SCALSPECTRA: A SCALABLE ANY-SPECTRAL FOUNDATION MODEL FOR EARTH OBSERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose ScalSpectra, a self-supervised framework for multimodal Earth observation (EO) foundation model pretraining. Although several pioneering efforts have emerged in this field, existing approaches still suffer from two major limitations. First, different EO sensors provide different sets of spectral bands, and the contribution of each band to the fused representation depends on which other bands are observed alongside it. Current tokenizers do not explicitly adapt band fusion to the observed band set, leaving this cross-band dependency unmodeled in the fused representation. Second, existing multimodal EO foundation models predominantly rely on masked image modeling and contrastive self-distillation objectives that require large-scale, high-quality data to learn dense semantic features, whereas spectral observations are relatively scarce, costly to acquire, and complex to process. Our framework is built on spectral-aware dynamic band tokenization (SADBT) and dual-teacher dense feature learning. SADBT encodes band content together with optical wavelengths or SAR polarization and orbit descriptors, and uses a configuration-aware query to fuse the available bands into physically informed patch embeddings, so a single model scales to any band combination. The dual-teacher framework combines multimodal local structural representation learning (MLSRL), which aligns masked student features with an unmasked exponential moving average teacher over all patch positions, and cross-modal semantic representation distillation (CMSRD), which transfers dense semantic knowledge from a frozen visual-prior teacher through aligned optical observations while retaining modality-specific evidence. We pretrain ScalSpectra on heterogeneous EO data sources spanning RGB, multispectral, hyperspectral, and SAR observations. Extensive evaluations across scene classification, semantic segmentation, change detection, and domain- generalized dense prediction demonstrate that ScalSpectra achieves the best overall performance across the evaluated benchmarks.

1 INTRODUCTION

Earth observation (EO) satellites measure the Earth's surface through sensing modalities such as optical, multispectral, hyperspectral, and synthetic aperture radar (SAR), supporting environmental monitoring, disaster management, urban planning, and related applications (Moreira et al., 2013; Rolf et al., 2024; Zhu et al., 2026). Building a multimodal remote-sensing foundation model that can process these heterogeneous modalities and learn robust, transferable representations for diverse downstream tasks is an important research goal (Wang et al., 2022; Zhu et al., 2026).

Despite recent pioneering efforts, existing multimodal remote-sensing foundation models still face two fundamental challenges: building physically aware spectral-band embeddings that scale to arbitrary band combinations, and learning dense representations with strong semantic discriminability. First, current input interfaces do not adequately combine band flexibility with physical awareness. Optical measurements depend on wavelength, whereas synthetic aperture radar (SAR) backscatter depends on polarization and imaging geometry (Moreira et al., 2013). Sensor-specific interfaces rely on predefined band configurations or modality branches (Fuller et al., 2023; Astruc et al., 2025; Tseng et al., 2025; Danish et al., 2026), while flexible-band methods support variable channel counts using physical metadata and band aggregation (Xiong et al., 2024; Prexl & Schmitt, 2024; Waldmann et al., 2025). However, these designs do not fully account for how each band's contribution to a patch embedding depends on the other bands available in the current observation. Second, current mul-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

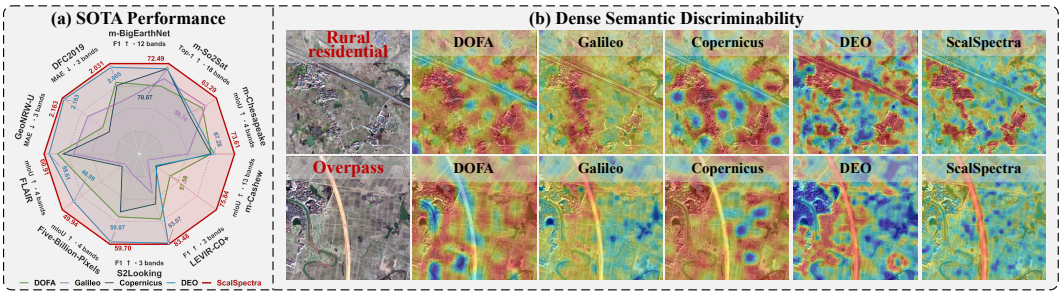

Figure 1: **ScalSpectra. (a)** Best performance across tasks with varying band counts. **(b)** Region-of-interest heatmaps show stronger semantic discrimination.

timodal remote-sensing foundation models are primarily pretrained with masked image modeling and contrastive self-distillation objectives (Cong et al., 2022; Nedungadi et al., 2024; Fuller et al., 2023; Guo et al., 2024; Wang et al., 2024; Waldmann et al., 2025). Visual foundation models obtain dense features with strong semantic discriminability by pretraining on massive, high-quality image corpora (Siméoni et al., 2025; Wolf et al., 2026). High-quality spectral observations are relatively scarce, costly to acquire, and complex to process. This data limitation constrains remote-sensing pretraining and makes it difficult to learn dense features with strong semantic discriminability.

To address these challenges, we propose ScalSpectra, a self-supervised framework for multimodal remote-sensing foundation model pretraining (Figure 2). Its design follows two principles: spectral-band embedding should jointly account for band content, physical attributes, and the current band combination, while dense feature learning should combine local structural supervision with semantic guidance. Accordingly, ScalSpectra introduces two complementary designs. First, spectral-aware dynamic band tokenization (SADBT) encodes optical wavelengths or SAR polarization and orbit descriptors together with band content, and summarizes the complete descriptor set into a spectral-adaptive query that guides fusion into patch embeddings. This configuration-aware construction supports physically informed embeddings for arbitrary band combinations. Second, dual-teacher dense feature learning combines multimodal local structural representation learning (MLSRL) with cross-modal semantic representation distillation (CMSRD). MLSRL aligns a masked student with an unmasked exponential moving average (EMA) multimodal teacher at every patch position, preserving modality-specific characteristics in the multimodal representation while maintaining local structure and anchoring visible features. CMSRD transfers dense semantic knowledge from a frozen visual-prior teacher through aligned optical views (Siméoni et al., 2025), providing semantic supervision that does not depend solely on scarce spectral observations. Together, the two teachers provide local structural supervision and dense semantic guidance for the same multimodal student. Figure 1 illustrates ScalSpectra's leading performance across tasks with varying input band counts (a) and stronger semantic discrimination in region-of-interest heatmaps (b).

We pretrain ScalSpectra on public remote-sensing datasets spanning RGB, multispectral, hyperspectral, and SAR observations. Extensive evaluations across these four modalities cover scene classification, semantic segmentation, change detection, and domain generalization for segmentation and height estimation. Together, these benchmarks probe scene-level separability, dense spatial transfer, temporal comparison, and robustness to sensor and regional shifts. ScalSpectra achieves the highest average GeoBench classification and segmentation scores, improves average segmentation mIoU by 6.49 points over the strongest baseline, and leads on all evaluated change-detection and domain-generalization benchmarks. Feature visualizations further show distinct responses to different land-cover regions before downstream fine-tuning, supporting the semantic discriminability of the pretrained dense features (Figure 4).

Our contributions are summarized as follows:

- We introduce spectral-aware dynamic band tokenization for physically informed patch embeddings across arbitrary band combinations.
- We design a dual-teacher framework that combines local structural alignment and cross-modal semantic distillation to learn spatially coherent and semantically discriminative dense features.

- We evaluate ScalSpectra on classification, segmentation, change detection, and domain generalization for segmentation and height estimation, demonstrating strong transfer performance and semantically informative pretrained features.

2 Related Work

Multimodal EO foundation models. Multimodal EO foundation models mainly use masked image modeling (MIM) or feature-space contrastive and self-distillation objectives. MIM-based methods learn representations by reconstructing masked regions of remote-sensing imagery, whereas feature-space methods align augmented, temporal, or cross-sensor views (Cong et al., 2022; Nedungadi et al., 2024; Fuller et al., 2023; Guo et al., 2024; Wang et al., 2024; Waldmann et al., 2025). To accommodate sensor diversity, generalist models use modality-specific encoders, band-group processing, shared backbones, or flexible-band tokenizers based on wavelength conditioning, physical metadata, single-band processing, and attention aggregation (Astruc et al., 2025; Tseng et al., 2025; Zhang et al., 2025; Xiong et al., 2024; Prexl & Schmitt, 2024; Sumbul et al., 2025; Li et al., 2026; Bao et al., 2024; Nguyen et al., 2023). These interfaces improve channel flexibility, but sensor-specific designs remain tied to predefined band configurations, which limits arbitrary band combinations. Flexible-band tokenizers accept variable channel counts but do not explicitly adapt fusion to the observed band set, even though each band's contribution depends on the other bands observed alongside it. ScalSpectra models it with physical descriptors and a dynamic fusion query.

Dense representation learning and visual-prior transfer. Modern self-supervised methods combine view-level alignment with masked or patch-level supervision: DINO and DINOv2 establish teacher–student self-distillation, iBOT adds masked-token prediction, and Galileo separates global and local EO objectives (Caron et al., 2021; Zhou et al., 2021; Oquab et al., 2023; Tseng et al., 2025). TIPSv2 further highlights the value of supervising both visible and masked tokens (Cao et al., 2026). In parallel, visual foundation models develop strong patch-level semantics through pretraining on massive, curated image corpora (Siméoni et al., 2025). High-quality spectral observations are relatively scarce and more costly to acquire, making it difficult for EO pretraining to produce dense features with comparable semantic discriminability. EO distillation and dual-teacher approaches have begun to transfer visual priors (Wang et al., 2025b; Wolf et al., 2026; Lv et al., 2026), but direct semantic transfer can risk suppressing modality-specific spectral or polarimetric evidence. A training design that supplies dense semantic guidance under limited spectral data while preserving local multimodal structure remains underexplored. ScalSpectra addresses this gap with MLSRL, which aligns all patch positions to an EMA multimodal teacher, and CMSRD, which transfers dense semantics from a frozen visual-prior teacher through aligned optical views.

3 Method

3.1 Overview

ScalSpectra is a self-supervised pretraining framework for multimodal EO foundation models. It maps heterogeneous observations to a shared feature space while preserving band identity and learning semantically discriminative dense features. As shown in Figure 2, ScalSpectra combines spectral-aware dynamic band tokenization to produce physical patch embeddings for arbitrary band combinations with dual-teacher dense feature learning for semantically discriminative features.

The remainder of this section follows this pipeline. Section 3.2 describes how heterogeneous views are constructed from co-located observations, Section 3.3 presents SADBT, which converts each view's band subset into spectrum-aware patch embeddings, and Section 3.4 first introduces the scene-level objective and then details the two dense objectives and the joint loss that updates a single student. Section 4 then evaluates the resulting encoder on downstream classification, segmentation, change detection, and height estimation.

3.2 Heterogeneous EO View Construction

Heterogeneous EO view construction turns acquisition variability into self-supervision for multimodal remote-sensing foundation model pretraining. Following global–local multi-crop learning

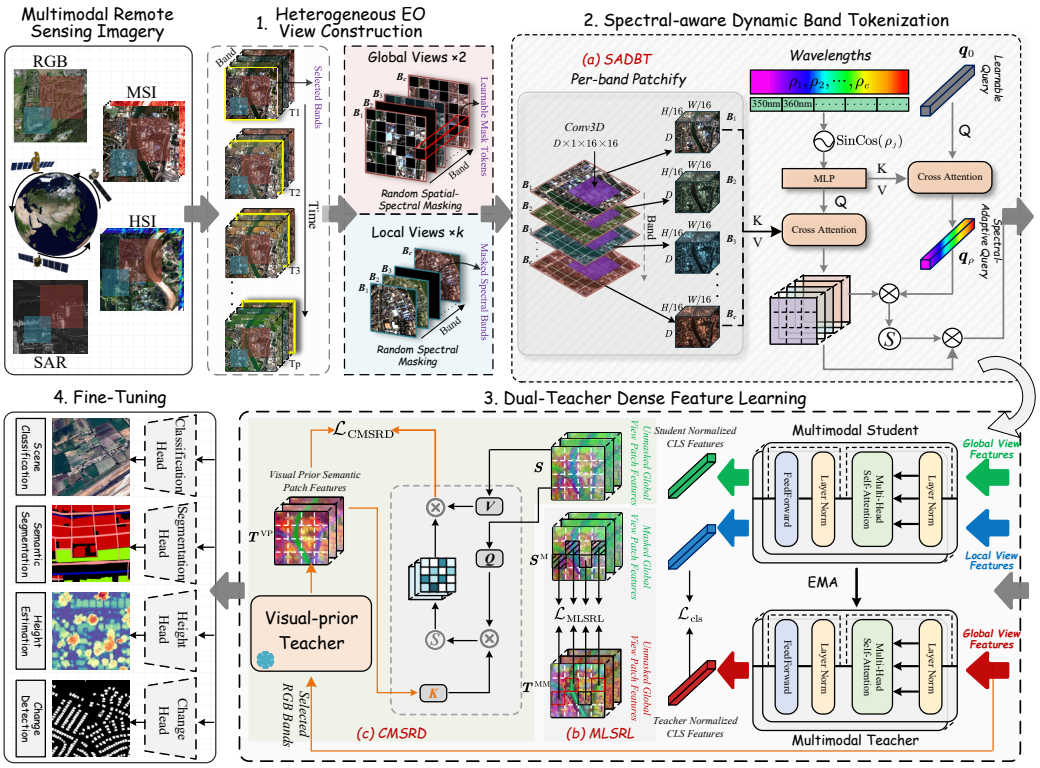

Figure 2: **Overview of ScalSpectra. 1.** Construct heterogeneous EO views across sensors, times, spatial extents, and bands. **2.** SADBT fuses variable bands into spectrum-aware patch tokens. **3.** Alongside scene-level learning, MLSRL learns local structure from the EMA multimodal teacher, while CMSRD transfers dense semantics from the frozen VP teacher. **4.** Adapt the pretrained encoder to classification, segmentation, height estimation, and change detection.

and co-located EO positives (Caron et al., 2021; Waldmann et al., 2025), we construct views for each geographic sample by varying the sensor, acquisition time, spatial support, and retained bands. Each view v is formed by selecting a sensor r_v, sampling c_v bands from an admissible range $[c_v^{\min}, c_v^{\max}]$ that varies by view type, and applying shared geometric augmentation with sensor-appropriate radiometric perturbations. For repeated observations, one timestamp is sampled per physical band and mapped to a common spatial grid before band selection. This controlled variability exposes the encoder to changing sensor configurations while preserving physical band identity through descriptors $\boldsymbol{\rho}_{v,j}$, enabling self-distillation across heterogeneous positive pairs.

