

Vision-Language Modeling Meets Remote Sensing: Models, Datasets and Perspectives

Xingxing Weng, Chao Pang, and Gui-Song Xia

Abstract—Vision-language modeling (VLM) aims to bridge the information gap between images and natural language. Under the new paradigm of first pre-training on massive image-text pairs and then fine-tuning on task-specific data, VLM in the remote sensing domain has made significant progress. The resulting models benefit from the absorption of extensive general knowledge and demonstrate strong performance across a variety of remote sensing data analysis tasks. Moreover, they are capable of interacting with users in a conversational manner. In this paper, we aim to provide the remote sensing community with a timely and comprehensive review of the developments in VLM using the two-stage paradigm. Specifically, we first cover a taxonomy of VLM in remote sensing: contrastive learning, visual instruction tuning, and text-conditioned image generation. For each category, we detail the commonly used network architecture and pre-training objectives. Second, we conduct a thorough review of existing works, examining foundation models and task-specific adaptation methods in contrastive-based VLM, architectural upgrades, training strategies and model capabilities in instruction-based VLM, as well as generative foundation models with their representative downstream applications. Third, we summarize datasets used for VLM pre-training, fine-tuning, and evaluation, with an analysis of their construction methodologies (including image sources and caption generation) and key properties, such as scale and task adaptability. Finally, we conclude this survey with insights and discussions on future research directions: cross-modal representation alignment, vague requirement comprehension, explanation-driven model reliability, continually scalable model capabilities, and large-scale datasets featuring richer modalities and greater challenges.

Index Terms—Remote Sensing, Vision-Language Modeling, Contrastive Learning, Visual Instruction Tuning, Diffusion Model

I. INTRODUCTION

VISION-language modeling (VLM) in remote sensing, aiming to bridge the information gap between remote sensing images and natural language, facilitates a deeper understanding of remote sensing scene semantics like the attributes of ground objects and their relationship, and enables more natural human interaction with intelligent remote sensing data analysis models or methods [17], [164]. Since the introduction of remote sensing tasks such as image captioning [62], visual question answering [54], text-image (or image-text) retrieval [166], and text-based image generation [165], VLM in remote sensing has achieved significant success, driven by advancements in deep learning.

X. Weng, C. Pang, and G.-S. Xia are with the School of Computer Science, Wuhan University, Wuhan 430072, China. C. Pang, and G.-S. Xia are also with the School of Artificial Intelligence, Wuhan University, Wuhan 430072, China.

Corresponding Author: Gui-Song Xia (guisong.xia@whu.edu.cn)

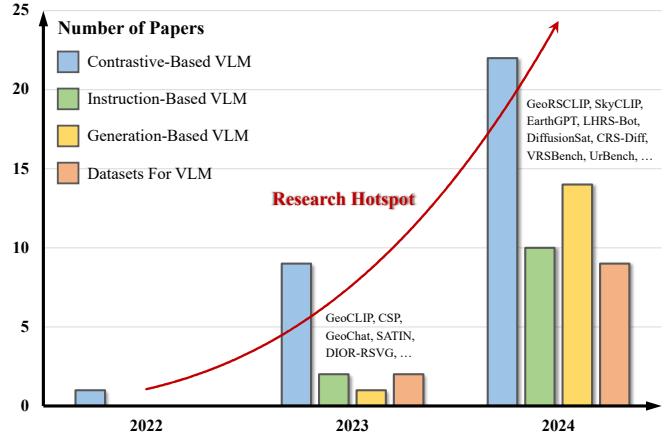


Fig. 1. The number of publications on visual-language modeling in remote sensing using the pre-training and fine-tuning paradigm.

Early works on VLM primarily emphasize the careful design of model architectures, followed by supervised training from scratch on small-scale datasets. For example, in image captioning research, many efforts [167]–[170] have been made to effectively combine convolutional neural networks (*e.g.* VGG [171] and ResNet [172]) with sequential models (*e.g.* LSTM [173] and Transformer [174]) before training on UCM-captions [62] and Sydney-captions [62] datasets. Under this classical construction paradigm, deep models often excel on test datasets but struggle to perform satisfactorily in large-scale deployments. Moreover, although these models are capable of describing image content, they fall short when tasked with answering questions about the images. In other words, they struggle to accomplish related tasks, such as visual question answering. The task-specific nature of these models seriously limits their applicability across diverse scenarios.

Recently, a new paradigm of pre-training followed by fine-tuning provides a promising solution to address the challenges mentioned above. The core idea is to first pre-train a model on massive image-text data, enabling it to capture general knowledge that covers a wide range of visual and textual concepts, along with their underlying correspondence. The pre-trained model is then fine-tuned on task-specific training data. The integration of general knowledge has been shown to enhance the model’s generalization ability in a single task [7], [8], while also making it more versatile and capable of handling a variety of downstream tasks [1], [3]. Consequently, vision-language modeling with this new paradigm has emerged as a prominent research focus in the field of remote sensing.

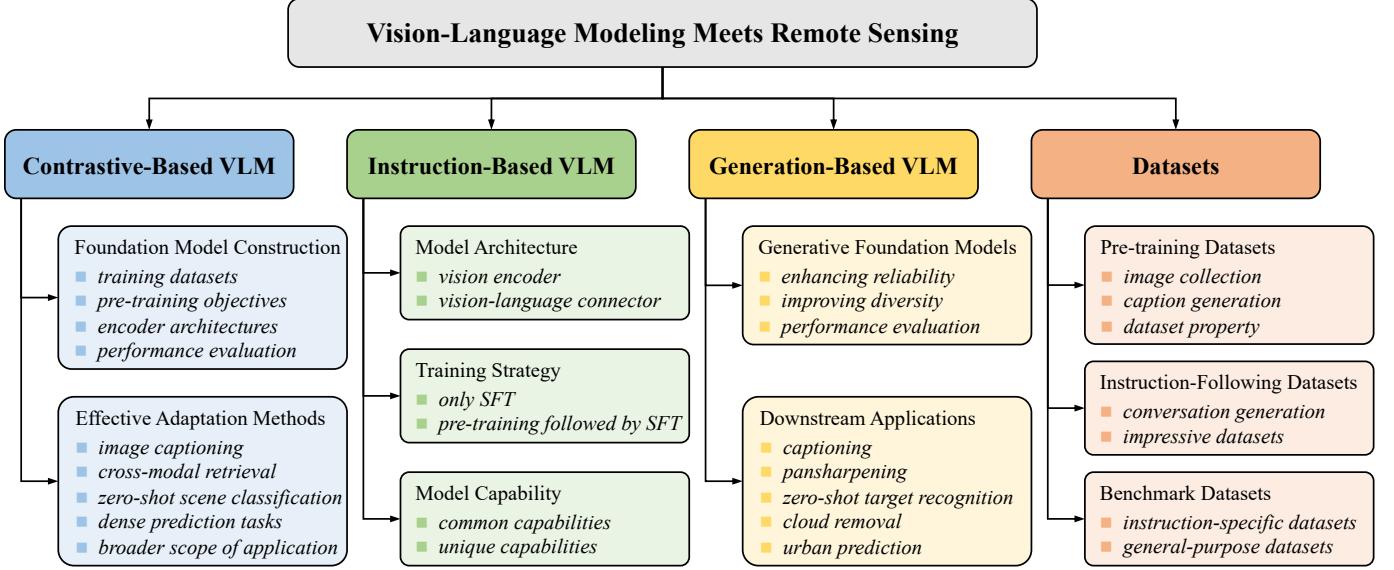


Fig. 2. Overview of this survey.

To date, significant progress has been achieved, as illustrated in Fig. 1. This includes works based on 1) contrastive learning [175], such as GeORSCLIP [7], SkyCLIP [8] and RemoteCLIP [2], which have driven substantial advancements in various cross-modal tasks and zero-shot image understanding tasks. 2) learning an implicit joint distribution between text and images, like RS-SD [7], DiffusionSat [38] and CRS-Diff [39], which allow for image generation from text prompts. 3) visual instruction tuning [201], such as GeoChat [3], LHRSS-Bot [9], and SkySenseGPT [11], which have demonstrated improved performance, diverse capabilities, and conversational interactions in remote sensing data analysis.

Despite these remarkable achievements, it is widely acknowledged that VLM remains an open challenge. Indeed, existing works have not yet achieved the level of remote sensing experts in processing remote sensing data. To provide clarity and motivation for further advances in the research community, several surveys have reviewed vision-language modeling in remote sensing. For instance, Li et al. [17] summarize vision-language models from an application perspective and suggest potential research opportunities. However, due to time constraints, they primarily concentrate on vision-only foundation models and early works. Zhou et al. [16] review recent developments but lack an in-depth analysis of key designs, which is significant for inspiring future research. Moreover, datasets, as a prerequisite of visual-language modeling research, have not been given adequate attention in existing surveys.

In this work, we aim to provide a timely and comprehensive review of the literature, with a focus on vision-language modeling based on the *pre-training and fine-tuning* paradigm in the field of remote sensing. Specifically, we cover: 1) a taxonomy of VLM in remote sensing, detailing commonly used network architectures and pre-training objectives for each category; 2) the latest advancements in contrastive-based, instruction-based, and generation-based vision-language modeling in remote sensing, highlighting key designs and down-

stream applications; 3) progress in datasets for VLM pre-training, fine-tuning, and evaluation; 4) several challenges and potential research directions. Fig. 2 presents an overview of this paper.

II. THE TAXONOMY OF VISUAL-LANGUAGE MODELING IN REMOTE SENSING

Under the *pre-training and fine-tuning* paradigm, vision-language modeling in remote sensing can be divided into three distinct groups based on their strategies for bridging the two modalities in the pre-training phase: contrastive learning, visual instruction tuning, and text-conditioned image generation. In this section, we present commonly used network architectures and pre-training objectives within each group.

Contrastive Learning: The motivation behind applying contrastive learning to vision-language modeling is training a model to map vision and language into a shared representation space, where an image and its corresponding text share similar representations while differing from other texts, which was first implemented by the pioneering work CLIP [86] in the field of computer vision. As illustrated in Fig. 3 (a), CLIP utilizes two independent encoders responsible for encoding visual and textual information. Given a batch of N image-text pairs $\{(x_i, y_i)\}_{i=1}^N$, the embeddings extracted by the two encoders followed by normalization are $\{\mathbf{e}_i^I, \mathbf{e}_i^T\}_{i=1}^N$. To achieve the similarity between visual and textual representations, CLIP is trained to maximize the cosine similarity of the embeddings of the i_{th} image and the corresponding i_{th} text ($i \in [N]$ with $[N] = \{1, 2, \dots, N\}$), while minimizing the cosine similarity of the embeddings of the i_{th} image and j_{th} text ($i, j \in [N]$, $i \neq j$). The loss is computed using InfoNCE [176] as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N} \sum_{i=1}^N (\log \sigma_\tau(s_{ii}^I, \{s_{ij}^I\}_{j=1}^N) + \log \sigma_\tau(s_{ii}^T, \{s_{ij}^T\}_{j=1}^N)), \quad (1)$$

where $s_{ij}^I = \mathcal{S}(\mathbf{e}_i^I, \mathbf{e}_j^T)$ and $s_{ij}^T = \mathcal{S}(\mathbf{e}_i^T, \mathbf{e}_j^I)$ are the similarity scores between the image and text embeddings, as computed by the function $\mathcal{S}(\mathbf{e}, \mathbf{e}') = \frac{\mathbf{e} \cdot \mathbf{e}'}{\|\mathbf{e}\| \|\mathbf{e}'\|}$. $\sigma_\tau(s_i, \{s_j\}_{j=1}^N) =$

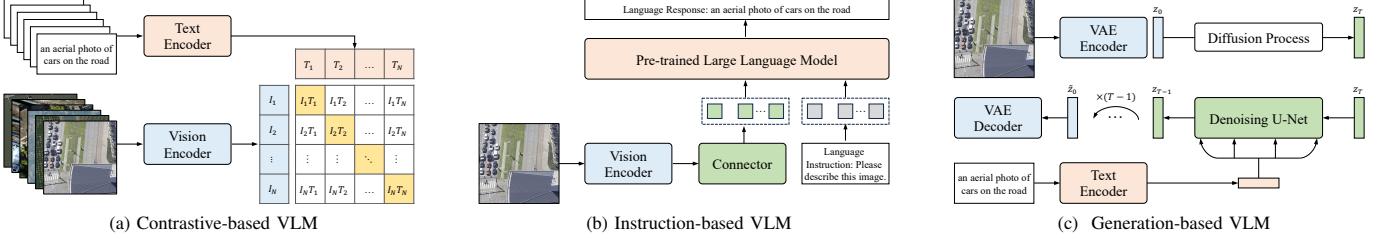


Fig. 3. Based on the strategies employed to bridge remote sensing images and natural language during the pre-training phase, vision-language modeling (VLM) in remote sensing can be categorized into three groups: (a) contrastive learning, (b) visual instruction tuning, and (c) text-conditioned image generation. This image is recreated by us based on [86], [201], [203].

$\exp(s_i/\tau)/(\sum_{j=1}^N \exp(s_j/\tau))$ is the softmax function, which normalizes the similarity score s_{ii}^I or s_{ii}^T over the corresponding sets $\{s_{ij}^I\}_{j=1}^N$ or $\{s_{ij}^T\}_{j=1}^N$. τ is the temperature.

The original CLIP model was trained on 400 million image-text pairs collected from the Internet and has demonstrated impressive results across various computer vision tasks [177]–[179]. These advancements spark interest in extending its capability to advance vision-language modeling in remote sensing. Two primary lines of research have been actively explored. The first, following the CLIP learning way, focuses on pre-training foundation models that are task-agnostic but specifically adapted for remote sensing domain. This includes efforts such as constructing large-scale image-text datasets [7], [8] and developing novel pre-training objectives [4], [26]. The second line explores effective adaptation of pre-trained CLIP models toward diverse downstream tasks, including image captioning [14], [19], zero-shot scene classification [20], [25], image-text retrieval [15], [31], etc.

Visual Instruction Tuning: Optimizing a model from scratch for image-text alignment is extremely resource-intensive due to the need for vast amounts of data and computational power. Fortunately, many pre-trained vision encoders and language models have been released. Pre-trained vision encoders can provide high-quality visual representations, while pre-trained language models, particularly large language models (LLMs), demonstrate advanced language understanding capabilities. As a result, recent works increasingly leverage these models to achieve image-text alignment through visual instruction tuning, as introduced by LLaVA [201] and MiniGPT-4 [129]. Fig.3 (b) illustrates a network architecture for this type of work, consisting of three key components: a pre-trained vision encoder, a connector, and a large language model. Specifically, the vision encoder compresses remote sensing images into compact visual representations, while the connector maps image embeddings into the word embedding space of the LLM. The LLM then receives both visual information and language instructions to perform reasoning tasks. Different from CLIP, which directly takes images and corresponding texts as input, this type of work preprocesses image-text pairs to instruction-following data. In this setup, each image is accompanied by a question, or language instruction, requiring the LLM to describe the image, while the corresponding text serves as the ground truth for the LLM’s predictions. Denote a batch of instruction-following data as $\{(x_i, q_i, y_i)\}_{i=1}^N$, where q_i is the question associated with the

i_{th} image. The pre-training objective is defined as:

$$\mathcal{L}_{VIT} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{L_i} \sum_{j=1}^{L_i} \log P(w_j|x_i, q_i, y_{i,<j}), \quad (2)$$

where L_i is the length of the caption $y_i = \{w_1, w_2, \dots, w_{L_i}\}$, $y_{i,<j}$ denotes the masked sequence that contain only the first $j-1$ words, and $P(w_j|x_i, q_i, y_{i,<j})$ is the conditional probability of generating the caption word w_j given the image x_i , question q_i and the previous words $y_{i,<j}$ in the caption. During pre-training, the vision encoder and large language model are typically kept frozen, with only the parameters of the connector being trainable.

In [129], [201], the authors demonstrated that pre-training with visual instruction tuning can align vision and language representations while preserving extensive knowledge. Since then, advances have been made by modifying network architectures and creating high-quality pre-training datasets [9], [202]. In addition to improving alignment, another line of research focuses on supervised fine-tuning, aiming to enable the model to perform a variety of remote sensing image analysis tasks and interact with users in a conversational manner. This includes efforts to generate task-specific instruction-following data [3], [5], [11], design novel training strategies [136], [202], and incorporate cutting-edge vision encoders [136].

Text-Conditioned Image Generation: Taking advantage of the advances in conditional image generation, a group of works [7], [38], [39] uses off-the-shelf generative models, primarily Stable Diffusion [203], to generate remote sensing images given text prompts, which essentially learns an implicit joint distribution between images and texts. As illustrated in Fig.3 (c), their network architecture comprises three main components: a text encoder, a variational autoencoder (VAE) [204], and a denoising U-Net. During training, the VAE encoder first transforms an image into a latent representation z_0 . Gaussian noise ϵ is then added to this latent representation at different timesteps t , resulting in $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ is a time-dependent scaling factor. Next, the conditioning representation c , extracted from the text encoder, is provided as input alongside z_t to the denoising U-Net, which predicts the noise added at timestep t . Finally, the VAE decoder upsamples the denoised latent representation to reconstruct the input image. Based on image-text pairs, the training objective is defined as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0, c, \epsilon, t} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2], \quad (3)$$

TABLE I
SUMMARY OF CONTRASTIVE-BASED VISION-LANGUAGE FOUNDATION MODELS IN REMOTE SENSING.

Model	Vision Encoder	Text Encoder	Training Dataset	Pre-training Objective	Evaluation Task	Public
GeoRSCLIP [7] [link]	ViT-B/32, ViT-H/14 [180]	Transformer [174]	RSSM [7]	InfoNCE loss [176]	Zero-Shot Scene Classification Image-Text Retrieval Semantic Localization	✓
RemoteCLIP [2] [link]	ResNet-50 [172], ViT-B/32, ViT-L/14 [180]	Transformer [174]	RET-3+SEG-4+DET-10 [2]	InfoNCE loss [176]	ZeroFew-Shot Scene Classification Image-Text Retrieval Object Counting Linear/k-NN Classification	✓
SkyCLIP [8] [link]	ViT-B/32, ViT-L/14 [180]	Transformer [174]	SkyScript [8]	InfoNCE loss [176]	Zero-Shot Scene Classification Image-Text Retrieval	✓
S-CLIP [18] [link]	ResNet-50 [172], ViT-B/32, ViT-B/16 [180]	Transformer [174]	RS-ALL [18] NWPU-RESISC45 [53]	InfoNCE loss [176] Pseudo-label losses [18]	Zero-Shot Fine-Grained Classification Zero-Shot Scene Classification	✓
Set-CLIP [26]	ResNet [172]	Transformer [174]	RS-ALL [18]	InfoNCE loss [176] SSL loss [184]	Image-Text Retrieval	✓
GRAFT [4] [link]	ViT-B/16 [180]	-	Internet-NAIP Dataset [4] Internet-Sentinel-2 Dataset [4]	Image/Pixel-level contrastive losses [4]	MK-MMD, SDD losses [26] Zero-Shot Scene Classification Zero-Shot Text-to-Image Retrieval	✓
CSP [37] [link]	ResNet-50 [172]	Sinusoidal transform+FcNet [181]	fMoW [117]	InfoNCE loss [176]	Zero-Shot Semantic Segmentation Zero-Shot Visual Question Answering	✓
GeoCLIP [36] [link]	ViT-L/14 [180]	Random Fourier features+MLP [185]	MP-16 [182]	InfoNCE loss [176] SSL loss [184]	Few-Shot Geo-Aware Scene Classification Image Geo-Localization	✓
SatCLIP [35] [link]	ResNet-18, ResNet-50 [172], ViT-S/16 [180]	Spherical harmonics+Siren [183]	S2-100K [35]	InfoNCE loss [176]	Text-Query Geo-Localization Geo-Aware Image Classification Air Temperature Prediction Elevation Prediction Median Income Estimation California Housing Price Estimation Population Density Estimation	✓
PIR-CLIP [133] [link]	ViT-B/32 [180]+ResNet-50 [172]	Transformer [174]	RSSM [7]	InfoNCE loss [176] Affiliation loss [133]	Ecoregion Classification Country Code Classification Species Recognition Image-Text Retrieval	✓

[link] directs to model websites. Image-text retrieval involves image-to-text and text-to-image retrieval tasks. In *Vision Encoder* column, we separate different architectures of vision encoders with commas, whereas in [133], the vision encoder is a combination of two architectures. *Public* refers to the availability of both code and model weights. S-CLIP has only open-sourced its code.

where $\epsilon_\theta(z_t, t, c)$ is the model's predicted noise, parameterized by θ . The application of diffusion models in remote sensing has shown rapid development, encompassing areas such as remote sensing image generation, enhancement, and interpretation [205]. As this paper focuses on vision-language modeling using diffusion models, we do not attempt to cover every instance of their application in remote sensing tasks. Instead, we specifically highlight works that integrate text conditions with diffusion models. For a more comprehensive overview of the advancements in diffusion models for remote sensing, please refer to [205].

Based on the fundamental principles outlined above, two major research groups have emerged. The first group aims to develop generative foundation models for various remote sensing images, including satellite [38], aerial [206], hyperspectral [207], and multi-resolution images [208]. The second group, in contrast, extends text-conditioned diffusion models to specific remote sensing tasks, such as image or change captioning [150], [154], pansharpening [148] and zero-shot target recognition [155].

III. CONTRASTIVE-BASED VISION-LANGUAGE MODELING

Most existing works on VLM fall into the group of employing contrastive learning. As mentioned previously, there are two main research directions being actively investigated, namely *foundation model construction* and *effective adaptation*. Specifically, foundation model construction concerns the large domain gap between natural and remote sensing images, aiming to learn visual representations with rich remote sensing scene semantics and well-aligned with textual representations. On the other hand, effective adaptation answers the question of how to leverage pre-trained CLIP models for specific

remote sensing tasks. In the following sections, we analyze the existing works from these two directions.

A. Foundation Model Construction

Table I summarizes contrastive-based vision-language models in remote sensing. To build foundation models, three key components need to be carefully designed: training datasets, pre-training objectives, and encoder architectures.

Training Datasets: Large-scale image-text datasets form the basis for constructing foundation models. Ready-made image-text datasets in remote sensing, *e.g.* UCM-captions and Sydney-captions, suffer from limited data volume and insufficient image diversity, rendering them inadequate for pre-training models to capture general knowledge of the domain. Recognizing the availability of numerous remote sensing image datasets, [7] and [2] use open-source datasets as the image source and develop image captioning methods to generate corresponding textual descriptions. Notably, [7] filter 11 commonly-used image-text datasets using RS-related keywords and caption 3 large-scale RS image datasets (Million-AID [100], fMoW [117], and BigEarthNet [113]) with the aid of the tuned BLIP2 model [118], resulting in the RSSM dataset, which contains over 5 million image-text pairs. Off-the-shelf vision-language models are indeed powerful tools for building large-scale image-text datasets due to their availability and ease of use, but ensuring captioning accuracy remains a significant challenge. To address this, [2] proposes a rule-based method called mask-to-box and box-to-caption, which converts pixel-wise or bounding box annotations into natural language captions. Another concern is that the semantic diversity of the generated captions is constrained by the limited number of predefined classes in open-source remote sensing image

datasets. Given this, Wang et al. [8] attempt to leverage rich semantic information contained in OpenStreetMap, allowing the textual descriptions to encompass not only a wide variety of object categories but also fine-grained subcategories and object attributes. Similar to [2], captions are assembled from object tags following predefined rules.

Pre-training Objectives: Instead of creating large-scale datasets, several works [4], [18], [26] explore new training objectives to facilitate model pre-training with few remote sensing image-text pairs. For instance, S-CLIP [18] introduces caption-level and keyword-level pseudo labels to fine-tune the original CLIP model on massive unpaired remote sensing images alongside a few image-text pairs. Denote a large number of unpaired images as $\{u_i\}_{i=1}^M$ ($M \gg N$). The caption-level pseudo-label $q_i^c \in \mathbb{R}^N$ is based on the assumption that the semantics of an unpaired image u_i can be represented as a combination of those of paired images. Thus, q_i^c represents a probability distribution over the captions of N paired images, derived from the relationships between unpaired and paired images, which are formulated as an optimal transport problem. The keyword-level pseudo label $q_i^k \in \mathbb{R}^K$ relies on the assumption that u_i shares keywords with visually similar images, representing the similarity between the embeddings of u_i and the keywords (drawn from the nearest paired image) $\{k_j\}_{j=1}^K$. The loss functions for two pseudo-labels are defined as:

$$\mathcal{L}_{\text{caption}} = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^N q_{ij}^c \log \sigma_\tau(s_{ij}^U, \{s_{il}^U\}_{l=1}^N), \quad (4)$$

$$\mathcal{L}_{\text{keyword}} = -\frac{1}{M} \sum_{i=1}^M \sum_{j=1}^K q_{ij}^k \log \sigma_\tau(s_{ij}^U, \{s_{il}^U\}_{l=1}^K). \quad (5)$$

In Eq.(4), $s_{il}^U = \mathcal{S}(\mathbf{e}_i^U, \mathbf{e}_l^T)$ refers to the similarity score of the unpaired image embedding \mathbf{e}_i^U and the caption embedding \mathbf{e}_l^T of the paired image. Meanwhile, in Eq.(5), $s_{il}^U = \mathcal{S}(\mathbf{e}_i^U, \mathbf{e}_l^k)$ denotes the similarity score between \mathbf{e}_i^U and the keyword embedding \mathbf{e}_l^k . Since few image-text pairs are involved in the fine-tuning process, the overall training objective also includes the InfoNCE loss, as shown in Eq.(6).

$$\mathcal{L}_{\text{S-CLIP}} = \mathcal{L}_{\text{InfoNCE}} + \frac{1}{2} (\mathcal{L}_{\text{caption}} + \mathcal{L}_{\text{keyword}}). \quad (6)$$

With a similar training data setup consisting of massive unpaired images and texts $\{u_i, v_i\}_{i=1}^M$ and limited image-text pairs $\{(x_i, y_i)\}_{i=1}^N$, Set-CLIP [26] transforms the representation alignment between images and texts into a manifold matching problem, developing a multi-kernel maximum mean discrepancy loss $\mathcal{L}_{\text{MK-MMD}}$ and a semantic density distribution loss \mathcal{L}_{SDD} . The loss $\mathcal{L}_{\text{MK-MMD}}$ constrains the consistency of whole representation distributions of images and texts, thereby achieving macro-level alignment, as formulated in Eq.(7).

$$\mathcal{L}_{\text{MK-MMD}} = \left\| \frac{1}{M+N} \sum_{i=1}^{M+N} \phi(\mathbf{e}_i^I) - \frac{1}{M+N} \sum_{i=1}^{M+N} \phi(\mathbf{e}_i^T) \right\|_{\mathcal{H}_{\text{RKHS}}}^2, \quad (7)$$

where $\phi(\cdot)$ is a linear combination of Gaussian and Polynomial kernel functions, used to map image and text embeddings $\mathbf{e}_i^I, \mathbf{e}_i^T$ into Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_{\text{RKHS}}$.

