CrossEarth-Gate: Fisher-Guided Adaptive Tuning Engine for Efficient Adaptation of Cross-Domain Remote Sensing Semantic Segmentation

Appendix

Contents

A. Overview

This Appendix provides additional details, in-depth analysis, and implementation specifics. In Sec. B, we elaborate on the trade-offs regarding storage and computational costs, clarify the motivation behind the unified RS module toolbox, and discuss potential avenues for future generalization. In Sec. C, we present a dynamic network analysis to visualize the explainability of the Fisher-guided mechanism, followed by complete ablation studies across various backbones, module complementarity analysis, computational analysis, and extensive qualitative comparisons of segmentation results. In Sec. D, we provide rigorous descriptions of the cross-domain benchmarks (CASID, ISPRS, LoveDA, RescueNet), detailed implementation specifications for the eight PEFT baselines, and the exact hyperparameter configurations for CrossEarth-Gate.

B. Discussion

B.1. Storage and Computing

The Fisher-guided adaptive selection mechanism of CrossEarth-Gate dynamically activates specific, lightweight modules from the Remote Sensing (RS) module toolbox during training to optimally address the multifaceted challenges, *i.e.*, spatial, semantic, and frequency shifts, prevalent in RS cross-domain adaptation. As our experiments demonstrate, this selective guidance of the gradient flow not only enhances parameter efficiency but also yields superior adaptation performance by preventing the overfitting seen in full fine-tuning and the performance degradation from naively training all modules at once.

However, this dynamic framework introduces a noteworthy consideration regarding the cumulative storage and computational costs. During training, the framework is highly efficient as it only activates and computes gradients for a subset of Top-k modules at any given iteration. However, as the adaptation progresses, the set of modules selected at different stages may expand. A "worst-case" scenario arises if the model, in its effort to adapt, eventually selects and updates all modules in the toolbox at least once. This would necessitate storing the parameters for the entire toolbox to constitute the final adapted model. Consequently, at inference time, the computational path would require passing the features through all of these activated modules. Fortunately, a core design principle of the RS module toolbox is its lightweight nature. Even in this hypothetical worst-case scenario, the total parameter overhead remains minimal. For our experiments, the cumulative size of the entire toolbox is approximately 14.4M. When contrasted with the 304.2M parameters of the frozen DINOv2-L [16] backbone, this overhead constitutes less than 5% of the backbone's size. This modest cost is an acceptable tradeoff for the significant and consistent performance gains and robust generalization achieved across 16 benchmarks.

Therefore, this opens avenues for future exploration of this tradeoff. For instance, a post-hoc pruning strategy could be implemented to permanently discard modules that received consistently low Fisher importance scores, thereby creating a more compact final model for inference. One could also investigate parameter-sharing techniques where a single, lightweight module (*e.g.*, one spatial adapter) is shared across multiple layers, with its activation still governed by the layer-specific Fisher-guided gate. This would further reduce the total storage footprint without sacrificing dynamic adaptation.

B.2. Toolbox Modules

A foundational observation of this paper is the failure of existing specialized Parameter-Efficient Fine-Tuning (PEFT) methods when applied to cross-domain RS adaptation. We identify that RS domain gaps are uniquely multifaceted, manifesting as a complex interplay of spatial shifts (*e.g.*, changes in object scale and structure), semantic shifts (*e.g.*, different class appearances), and frequency shifts (*e.g.*, textural noise or sensor-based artifacts). The core limitation of prior work is its "single-pathway" design. As illustrated in our analysis, methods like LoRA [7] (spatial), AdaptFormer [1] (semantic), and Earth-Adapter [9] (frequency) specialize in one functional pathway. This specialization is their critical weakness: each method may capture only one facet of the RS domain shifts while leaving others unaddressed. This results in the specific, catastrophic failures we observed, such as LoRA misclassifying high-frequency waves (a frequency-domain failure) or AdaptFormer shattering the spatial continuity of a road (a spatial-domain failure). This clear gap motivates the design of our RS module toolbox. Instead of proposing yet another specialized module, we establish a unified framework to comprehensively tackle these intertwined challenges by integrating modules that target each of the three distinct functional pathways. In this work, we instantiate the toolbox using representative PEFTs with spatial, semantic, and frequency models.

We acknowledge that this implementation, while proven effective, is a first step. The challenges in RS are vast, and other types of domain shifts (*e.g.*, temporal, atmospheric, or sensor-specific variations) may exist that are not fully captured by our current spatial-semantic-frequency-trichotomy. This provides two clear avenues for future work. Future research could identify and model these other potential domain gaps, leading to the development and integration of new module types into the toolbox. Moreover, the specific methods chosen to implement each module (LoRA, Adapter, Earth-Adapter) could be

substituted. One could explore other reparameterization PEFTs for the spatial module or different additive or prompt-based methods for the semantic module, potentially offering different performance-efficiency tradeoffs.

B.3. Generalization

In this paper, we have comprehensively validated the effectiveness of CrossEarth-Gate on the task of cross-domain semantic segmentation in RS. A key strength demonstrated in our ablations is the framework's versatility across several State-Of-The-Art (SOTA) Transformer-based [20] foundation models, including the DINOv2 series [16], SAM [11], SatMAE [2], and Scale-MAE [18]. Our method consistently outperforms baselines, confirming its robustness within this architectural class. However, we acknowledge two primary avenues for exploring the broader generalizability of our approach. Our current validation is confined to Transformer-based backbones. While these represent the SOTA, a significant direction for future work is to investigate the applicability of CrossEarth-Gate to other foundational architectures, such as modern Convolutional Neural Networks. This would test whether our core methodology is a universal adaptation strategy or one uniquely suited to the functional pathways (*e.g.*, Multi-Head Self-Attention (MSA) and Multi-Layer Perception (MLP) layers) of Transformers. Moreover, the multifaceted domain gaps we identify (spatial, semantic, and frequency) are fundamental challenges in RS, extending far beyond semantic segmentation. Our experiments focus on this single task. A valuable and logical next step is to apply and evaluate the CrossEarth-Gate framework on other critical cross-domain RS tasks, such as object detection, change detection, or land-cover classification. This would provide a more rigorous test of our framework's robustness and further confirm the hypothesis that dynamically addressing these three-domain shifts is essential for effective RS adaptation.

C. Additional Studies

C.1. Dynamic Network Analysis

To further investigate the explainability of our proposed Fisher-guided adaptive selection mechanism, we conduct an in-depth dynamic network analysis of the Temperate Monsoon (Tem) domain generalization experiment, as illustrated in Fig. 1a and 1c. Specifically, we record the evolution of module selection of each module type (Spatial, Semantic, and Frequency) across different layers of the DINOv2-L [16] backbone (visualized via bubble size and distribution) and aggregated importance intensity of different module types across layers (visualized via the heatmap). Furthermore, we provide another intensity heatmap for the Subtropical Monsoon (Sub) domain in the Fig. 1b.

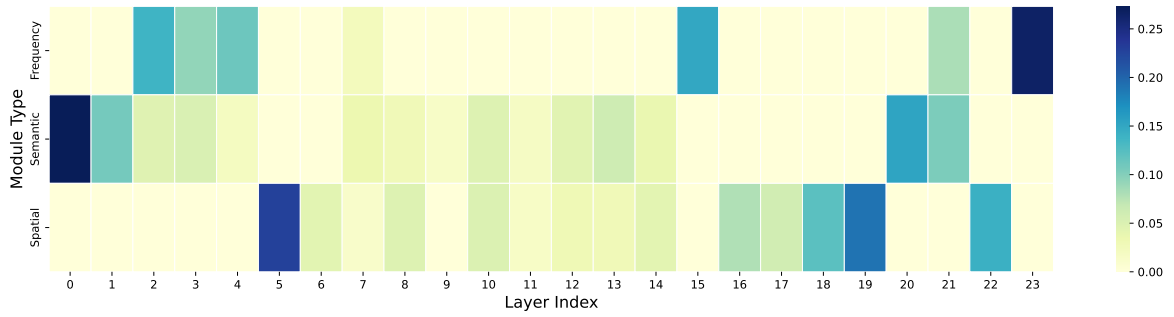
C.1.1. Layer-wise Module Distribution

As shown in Fig. 1a, the aggregated importance heatmap reveals a non-uniform, layer-dependent activation pattern. This hierarchy suggests that CrossEarth-Gate effectively decomposes the complex domain shift into distinct sub-problems, addressing them at the most appropriate network depths:

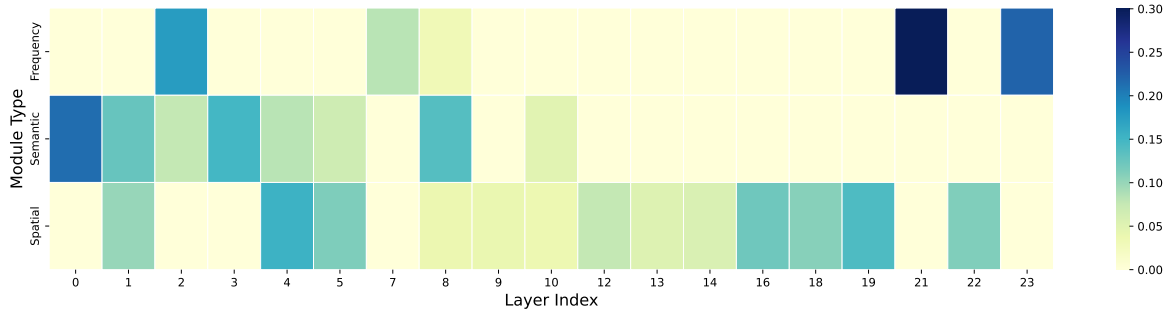
- **Shallow Layers:** In the shallow layers of the network, the model predominantly activates Semantic Modules. This indicates that during the initial stages of feature extraction, the adaptation process focuses on realigning fundamental semantic concepts such as the variations in vegetation tonality caused by climatic differences. The efficient feature projection capabilities of Adapters play a critical role in this low-level semantic alignment.
- **Middle Layers:** As information propagates to the middle layers, the selection shifts towards a hybrid configuration dominated by Spatial Modules. This suggests that after modifying semantics, the model focuses on adjusting to geometric shifts. Given that the intermediate layers of Transformer [20] architectures are typically responsible for modeling long-range dependencies and spatial context, the activation of LoRA implies a targeted fine-tuning of the self-attention mechanism. This effectively mitigates spatial shifts inherent to the Tem domain, such as variations in object scale and geometric layout.
- **Deep Layers:** A notable observation lies in the deep layers, where Frequency Modules exhibit a commanding presence. While deep features encapsulate highly abstract structural information, they are also most susceptible to frequency-domain artifacts arising from varying imaging conditions, such as spectral noise and textural distortion. The adaptive activation of frequency modules suggests the model possesses an implicit "awareness" of the necessity for spectral denoising and detail refinement prior to generating the final prediction.