3.3 Spectral-Aware Dynamic Band Tokenization

EO views contain variable band sets with modality-specific physical descriptors: optical bands use center wavelength, whereas SAR bands use transmit/receive polarization and orbit direction. As shown in Figure 2, spectral-aware dynamic band tokenization (SADBT) conditions band features on their physical descriptors and summarizes the observed descriptor set into a spectral-adaptive query that adapts band fusion to the current band combination. It maps any band subset to a fixed d-dimensional patch token while preserving band identity and cross-band interactions.

A view v contains c_v band images $\boldsymbol{B}_{v,j} := (\mathbf{X}_v)_{j,:,:}$, $j = 1, \ldots, c_v$. A shared 3D convolution with kernel and stride $1 \times 16 \times 16$ projects non-overlapping patches independently within each band, yielding $\boldsymbol{p}_{v,i,j} \in \mathbb{R}^d$ for patch i of band j. For an optical view, let $\boldsymbol{\rho}_{v,\text{wav}} = [\rho_{v,1}, \ldots, \rho_{v,c_v}]^\top$ collect the center wavelengths, where $\rho_{v,j}$ is the center wavelength of band j. We use $\boldsymbol{\rho}_{v,j}$ for the complete physical descriptor of band j: it reduces to the scalar $\rho_{v,j}$ for an optical band and comprises the acquisition attributes $\{\rho_{v,j,q}\}_{q\in\{\text{tx,rx,orb}\}}$ for a SAR band. For SAR, write $\boldsymbol{\eta}_{v,j,q} = \text{Emb}_q(\rho_{v,j,q})$ for $q \in \{\text{tx}, \text{rx}, \text{orb}\}$, where tx, rx, and orb denote transmit polarization, receive

polarization, and orbit direction, respectively. Each band descriptor is encoded as

$$\boldsymbol{g}_{v,j} = E_\rho(\boldsymbol{\rho}_{v,j}) = \begin{cases} \mathrm{MLP}_{\mathrm{wav}}(\mathrm{SinCos}(\rho_{v,j}))\,, & \text{optical,} \\ \boldsymbol{\eta}_{v,j,\mathrm{tx}} \| \boldsymbol{\eta}_{v,j,\mathrm{rx}} \| \boldsymbol{\eta}_{v,j,\mathrm{orb}}, & \text{SAR.} \end{cases} \quad \boldsymbol{g}_{v,j} \in \mathbb{R}^d. \tag{1}$$

Thus, both branches are parameterized to output a d-dimensional descriptor: the optical branch applies a learnable MLP projection to the sinusoidal encoding of each center wavelength, whereas the SAR branch concatenates learnable embeddings of the three acquisition attributes, with their concatenated dimension equal to d. Stack the band features and the resulting descriptor features as $\boldsymbol{P}_{v,i} = [\boldsymbol{p}_{v,i,1}^\top; \ldots; \boldsymbol{p}_{v,i,c_v}^\top]$ and $\boldsymbol{R}_v = [\boldsymbol{g}_{v,1}^\top; \ldots; \boldsymbol{g}_{v,c_v}^\top]$, both in $\mathbb{R}^{c_v \times d}$. Both matrices preserve band order, maintaining row-wise correspondence between image content and descriptors.

SADBT uses three attention blocks. First, the descriptors query the band features to produce $\widetilde{\boldsymbol{P}}_{v,i} = \mathrm{CrossAttn}_{\mathrm{phy}}(\boldsymbol{R}_v, \boldsymbol{P}_{v,i}, \boldsymbol{P}_{v,i}) \in \mathbb{R}^{c_v \times d}$, retaining one descriptor-conditioned feature for each available band. This makes band content aware of its wavelength or polarization–orbit semantics. Second, a learnable query $\boldsymbol{q}_0 \in \mathbb{R}^d$ summarizes the available descriptors into the d-dimensional spectral-adaptive query $\boldsymbol{q}_{\rho,v}^\top = \mathrm{CrossAttn}_{\mathrm{set}}(\boldsymbol{q}_0^\top, \boldsymbol{R}_v, \boldsymbol{R}_v) \in \mathbb{R}^{1 \times d}$. Finally, the spectral-adaptive query, which encodes the physical composition of the current view, attends to the descriptor-conditioned band features to produce the patch token:

$$\boldsymbol{e}_{v,i}^\top = \mathrm{softmax}_{\mathrm{band}}\left(\frac{\boldsymbol{q}_{\rho,v}^\top \widetilde{\boldsymbol{P}}_{v,i}^\top}{\sqrt{d}}\right) \widetilde{\boldsymbol{P}}_{v,i} \in \mathbb{R}^{1 \times d}, \tag{2}$$

Here, $\boldsymbol{q}_{\rho,v}^\top \widetilde{\boldsymbol{P}}_{v,i}^\top \in \mathbb{R}^{1 \times c_v}$ produces band-wise compatibility scores, which are normalized by softmax to aggregate $\widetilde{\boldsymbol{P}}_{v,i} \in \mathbb{R}^{c_v \times d}$. Since $\boldsymbol{q}_{\rho,v}$ summarizes the physical composition of the current view, the weights depend on both the available descriptors and the local band content, while the aggregation always returns one d-dimensional patch embedding.

3.4 Dual-Teacher Dense Feature Learning

After SADBT converts each heterogeneous view into spectrum-aware patch tokens, we first establish scene-level alignment as the basis for dense feature learning. Views sharing a geographic identity but differing in sensor, time, spatial support, or bands are aligned by contrastive self-distillation with an EMA teacher (Caron et al., 2021; Wu et al., 2025; Grill et al., 2020; He et al., 2020). The scene-level loss combines cosine compression and coding-rate expansion (Wu et al., 2025): $\mathcal{L}_{\mathrm{cls}} = \mathcal{L}_{\mathrm{comp}} - \mathcal{R}_\delta(\boldsymbol{\Gamma}_{\mathrm{het}})$, where $\mathcal{L}_{\mathrm{comp}} = -\alpha_{\mathrm{comp}} \mathbb{E}[\langle \boldsymbol{z}_{\mathrm{T}}^g, \boldsymbol{z}_{\mathrm{S}}^v \rangle]$ aligns normalized CLS features of positive pairs (g, v), while $\mathcal{R}_\delta(\boldsymbol{\Gamma}_{\mathrm{het}}) = \frac{1}{2} \log\det(\boldsymbol{I}_{d_z} + \frac{d_z}{\delta} \boldsymbol{\Gamma}_{\mathrm{het}})$ maximizes the log-determinant of the second-moment matrix $\boldsymbol{\Gamma}_{\mathrm{het}}$. However, scene-level learning mainly optimizes global CLS representations and provides limited guidance for learning the locally structured and semantically discriminative dense features required by multimodal EO foundation models, especially when high-quality spectral observations are scarce (Caron et al., 2021; Zhou et al., 2021). We therefore train the same multimodal student with two complementary dense objectives: MLSRL aligns every patch position with the EMA multimodal teacher to preserve local structure, while CMSRD distills dense semantics from a frozen visual-prior (VP) teacher through geometrically aligned optical views. The two objectives use separate teacher spaces but jointly update the same student.

To ground both dense objectives in the same geographic content, we construct their view sets from the same underlying geographic samples. Let $\mathbb{G}_{\mathrm{M}}$ be the masked global views used by MLSRL and $\mathbb{G}_{\mathrm{O}}$ the unmasked global views with aligned optical views for CMSRD. The student receives masked and unmasked patch grids from $\mathbb{G}_{\mathrm{M}}$ and $\mathbb{G}_{\mathrm{O}}$, respectively; all supervision compares geometrically corresponding positions, with N patches per global view.

Multimodal local structural representation learning (MLSRL). MLSRL is designed to preserve modality-specific characteristics and local semantic structure in the multimodal student while requiring it to use global context to recover masked content. A masked-only patch objective (Zhou et al., 2021) directly constrains masked positions but supervises visible positions only indirectly through their role in predicting masked content. Visible representations can therefore drift without remaining locally semantically anchored. We address this asymmetry by applying the EMA-teacher alignment to the complete patch grid. Masked positions encourage context-based recovery of missing local semantics, whereas visible positions are directly anchored to the corresponding features of

the unmasked EMA multimodal teacher. This dual supervision stabilizes the full dense representation and is particularly beneficial for multimodal remote-sensing pretraining with heterogeneous and incomplete observations.

For each $a \in \mathbb{G}_{\mathrm{M}}$, let $\boldsymbol{s}_{a,i}^{\mathrm{M}}$ denote the student patch feature from a masked multimodal global view and let $\boldsymbol{t}_{a,i}^{\mathrm{MM}}$ denote the corresponding feature from the EMA multimodal teacher, both in $\mathbb{R}^d$. The projection heads $\phi_{\mathrm{S}}, \phi_{\mathrm{T}} : \mathbb{R}^d \to \mathbb{R}^{d_z}$ map the backbone features into a shared projection space before normalization:

$$\widehat{\boldsymbol{z}}_{a,i}^{\mathrm{S}} = \mathrm{norm}_2\big(\phi_{\mathrm{S}}(\boldsymbol{s}_{a,i}^{\mathrm{M}})\big), \qquad \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{T}} = \mathrm{norm}_2\big(\phi_{\mathrm{T}}(\boldsymbol{t}_{a,i}^{\mathrm{MM}})\big). \tag{3}$$

MLSRL matches these features at corresponding spatial positions and averages their normalized negative cosine similarity:

$$\mathcal{L}_{\mathrm{MLSRL}} = -\frac{1}{|\mathbb{G}_{\mathrm{M}}|} \sum_{a \in \mathbb{G}_{\mathrm{M}}} \frac{1}{N} \sum_{i=1}^{N} \big\langle \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{S}}, \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{T}} \big\rangle, \tag{4}$$

Because the loss covers every patch position, visible and masked tokens jointly shape the student representation: visible tokens retain local semantic evidence, while masked tokens learn context-dependent recovery. The resulting dense features are continuous and spatially stable across the multimodal remote-sensing data distribution.

Cross-modal semantic representation distillation (CMSRD). CMSRD enables the multimodal student to learn dense semantic features from the frozen VP teacher. It takes unmasked multimodal views as student inputs and geometrically aligned optical observations as inputs to the VP teacher, transferring the teacher's patch-level dense semantic knowledge into the multimodal student while retaining modality-specific spectral and polarimetric evidence. This objective complements MLSRL, which learns local structural continuity through multimodal self-distillation.

