

# Earth-Adapter: Bridge the Geospatial Domain Gaps with a Frequency-Guided Mixture of Adapters

Xiaoxing Hu<sup>1\*</sup>, Ziyang Gong<sup>2\*</sup>, Yupei Wang<sup>1\*†</sup>, Yuru Jia<sup>3</sup>, Fei Lin<sup>4</sup>,  
Dexiang Gao<sup>5</sup>, Ke An<sup>1</sup>, Jianhong Han<sup>1</sup>, Zhuoran Sun<sup>1</sup>, Gen Luo<sup>6</sup>, Xue Yang<sup>2†</sup>

<sup>1</sup> Beijing Institute of Technology

<sup>2</sup> Shanghai Jiao Tong University

<sup>3</sup> KU Leuven

<sup>4</sup> Macau University of Science and Technology

<sup>5</sup> Peking University

<sup>6</sup> Shanghai Artificial Intelligence Laboratory

## Abstract

Vision Foundation Models (VFM), while powerful, often struggle in Remote Sensing (RS) segmentation tasks when combined with existing Parameter-Efficient Fine-Tuning (PEFT) methods. We observe that this limitation primarily arises from their inability to effectively handle the pervasive artifacts in RS imagery. To address this, we introduce Earth-Adapter, the first PEFT method specifically designed for RS artifact mitigation. Earth-Adapter introduces a novel Frequency-Guided Mixture of Adapters (MoA) approach, structured around a “divide and conquer” strategy. It first utilizes Discrete Fourier Transformation (DFT) to “divide” features into distinct frequency components, thereby effectively isolating artifact-related information from semantic signals. Subsequently, to “conquer” these artifacts, MoA independently optimizes features within different subspaces and dynamically assigns weights via a router to aggregate the refined representations. This enables adaptive refinement of the VFM’s representation space to mitigate the impact of artifacts. This simple yet highly effective PEFT method demonstrably mitigates artifacts and significantly enhances VFMs’ performance on RS segmentation tasks. Extensive experiments demonstrate Earth-Adapter’s effectiveness on in-domain semantic segmentation (SS), as well as Domain Adaptive (DA) and Domain Generalized (DG) semantic segmentation tasks. Compared with the baseline Rein, Earth-Adapter significantly improves mIoU by **1.2%** in SS, **9.0%** in DA, and **3.1%** in DG benchmarks.

## Introduction

Vision Foundation Models (VFMs) have made significant strides in recent years, driven by both natural language supervision paradigms (e.g., CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), SigLIP (Zhai et al. 2023)) and self-supervised paradigms (e.g., MAE (He et al. 2022a), DINOv2 (Oquab et al. 2023)). Pre-training on ultra-large-scale datasets empowers VFMs with robust zero-shot comprehension capabilities, leading to their extensive adoption

\*These authors contributed equally.

†Corresponding author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

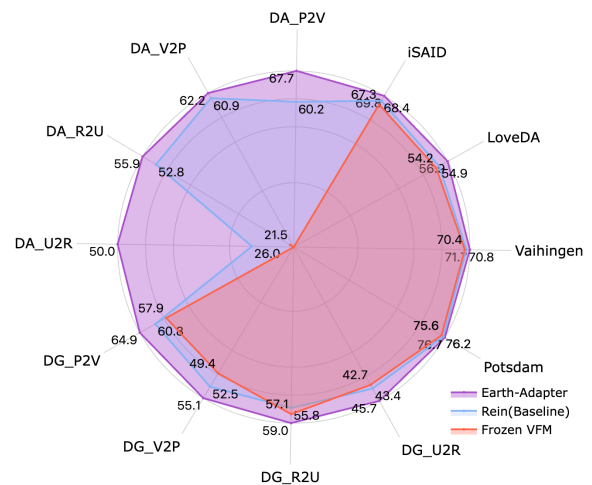


Figure 1: Performance across various remote sensing image segmentation benchmarks between Frozen VFM (DINOv2-L), Rein (Baseline) and the proposed Earth-Adapter.

across diverse tasks. These include classical computer vision tasks (classification (Zhou et al. 2022), detection (Zareian et al. 2021; Minderer et al. 2022), segmentation (Chen et al. 2024a)) as well as multimodal learning (Liu et al. 2023). Simultaneously, the conventional pre-train-then-fine-tune paradigm is evolving. When adapting VFMs to downstream tasks, the key challenge becomes how to efficiently preserve and unleash their inherent capabilities. Under this context, Parameter-Efficient Fine-Tuning (PEFT) methods (Hu et al. 2021; Jia et al. 2022; Houlsby et al. 2019a; Gong et al. 2024) emerge as the pivotal solution due to their superior parameter-performance trade-off.

Existing PEFT methods designed for language or nature imagery have made great progress with excellent works, such as LoRA (Hu et al. 2021), Visual Prompt Tuning (VPT) (Jia et al. 2022). However, most exhibit significant performance degradation in remote sensing (RS) segmentation tasks when integrated with DINOv2 (Oquab et al. 2023) (ex-

perimental evidence provided in later sections). We argue the *main reason lies in the influences of artifacts in the features*. As shown in Figure 2 (a), we show the PCA visualization of features of DINOv2-L, which contains obvious redundant artifacts. We have also observed that artifacts in natural images exhibit distinct characteristics compared to those in RS images. In natural images, artifacts typically surround foreground objects (Darcet et al. 2024), such as humans or animals, and the disturbances they cause are relatively limited. In contrast, RS images, due to their overhead perspective, lack centralized subjects and contain multiple coexisting multi-scale targets. For example, a single RS image may simultaneously include large-scale agricultural regions and fragmented road networks. As a result, artifacts are almost situated everywhere in RS images shown in Figure 2 (a), causing severe interference to pixel-level feature extraction, which is pivotal for segmentation tasks.

