Computational Visual Media
https://doi.org/10.26599/CVM.2025.9450490

Review Article

Remote sensing tuning: A survey

Dongshuo Yin[1], Ting-Feng Zhao[2], Deng-Ping Fan[2], Shutao Li[3], Bo Du[4], Xian Sun[5], and Shi-Min Hu[1] (✉)

Abstract Large models have accelerated the development of intelligent interpretation in remote sensing. Many remote sensing foundation models (RSFM) have emerged in recent years, sparking a new wave of deep learning in this field. Fine-tuning techniques serve as a bridge between remote sensing downstream tasks and advanced foundation models. As RSFMs become more powerful, fine-tuning techniques are expected to lead the next research frontier in numerous critical remote sensing applications. Advanced fine-tuning techniques can reduce the data and computational resource requirements during the downstream adaptation process. Current fine-tuning techniques for remote sensing are still in their early stages, leaving a large space for optimization and application. To elucidate the current development and future trends of remote sensing fine-tuning techniques, this survey offers a comprehensive overview of recent research. Specifically, this survey summarizes the applications and innovations of each work and categorizes recent remote sensing fine-tuning techniques into six types: adapter-based, prompt-based, reparameterization-based, hybrid methods, partial tuning, and improved tuning. In the final section, this survey suggests nine areas worth exploring in this field. Remote sensing fine-tuning methods in this survey can be found at `https://github.com/DongshuoYin/Remote-Sensing-Tuning-A-Survey`.

Keywords remote sensing; deep learning; foundation models; fine-tuning; pre-training

1 Introduction

Driven by the rapid development of big data and large model technologies, remote sensing research is undergoing a profound transformation [1, 2]. Recent advances in large language models [3] and vision foundation models [4] have underscored the importance of efficiently leveraging big data, which is as crucial as designing advanced models for deep learning tasks. Inspired by progress in natural language processing (NLP) and computer vision (CV), recent research hotspots in remote sensing are shifting from model designs to foundation models [1, 2] and fine-tuning techniques [19, 20]. Currently, there are over 50 research works on remote sensing foundation models (RSFM) covering various modalities and tasks through different pre-training methods [1]. Combined with advanced fine-tuning techniques, RSFMs will significantly enhance the performance of deep learning technologies in critical applications such as land surveying [5], agricultural monitoring [6], weather forecasting [7], and maritime navigation [8].

Fine-tuning techniques [9, 10] are designed to transfer the broad comprehension capabilities

1 BNRist, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: D. Yin, yinds@mail.tsinghua.edu.cn; S.-M. Hu, shimin@tsinghua.edu.cn (✉).
2 TKLNDST, College of Computer Science, Nankai University, Tianjin 300350, China. E-mail: T.-F. Zhao, 2112529@mail.nankai.edu.cn; D.-P. Fan, dengpfan@gmail.com.
3 College of Electrical and Information Engineering and the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha 410082, China. E-mail: shutao_li@hnu.edu.cn.
4 School of Computer Science, National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China. E-mail: dubo@whu.edu.cn.
5 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China. E-mail: sunxian@aircas.ac.cn.

Manuscript received: 2025-01-23; accepted: 2025-04-18

of foundation models to permit generalization by task-specific models. Figure 1 indicates the significant role of fine-tuning techniques in remote sensing. Research in NLP [11] and CV [12] has demonstrated that advanced fine-tuning techniques can significantly reduce data requirements and enhance model performance in few-shot scenarios. In fact, there are many few-shot objects, such as landfills [13] and power plants [14], in remote sensing images, which present challenges for remote sensing research [15]. Advanced fine-tuning techniques can accelerate the deployment of RSFM to remote sensing downstream tasks and enhance performance in few-shot scenarios. Furthermore, some fine-tuning techniques can impressively reduce memory and time requirements for hardware resources during training [16, 20], thus lowering costs for researchers conducting fine-grained studies in remote sensing.

This paper provides a comprehensive survey of recent fine-tuning work in remote sensing, including designs of fine-tuning methods for remote sensing scenarios and applications of advanced fine-tuning techniques to fine-grained remote sensing tasks. Based on the relationships between tuned parameters and models, we categorize existing remote sensing fine-tuning research into six types: adapter tuning, prompt tuning, reparameterized tuning, hybrid tuning, partial tuning, and improved tuning. Figure 2 presents the structural differences between five of these categories, omitting hybrid tuning. Table 1 summarizes the basic concepts of these six categories and associated work in remote sensing. The main contributions of this paper are as follows:

- a systematic review of fine-tuning research in remote sensing and a summary of the technical design and innovative application of each study, enabling researchers to quickly grasp the development of fine-tuning techniques in remote sensing,
- a classification of fine-tuning research in remote sensing based on the relationship between tuned parameters and the model, helping researchers to clearly understand the development trajectory of fine-tuning technologies in remote sensing, and
- suggestions for future research directions for fine-tuning in remote sensing.

2 Background

2.1 Remote sensing foundation models

Earth observation is important to contemporary society [13, 17]. Remote sensing technology [18]

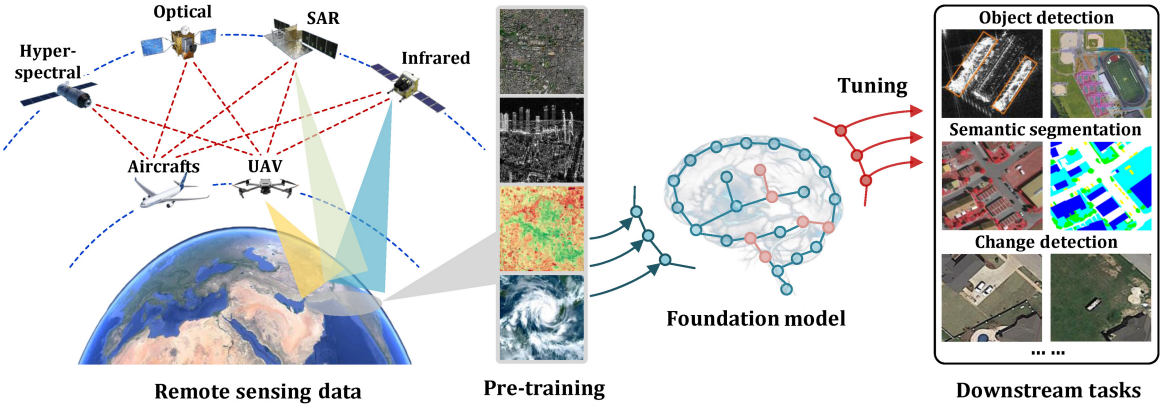

Fig. 1 In the era of large models, fine-tuning serves as the bridge connecting remote sensing foundation models and downstream tasks, and also acts as a catalyst for optimizing many remote sensing applications.

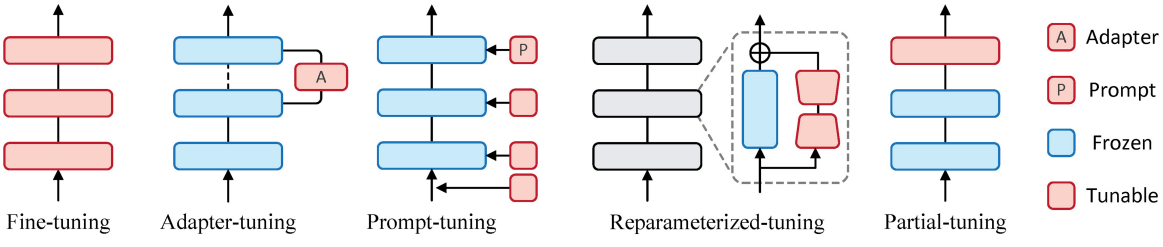

Fig. 2 Structural differences between the five paradigms.

Table 1 Brief description of the six categories of remote sensing fine-tuning

Category	Description	Related works
Adapter tuning	Inserting micro adapter structures into the backbone network to adapt it to downstream tasks	[19–34]
Prompt tuning	Adding learnable patches to the input or implicit layers to change the feature distribution of the backbone network	[35–49]
Reparameterized tuning	Fine-tuning the low-dimensional space of some parameters in the backbone network to fit downstream tasks without adding extra structures	[50–58]
Hybrid tuning	Methods based on at least two fine-tuning paradigms	[59–64]
Partial tuning	Fine-tuning some of the parameters in the backbone network	[65–68]
Improved tuning	Optimizing traditional fine-tuning in some way	[69–72]

not only facilitates a comprehensive understanding of changes on the Earth's surface, but also plays an essential role in disaster prevention, environmental monitoring, urban planning, and agricultural surveillance. As the volume of remote sensing data proliferates and the complexity of tasks intensifies, traditional processing methods are increasingly insufficient to meet the demands of modern remote sensing applications. Remote sensing foundation models (RSFM) [1, 2], developed through the integration of extensive multimodal data, enhance both the efficiency of training processes and the accuracy of predictions for downstream tasks. Furthermore, RSFMs enable the sharing and transfer of knowledge across diverse tasks, thereby significantly increasing the versatility and practicality of remote sensing applications.

In recent years, the field of remote sensing has experienced a period of rapid growth in the development of RSFMs. By June 2024, over 51 studies on RSFMs had been documented [1, 2]. Table 2 presents a summary of the RSFMs reported in recent reviews [1]. Representative works include RVSA [73], which focuses on architectural designs tailored for remote sensing images, RSP [74] and SatMAE [75] for supervised or self-supervised pre-training, MTP [76] for multitask learning, HyperSIGMA [77] for hyperspectral image processing, and GeoChat [78] for vision-text cross-modal understanding. These existing models exhibit considerable diversity in several ways, including architecture, data, tasks, and training paradigms. The relationship between RSFMs and fine-tuning methods is indispensable. During the pre-training phase, RSFMs assimilate knowledge from vast amounts of remote sensing data, thereby acquiring robust general understanding capabilities. Due to the variability of specific application scenarios, RSFMs often require fine-tuning with downstream data to optimize their performance for specific tasks. Fine-tuning methods are tools to refine RSFMs, enabling them to better address the particular needs of diverse tasks, thereby enhancing the models' accuracy and effectiveness. Consequently, it is crucial to study fine-tuning methods for remote sensing tasks. Better remote sensing tuning methods not only fully leverage the potential of foundation models, but also promote the extensive application and in-depth development of remote sensing technology.

2.2 Fine tuning

Prior to the emergence of large models, numerous traditional fine-tuning approaches were already in practice. Huang et al. [79] were the first to introduce the paradigm of pre-training combined with fine-tuning for synthetic aperture radar (SAR) image object classification, marking a significant step in applying fine-tuning techniques to remote sensing tasks. Subsequently, Liu et al. [80] developed a fine-tuning method that successfully facilitated the transfer of learning from computer vision to remote sensing. Researchers such as Bazi et al. [81], Zhang et al. [82], and Wu et al. [83] advanced this field by designing task-specific loss functions, thereby enhancing both the efficiency and performance of models during the fine-tuning process for remote sensing applications. Moreover, Huang et al. [84] and Zhao et al. [85] proposed highly efficient two-stage fine-tuning strategies tailored for few-shot object detection. Kim et al. [86] rethought the fine-tuning of feature backbones for remote sensing object detection.

Table 2 Existing remote sensing foundation models. SC/SS/OD/CD represent scene classification, semantic segmentation, object detection, and change detection, respectively. This table is adapted from Ref. [1]. The first column shows when they were first made public

Date	Architecture	Model name	Publication	Tasks
2021 Jun	ResNet-50	CMC-RSSR [91]	CVPRW2021	SC
2021 Oct	ResNet-50	SeCo [92]	ICCV2021	SC/CD
2021 Oct	ResNet-50	GeoKR [93]	TGRS2022	SC/SS/OD
2021 Dec	ResNet-34	MATTER [94]	CVPR2022	SC/SS/CD
2022 Mar	ResNet-50	GASSL [95]	ICCV2021	SC/SS/OD
2022 May	ViTAEv2-S	RSP [74]	TGRS2023	SC/SS/OD/CD
2022 Jun	ViT-S/8	DINO-MM [96]	IGRSS2022	SC
2022 Jun	Swin Transformer	Scheibenreif et al. [97]	CVPRW2022	SC/SS
2022 Jul	ViT/Swin Transformer	RingMo [98]	TGRS2023	SC/SS/OD/CD
2022 Aug	ResNet-50	GeCO [99]	TGRS2022	SC/SS/OD
2022 Sep	BYOL	RS-BYOL [100]	JSTAR2022	SC/SS
2022 Nov	ViT-B	CSPT [101]	RS2022	SC/OD
2022 Nov	ViT	RVSA [102]	TGRS2023	SC/SS/OD
2023 Jan	MAE-based Framework	SatMAE [75]	NeurIPS2022	SC/SS
2023 Apr	TOV	TOV [103]	JSTAR2023	SC/SS/OD
2023 Apr	Teacher–student Self-distillation	CMID [104]	TGRS2023	SC/SS/OD/CD
2023 Jun	CACo	CACo [105]	CVPR2023	SC/SS/CD
2023 Jun	ResNet-18	IaI-SimCLR [106]	CVPRW2023	SC
2023 Jul	EVA/Vicuna/Q-former	RSGPT [107]	arXiv2023	Visual-Language
2023 Aug	Teacher–Student	GFM [108]	ICCV2023	SC/SS/CD
2023 Aug	Swim Transformer	SatLasPretrain [109]	ICCV2023	SC/SS
2023 Sep	Multi-Branch	RingMo-Sense [110]	TGRS2023	SS
2023 Sep	ViT	Scale-MAE [111]	ICCV2023	SC/SS
2023 Sep	CNN-Transformer	RingMo-lite [112]	arXiv2023	SC/SS/OD/CD
2023 Sep	Multimodal SSL	DeCUR [113]	arXiv2023	SC/SS
2023 Oct	MSFE+MMFH	Feng et al. [114]	IGRSS2023	SC/SS/OD/CD
2023 Oct	ViT	FG-MAE [115]	JSTAR2024	SC/SS
2023 Oct	ViTAEv2-S	SAMRS [116]	NeurIPS2023	SS
2023 Nov	ViT	Prithvi [117]	arXiv2023	SS
2023 Nov	Multimodal Encoder	CROMA [118]	NeurIPS2023	SC/SS
2023 Nov	CLIP-ViT/Vicuna-v1.5	GeoChat [78]	CVPR2024	Visual-Language
2023 Dec	ViT	USat [119]	arXiv2023	SC
2024 Jan	EVA-CLIP/LLaMA2-chat	SkyEyeGPT [52]	ISPRS2024	Visual-Language
2024 Jan	ViT-B	Cross-Scale MAE [120]	arXiv2023	SC/SS
2024 Jan	Unet+Transformer	U-BARN [121]	JSTAR2024	SC/SS
2024 Jan	Autoregressive Transformer	EarthPT [122]	arXiv2023	SC
2024 Jan	Teacher–Student Network	GeRSP [123]	TGRS2024	SC/SS/OD
2024 Jan	Dual-Branch	SwiMDiff [124]	TGRS2024	SC/CD
2024 Jan	Generative ConvNet	SMLFR [125]	TGRS2024	SS/OD
2024 Feb	3D GPT	SpectralGPT [126]	TPAMI2024	SC/SS/CD
2024 Feb	MAE-based Framework	Presto [127]	arXiv2024	SS
2024 Mar	SatMAE	SatMAE++ [128]	CVPR2024	SC
2024 Mar	Joint-Embedding Predictive Architecture	SAR-JEPA [129]	ISPRS2024	SC
2024 Mar	ViT	FoMo-Bench [130]	arXiv2023	SC/SS/OD
2024 Mar	Factorized Multi-Modal Spatiotemporal Encoder	SkySense [131]	CVPR2024	SC/SS/OD/CD
2024 Mar	Multi-Modules	UPetu [132]	TGRS2024	SC/SS/CD
2024 Apr	Swin Transformer	msGFM [133]	CVPR2024	SC/SS
2024 Apr	DINO	DINO-MC [134]	CVPRW2024	SC/CD
2024 May	OFA-Net	OFA-Net [135]	IGARSS2024	SC/SS
2024 May	Shared Encoder, Task-Specific Decoders	MTP [76]	JSTAR2024	SC/SS/OD/CD
2024 May	ViT	BFM [136]	JSTAR2024	SS/OD
2024 May	MP-MAE	MMEarth [137]	arXiv2024	SC/SS
2024 May	ViT	CtxMIM [138]	arXiv2023	SC/SS/OD
2024 May	HiViT	SARATR-X [139]	arXiv2024	SC/OD
2024 May	LeMeViT	LeMeViT [140]	IJCAI2024	SC/SS/OD/CD
2024 Jun	Dynamic OFA	DOFA [141]	arXiv2024	SC
2024 Jun	ViT	HyperSIGMA [77]	arXiv2024	SC/SS/OD/CD
2024 Mar	CLIP-ViT/Vicuna-v1.5	SkySenseGPT [131]	arXiv2024	Visual-Language

Drawing inspiration from knowledge distillation [87], Zhang et al. [88] utilized the insights gained from large datasets through a teacher model to accelerate the fine-tuning of a student model with limited data. Additionally, Zhang et al. [89] applied fine-tuning techniques to UAV-based forest image classification, while Zhang et al. [90] optimized fine-tuning methods for diffusion models to significantly improve dehazing performance. Our paper primarily investigates new fine-tuning paradigms or optimization methods that have emerged in remote sensing in recent years.

2.3 Related surveys

In this section, we review relevant survey works. Ref. [142] introduced the development of parameter-efficient fine-tuning (PEFT) across general domains. Refs. [11, 143, 144] cataloged PEFT studies in natural language processing. Refs. [12, 145] examined fine-tuning methods within the domain of general vision research. Refs. [9, 146] systematically evaluated the performance and efficiency of various PEFT methods in language models. Ref. [147] discussed reparameterization techniques in language models. Ref. [148] outlined studies on prompt and adapter methods in vision-language tasks. Furthermore, Ref. [149] briefly explored the application of PEFT in areas such as text generation, medical imaging, protein modeling, and speech synthesis. Existing fine-tuning surveys primarily focused on domains other than remote sensing. To our knowledge, this is the first comprehensive survey on fine-tuning within the field of remote sensing. To clarify the development of existing fine-tuning techniques in remote sensing, this survey categorizes these methods and further considers their technical approaches. Also, this paper outlines the developmental history of typical technologies along a timeline and presents future directions.

2.4 Notation and formalization

2.4.1 Notation

This section represents the processes of pre-training and fine-tuning using formulae. After the fine-tuning formula, we explain the specific forms of fine-tuning paradigms within the formula. Table 3 presents the notation used, with explanations of the symbols. In remote sensing, D_{pre} generally refers to large datasets or their combinations. F_{pre} can be models like ResNet [150], ViT [151], Swin Transformer [152], etc. $T_{\text{pre}}(\cdot)$ includes supervised and self-supervised methods.

Table 3 Symbols used in formalization

Symbol	Definition
D_{ft}	Data for fine-tuning
D_{pre}	Data for pre-training
F_{ft}	Framework for fine-tuning
F_{pre}	Framework for pre-training
M_{ft}	Fine-tuned model
M_{pre}	Pre-trained model
M_f	Frozen part of the model
M_t	Fine-tuned part of the model
P	All parameters in fine-tuning
P_0	Parameters for pre-training
P_N	New parameters in fine-tuning
P_N^F	Parameters in the new framework
P_N^{PEFT}	Parameters of the PEFT module
$T_{\text{ft}}(\cdot)$	Training algorithm for fine-tuning
$T_{\text{pre}}(\cdot)$	Training algorithm for pre-training

M_{pre} may contain some classification head parameters and mask parameters, which may be discarded by some fine-tuning algorithms. D_{ft} is the target data. F_{ft} can be detection, segmentation, or multimodal frameworks. P_N includes adapters, prompts, LoRA modules, and parameters in neck/head parts. $T_{\text{ft}}(\cdot)$ fine-tunes some or all pre-trained parameters, depending on the fine-tuning paradigm. M_{ft} can be used for downstream tasks.

2.4.2 Pre-training

With the above notation, the training process can be represented as

$$M_{\text{pre}} = T_{\text{pre}}(F_{\text{pre}}(P_0), D_{\text{pre}}) \tag{1}$$

where P_0 represents parameters for pre-training.

