remote sensing

MDPI

Review

Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques

Lijie Tao [1], Haokui Zhang [1,*], Haizhao Jing [1], Yu Liu [2], Dawei Yan [1], Guoting Wei [3] and Xizhe Xue [4]

[1] School of Cybersecurity, Northwestern Polytechnical University, Xi'an 710072, China; taolijie11111@mail.nwpu.edu.cn (L.T.); haizhao_jing@mail.nwpu.edu.cn (H.J.); ydw4396@mail.nwpu.edu.cn (D.Y.)
[2] Zhejiang Lab, Hangzhou 311500, China; uniliu@zhejianglab.org
[3] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; weiguoting@njust.edu.cn
[4] Department of Aerospace and Geodesy, Technical University of Munich, 80333 Munich, Germany; xuexizhe@mail.nwpu.edu.cn
* Correspondence: hkzhang@nwpu.edu.cn

Abstract: Recently, the remarkable success of ChatGPT has sparked a renewed wave of interest in artificial intelligence (AI), and the advancements in Vision–Language Models (VLMs) have pushed this enthusiasm to new heights. Differing from previous AI approaches that generally formulated different tasks as discriminative models, VLMs frame tasks as generative models and align language with visual information, enabling the handling of more challenging problems. The remote sensing (RS) field, a highly practical domain, has also embraced this new trend and introduced several VLM-based RS methods that have demonstrated promising performance and enormous potential. In this paper, we first review the fundamental theories related to VLM, then summarize the datasets constructed for VLMs in remote sensing and the various tasks they address. Finally, we categorize the improvement methods into three main parts according to the core components of VLMs and provide a detailed introduction and comparison of these methods.

Keywords: vision–language models; remote sensing

check for updates

Academic Editor: Dino Ienco

Received: 8 October 2024
Revised: 25 December 2024
Accepted: 29 December 2024
Published: 6 January 2025

Citation: Tao, L.; Zhang, H.; Jing, H.; Liu, Y.; Yan, D.; Wei, G.; Xue, X. Advancements in Vision–Language Models for Remote Sensing: Datasets, Capabilities, and Enhancement Techniques. *Remote Sens.* **2025**, *17*, 162. https://doi.org/10.3390/rs17010162

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

Remote sensing technology can help people obtain various observation data from a distance, and it is widely applied in fields such as disaster monitoring [1], urban planning [2], agricultural management [3], surface observation [4], environmental monitoring [5], etc. Consequently, remote sensing image processing technology has been a key topic of focus for researchers both domestically and internationally. Over the past decade, with the development of artificial intelligence technology, significant breakthroughs have also been made in remote sensing image processing techniques. Specifically, various 3D convolutional neural networks (CNNs) [6] have been specifically designed for hyperspectral and multispectral image classification, as well as spectral reconstruction. U-Net [7] is utilized for semantic segmentation and land cover mapping in aerial imagery. A pyramid network has been proposed for multi-scale object detection in Special Administrative Region (SAR) images. The YOLO framework [8] is applied for small target detection in infrared remote sensing images. Additionally, AI techniques are also applied in fields such as denoising, fusion, enhancement, and compression of remote sensing images, achieving remarkable results. For specific arts in remote sensing, FFCA-YOLO [9] designed three plug-and-play modules, which are used to enhance and fuse features in the network, as well as to be spatial context aware, in order to improve local region perception and global correlation as much as

possible without increasing complexity. Zhang et al. [10] improved YOLOv5's perception and detection by integrating spatial-to-depth (SPD) elements, and the CoTC3 module to utilize contextual information. Liu et al. [11] proposed the MLPA method, which enhances cross-domain hyperspectral imaging (HSI) classification by aligning multi-level features to improve generalization and discriminative ability in the target domain. LBA-MCNet [12] enhanced ORSI-SOD by integrating localization, balancing, and affinity to improve boundary detection and foreground–background context modeling. MSC-GAN [13] leveraged MSC architecture, GFR modules, and channel enhancement to improve multitemporal cloud removal through efficient feature interaction and fusion. In summary, AI techniques have significantly advanced the processing of remote sensing images, elevating the capabilities in these fields. However, most of these methods primarily focus on image processing, and the models they employ are discriminative models. There are inherent limitations to this type of technology, such as the inability to incorporate some human common sense, and the trained models can only perform a single vision task.

Recently, due to advancements in data availability, computational power, and the introduction of the Transformer model [14], natural language processing (NLP) has made significant breakthroughs. By feeding tens of billions of text tokens and scaling the model to tens of billions of parameters, ChatGPT [15] achieved amazing performance in natural language understanding, language translation, text generation, question answering, etc. The impressive success of ChatGPT [15] has reignited enthusiasm for AI. Larger and more advanced models continue to emerge, showcasing unprecedented advancements in language understanding through their extensive knowledge and sophisticated reasoning abilities, demonstrating human-like thinking. Specifically, Google released the encoder-only model BERT [16], which pushed the GLUE [17] benchmark to 80.4% and achieved an accuracy of 86.7% on MultiNLI [18] and set a new record of 93.2 for the F1 score on the SQuAD v1.1 [19] question-answering test, surpassing human performance by 2.0 points. Meta released LLaMA [20], which contains a series of models for 7 billion parameters to 65 billion parameters, and it can beat GPT-3 [21] with 1/10 parameters. The development of large language models reveals new possibilities for the advancement of artificial intelligence technology, particularly in generative AI. Compared to discriminative AI, generative AI models approach problem-solving in a more flexible and open manner, allowing for the modeling of a broader range of issues. The development of large language models (LLMs) also offers new insights for addressing the challenges faced by discriminative models in the computer vision field.

The introduction of models such as CLIP [22] has effectively bridged the modality gap between images and language, creating opportunities for innovative integration. CLIP [22], with its innovative multimodal encoder–decoder structure utilizing BERT [16] as a text encoder and Vision Transformer (ViT) [23]/ResNet [24] as an image encoder, exemplifies the synergy between images and language vocabulary, leading to improved accuracy in classification and detection tasks through a deep understanding of the relationship between image content and textual vocabulary. GLIP [25] further enhances this consistency between language sentences and image content. Furthermore, the emergence of VLM has fundamentally addressed numerous constraints associated with traditional discriminative models in remote sensing image processing, thus catalyzing the transition to AI 2.0. Generally speaking, VLM is a type of model designed to integrate and understand both visual and textual information, enabling tasks such as Image Captioning, Visual Question Answering, and crossmodal retrieval by leveraging the relationships between images and language. Compared to previous discriminative models, VLM offers greater flexibility in modeling various AI tasks, allowing it to handle multiple tasks within a single framework. Additionally, its multimodal perceptual ability brings it closer to human perception. In

the past two years, both LLMs and VLMs have developed rapidly, with various stronger models being introduced, such as LLaMA [20], LLaVA [26], and GPT-4 [27], among others. The performance of related tasks has also been continually improved.

Similarly to previous discriminative AI technologies, which have significantly improved the field of remote sensing data processing, VLMs—one of the core technologies in AI 2.0—also bring new opportunities for this domain. VLMs, by leveraging their multimodal capabilities, have demonstrated significant progress across various tasks within remote sensing, including geophysical classification [28,29], object detection [30,31], and scene understanding [32,33]. By introducing LLaVA-1.5 [26] into the remote sensing field, Kuckreja et al. proposed GeoChat [30], which not only answers image-level queries but also accepts region inputs for region-specific dialogue [30]. Deng et al. designed ChangeChat, which utilizes multimodal instruction tuning to handle complex queries such as change captioning, category-specific quantification, and change localization [34]. Inspired by GPT-4 [27], Hu et al. developed a high-quality Remote Sensing Image Captioning dataset (RSICap) to facilitate the advancement of large VLMs in the remote sensing domain [35]. Wang et al. [36] proposed a zero-shot target recognition framework for SAR, which combines a vision–language model with a diffusion generative model to achieve the construction of 3D models. Changen2 [37] is a generative model that creates time-series images with semantic and changeable labels from single-temporal images for collecting and annotating large multitemporal remote sensing datasets. The publication growth trend in Figure 1 illustrates the development timeline of Visual–Language Models (VLMs) in the remote sensing (RS) field. Growth began after the release of CLIP [22], and in early 2023, following the successive releases of LLaVA [26] and GPT-4 [27], numerous high-performing VLMs started to emerge in the RS domain.

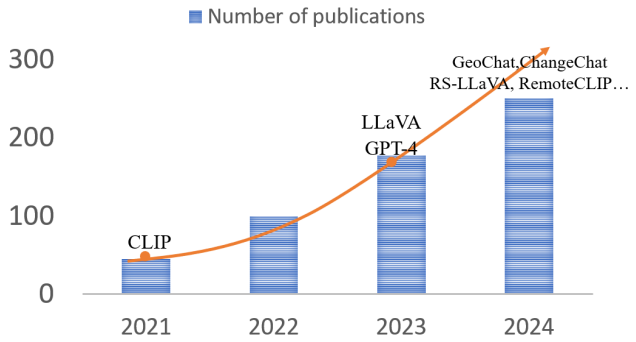

Figure 1. Number of publications for VLMs in RS (from Web of Science).

In practical applications, VLMs are beginning to demonstrate their potential in the remote sensing domain, albeit still in the early stages of adoption. A prominent example is GeoForge [38], which utilizes a customized GPT-4 model trained with additional geospatial vocabulary and knowledge to handle tasks like mapping administrative boundaries and geocoding major cities. For instance, users can interact with GeoForge to easily obtain the administrative boundaries of Bangladesh and the names of its five most important cities, and can further generate geocoded markers for these five cities based on the obtained information. Despite these advancements, the range of practical applications remains limited, and widespread industry implementation is yet to be realized. This early stage underscores the significant potential of VLMs like GeoForge to enhance remote sensing workflows while highlighting the need for ongoing development to overcome existing challenges and expand their real-world applicability.

Beyond these direct applications, VLMs also demonstrate potential advantages over traditional models in handling long-tail distributions and supporting multi-task learning. Pre-trained on extensive and diverse datasets, VLMs can achieve strong generalization

capabilities, enabling them to effectively generalize to rare categories and thus mitigate the challenges associated with long-tail distributions. In addition, thanks to their generative output paradigm, VLMs can simultaneously tackle multiple tasks—such as classification, question answering, and counting—within a single framework.

In this paper, we present a comprehensive review of VLM-based remote sensing applications; to ensure coverage of relevant research, we utilized Google Scholar and employed keywords such as "Visual language models for remote sensing", "Multimodal remote sensing datasets", and "Deep learning for remote sensing imagery", offering a thorough delineation of the principal methodologies and the most recent advancements in VLMs, particularly within the context of remote sensing. By providing an in-depth introduction to the most representative and influential works in this domain, the survey enables researchers to quickly acquire a nuanced understanding of the development trajectory of these methodologies, facilitating the classification and evaluation of different approaches. Furthermore, the survey extends its contribution by presenting a detailed classification of enhancement techniques, focusing on key components such as vision encoders, text encoders, and alignment between vision and language. This classification, which differs from existing surveys categorized by task type [39], not only systematically illustrates the research progress made within this specialized field but also serves as a valuable reference point, supporting and inspiring further innovative research endeavors by scholars aiming to explore and advance the boundaries of VLM applications in remote sensing.