To address artifacts and improve VFM performance in RS segmentation tasks, we propose a “**divide-and-conquer**” strategy. In the “**divide**” stage, based on the observation that high-frequency (HF) signals capture local details while low-frequency (LF) signals encode global structures, we apply Discrete Fourier Transformation (DFT) to separate HF and LF components, isolating artifacts in the HF domain. In the “**conquer**” stage, we design the Mixture of Adapters (MoA), which adjusts features in their respective frequency domains and then uses a router to dynamically assign weights for aggregating the corrected features. These aggregated features are further combined with the original frozen VFM features through skip connections and dynamic scaling coefficients, preserving the strong feature extraction ability of the underlying model.

This unique pipeline forms Earth-Adapter, a PEFT method tailored for RS semantic segmentation. We comprehensively evaluate Earth-Adapter on 12 established benchmarks across three settings—in-domain (SS), domain adaptive (DA), and domain generalized (DG)—and demonstrate its effectiveness. Compared with existing PEFT methods, Earth-Adapter achieves state-of-the-art (SOTA) performance on SS, DA and DG benchmarks and outperforming our baseline Rein (Wei et al. 2024) by **1.2%**, **9.0%** and **3.1%** mIoU, respectively. We also conduct an in-depth analysis of Earth-Adapter’s design and potential. The core contributions can be summarized as follows:

- We observe that VFMs are prone to artifact interference in the high-dimensional feature space of RS imagery, which hinders their adaptation to RS segmentation tasks.
- To address this, we propose Earth-Adapter, the first PEFT approach tailored specifically to mitigate artifact-related issues.
- We develop a Frequency-Guided Mixture of Adapters (MoA) that effectively isolates artifacts within distinct frequency subspaces, coupled with a dynamic router that adaptively fuses adapter outputs to optimize pixel-level VFM features for RS imagery.
- Extensive experiments across diverse RS segmentation benchmarks demonstrate the effectiveness of Earth-Adapter.

## Related Work

### RS Semantic Segmentation

Semantic segmentation (Long, Shelhamer, and Darrell 2015; Shelhamer, Long, and Darrell 2016; Chen et al. 2017) aims to classify every pixel in an image. RS segmentation is a critical task in numerous real-world applications, including land cover mapping, urban planning, and environmental monitoring (He et al. 2022b). Cross-domain task is crucial due to the diversity of data, and lots of DA methods have been applied to improve the generalization ability of segmentation models (Rui and Jintao 2020; Tong et al. 2020). Compared with DA, the number of DG research (Iizuka, Xia, and Yokoya 2023; Liang et al. 2024) in RS is much less than DA. Although these works (Zhu et al. 2023) make contributions, most of them only focus on specialized models rather than leveraging the power of VFMs, leading to limited cross-domain capabilities.

### Vision Foundation Models

Vision Foundation Models (VFMs) have made significant progress and have exerted considerable influence across various AI-related fields. Among them, vision-language foundation models—such as CLIP (Radford et al. 2021), ALIGN (Jia et al. 2021), SigLIP (Zhai et al. 2023), and InternViT (Chen et al. 2024b)—are trained on billions of image-text pairs to learn joint visual-textual representations. These models have been widely applied to a broad range of vision and multimodal tasks, including classification (Zhou et al. 2022), detection (Zareian et al. 2021; Minderer et al. 2022), segmentation (Chen et al. 2024a), image-text retrieval (Radford et al. 2021), image captioning (Dai et al. 2023), and large multimodal models (LMs) (Luo et al. 2024). On the other hand, self-supervised models can also learn rich visual representations from large-scale unlabeled data. MAE (He et al. 2022a) learns generalized visual features by reconstructing masked images from partial inputs. MoCo (Chen, Xie, and He 2021) employs a momentum encoder and contrastive learning to extract discriminative representations. DINOv2 (Oquab et al. 2023) combines self-distillation with contrastive learning to produce scalable, transferable features for various downstream tasks. *Beyond these various VFMs, a key challenge remains in effectively unlocking their potential to enhance performance on downstream tasks. Earth-Adapter is a pioneering work for RS segmentation tasks*

### PEFT in RS

PEFT approaches, such as adapter (Houlsby et al. 2019a,b; Hu et al. 2021) are proposed to fine-tune large pre-trained models to downstream tasks in Natural Language Processing (NLP) by introducing light-weight learnable parameters. Subsequently, the PEFT has begun to be transferred to vision (Yin et al. 2023; Agiza, Neseem, and Reda 2024) tasks and rapidly become the focus of researchers. Visual Prompt Tuning (VPT) (Jia et al. 2022) first introduced ‘prompt tuning’ into the vision area as learnable vectors. The thought of this paradigm is also similar to the adversarial re-programming (Elsayed, Goodfellow, and Sohl-Dickstein

