Large Vision-Language Models and Their Adaptation in Remote-Sensing Domains

SUN Qianru
School of Computing and Information Systems
Singapore Management University
qianrusun@smu.edu.sg

Date: 19 Dec 2025

1. Background and Motivation

Large-scale pre-trained visual foundation models, such as CLIP [13], Stable Diffusion [15], and DINO [2], have demonstrated remarkable capabilities in natural image domains. They achieve impressive performance in downstream tasks including object classification, detection, and semantic segmentation [27]. Their success stems from learning robust and generalizable representation spaces from billions of web images [16]. However, in data-scarce domains such as remote sensing (RS), training such foundation models from scratch remains impractical for three reasons. First, RS data exhibits high spectral diversity. Specifically, different spectral bands from optical remote sensing (ORS, 400-700nm) to synthetic aperture radar (SAR, 1mm-1m) capture fundamentally different physical properties of Earth's surface [4, 17]. Second, collecting and annotating RS images is difficult due to military restrictions, sensor availability, and high acquisition costs [5, 30]. Third, even existing attempts to train RS foundation models [3, 5] are constrained to small- or medium-scale models (*e.g.*, ViT-L with 300M parameters) and require prohibitive computational resources (*e.g.*, 80 × A100 GPUs with 80GB VRAM each).

Given these constraints, adapting pre-trained models from natural image domains to RS domains has emerged as a more practical approach. Parameter-Efficient Fine-Tuning (PEFT), particularly Low-Rank Adaptation (LoRA) [7], has become the *de-facto* standard for such adaptation. LoRA introduces a low-rank factorization of parameter changes (*i.e.*, $\Delta\theta := B \cdot A$ where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ with $r \ll \min(d, k)$). This formulation enables efficient adaptation with minimal trainable parameters, zero inference latency overhead, and modular deployment. PEFT is particularly appealing for RS domains because 1) it reduces computational and data requirements, 2) it preserves general knowledge learned from natural images [24], and 3) it mitigates overfitting in data-limited scenarios [18, 21]. Nevertheless, RS domains present unique challenges that standard PEFT methods fail to address. RS datasets inherently suffer from severe data imbalance. For example, the imbalance ratios of DOTA and ShipRSImageNet reach 86 and 112 respectively, significantly higher than natural image benchmarks such as CIFAR100-LT with a ratio of 50 [1, 23, 29]. Beyond single-task adaptation, real-world RS applications often require handling multiple tasks (*e.g.*, multi-spectral adaptation, multi-class recognition) simultaneously. This multi-task requirement introduces additional complexity because different tasks may conflict in their parameter updates.

Our research addresses these fundamental challenges of applying PEFT to data-scarce and imbalanced domains, with RS as a representative case study. We identify three levels of challenges when directly applying LoRA-based adaptation. **1) At the representation level**, the learned feature space exhibits strong bias towards head (majority) classes while neglecting tail (minority) classes [25]. Additionally, large domain gaps (*e.g.*, between natural images and SAR) cannot be bridged by simple low-rank updates. **2) At the optimization level**, PEFT methods prove highly sensitive to hyperparameters. Specifically, performance can vary by up to 86% across different configurations of scaling factors, layer positions, and insertion depths. Manual tuning is intractable because hyperparameter interactions exhibit non-monotonic effects. **3) At the scalability level**, extending multi-task LoRA to a large number of tasks (15-50+) induces catastrophic performance degradation. This failure occurs because of parameter and representation misalignment across tasks.

In the following sections, we present our recent works in addressing these challenges, progressively: 1) representation calibration and cross-modal data synthesis (Section 2.1), 2) automated hyperparameter optimization via meta-learning (Section 2.2), and 3) scalable multi-task adaptation (Section 2.3).

2. Key Observations and Methods

We organize our contributions by the three challenge levels (identified in Sec 1). **At the representation level** (Section 2.1), we solve two key problems: 1) feature bias in long-tailed adaptation, and 2) data scarcity in non-visible modalities. We solve the first through debLoRA [21] and the second through 2LoRA/pLoRA [20]. Both methods leverage unsupervised clustering to discover shared visual attributes; this common insight enables calibration without additional labels. **At the optimization level** (Section 2.2), we address PEFT methods' hyperparameter sensitivity issue through MetaPEFT [22], a meta-learning framework that automatically discovers optimal PEFT configurations. **At the scalability level** (Section 2.3), we address catastrophic failure in large-scale multi-task adaptation through mtLoRA.