C.1.2. Domain-Specific Activation Patterns

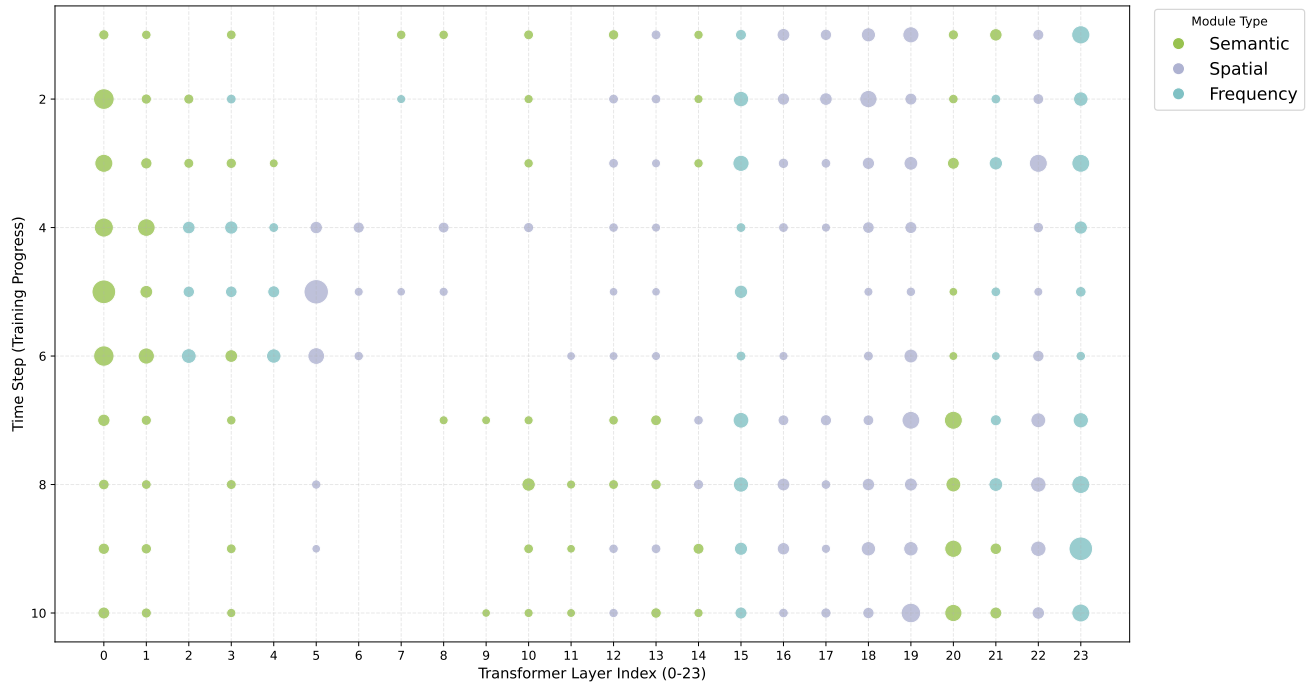
As shown in Fig. 1a and 1b, the Sub triggers a higher activation of Spatial modules and fewer Semantic modules compared to the Tem domain. This distinct activation pattern aligns with the specific characteristics of the Sub shift. This comparative analysis further substantiates that CrossEarth-Gate does not rely on a static or uniform adaptation strategy. Instead, it dynamically tailors its module topology to the unique, multifaceted requirements of each specific domain shift.



(a) Aggregated importance intensity of different module types across layers of Tem.



(b) Aggregated importance intensity of different module types across layers of Sub.



(c) The evolution of module selection and importance during training of Tem.

Figure 1. Dynamic network analysis. (a) The vertical axis represents training steps (flowing downwards), and the bubble size corresponds to importance score on the Tem. (b) Aggregated importance intensity of different module types across layers in the entire process on the Tem. (C) Aggregated importance intensity of different module types across layers in the entire process on the Sub.

C.1.3. The Evolution of Training Dynamics

The temporal evolution of module selection, as visualized in the Fig. 1c, demonstrates that CrossEarth-Gate operates as a dynamic system rather than a static ensemble. Early in the training phase, we observe a broader distribution of active modules with fluctuating Fisher scores, indicating an "exploration" phase where the gradient flow probes various pathways to identify the most effective adaptation routes. As training progresses, the selection stabilizes into the distinct functional zones described above. This temporal behavior confirms that the Fisher-guided mechanism successfully acts as a "gating" controller, as it does not merely select parameters randomly but progressively converges on a resource-efficient configuration that maximizes the reduction of task-specific loss. This ability to "learn how to adapt" ensures that the model's capacity is not wasted on redundant layers, but is instead dynamically allocated to where the domain gap is most severe.

The training patterns observed on Tem demonstrate that CrossEarth-Gate possesses both parameter efficiency and explainable adaptive capability. Rather than relying on a static, specialized fine-tuning paradigm, it employs a dynamic resource allocation strategy dictated by the specific hierarchical requirements of the domain shift. This adaptive, layer-wise governance is the core factor enabling CrossEarth-Gate to achieve SOTA performance in complex cross-domain tasks.

C.1.4. Qualitative Gradient Saliency

To further investigate the interpretability of our framework and explicitly link module activations to the learned visual representations, we compute gradient saliency maps for each module type with respect to the input pixels. As illustrated in Fig. 2, these visualizations confirm that the Spatial, Semantic, and Frequency modules dynamically attend to distinct, functionally relevant regions of the image, corroborating our multi-faceted adaptation hypothesis. Specifically, the Frequency modules exhibit high activation around high-frequency image components. Conversely, the Spatial modules concentrate their gradients on prominent geometric structures. Meanwhile, the Semantic modules target broader, class-specific regions characterized by distinct visual appearances.

Figure 2. Visualization of gradient saliency maps for the Frequency, Semantic, and Spatial modules.

C.2. Complete Ablation Studies

We provide the complete results of the ablation studies in Tab. 1 to further evaluate the generalizability of CrossEarth-Gate across diverse foundation model architectures and to rigorously dissect the contribution of each component within our framework. The experiments are conducted on the CASID [13] dataset, covering 12 domain generalization (DG) scenarios.

C.2.1. Generalizability Across Backbones

We evaluate CrossEarth-Gate against "Frozen" and "Full-Tuning" baselines across five distinct backbone architectures ranging from masked autoencoders [5] (SatMAE [2], Scale-MAE [18]) to segmentation-specific models (SAM [11]) and self-supervised vision transformers (DINOv2 series [16]). While it is intuitive to assume that updating all parameters would yield the highest adaptation capacity, our empirical results demonstrate a "Full-Tuning Paradox". As evidenced in the DINOv2-S and DINOv2-B experiments, Full-Tuning results in a precipitous performance decline compared to even the Frozen baseline (e.g., a drop of average performance from 59.6% to 49.3% average mIoU on DINOv2-S in the Tem source). We attribute this phenomenon to catastrophic forgetting, where unconstrained optimization on a smaller, domain-specific source dataset erodes the robust, generalizable feature representations acquired during large-scale pre-training. Furthermore, Full-Tuning appears highly susceptible to overfitting the specific noise patterns of the source domain, thereby hampering transferability to unseen target domains. In contrast, CrossEarth-Gate consistently surpasses the Frozen baseline while updating less than 2% of the parameters, effectively balancing plasticity (adaptation) and stability (generalization). Even on the SAM, which possesses a massive parameter count (632M) and strong zero-shot capabilities, CrossEarth-Gate improves the average mIoU by roughly 3-4% across tasks. This indicates that our Fisher-guided mechanism effectively locates and tunes the sparse "skill neurons" required for domain alignment, even in highly capable, frozen architectures.

Table 1. Complete ablation studies on CASID benchmarks across 12 DG experiments. We demonstrate the generalizability of CrossEarth-Gate across different backbones and compare the performance and trainable parameters of CrossEarth-Gate against versions with key components removed. Sub: Subtropical Monsoon. Tem: Temperate Monsoon. Tms: Tropical Monsoon. Trf: Tropical Rainforest.

