Any-Optical-Model: A Universal Foundation Model for Optical Remote Sensing

Xuyang Li[2,3], **Chenyu Li**[1*], **Danfeng Hong**[1]

[1]Southeast University, 211189 Nanjing, China
[2]Aerospace Information Research Institute, Chinese Academy of Sciences, 100094 Beijing, China
[3]School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, 100049 Beijing, China
lixuyang23@mails.ucas.ac.cn, chenyuli.erh@gmail.com, danfeng.hong@seu.edu.cn

Abstract

Optical satellites, with their diverse band layouts and ground sampling distances, supply indispensable evidence for tasks ranging from ecosystem surveillance to emergency response. However, significant discrepancies in band composition and spatial resolution across different optical sensors present major challenges for existing Remote Sensing Foundation Models (RSFMs). These models are typically pretrained on fixed band configurations and resolutions, making them vulnerable to real world scenarios involving missing bands, cross sensor fusion, and unseen spatial scales, thereby limiting their generalization and practical deployment. To address these limitations, we propose Any Optical Model (AOM), a universal RSFM explicitly designed to accommodate arbitrary band compositions, sensor types, and resolution scales. To preserve distinctive spectral characteristics even when bands are missing or newly introduced, AOM introduces a spectrum-independent tokenizer that assigns each channel a dedicated band embedding, enabling explicit encoding of spectral identity. To effectively capture texture and contextual patterns from sub-meter to hundred-meter imagery, we design a multi-scale adaptive patch embedding mechanism that dynamically modulates the receptive field. Furthermore, to maintain global semantic consistency across varying resolutions, AOM incorporates a multi-scale semantic alignment mechanism alongside a channel-wise self-supervised masking and reconstruction pretraining strategy that jointly models spectral-spatial relationships. Extensive experiments on over 10 public datasets, including those from Sentinel-2, Landsat, and HLS, demonstrate that AOM consistently achieves state-of-the-art (SOTA) performance under challenging conditions such as band missing, cross sensor, and cross resolution settings. These results highlight AOM as a crucial step toward building truly general-purpose RSFMs.

Introduction

Spectral remote sensing (RS) data (Hong et al. 2025) is a cornerstone of optical Earth observation, leveraging its ability to capture rich spectral and spatial information to deliver significant value in applications such as environmental monitoring, disaster response, and precision agriculture. Compared to single RGB imagery, spectral data offers higher spectral resolution and inherently exhibits two key characteristics: band heterogeneity and scale diversity. Different satellites vary significantly in band count, central wavelength, and bandwidth, while spatial resolutions span from sub-meter to hundred-meter scales, with differences reaching orders of magnitude (Li, Hong, and Chanussot 2024). In recent years, the integration of artificial intelligence in the RS field has led to the development of various optical foundation models to support downstream tasks such as land cover classification, change detection, and object recognition. However, existing RSFMs are typically pretrained on fixed band configurations and spatial scales, implicitly assuming complete band availability and constant resolution. This assumption limits their performance in real-world scenarios involving (1) missing or additional bands, (2) cross-satellite data, or (3) unseen resolutions, severely constraining their generalization and practical utility, as described in Figure 1.

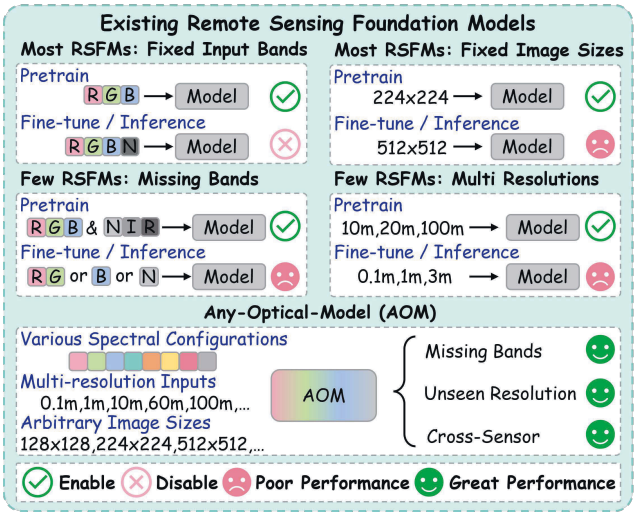

Figure 1: **Limitations of current remote sensing foundation models.** Mainstream models fail to adapt to varying spectral bands, spatial resolutions, and image sizes across pretraining and downstream tasks.

Current mainstream RSFMs (Guo et al. 2024; Astruc et al. 2025) typically process multispectral data by treating

*Corresponding author.

it as a monolithic input, overlooking differences in physical properties and complex inter-band interactions. Some models, such as DOFA (Xiong et al. 2024) and SpectralGPT (Hong et al. 2024), incorporate sophisticated designs that enable their structures to accommodate data inputs from different channels. However, in real-world scenarios, spectral data often exhibits significant heterogeneity due to sensor variations or missing bands. For instance, when certain bands are unavailable, models like DOFA struggle to effectively extract features from the remaining bands, resulting in degraded performance. Additionally, most existing models are designed for specific resolutions, employing single-scale patch embedding that fails to adapt to the texture and contextual modeling required for unseen resolutions. RS imagery spans a vast range of resolutions, from sub-meter to hundreds of meters. While recent studies have attempted to address scale variability, they primarily focus on high-resolution RGB imagery and do not adequately account for the inter-band spatial feature correlations and sensor-specific differences inherent in multispectral data (Reed et al. 2023; Noman et al. 2024). Scale variations can also disrupt global semantic understanding, leading to unstable performance in multi-scale scenarios. Consequently, models often require retraining or become ineffective when processing data from different sensors or resolutions. **How can we design a model that seamlessly adapts to arbitrary band compositions, sensor types, and resolution scales?**

