ORSATR-X: A Foundation Model based on Differential-and-Excitation Networks for Optical Remote Sensing Object Recognition

Supplementary Material

1. Overview

This material provides additional details of the proposed ORSATR-X, as well as experimental results that are omitted from the main body of this paper due to the page limit, which are organized as follows:

- Sec. 2 **Configurations of Pre-training and Fine-tuning**. Provides the full experiment configurations for pre-training and all downstream tasks, including scene classification, object detection, and semantic segmentation.
- Sec. 3 **Module Design: Background and Pseudocode**. Introduces the theoretical background of Weber's Law and the Weber Local Descriptor (WLD), and presents detailed pseudocode for the proposed WLA and MSAM modules.
- Sec. 4 **Feature Space Metric Definitions**. Provides detailed calculation formulas for the feature space metrics reported in Tab. 4(a) and Fig. 7 of the main text.

2. Configurations of Pre-training and Fine-tuning

This section presents the datasets and implementation details for both pre-training and fine-tuning.

Pre-training Configuration. We followed the default pre-training settings outlined in Tab. 1 (i). For the 50 epoch training scenario, we proportionally adjusted the warm-up phase to 5 epochs to maintain the same warmup ratio as the 100 epoch setting.

Scene Classification. To evaluate the learned representations, we conducted scene classification experiments on two common benchmarks: AID and NWPU-RESISC45, using a standard linear classifier protocol. The specific implementation details are consolidated in Tab. 1 (ii).

- NWPU-RESISC45 (RESISC-45). The RESISC-45 dataset includes 31,500 images (256×256 pixels) and features a broader GSD range of 0.5 to 30 meters. It encompasses 45 classes, each containing 700 images. In line with comparative literature, we evaluated our model under two standard data splits: training on 10% (90% test) and 20% (80% test) of the data.
- AID. The AID dataset consists of 10,000 images (600×600 pixels) with a GSD ranging from 0.5 to 8 meters. It contains 30 scene classes, with each class having between 220 and 400 samples. To align with the established protocol from MTP [4] and ensure a fair comparison, we adopted a training ratio (TR) of X% (with X=20), using the remainder for testing, as referenced in Table 1 of the main text.

Object Detection. We use the DIOR dataset to assess the performance of ORSATR-X and other RSFMs in horizontal object detection tasks. Following RVSA[3] and MTP [4], we employ Faster-RCNN as the detector, as detailed in Tab. 2 (iv).

- DIOR. This dataset consists of 23,463 visible remote sensing images with 192,472 object instances, annotated with horizontal bounding boxes across 20 common object classes. Each image of size 800×800 has a GSD ranging from 0.5 to 30 meters. The dataset is split into 5,862 training patches, 5,863 validation patches, and 11,738 test patches. Following RVSA[3] and MTP [4], we merge the training and validation sets for training, using the test set for evaluation. The high inter-class similarity and intra-class diversity pose significant challenges for precise localization and classification.

Remote sensing images include diverse objects such as buildings, vehicles, and bridges, which are densely distributed and vary in size, scale, and orientation. This makes object detection particularly challenging, especially for oriented object detection. To evaluate RSFMs on this task, we use the DIOR-R dataset and Oriented-RCNN as the detector, as detailed in Tab. 2 (iv).

- DIOR-R. This dataset uses the same images as DIOR but includes oriented bounding boxes, making it suitable for oriented object detection. Following RVSA[3] and MTP [4], we combine the training and validation sets for training, using the test set for evaluation.

For horizontal object detection and oriented object detection, we use MMDetection2 and MMRotate2 for implementation, respectively.

Semantic Segmentation. Semantic segmentation plays a crucial role in remote sensing by enabling the automatic delineation of land cover and land use. To conduct a comprehensive evaluation, we selected a prominent dataset that offers diversity in spatial resolution, spectral information, and semantic categories.

- Potsdam. A scene-level semantic segmentation dataset that contains 38 high-resolution aerial images with a GSD of 0.05 meters, each with a fixed size of 6000×6000 pixels. Following the setup in RingMoE [1], we conduct experiments using images that include near-infrared, red, and green spectral bands, focusing on five categories: "Impervious surface", "Building", "Low vegetation", "Tree", and "Car". The dataset is split into 24 images for training and 14 for testing.