Backbone	Method	Params (M)	Sub2Tem	Sub2Tms	Sub2Trf	Average	Tem2Sub	Tem2Tms	Tem2Trf	Average
SatMAE (Large) [2]	Frozen	0.0	22.4	24.6	26.1	24.4	22.3	18.9	**23.3**	21.5
	Full-Tuning	304.2	**26.7**	42.2	37.3	35.4	11.1	16.1	17.4	14.9
	CrossEarth-Gate	2.9-4.0	21.8	**46.8**	**42.9**	**37.1**	**29.2**	**21.6**	14.6	**21.8**
Scale-MAE (Large) [18]	Frozen	0.0	38.6	58.9	54.6	50.7	55.6	50.9	38.9	48.5
	Full-Tuning	304.2	38.7	64.0	55.8	52.8	62.7	59.2	**51.0**	57.6
	CrossEarth-Gate	2.9-3.6	**45.1**	**66.3**	**57.4**	**56.3**	**64.6**	**61.0**	48.7	**58.1**
SAM (Huge) [11]	Frozen	0.0	39.2	**64.0**	56.7	53.3	63.5	61.9	47.1	57.5
	Full-Tuning	631.2	**43.3**	60.6	60.2	54.7	63.1	60.2	51.2	58.2
	CrossEarth-Gate	4.0-4.9	42.9	63.7	**61.7**	**56.1**	**67.1**	**65.4**	**53.3**	**61.9**
DINOv2 (Small) [16]	Frozen	0.0	43.0	50.9	54.7	49.5	64.9	58.8	55.1	59.6
	Full-Tuning	22.1	33.9	**60.2**	54.2	49.4	56.5	49.6	41.9	49.3
	CrossEarth-Gate	0.7	**46.2**	58.4	**59.1**	**54.6**	**65.1**	**63.9**	**57.1**	**62.0**
DINOv2 (Base) [16]	Frozen	0.0	39.3	57.2	57.4	51.3	**67.2**	63.9	56.4	62.5
	Full-Tuning	86.6	36.3	59.3	54.8	50.2	58.6	54.9	47.5	53.7
	CrossEarth-Gate	1.4-1.8	**45.3**	**62.5**	**58.9**	**55.6**	**67.2**	**65.7**	**57.9**	**63.6**
DINOv2 (Large) [16]	Frozen	0.0	43.1	63.4	58.8	55.2	66.9	64.8	59.0	63.6
	Full-Tuning	304.2	38.6	62.3	58.3	53.0	58.7	55.8	43.8	52.8
	CrossEarth-Gate	3.0-4.4	50.1	**66.6**	**65.2**	**60.6**	68.0	**67.0**	**60.3**	**65.1**
	w/o Spatial	2.7-3.2	47.9	60.3	61.9	56.7	68.4	66.8	59.5	64.8
	w/o Semantic	4.2-4.5	51.3	**66.6**	60.2	59.4	**68.5**	66.2	57.9	64.2
	w/o Frequency	3.2-4.2	**51.4**	57.6	63.4	57.5	67.5	64.0	59.9	63.8
	w/o Selection	14.4	48.0	61.8	63.5	57.8	67.7	63.6	59.2	63.5
Backbone	Method	Params (M)	Tms2Sub	Tms2Tem	Tms2Trf	Average	Trf2Sub	Trf2Tem	Trf2Tms	Average
SatMAE (Large) [2]	Frozen	0.0	28.8	18.5	27.0	24.8	33.1	**22.0**	23.7	26.3
	Full-Tuning	304.2	42.0	**24.6**	39.1	35.2	37.7	20.0	**31.8**	29.8
	CrossEarth-Gate	2.9-4.0	**45.2**	24.1	**41.0**	**36.8**	**45.1**	18.9	31.2	**31.8**
Scale-MAE (Large) [18]	Frozen	0.0	64.8	31.5	56.4	50.9	59.4	34.5	58.8	50.9
	Full-Tuning	304.2	64.3	33.0	58.5	51.9	56.6	**42.8**	58.6	52.6
	CrossEarth-Gate	2.9-3.6	**67.3**	**37.6**	**59.6**	**54.9**	**62.4**	36.3	**62.0**	**53.8**
SAM (Huge) [11]	Frozen	0.0	64.9	37.7	57.2	53.2	63.3	34.2	62.6	53.4
	Full-Tuning	631.2	64.6	33.2	57.4	51.7	63.9	36.9	62.9	54.6
	CrossEarth-Gate	4.0-4.9	**65.8**	**43.4**	**59.9**	**56.3**	**67.2**	**43.3**	**67.3**	**59.3**
DINOv2 (Small) [16]	Frozen	0.0	63.1	**37.2**	55.5	51.9	63.5	33.2	61.5	52.7
	Full-Tuning	22.1	60.5	29.4	56.4	48.7	58.9	32.3	59.5	50.2
	CrossEarth-Gate	0.7	**64.6**	35.0	**58.6**	**52.7**	**64.8**	**38.1**	**64.6**	**55.8**
DINOv2 (Base) [16]	Frozen	0.0	64.8	36.1	57.4	52.8	**65.1**	40.2	58.0	54.4
	Full-Tuning	86.6	62.2	34.1	54.3	50.2	58.5	29.1	60.6	49.4
	CrossEarth-Gate	1.4-1.8	**65.0**	**38.0**	**61.2**	**54.7**	64.5	**45.6**	**62.3**	**57.4**
DINOv2 (Large) [16]	Frozen	0.0	65.3	37.0	60.9	54.4	68.0	43.0	66.7	59.2
	Full-Tuning	304.2	60.9	29.9	58.9	49.9	61.9	32.9	64.3	53.0
	CrossEarth-Gate	3.0-4.4	68.3	46.3	**63.3**	**59.3**	69.0	50.0	68.1	**62.4**
	w/o Spatial	2.7-3.2	**70.0**	45.3	60.5	58.6	68.8	49.3	67.9	61.9
	w/o Semantic	4.2-4.5	68.3	43.1	61.6	57.6	68.9	44.7	67.6	60.4
	w/o Frequency	3.2-4.2	67.7	**47.4**	62.9	**59.3**	**69.1**	45.5	**68.3**	60.9
	w/o Selection	14.4	69.5	41.2	61.2	57.3	68.3	**51.0**	63.9	61.1

C.2.2. Component Dissection

The ablation study on the DINOv2-Large backbone validates the necessity of our unified toolbox and selection mechanism. A critical observation is that the w/o Selection variant (activating all modules simultaneously) consistently underperforms the full CrossEarth-Gate. For example, in the Tem source experiments, removing selection drops the average mIoU from 65.1% to

63.5%. This empirically indicates that naively activating all adaptation pathways introduces gradient conflict and noise. The Fisher-guided selection acts as a necessary gate, ensuring that spatial, semantic, and frequency updates do not interfere with one another. The removal of specific modules (w/o Spatial/Semantic/Frequency) leads to performance degradation, though the impact varies by domain. Notably, the w/o spatial variant causes the most significant drop in the Sub source scenarios. This suggests that the shift from Subtropical Monsoon to other zones involves significant spectral and phenological changes (e.g., vegetation density, lighting conditions) but retains relatively consistent spatial layouts and structural geometries.

C.2.3. Analysis of Edge Cases and Limitations

While CrossEarth-Gate achieves state-of-the-art performance in the majority of scenarios, we observe isolated cases where baselines or ablated versions show marginal superiority. For example, in the benchmark of Tms2Sub using Dinov2-L, the variant without Spatial modules achieves a higher score (70.0%) than the full method (68.3%). The shift from Tropical Monsoon to Subtropical Monsoon involves climate zones with relatively similar object scales and geometric structures. Consequently, the domain gap may be predominantly textural (frequency) rather than geometric (spatial). In this context, the introduction of LoRA modules might have introduced unnecessary inductive bias, slightly disrupting the spatial integrity of the pre-trained features. As for Trf2Tem, the variant activating all modules performs best (51.0%). This is a scenario with an extreme domain shift. It is possible that the domain gap is so multifaceted and severe that the model requires every available degree of freedom to adapt, outweighing the negative impact of gradient conflict. However, this comes at the cost of tripling the trainable parameters, which reduces the method's efficiency. Therefore, despite these isolated outliers, the CrossEarth-Gate framework remains the superior strategy, as it provides the requisite plasticity to address multifaceted shifts on average, ensuring global robustness without requiring manual, domain-specific architecture tuning.

C.3. Module Complementarity and Synergy Analysis

Regarding module complementarity, we further conduct a granular decomposition ablation on the Sub domain in the Tab 2. The combination of all three modules outperforms single-module baselines, confirming that they address complementary domain shifts. However, the gain from naive combination is marginal, which implies that without explicit gating, "feature interference" suppresses the full potential of the toolbox. Randomly selecting modules helps mitigate interference slightly and improves performance. However, the Fisher-Guided mechanism is the key to unlocking true synergy, which boosts the performance significantly. By activating only the most information-rich pathway, it facilitates the toolbox modules to capture complementary aspects of the domain shift.

Table 2. Granular decomposition ablation study on the Sub domain.

Spatial	Semantic	Frequency	Random	Fisher	Tem	Tms	Trf	Avg.
✓					48.7	60.5	62.3	57.2
	✓				47.9	58.4	62.1	56.1
		✓			47.8	61.1	59.2	56.0
✓	✓				49.3	60.4	62.8	57.5
	✓	✓			48.5	61.0	62.2	57.2
✓		✓			48.6	61.7	62.4	57.6
✓	✓	✓			48.0	61.8	63.5	57.8
✓	✓	✓	✓		48.6	64.3	63.1	58.7
✓	✓	✓		✓	50.1	66.6	65.2	60.6

C.4. Additional Computational Analysis

As detailed in Appendix D, the selection process is performed 10 times during the 30,000 training iterations (update interval $N = 3000$, calculation samples $M = 100$) for CASID benchmarks. As shown in the Tab. 3, we quantify the computation costs using the CASID [13] benchmarks on the Sub souce domain. The total time consumed by the selection process constitutes only $\sim$3.7% of the training time. For practical deployment, the Fisher-guided selection is exclusively a training-time optimization. Once training is complete, the module topology is frozen as other PEFT methods do. Our inference throughput and computational cost are highly competitive with static PEFT baselines.