2.1. Representation Calibration and Synthesis

LoRA adaptation on imbalanced or cross-domain data produces biased or inadequate features. Specifically, fine-tuning on imbalanced datasets causes LoRA features bias towards head classes. Besides, LoRA fails to capture necessary feature transformations between large domain gaps (*e.g.*, from natural images to SAR). We address these two problems through two complementary approaches: debLoRA [21] calibrates biased features, and 2LoRA/pLoRA [20] synthesizes cross-modal data when training data is scarce, particularly for minority classes.

Learning De-Biased Representations for Remote-Sensing Imagery (debLoRA) [21]. We observe that LoRA features exhibit strong bias towards head classes in long-tailed RS datasets. Specifically, when adapting pre-trained models using LoRA, the learned feature space is dominated by discriminative features of head classes while tail class features are under-represented. We observe two manifestations of this bias: 1) head class representations over-expand their territory into the tail class space, *i.e.*, tail class validation samples are wrongly distributed in the head class's feature region, and 2) the center of the entire feature space shifts towards head classes due to the overwhelming number of head class training samples. This bias is particularly severe for PEFT methods because using fewer parameters restricts model capacity and forces the model to prioritize features from classes with more samples.

To address this without requiring additional data or labels, we propose debLoRA, an unsupervised learning approach that de-biases LoRA features through three steps: clustering, calibration, and training. First, we apply K-means clustering on all LoRA features regardless of class labels. Each obtained cluster center represents a visual attribute (*e.g.*, "streamlined tail" or "with wooden deck") shared by both head and tail classes. These cluster centers are more robust than original tail class representations because they leverage the diversity of head class samples while exhibiting reduced class imbalance. Second, for each tail class, we compute a de-biased center by weighted averaging all cluster centers, where weights are proportional to the fraction of that tail class in each cluster. This ensures the de-biased center is not dominated by head classes. Third, we calibrate tail class features by moving them closer to their de-biased centers and train a debLoRA module to learn this mapping. Extensive experiments on two RS adaptation scenarios (natural→ORS and ORS→SAR) demonstrate that debLoRA consistently outperforms prior methods, achieving up to +3.3% and +4.7% gains on tail classes for the respective scenarios while preserving head class performance.

Non-Visible Light Data Synthesis via Two-Stage Low-Rank Adaptation (2LoRA/pLoRA) [20]. To tackle the data scarcity issue for RS domain, especially for minor classes, in this work, we were the first to propose synthesizing data for non-visible light modalities (*e.g.*, SAR). In preliminary study, we observed direct fine-tuning or LoRA on Stable Diffusion (SD) with SAR images fails. We identify such fail stems from two key differences between regular images and SAR: *structure* (regular view vs. aerial view) and *modality* (RGB vs. radar). For example, when LoRA application strength increases from 0 to 1, the transition from regular-view images to aerial-view SAR images exhibits severe distortions.

To address this, we propose 2LoRA, a two-stage adaptation approach that decomposes the domain gap. In the first stage (view adaptation), we train an ORS LoRA module on large-scale optical remote sensing datasets to adapt SD from regular view to aerial view without changing the data modality. In the second stage (modality adaptation), we further train a SAR LoRA module on top of the ORS-adapted model to handle the RGB-to-SAR modality shift. Moreover, we found that 2LoRA still struggles with class imbalance in SAR datasets, *i.e.*, classes with more samples achieve higher generation quality. We hence propose pLoRA (prototype LoRA), which clusters SAR training samples based on visual attributes (*e.g.*, "in the harbor" or "moving") regardless of class labels, trains a separate LoRA module for each cluster, and combines them with weights designed to avoid bias towards major classes. Specifically, when generating images for a target minor class, we assign higher weights to clusters containing larger fractions of that class. Integrating the synthesized SAR data into downstream classification and segmentation tasks yields notable improvements, especially for minor classes (+4.0% tail-class accuracy), demonstrating the effectiveness of addressing data scarcity through cross-modal synthesis.