For each $a \in \mathbb{G}_{\mathrm{O}}$, let $\mathbf{X}_a^{\mathrm{MM}}$ be an unmasked multimodal global view and let $\mathbf{X}_a^{\mathrm{OPT}}$ be its geometrically aligned three-channel optical view. The multimodal student and the frozen VP teacher produce the corresponding dense patch grids

$$\boldsymbol{S}_a = f_{\boldsymbol{\theta}}^{\mathrm{patch}}(\mathbf{X}_a^{\mathrm{MM}}) \in \mathbb{R}^{N \times d}, \qquad \boldsymbol{T}_a^{\mathrm{VP}} = \mathrm{sg}\Big((f_{\boldsymbol{\phi}}^{\mathrm{VP}})^{\mathrm{patch}}(\mathbf{X}_a^{\mathrm{OPT}})\Big) \in \mathbb{R}^{N \times d_{\mathrm{VP}}}, \tag{5}$$

where sg denotes stop-gradient. CMSRD transfers the relations among VP patch features rather than treating each target as an independent regression problem. Specifically, we define $\boldsymbol{Q}_a = \phi_q(\boldsymbol{S}_a)$, $\boldsymbol{K}_a = \boldsymbol{T}_a^{\mathrm{VP}}$, and $\boldsymbol{V}_a = \phi_v(\boldsymbol{S}_a)$, where learnable maps $\phi_q, \phi_v : \mathbb{R}^d \to \mathbb{R}^{d_{\mathrm{VP}}}$ project student tokens into the VP feature space, so $\boldsymbol{Q}_a, \boldsymbol{K}_a, \boldsymbol{V}_a \in \mathbb{R}^{N \times d_{\mathrm{VP}}}$. The complete dot-product cross-attention is

$$\boldsymbol{C}_a = \mathrm{softmax}_{\mathrm{row}}\big(\boldsymbol{Q}_a \boldsymbol{K}_a^{\top}\big)\, \boldsymbol{V}_a \in \mathbb{R}^{N \times d_{\mathrm{VP}}}. \tag{6}$$

The softmax is applied row-wise over all VP teacher patches. Because the query and value come from the multimodal student while the key comes from the VP teacher, this operation explicitly establishes cross-modal patch relations. The resulting attention weights use the teacher's dense semantic geometry to reorganize student values, producing $\boldsymbol{C}_a$ without replacing the student's modality-specific evidence. We then normalize this teacher-guided student representation and its spatially corresponding VP target as $\widehat{\boldsymbol{z}}_{a,i}^{\mathrm{C}} = \mathrm{norm}_2(\boldsymbol{C}_{a,i})$ and $\widehat{\boldsymbol{z}}_{a,i}^{\mathrm{VP}} = \mathrm{norm}_2(\boldsymbol{T}_{a,i}^{\mathrm{VP}})$, where $\widehat{\boldsymbol{z}}_{a,i}^{\mathrm{C}}, \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{VP}} \in \mathbb{R}^{d_{\mathrm{VP}}}$.

$$\mathcal{L}_{\mathrm{CMSRD}} = -\frac{1}{|\mathbb{G}_{\mathrm{O}}|} \sum_{a \in \mathbb{G}_{\mathrm{O}}} \frac{1}{N} \sum_{i=1}^{N} \big\langle \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{C}}, \widehat{\boldsymbol{z}}_{a,i}^{\mathrm{VP}} \big\rangle. \tag{7}$$

Because the VP teacher is frozen, minimizing this objective transfers its dense semantic knowledge in two complementary ways: its keys shape cross-modal patch relations, and its spatial targets anchor the resulting student features in the VP semantic space. Only the multimodal student and its projections are updated.

Joint objective. The complete training objective combines scene-level cross-view learning with the two complementary dense objectives:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{MLSRL}} \mathcal{L}_{\mathrm{MLSRL}} + \lambda_{\mathrm{CMSRD}} \mathcal{L}_{\mathrm{CMSRD}}, \tag{8}$$

where the two coefficients balance local-structure learning and semantic distillation. Together, the three objectives update the same student tokenizer and backbone.

Table 1: GeoBench results: linear probing for classification and partial fine-tuning for segmentation. m-BigEarthNet uses F1; other classification datasets use Top-1 accuracy, and segmentation uses mIoU. Avg. is the arithmetic mean across the four datasets in each task. All metrics are in % and higher is better (↑). Best and second-best are bold and underlined.

Method	*Scene Classification*					*Semantic Segmentation*				
	m-EuroSAT	m-ForestNet	m-BigEarthNet	m-So2Sat	Avg.	m-PV4Ger	m-Chesapeake	m-Cashew	m-NeonTree	Avg.
CROMA	95.58	48.44	67.63	53.85	66.38	86.91	45.76	33.31	52.85	54.71
SenPaMAE	93.07	46.02	57.47	55.38	62.99	88.99	45.47	26.01	48.54	52.25
Prithvi-v2	96.99	52.87	69.64	54.87	68.59	92.65	55.43	<u>60.11</u>	51.05	64.81
DOFA	<u>97.29</u>	51.06	65.22	58.71	68.07	94.62	66.36	57.58	57.13	<u>68.92</u>
Galileo	96.79	42.30	67.67	<u>59.74</u>	66.63	85.51	59.23	48.84	47.31	60.22
Copernicus	96.29	48.64	<u>70.87</u>	54.77	67.64	94.74	66.29	52.30	54.07	66.85
DEO	94.98	**53.47**	70.83	55.58	<u>68.72</u>	<u>94.85</u>	<u>67.28</u>	51.87	**58.34**	68.09
ScalSpectra	**98.19**	<u>53.07</u>	**72.49**	**63.29**	**71.76**	**94.94**	**73.61**	**75.84**	<u>57.24</u>	**75.41**

Table 2: Change detection performance measured by Binary F1 (%, ↑) across four benchmarks covering different sensors and acquisition conditions. Best results are highlighted in **bold**.

Dataset	*Method*							
	CROMA	SenPaMAE	Prithvi-v2	DOFA	Galileo	Copernicus	DEO	ScalSpectra
CDD	86.66	84.28	89.37	91.90	89.07	91.11	92.38	**92.61**
LEVIR-CD+	50.45	47.90	64.04	74.14	64.57	68.59	83.07	**83.48**
DSIFN	54.60	58.54	60.52	65.43	61.43	65.20	70.54	**72.28**
S2Looking	30.46	44.42	42.96	52.46	41.21	51.13	59.07	**59.70**
Average	55.54	58.79	64.22	70.98	64.07	69.01	76.27	**77.02**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Pretraining data. We pretrain on heterogeneous EO data sources (Wang et al., 2025a; Velazquez et al., 2025; Braham et al., 2025; Nedungadi et al., 2024). OpticalRS-4M contributes RGB remote-sensing images. EarthView contributes high-resolution multispectral Satellogic imagery together with Sentinel-2 and Sentinel-1 observations. SpectralEarth contributes EnMAP hyperspectral images. MMEarth provides additional co-located multimodal observations. Dataset-specific radiometric normalization is retained before heterogeneous-view construction. The available band descriptors are passed to SADBT.

Downstream tasks. We evaluate five task groups. Scene classification uses GeoBench m-EuroSAT, m-ForestNet, m-BigEarthNet, and m-So2Sat, with m-BigEarthNet as the multilabel dataset (Lacoste et al., 2023). Semantic segmentation uses m-PV4Ger-Seg, m-Chesapeake, m-Cashew-Plant, and m-NeonTree; change detection uses CDD, LEVIR-CD+, DSIFN, and S2Looking (Shen et al., 2021). Domain-generalized segmentation covers Five-Billion-Pixels, FLAIR, and LoveDA (Garioud et al., 2023), while height estimation covers GAMUS→DFC2019, GeoNRW-U→GeoNRW-R, and SynRS3D→GeoNRW-U (Song et al., 2024).

Implementation details. The student and EMA teacher use a 12-block ViT-B/16 with RoPE, four storage tokens, and $d = 768$; the frozen VP teacher is DINOv3 ViT-L/16 with 1,024-dimensional patch features. We use two 224×224 global crops and six 96×96 local crops, with band-count ranges $[4, 13]$ and $[1, 4]$. AdamW training runs for 20 epochs on eight 80-GB A100 GPUs with batch size 128 per GPU and learning rate 2×10^{-3}; both dense losses have weight 0.5, and the EMA momentum follows a cosine schedule from 0.992 to 1.0.

4.2 TRANSFER TO MULTI-SENSOR DOWNSTREAM TASKS

Baselines and evaluation protocol. We compare ScalSpectra with seven EO foundation models: CROMA (Fuller et al., 2023), SenPaMAE (Prexl & Schmitt, 2024), Prithvi-v2 (Szwarcman et al., 2026), DOFA (Xiong et al., 2024), Copernicus-FM (Wang et al., 2025b), Galileo (Tseng et al., 2025), and DEO (Wolf et al., 2026). All models use ViT-Base backbones except Prithvi-v2 (ViT-Large). We use official weights and fix heads, resolution, and schedules within each benchmark.

Scene classification. We assess scene-level semantic separability by freezing the encoder and training a linear classifier on four GeoBench datasets. ScalSpectra achieves the highest average score (71.76), exceeding DEO by 3.04 points, and leads on three datasets (Table 1, left), with its largest gain on m-So2Sat (+3.55 over Galileo); DEO remains best on m-ForestNet.

Table 3: Domain generalization performance on semantic segmentation and monocular height estimation. The task decoder is trained on the source domain only and evaluated on the target domain without adaptation. Segmentation reports mIoU and mAcc in %; height estimation reports MAE and RMSE in metres. Best results are highlighted in **bold**.

Method	*Domain-Generalizable Semantic Segmentation*						*Domain-Generalizable Height Estimation*					
	Five-Billion-Pixels (Cross-sensor)		FLAIR (Cross-Regional)		LoveDA (Cross-Style)		GAMUS→DFC2019 (Cross-Regional)		GeoNRW-U→GeoNRW-R (Cross-Style)		SynRS3D→GeoNRW-U (Synthetic→Real)	
	mIoU ↑	mAcc ↑	mIoU ↑	mAcc ↑	mIoU ↑	mAcc ↑	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
CROMA	25.95	38.34	49.32	65.04	29.63	41.76	2.839	5.007	1.398	2.840	3.161	4.788
SenPaMAE	21.41	29.17	48.75	64.96	26.52	38.21	2.685	4.691	1.477	2.940	2.692	4.440
Prithvi-v2	37.99	48.87	55.09	69.86	43.55	57.50	2.476	4.344	1.125	2.295	2.336	3.987
DOFA	42.49	52.62	57.48	72.51	50.60	64.97	2.215	4.026	1.134	2.305	2.739	4.742
Galileo	33.66	45.24	53.40	68.62	40.31	53.21	2.437	4.342	1.146	2.330	2.492	4.454
Copernicus	34.42	45.23	56.06	70.98	50.56	65.68	2.249	4.067	1.135	2.323	2.857	4.803
DEO	46.99	60.47	59.61	73.90	55.29	70.76	2.065	3.792	1.029	2.119	2.183	3.852
ScalSpectra	**49.94**	**63.45**	**60.91**	**74.69**	**55.56**	**70.95**	**2.031**	**3.783**	**0.996**	**2.053**	**2.163**	**3.652**

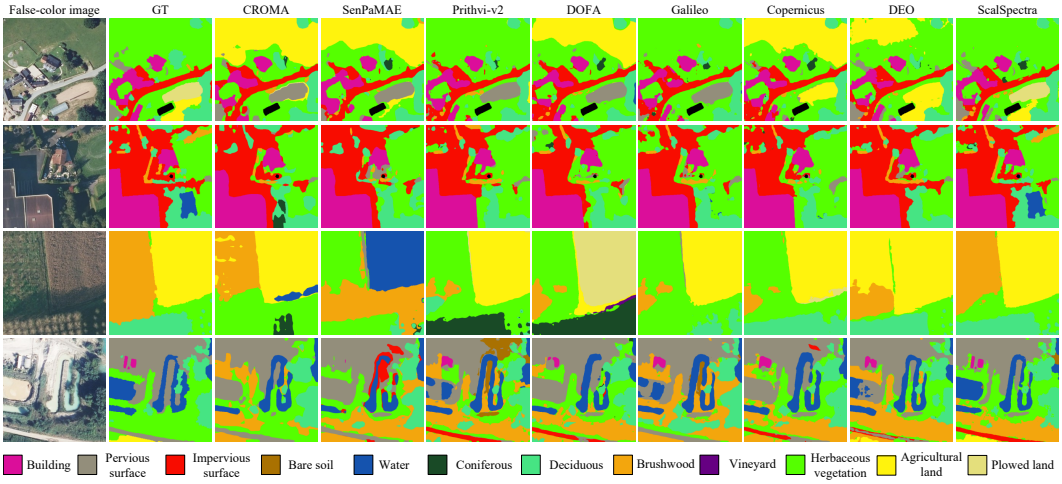

Figure 3: **FLAIR transfer.** Predictions from a source-trained decoder on disjoint target domains show that ScalSpectra better follows the annotated land-cover layout and separates adjacent classes.

Semantic segmentation. To evaluate dense prediction from frozen features, we train only a UPerNet segmentation head under the GeoBench partial-fine-tuning protocol. ScalSpectra has the highest average mIoU (75.41), exceeding DOFA by 6.49 points and leading on three datasets (Table 1, right). The largest gains are on m-Cashew (+15.73 over Prithvi-v2) and m-Chesapeake (+6.33 over DEO), consistent with the complementary MLSRL and CMSRD objectives.

Change detection. With the shared-weight Siamese backbone frozen, we extract pre- and post-event backbone features, fuse them by element-wise subtraction, and decode the difference with a UPerNet head trained only for binary change prediction. ScalSpectra leads all four benchmarks, averaging 77.02 binary F1 versus 76.27 for DEO (Table 2); its largest gains are on DSIFN (+1.74) and S2Looking (+0.63) under multi-source and off-nadir shifts.