Table 3. Comparison of efficiency metrics including parameters, training time, memory usage, FLOPs, and throughput on the Sub domain.

Method	Params (M)	Training		Inference	
		Time (min)	Mem (GB)	GFLOPs	Throughput (img/s)
Selection	14.4	16	16.1	N/A	N/A
Ours	3.0–4.4	437	12.8	1258	2.12
LoRA	6.4	441	12.3	1242	2.18
AdaptFormer	3.2	420	12.1	1255	2.14
Earth-Adapter	9.6	470	15.2	1281	1.83
Full-tuning	304.2	617	18.2	1242	2.18

C.5. Additional Qualitative Comparison

In this section, we provide a more comprehensive qualitative analysis of the segmentation results. We present visualizations across diverse benchmarks, including the cross-climate scenarios of CASID (Sub2Tms, Tem2Trf, Tms2Sub, Trf2Tem), the disaster adaptation scenarios of RescueNet (P(r)2Res, P(i)2Res), and the domain adaptation benchmarks (P2V, V2P, R2U, U2R). The core motivation of CrossEarth-Gate is to address the limitations of existing specialized PEFT methods, which typically focus on a single functional pathway (spatial, semantic, or frequency).

The CASID benchmarks (Fig. 3 - 6) highlight the acute limitations of single-pathway adaptation when facing simultaneous spectral and geometric shifts. For example, in the Sub2Tms scenarios, the domain shift involves significant changes in vegetation appearance and water surface texture (frequency artifacts). Methods like AdaptFormer and LoRA, which specialize in semantic features or spatial, struggle to distinguish between large water bodies and forests when high-frequency spectral noise is present. This results in substantial "hallucinations" of land over water. The Tem2Trf benchmarks introduce complex geometric shifts in road networks. As seen in Fig. 4, methods lacking strong spatial reasoning, such as AdaptFormer and Earth-Adapter, frequently produce fragmented road predictions, failing to maintain connectivity. As shown in Fig. 7 and 8, the transition from standard aerial imagery (Potsdam) to post-disaster scenes (RescueNet) presents a severe semantic challenge. While LoRA may preserve object boundaries, it misinterprets the texture of disaster debris, misclassifying "clutter" (red) as "vegetation" (green) in the P(i)2Res benchmark (Fig. 8). This represents a critical semantic error where the spatial adaptation is insufficient to correct the conceptual drift. The results for Domain Adaptation (DA) tasks (P2V, V2P, R2U, U2R), visualized in Fig. 9 - 12, further validate our approach in scenarios with access to unlabeled target data. In the transition between urban and rural domains (e.g., U2R in Fig. 12), existing methods struggle to reconcile the scale differences between dense urban clusters and sparse rural features. For instance, AdaptFormer struggles to maintain the structural integrity of roads, while Earth-Adapter fails to resolve the boundaries of dense, small-scale forest regions. These failures indicate that a fixed adaptation strategy cannot dynamically prioritize the spatial refinement needed for scale variations or the semantic tuning needed for layout changes.

In contrast to these baselines, CrossEarth-Gate effectively navigates these trade-offs. By leveraging Fisher Information to quantify the importance of each module dynamically, our engine activates the appropriate pathway for the specific shift at hand. Consequently, our method consistently produces segmentation maps that are closer to the ground-truth labels across all scenarios, demonstrating a robust and superior generalization capability.

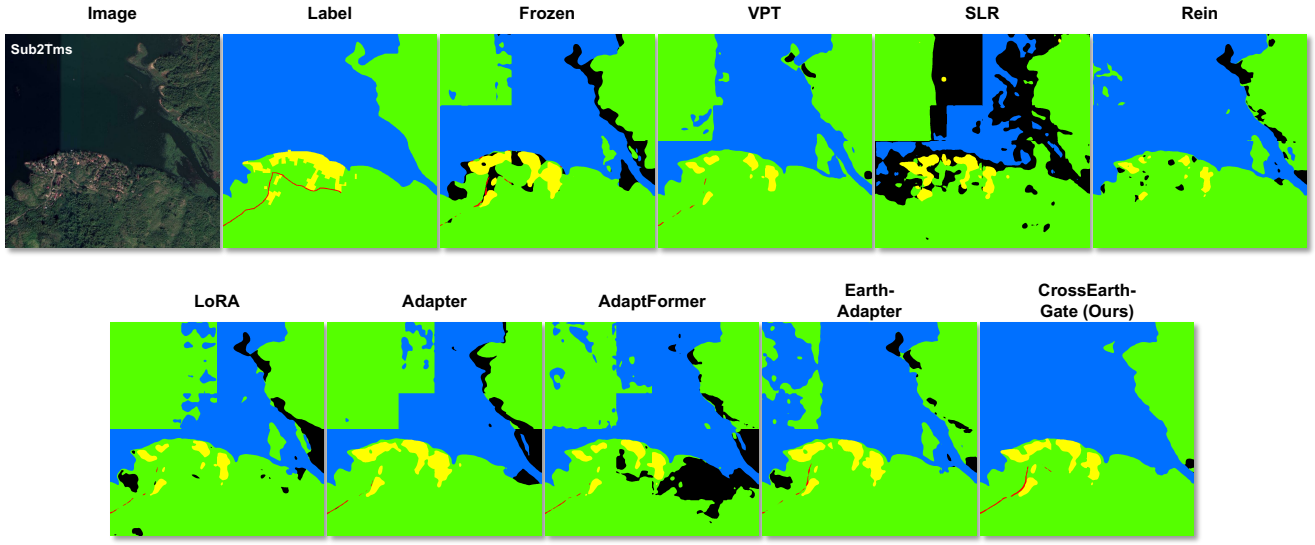


Figure 3. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of Sub2Tms on the CASID [13] dataset, where red is the road class, yellow is the building class, blue is the water class, green is the forest class, and black is the background class.

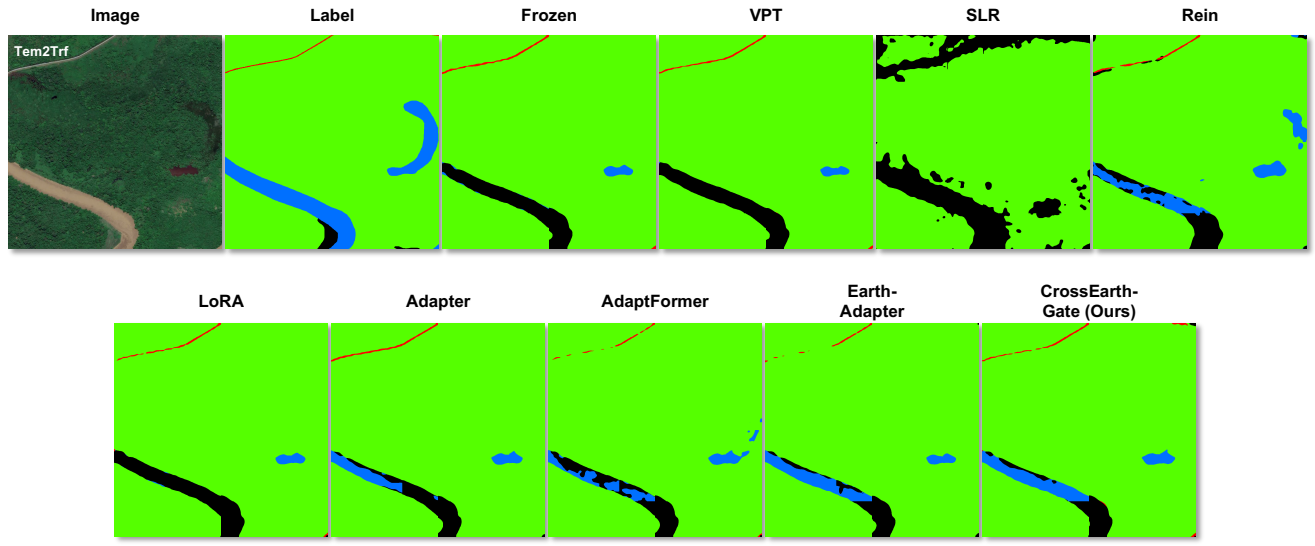


Figure 4. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of Tem2Trf on the CASID [13] dataset, where red is the road class, yellow is the building class, blue is the water class, green is the forest class, and black is the background class.

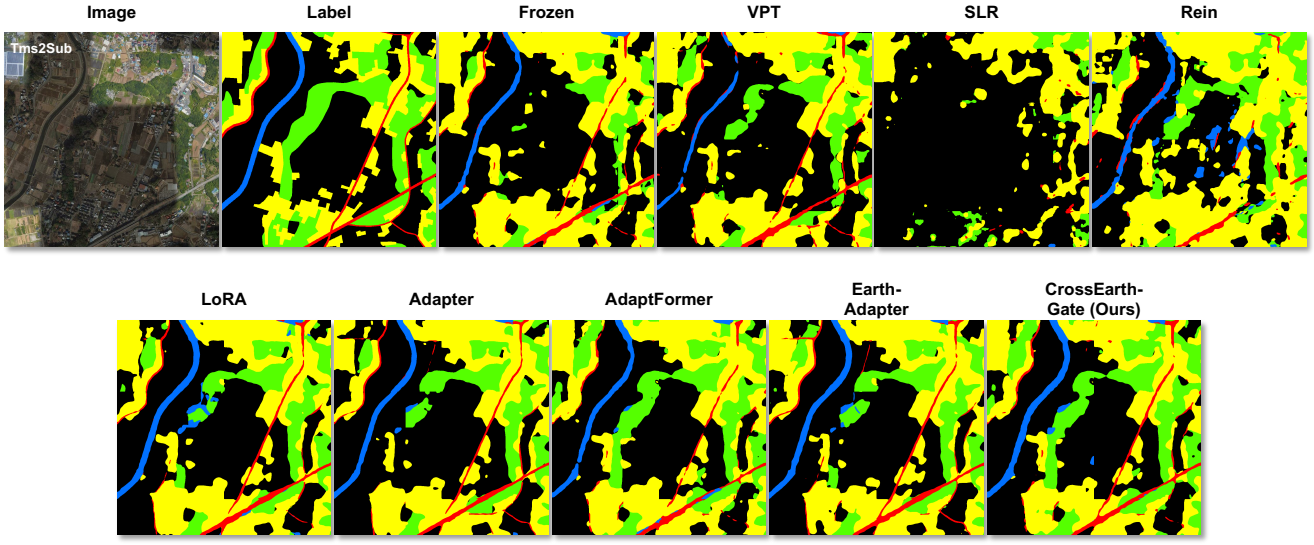


Figure 5. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of Tms2Sub on the CASID [13] dataset, where red is the road class, yellow is the building class, blue is the water class, green is the forest class, and black is the background class.