To overcome these limitations, we introduce Any-Optical-Model (***AOM***), an RSFM that unifies spectral and spatial feature learning within a single framework. First, to tolerate missing or additional bands, enhance cross-sensor transfer, and support arbitrary spectral configurations, *AOM* introduces a *Spectrum-independent Tokenizer* that performs channel-wise patch embedding augmented with channel index encoding, preserving per-band information while allowing nonlinear inter-band interactions in later layers. Second, to ensure robust feature capture across resolutions from sub-meter to hundred-meter scales, a *Multi-scale Adaptive Patch Embedding* based on pseudo-inverse resize dynamically adjusts its receptive field to accommodate resolutions from sub-meter to hundred-meter, allowing *AOM* to adapt its spatial granularity to the input resolution without sacrificing fine details or global context. Third, a self-supervised routine that masks and reconstructs individual channels teaches *AOM* fine-grained spectral signatures and local spatial structure, yielding resilience to incomplete data. Finally, to maintain semantic consistency across diverse resolutions and imaging conditions, a multi-scale alignment constraint regularizes global representations. We validate *AOM* on Geo-Bench and multiple mainstream datasets, including Sentinel-2, Landsat, and HLS. The results show that our model significantly outperforms existing models in downstream tasks, particularly in scenarios involving missing bands, cross-sensor data, and varying resolutions, showcasing its superior generalization.

Overall, our contributions in this work are as follows: (1) We propose *AOM*, a foundation model that adapts to arbitrary spectral channels, spatial resolutions, and sensor types, thereby overcoming the limitations of existing models. (2) We introduce spectrum-independent tokenizer to enhance cross-sensor generalization and a multi-scale patch embedding to enable resolution-adaptive feature extraction. (3) We develop channel-wise masked feature learning and reconstruction, alongside a global semantic alignment mechanism, to ensure multi-scale semantic consistency while improving the efficacy of self-supervised pretraining. (4) We validate *AOM* on over 10 datasets and benchmarks, showing its superior performance and robustness in diverse scenarios.

Related Work

Remote sensing foundation model. RSFMs are typically self-supervised networks pre-trained on large-scale satellite corpora and fine-tuned for diverse downstream tasks. Recent progress follows two main directions. One focuses on tailoring pretraining to RS-specific characteristics, such as multispectral inputs, small objects, and heterogeneous sensors. For instance, SatMAE (Cong et al. 2022) leverages masked autoencoding to preserve cross-band coherence, Skysense (Guo et al. 2024) fuses multi-source data with spatiotemporal cues via contrastive learning, SeaMo (Li et al. 2025b) incorporates seasonal attributes, and AnySat (Astruc et al. 2025) adapts I-JEPA (Assran et al. 2023) to handle resolution diversity. The other line integrates vision-language models with geospatial priors: GeoChat (Kuckreja et al. 2024) uses instruction tuning for scene understanding, while GeoPixel (Shabbir et al. 2025) introduces pixel-level decoding with adaptive partitioning. Despite their flexibility, these approaches primarily emphasize task-level customization, keeping core architectures static, potentially limiting generalization across the full spectrum of RS data diversity.

Channel vision transformer. Vision Transformer (ViT) (Dosovitskiy et al. 2020) has become the core architecture for foundation models by leveraging self-attention for global image understanding. However, standard ViTs are limited to predefined settings such as natural RGB images and struggle with multi-spectral data, where spectral bands vary across sensors, as seen in biomedical and remote sensing domains. To address this, several channel-aware variants have been proposed. ChannelViT (Bao, Sivanandan, and Karaletsos 2024) introduces per-channel tokens with hierarchical sampling; ChAda-ViT (Bourriez et al. 2024) adds inter-channel attention for improved spatial interaction; DiChaViT (Pham and Plummer 2024) employs a channel-diversity loss to retain band-specific features; and ChA-MAEViT (Pham, Caicedo, and Plummer 2025) enhances self-supervised learning through dynamic channel masking and memory tokens. While these methods improve multi-spectral handling, achieving broad adaptability across diverse imaging modalities remains unresolved.

Methodology

Overview of the Proposed AOM

To address the critical domain gap limitation in current RSFMs caused by spectral and spatial-scale discrepancies between pretraining and downstream tasks, we propose *AOM*, a novel architecture with three key innovations. As described in Figure 2, we first design a spectrum-independent

