SPEAR: Self-Supervised Sample Efficient Pixel-Level Multi-Modal Spectral Fusion for Earth Observation Applications

Rajiv Ranjan Udaiveer Singh Anjali Aggarwal Shashank Tamaskar
Plaksha University, India
{rajiv.ranjan, udaiveer.singh.ug23, anjali.aggarwal, shashank.tamaskar}@plaksha.edu.in

Dharmendra Saraswat
Purdue University, USA
saraswat@purdue.edu

Abstract

Self-supervised learning (SSL) has emerged as a powerful approach in computer vision and language modeling, yet its application to Earth observation (EO) remains challenging due to heterogeneous sensors, irregular sampling, and weak physical priors. We present SPEAR, a modular, pixel-level foundation framework that learns sensor-adaptive representations across optical, radar, and climate modalities. Each modality is pretrained through a dedicated SSL learner Spectral-MAE for multispectral imagery, BYOL-Denoise for radar backscatter, and Climate-MAE for environmental variables producing compact 32-D per-pixel embeddings conditioned on wavelength, spatial scale, and geolocation. A sensor- and scale-aware fusion head integrates these embeddings via concatenation or cross-attention, while learned zero-slot tokens enable seamless operation under missing-modality conditions. Leveraging pretraining on a heterogeneous multimodal dataset Sentinel-1, Sentinel-2, PlanetScope, and climate fields the model is evaluated on four representative benchmarks: Dynamic World land-cover, Sen1Floods11 flood detection, VIIRS Active Fire detection, and Sickle crop-type classification Despite using only approximately 1M parameters, achieves competitive accuracy, demonstrating strong performance and scalability in multi-sensor settings. Overall, SPEAR establishes a lightweight, sensor-awared foundation architecture for scalable, multi-sensor Earth analytics.

1. Introduction

Earth observation (EO) - Satellite imagery offers a rich multimodal view of Earth, yet supervised learning is hindered by limited labels, sensor heterogeneity, and inconsistent cross-sensor characteristics. Traditional remote sensing relied on handcrafted spectral indices such as NDVI [45] and classical machine learning methods such as SVMs [40] and random forests [4]. However, RS (Remote Sensing) and ML (Machine Learning) pipelines exhibit substantial degradation when exposed to spatial and temporal variability [31–34, 38, 39]. Deep learning methods, particularly CNNs [29], improved representation learning for land cover classification and flood detection, but they remain data-hungry and fully supervised. Masked image modeling (MIM) [18] methods, such as MAE [10, 17, 26], alleviate the need for paired samples by reconstructing masked inputs, but most remote-sensing variants operate on a single modality or fixed spectral groupings, lacking multi-sensor integration.

Self-supervised learning (SSL) [14, 43] has emerged as a powerful paradigm, demonstrating that large models can be pretrained on vast unlabeled datasets and then adapted to diverse downstream tasks. Recent remote sensing foundation models [20, 21] such as Presto [42], SatMAE [9], and S2MAE [24] represent early progress towards general-purpose EO models. However, major gaps remain. Most existing approaches lack sensor diversity, focusing primarily on optical imagery or simply stacking multispectral bands without exploiting their physical and spectral distinctions. They rarely address pixel-level pretraining, which is critical for dense prediction tasks such as classification and crop type mapping. Many also struggle under missing-modality conditions (e.g., clouds or radar unavailability) and neglect physics and geography aware priors, relying instead on architectures developed for natural images.

To address these challenges, we introduce SPEAR a Self-Supervised Pixel-level Multi-Modal Spectral fusion framework that learns sensor-adaptive, and wavelength-continuous representations. Each modality optical, radar, and climate is modeled by a dedicated SSL learner tailored to its physical characteristics, while a flexible fusion head unifies them into per-pixel embeddings. Unlike prior patch-level approaches, SPEAR operates directly on pixel-wise spectra, enabling scalable cross-sensor learning and robust inference under missing-modality conditions through learned zero-slot tokens.

Key Contributions and Novelty. The proposed SPEAR framework is a unified, sensor-awared, self-supervised foundation model for multimodal Earth observation. It jointly pretrains optical, radar, and climate modalities at the pixel level using wavelength and geo-aware encodings to learn compact, interpretable embeddings.

- **Spectral-MAE:** A wavelength-conditioned masked autoencoder for multispectral imagery (Sentinel-2 [12], PlanetScope [2]) that incorporates center wavelength (λ) and bandwidth ($\Delta\lambda$) metadata to produce 32-D continuous spectral embeddings.
- **Climate-MAE:** A geophysical autoencoder leveraging sinusoidal latitude-longitude encodings to capture spatial and climatological continuity from LST-day/night [27], precipitation, and elevation data.
- **BYOL-Denoise:** A radar-specific Bootstrap-Your-Own-Latent [13] variant with physics-based augmentations (speckle noise, polarization dropout, intensity jitter) and a denoising head for noise-invariant SAR features.
- **Sensor- and scale-aware fusion:**A flexible head for concatenation and cross-attention integration, conditioned on sensor identity and ground-sampling distance (GSD), ensuring robustness to spatial misalignment and missing modalities.
- **Scalability and efficiency:** Using lightweight 32-D embeddings and only $\sim$1M parameters, SPEAR achieves competitive performance.

In summary, SPEAR is the first sensor- and scale-aware, pixel-level self-supervised framework that unifies optical, radar, and climate modalities within a compact, physically grounded embedding space, able to understand under any modality combination. The following section reviews related advances in SSL and EO that motivate our design choices.

2. Related Work

Self-supervised learning (SSL) progresses primarily through contrastive, masked prediction, and bootstrapping strategies. Contrastive models such as MoCo [16] and SimCLR [8] maximize agreement between augmented views while contrasting others, often requiring large batches and carefully defined positives. Bootstrapping methods like BYOL [13] and DINO [7] remove explicit negatives by training student and teacher encoders via momentum updates, yielding stable, semantically rich embeddings; DINO particularly leverages ViTs [15] with multi-crop self-distillation. Masked image modeling (MIM) [18] takes a generative route where MAE-style models reconstruct masked patches using an asymmetric encoder–decoder, learning global structure from partial inputs. These three SSL families underpin modern visual foundation models such as CLIP [30], BEiT [1], and MAE variants, enabling strong cross-domain transfer. In remote sensing, SatMAE [9] introduced spectral-temporal masking for multispectral imagery, ScaleMAE [35] added scale-awareness via GSD encoding, Cross-Scale MAE [41] enforced multi-resolution consistency, and S2MAE [24] extended masked reconstruction to 3D spectral–spatial cubes. SpectralGPT [19] further scaled MIM for hyperspectral data using 3D tokens and multi-target decoding. Beyond purely spectral models, multimodal transformers such as ContextFormer [23] and FusAtNet [28] fuse optical, meteorological, or LiDAR features via cross-attention, highlighting the value of multi-sensor integration. More recent works like AlphaEarth [6] with physics-aware implicit decoders and TESSERA [11] with dual Sentinel-1/2 encoders, push toward unified global-scale embeddings robust to missing or cloudy observations. MMST-ViT [25] advances this direction by jointly modeling multi-modal spatial-temporal signals for climate-aware crop yield prediction, emphasizing long-range temporal reasoning and modality-conditioned attention. Despite these advances, most prior efforts specialize in isolated aspects - fusion, pixel-level encoding, or missing-data robustness. SPEAR instead unifies optical, radar, and climate modalities through physics-informed self-supervised pretraining, producing compact wavelength- and resolution-conditioned 32-D pixel embeddings that remain consistent across single-modality or multimodal inputs. The next section introduces our method in detail.