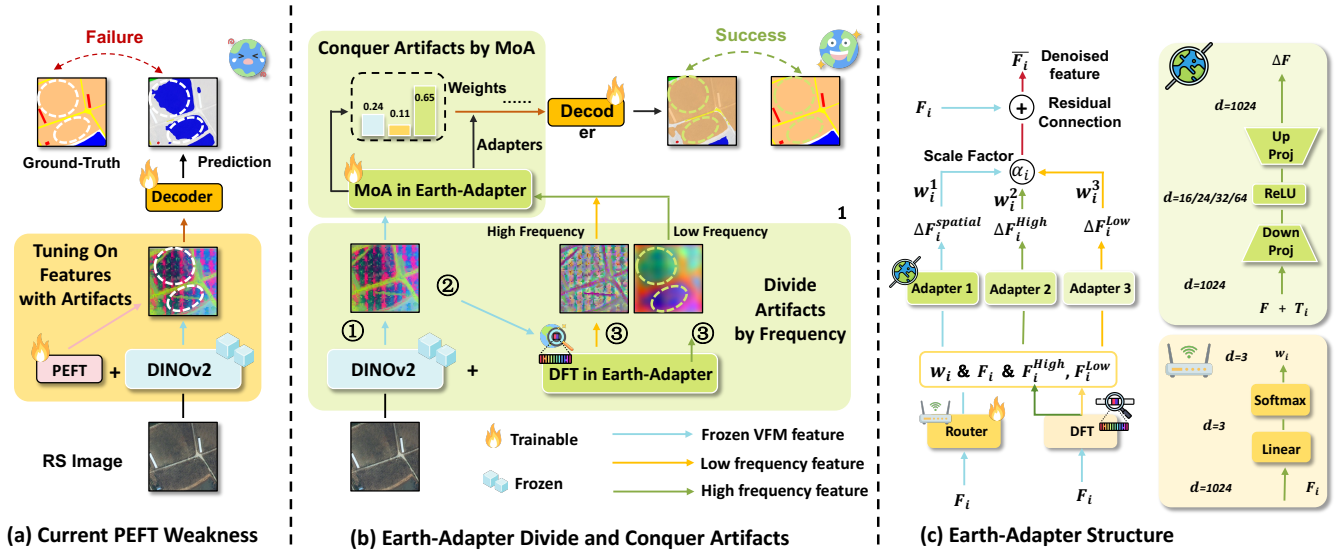


Figure 2: Motivation and Structure Details of Earth-Adapter (a) points out the artifact problems in existing PEFT methods. (b) illustrates how Earth-Adapter divides and conquers the artifacts by frequency-guided strategy and MoA framework. ①, ②, and ③ show the sequence of each step in the DFT operation. (c) introduces the details of the Earth-Adapter component structures.

2018). Currently, cross-domain works also focus on PEFT, such as (Ge et al. 2023) which is the first prompt tuning method in Domain Adaptation (DA) (Sun et al. 2023; Gao et al. 2022; Gong et al. 2024) and (Wei et al. 2024) is the first work leveraging PEFT to fine-tuning VFMs for Domain Generalization (DG) (Bi et al. 2024). In the RS field, the exploration of PEFT is also vigorous. For example, UPetu (Dong, Gu, and Liu 2024) addresses storage for dense prediction tasks. TEA (Hu et al. 2024) uses an adapter network and top-down guidance for detection and segmentation tasks. (Pu and Xu 2025) also leverage LoRA to solve oriented detection task. However, for the RS semantic segmentation tasks, the research of PEFT is relatively blank, and existing methods can not handle the influence of artifacts. *Earth-Adapter is the first PEFT designed to tackle artifacts, thereby unlocking the potential of VFMs for remote sensing segmentation tasks.*

## Method

### Preliminary

We first denote the input RS image as  $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$ , the corresponding label of  $\mathbf{x}$  as  $\mathbf{y} \in \mathbb{R}^{H \times W \times K}$ , and the semantic segmentation prediction of  $\mathbf{x}$  as  $\mathbf{y}' \in \mathbb{R}^{H \times W \times K}$ , where  $K$  represents the number of semantic classes.

Here, we'll illustrate our innovative Earth-Adapter overview.

The proposed Earth-Adapter is composed of two key components: the MoA and Router. MoA consists of a *Spatial Adapter*, a low-frequency (*LF Adapter*), a high-frequency (*HF Adapter*). We employ  $R_\xi$ , parameterized by  $\xi$ , to denote the router and  $E_\epsilon^i$ ,  $i \in \{1, 2, 3\}$ , parameterized by  $\epsilon$ , to represent the three adapter experts. For the network architecture, we adopt the framework introduced by Rein (Wei et al. 2024), integrating DINOv2-L (Oquab et al. 2023) as the

backbone, represented as  $f_\phi$  and parameterized by  $\phi$ , alongside Mask2Former (Cheng et al. 2022) as the decoder, denoted by  $f_\theta$  and parameterized by  $\theta$ . The following presents the details of Earth-Adapter, and we will elaborate on our training framework in the Training Framework section.

### Details of Earth-Adapter

Earth-Adapter aims to adapt VFMs to RS image semantic segmentation tasks with minimal learnable parameters.

**Optimization objective.** Before employing the Earth-Adapter, the optimization objective of fine-tuning VFMs is to identify a set of parameters that minimize the loss of the entire model on the downstream task:

$$\arg \min_{\theta} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} \mathcal{L}_{\text{seg}}(f_\theta(f_\phi(\mathbf{x})), \mathbf{y}), \quad (1)$$

where  $\mathcal{D}$  is the dataset and  $\mathcal{L}_{\text{seg}}$  is the loss function, here we use the default CE loss. After incorporating our Earth-Adapter, the optimization objective becomes:

$$\arg \min_{\theta, \epsilon, \xi} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D}} \mathcal{L}_{\text{seg}} \left( f_\theta \left( \underbrace{R_\xi \circ E_\epsilon}_{\text{Earth-Adapter}}(f_\phi^*(\mathbf{x})) \right), \mathbf{y} \right). \quad (2)$$

During training, the original parameters of the backbone (VFM) are kept frozen (denoted as  $f_\phi^*$ ). In the following description, we denote the segmentation network combined with Earth-Adapter as  $G_{\theta, \xi, \epsilon}$ .