4.3 Generalization of the Pretrained Representation under Domain Shift

To evaluate transfer without target-domain supervision, we train task decoders only on the source domain and evaluate directly on the target domain without adaptation or target-domain model selection. Table 3 reports segmentation and height estimation under six domain shifts.

Semantic segmentation. We assess semantic transfer across sensors (Five-Billion-Pixels), regions (FLAIR), and imaging styles (LoveDA) using a Mask2Former decoder. ScalSpectra achieves the highest mIoU and mAcc in all three settings, averaging 55.47/69.70 versus 53.96/68.38 for DEO. On Five-Billion-Pixels, the mIoU gain is 2.95 points over DEO and 7.45 over DOFA, supporting transfer across sensors and regions.

Monocular height estimation. We use a DPT decoder across regional, style, and synthetic-to-real shifts. ScalSpectra has the lowest MAE and RMSE on all three transfers, averaging 1.730/3.163 m versus 1.759/3.254 m for DEO; it reaches 0.996 m MAE on GeoNRW-U→GeoNRW-R and reduces RMSE by 0.200 m on SynRS3D→GeoNRW-U.

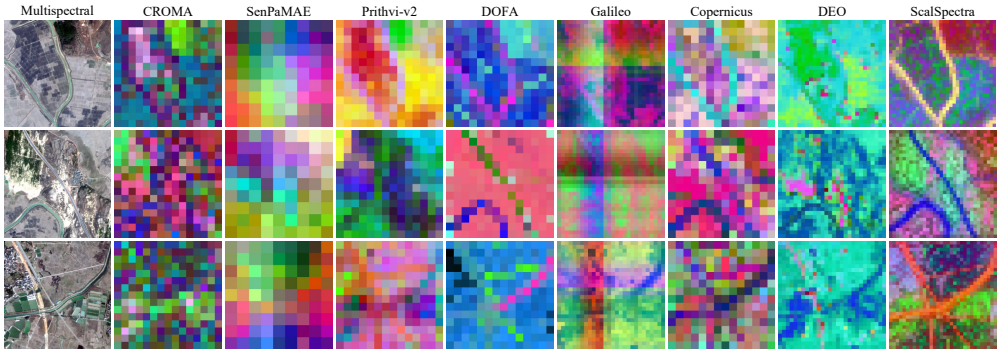

Figure 4: **Dense feature comparison.** PCA maps of frozen pretrained patch tokens show clearer separation of waterways and agricultural parcels with ScalSpectra.

4.4 Ablation Study

We ablate ScalSpectra on the four GeoBench classification and four segmentation datasets in Table 1. The first row is the baseline, combining coding-rate compression and expansion with iBOT and Panopticon's flexible patch embedding. This embedding applies a shared patchifier to each input channel and uses single-query cross-attention to reduce a variable number of channel tokens to one patch token, with spectral positional information based on channel wavelengths or SAR modes (Waldmann et al., 2025). It does not use SADBT's descriptor-conditioned, configuration-aware fusion. The second row replaces this embedding with SADBT, the third and fourth rows add MLSRL or CMSRD separately, and the final row combines both dense objectives.

Table 4: Key component ablation of ScalSpectra.

SADBT	MLSRL	CMSRD	Linear probing Avg.	Linear probing Δ	Segmentation Avg.	Segmentation Δ	Overall Avg.	Overall Δ
×	×	×	70.83	–	70.04	–	70.44	–
✓	×	×	71.73	↑0.90	70.18	↑0.14	70.96	↑0.52
✓	✓	×	70.74	↓0.09	73.17	↑3.13	71.96	↑1.52
✓	×	✓	71.08	↑0.25	74.11	↑4.07	72.60	↑2.16
✓	✓	✓	**71.76**	↑0.93	**75.41**	↑5.37	**73.59**	↑3.15

All deltas are relative to the baseline row. Replacing the baseline embedding with SADBT raises the linear-probing and segmentation averages by 0.90 and 0.14 points, showing the value of configuration-aware spectral fusion. Adding MLSRL then improves segmentation by 3.13 points while linear probing drops by 0.09 points, indicating stronger dense features at a small cost to global representations. CMSRD adds 4.07 and 0.25 points on segmentation and linear probing. Combining both gives the best scores, 71.76 and 75.41, supporting the effectiveness of the complete design.

4.5 Qualitative Analysis

Figure 3 shows cross-regional transfer on FLAIR, where ScalSpectra produces segmentation maps that agree more closely with the reference and mislabel fewer regions than the compared models. Its frozen features in Figure 4 form more clearly separated semantic regions than those of the compared remote-sensing foundation models, indicating that the learned dense features encode semantics more clearly. Additional dense feature comparisons are provided in Figure A1, and Figure A2–A10 add qualitative comparisons of change detection, semantic segmentation, and height estimation.

5 Conclusion

This paper presents ScalSpectra, a self-supervised multimodal EO framework that couples flexible band tokenization with dense semantic representation learning. SADBT encodes band content with physical descriptors and adapts its fusion query to the observed band set, producing physically informed fixed-dimensional patch tokens from variable band combinations. MLSRL aligns visible and masked student patch features with an unmasked EMA multimodal teacher, preserving modality-specific evidence and local structure, whereas CMSRD transfers dense semantic knowledge from a frozen visual-prior teacher through aligned optical observations. These objectives learn spatially coherent and semantically discriminative dense features: ScalSpectra attains the best average GeoBench classification and segmentation scores and the highest average F1 on all four change-detection benchmarks, and source-only domain generalization transfers competitively across sensors, regions, imaging styles, and synthetic-to-real shifts without target-domain adaptation. Evaluations over band configurations from three-channel RGB to multispectral and SAR observations validate the scalable spectral embedding design for arbitrary band combinations.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and restructuring of author-written text, assistance with LaTeX formatting and figure layout, and checks of manuscript clarity and terminology. We did not use generative AI tools to generate the research ideas, design the method, implement ScalSpectra or the experiments, generate or modify experimental data or results, or perform the analysis and interpretation of the experimental results; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to review manuscript organization, citation placement, and presentation choices. We reviewed all AI-assisted work and verified the affected text, formatting, citations, and code against the authors' records. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We specify the full training and evaluation protocol in the main text and the appendix. Section 3 defines the pretraining pipeline, and Section 4.1 reports the pretraining budget. Table A1 lists every pretraining source with its platform, modality and band count, sample size, ground sample distance, and temporal coverage, while Appendix C.1 gives the sampling procedure that turns them into heterogeneous views (Eqs. A1–A3). Appendix C.2 states the complete scene-level objective together with its coding-rate expansion, and Table A2 lists all pretraining hyperparameters, including the student and teacher backbones, crop and band ranges, loss weights, and optimizer settings. For evaluation, Appendix B documents the splits, input sizes, band counts, sensors, metrics, and decoder of every benchmark, Section 4.2 fixes the protocol used for the baseline comparison, and Section 4.3 defines the source-only setting under which the domain-generalization results were obtained. All datasets used in this work are publicly available and cited in the main text.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Anysat: One earth observation model for many resolutions, scales, and modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19530–19540, 2025.

Gerald Baier, Antonin Deschemps, Michael Schmitt, and Naoto Yokoya. Synthesizing optical and SAR imagery from land cover maps and auxiliary raster data. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–12, 2021.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel vision transformers: An image is worth 1 x 16 x 16 words. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CK5Hfb5hBG.

Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:16780–16797, 2025.

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29325–29335, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. TerraFM: A scalable foundation model for unified multisensor earth observation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=cBxuzdUDNx.

Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36: 16456–16482, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. 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 (CVPR)*, pp. 27662–27673. IEEE, 2024.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9726–9735, 2020.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.

Bertrand Le Saux, Naoto Yokoya, Ronny Hansch, Myron Brown, and Greg Hager. 2019 data fusion contest [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 7(1):103–105, 2019.

MA Lebedev, Yu V Vizilter, OV Vygolov, VA Knyaz, and A Yu Rubis. Change detection in remote sensing images using conditional adversarial networks. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(2):565–571, 2018.

Xuyang Li, Chenyu Li, and Danfeng Hong. Any-optical-model: A universal foundation model for optical remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6539–6547, 2026.

Chonghua Lv, Dong Zhao, Shuang Wang, Dou Quan, Ning Huyan, Nicu Sebe, and Zhun Zhong. Generalizable knowledge distillation from vision foundation models for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26656–26666, 2026.

Alberto Moreira, Pau Prats-Iraola, Marwan Younis, Gerhard Krieger, Irena Hajnsek, and Konstantinos P Papathanassiou. A tutorial on synthetic aperture radar. *IEEE Geoscience and remote sensing magazine*, 1(1):6–43, 2013.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182, 2024.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

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

Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *DAGM German Conference on Pattern Recognition*, pp. 317–331, 2024.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical – satellite data is a distinct modality in machine learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 42691–42706, 2024.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2Looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Jian Song, Hongruixuan Chen, Weihao Xuan, Junshi Xia, and Naoto Yokoya. SynRS3D: A synthetic dataset for global 3d semantic understanding from monocular remote sensing imagery. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117388–117425, 2024.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5569–5578, 2025.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multitemporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026.

Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 178–196, 2023.

Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. *arXiv preprint arXiv:2502.09356*, 2025.

Diego Velazquez, Pau Rodriguez López, Sergio Alonso, Josep M. Gonfaus, Jordi Gonzalez, Gerardo Richarte, Javier Marin, Yoshua Bengio, and Alexandre Lacoste. EarthView: A large scale remote sensing dataset for self-supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 1138–1147, 2025.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J. Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2195–2205, 2025.

Fengxiang Wang, Hongzhen Wang, Di Wang, Zonghao Guo, Zhenyu Zhong, Long Lan, Wenjing Yang, and Jing Zhang. Harnessing massive satellite imagery with efficient masked image modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6935–6947, 2025a.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.

Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pp. 286–303, 2024.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9888–9899, 2025b.

Filip Wolf, Blaž Rolih, and Luka Čehovin Zajc. Brewing stronger features: Dual-teacher distillation for multispectral earth observation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27815–27826, 2026.

Ziyang Wu, Jingyuan Zhang, Druv Pai, Xudong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma. Simplifying DINO via coding rate regularization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 68036–68059, 2025.

Zhitong Xiong, Sining Chen, Yi Wang, Lichao Mou, and Xiao Xiang Zhu. GAMUS: A geometry-aware multi-modal semantic segmentation benchmark for remote sensing data. *arXiv preprint arXiv:2305.14914*, 2023.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Chenxiao Zhang, Peng Yue, Deodato Tapete, Liangcun Jiang, Boyi Shangguan, Li Huang, and Guangchao Liu. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:183–200, 2020.

Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. Skysense v2: A unified foundation model for multi-modal remote sensing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9136–9146, 2025.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. *Communications Earth & Environment*, 7(1):103, 2026.

APPENDIX

This supplementary material complements the main paper with additional data descriptions, method details, analyses, and visualizations. It describes the pretraining data (A) and the downstream benchmarks and evaluation protocols (B). The latter include GeoBench scene classification and semantic segmentation (B.1), change detection (B.2), domain-generalized semantic segmentation (B.3), and domain-generalized monocular height estimation (B.4). Additional method details (C) cover heterogeneous view construction (C.1), the scene-level objective (C.2), and the pretraining hyperparameters (C.3). The supplement also provides an extended analysis of pretrained dense features (D) and qualitative results for change detection (E.1), domain-generalized semantic segmentation (E.2), and domain-generalized height estimation (E.3).

A PRETRAINING DATA

ScalSpectra is pretrained on heterogeneous Earth observation (EO) sources that jointly cover passive optical, multispectral, hyperspectral, and active microwave SAR observations. The mixture spans three axes of generalization. Spatial detail ranges from sub-meter commercial imagery to 30 m hyperspectral pixels, while spectral range extends from three-channel RGB to 202 narrow bands. Sensing physics includes reflective optical measurements described by centre wavelengths and SAR backscatter described by transmit/receive-polarization and orbit tuples. Table A1 summarizes the platform, modality and band count, sample size, ground sample distance (GSD), footprint, and temporal coverage of each source. Dataset-specific radiometric normalization is retained before heterogeneous-view construction, and every retained band is passed to SADBT together with its physical descriptor $\boldsymbol{\rho}_{v,j}$.

OpticalRS-4M. OpticalRS-4M provides the large-scale RGB component of the mixture. It is a $\sim$4 M-image visible-light subset of the OpticalRS-13M collection, whose images were gathered from public remote-sensing archives following diversity, richness, and scalability criteria. High-resolution scenes were sliced into sub-images of 64×64 to 1024×1024 pixels and de-duplicated with perceptual hashing followed by manual review (Wang et al., 2025a). Because it aggregates many sensors and acquisition conditions rather than a single platform, it contributes a broad pool of scene and object appearances, including fine-grained categories that are visually similar to each other. This diversity enriches the visual content seen during pretraining and provides varied scene structure for scene-level learning.

