Small but Mighty: Dynamic Wavelet Expert-Guided Fine-Tuning of Large-Scale Models for Optical Remote Sensing Object Segmentation

Yanguang Sun[1], **Chao Wang**[1], **Jian Yang**[2], **Lei Luo**[1] *

[1]PCA Lab, Nanjing University of Science and Technology, Nanjing, China
[2]PCA Lab, VCIP, College of Computer Science, Nankai University, Tianjin, China
Sunyg@njust.edu.cn, wchao0601@163.com, csjyang@nankai.edu.cn, luoleipitt@gmail.com

Abstract

Accurately localizing and segmenting relevant objects from optical remote sensing images (ORSIs) is critical for advancing remote sensing applications. Existing methods are typically built upon moderate-scale pre-trained models and employ diverse optimization strategies to achieve promising performance under full-parameter fine-tuning. In fact, deeper and larger-scale foundation models can provide stronger support for performance improvement. However, due to their massive number of parameters, directly adopting full-parameter fine-tuning leads to pronounced training difficulties, such as excessive GPU memory consumption and high computational costs, which result in extremely limited exploration of large-scale models in existing works. In this paper, we propose a novel dynamic wavelet expert-guided fine-tuning paradigm with fewer trainable parameters, dubbed WEFT, which efficiently adapts large-scale foundation models to ORSIs segmentation tasks by leveraging the guidance of wavelet experts. Specifically, we introduce a task-specific wavelet expert extractor to model wavelet experts from different perspectives and dynamically regulate their outputs, thereby generating trainable features enriched with task-specific information for subsequent fine-tuning. Furthermore, we construct an expert-guided conditional adapter that first enhances the fine-grained perception of frozen features for specific tasks by injecting trainable features, and then iteratively updates the information of both types of feature, allowing for efficient fine-tuning. Extensive experiments show that our WEFT not only outperforms 21 state-of-the-art (SOTA) methods on three ORSIs datasets, but also achieves optimal results in camouflage, natural, and medical scenarios.

Code — https://github.com/CSYSI/WEFT

Introduction

In recent years, object segmentation in optical remote sensing images (ORSIs) has become a research focus (Li et al. 2023a; Quan et al. 2024; Sun et al. 2025c), owing to its broad applications in areas such as urban planning, agricultural monitoring, disaster assessment, and military reconnaissance. Unlike common natural images , ORSIs are typically captured by sensors mounted on aircraft or satellites and are presented as bird's-eye views. As a result, targets in ORSIs often exhibit arbitrary orientations, drastic scale variations, and dense distributions across complex backgrounds, making their segmentation particularly challenging.

To achieve promising performance, a large number of deep learning–based ORSIs object segmentation methods (Li et al. 2019, 2022; Gong et al. 2023; Quan et al. 2024) have been proposed one after another. At the beginning, the architectures (as depicted in Fig. 1(a)) of these methods are primarily based on pre-trained convolutional encoders ($e.g.$, VGG16 (Simonyan and Zisserman 2014), or ResNet50 (He et al. 2016)), which extract initial features that are then enhanced through sophisticated optimization strategies (Li et al. 2022, 2023b; Liu et al. 2023; Quan et al. 2024; Lian et al. 2025). Although these methods achieve commendable performance by updating all parameters, the inherent limitation of convolutional networks ($i.e.$, local receptive fields), makes it difficult to further improve performance. Soon after that, Transformer architectures (Yuan et al. 2021; Wang et al. 2022) rose to prominence due to their superior modeling capacity and effectively mitigated this limitation.

Building on this, numerous Transformer-based methods (Li et al. 2023a; Dong et al. 2024; Sun, Yang, and Luo 2024; Zhao et al. 2024; Sun et al. 2025c) are introduced and achieve better performance in ORSIs object segmentation tasks. This performance advantage stems in part from global modeling, with another important factor being the model scale. For example, Transformer-based architectures ($e.g.$, Swin-B (Liu et al. 2021b), 87.77M parameters; PVTv2-B4 (Wang et al. 2022), 62.60M; and DPU-Former (Sun et al. 2025c), 38.89M) used in existing methods (Zhuge et al. 2023; Sun, Yang, and Luo 2024; Sun et al. 2025c) typically have two to three times more trainable parameters than their convolutional counterparts, or even more. Indeed, the scale of the above architectures is not particularly large, but rather moderate. However, larger-scale and deeper foundation models (Zhu et al. 2022; Oquab et al. 2023) have been explored in current architecture research and have demonstrated outstanding performance across visual tasks .

Inspired by this, we believe that rather than relying on sophisticated strategies to improve performance, it is more effective to embed a deeper and larger-scale foundation model to encode strong discriminative features from input images. With this goal in mind, we abandon small- and

*Corresponding author.

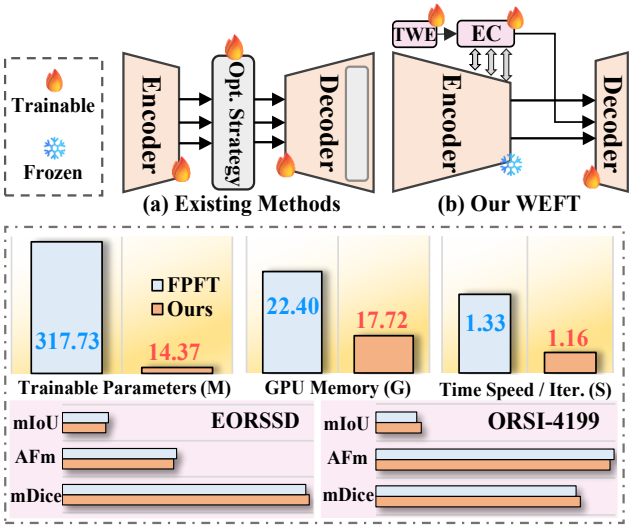

Figure 1: The top part shows architectural differences between existing methods. The bottom part presents a comparison of efficiency and performance between full-parameter fine-tuning (FPFT) and the proposed WEFT method, conducted under the same architecture.

moderate-scale foundation models in terms of framework design and instead introduce a deeper, larger-scale, and more parameter-rich foundation model ($e.g.$, UniPerceiver-L (Zhu et al. 2022) with 303.36M parameters). At first, we adopt the same training strategy ($i.e.$, full-parameter fine-tuning (FPFT)) as existing works (Li et al. 2023a; Quan et al. 2024; Sun et al. 2025c) to optimize the model. However, in Fig. 1, we find that this strategy struggles to cope with large-scale foundation models with massive parameters, as all parameters must be updated simultaneously during training. This significantly increases the computational burden during both the forward and backward passes and leads to a sharp increase in GPU memory consumption, making the optimization process especially difficult. In particular, when dealing with high-resolution inputs or large batch sizes, memory usage often approaches hardware limits, easily resulting in resource bottlenecks or even training interruption. This constitutes an important reason for the relatively limited exploration of large-scale models in ORSIs segmentation tasks.

Considering this key factor, we no longer use the typical FPFT strategy to train the network, and instead propose a novel dynamic wavelet expert-guided fine-tuning paradigm, termed WEFT. As shown in Fig. 1(b), we keep the large-scale model frozen with its parameters unchanged, and introduce a lightweight, trainable branch in parallel to supplement task-specific information that the frozen branch lacks, thereby enabling more effective adaptation to ORSIs tasks. From Fig. 1, the performance of our WEFT is almost the same as that under full-parameter fine-tuning, but our WEFT only requires 14.37M trainable parameters, which only accounts for 4.52% of the entire framework parameters. Moreover, it reduces training GPU memory consumption by approximately 26.41% and improves training speed (1 iter.) by 14.66%. Specifically, our WEFT framework consists of two lightweight components: the task-specific wavelet expert (TWE) extractor and the expert-guided conditional (EC) adapter. Technically, the TWE extractor adaptively integrates wavelet experts from different perspectives to obtain trainable features rich in task-oriented knowledge from input images. Subsequently, the EC adapter is used to supplement task-specific details in frozen features by merging trainable features, and further strengthens both types of features, enabling efficient fine-tuning. Extensive experiments on three ORSIs benchmark datasets demonstrate that our WEFT not only clearly surpasses 21 state-of-the-art (SOTA) approaches but also requires far few trainable parameters. Additionally, it exhibits strong generalization capabilities in camouflage, natural, and medical scenarios.