This survey is organized as shown in Figure 2. Section 2 presents a convincing introduction to foundation models, including Transformer, Vision Transformer, and the famous VLM—LLaVA. Section 3 outlines the datasets used in VLM for RS, which can be further divided manually into collected data, data combined with existing datasets, and data enhanced by VLMs and LLMs. Section 4 describes the capabilities of VLMs, specifically introducing the pure visual tasks and vision–language tasks. In Section 5, current VLM-based remote sensing data processing methods are introduced according to three major improvement directions, with a comparison and analysis of different approaches. Finally, several promising future directions are discussed in Section 6.

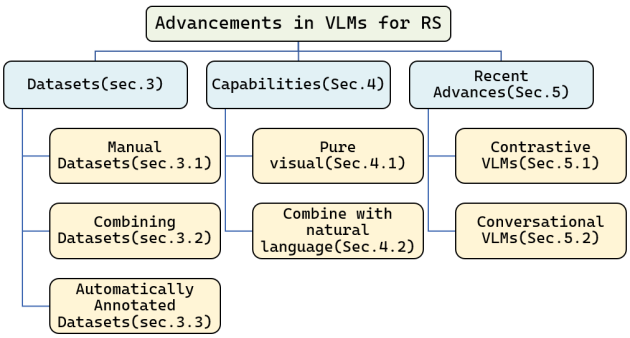

Figure 2. Organization of this survey.

2. Foundation Models

2.1. Transformer

The emergence of the Transformer has changed the paradigm in the NLP field and laid the foundation for the subsequent development of LLMs. As shown in Figure 3, the Transformer consists of an encoder and a decoder, each comprising an attention layer (also known as the token mixer layer) and a Feed Forward layer (FFN, sometimes referred to as the channel mixer layer). The attention layer allows for the exchange of information among all input tokens, while the Feed Forward layer facilitates information exchange across channels.The FFN layer consists of two linear layers. The width of the first linear

layer is set to 4× (the scale factor) that of the input tokens, while the second linear layer matches the width of the input tokens. The attention mechanism is formulated as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is a scaling factor, typically set to the dimension of the input feature head. Q, K, and V are variations of the input data, usually generated by a linear layer. For instance, $Q = W_q(x)$, $K = W_k(x)$, $V = W_v(x)$, where x denotes the input feature and W_q, W_k and W_v are learnable parameters of the linear layers. When Q, K and V are variations of the same input feature, it is referred to as self-attention. If Q varies from x_1 while K and V vary from x_2 (with x_1 and x_2 coming from different sources), it is called cross-attention.

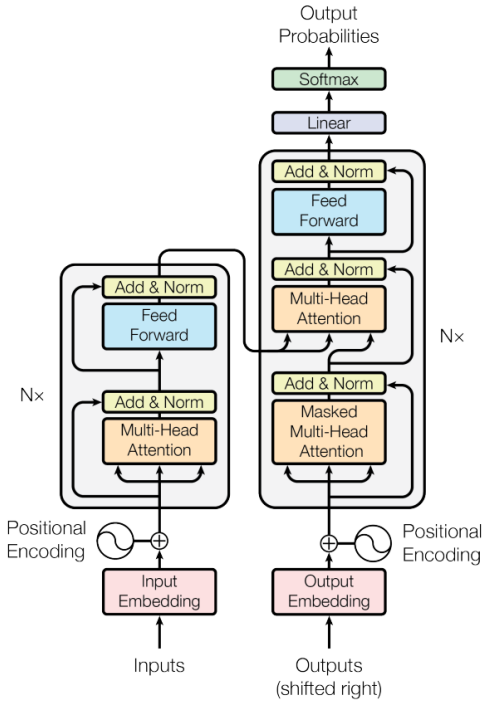

Figure 3. An illustration of Transformer [14].

The Transformer model offers several advantages over traditional RNNs and LSTMs. Firstly, it utilizes self-attention mechanisms, allowing it to process input sequences in parallel rather than sequentially, which significantly speeds up training. Secondly, the ability to capture long-range dependencies without the vanishing gradient problem enhances its effectiveness in understanding context. This efficiency and scalability enable the development of large language models (LLMs), as Transformers can handle vast amounts of data and complex patterns more effectively than RNNs and LSTMs, paving the way for advancements in natural language processing tasks.

2.2. Vision Transformer

Inspired by the success and popularity of transformer used in NLP, the Vision Transformer (ViT) model was proposed later and tailored for the computer vision field. As shown in Figure 4, ViT is composed of three components: Linear Projection of Flattened Patches (embedding layer), Transformer Encoder, and MLP Head. In summary, to enable NLP models to process images, the images are serialized by cropping them into uniform-sized patches. These patches are then transformed into features using a projection layer. To ensure the model is sensitive to the positional information of the image patches, position embeddings are added to the corresponding patch embeddings. Currently, there are

four types of position, embedding methods, including (1) absolute position embedding; (2) relative position embedding (RPE) [40]; (3) conditional position embedding (CPE) [41]; (4) rotary position embedding (RoPE) [42]. The ViT adopts the first one.

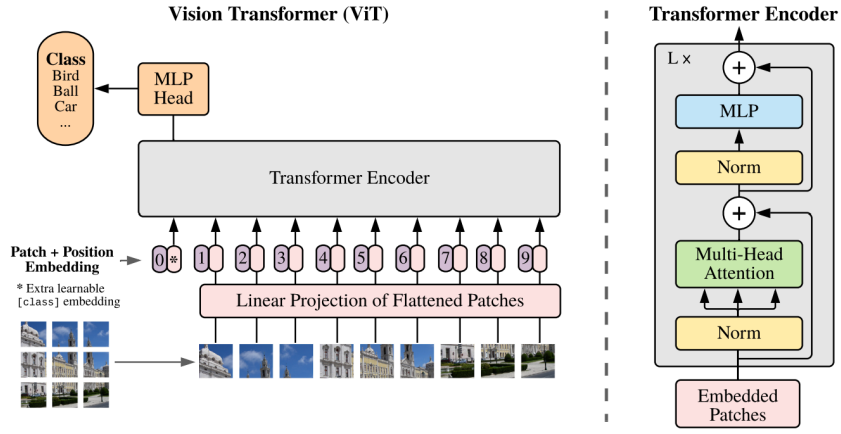

Figure 4. Model overview of Vision Transformer [23].

Compared to convolutional networks, ViT has a significant advantage in scaling-up ability. Its performance on image processing does not saturate quickly as the model size increases, allowing for better utilization of larger models and enhanced capabilities in handling more complex tasks. This characteristic has sparked a surge of research interest in the ViT architecture, leading to the continuous emergence of various improved versions of ViT models, such as SwinTransformer [43], Deit [44], etc. So far, the model paradigms in both the NLP and CV fields have been unified, which has laid the foundation for the emergence of VLMs.

2.3. Vision–Language Models

The introduction of Transformers has significantly advanced language processing, while the emergence of ViT has unified processing paradigms in both NLP and CV. Subsequently, CLIP [22] breaks down the remaining barriers between language and images. Building on these foundations, VLMs have been proposed by combining ViT and LLMs. A representative paradigm among these VLMs is LLaVA [26], whose architecture is illustrated in Figure 5.

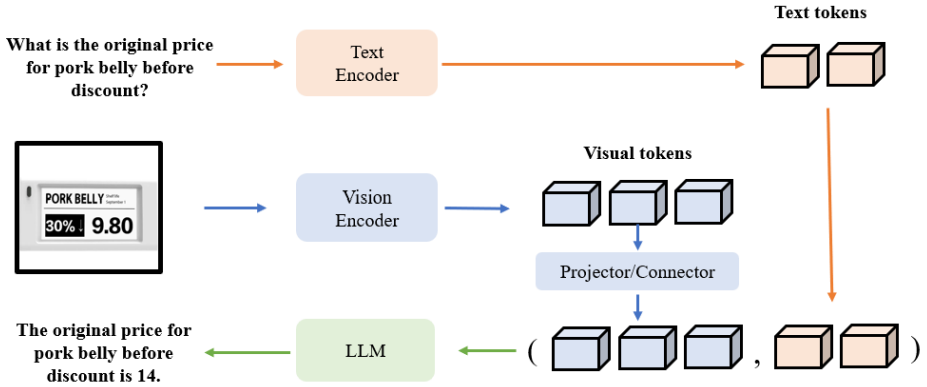

Figure 5. An illustration of LLaVA [26].

The core idea of LLaVA is to integrate visual tokens with textual tokens and feed them into an LLM to produce text-based answers. Specifically, the visual encoder extracts image features, and a projection layer converts these features into language-embedding tokens, bridging the gap between different modalities. These visual tokens are subsequently input into the LLM together with the text tokens, enabling multimodal interaction.

In practice, a pre-trained CLIP model is often employed as the visual encoder to extract visual tokens, because it has been trained extensively on large-scale image–text pairs. This training allows the CLIP visual encoder to map image features into a representation space comparable to that of text features, laying the groundwork for subsequent conversion into language-embedding tokens.

Although the merging approach is straightforward, the resulting VLM can handle a wider range of tasks and interact with users, enabling conversations based on the content of images. Figure 6 illustrates ten representative tasks of VLM. From the results, it is evident that the model can not only perceive the content of images but also possesses a certain degree of reasoning ability, as demonstrated in the example shown in the bottom left corner of the Figure 6.

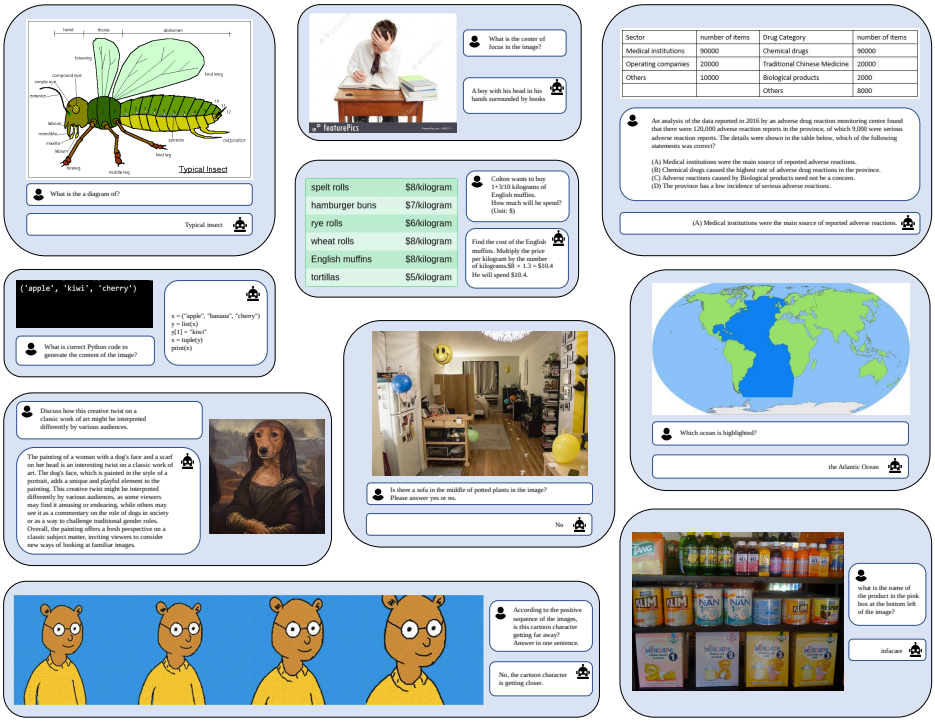

Figure 6. Ten representative tasks of VLM [45].