Table 1. **Detailed configurations of pre-training and fine-tuning.**

Task	(i) Pre-training	(ii) Scene Classification	
Dataset	Million-AID	AID	RESISC-45
Optimizer	AdamW	AdamW	AdamW
Input Size	224×224	600×600	224×224
Input channel	RGB	RGB	RGB
Base learning rate	1e-5	5e-3	2e-3
Learning rate scheduler	Cosine Annealing	Cosine Annealing	Cosine Annealing
Weight decay	0.05	0.1	0.05
Optimizer momentum	(0.9, 0.95)	(0.9, 0.999)	(0.9, 0.999)
Batch size	512	64	256
Max iteration/epoch	50 epoch	200 epoch	200 epoch
Warmup	linear	linear	linear
Warmup iteration/epoch	5 epoch	5 epoch	10 epoch
Drop path rate	-	0.1	0.1
Augmentation	RandomResizedCrop	RandomCrop, RandomErasing	RandomCrop, RandomErasing
Head/Detector	-	Linear Classifier	Linear Classifier
Loss function	L2 norm	SoftTargetCrossEntropy	SoftTargetCrossEntropy

Table 2. **Detailed configurations of pre-training and fine-tuning.**

Task	(iii) Semantic Segmentation	(iv) Object Detection	
Dataset	Potsdam	DIOR	DIOR-R
Optimizer	AdamW	AdamW	AdamW
Input Size	512×512	800×800	800×800
Input channel	RGB	RGB	RGB
Base learning rate	1e-6	1e-4	1e-4
Learning rate scheduler	Cosine Annealing	Multistep	Multistep
Weight decay	0.005	0.05	0.05
Batch size	8	8	8
Max iteration/epoch	80k iters	12 epoch	12 epoch
Warmup	linear	linear	linear
Warmup iteration/epoch	1.5k iters	0.5k iters	0.5k iters
Warmup ratio	3e-5	1e-3	1e-3
Drop path rate	0.1	0.1	0.1
Augmentation	RandomScaling (0.5 to 2.0), RandomCrop, RandomFlip	RandomFlip, AutoAugment	RandomFlip
Head/Detector	UperNet	Faster-RCNN	Oriented-RCNN
Loss function	CrossEntropy	CrossEntropy, L1, GIoU	CrossEntropy, SmoothL1,

Our experimental setup for this task leverages the MMSegmentation codebase, utilizing UperNet as segmentation head. The full suite of fine-tuning hyperparameters is documented in Tab. 2 (iii).

3. Module Design: Background and Pseudocode

3.1. Weber's Law and WLD

Weber's Law [2], a fundamental principle in psychophysics, states that the just-noticeable difference (JND) in a stimulus is proportional to the original stimulus intensity:

$$\frac{\Delta I}{I} = k, \tag{1}$$

where ΔI denotes the increment threshold, I is the initial stimulus intensity, and k is a constant known as the Weber fraction. This law implies that human perception is governed by *relative* rather than absolute changes in stimuli.

Building upon this principle, Chen *et al.* [2] proposed the Weber Local Descriptor (WLD), which consists of two complementary components: **differential excitation** (ξ) and **orientation** (θ).

Differential Excitation. For a pixel x_c with p neighbors $\{x_0, x_1, \ldots, x_{p-1}\}$, the differential excitation is defined as:

$$\xi(x_c) = \arctan\left[\sum_{i=0}^{p-1}\left(\frac{x_i - x_c}{x_c}\right)\right]. \tag{2}$$

This formulation first computes the sum of intensity ratios between each neighbor and the center pixel (following the Weber fraction), and then applies the arctangent function to bound the output within $(-\pi/2, \pi/2)$. Intuitively, a positive ξ indicates that the surroundings are brighter than the center pixel, while a negative ξ indicates the opposite.

In practice, this computation can be decomposed into two filtering operations. The first filter f_{00} computes the sum of differences between neighbors and the center pixel:

$$v_s^{00} = \sum_{i=0}^{p-1}(x_i - x_c), \tag{3}$$

which, for a 3×3 neighborhood ($p = 8$), corresponds to a convolution kernel:

$$f_{00} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{4}$$

The second filter f_{01} simply outputs the original pixel intensity $v_s^{01} = x_c$. The differential excitation is then:

$$\xi(x_c) = \arctan\left(\frac{v_s^{00}}{v_s^{01}}\right). \tag{5}$$

Orientation. The orientation component captures the gradient direction of each pixel:

$$\theta(x_c) = \arctan\left(\frac{v_s^{11}}{v_s^{10}}\right), \tag{6}$$

where v_s^{10} and v_s^{11} are the outputs of horizontal and vertical gradient filters f_{10} and f_{11}, respectively:

$$f_{10} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad f_{11} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{7}$$

3.2. Pseudocode for WLA and MSAM

We provide detailed pseudocode for the Weber Local Adapter (Alg. 1) and the Multi-scale Alignment Module (Alg. 2) to facilitate reproducibility.