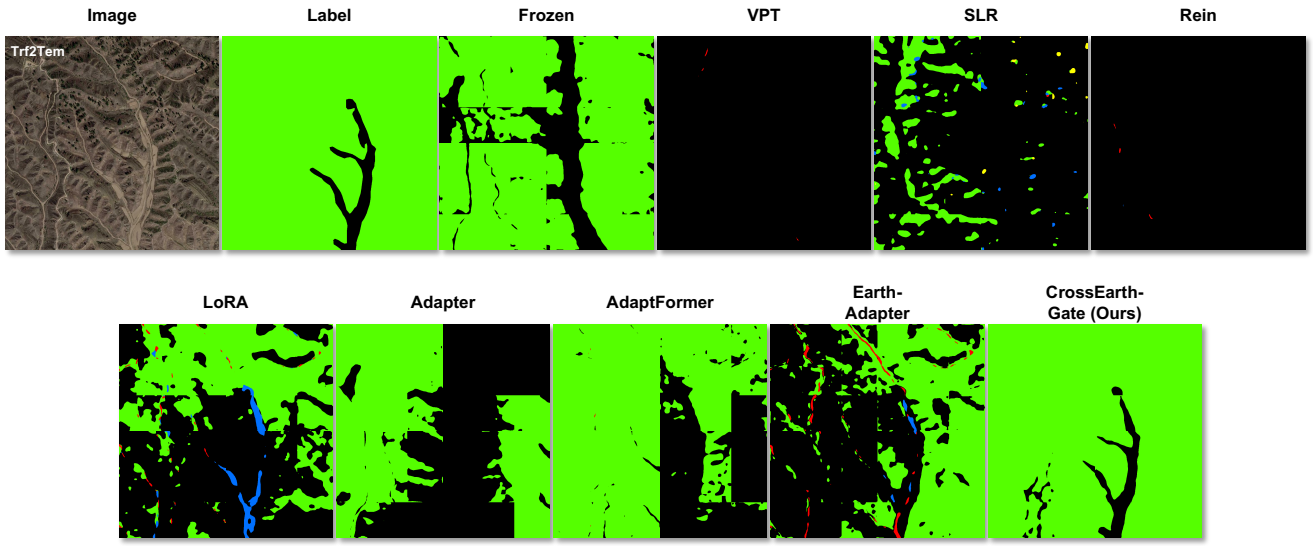


Figure 6. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of Trf2Tem on the CASID [13] dataset, where red is the road class, yellow is the building class, blue is the water class, green is the forest class, and black is the background class.

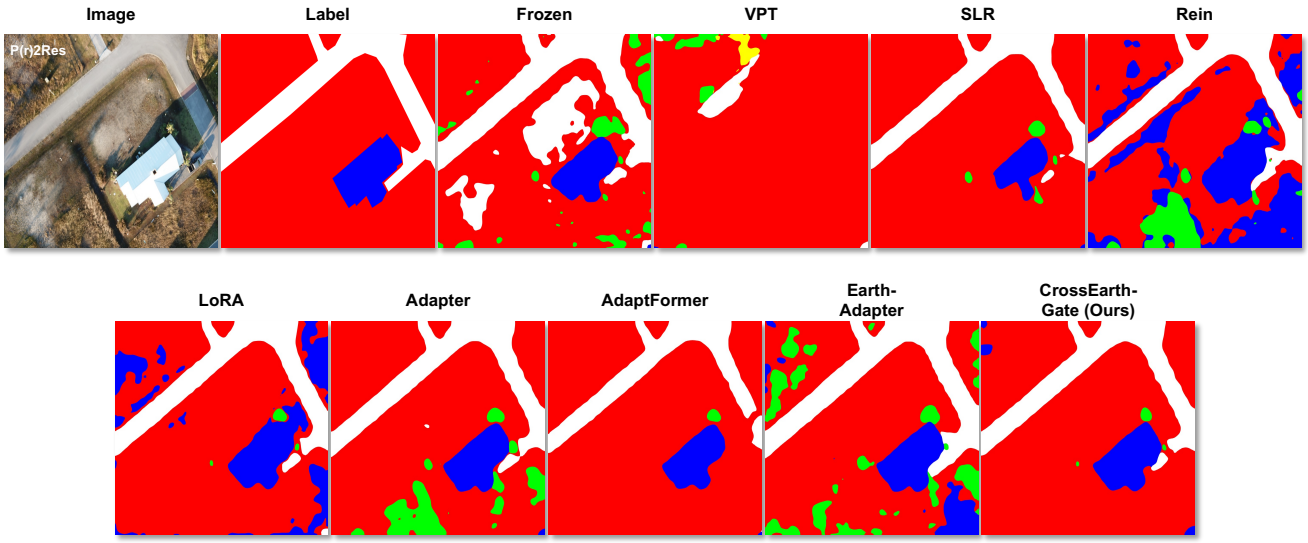


Figure 7. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of P(r)2Res on the the RescueNet [17] dataset, where white is the impervious surface class, red is the clutter class, blue is the building class, green is the vegetation class, and yellow is the car class.

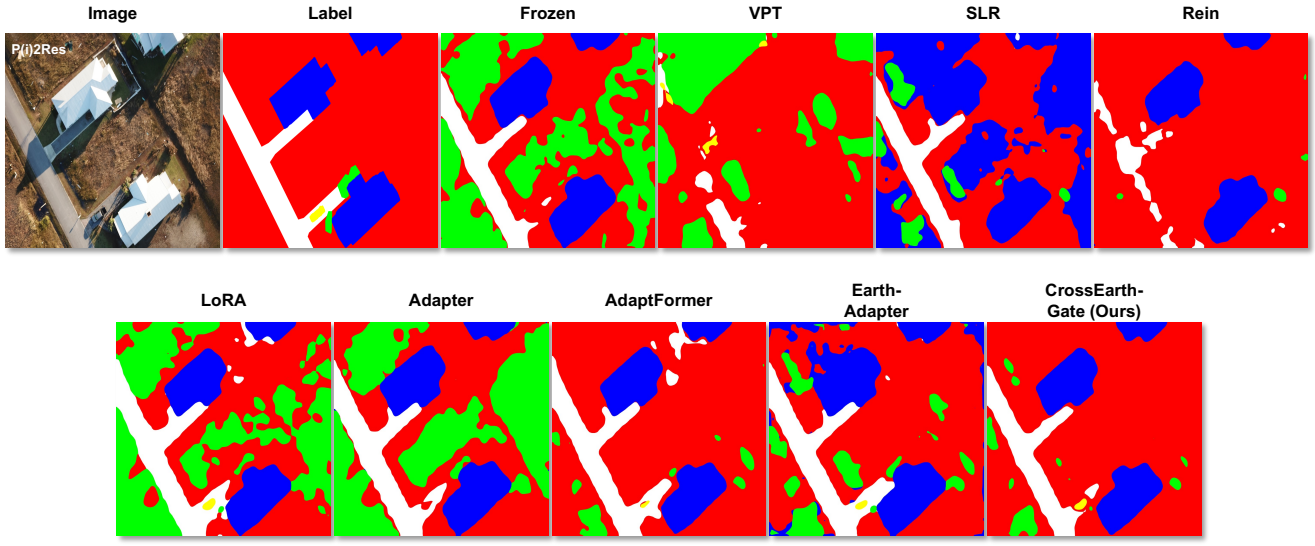


Figure 8. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of P(i)2Res on the the RescueNet [17] dataset, where white is the impervious surface class, red is the clutter class, blue is the building class, green is the vegetation class, and yellow is the car class.