Meanwhile, the loss \mathcal{L}_{SDD} refines the alignment between the two modalities by ensuring that their probability density distributions remain similar in the representation space. It can be formulated as:

$$\mathcal{L}_{\text{SDD}} = \frac{1}{2} (\Gamma(I, T) + \Gamma(T, I)), \quad (8)$$

where $I = \{\mathbf{e}_i^I\}_{i=1}^{M+N}$ and $T = \{\mathbf{e}_i^T\}_{i=1}^{M+N}$ refer to embedding distributions of images and texts, respectively. $\Gamma(\cdot, \cdot)$ represents the Kullback-Leibler divergence, which measures the dissimilarity between two distributions. The format of $\Gamma(I, T)$ is given by:

$$\Gamma(I, T) = \sum_{i=1}^{M+N} \frac{\kappa(\mathbf{e}_i^I, I)}{\sum_{j=1}^{M+N} \kappa(\mathbf{e}_j^I, I)} \log \frac{\kappa(\mathbf{e}_i^I, I) / \sum_{j=1}^{M+N} \kappa(\mathbf{e}_j^I, I)}{\kappa(\mathbf{e}_i^T, T) / \sum_{j=1}^{M+N} \kappa(\mathbf{e}_j^T, T)}, \quad (9)$$

with $\Gamma(T, I)$ defined similarly by swapping the roles of I and T . Here, $\kappa(\cdot, \cdot)$ is an exponential probability density function that estimates the density value for each embedding in the corresponding distribution. Additionally, the self-supervised contrastive loss \mathcal{L}_{SSL} is introduced to obtain robust feature representations for each modality independently, defined as:

$$\mathcal{L}_{\text{SSL}} = -\frac{1}{M+N} \sum_{i=1}^{M+N} \log \frac{\exp(\mathcal{S}(\mathbf{e}_i, \mathbf{e}_i^+)/\tau)}{\sum_{j=1}^{M+N} \exp(\mathcal{S}(\mathbf{e}_i, \mathbf{e}_j)/\tau)}, \quad (10)$$

where \mathbf{e}_i is an embedding, and \mathbf{e}_i^+ is the representation of the positive sample generated by augmentation techniques. Ultimately, the overall training objective is formulated as:

$$\mathcal{L}_{\text{Set-CLIP}} = \alpha \mathcal{L}_{\text{InfoNCE}} + \mu \mathcal{L}_{\text{SSL}} + \delta \mathcal{L}_{\text{MK-MMD}} + \eta \mathcal{L}_{\text{SDD}}, \quad (11)$$

where α, μ, δ and η are tunable hyperparameters to balance the loss terms.

Even a small set of image-text pairs requires expert knowledge to craft textual descriptions. To avoid textual annotations entirely, GRAFT [4] proposes utilizing co-located ground images as the bridge between satellite images and language. In doing so, a dataset of two million ground-satellite image pairs, denoted as $\{x_i, \{g_{ij}\}_{j=1}^{N_i}\}_{i=1}^N$, is collected to support model training. Building on this, a feature extractor is designed to map satellite images to the representation space of the CLIP model, which was trained on internet image-text pairs. Since a satellite image x_i can cover a large ground area and thus be associated with multiple ground images $\{g_{ij}\}_{j=1}^{N_i}$, the extractor is optimized using the following loss:

$$\mathcal{L}_{\text{GRAFT}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \log \frac{\exp(\mathcal{S}(\mathbf{e}_i^I, \mathbf{e}_{ij}^G)/\tau)}{\sum_{a=1}^N \sum_{b=1}^{N_a} \exp(\mathcal{S}(\mathbf{e}_i^I, \mathbf{e}_{ab}^G)/\tau)}, \quad (12)$$

where \mathbf{e}_{ij}^G is the embedding of the j_{th} ground image corresponding to the i_{th} satellite image. This loss focuses solely on aligning image-level representations between the two types of images while ignoring the fact that a ground image g_{ij} can be mapped to a specific pixel location p_{ij} in the corresponding satellite image x_i . Thus, GRAFT trains the feature extractor with an additional pixel-level contrastive loss as follows:

$$\mathcal{L}_{\text{GRAFT}} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{N_i} \sum_{j=1}^{N_i} \log \frac{\exp(\mathcal{S}(\mathbf{e}_{ij}^P, \mathbf{e}_{ij}^G)/\tau)}{\sum_{a=1}^N \sum_{b=1}^{N_a} \exp(\mathcal{S}(\mathbf{e}_{ij}^P, \mathbf{e}_{ab}^G)/\tau)}, \quad (13)$$

where \mathbf{e}_{ij}^P denotes the feature vector for pixel p_{ij} .

Encoder Architectures: Unlike natural images, where corresponding textual information typically describes the image content, the textual information for remote sensing images can be represented by their geographic coordinates (longitude and latitude), which are beneficial for tasks such as scene classification and object recognition [117]. Therefore, some works propose aligning representations from remote sensing images and their geographic coordinates, with particular attention given to the choice of location encoders [35]–[37]. A location encoder generally consists of a nonparametric functional positional encoding combined with a small neural network. In CSP [37], an existing 2D location encoder, namely Space2Vec’s grid proposed in [181], is utilized. This encoder maps image coordinates into high-dimensional representations using sinusoid transforms for position encoding, followed by fully connected ReLU layers. GeoCLIP [36], on the other hand, first applies equal earth projection to the image coordinates to reduce distortions inherent in standard geographic coordinate systems. It then adopts random Fourier features [185] to capture high-frequency details, varying the frequency to construct hierarchical representations. These hierarchical representations are processed through separate multi-layer perceptions (MLPs), followed by element-wise addition, resulting in a joint representation. This design enables the location encoder to effectively capture features of a specific location across multiple scales. In SatCLIP [35], the authors use a location encoder that combines spherical harmonics basis functions with sinusoidal representation networks to map geographic coordinates into latent representations [183].

In addition to adapting encoders for different types of textual information, another purpose of adjusting encoders is to improve the representations of visual concepts. Remote sensing images typically cover a large field of view and include a variety of objects. However, the corresponding textual descriptions often center on specific objects of interest and their relationships. Semantic noise, such as irrelevant objects and background, can interfere with the representation of key content in images, thereby obstructing the alignment between image and text representations. In [133], this problem is addressed by using prior knowledge of remote sensing scenes to instruct the pre-trained model in filtering out semantic noise before calculating the similarity between image and text embeddings. This is practically achieved by adding an instruction encoder and a transformer encoder layer on top of the vision encoder. The instruction encoder, pre-trained on the scene classification dataset AID [102], generates instruction embeddings, which are used to filter image embeddings via a soft belief strategy. The filtered embeddings are then activated using instruction information through the transformer encoder layer, producing relevant image embeddings for subsequent alignment.

Performance Evaluation: Zero-shot scene classification and image-text retrieval are commonly used tasks to assess foundation models’ capabilities to capture a wide range of visual and textual concepts, along with their correspondence. Table II lists the performance of contrastive-based vision-language foundation models on these two tasks. Note that it shows the best model performance in the original papers. For

zero-shot scene classification, three conclusions can be drawn: 1) AID [102] is the most widely used dataset for evaluation, with RemoteCLIP achieving the best accuracy of 87.9%; 2) Using the same vision encoder (ViT-L/14), SkyCLIP, which is pre-trained on images collected from Google Earth, outperforms RemoteCLIP only on PatternNet [95], a dataset sourced from Google Earth; 3) Comparing models pre-trained on limited image-text pairs (*i.e.* S-CLIP, SetCLIP, and GRAFT) with RemoteCLIP, their performance is somewhat comparable, showing the importance of exploring data-effective pre-training. For cross-modal retrieval, PIR-CLIP achieves state-of-the-art performance on RSICD [60] and RSITMD [61]. In comparison, S-CLIP and SetCLIP perform significantly worse, with a performance gap exceeding 20%. This suggests that existing foundation models pre-trained on limited image-text pairs are capable of acquiring general visual and textual concepts but may face challenges in learning the relationships between them.

B. Effective Adaptation Methods

Inspired by the outstanding performance of foundation models, numerous methods have been developed to effectively adapt pre-trained models for specific remote sensing tasks, including image captioning, cross-modal retrieval, zero-shot classification, dense prediction, etc. This section presents the development of adaptation methods within the context of different downstream tasks.

Image Captioning: aims to describe the content of a given image using natural language. Vision-language foundation models are well-suited for this task, as they align image and text representations. One can use the vision encoder of these models to extract representations of image content, which prepares the input for language models to generate captions [19], [29], [43]. Commonly used vision encoders for remote sensing image captioning are from CLIP, while language models for this task include GPT-2 [200], BERT [198], and OPT [197]. To improve caption accuracy, a few works are devoted to enhancing models’ visual representation capabilities or further aligning images and texts.

For example, regarding representation enhancement, VLCA [29] introduces external attention [199] to capture potential correlations between different images, improving visual representations. MVP [19] fuses visual representations from pre-trained vision-language models and pre-trained vision models through stacked transformer encoders. It also leverages CLIP’s vision encoder, followed by adaptive average pooling layers, to generate visual prefixes, which are concatenated with token embeddings. A BERT-based caption generator is subsequently developed to combine the fused visual representations and concatenated embeddings, enabling the generation of accurate captions. In contrast, BITA [43] focuses on improving the alignment of images and texts in the remote sensing domain by introducing an interactive Fourier transformer. In this design, learnable visual prompts are fed to the Fourier layer and interact with image embeddings from a pre-trained vision encoder, capturing the most relevant visual representations. Through contrastive learning, these visual

TABLE II
PERFORMANCE OF CONTRASTIVE-BASED VISION-LANGUAGE FOUNDATION MODELS.

Zero-Shot Scene Classification: Top-1 Accuracy																		
Model	Vision Encoder	AID [102]	EuroSAT [69]	fMoW [117]	Million-AID [100]	MLRSNet [103]	RESISC [53]	Optimal-31 [98]	PatternNet [95]	RSSCN7 [72]	RSC11 [93]	RSI-CB128 [101]	RSI-CB256 [101]	RSICD [60]	SIRI-WHU [94]	UCM [70]	WHU-RS19 [186]	
SkyCLIP [8]	ViT-L/14 [180]	71.70	51.33	27.12	67.45	-	70.94	-	80.88	-	-	-	50.09	-	-	-	-	
Set-CLIP [26]	ResNet-50 [172]	76.20	-	-	-	-	-	-	-	66.20	-	-	-	69.20	-	67.50	89.00	
GRAFT [4]	ViT-B/14 [180]	-	63.76	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GeoRSCLIP [7]	ViT-H/14 [180]	76.33	67.47	-	-	-	73.83	-	-	-	-	-	-	-	-	-	-	
S-CLIP [18]	ViT-B/16 [180]	85.20	-	-	-	-	-	-	-	76.30	-	-	-	79.50	-	82.30	93.90	
RemoteCLIP [2]	ViT-L/14 [180]	87.90	59.94	-	-	66.32	79.84	90.05	68.75	72.32	74.90	37.22	52.82	-	70.83	-	94.66	
Image-Text Retrieval: Recall@1, Recall@5, and Recall@10																		
Model	Vision Encoder	RSICD [60]			RSITMD [61]			UCM-captions [62]			Sydney-captions [62]							
Image-to-Text Retrieval	Encoder	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	R@1	R@5	R@10	Mean	
S-CLIP [18]	ResNet-50 [172]	4.20	18.40	-	-	-	-	-	-	11.60	45.70	-	-	14.90	50.00	-	-	
Set-CLIP [26]	ResNet-50 [172]	-	19.60	-	-	-	-	-	-	46.30	-	-	-	51.10	-	-	-	
SkyCLIP [8]†	ViT-L/14 [180]	-	-	-	23.70	-	-	-	30.75	-	-	-	-	72.22	-	-	-	
RemoteCLIP [2]	ViT-L/14 [180]	18.39	37.42	51.05	35.62	28.76	52.43	63.94	48.38	19.05	54.29	80.95	51.43	-	-	-	-	
GeoRSCLIP [7]	ViT-B/32 [180]	21.13	41.72	55.63	39.49	32.30	53.32	67.92	51.18	-	-	-	-	-	-	-	-	
PIR-CLIP [133]	+ResNet-50 [172]	27.63	45.38	55.26	42.76	45.58	65.49	75.00	62.02	-	-	-	-	-	-	-	-	
Text-to-Image Retrieval																		
S-CLIP [18]	ResNet-50 [172]	4.20	16.80	-	-	-	-	-	-	11.1	43.50	-	-	17.8	55.10	-	-	
Set-CLIP [26]	ResNet-50 [172]	-	17.40	-	-	-	-	-	-	44.10	-	-	-	55.20	-	-	-	
SkyCLIP [8]†	ViT-L/14 [180]	-	-	-	19.97	-	-	-	30.58	-	-	-	-	59.33	-	-	-	
RemoteCLIP [2]	ViT-L/14 [180]	14.73	39.93	56.58	37.08	23.76	59.51	74.73	52.67	17.71	62.19	93.90	57.93	-	-	-	-	
GeoRSCLIP [7]	ViT-B/32 [180]	15.59	41.19	57.99	38.26	25.04	57.88	74.38	52.43	-	-	-	-	-	-	-	-	
PIR-CLIP [133]	+ResNet-50 [172]	21.10	44.87	56.12	40.70	30.13	55.44	68.54	51.37	-	-	-	-	-	-	-	-	

RESISC is short for NWPU-RESISC45 [53]. *Mean* refers to the mean recall averaged of recall@1 (*R@1*), recall@5 (*R@5*) and recall@10 (*R@10*). † indicates that the training data corresponding to the test dataset was not used to pre-train the model.

representations are aligned with textual representations, also extracted by the Fourier-based transformer. The interactive Fourier transformer then connects the frozen vision encoder with the frozen language model, leveraging the language model’s generation and reasoning capabilities.

Beyond improving caption accuracy, generating detailed captions has also been explored [135], [297]. In [135], a two-stage instruction fine-tuning is proposed for the vision-to-language mapping layer to generate geographically detailed captions. The first stage, guided by the instruction “[region] Based on the provided region of the remote sensing image, describe the basic attributes of the main objects in that region.”, aligns geographic object regions with their attribute descriptions. The second stage, guided by the instruction “[image] Please provide a detailed description of this image.”, focuses on understanding the spatial distribution of geographic objects within images. In [297], image captioning is defined as the aggregation of information from multi-turn dialogues, where each turn serves to query the image content. In each turn, CLIP’s vision encoder extracts image features, which are then input into an auto-regressive language model [200] along with previous questions, answers, and the current question to generate the response. After several dialogue turns, GPT-3 [298] summarizes the dialogue information to produce an enriched textual image description.

Unlike previous works that caption a single image, Prompt-CC [42] utilizes pre-trained models to describe differences between bi-temporal images, a task known as change captioning. Based on bit-temporal visual representations from CLIP, Prompt-CC introduces an image-level classifier to detect the presence of changes, and a feature-level encoder to extract discriminative features that identify the specific changes.

Cross-Modal Retrieval: is the task of retrieving data from one modality by using queries from another modality. Based on vision-language foundation models, there is ongoing research that explores retrieval between images and text [31],

[131], [134], as well as between images captured with different imaging parameters [15]. Image-text retrieval involves the challenges of finding textual descriptions that correspond to given image queries and vice versa. Taking advantage of advances in vision-language foundation models, existing works use pre-trained encoders to encode images and text separately, mapping them into a shared representation space for similarity measurement. Typically, CISEN [31] adopts encoders from CLIP and GeoRSCLIP as the backbone for its retrieval network. Given the abundance of objects in remote sensing images, CISEN is trained in two stages to enhance visual representation. In the first stage, an image adapter [188] is trained to map global visual features into textual-like features. In the second stage, a feature pyramid network is used to integrate the textual-like features into local visual features, which are then utilized to enrich global visual features. Rather than focusing on effective representation, KTIR [131] aims to improve alignment by explicitly incorporating knowledge into text features. The knowledge is derived from textual information using off-the-shelf knowledge sources like RSKG [190] and ConceptNet [189]. Once converted to knowledge sentences, the knowledge is processed by the text encoder to extract knowledge features, which are then fused with text features via a single cross-attention layer.

Image-image retrieval involves searching for relevant images across different imaging parameters, such as matching RGB images with multispectral images, as explored in [15]. The authors use a pre-trained CLIP text encoder as the classification head, and fine-tune the CLIP vision encoder alongside a newly added multispectral-specific encoder. The training is conducted in stages: first, the CLIP vision encoder is fine-tuned on RGB images composited from multispectral images, followed by fine-tuning the new encoder on multispectral images. This encoder is guided to produce discriminative representations that can achieve accurate image classification, while ensuring its representations are similar to those of RGB

TABLE III
SUMMARY OF CONTRASTIVE-BASED VISION-LANGUAGE FOUNDATION MODELS APPLIED TO REMOTE SENSING TASKS.

Work	Task	Vision-Language Foundation Model	Adaptation
VLCA [29]	Image Captioning	CLIP ResNet-50, ViT [86]	Utilize external attention [199] to enhance visual representations extracted by CLIP's vision encoder.
MVP [19]	Image Captioning	CLIP ViT-B/16 [86]	Fuse visual representations from CLIP and pre-trained vision models.
BITA [43]	Image Captioning	CLIP ViT-L/14 [86]	Introduce the Fourier-based transformer to align images and texts in the remote sensing domain.
MGIMM [135]	Image Captioning	CLIP ViT-L/14 [86]	Region-level and image-level instruction tuning for generating captions with detailed geographical information.
CFD-C [297]	Image Captioning	CLIP ViT-B/16 [86]	Define image captioning as the aggregation of information from multi-turn dialogues
Prompt-CC [42]	Change Captioning	CLIP ViT-B/32 [86]	Develop an image-level classifier, a feature-level encoder on top of pre-trained vision encoder for detecting the presence of changes and extracting discriminative features.
CISEN [31]	Image-Text Retrieval	CLIP ResNet-50, ViT-B/32+Transformer [86] GeoRSCLIP ViT-B/32+Transformer [7]	Two-stage training to progressively fuse fine-grained semantic features to enrich visual representation.
KTIR [131]	Image-Text Retrieval	BLIP ViT-B/16+BERT [132]	Incorporate knowledge into text features for better image-text alignment.
Zavras et al. [15]	Zero-Shot RGB-Multispectral Retrieval Zero-Shot Scene Classification	CLIP ViT-B/32, ViT-B/16, ViT-L/14, ViT-L/14@336 [86]	Two-stage fine-tuning to achieve alignment between distinct image modalities in the CLIP representation space.
WEICOM [134]	Composed-to-Image Retrieval	CLIP ViT-L/14+Transformer [86] RemoteCLIP ViT-L/14+Transformer [2]	A weighting parameter for combining image-query and text-query retrieval results, achieving composed query.
Rahhal et al. [20]	Zero-Shot Scene Classification	Thirteen CLIP/Open-CLIP Models [86]	Tuning prompts by adding task-relevant context words, including "remote sensing image" and "scene".
Lan et al. [153]	Few-Shot Fine-Grained Ship Classification	CLIP ResNet-50, ViT-B/16+Transformer [86]	Inject domain priors to adapt frozen vision encoder to remote sensing scenes.
RS-CLIP [14]	Zero-Shot Scene classification	CLIP ViT-L/14+Transformer [48]	Design hierarchical, learnable prompts to capture rich task-specific knowledge.
DSVA [25]	Zero-Shot Scene Classification	CLIP ViT-B/32+Transformer [86]	Use pseudo labeling and curriculum learning to fine-tune the pre-trained model in multiple rounds.
Text2Seg [41]	Zero-Shot Semantic Segmentation	CLIP ViT-B/16+Transformer [86]	Employ CLIP to endow segmentation masks with semantic categories.
Lin et al. [40]	Semantic Segmentation	CLIP ViT-B/16+Transformer [86]	Modify CLIPSeg decoder [192] to receive CLIP joint image-text embeddings as input and generate segmentation masks.
SegEarth-OV [187]	Open-Vocabulary Semantic Segmentation	CLIP ViT-B/16+Transformer [86]	Develop SimFeatUp to robustly upsample low-resolution CLIP embeddings, Execute subtraction operations between patch embeddings and the class token embedding to alleviate global bias.
SCM [27]	Unsupervised Change Detection	CLIP [86]	Employ CLIP to identify objects of interest in bi-temporal images, helping filter out pseudo changes.
ChangeCLIP [13]	Change Detection	CLIP ResNet-50, ViT-B/16+Transformer [86]	Employ CLIP to construct, encode multi-modal input data for change detection, Design transformer-based decoder to combine vision-language features with image features, predicting change maps.
BAN [28]	Change Detection	CLIP ViT-B/16, ViT-L/14 [86] RemoteCLIP ViT-B/32, ViT-L/14 [2] GeoRSCLIP ViT-B/32, ViT-L/14 [7]	Design bridging modules to inject general knowledge from foundation models to existing change detection models.
Bazi et al. [23]	Visual Question Answering	CLIP ViT+Transformer [86]	Employ CLIP to encode images and questions.
Czernakowski et al. [21]	Cloud Presence Detection	CLIP [86]	Apply CLIP to this task via tuning prompts or adding linear classifier.
TGN [44]	Text-based Image Generation	CLIP [86]	Utilize CLIP to classify generated images, ensuring semantic class consistency.
AddressCLIP [196]	Image Address Localization	CLIP ViT-B/16+Transformer [86]	Align image with scene captions and address texts by contrastive learning, Introduce image-geography matching to constrain image features with the spatial distance.

images extracted by the CLIP vision encoder.

Considering the complexity of Earth's surface, single-modality queries, such as text queries, require users to fully articulate their needs to pinpoint relevant images. This raises the barrier to using retrieval models. To address this issue, Psomas et al. [134] introduce the composed-to-image retrieval task, which searches for remote sensing images based on a composed query of image and text. The retrieved images share the same scene or object category as the image query and reflect the attribute defined by the text query. To achieve this, the authors in [134] propose calculating similarity scores for the image and text query separately, then normalizing and combining these scores using a convex combination controlled by a weighting parameter λ , which adjusts the contribution of each modality. As shown in Fig. 4, when $\lambda = 0$, composed-to-image retrieval simplifies to image-to-image retrieval, while at $\lambda = 1$, it simplifies to text-to-image retrieval.

Zero-shot Scene Classification: challenges models to identify remote sensing images from scene categories that were not seen during training. The idea of applying vision-language foundation models to this task is straightforward: since these models align images and texts, zero-shot scene classification can be achieved by comparing image embeddings with text embeddings extracted by the text encoder,



Fig. 4. Illustration of composed image retrieval for remote sensing [134].

which takes as input textual descriptions specifying the unseen classes. By default, these descriptions follow the format "a photo of a [CLASS].", where the class token is replaced by the specific class name, such as "farmland", "forest" or "playground". The input text, also known as prompt, plays a crucial role in the performance of foundation models on downstream tasks. As a result, several works [14], [20] pay attention to prompt tuning and suggest adding task-relevant context words to improve performance. An example is provided by [20], where the authors replace "photo" with terms such as "remote sensing image", "top view image", "satellite image" and "scene". This slight modification results in more than a 5% increase in the accuracy of the CLIP model with the ViT-

B/32 vision encoder on UCM dataset [70]. However, as noted in other experiments from [20], prompts subjected to extensive tuning are not guaranteed to be optimal for the task and may not be suitable for other test datasets. To avoid manual prompt tuning, Lan et al. [153] model context words in a prompt as learnable vectors, which are combined with the class token embeddings before being input to the text encoder. Moreover, the class token embeddings for each category are organized across multiple levels of granularity, taking the form: “a photo of a ship, the primary type is [CLASS1], secondary type is [CLASS2], final type is [CLASS3].”, aimed at the task of few-shot fine-grained ship classification. Compared to hand-crafted prompts, these hierarchical, learnable prompts incorporate richer task-specific knowledge.