3. Proposed Method

Building on the gaps identified in previous studies, we now outline the design and components of the proposed SPEAR framework. The objective is to learn per-pixel embeddings for each sensor modality and fuse them into a unified representation. Let x denote the multispectral reflectance vector, u the climate variables, and r the two-channel SAR backscatter for a given pixel. Three self-supervised encoders are employed: Spectral-MAE for optical data, Climate-MAE for environmental variables, and BYOL-Denoise for radar. Each encoder produces a 32-dimensional embedding (z_{spec}, z_{clim}, z_{rad}), which the fusion head projects into a common latent space and integrates via concatenation or cross-attention.

3.1. SpectralMAE for multispectral imagery

Let $x \in \mathbb{R}^C$ denote the per-pixel reflectance vector of C spectral bands captured by a multispectral sensor (Sentinel-2 or PlanetScope). Each band b has a central wavelength λ_b

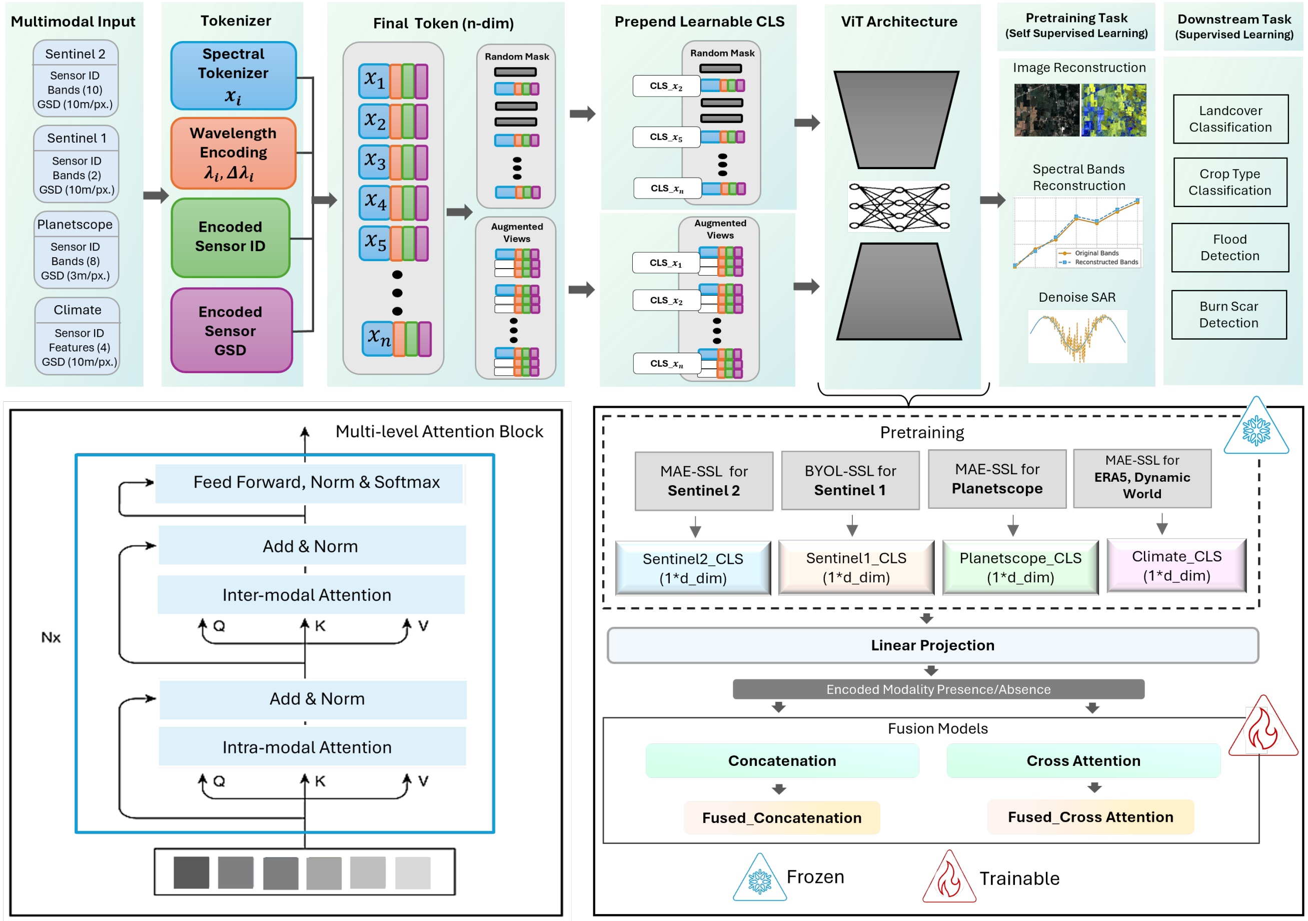


Figure 1. **Overview of the SPEAR framework.** Each modality-Sentinel-2, Sentinel-1, PlanetScope, and climate variables is tokenized with spectral, meta, and sensor-aware embeddings (wavelength λ, bandwidth $\Delta\lambda$, and ground sampling distance). Masked tokens are processed by a ViT encoder-decoder under three self-supervised objectives: spectral reconstruction (Spectral-MAE), radar denoising (BYOL-Denoise), and climate variable prediction (Climate-MAE). During fusion, pretrained modality-specific CLS tokens are linearly projected and integrated through either concatenation or cross-attention, conditioned on encoded modality-presence indicators. Frozen encoders (blue) and trainable fusion heads (red) together produce compact, 32-D pixel-level embeddings applicable to downstream tasks such as land-cover classification, crop-type classification , flood detection, and burn-scar detection.

and a full width at half maximum (FWHM [22]) $\Delta\lambda_b$. Both quantities are normalized to $[0, 1]$.

Spectral meta-embedding: To encode spectral locality and periodicity, we compute Fourier features

$$\phi_k(\lambda_b) = [\sin(2\pi k\hat{\lambda}_b),\ \cos(2\pi k\hat{\lambda}_b)], \quad k = 1, \dots, K, \tag{1}$$

and feed the concatenated vector $[\hat{\lambda}_b,\ \hat{\Delta\lambda}_b,\ \phi_1(\lambda_b), \dots, \phi_K(\lambda_b)]$ to a lightweight meta-embedding network:

$$m_b = \mathrm{MLP}_{\mathrm{meta}}\big([\hat{\lambda}_b, \hat{\Delta\lambda}_b, \phi_1, \dots, \phi_K]\big) \in \mathbb{R}^D. \tag{2}$$

This produces a wavelength-aware embedding m_b that conditions the model on sensor-specific spectral characteristics.