2.4.3 Fine-tuning

During the fine-tuning process, the parameters in the pre-trained model can be divided into a frozen part M_f and a fine-tuned part M_t, thus $M_{\text{pre}} = \{M_f, M_t\}$. The new parameters in fine-tuning can be divided into the parameters of the parameter-efficient fine-tuning module P_N^{PEFT} and the parameters within the new framework P_N^{F} (such as head, neck, etc.). Similarly, all parameters in fine-tuning can be divided into a frozen part P_f and a fine-tuned part P_t. Based on the above definitions, the fine-tuning process can be uniformly represented by Formula (2):

$$\begin{aligned} M_{\text{ft}} &= T_{\text{ft}}[F_{\text{ft}}(P), D_{\text{ft}}] \\ &= T_{\text{ft}}[F_{\text{ft}}(P_f, P_t), D_{\text{ft}}] \\ &= T_{\text{ft}}\{F_{\text{ft}}[(P_N, M_t), M_f], D_{\text{ft}}\} \\ &= T_{\text{ft}}\left\{F_{\text{ft}}[\left(P_N^{\text{PEFT}}, P_N^F, M_t\right), M_f], D_{\text{ft}}\right\} \end{aligned} \tag{2}$$

In fine-tuning paradigms, methods based on adapters, prompts, and reparameterization generally fix M_{pre} during training and fine-tune only P_N. Their training process can be simplified to the representation in Eq. (3):

$$M_{\text{ft}} = \text{FT}_{D_{\text{ft}}, M_{\text{pre}}}\left(P_N^{\text{PEFT}}, P_N^F\right) \tag{3}$$

where $\text{FT}(\cdot)$ represents the simplified forward and training processes. Partial tuning generally does not include additional PEFT modules, and its training process can be simplified as Formula (4):

$$M_{\text{ft}} = \text{FT}_{D_{\text{ft}}, M_f}\left(M_t, P_N^F\right) \tag{4}$$

Hybrid tuning combines the above methods.

3 Remote sensing tuning

3.1 Overview

Here, we first consider the timeline and technological development of fine-tuning techniques in remote sensing. Figure 3 shows the timeline of development of typical fine-tuning methods in remote sensing, including those introducing new paradigms and those obtaining significant conclusions. Figure 4 presents a taxonomical tree diagram, which introduces the developmental routes of different fine-tuning paradigms in remote sensing in terms of aspects such as design, optimization, and application. The remainder of this section will summarize and introduce existing fine-tuning work in remote sensing.

3.2 Adapter tuning

3.2.1 Concepts

Adapter tuning is a representative parameter-efficient fine-tuning method in the era of large models [144]. Adapters have their early origins in NLP. Houlsby et al. [153] proposed a simple yet effective simple adapter architecture that achieved excellent performance on several typical NLP tasks. Owing to its exceptional generalization capability, simple adapters have garnered significant interest within the realms of computer vision [154], remote sensing [19], medical imaging [155], and multi-modal learning [156]. The simple adapter's structure mainly consists of two linear projection layers, a nonlinear activation layer, and a skip-connection. Assuming that the input of the adapter is x and its output is y, the computational process of the adapter can be represented as Eq. (5):

$$y = U(\sigma(D(x))) + x \tag{5}$$

where $D(\cdot)$ and $U(\cdot)$ represent down-projection and up-projection respectively, and $\sigma(\cdot)$ denotes the activation function. The projection process can be expressed as

$$y = Wx + b \tag{6}$$

which means most parameters are in the W matrix.

Adapters are usually inserted into the pre-trained backbone network during training and inference phases. During the fine-tuning process, most works fix the parameters of the backbone network outside the adapters, while other work argues that fine-tuning all parameters gives better results. Both paradigms occur in the methods presented. In remote sensing, existing adapter fine-tuning techniques can be primarily divided into three branches. Methods like AiRs [19] and ACTNet [22] focus on designing more efficient adapters for remote sensing scenarios. Methods like SCD-SAM [21] and TEA [20] optimize existing adapter structures or tuning frameworks. Methods

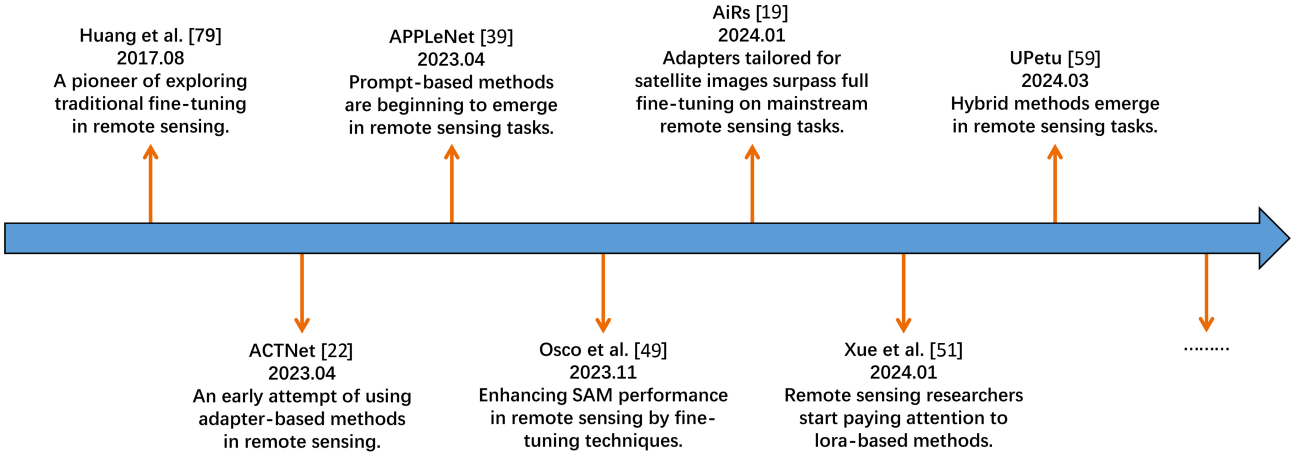

Fig. 3 Timeline of the emergence of several remote sensing fine-tuning paradigms.

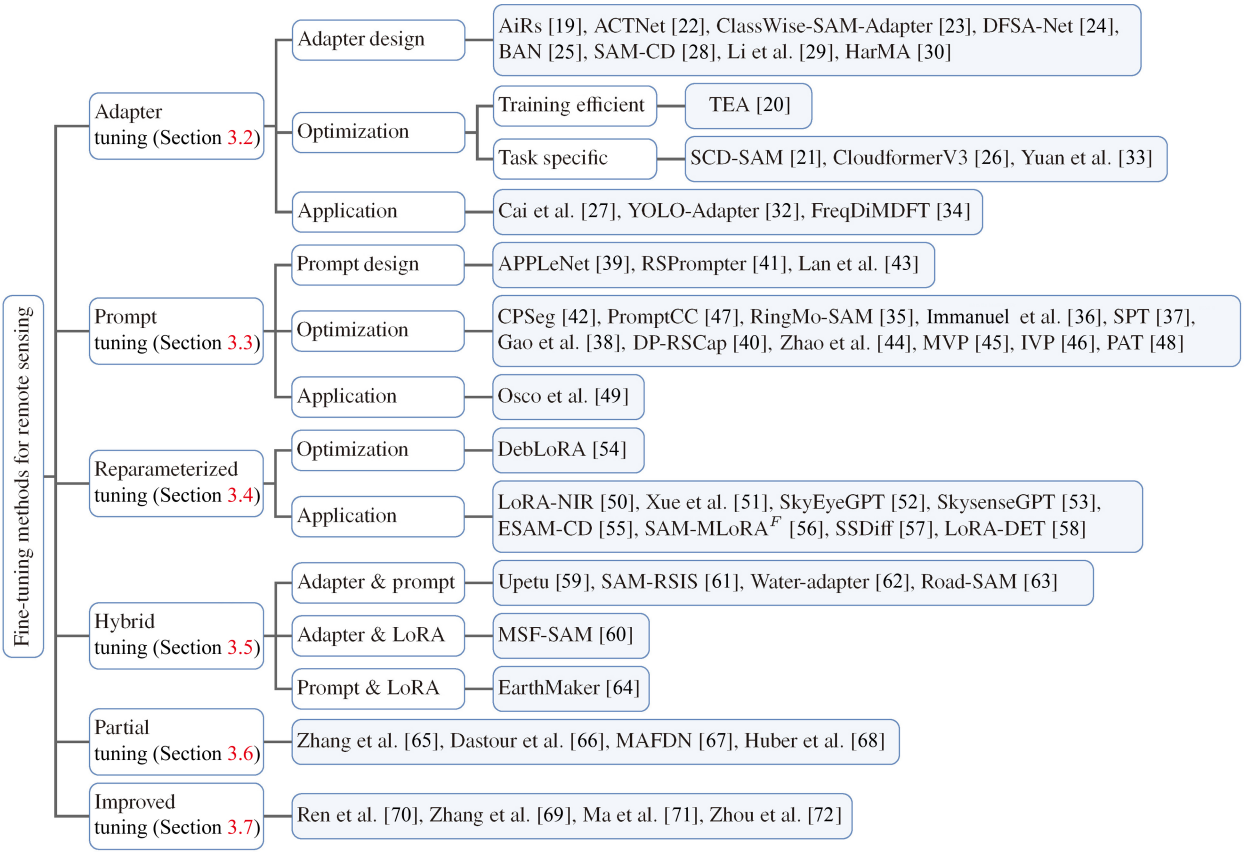

Fig. 4 Taxonomy of fine-tuning methods in remote sensing.

such as YOLO-adapter [32] and FreqDiMDFT [34] directly apply existing adapter structures to remote sensing tasks. We next present existing adapter tuning techniques in remote sensing.

3.2.2 AiRs

AiRs [19] is a customized adapter scheme for common remote sensing vision tasks, which goes beyond full fine-tuning on tasks such as object detection, semantic segmentation, scene classification, etc. As Fig. 5 shows, AiRs includes two independent adapter structures for remote sensing tasks, a spatial context adapter (SCA) and a semantic response adapter (SRA). AiRs freezes the pre-trained parameters during the training process. The computational process of the two adapter structures of AiRs can be represented as Eqs. (7) and (8):

$$h_0, h_1 = S(W_{\text{down}}(y)) \tag{7}$$

$$y = W_{\text{up}}(\gamma\phi\left(C\left(F_{\text{SCA/SRA}}\left(h_1\right), h_0\right)\right)) + y \tag{8}$$

where γ is a scaling factor, and $S(\cdot)$ and $C(\cdot)$ are the split and concatenation operations. h_0, h_1 are two split matrices, and the activation function is GeLU. $F_{\text{SCA/SRA}}(\cdot)$ is the internal operation of SCA or SRA, and $W(\cdot)$ denotes a down or up projection operation. AiRs reduces the parameter size of adapters by the split and concatenation operations.

Specifically, SCA introduces a separable convolution structure to change the bias of the pre-trained backbone network's understanding of remote sensing knowledge. The specific calculations used in its internal structure are as Eq. (9):

$$F_{\text{SCA}}(h_1) = sF_{\text{DWConv}}\left(h_1, W_k\right) + h_1 \tag{9}$$

where s is a scaling factor, W_k is the parameter of separable convolution, and $F_{\text{DWConv}}(\cdot)$ is the separable convolution operation. SRA utilizes an inverted bottleneck-like operator to optimize the feature extraction process on the pre-trained layer, and the computational performed by its internal structure is as Eq. (10):

$$F_{\text{SRA}}\left(h_1\right) = W_{\text{down}}^{\text{SRA}}(s\phi\left(W_{\text{up}}^{\text{SRA}}\left(LN\left(h_1\right)\right)\right)) + h_1 \tag{10}$$

where s is a scaling factor and $W_{\text{down/up}}^{\text{SRA}}$ is a small bottleneck structure inside the SRA structure.

AiRs systematically and experimentally illustrates that full fine-tuning is no longer optimal for remote

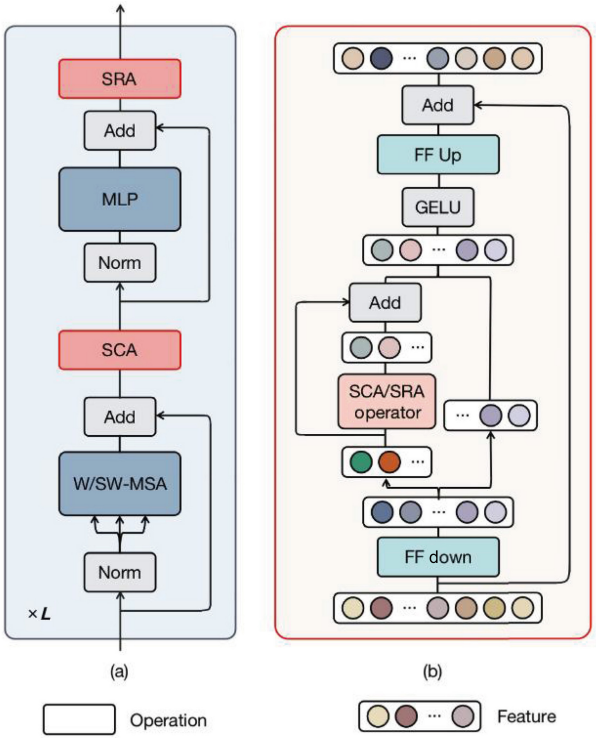

Fig. 5 Framework of AiRs. (a) Insertion position of the SCA and SRA. (b) Overview of AiRs modules. Reproduced with permission from Ref. [19], © IEEE 2024.

sensing visual tasks. Furthermore, the authors also claim that AiRs performs less well on dense small object scenes and datasets with fine-grained semantic labels.

3.2.3 TEA

Hu et al. [20] presented a training-efficient fine-tuning paradigm, TEA, which is able to achieve impressive performance while saving more than 57% of training memory. TEA avoids the gradient computation process for the pre-trained layer by use of a gradient highway, which is motivated by E^{3}VA [16]. Unlike E^{3}VA, TEA tailors the adapter structure and fine-tuning process specifically for the characteristics of remote sensing imagery, thereby improving the performance of training-efficient tuning methods on remote sensing tasks. TEA replaces the traditional adapter with GhostNetV2 [157]. Hu et al. analyzed the redundancy of the traditional adapter fine-tuning paradigm in gradient backpropagation and designed a parallel framework based on the gradient chain rule. The new framework can prevent adapter gradients from passing into the pre-trained layer. The authors also elaborate on this efficient training paradigm through mathematical derivations.

Furthermore, a top–down guidance mechanism is provided for TEA, considering the performance of the parallel structure. The authors argue that background information in remote sensing images helps the model to understand foreground objects. In other words, the performance of adapter tuning is enhanced if the model can capture high-level semantic information from the image using a larger receptive field. TEA borrows the idea of methods such as FPN [158]. Features from multiple stages of the backbone are fused and sent to the adapters for learning. This approach successfully improves the performance of TEA on several remote sensing visual tasks, including object detection, rotated object detection, semantic segmentation, and image classification.

3.2.4 SCD-SAM

Remote sensing images often exhibit substantial inter-class similarity and significant intra-class variation, which can lead to pronounced performance degradation when directly applying the segment anything model (SAM) [159] to semantic change detection (SCD) tasks in remote sensing imagery. To alleviate this limitation, SCD-SAM [21] introduces a semantic adapter designed to aggregate semantically oriented information about changing objects, thereby pioneering the adaptation of MobileSAM [160] to the SCD domain.

Specifically, SCD-SAM uses semantic adapters to optimize input processing, refine overlapping patch embeddings, and integrate multi-scale semantic features. The input image size is reduced by half from its original dimensions of 1024×1024 with a corresponding reduction in the patch stride, thus augmenting the model's capacity for fine-grained semantic understanding and its sensitivity to semantic information along patch boundaries. The adapter restructures the original feature representation derived from MobileSAM to better preserve the spatial and semantic information necessary for downstream tasks. In addition, SCD-SAM adapts the MobileSAM encoder into a four-stage architecture, enabling the extraction of multi-scale semantic change features, enhancing its capacity to detect and differentiate semantic variations for diverse land cover types. The necessity and efficacy of the proposed semantic adapter are rigorously demonstrated through ablation experiments, which highlight its contributions to performance improvements.

Experimental results underscore the effectiveness of SCD-SAM, showcasing its potential to migrate powerful vision foundation models to remote sensing applications and significantly advancing the field of SCD.

3.2.5 ACTNet

ACTNet [22] introduces ResAttn to accelerate training convergence of the Swin Transformer [152] when segmenting high-resolution remote sensing images. ResAttn uses a dual-branch attention mechanism to model the interdependencies between feature sets, thereby improving the model's global representation capabilities. This mechanism also mitigates the risk of vanishing gradients in deeper networks and enhances the model's sensitivity to small objects.

Integrating the concept of residual structure, ResAttn first generates tokens and queries by combining the outputs from the current Swin Transformer block and the preceding ResAttn module. These features are then fused and fed into a self-attention module, employing a multi-head self-attention mechanism. The outputs from the two branches are concatenated and subsequently passed through a feed-forward network (FFN). The FFN consists of two linear layers interspersed with activation functions, which augment the query representations to further refine feature embeddings. To preserve the structural integrity of the Swin Transformer, ResAttn employs the same downsampling strategy, thereby optimizing computational efficiency within the adapter module.

Experimental results demonstrate that ResAttn incurs only a minimal increase in training time, yet it significantly enhances the model's performance on downstream tasks. However, it is noteworthy that the paper also highlights that ACTNet's inference time is longer than the baseline.

3.2.6 ClassWise-SAM-Adapter

ClassWise-SAM-Adapter (CWSAM) [23] introduces a lightweight adapter designed for parameter-efficient fine-tuning of the SAM model on land cover classification tasks using satellite-borne SAR images.

Inspired by AdaptFormer [161], CWSAM integrates several simple yet effective adapters into individual blocks within the Vision Transformer (ViT) architecture. A lightweight adapter is incorporated after the multi-head attention (MHA) module, introducing a skip connection that merges with the initial features prior to the second MLP sub-block. Within the second MLP sub-block, an additional adapter operates in parallel with the original MLP layer to refine and enhance the output representations. Both adapters are simple adapters, denoted Adapter$(\cdot)$. The feature extraction process of the Transformer block in the i-th layer can be represented as

$$\begin{cases} x_i' = \text{Adapter}(\text{Attention}(\text{LN}(x_{i-1}))) + x_{i-1} \\ x_i = \text{MLP}_{\text{sam}}(\text{LN}(x_i')) + \text{Adapter}(\text{LN}(x_i')) + x_i' \end{cases} \tag{11}$$

Here x_i and x_{i-1} denote the output features of consecutive Transformer blocks, while x_i' represents the intermediate features from the first sub-block in the Transformer block. The function Attention$(\cdot)$ refers to the attention mechanism within the Transformer block, and LN$(\cdot)$ represents the Layer Normalization operation. $\text{MLP}_{\text{sam}}(\cdot)$ is the MLP block within the Transformer block.

By embedding these adapter structures, CWSAM enables the SAM model to transfer its representation capabilities from natural scenes to SAR imagery, thereby providing meaningful features for subsequent mask decoding.

3.2.7 DFSA-Net

DFSA-Net [24] introduces the disentangled low-rank adapter (DLA) to address two challenges in generalized few-shot semantic segmentation (GFS-Seg): degradation of base class performance and overfitting to new classes during fine-tuning. This approach safeguards the pretrained base parameters to mitigate base class degradation, while adaptively fine-tuning low-rank parameters to accommodate the representations of novel classes.

The DLA architecture ensures that the parameters of the backbone network and the pretrained base branch remain frozen, introducing a trainable adapter branch that facilitates efficient adaptation. This adapter branch restructures conventional convolutional layers by replacing the FPN, FAM, and semantic decoder modules by a low-rank adaptation layer. The learnable kernel weights $\widetilde{K}$ are approximated via a low-rank matrix decomposition, which can be mathematically expressed as

$$\widetilde{K} = \text{reshape}(X_1 Y_1 \odot X_2 Y_2) \tag{12}$$

Here, $X_1 \in \mathbb{R}^{C_{\text{out}} \times r_1}$, $X_2 \in \mathbb{R}^{C_{\text{out}} \times r_2}$, $Y_1 \in \mathbb{R}^{r_1 \times (C_{\text{in}} \times k \times k)}$, and $Y_2 \in \mathbb{R}^{r_2 \times (C_{\text{in}} \times k \times k)}$. The low-rank approximation of the learnable kernel $\widetilde{\boldsymbol{K}}$ is computed in $\mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times k \times k}$ as $r_1, r_2 \ll \min(C_{\text{in}}, C_{\text{out}})$.

Then output features can be computed as

$$F_{\text{output}} = \text{Conv}(\widetilde{\boldsymbol{K}}, F_{\text{input}}) + \text{Conv}(\boldsymbol{K}, F_{\text{input}}) \quad (13)$$

where $F_{\text{input}} \in \mathbb{R}^{C_{\text{in}} \times h \times w}$ and $F_{\text{output}} \in \mathbb{R}^{C_{\text{out}} \times h \times w}$.

To enhance the model's capacity to learn new classes, the DLA integrates a salience suppression mechanism during fine-tuning. This technique reduces the foreground salience of the base class while enabling the extraction of foreground features for the novel class branch by computing the difference between the predicted output and the foreground mask of the base class. This process is mathematically formulated as

$$\mathbf{M}_{\text{fg}}^{\text{novel}} = \delta_{\text{ada}}\big(F'_{\text{ada}}\big) - \delta\big(F'\big) \quad (14)$$

where δ_{ada} and F'_{ada} correspond to the adapter branch for the novel classes, while δ and F' are derived from the original convolutional branch for the base classes.