3. Datasets

Whether it is the previous discriminative models or the current generative models, they all fall under data-driven models. Therefore, VLM-based remote sensing image processing methods also require corresponding training data as support. Figure 7 compares the datasets used in previous discriminative models with those used in current VLMs. From the figure, we can observe that the most significant improvement is that VLM datasets include both image and language annotations.

In the remote sensing field, existing VLM datasets can be broadly categorized into three types based on their production approaches:

- Manual datasets, which are completely manually annotated by humans based on specific task requirements.
- Combined datasets, which are built by combining several existing datasets and adding part-new language annotations.
- Automatically annotated datasets. This type of data construction involves minimal human participation and relies on various multimodal models for filtering and annotating image data.

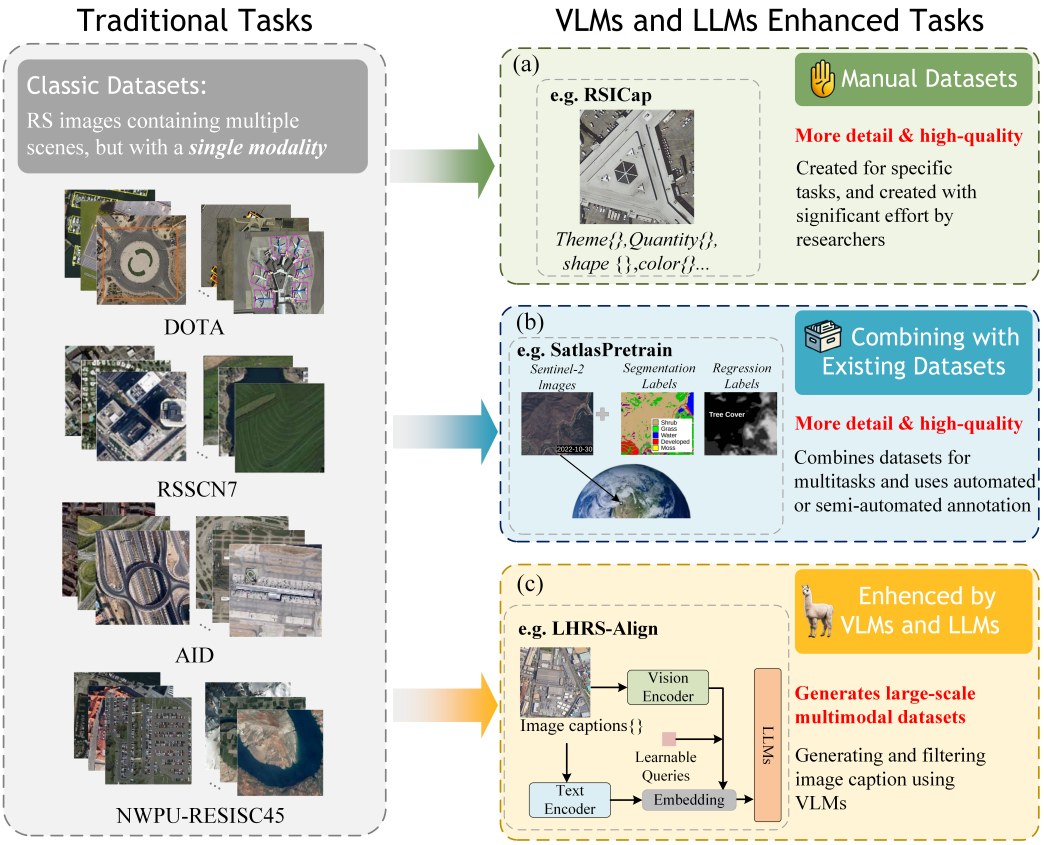

Figure 7. Comparison of traditional (**left**) and VLM datasets (**right**) in remote sensing.

In this section, we will delve into the construction methods and applicable task domains of existing remote sensing VLM datasets, offering insights into how these datasets support extensive applications and in-depth research within the field. The statistical information of current existing datasets are summarized in Table 1.

Table 1. Statistical information of datasets. K = thousands; M = million; SC = Scene Classification; VQA = Visual Question Answering; VG = Visual Grounding; RRSIS = Referring Remote Sensing Image Segmentation; IC = Image Captioning.

Method	Dataset	Year	Source	Image	Image Size	Task
Manual Datasets	HallusionBench [46]	2023	-	346	-	VQA
	RSICap [35]	2023	DOTA	2585	512	IC
	CRSVQA [47]	2023	AID	4639	600	VQA
Combining Datasets	SATIN [48]	2023	Million-AID, WHU-RS19, SAT-4, AID [1]	≈775 K	-	SC
	GeoPile [49]	2023	NAIP, RSD46-WHU, MLRSNet, RESISC45, PatternNet	600 K	-	-
	SatlasPretrain [50]	2023	UCM, BigEarthNet, AID, Million-AID, RESISC45, FMoW, DOTA, iSAID	856 K	512	-
	RSVGD [51]	2023	DIOR	17,402	800	VG
	RefsegRS [52]	2024	SkyScapes	4420	512	RRSIS
	SkyEye-968 K [53]	2024	RSICD, RSITMD, RSIVQA, RSVG [1]	968 K	-	-
	MMRS-1M [54]	2024	AID, RSIVQA, Syndney-Captions [1]	1 M	-	-
	RSSA [55]	2024	DOTA-v2, FAIR1M	44 K	512	-
	FineGrip [56]	2024	MAR20	2649	-	-
	RRSIS-D [57]	2024	-	17,402	800	RRSIS
	RingMo [58]	2022	Gaofen, GeoEye, WorldView, QuickBird [1]	≈2 M	448	-

Table 1. *Cont.*

Method	Dataset	Year	Source	Image	Image Size	Task
	GRAFT [59]	2023	NAIP, Sentinel-2	-	-	-
	SkySense [60]	2024	WorldView, Sentinel-1, Sentinel-2	21.5 M	-	-
	EarthVQA [61]	2024	LoveDA, WorldView	6000	-	VQA
	GeoSense [62]	2024	Sentinel-2, Gaofen, Landsat, QuickBird	≈9 M	224	-
Automatically Annotated Datasets	RS5M [31]	2024	LAION2B-en, LAION400M, LAIONCOCO [1]	5M	-	-
	SkyScript [32]	2024	Google Earth Engine, OpenStreetMap	2.6 M	-	-
	LHRS-Align [63]	2024	Google Earth Engine, OpenStreetMap	1.15 M	-	-
	GeoChat [30]	2024	SAMRS, NWPU-RESISC-45, LRBEN, Floodnet	318 K	-	-
	GeoReasoner [64]	2024	Google Street View, OpenStreetMap	70 K+	-	StreetView
	HqDC-1.4 M [55]	2024	Million-AID, CrowdAI, fMoW, CVUSA, CVACT, LoveDA	≈1.4 M	512	-
	ChatEarthNet [65]	2024	Sentinel-2, WorldCover	163,488	256	-
	VRSBench [66]	2024	DOTA-v2, DIOR	29,614	512	-
	FIT-RS [67]	2024	STAR	1800.8 K	512	-

[1] Only a subset of common datasets are enumerated.

3.1. Manual Datasets

Manually curated datasets, though limited in size, are typically of high quality and are specifically designed for tasks. There are three high-quality datasets made by hand, which are meticulously crafted to address specific challenges within the remote sensing domain.

HallusionBench [46], the first benchmark designed to examine visual illusion and knowledge hallucination of large vision–language models (LVLMs) and analyze the potential failure modes based on each hand-crafted example pair, consists of 455 visual–question control pairs, including 346 different figures and a total of 1129 questions on diverse topics. RSICap's [35] RSIEval comprises five expert-annotated captions with rich and high-quality information; thus, the model achieves satisfactory performance with fine-tuning. CRSVQA [47] employs annotators from different backgrounds, including geography, surveying, and mapping, to write complex questions that should reason the relationship between the objects or a concise question related to the image content. This can effectively prevent the problems of data redundancy and language bias in existing RSVQA datasets [68].

However, manual datasets are time-consuming and expensive to create, requiring domain experts to annotate data. Even with expert annotators, inconsistencies and errors may arise, affecting dataset quality. Additionally, the limited capacity for manual annotation results in smaller datasets, which can hinder model generalization. To ensure consistency, clear annotation guidelines and multiple annotators for cross-validation are essential. Regular audits by experts can maintain quality. Using semi-automated tools can improve efficiency, helping manage large-scale manual annotation projects.

3.2. Combining Datasets

As shown in Table 1, some datasets are built from scratch, whose sources are usually Sentinel-2, Gaofen, Google Earth Engine and so on. But most datasets are created by merging existing remote sensing datasets, which facilitates multi-task model creation by integrating various data types from traditional RS domain-specific datasets, and greatly eases labor compared to manual production and starting from scratch. The most utilized datasets are AID [69], NWPU-RESISC45 [70], UCM [71] for Scene Classification, FAIR1M [72], DIOR [73] for object detection, iSAID [74] for semantic segmentation, and LEVIR-CD [75] for Change Detection; more details about those datasets can be found in Section 4. Additionally, the well-known benchmark dataset Million-AID [76] is frequently included.

Combining datasets from different sources often leads to issues such as inconsistent formats, resolutions, and annotations, requiring extensive preprocessing. Misalignment of annotations can occur, and merging datasets from different contexts (e.g., regions or timeframes) may introduce gaps in data, reducing generalization. When combining datasets, it is crucial to standardize the data formats, resolutions, and annotation styles across the datasets to ensure consistency. Preprocessing steps such as data harmonization are necessary to align different datasets and eliminate inconsistencies. Careful curation ensures balanced representation across different data sources. Filling annotation gaps and validating the combined dataset help improve consistency and robustness.

3.3. Automatically Annotated Datasets

The explosive development of VLMs and LLMs has provided efficient methods for constructing large-scale remote sensing vision–language datasets. This category, which is rapidly becoming the mainstream, leverages the power of VLMs and LLMs to construct large-scale, high-quality datasets.