Most works employ and freeze pre-trained CLIP models. However, due to the significant domain gap between web images and remote sensing images, their performance tends to be limited. To mitigate this, one can inject remote sensing domain priors into the vision encoder [153], or fine-tune the entire model using pseudo labeling techniques [14]. Typically, Lan et al. [153] introduce a lightweight network that is trained on data from seen classes to capture the domain prior. This prior is then combined with image embeddings output by the vision encoder, allowing the pre-trained model to adapt better. On the other hand, RS-CLIP [14] leverages the strong transferability of the pre-trained model to automatically generate pseudo labels from unlabeled images, which are used to fine-tune the model. Additionally, a curriculum learning strategy is developed to gradually select more pseudo-labeled images for model training in multiple rounds, further boosting the model’s performance in zero-shot scene classification.

In addition to comparing the similarity between embeddings of images and unseen-class texts, DSVA [25] introduces a solution that uses pre-trained models to annotate attributes for each scene class, and predict scene categories by evaluating the similarity between attribute values derived from image embeddings and those associated with scene classes. For automatic attribute annotation, textual descriptions have the form of “This photo contains [ATTRIBUTE]”, where the attribute token is replaced by specific attribute names, such as “red”, “cement” or “rectangle”. Meanwhile, the attribute value for each class is calculated by measuring the similarity between embeddings of attribute text and images belonging to that class.

Dense Prediction Tasks: such as semantic segmentation and change detection, which produce pixel-level predictions for input images, have recently benefited from the application of vision-language foundation models. Typically, Text2Seg [41] uses a CLIP model to classify category-agnostic segmentation masks generated by the segment anything model (SAM) [191], enabling zero-shot semantic segmentation of remote sensing images. Lan et al. [40] instead modify the CLIPSeg decoder [192] to receive joint image-text embeddings from the CLIP model as input, producing a binary segmentation mask. The decoder is composed of transformer blocks, convolution layers, and linear interpolation layers.

Without the need for additional decoders or segmentation models, one can perform an upsampling operation on image

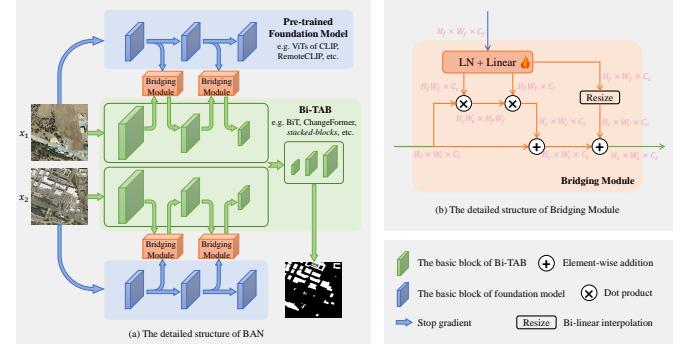


Fig. 5. Illustration of BAN architecture [28].

embeddings from CLIP, and compare the similarity between image patch embeddings and text embeddings to produce segmentation results. However, empirical findings in [187] suggest that for CLIP with a vision encoder based on ViT-B/16, the image embeddings are downsampled to 1/16 of the input image size. This downsampling leads to distorted object shapes and poorly fitting boundaries in segmentation masks. Furthermore, CLIP’s self-attention causes global information from the class token embedding to be attached to the patch embeddings, which significantly degrades performance in semantic segmentation. To handle these issues, SegEarth-OV [187] incorporates SimFeatUp on top of the CLIP vision encoder to restore lost spatial information in image embeddings. Subsequently, subtraction operations are executed between patch embeddings and the class token embeddings before similarity measurement, alleviating global bias in patch embeddings. In particular, SimFeatUp is a learnable upsample consisting of a single parameterized JBU operator. It is trained with a frozen CLIP model, a learnable downsample, and a lightweight content retention network. The training objective is to ensure that the image embeddings, after the up-downsampling process, remain similar to those from CLIP. Meanwhile, the image generated by the content retention network, which takes the upsampled image embeddings as input, should closely resemble the input image to CLIP.

When it comes to change detection, similar adaptation strategies used in segmentation tasks, such as combining SAM with CLIP and designing decoders, are applied. For instance, in SCM [27], CLIP is applied after SAM to identify objects of interest in bi-temporal images, helping to filter out pseudo changes. In ChangeCLIP [13], CLIP is used to construct and encode multi-modal input data for change detection tasks, while a transformer-based decoder combines bi-temporal vision-language features with image features to produce change maps. To be specific, the authors pre-define 56 common land cover categories in remote sensing images and use the CLIP model to identify bi-temporal images from these categories. Each image is paired with textual descriptions of the foreground and background, formatted as “remote sensing image foreground objects, [Predict Classes]” and “remote sensing image background objects”, respectively. This way allows bi-temporal texts to highlight the changing object, providing additional information for change detection.

Alternatively, only the vision encoders of foundation models are adopted, as most existing datasets provide only bi-temporal images for identifying changes. In BAN [28], the authors introduce bridging modules to inject general knowledge extracted from the vision encoders of foundation models into existing change detection models like BiT [193] and ChangeFormer [194]. Since the input image sizes for foundation models and change detection models may differ, and not all general knowledge contributes to predicting changes, the bridging modules are responsible for selecting, aligning, and injecting this knowledge. As shown in Fig.5, these modules consist of three main components: layer normalization and a linear layer to mitigate the distribution inconsistency between the image features from the two models, cross-attention to obtain valuable information from general knowledge, and bi-linear interpolation to solve the misalignment problem. The bridging modules are placed between the two encoders, executing multi-level knowledge injection.

Broader Scope of Application: In addition to remote sensing tasks previously discussed, there are several emerging applications of contrastive-based vision-language foundation models, including:

(1) *Visual Question Answering* attempts to provide answers to questions related to the content of images. In [23], CLIP is used to extract both visual and textual representations from images and questions. Two decoders, followed by two classifiers, are then developed to capture the intradependencies and interdependencies within and between these representations. The final answer is determined by averaging the predictions from both classifiers.

(2) *Cloud Presence Detection* involves identifying satellite images that are affected by clouds. The authors in [21] explore several strategies to adapt CLIP for this task: One approach, similar to zero-shot scene classification [20], involves selects prompts such as “This is a satellite image with clouds” and “This is a satellite image with clear sky”, and then classifying satellite images via similarity measurement. Another strategy follows CoOp [195], combining learnable context with the class token embeddings. Alternatively, the vision encoder can be employed by itself, with a linear classifier added on top. Unlike the first strategy, the other approaches necessitate training for prompts or classifiers.

(3) *Text-based Image Generation* refers to the task of creating images from textual descriptions, which can help mitigate the issue of class imbalance commonly found in remote sensing data. Typically, TGN [44] utilizes CLIP to classify generated images, ensuring semantic class consistency between input images and generated ones.

(4) *Image Address Localization* aims to predict the readable textual address where an image was taken [196]. Different from image geo-localization in [36], which predicts GPS coordinates (*i.e.*, latitude and longitude) from images, this task outputs semantic address text such as “Grant Street, Downtown”. Utilizing existing foundation models like CSP [37] and GeoCLP [36], a viable solution is predicting GPS coordinates with a foundation model and then converting them into readable addresses. However, this mapping from coordinates to addresses often presents ambiguities. A recent work, Address-

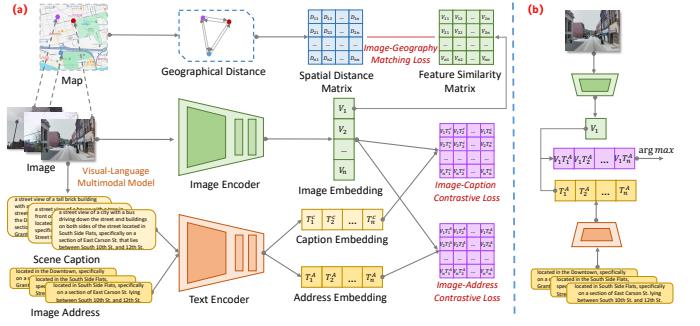


Fig. 6. Illustration of AddressCLIP framework [196].

CLIP [196], introduces a novel idea for performing image address localization in an end-to-end manner. In particular, AddressCLIP applies contrastive learning to align images with scene captions and address texts. Furthermore, image-geography matching is developed to bring features of geographically proximate images closer together while distancing features of images that are far apart geographically. Fig.6 shows the framework of AddressCLIP.

IV. INSTRUCTION-BASED VISION-LANGUAGE MODELING

Today, instruction-based vision-language modeling is advancing rapidly. Since 2023, over ten impressive vision-language models have emerged, as shown in Table IV. They are not only versatile, capable of performing a range of remote sensing image analysis tasks, but also able to interact with users in a conversational manner. This broadens the accessibility of intelligent models beyond experts in remote sensing, facilitating their widespread deployment and application. This section presents critical developments in terms of *model architecture*, and *training strategy*, *model capability*.

A. Model Architecture

As CLIP has been trained to align image and text representations, most works [3], [9]–[11], [202] directly employ its vision encoder. The commonly used large language models include the LLaMA family [57], [58] and its derivative Vicuna family [47]. Developed by Meta, the first-generation LLaMA [58] adopts transformer architectures ranging from 7B to 65B parameters, pre-trained on approximately 1.4 trillion tokens of publicly available text corpora. Despite its smaller scale, LLaMA outperforms proprietary models like GPT-3 [298]. Its successor, LLaMA 2 [57], introduces significant improvements, including expanded pre-training data (2 trillion tokens), architectural upgrades such as grouped-query attention [299], and a scaled-up 70B-parameters variant. Through supervised fine-tuning and reinforcement learning with human feedback, LLaMA 2 was further optimized for dialogue, yielding the LLaMA 2-Chat series [57]. Parallel to this development, the research community developed Vicuna [47] by fine-tuning LLaMA models on 70K conversations from ShareGPT. These models achieve more than 90% of ChatGPT’s quality.

Regarding model architecture, existing works are devoted to improving visual encoders to enhance visual perception, and

TABLE IV
SUMMARY OF INSTRUCTION-BASED VISION-LANGUAGE MODELS IN REMOTE SENSING.

Model	Vision Encoder	Connector	Large Language Model	Training Dataset	Training Strategy	Public
RSGPT [1]	EVA CLIP-G [56]	Q-Former [46] +Linear layer	Vicuna-7B/13B [47]	RSICap [1]	<ul style="list-style-type: none"> Load pre-trained weights of InstructBLIP [46] Fine-tune Q-Former and linear layer Load pre-trained weights of [56], [59] Employ LoRA [50] for linear layer and LLM in two stages 	✗
SkyEyeGPT [5]	EVA CLIP-G [56]	Linear layer	LLaMA 2-Chat [57]	SkyEye-968k [5]	<ul style="list-style-type: none"> Load pre-trained weights of [57], [67], [86] Train linear layer and LLM on LAION-400M and COCO Caption Employ bias tuning [233] for LLM on MMRS-1M Load pre-trained weights of [57], [67], [86] Train projection layer on COCO Caption and RSVP subset 	✗
EarthGPT [6]	DINOv2 ViT-L/14 [67] +CLIP ConvNeXt-L [86]	Linear layer	LLaMA 2-13B [57]	LAION-400M [209] COCO Caption [210] MMRS-1M [6]	<ul style="list-style-type: none"> Employ bias tuning [233] for LLM on MMRS-1M Load pre-trained weights of [57], [67], [86] Train projection layer on COCO Caption and RSVP subset 	✗
EarthMarker [136]	DINOv2 ViT-L/14 [67] +CLIP ConvNeXt-L [86]	Projection layer	LLaMA 2-13B [57]	COCO Caption [210] RSVP-3M [136] RefCOCO [211] RefCOCO+ [212]	<ul style="list-style-type: none"> Train projection layer on COCO Caption and RSVP subset Fine-tune attention layers of LLM on RefCOCO and RefCOCO+ Employ LoRA [50] for LLM on RSVP-3M Load pre-trained weights of [58], [67], [86] Train projection layer and employ LoRA [50] for LLM on COCO Caption Fine-tuning projection layer and employ LoRA and instruction adapters for LLM on MMShip Load pre-trained weights of [57], [86] 	✗
Popeye [24]	DINOv2 ViT-L/14 [67] +CLIP ViT-L/14 [86]	Projection layer	LLaMA-7B [58]	COCO Caption [210] MMShip [24]	<ul style="list-style-type: none"> Train projection layer and employ LoRA [50] for LLM on COCO Caption Fine-tuning projection layer and employ LoRA and instruction adapters for LLM on MMShip Load pre-trained weights of [57], [86] 	✗
LHRS-Bot [9] [link]	CLIP ViT-L/14 [86]	Vision perceiver	LLaMA 2-7B [57]	LHRS-Align [9] LHRS-Instruct [9] Multi-task Dataset [9] LLaVA-Instruct-158K [201]	<ul style="list-style-type: none"> Train vision perceiver on LHRS-Align Fine-tune vision perceiver and Employ LoRA [50] for LLM on LHRS-Instruct, multi-task dataset Fine-tune LLM on LHRS-Instruct, multi-task dataset and LLaVA-Instruct-158K subset Load pre-trained weights of [47], [86] Train vision encoder, MLP and LLM on VersaD Fine-tune MLP and LLM on VersaD-Instruct, HnStD and VariousRS-Instruct Load pre-trained weights of [47]–[49] Employ LoRA [50] for LLM Load pre-trained weights of [47], [86] Train MLP on a general image-language dataset Employ LoRA [50] for LLM Load pre-trained weights of [47], [49], [86] Fine-tune MLP and employ LoRA [50] for LLM Load pre-trained weights of [47], [86] Employ LoRA [50] for MLP and LLM Load pre-trained weights of [86], [214] Employ LoRA [50] for LLM 	✓
VHM [202] [link]	CLIP ViT-L/14 [86]	MLP	Vicuna-v1.5-7B [47]	VersaD [202] VersaD-Instruct [202] HnstD [202] VariousRS-Instruct [202]	<ul style="list-style-type: none"> Train vision encoder, MLP and LLM on VersaD Fine-tune MLP and LLM on VersaD-Instruct, HnStD and VariousRS-Instruct Load pre-trained weights of [47]–[49] Employ LoRA [50] for LLM Load pre-trained weights of [47], [86] Train MLP on a general image-language dataset Employ LoRA [50] for LLM Load pre-trained weights of [47], [49], [86] Fine-tune MLP and employ LoRA [50] for LLM Load pre-trained weights of [47], [86] Employ LoRA [50] for MLP and LLM Load pre-trained weights of [86], [214] Employ LoRA [50] for LLM 	✓
GeoChat [3] [link]	CLIP ViT-L/14 [86]	MLP	Vicuna-v1.5-7B [47]	Multimodal Instruction Dataset [3]	<ul style="list-style-type: none"> Load pre-trained weights of [47], [86] Employ LoRA [50] for LLM Load pre-trained weights of [47], [86] Train MLP on a general image-language dataset Employ LoRA [50] for LLM 	✓
RS-LLaVA [10] [link]	CLIP ViT-L@336 [86]	MLP	Vicuna-v1.5-7B/13B [47]	RS-Instructions [10]	<ul style="list-style-type: none"> Load pre-trained weights of [47], [86] Train MLP on a general image-language dataset Employ LoRA [50] for LLM 	✓
SkySenseGPT [11] [link]	CLIP ViT-L/14 [86]	MLP	Vicuna-v1.5 [47]	FIT-RS [11] Additional Instruction Dataset [11]	<ul style="list-style-type: none"> Load pre-trained weights of [47], [49], [86] Fine-tune MLP and employ LoRA [50] for LLM 	✓
IFShip [142]	CLIP ViT-L/14 [86]	MLP	Vicuna-13B [47]	TITANIC-FGS [142]	<ul style="list-style-type: none"> Load pre-trained weights of [47], [86] Employ LoRA [50] for MLP and LLM 	✗
TEOChat [213][link]	CLIP ViT-L/14 [86]	MLP	LLaMA 2 [57]	TEOchatas [213]	<ul style="list-style-type: none"> Load pre-trained weights of [86], [214] Employ LoRA [50] for LLM 	✓

[link] directs to model websites. Detailed information about training datasets can be found in Tables XII and XIII. Public refers to the availability of both code and model weights.

designing connectors to promote the alignment between the two modalities. Specifically,

Vision Encoder: Through a mask image modeling pre-text task, EVA [56] incorporates geometry and structure information into CLIP’s visual representations, leading to improved performances across a wide range of visual perception tasks. Consequently, RSGPT [1] and SkyEyeGPT [5] adopt EVA as their vision encoder. An alternative to complementing CLIP is utilizing the strengths of diverse vision encoders. CLIP learns visual representations through language supervision, which limits the information retained about the image. Captions only approximate the main content of images, often failing to present complex pixel-level details. Targeting this limitation, some works [6], [24], [136] propose combining vision encoders from CLIP and DINOv2 [67]. DINOv2 learns visual representation from images alone via self-supervised learning, enabling it to capture both image-level and pixel-level information. Moreover, given the varied object sizes in remote sensing images, these works refine visual representations by incorporating multi-scale information. In [6], a vision encoder with the convolutional neural network architecture is used to extract multi-scale visual features. In [24], [136], the input image is downsampled to different resolutions and then respectively fed into two vision encoders. The encoded visual features are transformed to the same dimension and concatenated channel-wise.

Vision-Language Connector: The linear layer and MLP are widely used as vision-language connectors, serving as

key components in most models [3], [5], [6], [10], [11], [142], [202], [213]. Differently, RSGPT [1] and LHRS-Bot [9] explore alternative connector architectures. Following InstructBLIP [46], RSGPT includes an instruction-aware query transformer (Q-Former) as an intermediate module between the vision encoder and LLM. The Q-Former is designed to extract task-relevant visual representations by interacting additional query embeddings with instruction and image embeddings. It achieves this through the attention mechanism: first, self-attention is applied to implement interaction between instruction and query embeddings, and then cross-attention is employed between query and image embeddings. The resulting output from the Q-Former, after passing through a linear layer, is then input into the LLM along with the instruction embeddings to generate responses. LHRS-Bot proposes incorporating multi-level image embeddings to sufficiently capture the semantic content of images. To mitigate the computational burden and the risk of overwhelming language information with excessive visual information, it introduces a set of learnable queries for each level of visual representation. These queries are used to summarize the semantic information of each level through stacked cross-attention and MLP layers. As a result, a dedicated visual perceiver is developed, with experimental results demonstrating that, compared to single-level visual representations and a two-layer MLP, multi-level representations paired with the visual perceiver improve the model’s performance in scene classification, visual question answering, and visual grounding tasks.

B. Training Strategy

Training instruction-based vision-language models typically involves two stages: pre-training for modality alignment and supervised fine-tuning (SFT) for following task-specific instructions.

Only SFT: Due to the lack of large-scale image-text datasets specifically designed for the remote sensing domain, most works target the SFT stage using carefully crafted instruction-following datasets [1], [3], [5], [11], [142], [213]. To preserve the general knowledge embedded in pre-trained vision encoders, the vision encoder is typically kept frozen during training, with the connector or the LLM undergoing fine-tuning. For instance, RSGPT [1] fine-tunes the connector, while GeoChat [3] and TEOChat [213] fine-tune the LLM. SkyEyeGPT [5], SkySenseGPT [11] and IFSHIP [142] fine-tune both the connector and LLM. To avoid the expense of full-parameter tuning, LoRA (Low-Rank Adaptation) [50] is often adopted, which introduces low-rank learnable matrices into the layers of the connector [5], [142] or LLM [3], [5], [11], [142], [213].

Pre-training followed by SFT: A couple of recent works have investigated how to implement the pre-training stage to boost model performance. Based on the choice of training data, they can be categorized into two groups: those that combine available image-text pairs from multiple domains for pre-training and those that direct attention toward creating large-scale RS image-text datasets. For combining available data, EarthGPT [6], Popeye [24] and RS-LLaVA [10] utilize natural image-text datasets, while EarthMarker [136] integrates data from both computer vision and remote sensing domains. COCO Caption [210] is a commonly used pre-training dataset. Unlike RS-LLaVA, which limits pre-training to the connector, the other three models perform pre-training for the connector and the LLM.

To address the domain gap between remote sensing images and natural images, researchers have developed large-scale RS image-text datasets, such as LHR-S-Align [9] and VersaD [202], both of which contain over one million training samples. Leveraging the LHR-S-Align dataset, Muhtar et al. [9] design a three-stage curriculum learning strategy for LHR-S-Bot. First, the vision perceiver is pre-trained on LHR-S-Align. Next, the vision perceiver and LLM are fine-tuned on LHR-S-Instruct subset and multi-task dataset, equipping the model to handle multimodal tasks. Finally, the LLM undergoes additional fine-tuning on all instruction data from LHR-S-Instruct, the multi-task dataset, and LLaVA-Instruct-158K subset to fully unlock LHR-S-Bot’s capabilities. Note that these models still keep the vision encoder frozen. In contrast, Pang et al. [202] unfreeze the vision encoder, MLP and LLM for pre-training on VersaD. They subsequently fine-tune the MLP and LLM on customized datasets, including VersaD-Instruct, HnStD and VariousRS-Instruct. Their experimental results show that the model pre-trained with RS data significantly outperforms the one only fine-tuned with RS data across multiple tasks, confirming the importance of incorporating extensive RS visual knowledge by pre-training.



Fig. 7. Examples of conversations between vision-language models and users.

C. Model Capability

Table V lists the capabilities of existing instruction-based vision-language models in remote sensing. Key observations include: 1) Most models are developed for general-purpose remote sensing data analysis [3], [5], [6], [9], [11], [136], [202], [213], with only Popeye [24] and IFSHIP [142] specifically tailored for remote sensing images of ships. These models primarily process optical images but have expanded to include SAR (Synthetic Aperture Radar) [6], [24] and infrared [6] images. 2) Most models support conversational interaction with users and can perform a variety of tasks [3], [5], [6], [9], [11], [136], [202], [213], ranging from single image analysis to temporal series analysis [213]. 3) The granularity of analyzed information has progressed from the image level to region level [11], [136], and even to the point level [136]. In this section, we examine the models’ performance on common capabilities and offer an in-depth exploration of unique capabilities.

Common Capabilities: As shown in Table V, most models are capable of performing remote sensing tasks such as image captioning, scene classification, visual question answering, and visual grounding. Accordingly, we present model performance on these tasks in Tables VI-VIII. To ensure a fair comparison, we report only the models’ performance on public datasets. Although multi-turn Conversation is also a common capability among existing models [3], [5], [6], [9], [11], [24], [136], [202], [213], it is challenging to evaluate quantitatively. Fig. 7 provides examples of conversations between various models and users.

(1) *Image Captioning:* Based on the input remote sensing image and the language instruction, vision-language models generate a description of image content. Examples of instruc-

TABLE V
CAPABILITY COMPARISONS OF INSTRUCTION-BASED VISION-LANGUAGE MODELS IN REMOTE SENSING.

Model	SIT Opt/SAR/IR	Cap. Img/Reg/Pt	CLS Img/Reg/Pt	VQA	VG	OD	ORR	MTC	Others
RSGPT [1]	✓ / X / X	✓ / X / X	X / X / X	✓	X	X	X	X	
SkyEyeGPT [5]	✓ / X / X	✓ / X / X	✓ / X / X	✓	✓	✓	X	✓	Video Captioning Referring Expression Generation
EarthGPT [6]	✓ / ✓ / ✓	✓ / ✓ / X	✓ / X / X	✓	✓	✓	X	✓	X
EarthMarker [136]	✓ / X / X	✓ / ✓ / ✓	✓ / ✓ / ✓	X	X	✓	✓	✓	X
Popeye [24]†	✓ / ✓ / X	✓ / X / X	X / X / X	X	X	✓	X	✓	X
LHRS-Bot [9]	✓ / X / X	✓ / X / X	✓ / X / X	✓	✓	X	✓	✓	Object Counting Object Attribute Recognition Image Property Recognition Object Counting Geometric Measurement Building Vectorizing Image Property Recognition Multi-Label Classification Honest Question Answering
VHM [202]	✓ / X / X	✓ / X / X	✓ / X / X	✓	✓	X	✓	✓	
GeoChat [3]	✓ / X / X	✓ / ✓ / X	✓ / X / X	✓	✓	X	X	✓	X
RS-LLaVA [10]	✓ / X / X	✓ / X / X	X / X / X	✓	X	X	X	X	
SkySenseGPT [11]	✓ / X / X	✓ / ✓ / X	✓ / X / X	✓	✓	✓	✓	✓	Multi-Label Classification Image/Region Scene Graph Generation
IFShip [142]†	✓ / X / X	✓ / X / X	✓ / X / X	✓	X	X	X	✓	X Change Detection Temporal Scene Classification Temporal Referring Expression Spatial Change Referring Expression Image/Region Change Question Answering
TEOChat [213]	✓ / X / X	✓ / ✓ / X	✓ / X / X	✓	✓	X	X	✓	

† indicates the model's capability tailored for remote sensing images of ships. *SIT*, *Cap.*, *CLS*, *VQA*, *VG*, *OD*, *ORR*, and *MTC* are short for *Supported Image Type*, *Captioning*, *Classification*, *Visual Question Answering*, *Visual Grounding*, *Object Detection*, *Object Relationship Reasoning*, and *Multi-Turn Conversation*, respectively. We present the *Opt/SAR/IR* column to indicate whether the model supports input of optical (Opt), synthetic aperture radar (SAR), or infrared (IR) images, and the *Img/Reg/Pt* column to denote the granularity of the analyzed information, which includes image level (Img), region level (Reg), or point level (Pt).

tions include “Describe this image in detail.” [1], [5] or “Please provide a one-sentence caption for the provided remote sensing image in detail.” [6]. From Table VI, we can observe that SkyEyeGPT [5] outperforms RSGPT [1] on the RSICD [60], UCM-captions [62] and Sydney-captions [62] datasets in terms of BLEU [225], METEOR [226], and ROUGE_L [227], though it falls short of RSGPT in CIDEr [228] scores. On the UCM-captions dataset, RS-LLaVA achieves BLEU scores comparable to SkyEyeGPT while surpassing it in all other metrics. On the NWPU-Captions [63] dataset, EarthGPT [6] achieves state-of-the-art results, with a remarkable improvement in CIDEr (nearly 30%) over EarthMarker [136].