3.2.8 BAN

Li et al. [25] introduced the bi-temporal adapter network (BAN), a novel approach that harnesses knowledge priors from foundational models to enhance change detection tasks. Specifically, BAN extracts high-level features from frozen foundational models (e.g., CLIP [162]) and employs a bridging module to efficiently select, align, and integrate these features into the bi-temporal adapter branch (Bi-TAB).

The bi-temporal adapter branch (Bi-TAB) presents a model-agnostic framework designed to extract both domain-specific and task-specific features. In addition to utilizing the original dual-time-phase images, the base model contributes an expansive generalized feature library, from which Bi-TAB can dynamically extract pertinent features. Using its Siamese architecture, Bi-TAB seamlessly integrates information from the base model into its backbone for the corresponding time phase, thereby enabling the use of virtually any remote sensing change detection (RSCD) model as a foundational element for Bi-TAB. To effectively select and align generic features with domain- or task-specific ones, the bridging module resamples the generalized domain knowledge via cross-domain dot-product attention, and subsequently injects the refined features into the RSCD domain representations.

3.2.9 CloudformerV3

CloudformerV3 [26] utilizes a multi-scale adapter to incorporate a priori knowledge from diverse channels into the model's backbone, thereby augmenting its ability to capture both foundational information and intricate multi-scale details within remote sensing imagery for cloud detection.

The conventional adapter [163] consists of three core components: spatial prior module, injector, and extractor. However, traditional adapters are ill-suited to hierarchical network architectures, limiting the model's capacity to effectively extract multi-scale features. To overcome this constraint, this work enhances the model's multi-scale perceptual capability by introducing a downsampling layer between the extractor and injector of the conventional adapter. Specifically, the features extracted by the extractor are first partitioned into multiple resolution levels via split and reshape operations, denoted Split(·) and Reshape(·). These features are subsequently downsampled using patch merging layers [164], denoted PatchMerging(·), which reduce both their spatial length and width. Finally, the downsampled features are flattened and concatenated by the functions Flatten(·) and Concat(·), preparing them for subsequent fusion. This modified adapter structure allows the model to more effectively assimilate image information relevant to cloud detection. Furthermore, the use of adapters facilitates the seamless integration of downstream tasks and supplementary a priori information into the model. The whole process can be formulated as

$$F_{\text{sp1}}^{i}, F_{\text{sp2}}^{i}, F_{\text{sp3}}^{i} = \text{Reshape}\Big(\text{Split}(F_{\text{sp}}^{i})\Big) \quad (15)$$

$$\hat{F}_{\text{sp}j}^{i} = \text{PatchMerging}(F_{\text{sp}j}), \quad j \in \{1, 2, 3\} \quad (16)$$

$$\hat{F}_{\text{sp}}^{i} = \text{Flatten}\Big(\text{Concat}(\hat{F}_{\text{sp1}}^{i}, \hat{F}_{\text{sp2}}^{i}, \hat{F}_{\text{sp3}}^{i})\Big) \quad (17)$$

where F_{sp}^{i} represents original features and $\hat{F}_{\text{sp}}^{i}$ denotes enhanced features.

3.2.10 Cai et al. module

Existing dehazing models often underperform in severe haze conditions due to their inability to effectively capture fine-grained details. To mitigate this limitation, Cai et al. [27] introduced an auxiliary self-attention module designed to enhance the model's capacity for fine detail extraction, while also incorporating an adapter to optimize the model's adaptability to additional structural components.

In this work, two self-attention modules are preceded by an adapter module, which functions as an initial mechanism to evaluate the haze

intensity within the image. This preliminary assessment allows for dynamic adjustment of the information flow, ensuring a more refined alignment with the subsequent processing stages. The proposed adapter module employs a convolutional layer for preprocessing, effectively alleviating the computational burden that would typically arise in directly applying fully connected layers to image data. Following this step, the processed features are passed through a ReLU activation function before being directed into the subsequent attention computation modules. Experimental results validate that the integration of the adapter module enables the model to more seamlessly adapt to the new architecture, thereby enhancing overall dehazing performance.

3.2.11 SAM-CD

SAM-CD [28] introduces a convolutional adapter designed to extract task-specific change information, thereby fine-tuning FastSAM [165] to focus on distinct ground objects within the context of remote sensing image change detection. This approach leverages the powerful visual perception capabilities of vision foundation models to enhance the performance of high-resolution remote sensing image change detection.

In alignment with FastSAM's convolution-based architecture, SAM-CD is similarly constructed using convolutional operations. Upon obtaining multi-scale features extracted by FastSAM, the features undergo processing according to the procedure outlined by Eq. (18):

$$f_i^* = \alpha(f_i) = \gamma\{\mathrm{bn}[\mathrm{conv}(f_i)]\} \tag{18}$$

where $\mathrm{conv}(\cdot)$ denotes a 1×1 conv layer, $\mathrm{bn}[\cdot]$ denotes a batch normalization function, and $\gamma(\cdot)$ is a ReLU function.

Given that remote sensing images typically involve fewer categories, the number of feature channels is reduced to mitigate feature redundancy. Moreover, since low-level features are particularly crucial for object segmentation, feature fusion is performed using a structure akin to UNet [166]:

$$d_1 = f_1, \quad d_{i+1} = \mathrm{conv}[f_{i+1}, \mathrm{upsample}(d_i)] \tag{19}$$

where $\mathrm{upsample}(\cdot)$ represents the upsampling layer, $[\cdot;\cdot]$ denotes concatenation along the channel dimension and d_i is the i-th feature layer.

3.2.12 Li et al. module

Li et al. [29] applied the ViT-Adapter [167] to algal bed area segmentation from remote sensing imagery, demonstrating favorable results across several ecological regions.

The adapter architecture is composed of three primary components: the spatial prior (SP) module, the spatial feature injector (SFI) module, and the multi-scale feature extractor (MFE). The SP module first captures spatial features from the input remote sensing images. Subsequently, the SFI module injects these spatial features into the Vision Transformer architecture, allowing the model to incorporate crucial spatial information. The output from the ViT is then passed to the MFE, which generates hierarchical features at multiple scales. These components of the ViT-Adapter architecture enable the model to adapt effectively to varying algal bed states, illumination conditions, and diverse coastal ecosystems, thereby significantly enhancing the segmentation accuracy of the model. This method can help evaluate the potential of an algal bed for CO_2 sequestration, contributing to blue carbon initiatives and efforts to combat climate change.

3.2.13 HarMA

Huang [30] introduced the Harmonized Transfer Learning and Modal Alignment (HarMA) framework, which enables multi-modal transfer learning to simultaneously satisfy task constraints, modality alignment, and intra-modal unified alignment. It also reduces training overhead through parameter-efficient fine-tuning.

Inspired by the human brain's use of shared regions for processing visual and linguistic stimuli, HarMA designs hierarchical multi-modal adapters to model the visual-linguistic semantic space from bottom–up. The proposed multi-modal adapters inherit the alignment of multi-modal contextual representations by sharing the weights of I-MSA and MM-MSA, while also introducing a dynamic gating mechanism. Distinct features are first enhanced through the I-MSA module to refine their representations, and then passed into the shared-parameter-weighted MMS-Adapter for further interaction. HarMA's simplicity allows it to be seamlessly integrated into most models. By leveraging efficient fine-tuning, the framework significantly improves multi-modal retrieval performance, even surpassing the results of full fine-tuning.

3.2.14 PanAdapter

To overcome the constraints imposed by limited

dataset size in the task of pan-sharpening, PanAdapter [31] introduces a sophisticated two-stage fine-tuning framework for domain adaptation, facilitating a more efficient exploitation of the high-level semantic information encapsulated within the pre-trained models in the domain of image restoration. By leveraging the remarkable generalization prowess of these pre-trained models in conjunction with the fine-tuning strategy, PanAdapter achieves cutting-edge performance across a range of benchmark pan-sharpening datasets.

In the first phase, PanAdapter refines the pre-trained convolutional neural network (CNN) architecture by integrating a local prior extraction block at various intermediate stages, thereby capturing spectral and spatial priors at multiple hierarchical levels. During the second phase, a cascaded dual-branch adapter is employed to fuse the spatial and spectral priors derived from the first stage through multi-scale feature interaction. These enriched priors are subsequently injected into the pre-trained ViT model for additional fine-tuning, thereby enhancing the model's performance.

3.2.15 YOLO-Adapter

YOLO-Adapter [32] introduces a lightweight multimodal adapter designed to dynamically facilitate multimodal alignment and confidence estimation. This model addresses the challenge of misalignment in visible-infrared object detection, specifically tackling complex biases such as translation, scaling, and rotation.

The adapter consists of several compact yet efficient layers, seamlessly integrated between the shallow feature representations of the visible and infrared branches. These layers are engineered to concurrently predict the alignment parameters and confidence weights across the two modalities. In particular, this work designed a simple mapping operation and proposed a feature contrastive learning loss that imposes a regularizing constraint on the learning process. This loss function serves to minimize discrepancies between visible and infrared features within the hyperbolic geometric space, thereby diminishing the representational divergence between the modalities. Results substantiate that YOLO-Adapter achieves substantial performance improvements in the visible-infrared object detection task.

3.2.16 Yuan et al. module

Yuan et al. [33] introduced a highly efficient fine-tuning framework tailored for remote sensing image-text retrieval. The proposed multimodal remote sensing adapter, MRS-Adapter, in conjunction with the hybrid multimodal contrastive (HMMC) learning objective, substantially enhances the performance of fine-tuning models for remote sensing image-text retrieval, offering novel insights and approaches for tasks pertaining to remote sensing image–text correspondence.

In particular, the MRS-Adapter refines the cross-modal adapter architecture by removing the skip connection and linking it solely in parallel with the FFN module. It operates on the feature representations obtained from the transformer blocks of both modalities following the MHA operation. The MRS-Adapter employs the traditional adapter framework of down-projection, nonlinear activation, and up-projection. To enable efficient cross-modal parameter sharing, a modality-shared upward projection is introduced, which bridges the modality-specific linear layers of the original projections, facilitating shared information transfer between modalities.

3.2.17 FreqDiMFT

To address the limitations of existing multimodal fine-tuning strategies, which predominantly focus on natural scene datasets, Zhang et al. [34] proposed a frequency-based multimodal fine-tuning strategy (FreqDiMFT). Specifically, the strategy incorporates local–global frequency distribution information within the visual branch to adapt it to the high inter-class similarity and intra-class diversity inherent in remote sensing images. To further enhance the model's generalization ability, FreqDiMFT introduces an adaptive feature refinement module designed for transformers, which filters out redundant features induced by domain discrepancies.

The frequency distribution integration module, designated FreqDiMFT, first maps the features to the amplitude-frequency domain. It then employs distributed average pooling techniques to reshape the features into a two-dimensional vector format. Finally, a fully connected layer and bottleneck structure are used to transform the distribution features into the feature space of CLIP. This work provided a novel paradigm for multimodal fine-tuning in the remote

sensing domain, inspiring future design approaches in this field.

3.2.18 Summary

At the architectural optimization level, AiRs [19] constructs dual adapters for spatial context and semantic response, reducing parameter redundancy via feature splitting mechanisms, TEA [20] innovates a gradient highway structure combined with multi-stage feature fusion to significantly enhance training efficiency, and ACTNet [22] embeds separable convolutions within Swin Transformer to enhance small-target perception capabilities. For multimodal adaptation, CWSAM [23] achieves cross-modal migration of SAM to SAR images through a lightweight dual-branch architecture, HarMA [30] employs hierarchical dynamic gating mechanisms to optimize visual-linguistic feature alignment, FreqDiMFT [34] pioneers a frequency-domain integration module, leveraging amplitude-frequency mapping to strengthen domain adaptability. In task-oriented design, BAN [25] develops a dual-temporal feature alignment network for precise temporal change detection, CloudformerV3 [26] enhances cloud layer detail characterization through multiscale adapters, and Li's algal bed segmentation method [29] improves coastline segmentation accuracy via spatial prior modules. For real-time optimization, SAM-CD [21] inherites FastSAM's lightweight advantages while balancing detection speed and precision through UNet-style feature fusion, and YOLO-Adapter [32] establishes a cross-modal dynamic alignment framework using hyperbolic space contrastive learning. Additionally, DFSA-Net [24] proposes decoupled low-rank adaptation strategies, PanAdapter [31] builds a two-stage spectral-spatial prior optimization framework, and Yuan's method [33] develops parameter-shared up-projection mechanisms, collectively expanding the application boundaries of adapter technology through feature decoupling, multiscale interaction, and cross-modal compression.

3.3 Prompt tuning

3.3.1 Concepts

Early prompt tuning techniques showed impressive performance in NLP by introducing a few parameters into the input or hidden layers, enabling pre-trained language models to modify their outputs based on few-shot data [168]. VPT [10] pioneered parameter-efficient visual fine-tuning. Shallow VPT introduced trainable patches as prompts for the new task into the input, altering the input distribution. This approach allows the pre-trained visual model to adjust the output feature distribution based on few-shot image data, thereby enhancing the performance on new tasks. Deep VPT incorporated trainable parameters into the hidden layers of the model, gradually altering the visual feature distribution to better adapt to few-shot data.

Let m image patches $\{I_j \in \mathbb{R}^{3\times h\times w} | j \in \mathbb{N}, 1 \leqslant j \leqslant m\}$ be the input to the Vision Transformer [151], which become hidden space vectors after embedding:

$$e_0^j = \text{Embed}(I_j) \tag{20}$$

Let the embedding output of the i-th layer be denoted $E_i = \left\{e_i^j \in \mathbb{R} | j \in \mathbb{N}, 1 \leqslant j \leqslant m\right\}$, which serves as the input to the $(i+1)$-th layer L_{i+1}. Along with an additional learnable class label [CLS], the entire Transformer can be represented as Eq. (21):

$$\begin{cases} [x_i, E_i] = L_i([x_{i-1}, E_{i-1}]) \\ y = \text{Head}(x_N) \end{cases} \tag{21}$$

where x_i denotes the i-th embedding of [CLS], and Head represents the trainable classification head structure. Shallow VPT only introduces learnable parameters P into the first layer, while keeping other backbone parameters fixed. The calculation process can be expressed as Eq. (22):

$$\begin{cases} [x_1, Z_1, E_1] = L_1([x_0, P, E_0]) \\ [x_i, Z_i, E_i] = L_i([x_{i-1}, Z_{i-1}, E_{i-1}]) \\ y = \text{Head}(x_N) \end{cases} \tag{22}$$

where Z denotes the hidden vectors of P in different layers. Deep VPT adds learnable parameters P_i in each layer. The calculation process can be expressed as Eq. (23):

$$\begin{cases} [x_i, \_, E_i] = L_i([x_{i-1}, P_{i-1}, E_{i-1}]) \\ y = \text{Head}(x_N) \end{cases} \tag{23}$$

In remote sensing images, the positions of many objects are unpredictable, and the cost of data acquisition is high. As a result, there are many few-shot tasks in remote sensing. Recent studies claim that prompt tuning is better suited for few-shot data [169], so is beneficial for few-shot scenarios in remote sensing. We now focus on the application of prompt tuning in remote sensing.

3.3.2 RingMo-SAM

The multimodal remote sensing segmentation model

RingMo-SAM [35] elevates the performance of SAM in segmenting arbitrary objects within both optical and SAR imagery. This is achieved through the design of an encoder that supports multiple bounding box prompts and SAR feature prompts, followed by the fine-tuning of these parameters. The proposed method helps to improve segmentation accuracy and object classification. The proposed model investigates the potential of utilizing remote sensing image features as prompts, thereby enhancing the performance of vision foundation models in multimodal remote sensing tasks through fine-tuning.

In the proposed framework, multiple prompt boxes are seamlessly integrated into the sparse encoding process, working collaboratively to enhance the decoding phase and optimize segmentation precision. Additionally, RingMo-SAM incorporates SAR polarization decomposition prompts, where the extracted polarization features provide essential object-related information. These features are strategically utilized in the prompt encoding, significantly boosting the model's segmentation capabilities.

3.3.3 Immanuel and Sinulingga module

Immanuel and Sinulingga [36] introduced a learnable prompt to fine-tune SegGPT [170] to adapt it to few-shot new classes in segmentation tasks. Considering the multi-scale characteristics inherent in remotely sensing objects, they devised patch-level predictions and proposed a patch stitching technique to mitigate the issue of boundary discontinuities between adjacent patches.

The model's parameters are frozen after training on the base classes, and only the learnable prompt $\boldsymbol{Z}$ undergoes optimization. A distinct prompt set, $\{\boldsymbol{Z}^i\}_{i=1}^N$ is then generated for N new classes, with each prompt being explicitly associated with its corresponding class and trained by sampling from the respective class examples. This method not only preserves segmentation performance on the base classes but also reduces the number of additional parameters. During inference on new classes, the model merely integrates the relevant prompt into the framework, enabling efficient predictions without necessitating further retraining.

3.3.4 SPT

SPT [37] represents the pioneering approach to adapting visual-language models for the joint classification of remote sensing hyperspectral and LiDAR imagery. By integrating spectral-driven visual prompts into the visual encoder and employing learnable textual prompts, the model is fine-tuned to bolster the generalization capacity of the visual-language framework.

In remote sensing images, features often exhibit spatial similarity, which means extracting spatial features alone is insufficient for accurately classifying complex ground objects. To leverage the spectral vectors in hyperspectral imagery, the spectral prompt learner (SPL) utilizes 1D-CNN to process the spectral vectors of each pixel v_i, generating independent sampled spectral prompt vectors (SPVs).

$$P_i^{(\mathrm{S})} = f_{\theta^{(\mathrm{SPL})}}(v_i) \tag{24}$$

where $\theta^{(\mathrm{SPL})}$ denotes the learnable parameters of SPL. These SPVs are then injected into the visual encoder alongside the visual embeddings, enabling SPT to extract spatial-spectral fused features that cater to different modality preferences based on downstream tasks.

Moreover, recognizing that using only category names for text embedding is inadequate to fully capture the class descriptions, SPT provides a learnable and class-related textual prompt to enrich the representation of each class.

3.3.5 Gao et al. module

Gao et al. [38] pioneered the exploration of source-free adaptation segmentation for remote sensing images by combining vision foundation models with prompt learning. To more effectively transfer prior knowledge from vision foundation models to diverse target remote sensing tasks, they proposed an attention-guided prompt fine-tuning strategy to transfer model priors that are most relevant to downstream tasks across different layers and locations.

This approach utilizes attention matrices to dynamically adjust the correlation between the prompt and the layers, thereby mining features that are more relevant and discriminative to the task. Specifically, learnable prompts initialized using the Xavier method are denoted $\boldsymbol{P} = \{p_1, \cdots, p_N\} \in \mathbb{R}^{N\times D}$, and the classification token is denoted z_{cls}. In the l-th layer, the input and output prompt tokens are denoted $\boldsymbol{Z}_p^{l-1}$ and $\tilde{\boldsymbol{Z}}_p^{l}$. The attention matrix is represented as $W = \{\boldsymbol{w}^1, \cdots, \boldsymbol{w}^{L-1}\}$ where L is the total number of layers. The entire attention-guided prompt fine-tuning process can be formalized as

$$\left[z_{\text{cls}}^{l}, \tilde{\boldsymbol{Z}}_{p}^{l}, \boldsymbol{Z}^{l}\right] = \text{Layer}^{l}\left(\left[z_{\text{cls}}^{l-1}, \boldsymbol{Z}_{p}^{l-1}, \boldsymbol{Z}^{l-1}\right]\right) \quad (25)$$

$$\boldsymbol{Z}_{p}^{l} = \boldsymbol{w}^{l} \odot \boldsymbol{Z}_{p}^{l-1} + \left(\mathbf{1} - \boldsymbol{w}^{l}\right) \odot \tilde{\boldsymbol{Z}}_{p}^{l} \quad (26)$$

where the symbol $\odot$ denotes element-wise multiplication.

3.3.6 APPLeNet

Singha et al. [39] proposed the visual attention parameterized prompts learning network (APPLeNet) to address the few-shot generalization task for remote sensing images. This framework leverages the multi-scale features extracted by the CLIP encoder, decoupling the visual style and content priors within the domain generalization task. Additionally, it introduces an attention-driven lightweight injection module to better exploit both visual features and style elements.

Given the multi-scale features f_v, the model first compresses the spatial dimensions of each channel using global average pooling (GAP), resulting in a compact representation $\hat{f}_v^l(\hat{x}) \in \mathbb{R}^{C\times 1}$ where C is the number of channels. These compressed features are then concatenated to form the input $\hat{F}(x) = [\hat{f}_v^1(x); \cdots; \hat{f}_v^{\hat{L}}(x)]$. The average feature estimate is utilized to capture the style elements of a specific domain, producing a representation that encompasses both multi-scale content and style elements $\mu_i = f_v(X^i)$. The attention module within each block of the layer is denoted $\mathcal{A}_q(\cdot)$ where q is the number of attention blocks. Therefore, the output feature for each layer can be computed as Eq. (27):

$$\mathcal{O}_q = \begin{cases} F(x) \odot \mathcal{A}_q(F(x)) + F(x), & \text{if } q = 1 \\ \mathcal{O}_{q-1} \odot \mathcal{A}_q(\mathcal{O}_{q-1}) + \mathcal{O}_{q-1}, & \text{otherwise} \end{cases} \quad (27)$$

After passing through M lightweight mapping layers, M visual tokens $\{v_1, \cdots, v_M\}$ are obtained. Upon fusing each with the corresponding m-th language token c_m, the result yields a learnable prompt:

$$t_y = \{[v_1 + c_1], \cdots, [v_{\mathcal{M}} + c_{\mathcal{M}}], [\text{CLS}_y]\} \quad (28)$$

where $[\text{CLS}_y]$ denotes the word embeddings for class y.