The main contributions can be summarized as follows:

- A novel dynamic wavelet expert-guided fine-tuning (WEFT) is proposed to efficiently adapt large-scale foundation models to object segmentation tasks in ORSIs.
- A lightweight task-specific wavelet expert (TWE) extractor is introduced to obtain discriminative trainable features by adaptively combining various wavelet experts.
- An efficient expert-guided conditional (EC) adapter is designed to reconstruct the internal information within trainable and frozen features through conditional optimization.

Related Works

Optical remote sensing object segmentation. The primary objective of ORSIs object segmentation tasks is to segment meaningful remote sensing targets. Initially, convolution-based models (Li et al. 2019, 2022) are widely proposed, with various optimization strategies ($e.g.$, attention (Zhang et al. 2021), boundary auxiliary (Zhou et al. 2022b; Gong et al. 2023), and multi-scale enhancement (Quan et al. 2024; Sun et al. 2025a)) designed to optimize initial features, yielding promising results. However, due to the limited local perception of convolutional encoders (He et al. 2016), these methods face performance bottlenecks. Considering this factor, GeleNet (Li et al. 2023a), TLCKDNet (Dong et al. 2024), and DPU-Former (Sun et al. 2025c) introduced moderate-scale Transformer-based encoders (such as PVTv2 (Wang et al. 2021)) in their design and achieve excellent accuracy by updating all parameters. We believe that architectures based on large-scale models ($e.g.$, UniPerceiver (Zhu et al. 2022)) are more advantageous for tackling challenging ORSIs tasks. However, the FPFT strategy adopted by existing works struggles to deploy these models effectively, primarily due to their massive parameters.

Fine-tuning of large-scale foundation models. Large-scale models (Zhu et al. 2022; Oquab et al. 2023) refer to foundation architectures with deep structures and large number of parameters, pre-trained on massive datasets. Their powerful modeling capabilities provide strong support for achieving excellent performance in downstream tasks. However, the application of these models is often challenging due to their enormous parameters, which leads to significant computational overhead and makes training difficult when using the FPFT strategy. Recently, several fine-tuning methods (Hu et al. 2022; Jia et al. 2022) have been proposed to adapt

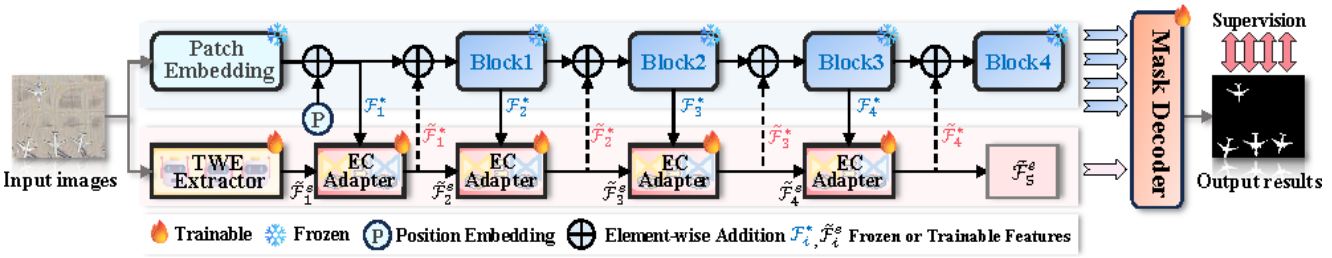

Figure 2: Overview of our WEFT framework. The overall architecture comprises a frozen foundation model, a Task-specific Wavelet Expert (TWE) extractor, four Expert-guided Conditional (EC) adapters, and a mask decoder. In the training process, the proposed WEFT method requires only 14.37M trainable parameters while achieving excellent performance.

large-scale models to visual tasks. Specifically, LoRA (Hu et al. 2022) achieved parameter fine-tuning by inserting low-rank trainable matrices into pre-trained models. VPT (Jia et al. 2022) embedded some small learnable prompts in a frozen backbone to allow adaptation to recognition tasks.

Unlike the above methods, we propose an innovative fine-tuning paradigm based on dynamic wavelet expert guidance to alleviate the training difficulties encountered when applying large-scale models to ORSIs object segmentation tasks.

Methodology

Overall Framework

The overall framework of our WEFT is illustrated in Fig. 2, which consists of four key components: **(a)** UniPerceiver-L (Zhu et al. 2022) foundation model with frozen parameters, **(b)** Task-specific wavelet expert (TWE) extractor, **(c)** Expert-guided conditional (EC) adapter, and **(d)** mask decoder (Cheng et al. 2022). Specifically, for an input image $\mathcal{I}_m \in \mathbb{R}^{3\times H\times W}$, we adopt a dual-branch architecture ($i.e.$, the frozen UniPerceiver-L network and the trainable TWE extractor) to simultaneously extract frozen and trainable features. Subsequently, the trainable features $\{\mathcal{F}_i^{\diamond}\}_{i=1}^{4}$ and the frozen features $\{\mathcal{F}_i^{*}\}_{i=1}^{5}$ are fed into the EC adapter, where information guidance and reinforcement are achieved through collaborative operations of deformable attention (Zhu et al. 2021), edge-aware subspace token optimizer (ESTO), and spatial-aware expert enhancer (SEE). Ultimately, the optimized features are input into a lightweight mask decoder to generate the final segmentation results.

Task-specific Wavelet Expert Extractor

The task-specific wavelet expert (TWE) extractor is designed to dynamically acquire wavelet experts that are rich in task-related knowledge through well-designed wavelet convolutions (Finder et al. 2024), thereby providing strong informational support for subsequent fine-tuning. Not only does it provide task-specific information to the frozen foundation model, but it also supplements local details within the Transformer structure, refining object boundaries.

Specifically, as shown in Fig. 3, we first perform a downsampling ($\mathsf{DS}(\cdot)$) on the image $\mathcal{I}_m$, $i.e.$, $f_m = \mathsf{DS}(\mathcal{I}_m)$, and then model various wavelet experts $\{\mathrm{E}_n^{\diamond}\}_{n=1}^{7}$, which are obtained through wavelet convolutions (Finder et al. 2024) from different perspectives and enriched with task-specific

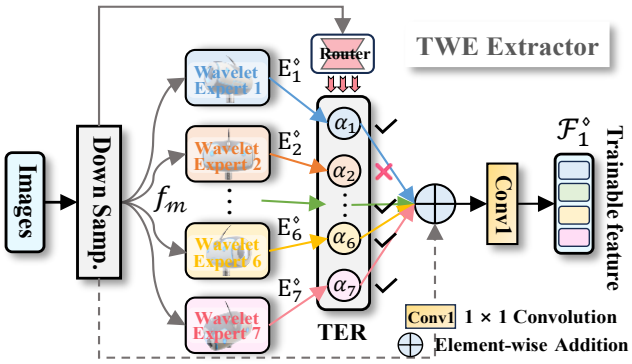

Figure 3: Details of the first stage in TWE extractor.

