

Tree Mapping with Limited Data: Fine-Tuning Foundation Models for Multimodal Fusion

Xiaoyan Lu

Hong Kong Polytechnic University, Hung Hom, HongKong
xiaoyan07.lu@polyu.edu.hk

Qihao Weng

Hong Kong Polytechnic University, Hung Hom, HongKong
qihao.weng@polyu.edu.hk

Abstract

Tree mapping plays a crucial role in remote sensing for ecological monitoring and resource management. Achieving accurate tree mapping relies on labeled training data, however, annotating geometrically complex trees in large-scale remote sensing imagery is time-consuming and labor-intensive. We propose a multimodal fusion framework that leverages fine-tuned foundation models to enable data-efficient tree mapping. To enhance spatial understanding and structural perception, we introduce depth information as an auxiliary modality alongside high-resolution RGB remote sensing imagery. By leveraging the complementary strengths of depth and visual data, our method mitigates the limitations of unimodal inputs. Experimental results demonstrate that integrating depth information significantly improves recognition accuracy and boundary delineation, particularly when training samples are scarce. This work highlights the potential of depth-aware multimodal learning to boost performance in data-constrained scenarios, offering a promising direction for scalable and cost-efficient environmental monitoring.

1. Introduction

Tree mapping is a fundamental task in remote sensing image analysis, supporting applications such as environmental change detection, biomass estimation, and green resource management [1]. However, achieving high-precision tree mapping across large spatial extents remains challenging. One of the primary obstacles lies in the high cost and effort required for collecting large-scale annotated training samples. Manual labeling of remote sensing imagery is both time-consuming and labor-intensive, particularly in complex and heterogeneous natural environments. These limitations restrict the scalability of traditional supervised learning approaches, which often underperform in scenarios with limited or regionally biased training data.

The advent of visual foundation models introduces a promising paradigm for addressing these challenges in remote sensing tasks. Such models, pretrained on large-

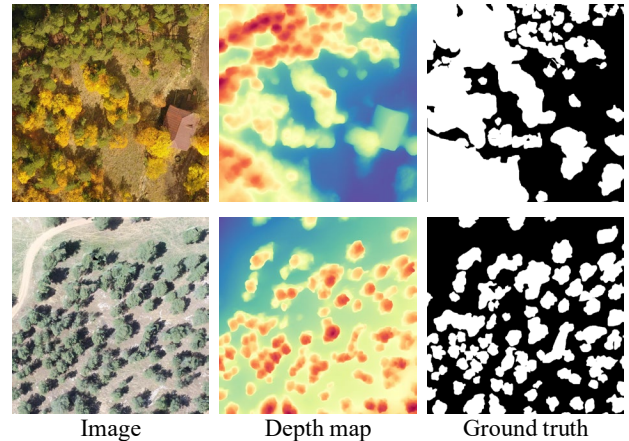


Figure 1: Trees exhibit distinct height information, which serves as a key differentiator compared to other ground objects.

scale datasets, exhibit strong generalization capabilities across domains and tasks. This work leverages two state-of-the-art visual foundation models: Segment Anything Model (SAM) [2], a promptable segmentation model, and Depth-Anything-V2 [3], a depth estimation model capable of inferring fine-grained 3D structural information from 2D imagery. Given the distinctive vertical structure of trees relative to surrounding terrain, depth emerges as a crucial cue for accurate segmentation, as shown in Figure 1. By integrating SAM’s visual representation capabilities with the 3D spatial cues extracted by Depth-Anything-V2, a depth-aware multimodal framework is developed to enhance tree mapping performance, particularly in data-scarce conditions.

The main contributions of this work are summarized as follows:

1. A multimodal visual-depth fusion framework is proposed for tree mapping, which leverages SAM and Depth-Anything-V2 foundation models to integrate visual features from RGB remote sensing imagery with structural cues derived from depth features.
2. A depth-conditioned network architecture is designed to explicitly integrate and amplifies depth features, enhancing structural perception and enabling more precise tree mapping.
3. Our framework exhibits strong performance under limited training data scenarios, highlighting the

effectiveness of foundation model adaptation and multimodal learning in mitigating data scarcity in large-scale remote sensing tasks.