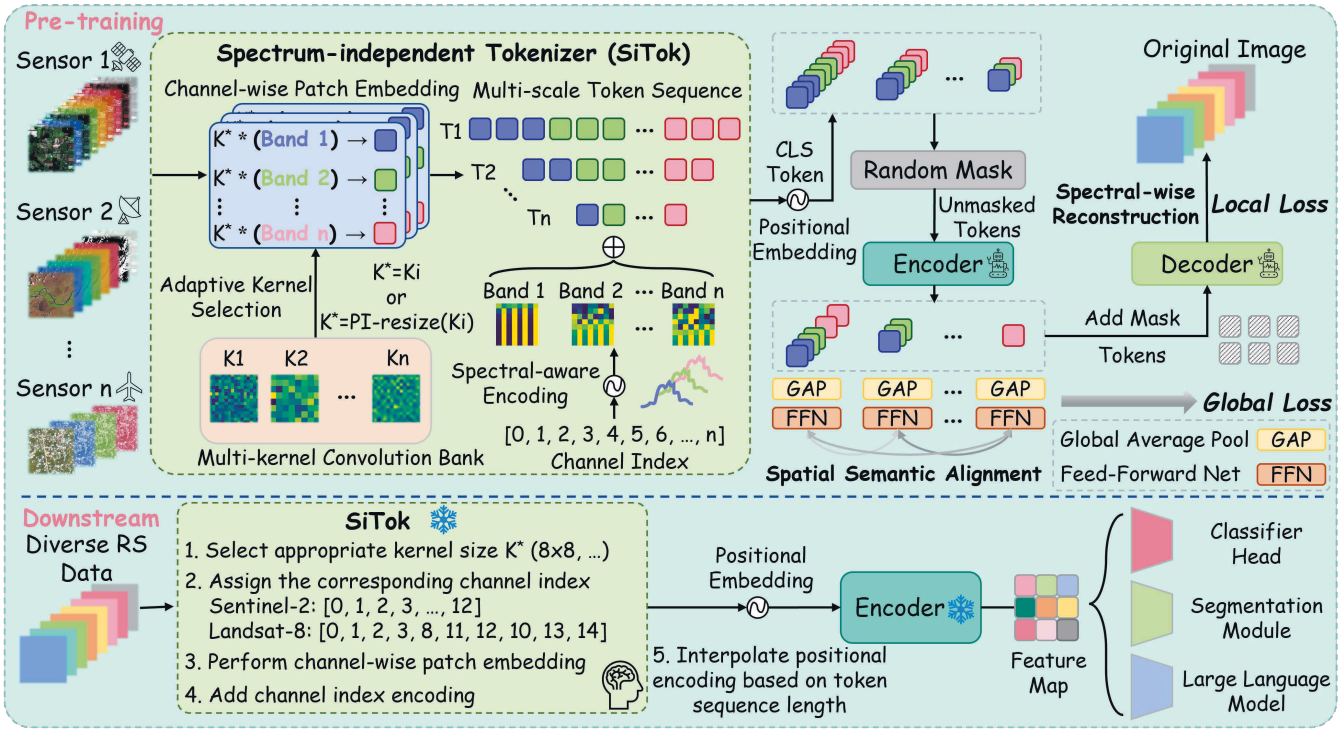

Figure 2: **An illustration of the proposed model.** AOM unifies spectral and spatial modeling through channel-wise patch embedding, adaptive multi-scale extraction, spectral-wise masking & reconstruction, and multi-scale alignment, enabling flexible band configurations and robust performance across resolutions.

tokenizer that processes each spectral band individually through dedicated tokenization while incorporating channel index encoding for spectrum-aware representation. Second, to handle varying image resolutions and sizes, we develop a multi-scale patch embedding strategy based on pseudo-inverse resize transformation (Beyer et al. 2023), enabling adaptive patch representation across different spatial scales. Third, for effective large-scale pretraining, we enhance the Masked Autoencoder framework with two specialized components: (1) channel-wise reconstruction to capture spectral sequential properties, and (2) semantic consistency constraints across different scale embeddings.

Spectrum-independent Tokenizer (SiTok)

Conventional RSFMs typically process input images $\mathbf{I} \in \mathbb{R}^{C\times H\times W}$ with a fixed number of spectral channels C, limiting their flexibility for multi-modal applications. To enable robust processing of arbitrary spectral configurations, including handling missing or additional channels, we propose a novel *Spectrum-independent Tokenizer* (**SiTok**) with two components.

Channel-wise patch embedding. For each spectral channel $\mathbf{S}_i \in \mathbb{R}^{1\times H\times W}$, we apply a shared convolutional kernel $\mathbf{K} \in \mathbb{R}^{P\times P\times 1\times D}$ to extract patch embeddings: $\mathbf{I}_P = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_C\}$, where $\mathbf{I}_i \in \mathbb{R}^{(N_H\times N_W)\times D}$. Here, P denotes the patch size, D denotes the embedding dimension, and $N_H = H/P$, $N_W = W/P$ represent the spatial dimensions of the token grid. This design ensures consistent processing regardless of the input's channel count.

Spectral-aware encoding. We augment each channel's tokens with spectral-aware positional encodings: $\mathrm{PE}_{\text{channel}} = \{\mathrm{PE}_1, \mathrm{PE}_2, \ldots, \mathrm{PE}_C\}$, $\mathrm{PE}_i \in \mathbb{R}^{1\times D}$, where each PE_i is generated via sinusoidal encoding of the channel index i. The encoding is then broadcast to $\mathbb{R}^{(N_H\times N_W)\times D}$ and added to the corresponding channel: $\mathbf{I}_i + \mathrm{PE}_i \rightarrow \mathbf{I}_i$. This encoding preserves spectral ordering information while maintaining permutation invariance.

The final token sequence $\mathbf{T} \in \mathbb{R}^{(C\times N_H\times N_W)\times D}$ concatenates the tokens from all channels and forms a spectrally-aware representation suitable for transformer processing. Compared to fixed-channel tokenizers, our approach offers two key advantages: (1) inherent adaptability to varying channel configurations, (2) preservation of spectral characteristics through explicit encoding.

Multi-scale Adaptive Patch Embedding (MAPE)

Conventional patch embedding layers employ a single convolutional kernel $\mathbf{K}_{\text{old}} \in \mathbb{R}^{P\times P\times C\times D}$ with a fixed patch size P. Since RS images span a broad range of spatial resolutions and scene scales, a single P cannot capture both fine-grained textures and coarse contextual structures. Previous work (Li et al. 2025a), therefore, resizes the kernel via the pseudo-inverse resize (PI-resize) operator $\mathbf{K}_{\text{new}} = \text{PI-resize}(\mathbf{K}_{\text{old}}, r)$, but large scale factors r introduce interpolation errors that degrade performance. To overcome this limitation, we introduce *Multi-scale Adaptive Patch Embed-*

ding (**MAPE**), which extends SiTok with the following components.