knowledge. Unlike standard convolutions, wavelet convolution is more lightweight and increases diversity by modeling features from four directions ($i.e.$, HH, HL, LH, and LL). This entire process can be formulated as follows:

$$\begin{aligned}
&\{\mathrm{E}_n^{\diamond}\}_{n=1}^{7} = \mathcal{C}_1(\mathsf{Cat}[\tilde{f}_n^1, \tilde{f}_n^2, \tilde{f}_n^3, \tilde{f}_n^4]) + f_m,\\
&\tilde{f}_n^k = \mathcal{WC}_{2n-1}(f_n^k + \tilde{f}_n^{k-1}), \{f_n^k\}_{k=1}^{4} = \mathsf{Split}(f_m),\\
&\mathcal{WC}_{2n-1}(x) = \mathsf{IWT}(\mathcal{DC}_{2n-1}(w_c, x_{hh}, x_{hl}, x_{lh}, x_{ll})),
\end{aligned} \tag{1}$$

where $\mathcal{C}_1(\cdot)$ is a 1×1 convolution, $\mathsf{Cat}[\cdot]$ and "+" present concatenation and element-wise addition, $\mathcal{WC}_{2n-1}(\cdot)$ and $\mathcal{DC}_{2n-1}(\cdot)$ represent a wavelet convolution and a depthwise convolution with kernel size $2n$-1×$2n$-1, $\mathsf{Split}(\cdot)$ is a separation operation that divides the feature into four sub-features along the channels. $\mathsf{IWT}(\cdot)$ is the inverse wavelet transform. "$(x_{hh}, x_{hl}, x_{lh}, x_{ll})$" are obtained from the wavelet transform of x, denoted as "$\mathsf{WT}(x)$", "w_c" is a weight tensor. Based on this, we obtain multiple wavelet experts $\{\mathrm{E}_n^{\diamond}\}_{n=1}^{7}$ with varying receptive fields. Due to differences in their receptive fields, these experts carry distinct task-relevant knowledge, each specializing in object information at a specific scale. In ORSIs, the scale variation of targets is often significant. We argue that not all wavelet experts $\{\mathrm{E}_n^{\diamond}\}_{n=1}^{7}$ are equally beneficial. For small objects, an excessively large receptive field may introduce ambiguity, whereas for large objects, an overly small receptive field may lead to incomplete comprehension. Only those with appropriate receptive fields are effective in contributing to model fine-tuning.

Therefore, we propose a **Top-k Expert Router (TER)**, which aims to dynamically select appropriate wavelet experts $\{\mathrm{E}_n^{\diamond}\}_{n=1}^{7}$ to achieve optimal matching of remote sensing targets with varying types and scales. Technically, we first obtain a set of learnable weights α=$\{\alpha_1, \alpha_2, \dots, \alpha_7\}$ based on the input feature f_m, as shown in:

$$\alpha = \mathbf{Sof}(w_g(\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} f_m[:,:,i,j]) + b_g), \quad (2)$$

where $\mathbf{Sof}(\cdot)$ presents the Softmax, "w_g" and "b_g" denote the learned weight tensor and bias vector from a linear layer, "(i, j)" are indices over the spatial dimension. Furthermore, we select the top-4 wavelet experts indices with the highest scores from the 7 weights generated by the gating network to form the set $\mathcal{T}$, *i.e.*, $\mathcal{T} = \mathbf{TopK}(\alpha, 4) \subset \{1, 2, ..., 7\}, |\mathcal{T}|$ = 4, where $\mathbf{TopK}(\cdot)$ is an operator that selects the indices of the k largest values from a given vector. The corresponding weights are then normalized to ensure a valid probability distribution, which is used as the final fusion coefficients $\tilde{\alpha}$=$\{\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4\}$. These coefficients are applied to the outputs of the selected wavelet experts to generate the final trainable feature $\tilde{\mathcal{F}}_1^{\diamond}$ via a weighted sum:

$$\mathcal{F}_1^{\diamond} = \mathcal{C}_1(f_m + \sum_{u\in\mathcal{T}} \tilde{\alpha}_u \cdot \mathrm{E}_u^{\diamond}),\ \tilde{\alpha}_u = \frac{\mathsf{exp}(\alpha_u)}{\sum_{v\in\mathcal{T}} \mathsf{exp}(\alpha_v)}, \quad (3)$$

where "$\cdot$" is an element-wise multiplication, "u" denotes the index of a selected expert in the top-k set $\mathcal{T}$, and "v" is used to compute the denominator over the selected experts.

In the subsequent process, the proposed TWE extractor adopts a classical hierarchical structure, which takes the generated trainable feature $\mathcal{F}_1^{\diamond}$ as input and progressively produces three multi-scale features $\{\mathcal{F}_2^{\diamond}, \mathcal{F}_3^{\diamond}, \mathcal{F}_4^{\diamond}\}$ enriched with task-related information through identical operations.

Expert-guided Conditional Adapter

The expert-guided conditional (EC) adapter is constructed to complement the task-oriented details of frozen features $\{\mathcal{F}_i^*\}_{i=1}^{5}$ ($\mathcal{F}_i^* \in \mathbb{R}^{\frac{HW}{16^2}\times C}$) through trainable features $\{\mathcal{F}_i^{\diamond}\}_{i=1}^{4}$ ($\mathcal{F}_i^{\diamond} \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C}$) enriched with expert knowledge, and to enable iterative updates between the trainable and frozen features. Specifically, for trainable features $\{\mathcal{F}_i^{\diamond}\}_{i=1}^{4}$ based on wavelet experts, we select three scale features with the same scale as and adjacent to the frozen feature for fine-tuning progress, *i.e.*, $\tilde{\mathcal{F}}_1^e$= $\mathsf{Cat}[\mathsf{RS}(\mathcal{F}_2^{\diamond}), \mathsf{RS}(\mathcal{F}_3^{\diamond}), \mathsf{RS}(\mathcal{F}_4^{\diamond})]$, where $\mathsf{RS}(\cdot)$ denotes a reshaping of the tensor, $\tilde{\mathcal{F}}_1^e \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times C}$. Compared to a single scale, it contains more task-oriented knowledge. Technically, during the first stage fine-tuning, with $\tilde{\mathcal{F}}_1^e$ and $\mathcal{F}_1^*$ from the patch embedding as initial inputs, we first employ deformable attention (Zhu et al. 2021) to inject task-specific information into the frozen feature $\mathcal{F}_1^*$, as follows:

$$\hat{\mathcal{F}}_1^* = DeformAttn(\mathbf{LN}(\mathcal{F}_1^*), \mathbf{LN}(\tilde{\mathcal{F}}_1^e)), \quad (4)$$

where $\mathbf{LN}(\cdot)$ denotes the layer normalization to stabilize the input distribution. Although deformable attention (Zhu et al. 2021) effectively injects task-specific knowledge, it focuses mainly on semantic alignment between feature sources, without explicitly modeling structural details within the fused representation. As a result, these tokens may suffer from semantic ambiguity and insufficient structural awareness, which can limit the accuracy of segmentation.

Algorithm 1: Edge-aware Subspace Token Optimizer

Input: $\hat{\mathcal{F}}_1^* \in \mathbb{R}^{N\times C}$, subspaces $H > 1$, temp. ϱ, weight λ

Normalize $\hat{\mathcal{F}}_1^*$ along channel dimension;
Split into H subspaces: $\mathrm{Norm}(\hat{\mathcal{F}}_1^*) \to \mathbf{F}_h\ (h = 1, \dots, H)$;
Compute similarity: $\mathbf{S}_h \leftarrow \mathbf{F}_h\mathbf{F}_h^{\top}/\sqrt{d}\cdot\varrho$;
Attention: $\mathcal{A}_h \leftarrow \mathrm{Softmax}(\mathbf{S}_h)$;
Aggregate & merge heads: $\mathbf{T}_1^* \leftarrow [\mathcal{A}_h \cdot \mathbf{X}_h]_{h=1}^{H}$;
Compute channel-wise variance: $\mathbf{V} \leftarrow \mathrm{Var}(\hat{\mathcal{F}}_1^*)$;
Edge mask: $\mathbf{M} \leftarrow \sigma\big((\mathbf{V} - \overline{\mathbf{V}})/(\sigma_{\mathbf{v}} + \epsilon)\big)$;
Apply mask: $\widetilde{\mathbf{T}}_1^* \leftarrow (\mathbf{1} + \lambda\mathbf{M}) \cdot \mathbf{T}_1^*$;
Gate & Residual: $\widetilde{\mathcal{F}}_1^* \leftarrow \delta \cdot \widetilde{\mathbf{T}}_1^* + \hat{\mathcal{F}}_1^*$;