**Mixture of Adapters.** Let  $\mathbf{F}_i \in \mathbb{R}^{(hw) \times c}$  denote the visual feature at the  $i$ -th layer of the backbone, where  $hw$  represents the token sequence length and  $c$  denotes the token dimension. Our frequency adaptation operates through parallel processing streams: The first *Spatial Adapter* employs

a low-rank projection to refine the spatial feature:

$$\Delta \mathbf{F}_i^{spatial} = \text{Adapter}_1^i(\mathbf{F}_i^T), \quad (3)$$

where  $\Delta \mathbf{F}_i^{spatial}$  means the processed  $i$ -th layer Spatial-Domain features, and  $\text{Adapter}_1$  represents a nonlinear mapping layer composed of two low-rank matrices and an activation:

$$\text{Adapter}_i = \mathbf{W}_{up}(\text{Relu}(\mathbf{W}_{down}(\cdot))). \quad (4)$$

The *Frequency Adapters* consist of a *HF Adapter* (high-frequency subspace) and a *LF Adapter* (low-frequency subspace), which fine-tune features in specific frequency subspaces derived from a 2D Discrete Fourier Transform (DFT) decomposition. We first reshape  $\mathbf{F}^{spatial}$  to  $(C, H, W)$  and apply DFT on spatial features as  $\mathcal{FT}(\mathbf{F}^{spatial})$ . When splitting the frequency domains, we employ a fixed frequency cutoff  $\rho$  to decompose the spectrum into high and low frequency components. Subsequently, these components are transformed back into features via the Inverse Fourier Transform (IFT), yielding *LF* and *HF* features:

$$\mathbf{F}_i^{low} = \mathcal{FT}^{-1}(\mathbf{M} \odot \mathcal{FT}(\mathbf{F}^{spatial})), \quad (5)$$

$$\mathbf{F}_i^{high} = \mathcal{FT}^{-1}((1 - \mathbf{M}) \odot \mathcal{FT}(\mathbf{F}^{spatial})). \quad (6)$$

The frequency mask  $\mathbf{M} \in \{0, 1\}^{H \times W}$  is defined by:

$$\mathbf{M}(u, v) = \begin{cases} 1 & , \text{if } \max(|u - \frac{H}{2}|, |v - \frac{W}{2}|) \leq \rho \frac{H}{2} \\ 0 & , \text{otherwise} \end{cases} \quad (7)$$

Thereafter, the *LF* and *HF* features are independently passed through two distinct low-rank linear projection layers, generating frequency adaptation adjustments:

$$\Delta \mathbf{F}_i^{low} = \text{Adapter}_2^i(\mathbf{F}_i^{low}), \quad (8)$$

$$\Delta \mathbf{F}_i^{high} = \text{Adapter}_3^i(\mathbf{F}_i^{high}). \quad (9)$$

The *Frequency Adapters* shares the same structure as the *Spatial Adapter*, as defined in Equation 4.

**Dynamic Router.** We implement dynamic feature aggregation through a Router that learns optimal combinations of feature adjustment according to the original visual features. The Router weight is computed by channel-wise attention:

$$\mathbf{w}_i = \text{Softmax}(R_\xi(\mathbf{F}_i)), \quad (10)$$

where  $\mathbf{w}_i$  represents weights for *spatial*, *LF*, and *HF* components. The final feature adjustment is calculated as:

$$\Delta \mathbf{F}_i = \alpha_i \sum_{k=1}^3 \mathbf{w}_i^{(k)} \Delta \mathbf{F}_i^{(k)}, \quad (11)$$

where  $\alpha_i$  is a learnable scaling parameter with small initial value, and  $k \in \{spatial, low, high\}$ . The frozen features and the refined features are fused via a skip connection:

$$\bar{\mathbf{F}}_i = \mathbf{F}_i + \Delta \mathbf{F}_i. \quad (12)$$

$\bar{\mathbf{F}}_i$  is then forwarded to the subsequent Transformer block to continue the layer-wise processing.

## Training Framework

In this work, we explore the application of Earth-Adapter across three distinct semantic segmentation subtasks: semantic segmentation (SS), domain adaptive (DA) semantic segmentation and domain generalized (DG) semantic segmentation. For SS and DGSS, we train Earth-Adapter with Mask2former (Cheng et al. 2022) in end-to-end supervised learning as described in Equation 2. For DA semantic segmentation, we use the self-training framework introduced in DACS (Tranheden et al. 2021) without using target domain labels. DACS employs an EMA teacher to generate pseudo labels for target-domain images, mixes target and source samples to obtain the mixed domain  $(\mathbf{x}_{mix}, \mathbf{y}_{mix})$  and then jointly trains on both the source and mixed domains. The optimization objective for the DASS is defined by:

$$\arg \min_{\theta, \epsilon, \xi} \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{D} \cup \mathcal{D}_{mix}} \mathcal{L}_{seg}(G_{\theta, \xi, \epsilon}(\mathbf{x}), \mathbf{y}) + \lambda_{da} \mathcal{L}_{seg}(G_{\theta, \xi, \epsilon}(\mathbf{x}_{mix}), \mathbf{y}_{mix}). \quad (13)$$

$\lambda_{da}$  is the weight parameter for DA loss.

## Experiments

### Experimental setup

**Datasets and Benchmarks.** We conduct all experiments on several widely used RS image segmentation datasets: Potsdam (Markus Gerke 2014), Vaihingen (Markus Gerke 2014), LoveDA (Wang et al. 2021), and iSAID (Waqas Zamir et al. 2019). Following the standard MMSegmentation configurations, we split each dataset into training and validation sets, apply cropping, and construct four semantic segmentation (SS) benchmarks: **Potsdam(P)**, **Vaihingen(V)**, **LoveDA(L)**, **iSAID(i)**, along with four domain adaptation (DA) and four domain generalization (DG) tasks: **Potsdam** to **Vaihingen** (P2V), **Vaihingen** to **Potsdam** (V2P), **Rural** to **Urban** (R2U), **Urban** to **Rural** (U2R). In our experimental setup, all images are cropped to a size of  $512 \times 512$ .

**Implementation Details.** We use MMSegmentation as our training and evaluation framework. We employ Mask2former (Cheng et al. 2022) as our decoder, which is a highly efficient and effective algorithm for semantic segmentation tasks. For backbone, we utilize the Dinov2-Large (Oquab et al. 2023) model to extract features, which are subsequently utilized as input for the Mask2former. During training, we utilize AdamW (Loshchilov and Hutter 2019) as the optimizer, with a learning rate of  $1e-5$  for the backbone,  $1e-4$  for the decoder, and  $1e-4$  for the relevant parameters in PEFT. Within the DACS training framework, we set  $\lambda_{da}$  to 0.5.