3.3.7 DP-RSCap

DP-RSCap [40] proposes an entity-concept prompt extractor to capture entity information from images, along with a scene-class prompt generator to predict scene categories and obtain scene-relevant semantic information. Based on these two prompts, DP-RSCap facilitates the exchange and alignment of information across different modalities, thereby improving the quality of remote sensing image interpretation.

The entity-concept prompt extractor initially employs NLTK (a grammar tool) to construct a preselected entity space, which is then organized into prompts of the form "An image contains {entity concept}". Subsequently, the CLIP encoder is utilized to obtain the visual representation of the image F_v, as well as the textual representation of the prompt F_t. The correlation between these two representations is assessed to identify M entity concepts, from which the entity prompt "There are $v_1, \cdots, v_M$ in the image" is constructed.

The scene class prompt generator uses the output of the final layer of the CLIP visual encoder and applies downsampling to obtain multi-scale visual features $V^v = \{V_1, \cdots, V_L\}$. These visual features are then enhanced using a simple transformer block and MHA mechanisms: $V^{\text{en}} = \{V_1^{\text{en}}, \cdots, V_L^{\text{en}}\}$. Finally, a global semantic representation is obtained via average pooling, which is subsequently used to predict the probability value of scene category P_{cls}:

$$V_g = \text{FC}\left(\text{cascade}(V_{1,g}, \ldots, V_{L,g})\right) \quad (29)$$

$$P_{\text{cls}} = \text{Softmax}\left(\text{FC}\left(\text{dropout}(V_g)\right)\right) \quad (30)$$

where $\text{cascade}(\cdot,\cdot)$ denotes the cascade operation.

3.3.8 RSPrompter

RSPrompter [41] introduces an automated framework for generating category-specific prompts tailored to individual instances, thereby enhancing the performance of SAM in remote sensing instance segmentation. The method uses a multiscale feature enhancer to augment the visual features extracted by the SAM encoder. Based on these enriched semantic features, the anchor-based prompter constructs prompt embeddings for the SAM mask decoder.

After obtaining the multiscale-enhanced feature representations, the anchor-based prompter employs an anchor-based region proposal network (RPN) to generate candidate object bounding boxes. These proposals are subsequently subjected to RoI pooling, yielding refined feature representations. From these representations, three perceptual heads are derived: the semantic head, the localization head, and the prompt head. The prompt head is dedicated to generating the prompt embeddings required by the SAM mask decoder. The whole process is shown in Fig. 6.

The paper also introduces a query-based prompter to streamline the process of the anchor-based prompter, which is composed of a transformer encoder and decoder. The encoder is employed to extract high-dimensional semantic features, while the decoder is

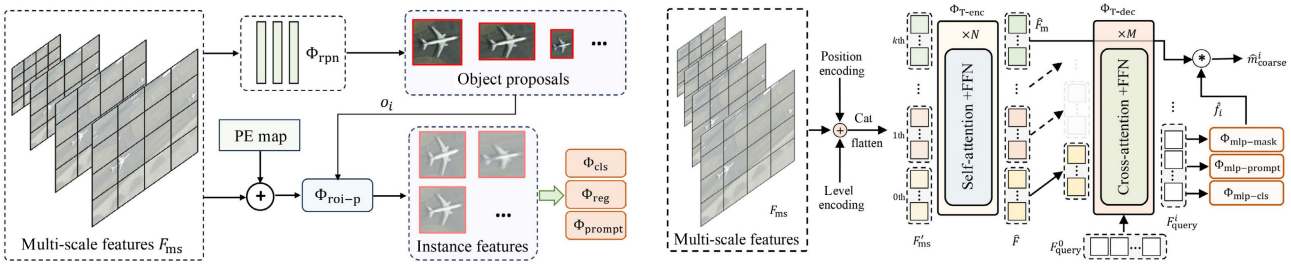

Fig. 6 Anchor-based prompter and query-based prompter in RSPrompter. Reproduced with permission from Ref. [41], © IEEE 2024.

responsible for transforming the pre-defined learnable queries into prompt embeddings.

3.3.9 CPSeg

Li [42] proposed a framework for semantic segmentation of remote sensing images based on vision-language models (VLMs): chain-of-thought language prompting for finer-grained semantic segmentation (CPSeg). By sequentially and logically prompting the vision-language model, this approach encourages a chain-of-thought reasoning process, thereby enhancing the model's performance in semantic segmentation tasks.

The method begins by querying the model as to whether a specific object exists within the image. If the model detects the object, the next query focuses on determining how many of that object are in the image. The text prompts generated throughout this reasoning process are then consolidated into a text encoder to guide the segmentation model to better understand the image. Specifically, this work constructs a chain of thought as follows: $C = c_1, \cdots, c_m$. Each thought c_i corresponds to a sentence s_i, and each reasoning step involves a text encoder $T(c_i)$ and a pixel-level segmentation function $f_{c_i}(p_i)$, ultimately generating m segmentation decisions $D = d_1, \cdots, d_m$. It is noteworthy that this model demonstrates promising results exclusively on flood disaster analysis tasks, and its transferability to and efficacy on other datasets have yet to be thoroughly assessed.

3.3.10 Lan et al. module

Directly fine-tuning VLMs for fine-grained ship classification in remote sensing (RS-FGSC) tasks may lead to overfitting on the base class, resulting in suboptimal performance when classifying a new class. To address this issue, Lan et al. [43] proposed a hierarchical, multi-granularity prompt fine-tuning approach that integrates prior knowledge of ship information extracted by a lightweight model as a bias term. This strategy enhances the model's generalization ability and facilitates better discrimination of complex backgrounds.

The framework employs a hierarchical, multi-granularity tri-class prompt to optimize the CLIP text prompt: $c_i = \{c_i^p, c_i^s, c_i^f\}$, thereby guiding the model to achieve a more granular perception of the image. Building upon this, M context vectors are introduced as learnable text prompts: $V = \{v_1, \cdots, v_M\}$. These vectors are fused with the feature vectors encoded by a lightweight remote sensing encoder, creating a comprehensive representation:

$$v_m(x) = v_m + \delta \tag{31}$$

This fusion enables the context vectors to encapsulate both the features encoded by the CLIP model and the prior knowledge of ships in remote sensing images.

3.3.11 Zhao et al. module

Zhao et al. [44] explored the feasibility of using prompt learning for continual learning in remote sensing scene classification tasks. They also investigated the effects of prefix tuning, prompt tuning, and other fine-tuning components on the effectiveness of prompt learning.

The proposed Pro-T model introduces the prompt parameter before the input tokens, effectively adding p to the queries, keys, and values h_Q, h_K, and h_V within the MHA layers:

$$f_p^{\mathrm{Pro-T}}(p, h) = \mathrm{MSA}([p; h_Q], [p; h_K], [p; h_V]) \tag{32}$$

Here, the operation $[\cdot;\cdot]$ denotes concatenation in the sequence length dimension. In contrast, the Pre-T model splits p into p_K, $p_V \in \mathbb{R}^{L_p/2 \times D}$respectively, while keeping unchanged:

$$f_p^{\mathrm{Pre-T}}(p, h) = \mathrm{MSA}(h_Q, [p; h_K], [p; h_V]) \tag{33}$$

Experimental results demonstrate that the prompt provides essential information, assisting the model to better understand the image content and overcoming the issue of catastrophic forgetting.

3.3.12 MVP

To address the overfitting and storage issues associated with full fine-tuning in the remote sensing domain, meta visual prompt (MVP) tuning [45] integrates the concept of prompt tuning into a meta-learning framework and applies it to few-shot remote sensing scene classification tasks.

Consistent with VPT [10], a set of prompt tokens is incorporated into the input of ViT layers; the ViT structure can be represented as

$$[\mathrm{CLS}_i, P_i, E_i] = L_i\left([\mathrm{CLS}_{i-1}, P_{i-1}, E_{i-1}]\right) \tag{34}$$

$$f_{\theta'}(x) = \mathrm{CLS}_N \tag{35}$$

where θ' denotes the parameters of ViT. During both meta-training and meta-finetuning, only the prompt parameters θ_p are updated. Finally, the classification task is formulated as

$$\theta^* = \mathrm{argmax}_{\theta^P} \sum_x \log p\left(y|f_{\theta'}(x); \theta^P\right) \tag{36}$$

where θ_* represents the optimal value for prompt parameters θ_p. Results demonstrate that this method achieves impressive performance across various experimental configurations.

3.3.13 IVP

Instance-aware visual prompting (IVP) [46] represents the first application of prompt learning to remote sensing scene classification. IVP utilizes an instance-level prompt generator, Meta-Net, to aggregate contextual information from the image after it has been embedded into patches. This aggregated information is then combined with the image's patch embeddings and fed into a pretrained model for encoding, thereby facilitating the extraction of instance-specific features with enhanced precision.

Specifically, the IVP-shadow approach begins by passing the image tokens through a GAP layer to adapt it to variations in the data distribution:

$$G^{(k)} = \mathrm{AveragePooling}\left(e_1^{(k)}, \cdots, e_N^{(k)}\right) \tag{37}$$

Subsequently, a bottleneck-structured feedforward layer is employed to extract relevant feature representations:

$$\widehat{G}^{(k)} = W_{\mathrm{up}}\left(\mathrm{ReLU}\left(W_{\mathrm{down}} G^{(k)} + b_1\right)\right) + b_2 \tag{38}$$

$$\left(p_l^{(k)}, \cdots, p_M^{(k)}\right) = \mathrm{Reshape}\left(\widehat{G}^{(k)}\right) \tag{39}$$

Then, these representations, along with the original image features and class tokens, are input into a transformer layer to obtain:

$$\begin{aligned}&\left(c^{(k+1)}, p_1^{(k+1)}, \cdots, p_M^{(k+1)}, e_1^{(k+1)}, \cdots, e_N^{(k+1)}\right)\\ &\quad = L_k\left(c^{(k)}, p_1^{(k)}, \cdots, p_M^{(k)}, e_1^{(k)}, \cdots, e_N^{(k)}\right)\end{aligned} \tag{40}$$

Furthermore, the IVP-deep method uses an adaptive max-pooling layer to optimize Meta-Net, thereby reducing computational overhead:

$$\left(\hat{e}_1^{(k)}, \cdots, \hat{e}_N^{(k)}\right) = \mathrm{ReLU}\left(W_{\mathrm{down}}\left(e_1^{(k)}, \cdots, e_N^{(k)}\right) + b_1\right) \tag{41}$$

$$\left(\hat{p}_1^{(k)}, \cdots, \hat{p}_M^{(k)}\right) = \mathrm{AdaptiveMaxPool}\left(\hat{e}_1^{(k)}, \cdots, \hat{e}_N^{(k)}\right) \tag{42}$$

$$\left(p_1^{(k)}, \cdots, p_M^{(k)}\right) = W_{\mathrm{up}}\left(\hat{p}_1^{(k)}, \cdots, \hat{p}_M^{(k)}\right) + b_2 \tag{43}$$

3.3.14 PromptCC

PromptCC [47] decouples the task of remote sensing image change captioning into two distinct problems: whether a change has occurred and where the change has occurred. To more effectively address the latter issue, PromptCC integrates prompt learning into a pretrained large language model and employs a multi-prompt learning strategy. This strategy generates unified prompts, along with class-specific prompts based on the image classification results. By leveraging these prompts and the extracted visual features, the LLM can produce more accurate descriptions of changes within the image.

In the multi-prompt learning strategy, the unified prompts can be regarded as global, task-dependent prompts that are applicable at any stage of the inference process. The class-specific prompts, derived by fusing two learnable prompt embeddings $P_{C_0} \in \mathbb{R}^{1\times d_T}$ and $P_{C_1} \in \mathbb{R}^{1\times d_T}$ where d_T is the dimension of the textual embedding, assist the LLM in determining whether a change is present. Results show the robust performance of PromptCC, as well as the effectiveness of its decoupling paradigm.

3.3.15 PAT

Inspired by human visual perception, Bi et al. [48] proposed a novel prompt-driven paradigm, prompt and transfer (PAT). It constructs a dynamic, class-aware prompting framework, which enables the precise transfer of class-related semantic information from the image to the prompt, allowing the model's encoder to focus on the target class of the current task.

PAT leverages a traditional pre-trained vision-language model to encode class information, embedding it into several randomly initialized

embeddings, thereby ensuring that the prompt is initially category-aware for the specific task. PAT also introduces two key components: semantic prompt transfer (SPT) and a part mask generator (PMG) to further enhance the category awareness of the prompt. SPT establishes a semantic transfer between feature tokens and the prompt, enabling effective communication of semantic information. PAT achieves competitive performance in few-shot segmentation within the remote sensing domain and also performs well on several other tasks.

3.3.16 Osco et al. module

To augment the model's performance, Osco et al. [49] introduced a sophisticated automated technique that synergistically combines a text-prompt-derived general exemplar with one-shot learning. This innovative strategy enables the model to leverage pre-existing knowledge encoded in VFMs, thereby achieving efficient task adaptation with minimal task-specific annotations. Specifically, the one-shot learning framework facilitates rapid model adaptation by utilizing a single reference example to fine-tune the prompt, reducing the dependency on large annotated datasets. This approach has significant advantages in tasks with restricted samples.

3.3.17 Summary

In the development of prompt-based tuning techniques for remote sensing, researchers have advanced multimodal understanding, few-shot adaptation, and task-specific optimization through three key technical directions. Cross-modal prompt fusion strategies integrate diverse data sources: RingMo-SAM [35] combines optical/SAR polarization decomposition prompts for arbitrary object segmentation, while SPT [37] embeds hyperspectral/LiDAR features into vision-language frameworks to enhance land cover classification. Dynamic prompt generation mechanisms leverage attention guidance and meta-learning: Gao et al. [38] employed attention-weighted layer correlation to mine task-relevant features, while MVP [45] dynamically generates prompts via meta-features for few-shot scene classification. Task-driven architectures address specialized challenges: RSPrompter [41] uses anchor/query dual prompters for instance-level SAM adaptation, CPSeg [42] implements chain-of-thought language prompting for explainable fine-grained segmentation, and PromptCC [47] develops unified and class-specific prompts to optimize change captioning consistency. Few-shot optimization breakthroughs include Lan et al.'s method [43], which integrates lightweight ship priors for cross-domain classification, and Osco et al.'s approach [49], which automates prompt generation through one-shot learning. Additionally, interpretability enhancement is achieved via PAT's [48] category-aware semantic transfer and IVP's [46] instance-specific feature pooling.

3.4 Reparameterized tuning

Reparameterized tuning is one of the most popular efficient fine-tuning paradigms in recent years. LoRA [171], representative of reparameterized tuning, made significant progress in NLP fine-tuning tasks and has been widely applied to computer vision [172], remote sensing [173], medicine [174], and multimodal tasks [175]. It is worth noting that the trainable structure of LoRA can be merged into the original backbone network after training. Therefore, LoRA does not introduce additional computational costs due to new structures during inferencing, which is an advantage over most adapter-based and prompt-based fine-tuning methods. LoRA trains a structure that merges into the pre-trained matrix $W_0 \in \mathbb{R}^{d\times k}$. The LoRA structure mainly comprises two matrices: a matrix $A \in \mathbb{R}^{r\times k}$ initialized by a Gaussian distribution, and a matrix $B \in \mathbb{R}^{d\times r}$ initialized to zero, where r is the internal dimension. Assuming the input is $X \in \mathbb{R}^k$ and the output is $Y \in \mathbb{R}^d$, the reparameterization process of LoRA can be expressed as Eq. (44):

$$y = W_0 x + \Delta W x = W_0 x + BAx \tag{44}$$

The concept of reparameterization is similar to residual connections [150], and its design is simple yet effective. Reparameterization methods can achieve good performance in visual tasks with relatively small sample sizes [169], and remote sensing scenarios often involve few-shot situations. Next, we introduce reparameterization fine-tuning techniques in remote sensing, most of which are applications of LoRA or its variants.

3.4.1 LoRA-NIR

LoRA-NIR [50] leverages the LoRA technique to fine-tune a ViT pre-trained on the RGB domain for application to the near-infrared (NIR) spectrum. This approach advances the use of NIR images, enhancing crop semantic segmentation and enabling more accurate plant health monitoring.

LoRA effectively addresses the domain adaptation

challenge between the RGB and NIR domains. In this approach, LoRA layers are applied to the query and value projection layers of each transformer block to fine-tune the backbone model. Experimental results demonstrate that the model fine-tuned with LoRA in the NIR domain outperforms its RGB domain counterpart, highlighting LoRA's potential as a powerful technique for fine-tuning in NIR-based segmentation tasks.

3.4.2 Xue et al. module

Xue et al. [51] introduced a novel application of LoRA for transferring the SAM model to aerial imagery, aiming to address the task of land cover classification. While the integration of the SAM encoder inevitably introduces some inference latency, this work pioneers an efficient fine-tuning methodology for adapting vision foundation models to remote sensing tasks involving aerial imagery.

Specifically, the authors integrate LoRA layers into each transformer block of the SAM encoder, enabling fine-tuning with a reduced parameter set while tailoring the model to the downstream tasks associated with aerial image analysis. The approach yields promising results on the ISPRS Vaihingen and Potsdam datasets [176], demonstrating the model's superior performance and highlighting its potential for advancing land cover classification in remote sensing.

3.4.3 SkyEyeGPT

SkyEyeGPT [52] introduced a sophisticated multimodal large model tailored for remote sensing language-vision comprehension, accompanied by a meticulously designed remote sensing multimodal fine-tuning dataset that incorporates both single-task and multi-task dialogue instructions.

In the training process, the authors use LoRA to fine-tune the alignment layers of the model, thereby optimizing the coalescence of remote sensing visual features derived from the vision encoder with linguistic features from the large language model. In addition, the model's language decoder is fine-tuned to further enhance its efficacy in adapting to a range of downstream tasks. Experiments systematically explored the effect of LoRA's rank on the results, providing lessons for similar studies.

3.4.4 SkySenseGPT

To comprehend the intricate semantic relationships within complex remote sensing scenarios, Luo et al. [53] introduced a fine-grained, large-scale instruction-tuning dataset, FIT-RS. Beyond image understanding tasks, this dataset encompasses multi-tiered challenges, spanning from object relationship inference to scene image generation, thereby holistically augmenting the fine-grained, multi-dimensional interpretative capabilities of remote sensing multimodal large models.

Building upon this dataset, the authors also proposed SkySenseGPT, a model comprising a visual encoder, a multimodal projector, and a large language model. During the instruction-tuning phase, the parameters of the visual encoder are frozen, the projector undergoes fine-tuning, and LoRA is employed to optimize the LLM.

3.4.5 DebLoRA

Due to the inherent class imbalance in remote sensing datasets, fine-tuned models often exhibit pronounced class bias, leading to suboptimal performance across different categories. To mitigate this issue, Tian et al. [54] introduced De-biased LoRA (DebLoRA) to address class bias and seamlessly integrate with any LoRA variant.

Building on the foundations of LoRA and cLoRA [177], DebLoRA employs an unsupervised strategy to cluster the biased feature space Z. Specifically, the method uses K-means clustering to identify K distinct cluster centers within the feature space. The representations of the tail classes are then reformulated as a weighted combination of these cluster centers, effectively reducing the influence of bias. Following this, each tail class is recalibrated within the de-biased feature space Z', which is subsequently used to train the DebLoRA module. This approach demonstrates substantial improvements in both head and tail class performance across tasks, including the transfer from natural images to optical remote sensing images, and from optical to multispectral remote sensing images. By mitigating class bias, DebLoRA enables more balanced and accurate model performance across all class distributions.

3.4.6 ESAM-CD

ESAM-CD [55] employs LoRA-based fine-tuning of the EfficientSAM encoder to enhance its performance in remote sensing image change detection tasks. Specifically, LoRA layers are inserted into the query and key projection layers of the transformer blocks, thereby influencing the attention mechanism's score

to direct the model's focus towards areas with significant changes. This approach may be formalized as

$$\mathrm{Att}(Q,K,V)=\mathrm{Softmax}\left(QK^{\mathrm{T}}/\sqrt{d}\right) \tag{45}$$

$$Q=M_qF_{\mathrm{SAM}}+B_qA_qF_{\mathrm{SAM}} \tag{46}$$

$$K=M_kF_{\mathrm{SAM}} \tag{47}$$

$$V=M_vF_{\mathrm{SAM}}+B_vA_vF_{\mathrm{SAM}} \tag{48}$$

where M_q, M_k, and M_v represent the weight matrices of the frozen mapping layers in EfficientSAM, F_{SAM} denotes the features extracted by EfficientSAM, and B_q, A_q, B_v, and A_v are the trainable LoRA layer parameters.