RS5M [31] employs the general multimodal model BLIP2 [77] to generate image captions and then uses CLIP [22] to select the top five highest-scoring captions. SkyScript [32] matches Google Images with the OpenStreetMap (OSM) database and selects relevant image attribute values from OSM to concatenate them with CLIP [22] as image captions. LHRS-Align [63] uses a similar approach to extract multiple attribute values from OSM and then summarizes them using Vicuna-13B [78] to produce long and fluid captions. GeoChat [30] specifically provide system instructions as prompts that ask Vicuna [78] to generate multi-round question and answer pairs. GeoReasoner [64] obtains image–text data pairs for geolocation from the communities of the two games, utilizes BERT [16] to clean and filter text that lacks geolocation-specific information, and then uses CLIP [22] to align the image and text to construct a highly locatable Street View dataset. HqDC-1.4 M [55] utilizes the "gemini-1.0-pro-vision" [79] to generate descriptions for images from multiple public RS datasets, thereby obtaining a dataset of image–text pairs to serve as the pre-training data for RSVLMs. ChatEarthNet [65] carefully designs prompts that embed semantic information from the land cover maps to make ChatGPT-3.5 and ChatGPT-4v achieve the best results on satellite imagery for the caption generation. VRSBench [66] did a similar job and spent thousands of hours manually verifying the generated annotations. FIT-RS [67] used the efficient TinyLLaVA-3.1B [80] to swiftly generate concise background descriptions for numerous remote sensing imageries (RSIs), then combined them as the prompt to obtain detailed and fluent sentences using GPT-4/GPT-3.5. These are examples which utilize advanced models like BLIP2 [77] and CLIP [22] to generate detailed image–text pairs, supporting complex applications such as disaster monitoring, urban planning, and environmental analysis.

Automatically generated annotations can also introduce errors, as models may misinterpret features or be biased due to limited training data. The lack of human oversight means that errors may persist, resulting in lower-quality annotations, especially for complex or sparse data. To mitigate errors in automatically annotated datasets, models should be carefully evaluated and fine-tuned on high-quality ground-truth data. Implementing hybrid annotation strategies, where human oversight is incorporated into the process, can help identify and correct errors that may go unnoticed in fully automated processes. Data augmentation techniques can be used to enhance the dataset and address limitations in model performance, especially in cases of sparse or ambiguous data. Regular validation of automatic annotations against expert-annotated data can also help ensure the quality of the dataset. Additionally, models should be continually retrained and updated to address evolving challenges in remote sensing and improve annotation accuracy.

3.4. Summary

The existing three types of datasets each have their own advantages and disadvantages, as detailed below:

- Manually annotated datasets generally have higher quality and are most closely aligned with specific problems. But these datasets, typically consisting of only a few hundred or thousand images, are limited in scale compared to combined datasets and VLM datasets, which can reach tens of millions of images. The size of a dataset is crucial for model training, as larger and richer datasets yield more accurate and versatile models. Manually annotated datasets can be used to fine-tune large models, but they struggle to support training large models from scratch.
- Combined datasets. By merging existing datasets, large-scale datasets can be constructed at a low cost. This type of data can easily reach millions of entries. However, datasets created in this manner typically have lower annotation quality compared to manually annotated datasets and may not fully align with specific tasks. Nonetheless, they can be utilized for model pre-training.
- Automatically annotated datasets. Mainstream VLM dataset production methods leverage state-of-the-art LLMs and VLMs to generate high-quality captions or annotations for large-scale remote sensing images. This approach produces flexible image–text pairs, including long sentences, comments, and local descriptions, without requiring human supervision, thus meeting the needs of various tasks. Researchers may need to craft prompts and commands to guide LLMs and VLMs in generating the desired texts.

Overall, automatically annotated datasets leverage the advantages of pre-trained large models, resulting in datasets that are both quantitatively and qualitatively substantial. As a result, this type of dataset has become quite mainstream in the field of remote sensing.

4. Capabilities

In various tasks within the field of remote sensing, VLMs have demonstrated powerful capabilities. These capabilities can be broadly classified into two categories: pure visual tasks and vision–language tasks. Pure visual tasks focus on analyzing remote sensing images to extract meaningful information about the Earth's surface. Vision–language tasks, which combine natural language processing with visual data analysis, have opened new avenues for remote sensing applications. The zero-shot and few-shot tasks are SC, OD, SS and IR, which are described in more detail below. Figure 8 illustrates each capability with a simple example, particularly highlighting two emerging tasks along with their source tasks.

4.1. Pure Visual

- Scene Classification (SC): RS Scene Classification entails the classification of satellite or aerial images into distinct land-cover or land-use categories with the objective of deriving valuable insights about the Earth's surface. It offers valuable information regarding the spatial distribution and temporal changes in various land-cover categories, such as forests, agricultural fields, water bodies, urban areas, and natural landscapes. The three most commonly used datasets in this field are AID [69], NWPU-RESISC45 [70] and UCM [71], which contain 10,000, 31,500, and 2100 images, and 30, 45, and 21 categories, respectively.



Figure 8. Examples of tasks handled by VLMs in remote sensing field.

- Object Detection (OD): RSOD aims to determine whether or not objects of interest exist in a given RSI and return the category and position of each predicted object. In this domain, the most widely used datasets are DOTA [81] and DIOR [73]. DOTA [81] contains 2806 aerial images which are annotated by experts in aerial image interpretation, with respect to 15 common object categories. DIOR [73] contains 23,463 images and 192,472 instances, covering 20 object classes, with much more detail than DOTA [81].
- Semantic Segmentation (SS): Semantic segmentation is a computer vision task in which the goal is to categorize each pixel in an image into a class or object. For RS images, it plays an irreplaceable role in disaster assessment, crop yield estimation, and land change monitoring. The most widely used datasets are ISPRS Vaihingen and Potsdam, which contain 33 true orthophoto (TOP) images and 38 image tiles with a spatial resolution of 5 cm, respectively. Another widely used dataset iSAID [74] contains 2806 aerial images, which were mainly collected from Google Earth.
- Change Detection (CD): Change Detection in remote sensing refers to the process of identifying differences in the structure of objects and phenomena on Earth surface by analyzing two or more images taken at different times. It plays an important role in urban expansion, deforestation, and damage assessment. LEVIR-CD [75], AICD [82] and the Google Data Set focus on changes in buildings. Other types of changes such as roads and groundwork are commonly included in CD datasets.
- Object Counting (OC): RSOC aims to automatically estimate the number of object instances in an RSI. It plays an important role in many areas, such as urban planning, disaster response and assessment, environment control, and mapping. RemoteCount [29] is a manual dataset for object counting in remote sensing imagery

consisting of 947 image–text pairs and 13 categories, which are mainly selected from the validation set of the DOTA dataset.

4.2. Combine with Natural Language

- Image Retrieval (IR): IR from RS aims to retrieve RS images of interest from massive RS image repositories. With the launch of more and more Earth observation satellites and the emergence of large-scale remote sensing datasets, RS image retrieval can prepare auxiliary data or narrow the search space for a large number of RS image processing tasks.
- Visual Question Answering (VQA): VQA is a task that seeks to provide answers to free-form and open-ended questions about a given image. As the questions can be unconstrained, a VQA model applied to remote sensing data could serve as a generic solution to classical problems but also very specific tasks involving relations between objects of different nature [83]. VQA was first proposed by Anto et al. [84] in 2015, and then applied to the RS domain by RSVQA [68] in 2020. This work contributed an RSVQA framework and two VQA datasets, RSVQA-LR and RSVQA-HR, which are widely utilized in further works.
- Image Captioning (IC): IC aims to generate natural language descriptions that summarize the content of an image. It requires representing the semantic relationships among objects and generating an exact and descriptive sentence, so it is more complex and challenging than image detection, classification, and segmentation tasks. In this domain, the commonly used datasets are Syndney-Captions [85], RSICD [86], NWPU-Captions [87], and UCM-Captions [85]. Recently, CapERA [88] provided UAV video with diverse textual descriptions to advance visual–language-understanding tasks.
- Visual Grounding (VG): RSVG aims to localize specific objects in remote sensing images with the guidance of natural language. It was first introduced in [89] in 2022. In 2023, Yang [51] et al. not only built the new large-scale benchmark of RSVG based on detection in the DIOR dataset, termed RSVGD, but designed a novel transformer-based MGVLF module to address the problems of scale variation and cluttered background in RS images.
- Remote Sensing Image Change Captioning (RSICC): A new task aiming to generate human-like language descriptions for the land-cover changes in multitemporal RS images. It is a combination of IC and CD tasks, offering important application prospects in damage assessment, environmental protection, and land planning. Chenyang et al. [90] first introduced the CC task into the RS domain in 2022 and proposed a large-scale dataset LEVIR-CC, containing 10,077 pairs of bitemporal RS images and 50,385 sentences describing the differences between the images. In 2023, they proposed a pure Transformer-based model [91] to further improve the performance of the RSICC task.
- Referring Remote Sensing Image Segmentation (RRSIS): RRSIS provides a pixel-level mask of desired objects based on the content of given remote sensing images and natural language expressions. It was first proposed by Zhenghang et al. [52] in 2024, who created a new dataset, called RefSegRS, for this task. In the same year, Sihan et al. [57] curated an expansive dataset comprising 17,402 image–caption–mask triplets, called RRSIS-D, and proposed a series of models to meet the ubiquitous rotational phenomena in RRSIS.

The invention of VLMs has greatly changed the remote sensing field, which mainly focuses on geospatial analysis. With the development and usage of various sensors, conventional vision models cannot handle such complex data types, while VLMs could enhance the accuracy and efficiency of remote sensing analysis. Specifically, the combination of

visual and linguistic data within VLMs allows more subtle and sophisticated interpretations of complex remote sensing datasets. By utilizing their multimodal capabilities, VLMs have shown significant progress across various tasks within remote sensing, including object detection, geophysical classification, and scene understanding. Meanwhile, they provide more accurate and smart solutions for practical applications such as disaster monitoring, urban planning, and agricultural management.

5. Recent Advances

Currently, there are two main types of models that integrate vision and language: one is based on contrastive learning structures, such as the CLIP series, and the other is based on large language models that fuse visual features, like LLaVA, as shown in Figure 9. Accordingly, recent improvements can also be categorized into two major types: enhancements within the contrastive learning framework and improvements within the conversational framework.

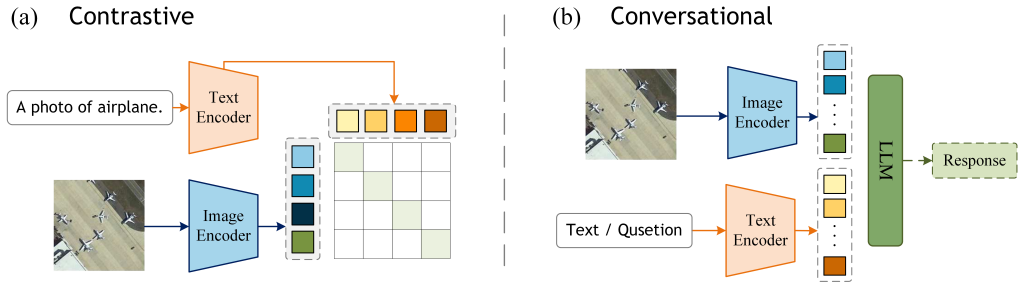

Figure 9. Fundamental architecture of contrastive and conversational Vision–Language Models.