2. Related work

2.1. Foundation model

Recent advances in artificial intelligence have been marked by the development of powerful foundation models with strong generalization capabilities. SAM [2] is a general-purpose segmentation model trained on over one billion masks from diverse domains. It supports zero-shot generalization and produces high-resolution segmentations with precise boundary delineation. Concurrently, Depth-Anything-V2 [3] addresses monocular depth estimation through a vision-language pretraining framework and a heterogeneous set of depth datasets. This approach enables the generation of structurally consistent, high-resolution depth maps across domains without requiring task-specific fine-tuning. When combined, SAM and Depth-Anything-V2 allow for the effective fusion of semantic and geometric cues, establishing a strong foundation for a broad range of visual perception tasks.

In the domain of remote sensing, several foundation models have been introduced, primarily built upon the Masked Autoencoder (MAE) [4] architecture and trained in a self-supervised manner. This training paradigm reconstructs missing portions of the input from partial observations, enabling the extraction of semantically meaningful representations. For instance, SatMAE [5] adopts a masked autoencoding strategy tailored to multi-spectral and temporal satellite imagery, facilitating robust and transferable feature learning across diverse sensors and acquisition conditions. Extending this framework, ScaleMAE [6] introduces scale-awareness by employing multi-resolution training, thereby supporting representation learning that is invariant to spatial scale variations. Cross-Scale MAE [7] further advances this approach by integrating cross-resolution attention mechanisms to align hierarchical features, thereby improving the model's ability to capture multiscale spatial patterns. Additionally, SatlasPretrain [8] offers a large-scale and diverse remote sensing dataset alongside a dedicated pretraining framework, yielding a generalized visual encoder suitable for a wide range of downstream geospatial tasks.

2.2. Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) strategies have become a practical solution for adapting large pretrained models to downstream tasks, especially under limited computational budgets [9]. Originating in the field of natural language processing (NLP), PEFT methods have seen widespread adoption in large language models (LLMs)

due to their balance of efficiency and performance. Unlike traditional fine-tuning approaches that update all network parameters, PEFT techniques focus on modifying only a small, task-relevant subset while keeping most pretrained weights unchanged, thereby reducing both memory consumption and training cost. These techniques can be broadly classified into three categories [10]: addition-based (e.g., inserting lightweight modules such as adapters [11] within the transformer blocks), selection-based (e.g., tuning only selected layers or parameters), and reparameterization-based methods such as LoRA [12], which introduces trainable low-rank decompositions into attention components. In addition, prompt tuning—another widely adopted approach—learns compact task-specific embeddings that are prepended to the model's input. These strategies have achieved strong results in both vision and language domains, particularly under few-shot or resource-limited conditions, and are increasingly being explored for remote sensing tasks where annotated datasets are scarce and expensive to collect.

3. Methodology

Figure 2 presents the overall multimodal visual-depth fusion framework. The encoder consists of the pre-trained SAM ViT-B architecture with the integration of adapter layers, while the decoder utilizes the LinkNet decoder architecture. For the input RGB aerial image, the SAM ViT-B model is employed to extract visual features, while depth features are extracted using the Depth Anything Model (ViT-L). The depth features, prior to classification in the Depth Anything Model, are injected into each block of the SAM encoder. The fusion of depth and visual features is facilitated through the adapter structure. Importantly, only the adapter and decoder layers are updated during training, while all other parameters are frozen.

The visual-depth fusion is facilitated by two types of lightweight adapters, which explicitly enable cross-modal interaction. **Depth-Linear Adapter (DLA)**: Prior to the self-attention and MLP sub-layers of each block, the depth feature is projected to the visual embedding dimension (768) using a two-layer MLP (Linear, GELU, Linear), followed by Layer Normalization. The resulting feature vector is added element-wise to the visual token stream, allowing the depth cues to bias both attention and feed-forward network with negligible overhead. To adaptively inject the depth feature, we introduce a learnable coefficient γ for each DLA module. The coefficient γ is initially set to zero, ensuring that the depth features exert no influence during the initial training step. As training progresses, the network learns to adjust the coefficient γ , progressively weighting the depth features and adaptively incorporating depth information into the visual representations.