Multi-scale convolutional bank. MAPE maintains a bank of n convolutions with different receptive fields: $\mathcal{K} = \{\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n\}, \quad \mathbf{K}_i \in \mathbb{R}^{P_i \times P_i \times C \times D}$, where P_i is the kernel size of the i-th branch. Small P_i ($P_i \to 1$) excels at capturing high-frequency details, whereas large P_i ($P_i \uparrow$) gathers global context; the model thus adapts its computational footprint and receptive field to the input.

Adaptive kernel selection. For imagery that demands fine spatial detail, AOM employs a small convolutional kernel; conversely, applications that emphasise global context or higher throughput adopt a larger kernel. Given a target patch size P_t, the closest kernel in the bank is

$$i^\star \;=\; \arg\min_i |P_i - P_t|, \tag{1}$$

and the effective weights are

$$\mathbf{K}^\star = \begin{cases} \mathbf{K}_{i^\star}, & P_{i^\star} = P_t, \\ \text{PI-resize}\big(\mathbf{K}_{i^\star}, r\big), & P_{i^\star} \neq P_t, \end{cases} \qquad r = \frac{P_t}{P_{i^\star}}. \tag{2}$$

PI-resize definition. Let $\boldsymbol{\omega} = \text{vec}(\mathbf{K_i})$ vectorise the $P \times P$ spatial dimensions and $B^r \in \mathbb{R}^{(rP)^2 \times P^2}$ denote the bilinear-resize matrix. PI-resize projects $\boldsymbol{\omega}$ through the Moore–Penrose pseudo-inverse of B^r:

$$\text{PI-resize}(\mathbf{K_i}, r) \;=\; \text{reshape}\Big((B^{r\,\top})^{+}\,\boldsymbol{\omega}\Big), \tag{3}$$

yielding the unique solution that minimises the expected response distortion $\mathbb{E}_{\mathbf{x}}\big[\langle \mathbf{x}, \mathbf{K_i}\rangle - \langle B^r\mathbf{x}, \mathbf{K_i}'\rangle\big]^2$. Because PI-resize is applied only to the nearest kernel, the approximation error is bounded, ensuring stable feature extraction for arbitrary P_t.

Integration with SiTok. **MAPE** is a plug-and-play module that can augment any tokenizer. To insert it into **SiTok** we instantiate the multi-scale convolutional bank for a single-channel input, $\mathcal{K} = \{\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n\}, \mathbf{K}_i \in \mathbb{R}^{P_i \times P_i \times 1 \times D}$. Adaptive kernel selection provides the resolution-matched weights $\mathbf{K}^\star$, which are convolved with the input image $\mathbf{I}$ to yield channel-wise patch tokens; these tokens are subsequently enriched by the *Spectral-aware Encoding* inherited from **SiTok**.

Semantic Alignment Pretraining Task

The data-ingest module of *AOM* couples **SiTok** with our **MAPE**. To exploit large-scale RS corpora without manual labels, We design a dual-objective self-supervised approach combining Masked Autoencoding and Contrastive Learning to capture both local spatial-spectral patterns and global semantic alignment.

For an input image $\mathbf{I} \in \mathbb{R}^{C \times H \times W}$, the **SiTok** and **MAPE** modules generate n parallel token sequences at different scales: $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n\}$, where $\mathbf{T}_i \in \mathbb{R}^{(N_i \times D)}$, where $N_i = C \times (H/P_i) \times (W/P_i)$ and P_i denotes the patch size at scale i. We then randomly mask portions of each token sequence with masking ratio $m \in (0, 1)$, preserving only the unmasked tokens $\mathbf{T}_i^{\text{vis}} \in \mathbb{R}^{(N_i \times (1-m)) \times D)}$ for encoder processing:

$$\begin{aligned} \mathbf{E}_i &= \text{Encoder}(\mathbf{T}_i^{\text{vis}}), \\ \text{where} \quad \mathbf{E}_i &\in \mathbb{R}^{(N_i \times (1-m)) \times D}. \end{aligned} \tag{4}$$

The encoded features undergo parallel optimization through two complementary learning objectives:

Masked spectral–spatial reconstruction. For every scale i we instantiate an individual decoder Decoder_i. We first augment the encoder output $\mathbf{E}_i$ with learnable mask tokens $\mathbf{M} \in \mathbb{R}^{(N_i \times m) \times D}$ and then feed the concatenated sequence into the corresponding decoder: $\hat{\mathbf{T}}_i = \text{Decoder}_i\big(\text{Concat}(\mathbf{E}_i, \mathbf{M})\big)$. The reconstruction loss is the pixel-wise mean-squared error (MSE) between the predicted and ground-truth patches across all spectral channels:

$$\mathcal{L}_{\text{Recon}} = \frac{1}{n}\sum_{i=1}^{n} \big\|\hat{\mathbf{T}}_i - \mathbf{T}_i^{\text{masked}}\big\|_2^2. \tag{5}$$

Multi-scale semantic alignment. To ensure that tokens derived from different patch scales encode consistent semantics, we introduce a multi-scale alignment branch. We apply global average pooling (GAP) followed by non-linear projection heads $g_i(\cdot)$:

$$\mathbf{H}_i = g_i(\text{GAP}(\mathbf{E}_i)), \quad \mathbf{H}_i \in \mathbb{R}^d \tag{6}$$