Through this method, the robust image understanding capabilities of EfficientSAM are preserved, and the model is able to better capture remote sensing characteristics. Results on the WHU-CD [178] and LEVIR-CD [179] change detection datasets demonstrate that ESAM-CD performs better than many weakly supervised methods.

3.4.7 SAM-MLoRA^F

As Fig. 7 shows, the SAM-MLoRA^F [56] framework employs multiple LoRA fine-tuning modules to transfer the SAM model to urban man-made object extraction, effectively harnessing the segmentation capabilities of SAM on remote sensing datasets. Through the integration of a limited number of trainable parameters and the employment of both supervised and unsupervised fine-tuning strategies, this approach adeptly transfers the segmentation capabilities of the SAM model to remote sensing datasets.

To alleviate the overfitting challenges commonly encountered with high-rank LoRA in conventional frameworks, the SAM-MLoRA^F architecture employs an approximation of high-rank LoRA through the integration of multiple LoRA components. This strategy not only sustains the fine-tuning efficacy of SAM, but also reduces the parameter overhead introduced by the LoRA layers. Specifically, SAM-MLoRA^F strategically incorporates LoRA blocks in parallel within the self-attention and MLP layers of the ViT architecture, enhancing both efficiency and adaptability.

3.4.8 SSDiff

SSDiff [57] addresses the pansharpening problem by decomposing the generalized sharpening process into two subspaces: spatial and spectral components. This decomposition facilitates the independent learning of spatial details and spectral features. Building upon this framework, SSDiff introduces an alternating projection fusion module (APFM) to fuse the features extracted by the two branches, and a frequency modulation inter-branch module (FMIM) to equilibrate the frequency distribution between the two branches.

However, maintaining a balanced training process between the two branches proves to be a significant challenge. To overcome this, SSDiff proposes an innovative solution that uses LoRA to alternately

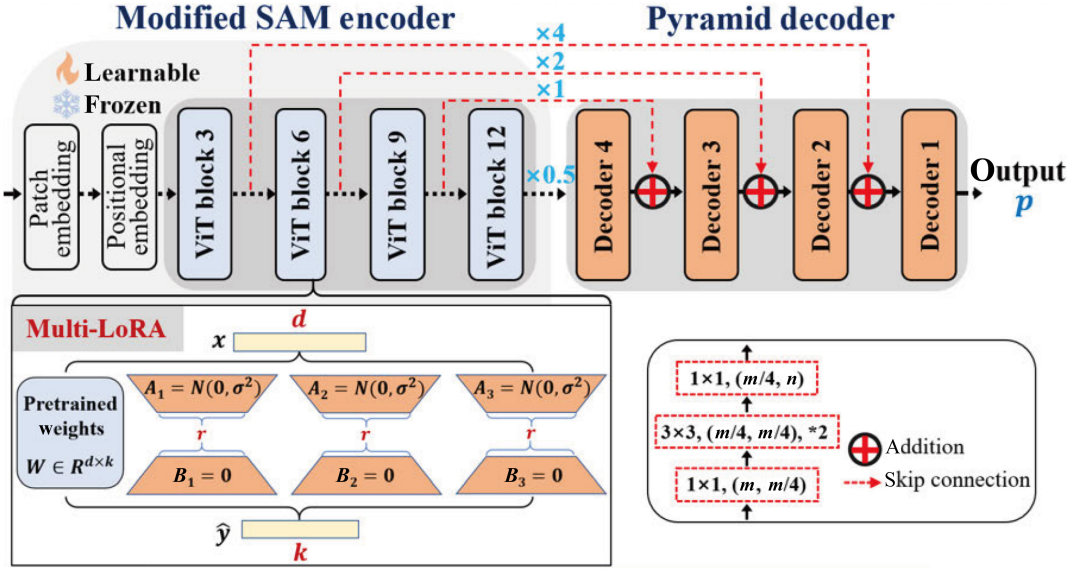

Fig. 7 Encoder and decoder of SAM-MLoRA^F. Reproduced with permission from Ref. [56], © IEEE 2024.

fine-tune the two branches: when the spatial branch is fine-tuned, gradient backpropagation for the spectral branch's parameters is decoupled; conversely, when the spectral branch is fine-tuned, gradient backpropagation for the spatial branch is separated. This block-wise gradient decoupling mechanism effectively ensures a balanced fine-tuning process for both branches.

3.4.9 LoRA-Det

LoRA-Det [58] is specifically designed to address the bandwidth limitations inherent in satellite onboard object detection systems, where extensive model updates are often impractical. This approach facilitates highly efficient fine-tuning by updating only a minimal subset of the model's parameters, thereby reducing computational overhead—a crucial factor for real-time image processing in space-borne environments.

The primary innovation of LoRA-Det lies in its hybrid fine-tuning strategy, which synergistically integrates PEFT with full fine-tuning. This strategy enables the model to achieve 97%–100% of the performance of full fine-tuning while updating only 12.4% of the overall parameters. By incorporating LoRA into both the transformer backbone and the detection head, the method not only preserves detection accuracy but also minimizes computational burdens. Additionally, a low-rank approximation technique is employed to optimally select the rank of the LoRA matrices, further enhancing the efficiency of the adaptation process.

Empirical evaluations conducted across various remote sensing datasets demonstrate that LoRA-Det achieves a substantial reduction in the number of trainable parameters, while maintaining near-optimal detection accuracy. This efficiency accelerates the model's training iterations and bolsters its generalization capabilities, making LoRA-Det particularly suitable for satellite systems operating under strict computational and bandwidth constraints. The proposed approach offers a scalable and resource-efficient solution that is adept at balancing performance and computational efficiency for real-time remote sensing image interpretation.

3.4.10 Summary

Reparameterized tuning has proved to be useful in remote sensing. Existing approaches can be categorized into key areas spanning cross-domain adaptation, multimodal integration, class imbalance mitigation, task-specific innovation, and resource-aware optimization. Cross-domain adaptation is exemplified by LoRA-NIR [50], where LoRA layers inserted into ViT's query and value projections enable effective migration from RGB to near-infrared (NIR) domains for crop health monitoring, while Xue et al. [51] demonstrated SAM's adaptation to aerial imagery via LoRA-enhanced transformer blocks. Multimodal integration is addressed in SkyEyeGPT [52] and SkySenseGPT [53], which leverage LoRA to align remote sensing visual features with linguistic embeddings, supporting instruction-driven fine-grained interpretation. For class imbalance mitigation, DebLoRA [54] introduces unsupervised clustering to recalibrate tail-class representations within a de-biased feature space, achieving balanced performance across optical and multispectral domains. Task-specific innovations include ESAM-CD [55], which integrates LoRA into EfficientSAM's attention layers to prioritize change-sensitive regions, and SAM-MLoRA-F [56], which approximates high-rank adaptation through parallel low-rank modules to prevent overfitting in urban object extraction. Resource-aware optimization is exemplified by SSDiff [57]'s gradient-decoupled LoRA tuning for spatial-spectral pansharpening and LoRA-Det [58]'s hybrid strategy for satellite-based detection under bandwidth constraints.

3.5 Hybrid tuning

3.5.1 Combinations

The aforementioned methods represent different fine-tuning paradigms, each with its own characteristics. Various combinations of these paradigms can yield impressive results for specific tasks. This section introduces several hybrid methods in remote sensing. We finish by summarizing existing and non-existent hybrid methods.

3.5.2 Upetu

As Fig. 8 shows, UPetu [59] introduces a unified PEFT framework tailored for dense prediction tasks in remote sensing, addressing the limitations of current PEFT methods that are mainly designed for classification tasks. One key component of this framework is the efficient quantization adapter module (EQAM), which strengthens the alignment between fine-grained feature representations and task-

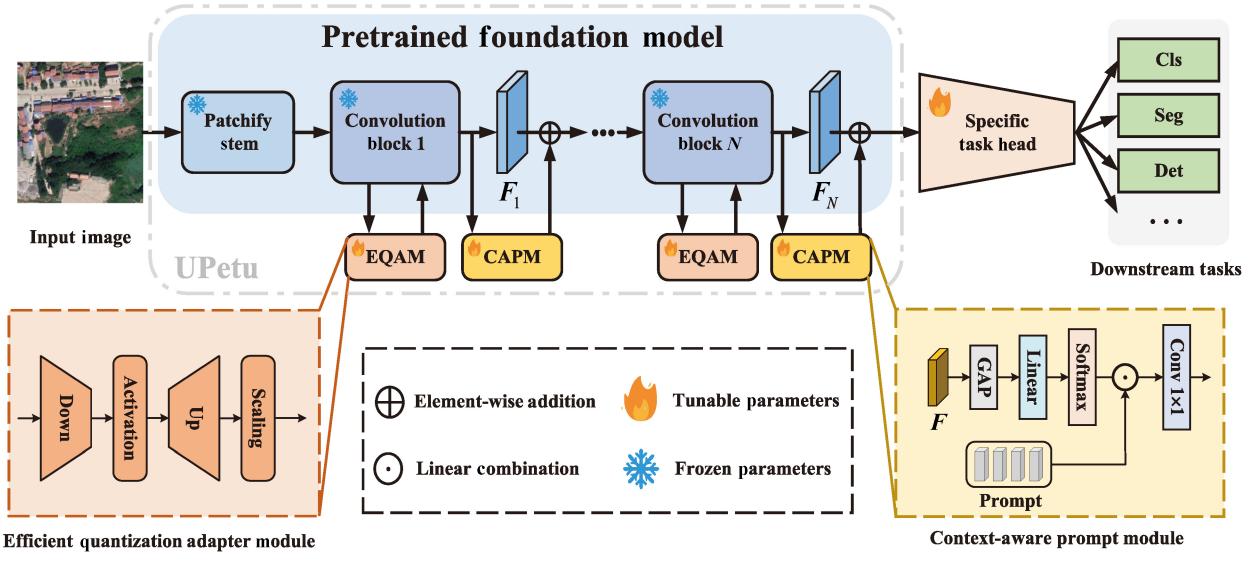

Fig. 8 Framework of UPetu. Reproduced with permission from Ref. [59], © IEEE 2024.

specific knowledge prior through the incorporation of a quantized linear (Q-Linear) layer alongside a nonlinear activation function. Another key component of this framework is the context-aware prompt module (CAPM), which integrates learnable prompts into multi-scale features, enabling the model to extract richer contextual information.

The proposed EQAM consists of two Q-Linear layers with a ReLU activation in between, along with flattening, scaling, and reshaping operations. Specifically, the Q-Linear layer within EQAM is designed to achieve an optimal trade-off between efficiency and accuracy. EQAM employs a clustering-based quantization strategy to minimize the bit width of the linear layer weights. Assuming the weight matrix follows a Gaussian distribution, denormalization restores quantized weights to their original mean and variance, preserving feature representation fidelity:

$$w_i' = \frac{w_i - \mu}{\sigma} \tag{49}$$

$$\hat{w}_i' = \mathcal{Q}(w_i') = c_j, \quad \text{if} \quad w_i' \in \mathcal{U}_j \tag{50}$$

$$\hat{w}_i = \hat{w}_i' \cdot \sigma + \mu \tag{51}$$

where $\{c_1, \cdots, c_n\}$ and $\{\mathcal{U}_1, \cdots, \mathcal{U}_n\}$ represent the codebook and the partitioned subsets of the real number space respectively, $\mu = \text{MEAN}(\{w_i\}_{i=1}^m)$, $\sigma = \text{STD}(\{w_i\}_{i=1}^m)$. $\mathcal{Q}(\cdot)$ is the quantization function. Within the quantization process, only the quantization operation $\mathcal{Q}(\cdot)$ is non-differentiable. To address this, the paper employs a straight-through estimator (STE) to approximate the gradients.

To address the limitations of VPT in complex remote sensing scenarios, where contextual information and inter-feature relationships are not sufficiently considered, CAPM first compresses the feature map $\boldsymbol{F} \in \mathbb{R}^{H \times W \times C}$ to obtain prompt weights:

$$\varepsilon = [\varepsilon_1, \cdots, \varepsilon_L] = \text{Softmax}(\text{Linear}(\text{GAP}(\boldsymbol{F}_2))) \tag{52}$$

These weights are then incorporated into the prompt components to derive $\boldsymbol{P}_w$, which dynamically adjusts according to the input:

$$\boldsymbol{P}_w = \sum_{l=1}^{L} \varepsilon_l \boldsymbol{P}_l \tag{53}$$

Finally, through an upsampling operation and skip connections, the prompts are fused back into the original feature:

$$\boldsymbol{F}_2' = \boldsymbol{F}_2 + \text{Conv}_{1\times1}(\text{Upsample}(\boldsymbol{P}_w)) \tag{54}$$

This process facilitates the extraction of task-relevant features, thereby enhancing the model's learning capacity.

3.5.3 MSF-SAM

Song et al. [60] proposed MSF-SAM, a sophisticated multistage fine-tuning methodology for SAM tailored to multispectral remote sensing crop detection. This approach surpasses other contemporary state-of-the-art techniques across a range of segmentation performance metrics for diverse crop types.

MSF-SAM incorporates a prefix adapter in its first stage, which is comprised of head convolution, depthwise convolution, and point-wise convolution. This adapter efficiently extracts salient crop

features and low-level semantic representations from multispectral imagery. These features are then compressed into three-channel, low-dimensional embeddings to align with the requirements of subsequent image encoding and the fine-tuning of the model's encoder. Notably, the depthwise convolution utilizes large kernel sizes to capture global spatial dependencies more effectively, while the point-wise convolution fosters inter-channel integration, enhancing feature interaction between the different spectral channels.

The low-level semantic information extracted in the first stage is used to guide the low-rank fine-tuning process in the second stage. In the second stage, LoRA layers are inserted into each transformer block within the image encoder. These LoRA-based fine-tuning layers encourage the model to focus more on high-level semantic information related to the crops while also reducing inferencing latency. Additionally, the authors claim that MAF-SAM exhibits strong temporal transferability, achieving good segmentation performance across different growth stages of crops.

3.5.4 SAM-RSIS

The application of SAM to remote sensing tasks encounters three prominent challenges: substantial domain divergence, SAM's limited capacity to model geographic and semantic information, and inadequate segmentation of complex and smaller objects.

To overcome these challenges, SAM-RSIS [61] performs fine-tuning at both the feature extraction and mask decoding stages to thoroughly adapt SAM to the instance segmentation task in remote sensing images. Specifically, in the first stage, inspired by ViT-Adapter [167], SAM-RSIS introduces a multi-scale adapter designed to effectively address the multi-scale object problem inherent in remote sensing images. This adapter integrates a spatial prior module to encode spatial information, along with four distinct scale extractors that collaborate with the ViT to more effectively capture multi-scale features in remote sensing imagery. In the second stage, the method fine-tunes the mask decoder using the prompt boxes generated by an object detector and the multi-scale features output by the adapter. Extensive empirical evidence underscores the superiority of the proposed fine-tuning approach over other SAM-based methods, demonstrating enhanced performance and robust generalization capacity.

3.5.5 Water-Adapter

The surface water extraction (SWE) task at ultra-high resolution poses considerable challenges due to the inherent multispectral variability of water surfaces. To mitigate this, Feng et al. proposed the Water-Adapter [62] to fine-tune the SAM for the SWE task. They incorporate learnable adapters into the SAM encoder, enabling SAM to capture domain-specific knowledge from remote sensing images. They also employ an explicit visual prompting (EVP) mechanism to integrate low-frequency components in water bodies.

In particular, the Water-Adapter introduces a series of simple yet effective adapter blocks within a ViT architecture. Borrowing from the approach in AdaptFormer [161], one adapter is inserted in parallel with the MLP, while another adapter is positioned subsequent to the MHA mechanism. This strategic placement allows for the fine-tuning of the output following the attention process, further refining the model's representation and enabling more precise extraction of domain-specific information. In addition, the EVP module leverages an adapter to integrate the features obtained after patch embedding with the low-frequency features extracted in the previous stage, which can be formulated as Eq. (55):

$$P^i = \mathrm{MLP}_{\mathrm{up}}(\mathrm{GELU}(\mathrm{MLP}^i_{\mathrm{tune}}(F_{\mathrm{pe}} + F_{\mathrm{lfc}}))) \quad (55)$$

Here, F_{pe} denotes the output from patch embedding tuning, while F_{lfc} represents the result obtained from tuning the low-frequency components. The activation function used is GELU. The component $\mathrm{MLP}^i_{\mathrm{tune}}$ is a linear layer responsible for generating distinct prompts within each adapter. Meanwhile, $\mathrm{MLP}_{\mathrm{up}}$ serves as a shared up-projection layer, applied uniformly across all adapters. P^i is the output prompt.

3.5.6 RoadSAM

In addition to the aforementioned Water-Adapter, Fang et al. [63] also applied the approach to road extraction tasks and introduced RoadSAM. Experiments on two road extraction datasets validated the effectiveness of RoadSAM.

Unlike Water-Adapter, RoadSAM explores multiple adapter insertion strategies: serial adapter, parallel adapter, and mixed adapter. These adapters share the same structure but differ in their insertion methods. Another difference lies in the information requirements of the tasks: while the SWE task

relies more on low-frequency information from water surfaces, high-frequency information is more critical for road extraction. Therefore, the low-frequency component in the EVP mechanism was replaced by high-frequency information to better aggregate the high-frequency features of roads in remote sensing images.

3.5.7 EarthMaker

EarthMaker [64] was the first large multimodal language model in the remote sensing domain to support visual prompting. It enables multi-granularity remote sensing image interpretation, including image-level, region-level, and point-level analysis. This paper also introduced RSVP, a multimodal, fine-grained, visual prompting dataset for remote sensing, extending the scope of existing datasets.

EarthMaker employs a unified visual encoder, integrating DINOv2-ViT L/14 [180] and CLIP-ConvNeXt [162] with a mixture of visual experts (MoV) [181], utilized for encoding multi-scale images and visual prompts, facilitating better understanding of the relationship between the two. To align the dimensionality of the prompts with that of the images, the prompts are repeated three times, after which the transformed prompts are encoded through MoV.

To support EarthMaker in performing image interpretation at varying levels of granularity, the authors also proposed a cross-domain, multi-stage training strategy. In the final stage, LoRA is applied to fine-tune the MHA mechanism within the transformer blocks, enabling the model to focus more effectively on user instructions.

3.6 Partial tuning

3.6.1 Approach

The above methods introduce new structures during the fine-tuning process. In certain scenarios, fine-tuning parts of the pre-trained model's parameters (such as the last few layers, biases, normalization layers, etc.) can also achieve good performance. Here, we introduce several partial tuning methods existing in remote sensing.

3.6.2 Zhang et al. module

Zhang et al. [65] tackled the challenge of few-shot automatic object recognition in SAR imagery, a field traditionally hindered by significant data dependency. To address this, the authors proposed a transfer learning-based approach to enhance performance with limited samples. The method begins by generating SAR-style image data using a style conversion model, enabling cross-domain data augmentation to mitigate the lack of SAR training data. During model training, the deep Brownian distance covariance pooling layer is introduced to refine feature representation by measuring differences in joint and marginal distributions of features. For fine-tuning, only the classifier is updated using a small amount of new data, while the model structure remains frozen. A knowledge distillation technique further strengthens the model by iteratively refining learned representations. Experimental results on the MSTAR dataset [182] demonstrate 80% accuracy in a 10-way 10-shot scenario.

3.6.3 Dastour and Hassan module

Dastour and Hassan [66] focused on leveraging deep transfer learning to tackle challenges in land use/land cover (LULC) classification, particularly the limitations posed by insufficient and imbalanced training data in remote sensing applications. The authors fix the backbone network during training and only train the new head structures. A comprehensive evaluation was performed on thirty-nine deep transfer learning models to assess their performance under consistent conditions. Among them, ResNet50, EfficientNetV2B0, and ResNet152 achieved superior results in terms of accuracy and kappa scores. ResNet50 also attains an impressive f1-score of 0.967 on the test set. These findings underscore the potential of deep transfer learning in enhancing LULC classification accuracy, offering practical guidance for future research in this domain.

3.6.4 MAFDN

The morphologically augmented fine-tuned DenseNet-121 (MAFDN) [67] employs advanced morphological techniques–such as erosion, dilation, blurring, and contrast enhancement to automate high-resolution LULC classification in IoT-enabled smart cities. By addressing challenges like noisy and heterogeneous data, MAFDN improves spatial pattern extraction and expands the training dataset. During fine-tuning, the pre-trained DenseNet-121 undergoes a key modification: the final fully connected layer, originally designed for 1000-class classification, is replaced by a new fully connected layer for 15 classes, optimized using the ADAM algorithm. This change significantly boosts the model's performance for LULC classification. Comparative results with

state-of-the-art methods show that MAFDN not only improves classification accuracy but also holds great potential for sustainable resource management and more personalized, data-driven urban planning services.