### Main Result

As depicted in Table 1, we conduct a comparative analysis of the existing mainstream PEFT methods, such as Adapter (Houlsby et al. 2019a), LoRA (Hu et al. 2021), and VPT (Jia et al. 2022), against our Earth-Adapter across eight cross-domain benchmarks, with state-of-the-art

Methods	DASS				Avg.	DGSS				Avg.
	P2V	V2P	R2U	U2R		P2V	V2P	R2U	U2R	
<i>Previous SOTA methods</i>										
DAFormer	64.4	54.8	52.7	42.5	53.6	42.4	41.4	54.2	39.9	44.5
HRDA	<u>67.6</u>	58.6	53.2	35.3	53.7	33.1	31.1	54.2	39.8	39.6
<i>PEFT-based methods with VFM</i>										
Frozen	21.0	7.2	21.5	11.8	15.4	57.9	49.4	57.1	42.7	51.8
Frozen (with register)	11.3	28.2	36.4	16.1	23.0	59.0	51.1	58.9	44.9	53.5
Full Fine-Tune	11.3	16.5	23.1	10.6	15.4	12.6	19.6	33.1	19.7	21.3
Adapter	66.4	59.3	<b>55.9</b>	46.2	57.0	56.0	47.4	<u>57.6</u>	41.8	50.7
LoRA	17.8	18.8	24.5	26.0	15.7	20.3	25.6	29.3	21.9	24.3
VPT	66.2	59.3	<u>55.3</u>	<u>48.0</u>	<u>57.2</u>	59.2	52.3	57.4	<u>44.9</u>	<u>53.5</u>
AdaptFormer	12.9	15.3	25.1	22.2	18.9	14.0	19.3	31.0	15.8	20.0
Rein (baseline)	60.2	<u>60.9</u>	52.8	26.0	50.0	<u>60.8</u>	<u>52.5</u>	55.8	43.4	53.1
<b>Earth-Adapter (Ours)</b>	<b>67.7</b>	<b>62.2</b>	<b>55.9</b>	<b>50.0</b>	<b>59.0</b>	<b>64.9</b>	<b>55.1</b>	<b>59.0</b>	<b>45.7</b>	<b>56.2</b>
$\Delta$ over baseline method	<b>+7.5</b>	<b>+1.3</b>	<b>+3.1</b>	<b>+24.0</b>	<b>+9.0</b>	<b>+4.1</b>	<b>+2.6</b>	<b>+3.2</b>	<b>+2.3</b>	<b>+3.1</b>

Table 1: Performance (mIoU%) comparison between previous SOTA methods and PEFT-based methods on DA and DG benchmarks. **Bold** indicates the best performance. Underlined results denote the second-best. The last row shows improvements over the baseline.

Method	P	V	L	i	Avg.
<i>Traditional Methods</i>					
DeepLabV3+ (R-101)	74.8	69.7	51.4	53.4	62.3
PSPNet (R-101)	74.7	69.2	48.5	57.5	62.5
UperNet (Swin-B)	75.5	69.4	55.6	67.9	67.1
Segformer (MiT-B5)	75.3	68.4	55.5	66.8	66.5
Mask2Former (Swin-B)	75.7	<u>71.2</u>	54.7	64.0	66.4
<i>PEFT-based methods with VFM</i>					
Frozen	75.6	70.4	54.2	67.3	66.9
Frozen (with register)	75.7	70.8	<u>56.0</u>	67.5	67.5
Full Fine-Tune	70.5	64.6	43.9	12.0	47.7
LoRA	71.9	64.6	51.0	48.5	59.0
VPT	75.9	70.6	54.6	<u>68.5</u>	67.4
AdaptFormer	65.4	56.0	42.0	<u>7.5</u>	42.8
Rein	<u>76.2</u>	70.8	54.9	68.4	67.6
<b>Earth-Adapter (Ours)</b>	<b>76.7</b>	<b>71.7</b>	<b>56.9</b>	<b>69.8</b>	<b>68.8</b>
$\Delta$ over baseline method	<b>+0.5</b>	<b>+0.9</b>	<b>+2.0</b>	<b>+1.4</b>	<b>+1.2</b>

Table 2: Performance (mIoU%) comparison between previous SOTA methods and PEFT-based methods on SS benchmarks. **Bold** indicates the best performance. Underlined results denote the second-best. The last row shows improvements over the baseline.

method Rein (Wei et al. 2024) in natural scene as our baseline. These benchmarks encompass four DA scenarios (P2V, V2P, R2U, and U2R) and four DG scenarios that mirror the DA ones.

Starting with the DA experiments, many existing PEFT methods suffer from severe performance degradation when compared to DA-specialized models. For example, LoRA achieves only 17.8% mIoU on the P2V (DA) benchmark for RS images, even lower than the Frozen DINOv2-L back-

<i>General VFMs</i>					
Backbone	Methods	Params	P2V(DG)	V2P(DG)	Avg.
DINOv2-S	Frozen	0.0M	44.3	43.3	43.8
	Earth-Adapter	2.6M*9.6M	<b>47.5</b>	<b>44.8</b>	<b>46.2</b>
DINOv2-B	Frozen	0.0M	51.4	49.5	50.5
	Earth-Adapter	2.6M*9.6M	<b>53.0</b>	<b>51.8</b>	<b>52.4</b>
DINOv2-L	Frozen	0.0M	57.9	49.4	53.7
	Earth-Adapter	2.6M*9.6M	<b>64.9</b>	<b>55.1</b>	<b>60.0</b>
<i>RS-pretrained VFMs</i>					
Backbone	Methods	Params	P2V(DA)	V2P(DA)	Avg.
MTP-L	Frozen	0.0M	32.0	33.0	32.5
	Earth-Adapter	2.6M*9.6M	<b>40.8</b>	<b>41.0</b>	<b>40.9</b>
ScaleMAE-L	Frozen	0.0M	20.0	19.8	19.9
	Earth-Adapter	2.6M*9.6M	<b>20.2</b>	<b>29.6</b>	<b>24.9</b>
DOFA-L	Frozen	0.0M	9.9	13.7	11.8
	Earth-Adapter	2.6M*9.6M	<b>26.2</b>	<b>33.2</b>	<b>20.7</b>

