU by taking advantage of efficient unstructured pruning and bit-width reduction. *J. Comput. Sci.* **2021**, *67*, 101971. [\[CrossRef\]](#)
169. Hershcovitch, M.; Choshen, L.; Wood, A.; Enmouri, I.; Chin, P.; Sundararaman, S.; Harnik, D. Lossless and near-lossless compression for foundation models. *arXiv* **2024**, arXiv:2404.15198.
170. Xing, X.; Du, L.; Wang, X.; Zeng, X.; Wang, Y.; Zhang, Z.; Zhang, J. BiPFT: Binary Pre-trained Foundation Transformer with Low-rank Estimation of Binarization Residual Polynomials. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.

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.