EarthView: Satellogic. EarthView combines a curated selection of Sentinel-2 and Sentinel-1 imagery with the first public release of commercial Satellogic imagery (Velazquez et al., 2025). The Satellogic portion contains 2.97 M 384×384 patches at 1 m GSD with four registered bands (red, green, blue, near-infrared). The imagery covers all continents and was acquired in the second half of 2022. The dataset was constructed from 3,758 captures, all with cloud cover below 30%. We use it for spatial detail: at 1 m the model must resolve building outlines, roads, field boundaries, and small isolated objects, which is precisely the regime in which patch-level objectives have to remain locally discriminative.

EarthView: multi-temporal Sentinel-2 and Sentinel-1. The Sentinel portion of EarthView is explicitly multi-temporal: about 1.05 M co-located footprints of 3.84×3.84 km carry *series* of acquisitions rather than single snapshots. Sentinel-2 contributes 13 multispectral bands at 10/20/60 m with up to ten revisits per location, while Sentinel-1 provides two polarization channels (VV, VH) derived from Level-1 Ground Range Detected (GRD) products, resampled to 10 m with up to nine revisits. The collection spans 2017–2022 and deliberately mixes densely sampled dates with seasonal intervals. We exploit this property at the view level. For every physical band c of sensor r, we sample one timestamp from its admissible set $\mathbb{T}_c^{(r)}$ and resample the selected acquisition onto the common spatial grid before band subsampling (Eq. A1). Two views of the same geographic sample can therefore differ in date, season, and illumination, and a single view may even combine asynchronous acquisitions of different bands. We treat this as a *controlled temporal perturbation* rather than as a synchronized observation: scene-level learning has to infer semantics that are stable under acquisition date, while masked local-structure learning has to recover content from context observed under a different temporal state. The purpose is invariance to time-dependent radiometric variation without discarding physically meaningful temporal information, which is

Table A1: Pretraining data sources. Every value is taken from the cited dataset paper, and an em dash marks a statistic that the corresponding paper does not report. Patches are stored image chips and locations are distinct geographic sites, so the two counts are not additive. The two EarthView Sentinel rows cover the same 1.05 M footprints, and OpticalRS-4M is the ∼4 M-image visible-light subset of the OpticalRS-13M collection used here. Footprints follow from the reported patch size and ground sample distance (GSD). SpectralEarth additionally provides 415 K unique locations, and MMEarth provides 12 aligned modalities (46 bands in total) at every location. MSI = multispectral imager, HSI = hyperspectral imager, RGBN = red, green, blue, near-infrared. Sources: [a] Wang et al. (2025a); [b] Velazquez et al. (2025); [c] Braham et al. (2025); [d] Nedungadi et al. (2024).

Pretraining source	Platform, modality (bands)	Samples	GSD (m)	Patch size, footprint	Temporal coverage
OpticalRS-4M[a]	Public satellite RGB (3)	∼4 M images	—	64^2–1024^2 px —	—
EarthView-Satellogic[b]	Satellogic RGBN (4)	2.97 M patches	1	384^2 px 384 m × 384 m	Jul–Dec 2022 1–5 revisits
EarthView-Sentinel-2[b]	Sentinel-2 MSI (13)	1.05 M patches	10/20/60	384^2 px 3.84 × 3.84 km	2017–2022 10 revisits
EarthView-Sentinel-1[b]	Sentinel-1 SAR (2: VV, VH)	1.05 M patches	10	384^2 px 3.84 × 3.84 km	2017–2022 3–9 revisits
SpectralEarth[c]	EnMAP HSI (202)	539 K patches	30	128^2 px 3.84 × 3.84 km	Apr 2022–Apr 2024 17.5% revisited
MMEarth[d]	Sentinel-2 MSI (12) + Sentinel-1 SAR (8)	1.2 M locations	10	128^2 px 1.28 × 1.28 km	2017–2020 —

what change-detection transfer requires, since pre- and post-event images of the same location differ in exactly such seasonal and illumination effects.

SpectralEarth (EnMAP). SpectralEarth contributes the hyperspectral component of the mixture (Braham et al., 2025). It was constructed from 11,636 EnMAP tiles selected with cloud coverage below about 10% between April 2022 and April 2024. The tiles were processed with the EnMAP L2A processor and split into 128×128-pixel patches at 30 m GSD. Bands 127–141 and 161–167 were excluded because they are dominated by water absorption and frequently contain missing values, leaving 202 of the original 224 bands. The resulting collection contains 415,153 locations and 538,974 patches. It also includes 73,307 non-overlapping locations with multiple timestamps (Braham et al., 2025). The retained 202 bands are represented as individual channels with their centre wavelengths. During pretraining, each view samples a subset of these channels according to the global- or local-crop band range, giving SADBT narrow-band inputs and fine-grained wavelength descriptors at varying spectral positions within EnMAP's observed coverage.

MMEarth: co-located Sentinel-2 and Sentinel-1. MMEarth contributes observations that are *exactly co-located* rather than merely drawn from the same region (Nedungadi et al., 2024). It contains 1.2 M locations sampled between 2017 and 2020, and provides 12 geo-aligned modalities. Each pixel-level modality is a 128×128 raster covering 1.28×1.28 km, and all of them were reprojected to the Sentinel-2 10 m grid, with time-critical modalities collected around the Sentinel-2 acquisition date. In particular, the Sentinel-2 modality provides 12 multispectral bands (L1C and L2A), and the Sentinel-1 modality provides eight SAR channels, i.e. VV, VH, HV, and HH in both ascending and descending orbits. For ScalSpectra, these co-located optical and SAR observations form heterogeneous positive pairs for scene-level learning, and the shared SADBT tokenizes both modalities with their physical descriptors so that band identity (wavelength versus polarization and orbit) is preserved across modalities. The scene-level objective enforces semantic compatibility without requiring radiometric agreement, allowing this cross-sensor supervision to complement the multi-temporal single-sensor views.

B Downstream Datasets and Evaluation Protocols

We evaluate five task groups: scene classification, semantic segmentation, change detection, domain-generalized semantic segmentation, and domain-generalized monocular height estimation. Together, these tasks probe scene-level separability, spatial transfer across resolutions and band counts, temporal comparison, and robustness to sensor, regional, and style shifts without target-

domain supervision. All decoders are trained only on the source splits. In the domain-generalization experiments, we evaluate the target domain directly without adaptation or target-domain model selection, as described in Section 4.3.

B.1 Scene Classification and Semantic Segmentation (GeoBench)

The GeoBench classification and segmentation tasks harmonize heterogeneous public datasets under a common schema while retaining diversity in sensors, spatial resolutions, and band counts (Lacoste et al., 2023). We use four tasks for each problem. Their patch sizes, label spaces, data splits, band counts, GSDs, and sensors are summarized below, along with the aspect of the representation that each task probes.

Scene classification. Classification freezes the encoder and trains a linear classifier to measure the linear separability of its pretrained representation. The four tasks cover GSDs from 0.1 to 15 m, 6–18 input channels, and both single- and multi-sensor settings. m-EuroSAT provides 2000/1000/1000 training, validation, and test 64×64 patches with 13 Sentinel-2 bands at 10 m and ten land-cover classes. m-ForestNet provides 6464/989/993 332×332 Landsat-8 patches with six bands at 15 m and twelve classes. These are single-sensor tasks. m-BigEarthNet is a multi-label task. Its 20000/1000/1000 120×120 Sentinel-2 patches use 12 bands at 10 m and carry several co-occurring labels from 43 classes, testing whether multiple semantic dimensions remain jointly decodable. m-So2Sat pairs six Sentinel-1 channels with twelve Sentinel-2 bands over 19992/986/986 32×32 patches at 10 m and 17 classes. It is the only classification task that requires the encoder to consume co-located optical and SAR inputs.

Semantic segmentation. Segmentation follows the GeoBench partial-fine-tuning protocol. The backbone is frozen and only a UPerNet segmentation head is trained. The four tasks span submeter aerial imagery to 10 m multispectral imagery and cover different dense-prediction settings. m-PV4Ger-Seg contains 3000/403/403 320×320 aerial RGB patches at 0.1 m for binary building segmentation. m-Chesapeake uses 3000/1000/1000 256×256 RGBN patches at 1 m to segment seven land-cover classes. m-Cashew-Plant is a 10 m Sentinel-2 task with 13 bands and 1350/400/50 256×256 patches across seven classes, where individual cashew trees occupy only a few pixels. m-NeonTree combines 0.1 m RGB with hyperspectral and elevation layers in 270/94/93 400×400 patches to delineate individual tree crowns with binary labels.

B.2 Change Detection Benchmarks

Change detection is evaluated with a frozen shared-weight Siamese backbone and a UPerNet decoder trained on the task data. The pre- and post-event feature maps are fused by subtraction, and binary F1 is reported. We use the official training, validation, and test splits for all four benchmarks. CDD retains its official 256×256 image size, while the other three datasets are cropped to 512×512 inputs for our experiments.

CDD: seasonal appearance change. The Change Detection Dataset contains bitemporal Google Earth images acquired in different seasons, with spatial resolutions between 3 cm and 1 m and annotations that deliberately *exclude* change caused by season and brightness. From its 11 season-varying source pairs, it provides 10,000/3,000/3,000 training, validation, and test 256×256 patches (Lebedev et al., 2018). Its difficulty is therefore pseudo-change: vegetation phenology, soil moisture, and illumination alter the appearance of unchanged objects, so a model that compares raw bi-temporal appearance produces spatially scattered false positives. This benchmark rewards representations that encode stable semantic structure rather than radiometry, which is what the multi-temporal views of Section A encourage.

LEVIR-CD+: long-term building change. LEVIR-CD+ extends LEVIR-CD with 985 image pairs of 1024×1024 pixels at 0.5 m GSD, acquired between 2002 and 2020 and containing roughly 80,000 building instances. We use the official training, validation, and test splits and crop the images to 512×512 inputs for our experiments (Chen & Shi, 2020). Changes such as construction, expansion, and demolition occupy only a small fraction of the pixels, and the changed objects are often small relative to the image. The long temporal baseline introduces substantial radiometric and structural differences between the two acquisitions. The benchmark is characterized by severe class imbalance and pronounced variation in appearance and structure across dates.

DSIFN: fine-grained change under radiometric distortion. The DSIFN benchmark consists of multi-source bi-temporal Google Earth imagery over several Chinese cities and contains 4,000 pairs of 512×512 images. We use its official training, validation, and test splits and retain the 512×512 input size. The benchmark focuses on fine-grained object-level changes under heterogeneous resolution and radiometric conditions (Zhang et al., 2020). Because the changed objects are subtle and the two dates may originate from different sources, the model has to combine local evidence with semantic context instead of relying on a global appearance difference. ScalSpectra attains its largest change-detection gain here (+1.74 F1 over the strongest baseline, Table 2).

S2Looking: viewing-geometry-induced pseudo-change. S2Looking contains 5000 bitemporal pairs of rural areas distributed globally, captured by side-looking satellites with 0.5–0.8 m GSD and clearly different off-nadir angles between the two acquisitions. We use its official training, validation, and test sets, containing 3500/500/1000 pairs, and crop the images to 512×512 inputs (Shen et al., 2021). Different off-nadir angles can shift the apparent positions of unchanged buildings between the two dates, so direct bi-temporal comparison may mistake geometric displacement for change. S2Looking therefore probes robustness to acquisition geometry. This challenge is related to heterogeneous-view construction during pretraining, where views of the same geographic sample can differ in spatial support, sensor, and acquisition time. The views are not required to be radiometrically or geometrically identical, encouraging the encoder to interpret local differences in their broader context.

B.3 Domain-Generalized Semantic Segmentation

For domain generalization we train a Mask2Former decoder on the source domain only and evaluate it on the target domain without adaptation or target-domain model selection, reporting mIoU and mAcc. The three benchmarks isolate three different kinds of distribution shift, and each provides four-band (blue, green, red, near-infrared) inputs except LoveDA, which is RGB.

Five-Billion-Pixels: cross-sensor shift. The source domain consists of Gaofen-2 multispectral images (4 m GSD), and the target domains contain the same land-cover classes observed by PlanetScope (3.7–4.1 m), Gaofen-1 (8 m), and Sentinel-2 (10 m). The underlying scene content is similar, but the sensor response, the effective GSD, and the resampling differ (Tong et al., 2023). This setting evaluates cross-sensor transfer under a fixed land-cover taxonomy without target-domain supervision.

FLAIR: cross-regional shift. FLAIR provides 0.2 m aerial imagery over France with four bands (blue, green, red, near-infrared) and 13 land-cover classes, acquired over 50 spatially separated domains between 2018 and 2021 (Garioud et al., 2023). Source and target domains are geographically disjoint, so the shift combines different landscape composition with different acquisition campaigns while the sensor stays fixed. This isolates regional and seasonal variation from sensor variation. The very high 0.2 m resolution also requires the decoder to exploit fine spatial detail. ScalSpectra improves mIoU by 1.30 points over the strongest baseline here, the second-largest margin among the three segmentation benchmarks (Table 3).