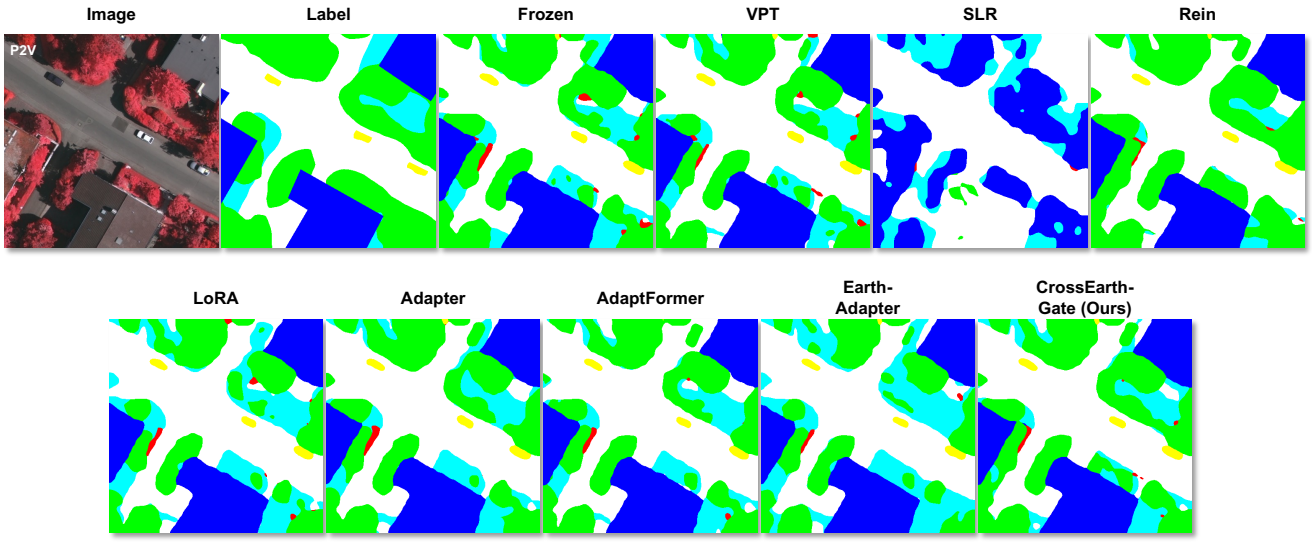


Figure 9. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of P2V on the the Vaihingen dataset, where white is the impervious surface class, red is the clutter class, blue is the building class, cyan is the low vegetation, green is the tree class, and yellow is the car class.

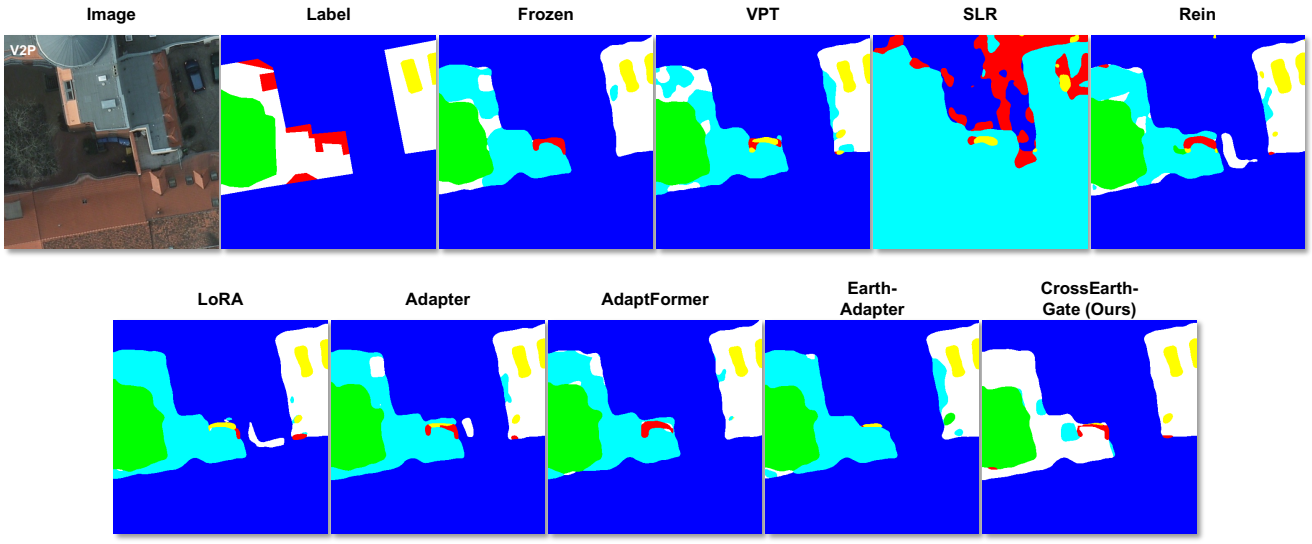


Figure 10. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of P2V on the the Potsdam dataset, where white is the impervious surface class, red is the clutter class, blue is the building class, cyan is the low vegetation, green is the tree class, and yellow is the car class.

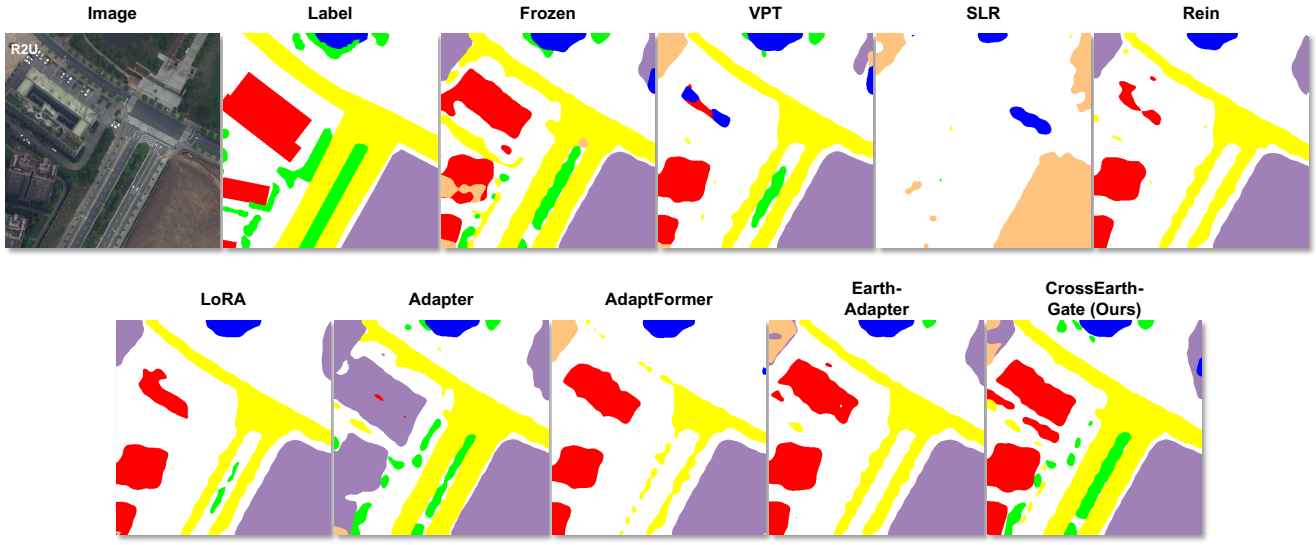


Figure 11. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of R2U on the the LoveDA [21] dataset, where white is the background class, red is the building class, yellow is the road class, blue is the water class, purple is the barren class, green is the forest class, and brown is the agriculture class.

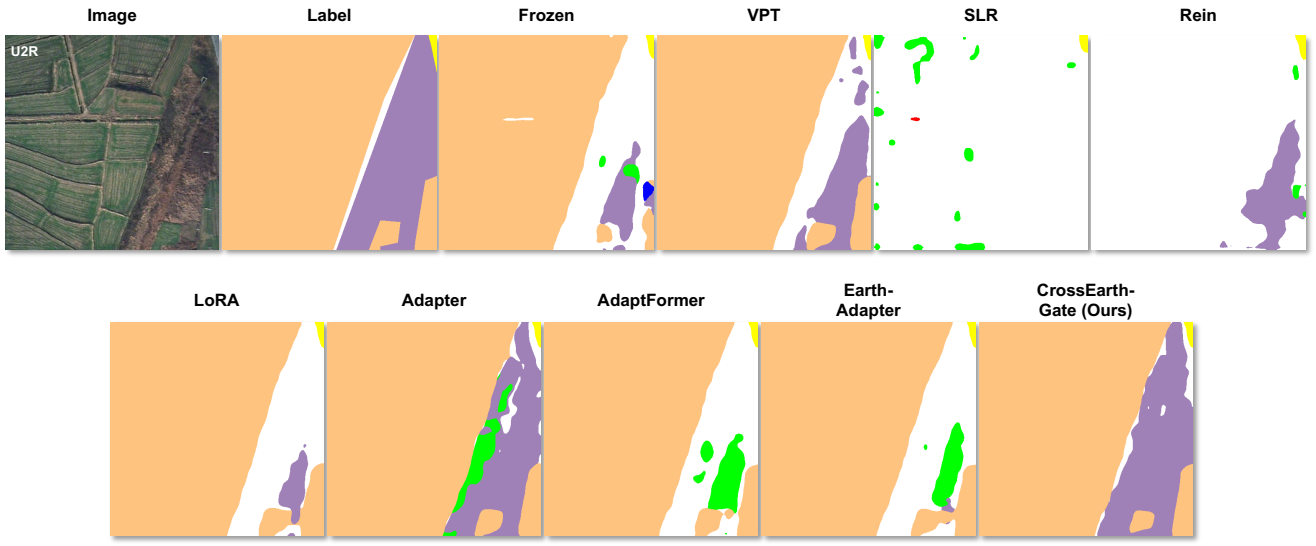


Figure 12. Complete visualizations of predicted segmentation maps of PEFT methods. These samples are collected from the domain generalization benchmarks of U2R on the the LoveDA [21] dataset, where white is the background class, red is the building class, yellow is the road class, blue is the water class, purple is the barren class, green is the forest class, and brown is the agriculture class.

Table 4. Detailed statistics and configuration of the cross-domain remote sensing benchmarks, including Domain Adaptation (DA) and Domain Generalization (DG). We categorize the experimental settings by the type of domain gap (Unseen Region, Spectral Band, and Climate) and provide specific details regarding the source and target domains, data splits (train/test numbers), image resolutions, and category counts for LoveDA [21], Potsdam, Vaihingen, RescueNet [17], and CASID [13] datasets.