Tokenization and masking: Each scalar reflectance x_b is linearly projected $v_b = Wx_b$ and combined with the meta-embedding to form

$$t_b = v_b + m_b. \tag{3}$$

The sequence of tokens $T = [t_1, \dots, t_C] \in \mathbb{R}^{C\times D}$ is prepended with a learnable CLS token and passed to a transformer encoder f_θ with L layers and H heads. During pretraining, a random fraction r of the bands is masked: visible tokens $T_\mathcal{K}$ and the CLS token are encoded, while a transformer decoder g_ψ reconstructs the masked reflectances:

$$\hat{x}_b = g_\psi(f_\theta([\mathrm{CLS}, T_\mathcal{K}]), m_b), \quad b \in \mathcal{M}. \tag{4}$$

The reconstruction objective is defined over the masked spectral bands. Let $\mathcal{M}$ denote the set of masked band indices. The loss is computed as

$$\mathcal{L}_{\text{spec}} = \frac{1}{|\mathcal{M}|} \sum_{b=1}^{C} \mathbf{1}_{\{b \in \mathcal{M}\}} \, \ell(\hat{x}_b, x_b), \tag{5}$$

where ℓ is either the mean absolute error (MAE) or mean squared error (MSE). The CLS embedding $h_{\text{CLS}} \in \mathbb{R}^D$ becomes the final spectral descriptor z_{spec}.

Optical Sensors. Sentinel-2 provides C=10 spectral bands and PlanetScope SuperDove C=8 bands, both trained with a mask ratio of r_{S2}=0.6. The two models share weights in the meta-embedding network for spectral alignment but use separate linear projection layers, $W_{\text{val}}^{(\text{S2})}$ and $W_{\text{val}}^{(\text{P})}$, to account for differing radiometric scales. The resulting CLS embeddings, z_{S2} and z_{P}, act as sensor-invariant spectral descriptors for downstream fusion.

3.2. ClimateMAE for environmental variables

Environmental variables such as land surface temperature (day/night), precipitation, and elevation provide crucial context for interpreting spectral signatures. Let $u \in \mathbb{R}^{C_c}$ be the vector of climate features. We augment u with geographic coordinates (φ, λ) (latitude and longitude in degrees). To encode geographic periodicity we use sinusoidal embeddings:

$$\begin{aligned} \text{pos}(\varphi) &= [\sin(2\pi\varphi/180),\, \cos(2\pi\varphi/180)], \\ \text{pos}(\lambda) &= [\sin(2\pi\lambda/180),\, \cos(2\pi\lambda/180)]. \end{aligned} \tag{6}$$

The four values are projected via a linear layer to a D-dimensional location embedding e_{loc}. The climate variables are projected to the same dimension using a learnable matrix W_c, giving $e_{\text{clim}} = W_c u$. We concatenate e_{clim} and e_{loc} and feed the result to an MLP with dropout to obtain h_{clim}. Masked climate features are zero-filled before encoding, as this empirically stabilized training for low-dimensional inputs, optimizing

$$\mathcal{L}_{\text{climate}} = \frac{1}{|\mathcal{M}_c|} \sum_{j \in \mathcal{M}_c} \ell\big(\hat{u}_j, u_j\big). \tag{7}$$

The embedding h_{clim} of the dimension $D = 32$ is the climate representation z_{clim}. For consistency, we employ a transformer encoder of depth 6, although with a single token it reduces to a feed-forward network. The pre - training uses ERA5 reanalysis and remote sensing products matched to each pixel during the training period. A geospatial split is used to prevent leakage across regions (0.9° × 0.9° grid blocks; validation blocks randomly selected until $\approx 10\%$ of samples).

3.3. BYOL-Denoise for Sentinel-1 radar

Radar backscatter in VV and VH polarizations encodes surface roughness, moisture and structure. However, SAR images suffer from speckle noise and calibration variability. We adopt Bootstrap-Your-Own-Latent (BYOL) to learn radar embeddings invariant to such artifacts. Let $r \in \mathbb{R}^2$ be the log-scaled VV/VH pair for a pixel. We generate two augmented views r_1, r_2 using: (i) random channel dropout (with probability p_d), (ii) multiplicative jitter $r \cdot \exp(\varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, (iii) additive Gaussian noise, and (iv) random scaling. An online encoder f (a small MLP) maps each view to a latent vector $y_i = f(r_i) \in \mathbb{R}^D$, followed by a projection head g. A target encoder f' with parameters updated by exponential moving average (EMA) produces target representations y_i'. BYOL minimizes the cosine distance between online and target representations:

$$\mathcal{L}_{\text{BYOL}} = 1 - \frac{\langle g(y_1), y_2' \rangle}{\|g(y_1)\| \, \|y_2'\|} + 1 - \frac{\langle g(y_2), y_1' \rangle}{\|g(y_2)\| \, \|y_1'\|}. \tag{8}$$

To encourage the encoder to preserve physical information, we add a denoising head h that maps the latent representation back to the clean input: $\hat{r}_i = h(y_i)$. The denoising loss is an ℓ_2 penalty $\mathcal{L}_{\text{denoise}} = \|\hat{r}_1 - r\|^2 + \|\hat{r}_2 - r\|^2$. The total radar loss is

$$\mathcal{L}_{\text{radar}} = \mathcal{L}_{\text{BYOL}} + \alpha \mathcal{L}_{\text{denoise}}, \tag{9}$$

where α balances invariance and reconstruction. We schedule the EMA momentum of the target network from 0.985 to 0.996 over training epochs. The radar encoder outputs a $D = 32$ radar embedding z_{S1}.

3.4. Multi-modal Fusion

Let $z_{\text{S2}}, z_{\text{P}}, z_{\text{S1}}$, and z_{Clim} denote the 32-dimensional embeddings from Sentinel-2, PlanetScope, Sentinel-1, and Climate modalities. Each embedding is projected via a modality-specific transformation $P_m : \mathbb{R}^{32} \to \mathbb{R}^{64}$ using a linear layer with LayerNorm and GELU. Missing modalities are handled through learned *zero-slot tokens* that replace unavailable CLS embeddings, while a binary presence mask informs the transformer to process a fixed set of tokens real or placeholder without retraining or padding. This design ensures stable fusion and consistent outputs under sensor dropout, as validated by the missing-modality results in Table 3. Two fusion strategies are explored: (i) concatenation-based aggregation and (ii) cross-attention-based integration enabling inter-modality interaction.

3.4.1. Fusion via concatenation

The simplest fusion concatenates the projected embeddings:

$$\begin{aligned} f = \big[P_{\text{spec}}(z_{\text{S2}}) \| P_{\text{rad}}(z_{\text{S1}}) \| P_{\text{clim}}(z_{\text{Clim}}) \| \\ P_{\text{planet}}(z_{\text{P}})\big] \in \mathbb{R}^{4d} \end{aligned} \tag{10}$$

Absent modalities contribute zeros. A shallow classifier (e.g. random forest) is trained on f to predict the downstream labels. For robustness to missing inputs, a binary presence mask is optionally concatenated to the fused embedding before classification. Each entry in the mask indicates whether the corresponding modality (Sentinel-2, Sentinel-1, PlanetScope, or Climate) is available for that sample.