(2) *Scene Classification*: Table VII presents evaluations on 10 remote sensing scene classification datasets under both fully supervised and zero-shot settings. The absorption of extensive remote sensing visual knowledge improves model accuracy and generalization in scene classification tasks. Specifically, in the supervised setting, VHM [202] and EarthGPT [6] boost classification performance to over 93% on the NWPU-RESISC45 [53] dataset, which includes 45 scene categories with image spatial resolutions ranging from 0.2m to more than 30m. For zero-shot scene classification, LHRS-Bot [9], VHM, and SkySenseGPT [11] demonstrate impressive generalization, achieving accuracy above 91% on the AID [102] dataset and over 93% on the WHU-RS19 [71] dataset. SkySenseGPT notably achieves the highest accuracy, reaching 92.25% and 97.02% on these datasets, respectively. Despite these remarkable results, low-resolution and fine-grained scene

classification remain significant challenges. The EuroSAT [69] dataset, used for land use and land cover classification, has an image spatial resolution of 10m, while fMoW [117] contains over 1 million images across 63 scene categories. On these datasets, LHRS-Bot [9] achieves an accuracy below 57%.

(3) *Visual Question Answering*: RSVQA-HR [54] and RSVQA-LR datasets are widely adopted to assess model performance in visual question answering tasks, as shown in Table VII. Under a supervised setting, RSGPT [1] establishes strong performance baselines on both datasets. Subsequent models, including EarthGPT [6], GeoChat [3], VHM [202], and SkySenseGPT [11], have progressively approached these benchmarks in both supervised and zero-shot settings. The latest model, SkySenseGPT, narrows the zero-shot performance gap to approximately 13%, coming closest to the baseline on test set 2 of the RSVQA-HR dataset and surpassing RSGPT on the RSVQA-LR dataset in the supervised setting.

(4) *Visual Grounding*: aims to locate specific objects within an image based on a natural language expression. Vision-language models accomplish this task by providing the coordinates of the target object. Table VIII presents the performance of several models on RSVG [66] and DIOR-RSVG [33] datasets, using an intersection over union (IoU) threshold of 0.5 adopted as the evaluation metric. LHRS-Bot [9] achieves the highest score on the RSVG test set, while SkyEyeGPT [5] leads on DIOR-RSVG test set. Due to differences in metric calculations, VHM [202] achieves a score of 57.17% on DIOR-RSVG.

TABLE VI
PERFORMANCE OF INSTRUCTION-BASED VISION-LANGUAGE MODELS ON IMAGE CAPTIONING.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
	RSICD [60]						
RSGPT [1]	70.32	54.23	44.04	36.83	30.10	53.34	102.94
SkyEyeGPT [5]	86.71	76.66	67.31	59.99	35.35	62.63	83.65
	UCM-captions [62]						
RSGPT [1]	86.12	79.14	72.31	65.74	42.21	78.34	333.23
RS-LLaVA [10]	90.00	84.88	80.30	76.03	49.21	85.78	355.61
SkyEyeGPT [5]	90.71	85.69	81.56	78.41	46.24	79.49	236.75
	Sydney-captions [62]						
RSGPT [1]	82.26	75.28	68.57	62.23	41.37	74.77	273.08
SkyEyeGPT [5]	91.85	85.64	80.88	77.40	46.62	77.74	181.06
	NWPU-Captions [63]						
EarthMarker [136]	84.40	73.10	62.90	54.30	37.50	70.00	162.90
EarthGPT [6]	87.10	78.70	71.60	65.50	44.50	78.20	192.60

BLEU, *METEOR*, *ROUGE_L*, and *CIDEr* are short for *BiLingual Evaluation Understudy* [225], *Metric for Evaluation of Translation with Explicit ORdering* [226], *Recall-Oriented Understudy for Gisting Evaluation* [227], *Consensus-based Image Description Evaluation* [228].

Unique Capabilities: Current research seeks to develop versatile vision-language models capable of handling various remote sensing image analysis tasks in a conversational manner. For instance, EarthMarker [136] supports not only language instructions but also visual prompts (*e.g.* boxes and points), enabling the model to perform fine-grained image understanding, such as region/point-level captioning and referring object classification. TEOChat [213] excels in analyzing time-series remote sensing images, detecting changes of interest (in the form of the bounding box), and answering questions related to changes. Coincidentally, LHRs-Bot [9] and VHM [202] both feature the ability to qualitatively recognize object attributes (*e.g.* color) and image properties (*e.g.* resolution and modality), while also quantitatively counting objects. Additionally, VHM offers insights into model honesty, which is vital for applications such as national defense security. Beyond single-object analysis, relationship analysis between objects is increasingly recognized as essential for understanding complex remote sensing scenes, drawing attention in recent models [9], [11], [136], [202]. This section delves into these unique capabilities, with a focus on their implementations as detailed.

(1) *Fine-grained Image Understanding:* Region-level image understanding is challenging but achievable. One can include the coordinates of the target region in language instructions to direct the model’s attention to a specific local region, as demonstrated in EarthGPT [6] and GeoChat [3]. While using precise coordinates is effective, it lacks flexibility and is challenging to extend to the point level. Differently, EarthMarker [136] leverages visual prompting marks to guide the model to interpret specific regions or points. To ensure the model understands the relationship between visual prompts and the whole image, the visual prompts share the same visual encoder and projection layer with the input image. Their embeddings are combined with instruction embeddings before being fed into the LLM. Visual Prompts allow EarthMarker to perform multi-granularity RS image interpretation at the image, region, and point levels.

(2) *Time-Series Image Analysis:* Change detection, the pro-

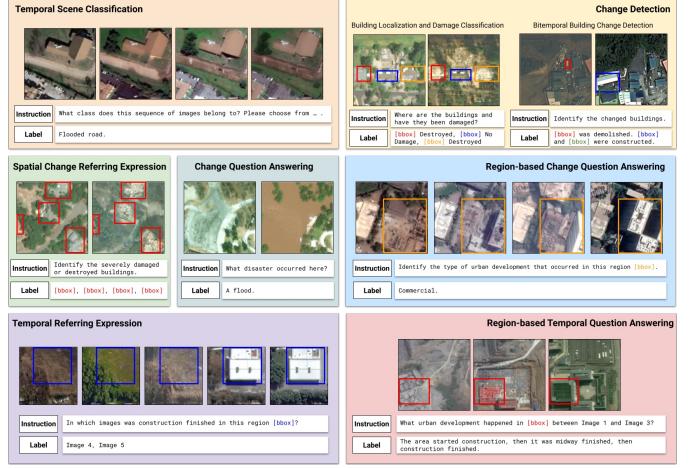


Fig. 8. Examples showcasing the capabilities of TEOChat [213] in time-series image analysis.

cess of identifying changes on the Earth’s surface from sequences of remote sensing images captured over the same area, plays an important role in many applications such as urban planning [230] and war damage assessment [229]. The work TEOChat [213] pioneers change detection in a conversation manner. Given a sequence of remote sensing images, TEOChat employs a shared vision encoder to generate embeddings for each image. These embeddings are then projected to the input space of the LLM via a 2-layer MLP. The LLM processes these projected image embeddings alongside instruction embeddings to produce change detection results in the form of bounding boxes. To train TEOChat, the authors of [213] develop TEOChatlas, the first instruction-following dataset tailored for time-series image analysis tasks. Leveraging this dataset, TEOChat extends its capabilities beyond change detection to include tasks such as change question answering, temporal scene classification, temporal referring expression, and spatial change referring expression. Fig. 8 provides specific examples illustrating each of these tasks.

(3) *From Qualitative Recognition to Quantitative Analysis:* Existing models mostly concentrate on qualitative recognition tasks, excelling at describing the attributes or categories of images and objects in a non-numerical manner. In contrast, LHRs-Bot [9] and VHM [202] broaden their capabilities to include quantitative image analysis, answering “how many” questions. LHRs-Bot is equipped to count objects in an image and estimate image resolution, using single-choice questions with 2 to 4 candidate answers. This question format requires a certain level of user expertise, as the user needs to predefined several candidate answers, one of which must be correct. VHM goes a step further by not only handling these tasks but also measuring object size. Unlike LHRs-Bot, all quantitative tasks in VHM adopt open-ended question formats, which may be more suitable for practical application. On the test set of the VariousRS-Instruct [202], VHM achieves mean absolute errors of 0.24 for mage resolution estimation, 6.75 for object counting, and 12.82 for geometric measurement. These results highlight the potential of vision-language models to advance quantitative analysis in remote sensing images. The

TABLE VII
PERFORMANCE OF INSTRUCTION-BASED VISION-LANGUAGE MODELS ON SCENE CLASSIFICATION AND VISUAL QUESTION ANSWERING.

Model	Scene Classification: Top-1 Accuracy									
	RESISC [53]	CLRS [99]	NaSC-TG2 [92]	UCM [70]	AID [102]	WHU-RS19 [71]	SIRI-WHU [94]	EuroSAT [69]	METER-ML [224]	fMoW [117]
GeoChat [3]	-	-	-	84.43†	72.03†	-	-	-	-	-
EarthMarker [136]	-	-	-	86.52†	77.97†	-	-	-	-	-
EarthGPT [6]	93.84	77.37†	74.72†	-	-	-	-	-	-	-
LHRS-Bot [9]	83.94	-	-	-	91.26†	93.17†	62.66†	51.40†	69.81	56.56
VHM [202]	94.54	-	-	-	91.70†	95.80†	70.88†	-	72.74	-
SkySenseGPT [11]	-	-	-	-	92.25†	97.02†	74.75†	-	-	-

Model	Visual Question Answering									
	RSVQA-HR [54] Test Set 1			RSVQA-HR [54] Test Set 2			RSVQA-LR [54]			
Model	Presence	Comparison	Avg. Acc.	Presence	Comparison	Avg. Acc.	Presence	Comparison	Rural/Urban	Avg. Acc.
SkyEyeGPT [5]	84.95	85.63	85.29	83.50	80.28	81.89	88.93	88.63	75.00	84.19
RSGPT [1]	91.86	92.15	92.00	89.87	89.68	89.78	91.17	91.70	94.00	92.29
LHRS-Bot [9]	-	-	-	-	-	-	88.51	90.00	89.07	89.19
RS-LLaVA [10]	-	-	-	-	-	-	92.27	91.37	95.00	88.10
GeoChat [3]	-	-	-	58.45†	83.19†	72.30†	91.09	90.33	94.00	90.70
EarthGPT [6]	-	-	-	62.77†	79.53†	72.06†	-	-	-	-
VHM [202]	-	-	-	64.00†	83.50†	73.75†	90.11	89.89	88.00	89.33
SkySenseGPT [11]	-	-	-	69.14†	84.14†	76.64†	91.07	92.00	95.00	92.69

RESISC, Avg. Acc. are short for NWPU-RESISC45 [53] and Average Accuracy, respectively. † indicates zero-shot performance.

TABLE VIII
PERFORMANCE OF INSTRUCTION-BASED VISION-LANGUAGE MODELS ON VISUAL GROUNDING.

Model	RSVG [66]		DIOR-RSVG [33]	
	Val	Test	Test	Test
SkyEyeGPT [5]	69.19	70.50	88.59	
EarthGPT [6]	-	-	76.65	
LHRS-Bot [9]	-	73.45	88.10	
VHM [202]	-	-	56.17	

Evaluation metric adopts Acc@0.5, which means that the intersection over union (IoU) of the predicted box is at least 0.5 with the ground truth bounding box.

successful implementation of such analysis relies heavily on the development of specialized instruction-following datasets, which are discussed in detail in Section VI-B.

(4) *Endowing Models with Honesty*: Most instruction-following datasets designed for remote sensing tasks exclusively contain factual questions that query real objects within the images and are accompanied by affirmative answers. This makes vision-language models susceptible to lying, such as providing an affirmative answer regarding a category of a non-existent object. A dishonest vision-language model is particularly problematic in remote sensing applications, especially in critical areas like national defense security, as it may lead to serious negative consequences, including inaccurate intelligence gathering. Given this, Pang et al. [202] are the first to offer insights into the honesty of models. They create HnStD, an RS-specific honest dataset comprising questions with factual and deceptive categories. By utilizing it as an additional instruction-following dataset, VHM is endowed with honesty. As shown in Fig. 9, VHM refuses to answer questions regarding the color or position of non-existent objects.

(5) *Object Relationship Understanding*: Remote sensing images often encompass a vast number of objects, owing to their long-distance imaging. Comprehensively understanding the relationships between objects within an image is essential for interpreting complex remote-sensing scenes [231], [232]. Consequently, the latest models, i.e. LHRS-Bot [9],

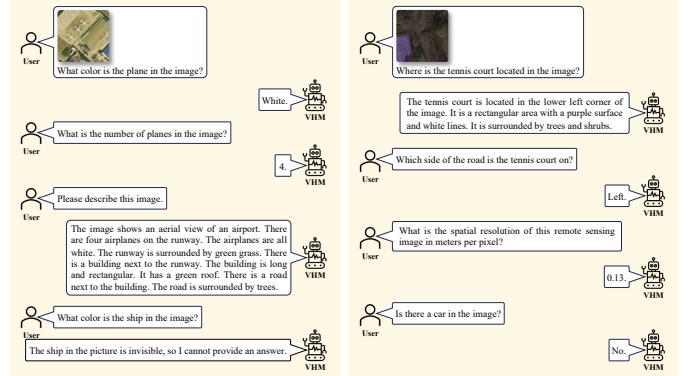


Fig. 9. Conversations between users and VHM [202]

VHM [202], EarthMarker [136] and SkySenseGPT [11], extend their capabilities to understand relationships between objects in an image. Specifically, LHRS-Bot and VHM target spatial relationships, using simple rules like “top” or “top right corner” to indicate object relationships. EarthMarker, on the other hand, explores functional relationships, such as the “overpass” and “toll station” potentially contributing to the transportation infrastructure around the “stadium”. Meanwhile, SkySenseGPT describes semantic relationships, for example, the “crane” is over the “boat”. This promising understanding of relationships benefits from the design of specialized instruction-following datasets, as discussed in Section VI-B.

V. GENERATION-BASED VISION-LANGUAGE MODELING

Similar to research on contrastive-based vision-language modeling, generation-based vision-language modeling follows two major research directions: the construction of foundation models concerning the characteristics of remote sensing images, and their application to promote various remote sensing data analysis tasks. This section first presents the development of *generative foundation models* and then discusses representative *downstream applications*.

TABLE IX
SUMMARY OF GENERATION-BASED VISION-LANGUAGE FOUNDATION MODELS IN REMOTE SENSING.

Model	Diffusion Model	Condition Encoder	Text	Condition	Training Dataset	GenImgType	Public
				Metadata	Image		
RS-SD [7] [link]	Stable Diffusion [203]	CLIP Transformer [86]	✓	x	x	1% RS5M [7]	Optical ✓
DiffusionSat [38] [link]	Stable Diffusion [203]	CLIP Transformer [86] Sinusoidal Projection+MLP 3D ControlNet [38]	✓	Latitude Longitude GSD Cloud Cover Imaging Time	Satellite Image	fMoW [117] Satlas-small [119] SpaceNet [130]	Optical ✓
CRS-Diff [39] [link]	Stable Diffusion [203]	CLIP Transformer [86] MLP ControlNet [234] FFN	✓	Latitude Longitude GSD Cloud Cover Imaging Time	HED MLSD Depthmap Sketch Road Map Seg. Mask Deep Feature	RSICD [60] fMoW [117] Million-AID [100]	Optical ✓
GeoSynth [223] [link]	Stable Diffusion [203]	CLIP Transformer [86] SatCLIP Location Encoder [35] CoordNet [223] ControlNet [234]	✓	Latitude Longitude	OSM Image	Satellite-OSM Dataset [223]	Optical ✓
MetaEarth [208]	DDPM [235]	Sinusoidal Projection+MLP RRDBNet [236]+Upsampling CLIP Transformer [86]	x	Resolution	Low-Resolution Image	Multi-Resolution Dataset [208]	Optical ✗
Text2Earth [293]	Stable Diffusion [203]	Projection Layer VAE Encoder [204]	✓	Resolution	Masked Image	Git-10M [293]	Optical ✗
HSIGene [207] [link]	Stable Diffusion [203]	CLIP Transformer [86] ControlNet [234] FFN	✓	x	HED MLSD Sketch Seg. Mask Deep Feature	Xiongan [219] Chikusei [218] DFC2013 [221] DFC2018 [222] Heihe [220]	Hyperspectral ✓
GPG2A [206] [link] RSDiff [145]	Improved DDPM [146] Imagen [147]	CLIP Transformer [86] ControlNet [234] T5 [237]	✓	x	Ground Image	VIGORv2 [206]	Optical ✓
			✓	x	x	RSICD [60]	Optical ✗

[link] directs to model websites. GenImgType refers to the type of generated image, e.g. optical or hyperspectral images. GSD, HED, MLS, Seg. Mask, and OSM stand for Ground Sampling Distance, Holistically-nested Edge Detection, Multiscale Line Segment Detection, Segmentation Mask, and OpenStreetMap, respectively. Public refers to the availability of both code and model weights.

A. Generative Foundation Models

Building an effective generative foundation model is a formidable task because one needs to consider how to ensure the reliability and diversity of the generated images.

Enhancing Reliability: Text, often in the form of descriptions of ground objects within images [7], [39], [145], [206], [223], [293] or semantic categories of images [38], [207], has been a common condition for conditional image generation. For instance, the text “a satellite view of San Francisco showing the bay, a street and a bridge” is used to guide RS-SD [7], while the text “a fmow satellite image of a car dealership in the United States of America” constrains DiffusionSat [38]. However, it is evident that these textual descriptions alone struggle to fully encapsulate the variety of objects and intricate relationships present within a satellite image. The lack of sufficient constraint information poses a challenge to generating reliable images. To address this challenge, additional conditions, such as metadata or images, are increasingly utilized to constrain the generation process.

(1) *Metadata:* Metadata such as latitude, longitude, ground sampling distance, cloud cover, and imaging time (year, month, and day) are adopted in both DiffusionSat [38] and CRS-Diff [39]. In contrast, MetaEarth [208] and Text2Earth [293] focus on spatial resolution, and GeoSynth [223] on geographic location (latitude and longitude). Compared to text conditions, metadata is more easily available, as it is inherently embedded within remote sensing images. Furthermore, it allows generative foundation models to be trained on large-scale image datasets, benefiting from the diverse geographic distribution of these datasets. Consequently, the key problem lies in addressing how to inject metadata conditions into diffusion models. Identical to the

encoding of diffusion timesteps, DiffusionSat and MetaEarth process metadata values through sinusoidal encoding, followed by MLPs. The resulting metadata embeddings are added with timestep embeddings before being fed into the diffusion model. CRS-Diff maps metadata values to a fixed range and encodes them using different MLPs. The metadata embeddings are concatenated with text and content embeddings, providing global control information. For latitude and longitude, an alternative approach is to utilize a pre-trained location encoder. An example is [223] where the authors employ SatCLIP [35] to extract location embeddings and use CoordNet, which takes location and timestep embeddings as input, to integrate these conditions into the diffusion model. CoordNet consists of 13 multi-head cross-attention blocks, each including a zero-initialized feed-forward layer.

(2) *Image:* Metadata, such as geographic location, determines the visual appearance of objects in generated images to some extent. For example, the architectural styles of Chinese and European buildings are notably distinct. However, metadata primarily imposes constraints at a macro level, allowing considerable flexibility in the generated objects, which can result in uncontrollable object shapes. As a result, recent works have investigated the use of image-form conditions to enable more precise control over the image generation process. Based on the information conveyed by the image, image conditions are roughly split into low-level visual conditions and high-level semantic conditions. Low-level visual conditions pertain to the geometric information of images, such as edges, line segments, and sketches, which constrain the shape and structure of objects in generated images, as demonstrated in works [39], [207]. Off-the-shelf algorithms or models, including HED [238], LETR [239] and the model proposed

in [240], are employed to obtain these image conditions.

High-level semantic conditions, as the name implies, refer to the semantic information of images, providing constraints on object categories and their relationships in generated images. These conditions can be further categorized into two sub-groups. The first group [38], [206], [208], [293] conditions on associated remote sensing images, exemplified by works such as MetaEarth [208], which uses low-resolution images to prompt the diffusion model to produce high-resolution images, and GPG2A [206], which synthesizes aerial images guided by layout maps derived from corresponding ground images. The second group [39], [207], [223] leverages abstract representations of image content, as seen in works like CRS-Diff [39] and GeoSynth [223]. CRS-Diff utilizes road maps, segmentation masks, and deep features to constrain the layout and objects within generated images, while GeoSynth employs OpenStreetMap (OSM) images to achieve similar constraints on layout and object placement.

Image-form conditions are typically injected into diffusion models using ControlNet [234]. ControlNet replicates the encoder blocks and the middle block of Stable Diffusion's U-Net and incorporates several zero convolutions. It processes both conditioning representations and noisy latent representations as inputs, with its outputs added to the decoder blocks and middle block of Stable Diffusion's U-Net. In GeoSynth [223] and GPG2A [206], ControlNet is used for condition injection with different inputs: GeoSynth provides layout maps derived from ground images, while GPG2A inputs OSM images, text, and diffusion timesteps. To enhance multi-condition injection, DiffusionSat [38] extends ControlNet into a 3D version capable of accepting a sequence of satellite images. This 3D ControlNet retains the replicated encoder and middle blocks, with each followed by a temporal layer consisting of a 3D zero-convolution and a temporal, pixel-wise transformer. Following Uni-ControlNet [241], the authors in [39], [207] perform multi-scale condition injection and use attentional feature fusion [242] to combine conditioning and latent representations. Unlike these works relying on ControlNet, MetaEarth and Text2Earth directly concatenate latent and image condition representations. MetaEarth encodes low-resolution images via RRDBNet [236] with upsampling and convolution layers, while Text2Earth processes masked images using a VAE encoder [204].

Improving Diversity: Herein, diversity refers to two aspects: first, the generated objects exhibit a variety of semantic categories, with each category featuring varied visual characteristics; and second, the generated images capture a broad range of variations in imaging condition (*e.g.* season and illumination) and imaging sensors (*e.g.* spatial resolution and viewpoint). By conditioning on metadata, the diversity of the generated data is improved as shown in Fig. 10. DiffusionSat [38] modifies geographic coordinates, imaging time, and ground sampling distance, resulting in varied stadiums, different scenes, and multi-resolution images, respectively. In addition to considering the perspective of conditions, some works [145], [208] explore new frameworks for generating multi-resolution images. In [145], cascaded diffusion models [147] are employed, where one model gen-

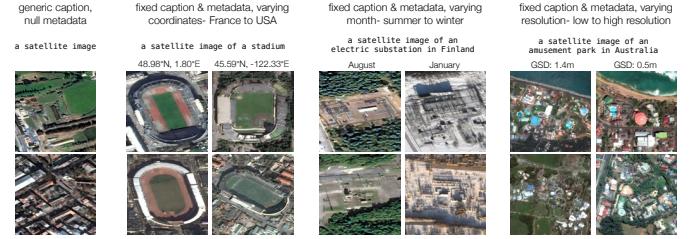


Fig. 10. Examples of remote sensing images generated by DiffusionSat [38].

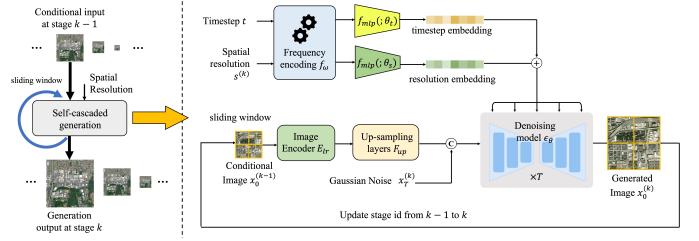


Fig. 11. Illustration of MetaEarth architecture [208].

erates low-resolution images conditioned on text embeddings, and the other super-resolves these low-resolution images. This framework achieves the gradual generation of high-resolution satellite images purely from textual descriptions. In [208], a resolution-guided self-cascading framework is proposed to generate images at various resolutions, instead of being limited to the two resolutions described in [145]. As illustrated in Fig. 11, the generation process unfolds over multiple stages, with each stage conditioned on the low-resolution image output from the preceding stage and its corresponding spatial resolution. The resolution increases at a fixed ratio with each stage. Considering the resolution gap between adjacent stages, the image condition is injected through a sequential process involving an image encoder, followed by several upsampling and convolution layers. Building on this framework, the authors introduce sliding windows with overlaps and a noise sampling strategy to generate continuous, unbounded scenes.