Domain Gap	Benchmark	Dataset	Source Domain	Target Domain	Abbreviation	Train Number	Test Number	Image Size	Categories
Unseen Region	DA	LoveDA[21]	LoveDA-Urban	LoveDA-Rural	U2R	5464	3968	512×512	7
			LoveDA-Rural	LoveDA-Urban	R2U	5464	2708		7
	DG	Potsdam, RescueNet [17] (Disaster Assessment)	Potsdam (RGB)	RescueNet (RGB)	P(r) 2 Res	3456	449	512×512	5
Unseen Region and Spectral Band	DA	Potsdam, Vaihingen	Potsdam (RGB)	Vaihingen (IRRG)	P2V	3456	398	512×512	6
			Vaihingen (IRRG)	Potsdam (RGB)	V2P	3456	2016		6
	DG	Potsdam, RescueNet [17] (Disaster Assessment)	Potsdam (IRRG)	RescueNet (RGB)	P(i) 2 Res	3456	449	512×512	5
Unseen Region and Climate	DG	CASID [13]	Subtropical Monsoon	Temperate Monsoon	Sub2Tem	4900	2075	1024×1024	5
				Tropical Monsoon	Sub2Tms		1650		5
				Tropical Rainforest	Sub2Trf		1550		5
			Temperate Monsoon	Subtropical Monsoon	Tem2Sub	5025	2200		5
				Tropical Monsoon	Tem2Tms		1650		5
				Tropical Rainforest	Tem2Trf		1550		5
			Tropical Monsoon	Subtropical Monsoon	Tms2Sub	3400	2200		5
				Temperate Monsoon	Tms2Tem		2075		5
				Tropical Rainforest	Tms2Trf		1550		5
			Tropical Rainforest	Subtropical Monsoon	Trf2Sub	3700	2200		5
				Temperate Monsoon	Trf2Tem		2075		5
				Tropical Monsoon	Trf2Tms		1650		5

D. Additional Experimental Details

In this section, we provide a comprehensive breakdown of the datasets and experiments utilized to benchmark CrossEarth-Gate. To ensure a rigorous evaluation of the multifaceted domain shifts inherent to remote sensing, we select datasets representing distinct challenges: climate variation, spectral band discrepancies, urban-rural transitions, and post-disaster alterations. The detailed statistics and configurations for each benchmark, including image resolution, sample sizes, and semantic categories, are summarized in Table 4. All models are fine-tuned using the AdamW [14] optimizer, employing a weight decay of 0.05 and betas as 0.9 and 0.999. We employ the data augmentation techniques following [9, 23]. The parameter fluctuation in the all table represents the range of the active trainable parameter set.

D.1. Dataset Introduction

D.1.1. CASID

The Climate-Aware Satellite Image Dataset (CASID) [13] represents a pioneering advancement in the field of RS. It is recognized as the first dataset explicitly engineered to mitigate domain shift challenges stemming from climatic diversity. By systematically capturing distinct environmental variances, CASID enables the development of models that are robust across shifting geographic and meteorological domains. The dataset is stratified across four primary climate zones: Subtropical Monsoon (Sub), Temperate Monsoon (Tem), Tropical Monsoon (Tms), and Tropical Rainforest (Trf). To facilitate precise land-cover segmentation, CASID provides pixel-wise annotations across five semantic categories. These classes capture the essential morphological and spectral features required for earth observation tasks, including Background, Building, Forest, Road, and Water. We divide the dataset into training and validation sets and crop images to 1024×1024 resolutions, following the data processing protocols of [4]. For the CASID DG experiment, we fine-tune the model for 30k iterations with a batch size of 1.

D.1.2. ISPRS Potsdam and Vaihingen

The ISPRS Potsdam and Vaihingen datasets are widely recognized as foundational benchmarks in the field of RS semantic segmentation. Curated by the International Society for Photogrammetry and Remote Sensing (ISPRS), these datasets consist of High Spatial Resolution (HSR) aerial imagery captured over their respective namesake cities in Germany. Due to their distinct environmental and spectral characteristics, these datasets have become the standard for evaluating Domain Adaptation (DA) techniques [9, 12, 15, 22, 24, 25]. They present a robust challenge for algorithms attempting to bridge the domain gap between differing urban landscapes and sensor configurations. The annotations cover six common land-cover categories: Impervious surfaces, Building, Low vegetation, Tree, Car, and Clutter. We utilize the RGB channels of Potsdam as the

source domain and the IR-R-G channels of Vaihingen as the target domain (P2V), and vice-versa (V2P) with a cropped size of 512 $\times$ 512, following [9]. For the DA experiment, we fine-tune the model for 20k iterations with a batch size of 2.

D.1.3. LoveDA

The LoveDA (Land-cOVEr Domain Adaptive semantic segmentation) [21] dataset is a large-scale HSR benchmark designed to advance both semantic segmentation and transferable learning. Unlike traditional datasets that focus solely on semantic representation, LoveDA explicitly addresses DA challenges by capturing the diverse stylistic and structural differences between developed and undeveloped geographic areas. The dataset comprises 5,987 HSR images collected from three different Chinese cities (Nanjing, Changzhou, and Wuhan). These images cover 536.15 km^2 at a spatial resolution of 0.3 meters. Specifically, LoveDA is uniquely structured around two distinct domains, which introduce significant challenges regarding multi-scale objects and inconsistent class distributions: Urban domain, characterized by high population density, neatly arranged buildings of various shapes, and wide roads, and Rural domain, characterized by natural elements, disordered building layouts, small-scale agricultural zones, and narrow roads. We follow the protocol of [9] to establish two DA settings: Urban-to-Rural (U2R) and Rural-to-Urban (R2U). The evaluation is performed across seven semantic categories: Background, Building, Road, Water, Barren, Forest, and Agriculture. For the DA experiment, we fine-tune the model for 20k iterations with a batch size of 2.

D.1.4. RescueNet

RescueNet [17] stands as a cutting-edge, high-resolution benchmark designed to advance the capabilities of Unmanned Aerial Vehicles (UAVs) in post-disaster scenarios. Unlike satellite-based datasets that often suffer from low resolution, RescueNet consists of low-altitude aerial imagery captured specifically after Hurricane Michael. The primary objective of RescueNet is to overcome the limitations of existing datasets by providing: (1) holistic scene understanding, which moves beyond simple building detection to provide pixel-level annotations for the entire scene, including infrastructure and natural elements; and (2) granular damage assessment, which introduces detailed severity classifications for buildings and roads, enabling rescue teams to prioritize efforts based on precise damage levels. RescueNet contains 11 categories, including Background, Water, Building-No-Damage, Building-Medium-Damage, Building-Major-Damage, Building-Total-Destruction, Vehicle, Road-Clear, Road-Blocked, Tree, and Pool. In our experiments, we utilize RescueNet solely as an unseen target domain, using the RGB (P(r)2Res) and IR-R-G (P(i)2Res)channels Potsdam dataset as the source, following the protocol from [4]. This setup assesses the model's ability to maintain semantic integrity when transferring from organized urban environments to chaotic, post-disaster scenes. The evaluation focuses on five shared classes: Impervious surfaces, Building, Tree, Car, and Clutter. For the DG experiment, we fine-tune the model for 30k iterations with a batch size of 1.

D.2. PEFT Baseline Implementation

We implement eight PEFT baseline methods to rigorously benchmark the performance of our proposed CrossEarth-Gate. The default hyperparameters for these methods, including rank configurations, prompt lengths, and scaling factors, are detailed in Table 5. Below, we briefly describe the implementation specifics for each baseline.

D.2.1. VPT

Visual Prompt Tuning (VPT) [10] is utilized to adapt the frozen foundation model by injecting a sequence of learnable parameters, known as soft prompts, directly into the input sequence. Instead of fine-tuning the backbone weights, VPT prepends these trainable tokens to the sequence of image patch embeddings. These prompts interact with the image features via the Transformer's self-attention mechanisms, effectively encoding task-specific context to steer the model's representation. We implement the "Deep" version of VPT, where learnable prompt tokens are inserted into the sequence at every Transformer layer. As indicated in Table 5, we set the prompt length to 128 tokens and disable dropout to maintain signal integrity during the adaptation of geospatial features.

D.2.2. SLR

We implement the Scaled Low-Rank (SLR) adapter method as part of our PEFT baseline, which introduces a small number of parameters to enhance pre-trained transformer models for domain adaptation, especially in remote sensing applications. The SLR method focuses on efficiently adapting foundation models without retraining all parameters. It achieves this by introducing low-rank matrices and learnable scaling vectors into the existing transformer layers, specifically targeting the linear transformations in the Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP) components. The key advantage of SLR is its ability to scale the model's capacity with minimal computational overhead by adjusting only the scaling parameters and the low-rank matrices. We set the bottleneck dimension to 16, focusing on parameter efficiency while allowing sufficient capacity for modality alignment.

Table 5. Default hyperparameters of PEFT baselines.

Method	Hyperparameter	Meaning	Value
VPT [10]	Prompt length	The number of soft prompt tokens	128
	Dropout	Dropout rate	0
SLR [19]	Bottleneck dimension	The hidden dimension of the adapter	16
Rein [23]	Prompt length	The number of soft prompt tokens	100
	Rank	The rank of the low rank matrix	16
	Link to query	Whether to link the token to query	True
	Scale	The initial scale of the adapter feather	0.001
	Dropout	Dropout rate	0.1
LoRA [7]	Rank	The rank of the low rank matrix	64
	Alpha	The alpha value	1
	Dropout	Dropout rate	0
Adapter [6]	Bottleneck dimension	The hidden dimension of the adapter	64
	Dropout	Dropout rate	0.1
AdaptFormer [1]	Bottleneck dimension	The hidden dimension of the adapter	64
	Dropout	Dropout rate	0.1
	Scale	The initial scale of the adapter feather	0.1
Earth-Adapter [8]	Bottleneck dimension	The hidden dimension of the adapter	16 on R2U 32 on U2R 64 on other benchmarks
	Scale	The initial scale of the adapter feather	0.1
	Cut off ratio	The threshold to decompose frequncy	0.2 on V2R & V2P 0.3 on other benchmarks
	Frequency layer	The layer to apply frequency Adapters	[0-2] on P2V [21-23] on V2P & R2U [18-23] on U2R [0-23] on other benchmarks

D.2.3. Rein

Rein [23] is a robust fine-tuning method specifically designed to harness Vision Foundation Models for domain generalization tasks. Unlike standard adapters that insert modules within the Transformer blocks, Rein refines the output feature maps of each frozen backbone layer. It introduces a set of learnable tokens that interact with the image features through a dot-product similarity mechanism to generate instance-level feature refinements. To maintain parameter efficiency, the method employs a low-rank decomposition for the token sequences ($T = A \times B$) and shares the MLP weights across all layers. Following the default settings of the official code, we configure the token length to 100 and the low-rank dimension to 16. Additionally, we enable the explicit linkage between these learnable tokens and the object queries in the decoder to further enhance instance discrimination.