5.1. Advancements in Contrastive VLMs

Contrastive VLMs primarily consist of an image encoder and a text encoder, with the core challenge being the alignment of features extracted by both encoders. As shown in Table 2, the image encoder, utilizing a CNN or ViT, encodes images into feature vectors. Concurrently, the text encoder processes textual descriptions into corresponding feature vectors. The model is trained to maximize the similarity between positive image–text pairs while minimizing the similarity between negative pairs. CLIP is a pioneering model in vision–language pre-training, employing InfoNCE loss to align text and image features in a shared space. Currently, in the field of remote sensing, contrastive VLMs focus on transferring the image–text alignment capabilities established in RGB images and text to remote sensing images and text, which are then utilized to address various downstream tasks, such as Change Detection and remote sensing Image Captioning.

Table 2. Architecture of contrastive VLMs. SC = Scene Classification; VG = Visual Grounding; IC = Image Captioning; CD = Change Detection; IR = Image Retrieval. For each model, we indicate the LLM used in its best configuration as shown in the original paper.

Model	Year	Image Encoder	Text Encoder	Capability
RemoteCLIP [29]	2024	ViT-B (CLIP)	Transformer	SC, IC, IR, OC
CRSR [92]	2024	ViT-B (CLIP)	-	IC
ProGEO [93]	2024	ViT-B (CLIP)	BERT	VG
GRAFT [59]	2023	ViT-B (CLIP)	-	SC, VQA, IR
GeoRSCLIP [31]	2024	ViT-H (CLIP)	-	SC, IR
ChangeCLIP [94]	2024	CLIP	-	CD
APPLeNet [95]	2023	CLIP	-	SC, IC
MGVLF [51]	2023	ResNet-50	BERT	VG

To address the issue of requiring additional annotated data and fine-tuning when building foundational models using self-supervised learning and masked image modeling

methods in the remote sensing domain, Liu et al. [29] proposed the first vision–language foundational model for remote sensing, RemoteCLIP. This model is designed to learn robust visual features with rich semantics and aligned text embeddings for seamless downstream applications. They created a combined image–caption dataset and fine-tuned the CLIP model on this newly constructed dataset through continual learning, thereby tackling the scarcity of pre-trained data. RemoteCLIP exhibits powerful multi-task capabilities and can be applied to a range of downstream tasks, including zero-shot image classification, linear probing, k-NN classification, few-shot classification, image-text retrieval, and object counting. Similar to [29], GeoRSCLIP builds a remote sensing dataset consisting of 5 million image–text pairs by filtering existing image–text pair datasets and generating captions for image-only datasets. It then fine-tunes CLIP on this proposed dataset [31]. CRSR [92] incorporates a Transformer Mapper network, utilizing an attention mechanism and learnable semantic predictive queries to provide a comprehensive representation of image features. This approach effectively analyzes and understands more semantic details present in RSI. ProGEO proposes adding more descriptions to a geographic image-based pre-trained CLIP, using these text descriptions to enhance the image encoder by aligning features from the image encoder with the features of the descriptions [93]. This approach enables the image encoder to generate features with more details, leading to improvements in visual geo-localization. CRAFT proposes training a satellite image model to align its embedding with the CLIP embedding of co-located internet images, which sidestep the need for textual annotations for training remote sensing VLMs [59]. ChangeCLIP [94] is the first to apply a multimodal vision–language scheme to Remote Sensing Change Detection tasks. It introduces a Differentiable Feature Calculation (DFC) layer, enabling the network to focus on learning the changed features of bitemporal remote sensing images rather than semantic categories. APPLeNet designs an injection module to fuse image features from different levels and interact with text to generate better visual tokens [95].

In addition, S-CLIP [96] and RS-CLIP [97] enhance the performance of CLIP in remote sensing by improving the pseudo-label strategy. The former divides pseudo-labels into title-level and keyword-level categories, combining their advantages to ensure the model's generalization ability. RS-CLIP automatically generates pseudo-labels from unlabeled data and employs a curriculum learning strategy to gradually select more pseudo-samples for model fine-tuning.

5.2. Advancements in Conversational VLMs

Currently, conversational VLMs are particularly noteworthy for their ability to address complex remote sensing tasks by transforming intricate visual features into a format that LLMs can comprehend. By integrating advanced image encoders that adeptly capture subtle variations in remote sensing imagery with text encoders optimized for efficiently processing language instructions, these models demonstrate remarkable performance, particularly in few-shot learning scenarios. The alignment layer is pivotal in establishing connections between remote sensing imagery and LLMs. Techniques such as utilizing MLPs for feature projection and incorporating learnable query embeddings for task-specific image feature extraction have proven effective in elevating the overall performance of VLMs.

Leveraging the existing API of pre-trained LLMs, conversational VLMs extract features from image data using a visual encoder. These features are then transformed into a format that LLMs can interpret, enabling the completion of various vision–language tasks. To enhance task performance, researchers meticulously design each component, which is described in detail below. Figure 10 clearly shows the model details of conversational VLMs.

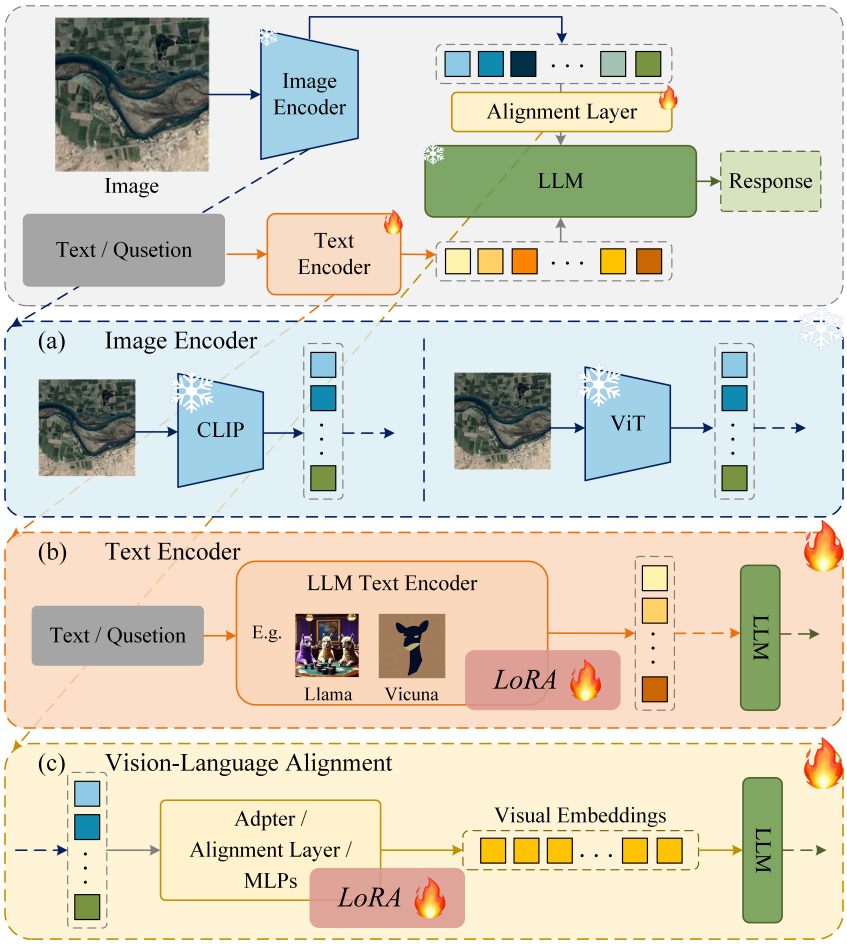

Figure 10. Enhancement techniques. Fire denotes fine-tuning, while snowflake denotes frozen.

Image Encoder: Table 3 presents the visual encoders commonly used in current VLM implementations. Most studies utilize the standard CLIP image encoder [22], but RS-GPT [35] and SkyEyeGPT [53] have explored other variants, such as EVA-CLIP encoder [98], which is trained with improved training techniques.

Additionally, RS-LLaVa [99] adapts the LLaVA framework for the remote sensing domain by using a domain-specific image encoder and training it on specialized RS datasets. Through instruction tuning and precise visual–language alignment, RS-LLaVa gains a deeper understanding of geospatial semantics, allowing it to provide contextually relevant, detail-oriented responses tailored to the complexities of remote sensing imagery. EarthGPT [54] employs a Vision Transformer as its image encoder to extract multi-layered features that capture subtle differences in RS images. It then integrates a CNN backbone to incorporate multi-scale local details into the visual representation, delivering more comprehensive analytical capabilities. GeoChat [30] builds on the CLIP-ViT(L-14) backbone, interpolating positional encodings at higher resolutions (512 × 512) to better discern fine-grained details. This approach enhances its capacity to identify small objects and subtle characteristics in remote sensing imagery.

Table 3. Architecture of conversational VLMs. ∗: frozen; ▲: fine-tuning; SC = Scene Classification; VQA = Visual Question Answering; VG = Visual Grounding; IC = Image Captioning; OD = object detection; IR = Image Retrieval; RSICC = Remote Sensing Image Change Captioning. For each model, we indicate the LLM used in its best configuration as shown in the original paper.

Model	Year	Image Encoder	Text Encoder	Capability
RS-LLaVA [99]	2024	ViT-L(CLIP) *	Vicuna-v1.5-13B *	IC, VQA
H2RSVLM [55]	2024	ViT-L (CLIP) *	Vicuna-v1.5-7B *	SC, VQA, VG
SkySenseGPT [67]	2024	ViT-L (CLIP) *	Vicuna-v1.5-7B ▲	SC, IC, VQA, OD
GeoChat [30]	2024	ViT-L (CLIP) *	Vicuna-v1.5-7B ▲	SC, VQA, VG Region Caption
RSGPT [35]	2023	ViT-G (EVA) *	Vicuna-v1.5-7B *	IC, VQA
SkyEyeGPT [53]	2024	ViT-G (EVA) *	LLaMA2-7B *	IC, VQA, VG
RS-CapRet [100]	2024	ViT-L (CLIP) ▲	LLaMA2-7B *	IC, IR
LHRS-Bot [63]	2024	ViT-L (CLIP) *	LLaMA2-7B *	SC, VQA, VG
EarthGPT [54]	2024	ViT-L (DINOv2) * ConvNeXt-L (CLIP) *	LLaMA2-7B ▲	SC, VQA, IC, VG, OD
Liu et al. [91]	2023	ViT-B(CLIP) *	GPT-2 *	RSICC

Text Encoder: Training large language models (LLMs) from scratch is highly costly. Therefore, conversational Vision–Language Models (VLMs) utilize pre-trained LLMs as text encoders. These LLMs process visual elements extracted by the visual encoder and converted by middleware, along with the researcher's questions or language instructions, to generate the desired results. Currently, the academic community is focusing on the LLaMA [26] and Vicuna [78] series, which serve as text encoders for most conversational LLMs. Table 3 provides a summary of these models. Common approaches include directly using LLMs or efficiently fine-tuning them with Low-Rank Adaptation (LoRA) [101].

Liu et al. [91] designed a complex image classifier and proposed a multi-prompt learning strategy to generate multiple learnable prompt embeddings as prefixes for LLMs to enhance their utilization for the RSICC task. In addition, RS-LLaVA [99] applies the LLaVA framework to remote sensing data, incorporating a Vicuna-v1.5-13B LLM along with domain-specific image encoders and specialized training corpora, enabling the model to better capture geospatial semantics. Similarly, RSGPT [35] utilizes a Vicuna-v1.5-7B LLM, showing that while large language models generally yield improved performance in remote sensing tasks, considerations regarding training time and resource requirements remain central to practical deployments.