3.6.5 Huber et al. module

Huber et al. [68] applied deep transfer learning for yield prediction using remote sensing data, overcoming the challenge of limited ground truth by transferring knowledge from data-rich to data-scarce regions. The proposed framework includes unique histogram-based preprocessing and fine-tuning techniques, such as L^2-SP, BSS, and layer freezing, which optimize the loss function to address issues like catastrophic forgetting. L^2-SP and BSS work by refining the model through targeted loss function adjustments, allowing the model to focus more effectively on relevant features while avoiding overfitting. Additionally, the method employs the Optuna hyperparameter optimization framework [183] to tune hyperparameters, ensuring an optimal configuration for better model performance. Gaussian processes are also used to capture spatiotemporal patterns. The method improved soybean yield prediction in Argentina, achieving a 19% reduction in RMSE and a 39% increase in R^2 compared to models without transfer learning, demonstrating its potential for accurate yield forecasting, particularly in developing countries.

3.7 Improved tuning

3.7.1 Topic

This section introduces several special fine-tuning optimization approaches, including metric learning and data filtering. In addition to the methods mentioned here, there are many interesting optimization strategies waiting to be explored.

3.7.2 Zhang et al. module

Zhang et al. [69] proposed a two-stage fine-tuning method in generalized few-shot scenarios. In the first stage, they train all the parameters of the base detector on the base class sub-datasets to better establish general knowledge. In the second stage, they continue to fine-tune on few-shot datasets. Faster RCNN is chosen as the detection framework and initialized according to TFA [184]. The authors find that fine-tuning only the detector while freezing other parts performs poorly on few-shot datasets, so they chose to freeze only the backbone network and fine-tune the RPN and detector. Considering the limited samples in the second stage, the authors designed a metric-based discriminative loss to optimize the fine-tuning process. Additionally, the authors introduced dynamic freezing mechanisms and knowledge distillation techniques to overcome the catastrophic forgetting problem when introducing new tasks. The proposed method achieved significant improvements in multiple remote sensing few-shot tasks. This work demonstrates that optimizing training settings, such as freezing mechanisms and loss functions, during fine-tuning, can further enhance performance.

3.7.3 Ren et al. module

Ren et al. [70] optimized the fine-tuning process from a data perspective. The authors believed that data quality significantly impacts fine-tuning. Anomalous data affects performance, while redundant data increases the time needed for fine-tuning. They proposed a method that used only one-third of the remote sensing data for fine-tuning, but performance metrics only decreased by 1%, while the training time was reduced by nearly 70%.

Additionally, Wei et al. [185] achieved better results using only 6% of the data compared to methods using 100% in language tasks. Data greatly influences fine-tuning results. Optimizing the fine-tuning process from a data perspective can lead to impressive results and conclusions.

Additionally, some methods optimized models for natural vision scenes based on boundary constraints [71] or the multi-scale characteristics of remote sensing [72]. These concepts can be widely applied in further fine-tuning scenarios.

3.8 Comparison of paradigms

After introducing several fine-tuning paradigms, we summarize the differences between them here. *Full fine-tuning* trains all the parameters of the backbone, causing significant changes to the foundation model. For large pre-trained models, full fine-tuning may weaken their original general understanding ability and even lead to overfitting to the new data. *Adapter tuning* introduces more modules and parameters than other parameter-efficient fine-tuning methods, thus bringing more inferencing costs, but achieving relatively better performance. *Prompt tuning* introduces fewer new parameters and has

minimal impact on the model structure. It can bring impressive results in multimodal tasks, but its performance improvement is limited in single-modal tasks. The biggest advantage of *reparameterized tuning* is that it has no additional inferencing cost, making it cost-effective for developers. However, its performance improvement in visual tasks is limited. *Hybrid tuning* can combine the advantages of multiple fine-tuning methods but also introduces additional computational costs. *Partial tuning* has very low training costs (e.g., only training the last few layers of the backbone network) but usually performs poorly. *Improved tuning* can only be used in certain scenarios.

4 Datasets and metrics

4.1 Datasets and applications

Table 4 presents a summary of commonly used datasets in remote sensing fine-tuning. The first column details the modality distribution of all datasets. For modality, remote sensing fine-tuning works focus mainly on optical satellite images, SAR images, and multimodal data. Additionally, this field involves panchromatic images, point clouds, multispectral, and hyperspectral modalities. The third column outlines the tasks associated with these datasets. Most datasets used in remote sensing fine-tuning are geared towards tasks such as classification, object detection, semantic segmentation, change detection, and instance segmentation. Beyond these common tasks, some works also address cloud removal, image description, and pansharpening. The fourth and fifth columns show the number of categories and instances contained in all datasets. While some datasets have a small number of instances suitable for fine-grained studies in the few-shot domain, others have extremely uneven sample distributions across various classes, highlighting the need for research into long-tail distribution methods. The sixth column shows the resolution of the datasets. Some datasets have a limited resolution distribution, restricting the generalization ability of fine-tuned models, whereas others span a broad range of resolutions. Designing efficient tuning methods to leverage the advantages of multi-resolution data is a significant challenge. Most existing works on remote sensing fine-tuning are conducted on a single modality or a single task type. However, with the development of remote sensing foundation models, research into cross-modality and cross-task fine-tuning is expected to show higher generality and practicality.

4.2 Metrics

4.2.1 Targets

Evaluation metrics for fine-tuning algorithms primarily focus on three issues: performance, number of parameters, and computational cost. This subsection elaborates on each aspect.

4.2.2 Performance

Performance metrics quantify the practical effectiveness of fine-tuning algorithms. As these algorithms are typically task-specific, the selection of evaluation metrics depends on the application. For instance, object detection tasks commonly employ the average precision (AP) metric for bounding box evaluation, semantic segmentation utilizes mean intersection over union (mIoU), and image classification adopts top-k accuracy as standard benchmarks.

4.2.3 Number of parameters

Parameter efficiency represents a critical focus in parameter-efficient fine-tuning research, significantly influencing algorithmic performance, computational costs, and comparative fairness. Key metrics in this dimension include: (i) parameter count for additional architectural components (and their proportion relative to the base model), (ii) number of trainable parameters (and their proportion), and (iii) total parameter count for the full model. Additional architectural parameters typically refer to those introduced by structures like adapters or prompts. Trainable parameters encompass all modifiable elements during fine-tuning, including both parameter-efficient components (e.g., LoRA matrices) and partially adjusted pretrained parameters.

4.2.4 Computational cost

For large-scale models or high-concurrency deployment scenarios, computational cost constitutes a primary concern for researchers and developers. This dimension comprises two components: (i) training cost: time taken and GPU memory consumption during optimization, and (ii) inferencing cost: latency and memory requirements during deployment. Different fine-tuning methods exhibit distinct cost profiles. Partial-tuning approaches may reduce training costs by 50% compared to full fine-tuning

Table 4 A summary of datasets used in existing remote sensing tuning works, including modalities, tasks, and details of each dataset

Modality	Dataset	Task	Classes	Images	Resolution
Optical satellite imagery	Haze1K [186]	Cloud removal	3	1200	3.2 m
	RICE1 [187]	Cloud removal	2	500	—
	RICE2 [187]	Cloud removal	3	450	—
	LEVIR-CD [179]	Change detection	20	637	0.5 m
	LandSAT-CD [188]	Change detection	4	2385	30 m
	S2Looking [189]	Change detection	2	5000	0.5 m–0.8 m
	WHU-CD [178]	Change detection	2	7620	0.2 m
	NWPU-RESISC45 [190]	Classification	45	31,500	0.2 m–30 m
	UCM [191]	Classification	21	2100	0.3 m
	ISPRS [176]	Classification	6	71	—
	EuroSAT [192]	Classification	10	2700	0 m–10 m
	AID [193]	Classification	30	10,000	30 m
	NWPU-VHR10 [194]	Instance segmentation	10	800	0.5 m–2 m
	WHU Aerial Building Dataset [178]	Instance segmentation	6	8189	0.3 m
	DIOR [195]	Object detection	20	23,463	0.5 m–30 m
	DOTA [196]	Object detection	15	2806	—
	HRSC [197]	Object detection	4	1061	0.4 m–2 m
	UCAS-AOD [198]	Object detection	3	310	30 m
	NWPU-VHR10 [194]	Object detection	10	800	0.5 m–2 m
	ISPRS Potsdam [176]	Semantic segmentation	5	38	GSD=5 cm
	ISPRS Vaihingen [176]	Semantic segmentation	5	33	GSD=9 cm
	LoveDA Urban [199]	Semantic segmentation	7	5987	0.3m
	NWPU-VHR10 [194]	Semantic segmentation	10	800	0.5 m–2 m
	NWPU-RESISC [190]	Semantic segmentation	45	31,500	0.2 m–30 m
	iSAID [200]	Semantic segmentation	15	2806	—
Synthetic aperture radar (SAR) imagery	MSTAR [182]	Classification	10	17,658	0.3 m
	FUSAR-Ship [201]	Classification	15	5000+	0.5 m–500 m
	SRSDD [202]	Classification	6	30	1 m
	FUSAR-Map1.0 [203]	Classification	12	610	1 m–4 m
	FUSAR-Map2.0 [204]	Classification	10	738	1 m
	SSDD [205]	Instance segmentation	1	1160	1 m–15 m
	FUSAR-Ship [201]	Object detection	15	5000+	0.5 m–500 m
	SRSDD [202]	Object detection	6	666	1 m
	AIR-PolSAR-Seg [206]	Semantic segmentation	6	2000	8 m
	PolSAR-ZG [207]	Semantic segmentation	6	6	5 m
	FUSAR-Map1.0 [203]	Semantic segmentation	12	610	1 m–4 m
	FUSAR-Map2.0 [204]	Semantic segmentation	10	738	1 m
Text & imagery	FloodNet [208]	Semantic segmentation	9	3200	1.5 cm (UAV)
	AID [193]	Image caption	30	10,000	30 m
	MMBench [209]	Image caption	1	2974 questions	—
	MME [210]	Image caption	2	2194 questions	—
	SEEDBench [211]	Image caption	4	24,371 questions	—
	Fit-RS [53]	Image caption	11	1800.8k	—
	SkyEye-968k [52]	Image caption	2	968k	—
	RSICD [212]	Image caption	30	10,921	—
	RSITMD [213]	Image caption	32	4743	30 m
	UCM [191]	Image caption	21	2100	0.3 m
Panchromatic (PAN) imagery	Pancollection [214]	Pansharpening	3	26,107	0.3 m–2 m
Point cloud	Harbor of Tobermory [215]	Classification	7	7,181,982	—
	University of Houston [215]	Classification	7	4,436,470	—
Multi-spectral imagery	DSTL [216]	Semantic segmentation	10	—	0.31 m–1.24 m
	RIT-18 [217]	Semantic segmentation	18	21	0.047 m (UAV)
Hyper-spectral imagery	Houston 2013 Dataset [218]	Classification	15	15,029	—
	MUUFL Dataset [219]	Classification	11	53,587	—
	Trento Dataset [220]	Classification	6	30,214	—

while maintaining equivalent inferencing costs. LoRA-based methods achieve training costs comparable to adapter-based approaches, yet demonstrate significantly lower inferencing overhead. Task-specific requirements substantially influence the prioritization of these cost factors. Research applications might emphasize training efficiency, whereas production systems typically prioritize inferencing costs.

5 Future directions

5.1 Few-shot scenarios

Previous studies [154] have suggested that advanced fine-tuning techniques can effectively leverage foundation models to address few-shot issues. In remote sensing, categories such as landfills [13], wildfires [221], power plants [14], and slums [222] commonly face few-shot problems. Most existing fine-tuning work in remote sensing has been conducted on multi-category datasets, while in-depth exploration of specific categories (e.g., water [62]) is relatively rare. However, many remote sensing studies focusing on a single category, like wildfires [223, 224], have achieved significant successes. By applying fine-tuning techniques, we can enhance the automatic identification efficiency of specific objects and further explore the potential correlations between the spatial and temporal distributions of these objects and social and economic factors.

Additionally, the relationship between sample size and fine-tuning efficiency is also worth exploring. Capturing and annotating many remote sensing objects are both very costly. Therefore, achieving improved performance with as little data as possible is crucial. Currently, there is no systematic research in remote sensing that specifically addresses this issue. Research in this area can effectively reduce the data requirements for fine-tuning on downstream tasks and lower the computational costs during the model training process.

5.2 Further remote sensing tasks

Existing fine-tuning efforts in remote sensing largely focus on common tasks such as classification, detection, segmentation, and change detection. Additionally, there are numerous other tasks in the field that require advanced fine-tuning algorithms, including super-resolution [225], image restoration [226], cloud removal [227], image registration [228], object tracking [229], and trajectory prediction [230]. With the advent of powerful remote sensing foundation models (RSFMs), fine-tuning techniques have the potential to revolutionize many more remote sensing tasks in the future.

5.3 Validation on more RSFMs

Many fine-tuning efforts have utilized foundation models trained on general image datasets such as ImageNet. In the past two years, numerous foundation models specifically designed for remote sensing have been proposed. Theoretically, RSFMs can achieve better performance on remote sensing tasks than general foundation models. When combined with advanced fine-tuning techniques, the advantages of RSFMs are expected to be further amplified. We encourage researchers to conduct more testing on and optimization of various RSFMs.

5.4 Design with remote sensing characteristics

Our investigation shows that many remote sensing fine-tuning works directly apply existing parameter-efficient fine-tuning (PEFT) techniques to remote sensing tasks. In fact, most existing fine-tuning techniques were designed for specific scenarios, such as plain adapters [153] and LoRA [171] in NLP, and Mona [169] in CV. These existing techniques cannot leverage the prior characteristics of remote sensing images, leading to limited performance. For example, optical remote sensing images feature characteristics such as small objects, high density, and large scale, while SAR images have unique scattering properties. Fine-tuning techniques designed with these characteristics should demonstrate stronger performance.

5.5 Further fine-tuning paradigms

Remote sensing fine-tuning techniques are currently at an early and rapidly developing stage. Existing works are mostly based on paradigms such as adapters, prompts, and reparameterization. NLP and CV fields have seen the emergence of new fine-tuning techniques, including parameter selection [231, 232], PEFT pruning [233], PEFT quantization [234], and PEFT knowledge distillation [235, 236]. These novel techniques may offer new insights for remote sensing tasks.

5.6 Further hybrid methods

In Fig. 9, we present both existing and unexplored hybrid methods. Currently, there are three

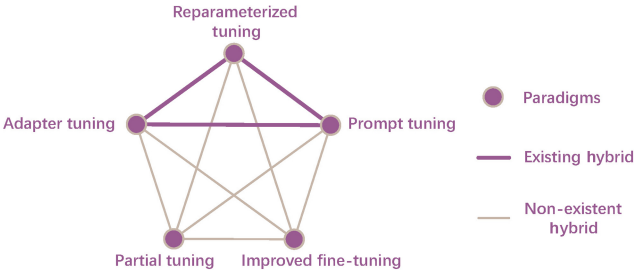

Fig. 9 Existing and unexplored hybrid tuning methods. Purple thick lines represent existing hybrids, and gray thin lines represent those yet to be explored.

combined approaches: adapters+reparameterization, adapters+prompts, and prompts+reparameterization. Each fine-tuning paradigm has its own strengths and weaknesses. New hybrid methods may significantly impact specific remote sensing tasks in the future.

5.7 Theoretical analysis of PEFT

The principles by which PEFT techniques surpass full fine-tuning are currently not well understood. Previous work [9] systematically analyzed the theoretical principles of PEFT in NLP across more than 100 tasks, yielding significant research findings. There are also hundreds of important tasks in the remote sensing field. It is worthwhile to explore the unique theoretical principles of remote sensing PEFT methods in the future.

5.8 Setting optimization

PEFT methods are sensitive to settings, including optimizers, learning rates, and loss functions. It is essential to investigate how to better harness the potential of PEFT methods in remote sensing tasks. Additionally, PEFT techniques contain numerous hyperparameters, such as the intermediate dimensions of adapters and the rank in LoRA, which can influence the performance gains of RSFM in downstream tasks. We encourage researchers to design a series of detailed settings that can improve the performance of PEFT methods in general remote sensing scenarios.

5.9 Scaling law

The concept of scaling law [237, 238] has frequently been mentioned in the domain of large models recently, emphasizing the correlation between performance and the size of models and amount of data. In remote sensing, robust foundation models need to demonstrate their performance on downstream tasks through fine-tuning techniques. Therefore, performance is influenced not only by the foundation models themselves, but also by the fine-tuning techniques. Investigating the scaling laws of remote sensing foundation models with fine-tuning techniques is an interesting and significant area for future research.

6 Conclusions

In the era of large models and big data, foundation models and fine-tuning techniques are leading new trends in remote sensing research. This survey has systematically reviewed fine-tuning techniques in remote sensing. Existing techniques are categorized based on the relationship between the tuned parameters and the pre-trained models to help readers understand the technological trajectory. The survey concludes by highlighting nine directions worth exploring in this field. We hope this survey encourages researchers to utilize fine-tuning techniques to improve their deep learning results across various remote sensing tasks. We also hope researchers discover more interesting directions for remote sensing fine-tuning through this survey.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62495061, 62495064, and 62476143), the Tsinghua–Tencent Joint Laboratory for Internet Innovation Technology, and the Shuimu Tsinghua Scholar Program.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. The author Shi-Min Hu is the Editor-in-Chief of this journal.

References

[1] Lu, S.; Guo, J.; Zimmer-Dauphinee, J. R.; Nieusma, J. M.; Wang, X.; VanValkenburgh, P.; Wernke, S. A.; Huo, Y. Vision foundation models in remote sensing: A survey. *arXiv preprint* arXiv:2408.03464, 2024.

[2] Xiao, A.; Xuan, W.; Wang, J.; Huang, J.; Tao, D.; Lu, S.; Yokoya, N. Foundation models for remote sensing and earth observation: A survey. *arXiv preprint* arXiv:2410.16602, 2024.

[3] Min, B.; Ross, H.; Sulem, E.; Ben Veyseh, A. P.;

Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; Roth, D. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys* Vol. 56, No. 2, Article No. 30, 2024.

[4] Liu, X.; Zhou, T.; Wang, C.; Wang, Y.; Wang, Y.; Cao, Q.; Du, W.; Yang, Y.; He, J.; Qiao, Y.; et al. Toward the unification of generative and discriminative visual foundation model: A survey. *The Visual Computer* Vol. 41, No. 5, 3371–3412, 2025.

[5] Macarringue, L. S.; Bolfe, É. L.; Pereira, P. R. M. Developments in land use and land cover classification techniques in remote sensing: A review. *Journal of Geographic Information System* Vol. 14, No. 1, 1–28, 2022.

[6] Olson, D.; Anderson, J. Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. *Agronomy Journal* Vol. 113, No. 2, 971–992, 2021.

[7] Ren, X.; Li, X.; Ren, K.; Song, J.; Xu, Z.; Deng, K.; Wang, X. Deep learning-based weather prediction: A survey. *Big Data Research* Vol. 23, Article No. 100178, 2021.

[8] Qiao, D.; Liu, G.; Lv, T.; Li, W.; Zhang, J. Marine vision-based situational awareness using discriminative deep learning: A survey. *Journal of Marine Science and Engineering* Vol. 9, No. 4, Article No. 397, 2021.

[9] Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C. M.; Chen, W.; et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* Vol. 5, No. 3, 220–235, 2023.

[10] Jia, M.; Tang, L.; Chen, B. C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S. N. Visual prompt tuning. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13693.* Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 709–727, 2022.

[11] Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Chen, Y.; Chan, C. M.; Chen, W.; et al. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint* arXiv:2203.06904, 2022.

[12] Yu, B. X. B.; Chang J.; Wang H.; Liu L.; Wang S.; Wang Z.; Lin J.; Xie L.; Li H.; Lin Z.; et al. Visual tuning. *ACM Computing Surveys* Vol. 56, No. 12, 1–38, 2024.

[13] Sun, X.; Yin, D.; Qin, F.; Yu, H.; Lu, W.; Yao, F.; He, Q.; Huang, X.; Yan, Z.; Wang, P.; et al. Revealing influencing factors on global waste distribution via deep-learning based dumpsite detection from satellite imagery. *Nature Communications* Vol. 14, No. 1, Article No. 1444, 2023.

[14] Yin, W.; Diao, W.; Wang, P.; Gao, X.; Li, Y.; Sun, X. PCAN: Part-based context attention network for thermal power plant detection in remote sensing imagery. *Remote Sensing* Vol. 13, No. 7, Article No. 1243, 2021.

[15] Sun, X.; Wang, B.; Wang, Z.; Li, H.; Li, H.; Fu, K. Research progress on few-shot learning for remote sensing image interpretation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 14, 2387–2402, 2021.

[16] Yin, D.; Han, X.; Li, B.; Feng, H.; Bai, J. Parameter-efficient is not sufficient: Exploring parameter, memory, and time efficient adapter tuning for dense predictions. In: Proceedings of the 32nd ACM International Conference on Multimedia, 1398–1406, 2024.