LoveDA: cross-style shift. LoveDA contains 0.3 m Google Earth imagery with three RGB bands and seven land-cover classes, and its official rural-to-urban split makes the domain gap one of scene style rather than of sensor or region: rural source images are dominated by agricultural parcels and scattered settlements, whereas urban target images contain dense building blocks and street networks (Wang et al., 2021). The same classes therefore appear with very different shapes, sizes, and spatial context, and a model that memorizes rural layouts fails. Because the encoder is never fine-tuned on urban data, this benchmark measures whether the pretrained dense features preserve object-level semantics that are shared across styles.

B.4 Domain-Generalized Height Estimation

Height estimation is evaluated with a DPT decoder trained on labeled source images and height maps and evaluated directly on an unseen target domain, reporting mean absolute error (MAE) and root mean squared error (RMSE) in metres. The three transfers below cover cross-regional, cross-style, and synthetic-to-real shifts, each evaluated without target-domain supervision.

GAMUS → DFC2019: cross-regional shift. The source domain is the GAMUS benchmark, which provides co-registered RGB imagery and normalized digital surface models in 11,507 1024×1024 tiles from five U.S. cities. For this transfer, we use the Washington D.C., Philadelphia, and New York City subsets and crop each 1024×1024 tile into non-overlapping 512×512 patches (Xiong et al., 2023). The target domain is DFC2019, which provides WorldView-3 imagery with LiDAR-derived height references over Jacksonville and Omaha (Le Saux et al., 2019). The source and target domains therefore differ in city and height distribution while sharing the same urban height-regression task. The decoder is trained only on the source domain and evaluated directly on the target domain.

GeoNRW-U → GeoNRW-R: cross-style shift. GeoNRW provides orthorectified aerial photographs at 1 m GSD with LiDAR-derived elevation over North Rhine-Westphalia, Germany, distributed as 1000×1000 tiles after resampling to a 1 m grid (Baier et al., 2021). Following the real-world height-estimation protocol of SynRS3D, we split the dataset into an urban (GeoNRW-U) and a rural (GeoNRW-R) subset. The two subsets differ in building density, vegetation coverage, and height statistics, which creates a cross-style transfer setting under changes in scene composition and height distribution. Compared with the strongest baseline, ScalSpectra lowers MAE by 0.033 m and RMSE by 0.066 m on this transfer (Table 3).

SynRS3D → GeoNRW-U: synthetic-to-real shift. SynRS3D contains 69,667 high-resolution synthetic images generated from controllable 3D urban scenes across six city styles, with eight land-cover categories and pixel-accurate height maps, and GSDs from 0.09 m to 1 m (Song et al., 2024). SynRS3D provides dense height supervision from controllable synthetic scenes, whereas GeoNRW-U consists of real aerial imagery. This transfer therefore evaluates robustness to the simulation-to-reality gap and tests whether source-trained representations remain useful when the target appearance and scene statistics change.

C Additional Method Details

C.1 Heterogeneous EO View Construction

Heterogeneous view construction turns acquisition variability into self-supervision. Section 3.2 introduces this mechanism at the pipeline level. Here we make the sampling procedure explicit because the treatment of repeated observations and band subsets determines what the scene-level and dense objectives actually observe.

A geographic sample is a location for which at least one sensor $r \in \mathbb{S}$ provides measurements. Each sensor exposes a set of physical channels $\mathbb{C}_r$, and channel c carries the descriptor $\boldsymbol{\rho}_c^{(r)}$, i.e. a centre wavelength for an optical band or a transmit/receive-polarization and orbit tuple for a SAR channel. Repeated acquisitions of channel c of sensor r are indexed by the timestamp set $\mathbb{T}_c^{(r)}$, so a location may provide several observations of the same physical band.

When repeated observations exist, we sample one timestamp per physical channel independently and resample the selected acquisition onto a common spatial grid:

$$t_c^{(r)} \sim \mathrm{Unif}\Big(\mathbb{T}_c^{(r)}\Big), \qquad \bar{\boldsymbol{X}}_c^{(r)} = \mathrm{Resample}_c^{(r)}\Big(\boldsymbol{X}_{t_c^{(r)},c}^{(r)}\Big), \qquad \bar{\mathbb{X}}^{(r)} = \Big\{\Big(\bar{\boldsymbol{X}}_c^{(r)}, \boldsymbol{\rho}_c^{(r)}\Big) : c \in \mathbb{C}_r\Big\}. \tag{A1}$$

Here $\mathrm{Resample}_c^{(r)}$ is the identity for sources that are already co-registered, such as the MMEarth modalities, which are all re-projected to the Sentinel-2 10 m grid, and it is a sensor-specific interpolation for the multi-temporal EarthView Sentinel-2 and Sentinel-1 series. Because the timestamp is drawn per channel rather than once per view, $\bar{\mathbb{X}}^{(r)}$ may combine asynchronous acquisitions of different channels. We therefore treat it as a controlled temporal perturbation rather than as a synchronized observation: it removes the assumption that all bands of a view were measured simultaneously, which is exactly the assumption that breaks down when a model is transferred to change detection or to multi-temporal EO products.

Each crop v then selects a single sensor r_v and draws only channels belonging to that sensor, so that a view never mixes modalities inside its band dimension. For crop type $\nu(v) \in \{\mathrm{G}, \mathrm{L}\}$ (global or local), the admissible band count is clipped to the sensor's actual channel set,

$$c_v^{\min} = \min\Big(k_{\nu(v)}^{\min}, |\mathbb{C}_{r_v}|\Big), \qquad c_v^{\max} = \min\Big(k_{\nu(v)}^{\max}, |\mathbb{C}_{r_v}|\Big), \qquad c_v \sim \mathrm{Unif}\big\{c_v^{\min}, \ldots, c_v^{\max}\big\}, \tag{A2}$$

and a subset $\mathbb{B}_v \subseteq \mathbb{C}_{r_v}$ of size c_v is drawn uniformly at random. Global crops use the wider range $(k_{\mathrm{G}}^{\min}, k_{\mathrm{G}}^{\max}) = (4, 13)$, which covers every sensor in the mixture, whereas local crops use $(k_{\mathrm{L}}^{\min}, k_{\mathrm{L}}^{\max}) = (1, 4)$, so that local views frequently expose the encoder to very sparse band subsets. The resulting view and descriptor sequence are

$$\mathbf{X}_v = \mathrm{Aug}_v\Big(\mathrm{Stack}_{j=1}^{c_v} \bar{\boldsymbol{X}}_{c_{v,j}}^{(r_v)}\Big), \qquad \boldsymbol{\rho}_{v,j} = \boldsymbol{\rho}_{c_{v,j}}^{(r_v)}, \qquad j = 1, \ldots, c_v, \tag{A3}$$

where Aug_v applies resize-cropping, flipping, and sensor-appropriate photometric jitter, and $c_{v,1}, \ldots, c_{v,c_v}$ denote the selected channels. The channel set and its descriptors are permuted jointly, so the stored view preserves the row-wise correspondence between image content and physical identity that SADBT relies on.

Three properties of Eq. A1–A3 drive the rest of the framework. First, the two global crops of a geographic sample are drawn from *different* sensors whenever the sample supports it, so the positive pairs used by scene-level learning are never trivially identical. Second, band subsampling changes the input width without changing the tokenizer parameters, so the same student processes three-band RGB views and variable-size subsets of multispectral and hyperspectral channels, including subsets drawn from the 202 retained EnMAP bands. Third, the descriptors $\boldsymbol{\rho}_{v,j}$ survive every sampling and augmentation step, which is what allows SADBT to recover physical identity after the band configuration has changed.

C.2 Scene-Level Representation Learning across Heterogeneous Observations

Scene-level learning has to organize one shared embedding space in which observations of the same geographic sample are semantically compatible even though their sensors, acquisition dates, spatial supports, and band configurations differ. Section 3.4 states the objective. This subsection gives its complete form, clarifies how the expansion features are formed in practice, and explains the role of its two terms.

Let $f_{\boldsymbol{\theta}}$ denote the student encoder and $h_{\boldsymbol{\theta}}$ its projection head, and let $f_{\bar{\boldsymbol{\theta}}}$ and $h_{\bar{\boldsymbol{\theta}}}$ be the corresponding exponential moving average (EMA) teacher. For a global teacher view g and a student view v of the same geographic sample, the ℓ_2-normalized class-token features are

$$\boldsymbol{z}_{\mathrm{S}}^{v} = \mathrm{norm}_2(h_{\boldsymbol{\theta}}(f_{\boldsymbol{\theta}}(\mathbf{X}_{\mathrm{S}}^{v})_{\mathrm{cls}})), \qquad \boldsymbol{z}_{\mathrm{T}}^{g} = \mathrm{norm}_2\big(h_{\bar{\boldsymbol{\theta}}}\big(f_{\bar{\boldsymbol{\theta}}}(\mathbf{X}_{\mathrm{T}}^{g})_{\mathrm{cls}}\big)\big), \tag{A4}$$

with $\mathrm{norm}_2(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|_2$ and $\boldsymbol{z}_{\mathrm{S}}^{v}, \boldsymbol{z}_{\mathrm{T}}^{g} \in \mathbb{R}^{d_z}$. Positive pairs are defined by geographic identity rather than by radiometric similarity: writing $\mathbb{P}_{\mathrm{het}}(\boldsymbol{x})$ for the set of view pairs drawn from sample $\boldsymbol{x}$, the compression term is

$$\mathcal{L}_{\mathrm{comp}} = -\alpha_{\mathrm{comp}}\, \mathbb{E}_{\boldsymbol{x}}\, \mathbb{E}_{(g,v)\sim\mathbb{P}_{\mathrm{het}}(\boldsymbol{x})} \left[\langle \boldsymbol{z}_{\mathrm{T}}^{g}, \boldsymbol{z}_{\mathrm{S}}^{v} \rangle\right], \qquad \alpha_{\mathrm{comp}} > 0, \tag{A5}$$

which pulls together the projected class tokens of heterogeneous views of the same location. Equivalently, Eq. A5 maximizes the cosine similarity between teacher and student views, so it imposes semantic agreement and says nothing about raw measurements.

Compression alone has a degenerate optimum: mapping every input to a single direction satisfies Eq. A5 for all pairs. We therefore regularize the geometry of the features used by the expansion term. For each global view g and sample i, let $\bar{\boldsymbol{z}}_i^{g} = (\boldsymbol{z}_{\mathrm{S},i}^{g} + \boldsymbol{z}_{\mathrm{T},i}^{g})/2$ be the midpoint of the normalized student and corresponding EMA-teacher class-token features. Let $\mathbb{Z}_{\mathrm{het}}^{g} = \{\bar{\boldsymbol{z}}_i^{g}\}_{i=1}^{M}$ denote the midpoint features for global view g, and let

$$\boldsymbol{\Gamma}_{\mathrm{het}}^{g} = \frac{1}{M} \sum_{i=1}^{M} \bar{\boldsymbol{z}}_i^{g} (\bar{\boldsymbol{z}}_i^{g})^{\top} \in \mathbb{R}^{d_z \times d_z} \tag{A6}$$

be its second-moment matrix. Following the coding-rate view of self-supervised learning (Wu et al., 2025), we define the coding-rate expansion with coding resolution $\delta > 0$ as

$$\mathcal{R}_{\delta}(\boldsymbol{\Gamma}_{\mathrm{het}}^{g}) = \frac{1}{2} \log\det\left(\boldsymbol{I}_{d_z} + \frac{d_z}{\delta} \boldsymbol{\Gamma}_{\mathrm{het}}^{g}\right), \tag{A7}$$

and the complete scene-level objective is

$$\mathcal{L}_{\mathrm{cls}} = \mathcal{L}_{\mathrm{comp}} - \frac{1}{|\mathbb{V}_{\mathrm{G}}|} \sum_{g \in \mathbb{V}_{\mathrm{G}}} \mathcal{R}_{\delta}(\boldsymbol{\Gamma}_{\mathrm{het}}^{g}). \tag{A8}$$

Table A2: Details of the pretraining hyperparameters. Both dense objectives share the same weight, and both are warmed up over the first two epochs. Additional optimization settings (drop path, gradient clipping, and bfloat16 mixed precision) are described in the text.