D.2.4. LoRA

Low-Rank Adaptation (LoRA) [7] freezes the pre-trained model weights and injects trainable rank decomposition matrices into the self-attention layers. We specifically target the Query and Value projection matrices in the MSA blocks. We set the rank r to 64 and the scaling factor α to 1, ensuring that the low-rank updates can sufficiently capture spatial dependencies without introducing excessive parameters.

D.2.5. Adapter

The Adapter [6] method, originally proposed for efficient natural language processing transfer learning, introduces small bottleneck modules between the pre-trained layers of a Transformer model. Unlike full fine-tuning, which updates all pa-

rameters, Adapters freeze the pre-trained weights and only train the newly added parameters, typically comprising a very small percentage of the original model size. In the standard architecture, two adapter modules are inserted per Transformer layer: one after the multi-head attention projection and another after the feed-forward network. Each adapter consists of a down-projection to a low-dimensional bottleneck, a non-linear activation, and an up-projection back to the original dimension, along with a skip-connection to facilitate identity initialization. We set the bottleneck dimension to 64 and apply a dropout rate of 0.1.

D.2.6. AdaptFormer

AdaptFormer [1] is a method designed to efficiently adapt Vision Transformers [3] (ViTs) to downstream tasks by replacing the original MLP block with a modified "AdaptMLP" module. This module introduces a lightweight bottleneck structure in parallel to the frozen MLP layers. Specifically, the AdaptMLP consists of two branches: the original frozen MLP and a trainable branch comprising a down-projection layer, a non-linear activation (ReLU), and an up-projection layer. The trainable branch processes the input features and scales the output by a learnable factor s before adding it to the frozen MLP output via a residual connection. This design allows the model to learn task-specific features with a minimal number of additional parameters. In our experiments, we set the bottleneck dimension to 64 and initialize the scaling factor s to 0.1.

D.2.7. Earth-Adapter

Earth-Adapter [9] is designed to mitigate the impact of high-frequency artifacts prevalent in RS imagery. It employs a Mixture of Frequency Adaptation mechanisms, which integrates a Mixture of Adapters (MoA) with the Discrete Fourier Transform (DFT). The module operates in parallel to the frozen backbone layers. Specifically, the input features are processed by three expert branches: a standard Spatial Adapter, a Low-Frequency (LF) Adapter, and a High-Frequency (HF) Adapter. The frequency adapters utilize DFT to decompose features based on a learnable cutoff frequency ρ, separating artifacts (typically high-frequency) from global semantic structures (low-frequency). A dynamic router then assigns weights to these experts to fuse the features.

To ensure optimal performance, we strictly follow the hyperparameter configurations derived from the extensive ablation studies reported in the original paper for the Domain Adaptation (DA) benchmarks. For Potsdam $\rightarrow$ Vaihingen (P2V), we set the bottleneck dimension to 64, cutoff frequency to 0.3, and apply frequency adapters to shallow layers [0-2]. For Vaihingen $\rightarrow$ Potsdam (V2P), we use dimension 32, cutoff 0.2, and deep layers [21-23]. For Rural $\rightarrow$ Urban (R2U), we use dimension 16, cutoff 0.3, and layers [21-23]. For Urban $\rightarrow$ Rural (U2R), we use dimension 32, cutoff 0.2, and layers [18-23]. For all Domain Generalization (DG) experiments, we utilize the default robust setting with a bottleneck dimension of 64, a cutoff frequency of 0.3, and layers [0-23]. We follow the official code to implement the Earth-Adapter, with the initial scaling factor set to 0.1 across all experiments.

D.3. Proposed Methods

The pseudo-code for CrossEarth-Gate is described in Algorithm 1. Furthermore, we provide the default hyperparameters for our proposed CrossEarth-Gate in Table 6. The implementation is based on the MMEngine framework. CrossEarth-Gate introduces a dynamic selection mechanism that activates a specific subset of modules from the RS Module Toolbox based on Fisher Information. The importance scores are calculated using the squared gradients accumulated over a defined number of steps, normalized relatively across module types to ensure balanced selection.

Table 6. Default hyperparameters of CrossEarth-Gate.

Method	Hyperparameter	Meaning	Value
CrossEarth-Gate	Active nodules	The number of selected modules	18
	Semantic dimension	The hidden dimension of the semantic module	64
	Spatial dimension	The hidden dimension of the spatial module	64
	Frequency dimension	The hidden dimension of the frequency module	32
	Selection number	The number of the selections	10
	Accumulation steps	Steps to accumulate Fisher information	50 on DA bechmarks 100 on DG bechmarks

Algorithm 1 Pseudo-code for CrossEarth-Gate

Input: Pre-trained foundation model α, RS module toolbox ζ (Spatial, Semantic, Frequency), training dataset D.
Parameters: Evaluation samples M, update interval N, maximum active modules k.

Initialize: Insert all modules ζ into every block of α and set $Iter \leftarrow 0$.
Set State: $EvaluatePhase \leftarrow$ True.
while training not converged **do**
  Zero gradients.
  Sample batch $(\mathbf{X}, \mathbf{Y})$ from D.
  Compute task loss $\mathcal{L}$ and backpropagate to obtain gradients.
  if $EvaluatePhase$ is True **then**
    for each module parameter ζ_i^z in layer i of type z **do**
      Accumulate Fisher Info: $\hat{F}_{\zeta_i^z} \mathrel{+}= (\nabla_{\zeta_i^z} \mathcal{L})^2$
    end for
    if evaluated M samples **then**
      for each module type z and layer i **do**
        Aggregate score: $\hat{S}_i^z \leftarrow \sum_{\zeta_i^z} \hat{F}_{\zeta_i^z}$
      end for
      for each module type z **do**
        Relative score: $S_i^z \leftarrow \hat{S}_i^z \,/\, \sum_j \hat{S}_j^z$
      end for
      Activate Top-k modules with the highest S_i^z and freeze all other modules.
      $EvaluatePhase \leftarrow$ False, Reset $\hat{F} \leftarrow 0$.
    end if
  else
    Update weights of active modules.
    if $Iter \mod N == 0$ **then**
      Activate gradients for all modules in toolbox.
      $EvaluatePhase \leftarrow$ True.
    end if
    $Iter \leftarrow Iter + 1$
  end if
end while

References

[1] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022. 3, 17, 18

[2] 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. 4, 6, 7

[3] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 18

[4] Ziyang Gong, Zhixiang Wei, Di Wang, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Naoto Yokoya, et al. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *arXiv preprint arXiv:2410.22629*, 2024. 15, 16

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 6

[6] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 17

[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. *The International Conference on Learning Representations*, 1(2):3, 2022. 3, 17

[8] Jiajun Hu, Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Learn to preserve and diversify: Parameter-efficient group with orthogonal regularization for domain generalization. In *European Conference on Computer Vision*, pages 198–216. Springer, 2024. 17

[9] Xiaoxing Hu, Ziyang Gong, Yupei Wang, Yuru Jia, Gen Luo, and Xue Yang. Earth-adapter: Bridge the geospatial domain gaps with mixture of frequency adaptation. *arXiv preprint arXiv:2504.06220*, 2025. 3, 15, 16, 18

[10] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 16, 17

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 4, 6, 7

[12] Chenbin Liang, Bo Cheng, Baihua Xiao, Yunyun Dong, and Jinfen Chen. Multilevel heterogeneous domain adaptation method for remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023. 15

[13] Songlin Liu, Linwei Chen, Li Zhang, Jun Hu, and Ying Fu. A large-scale climate-aware satellite image dataset for domain adaptive land-cover semantic segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:98–114, 2023. 6, 8, 10, 11, 15

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 15

[15] Xianping Ma, Xiaokang Zhang, Zhiguo Wang, and Man-On Pun. Unsupervised domain adaptation augmented by mutually boosted attention for semantic segmentation of vhr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 15

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 4, 6, 7

[17] Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy. Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1):913, 2023. 12, 15, 16

[18] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 4, 6, 7

[19] Linus Scheibenreif, Michael Mommert, and Damian Borth. Parameter efficient self-supervised geospatial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27841–27851, 2024. 17

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[21] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 14, 15, 16

[22] Luhan Wang, Pengfeng Xiao, Xueliang Zhang, and Xinyang Chen. A fine-grained unsupervised domain adaptation framework for semantic segmentation of remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:4109–4121, 2023. 15

[23] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. 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*, pages 28619–28630, 2024. 15, 17

[24] Fahong Zhang, Yilei Shi, Zhitong Xiong, Wei Huang, and Xiao Xiang Zhu. Pseudo features-guided self-training for domain adaptive semantic segmentation of satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 15

[25] Ying Zhao, Shuang Li, Chi Harold Liu, Yuqi Han, Hao Shi, and Wei Li. Domain adaptive remote sensing scene recognition via semantic relationship knowledge transfer. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 15