2.2. Automated Hyperparameter Optimization

During our research, we identifeid another key challenge in PEFT methods: PEFT methods' performance is highly sensitive to hyperparameter configurations, and require heavy manual hyperparameter tuning. We tackle this challenge by proposing an automatic meta-learning framework, MetaPEFT.

Meta-Learning Hyperparameters for Parameter Efficient Fine-Tuning (MetaPEFT) [22]. In our comprehensive preliminary study of PEFT methods, we identified a critical but overlooked challenge: PEFT hyperparameters exhibit complex, non-monotonic interactions that severely impact performance. Specifically, we study three key hyperparameters for additive PEFT methods: 1) intra-block position (*i.e.*, Q/K/V projection layers, attention output, or FFN), 2) block depth (*i.e.*, which transformer blocks to adapt), and 3) scaling factor (*i.e.*, the magnitude of PEFT updates). Our experiments reveal that accuracy varies by up to 86% across scaling factors (from 8.1% to 91.1%), 4.0% across block depths, and 2.4% across intra-block positions. More critically, combining individually optimal hyperparameters often leads to unexpected performance degradation, *e.g.*, using the optimal intra-block position (FFN) with the optimal block depth (depth 11) results in a 0.6% accuracy drop. This poses a Mixed Integer Non-Linear Programming (MINLP) problem with complexity $\mathcal{O}(L|\mathcal{S}|N_\alpha)$, where L is the number of layers, $|\mathcal{S}|$ is the number of insertion positions, and N_α is the number of scaling factor candidates, which makes exhaustive search computationally infeasible.

To address this, we propose MetaPEFT with two key designs. First, we introduce a *unified modulator* γ that combines the discrete positional indicator and continuous scaling factor into a single differentiable scalar applied to each PEFT position: when $\gamma \approx 0$, the PEFT module is effectively deactivated; when $\gamma > 0$, it controls both activation and update magnitude. This transforms the MINLP problem into continuous optimization amenable to gradient descent. Second, we optimize this modulator through a *bi-level optimization framework*: the inner loop trains PEFT parameters with the modulator fixed, while the outer loop updates the modulator on a dynamically sampled subset of training data. This dynamic sampling serves as implicit regularization, exposing optimization to diverse data subsets and particularly benefiting tail classes where overfitting is severe. Experiments across three transfer-learning scenarios and five datasets demonstrate that MetaPEFT achieves state-of-the-art performance, with up to 3.5× higher accuracy improvement for tail classes compared to head classes, all without manual hyperparameter tuning.

2.3. Scalable Multi-Task Adaptation

The methods above handle single-task or few-task scenarios. However, real-world deployments, especially for RS domain, usually require adapting to 15-50+ tasks simultaneously. At this scale, existing multi-task LoRA methods fail catastrophically. In this section, we presents mtLoRA, a principled solution for large-scale multi-task low-rank adaptation.

Scalable Multi-Task Low-Rank Adaptation (mtLoRA). Scaling multi-task LoRA to large numbers of tasks induces catastrophic failure, *e.g.*, accuracy drops from 88.2% to 2.0% on DOTA when scaling from 5 to 15 tasks. The core challenges are two kinds of misalignment: *parameter misalignment* (conflicting weight updates across LoRA modules) and *representation misalignment* (divergent output features). Existing solutions use orthogonal regularization to address parameter conflicts and dynamic routing to handle representation divergence, but we find they face a fundamental trade-off: strengthening regularization to reduce inter-task conflict inadvertently suppresses the essential feature discrimination required for effective routing. Specifically, when increasing regularization strength λ on a multi-task NLP scenario, accuracy initially improves (+1.7% at $\lambda = 0.25$) but then drops by 1.8% as routing uncertainty increases. We identify two root causes for this trade-off. **First, uniform regularization disrupts knowledge sharing across tasks.** We discover that shared underlying knowledge concentrates in high singular-value (SV) components: the top-20% SV components contain 89% of inter-task alignment while encoding 54% of total singular values. Uniform regularization treats all spectral components equally, forcing orthogonality on high-SV components and thereby disrupting essential knowledge sharing. **Second, applying LoRA at the component-level amplifies gradient conflicts.** When multiple LoRA modules adapt individual weight matrices (*e.g.*, W_q, W_v), their gradients exhibit stronger misalignment (average cosine similarity of -0.054); block-level adaptation reduces this conflict by 76%.