Vision–Language Alignment: In the architecture of conversational Vision–Language Models, the middle connection layer plays a crucial role. It acts as a bridge between remote sensing images and large language models, projecting image information into a space that LLMs can comprehend, enabling them to understand complex remote sensing tasks. A common approach to bridge the modal gap is to use Multi-Layer Perceptrons (MLPs), which employ a single linear layer [53,54] or double layers [55,67,99] to project visual tokens and align feature dimensions with word embeddings. Additionally, some works focus on designing connectors with intricate structures, offering valuable insights for future research. SkyEyeGPT [53] is introduced as a unified multimodal large language model specifically designed for remote sensing vision–language understanding, with the objective of exploring the application of VLMs in the remote sensing field. The authors carefully create a remote sensing multimodal instruction-following dataset, which includes both single-task and multi-task dialogue instructions. After manual verification, the dataset consists of a high-quality remote sensing instruction-following corpus with a total of 968 K samples. SkyEyeGPT projects remote sensing visual features into the language domain via an alignment layer, and these features, along with task-specific instructions, are fed into a remote sensing decoder based on LLM to predict answers for open-ended remote sensing tasks.

SkyEyeGPT employs a two-stage fine-tuning approach to enhance instruction-following and multi-turn dialogue capabilities at different granularities. Notably, it achieves excellent performance across a variety of tasks without requiring additional encoding modules.

One reason current general VLMs perform poorly in the remote sensing field is that the visual encoder of LLMs usually uses CLIP, but CLIP's ViT lacks spatial awareness, affecting VLM's spatial understanding ability [102]. To address this issue, H2RSVLM [55] and SkySenseGPT [67] enhance the visual encoder's perceptual ability at the source by creating higher-quality, large-scale, and detailed image–text pairs in remote sensing datasets. EarthGPT [54] expands the input images to include SAR and infrared images. Furthermore, some research endeavors opt for designing elaborate vision–language connection structures to enhance the multimodal perception capabilities of VLM. GeoChat [30] constructs bounding boxes for text representations to express their spatial coordinates, enabling it to perform object grounding or accept region inputs, thus having reasoning capabilities for remote sensing images. RSGPT [35], under the guidance of InstructBLIP [103], inserts an annotation-aware query transformer (Q-Former) to enhance the alignment of visual and textual features. Additionally, the annotation-aware mechanism interacts with query embeddings through self-attention layers, allowing RSGPT to extract comprehensive visual features crucial for RS vision–language tasks. LHRS-Bot [63] introduces a set of learnable queries for each level of image features. These queries aim to summarize the semantic information of each level of visual features through a series of stacked cross-attention and MLP layers.

5.3. Other Models

In addition to the aforementioned approaches, as shown in Table 4, several other advancements have been made to address domain-specific tasks in remote sensing.

Txt2Img [104] employs modern Hopfield layers for hierarchical prototype learning of text and image embeddings, enabling the generated remote sensing images to better preserve the shape information of ground objects while providing more detailed texture and boundary information. CPSeg [105] employs a novel "Chain-of-Thought" approach using language prompting to enhance semantic segmentation in remote sensing imagery, achieving superior performance, particularly in flood scenarios, through improved context-aware understanding and integration of textual descriptions. SHRNet [106] utilizes a hash-based architecture to extract multi-scale image features, while QRN improves alignment efficiency and enhances the visual–spatial reasoning ability of remote sensing Visual Question Answering systems. MGeo [107], a multimodal geographic language model, introduces the geographic environment (GC) as a new modality to enhance query–POI (Point of Interest) matching, outperforming baselines even without user geolocation, while its GC modeling holds potential for further improvements with extensions like POI images. GeoCLIP [108], the first CLIP-inspired image-to-GPS retrieval approach, introduces a novel Location Encoder that models the Earth as a continuous function using positional encoding and hierarchical representations. By aligning image features with GPS locations, GeoCLIP achieves competitive geo-localization performance, even with limited training data, and enables text-based geo-localization via its CLIP backbone. SpectralGPT [109], a spectral RS foundational model based on MAE [110], leverages 3D token generation and multi-target reconstruction to capture spatial–spectral patterns. Trained on one million spectral images with 600 M+ parameters, it achieves 99.21% for single-label Scene Classification on EuroSAT [111] and also demonstrates strong performance in semantic segmentation and Change Detection, advancing spectral RS big data applications. TEMO [112] introduces a text-modal knowledge extractor and a crossmodal assembly module to fuse text and visual features, reducing classification confusion in novel classes and enhancing few-shot

object detection in remote sensing imagery. By leveraging a mask strategy, separation loss, and improved word embeddings to capture long-term dependencies, TEMO achieves 42.8%, 75.1% and 40.8% across DIOR [73], NWPU [113], and FAIR1M [72] in 10-shot scenarios, respectively.

Table 4. Summary of other VLMs in Section 5.3.

Model	Year	Enhancement Technique
Txt2Img [104]	2023	Employs modern Hopfield layers for hierarchical prototype learning of text and image embeddings.
CPSeg [105]	2024	Utilizes a "chain of thought" process that leverages textual information associated with the image.
SHRNet [106]	2023	Utilizes a hash-based architecture to extract multi-scale image features.
MGeo [107]	2023	Introduces the geographic environment (GC) extracting multimodal correlations for accurate query–POI matching.
GeoCLIP [108]	2023	Via an image-to-GPS retrieval approach by explicitly aligning image features with corresponding GPS locations.
SpectralGPT [109]	2024	Provides a foundational model for the spectral data based on the MAE [110] architecture.
TEMO [112]	2023	Captures long-term dependencies from description sentences of varying lengths, generating text-modal features for each category.

5.4. Performance Comparison

In this section, we conduct a comparative analysis of recent Visual–Language Models, evaluating their performance on five widely recognized datasets and common visual–language tasks including Visual Question Answering and Image Captioning. VLMs have gained significant attention in the remote sensing domain over the past two years, with researchers exploring their applications in various tasks. However, models tailored for specialized domains, such as SpectralGPT [109] which is designed for spectral remote sensing, face challenges in cross-model comparisons due to unique datasets and tasks. Despite efforts to build versatile foundational models capable of handling diverse downstream tasks, the remote sensing field still lacks a unified benchmark for comprehensive performance evaluation. To address this gap, we focus on shared datasets and tasks to provide a consistent and meaningful comparison of VLM performance.

Table 5 summarizes the evaluation results for five prevalent datasets (e.g., AID [69], RSVGD [51]), highlighting the best-performing VLMs. Due to differing implementations of VLM pre-training methods, it is essential to identify standout models. Among the contrastive approaches, RemoteCLIP [29] emerges as a top performer, particularly excelling in WHU-RS19 [114]. Notably, GeoRSCLIP [31] demonstrates strong few-shot capabilities, achieving competitive results on NWPU-RESISC45 [70]. In comparison, conversational models, such as SkySenseGPT [67] and SkyEyeGPT [53], consistently achieve superior performance in multiple datasets, with SkySenseGPT leading in AID [69] and WHU-RS19 [114]. Unlike contrastive models, conversational methods also support Visual Grounding tasks, highlighting their broader applicability.

Table 5. Performance of VLM pre-training methods on common remote sensing datasets; bold indicates the best performance. ∗: few-shot prediction.

Category	Model	AID [69]	RSVGD [51]	NWPU-RESISC45 [70]	WHU-RS19 [114]	EuroSAT [111]
Contrastive	RemoteCLIP [29]	91.30	-	79.84	96.12	60.21
	GeoRSCLIP [31]	76.33 *	-	73.83 *	-	**67.47**
	APPLeNet [95]	-	-	72.73 *	-	-
	GRAFT [59]	-	-	-	-	63.76
Conversational	SkySenseGPT [67]	**92.25**	-	-	**97.02**	-
	LHRS-Bot [63]	91.26	88.10	-	93.17	51.40
	H2RSVLM [55]	89.33	48.04 *	-	97.00	-
	GeoChat [30]	72.03 *	-	-	-	-
	SkyEyeGPT [53]	-	**88.59**	-	-	-
	EarthGPT [54]	-	81.54	**93.84**	-	-
Others	RSVG [51]	-	78.41	-	-	-

In Tables 6 and 7, we delve deeper into two frequently encountered visual–language tasks: Visual Question Answering (VQA) and Image Classification (IC). The results clearly illustrate that the conversational model, SkySenseGPT [67], maintains exceptional performance across both tasks. Notably, none of the contrastive VLMs proved effective for the VQA task, further underscoring the advantages of conversational methods. Additionally, in the IC task, the prevalence of conversational VLMs exceeds that of their contrastive counterparts.

Table 6. Performance of VLM pre-training methods on the Visual Question Answering task; bold indicates the best performance. ∗: few-shot prediction.

Model	RSVQA-LR [68]	RSVQA-HR [68]	RSIVQA [115]
SkySenseGPT [67]	**92.69**	76.64 *	-
RSGPT [35]	92.29	92.00	-
GeoChat [30]	90.70	72.30 *	-
LHRS-Bot [63]	89.19	**92.55**	-
H2RSVLM [55]	89.12	74.35 *	-
RS-LLaVA [99]	88.56	-	-
SkyEyeGPT [53]	88.23	86.87	-
EarthGPT [54]	-	72.06 *	-
SHRNet [106] [1]	85.85	85.39	**84.46**
MQVQA [47] [1]	-	-	82.18

[1] Only these are from the other part; the rest are conversational.

In conclusion, from Tables 2 and 3, it is evident that a key architectural difference between conversational VLMs and contrastive VLMs lies in the use of pre-trained LLMs (e.g., Vicuna [78], LLaMA [20]) in conversational VLMs. These LLMs process visual elements extracted by visual encoders and transformed by middleware, along with user queries or textual instructions. This enables conversational VLMs to handle more diverse and complex visual–language tasks, such as VQA and IC. As shown in Tables 6 and 7, conversational VLMs consistently outperform other types, further unlocking the potential of language to enhance reasoning in complex scenarios. Consequently, we advocate for adopting conversational model architectures as the preferred choice for future advancements in visual–language modeling.

Table 7. Performance of VLM pre-training methods on image caption task; bold indicates the best performance.