Table 3: Ablation studies of Earth-Adapter on different VFMs. Results validate that Earth-Adapter’s effectiveness on diverse backbone networks.

bone. In contrast, our Earth-Adapter demonstrates much stronger domain adaptation capabilities. It surpasses the baseline Rein by 7.5% mIoU on P2V, 3.1% mIoU on R2U, 24.0% mIoU on U2R, and achieves an average improvement of 9.0% mIoU across all DA benchmarks. Shifting to the DG experiments, the issue of performance degradation remains, although the trend is somewhat mitigated. Unlike in the DA setting, the Frozen model highlights the potential generalizability of DINOv2-L. However, FFT, LoRA, and AdaptFormer continue to show clear limitations. Consistent with the DA benchmarks, Earth-Adapter achieves state-of-the-art (SOTA) performance in DG, surpassing the baseline Rein and Adapter by 3.1% and 4.4% mIoU, respectively.

We also conducted comparisons on four SS benchmarks,

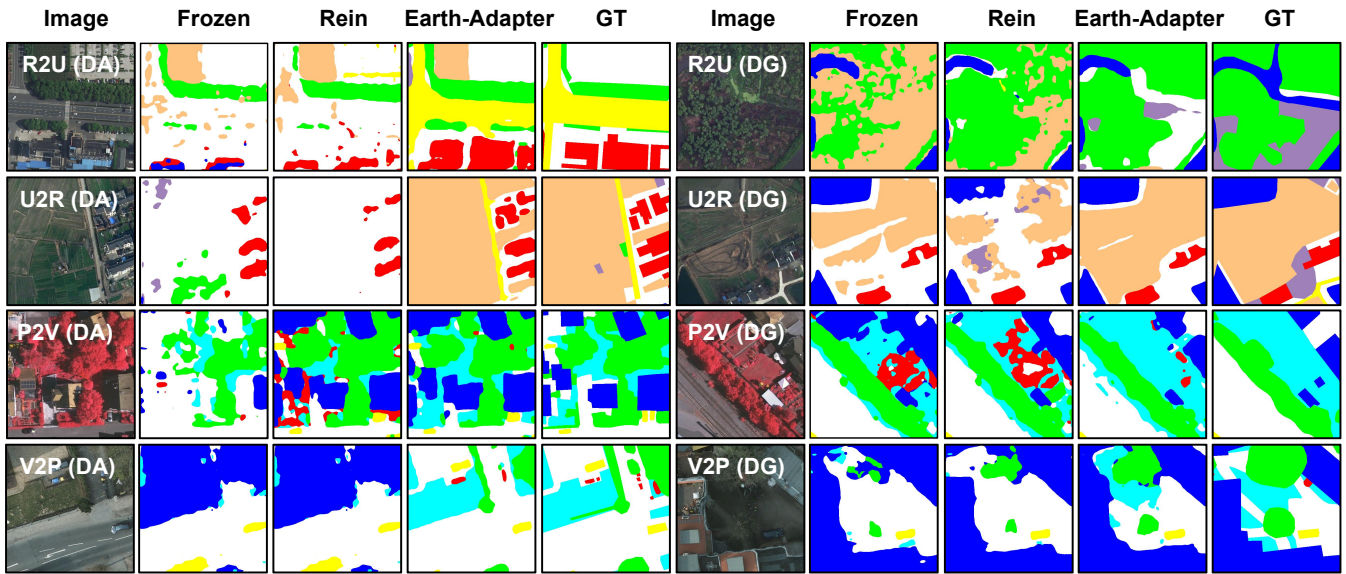


Figure 3: Visualization of Predicted Segmentation Maps We Compare Earth-Adapter with the Frozen DINOv2-L backbone and our baseline Rein on eight cross-domain benchmarks. For the Potsdam and Vaihingen color map, white is the Impervious surface, red is the clutter, blue is the building, Cyan is the low vegetation, green is the tree, and yellow is the car. For LoveDA color map, red is the building, yellow is the road, blue is the water, purple is the barren, green is the forest, brown is the agriculture.

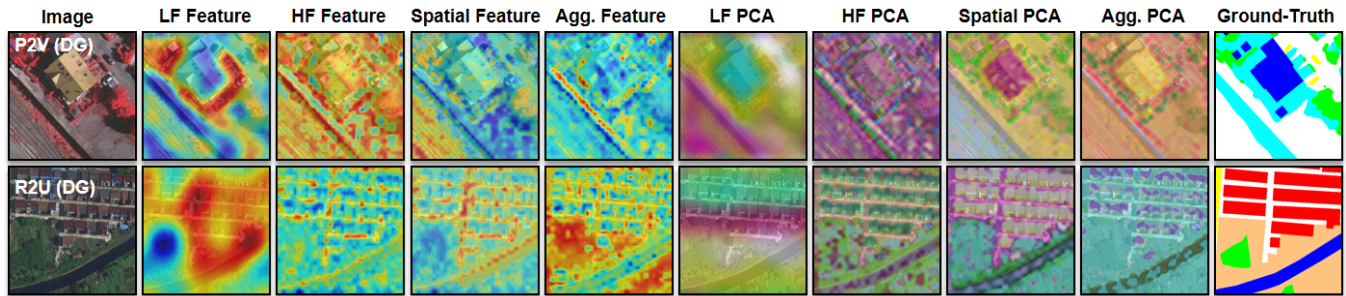


Figure 4: Visualization and PCA of Adapters’ Feature Maps. ‘Agg. Feature’ represents the aggregated adapters’ features. ‘PCA’ represents the Principal Component Analysis of features. All visualizations represent feature maps, not heatmaps. Thus only the semantic boundaries within the features should be focused rather than color intensities.

with the results summarized in Table 2. Regarding PEFT-related methods, conventional approaches such as LoRA and AdaptFormer struggle to adapt to the characteristics of remote sensing images, often disrupting the representations of VFMs during fine-tuning and ultimately leading to poor performance. In contrast, our Earth-Adapter consistently achieves the best performance among all PEFT methods. We attribute this advantage to its unique design, which mitigates high-dimensional artifacts and effectively “denoises” semantic representations, thereby enabling more accurate pixel-level feature extraction and superior semantic segmentation performance. Overall, our Earth-Adapter surpasses the baseline Rein by an average of 1.2% mIoU, highlighting its effectiveness.