3.4.2. Fusion via Cross-Attention

We adopt a cross-attention fusion module that enables each modality to query information from all others. Let the projected modality embeddings be

$$U = [\, u_{\text{S2}},\, u_{\text{S1}},\, u_{\text{Clim}},\, u_{\text{P}} \,] \in \mathbb{R}^{M \times d},$$

where missing modalities are represented by zeros.

For each modality m, the query is

$$q_m = u_m W_Q, \tag{11}$$

with shared keys/values $K = UW_K,\ V = UW_V$. The cross-attention output is

$$\text{Attn}_m(U) = \text{Softmax}\left(\frac{q_m K^\top}{\sqrt{d_k}}\right) V, \tag{12}$$

where $d_k = d/H$ and H denotes attention heads. (Multi-head attention follows the standard formulation.)

Each attended vector is refined via a modality-specific feed-forward network ϕ_m (LayerNorm → Linear → GELU):

$$\tilde{u}_m = u_m + \phi_m(\text{Attn}_m(U))\,. \tag{13}$$

We concatenate the refined outputs

$$f = [\, \tilde{u}_{\text{S2}},\, \tilde{u}_{\text{S1}},\, \tilde{u}_{\text{Clim}},\, \tilde{u}_{\text{P}} \,] \in \mathbb{R}^{Md},$$

flattening f to obtain the fused embedding for the classifier. Cross-attention enables *early fusion*, allowing modalities to reweight each other (e.g., radar attending to climate in cloudy conditions). A binary modality mask is appended to f to indicate which modalities are present, ensuring robustness to missing inputs.

3.5. Implementation and training details

Each modality encoder is pre-trained independently on unlabeled data. SpectralMAE is trained on millions of Sentinel-2 and PlanetScope pixels using AdamW with learning rate 1×10^{-4}, weight decay 0.05. Mask ratios of 0.6 are used for Sentinel-2 and PlanetScope, respectively, and four transformer layers with four attention heads are employed. ClimateMAE uses a similar optimizer and training schedule; half of the climate variables are masked at random per sample. BYOL-Denoise for radar is trained using AdamW with learning rate 3×10^{-4} for 100 epochs and EMA momentum ramping from 0.985 to 0.996. After pre-training, encoders are frozen and only the projection layers and the downstream classifier are trained on labelled data. In practice we use random forests with 500 trees and balanced class weights for classification tasks, and pixel-wise classification with connected-component smoothing for segmentation tasks.

4. Experimental Setup

We evaluate the proposed modular pretraining and fusion pipeline on multiple downstream tasks. This section details the implementation and training of each component: (i) spectral MAEs for Sentinel-2 and PlanetScope, (ii) Climate-MAE, (iii) Sentinel-1 BYOL-Denoise, and (iv) fusion and classification head for downstream evaluation.

4.1. Pretraining Dataset

Dataset Overview. Our pretraining corpus is a large-scale, pixel-level Earth Observation (EO) dataset spanning all of India. It combines multi-modal observations from Sentinel-1, Sentinel-2, PlanetScope, and Climate, capturing spectral and spatio-temporal variability across diverse landscapes. The dataset spans 2020–2024 for Sentinel-1, Sentinel-2, and Climate, while PlanetScope provides high-resolution imagery for 2020–2021. Uniform sampling across agricultural, forested, urban, and barren regions yields a multi-year, multi-sensor, and multi-resolution corpus supporting robust and generalizable pixel-level pretraining. Sentinel-1 offers dual-polarization SAR data (VV, VH), and the Climate modality includes four geophysical variables: daytime and nighttime LST, precipitation, and elevation.

4.2. Implementation Details

We implement modality-specific self-supervised learners following a unified transformer-based design. All MAE variants use embedding dimension D=32, a 6-layer encoder with 4 attention heads, and a 3-layer decoder (D_d=32). The mask ratio is fixed at r=0.60 to promote cross-feature reconstruction. Models are trained with the AdamW optimizer (learning rate 5×10^{-4}, weight decay 1×10^{-2}), batch size 2048, cosine annealing with 10-epoch warmup, and early stopping (patience 15). Inputs are normalized using a *RobustScaler* fitted to the 10th–90th percentile range.

Sentinel-2 and PlanetScope Spectral-MAE. We construct pretraining datasets from Sentinel-2 (C=10 bands) and PlanetScope SuperDove (C=8 bands) imagery across India. For Sentinel-2, ∼5.1M pixel samples are extracted, with 2.7M balanced samples (0.3M per class) drawn from Dynamic World land-cover labels to ensure uniform class distribution. PlanetScope uses ∼5.5M pixels coregistered with Sentinel-2 and the same balanced sampling strategy. Both models share weights in the meta-embedding network for spectral alignment but maintain independent projection layers $W_{\text{val}}^{(\text{S2})}$ and $W_{\text{val}}^{(\text{P})}$ to account for differing radiometric scales. Each produces a 32-D CLS embedding (z_{S2}, z_{P}) as a sensor-invariant spectral descriptor for downstream fusion.

Climate-MAE. The Climate-MAE ingests four variables elevation, daytime and nighttime Land Surface Temperature (LST), and total precipitation along with geolocation (lat, lon) encoded via sinusoidal functions. Data are scaled with a *RobustScaler*. The model uses the same MAE configuration described above. Spatial leakage is prevented by assigning 0.9° latitude/longitude blocks exclusively to train or validation splits. Training runs for 25 epochs, with early stopping based on validation MAE.

Sentinel-1 BYOL-Denoise. Each Sentinel-1 sample includes dual-polarization SAR channels (VV, VH). We adopt a BYOL framework augmented with a denoising head, generating two augmented views (v_1, v_2) per input through a combination of channel dropout (p_{drop}=0.2), multiplicative log-jitter ($x \leftarrow x \times \exp(\epsilon)$, $\epsilon \sim \mathcal{N}(0, 0.1^2)$), random elementwise scaling $\mathcal{U}(0.97, 1.03)$, and additive Gaussian noise $\eta \sim \mathcal{N}(0, 0.02^2)$. The encoder is a 4-layer ViT (D=32, 8 heads) with an MLP projector and predictor (dimension 64). A momentum encoder updates target weights via an EMA coefficient β_t linearly ramped from 0.985 to 0.996 over the first five epochs. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{BYOL}} + 0.1\, \mathcal{L}_{\text{denoise}}, \tag{14}$$

where $\mathcal{L}_{\text{BYOL}}$ is the mean-squared error between online predictions and target projections, and $\mathcal{L}_{\text{denoise}}$ is the reconstruction loss between the denoised and clean inputs. We train using AdamW (lr $3{\times}10^{-4}$, weight decay $1{\times}10^{-4}$), batch size 1024, linear warmup for 10k steps followed by cosine decay, and early stopping (patience 10).