Parameter	Value
Student / EMA teacher	ViT-B/16, 12 blocks, $d = 768$, 4 storage tokens
Frozen VP teacher	DINOv3 ViT-L/16, layer 23
Global crops	2, 224×224, 4–13 bands per view
Local crops	6, 96×96, 1–4 bands per view
Masked global views	probability 0.5, mask ratio [0.1, 0.5]
EMA teacher momentum	cosine schedule $0.992 \rightarrow 1.0$
Scene loss	cosine compression + coding-rate expansion, $\delta = 0.05$, $\alpha_{\mathrm{comp}} = 1$
Dense loss weights	$\lambda_{\mathrm{MLSRL}} = \lambda_{\mathrm{CMSRD}} = 0.5$
Optimizer	AdamW, weight decay $0.04 \rightarrow 0.4$
Learning rate	2×10^{-3}, cosine decay, 2-epoch warmup
Batch size / epochs	128 per GPU $\times$ 8 GPUs / 20 epochs

The determinant in Eq. A7 is the product of the eigenvalues of $\boldsymbol{I}_{d_z} + (d_z/\delta)\boldsymbol{\Gamma}^g_{\mathrm{het}}$. Maximizing it penalizes low-rank collapse by encouraging the features to occupy more than a single direction. This per-view formulation reflects the implementation, which computes the second moment separately from the midpoint features of each global view and averages the coding-rate terms across views. Expansion therefore discourages representational collapse but does not by itself enforce separability by location, sensor, or band configuration. Compression aligns positive heterogeneous views, while expansion discourages collapse. Together, the terms support semantic compatibility *without* requiring radiometric equality, as for co-located Sentinel-2 and Sentinel-1 views in MMEarth. The log-determinants are evaluated using a Cholesky factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$, for which $\log\det(\boldsymbol{A}) = 2\sum_i \log L_{ii}$.

C.3 Pretraining Hyperparameters

Table A2 lists the pretraining configuration. The student and the EMA teacher share the same 12-block ViT-B/16 backbone with RoPE, four storage tokens, and $d = 768$. The frozen visual-prior (VP) teacher is DINOv3 ViT-L/16. We use its last-layer patch tokens ($d_{\mathrm{VP}} = 1024$), and gradients never flow into the teacher. The tokenizer uses the physical descriptors and the dynamic, configuration-aware query of Section 3.3. Views follow Eq. A3: two global crops with the wider band range and six local crops with the narrower one, where the two global crops are drawn from different sensors whenever the sample permits. Masks are applied only to global views, with masking ratios in $[0.1, 0.5]$. The scene objective uses coding resolution $\delta = 0.05$ and compression coefficient $\alpha_{\mathrm{comp}} = 1$, and both dense objectives are weighted equally. All runs use AdamW with a cosine learning-rate schedule, layer-wise learning rate decay, and bfloat16 mixed precision.

D Additional Dense Feature Analysis

Figure A1 provides additional visualizations of multispectral features from Five-Billion-Pixels. For each displayed scene and method, we extract the patch tokens of the *frozen* pretrained encoder, i.e. before any downstream fine-tuning, and visualize the first three principal components of the patch features as an RGB image. All methods capture the coarse scene layout, while their retained fine structure differs. Several baselines (CROMA, SenPaMAE, Prithvi-v2, and Copernicus) produce spatially diffuse responses across adjacent parcels or fragmented maps with isolated high-contrast patches. By comparison, the component maps from ScalSpectra more clearly delineate waterways, agricultural parcels, road networks, and built-up blocks, while retaining thin features such as canals and field boundaries. The corresponding downstream segmentation results are reported in Table 3.

E Additional Qualitative Results

E.1 Change Detection

Figure A2–A5 show the pre-event and post-event images, the ground-truth change mask, and the predictions of all compared models on the four change-detection benchmarks. In each figure the

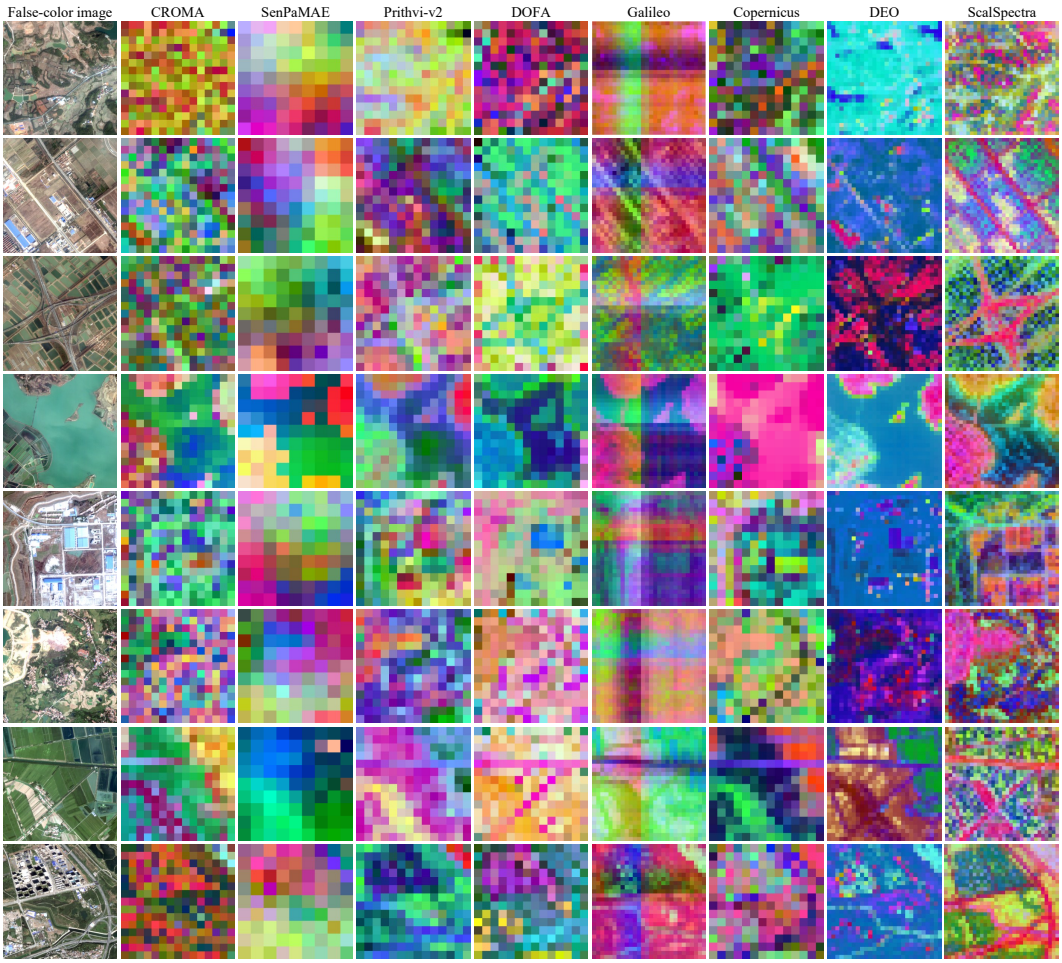

Figure A1: Additional visualizations of multispectral features from Five-Billion-Pixels. Each column after the false-color input shows the first three principal components of the patch tokens of a frozen pretrained encoder, visualized as RGB. No downstream fine-tuning or label supervision is used. In the displayed scenes, ScalSpectra more clearly delineates waterways, agricultural parcels, and built-up blocks and retains thin structures such as canals, roads, and field boundaries. The figure provides additional multispectral feature visualizations beyond Figure 4.

first three columns are the inputs and the annotation, and the remaining columns correspond to CROMA, SenPaMAE, Prithvi-v2, DOFA, Galileo, Copernicus, DEO, and ScalSpectra.

On CDD, the dense semantically discriminative features learned by ScalSpectra improve the separation of true changes from seasonal and illumination variation, leading to more reliable localization of changed regions. On LEVIR-CD+, these features improve the localization of fine-grained building changes, which are small and sparse in the image. On DSIFN, the same features improve the detection of fine-grained, low-contrast changes under radiometric and resolution differences between multi-source acquisitions. On S2Looking, they improve change localization under off-nadir viewing differences by preserving semantic structure despite geometry-induced displacement. Quantitatively, ScalSpectra achieves the highest F1 on all four benchmarks and their average in Table 2.

E.2 Domain-Generalized Semantic Segmentation

Figure A6 shows the cross-sensor setting of Five-Billion-Pixels, in which the Mask2Former decoder is trained on the Gaofen-2 source domain and applied to imagery acquired by other sensors (PlanetScope, Gaofen-1, and Sentinel-2) without any target-domain supervision. Across the shown scenes, CROMA and Galileo produce large nearly uniform regions in the first row, while some other baselines show fragmented responses near class boundaries. The ScalSpectra predictions in these examples retain connected road and water networks and follow the parcel layout more closely.

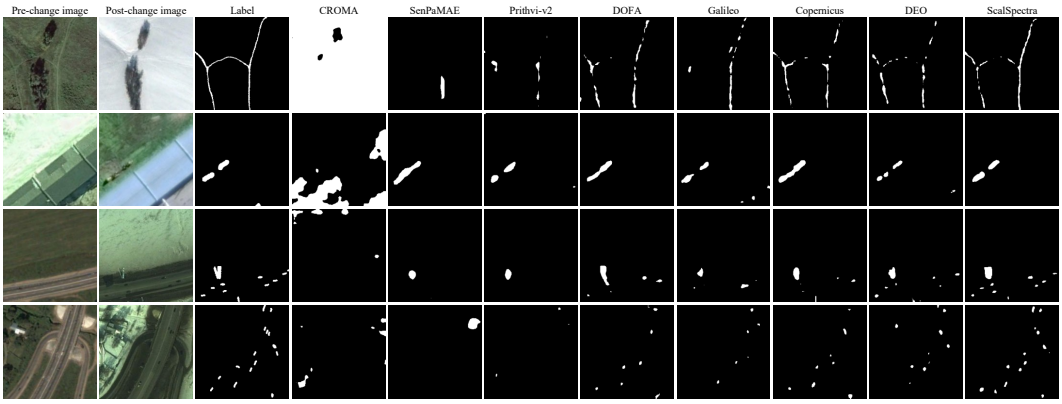

Figure A2: Qualitative change detection on CDD. Columns from left to right: pre-change image, post-change image, ground-truth change mask, and the predictions of CROMA, SenPaMAE, Prithvi-v2, DOFA, Galileo, Copernicus, DEO, and ScalSpectra. White denotes change. The benchmark emphasizes seasonal appearance variation, which produces pseudo-change in unchanged regions.

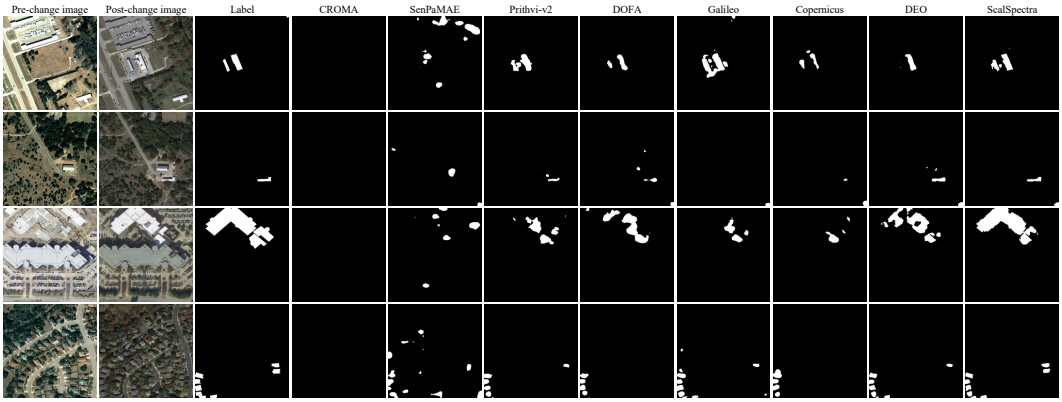

Figure A3: Qualitative change detection on LEVIR-CD+. White denotes change. The benchmark targets building construction and demolition over a long temporal baseline, so the changed objects are small and sparse relative to the image.

Figure A7 shows the rural-to-urban generalization setting of LoveDA, in which the same decoder is trained on rural source domains and applied to unseen urban target domains. In the displayed examples, some baseline predictions omit parts of the road network, merge water with neighbouring classes, or differ in the extent of forest regions. The ScalSpectra predictions retain more of the visible road grid and water boundaries in these scenes.

E.3 Domain-Generalized Height Estimation

Figure A8–A10 show predicted height maps for the three domain-generalization transfers of Table 3, together with the input image and the LiDAR-derived reference. Colour encodes height, with warmer colours corresponding to taller structures.

In the displayed GAMUS→DFC2019 examples (Figure A8), some baseline predictions smooth roof planes, merge nearby structures, or show responses along street canyons. In the second and third rows, the ScalSpectra predictions retain clearer height transitions around the large isolated building and the irregular building complex.

For GeoNRW-U→GeoNRW-R (Figure A9), the decoder is trained on urban scenes and evaluated on rural scenes. In the displayed examples, some baseline predictions show building-like responses in vegetation or reduced canopy-height variation, while the ScalSpectra maps more closely follow the reference in the forested and built regions.