Algorithm 1 Weber Local Adapter (WLA)

Require: Input features $\mathbf{F}_i \in \mathbb{R}^{C\times H\times W}$ from the i-th Transformer block
Ensure: Output features $\mathbf{F}^i_{\text{WLA}} \in \mathbb{R}^{C\times H\times W}$

// —- Center-surround contrast branch (CDC) —-
$\mathbf{K}^i_{cs} \leftarrow$ Learnable depthwise kernel $\in \mathbb{R}^{C\times 1\times 3\times 3}$ with center fixed to -1
$\mathbf{F}^i_{cs} \leftarrow \alpha^i \cdot \text{DWConv}(\mathbf{F}_i;\ \mathbf{K}^i_{cs})$ ▷ **Eq. (1)**
$\mathbf{F}^{i,+}_{cs} \leftarrow \text{ReLU}(\mathbf{F}^i_{cs});\ \ \mathbf{F}^{i,-}_{cs} \leftarrow \text{ReLU}(-\mathbf{F}^i_{cs})$ ▷ **Eq. (2)**
$\mathbf{F}^i_{\text{polar}} \leftarrow \text{Concat}[\mathcal{W}^{i,+}(\mathbf{F}^{i,+}_{cs}),\ \mathcal{W}^{i,-}(\mathbf{F}^{i,-}_{cs})]$ ▷ **Eq. (3)**, $C/2{+}C/2{=}C$

// —- Directional gradient branch (HDC/VDC) —-
Construct anti-symmetric kernels $\mathbf{K}^i_h, \mathbf{K}^i_v$ ▷ **Eq. (4)**
$\mathbf{F}^i_h \leftarrow \text{Conv2D}(\mathbf{F}_i;\ \mathbf{K}^i_h);\ \ \mathbf{F}^i_v \leftarrow \text{Conv2D}(\mathbf{F}_i;\ \mathbf{K}^i_v)$
$\mathbf{F}^i_{\text{grad}} \leftarrow \mathcal{W}^i_{\text{grad}}(\text{Concat}[\mathbf{F}^i_h,\ \mathbf{F}^i_v])$ ▷ **Eq. (5)**, $2C \to C$

// —- Adaptive branch fusion —-
$\mathbf{F}^{i,r}_{\text{polar}} \leftarrow \mathcal{W}^i_l(\mathbf{F}^i_{\text{polar}});\ \ \mathbf{F}^{i,r}_{\text{grad}} \leftarrow \mathcal{W}^i_r(\mathbf{F}^i_{\text{grad}})$ ▷ **Eq. (6)**, $C \to C/2$
$\mathbf{F}^i_{\text{cat}} \leftarrow \text{Concat}[\mathbf{F}^{i,r}_{\text{polar}},\ \mathbf{F}^{i,r}_{\text{grad}}]$
$\mathbf{S}^i \leftarrow \text{Concat}[\text{MeanPool}(\mathbf{F}^i_{\text{cat}}),\ \text{MaxPool}(\mathbf{F}^i_{\text{cat}})]$
$\mathbf{A}^i \leftarrow \sigma(\text{Conv}_{7\times 7}(\mathbf{S}^i))$ ▷ **Eq. (7)**, $\mathbf{A}^i \in \mathbb{R}^{2\times H\times W}$
$\mathbf{F}^i_{\text{fused}} \leftarrow \mathbf{A}^i[0] \odot \mathbf{F}^{i,r}_{\text{polar}} + \mathbf{A}^i[1] \odot \mathbf{F}^{i,r}_{\text{grad}}$ ▷ **Eq. (8)**

// —- Output: multiplicative modulation —-
$\mathbf{F}^i_{\text{WLA}} \leftarrow \mathbf{F}_i \odot \mathcal{W}^i_{\text{out}}(\mathbf{F}^i_{\text{fused}})$ ▷ **Eq. (9)**, $C/2 \to C$
return $\mathbf{F}^i_{\text{WLA}}$

4. Feature Space Metric Definitions

This section provides detailed calculation formulas for the feature space metrics reported in **Tab. 4(a)** and **Fig. 7** of the main text.

Convex Hull Area. Measures spatial coverage of features in 2D PCA space. Given N projected points $\{\mathbf{z}_i\}_{i=1}^N$ where $\mathbf{z}_i \in \mathbb{R}^2$, we compute the convex hull $\mathcal{H}$ and calculate its area using the Shoelace formula with vertices $\{\mathbf{v}_j = (x_j, y_j)\}_{j=1}^M$:

$$\text{Hull Area} = \frac{1}{2}\left|\sum_{j=1}^{M}(x_j y_{j+1} - x_{j+1} y_j)\right| \tag{8}$$

where indices wrap (i.e., $\mathbf{v}_{M+1} = \mathbf{v}_1$). Larger areas indicate more diverse and expressive features.