Performance Evaluation: With these efforts, existing generative foundation models have demonstrated the capability to generate optical remote sensing images [7], [38], [39], [145], [206], [208], [223] or hyperspectral remote sensing images [207]. Their performance is typically evaluated using metrics such as Fréchet Inception Distance [243] and Inception Score [244]. For text-conditioned models, CLIP [86] is employed to measure the similarity between generated images and their corresponding textual descriptions, as detailed in Table X. In addition to the direct evaluations, some works assess model performance by applying them or the generated images to specific downstream tasks and measuring their impact [38], [39], [206]–[208]. For example, DiffusionSat [38] is applied to tasks such as super-resolution and temporal generation, and in-painting, while CRS-Diff [39] uses its generated images to augment training datasets of the road detection model.

B. Downstream Applications

Generative foundation models serve as powerful and versatile tools for advancing various remote sensing tasks. Their

TABLE X
PERFORMANCE OF GENERATIVE FOUNDATION MODELS.

Model	Condition	Dataset	FID	IS	CLIP-Score
DiffusionSat [38]	Text+Metadata	fMoW [117]	15.80	6.69	17.20
	Text		50.72	18.39	20.33
	HED		30.18	-	-
	MLSD		55.75	-	-
CRS-Diff [39]	Depthmap	RSICD [60]	54.40	-	-
	Sketch		68.29	-	-
	Seg. Mask		70.05	-	-
	Road Map		94.16	-	-
GeoSynth [223]	Deep Feature	Satellite-OSM [223]	44.93	-	-
	Coordinate+OSM Image		11.90	-	30.30
	Text+OSM Image		12.97	-	29.80
	MLSD		-	-	79.27
HSIGene [207]	HED+MLSD	AID [102]	-	-	81.52
	HED+MLSD+Sketch		-	-	81.40
RSDiff [145]	HED+MLSD+Sketch+Seg. Mask	RSICD [60]	-	-	81.83
	Text		66.49	7.22	-
Text2Earth [293]	Text	RSICD [60]	24.49	-	-

FID and *IS* stand for *Fréchet Inception Distance* [243], *Inception Score* [244]. *CLIP-Score* [86] measures the similarity between generated images and corresponding textual descriptions. *HED*, *MLSD*, *Seg. Mask*, and *OSM* are short for *Holistically-nested Edge Detection*, *Multiscale Line Segment Detection*, *Segmentation Mask*, and *OpenStreetMap*, respectively.

exceptional image generation capabilities have been utilized to tackle challenges such as limited image availability [250] and the high costs associated with annotation [251], [252]. Additionally, the conditionally controllable nature of these models makes them particularly suitable for remote sensing image enhancement tasks. Examples include tasks like super-resolution [245], [246], where low-resolution images serve as conditions, and cloud removal [253], [254], where SAR images are used as conditions. Moreover, their advantages in learning diverse remote sensing image distribution equip them to handle the complexity of remote sensing scenes, improving the accuracy of remote sensing image interpretation tasks such as change detection [247], [248] and land cover classification [249]. These advancements predominantly center on the ability of generative foundation models to model image distribution, which has been comprehensively reviewed in [205]. In this discussion, however, we shift the focus to recently emerging applications that show how the joint distribution between images and texts in generative foundation models promotes advancements in remote sensing tasks. This includes tasks such as image or change captioning [150], [154], pansharpening [148], and zero-shot target recognition [155], as illustrated in Table XI. This section delves into representative implementations in conjunction with specific tasks.

Captioning: Recent works [150], [154] propose generating textual descriptions conditioned on visual features, rather than generating images from textual descriptions, thus applying generative models for captioning remote sensing images. In [150], the authors focus on extracting both global and local features of remote sensing images to enhance the visual conditions for the generative model, addressing challenges such as intra-class diversity, inter-class similarity, and varying object sizes. They then improve the interaction between noisy representations and visual conditions in the decoder by sequentially interacting with global and local conditions, aiming for generated captions that better align with the image content. In [154], difference features of bi-temporal images condition the generative model to generate change captions from standard Gaussian noise. These difference features are integrated

into the generative model using cross-attention, followed by stacking self-attention. Inspired by the success of generative models in change detection [261], the authors in [255] employ these models as a powerful feature extractor for multi-level, multi-timestep features of bi-temporal images. A difference encoder, based on a time-channel-spatial attention mechanism, is then utilized to extract discriminative information from the features of bi-temporal images, and a decoder generates change captions from this information.

Pansharpening: is a process of fusing high-resolution panchromatic images with low-resolution multispectral images to produce high-resolution multispectral images. The idea of applying generative foundation models to this task involves formulating it as an image generation problem conditioned on panchromatic and multispectral images. To enhance the generalizability of the pansharpening model and enable it to uniformly handle multispectral images from different satellites, TMDiff [148] incorporates textual descriptions, such as “Text Prompts of GaoFen-2 Satellite”, as identifiers to specify the satellites that capture multispectral images. These descriptions, combined with panchromatic-multispectral image pairs, jointly guide the generation process. Specifically, the textual descriptions are processed by a physics-informed HyperNet to get text embeddings, which are then used to modulate the denoising network. Meanwhile, the image pair is injected into the decoder of the denoising network via a condition encoder that shares the same architecture as the network’s encoder. To address the variability in spectral band configurations among different satellites, TMDiff, built on the U-Net architecture, organizes each block with stacked modulated ResBlock, swish activation layer, and frequency-aware downsampling (or upsampling). Compared to PanDiff [257], which relies solely on image conditions, TMDiff demonstrates impressive generalization capabilities.

Zero-shot SAR Target Recognition: Due to the high costs associated with SAR imaging and the limited azimuth range for capturing images, SAR datasets are often small, making it essential to explore SAR target recognition in zero-shot settings. From the perspective of SAR image simulation, adopting generative foundation models for this task is a natural solution. Typically, Wang et al. [155] employ generative models to create 3D models of targets conditioned on semantic information, such as category and structure of targets. Since most existing generative models are trained on RGB natural images or optical remote sensing images, the authors adopt a two-step process: they first use the generative model to generate optical images based on the target semantic information, and then transform these optical images into 3D models using TripoSR [258], rather than directly generating 3D models or SAR images.

Cloud Removal: Passive remote sensing is susceptible to cloud interference. By conditioning on cloud-contaminated images and SAR images, generative models have been employed for cloud removal of remote sensing images [253], [254]. Such solutions require geographically aligned multimodal image pairs, which are challenging to obtain in practical applications. To overcome this limitation, Czernawski et al. [259] propose a novel approach that uses historical edge information

TABLE XI
SUMMARY OF GENERATION-BASED VISION-LANGUAGE FOUNDATION MODELS APPLIED TO REMOTE SENSING TASKS.

Work	Task	Generative Foundation Model	Adaptation
VCC-DiffNet [150]	Image Captioning	D3PM [151]	Extract local and global features to enhance visual conditions, Enhance interactions between noisy representations with visual condition by sequentially interacting global and local conditions. Design cross-mode feature fusion module to inject difference features of bi-temporal images,
Diffusion-RSCC [154]	Change Captioning	Self-Design	Design stacking self-attention module to recreate change captions from Gaussian noise iteratively.
MADiffCC [255]	Change Captioning	SR3 [256], [261]	Use a generative model to extract multi-level, multi-timestep features of bi-temporal images, Design an encoder using the time-channel-spatial attention to obtain discriminative information, Design a decoder guided by gated multi-head cross-attention for generating change captions.
TMDiff [148]	Pansharpening	SR3 [256]	Use text and panchromatic-multispectral image pairs together to condition the generation process, enhancing generalization.
Wang et al. [155]	Zero-shot SAR Target Recognition	Stable Diffusion [203]	Generate optical images conditioned on target semantic information, aiding in the creation of 3D SAR target models.
Czerkawski et al. [259]	Cloud Removal	Stable Diffusion [203]	Use text and historical edge information to constrain cloud-free image generation.
UP-Diff [149]	Urban Prediction	Stable Diffusion [203]	Use text, current urban layouts and planned change maps as conditions, guiding the model to generate future urban layouts.

and textual descriptions (*e.g.* a cloud-free satellite image) to constrain the generation process. Experimental results suggest that general-purpose generative models may not be directly applicable for cloud removal of remote sensing images, as they tend to generate undesirable artifacts in cloud-contaminated areas.

Urban Prediction: aims to forecast future urban layouts based on current urban layouts and planned change maps, providing support for urban planning. The pioneering work, UP-Diff [149], incorporates ConvNeXt [260] to encode current urban layouts and planned change maps into embeddings, which are subsequently injected into the decoder of diffusion models through cross-attention layers. Additionally, the text condition, derived from encoding “A remote sensing photo”, guides the model to generate outputs in a satellite image style.

VI. DATASETS

Large-scale datasets are an essential prerequisite for vision-language research under the two-stage paradigm. Thus, a considerable amount of current research [1], [7]–[9], [11], [12], [22], [156], [202] focuses on dataset construction. In this section, we summarize these efforts, detailing the properties of datasets and their creation methods, with the aim of providing both convenience and inspiration for further research. Based on their usage, existing datasets can be broadly categorized into three groups: *pre-training datasets*, *instruction-following datasets*, and *benchmark datasets*.

A. Pre-training Datasets

Pre-training datasets, which consist of remote sensing images and their corresponding texts, play a crucial role in infusing the model with a broad range of visual and language concepts. This requires the images in pre-training datasets to be sufficiently diverse and rich. Currently, there are two alternative sources for collecting images: one is combining

various open-source remote sensing image datasets [1], [2], [7], [22], [29], [202], [293], and the other is utilizing public geographic databases [8], [9], [31], [143], [293]. Once the images are collected, corresponding textual descriptions can be generated through manual annotation [1], [31]. Although manual annotation is easy to implement and highly accurate, the complexity and diversity of remote sensing images make it costly, which can significantly limit the dataset size. As a result, there is a growing shift toward automatic image captioning using rule-based methods [2], [8] or off-the-shelf models [7], [9], [22], [143], [202], [293]. This section begins by presenting the image collection strategies of existing pre-training datasets, followed by an introduction to their caption generation methods.

Image Collection: Object detection datasets in remote sensing typically feature diverse ground objects, such as DOTA [45] containing 1,793,658 instances across 18 categories, and DIOR [51] including 192,472 instances spanning 20 categories. These datasets have quickly garnered attention and have been instrumental in constructing pre-training datasets, leading to the development of RSICap [1] and DIOR-Captions [29]. Compared to relying on a single dataset, combining multiple datasets further enriches the diversity of images and objects while significantly increasing the dataset size. For instance, RET-3+SEG-4+DET-10 [2] integrates three datasets for image retrieval, four for image segmentation, and ten for object detection. VersaD [202] further emphasizes multi-resolution and multi-domain coverage, drawing from datasets such as CVUSA [264] (0.08m resolution), Million-AID [100] (0.5m to 153m resolution, covering diverse geographies), and LoveDA [126] (featuring urban and rural environments). RS5M [7] prioritizes large-scale image datasets, incorporating sources like BigEarthNet [113] (590,326 images), fMoW [117] (1,047,691 images), and Million-AID (over 1 million images). Moreover, RS5M integrates PUB11, a dataset of over 3 million remote sensing-related image-text pairs

TABLE XII
SUMMARY OF PRE-TRAINING DATASETS FOR VISION-LANGUAGE MODELING IN REMOTE SENSING.

Dataset	#Pairs	Image Source		Image Size	Image Resolution (m)	Caption Generation	#Captions Per Image	Avg. Cap. Length	Public	
RSICap [1] [link]	2,585	DOTA-v1.5 [45]		512×512	-	Manual Annotation	1	60	✓	
DIOR-Captions [29]	16,565	DIOR [51]		800×800	-	Manual Annotation	2	11	✗	
ChatEarthNet [22] [link]	173,488	SatlasPretrain [119]		256×256	10	ChatGPT-3.5 [267] ChatGPT-4V [267]	1~2	155 / 90	✓	
		AUAIR [120]	CARPK [121]	DIOR [51]						
		DOTA [45]	HRRSD [76]	HRSC [122]						
		iSAID [125]	LEVR [123]	LoveDA [126]	224×224					
		Potsdam [127]	RSICD [60]	RSITMD [61]	~1920×1080					
		RSOD [96]	Stanford [124]	UCM [70]		M2C+B2C Generation [2]	5	-	✓	
		Vaihingen [128]	VisDrone [78]							
		CrowdAI [262]	CVACT [263]	CVUSA [264]	512×512	0.08~153	Gemini-Vision [265]	1	369	✓
		fMoW [117]	LoveDA [126]	Million-AID [100]						
		BigEarthNet [113]	fMoW [117]	PUB11 [7]			BLIP-2 (6.7B) [118]	1~5	49 / 87	✓
		Million-AID [100]								
		DIOR [51]	GeoPile [295]	Google Earth						
		Million-AID [100]	RSICB [296]	SkyScript [8]		0.5~128	ChatGPT-4o [267]	1	52	✓
		SSL4EO-S12 [294]								
LuoJiaHOG [31]	94,856	Google Maps		1280×1280	-	Manual Annotation	1	124	✗	
LHRS-Align [9] [link]	1,150,000	Google Earth		768×768	1	Vicuna-v1.5 (13B) [47]	1	-	✓	
RSTeller [143] [link]	2,539,256	Google Earth Engine		448×448	0.6	Mixtral-7B [144]	2~5	54	✓	
SkyScript [8] [link]	2,600,000	Google Earth Engine		-	0.1~30	Rule-based Assembly [8]	2	-	✓	

[link] directs to dataset websites. The average caption length (Avg. Cap. Length) in ChatEarthNet is calculated for captions generated by ChatGPT-3.5 and ChatGPT-4V separately, and in RS5M for PUB11 and other remote sensing datasets separately.

filtered from 11 computer vision datasets, including LAION-400M [209] and CC3M [266]. PUB11 creation involved sequential steps such as invalid image checking, deduplication, filtering using a CLIP model, and employing a remote sensing image detector. These pre-training datasets support the development of vision-language models beyond specific airborne or satellite sensors. An exception is ChatEarthNet [22], which is tailored for satellite image analysis and sources images from the Sentinel-2 collected in SatlasPretrain [119].

In addition to open-source image datasets, public geographic databases contribute to the extensive collection of remote sensing images [8], [9], [31], [143], [293]. For example, LuoJiaHOG [31] derives its images from Google Maps, determining global sampling points through spatial analysis and the evaluation of landscape indices. This method ensures the inclusion of images representing diverse topographies and varying economic conditions across countries and regions. Leveraging a customized OpenStreetMap (OSM) database, LHRS-Align [9] collects remote sensing images from Google Earth, ensuring their centers are aligned with OSM features. These images then undergo a series of processing steps, including resizing to a uniform size, deduplication to remove images dominated by vast ocean areas or obscured by clouds, and pruning using a trained network. The final dataset encompasses images from 9,259 cities across 129 countries. Similarly, Git-10M [293] consists mainly of remote sensing images from Google Earth, which are gathered from both randomly selected and manually curated regions worldwide. Manual screening is performed to discard redundant scenes, and an image enhancement model is applied to all collected images, thereby improving dataset quality. SkyScript [8] and RSTeller [143] are sourced from Google Earth Engine. Among these, the authors of SkyScript carefully choose 10 image collections from Google Earth Engine, such as SWISSIMAGE 10cm RGB imagery and Landsat 9 C2 T1 TOA Reflectance, forming a multi-source, multi-resolution image pool. In contrast, RSTeller relies solely on one image collection, the National Agriculture Imagery Program, which provides aerial images covering most of the continental United States and parts of Hawaii, with a ground

sampling distance of 0.6m.

Caption Generation: Texts associated with remote sensing images in pre-training datasets, also referred to as image captions, are typically human-understandable sentences that describe various aspects of the images. These descriptions often include information about image properties, overall scenes, and specific local objects. In existing pre-training datasets, only RSICap [1] and DIOR-Captions [29] rely entirely on manual annotation to generate image captions. In [1], five remote sensing experts carried out the annotation process following predefined principles. The principles consist of three points: first, describing image attributes; second, describing object attributes; and third, providing a description of the overall scene before detailing specific objects. In [29], the authors refer to bounding boxes from the original image dataset to describe the most prominent or abundant objects. Although manual annotation guarantees high-quality image-text pairs, it is a time-intensive and laborious process, particularly for large-scale datasets. To address this, later datasets have adopted automated methods, employing rules or models to generate captions. This approach significantly reduces the cost of dataset creation while greatly expanding its scale.

(1) *Rule-based Captioning*: Motivated by the fact that open-source image datasets are accompanied by image semantic information in the form of bounding boxes, segmentation masks, and class names, Liu et al. [2] propose the box-to-caption (B2C) and mask-to-box (M2B) methods to convert heterogeneous annotations into natural language captions. In detail, B2C, designed for bounding box annotations, generates five distinct captions for each image by considering the location (bounding box center) and the number of objects. Meanwhile, M2B, tailored for segmentation mask annotations, converts segmentation masks into bounding boxes by first extracting contours for each class and then sorting the contour points to define the bounding boxes. By applying M2B followed by B2C, it becomes convenient to create image-text datasets from image segmentation datasets. With a similar motivation, Wang et al. [8] leverage semantic information from OSM to caption remote sensing images from Google Earth Engine, as the two

can be linked through geo-coordinates. Due to its uncurated nature, semantic information from OSM is rich but often messy. Consequently, not all this information is suitable for describing images, especially if it is not visually discernible. To address this, the authors leverage CLIP embeddings of semantic information as input and apply a binary logistic regression model to filter out invisible information. After filtering, a description for each object is crafted by assembling OSM semantic data using connecting words such as “of”, “is”, “and”. Each image includes two captions: one describing the object used to determine the image boundary, and another describing multiple objects within the image by assembling captions of individual objects based on their geospatial relationships. To reduce noisy image-text pairs, they finally use a CLIP model to estimate the similarity between images and texts, and perform image filtering.

(2) *Model-based Captioning*: Numerous large language models are available for generating remote sensing image captions, including ChatGPT series [267], Gemini [265], Vicuna [47], and Mixtral [144]. To effectively apply these models, the key research focus lies in prompt design, as caption quality heavily depends on the prompt. The fundamental principle of prompt design is to align with the input requirements of the LLM being used. For multimodal LLMs such as ChatGPT-4V, Gemini-Vision, BLIP-2 [118], and MiniGPT-4 [129], which can process both text and image inputs, the prompt can simply instruct the model to generate a caption like “generate a description for the image’s visual content” [7]. However, to improve caption quality, prompts are optimized by adding constraints on the model’s response [22], [31], [202] or incorporating semantic information about the image [22], [31], [293]. For instance, VersaD [202] defines prompts to ensure the captions generated by Gemini-Vision encompass information about image properties, object attributes, and scene context, while preventing the model from describing uncertain objects. Furthermore, it imposes constraints on the format of the model’s response. LuoJiaHOG [31] prompts MiniGPT-4 to follow principles such as describing object attributes, reducing vague words, using “next” instead of “up”, adding synonyms, and more. Incorporating image semantic information into the prompt helps guide the model’s focus toward the image content of interest, while enriching the captions. For example, in ChatEarthNet [22], the distribution and proportion of various land cover types in the image, derived from the European Space Agency (ESA)’s WorldCover project, are embedded into the prompt to guide ChatGPT-4V in describing the land cover types of interest. In LuoJiaHOG, manually corrected OSM semantics are provided to MiniGPT-4, allowing the model to describe objects of interest within the images.

For models like ChatGPT-3.5, Vicuna-v1.5, and Mixtral, which primarily handle text inputs, providing image semantic information in the prompt allows the models to simulate “seeing” the image and generate captions. LHRS-Align [9] and RSTeller [143] both utilize OSM semantics to assist the model in understanding the image. Since the quality of the semantic information directly impacts the quality of the captions, a significant portion of the effort in caption generation for these two datasets is devoted to carefully processing the semantic

data. For example, similar to SkyScript [8], LHRS-Align filters out invisible semantic information using rules, followed by manual inspection. Additionally, LHRS-Align removes duplicate semantics for each image and applies a threshold to ensure semantic balance across the entire dataset. Given that an image is often associated with abundant OSM semantic data, providing all of it to the LLM can be overwhelming, making it difficult for the model to generate accurate captions. Therefore, RSTeller focuses on describing the primary object in each image, defined by the largest size or longest length, and refers to the OSM Wiki to interpret the OSM semantics, thereby reducing the ambiguity of the semantic information. Based on the processed OSM semantic data, LHRS-Align and RSTeller use custom templates to integrate semantic information into the prompt. To enhance the model’s understanding of the captioning task, a few examples are provided to the model.

Based on the captions generated by models, some techniques are proposed to enrich the captions further. In RS5M [7], the meta-information of images, *e.g.* longitude, latitude, and timestamp, is structured into readable sentences and combined with the generated captions. Additionally, RS5M includes rotation-invariant captions, which are those that exhibit the most stable similarity to the image features, regardless of their rotation, among the generated captions. In RSTeller, the LLM is guided to create multiple revisions from the generated captions with different tones, resulting in each image being accompanied by at least two captions, and up to five in total.

Dataset Property: Table XII provides an overview of existing pre-training datasets for vision-language modeling in remote sensing, with several key points worth noting:

(1) *Geographic Coverage*: The datasets RS5M, VersaD, and ChatEarthNet benefit from large-scale remote sensing image datasets, while SkyScript, LHRS-Align, LuoJiaHOG, and Git-10M derive from public geographic databases. Each of these datasets provides relatively comprehensive global coverage, with Git-10M being the largest (over 10 million image-text pairs). In contrast, RSTeller’s coverage is primarily limited to the continental United States and parts of Hawaii, potentially introducing geographic bias in trained models.

(2) *Scene Diversity*: Large-scale datasets such as RS5M, LHRS-Align, and SkyScript have, to some extent, ensured diversity in remote sensing scenes. Taking this a step further, the creators of LuoJiaHOG employ geospatial analysis to collect images from global regions with varied topography and different development levels. Meanwhile, the Git-10M team manually selects specific areas to ensure comprehensive coverage of representative scenes, including urban areas, forests, mountains, and deserts.

(3) *Caption Quality*: Manually annotated or rule-generated image captions typically achieve high accuracy. For instance, manual inspection of 1,000 randomly sampled image-text pairs from SkyScript demonstrates 96.1% precision. Conversely, model-generated captions inevitably contain errors despite considerable efforts to mitigate noise injection. For smaller-scale datasets like ChatEarthNet, manual correction can significantly improve caption quality. However, this approach becomes prohibitively expensive for million-scale datasets such as VersaD and RS5M, making caption errors unavoidable.

Surprisingly, the authors of VersaD experimentally verified that models pre-trained on the noisy dataset VersaD (82.3%) outperformed those trained on the more accurate dataset SkyScript, suggesting that rich-content and long captions may make models less sensitive to noise.

(4) *Distinctive Characteristics*: Current pre-training datasets predominantly contain English text, with DIOR-Captions being the only exception that provides bilingual captions in both Chinese and English. This enables the exploration of cross-lingual vision-language alignment. Moreover, while most datasets focus on optical images, ChatEarthNet stands out by facilitating a deeper understanding of multispectral images by offering satellite images with nine specific bands, each paired with textual descriptions.

B. Instruction-Following Datasets

Instruction-following datasets are specifically designed for the supervised fine-tuning of instruction-based vision-language models, allowing them to perform specific remote sensing tasks. For contrastive or generation-based vision-language models, the emphasis during task-specific applications lies more on designing network architectures than on developing new datasets, as discussed in Sections III-B and V-B. Therefore, in the fine-tuning phase of vision-language modeling, we exclusively summarize instruction-following datasets. This type of dataset is comprised of images paired with conversations, structured as instructions (or questions) and answers. Its construction involves image collection and conversation generation, with images primarily sourced from open-source remote sensing image datasets, as shown in Table XIII. The main challenge lies in generating conversations, which is addressed in this section, followed by an overview of notable instruction-following datasets.

Conversation Generation: Since sourced image datasets come with task-specific annotations, the most intuitive idea is to convert these annotations into a dialogue format. Currently, three methods are commonly used for this conversion: template-based transformation, large model assistance, and manual annotation. Below, we demonstrate each method with examples from specific tasks.

(1) *Template-Based Transformation*: To create task-specific conversations, open-source image datasets for the given task are typically used. Answers are derived directly from dataset annotations, while instructions are either manually defined or randomly selected from a pool generated by large language models. This pool contains instructions tailored to specific tasks, conveying the same meanings with varied phrasing. For example, in MMRS-1M [6], classification datasets such as NWPU-RESISC45 [53] and EuroSAT [69], along with object detection datasets like DIOR [51] and DOTA [45], are selected to create instruction-following data for classification and detection tasks. For scene classification, the authors design instructions using the template “What is the category of this remote sensing image? Answer the question using a single word or phrase. Reference categories include category 1, ..., and category n ”. For object detection, the template is “Detect all objects shown in the remote sensing image and

describe using oriented bounding boxes”. In SkyEye-968k [5], corresponding datasets are selected for image captioning and visual grounding tasks, including UCM-captions [62], Sydney-captions [62], RSVG [66], and DIOR-RSVG [33]. Separate instruction pools are constructed for each task. The image captioning pool includes instructions such as “Briefly describe this image” or “Provide a concise depiction of this image.” Meanwhile, the visual grounding pool contains instructions like “Give me the location of referring expression” or “Where is referring expression?”. Based on task-specific instruction-following data, multi-turn conversations can be generated by mixing data from different tasks [5], [6].