Linear Adapter (LA): After each sub-layer, a bottleneck adapter (Down-192→ ReLU → Up-768) learns task-specific residuals, while the frozen SAM weights preserve general segmentation knowledge. We initialize

the Down/Up projections with a scaled identity, such that, at epoch 0, the adapter behaves as a near-no-op, and gradually adjusts the representation during fine-tuning.

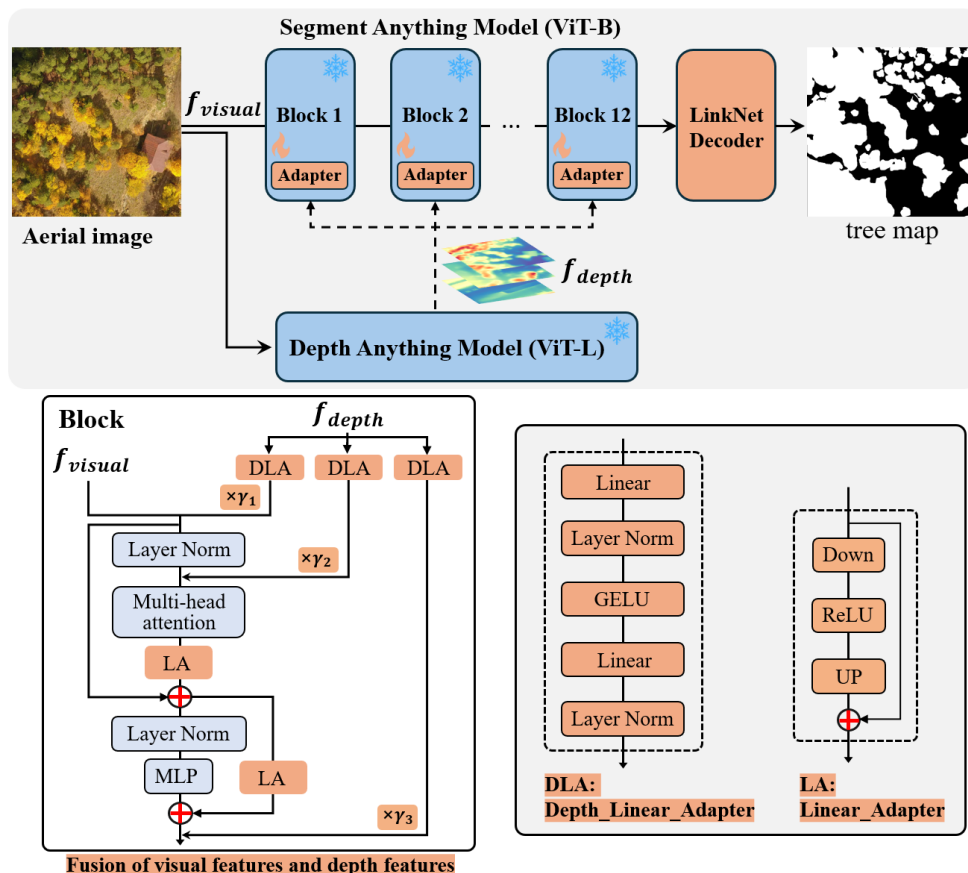


Figure 2. The proposed multimodal visual-depth fusion framework, where each block of SAM integrates visual features extracted by SAM and depth features obtained from Depth-Anything-V2.

4. Experiments

4.1. Implementation Details

Experiments are conducted on the OAM-TCD dataset, which consists of 5,072 high-resolution aerial images, each sized at 2048×2048 pixels with a spatial resolution of 0.1 m/pixel. In this study, all images are uniformly cropped into non-overlapping patches of 1024×1024 pixels. To simulate limited-data scenarios, 0.25% (42 samples) and 0.5% (84 samples) of the training set are randomly selected for model training. Evaluation is performed using 750 high-quality, manually annotated samples from the test set.

All experiments are performed on a single NVIDIA A100-SXM4 GPU with 80 GB of memory. Model training is conducted using the AdamW optimizer with a learning rate of 1e−5, a weight decay of 5e−4, a batch size of 4, and an input crop size of 768×768 pixels. The loss functions used in this study were Binary Cross-Entropy (BCE) and

Dice coefficient loss. The evaluation metrics included Precision (P), Recall (R), and F1 Score.