[17] Anderson, K.; Ryan, B.; Sonntag, W.; Kavvada, A.; Friedl, L. Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science* Vol. 20, No. 2, 77–96, 2017.

[18] Zhu, X. X.; Tuia, D.; Mou, L.; Xia, G. S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* Vol. 5, No. 4, 8–36, 2017.

[19] Hu, L.; Yu, H.; Lu, W.; Yin, D.; Sun, X.; Fu, K. AiRs: Adapter in remote sensing for parameter-efficient transfer learning. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5605218, 2024.

[20] Hu, L.; Lu, W.; Yu, H.; Yin, D.; Sun, X.; Fu, K. TEA: A training-efficient adapting framework for tuning foundation models in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5648118, 2024.

[21] Mei, L.; Ye, Z.; Xu, C.; Wang, H.; Wang, Y.; Lei, C.; Yang, W.; Li, Y. SCD-SAM: Adapting segment anything model for semantic change detection in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5626713, 2024.

[22] Zhang, Z.; Liu, F.; Liu, C.; Tian, Q.; Qu, H. ACTNet: A dual-attention adapter with a CNN-transformer network for the semantic segmentation of remote sensing imagery. *Remote Sensing* Vol. 15, No. 9, Article No. 2363, 2023.

[23] Pu, X.; Jia, H.; Zheng, L.; Wang, F.; Xu, F. ClassWise-SAM-adapter: Parameter-efficient fine-tuning adapts

segment anything to SAR domain for semantic segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 18, 4791–4804, 2025.

[24] Wang, Q.; Yin, J.; Jiang, H.; Feng, J.; Zhang, G. Disentangled foreground-semantic adapter network for generalized aerial image few-shot semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5641114, 2024.

[25] Li, K.; Cao, X.; Meng, D. A new learning paradigm for foundation model-based remote-sensing change detection. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5610112, 2024.

[26] Zhang, Z.; Tan, S.; Zhou, Y. CloudformerV3: Multi-scale adapter and multi-level large window attention for cloud detection. *Applied Sciences* Vol. 13, No. 23, Article No. 12857, 2023.

[27] Cai, Z.; Ning, J.; Ding, Z.; Duo, B. Additional self-attention transformer with adapter for thick haze removal. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 6004705, 2024.

[28] Ding, L.; Zhu, K.; Peng, D.; Tang, H.; Yang, K.; Bruzzone, L. Adapting segment anything model for change detection in VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5611711, 2024.

[29] Li, G.; Togo, R.; Maeda, K.; Sako, A.; Yamauchi, I.; Hayakawa, T.; Nakamae, S.; Ogawa, T.; Haseyama, M. Algal bed region segmentation based on a ViT adapter using aerial images for estimating CO_2 absorption capacity. *Remote Sensing* Vol. 16, No. 10, Article No. 1742, 2024.

[30] Huang, T. Efficient remote sensing with harmonized transfer learning and modality alignment. *arXiv preprint* arXiv:2404.18253, 2024.

[31] Wu, R.; Zhang, Z.; Deng, S.; Duan, Y.; Deng, L. J. PanAdapter: Two-stage fine-tuning with spatial-spectral priors injecting for pansharpening. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 39, No. 8, 8450–8459, 2025.

[32] Fu, H.; Liu, H.; Yuan, J.; He, X.; Lin, J.; Li, Z. YOLO-adaptor: A fast adaptive one-stage detector for non-aligned visible-infrared object detection. *IEEE Transactions on Intelligent Vehicles* DOI: 10.1109/TIV.2024.3393015, 2024.

[33] Yuan, Y.; Zhan, Y.; Xiong, Z. Parameter-efficient transfer learning for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5619014, 2023.

[34] Zhang, J.; Rao, Y.; Huang, X.; Li, G.; Zhou, X.; Zeng, D. Frequency-aware multi-modal fine-tuning for few-shot open-set remote sensing scene classification. *IEEE Transactions on Multimedia* Vol. 26, 7823–7837, 2024.

[35] Yan, Z.; Li, J.; Li, X.; Zhou, R.; Zhang, W.; Feng, Y.; Diao, W.; Fu, K.; Sun, X. RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5625716, 2023.

[36] Immanuel, S. A.; Sinulingga, H. R. Learnable prompt for few-shot semantic segmentation in remote sensing domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2755–2761, 2024.

[37] Kong, Y.; Cheng, Y.; Chen, Y.; Wang, X. Joint classification of hyperspectral image and LiDAR data based on spectral prompt tuning. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5521312, 2024.

[38] Gao, K.; You, X.; Li, K.; Chen, L.; Lei, J.; Zuo, X. Attention prompt-driven source-free adaptation for remote sensing images semantic segmentation. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 6012105, 2024.

[39] Singha, M.; Jha, A.; Solanki, B.; Bose, S.; Banerjee, B. APPLeNet: Visual attention parameterized prompt learning for few-shot remote sensing image generalization using CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2024–2034, 2023.

[40] Wang, L.; Qiu, H.; Zhang, M.; Meng, F.; Wu, Q.; Li, H. DP-RSCAP: Dual prompt-based scene and entity network for remote sensing image captioning. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 7129–7132, 2024.

[41] Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; Shi, Z. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 4701117, 2024.

[42] Li, L. CPSeg: Finer-grained image semantic segmentation via chain-of-thought language prompting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 502–511, 2024.

[43] Lan, L.; Wang, F.; Zheng, X.; Wang, Z.; Liu, X. Efficient prompt tuning of large vision-language model for fine-grained ship classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 63, Article No. 5600810, 2024.

[44] Zhao, L.; Xu, L.; Zhao, L.; Zhang, X.; Wang, Y.;

Ye, D.; Peng, J.; Li, H. Continual learning for remote sensing image scene classification with prompt learning. *IEEE Geoscience and Remote Sensing Letters* Vol. 20, Article No. 6012005, 2023.

[45] Zhu, J.; Li, Y.; Yang, K.; Guan, N.; Fan, Z.; Qiu, C.; Yi, X. MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5610413, 2024.

[46] Fang, L.; Kuang, Y.; Liu, Q.; Yang, Y.; Yue, J. Rethinking remote sensing pretrained model: Instance-aware visual prompting for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5626713, 2023.

[47] Liu, C.; Zhao, R.; Chen, J.; Qi, Z.; Zou, Z.; Shi, Z. A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5622018, 2023.

[48] Bi, H.; Feng, Y.; Diao, W.; Wang, P.; Mao, Y.; Fu, K.; Wang, H.; Sun, X. Prompt-and-transfer: Dynamic class-aware enhancement for few-shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 47, No. 1, 131–148, 2025.

[49] Osco, L. P.; Wu, Q.; de Lemos, E. L.; Gonçalves, W. N.; Ramos, A. P. M.; Li, J.; Marcato, J. The Segment Anything Model (SAM) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation* Vol. 124, Article No. 103540, 2023.

[50] Ulku, I.; Ozgur Tanriover, O.; Akagündüz, E. LoRA-NIR: Low-rank adaptation of vision transformers for remote sensing with near-infrared imagery. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 5004505, 2024.

[51] Xue, B.; Cheng, H.; Yang, Q.; Wang, Y.; He, X. Adapting segment anything model to aerial land cover classification with low-rank adaptation. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 2502605, 2024.

[52] Zhan, Y.; Xiong, Z.; Yuan, Y. SkyEyeGPT: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 221, 64–77, 2025.

[53] Luo, J.; Pang, Z.; Zhang, Y.; Wang, T.; Wang, L.; Dang, B.; Lao, J.; Wang, J.; Chen, J.; Tan, Y.; et al. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint* arXiv:2406.10100, 2024.

[54] Tian, Z.; Chen, Z.; Sun, Q. Learning de-biased representations for remote-sensing imagery. *arXiv preprint* arXiv:2410.04546, 2024.

[55] Wang, M.; Zhou, L.; Zhang, K.; Li, X.; Hao, M.; Ye, Y. ESAM-CD: Fine-tuned EfficientSAM network with LoRA for weakly supervised remote sensing image change detection. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 4708616, 2024.

[56] Lu, X.; Weng, Q. Multi-LoRA fine-tuned segment anything model for urban man-made object extraction. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5637519, 2024.

[57] Zhong, Y.; Wu, X.; Deng, L. J.; Cao, Z. SSDiff: Spatial-spectral integrated diffusion model for remote sensing pansharpening. *arXiv preprint* arXiv:2404.11537, 2024.

[58] Pu, X.; Xu, F. Low-rank adaption on transformer-based oriented object detector for satellite onboard processing of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 63, Article No. 5202213, 2025.

[59] Dong, Z.; Gu, Y.; Liu, T. UPetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5616613, 2024.

[60] Song, B.; Yang, H.; Wu, Y.; Zhang, P.; Wang, B.; Han, G. A multispectral remote sensing crop segmentation method based on segment anything model using multistage adaptation fine-tuning. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 4408818, 2024.

[61] Luo, M.; Zhang, T.; Wei, S.; Ji, S. SAM-RSIS: Progressively adapting SAM with box prompting to remote sensing image instance segmentation. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 4413814, 2024.

[62] Feng, W.; Guan, F.; Tu, J.; Sun, C.; Xu, W. Water-Adapter: Adapting the segment anything model for surface water extraction in optical very-high-resolution remotely sensed imagery. *Remote Sensing Letters* Vol. 15, No. 11, 1132–1142, 2024.

[63] Feng, W.; Guan, F.; Sun, C.; Xu, W. Road-SAM: Adapting the segment anything model to road extraction from large very-high-resolution optical remote sensing images. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 6012605, 2024.

[64] Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Li, J.; Mao, X. EarthMarker: A visual prompting multimodal large language model for remote sensing. *IEEE*

Transactions on Geoscience and Remote Sensing Vol. 63, Article No. 3523505, 2025.

[65] Zhang, C.; Dong, H.; Deng, B. Improving pre-training and fine-tuning for few-shot SAR automatic target recognition. *Remote Sensing* Vol. 15, No. 6, Article No. 1709, 2023.

[66] Dastour, H.; Hassan, Q. K. A comparison of deep transfer learning methods for land use and land cover classification. *Sustainability* Vol. 15, No. 10, Article No. 7854, 2023.

[67] Sahu, M.; Dash, R.; Kumar Mishra, S.; Humayun, M.; Alfayad, M.; Assiri, M. A deep transfer learning model for green environment security analysis in smart city. *Journal of King Saud University - Computer and Information Sciences* Vol. 36, No. 1, Article No. 101921, 2024.

[68] Huber, F.; Inderka, A.; Steinhage, V. Leveraging remote sensing data for yield prediction with deep transfer learning. *Sensors* Vol. 24, No. 3, Article No. 770, 2024.

[69] Zhang, T.; Zhang, X.; Zhu, P.; Jia, X.; Tang, X.; Jiao, L. Generalized few-shot object detection in remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 195, 353–364, 2023.

[70] Ren, Y.; Zhang, T.; Han, Z.; Li, W.; Wang, Z.; Ji, W.; Qin, C.; Liang, C.; Jiao, L. A novel adaptive fine-tuning algorithm for multimodal models: Self-optimizing classification and selection of high-quality datasets in remote sensing. *arXiv preprint* arXiv:2409.13345, 2024.

[71] Ma, X.; Wu, Q.; Zhao, X.; Zhang, X.; Pun, M. O.; Huang, B. SAM-assisted remote sensing imagery semantic segmentation with object and boundary constraints. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5636916, 2024.

[72] Zhou, X.; Liang, F.; Chen, L.; Liu, H.; Song, Q.; Vivone, G.; Chanussot, J. MeSAM: Multiscale enhanced segment anything model for optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5623515, 2024.

[73] Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5607315, 2022.

[74] Wang, D.; Zhang, J.; Du, B.; Xia, G. S.; Tao, D. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5608020, 2022.

[75] Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D. B.; Ermon, S. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. *arXiv preprint* arXiv:2207.08051, 2023.

[76] Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing remote sensing foundation model via multitask pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 17, 11632–11654, 2024.

[77] Wang, D.; Hu, M.; Jin, Y.; Miao, Y.; Yang, J.; Xu, Y.; Qin, X.; Ma, J.; Sun, L.; Li, C.; et al. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* DOI: 10.1109/TPAMI.2025.3557581, 2025.

[78] Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; Khan, F. S. GeoChat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27831–27840, 2024.

[79] Huang, Z.; Pan, Z.; Lei, B. Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. *Remote Sensing* Vol. 9, No. 9, Article No. 907, 2017.

[80] Liu, X.; Chi, M.; Zhang, Y.; Qin, Y. Classifying high resolution remote sensing images by fine-tuned VGG deep networks. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 7137–7140, 2018.

[81] Bazi, Y.; Al Rahhal, M. M.; Alhichri, H.; Alajlan, N. Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification. *Remote Sensing* Vol. 11, No. 24, Article No. 2908, 2019.

[82] Zhang, T.; Zhuang, Y.; Chen, H.; Chen, L.; Wang, G.; Gao, P.; Dong, H. Object-centric masked image modeling-based self-supervised pretraining for remote sensing object detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 16, 5013–5025, 2023.

[83] Wu, J.; Lang, C.; Cheng, G.; Xie, X.; Han, J. Retentive compensation and personality filtering for few-shot remote sensing object detection. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 34, No. 7, 5805–5817, 2024.

[84] Huang, X.; He, B.; Tong, M.; Wang, D.; He, C. Few-shot object detection on remote sensing images via shared attention module and balanced fine-tuning strategy. *Remote Sensing* Vol. 13, No. 19, Article No. 3816, 2021.

[85] Zhao, Z.; Tang, P.; Zhao, L.; Zhang, Z. Few-shot object

detection of remote sensing images via two-stage fine-tuning. *IEEE Geoscience and Remote Sensing Letters* Vol. 19, Article No. 8021805, 2021.

[86] Kim, Y.; Park, J.; Kim, S.; Jeon, M. Rethinking feature backbone fine-tuning for remote sensing object detection. *arXiv preprint* arXiv:2407.15143, 2024.

[87] Gou, J.; Yu, B.; Maybank, S. J.; Tao, D. Knowledge distillation: A survey. *International Journal of Computer Vision* Vol. 129, No. 6, 1789–1819, 2021.

[88] Zhang, X.; Li, X.; Chen, G.; Liao, P.; Wang, T.; Yang, H.; He, C.; Zhou, W.; Sun, Y. A deep transfer learning framework using teacher–student structure for land cover classification of remote-sensing imagery. *IEEE Geoscience and Remote Sensing Letters* Vol. 21, Article No. 6006405, 2023.

[89] Zhang, M.; Yin, D.; Li, Z.; Zhao, Z. Improved identification of forest types in the loess plateau using multi-source remote sensing data, transfer learning, and neural residual networks. *Remote Sensing* Vol. 16, No. 12, Article No. 2096, 2024.

[90] Zhang, J.; Zhang, L. Clouds and haze co-removal based on weight-tuned overlap refinement diffusion model for remote sensing images. In: Proceedings of the IEEE International Conference on Image Processing, 1635–1641, 2024.

[91] Stojnic, V.; Risojevic, V. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1182–1191, 2021.

[92] Mañas, O.; Lacoste, A.; Giró-i-Nieto, X.; Vazquez, D.; Rodríguez, P. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9414–9423, 2021.

[93] Li, W.; Chen, K.; Chen, H.; Shi, Z. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 60, Article No. 5405516, 2021.

[94] Akiva, P.; Purri, M.; Leotta, M. Self-supervised material and texture representation learning for remote sensing tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8203–8215, 2022.

[95] Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-aware self-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 10181–10190, 2021.

[96] Wang, Y.; Albrecht, C. M.; Zhu, X. X. Self-supervised vision transformers for joint SAR-optical representation learning. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 139–142, 2022.

[97] Scheibenreif, L.; Hanna, J.; Mommert, M.; Borth, D. Self-supervised vision transformers for land-cover segmentation and classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 1422–1431, 2022.

[98] Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5612822, 2022.

[99] Li, W.; Chen, K.; Shi, Z. Geographical supervision correction for remote sensing representation learning. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 60, Article No. 5411520, 2022.

[100] Jain, P.; Schoen-Phelan, B.; Ross, R. Self-supervised learning for invariant representations from multi-spectral and SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 15, 7797–7808, 2022.

[101] Zhang, T.; Gao, P.; Dong, H.; Zhuang, Y.; Wang, G.; Zhang, W.; Chen, H. Consecutive pre-training: A knowledge transfer learning strategy with relevant unlabeled data for remote sensing domain. *Remote Sensing* Vol. 14, No. 22, Article No. 5675, 2022.

[102] Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, 1–15, 2023.

[103] Tao, C.; Qi, J.; Zhang, G.; Zhu, Q.; Lu, W.; Li, H. TOV: The Original Vision model for optical remote sensing image understanding via self-supervised learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 16, 4916–4930, 2023.

[104] Muhtar, D.; Zhang, X.; Xiao, P.; Li, Z.; Gu, F. CMID: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, Article No. 5607817, 2023.

[105] Mall, U.; Hariharan, B.; Bala, K. Change-aware sampling and contrastive learning for satellite images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5261–5270, 2023.

[106] Prexl, J.; Schmitt, M. Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2136–2144, 2023.

[107] Hu, Y.; Yuan, J.; Wen, C.; Lu, X.; Liu, Y.; Li, X. RSGPT: A remote sensing vision language model and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 224, 272–286, 2025.

[108] Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; Chen, C. Towards geospatial foundation models via continual pretraining. *arXiv preprint* arXiv:2302.04476, 2023.

[109] Bastani, F.; Wolters, P.; Gupta, R.; Ferdinando, J.; Kembhavi, A. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 16772–16782, 2023.

[110] Yao, F.; Lu, W.; Yang, H.; Xu, L.; Liu, C.; Hu, L.; Yu, H.; Liu, N.; Deng, C.; Tang, D.; et al. RingMo-sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 61, 1–21, 2023.

[111] Reed, C. J.; Gupta, R.; Li S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4088–4099, 2023.

[112] Wang, Y.; Zhang, T.; Zhao, L.; Hu, L.; Wang, Z.; Niu, Z.; Cheng, P.; Chen, K.; Zeng, X.; Wang, Z.; et al. RingMo-lite: A remote sensing lightweight network with CNN-transformer hybrid framework. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, 1–20, 2024.

[113] Wang, Y.; Albrecht, C. M.; Ali Braham, N. A.; Liu, C.; Xiong, Z.; Zhu, X. X. Decoupling common and unique representations for multimodal self-supervised learning. *arXiv preprint* arXiv:2309.05300, 2023.

[114] Feng, Y.; Wang, P.; Diao, W.; He, Q.; Hu, H.; Bi, H.; Sun, X.; Fu, K. A self-supervised cross-modal remote sensing foundation model with multi-domain representation and cross-domain fusion. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 2239–2242, 2023.

[115] Wang, Y.; Hernández, H. H.; Albrecht, C. M.; Zhu, X. X. Feature guided masked autoencoder for self-supervised learning in remote sensing. *arXiv preprint* arXiv:2310.18653, 2024.

[116] Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model. *arXiv preprint* arXiv:2305.02034, 2023.

[117] Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint* arXiv:2310.18660, 2023.

[118] Fuller, A.; Millard, K.; Green, J. R. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *arXiv preprint* arXiv:2311.00566, 2023.

[119] Irvin, J.; Tao, L.; Zhou, J.; Ma, Y.; Nashold, L.; Liu, B.; Ng, A. Y. USat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint* arXiv:2312.02199, 2023.

[120] Tang, M.; Cozma, A.; Georgiou, K.; Qi, H. Cross-scale MAE: A tale of multiscale exploitation in remote sensing. In: Proceedings of the 37th International Conference on Neural Information Processing Systems, Article No. 879, 2023.

[121] Dumeur, I.; Valero, S.; Inglada, J. Self-supervised spatio-temporal representation learning of satellite image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 17, 4350–4367, 2024.

[122] Smith, M. J.; Fleming, L.; Geach, J. E. EarthPT: A time series foundation model for Earth observation. *arXiv preprint* arXiv:2309.07207, 2024.

[123] Huang, Z.; Zhang, M.; Gong, Y.; Liu, Q.; Wang, Y. Generic knowledge boosted pretraining for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, 1–13, 2024.

[124] Tian, J.; Lei, J.; Zhang, J.; Xie, W.; Li, Y. SwiMDiff: Scene-wide matching contrastive learning with diffusion constraint for remote sensing image. *arXiv preprint* arXiv:2401.05093, 2024.

[125] Dong, Z.; Gu, Y.; Liu, T. Generative ConvNet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 5603816, 2024.

[126] Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 46, No. 8, 5227–5244, 2024.

[127] Tseng, G.; Cartuyvels, R.; Zvonkov, I.; Purohit, M.; Rolnick, D.; Kerner, H. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint* arXiv:2304.14065, 2023.

[128] Noman, M.; Naseer, M.; Cholakkal, H.; Anwar, R. M.; Khan, S.; Khan, F. S. Rethinking transformers pre-training for multi-spectral satellite imagery.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27811–27819, 2024.