The alignment loss encourages feature similarity across scales using InfoNCE:

$$\mathcal{L}_{\text{Align}} = -\sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \log \frac{\exp(s(\mathbf{H}_i, \mathbf{H}_j)/\tau)}{\sum_{k=1}^{n}\exp(s(\mathbf{H}_i, \mathbf{H}_k)/\tau)} \tag{7}$$

where $s(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

Composite objective function. The total training objective combines both losses with balanced weighting:

$$\mathcal{L}_{\text{Total}} = \lambda_1\mathcal{L}_{\text{Recon}} + \lambda_2\mathcal{L}_{\text{Align}} \tag{8}$$

where λ_1 and λ_2 control the relative importance of reconstruction fidelity versus cross-scale semantic alignment. This dual supervision compels *AOM* to model local spectral-spatial correlations through channel-wise reconstruction while simultaneously aligning global semantics across multiple patches scales, providing strong generalization for downstream RS tasks.

Apply AOM on Diverse RS Data

After pretraining, *AOM* is deployed to diverse optical RS datasets (Fig. 2). Given the spatial resolution and image size, we first select a suitable patch-embedding kernel to balance detail and computation. We then assign channel indices according to the sensor's spectral configuration (e.g., Sentinel-2's 13 bands are indexed 0–12, and RGB or Landsat data select their corresponding indices). Since *AOM* models channels independently, missing or newly added bands do not affect its ability to exploit previously learned ones. Patch embedding is applied with the chosen kernel, channel-index encoding is added, and the positional encoding is interpolated to match the resulting token sequence. These tokens are finally fed into the pretrained encoder for feature extraction in downstream tasks such as classification, segmentation, or as visual input to a large language model.

Model	m-pv4ger-seg (RGB) 320 × 320	m-nz-cattle (RGB) 500 × 500	m-neonTree (RGB) 400 × 400	m-cashew-plant. (Sentinel-2) 256 × 256	m-SA-crop-type (Sentinel-2) 256 × 256	m-chesapeake (RGBN) 256 × 256	Overall
Random init.	76.6	67.5	46.9	37.1	26.7	38.9	48.95
MAE (single)	85.4	70.1	53.0	45.7	27.5	42.1	53.96
Scale-MAE	78.4	72.7	51.0	—	—	48.9	—
GFM	84.8	71.8	51.1	—	—	54.1	—
SenPaMAE	81.3	75.3	50.5	48.7	28.1	51.7	55.93
CROMA	—	—	—	<u>55.6</u>	**31.4**	—	—
AnySat	<u>90.2</u>	76.9	51.8	52.3	29.5	<u>55.3</u>	59.33
DOFA	89.1	<u>77.8</u>	<u>53.2</u>	53.8	29.0	53.8	<u>59.45</u>
AOM (ours)	**91.5** (+1.3)	**80.2** (+2.4)	**53.7** (+0.5)	**68.3** (+12.7)	<u>31.0</u>	**59.2** (+3.9)	**63.98** (+4.53)

Table 1: Partial fine-tuning results (mIoU) on the six Geo-Bench semantic-segmentation tasks. All backbones are frozen, and a UPerNet head is trained for 20 epochs. **Bold** numbers denote the best performance, <u>underlined</u> numbers mark the second-best, and the values in parentheses indicate the absolute improvement of our method over the respective second-best model.

Experiments

Our experimental section begins with the pretraining configuration detailing the multi-source datasets and optimization strategy. We then comprehensively assess *AOM* on the Geo-Bench benchmark and diverse cross-sensor datasets, demonstrating superior performance across all metrics. Rigorous ablation studies further validate the model's robustness to varying image sizes, spatial resolutions, and spectral configurations, consistently outperforming existing RSFMs across all test scenarios.

Pretraining Configurations

Pretraining dataset. Our pretraining dataset comprises multi-source remote sensing imagery carefully selected to enable the model to learn cross-sensor characteristics during pretraining. We employ a diverse collection of multi-sensor remote sensing datasets for pretraining, including: (1) Sentinel-2 imagery from SSL4EO-S12 (1.004 million samples at 10-60m resolution) (Wang et al. 2023), (2) Landsat-8 data from Activefire (146k samples at 30-100m resolution) (de Almeida Pereira et al. 2021), and (3) high-resolution RGB collections from GeoPile (Mendieta et al. 2023), fMoW (Christie et al. 2018), and OpenEarthMap (Xia et al. 2023) (108k samples at 0.1-30m resolution). This combined dataset spans resolutions from 0.1m to 100m, totaling approximately 1.56 million samples across optical, multi-spectral, and high-resolution domains to ensure a comprehensive representation of remote sensing scenarios.

Pretraining details. During the pertaining, we run 220 epochs on the pretraining corpus with a batch size of 1024 and a base learning rate of 1×10^{-4}. The multi-scale convolutional bank is initialized with kernel sizes $\{16, 32, 64\}$, while training sequentially cycles through patch sizes $\{16, 24, 32, 48, 64\}$, hence, five independent decoders are employed for reconstruction. Our *adaptive kernel selection* mechanism dynamically resizes each convolutional kernel so that its receptive field is always aligned with the currently sampled patch size. The encoder of *AOM* is based on the ViT-Base architecture, with 4 decoder layers applied. The image masking ratio is set to 75%. For the loss function, since the multi-scale features extracted by *AOM* are inherently positive pairs, we set the temperature parameter of the InfoNCE loss to 0.5 to strengthen similarity learning. Furthermore, because the overall magnitude of InfoNCE loss is lower than that of MSE loss, we assign weights of 0.8 and 0.2 to the InfoNCE and MSE losses, respectively, to balance their contributions during training. For data augmentation, we employ only simple augmentations, including random horizontal flipping and random cropping. Notably, the cropped images are resized back to the original sizes of each dataset. This means that the input image size during pretraining is not fixed, but rather varies according to the native resolution of each dataset image.