Based on these insights, we propose mtLoRA with three key innovative designs: 1) *spectral-aware regularization* that selectively orthogonalizes low-SV noise components while preserving high-SV shared knowledge, using a weighting function $w(\sigma) = \exp(-\sigma/\bar{\sigma})$; 2) *fine-grained routing* that assigns dimension-specific weights (a vector $\Pi_i \in \mathbb{R}^g$ per LoRA) instead of scalar weights, allowing different feature subspaces to use different expert combinations; and 3) *block-level adaptation* that applies LoRA as a parallel path to entire attention/FFN blocks rather than individual matrices, mitigating gradient conflict amplification. On four large-scale benchmarks (DOTA, iNat2018, Dolly-15k, BBH with 15-27 tasks), mtLoRA achieves state-of-the-art performance (+2.8% average accuracy over prior methods) while using 47% fewer parameters and 24% less training time, demonstrating that scalable multi-task LoRA is practical for real-world deployments.

3. Future Research Directions

Our four works establish a systematic framework for adapting foundation models to data-scarce RS domains, addressing challenges at the representation, optimization, and scalability levels. Building on these foundations, we identify two promising directions for future investigation: 1) developing a unified RS agent system that integrates our PEFT methods with interactive reasoning capabilities, and 2) extending our techniques to spectrum-specific foundation models that capture the unique physical properties of different RS modalities.

3.1. Towards Unified RS Agent System

Our current research has addressed representation bias (debLoRA), cross-modal synthesis (pLoRA), hyperparameter optimization (MetaPEFT), and multi-task scaling (mtLoRA) as separate methods. However, in real-world use cases, RS researchers need to manually select and chain these methods for complex RS analysis tasks. A unified agent system would automate this orchestration and enable interactive, multi-step reasoning.

We envision a **Mixture-of-RS-Experts (MoRE)** architecture where our PEFT methods serve as specialized experts with dynamic routing. Each expert would specialize in distinct adaptation challenges: debLoRA for imbalanced scenarios, pLoRA for data-scarce modalities, and mtLoRA for multi-task coordination. A lightweight gating network would route inputs to relevant experts based on task requirements. This architecture enables efficient processing through sparse activation while maintaining task-specific specialization.

Beyond expert organization, we aim to develop an **interactive RS agent** that interprets high-level user queries and orchestrates multi-step analysis workflows. Drawing from recent advances in Vision-Language Models [11], the agent would adopt a Reason-Act (ReAct) framework [26] for dynamic, iterative reasoning. For example, given the query "identify new ships in this port since last week," the agent would decompose this into sub-tasks: first invoke change detection experts, then route ship detection to relevant regions, and finally aggregate results. The agent would also incorporate self-reflection mechanisms to evaluate its findings and quantify prediction confidence, addressing the trustworthiness concerns in high-stakes RS applications. Direct Preference Optimization (DPO) [14] could align the agent's behavior with human expert preferences.

3.2. Spectrum-Specific Foundation Models

Our current PEFT methods adapt general-purpose foundation models (*e.g.*, CLIP, Stable Diffusion) pre-trained on natural RGB images. However, RS data spans diverse electromagnetic spectrums with fundamentally different physical properties: optical (400–700nm), near-infrared (700–1400nm), thermal infrared (8–14μm), and microwave/SAR (1mm–1m) [12, 17]. Each spectrum captures distinct surface properties. Optical bands measure reflected solar radiation; thermal bands detect emitted heat; SAR measures backscattered radar signals sensitive to surface roughness, dielectric properties, and geometric structure. General-purpose foundation models trained on RGB images cannot capture these spectrum-specific characteristics.

Existing geospatial foundation models [6, 8, 9, 19] have begun addressing this gap. However, they primarily process multi-spectral optical data or amplitude-only SAR data. For SAR, this means discarding phase information from Single Look Complex (SLC) raw data and retaining only Ground Range Detected (GRD) amplitude. This limitation prevents models from learning physical scattering mechanisms essential for interferometric applications (*e.g.*, surface deformation monitoring) and polarimetric analysis (*e.g.*, land cover classification based on scattering decomposition).