(2) *Large Model Assistance*: Simple template-based transformations are insufficient for generating instruction-following data for complex reasoning tasks, which equip vision-language models with higher-order cognitive abilities, such as making decisions and identifying relationships. Additionally, multi-turn conversations created through multi-task mixing after template transformation often lack diversity. To address these issues, large language models are utilized to generate instruction-following data. By providing a few manually defined in-context examples, the models learn to create high-quality instruction-answer pairs based on image captions and other related information. For instance, Kuckreja et al. [3] prompt Vicuna [47] to generate 30,000 detailed image descriptions, 65,000 multi-turn conversations, and 10,000 complex reasoning based on short descriptions supplied. Similarly, Muhtar et al. [9] use Vicuna [47] to produce instruction-following data based on image captions from RSITMD [61] and NWPU-Captions [63]. To ensure that the generated conversations focus on the visual content of the images, their prompts are designed to generate questions that inquire about object types, actions, locations, relative positions between objects, and more. However, considering the limitations of RSITMD and NWPU-Captions, where captions are short and lack detailed content, the authors supplemented their data with additional image-caption pairs from LHRS-Align [9]. GPT-4 is prompted with image captions, object bounding boxes, and object attributes, to generate detailed image descriptions, complex reasoning, and multi-turn conversations that incorporate questions about object locations and counts.

(3) *Manual Annotation*: In constructing instruction-following datasets, manual annotation primarily serves to provide accurate information that supports template-based or large model-assisted conversation generation methods, rather than directly creating conversations. For example, in FIT-RS [11], an instruction-following dataset for understanding semantic relationships between objects, Luo et al. [11] manually annotate very high-resolution remote sensing images with detailed scene graph labels, laying the groundwork for generating conversations focused on relationship comprehension tasks. In TITANIC-FGS [142], the authors manually summarize the common and private features of fine-grained objects, which are used to populate predefined templates for creating object descriptions. These descriptions are subsequently used to prompt GPT-4, enabling the generation of multi-turn conversations that simulate human-like logical decision-making.

The three methods can be combined to facilitate the con-

TABLE XIII
SUMMARY OF INSTRUCTION-FOLLOWING DATASETS FOR FINE-TUNING VISION-LANGUAGE MODELS IN REMOTE SENSING.

Dataset	Task	Data Source						#Sample Train / Test	Public
MMShip [24] Hnstd [202] [link]	Object Detection Honest Question Answering	DOSR [89] DOTA v2 [45]	DOTA ship subset [45] FAIRIM [52]	HRSID [90]	SSDD [79]			81,000 / - 45,000 / 1,642	✗
VersaD-Instruct [202] [link]	Complex Reasoning Multi-Turn Conversation	DIOR [51]	DOTA v2 [45]	FAIRIM [52]				30,000 / -	✓
RS-Instructions [10]	Image Captioning Visual Question Answering Visual Grounding	RSIVQA [65]	RSVQA-LR [54]	UAV [87]	UCM-captions [62]			5506 / 1552	✓
RS-GPT4V [30]	Complex Reasoning Image/Region Captioning Visual Question Answering Visual Grounding	DIOR-RSVG [33] RSIVQA [65]	FloodNet [55] RSVQA-HR [54]	NWPU-Captions [63] RSVQA-LR [54]	RSICD [60] Sydney-captions [62]	RSITMD [61] UCM-captions [62]	991,206 / 258,419	✗	
SkyEye-968k [5] [link]	Image/Video Captioning Visual Question Answering Multi-Turn Conversation Visual Grounding	CapERA [64] NWPU-Captions [63] RSVQA-HR [54] UCM-captions [62]	DIOR-RSVG [33] RSICD [60] RSVQA-LR [54] UCM-Conversa [5]	DIOR-Conversa [5] RSITMD [61] RSVG [66]	DOTA-Conversa [5] RSIVQA [65] Sydney-captions [62]	ERA-VQA [5] RSPG [5] Sydney-Conversa [5]	968,000 / -	✓	
Multi-task Dataset [9] [link]	Scene Classification Image Captioning Visual Question Answering Visual Grounding	DIOR-RSVG [33] RSITMD [61]	fMoW [117] RSVG [66]	METER-ML [224] RSVQA-HR [54]	NWPU-RESISC45 [53] RSVQA-LR [54]	RSICD [60] UCM-captions [62]	42,322 / -	✓	
TITANIC-FGS [142]	Fine-Grained Ship Classification Ship Image Captioning Ship Image Visual Question Answering Multi-Turn Conversation	Google	Baidu					16,876 / 2,053	✗
RSVP-3M [136]	Scene/Region/Point Classification Image/Region/Point Captioning Object Relationship Reasoning Multi-Turn Conversation	DIOR-RSVG [33] Hi-UCP [138] LEVIIR [123] Optimal-31 [98] SOTA [140] VisDrone [78]	DOSR [89] HRRSD [76] MAR20 [137] Potsdam [127] SOTA [140] WHU [141]	DOTA v2 [45] HRSC2016 [122] NWPU-Captions [63] RSI46-WHU [96] UAVid [139] WHU-RS19 [71]	FAIRIM [52] iSARD [125] NWPU-RESISC45 [53] RSITMD [61] UCAS-AOD [77]	FAST [140] Kuckreja et al. [3] NWPUVHR10 [75] RSOD [96] Vaihingen [128]	3,648,884 / -	✗	
Kuckreja et al. [3] [link]	Visual Grounding Complex Reasoning Scene Classification Image/Region Captioning Visual Question Answering Multi-Turn Conversation Object Counting Image Captioning	DIOR [51] RSVQA [54]	DOTA [45]	FAIRIM [52]	FloodNet [55]	NWPU-RESISC45 [53]	306,000 / 12,000	✓	
LHRS-Instruct [9] [link]	Complex Reasoning Object Attribute Recognition Object Relationship Analysis Multi-Turn Conversation	LHRS-Align [9]	NWPU-Captions [63]	RSITMD [61]			39,800 / -	✓	
MMRS-1M [6] [link]	Visual Grounding Object Detection Scene Classification Image/Region Captioning Visual Question Answering Multi-Turn Conversation	Aerial-mancar [83] DOTA [45] FGSCR-42 [74] Infrared-security [82] RSICD [60] RSVQA-LR [54] UCM [70]	AIR-SARShip-2.0 [75] Double-light-vehicle [84] FloodNet [55] Infrared-security [82] NWPU-Captions [63] RSITMD [61] Sea-shipping [81] UCM-captions [62]	CRSVQA [68] DSCB [73] HTU-UAV [80] NWPU-Captions [63] NWPU-RESISC45 [53] RSIVQA [65] SSDD [79] WHU-RS19 [71]	DIOR [51] EuroSAT [69] HRSSD [76] NWPUVHR10 [75] RSOD [96] Sydney-captions [62] VisDrone [78]	DIOR-RSVG [33] FAIRIM [52] HRSID [90] Oceanic ship [85] RSSCN7 [72] UCAS-AOD [77]	1,005,842 / -	✓	
PIT-RS [11] [link]	Visual Grounding Object Detection Image/Region Captioning Visual Question Answering Multi-Label Classification Object Relationship Reasoning Image/Region Scene Graph Generation Multi-Turn Conversation	extended STAR [88]					1,440,681 / 360,170	✓	
VariousRS-Instruct [202] [link]	Visual Grounding Object Counting Scene Classification Building Vectorizing Geometric Measurement Image Property Recognition Visual Question Answering Multi-Label Classification Visual Grounding Change Detection Scene Classification Image/Region Captioning Visual Question Answering Temporal Scene Classification Temporal Referring Expression Spatial Change Referring Expression Image/Region Change Question Answering	BANDON [268] FAIRIM [52] MSAR [270] RSVQA-LR [54]	CrowdAI [262] FBP [272] MtS-WH [269] UCM-captions [62]	DeepGlobe [271] fMoW [117] NWPU-RESISC45 [53]	DIOR-RSVG [33] GID [97] Potsdam [127]	DOTA v2 [45] METER-ML [224] RSITMD [61]	76,445 / 13,020	✓	
TEOChatlas [213] [link]		fMoW [117]	Kuckreja et al. [3]	QFabric [217]	S2Looking [216]	xBD [215]	554,071 / -	✓	

[link] directs to dataset websites.

struction of instruction-following datasets [11], [136], with the goal of creating rich and diverse samples across a variety of tasks. To help vision-language models better distinguish between tasks, a common practice is to incorporate task-specific identifiers into the instructions, as demonstrated in [3], [5], [9], [202]. One example is [5] where the authors introduce task identifiers such as “[caption]”, “[vqa]”, and “[refer]” to specify captioning, visual question answering, and visual grounding tasks, respectively.

Impressive Datasets: Table XIII provides an overview of existing instruction-following datasets. MMShip [24] and TITANIC-FGS [142] are specifically tailored for ship image analysis, whereas the others are designed for general-purpose remote sensing data analysis. Most datasets include instruction data for tasks such as captioning, visual question answering, visual grounding, and multi-turn conversations [3], [5], [6], [9], [11], [30], [30], [142], [213]. Among these, MMRS-1M [6] stands out for featuring instruction data that involves

not only optical remote sensing images but also synthetic aperture radar and infrared images. To achieve this, three SAR image datasets, *i.e.* AIR-SARShip-2.0 [75], SSDD [79] and HRISD [90], along with six infrared image datasets (HTU-UAV [80], Sea-shipping [81], Infrared-security [82], Aerial-mancar [83], Double-light-vehicle [84] and Oceanic ship [85]) are incorporated into the creation of instruction-following data. Beyond supporting common tasks, some datasets incorporate innovative instruction data that endow vision-language models with impressive capabilities. These include fine-grained image understanding, time-series image analysis, quantitative analysis, honest question answering, and object relationship comprehension. The remainder of this section offers a detailed discussion of these datasets.

(1) *RSVP-3M* [136] is the first visual prompting instruction dataset for remote sensing, with samples consisting of images, visual prompts, and conversations. The visual prompts take the form of masks that match the size of the correspond-

ing image. These masks include bounding boxes or points, highlighting regions or points of interest specified by the user within the image. For a specific task, visual prompts in the instructions are described as “each marked point” or “each marked region”, directing the model to focus on the indicated areas and perform the corresponding analysis. For example, in point-level classification, the instruction can be “Please identify the category of each marked point in the image.”, with the model’s response formatted as “<Mark 1>: Label 1\n <Mark 2>: Label 2\n”. Similarly, for region-level classification, instructions like “Please identify the category of each marked region in the image.” yield “<Region 1>: Label 1\n <Region 2>: Label 2\n”. To preserve the model’s image-level capabilities, image-level visual prompts are represented as bounding boxes with dimensions [0, 0, width, height]. Based on these definitions, RSVP-3M source images from 28 existing datasets spanning tasks such as image classification, image caption, object detection, and instance segmentation. Ground truth bounding boxes or masks from these datasets are used as visual prompts. Furthermore, GPT-4V is employed to automatically generate instruction data, resulting in over 3 million samples. This large-scale dataset empowers EarthMarker [136] with multi-granularity understanding of remote sensing images across image, region, and point levels for tasks like classification, captioning, and relationship analysis.

(2) *TEOChatlas* [213] seeks to unlock the potential of vision-language models for time-series image analysis. It provides instruction data for a variety of temporal tasks, including temporal scene classification, change detection (represented by bounding boxes), temporal referring expression, spatial change referring expression, and image/region change question answering. These tasks span two real-world applications: disaster response and urban development monitoring. Fig. 8 illustrates examples of each task. To ensure diverse image sequence lengths and image sources, TEOChatlas integrates datasets such as xBD [215], S2Looking [216], QFabric [217], and fMoW [117], which cover bi-temporal, penta-temporal, and multitemporal sequences. These datasets are sourced from six different sensors, namely WordView-2/3, Sentinel-2, GaoFen, SuperView, and BeiJing-2. Building on this diverse data foundation, conversation generation is assisted by GPT-4o [267], resulting in 245,210 samples tailored for temporal tasks. In addition, TEOChatlas incorporates 308,861 samples from the instruction-following dataset of GeoChat [3], which focuses on single-image analysis. To help the model differentiate between single-image tasks and temporal tasks, task-specific instructions are supplemented with prompts that explicitly specify the input consists of a sequence of images and, optionally, indicate the resolution and sensor name of the input images.

(3) *LHRS-Instruct* [9], *VariousRS-Instruct* [202] initiate the exploration of applying vision-language models for quantitative image analysis. In LHRS-Instruct, GPT-4 is prompted with image captions and detailed information about the types and coordinates of objects within the image. This allows GPT-4 to generate conversations centered on the number of objects, thereby equipping LHRS-Bot [9] with the capability to perform object counting. VariousRS-Instruct, on the other hand, extends its functionality beyond object counting by

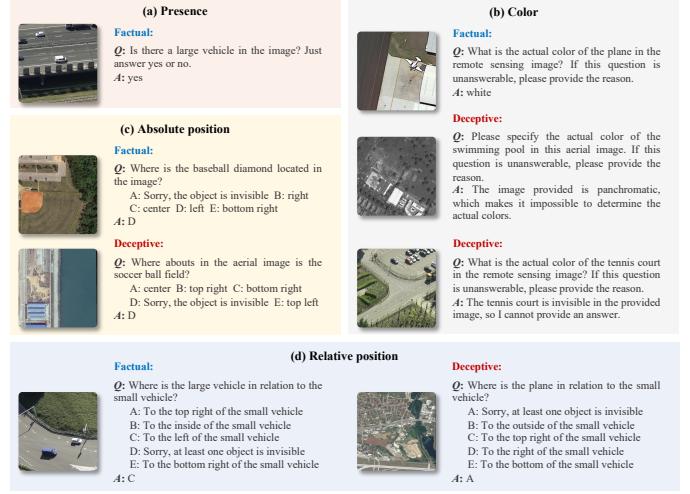


Fig. 12. Samples in the HnstD dataset [202].

incorporating instruction data designed for geometric measurement tasks, aimed at estimating the length and width of objects within images. This instruction data is derived from the DOTA [45] and FAIR1M [52] datasets through a process of template-based conversion. Ground truth for object sizes is determined using image resolution and the sizes of objects’ bounding boxes. Beyond quantitative tasks, VariousRS-Instruct also introduces new qualitative tasks such as building vectorizing and multi-label classification. These advancements contribute to the evolution of a versatile vision-language model for comprehensive remote sensing image analysis.

(4) *HnstD* [202] aims to enhance the honesty of vision-language models. It incorporates four recognition tasks: identifying the relative positions between objects, their presence, color, and absolute positions. Except for the presence task, all others include both factual and deceptive instructions. Specifically, deceptive instructions regarding object color arise from either the absence of objects or their presence in panchromatic images, while those for relative and absolute positions stem from the absence of objects. This combination of factual and deceptive instructions helps prevent models from producing affirmative answers to unreasonable user queries. To construct HnstD, template-based transformations are applied to images sourced from the DOTA [45] and FAIR1M [52] datasets. The model’s response formats vary by task: “yes” or “no” for the presence task, free-form text for the color task, and option selection from candidates for relative and absolute position tasks. Fig. 12 presents samples from the HnstD dataset.

(5) *FIT-RS* [11] is designed to advance vision-language models’ ability to comprehend fine-grained semantic relationships, introducing tasks such as object relationship reasoning and scene graph generation at both region and image levels. Object relationship reasoning is further divided into relation detection and relation reasoning. Relation detection requires the model to predict relationships between targets based on the provided categories and locations of subjects and objects, while relation reasoning demands the model to additionally infer the categories of the subject or object along with their relationships. Region-level scene graph generation involves

describing an object’s relative size, position, and visibility, followed by generating all subject-relation-object triples. Image-level scene graph generation extends this by creating scene graphs for entire images, appending isolated objects (with no relationships) to the answer in an object detection format. These tasks challenge vision-language models to accurately interpret relationships between targets and distinguish between subjects and objects within those relationships. To construct FIT-RS, the scene graph generation dataset STAR [88], containing over 400,000 triplets across 48 object categories and 58 semantic relationship categories, serves as the data source. Template-based transformations are applied to generate task-specific instruction data. Similar to most instruction-following datasets, FIT-RS also supports tasks such as captioning and visual question answering. In total, it comprises 1,800,851 samples.

C. Benchmark Datasets

Benchmark datasets are essential for evaluating and fairly comparing the performance of different models. To foster research in vision-language modeling for remote sensing, efforts have been made to develop benchmark datasets that integrate remote sensing images with text. Consequently, datasets like UCM-captions [62], RSICD [60], and RSVQA [54] have been established and are widely utilized to assess models’ abilities in visual perception, text understanding, and image-text alignment [2], [7], [9], [202]. However, the limited scale and task-specific design of these existing datasets have proven inadequate for a comprehensive evaluation of modern vision-language models. These advanced models, benefiting from the two-stage training paradigm and large-scale training datasets, excel at handling diverse remote sensing image analysis tasks in a conversational manner. In response to this, researchers have begun crafting large-scale benchmark datasets that encompass diverse tasks and are carefully designed to align with the input-output requirements of models, aiming to challenge vision-language models and push the boundaries of their capabilities. This section highlights these new research advancements [12], [32], [156], [273], [274], categorizing benchmark datasets into two types: instruction-specific datasets and general-purpose datasets, based on the organization of their samples. Note that remote sensing image datasets, such as AID [102] and NWPU-RESISC45 [53], which have been adapted for evaluating the performance of vision-language models through appropriate transformations, are comprehensively reviewed in [100], [288].

Instruction-Specific Datasets: This type of dataset is specifically designed for instruction-based vision-language models, with samples typically presented as image-instruction-answer pairs. Its creation follows a similar process to instruction-following datasets, involving image collection and conversation generation. As summarized in Table XIV, most benchmark datasets source their images from open-source image datasets and leverage methods such as template transformation, assistance from large language models, and human annotation to produce high-quality conversations. Detailed methods for conversation generation are provided in Section VI-B, while this section focuses on dataset properties,

including task formulations, the number of samples, and question format. From Table XIV, three key observations can be made. First, to align with the development trend of multifunctional capabilities in vision-language models, benchmark datasets now include an increasing variety of tasks, extending beyond common ones like image captioning and visual question answering. They now encompass more complex reasoning tasks, as well as expansions from single-image analysis to temporal or cross-view image analysis. Second, the scale of these datasets is increasing, often comprising tens of thousands of samples. However, as these samples are spread across an ever-growing variety of tasks, only a few hundred samples are typically available for each specific task. Third, question formats in the datasets can be categorized into three types: single-choice, multiple-choice, and open-ended. For single-choice and multiple-choice questions, each question is accompanied by a set of candidate answers, facilitating objective assessment of model performance. In contrast, open-ended questions allow models to generate answers freely, closely reflecting real-world scenarios where users may not know the answer in advance. Below, we present a detailed discussion of these datasets.

(1) *RSIEval* [1], *VRSBench* [12] both focus on common tasks. RSIEval, which relies entirely on human annotation, is limited in scale, with only 100 samples for image captioning and 936 samples for visual question answering. Such a small dataset may be insufficient for evaluating a model’s practicality and robustness. In contrast, VRSBench leverages a semi-automatic creation pipeline, substantially increasing the dataset size. It contains 29,614, 52,472, and 123,221 samples for image captioning, visual grounding, and visual question answering, respectively, with LLaVA-1.5 [49], GeoChat [3], and GPT-4V achieving the best performance in each task.

(2) *FIT-RSRC* [11], *LHRS-Bench* [9] challenge a model’s capability to understand relationships between objects in images. Specifically, FIT-RSRC examines the comprehension of semantic relationships among objects, using terms like “run along” and “around” to describe these relationships. It features four types of questions: querying the relationship between two objects, the existence of a specific relationship, the subject of a relationship, and the object of a relationship. Furthermore, each type of question contains unanswerable variants to assess the model’s robustness and veracity. LHRS-Bot, on the other hand, targets spatial relationships, describing them using terms like “top”, “middle” and “bottom”. It also includes tasks such as object counting, object attribute recognition, image property recognition, and visual reasoning. However, with only 690 samples in total, and some tasks containing merely a few dozen samples, its scale is highly inadequate. On these datasets, SkySenseGPT [11] and LHRS-Bot [9] demonstrate the best performance, respectively.

(3) *COREval* [273] expands the evaluation of model capabilities from perception to reasoning. In terms of perception, it encompasses image-level comprehension, single-instance identification, and cross-instance discernment. For reasoning, it emphasizes inferring attributes of scenes or instances, such as the CO₂ emissions or population density of a scene, the height of a building, or the imaging season of an image.

TABLE XIV
SUMMARY OF BENCHMARK DATASETS FOR VISION-LANGUAGE MODELS IN REMOTE SENSING.

Dataset	Data Source	Task		#Sample	Question Format	Public		
FIT-RSRC [11] [link]	-	Relation Comprehension	-	Single-choice	✓			
RSIEval [1] [link]	DOTA-v1.5 [45]	Image Captioning	100	Open-ended				
VRSBench [12] [link]	DIOR [51]	Visual Question Answering	936	Open-ended	✓			
LHRS-Bench [9]† [link]	Google Earth	Image Captioning	29,614					
COREval [273]†	Google Earth SDGSAT-1 SWISSIMAGE	Visual Grounding	52,472	Open-ended	✓			
VLEO-Bench [34] [link]	Aerial Landmark [34] COWC [115] NeonTreeEvaluation [114] xBD [215]	Object Counting	123,221					
GEOBench-VLM [274]† [link]	AiRound [275] Deforestation [279] FAIRIM [52] fMoW [117] GeoNRW [285] NASA Marine Debris [280] PatternNet [95] So2Sat [286]	Animal Detection [116] DIOR-RSVG [33] PatternNet [95]	BigEarthNet [113] fMoW-WILDS [112] RSICD [60]	Image-Level Comprehension Single-choice Cross-Instance Discernment Attribute Reasoning Assessment Reasoning Common Sense Reasoning Location Recognition Image Captioning LULC Classification Multi-Label LULC Classification Visual Grounding Object Counting Change Detection	3,257 1,244 562 300 400 500 602 1,009 3,000 1,000 - 2,239	Single-choice Single-choice Open-ended Single-choice Single-choice Single-choice Single-choice Open-ended Single-choice Multiple-choice Open-ended Open-ended Open-ended	✓	
UrBench [156]†	Cityscapes [157] IM2GPS [160]	Google Earth Google Street View	MTSD [158] VIGOR [159]	Scene Understanding Object Classification Object Localization and Detection Event Detection Caption Generation Semantic Segmentation Temporal Understanding Non-Optical Imagery	10,000	Single-choice Single-choice Single-choice Single-choice Open-ended Open-ended Single-choice -	✓	
GeoText-1652 [162] [link] DIOR-RSVG [33] [link] RemoteCount [2]	University-1652 [161] DIOR [51] DOTA [45]	AID [102] BCS Scenes [111] EuroSAT [69] MLRSNet [103]	AWTP [105] Canadian Cropland [108]	BC Scenes [110] CLRS [99] GID [97] MultiScene [104]	Image-Text Retrieval Visual Grounding Object Counting	1,652 38,320 947	N/A N/A N/A	✓
SATIN [32] [link]	NWPU-RESISC45 [53] Post Hurricane [106] RSI-CB256 [101] SAT-6 [91] UCM [70]	Optimal-31 [98] RSC11 [93] RSSCN7 [72] SIRI-WHU [94] USTC-SmokeRS [109]	PatternNet [95] RSD46-WHU [96] SAT-4 [91] SISI [107] WHU-RS19 [71]	Land Cover Classification Land Use Classification NaSC-TG2 [92] Hierarchical Land Use Classification Complex Scene Classification Rare Scene Classification False Colour Scene Classification	201,000 260,205 34,000 135,261 105,840 39,326	N/A	✓	

[link] directs to dataset websites. † indicates that the dataset's task follows a hierarchical taxonomy. In this table, only the broad dimensions of task design are displayed, rather than specific tasks.

These two evaluations are divided into 6 sub-dimensions and 22 specific tasks, with a total of 6,263 samples. Recognizing the significant regional intra-class variations in remote sensing images, COREval incorporates images sourced from multiple satellites (Landsat-8, SDGSAT-1, and Sentinel-1/2) and geographic databases (Google Earth, Natural Earth, and SWISSIMAGE), offering geographic coverage of 50 cities spanning six continents. On this dataset, 13 open-source vision-language models from both general and remote sensing domains are evaluated. The experimental results reveal that while existing models perform well in image-level comprehension, they struggle with fine-grained instance perception and complex reasoning tasks.

(4) *VLEO-Bench* [34], *GEOBench-VLM* [274], from the perspective of remote sensing applications (*e.g.* urban monitoring and disaster management), integrate remote sensing image analysis tasks such as temporal analysis of changes, counting objects, and understanding relationships between objects. In VLEO-Bench, three types of tasks are included: scene understanding, which tests a model's ability to combine high-level image semantics with knowledge expressed in language; object localization and counting, which evaluates

fine-grained perception; and change detection, which assesses a model's capability to identify differences between multiple images. Experimental results on VLEO-Bench show that although state-of-the-art models like GPT-4V achieve strong performance in scene understanding, their poor spatial reasoning limits their effectiveness in object localization, object counting, and change detection tasks. GEOBench-VLM, on the other hand, includes over 10,000 questions across 31 tasks, categorized into 8 broad categories. Compared to VLEO-Bench, GEOBench-VLM offers a more diverse range of tasks, including unique ones like referring expression segmentation and non-optical imagery analysis. Additionally, unlike VLEO-Bench, which confines change detection to counting damaged buildings and estimating damage severity, GEOBench-VLM's temporal understanding covers a wider range of tasks. These include detecting the presence of changes, reasoning about change causes, assessing disaster impacts, and classifying crop types based on long-term time-series images. Detailed experiments with 10 vision-language models show that none excels across all tasks in GEOBench-VLM. Significant effort is still required to develop remote sensing-specific models capable of addressing challenging tasks like referring expression

segmentation.