Model	SPARCS (Landsat-8)	HLS Burn Scars (HLS)
SatMAE (NeurIPS'22)	49.9	81.1
CROMA (NeurIPS'23)	52.3	<u>82.4</u>
SpectralGPT (TPAMI'24)	<u>57.6</u>	80.5
DOFA (ArXiv'24)	55.4	80.6
SeaMo (INF'25)	51.7	81.8
AOM (ours)	**68.5** (+10.9)	**85.4** (+3.0)

Table 2: Partial fine-tuning results (mIoU) on two cross-sensor segmentation tasks. **Bold** marks the best, <u>underline</u> denotes the second-best, and parentheses report AOM's absolute gain over the latter.

Downstream Evaluation

Geo-Bench evaluation. We benchmark our model on the six semantic-segmentation datasets of **Geo-Bench** (Lacoste et al. 2023), a suite that deliberately covers multiple sensors. Following the official protocol, we freeze the backbone for every method and train a UPerNet segmentation head (Xiao et al. 2018) for 20 epochs. As shown in Table 1, our approach establishes new state-of-the-art (SOTA) results on five out of six datasets and delivers sizable gains over the second-

Model	UCM (RGB)	BigEarthNet (Sentinel-2)
SatMAE (NeurIPS'22)	85.17	79.36
CROMA (NeurIPS'23)	—	83.41
SpectralGPT (TPAMI'24)	82.42	81.05
DOFA (ArXiv'24)	90.09	82.45
AnySat (CVPR'25)	88.92	82.79
AOM (ours)	**93.57** (+3.48)	**85.02** (+1.61)

Table 3: Linear probing (LP) accuracy (%) on two classification datasets: overall accuracy on UCM and mAP on BigEarthNet. **Bold** marks the best score, underline the second-best; numbers in parentheses give AOM's absolute gain over the latter.

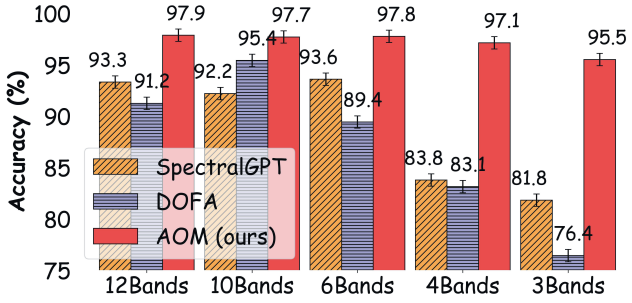

Figure 3: **Linear probing accuracy (%) on EuroSAT across different band combinations.** EuroSAT images are captured by Sentinel-2 and contain 13 spectral bands (indexed 0–12). The y-axis represents classification accuracy, while the x-axis indicates the number of spectral bands used during training.

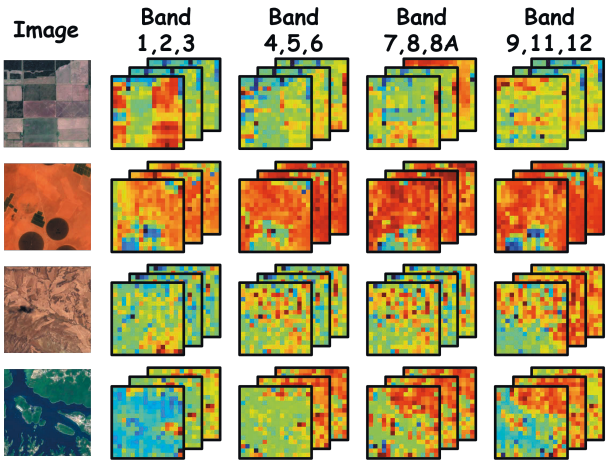

Figure 4: **Visualization of feature maps from different Sentinel-2 spectral bands through SiTok.** The results demonstrate the model's ability to extract features independently from each band.

best model; for instance, it improves the cashew-plantation

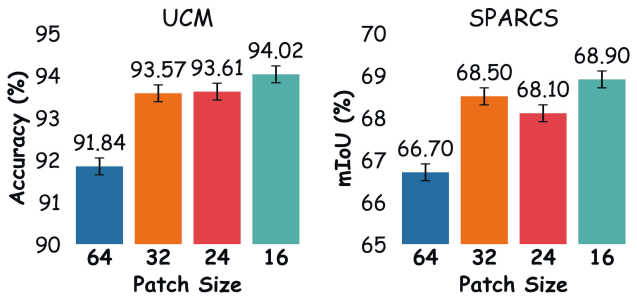

Figure 5: **Results on two datasets across different patch size configurations.** AOM maintains stable accuracy and mIoU across patch granularities, demonstrating robust performance in both classification and segmentation.

score by **+12.7 mIoU**. These findings underscore the proposed model's robustness to both resolution changes and cross-sensor shifts. Complete training details and additional metrics are reported in the supplementary material.

Diverse cross-sensor datasets. While Geo-Bench primarily covers segmentation tasks for RGB and Sentinel-2 imagery, many RSFMs are evaluated across various satellite datasets and tasks. To evaluate our model's generalization capability across different optical sensors, we conduct experiments on two representative datasets: (1) **SPARCS**: Landsat-8 cloud and cloud shadow detection (Hughes and Kennedy 2019). (2) **HLS Burn Scars**: Harmonized Landsat-Sentinel (HLS) burn scar identification (Phillips et al. 2023). Following standard practice for fair comparison with existing RSFMs, we adopt a partial fine-tuning strategy where only the segmentation head is fine-tuned. This protocol effectively evaluates the model's transfer learning capability. As shown in Table 2, our model achieves significant improvements over existing methods on both datasets, demonstrating superior cross-sensor generalization.