Output: optimized frozen features $\widetilde{\mathcal{F}}_1^* \in \mathbb{R}^{N\times C}$

Considering these problems, we design an **Edge-aware Subspace Token Optimizer (ESTO)**, which uses subspace-based token-to-token attention to model intra-feature relationships and incorporates an edge-aware modulation mechanism to further enhance structurally significant regions in tokens, as in Algorithm 1. Technically, given the input feature $\hat{\mathcal{F}}_1^* \in \mathbb{R}^{N\times C}$, we first apply L_2 normalization along the channel dimension to stabilize the similarity computation across tokens, and then we reshape the normalized feature into H subspaces, each with a dimension of $d = C/H$. To be specific, we obtain the $\mathbf{F}_h \in \mathbb{R}^{H\times N\times d}$ by reshaping the normalized feature into shape (N, H, d) and transposing the first two dimensions. This subspace formulation enables each head to independently capture token-to-token relationships, providing diverse contextual perspectives across different parts of the feature space. For each subspace, we compute the pairwise similarity between tokens using the scaled dot product, and the formula can be expressed as:

$$\mathbf{S}_h = \frac{\mathbf{F}_h(\mathbf{F}_h)^{\top}}{\sqrt{d}\cdot\varrho},\ \mathbf{S}_h \in \mathbb{R}^{H\times N\times N}, \quad (5)$$

where "$\top$" denotes a transpose operation, "$\sqrt{d}$ " presents the scaling factor, "ϱ" is a temperature hyperparameter that adjusts the sharpness of the similarity distribution. The attention weights $\mathcal{A}_h$ are then obtained by applying softmax to the similarity matrix along the last dimension, *i.e.*, $\mathcal{A}_h = \mathbf{Sof}(\mathbf{S}_h)$. The attention weights are applied to the original (unnormalized) input feature $\hat{\mathcal{F}}_1^*$, reshaped into subspace form $\mathbf{X}_h \in \mathbb{R}^{H\times N\times d}$, to obtain refined subspace representations. These outputs from all subspaces are then concatenated to form the optimized token $\mathbf{T}_1^*$, as follows:

$$\mathbf{T}_1^* = \mathsf{Cat}[\mathcal{A}_1\mathbf{X}_1, \mathcal{A}_2\mathbf{X}_2, \dots, \mathcal{A}_H\mathbf{X}_H] \in \mathbb{R}^{N\times C}. \quad (6)$$

Furthermore, we introduce an edge-aware modulation mechanism based on channel-wise variance to refine structural details, particularly around object boundaries. We observe that tokens with higher variance typically correspond to more informative or structurally salient regions ($e.g.$, edges, and contours). Based on this observation, we estimate a soft edge-aware mask $\mathbf{M}$ using the channel-wise variance $\mathbf{V}$ of the optimized token $\mathbf{T}_1^*$. We then update the token representation using the obtained edge mask $\mathbf{M}$, formulated as:

$$\begin{aligned} \widetilde{\mathbf{T}}_1^* &= (\mathbf{1} + \lambda \cdot \mathbf{M}) \cdot \mathbf{T}_1^*,\ \mathbf{M} = \mathbf{Sig}\left(\frac{\mathbf{V} - \overline{\mathbf{V}}}{\sigma_{\mathbf{v}} + \epsilon}\right), \\ \mathbf{V} &= \mathsf{Var}(\hat{\mathcal{F}}_1^*[n,:]) = \frac{1}{C}\sum_{c=1}^{C}\left(\hat{\mathcal{F}}_1^*[n,c] - \bar{f}_n\right)^2, \end{aligned} \tag{7}$$

where "$\mathbf{1}$" is a matrix of all ones with the same dimensions as "$\mathbf{M}$", $\mathbf{Sig}(\cdot)$ is the Sigmoid function, "$-$" is the element-wise subtraction, $\overline{\mathbf{V}}$ and $\sigma_{\mathbf{v}}$ represent the mean and standard values of variance, λ denotes the hyperparameter of modulating edge mask intensity, $\bar{f}_n(\frac{1}{C}\sum_{c=1}^{C}\hat{\mathcal{F}}_1^*[n,c])$ present the mean value of the n-th token in all channels. Additionally, we introduce a gating mechanism and residual connections to adaptively control the refinement strength and preserve the original information to generate the discriminative frozen feature $\widetilde{\mathcal{F}}_1^*$, which can be defined as:

$$\widetilde{\mathcal{F}}_1^* = \delta \cdot \widetilde{\mathbf{T}}_1^* + \hat{\mathcal{F}}_1^*,\ \delta = \mathbf{Sig}(w_g\bar{\mathcal{F}}_1^* + b_g), \tag{8}$$

where δ denotes a gating coefficient, $\bar{\mathcal{F}}_1^*$ represents a global mean of $\hat{\mathcal{F}}_1^*$. Subsequently, the optimized feature $\widetilde{\mathcal{F}}_1^*$ is injected into the first frozen Transformer block to obtain the feature $\mathcal{F}_2^*$, which is used for fine-tuning in the next block.

Spatial-aware Expert Enhancer (SEE). During the interactive fine-tuning process, we introduce a spatial-aware expert enhancer (SEE) to enhance and update the information in trainable features, aiming to strengthen their perception of spatial structures. Specifically, we employ three distinct branches ($i.e.$, Directional Laplacian Filter, Adaptive Max-pooling, and Multi-scale Operation) to capture spatial structural cues from the trainable features. These cues are then dynamically weighted through learnable parameters, allowing the model to selectively acquire spatial information as needed. Technically, the input feature $\tilde{\mathcal{F}}_1^e$ is restored to three spatial scales and reshaped as feature maps $\{\mathcal{F}_2^\diamond, \mathcal{F}_3^\diamond, \mathcal{F}_4^\diamond\}$. In the first branch, we adopt a directional Laplacian filter to capture structural information by enhancing second-order variations along a specific spatial axis, which emphasizes high-frequency components, such as boundaries and textures. For each scale $i \in \{2,3,4\}$, the DLF-enhanced feature $\mathcal{F}_i^d$ can be defined as:

$$\mathcal{F}_i^d = \sum_{p=-1}^{1}\sum_{q=-1}^{1}\mathbf{K}[p,q] \cdot \breve{\mathcal{F}}_i^\diamond[c, h+p, w+q], \tag{9}$$

where $\mathbf{K} \in \mathbb{R}^{3\times3}$ is a fixed-directional Laplacian kernel that approximates second-order derivatives. $\breve{\mathcal{F}}_i^\diamond$ is the input feature after reflection padding, $p,q \in \{-1,0,1\}$ are the relative spatial offsets. In the second branch, we apply adaptive max-pooling to extract global structural features, which often highlight salient spatial patterns and regions, $i.e.$, $\mathcal{F}_i^a$ = $\max_{(h,w)\in\Omega_i}\mathcal{F}_i^\diamond[c,h,w]$, where $\max_{(h,w)\in\Omega_i}$ represents adaptive max-pooling over the full spatial domain. In the third branch, we use a multi-scale operation ($\mathcal{MSO}(\cdot)$), which consists of depthwise convolutions with different kernels ($i.e.$, 3, 5, and 7), to extract local context at different scales, enhancing both fine and coarse structural patterns, $i.e.$, $\mathcal{F}_i^m = \sum_{n=1}^{3}\mathcal{MSO}_{2n+1}(\mathcal{F}_i^\diamond)$, where $2n+1$ represents the kernel size of depthwise convolutions within the multi-scale operation. Considering the differences in structural information captured by each branch, we employ dynamic weights to control the outputs of the three branches to produce the powerful trainable feature $\tilde{\mathcal{F}}_2^e$ for subsequent fine-tuning. The progress is formulated as follows:

$$\begin{aligned} \tilde{\mathcal{F}}_2^e &= \mathsf{Cat}[\mathsf{RS}(\tilde{\mathcal{F}}_2^\diamond), \mathsf{RS}(\tilde{\mathcal{F}}_3^\diamond), \mathsf{RS}(\tilde{\mathcal{F}}_4^\diamond)] + \tilde{\mathcal{F}}_1^e, \\ \tilde{\mathcal{F}}_i^\diamond &= \mathcal{F}_i^\diamond + \textstyle\sum_{z\in\{d,a,m\}} w_z \cdot \mathcal{F}_i^z,\ \ i = 2,3,4, \end{aligned} \tag{10}$$

where w_d, w_a, and w_m are different learnable weights used to flexibly output the structural information extracted by different branches, and their sum is equal to 1. Similarly, in the second stage of fine-tuning, the generated trainable and frozen features $\tilde{\mathcal{F}}_2^e$ and $\mathcal{F}_2^*$ are used as inputs to our EC adapter, where the same operations are applied to optimize the generation of features $\tilde{\mathcal{F}}_3^e$ and $\mathcal{F}_3^*$. As depicted in Fig. 2, the entire fine-tuning process lasted for four stages.

Loss Function

After the completion of the four fine-tuning stages, the optimized features are passed into a lightweight Transformer-based mask decoder (Cheng et al. 2022), comprising 3.19 million trainable parameters. The WEFT method is fine-tuned under the supervision of a composite loss function that integrates binary cross-entropy ($\mathcal{L}_{bce}$) and Dice coefficient ($\mathcal{L}_{dice}$) losses, which can be formulated as follows:

$$\mathcal{L}_{all} = \beta\mathcal{L}_{bce} + \gamma\mathcal{L}_{dice}, \tag{11}$$

where β and γ denote the hyperparameters set to 5 and 2.

Experiments

Experimental Settings

Datasets. We perform experiments separately on three ORSIs datasets: ORSSD (Li et al. 2019), comprising 600 training images and 200 testing images; EORSSD (Zhang et al. 2021), consisting of 1,400 training images and 600 testing images; and ORSIs-4199 (Tu et al. 2022), which includes 2,000 training images and 2,199 testing images.

Evaluation Metrics. We employ five evaluation metrics to assess the performance of our WEFT method, including mean Intersection over Union (**mIoU**), average F-measure ($\mathbf{AF}_m$), mean Dice coefficient (**mDice**), structural similarity measure ($\mathbf{S}_m$), and mean absolute error (**MAE**).

Implementation details. We implement our WEFT model based on the PyTorch framework and train it using four NVIDIA RTX 4090 GPUs with 24GB of memory. During training, input images are resized to 512 $\times$ 512, with a batch size of 6 and an initial learning rate of 5e-5. The model is optimized using the AdamW optimizer over 80K iterations.

Methods	Trainable Params (M)	ORSSD (200 images)					EORSSD (600 images)					ORSIs-4199 (2199 images)				
		mIoU ↑	AF_m ↑	mDice ↑	S_m ↑	MAE ↓	mIoU ↑	AF_m ↑	mDice ↑	S_m ↑	MAE ↓	mIoU ↑	AF_m ↑	mDice ↑	S_m ↑	MAE ↓
$LVNet_{19}$	-	0.7276	0.7506	0.7897	0.8730	0.0207	0.6734	0.6306	0.7167	0.8355	0.0145	-	-	-	-	-
VST_{21}	44.48	0.8531	0.8262	0.8895	0.9267	0.0094	0.8125	0.7089	0.8418	0.8829	0.0067	0.7809	0.7947	0.8444	0.8733	0.0281
$DAFNet_{21}$	29.35	0.8234	0.7876	0.8739	0.9119	0.0113	0.8002	0.6423	0.8297	0.8830	0.0060	-	-	-	-	-
$PA\text{-}KRN_{21}$	141.06	0.8382	0.8548	0.8955	0.9145	0.0139	0.8047	0.7993	0.8655	0.8802	0.0104	0.7371	0.8200	0.8167	0.8427	0.0382
$EMFINet_{22}$	95.09	0.8350	0.8617	0.8864	0.9267	0.0109	0.8004	0.7984	0.8507	0.8891	0.0084	0.7598	0.8186	0.8335	0.8612	0.0330
$MCCNet_{22}$	67.65	0.8554	0.8957	0.9036	0.9334	0.0087	0.8169	0.8137	0.8724	0.8954	0.0066	0.7768	0.8592	0.8475	0.8682	0.0316
$MJRBM_{22}$	43.54	0.8169	0.8022	0.8529	0.9097	0.0163	0.7931	0.7066	0.8217	0.8786	0.0099	0.7466	0.7995	0.8128	0.8530	0.0374
$ERPNet_{22}$	56.48	0.8224	0.8356	0.8687	0.9153	0.0135	0.7887	0.7554	0.8273	0.8812	0.0089	0.7533	0.8024	0.8224	0.8606	0.0357
$ESGNet_{23}$	-	0.8380	0.8860	0.8929	0.9243	0.0098	0.8023	0.8557	0.8624	0.8880	0.0070	0.7818	0.8740	0.8580	0.8730	0.0289
$GeleNet_{23}$	25.50	0.8575	0.9038	0.9070	0.9326	0.0080	0.8017	0.8528	0.8669	0.8894	0.0055	0.7885	0.8681	0.8614	0.8766	0.0264
$ICON\text{-}P_{23}$	65.68	0.8403	0.8444	0.8964	0.9162	0.0116	0.8122	0.8065	0.8792	0.8821	0.0073	0.7788	0.8531	0.8523	0.8692	0.0282
$ACCoNet_{23}$	102.55	0.8539	0.8806	0.9026	0.9335	0.0088	0.8118	0.7969	0.8651	0.8898	0.0074	0.7778	0.8581	0.8484	0.8711	0.0314
$SRAL_{23}$	31.90	0.8467	0.8514	0.9020	0.9225	0.0107	0.8211	0.8146	0.8824	0.8869	0.0069	0.7705	0.8230	0.8451	0.8678	0.0307
$TLCKDNet_{24}$	52.09	0.8689	0.8719	0.9142	0.9316	0.0082	0.8380	0.7969	0.8895	0.8955	0.0056	-	-	-	-	-
$SOLNet_{24}$	**6.50**	0.8302	0.8925	0.8885	0.9195	0.0111	0.7792	0.8392	0.8467	0.8814	0.0078	-	-	-	-	-
$UDCNet_{24}$	72.20	0.8680	0.8932	0.9131	0.9381	0.0068	0.8073	0.8211	0.8697	0.8957	0.0056	0.7913	0.8648	0.8640	0.8744	0.0266
$SFANet_{24}$	25.10	0.8593	0.8984	0.9077	0.9345	0.0077	0.8136	0.8492	0.8720	0.8955	0.0058	0.7774	0.8647	0.8509	0.8703	0.0292
$ADSTNet_{24}$	62.09	0.8459	0.8979	0.8980	0.9296	0.0086	0.8046	0.8532	0.8651	0.8912	0.0065	0.7693	0.8655	0.8460	0.8647	0.0318
$BCARNet_{25}$	24.00	0.8600	0.9073	0.9088	0.9361	0.0071	0.8248	0.8740	0.8821	0.8969	0.0054	0.7795	0.8666	0.8502	0.8694	0.0306
$LGIPNet_{25}$	65.80	0.8572	0.8924	0.9051	0.9348	0.0082	0.8154	0.8560	0.8721	0.8938	0.0065	0.7899	0.8280	0.8583	0.8754	0.0288
$DPU\text{-}Former_{25}$	44.20	0.8728	0.9024	0.9163	0.9312	0.0062	0.8268	0.8461	0.8811	**0.9011**	0.0056	0.7961	0.8816	0.8677	0.8769	0.0263
Ours	14.37	**0.8964**	**0.9213**	**0.9394**	**0.9383**	**0.0056**	**0.8621**	**0.8810**	**0.9188**	0.9006	**0.0048**	**0.7999**	**0.8826**	**0.8696**	**0.8772**	**0.0238**

Table 1: Comparison with 21 state-of-the-arts (SOTA) methods on three widely-utilized ORSIs datasets. The best results are highlighted in **best**. "↑" and "↓" respectively indicate higher-is-better and lower-is-better performance.