(5) *UrBench* [156] concentrates on exploring the potential applications of vision-language models in urban scenarios. It comprises 11,600 questions across 14 tasks, spanning four dimensions: geo-localization, scene reasoning, scene understanding, and object understanding. Unlike the previously mentioned datasets that only include single-view questions in satellite or aerial images, UrBench also incorporates cross-view questions, in which each question pairs images of the same scenario captured from satellite and street views. These questions involve tasks such as image retrieval, orientation identification, camera localization, road understanding, object attribute recognition, and object matching. Meanwhile, they encompass region-level questions, which examine a model’s ability to understand urban scenarios at a region level, and role-level questions, which evaluate its potential to assist humans in daily life. Experiments with 21 general-purpose models on UrBench reveal that current models still lag significantly behind human experts in urban environments. They struggle to understand multi-view image relations, and their performance varies inconsistently across different views.

General-Purpose Datasets: are designed to evaluate various types of vision-language models. These datasets typically target specific remote sensing multimodal tasks, with samples consisting of images and task ground truth presented in text form. As summarized in [17], commonly used general-purpose datasets for vision-language model evaluation have been well-documented. This section, therefore, focuses on the latest research advancements from the past two years, as outlined in Table XIV. GeoText-1652 [162] is a language-guided geo-localization benchmark built on the University-1652 image dataset [161]. Each image includes detailed image-level descriptions as well as region-level brief descriptions with corresponding bounding boxes. DIOR-RSVG [33] is designed for visual grounding toward remote sensing. Through an automatic expression generation method, it contains 38,320 image-expression-box triplets. The large scale of this dataset makes it a popular choice for evaluating vision-language models. RemoteCount [2] evaluates object counting abilities. Each image is paired with a human-annotated caption, such as “a photo of 9 tennis courts”. As a result, this dataset is small in scale, containing only 947 samples. To meet the input requirements of contrastive learning-based foundation models, the original caption is augmented with nine additional captions by replacing the number in the caption with all numbers from 1 to 10. SATIN [32] sources images from 27 remote sensing datasets and is designed to evaluate the classification capabilities of models for satellite images. This dataset covers six classification tasks: land cover, land use, hierarchical land use, complex scenes, rare scenes, and false colour scenes. Among these, the hierarchical land use task tests the ability to classify land use across varying levels of granularity, while the complex scene task leverages the large view fields of remote sensing images to assess the capability of identifying multiple land use types within a single image. Experiments with 40 vision-language models on SATIN in a zero-shot setting show that this dataset poses a significant challenge, with even models trained on billions of natural

images achieving an accuracy of just over 50%.

VII. CONCLUSION AND FUTURE DIRECTIONS

From the perspectives of models and datasets, we have covered the advancements in vision-language modeling for remote sensing, knowing how remote sensing images and natural language can be effectively bridged, which remote sensing tasks existing vision-language models can address, and which datasets are suitable for developing and testing vision-language models. Naturally, this raises two important questions: 1) Are existing vision-language models adequate for practical applications? 2) If not, which directions are worth pursuing to advance this field further? The answer to the first question is, unsurprisingly, no. Vision-language modeling remains a highly challenging task and is far from meeting practical needs. In this section, we aim to share insights on future research directions from two perspectives: models and datasets.

Effective representation and alignment for cross-modal data: A growing trend seeks to advance vision-language models to accommodate a wide range of remote sensing images, including optical, SAR, and infrared, thereby enabling the acquisition of more comprehensive information about the Earth’s surface [6], [24]. However, in applications such as disaster risk assessment, these models may need to integrate additional information sources beyond remote sensing images, such as geospatial vector data and social media, to perform complex reasoning [289], [290]. Geospatial vector data presents complex data structures in the form of points, polylines, polygons, and networks. Meanwhile, social media encompasses texts in various languages (*e.g.*, Chinese, English, and French), diverse types of images (*e.g.*, photographs and emojis), videos, and more. This complexity and diversity pose challenges for models in comprehending the embedded information and associating it with remote sensing data. Consequently, there is a pressing need for effective representation and alignment across a broader scope of cross-modal data.

Requirements arbitrarily described in natural language: Existing instructions for prompting vision-language models are usually definite, such as using task identifiers to specify particular remote sensing tasks [3], [5] or providing candidate answers [9]. However, in more realistic and practical scenarios, requirements described in natural language tend to be vague and complex, involving the sequential execution of multiple tasks. For example, given an instruction like “Please assess the water quality in the indicated area of the image”, the model must be capable of decomposing water quality assessment into two sub-tasks: water detection and quantitative retrieval, and then accomplish each task in sequence. This creates the need to advance further models’ language understanding capability to adapt to users’ arbitrary or flexible demands.

Enhancing the reliability of model answers via expert explanations: Existing vision-language models take language instructions and visual representations as input, producing image analysis results in the form of natural language. However, these models typically do not provide expert explanations for their answers, leading users to doubt the reliability of the

outputs, especially in tasks requiring complex reasoning rather than simple recognition. For instance, in precision agriculture, where the model may be used to guide pesticide spraying schedules, the lack of explanations regarding crop diseases and pest conditions makes it hard to convince farmers to follow the recommendations. Consequently, expert explanations accompanying image analysis results are necessary, as they not only enhance the reliability and interpretability of the model’s answers but also offer insight into the decision-making process, thereby fostering user trust. A recent initiative [142] has begun exploring the feasibility of developing the vision-language model capable of performing fine-grained ship classification while providing reasoning behind its classification. This is accomplished by integrating domain knowledge into the construction of the instruction-following dataset.

Continually adapting vision-language models: It is well known that new remote sensing images are collected daily from around the world, and human demands may change accordingly. This necessitates continually adapting vision-language models to these dynamic changes, rather than relying solely on one-time pre-training and fine-tuning. Combining newly collected data with previously gathered data for continual learning is a straightforward and effective approach. However, the growing volume of data poses significant challenges in terms of computational and storage costs. Training the model exclusively on new data risks catastrophic forgetting [291], [292]. Therefore, it is crucial for the research community to explore effective learning strategies that enable vision-language models to learn new knowledge from new data while maintaining old knowledge.

More diverse and rich remote sensing multimodal dataset: The original CLIP model was trained on 400 million image-text pairs, whereas the largest remote sensing pre-training dataset (Git-10M [293]) contains only 10 million pairs, predominantly limited to optical images. This data scarcity restricts models’ capacity to capture a broad range of visual concepts, a challenge that is particularly pressing in remote sensing. Variations in imaging conditions, sensor parameters, and geographic locations result in highly varied visual characteristics of objects in remote sensing images [100]. Therefore, substantial efforts are required to create image-text datasets that encompass diverse ground objects and a rich variety of remote sensing image modalities.

Challenging and application-specific benchmarks: Most benchmark datasets [1], [12] for model evaluation are limited to a narrow range of remote sensing tasks (*e.g.*, visual question answering and image captioning), whereas contemporary vision-language models [11], [202], [213] in remote sensing demonstrate versatile capabilities across diverse image analysis tasks. Furthermore, while some benchmarks include multiple tasks [273], [274], each task typically has a limited number of samples. Such datasets are insufficient for thoroughly testing and comparing the performance of different vision-language models. Beyond general-purpose benchmarks, exploring application-specific benchmarks is also a promising research direction, as some efforts have focused on developing versatile vision-language models tailored to specific real-world applications [24], [142].

REFERENCES

- [1] Y. Hu, J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li, “Rsgpt: A remote sensing vision language model and benchmark,” *ISPRS J. of Photogrammetry Remote Sens.*, vol. 224, pp. 272–286, 2025.
- [2] F. Liu, D. Chen, Z. Guan, X. Zhou, J. Zhu, Q. Ye, L. Fu, and J. Zhou, “Remoteclip: A vision language foundation model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [3] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, “Geochat: Grounded large vision-language model for remote sensing,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 27831–27840, 2024.
- [4] U. Mall, C. P. Phoo, M. K. Liu, C. Vondrick, B. Hariharan, and K. Bala, “Remote sensing vision-language foundation models without annotations via ground remote alignment,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, pp. 1–13, 2024.
- [5] Y. Zhan, Z. Xiong, and Y. Yuan, “Skyeyept: Unifying remote sensing vision-language tasks via instruction tuning with large language model,” *ISPRS J. of Photogrammetry Remote Sens.*, vol. 221, pp. 64–77, 2025.
- [6] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, and X. Mao, “Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–20, 2024.
- [7] Z. Zhang, T. Zhao, Y. Guo, and J. Yin, “Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–23, 2024.
- [8] Z. Wang, R. Prabha, T. Huang, J. Wu, and R. Rajagopal, “Skyscript: A large and semantically diverse vision-language dataset for remote sensing,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, pp. 5805–5813, 2024.
- [9] D. Muhtar, Z. Li, F. Gu, X. Zhang, and P. Xiao, “Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 440–457, 2024.
- [10] Y. Bazi, L. Bashmal, M. M. Al Rahhal, R. Ricci, and F. Melgani, “Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery,” *Remote Sens.*, vol. 16, no. 9, p. 1477, 2024.
- [11] J. Luo, Z. Pang, Y. Zhang, T. Wang, L. Wang, B. Dang, J. Lao, J. Wang, J. Chen, Y. Tan, et al., “Skysensegt: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding,” *arXiv:2406.10100*, 2024.
- [12] X. Li, J. Ding, and M. Elhoseiny, “Vrsbench: A versatile vision-language benchmark dataset for remote sensing image understanding,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, 2024.
- [13] S. Dong, L. Wang, B. Du, and X. Meng, “Changeclip: Remote sensing change detection with multimodal vision-language representation learning,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 208, pp. 53–69, 2024.
- [14] X. Li, C. Wen, Y. Hu, and N. Zhou, “Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 124, p. 103497, 2023.
- [15] A. Zavras, D. Michail, B. Demir, and I. Papoutsis, “Mind the modality gap: Towards a remote sensing vision-language model via cross-modal alignment,” *arXiv:2402.09816*, 2024.
- [16] Y. Zhou, L. Feng, Y. Ke, X. Jiang, J. Yan, X. Yang, and W. Zhang, “Towards vision-language geo-foundation model: A survey,” *arXiv:2406.09385*, 2024.
- [17] X. Li, C. Wen, Y. Hu, Z. Yuan, and X. X. Zhu, “Vision-language models in remote sensing: Current progress and future trends,” *IEEE Geosci. Remote Sens. Mag.*, vol. 12, no. 2, pp. 32–66, 2024.
- [18] S. Mo, M. Kim, K. Lee, and J. Shin, “S-clip: Semi-supervised vision-language learning using few specialist captions,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 61187–61212, 2023.
- [19] S. Wang, Q. Lin, X. Ye, Y. Liao, D. Quan, Z. Jin, B. Hou, and L. Jiao, “Multi-view feature fusion and visual prompt for remote sensing image captioning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [20] M. M. Al Rahhal, Y. Bazi, H. Elgibreen, and M. Zuair, “Vision-language models for zero-shot classification of remote sensing images,” *Appl. Sci.*, vol. 13, no. 22, p. 12462, 2023.
- [21] M. Czerkawski, R. Atkinson, and C. Tachtatzis, “Detecting cloud presence in satellite images using the rgb-based clip vision-language model,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 5170–5173, 2023.

- [22] Z. Yuan, Z. Xiong, L. Mou, and X. X. Zhu, "Chatearthnet: a global-scale image-text dataset empowering vision-language geo-foundation models," *Earth Syst. Sci. Data Discuss.*, vol. 17, no. 3, pp. 1245–1263, 2025.
- [23] Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, M. A. Al Zuair, and F. Melgani, "Bi-modal transformer-based approach for visual question answering in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
- [24] W. Zhang, M. Cai, T. Zhang, G. Lei, Y. Zhuang, and X. Mao, "Popeye: A unified visual-language model for multi-source ship detection from remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 20050–20063, 2024.
- [25] W. Xu, J. Wang, Z. Wei, M. Peng, and Y. Wu, "Deep semantic-visual alignment for zero-shot remote sensing image scene classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 140–152, 2023.
- [26] Z. Song, Z. Zang, Y. Wang, G. Yang, K. Yu, W. Chen, M. Wang, and S. Z. Li, "Set-clip: Exploring aligned semantic from low-alignment multimodal data through a distribution view," *arXiv:2406.05766*, 2024.
- [27] X. Tan, G. Chen, T. Wang, J. Wang, and X. Zhang, "Segment change model (scm) for unsupervised change detection in vhr remote sensing images: a case study of buildings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 8577–8580, 2024.
- [28] K. Li, X. Cao, and D. Meng, "A new learning paradigm for foundation model-based remote-sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.
- [29] T. Wei, W. Yuan, J. Luo, W. Zhang, and L. Lu, "Vlca: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning," *J. Syst. Eng. Electron.*, vol. 34, no. 1, pp. 9–18, 2023.
- [30] L. Xu, L. Zhao, W. Guo, Q. Li, K. Long, K. Zou, Y. Wang, and H. Li, "Rs-gpt4v: A unified multimodal instruction-following dataset for remote sensing image understanding," *arXiv:2406.12479*, 2024.
- [31] Y. Zhao, M. Zhang, B. Yang, Z. Zhang, J. Kang, and J. Gong, "Luojiahog: A hierarchy oriented geo-aware image caption dataset for remote sensing image-text retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 222, pp. 130–151, 2025.
- [32] J. Roberts, K. Han, and S. Albanie, "Satin: A multi-task meta-dataset for classifying satellite imagery using vision-language models," *arXiv:2304.11619*, 2023.
- [33] Y. Zhan, Z. Xiong, and Y. Yuan, "Rsvg: Exploring data and models for visual grounding on remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–13, 2023.
- [34] C. Zhang and S. Wang, "Good at captioning bad at counting: Benchmarking gpt-4v on earth observation data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 7839–7849, 2024.
- [35] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, pp. 4347–4355, 2025.
- [36] V. V. Cepeda, G. K. Nayak, and M. Shah, "GeoCLIP: Clip-inspired alignment between locations and images for effective worldwide geolocation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 8690–8701, 2023.
- [37] G. Mai, N. Lao, Y. He, J. Song, and S. Ermon, "Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 23498–23515, 2023.
- [38] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, "Diffusionsat: A generative foundation model for satellite imagery," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2023.
- [39] D. Tang, X. Cao, X. Hou, Z. Jiang, J. Liu, and D. Meng, "Crs-diff: Controllable remote sensing image generation with diffusion model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.
- [40] Y. Lin, K. Suzuki, and S. Sogo, "Practical techniques for vision-language segmentation model in remote sensing," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 48, pp. 203–210, 2024.
- [41] J. Zhang, Z. Zhou, G. Mai, M. Hu, Z. Guan, S. Li, and L. Mu, "Text2seg: Zero-shot remote sensing image semantic segmentation via text-guided visual foundation models," in *Proc. ACM SIGSPATIAL Int. Workshop AI Geographic Knowl. Discov. (AI GeoKD)*, p. 63–66, 2024.
- [42] C. Liu, R. Zhao, J. Chen, Z. Qi, Z. Zou, and Z. Shi, "A decoupling paradigm with prompt learning for remote sensing image change captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.
- [43] C. Yang, Z. Li, and L. Zhang, "Bootstrapping interactive image-text alignment for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.
- [44] H. Tang, W. Zhao, G. Hu, Y. Xiao, Y. Li, and H. Wang, "Text-guided diverse image synthesis for long-tailed remote sensing object classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
- [45] J. Ding, N. Xue, G.-S. Xia, X. Bai, W. Yang, M. Y. Yang, S. Belongie, J. Luo, M. Datcu, M. Pelillo, *et al.*, "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, 2021.
- [46] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023.
- [47] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023.
- [48] Y. Tay, M. C. Phan, L. A. Tuan, and S. C. Hui, "Learning to rank question answer pairs with holographic dual lstm architecture," in *Proc. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. (SIGIR)*, pp. 695–704, 2017.
- [49] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 26296–26306, 2024.
- [50] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2022.
- [51] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [52] X. Sun, P. Wang, Z. Yan, F. Xu, R. Wang, W. Diao, J. Chen, J. Li, Y. Feng, T. Xu, *et al.*, "Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.
- [53] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.
- [54] S. Lobry, D. Marcos, J. Murray, and D. Tuia, "Rsvqa: Visual question answering for remote sensing data," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8555–8566, 2020.
- [55] M. Rahnemoonfar, T. Chowdhury, A. Sarkar, D. Varshney, M. Yari, and R. R. Murphy, "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.
- [56] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19358–19369, 2023.
- [57] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv:2307.09288*, 2023.
- [58] H. Touvron, T. Lavigil, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.
- [59] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv:2310.09478*, 2023.
- [60] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, 2017.
- [61] Z. Yuan, W. Zhang, K. Fu, X. Li, C. Deng, H. Wang, and X. Sun, "Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
- [62] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput. Inf. Telecom. Syst. (CITS)*, pp. 1–5, 2016.
- [63] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpucaptions dataset and mlca-net for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022.
- [64] L. Bashmal, Y. Bazi, M. M. Al Rahhal, M. Zuair, and F. Melgani, "Capera: Captioning events in aerial videos," *Remote Sens.*, vol. 15, no. 8, p. 2139, 2023.
- [65] X. Zheng, B. Wang, X. Du, and X. Lu, "Mutual attention inception network for remote sensing visual question answering," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

- [66] Y. Sun, S. Feng, X. Li, Y. Ye, J. Kang, and X. Huang, "Visual grounding in remote sensing images," in *Proc. ACM Int. Conf. Multimedia*, pp. 404–412, 2022.
- [67] M. Oquab, T. Dariseti, T. Moutakanni, H. Vo, M. Szafraniec, V. Khaldov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, pp. 1–31, 2024.
- [68] M. Zhang, F. Chen, and B. Li, "Multistep question-driven visual question answering for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
- [69] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [70] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. (SIGSPATIAL)*, pp. 270–279, 2010.
- [71] D. Dai and W. Yang, "Satellite image classification via two-layer sparse coding with biased image representation," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 1, pp. 173–176, 2010.
- [72] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, 2015.
- [73] Y. Di, Z. Jiang, H. Zhang, and G. Meng, "A public dataset for ship classification in remote sensing images," in *Proc. SPIE 11155, Image Signal Process. Remote Sens. XXV*, vol. 11155, pp. 515–521, 2019.
- [74] Y. Di, Z. Jiang, and H. Zhang, "A public dataset for fine-grained ship classification in optical remote sensing images," *Remote Sens.*, vol. 13, no. 4, p. 747, 2021.
- [75] K. Li, G. Cheng, S. Bu, and X. You, "Rotation-insensitive and context-augmented object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2337–2348, 2017.
- [76] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, 2019.
- [77] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 3735–3739, 2015.
- [78] P. Zhu, L. Wen, D. Du, X. Bian, H. Fan, Q. Hu, and H. Ling, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, 2021.
- [79] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, Y. Xu, X. Ke, T. Zeng, H. Su, *et al.*, "Sar ship detection dataset (ssdd): Official release and comprehensive data analysis," *Remote Sens.*, vol. 13, no. 18, p. 3690, 2021.
- [80] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi, "Hit-uav: A high-altitude infrared thermal dataset for unmanned aerial vehicle-based object detection," *Sci. Data*, vol. 10, no. 1, p. 227, 2023.
- [81] InfiRay, "Sea-shipping," 2021.
- [82] InfiRay, "Infrared-security," 2021.
- [83] InfiRay, "Aerial-mancar," 2021.
- [84] InfiRay, "Double-light-vehicle," 2021.
- [85] C. for Optics Research and E. of Shandong University, "Oceanic-ship," 2020.
- [86] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8748–8763, 2021.
- [87] G. Hoxha and F. Melgani, "A novel svm-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
- [88] Y. Li, L. Wang, T. Wang, X. Yang, J. Luo, Q. Wang, Y. Deng, W. Wang, X. Sun, H. Li, B. Dang, Y. Zhang, Y. Yu, and Y. Junchi, "Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, pp. 1832–1849, 2024.
- [89] Y. Han, X. Yang, T. Pu, and Z. Peng, "Fine-grained recognition for oriented ship against complex scenes in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–18, 2021.
- [90] S. Wei, X. Zeng, Q. Qu, M. Wang, H. Su, and J. Shi, "Hrsid: A high-resolution sar images dataset for ship detection and instance segmentation," *IEEE Access*, vol. 8, pp. 120234–120254, 2020.
- [91] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: a learning framework for satellite imagery," in *Proc. ACM SIGSPATIAL Int. Conf. Adv. Geogr. Inf. Syst. (SIGSPATIAL)*, pp. 1–10, 2015.
- [92] Z. Zhou, S. Li, W. Wu, W. Guo, X. Li, G. Xia, and Z. Zhao, "Nasc-tg2: Natural scene classification with tiangong-2 remotely sensed imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3228–3242, 2021.
- [93] L. Zhao, P. Tang, and L. Huo, "Feature significance-based multibag-of-visual-words model for remote sensing image scene classification," *J. Appl. Remote Sens.*, vol. 10, no. 3, p. 035004, 2016.
- [94] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, 2015.
- [95] W. Zhou, S. Newsam, C. Li, and Z. Shao, "Patternet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, 2018.
- [96] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, 2017.
- [97] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, p. 111322, 2020.
- [98] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, 2018.
- [99] H. Li, H. Jiang, X. Gu, J. Peng, W. Li, L. Hong, and C. Tao, "Clrs: Continual learning benchmark for remote sensing image scene classification," *Sensors*, vol. 20, no. 4, p. 1226, 2020.
- [100] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4205–4230, 2021.
- [101] H. Li, X. Dou, C. Tao, Z. Wu, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large-scale remote sensing image classification benchmark using crowdsourced data," *Sensors*, vol. 20, no. 6, p. 1594, 2020.
- [102] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, 2017.
- [103] X. Qi, P. Zhu, Y. Wang, L. Zhang, J. Peng, M. Wu, J. Chen, X. Zhao, N. Zang, and P. T. Mathiopoulos, "Mirsnet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 337–350, 2020.
- [104] Y. Hua, L. Mou, P. Jin, and X. X. Zhu, "Multiscene: A large-scale dataset and benchmark for multiscene recognition in single aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.
- [105] A. D. G. S.A., "Airbus wind turbine patches," 2021.
- [106] Q. D. Cao and Y. Choe, "Deep learning based damage detection on post-hurricane satellite imagery," *arXiv:1807.01688*, 2018.
- [107] R. Hammell, "Ships in satellite imagery," 2018.
- [108] A. A. B. Jacques, A. B. Diallo, and E. Lord, "Towards the creation of a canadian land-use dataset for agricultural land classification," in *Proc. Can. Symp. Remote Sens.: Understanding Our World: Remote Sens. Sustain. Future*, vol. 4, p. 6, 2021.
- [109] R. Ba, C. Chen, J. Yuan, W. Song, and S. Lo, "Smokenet: Satellite smoke scene detection using convolutional neural network with spatial and channel-wise attention," *Remote Sens.*, vol. 11, no. 14, p. 1702, 2019.
- [110] O. A. Penatti, K. Nogueira, and J. A. Dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 44–51, 2015.
- [111] K. Nogueira, J. A. Dos Santos, T. Fornazari, T. S. F. Silva, L. P. Morelato, and R. d. S. Torres, "Towards vegetation species discrimination by using data-driven descriptors," in *Proc. IAPR Workshop Pattern Recognit. Remote Sens. (PRRS)*, pp. 1–6, 2016.
- [112] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 5637–5664, 2021.
- [113] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 5901–5904, 2019.