### Ablation and Analysis

**Ablation of Different Backbones.** Beyond the DINOv2-Large backbone used in the main results (Tables 1 and 2), we further evaluated other DINOv2 variants. Results for ViT-Small and ViT-Base (Table 3, rows 3–8) show that Earth-Adapter consistently improves performance across backbones of different scales. We also tested VFMs pre-trained on remote sensing data, including MTP-Large (Wang et al. 2024), ScaleMAE-Large (Reed et al. 2023), and DOFA-Large (Xiong et al. 2024) (Table 3, rows 11–16). Earth-Adapter again delivers stable improvements in segmentation performance. Finally, we observed that DINOv2 models outperform those pre-trained on remote sensing data, likely because the latter are trained on smaller-scale datasets and thus remain less mature.

**Component Effectiveness Ablation.** In Table 4, ablation studies on MoA’s components and adapter numbers show

Backbone	Methods	Params ↓	Train Speed (s/iter) ↓	Infer Speed (s/iter) ↓	mIoU (%) ↑
DINOv2-L	Full	304.2M	0.80	0.11	15.4
	Rein	3.0M	0.68	0.12	50.0
	Earth-Adapter	2.6M~9.6M	0.71	0.12	<b>59.0 (+9.0)</b>

Table 4: Comparison on Parameters, Speed, and Performance between FFT, Rein, and Earth-Adapter. All results are the average score of DA benchmarks, demonstrating the excellent trade-off of Earth-Adapter.

Adapter	+ HF	+ LF	P2V (DG)	V2P (DG)	Avg.
1	-	-	61.0	52.8	56.9 (+0.0)
-	1	-	59.4	50.9	50.9 (-6.0)
-	-	1	59.1	51.5	55.3 (-1.6)
2	-	-	61.6	51.4	56.5 (-0.4)
3	-	-	64.0	52.7	58.4 (+1.5)
4	-	-	61.4	53.3	57.4 (+0.5)
5	-	-	62.5	52.5	57.5 (+0.6)
1	1	1	<b>64.9</b>	<b>55.1</b>	<b>60.0 (+3.1)</b>
2	1	1	61.3	54.7	58.0 (+1.1)
3	1	1	61.6	54.6	58.1 (+1.2)

Table 5: Ablation of Adapters’ Combination. The ‘Adapter’ means the number of vanilla Adapter that accepts *Spatial* features. The ‘+ HF’ and ‘+ LF’ represent the number of adapters that accept high and low-frequency features as input. The ‘+’ in the bracket represents the gain compared with one Adapter.

that using only *HF* or *LF* adapters decreases performance compared to spatial adapters, with *HF* causing a larger drop. This aligns with PCA visualizations in 2 (b) and 4, where *HF* features have more artifacts and noise, while *LF* features are smooth with clear global semantics. Thus, using only *HF* severely damages performance. Although only using *LF* features outperforms *HF* features (50.9% mIoU), it still underperforms *spatial* features (56.9% mIoU) due to lacking semantic details. After that, we also conduct more experiments on the number of *Spatial* adapters, and reveal the best composition is the one *spatial* adapter with one *HF* and *LF* adapters.

**Prediction Visualization Analysis.** In Figure 3, we visualize the predicted segmentation map of the Frozen DINOv2-L, Rein, and Earth-Adapter. It is worth noting that in the U2R (DG) example, Rein’s prediction is worse than the original backbone, which means Rein can not adapt to RS image well, which leads to a negative effect on the backbone features. In contrast, on the backbone’s basement, the Earth-Adapter keeps the good details of backbone features, and further optimizes the representation of agriculture class. And in other experiments, Earth-Adapter also exhibits better prediction than Rein and Frozen backbone, showing a higher performance upperbound.

**Feature Visualization Analysis.** In Figure 4, we visualize features captured by three adapters on P2V (DG) and R2U (DG) benchmarks. As discussed in the introduction, *LF* features focus on coarse-grained and global semantics, while *HF* features show detailed representations. The aggregated

Method	P2V(DG)	V2P(DG)	Avg.
w/o DR	63.6	53.2	58.4
w/ DR	<b>64.9</b>	<b>55.1</b>	<b>60.0(+1.6)</b>

Table 6: Ablation of Dynamic Router.

features are a weighted summation of the three different frequency features. PCA visualizations clearly show each feature’s characteristics. Take P2V in the figure as an example, artifacts are almost filtered into high-frequency features (shown in ‘*HF* PCA’), and dynamic fusion of all features ensures the final aggregated features’ PCA maintains clear semantic edges and successfully filters out artifacts. These visualizations further enhance Earth-Adapters’ explainability.

**Performance-Speed Trade-Off Analysis.** We compare the parameters, speed, and performance between FFT, Rein, and Earth-Adapter in Table 4. Significantly, at the same time of close training and inference speed with Rein, Earth-Adapter achieves better performance (9.0% mIoU improvement), with a smaller parameter scale. This further confirms the efficiency of Earth-Adapter, which can be attributed to its simple yet effective design methodology, making it particularly well-suited for semantic segmentation tasks on remote sensing images.