Fusion and downstream training. Once pretrained, each encoder produces a 32-dimensional CLS embedding (z_{S2}, z_{P}, z_{clim}, z_{S1}). We evaluate two fusion strategies:

Concatenation-based fusion. Each z_m is projected to $\tilde{z}_m = P_m(z_m) \in \mathbb{R}^{64}$ via a ModalityProjector MLP (Linear→LayerNorm→GELU→Dropout). The fused feature is $f = [\tilde{z}_{\text{S2}}, \tilde{z}_{\text{PS}}, \tilde{z}_{\text{clim}}, \tilde{z}_{\text{S1}}]$. Missing modalities are replaced by learned zero vectors to ensure a constant dimension of 256.

Cross-attention-based fusion. We project each z_m to $\tilde{z}_m \in \mathbb{R}^{64}$, stack them as $U \in \mathbb{R}^{4\times 64}$, and apply modality-wise multi-head attention: for modality m, query $Q = \tilde{z}_m W_Q$, keys $K = UW_K$, and values $V = UW_V$. The attended representation is $\hat{z}_m = \text{softmax}(QK^\top/\sqrt{d})\,V$. A residual MLP refines each modality: $w_m = \phi(\hat{z}_m) + \tilde{z}_m$. The fused vector is $f = [w_{\text{S2}}, w_{\text{P}}, w_{\text{clim}}, w_{\text{S1}}] \in \mathbb{R}^{256}$.

For downstream tasks, we normalize via a RobustScaler, and train a *RandomForestClassifier* with 500 trees, maximum depth 35, and class-balanced sampling. The dataset is split into 80% training and 20% test sets with stratification; hyperparameters are tuned on a validation split. All experiments are run on NVIDIA RTX A6000 GPU with mixed precision and early stopping.

5. Results and Discussion

Pretrained Models. We use four modality-specific encoders, each pretrained independently to capture domain-relevant spatial and spectral features: (i) Sentinel-2 (S2) optical imagery for vegetation and land-cover representation; (ii) Sentinel-1 (S1) radar backscatter for surface roughness and soil moisture estimation; (iii) Climate gridded meteorological variables for long-term environmental context; and (iv) PlanetScope (P) high-resolution imagery for fine-scale texture details. All encoders are trained using self-supervised objectives on their respective datasets and later used for downstream evaluation. S2 and P correspond to Spectral-MAE models, S1 to the BYOL-Denoise model, and Climate to the Climate-MAE. Pretraining performance for all encoders is summarized in Table 1.

Table 1. Validation reconstruction metrics across modalities and embedding dimensions (32, 64, 128). Metrics are computed on held-out validation data for each modality.

Model	Dim	Bands	R^2 ↑	F1 ↑	MSE ↓	MAE ↓
S2	**32**	10	0.9671	0.9764	0.00080	0.0128
	64	10	0.9237	0.9455	0.0022	0.0245
	128	10	0.9215	0.9475	0.0022	0.0250
S1	**32**	2	0.9245	0.9444	0.0305	0.0291
	64	2	0.9199	0.9389	0.0196	0.0213
	128	2	0.9267	0.9290	0.0240	0.0242
P	**32**	8	0.9616	0.9500	0.00057	0.0101
	64	8	0.9289	0.9436	0.0010	0.0151
	128	8	0.9389	0.9398	0.0009	0.0147
Climate	**32**	4	0.9201	0.9799	0.0137	0.0354
	64	4	0.8275	0.9650	0.0263	0.0713
	128	4	0.7521	0.9669	0.0371	0.0894

The 32-dimensional embeddings deliver consistently strong performance across all modalities. Consequently, all experiments adopt 32-dimensional embeddings for efficiency. For spectral modalities (S2 and P), incorporating band metadata, central wavelength and FWHM into the spectral meta-embedding is empirically beneficial rather than heuristic: for S2, R^2 improves from 0.8682 to 0.9671, and for P from 0.9349 to 0.9616. Unlike unified multimodal models with shared attention, our modular pretraining independently optimizes each modality for domain-specific spectral and spatial features, producing compact, well-aligned representations without higher-dimensional overhead. Table 2 reports trainable parameters and FLOPs per forward pass for this configuration, highlighting the lightweight yet expressive design of all encoders, with Spectral-MAEs remaining below 6.5M FLOPs.

Table 2. Computational and architectural complexity of modality-specific encoders. Reported for the final pretrained models used in multimodal fusion experiments. The notation 6E–3D indicates a Transformer architecture with 6 encoder layers and 3 decoder layers.

Model	Architecture Type	Trainable Parameters	FLOPs (In millions)
S2	Transformer (6E–3D)	1,277,473	6.481
P	Transformer (6E–3D)	1,277,473	5.263
Climate	Transformer (6E–3D)	1,673,345	3.944
S1	ViT + MLP Denoiser	116,226	0.299

5.1. Fusion-Based Classification

We evaluate the discriminative capability of multimodal embeddings using two classifier heads a lightweight Multi-Layer Perceptron (MLP) and a Random Forest (RF) under both concatenation and cross-attention fusion strategies.

5.1.1. Classification with MLP Head

A lightweight MLP classifier is trained on top of frozen encoder embeddings from the best pretrained checkpoints. Training uses AdamW (learning rate $3{\times}10^{-4}$, weight decay $5{\times}10^{-4}$), label-smoothed cross-entropy loss (ϵ=0.05), dropout 0.1, batch size 128, and runs for 100 epochs with balanced class sampling. As shown in Table 3, cross-attention consistently surpasses concatenation, confirming its ability to enhance inter-modality interactions even with a small classifier.

5.1.2. Classification with Random Forest Head

A Random Forest (RF) head evaluates the generality of the learned embeddings, leveraging its robustness to heterogeneous fused features. The results in Table 3 indicate that concatenation-based fusion slightly outperforms cross-attention for most modality combinations, with a peak accuracy of 90.23% when all four modalities (*S2, S1, Climate, Planet*) are fused. Overall, concatenation tends to suit non-parametric models like RF, whereas cross-attention is more effective with parametric heads such as the MLP.

Table 3. Comparison of fusion strategies across different classification heads (MLP and RF). Each configuration uses balanced samples across eight classes. Accuracy values (%) are reported for both concatenation and cross-attention strategies.

Head	Modalities	Samples	Accuracy (%) with Concatenation	Accuracy (%) with Cross-Attention
MLP	S2	5000	75.84	**78.15**
	S2 + Climate	5000	73.67	**79.68**
	S2 + S1 + Climate	5000	79.85	**83.66**
	S1 + S2 + P + Climate	5000	82.02	**83.51**
RF	S2	5000	**82.38**	82.00
	S2 + Climate	5000	**89.97**	88.62
	S2 + S1 + Climate	5000	**90.13**	88.82
	S2 + S1 + P + Climate	5000	**90.23**	84.40

5.2. Comparison with Benchmark Models (MLP Head)

To ensure a fair assessment, we further compare our multimodal fusion results against existing pretrained models that use identical modality inputs (Table 4). All models follow the same fine-tuning protocol described in Section 5.1.1, loading pretrained encoders, freezing them, and training the same lightweight MLP classification head on extracted embeddings. We compare our approach against four benchmark models - CNN baseline [29], SatMAE++ [9], ViT-S/16 (SSL4EO) [44], and Presto [42].