- [114] B. G. Weinstein, S. J. Graves, S. Marconi, A. Singh, A. Zare, D. Steward, S. A. Bohlman, and E. P. White, "A benchmark dataset for canopy crown detection and delineation in co-registered airborne rgb, lidar and hyperspectral imagery from the national ecological observation network," *PLoS comput. biol.*, vol. 17, no. 7, p. e1009180, 2021.
- [115] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 785–800, 2016.
- [116] J. A. Eikelboom, J. Wind, E. van de Ven, L. M. Kenana, B. Schroder, H. J. de Knegt, F. van Langevelde, and H. H. Prins, "Improving the precision and accuracy of animal population estimates with aerial image object detection," *Methods Ecol. Evol.*, vol. 10, no. 11, pp. 1875–1887, 2019.
- [117] G. Christie, N. Fendley, J. Wilson, and R. Mukherjee, "Functional map of the world," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6172–6180, 2018.
- [118] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 19730–19742, 2023.
- [119] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 16772–16782, 2023.
- [120] S. Vujasinović, S. Becker, T. Breuer, S. Bullinger, N. Scherer-Negenborn, and M. Arens, "Integration of the 3d environment for uav onboard visual object tracking," *Appl. Sci.*, vol. 10, no. 21, p. 7622, 2020.
- [121] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 4145–4153, 2017.
- [122] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. Int. Conf. Pattern Recognit. Appl. Methods (ICPRAM)*, vol. 1, pp. 324–331, 2017.
- [123] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, p. 1662, 2020.
- [124] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 549–565, 2016.
- [125] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai, "isaid: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 28–37, 2019.
- [126] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, 2021.
- [127] I. S. for Photogrammetry and R. Sensing, "2d semantic labeling contest - potsdam," 2012.
- [128] I. S. for Photogrammetry and R. Sensing, "2d semantic labeling - vaihingen data," 2012.
- [129] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.
- [130] A. Van Etten, D. Lindenbaum, and T. M. Bacastow, "Spacenet: A remote sensing dataset and challenge series," *arXiv:1807.01232*, 2018.
- [131] L. Mi, X. Dai, J. Castillo-Navarro, and D. Tuia, "Knowledge-aware text-image retrieval for remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
- [132] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 12888–12900, 2022.
- [133] J. Pan, M. Ma, Q. Ma, C. Bai, and S. Chen, "Pir: Remote sensing image-text retrieval with prior instruction representation learning," *arXiv:2405.10160*, 2024.
- [134] B. Psomas, I. Kakogeorgiou, N. Efthymiadis, G. Tolias, O. Chum, Y. Avrithis, and K. Karantzalos, "Composed image retrieval for remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 8526–8534, 2024.
- [135] C. Yang, Z. Li, and L. Zhang, "Mgimm: Multi-granularity instruction multimodal model for attribute-guided remote sensing image detailed description," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
- [136] W. Zhang, M. Cai, T. Zhang, Y. Zhuang, J. Li, and X. Mao, "Earthmarker: A visual prompting multimodal large language model for remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–19, 2024.
- [137] Y. Wenqi, C. Gong, W. Meijun, Y. Yanqing, X. Xingxing, Y. Xiwen, and H. Junwei, "Mar20: A benchmark for military aircraft recognition in remote sensing images," *Natl. Remote Sens. Bull.*, vol. 27, no. 12, pp. 2688–2696, 2024.
- [138] S. Tian, A. Ma, Z. Zheng, and Y. Zhong, "Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery," *arXiv:2011.03247*, 2020.
- [139] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 165, pp. 108–119, 2020.
- [140] D. Wang, J. Zhang, B. Du, M. Xu, L. Liu, D. Tao, and L. Zhang, "Samrs: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track*, vol. 36, pp. 8815–8827, 2023.
- [141] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, 2018.
- [142] M. Guo, M. Wu, Y. Shen, H. Li, and C. Tao, "Ifship: A large vision-language model for interpretable fine-grained ship classification via domain knowledge-enhanced instruction tuning," *arXiv:2408.06631*, 2024.
- [143] J. Ge, Y. Zheng, K. Guo, and J. Liang, "Rsteller: Scaling up visual language modeling in remote sensing with rich linguistic semantics from openly available data and large language models," *arXiv:2408.14744*, 2024.
- [144] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. I. Casas, E. B. Hanna, F. Bressand, et al., "Mixtral of experts," *arXiv:2401.04088*, 2024.
- [145] A. Sebaq and M. ElHelw, "Rsdiff: Remote sensing image generation from text using diffusion model," *Neural Comput. Appl.*, vol. 36, pp. 23103–23111, 2024.
- [146] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 8162–8171, 2021.
- [147] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 36479–36494, 2022.
- [148] Y. Xing, L. Qu, S. Zhang, J. Feng, X. Zhang, and Y. Zhang, "Empower generalizability for pansharpening through text-modulated diffusion model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.
- [149] Z. Wang, Z. Hao, Y. Zhang, Y. Feng, and Y. Guo, "Up-diff: Latent diffusion model for remote sensing urban prediction," *IEEE Geosci. Remote Sens. Lett.*, vol. 22, pp. 1–5, 2024.
- [150] Q. Cheng, Y. Xu, and Z. Huang, "Vcc-difffnet: Visual conditional control diffusion network for remote sensing image captioning," *Remote Sens.*, vol. 16, no. 16, p. 2961, 2024.
- [151] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg, "Structured denoising diffusion models in discrete state-spaces," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, pp. 17981–17993, 2021.
- [152] K. El Khoury, M. Zanella, B. Gérin, T. Godelaine, B. Macq, S. Mahmoudi, C. De Vleeschouwer, and I. B. Ayed, "Enhancing remote sensing vision-language models for zero-shot scene classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1–5, 2025.
- [153] L. Lan, F. Wang, X. Zheng, Z. Wang, and X. Liu, "Efficient prompt tuning of large vision-language model for fine-grained ship classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–10, 2024.
- [154] X. Yu, Y. Li, and J. Ma, "Diffusion-rscc: Diffusion probabilistic model for change captioning in remote sensing images," *arXiv:2405.12875*, 2024.
- [155] J. Wang, H. Sun, T. Tang, Y. Sun, Q. He, L. Lin, and K. Ji, "Leveraging visual language model and generative diffusion model for zero-shot sar target recognition," *Remote Sens.*, vol. 16, no. 16, p. 2927, 2024.
- [156] B. Zhou, H. Yang, D. Chen, J. Ye, T. Bai, J. Yu, S. Zhang, D. Lin, C. He, and W. Li, "Urbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, pp. 10707–10715, 2025.
- [157] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset

- for semantic urban scene understanding,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3213–3223, 2016.
- [158] C. Ertler, J. Mislej, T. Ollmann, L. Porzi, G. Neuhold, and Y. Kuang, “The mapillary traffic sign dataset for detection and classification on a global scale,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 68–84, 2020.
- [159] S. Zhu, T. Yang, and C. Chen, “Vigor: Cross-view image geo-localization beyond one-to-one retrieval,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3640–3649, 2021.
- [160] J. Hays and A. A. Efros, “Im2gps: estimating geographic information from a single image,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1–8, 2008.
- [161] Z. Zheng, Y. Wei, and Y. Yang, “University-1652: A multi-view multi-source benchmark for drone-based geo-localization,” in *Proc. ACM Int. Conf. Multimedia*, pp. 1395–1403, 2020.
- [162] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, “Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 213–231, 2025.
- [163] Y. Zeng, X. Zhang, and H. Li, “Multi-grained vision language pre-training: Aligning texts with visual concepts,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 25994–26009, 2022.
- [164] L. Bashmal, Y. Bazi, F. Melgani, M. M. Al Rahhal, and M. A. Al Zuair, “Language integration in remote sensing: Tasks, datasets, and future directions,” *IEEE Geosci. Remote Sens. Mag.*, vol. 11, pp. 63–93, 2023.
- [165] M. B. Bejiga, F. Melgani, and A. Vascotto, “Retro-remote sensing: Generating images from ancient texts,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 3, pp. 950–960, 2019.
- [166] T. Abdullah, Y. Bazi, M. M. Al Rahhal, M. L. Mekhalfi, L. Rangarajan, and M. Zuair, “Textrs: Deep bidirectional triplet network for matching text to remote sensing images,” *Remote Sens.*, vol. 12, no. 3, p. 405, 2020.
- [167] U. Zia, M. M. Riaz, and A. Ghafoor, “Transforming remote sensing images to textual descriptions,” *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, p. 102741, 2022.
- [168] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, “Vaa: Visual aligning attention model for remote sensing image captioning,” *IEEE Access*, vol. 7, pp. 137355–137364, 2019.
- [169] Y. Li, S. Fang, L. Jiao, R. Liu, and R. Shang, “A multi-level attention model for remote sensing image captions,” *Remote Sens.*, vol. 12, no. 6, p. 939, 2020.
- [170] R. Zhao, Z. Shi, and Z. Zou, “High-resolution remote sensing image captioning based on structured attention,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
- [171] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv:1409.1556*, 2014.
- [172] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.
- [173] J. Schmidhuber, S. Hochreiter, et al., “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [174] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017.
- [175] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, pp. 1735–1742, 2006.
- [176] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv:1807.03748*, 2018.
- [177] Y. Li, Y. Pan, T. Yao, and T. Mei, “Comprehending and ordering semantics for image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17990–17999, 2022.
- [178] H. Wang, Y. Li, H. Yao, and X. Li, “Clipn for zero-shot ood detection: Teaching clip to say no,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1802–1812, 2023.
- [179] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, “Zegclip: Towards adapting clip for zero-shot semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11175–11185, 2023.
- [180] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021.
- [181] G. Mai, K. Janowicz, B. Yan, R. Zhu, L. Cai, and N. Lao, “Multi-scale representation learning for spatial feature distributions using grid cells,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2020.
- [182] M. Larson, M. Soleymani, G. Gravier, B. Ionescu, and G. J. Jones, “The benchmarking initiative for multimedia evaluation: Mediaeval 2016,” *IEEE MultiMedia*, vol. 24, no. 1, pp. 93–96, 2017.
- [183] M. Rußwurm, K. Klemmer, E. Rolf, R. Zbinden, and D. Tuia, “Geographic location encoding with spherical harmonics and sinusoidal representation networks,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.
- [184] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1597–1607, 2020.
- [185] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singh, R. Ramamoorthi, J. Barron, and R. Ng, “Fourier features let networks learn high frequency functions in low dimensional domains,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 7537–7547, 2020.
- [186] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, and H. Maître, “Structural high-resolution satellite image indexing,” in *ISPRS TC VII Symposium-100 Years ISPRS*, vol. 38, pp. 298–303, 2010.
- [187] K. Li, R. Liu, X. Cao, D. Meng, and Z. Wang, “Segearth-ov: Towards-free open-vocabulary segmentation for remote sensing images,” *arXiv:2410.01768*, 2024.
- [188] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, “Clip-adapter: Better vision-language models with feature adapters,” *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024.
- [189] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, pp. 4444–4451, 2017.
- [190] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, “Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 145–158, 2021.
- [191] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., “Segment anything,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 4015–4026, 2023.
- [192] T. Lüddecke and A. Ecker, “Image segmentation using text and image prompts,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7086–7096, 2022.
- [193] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
- [194] W. G. C. Bandara and V. M. Patel, “A transformer-based siamese network for change detection,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 207–210, 2022.
- [195] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [196] S. Xu, C. Zhang, L. Fan, G. Meng, S. Xiang, and J. Ye, “Addressclip: Empowering vision-language models for city-wide image address localization,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 76–92, 2024.
- [197] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., “Opt: Open pre-trained transformer language models,” *arXiv:2205.01068*, 2022.
- [198] J. D. M.-W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (NAACL-HLT)*, vol. 1, pp. 4171–4186, 2019.
- [199] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu, “Beyond self-attention: External attention using two linear layers for visual tasks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5436–5447, 2022.
- [200] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [201] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 34892–34916, 2024.
- [202] C. Pang, X. Weng, J. Wu, J. Li, Y. Liu, J. Sun, W. Li, S. Wang, L. Feng, G.-S. Xia, et al., “Vhm: Versatile and honest vision language model for remote sensing image analysis,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, pp. 6381–6388, 2025.
- [203] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10684–10695, 2022.
- [204] D. P. Kingma, “Auto-encoding variational bayes,” *arXiv:1312.6114*, 2013.

- [205] Y. Liu, J. Yue, S. Xia, P. Ghamisi, W. Xie, and L. Fang, “Diffusion models meet remote sensing: Principles, methods, and perspectives,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–22, 2024.
- [206] A. Arrabi, X. Zhang, W. Sultan, C. Chen, and S. Wshah, “Cross-view meets diffusion: Aerial image synthesis with geometry and text guidance,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 5356–5366, 2025.
- [207] L. Pang, D. Tang, S. Xu, D. Meng, and X. Cao, “Hsigene: A foundation model for hyperspectral image generation,” *arXiv:2409.12470*, 2024.
- [208] Z. Yu, C. Liu, L. Liu, Z. Shi, and Z. Zou, “Metaearth: A generative foundation model for global-scale remote sensing image generation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 3, 2025.
- [209] C. Schuhmann, R. Kaczmarczyk, A. Komatsuaki, A. Katta, R. Vencu, R. Beaumont, J. Jitsev, T. Coombes, and C. Mullis, “Laion-400m: Open dataset of clip-filtered 400 million image-text pairs,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Workshop Datacentric AI*, no. FZJ-2022-00923, 2021.
- [210] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, “Microsoft coco captions: Data collection and evaluation server,” *arXiv:1504.00325*, 2015.
- [211] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, “Referitgame: Referring to objects in photographs of natural scenes,” in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, pp. 787–798, 2014.
- [212] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, “Modeling context in referring expressions,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 69–85, 2016.
- [213] J. A. Irvin, E. R. Liu, J. C. Chen, I. Dormoy, J. Kim, S. Khanna, Z. Zheng, and S. Ermon, “Teochat: A large vision-language assistant for temporal earth observation data,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2025.
- [214] B. Lin, Y. Ye, B. Zhu, C. Jiaxi, M. Ning, P. Jin, and L. Yuan, “Video-llava: Learning united visual representation by alignment before projection,” in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, pp. 5971–5984, 2024.
- [215] R. Gupta, B. Goodman, N. Patel, R. Hosfelt, S. Sajeev, E. Heim, J. Doshi, K. Lucas, H. Choset, and M. Gaston, “Creating xbd: A dataset for assessing building damage from satellite imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 10–17, 2019.
- [216] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, “S2looking: A satellite side-looking dataset for building change detection,” *Remote Sens.*, vol. 13, no. 24, p. 5094, 2021.
- [217] S. Verma, A. Panigrahi, and S. Gupta, “Qfabric: Multi-task change detection dataset,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 1052–1061, 2021.
- [218] N. Yokoya and A. Iwasaki, “Airborne hyperspectral data over chikusei,” *Space Application Laboratory, University of Tokyo*, vol. 5, no. 5, p. 5, 2016.
- [219] C. Yi, L. Zhang, X. Zhang, W. Yueming, Q. Wenchao, T. Senlin, and P. Zhang, “Aerial hyperspectral remote sensing classification dataset of xiongan new area (matiwan village),” *Natl. Remote Sens. Bull.*, vol. 24, no. 11, pp. 1299–1306, 2020.
- [220] X. Li, S. Liu, Q. Xiao, M. Ma, R. Jin, T. Che, W. Wang, X. Hu, Z. Xu, J. Wen, et al., “A multiscale dataset for understanding complex eco-hydrological processes in a heterogeneous oasis system,” *Sci. Data*, vol. 4, no. 1, pp. 1–11, 2017.
- [221] I. G. D. F. Contest, “Houston hyperspectral dataset,” 2013.
- [222] I. G. D. F. Contest, “Houston hyperspectral dataset,” 2018.
- [223] S. Sastry, S. Khanal, A. Dhakal, and N. Jacobs, “Geosynth: Contextually-aware high-resolution satellite image synthesis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 460–470, 2024.
- [224] B. Zhu, N. Lui, J. Irvin, J. Le, S. Tadwalkar, C. Wang, Z. Ouyang, F. Y. Liu, A. Y. Ng, and R. B. Jackson, “Meter-ml: A multi-sensor earth observation benchmark for automated methane source mapping,” *arXiv:2207.11166*, 2022.
- [225] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, pp. 311–318, 2002.
- [226] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summar.*, pp. 65–72, 2005.
- [227] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, pp. 74–81, 2004.
- [228] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4566–4575, 2015.
- [229] S. Holail, T. Saleh, X. Xiao, J. Xiao, G.-S. Xia, Z. Shao, M. Wang, J. Gong, and D. Li, “Time-series satellite remote sensing reveals gradually increasing war damage in the gaza strip,” *Natl. Sci. Rev.*, vol. 11, no. 9, p. nwae304, 2024.
- [230] X. Huang, Y. Cao, and J. Li, “An automatic change detection method for monitoring newly constructed building areas using time-series multi-view high-resolution optical satellite images,” *Remote Sens. Environ.*, vol. 244, p. 111802, 2020.
- [231] Y. Yang, X. Tang, Y.-M. Cheung, X. Zhang, and L. Jiao, “Sagn: Semantic-aware graph network for remote sensing scene classification,” *IEEE Trans. Image Process.*, vol. 32, pp. 1011–1025, 2023.
- [232] Z. Lin, F. Zhu, Y. Kong, Q. Wang, and J. Wang, “Srsg and s2sg: a model and a dataset for scene graph generation of remote sensing images from segmentation results,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.
- [233] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, et al., “Llama-adapter v2: Parameter-efficient visual instruction model,” *arXiv:2304.15010*, 2023.
- [234] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3836–3847, 2023.
- [235] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.
- [236] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018.
- [237] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [238] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 1395–1403, 2015.
- [239] Y. Xu, W. Xu, D. Cheung, and Z. Tu, “Line segment detection using transformers without edges,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4257–4266, 2021.
- [240] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, “Learning to simplify: fully convolutional networks for rough sketch cleanup,” *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.
- [241] S. Zhao, D. Chen, Y.-C. Chen, J. Bao, S. Hao, L. Yuan, and K.-Y. K. Wong, “Uni-controlnet: All-in-one control to text-to-image diffusion models,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, pp. 11127–11150, 2024.
- [242] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, “Attentional feature fusion,” in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 3560–3569, 2021.
- [243] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, pp. 6629–6640, 2017.
- [244] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 29, pp. 2234–2242, 2016.
- [245] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, “Ediffsr: An efficient diffusion probabilistic model for remote sensing image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2023.
- [246] F. Meng, Y. Chen, H. Jing, L. Zhang, Y. Yan, Y. Ren, S. Wu, T. Feng, R. Liu, and Z. Du, “A conditional diffusion model with fast sampling strategy for remote sensing image super-resolution,” *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024.
- [247] J. Jia, G. Lee, Z. Wang, L. Zhi, and Y. He, “Siamese meets diffusion network: Smdnet for enhanced change detection in high-resolution rs imagery,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8189–8202, 2024.
- [248] Y. Wen, Z. Zhang, Q. Cao, and G. Niu, “Transc-gd-cd: Transformer-based conditional generative diffusion change detection model,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 7144–7158, 2024.
- [249] N. Chen, J. Yue, L. Fang, and S. Xia, “Spectraldiff: A generative framework for hyperspectral image classification with diffusion models,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

- [250] L. Liu, B. Chen, H. Chen, Z. Zou, and Z. Shi, "Diverse hyperspectral remote sensing image synthesis with diffusion models," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.
- [251] C. Zhao, Y. Ogawa, S. Chen, Z. Yang, and Y. Sekimoto, "Label freedom: Stable diffusion for remote sensing image semantic segmentation data generation," in *Proc. IEEE Int. Conf. Big Data (BigData)*, pp. 1022–1030, 2023.
- [252] Z. Yuan, C. Hao, R. Zhou, J. Chen, M. Yu, W. Zhang, H. Wang, and X. Sun, "Efficient and controllable remote sensing fake sample generation based on diffusion model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
- [253] X. Zhao and K. Jia, "Cloud removal in remote sensing using sequential-based diffusion models," *Remote Sens.*, vol. 15, no. 11, p. 2861, 2023.
- [254] R. Jing, F. Duan, F. Lu, M. Zhang, and W. Zhao, "Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery," *Remote Sens.*, vol. 15, no. 9, p. 2217, 2023.
- [255] Y. Yang, T. Liu, Y. Pu, L. Liu, Q. Zhao, and Q. Wan, "Remote sensing image change captioning using multi-attentive network with diffusion model," *Remote Sens.*, vol. 16, no. 21, p. 4083, 2024.
- [256] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, 2022.
- [257] Q. Meng, W. Shi, S. Li, and L. Zhang, "Pandiff: A novel pansharpening method based on denoising diffusion probabilistic model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–17, 2023.
- [258] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "Triposr: Fast 3d object reconstruction from a single image," *arXiv:2403.02151*, 2024.
- [259] M. Czerkawski and C. Tachtatzis, "Exploring the capability of text-to-image diffusion models with structural edge guidance for multi-spectral satellite image inpainting," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.
- [260] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11976–11986, 2022.
- [261] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "Ddpm-cd: Remote sensing change detection using denoising diffusion probabilistic models," *arXiv:2206.11892*, 2022.
- [262] C. M. Challenge, "Crowdai," 2018.
- [263] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5624–5633, 2019.
- [264] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocation with aerial reference imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3961–3969, 2015.
- [265] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al., "Gemini: a family of highly capable multimodal models," *arXiv:2312.11805*, 2023.
- [266] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. Annu. Meet. Assoc. Comput. Linguist. (ACL)*, pp. 2556–2565, 2018.
- [267] OpenAI, "Chatgpt," 2024.
- [268] C. Pang, J. Wu, J. Ding, C. Song, and G.-S. Xia, "Detecting building changes with off-nadir aerial images," *Sci. China Inf. Sci.*, vol. 66, no. 4, p. 140306, 2023.
- [269] C. Wu, L. Zhang, and L. Zhang, "A scene change detection framework for multi-temporal very high resolution remote sensing images," *Signal Process.*, vol. 124, pp. 184–197, 2016.
- [270] R. Xia, J. Chen, Z. Huang, H. Wan, B. Wu, L. Sun, B. Yao, H. Xiang, and M. Xing, "Crtranssar: A visual transformer based on contextual joint representation learning for sar ship detection," *Remote Sens.*, vol. 14, no. 6, p. 1488, 2022.
- [271] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pp. 172–181, 2018.
- [272] X.-Y. Tong, G.-S. Xia, and X. X. Zhu, "Enabling country-scale land cover mapping with meter-resolution satellite imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 178–196, 2023.
- [273] X. An, J. Sun, Z. Gui, and W. He, "Coreval: A comprehensive and objective benchmark for evaluating the remote sensing capabilities of large vision-language models," *arXiv:2411.18145*, 2024.
- [274] M. S. Danish, M. A. Munir, S. R. A. Shah, K. Kuckreja, F. S. Khan, P. Fraccaro, A. Lacoste, and S. Khan, "Geobench-vlm: Benchmarking vision-language models for geospatial tasks," *arXiv:2411.19325*, 2024.
- [275] G. Machado, E. Ferreira, K. Nogueira, H. Oliveira, M. Brito, P. H. T. Gama, and J. A. dos Santos, "Airound and cv-brct: Novel multiview datasets for scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 488–503, 2020.
- [276] J. Liu, W. Zhou, H. Guan, and W. Zhao, "Similarity learning for land use scene-level change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 6501–6513, 2024.
- [277] S. Shen, S. Seneviratne, X. Wanyan, and M. Kirley, "Firerisk: A remote sensing dataset for fire risk assessment with benchmarks using supervised and self-supervised learning," in *Proc. Int. Conf. Digit. Image Comput.: Tech. Appl. (DICTA)*, pp. 189–196, 2023.
- [278] S. F. Agency, "Forest damages-larch casebearer," 2021.
- [279] CSE499DeforestationSatellite, "Deforestation-satellite-imagery dataset," 2024.
- [280] A. Shah, L. Thomas, and M. Maskey, "Marine debris dataset for object detection in planetscope imagery," 2021.
- [281] J. Shermeyer, T. Hossler, A. Van Etten, D. Hogan, R. Lewis, and D. Kim, "Rareplanes: Synthetic data takes flight," 2020.
- [282] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 4872–4881, 2021.
- [283] C. Tundia, R. Kumar, O. Damani, and G. Sivakumar, "Fpcd: An open aerial vhr dataset for farm pond change detection," *arXiv:2302.14554*, 2023.
- [284] X. Zhang, W. Yu, M.-O. Pun, and W. Shi, "Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 197, pp. 1–17, 2023.
- [285] G. Baier, A. Deschamps, M. Schmitt, and N. Yokoya, "Synthesizing optical and sar imagery from land cover maps and auxiliary raster data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.
- [286] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, et al., "So2sat lcz42: A benchmark data set for the classification of global local climate zones [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 3, pp. 76–89, 2020.
- [287] D. R. Cambrin and P. Garza, "Quakeset: A dataset and low-resource models to monitor earthquakes through sentinel-1," in *Proc. Int. IS-CRAM Conf.*, 2024.
- [288] Z. Xiong, F. Zhang, Y. Wang, Y. Shi, and X. X. Zhu, "Earthnets: Empowering artificial intelligence for earth observation," *IEEE Geosci. Remote Sens. Mag.*, vol. Early Access, pp. 2–36, 2024.
- [289] J. F. Rosser, D. G. Leibovici, and M. J. Jackson, "Rapid flood inundation mapping using social media, remote sensing and topographic data," *Nat. Hazards*, vol. 87, pp. 103–120, 2017.
- [290] J. Li, Z. He, J. Plaza, S. Li, J. Chen, H. Wu, Y. Wang, and Y. Liu, "Social media: New perspectives to improve remote sensing for emergency response," *Proc. IEEE*, vol. 105, no. 10, pp. 1900–1912, 2017.
- [291] Y. Zhai, S. Tong, X. Li, M. Cai, Q. Qu, Y. J. Lee, and Y. Ma, "Investigating the catastrophic forgetting in multimodal large language model fine-tuning," in *Conf. Parsimony Learn.*, vol. 234, pp. 202–227, 2024.
- [292] D. Zhu, Z. Sun, Z. Li, T. Shen, K. Yan, S. Ding, C. Wu, and K. Kuang, "Model tailor: Mitigating catastrophic forgetting in multi-modal large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 62581–62598, 2024.
- [293] C. Liu, K. Chen, R. Zhao, Z. Zou, and Z. Shi, "Text2earth: Unlocking text-driven remote sensing image generation with a global-scale dataset and a foundation model," *arXiv:2501.00895*, 2025.
- [294] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 3, pp. 98–106, 2023.
- [295] M. Mendieta, B. Han, X. Shi, Y. Zhu, and C. Chen, "Towards geospatial foundation models via continual pretraining," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 16806–16816, 2023.
- [296] H. Li, X. Dou, C. Tao, Z. Hou, J. Chen, J. Peng, M. Deng, and L. Zhao, "Rsi-cb: A large scale remote sensing image classification benchmark via crowdsource data," *arXiv:1705.10450*, 2017.
- [297] R. Ricci, Y. Bazi, and F. Melgani, "Machine-to-machine visual dialoguing with chatgpt for enriched textual image description," *Remote Sens.*, vol. 16, no. 3, p. 441, 2024.
- [298] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

- [299] J. Ainslie, J. Lee-Thorp, M. De Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, “Gqa: Training generalized multi-query transformer models from multi-head checkpoints,” in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, pp. 4895–4901, 2023.