**Ablation of Dynamic Router.** As shown in Tab. 6, the use of the dynamic router leads to better results, yielding an average improvement of 1.6% mIoU across the two DG benchmarks. Compared with static weights (where we assign a weight of 1/3 to each static route), the dynamic router can adaptively adjust the feature allocation weights, enabling more effective representation optimization.

## Conclusion

Vision Foundation Models (VFM) excel across diverse visual tasks, and pairing them with Parameter-Efficient Fine-Tuning (PEFT) is an effective way to adapt them to downstream applications. However, existing PEFT methods often fall short on remote sensing (RS) semantic segmentation because they fail to suppress the pervasive artifacts in VFM deep features for RS imagery. To address this limitation, we introduce Earth-Adapter, a simple yet effective PEFT framework tailored for RS segmentation. It employs a frequency-guided mixture of adapters (MoA) that isolates artifacts into the high-frequency subspace, fine-tunes features within each subspace, and adaptively fuses them through a learnable router.

## Acknowledgments

This work was partly supported by National Natural Science Foundation of China (62301046, 62506229) and Natural Science Foundation of Shanghai (25ZR1402268).

## References

- Agiza, A.; Neseem, M.; and Reda, S. 2024. MTLORA: Low-Rank Adaptation Approach for Efficient Multi-Task Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16196–16205.
- Bi, Q.; Yi, J.; Zheng, H.; Zhan, H.; Huang, Y.; Ji, W.; Li, Y.; and Zheng, Y. 2024. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37: 94047–94072.
- Chen, K.; Liu, C.; Chen, H.; Zhang, H.; Li, W.; Zou, Z.; and Shi, Z. 2024a. RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.
- Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.
- Chen, X.; Xie, S.; and He, K. 2021. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9640–9649.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Intervl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.
- Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girshick, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.
- Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.
- Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*.
- Dong, Z.; Gu, Y.; and Liu, T. 2024. Upetu: A unified parameter-efficient fine-tuning framework for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*.
- Elsayed, G. F.; Goodfellow, I.; and Sohl-Dickstein, J. 2018. Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*.
- Gao, Y.; Shi, X.; Zhu, Y.; Wang, H.; Tang, Z.; Zhou, X.; Li, M.; and Metaxas, D. N. 2022. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*.
- Ge, C.; Huang, R.; Xie, M.; Lai, Z.; Song, S.; Li, S.; and Huang, G. 2023. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*.
- Gong, Z.; Li, F.; Deng, Y.; Bhattacharjee, D.; Ma, X.; Zhu, X.; and Ji, Z. 2024. CoDA: Instructive chain-of-domain adaptation with severity-aware visual prompt tuning. In *European Conference on Computer Vision*, 130–148. Springer.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022a. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; and Xue, Y. 2022b. Swin Transformer Embedding UNet for Remote Sensing Image Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.
- Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019a. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.
- Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019b. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2790–2799. PMLR.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Hu, L.; Lu, W.; Yu, H.; Yin, D.; Sun, X.; and Fu, K. 2024. TEA: A training-efficient adapting framework for tuning foundation models in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.
- Iizuka, R.; Xia, J.; and Yokoya, N. 2023. Frequency-Based Optimal Style Mix for Domain Generalization in Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.
- Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.
- Liang, C.; Li, W.; Dong, Y.; and Fu, W. 2024. Single Domain Generalization Method for Remote Sensing Image Segmentation via Category Consistency on Domain Randomization. *IEEE Transactions on Geoscience and Remote Sensing*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

- Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.
- Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- Luo, G.; Yang, X.; Dou, W.; Wang, Z.; Dai, J.; Qiao, Y.; and Zhu, X. 2024. Mono-intervl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*.
- Markus Gerke, I. 2014. Use of the stair vision library within the ISPRS 2D semantic labeling benchmark (Vaihingen). *Use of the stair vision library within the isprs 2d semantic labeling benchmark (vaihingen)*.
- Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. 2022. Simple open-vocabulary object detection. In *European conference on computer vision*, 728–755. Springer.
- Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.
- Pu, X.; and Xu, F. 2025. Low-Rank Adaption on Transformer-based Oriented Object Detector for Satellite Onboard Processing of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmlR.
- Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. 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.
- Rui, Z.; and Jintao, L. 2020. A survey on algorithm research of scene parsing based on deep learning. *Journal of Computer Research and Development*, 57(4): 859–875.
- Shelhamer, E.; Long, J.; and Darrell, T. 2016. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4): 640–651.
- Sun, J.; Ibrahim, M.; Hall, M.; Evtimov, I.; Mao, Z. M.; Ferrer, C. C.; and Hazirbas, C. 2023. VPA: Fully Test-Time Visual Prompt Adaptation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5796–5806.
- Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; and Zhang, L. 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237: 111322.
- Tranheden, W.; Olsson, V.; Pinto, J.; and Svensson, L. 2021. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1379–1389.
- Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. 2024. MTP: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*.
- Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.-S.; and Bai, X. 2019. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.
- Wei, Z.; Chen, L.; Jin, Y.; Ma, X.; Liu, T.; Ling, P.; Wang, B.; Chen, H.; and Zheng, J. 2024. Stronger Fewer & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 28619–28630.
- Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A. J.; Hanna, J.; Borth, D.; Papoutsis, I.; Le Saux, B.; Camps-Valls, G.; and Zhu, X. X. 2024. Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv e-prints*, arXiv-2403.
- Yin, D.; Yang, Y.; Wang, Z.; Yu, H.; Wei, K.; and Sun, X. 2023. 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.
- Zareian, A.; Rosa, K. D.; Hu, D. H.; and Chang, S.-F. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14393–14402.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.
- Zhu, J.; Guo, Y.; Sun, G.; Yang, L.; Deng, M.; and Chen, J. 2023. Unsupervised domain adaptation semantic segmentation of high-resolution remote sensing imagery with invariant domain-level prototype memory. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–18.