A promising direction is to develop **spectrum-aware foundation models** tailored to specific RS modalities. For SAR, this means training on complex-valued SLC data that preserves both amplitude and phase. The model architecture would process real and imaginary components jointly, learning representations invariant to SAR-specific noise (*e.g.*, speckle) while capturing phase coherence patterns. For hyperspectral data, models would learn spectral signatures across hundreds of contiguous bands, capturing diagnostic absorption features for material identification. Self-supervised objectives could be adapted to each spectrum: masked autoencoding for spatial-spectral reconstruction, contrastive learning across temporal acquisitions, or physics-informed losses encoding known scattering models.

Our PEFT techniques would then serve as the adaptation layer between these spectrum-specific foundation models and downstream tasks. The spectral-aware regularization in mtLoRA is particularly relevant: it could preserve physics-informed representations in high singular-value components while allowing task-specific adaptation in lower components. The clustering-based attribute discovery in debLoRA could identify shared physical scattering properties across different SAR applications. This direction extends beyond RS to other specialized domains with spectrum-specific data: medical imaging [10, 28] (CT, MRI, ultrasound with distinct imaging physics), industrial inspection (X-ray, thermal imaging), and scientific sensing (radio astronomy, seismic data).

References

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[3] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1

[4] Africa Ixmuca Flores-Anderson, Kelsey E Herndon, Rajesh Bahadur Thapa, and Emil Cherrington. The sar handbook: comprehensive methodologies for forest monitoring and biomass estimation. Technical report, NASA SERVIR Global Program, 2019. 1

[5] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. *arXiv preprint arXiv:2312.10115*, 2023. 1

[6] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *arXiv preprint arXiv:2311.07113*, 2023. 4

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[8] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, Daiki Kimura, Naomi Simumba, Linsong Chu, S. Karthik Mukkavilli, Devyani Lambhate, Kamal Das, Ranjini Bangalore, Dario Oliveira, Michal Muszynski, Kumar Ankur, Muthukumaran Ramasubramanian, Iksha Gurung, Sam Khallaghi, Hanxi (Steve) Li, Michael Cecil, Maryam Ahmadi, Fatemeh Kordi, Hamed Alemohammad, Manil Maskey, Raghu Ganti, Kommy Weldemariam, and Rahul Ramachandran. Foundation Models for Generalist Geospatial Artificial Intelligence. *Preprint Available on arxiv:2310.18660*, 2023. 4

[9] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[10] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2022. 4

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 4

[12] Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013. 4

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[14] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 4

[15] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2021. 1

[16] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, abs/2210.08402, 2022. 1

[17] Gary A Shaw and Hsiaohua K Burke. Spectral imaging for remote sensing. *Lincoln laboratory journal*, 14(1):3–28, 2003. 1, 4

[18] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. *arXiv preprint arXiv:2309.10019*, 2023. 1

[19] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arévolo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabe Moreno. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. *arXiv preprint arXiv:2412.02732*, 2024. 4

[20] Zichen Tian, Zhaozheng Chen, and Qianru Sun. Non-visible light data synthesis and application: A case study for synthetic aperture radar imagery. *arXiv preprint arXiv:2311.17486*, 2023. 2

[21] Zichen Tian, Zhaozheng Chen, and Qianru Sun. Learning de-biased representations for remote-sensing imagery. In *Advances in Neural Information Processing Systems*, pages 57970–57992. Curran Associates, Inc., 2024. 1, 2

[22] Zichen Tian, Yaoyao Liu, and Qianru Sun. Meta-learning hyperparameters for parameter efficient fine-tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23037–23047, 2025. 2, 3

[23] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1

[24] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024. 1

[25] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020. 1

[26] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022. 4

[27] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[28] Shaoting Zhang and Dimitris N. Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *ArXiv*, abs/2306.05705, 2023. 4

[29] Zhengning Zhang, Lin Zhang, Yue Wang, Pengming Feng, and Ran He. Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:8458–8472, 2021. 1

[30] Xiao Xiang Zhu, Sina Montazeri, Mohsin Ali, Yuansheng Hua, Yuanyuan Wang, Lichao Mou, Yilei Shi, Feng Xu, and Richard Bamler. Deep learning meets sar: Concepts, models, pitfalls, and perspectives. *IEEE Geoscience and Remote Sensing Magazine*, 9 (4):143–172, 2021. 1