Table 4. Modality-matched comparison of fine-tuned performance at 50k samples. Each model uses the same input modalities and balanced class distribution. Dimensions denote the size of the final feature representation used for classification.

Model	Modalities	Samples	Dimension	Accuracy (%) with CLS	Accuracy (%) without CLS
SatMAE++	S2	50,000	768	–	69.42
SPEAR (Concat)	S2	50,000	260	67.77	75.84
SPEAR (Cross)	S2	50,000	260	**67.92**	**78.15**
ViT-S/16 (SSL4EO)	S1, S2	50,000	768	–	74.76
SPEAR (Concat)	S1, S2	50,000	260	75.60	73.67
SPEAR (Cross)	S1, S2	50,000	260	**79.73**	**79.68**
CNN	S1, S2, Climate	50,000	64	–	74.80
SPEAR (Concat)	S1, S2, Climate	50,000	260	85.76	79.85
SPEAR (Cross)	S1, S2, Climate	50,000	260	**86.00**	**83.66**
Presto	S1, S2, Climate	50,000	128	–	61.66
SPEAR (Concat)	S1, S2, Climate	50,000	260	85.76	79.85
SPEAR (Cross)	S1, S2, Climate	50,000	260	**86.00**	**83.66**

5.3. Downstream Dataset and Evaluation

To assess the effectiveness of the pretrained multimodal representations, classification experiments conducted using a Random Forest (RF) classifier to evaluate feature transferability across diverse Earth observation (EO) domains.

Datasets and Evaluation. We evaluate on four benchmark datasets Sickle [36], Sen1Floods11 [3], VIIRS Active Fire [37], and Dynamic World Landcover [5] covering agricultural, flood, fire, and global land-cover classification tasks. As summarized in Table 5, SPEAR deliv-

Table 5. Cross-dataset downstream classification performance. Each model is evaluated across multiple benchmark datasets using class-balanced samples. Metrics represent classification accuracy (%). "–" indicates cases where fusion does not apply.

Model	Dataset	Modalities	Samples	Classes	Fusion Method	Accuracy (%)
RF	Sickle	S2	200	5	–	62.60
ViT-S/16 (SSL4EO)	Sickle	S2	200	5	–	71.50
SatMAE++	Sickle	S2	200	5	–	71.25
SPEAR (Ours)	Sickle	S2	200	5	Concatenation	**73.5**
SPEAR (Ours)	Sickle	S2	200	5	Cross-Attention	71.0
RF	Sen1Floods11	S2	1,000	2	–	88.87
ViT-S/16 (SSL4EO)	Sen1Floods11	S2	1,000	2	–	96.00
SatMAE++	Sen1Floods11	S2	1,000	2	–	96.25
SPEAR (Ours)	Sen1Floods11	S2	1,000	2	Concatenation	**97.50**
SPEAR (Ours)	Sen1Floods11	S2	1,000	2	Cross-Attention	97.23
RF	VIIRS Fire	S2	5,100	2	–	91.22
ViT-S/16 (SSL4EO)	VIIRS Fire	S2	5,100	2	–	98.91
SatMAE++	VIIRS Fire	S2	5,100	2	–	98.86
SPEAR (Ours)	VIIRS Fire	S2	5,100	2	Concatenation	**99.23**
SPEAR (Ours)	VIIRS Fire	S2	5,100	2	Cross-Attention	98.18
RF	VIIRS Fire	S1+S2	5,100	2	–	90.21
ViT-S/16 (SSL4EO)	VIIRS Fire	S1+S2	5,100	2	–	98.95
SPEAR (Ours)	VIIRS Fire	S1+S2	5,100	2	Concatenation	**99.64**
SPEAR (Ours)	VIIRS Fire	S1+S2	5,100	2	Cross-Attention	98.45
RF	Dynamic World Landcover	S1+S2+Climate	5,000	9	–	80.67
Presto	Dynamic World Landcover	S1+S2+Climate	5,000	9	–	84.49
SPEAR (Ours)	Dynamic World Landcover	S1+S2+Climate	5,000	9	Concatenation	**88.60**
SPEAR (Ours)	Dynamic World Landcover	S1+S2+Climate	5,000	9	Cross-Attention	87.82

ers consistently strong performance across diverse sensing conditions and tasks, while remaining robust to missing modalities. Its modular design enables operation with one to four modalities (optical, radar, climate, and PlanetScope) without retraining or performance degradation. Furthermore, the model is highly sample-efficient, leveraging self-supervised embeddings to deliver reliable results under limited labeled data. Overall, SPEAR is both modality-agnostic and data-efficient, achieving scalable and consistent performance across diverse EO benchmarks.

6. Conclusion and Future Work

We presented SPEAR, a self-supervised, pixel-level, multimodal foundation framework that unifies optical, radar, and climate data through physics-aware pretraining. Unlike patch-based Earth observation models, SPEAR learns wavelength- and scale-conditioned embeddings directly at the pixel level, capturing fine spectral and geophysical structure while remaining robust to missing modalities. Three lightweight learners Spectral-MAE, BYOL-Denoise, and Climate-MAE produce compact 32-D embeddings that are fused via concatenation or cross-attention into a unified, sensor- and scale-aware representation. Despite its small size ($\sim$1M parameters), SPEAR achieves competitive accuracy with existing EO SSL models across diverse downstream tasks, including land-cover, flood, and wildfire classification, and operates flexibly with any subset of modalities.

Future work will explore temporal self-supervision and dynamic fusion using time-series data (e.g., Sentinel-2 and ERA5), expand pretraining to continental and global scales, and incorporate additional modalities such as LiDAR and hyperspectral data. These extensions aim toward a universal, physics-guided foundation model for scalable, multi-sensor Earth observation.