Intra-class Distance. Measures feature compactness as average distance to centroid. For features $\{\mathbf{f}_i\}_{i=1}^N$ with centroid $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^N \mathbf{f}_i$:

$$\text{Intra-dist.} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{f}_i - \boldsymbol{\mu}\|_2 \tag{9}$$

Algorithm 2 Multi-scale Alignment Module (MSAM)

Require: Input features $\mathbf{F}^i_{\text{WLA}} \in \mathbb{R}^{C\times H\times W}$ from the i-th layer's WLA
Ensure: Output features $\mathbf{F}^i_{\text{MSAM}} \in \mathbb{R}^{C\times H\times W}$

// —- Multi-scale feature extraction —-
$\mathbf{F}^i_k \leftarrow \text{DWConv}_k(\mathbf{F}^i_{\text{WLA}}),\quad k \in \{3,5,7\}$ ▷ **Eq. (10)**

// —- Dynamic scale weighting —-
$\mathbf{z}^i \leftarrow \text{GAP}(\mathbf{F}^i_{\text{WLA}})$ ▷ $\mathbb{R}^{C\times 1\times 1}$
$\mathbf{w}^i_s \leftarrow \text{Softmax}(\mathcal{G}^i(\mathbf{z}^i))$ ▷ MLP: $C \to C/4 \to 3$
$\mathbf{F}^i_{\text{scale}} \leftarrow \sum_k \mathbf{w}^{i,(k)}_s \cdot \mathbf{F}^i_k$ ▷ **Eq. (11)**

// —- Spatial attention —-
$\mathbf{A}^i_{\text{spatial}} \leftarrow \sigma(\text{Conv}_{1\times1}(\text{GELU}(\text{LN}(\text{Conv}_{1\times1}(\mathbf{F}^i_{\text{scale}})))))$ ▷ **Eq. (12)**, $C \to C/8 \to 1$

// —- Dual residual connections —-
$\mathbf{F}^i_{\text{att}} \leftarrow \mathbf{A}^i_{\text{spatial}} \odot \mathbf{F}^i_{\text{scale}} + \mathbf{F}^i_{\text{WLA}}$ ▷ **Eq. (13)**, Residual 1
$\mathbf{F}^i_{\text{MSAM}} \leftarrow \text{GELU}(\text{LN}(\mathcal{P}^i(\mathbf{F}^i_{\text{att}}))) + \mathbf{F}^i_{\text{WLA}}$ ▷ **Eq. (13)**, Residual 2
return $\mathbf{F}^i_{\text{MSAM}}$

Higher values reflect greater internal feature variation and richer representations.

Effective Dimensionality. Quantifies feature space complexity using PCA variance. Given eigenvalues $\{\sigma_i^2\}_{i=1}^D$ (sorted descending), the participation ratio is:

$$\text{Eff. dim.} = \frac{(\sum_{i=1}^D \sigma_i^2)^2}{\sum_{i=1}^D (\sigma_i^2)^2} = \frac{1}{\sum_{i=1}^D \rho_i^2} \tag{10}$$

Equivalently, using variance ratios $\rho_i = \sigma_i^2 / \sum_j \sigma_j^2$:

$$\text{Eff. dim.} = \exp\left(-\sum_{i=1}^D \rho_i \log \rho_i\right) = \frac{1}{\sum_{i=1}^D \rho_i^2} \tag{11}$$

Range: $[1, D]$. Higher values indicate more evenly distributed variance and less redundant features. The full model's lower Eff. dim. suggests concentrated variance in key directions, reflecting efficient feature learning.

References

[1] Hanbo Bi, Yingchao Feng, Boyuan Tong, Mengyu Wang, Haichen Yu, Yongqiang Mao, Hao Chang, Wenhui Diao, Peijin Wang, Yue Yu, Hanyang Peng, Yehong Zhang, Kun Fu, and Xian Sun. RingMoE: Mixture-of-Modality-Experts Multi-Modal Foundation Models for Universal Remote Sensing Image Interpretation. *arXiv preprint arXiv:2504.03166*, 2025. 2

[2] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikäinen, Xilin Chen, and Wen Gao. WLD: A Robust Local Image Descriptor. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010. 3, 4

[3] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing Plain Vision Transformer Toward Remote Sensing Foundation Model. *IEEE Trans. Geosci. Remote Sens.*, 2023. 2

[4] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, and Liangpei Zhang. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, pages 1–24, 2024. 2