[129] Li, W.; Yang, W.; Liu, T.; Hou, Y.; Li, Y.; Liu, Z.; Liu, Y.; Liu, L. Predicting gradient is better: Exploring self-supervised learning for SAR ATR with a joint-embedding predictive architecture. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 218, 326–338, 2024.

[130] Bountos, N. I.; Ouaknine, A.; Rolnick, D. FoMo-Bench: A multimodal, multi-scale and multi-task Forest Monitoring Benchmark for remote sensing foundation models. *arXiv preprint* arXiv:2312.10114, 2024.

[131] Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 27662–27673, 2024.

[132] Dong, Z.; Gu, Y.; Liu, T. UPetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, 1–13, 2024.

[133] Han, B.; Zhang, S.; Shi, X.; Reichstein, M. Bridging remote sensors with multisensor geospatial foundation models. *arXiv preprint* arXiv:2404.01260, 2024.

[134] Wanyan, X.; Seneviratne, S.; Shen, S.; Kirley, M. Extending global–local view alignment for self-supervised learning with remote sensing imagery. *arXiv preprint* arXiv:2303.06670, 2023.

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

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

[137] Nedungadi, V.; Kariryaa, A.; Oehmcke, S.; Belongie, S.; Igel, C.; Lang, N. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In: *Computer Vision – ECCV 2024. Lecture Notes in Computer Science, Vol. 15122*. Sun, Q.; Zhou, H.; Zhou, W.; Li, L.; Li, H. Eds. Springer Cham, 164–182, 2025.

[138] Zhang, M.; Liu, Q.; Wang, Y. CtxMIM: Context-enhanced masked image modeling for remote sensing image understanding. *arXiv preprint* arXiv:2310.00022, 2023.

[139] Li, W.; Yang, W.; Hou, Y.; Liu, L.; Liu, Y.; Li, X. S SARATR-X: Toward building a foundation model for SAR target recognition. *arXiv preprint* arXiv:2405.09365, 2024.

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

[141] Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A. J.; Hanna, J.; Borth, D.; Papoutsis, I.; Le Saux, B.; Camps-Valls, G.; Zhu, X. X. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint* arXiv:2403.15356, 2024.

[142] Wang, L.; Chen, S.; Jiang, L.; Pan, S.; Cai, R.; Yang, S.; Yang, F. Parameter-efficient fine-tuning in large models: A survey of methodologies. *arXiv preprint* arXiv:2410.19878, 2024.

[143] Zheng, H.; Shen, L.; Tang, A.; Luo, Y.; Hu, H.; Du, B.; Wen, Y.; Tao, D. Learning from models beyond fine-tuning. *arXiv preprint* arXiv:2310.08184, 2023.

[144] Han, Z.; Gao, C.; Liu, J.; Zhang, J.; Zhang, S. Q. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint* arXiv:2403.14608, 2024.

[145] Xin, Y.; Yang, J.; Luo, S.; Zhou, H.; Du, J.; Liu, X.; Fan, Y.; Li, Q.; Du, Y. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint* arXiv:2402.02242, 2024.

[146] Xu, L.; Xie, H.; Qin, S. J.; Tao, X.; Wang, F. L. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *arXiv preprint* arXiv:2312.12148, 2023.

[147] Chen, Z.; Liu, Z.; Wang, K.; Lian, S. Reparameterization-based parameter-efficient fine-tuning methods for large language models: A systematic survey. In: *Natural Language Processing and Chinese Computing. Lecture Notes in Computer Science, Vol. 15361*. Wong, D. F.; Wei, Z.; Yang, M. Eds. Springer Cham, 107–118, 2025.

[148] Xing, J.; Liu, J.; Wang, J.; Sun, L.; Chen, X.; Gu, X.; Wang, Y. A survey of efficient fine-tuning methods for Vision-Language Models—Prompt and adapter. *Computers & Graphics* Vol. 119, No. C, 2024.

[149] Balne, C. C. S.; Bhaduri, S.; Roy, T.; Jain, V.; Chadha, A. Parameter efficient fine tuning: A comprehensive analysis across applications. *arXiv preprint* arXiv:2404.13506, 2024.

[150] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv preprint* arXiv:1512.03385, 2015.

[151] Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:2010.11929, 2020.

[152] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint* arXiv:2103.14030, 2021.

[153] Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; de Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. Parameterefficient transfer learning for NLP. *arXiv preprint* arXiv:1902.00751, 2019.

[154] Yin, D.; Yang, Y.; Wang, Z.; Yu, H.; Wei, K.; Sun, X. 1% vs 100%: Parameter-efficient low rank adapter for dense predictions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 20116–20126, 2023.

[155] Shen, J.; Wang, W.; Chen, C.; Jiao, J.; Liu, J.; Zhang, Y.; Song, S.; Li, J. Med-tuning: A new parameter-efficient tuning framework for medical volumetric segmentation. *arXiv preprint* arXiv:2304.10880, 2023.

[156] Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint* arXiv:2110.04544, 2025.

[157] Tang, Y.; Han, K.; Guo, J.; Xu, C.; Xu, C.; Wang, Y. GhostNetv2: Enhance cheap operation with long-range attention. *arXiv preprint* arXiv:2211.12905, 2022.

[158] Lin, T. Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 936–944, 2017.

[159] Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W. Y.; et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015–4026, 2023.

[160] Zhang, C.; Han, D.; Qiao, Y.; Kim, J. U.; Bae, S. H.; Lee, S.; Hong, C. S. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv preprint* arXiv:2306.14289, 2023.

[161] Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; Luo, P. AdaptFormer: Adapting vision transformers for scalable visual recognition. *arXiv preprint* arXiv:2205.13535, 2022.

[162] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. *arXiv preprint* arXiv:2103.00020, 2021.

[163] Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; Qiao, Y. Vision transformer adapter for dense predictions. *arXiv preprint* arXiv:2205.08534, 2022.

[164] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9992–10002, 2021.

[165] Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast segment anything. *arXiv preprint* arXiv:2306.12156, 2023.

[166] Fratta, L. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint* arXiv:1505.04597, 2015.

[167] Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; Qiao, Y. Vision transformer adapter for dense predictions. *arXiv preprint* arXiv:2205.08534, 2022.

[168] Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* Vol. 55, No. 9, Article No. 195, 2023.

[169] Yin, D.; Hu, L.; Li, B.; Zhang, Y.; Yang, X. 5%>100%: Breaking performance shackles of full fine-tuning on visual recognition tasks. *arXiv preprint* arXiv:2408.08345, 2024.

[170] Wang, X.; Zhang, X.; Cao, Y.; Wang, W.; Shen, C.; Huang, T. SegGPT: Segmenting everything in context. *arXiv preprint* arXiv:2304.03284, 2023.

[171] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-rank adaptation of large language models. *arXiv preprint* arXiv:2106.09685, 2021.

[172] Kesim, E.; Helli, S. S. Multi LoRA Meets Vision: Merging multiple adapters to create a multi task model. *arXiv preprint* arXiv:2411.14064, 2024.

[173] Demetri, S.; Zúñiga, M.; Picco, G. P.; Kuipers, F.; Bruzzone, L.; Telkamp, T. Automated estimation of link quality for LoRa: A remote sensing approach. In: Proceedings of the 18th International Conference on Information Processing in Sensor Networks, 145–156, 2019.

[174] Shi, Z.; Kim, J.; Li, W.; Li, Y.; Pfister, H. MoRA: LoRA guided multi-modal disease diagnosis withMissing modality. In: *Medical Image Computing and Computer Assisted Intervention. Lecture Notes in Computer Science, Vol. 15003.* Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; Schnabel, J. A. Eds. Springer Cham, 273–282, 2024.

[175] Smith, J. S.; Hsu, Y. C.; Zhang, L.; Hua, T.; Kira, Z.; Shen, Y.; Jin, H. Continual diffusion: Continual customization of text-to-image diffusion with C-LoRA. *arXiv preprint* arXiv:2304.06027, 2023.

[176] Rottensteiner, F.; Sohn, G.; Jung, J.; Gerke, M.; Baillard, C.; Benitez, S.; Breitkopf, U. The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* Vol. I–3, 293–298, 2012.

[177] Yue, Z.; Zhou, P.; Hong, R.; Zhang, H.; Sun, Q. Few-shot learner parameterization by diffusion time-steps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 23263–23272, 2024.

[178] Ji, S.; Wei, S.; Lu, M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 57, No. 1, 574–586, 2019.

[179] Chen, H.; Shi, Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing* Vol. 12, No. 10, Article No. 1662, 2020.

[180] Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F. DINOv2: Learning robust visual features without supervision. *arXiv preprint* arXiv:2304.07193, 2024.

[181] Liu, D.; Zhang, R.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; Zhang, K. SPHINX-X: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint* arXiv:2402.05935, 2024.

[182] Keydel, E. R.; Lee, S. W.; Moore, J. T. MSTAR extended operating conditions: A tutorial. In: Proceedings of the SPIE 2757, Algorithms for Synthetic Aperture Radar Imagery III, 228–242, 1996.

[183] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631, 2019.

[184] Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; Yu, F. Frustratingly simple few-shot object detection. *arXiv preprint* arXiv:2003.06957, 2020.

[185] Wei, L.; Jiang, Z.; Huang, W.; Sun, L. InstructionGPT-4: A 200-instruction paradigm for fine-tuning MiniGPT-4. *arXiv preprint* arXiv: 2308.12067, 2023.

[186] Huang, B.; Li, Z.; Yang, C.; Sun, F.; Song, Y. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1795–1802, 2020.

[187] Lin, D.; Xu, G.; Wang, X.; Wang, Y.; Sun, X.; Fu, K. A remote sensing image dataset for cloud removal. *arXiv preprint* arXiv:1901.00600, 2019.

[188] Wu, Y.; Wang, Y.; Li, Y.; Xu, Q. Optical satellite image change detection via transformer-based Siamese network. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, 1436–1439, 2022.

[189] Shen, L.; Lu, Y.; Chen, H.; Wei, H.; Xie, D.; Yue, J.; Chen, R.; Lv, S.; Jiang, B. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing* Vol. 13, No. 24, Article No. 5094, 2021.

[190] Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE* Vol. 105, No. 10, 1865–1883, 2017.

[191] Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 270–279, 2010.

[192] Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 12, No. 7, 2217–2226, 2019.

[193] Xia, G. S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 55, No. 7, 3965–3981, 2017.

[194] Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 54, No. 12, 7405–7415, 2016.

[195] Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 159, 296–307, 2020.

[196] Xia, G. S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3974–3983, 2018.

[197] Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A high resolution

optical satellite image dataset for ship recognition and some new baselines. In: Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, 324–331, 2017.

[198] Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation robust object detection in aerial images using deep convolutional neural network. In: Proceedings of the IEEE International Conference on Image Processing, 3735–3739, 2015.

[199] Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint* arXiv:2110.08733, 2021.

[200] Zamir, S. W.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Khan, F. S.; Zhu, F.; Shao, L.; Xia, G. S.; Bai, X. iSAID: A large-scale dataset for instance segmentation in aerial images. *arXiv preprint* arXiv:1905.12886, 2019.

[201] Hou, X.; Ao, W.; Song, Q.; Lai, J.; Wang, H.; Xu, F. FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition. *Science China Information Sciences* Vol. 63, No. 4, Article No. 140303, 2020.

[202] Lei, S.; Lu, D.; Qiu, X.; Ding, C. SRSDD-v1.0: A high-resolution SAR rotation ship detection dataset. *Remote Sensing* Vol. 13, No. 24, Article No. 5104, 2021.

[203] Shi, X.; Fu, S.; Chen, J.; Wang, F.; Xu, F. Object-level semantic segmentation on the high-resolution Gaofen-3 FUSAR-map dataset. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 14, 3107–3119, 2021.

[204] Zheng, N. R.; Yang, Z. A.; Shi, X. Z.; Zhou, R. Y.; Wang, F. Land cover classification of synthetic aperture radar images based on encoder: Decoder network with an attention mechanism. *Journal of Applied Remote Sensing* Vol. 16, No. 1, Article No. 014520, 2022.

[205] Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sensing* Vol. 13, No. 18, Article No. 3690, 2021.

[206] Wang, Z.; Zeng, X.; Yan, Z.; Kang, J.; Sun, X. AIR-PolSAR-seg: A large-scale data set for terrain segmentation in complex-scene PolSAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 15, 3830–3841, 2022.

[207] Zeng, X.; Wang, Z.; Sun, X.; Chang, Z.; Gao, X. DENet: Double-encoder network with feature refinement and region adaption for terrain segmentation in PolSAR images. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 60, Article No. 5217419, 2021.

[208] Rahnemoonfar, M.; Chowdhury, T.; Sarkar, A.; Varshney, D.; Yari, M.; Murphy, R. R. FloodNet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access* Vol. 9, 89644–89654, 2021.

[209] Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. MMBench: Is your multi-modal model an all-around player? *arXiv preprint* arXiv:2307.06281, 2023.

[210] Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *National Science Review* Vol. 11, No. 12, Article No. nwae403, 2024.

[211] Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; Shan, Y.; Li, B.; Ge, Y.; Ge, Y.; Wang, G.; et al. SEED-bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint* arXiv:2307.16125, 2023.

[212] Lu, X.; Wang, B.; Zheng, X.; Li, X. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 56, No. 4, 2183–2195, 2018.

[213] Yuan, Z.; Zhang, W.; Fu, K.; Li, X.; Deng, C.; Wang, H.; Sun, X. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 60, Article No. 4404119, 2021.

[214] Deng, L. J.; Vivone, G.; Paoletti, M. E.; Scarpa, G.; He, J.; Zhang, Y.; Chanussot, J.; Plaza, A. Machine learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine* Vol. 10, No. 3, 279–315, 2022.

[215] Wang, Q.; Gu, Y. A discriminative tensor representation model for feature extraction and classification of multispectral LiDAR data. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 58, No. 3, 1568–1586, 2020.

[216] Dstl Satellite Imagery Feature Detection. 2016. Available at https://www.kaggle.com/competitions/dstl-satellite-imagery-feature-detection

[217] Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 145, 60–77, 2018.

[218] Debes, C.; Merentitis, A.; Heremans, R.; Hahn, J.;

Frangiadakis, N.; van Kasteren, T.; Liao, W.; Bellens, R.; Pižurica, A.; Gautama, S.; et al. Hyperspectral and LiDAR data fusion: Outcome of the 2013 GRSS data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 7, No. 6, 2405–2418, 2014.

[219] Gader, P.; Zare, A.; Close, R.; Aitken, J.; Tuell, G. MUUFL Gulfport hyperspectral and LiDAR airborne data set. 2016. Available at https://github.com/GatorSense/MUUFLGulfport

[220] Hong, D.; Gao, L.; Hang, R.; Zhang, B.; Chanussot, J. Deep encoder–decoder networks for classification of hyperspectral and LiDAR data. *IEEE Geoscience and Remote Sensing Letters* Vol. 19, Article No. 5500205, 2020.

[221] Wang, G.; Li, H.; Ye, S.; Zhao, H.; Ding, H.; Xie, S. RFWNet: A multiscale remote sensing forest wildfire detection network with digital twinning, adaptive spatial aggregation, and dynamic sparse features. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 62, Article No. 4708523, 2024.

[222] Liu, R.; Kuffer, M.; Persello, C. The temporal dynamics of slums employing a CNN-based change detection approach. *Remote Sensing* Vol. 11, No. 23, Article No. 2844, 2019.

[223] Chen, B.; Wu, S.; Jin, Y.; Song, Y.; Wu, C.; Venevsky, S.; Xu, B.; Webster, C.; Gong, P. Wildfire risk for global wildland–urban interface areas. *Nature Sustainability* Vol. 7, No. 4, 474–484, 2024.

[224] Bousfield, C. G.; Lindenmayer, D. B.; Edwards, D. P. Substantial and increasing global losses of timber-producing forest due to wildfires. *Nature Geoscience* Vol. 16, No. 12, 1145–1150, 2023.

[225] Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews* Vol. 232, Article No. 104110, 2022.

[226] Rasti, B.; Chang, Y.; Dalsasso, E.; Denis, L.; Ghamisi, P. Image restoration for remote sensing: Overview and toolbox. *IEEE Geoscience and Remote Sensing Magazine* Vol. 10, No. 2, 201–230, 2022.

[227] Wang, Z.; Zhao, L.; Meng, J.; Han, Y.; Li, X.; Jiang, R.; Chen, J.; Li, H. Deep learning-based cloud detection for optical remote sensing images: A survey. *Remote Sensing* Vol. 16, No. 23, Article No. 4583, 2024.

[228] Paul, S.; Pati, U. C. A comprehensive review on remote sensing image registration. *International Journal of Remote Sensing* Vol. 42, No. 14, 5396–5432, 2021.

[229] Zhang, Z.; Wang, C.; Song, J.; Xu, Y. Object tracking based on satellite videos: A literature review. *Remote Sensing* Vol. 14, No. 15, Article No. 3674, 2022.

[230] Cheng, H.; Liu, M.; Chen, L.; Broszio, H.; Sester, M.; Yang, M. Y. GATraj: A graph- and attention-based multi-agent trajectory prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 205, 163–175, 2023.

[231] Zhang, Z.; Zhang, Q.; Gao, Z.; Zhang, R.; Shutova, E.; Zhou, S.; Zhang, S. Gradient-based parameter selection for efficient fine-tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 28566–28577, 2024.

[232] He, H.; Cai, J.; Zhang, J.; Tao, D.; Zhuang, B. Sensitivity-aware visual parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 11825–11835, 2023.

[233] Zhang, M.; Chen, H.; Shen, C.; Yang, Z.; Ou, L.; Yu, X.; Zhuang, B. LoRAPrune: Structured pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint* arXiv:2305.18403, 2023.

[234] Kim, J.; Lee, J. H.; Kim, S.; Park, J.; Yoo, K. M.; Kwon, S.; Lee, D. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *arXiv preprint* arXiv:2305.14152, 2023.

[235] Liu, J.; Xiao, G.; Li, K.; Lee, J. D.; Han, S.; Dao, T.; Cai, T. BitDelta: Your fine-tune may only be worth one bit. *arXiv preprint* arXiv:2402.10193, 2024.

[236] Duan, Y.; Li, L.; Zhai, Z.; Yao, J. In-context learning distillation for efficient few-shot fine-tuning. *arXiv preprint* arXiv:2412.13243, 2024.

[237] Aghajanyan, A.; Yu, L.; Conneau, A.; Hsu, W.-N.; Hambardzumyan, K.; Zhang, S.; Roller, S.; Goyal, N.; Levy, O.; Zettlemoyer, L. Scaling laws for generative mixed-modal language models. *arXiv preprint* arXiv:2301.03728, 2023.

[238] Gao, L.; Schulman, J.; Hilton, J. Scaling laws for reward model overoptimization. *arXiv preprint* arXiv:2210.10760, 2022.

Dongshuo Yin is currently a post-doctoral researcher in computer science at Tsinghua University. He received his Ph.D. degree from the Aerospace Information Research Institute, the University of the Chinese Academy of Sciences, in 2024. His research interests include computer vision, remote sensing image interpretation, parameter-efficient fine-tuning, and video editing.

Ting-Feng Zhao is currently an undergraduate senior in the College of Computer Science, Nankai University, under the supervision of Prof. Deng-Ping Fan. His research interests include deep learning, image processing, and computer vision.

Deng-Ping Fan joined the Department of Nankai International Advanced Research Institute (SHENZHEN-FUTIAN) as a faculty member in 2024. He was a full professor and deputy director of the Media Computing Lab (MC Lab) at the College of Computer Science, Nankai University, China. Before that, he was postdoctoral researcher, working with Prof. Luc Van Gool in the Computer Vision Lab at ETHZurich. He was one of the core technical members of the TRACE-Zurich project on automated driving.

Shutao Li received his B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively. In 2001, he joined the College of Electrical and Information Engineering, Hunan University, where he is currently a full professor. He has authored or co-authored over 200 refereed articles. His current research interests include image processing, pattern recognition, and artificial intelligence. He is a member of the editorial board of *Information Fusion and Sensing and Imaging*. He is an associate editor of *IEEE TGRS* and *IEEE TIM*.

Bo Du is currently a professor in the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University. He is also the director of the National Engineering Research Center for Multimedia Software, Wuhan University. His major research interests include machine learning, computer vision, and image processing. He regularly serves as a senior PC member of IJCAI and AAAI. He has served as an area chair for ICPR.

Xian Sun received his B.Sc. degree from the Beijing University of Aeronautics and Astronautics, China, in 2004, and M.Sc. and Ph.D. degrees from the Institute of Electronics, the Chinese Academy of Sciences, Beijing, in 2009, all in electronic information engineering. He is currently a professor in the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Shi-Min Hu received his Ph.D. degree from Zhejiang University in 1996. He is currently a professor in the Department of Computer Science and Technology, Tsinghua University. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer-aided geometric design. He is the Editor-in-Chief of *Computational Visual Media*, and is on the editorial boards of several other journals, including *Computer Aided Design* and *Computer & Graphics*. He is a senior member of ACM, and a Fellow of IEEE, CCF, and SMA.

To submit a manuscript, please go to `https://jcvm.org`.