4.2. Experimental Results

In this section, we evaluate the performance of various methods under limited training data conditions, specifically with only 42 and 84 samples. All models adopt the LinkNet decoder architecture to reconstruct tree segmentation maps. Among them, SatMAE-LinkNet, ScaleMAE-LinkNet, CrossScaleMAE-LinkNet, SwinV2-LinkNet, SAM, and SAM_Fuse are fully fine-tuned by updating all model parameters during training. In contrast, methods based on Adapter, MLORA, and our proposed SAM_Adapter_Fuse freeze the pre-trained model weights and only update the parameters of the newly introduced modules. This leads to a significant reduction in the number of trainable parameters compared to full fine-tuning.

As shown in Table 1 and Table 2, under limited sample settings, the performance of the SAM_Adapter-based method is comparable to, and in many cases surpasses, that of the fully fine-tuned SAM. For instance, SAM_Fuse achieves an F1-score of 90.52 and 92.09, while SAM_Adapter_Fuse attains a higher F1-score of 91.33 and 92.40, respectively, with 42 and 84 training samples. This improvement can be attributed to the reduced risk of overfitting when training with limited samples, as the full fine-tuning of large models may lead to over-parameterization in data-scarce scenarios.

Compared to methods that directly combine the depth map and RGB aerial image as inputs to the network, such as SAM_Adapter (RGB, Depth), the proposed depth-aware method, SAM_Adapter_Fuse, can seamlessly integrate depth information from scratch. With 42 and 84 training samples, SAM_Adapter (RGB, Depth) achieves F1 scores of 91.13 and 91.76, respectively. However, it requires two steps: first, obtaining the depth map using Depth-Anything-V2, and then concatenating it with the RGB image. In contrast, SAM_Adapter_Fuse only requires the

input of the RGB image, enabling end-to-end automatic learning and integration of depth information, thereby achieving superior F1 scores of 91.33 and 92.40, respectively.

Figures 3 and 4 present two visual examples. As depicted in Figure 3, the red-boxed region highlights an area where SAM_Adapter (RGB) fails to detect the tree due to significant color and texture differences between the tree and its surrounding environment. In contrast, the Depth Map obtained from Depth-Anything-V2 clearly reveals the height differences between this region and its surroundings, enabling SAM_Adapter (Depth) to correctly identify the tree. However, simply concatenating the RGB and depth map in SAM_Adapter (RGB, Depth) does not fully exploit the depth map's advantages, resulting in the failure to detect the tree. Our proposed SAM_Adapter_Fuse, which automatically learns and integrates depth information, effectively leverages the depth map's benefits and successfully identifies the tree. Figure 4 demonstrates a similar pattern in the red-boxed region.

Methods	Input		Trainable Params (M)	Ref.	Pre-training data	Backbone	P	R	F1
	RGB	Depth Map							
SatMAELinkNet [5, 13]	✓		311.45	NeurIPS'22	fMoW-Sentinel-2	ViT-L	90.73	77.45	83.57
ScaleMAELinkNet [6, 13]	✓		311.45	ICCV'23	fMoW-RGB	ViT-L	90.18	84.50	87.25
CrossScaleMAELinkNet [7, 13]	✓		311.45	NeurIPS'23	fMoW-RGB	ViT-L	90.48	74.75	81.87
SAM_MLoRA ($r = 32, n = 3$) [14]	✓		16.88	TGRS'24	SA-1B	ViT-B	92.92	86.09	89.37
SwinV2LinkNet_MLoRA ($r = 32, n = 3$) [8, 13, 14]	✓		22.52	ICCV'23	Satlas-Sentinel-2	SwinV2-B	91.99	87.84	89.87
SwinV2LinkNet (RGB) [8, 13]	✓		88.86	ICCV'23	Satlas-Sentinel-2	SwinV2-B	92.70	86.52	89.50
SwinV2LinkNet (Depth) [8, 13]		✓	88.85	ICCV'23	Satlas-Sentinel-2	SwinV2-B	91.27	86.01	88.56
SAM_Adapter (RGB) [11]	✓		9.83	MedIA'25	SA-1B	ViT-B	92.09	88.43	90.22
SAM_Adapter (Depth) [11]		✓	9.83	MedIA'25	SA-1B	ViT-B	89.46	89.31	89.39
SAM_Adapter (RGB, Depth) [11]	✓	✓	9.83	MedIA'25	SA-1B	ViT-B	92.40	88.19	90.25
SAM (RGB) [2, 13]	✓		91.01	ICCV'23	SA-1B	ViT-B	92.43	86.53	89.38
SAM (Depth) [2, 13]		✓	90.61	ICCV'23	SA-1B	ViT-B	91.50	88.81	90.14
SAM (RGB, Depth) [2, 13]	✓	✓	91.20	ICCV'23	SA-1B	ViT-B	92.91	89.42	91.13
SAM_Fuse (Ours)	✓		172.99	ICCV'23	SA-1B	ViT-B	93.96	87.31	90.52
SAM_Adapter_Fuse (Ours)	✓		77.63	ICCV'23	SA-1B	ViT-B	93.52	89.23	91.33