Dataset	Model	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEH	CIDEr
UCM-Captions [71]	CRSR [1] [92]	90.60	85.61	81.22	76.81	**49.56**	**85.86**	**380.69**
	SkyEyeGPT [53]	**90.71**	**85.69**	**81.56**	**78.41**	46.24	79.49	236.75
	RS-LLaVA [99]	90.00	84.88	80.30	76.03	49.21	85.78	355.61
	RSGPT [35]	86.12	79.14	72.31	65.74	42.21	78.34	333.23
	RS-CapRet [100]	84.30	77.90	72.20	67.00	47.20	81.70	354.80
Syndney-Captions [85]	CRSR [1] [92]	79.94	74.4	69.87	66.02	41.50	74.88	**289.00**
	SkyEyeGPT [53]	**91.85**	**85.64**	**80.88**	**77.40**	**46.62**	**77.74**	181.06
	RSGPT [35]	82.26	75.28	68.57	62.23	41.37	74.77	273.08
	RS-CapRet [100]	78.70	70.00	62.80	56.40	38.80	70.70	239.20
RSICD [86]	CRSR [1] [92]	81.92	71.71	63.07	55.74	**40.15**	**71.34**	**306.87**
	SkyEyeGPT [53]	**87.33**	**77.70**	**68.90**	**61.99**	36.23	63.54	89.37
	RSGPT [35]	70.32	54.23	44.02	36.83	30.1	53.34	102.94
	RS-CapRet [100]	72.00	59.90	50.60	43.30	37.00	63.30	250.20
NWPU-Captions [87]	RS-CapRet [100]	**87.10**	**78.70**	**71.70**	**65.60**	43.60	77.60	**192.90**
	EarthGPT [54]	**87.10**	**78.70**	71.60	65.50	**44.50**	**78.20**	192.60

[1] Only CRSR is contrastive; the rest are conversational.

6. Conclusions and Future Work

This review provides a systematic overview of Vision–Language Models (VLM) in the field of remote sensing. It introduces milestone works in the development of VLM itself, discusses the current research status of VLM in remote sensing, outlines the tasks involved, and offers a brief comparison of different VLM methods in various remote sensing applications. Overall, VLM has once again facilitated advancements in data processing technologies within the field of remote sensing. Current work has clearly demonstrated the potential of VLM applications in this area. However, there are still some shortcomings in current VLM technologies and their applications in remote sensing. Future work could consider the following directions:

- **Addressing regression problems.** Currently, the mainstream approach in VLM typically involves using a tokenizer to vectorize text, aligning the visual information with word vectors, and then feeding these into LLMs. However, this method fails to capture numerical relationships effectively, which leads to problems in regression tasks. For instance, a number like '100' may be tokenized as '1', '00', or '100', causing loss of precision and affecting regression accuracy. Future work should focus on enhancing VLMs for regression tasks by addressing this limitation. One promising direction is to develop specialized tokenizers for numerical values, which would ensure more accurate representation of numbers in text. Additionally, exploring the integration of task-specific regression heads into the existing VLM framework can help the model achieve a more efficient inference process. For example, REO-VLM [116] designs a regression head with a four-layer MLP-mixer-like structure that integrates knowledge embeddings derived from LLM hidden tokens with complex visual tokens. This innovative approach enables VLMs to effectively handle regression tasks such as Above-Ground Biomass (AGB) [117,118] estimation in Earth observation (EO) applications. Task-specific regression heads enable the model to directly output data in formats that meet the specific requirements of the task, rather than being limited to regression on text tokens. Moreover, these heads can be dynamically adjusted based on the needs of different tasks, making the regression output more accurate and efficient. Introducing the Mixture of Experts (MOE) architecture to improve the adaptability and accuracy of regression tasks is also an excellent choice. When deployed in VLMs, it can use a gating mechanism to dynamically match the appropriate expert modules

for different tasks, enabling the model to effectively handle multi-task learning with sufficient capability.

- **Aligning with the structural characteristics of remote sensing images.** Current VLMs in remote sensing largely adopt frameworks developed for conventional computer vision tasks and rely on models pre-trained exclusively on RGB imagery, thereby failing to fully account for the distinctive structural and spectral characteristics of remote sensing data. This limitation is particularly evident for modalities such as SAR and HSI, where crucial features remain poorly captured by general-purpose extractors. As a result, while existing VLMs may perform reasonably well on RGB data, they struggle with SAR and HSI inputs and often depend on superficial fine-tuning with limited datasets. To address this gap, future research should emphasize both the design of specialized feature extractors tailored to the complexities of multispectral and SAR data, and the development of algorithms capable of handling multidimensional inputs. REO-VLM [116] makes effective attempts by introducing spectral recombination and pseudo-RGB strategies. It splits the high-dimensional MS image into multiple three-channel images and generates a third channel for the SAR images, which typically contain two polarization channels, thus creating a pseudo-RGB representation compatible with the pre-trained visual encoder. Expanding and diversifying remote sensing datasets to include large-scale image–text pairs from multiple modalities (e.g., RGB, SAR, HSI, LiDAR) will be essential for training VLMs that can achieve robust visual–text alignment across all data types. By integrating richer and more comprehensive feature representations, as well as exploring data fusion strategies, it will be possible to enhance VLM performance in a wide range of complex remote sensing applications, including land-use classification, urban monitoring, and disaster response.
- **Multimodal output.** Early methods, such as classification models, were constrained by their output format and thus limited to a single task. Although most current VLMs, which produce text outputs, have broadened their scope and can effectively handle numerous tasks, they still fall short for dense prediction tasks like segmentation and Change Detection. To overcome this limitation, future research should focus on enabling VLMs to produce truly multimodal outputs—such as images, videos, and even 3D data. Recent studies have begun exploring the enhancement of traditional proprietary models by adding a text branch as an auxiliary component, thereby fostering richer multimodal feature interaction and improving image feature extraction and interpretation while preserving existing output heads. Building on this idea, another promising direction is to use a VLM as an agent that interfaces with multiple specialized model output heads and activates the most suitable one based on the task identified, thus effectively combining the reasoning power of LLMs with the expertise of dedicated models. In addition, a recent insightful approach is to use next-token prediction to achieve multimodal output. For example, Emu3 [119] adopts a completely next-token prediction-based architecture, avoiding the complex diffusion models or combinatorial methods used in the past. It innovatively uses visual tokens as part of the output to generate images and videos, and outperforms well-known open-source models.
- **Multitemporal Regression and Trend Inference.** Current VLM research in the remote sensing field often focuses solely on static images, overlooking the potential of multitemporal data. Since remote sensing data can capture information that changes over time, future research should explore how to extend VLMs to handle multitemporal regression tasks. This involves not only spatial feature recognition but also the ability to infer temporal trends and dynamics. A key research direction is to develop methods that utilize historical remote sensing data to analyze and predict long-term environ-

mental or ecological changes. This is particularly important for applications such as climate change monitoring, land-use change prediction, and ecosystem management, where temporal patterns play a crucial role. Remote sensing images with temporal continuity can be encoded as sequences, providing multitemporal data as input for training VLMs. By adding multitemporal regression capabilities to VLMs, the model could provide more context-aware predictions, improving the accuracy of long-term environmental assessments.

Author Contributions: Conceptualization, H.Z.; investigation, L.T. and H.J.; data curation, L.T.; writing—original draft preparation, L.T., H.Z., H.J., Y.L. and X.X.; writing—review and editing, L.T., H.Z., D.Y. and G.W.; visualization, L.T., H.Z. and H.J.; supervision, H.Z.; project administration, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 62401471; in part by 2024 Gusu Innovation and Entrepreneurship Leading Talents Program (Young Innovative Leading Talents) under Grant ZXL2024333.

Data Availability Statement: No new data were created or analyzed in this study.

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

References

1. Yang, K. Progress, Challenge and Prospect for Remote Sensing Monitoring of Flood and Drought Disasters in China. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 4280–4283. [CrossRef]
2. Shahtahmassebi, A.R.; Li, C.; Fan, Y.; Wu, Y.; Lin, Y.; Gan, M.; Wang, K.; Malik, A.; Blackburn, G.A. Remote sensing of urban green spaces: A review. *Urban For. Urban Green.* **2021**, *57*, 126946. [CrossRef]
3. 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.
4. Meng, L.; Yan, X.H. Remote Sensing for Subsurface and Deeper Oceans: An overview and a future outlook. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 72–92. [CrossRef]
5. Parra, L. Remote sensing and GIS in environmental monitoring. *Appl. Sci.* **2022**, *12*, 8045. [CrossRef]
6. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. *arXiv* **2015**, arXiv:1412.0767.
7. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
9. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for small object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
10. Zhang, J.; Li, S.; Long, T. Enhanced Target Detection: Fusion of SPD and CoTC3 Within YOLOv5 Framework. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 3000114. [CrossRef]
11. Liu, H.; He, J.; Li, Y.; Bi, Y. Multilevel Prototype Alignment for Cross-Domain Few-Shot Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4400115. [CrossRef]
12. Xie, Y.; Liu, S.; Chen, H.; Cao, S.; Zhang, H.; Feng, D.; Wan, Q.; Zhu, J.; Zhu, Q. Localization, Balance, and Affinity: A Stronger Multifaceted Collaborative Salient Object Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 4700117. [CrossRef]
13. Zhou, H.; Wang, Y.; Liu, W.; Tao, D.; Ma, W.; Liu, B. MSC-GAN: A Multistream Complementary Generative Adversarial Network with Grouping Learning for Multitemporal Cloud Removal. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5400117. [CrossRef]
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762.
15. OpenAI. ChatGPT. 2023. Available online: https://chatgpt.com/ (accessed on 5 October 2024).
16. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NaacL-HLT, Minneapolis, MN, USA, 2–7 June 2019.

17. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Lewis, M.; Zettlemoyer, L. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 3531–3541.
18. Williams, A.; Nangia, N.; Bowman, S.R. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv* **2018**, arXiv:1704.05426.
19. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv* **2016**, arXiv:1606.05250.
20. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
21. OpenAI. GPT-3: Language Models Are Few-Shot Learners. 2020. Available online: https://openai.com/research/gpt-3 (accessed on 5 October 2024).
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 International Conference on Machine Learning, Online, 18–24 July 2021.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Li, L.H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.N.; et al. 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.
26. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36* , 34892–34916.
27. OpenAI. GPT-4. 2023. Available online: https://openai.com/research/gpt-4 (accessed on 5 October 2024).
28. Al Rahhal, M.M.; Bazi, Y.; Elgibreen, H.; Zuair, M. Vision-Language Models for Zero-Shot Classification of Remote Sensing Images. *Appl. Sci.* **2023**, *13*, 12462. [CrossRef]
29. Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; Zhou, J. Remoteclip: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5622216. [CrossRef]
30. Kuckreja, K.; Danish, M.S.; Naseer, M.; Das, A.; Khan, S.; 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, Seattle, WA, USA, 16–22 June 2024; pp. 27831–27840.
31. 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* **2024**, arXiv:2306.11300. [CrossRef]
32. 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; Volume 38, pp. 5805–5813.
33. Wang, L.; Dong, S.; Chen, Y.; Meng, X.; Fang, S. MetaSegNet: Metadata-collaborative Vision-Language Representation Learning for Semantic Segmentation of Remote Sensing Images. *arXiv* **2023**, arXiv:2312.12735. [CrossRef]
34. Deng, P.; Zhou, W.; Wu, H. ChangeChat: An Interactive Model for Remote Sensing Change Analysis via Multimodal Instruction Tuning. *arXiv* **2024**, arXiv:2409.08582.
35. Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Li, X. Rsgpt: A remote sensing vision language model and benchmark. *arXiv* **2023**, arXiv:2307.15266.
36. Wang, J.; Sun, H.; Tang, T.; Sun, Y.; He, Q.; Lei, L.; Ji, K. Leveraging Visual Language Model and Generative Diffusion Model for Zero-Shot SAR Target Recognition. *Remote Sens.* **2024**, *16*, 2927. [CrossRef]
37. Zheng, Z.; Ermon, S.; Kim, D.; Zhang, L.; Zhong, Y. Changen2: Multi-Temporal Remote Sensing Generative Change Foundation Model. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–17. [CrossRef]
38. Ageospatial. GeoForge. 2024. Available online: https://medium.com/@ageospatial/geoforge-geospatial-analysis-with-large-language-models-geollms-2d3a0eaff8aa (accessed on 20 February 2024).
39. Li, X.; Wen, C.; Hu, Y.; Yuan, Z.; Zhu, X.X. Vision-Language Models in Remote Sensing: Current progress and future trends. *IEEE Geosci. Remote Sens. Mag.* **2024**, *12*, 32–66. [CrossRef]
40. Shaw, P.; Uszkoreit, J.; Vaswani, A.; Parmar, N.; Kiros, R.; Uszkoreit, J.; Jones, L. Self-Attention with Relative Position Representations. In Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 5687–5695.