7. Acknowledgement

The authors acknowledge CNH Industrial for providing the collaboration and field resources essential to this research.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[2] Sana Basheer, Xiuquan Wang, Rana Ali Nawaz, Tianze Pang, Toyin Adekanmbi, and Muhammad Qasim Mahmood. A comparative analysis of planetscope 4-band and 8-band imageries for land use land cover classification. *Geomatica*, 76(2):100023, 2024. 2

[3] Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1floods11: A georeferenced dataset to train and test deep learning flood algorithms for sentinel-1. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 210–211, 2020. 7, 12

[4] Leo Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001. 1

[5] Christopher F Brown, Steven P Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J Pasquarella, Robert Haertel, Simon Ilyushchenko, et al. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific data*, 9(1):251, 2022. 7, 12

[6] Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. 2

[7] 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*, pages 9650–9660, 2021. 2

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 2

[9] 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. 1, 2, 7, 12

[10] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 1

[11] Zhengpeng Feng, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline Lisaius, Markus Immitzer, James Ball, Clement Atzberger, David A Coomes, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*, 2025. 2

[12] Ferran Gascon, Enrico Cadau, Olivier Colin, Bianca Hoersch, Claudia Isola, B López Fernández, and Philippe Martimort. Copernicus sentinel-2 mission: products, algorithms and cal/val. In *Earth observing systems XIX*, pages 455–463. SPIE, 2014. 2

[13] 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. 2

[14] Jie Gui, Tuo Chen, Jing Zhang, Qiong Cao, Zhenan Sun, Hao Luo, and Dacheng Tao. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9052–9071, 2024. 1

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[16] 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*, pages 9729–9738, 2020. 2

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1

[18] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey. *International Journal of Computer Vision*, pages 1–47, 2025. 1, 2

[19] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *arXiv preprint arXiv:2311.07113*, 2023. 2

[20] Ziyue Huang, Hongxi Yan, Qiqi Zhan, Shuai Yang, Mingming Zhang, Chenkai Zhang, YiMing Lei, Zeming Liu, Qingjie Liu, and Yunhong Wang. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*, 2025. 1

[21] Chunlei Huo, Keming Chen, Shuaihao Zhang, Zeyu Wang, Heyu Yan, Jing Shen, Yuyang Hong, Geqi Qi, Hongmei Fang, and Zihan Wang. When remote sensing meets foundation model: A survey and beyond. *remote sensing*, 17(2), 2025. 1

[22] Renjie Ji, Xue Wang, Chao Niu, Wen Zhang, Yong Mei, and Kun Tan. Specaware: A spectral-content aware foundation model for unifying multi-sensor learning in hyperspectral remote sensing mapping. *arXiv preprint arXiv:2510.27219*, 2025. 3

[23] A Burakhan Koyuncu, Han Gao, Atanas Boev, Georgii Gaikov, Elena Alshina, and Eckehard Steinbach. Contextformer: A transformer with spatio-channel attention for context modeling in learned image compression. In *European conference on computer vision*, pages 447–463. Springer, 2022. 2

[24] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24088–24097, 2024. 1, 2

[25] Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, et al. Mmst-vit: Climate change-aware crop yield prediction via multi-modal spatial-temporal vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5774–5784, 2023. 2

[26] Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023. 1

[27] Forough Marzban, Sahar Sodoudi, and René Preusker. The influence of land-cover type on the relationship between ndvi–lst and lst-t air. *International Journal of Remote Sensing*, 39(5):1377–1398, 2018. 2

[28] Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 92–93, 2020. 2

[29] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015. 1, 7, 12

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[31] Rajiv Ranjan and Shashank Tamaskar. Prediction of sugarcane sucrose content and optimal harvest date using multispectral time series image processing of satellite data. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 4239–4244. IEEE, 2024. 1

[32] Rajiv Ranjan, Tejasvi Birdh, Nandan Mandal, Dinesh Kumar, and Shashank Tamaskar. Evaluating sugarcane yield variability with uav-derived cane height under different water and nitrogen conditions. In *International Conference on Pattern Recognition*, pages 395–407. Springer, 2024.

[33] Rajiv Ranjan, Ying-Jung Chen, Shashank Tamaskar, and Anupam Sobti. Stubble (crop residue) burning detection through satellite images using geospatial foundation model: A case study in punjab, india. In *NeurIPS 2024 Workshop on Tackling Climate Change with Machine Learning*, 2024.

[34] Rajiv Ranjan, Anit Upadhyaya, Dinesh Kumar, and Shashank Tamaskar. Yield forecasting of sugarcane using machine learning and uav-derived canopy height. In *2024 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)*, pages 1–4. IEEE, 2024. 1

[35] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4088–4099, 2023. 2

[36] Depanshu Sani, Sandeep Mahato, Sourabh Saini, Harsh Kumar Agarwal, Charu Chandra Devshali, Saket Anand, Gaurav Arora, and Thiagarajan Jayaraman. Sickle: A multisensor satellite imagery dataset annotated with multiple key cropping parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5995–6004, 2024. 7, 12

[37] Wilfrid Schroeder, Patricia Oliva, Louis Giglio, and Ivan A Csiszar. The new viirs 375 m active fire detection data product: Algorithm description and initial assessment. *Remote Sensing of Environment*, 143:85–96, 2014. 7, 12

[38] Sambal Shikhar, Rajiv Ranjan, Aman Sa, Anshika Srivastava, Yash Srivastava, Dinesh Kumar, Shashank Tamaskar, and Anupam Sobti. Evaluation of computer vision pipeline for farm-level analytics: A case study in sugarcane. In *Proceedings of the 7th ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, pages 238–247, 2024. 1

[39] Karan Singh, Rajiv Ranjan, Sushil Ghildiyal, Shashank Tamaskar, and Neeraj Goel. 3d-ctm: Unsupervised crop type mapping based on 3d convolutional autoencoder and satellite image time series. *IEEE Geoscience and Remote Sensing Letters*, 2024. 1

[40] Shan Suthaharan. Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, pages 207–235. Springer, 2016. 1

[41] Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066, 2023. 2

[42] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023. 1, 7, 12

[43] 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. 1

[44] Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. Ssl4eo-s12: A large-scale multimodal, multitemporal dataset for self-supervised learning in earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11(3):98–106, 2023. 7, 12

[45] Jinru Xue and Baofeng Su. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of sensors*, 2017(1):1353691, 2017. 1

Appendix

A.1. Pretraining Dataset

Dynamic World Land Cover Classes. The pretraining and evaluation datasets for all modalities use land-cover annotations derived from the **Dynamic World (DW)** dataset, a global 10 m resolution product from Google Earth Engine that provides near real-time land-cover classification from Sentinel-2 imagery. Dynamic World defines ten surface classes: *(0) Water, (1) Trees, (2) Grass, (3) Flooded Vegetation, (4) Crops, (5) Shrub and Scrub, (6) Built Area, (7) Bare Ground* and *(8) Snow and Ice*. Each pixel in our dataset is associated with one of these classes, ensuring consistent sampling and class-balanced training across modalities. Figure 2 illustrates the spatial distribution of Sentinel-2 and PlanetScope samples across India and neighbouring regions.

Note: These land-cover classes form the shared label space used for downstream multimodal fusion and evaluation, allowing for consistent comparison across optical, radar, and climate encoders before fusion in the results section.

Table 6. **Spectral bands for Sentinel-2 and PlanetScope.** Central wavelength and bandwidth (nm) used in all pretraining experiments.