Table 1. Performance of different models with 42 training samples.

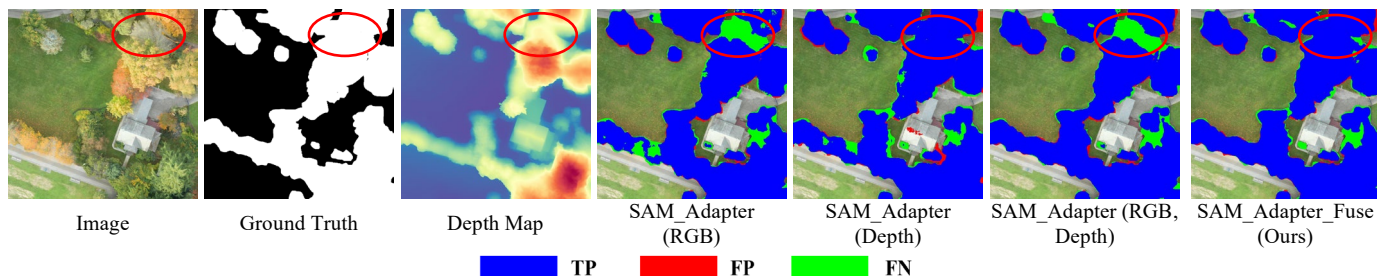


Figure 3. Visualization of the performance of various SAM frameworks on the test set, utilizing different modality data, with a training set comprising 42 samples.

Methods	Input		Trainable Params (M)	Ref.	Pre-training data	Backbone	P	R	F1
	RGB	Depth Map							
SatMAELinkNet [5, 13]	✓		311.45	NeurIPS'22	fMoW-Sentinel-2	ViT-L	91.18	82.36	86.55
ScaleMAELinkNet [6, 13]	✓		311.45	ICCV'23	fMoW-RGB	ViT-L	91.02	86.80	88.86
CrossScaleMAELinkNet [7, 13]	✓		311.45	NeurIPS'23	fMoW-RGB	ViT-L	89.79	78.68	83.87
SAM_MLoRA ($r = 32, n = 3$) [14]	✓		16.88	TGRS'24	SA-1B	ViT-B	90.12	93.85	91.95
SwinV2LinkNet_MLoRA ($r = 32, n = 3$) [8, 13, 14]	✓		22.52	ICCV'23	Satlas-Sentinel-2	SwinV2-B	90.68	91.97	91.32
SwinV2LinkNet (RGB) [8, 13]	✓		88.86	ICCV'23	Satlas-Sentinel-2	SwinV2-B	92.46	90.26	91.34
SwinV2LinkNet (Depth) [8, 13]		✓	88.85	ICCV'23	Satlas-Sentinel-2	SwinV2-B	90.98	87.94	89.44
SAM_Adapter (RGB) [11]	✓		9.83	MedIA'25	SA-1B	ViT-B	91.61	92.15	91.88
SAM_Adapter (Depth) [11]		✓	9.83	MedIA'25	SA-1B	ViT-B	88.70	92.38	90.51
SAM_Adapter (RGB, Depth) [11]	✓	✓	9.83	MedIA'25	SA-1B	ViT-B	92.05	91.82	91.93
SAM (RGB) [2, 13]	✓		91.01	ICCV'23	SA-1B	ViT-B	92.63	89.42	91.00
SAM (Depth) [2, 13]		✓	90.61	ICCV'23	SA-1B	ViT-B	90.77	90.58	90.67
SAM (RGB, Depth) [2, 13]	✓	✓	91.20	ICCV'23	SA-1B	ViT-B	93.57	90.03	91.76
SAM_Fuse (Ours)	✓		172.99	ICCV'23	SA-1B	ViT-B	92.81	91.37	92.09
SAM_Adapter_Fuse (Ours)	✓		77.63	ICCV'23	SA-1B	ViT-B	90.22	94.68	92.40

Table 2. Performance of different models with 84 training samples.