Methods	Trainable Params (M)	CAMO		COD10K		NC4K	
		mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑
$FSPNet_{23}$	273.79	.7367	.8186	.6771	.7579	.7612	.8308
$VSCode_{24}$	117.41	.7730	.8463	.7264	.8046	.7927	.8573
$FSEL_{24}$	67.13	.7996	.8742	.7440	.8243	.7997	.8683
$ZoomXNet_{24}$	84.78	.8090	.8727	.7795	.8429	.8141	.8741
RUN_{25}	-	.6712	.7661	.6500	.7451	.7249	.8095
$DPU\text{-}Former_{25}$	44.20	.7814	.8582	.7394	.8175	.7904	.8582
Ours	**14.37**	**.8308**	**.8972**	**.7984**	**.8707**	**.8362**	**.8964**

Methods	Trainable Params (M)	PASCAL-S		HKU-IS	
		mIoU ↑	mDice ↑	mIoU ↑	mDice ↑
$ICON\text{-}P_{23}$	65.68	.8112	.8730	.8924	.9332
$VSCode_{24}$	117.41	.8177	.8744	.9004	.9359
$MDSAM_{24}$	16.78	.8138	.8721	.9033	.9400
$VST{+}{+}_{24}$	112.23	.8232	.8782	-	-
$FSEL_{24}$	67.13	.8227	.8812	.9000	.9399
$DPU\text{-}Former_{25}$	44.20	.8093	.8709	.8921	.9336
Ours	**14.37**	**.8359**	**.8923**	**.9140**	**.9510**

Methods	Trainable Params (M)	CVC-300		Kvasir	
		mIoU ↑	mDice ↑	mIoU ↑	mDice ↑
$DCRNet_{22}$	28.73	.7974	.8029	.8384	.8881
$CFANet_{23}$	25.24	.8320	.8599	.8698	.9170
$FSEL_{24}$	67.13	.8176	.8828	.8558	.8054
$LSSNet_{24}$	35.94	.8154	.8835	.8660	.9111
$MEGANet_{24}$	44.19	.8214	.8899	.8619	.9135
$DPU\text{-}Former_{25}$	44.20	.8414	.9024	.8609	.9123
Ours	**14.37**	**.8502**	**.9121**	**.8875**	**.9329**

Table 2: Extended applications in **camouflage**, **natural**, and **medical** scenarios. Quantitative results with 13 state-of-the-arts segmentation models on three camouflaged object detection, two salient object detection, and two polyp segmentation datasets.

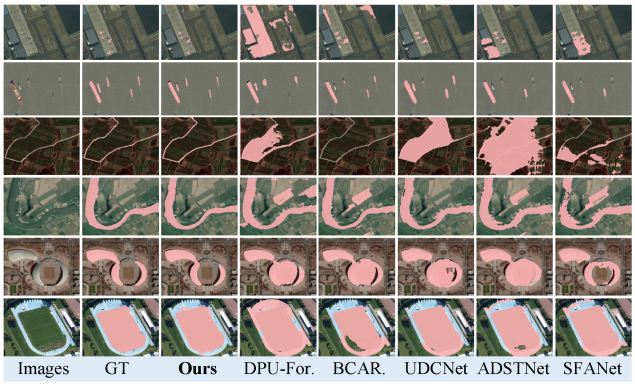

Figure 4: Visual results of WEFT and existing models.

Num.	Component Setting				EORSSD (600 images)			ORSIs-4199 (2199 images)		
	Base.	TWE	ESTO	SEE	mIoU ↑	AF_m ↑	mDice ↑	mIoU ↑	AF_m ↑	mDice ↑
(a)	✓				0.7691	0.8193	0.8521	0.7189	0.8153	0.8096
(b)	✓	✓			0.8253	0.8604	0.8934	0.7528	0.8441	0.8373
(c)	✓		✓		0.8196	0.8637	0.8864	0.7715	0.8596	0.8481
(d)	✓			✓	0.8048	0.8411	0.8785	0.7451	0.8403	0.8310
(e)	✓	✓	✓		0.8323	0.8668	0.8949	0.7864	0.8764	0.8608
(f)	✓		✓	✓	0.8312	0.8579	0.8964	0.7902	0.8705	0.8635
(g)	✓	✓		✓	0.8275	0.8597	0.8937	0.7939	0.8756	0.8649
(h)	✓	✓	✓	✓	**0.8621**	**0.8810**	**0.9188**	**0.7999**	**0.8826**	**0.8696**

Table 3: Ablation analysis of individual components in our WEFT framework on EORSSD and ORSIs-4199 datasets.

Comparison with State-of-the-Arts

We conduct comprehensive comparisons between our WEFT and 21 ORSIs object segmentation approaches, including LVNet, VST, DAFNet, PA-KRN, EMFINet, MCCNet, MJRBM, ERPNet, ESGNet, GeleNet, ICON, ACCoNet, SRAL, TLCKDNet, SOLNet, UDCNet, SFANet, ADSTNet, BCARNet, LGIPNet, and DPU-Former. To ensure fairness, all predicted results are obtained directly from the authors or reproduced using available open-source code.

Quantitative results. Table 1 summarizes the performance of our WEFT model compared to 21 SOTA methods. Specifically, our "**mIoU**" and "**mDice**" metrics outperform the second-best method by substantial margins 2.70% and 2.52% on the ORSSD dataset and 2.88% and 3.29% on the EORSSD dataset. Moreover, under "**MAE**" metric, our method surpasses the second-best approach by 10.71%, 12.50%, and 10.50% on three datasets, respectively. Likewise, other indicators exhibit considerable competitive advantages. In addition, the proposed WEFT model shows a clear advantage in terms of trainable parameters. The entire architecture contains only **14.37 M trainable parameters**, which facilitates the training and deployment of large-scale models in ORSIs tasks. Overall, the quantitative results confirm the effectiveness of our model in achieving high accuracy while maintaining computational efficiency.

Qualitative results. Fig. 4 illustrates the visual segmentation results in various remote sensing scenarios, including aircraft (1^{st} rows), ships (2^{nd} row), highway (3^{rd} row), river

Num.	Expert Allocation	EORSSD (600 images) mIoU ↑	AF_m ↑	mDice ↑	ORSIs-4199 (2199 images) mIoU ↑	AF_m ↑	mDice ↑
(a)	Exp.=1	0.7951	0.8464	0.8726	0.7278	0.8281	0.8156
(b)	Exp.=2	0.8039	0.8501	0.8799	0.7499	**0.8491**	0.8345
(c)	Exp.=6	0.8138	0.8501	0.8867	**0.7544**	0.8437	0.8371
(d)	Exp.=4	**0.8253**	**0.8604**	**0.8934**	0.7528	0.8441	**0.8373**

Table 4: Ablation analysis of the number of wavelet experts.