To assess our model's representation learning capability, we evaluate through linear probing (LP) on two standard RS classification benchmarks: (1) **UCM**: RGB land use classification (Yang and Newsam 2010). (2) **BigEarthNet**: Multi-label Sentinel-2 scene classification (Sumbul et al. 2021). We freeze the backbone weights and only train a linear classifier head. This LP scheme effectively measures the quality of learned representations without fine-tuning. As shown in Table 3, our model achieves SOTA performance on both datasets, demonstrating superior representation learning capability across different sensor modalities. More data information, experimental results, implementation details, and visualizations are provided in the supplementary material.

Ablation Studies

Robustness to arbitrary band combinations. As mentioned in the introduction, existing RSFMs suffer a significant drop in feature extraction capability when the spectral bands in downstream tasks differ from those used during pretraining. To evaluate this issue, we freeze the backbone of *AOM* and fine-tune only the classification head under various

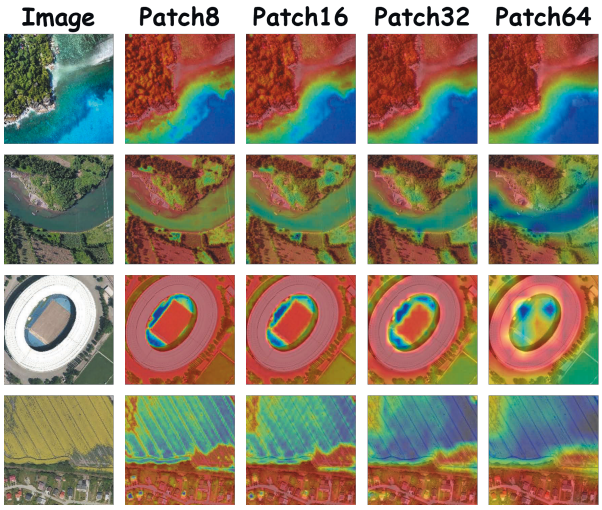

Figure 6: **Visualizing key regions of interest from AOM's final output feature map.** The results show that AOM preserves global semantics across different patch sizes.

band combinations of Sentinel-2. *AOM* demonstrates superior robustness across arbitrary band combinations in EuroSAT (Helber et al. 2019) linear probing, consistently outperforming existing methods under all tested configurations. As shown in Figure 3, it achieves 2.28-19.09% higher accuracy than competitors, maintaining 95.50% accuracy even with only 3 available bands, while others degrade significantly. The model shows particular strength in challenging scenarios, proving its unique capability to handle flexible spectral inputs without performance compromise. Figure 4 shows the visualized feature maps of Sentinel-2 images processed by **SiTok**, illustrating the model's ability to extract distinct features from each channel.

Ablation on patch sizes. As shown in Figure 5, across the partial fine-tuning patch-size ablation on UCM and SPARCS, *AOM*'s accuracy and mIoU remain remarkably stable as the patch granularity shifts from coarse to fine, underscoring the model's ability to preserve performance and hence robustness over a wide range of spatial resolutions for classification and segmentation tasks. Figure 6 visualizes the patch embeddings produced by *AOM* at multiple scales.

Ablation on local vs. global semantic objectives. To disentangle the effect of the loss components, we compare (i) a local per-spectral reconstruction objective that minimises the patch-wise MSE, and (ii) the same MSE term augmented with an InfoNCE loss that enforces multi-scale, global semantic consistency across patches. Table 4 shows that adding the InfoNCE term consistently boosts accuracy on all three datasets. This confirms that local spectral fidelity alone is insufficient: incorporating global, cross-scale cues produces stronger and more transferable representations.

Spectral encoding ablation. To incorporate prior information into the channel token sequence after patch embedding, several strategies can be considered. One approach involves embedding the central wavelength of each spectral band, while another encodes the channel index in sequential order (i.e., 0, 1, 2, 3, ...). Though prior works have typically adopted wavelength embeddings, our experiments reveal minimal performance differences between the two methods, as described in Figure 5. Given its simplicity and ease of use, particularly in downstream applications where exact wavelength metadata may be unavailable, we adopt channel index encoding as a more practical alternative.

Loss	Effects	Acc. (%)↑		mIoU (%)↑
		EuroSAT	UCM	SPARCS
MSE	Local	96.45	91.28	66.7
MSE&InfoNCE	Local&Global	**97.87**	**93.57**	**68.5**

Table 4: Effect of loss design on model performance.

Encoding settings	Acc. (%)↑		mIoU (%)↑
	EuroSAT	UCM	SPARCS
Wavelength	97.72	92.81	67.7
Channel Index	**97.87**	**93.57**	**68.5**

Table 5: Effect of spectral encoding on model performance.

Conclusion

In this paper, we propose Any-Optical-Model (AOM), a universal foundation model for remote sensing designed to handle arbitrary spectral bands, resolutions, and image sizes. The key innovations of AOM are: (1) a spectrum-independent tokenizer that ensures adaptability to varying band configurations while preserving spectral information; (2) a multi-scale adaptive patch embedding mechanism that dynamically adjusts to diverse spatial scales; and (3) a self-supervised pretraining strategy with channel-wise masking and multi-scale semantic alignment to enhance robustness and consistency. We conducted extensive experiments on over 10 datasets demonstrating that AOM outperforms existing RSFMs in downstream tasks, particularly under challenging conditions such as missing bands, cross-sensor data, and varying resolutions. These results establish AOM as a significant advancement toward truly universal RSFMs.