41. Zhang, Y.; Yu, X.; Yang, J.; Zhang, Y.; Hu, X. Conditional Position Embedding for Neural Machine Translation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, WA, USA, 10–15 July 2022; pp. 1641–1651.
42. Su, J.; Zhang, H.; Li, X.; Zhang, J.; Li, Y. RoFormer: Enhanced Transformer with Rotary Position Embedding. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), Association for Computational Linguistics, Online, 1–6 August 2021 ; pp. 1915–1925.
43. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Hu, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
44. Touvron, H.; Cord, M.; Sablayrolles, A.; Dehghani, G. Training data-efficient image transformers & distillation through attention. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
45. Yan, D.; Li, P.; Li, Y.; Chen, H.; Chen, Q.; Luo, W.; Dong, W.; Yan, Q.; Zhang, H.; Shen, C. TG-LLaVA: Text Guided LLaVA via Learnable Latent Embeddings. *arXiv* **2024**, arXiv:2409.09564.
46. Liu, F.; Guan, T.; Li, Z.; Chen, L.; Yacoob, Y.; Manocha, D.; Zhou, T. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv* **2023**, arXiv:2310.14566.
47. Zhang, M.; Chen, F.; Li, B. Multistep Question-Driven Visual Question Answering for Remote Sensing. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4704912. [CrossRef]
48. Roberts, J.; Han, K.; Albanie, S. Satin: A multi-task metadataset for classifying satellite imagery using vision-language models. *arXiv* **2023**, arXiv:2304.11619.
49. Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; Chen, C. Towards geospatial foundation models via continual pretraining. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 16806–16816.
50. 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; pp. 16772–16782.
51. 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]
52. Yuan, Z.; Mou, L.; Hua, Y.; Zhu, X.X. Rrsis: Referring remote sensing image segmentation. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5613312. [CrossRef]
53. 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.
54. 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. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5917820.
55. Pang, C.; Wu, J.; Li, J.; Liu, Y.; Sun, J.; Li, W.; Weng, X.; 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.
56. Zhao, D.; Yuan, B.; Chen, Z.; Li, T.; Liu, Z.; Li, W.; Gao, Y. Panoptic perception: A novel task and fine-grained dataset for universal remote sensing image interpretation. *IEEE Trans. Geosci. Remote. Sens.* **2024**, *62*, 5620714. [CrossRef]
57. Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; Ji, R. Rotated multi-scale interaction network for referring remote sensing image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26658–26668.
58. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5612822. [CrossRef]
59. 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.
60. 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, 16–22 June 2024; pp. 27672–27683.
61. Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; Zhong, Y. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 5481–5489.
62. 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]
63. 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.

64. Li, L.; Ye, Y.; Jiang, B.; Zeng, W. GeoReasoner: Geo-localization with Reasoning in Street Views using a Large Vision-Language Model. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
65. Yuan, Z.; Xiong, Z.; Mou, L.; Zhu, X.X. Chatearthnet: A global-scale, high-quality image-text dataset for remote sensing. *arXiv* **2024**, arXiv:2402.11325.
66. Li, X.; Ding, J.; Elhoseiny, M. VRSBench: A Versatile Vision-Language Benchmark Dataset for Remote Sensing Image Understanding. *arXiv* **2024**, arXiv:2406.12384.
67. 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.
68. Lobry, S.; Marcos, D.; Murray, J.; Tuia, D. RSVQA: Visual Question Answering for Remote Sensing Data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 8555–8566. [CrossRef]
69. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
70. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
71. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November 2010; GIS '10; pp. 270–279. [CrossRef]
72. Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
73. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
74. Zamir, S.W.; Arora, A.; Gupta, A.; Khan, S.H.; Sun, G.; Khan, F.S.; Zhu, F.; Shao, L.; Xia, G.; Bai, X. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. *arXiv* **2019**, arXiv:1905.12886.
75. Chen, H.; Shi, Z. A Spatial-Temporal Attention-Based Method and a New Dataset for Remote Sensing Image Change Detection. *Remote Sens.* **2020**, *12*, 1662. [CrossRef]
76. Long, Y.; Xia, G.S.; Li, S.; Yang, W.; Yang, M.Y.; Zhu, X.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]
77. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, PMLR, Hangzhou, China, 17–19 February 2023; pp. 19730–19742.
78. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 46595–46623.
79. Google. Gemini-1.0-Pro-Vision. 2024. Available online: https://ai.google.dev/docs/gemini_api_overview (accessed on 5 October 2024).
80. Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; Huang, L. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. *arXiv* **2024**, arXiv:2402.14289.
81. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983. [CrossRef]
82. Han, P.; Ma, C.; Li, Q.; Leng, P.; Bu, S.; Li, K. Aerial image change detection using dual regions of interest networks. *Neurocomputing* **2019**, *349*, 190–201. [CrossRef]
83. Lobry, S.; Tuia, D. Chapter 11 - Visual question answering on remote sensing images. In *Advances in Machine Learning and Image Analysis for GeoAI*; Prasad, S., Chanussot, J., Li, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2024; pp. 237–254. [CrossRef]
84. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. VQA: Visual Question Answering. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2425–2433. [CrossRef]
85. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 6–8 July 2016; pp. 1–5. [CrossRef]
86. Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring Models and Data for Remote Sensing Image Caption Generation. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2183–2195. [CrossRef]
87. Cheng, Q.; Huang, H.; Xu, Y.; Zhou, Y.; Li, H.; Wang, Z. NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5629419. [CrossRef]

88. Bashmal, L.; Bazi, Y.; Al Rahhal, M.M.; Zuair, M.; Melgani, F. Capera: Captioning events in aerial videos. *Remote Sens.* **2023**, *15*, 2139. [CrossRef]
89. Sun, Y.; Feng, S.; Li, X.; Ye, Y.; Kang, J.; Huang, X. Visual Grounding in Remote Sensing Images. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; MM '22; pp. 404–412. [CrossRef]
90. Liu, C.; Zhao, R.; Chen, H.; Zou, Z.; Shi, Z. Remote Sensing Image Change Captioning with Dual-Branch Transformers: A New Method and a Large Scale Dataset. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633520. [CrossRef]
91. Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 5622018. [CrossRef]
92. Li, Z.; Zhao, W.; Du, X.; Zhou, G.; Zhang, S. Cross-modal retrieval and semantic refinement for remote sensing image captioning. *Remote Sens.* **2024**, *16*, 196. [CrossRef]
93. Mao, C.; Hu, J. ProGEO: Generating Prompts through Image-Text Contrastive Learning for Visual Geo-localization. *arXiv* **2024**, arXiv:2406.01906.
94. Dong, S.; Wang, L.; Du, B.; Meng, X. ChangeCLIP: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 53–69. [CrossRef]
95. Singha, M.; Jha, A.; Solanki, B.; Bose, S.; Banerjee, B. Applenet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
96. Mo, S.; Kim, M.; Lee, K.; Shin, J. S-clip: Semi-supervised vision-language learning using few specialist captions. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 61187–61212.
97. Li, X.; Wen, C.; Hu, Y.; Zhou, N. RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103497. [CrossRef]
98. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv* **2023**, arXiv:2303.15389.
99. 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]
100. Silva, J.D.; Magalhães, J.; Tuia, D.; Martins, B. Large language models for captioning and retrieving remote sensing images. *arXiv* **2024**, arXiv:2402.06475.
101. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
102. Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; Xie, S. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
103. Dai, W.; Li, J.; LI, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: Nice, France, 2023; Volume 36, pp. 49250–49267.
104. Xu, Y.; Yu, W.; Ghamisi, P.; Kopp, M.; Hochreiter, S. Txt2Img-MHN: Remote sensing image generation from text using modern Hopfield networks. *IEEE Trans. Image Process.* **2023**, *32*, 5737–5750. [CrossRef]
105. Li, L. Cpseg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 513–522.
106. Zhang, Z.; Jiao, L.; Li, L.; Liu, X.; Chen, P.; Liu, F.; Li, Y.; Guo, Z. A Spatial Hierarchical Reasoning Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4400815. [CrossRef]
107. Ding, R.; Chen, B.; Xie, P.; Huang, F.; Li, X.; Zhang, Q.; Xu, Y. Mgeo: Multi-modal geographic language model pre-training. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taiwan, China, 23–27 July 2023; pp. 185–194.
108. Vivanco Cepeda, V.; Nayak, G.K.; Shah, M. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Adv. Neural Inf. Process. Syst.* **2023**, *36* , 8690–8701.
109. 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. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 5227–5244. [CrossRef] [PubMed]
110. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021.
111. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2217–2226. [CrossRef]
112. Lu, X.; Sun, X.; Diao, W.; Mao, Y.; Li, J.; Zhang, Y.; Wang, P.; Fu, K. Few-Shot Object Detection in Aerial Imagery Guided by Text-Modal Knowledge. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5604719. [CrossRef]
113. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]

114. Dai, D.; Yang, W. Satellite Image Classification via Two-Layer Sparse Coding with Biased Image Representation. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 173–176. [CrossRef]
115. Zheng, X.; Wang, B.; Du, X.; Lu, X. Mutual Attention Inception Network for Remote Sensing Visual Question Answering. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5606514. [CrossRef]
116. Xue, X.; Wei, G.; Chen, H.; Zhang, H.; Lin, F.; Shen, C.; Zhu, X.X. REO-VLM: Transforming VLM to Meet Regression Challenges in Earth Observation. *arXiv* **2024**, arXiv:2412.16583.
117. Lang, N.; Jetz, W.; Schindler, K.; Wegner, J.D. A high-resolution canopy height model of the Earth. *Nat. Ecol. Evol.* **2023**, *7*, 1778–1789. [CrossRef] [PubMed]
118. Sialelli, G.; Peters, T.; Wegner, J.D.; Schindler, K. AGBD: A Global-scale Biomass Dataset. *arXiv* **2024**, arXiv:2406.04928.
119. Wang, X.; Zhang, X.; Luo, Z.; Sun, Q.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Li, Z.; Yu, Q.; et al. Emu3: Next-Token Prediction is All You Need. *arXiv* **2024**, arXiv:2409.18869.

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.