Num.	Subspace Setting	EORSSD (600 images) mIoU ↑	AF_m ↑	mDice ↑	ORSIs-4199 (2199 images) mIoU ↑	AF_m ↑	mDice ↑
(a)	H = 2	0.8129	0.8541	0.8838	0.7705	0.8567	**0.8484**
(b)	H = 8	0.8130	0.8588	0.8832	0.7628	0.8520	0.8436
(c)	H = 16	0.8124	0.8559	0.8828	0.7634	0.8533	0.8445
(d)	H = 4	**0.8196**	**0.8637**	**0.8864**	**0.7715**	**0.8596**	0.8481

Table 5: Ablation analysis of the number of subspaces.

(4^{th} row), building (5^{th} row) and court (6^{th} row). As shown in Fig. 4, thanks to the efficient fine-tuning of large-scale models, our method achieves more accurate and complete segmentation of remote sensing targets compared to recent methods ($e.g.$, DPU-Former(Sun et al. 2025c), BCARNet (Gu et al. 2025) and SFANet (Quan et al. 2024)).

Extended Applications

To further validate the generalizability of our WEFT framework, we follow the same settings to conduct additional experiments on camouflaged object detection (COD), salient object detection (SOD), and polyp segmentation (PS) tasks. Table 2 compares our method with multiple existing methods on 7 segmentation datasets . In Table 2, it is evident that our WEFT achieves outstanding performance across various image types, which can be attributed to the strong modeling capacity of large-scale foundation models.

Ablation Study

Effect of each component. Table 3 presents the quantitative results for each individual component of our WEFT framework. To be specific, "**Base.**" (Table 3(a)) consists of a frozen UniPerceiver-L network, combined with a trainable mask decoder. Table 3(b) shows the effectiveness of our "**TWE**" extractor, demonstrating that integrating dynamic wavelet experts with knowledge guidance improves the adaptability of frozen frameworks to ORSIs tasks. Moreover, as illustrated in Table 3 (c) and (e), "**ESTO**" yields notable improvements in segmentation precision by enabling subspace token refinement and boundary-aware enhancement. Furthermore, we integrate the designed "**SEE**" into the WEFT (as shown in Table 3 (d), (f), and (g)), which further strengthens the spatial perception of experts during the fine-tuning process, thereby improving the performance. Additionally, in Fig. 5, we provide visualized results obtained by progressively incorporating each component ($i.e.$, TWE, ESTO, and SEE), illustrating that the predicted results increasingly resemble the ground truth. In conclusion, each component is essential and collectively contributes to improvements over the baseline by 12.09% and 11.27% in terms of the "**mIoU**" metric on two ORSIs datasets.

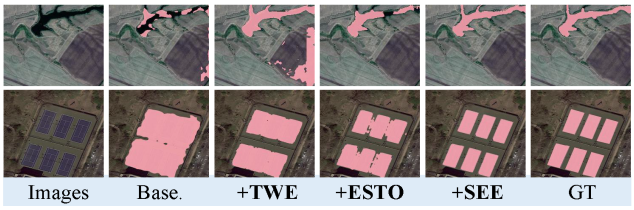

Figure 5: Visual results of each component in WEFT.

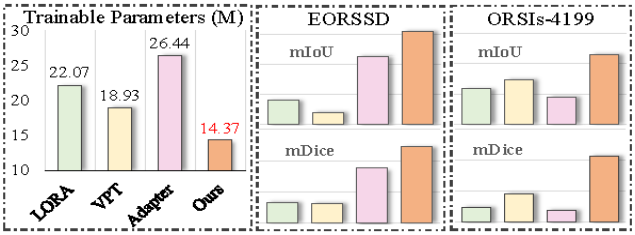

Figure 6: Comparison of different fine-tuning strategies.

Effect of expert allocation. We perform an ablation analysis on the number of experts within our "**TWE**" extractor. The results from Table 4 (a)–(d) indicate that employing multiple wavelet experts significantly outperforms the use of a single wavelet expert. However, the lower performance with six experts compared to four suggests that not all experts contribute positively during fine-tuning, indirectly proving the value of our TER strategy. Accordingly, we select four wavelet experts for the final configuration.

Effect of subspace setting. In Table 5, we present the experimental results obtained using different numbers of subspaces in the proposed "**ESTO**" component. It can be observed that optimal performance is achieved when the number of subspaces is set to four. Based on this, we adopt four subspaces in our subsequent experimental settings.

Effect of fine-tuning strategy. Fig. 6 provides a detailed comparative analysis of different fine-tuning strategies, with all experiments conducted under the same framework. These results demonstrate that, compared to the LORA (Hu et al. 2022), VPT (Jia et al. 2022), and Adapter (Chen et al. 2023) strategies, our WEFT not only significantly reduces trainable parameters, but also achieves superior performance on ORSIs object segmentation. This advantage is attributed to the lightweight and efficient design of all its components.

Conclusions

In this paper, we propose a novel wavelet expert-guided fine-tuning (WEFT), which efficiently adapts frozen large-scale models to ORSIs segmentation tasks. First, we design a TWE extractor that models wavelet experts enriched with task-specific knowledge, providing effective support for fine-tuning. Second, we construct an EC adapter that enables integration between frozen and trainable features, allowing both types of information to be updated. Extensive experiments demonstrate that our WEFT outperforms 35 SOTA models across 10 object segmentation datasets.

Acknowledgments

This work was supported in part by the National Science Fund of China (No. 62276135 and U24A20330).

References

Bernal, J.; Sánchez, J.; and Vilarino, F. 2012. Towards automatic polyp detection with a polyp appearance model. *PR*, 45(9): 3166–3182.

Bui, N.-T.; Hoang, D.-H.; Nguyen, Q.-T.; Tran, M.-T.; and Le, N. 2024. MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation. In *WACV*, 7985–7994.

Chen, Z.; Duan, Y.; Wang, W.; He, J.; Lu, T.; Dai, J.; and Qiao, Y. 2023. Vision transformer adapter for dense predictions. *ICLR*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 1290–1299.

Dong, P.; Wang, B.; Cong, R.; Sun, H.-H.; and Li, C. 2024. Transformer with large convolution kernel decoder network for salient object detection in optical remote sensing images. *CVIM*, 240: 103917.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020. Camouflaged object detection. In *CVPR*, 2777–2787.

Finder, S. E.; Amoyal, R.; Treister, E.; and Freifeld, O. 2024. Wavelet convolutions for large receptive fields. In *ECCV*, 363–380.

Gao, S.; Zhang, P.; Yan, T.; and Lu, H. 2024. Multi-scale and detail-enhanced segment anything model for salient object detection. In *ACM MM*, 9894–9903.

Gong, A.; Nie, J.; Niu, C.; Yu, Y.; Li, J.; and Guo, L. 2023. Edge and skeleton guidance network for salient object detection in optical remote sensing images. *TCSVT*, 33(12): 7109–7120.

Gu, Y.; Chen, S.; Sun, X.; Ji, J.; Zhou, Y.; and Ji, R. 2025. Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration. *PR*, 164: 111478.

He, C.; Zhang, R.; Xiao, F.; Fang, C.; Tang, L.; Zhang, Y.; Kong, L.; Fan, D.-P.; Li, K.; and Farsiu, S. 2025. RUN: Reversible Unfolding Network for Concealed Object Segmentation. *ICML*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Z.; Dai, H.; Xiang, T.-Z.; Wang, S.; Chen, H.-X.; Qin, J.; and Xiong, H. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers. In *CVPR*, 5557–5566.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-seg: A segmented polyp dataset. In *MM*, 451–462.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*, 709–727.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *CVIM*, 184: 45–56.

Li, C.; Cong, R.; Hou, J.; Zhang, S.; Qian, Y.; and Kwong, S. 2019. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *TGRS*, 57(11): 9156–9166.

Li, G.; Bai, Z.; Liu, Z.; Zhang, X.; and Ling, H. 2023a. Salient Object Detection in Optical Remote Sensing Images Driven by Transformer. *TIP*, 32: 5257–5269.