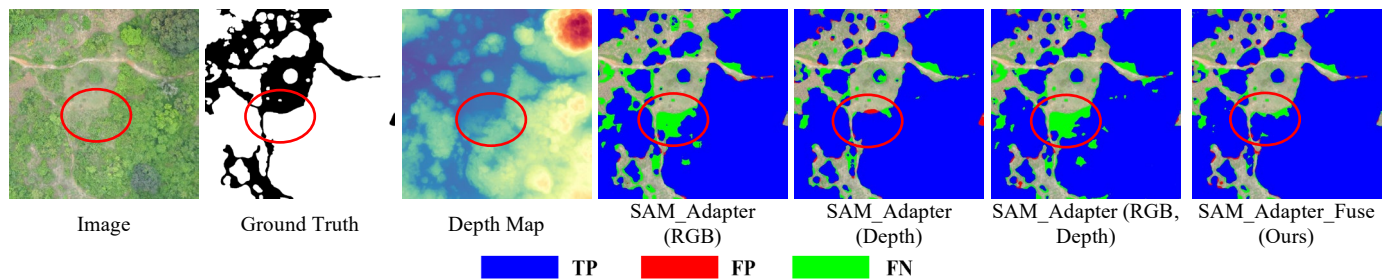


Figure 4. Visualization of the performance of various SAM frameworks on the test set, utilizing different modality data, with a training set comprising 84 samples.

5. Conclusion

This study presents a depth-aware multimodal fusion framework that integrates high-resolution RGB imagery with auxiliary depth cues to achieve data-efficient tree mapping in large-scale remote sensing scenes. By fine-tuning robust, pre-trained foundation encoders with lightweight adapters, the proposed architecture injects geometric context into each transformer layer, thereby enhancing recognition accuracy and boundary fidelity, particularly when labeled samples are limited. Our experiments yield two key findings: (1) depth information offers complementary structural guidance that conventional 2-D appearance cues lack; and (2) the adapter-based design achieves these gains while updating few parameters, greatly reducing annotation and computational costs. In practical terms, the framework reduces the barrier to obtaining reliable tree inventories, enabling stakeholders to monitor forest health, biomass, and biodiversity with fewer manual annotations. Future work will investigate unsupervised depth estimation to

eliminate the need for pre-computed depth maps, expand the method to multi-temporal change detection, and incorporate additional modalities such as hyperspectral data to further improve species-level tree discrimination.

References

- [1] J. Zheng, S. Yuan, W. Li, H. Fu, L. Yu, and J. Huang, "A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future," *IEEE Geoscience and Remote Sensing Magazine*, 2024.
- [2] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015-4026.
- [3] L. Yang *et al.*, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875-21911, 2024.

- [4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000-16009.
- [5] Y. Cong *et al.*, "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," *Advances in Neural Information Processing Systems*, vol. 35, pp. 197-211, 2022.
- [6] C. J. Reed *et al.*, "Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4088-4099.
- [7] M. Tang, A. Cozma, K. Georgiou, and H. Qi, "Cross-Scale MAE: A tale of multiscale exploitation in remote sensing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20054-20066, 2023.
- [8] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "SatlasPretrain: A large-scale dataset for remote sensing image understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16772-16782.
- [9] N. Ding *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220-235, 2023.
- [10] V. Lialin, V. Deshpande, and A. Rumshisky, "Scaling down to scale up: A guide to parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.15647*, 2023.
- [11] J. Wu *et al.*, "Medical sam adapter: Adapting segment anything model for medical image segmentation," *Medical Image Analysis*, vol. 102, p. 103547, 2025.
- [12] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [13] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017: IEEE, pp. 1-4.
- [14] X. Lu and Q. Weng, "Multi-LoRA fine-tuned segment anything model for urban man-made object extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1-19, 2024.