Limitations and future work. Although existing studies on AOM have demonstrated its effectiveness across multiple sensors and tasks, there remain two key challenges regarding its generalization and universality. First, its robustness to out-of-distribution data such as hyperspectral imagery and SAR still requires rigorous evaluation. Second, its performance on a broader range of tasks, including object detection and temporal prediction, remains underexplored. To this end, future work will extend AOM to more diverse sensors and tasks, investigate more effective channel-embedding strategies, and further examine its capability to extract features across a wider range of spatial resolutions.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2022YFB3903401, the National Natural Science Foundation of China under Grant 42271350, and by the International Partnership Program of the Chinese Academy of Sciences under Grant No.313GJHZ2023066FN.

References

Assran, M.; Duval, Q.; Misra, I.; Bojanowski, P.; Vincent, P.; Rabbat, M.; LeCun, Y.; and Ballas, N. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15619–15629.

Astruc, G.; Gonthier, N.; Mallet, C.; and Landrieu, L. 2025. AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19530–19540.

Bao, Y.; Sivanandan, S.; and Karaletsos, T. 2024. Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words. In *The Twelfth International Conference on Learning Representations*.

Beyer, L.; Izmailov, P.; Kolesnikov, A.; Caron, M.; Kornblith, S.; Zhai, X.; Minderer, M.; Tschannen, M.; Alabdulmohsin, I.; and Pavetic, F. 2023. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14496–14506.

Bourriez, N.; Bendidi, I.; Cohen, E.; Watkinson, G.; Sanchez, M.; Bollot, G.; and Genovesio, A. 2024. Chadavit: Channel adaptive attention for joint representation learning of heterogeneous microscopy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11556–11565.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6172–6180.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

de Almeida Pereira, G. H.; Fusioka, A. M.; Nassu, B. T.; and Minetto, R. 2021. Active fire detection in Landsat-8 imagery: A large-scale dataset and a deep-learning study. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178: 171–186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. 2024. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27672–27683.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226.

Hong, D.; Li, C.; Yokoya, N.; Zhang, B.; Jia, X.; Plaza, A.; Gamba, P.; Benediktsson, J. A.; and Chanussot, J. 2025. Hyperspectral imaging. *arXiv preprint arXiv:2508.08107*.

Hong, D.; Zhang, B.; Li, X.; Li, Y.; Li, C.; Yao, J.; Yokoya, N.; Li, H.; Ghamisi, P.; Jia, X.; et al. 2024. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5227–5244.

Hughes, M. J.; and Kennedy, R. 2019. High-quality cloud masking of Landsat 8 imagery using convolutional neural networks. *Remote Sensing*, 11(21): 2591.

Kuckreja, K.; Danish, M. S.; Naseer, M.; Das, A.; Khan, S.; and Khan, F. S. 2024. GeoChat: Grounded Large Vision-Language Model for Remote Sensing. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lacoste, A.; Lehmann, N.; Rodriguez, P.; Sherwin, E.; Kerner, H.; Lütjens, B.; Irvin, J.; Dao, D.; Alemohammad, H.; Drouin, A.; et al. 2023. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093.

Li, X.; Hong, D.; and Chanussot, J. 2024. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24088–24097.

Li, X.; Li, C.; Ghamisi, P.; and Hong, D. 2025a. Fleximo: A flexible remote sensing foundation model. *arXiv preprint arXiv:2503.23844*.

Li, X.; Li, C.; Vivone, G.; and Hong, D. 2025b. SeaMo: A season-aware multimodal foundation model for remote sensing. *Information Fusion*, 103334.

Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; and Chen, C. 2023. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16806–16816.

Noman, M.; Naseer, M.; Cholakkal, H.; Anwer, R. M.; Khan, S.; and Khan, F. S. 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27811–27819.

Pham, C.; Caicedo, J. C.; and Plummer, B. A. 2025. ChA-MAEViT: Unifying Channel-Aware Masked Autoencoders and Multi-Channel Vision Transformers for Improved Cross-Channel Learning. *arXiv preprint arXiv:2503.19331*.

Pham, C.; and Plummer, B. 2024. Enhancing feature diversity boosts channel-adaptive vision transformers. *Advances in Neural Information Processing Systems*, 37: 89782–89805.

Phillips, C.; Roy, S.; Ankur, K.; and Ramachandran, R. 2023. HLS Foundation Burnscars Dataset.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Shabbir, A.; Zumri, M.; Bennamoun, M.; Khan, F. S.; and Khan, S. 2025. GeoPixel: Pixel Grounding Large Multimodal Model in Remote Sensing. *arXiv preprint arXiv:2501.13925*.

Sumbul, G.; De Wall, A.; Kreuziger, T.; Marcelino, F.; Costa, H.; Benevides, P.; Caetano, M.; Demir, B.; and Markl, V. 2021. BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 9(3): 174–180.

Wang, Y.; Braham, N. A. A.; Xiong, Z.; Liu, C.; Albrecht, C. M.; and Zhu, X. X. 2023. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3): 98–106.

Xia, J.; Yokoya, N.; Adriano, B.; and Broni-Bediako, C. 2023. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6254–6264.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434.

Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A. J.; Hanna, J.; Borth, D.; Papoutsis, I.; Saux, B. L.; Camps-Valls, G.; and Zhu, X. X. 2024. Neural plasticity-inspired multimodal foundation model for Earth observation. *arXiv preprint arXiv:2403.15356*.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.