Li, G.; Liu, Z.; Lin, W.; and Ling, H. 2022. Multi-content complementation network for salient object detection in optical remote sensing images. *TGRS*, 60: 1–13.

Li, G.; Liu, Z.; Zeng, D.; Lin, W.; and Ling, H. 2023b. Adjacent context coordination network for salient object detection in optical remote sensing images. *TCYB*, 53(1): 526–538.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *CVPR*, 5455–5463.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *CVPR*, 280–287.

Li, Z.; Miao, Y.; Li, X.; Li, W.; Cao, J.; Hao, Q.; Li, D.; and Sheng, Y. 2024. Speed-oriented lightweight salient object detection in optical remote sensing images. *TGRS*.

Lian, J.; Du, X.; Liu, J.; Hui, L.; and Yang, J. 2025. Cross-Modal Driven Object Restoration for 3D Point Cloud Backdoor Defense. *TIFS*, 20: 11006–11018.

Lian, J.; Wang, D.-H.; Wu, Y.; and Zhu, S. 2024. Multi-Branch Enhanced Discriminative Network for Vehicle Re-Identification. *TITS*, 25(2): 1263–1274.

Liu, N.; Luo, Z.; Zhang, N.; and Han, J. 2024. Vst++: Efficient and stronger visual saliency transformer. *TPAMI*.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021a. Visual saliency transformer. In *ICCV*, 4722–4732.

Liu, Y.; Xiong, Z.; Yuan, Y.; and Wang, Q. 2023. Distilling Knowledge From Super-Resolution for Efficient Remote Sensing Salient Object Detection. *TGRS*, 61: 1–16.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 10012–10022.

Luo, Z.; Liu, N.; Zhao, W.; Yang, X.; Zhang, D.; Fan, D.-P.; Khan, F.; and Han, J. 2024. VSCode: General Visual Salient and Camouflaged Object Detection with 2D Prompt Learning. In *CVPR*, 17169–17180.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, 11591–11601.

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*.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2024. Zoomnext: A unified collaborative pyramid network for camouflaged object detection. *TPAMI*.

Quan, Y.; Xu, H.; Wang, R.; Guan, Q.; and Zheng, J. 2024. ORSI Salient Object Detection via Progressive Semantic Flow and Uncertainty-Aware Refinement. *TGRS*, 62: 1–13.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*.

Sun, L.; Liu, H.; Wang, X.; Zheng, Y.; Chen, Q.; Wu, Z.; and Fu, L. 2025a. Local-Global Information Perception Network for Salient Object Detection in Optical Remote Sensing Images. *TGRS*.

Sun, Y.; Lian, J.; Yang, J.; and Luo, L. 2025b. ControllableLPMoE: Adapting to Challenging Object Segmentation via Dynamic Local Priors from Mixture-of-Experts. In *ICCV*, 22327–22337.

Sun, Y.; Xia, C.; Gao, X.; Ge, B.; Zhang, H.; and Li, K.-C. 2022a. Emcenet: efficient multi-scale context exploration network for salient object detection. In *ICIP*, 1066–1070.

Sun, Y.; Xia, C.; Gao, X.; Yan, H.; Ge, B.; and Li, K.-C. 2022b. Aggregating dense and attentional multi-scale feature network for salient object detection. *DSP*, 130: 103747.

Sun, Y.; Xu, C.; Yang, J.; Xuan, H.; and Luo, L. 2024a. Frequency-spatial entanglement learning for camouflaged object detection. In *ECCV*, 343–360.

Sun, Y.; Xuan, H.; Yang, J.; and Luo, L. 2024b. Glconet: Learning multisource perception representation for camouflaged object detection. *TNNLS*.

Sun, Y.; Yan, J.; Qian, J.; Xu, C.; Yang, J.; and Luo, L. 2025c. Dual-Perspective United Transformer for Object Segmentation in Optical Remote Sensing Images. *IJCAI*.

Sun, Y.; Yang, J.; and Luo, L. 2024. United Domain Cognition Network for Salient Object Detection in Optical Remote Sensing Images. *TGRS*, 62: 3497579.

Tu, Z.; Wang, C.; Li, C.; Fan, M.; Zhao, H.; and Luo, B. 2022. ORSI salient object detection via multiscale joint region and boundary model. *TGRS*, 60: 1–13.

Wang, C.; Fang, W.; Li, X.; Yang, J.; and Luo, L. 2025. Msod: A large-scale multi-scene dataset and a novel diagonal-geometry loss for sar object detection. *TGRS*.

Wang, W.; Sun, H.; and Wang, X. 2024. Lssnet: A method for colon polyp segmentation based on local feature supplementation and shallow feature supplementation. In *MICCAI*, 446–456.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 568–578.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. Pvt v2: Improved baselines with pyramid vision transformer. *CVM*, 8(3): 415–424.

Xia, C.; Sun, Y.; Gao, X.; Ge, B.; and Duan, S. 2022. DMINet: dense multi-scale inference network for salient object detection. *TVC*, 38(9): 3059–3072.

Xia, C.; Sun, Y.; Li, K.-C.; Ge, B.; Zhang, H.; Jiang, B.; and Zhang, J. 2024. Rcnet: Related context-driven network with hierarchical attention for salient object detection. *ESWA*, 237: 121441.

Xu, B.; Liang, H.; Liang, R.; and Chen, P. 2021. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *AAAI*, volume 35, 3004–3012.

Yin, Z.; Liang, K.; Ma, Z.; and Guo, J. 2022. Duplex contextual relation network for polyp segmentation. In *ISBI*, 1–5.

Yu, Z.; Dai, J.; Zhang, Y.; Yang, J.; and Luo, L. 2025a. SSAIM: Not All Self-Attentions Contain Effective Spatial Structure in Diffusion Models for Text-to-Image Editing. In *ACM MM*, 9472–9480.

Yu, Z.; Jin, J.; Zhao, J.; Fu, Z.; and Yang, J. 2025b. TtfDiffusion: Training-free and text-free image editing in diffusion models with structural and semantic disentanglement. *Neurocomputing*, 619: 129159.

Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.-H.; Tay, F. E.; Feng, J.; and Yan, S. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, 558–567.

Zhang, Q.; Cong, R.; Li, C.; Cheng, M.-M.; Fang, Y.; Cao, X.; Zhao, Y.; and Kwong, S. 2021. Dense attention fluid network for salient object detection in optical remote sensing images. *TIP*, 30: 1305–1317.

Zhao, J.; Jia, Y.; Ma, L.; and Yu, L. 2024. Adaptive Dual-Stream Sparse Transformer Network for Salient Object Detection in Optical Remote Sensing Images. *JSTARS*, 17: 5173–5192.

Zhou, T.; Zhou, Y.; He, K.; Gong, C.; Yang, J.; Fu, H.; and Shen, D. 2023. Cross-level feature aggregation network for polyp segmentation. *PR*, 140: 109555.

Zhou, X.; Shen, K.; Liu, Z.; Gong, C.; Zhang, J.; and Yan, C. 2022a. Edge-Aware Multiscale Feature Integration Network for Salient Object Detection in Optical Remote Sensing Images. *TGRS*, 60: 1–15.

Zhou, X.; Shen, K.; Weng, L.; Cong, R.; Zheng, B.; Zhang, J.; and Yan, C. 2022b. Edge-Guided Recurrent Positioning Network for Salient Object Detection in Optical Remote Sensing Images. *TCYB*, 53(1): 539–552.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*.

Zhu, X.; Zhu, J.; Li, H.; Wu, X.; Li, H.; Wang, X.; and Dai, J. 2022. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 16804–16815.

Zhuge, M.; Fan, D.-P.; Liu, N.; Zhang, D.; Xu, D.; and Shao, L. 2023. Salient Object Detection via Integrity Learning. *TPAMI*, 45(3): 3738–3752.