For SynRS3D→GeoNRW-U (Figure A10), the decoder is trained with synthetic height supervision and evaluated on real aerial imagery. In the displayed examples, some baseline maps show diffuse responses or blurred roof-to-ground transitions, while the ScalSpectra maps retain height contrast

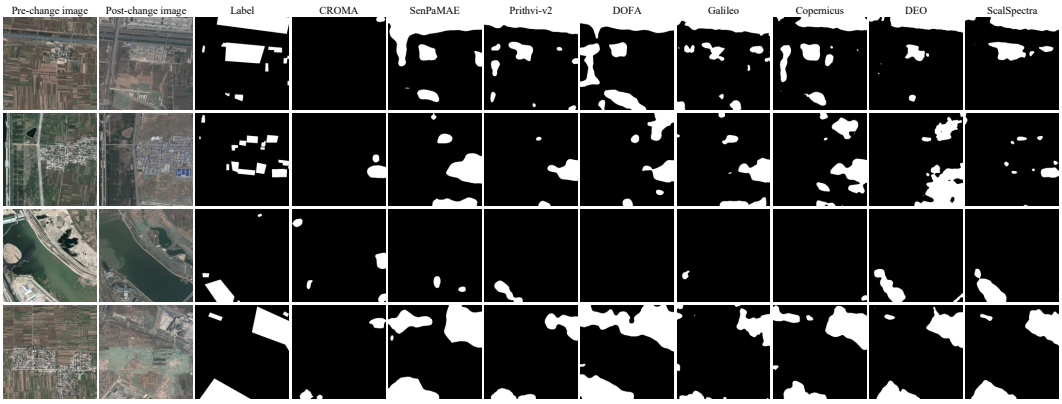

Figure A4: Qualitative change detection on DSIFN. White denotes change. The benchmark contains fine-grained object changes in multi-source imagery, where the two acquisitions differ in both radiometry and resolution, so the model cannot rely on a global appearance difference.

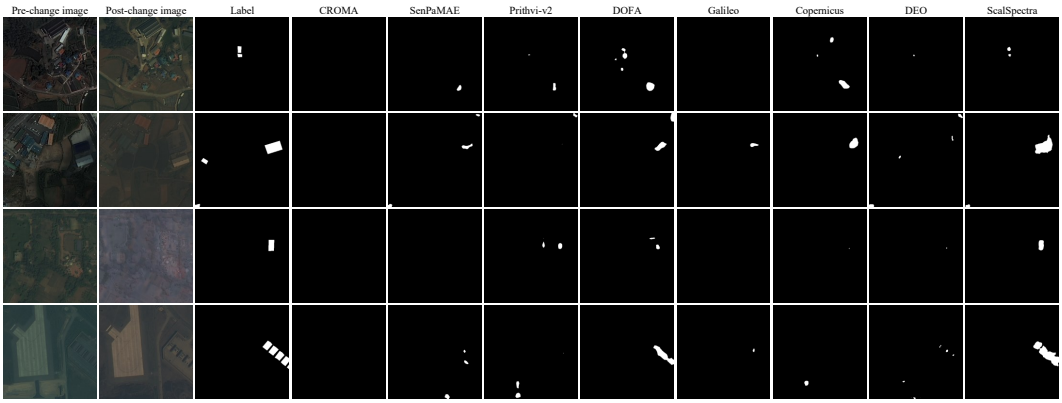

Figure A5: Qualitative change detection on S2Looking. White denotes change. The two acquisitions are captured under clearly different off-nadir angles, so unchanged buildings are displaced between dates and naive appearance comparison produces pseudo-change.

around buildings. Table 3 reports the lowest MAE and RMSE for ScalSpectra on all three height-estimation transfers.

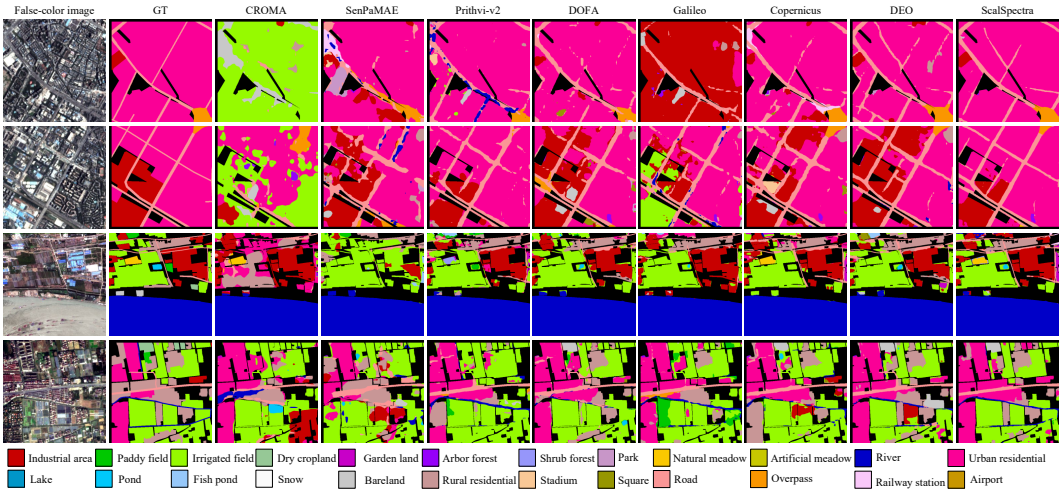

Figure A6: Domain-generalized semantic segmentation under the cross-sensor shift of Five-Billion-Pixels. The decoder is trained on the Gaofen-2 source domain and evaluated on imagery acquired by other sensors (PlanetScope, Gaofen-1, and Sentinel-2) without target-domain supervision. Columns from left to right: false-color input, ground truth, and the predictions of CROMA, SenPaMAE, Prithvi-v2, DOFA, Galileo, Copernicus, DEO, and ScalSpectra. The predictions show fine-grained land-cover categories and thin structures such as roads and canals under an unseen sensor response.

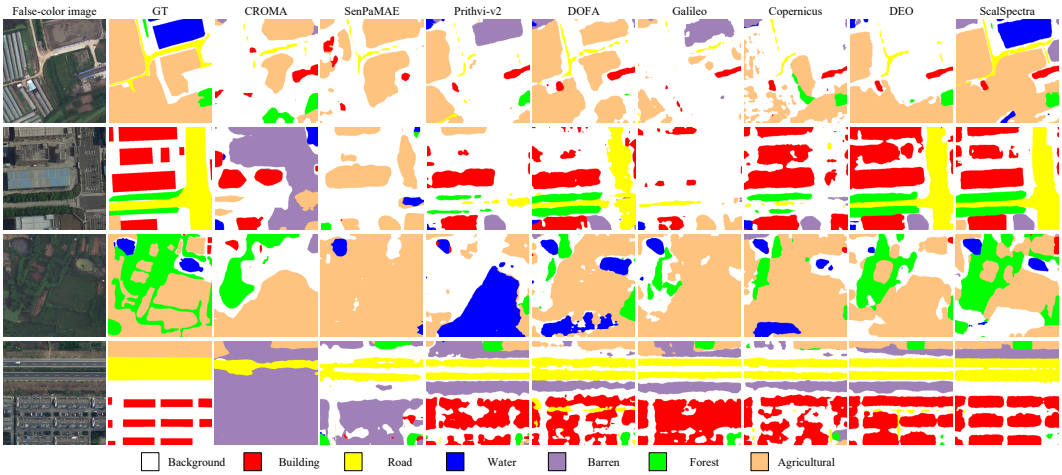

Figure A7: Domain-generalized semantic segmentation on LoveDA (rural→urban, cross-style). Columns from left to right: false-color input, ground truth, and the predictions of CROMA, SenPaMAE, Prithvi-v2, DOFA, Galileo, Copernicus, DEO, and ScalSpectra. The decoder is trained on the source domain only.

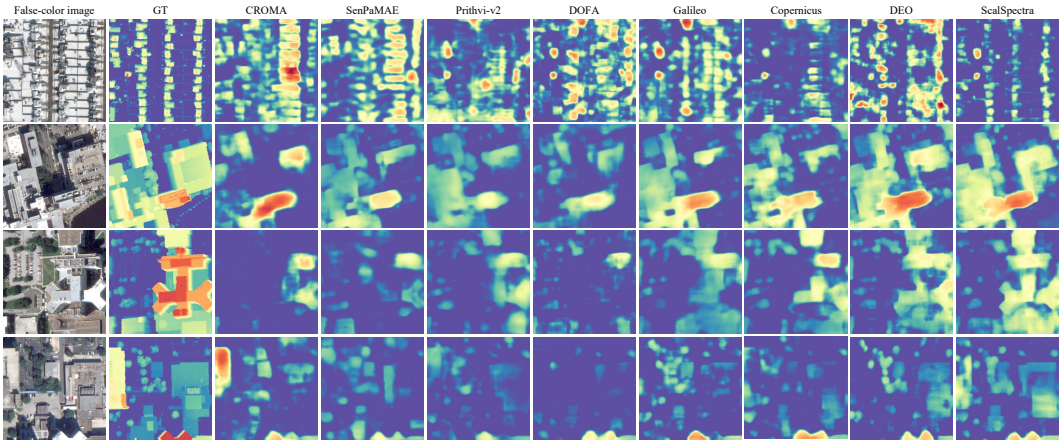

Figure A8: Domain-generalized monocular height estimation on GAMUS→DFC2019 (cross-regional). Columns from left to right: false-color input, LiDAR-derived reference height map, and the predictions of CROMA, SenPaMAE, Prithvi-v2, DOFA, Galileo, Copernicus, DEO, and ScalSpectra. Colour encodes height, and warmer colours denote taller structures, so height discontinuities between buildings and their surroundings remain visible.

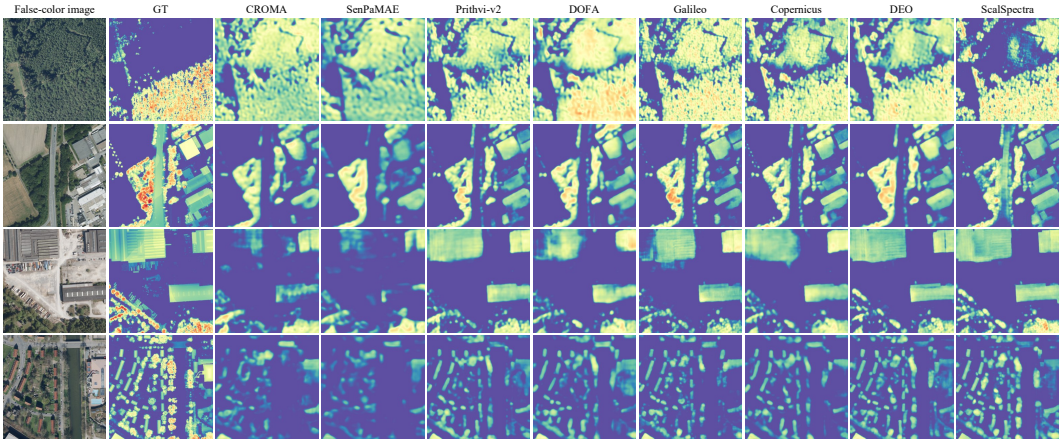

Figure A9: Domain-generalized monocular height estimation on GeoNRW-U→GeoNRW-R (cross-style): the decoder is trained on urban scenes and evaluated on rural scenes. Colour encodes height, and warmer colours denote taller structures, so the height discontinuities between individual buildings and the surrounding terrain remain clearly visible.

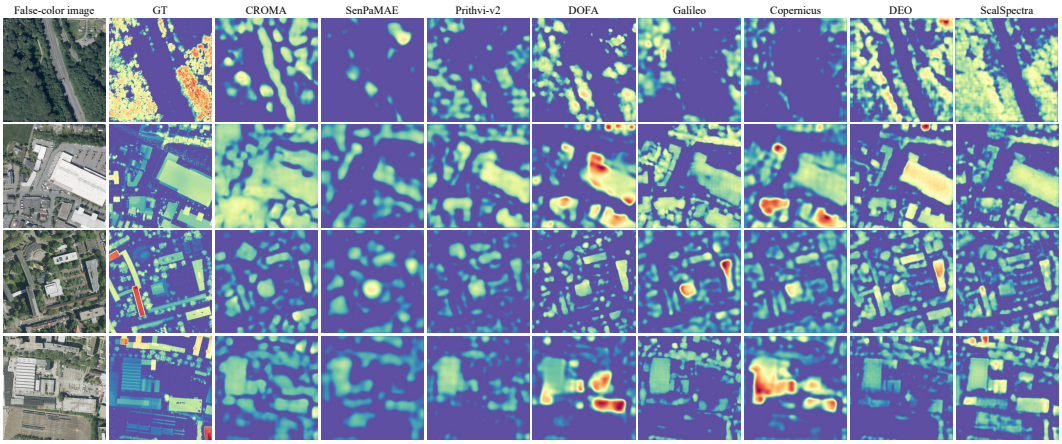

Figure A10: Synthetic-to-real height estimation, SynRS3D→GeoNRW-U: the decoder is trained on synthetic height supervision and evaluated on real aerial imagery over an unseen urban domain, where warmer colours denote taller structures, and height discontinuities remain clearly visible.