Sentinel-2 Band	Description	Center	BW
B2	Blue	490	65
B3	Green	560	35
B4	Red	665	30
B5	Red Edge 1	705	15
B6	Red Edge 2	740	15
B7	Red Edge 3	783	20
B8	NIR	842	115
B8A	NIR (narrow)	865	20
B11	SWIR 1	1610	90
B12	SWIR 2	2190	180

PlanetScope Band	Description	Center	BW
B1	Coastal Blue	443	20
B2	Blue	490	50
B3	Green I	531	14
B4	Green II	565	31
B5	Red	665	31
B6	Red Edge	705	15
B7	NIR I	865	40
B8	NIR II	945	40

A.2. Implementation Details

A.2.1. SpectralMAE for Multispectral Imagery

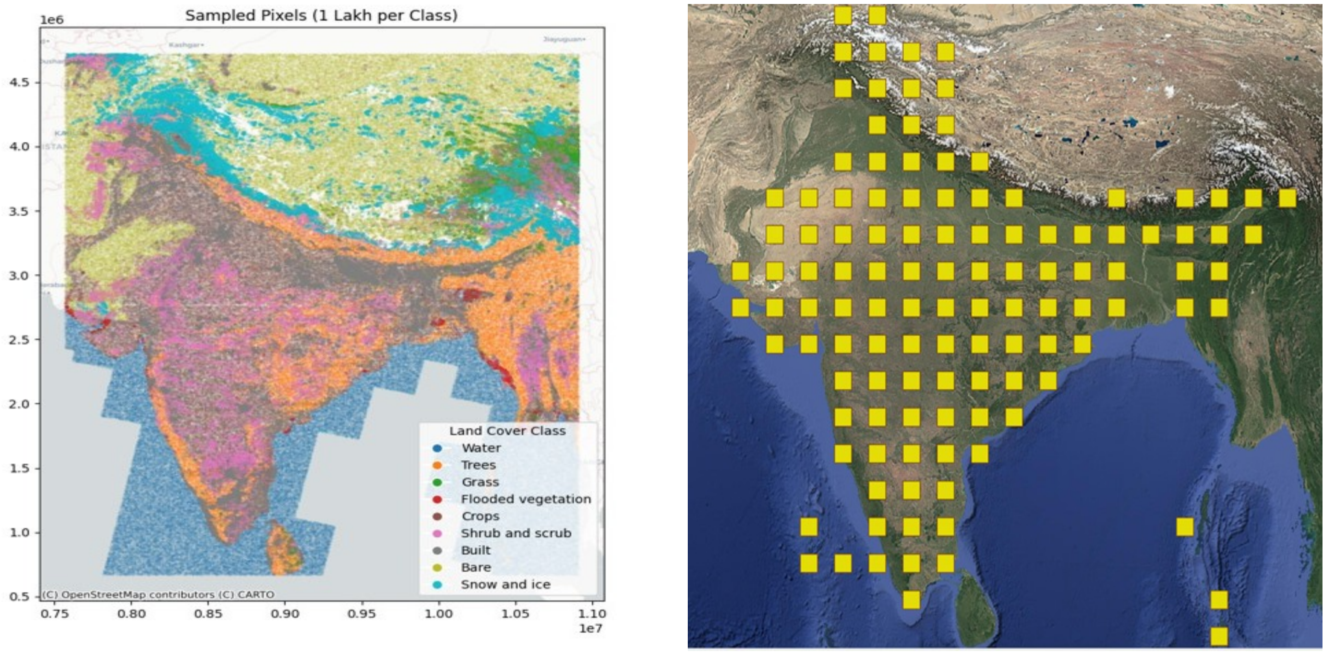
The implementation follows a fully MAE-compliant setup with several practical details. Each training batch samples masked and visible bands through random noise ranking, ensuring stochastic masking patterns per iteration. Learnable encoder and decoder mask tokens, as well as the CLS token, are initialized using truncated normal noise (σ=0.02). All transformer modules operate under mixed-precision training (`torch.amp.autocast`) with gradient scaling, and matrix multiplication precision is explicitly set to `high` for numerical stability on GPUs. Masked reconstruction loss is applied only to hidden spectral positions using boolean masking, preventing gradients from unmasked inputs. Training batches of 2048 samples are loaded with pinned memory and four worker threads; all remaining samples are retained (`drop_last=False`). Early stopping monitors validation loss with a patience of 15 epochs and a minimum improvement threshold of 10^{-6}. The best checkpoint is reloaded and only the encoder weights are exported for downstream tasks. A complete training report is automatically generated, including CSV logs, a PDF summary with loss curves, and a t-SNE visualization of CLS embeddings (perplexity = 35). These implementation components ensure reproducible and hardware-efficient pretraining consistent with MAE principles.

A.2.2. ClimateMAE for Environmental Variables

Several low-level implementation features support stable and geographically consistent pretraining. Masked feature indices are resampled independently for every batch using uniform random scores, and masking is performed by direct zero-filling of the corresponding feature values (rather than replacing them with a learnable token). Both the encoder and decoder initialize learnable parameters, including the residual mask token and per-feature query embeddings, with truncated normal noise (σ=0.02). Latitude and longitude are passed in raw degrees to preserve periodic structure and are never standardized by the feature scaler. Mixed-precision training (`torch.amp.autocast`) with dynamic gradient scaling is enabled throughout, and matrix multiplication precision is set to `high` for reproducible GPU behavior. The geographic split employs integer block identifiers computed from $\lfloor \text{lat}/0.9° \rfloor$ and $\lfloor \text{lon}/0.9° \rfloor$ to ensure spatially disjoint training and validation sets; a diagnostic check reports mean and median nearest-neighbor distances (in km) to verify spatial independence. Training and validation histories are streamed to a CSV log, and the best checkpoint is retained based on validation MAE improvement exceeding 10^{-6}. After convergence, the encoder weights and the full set of CLS embeddings are exported to disk for downstream use, enabling later retrieval without rerunning pretraining.

A.2.3. BYOL-Denoise for Sentinel-1 Radar

The training pipeline follows a fully deterministic and mixed-precision implementation. Model parameters of the target (EMA) branch are initialized as exact copies of



(a) Sentinel-2 sampled pixels - visualization of 100,000 randomly sampled pixels per Dynamic World class across India and neighbouring regions, colored by class label. Data sourced from Google Earth Engine.

(b) PlanetScope sampled grids - each grid represents uniformly sampled Planet imagery covering identical geographic regions. Planet samples are spatially gridded rather than pixel-wise. Data sourced from Planet Explorer.

Figure 2. **Dynamic World land-cover sampling across modalities.** (a) Sentinel-2 pixel-wise sampling and (b) PlanetScope grid-based sampling across India and neighbouring countries. These samples represent the nine Dynamic World land-cover classes and are used to ensure consistent cross-modal pretraining and evaluation.

the online network and updated per-step using exponential moving average momentum β_t, scheduled by a linear ramp from 0.985 to 0.996 during the first five epochs. All learnable tokens and positional embeddings are initialized with truncated normal noise (σ=0.02), and linear projection layers use Xavier uniform initialization. Training employs automatic mixed precision (`torch.amp.autocast`) with dynamic gradient scaling and explicit gradient clipping ($\|g\|_2 \leq 1.0$) for stability. A one-epoch linear warm-up scheduler precedes cosine annealing down to $\eta_{\min}$=10^{-5}, implemented via a sequential scheduler stack. Per-batch augmentations are generated on-the-fly and differ across the two BYOL views, ensuring stochastic decorrelation even within the same mini-batch. Early stopping monitors validation loss improvement over five epochs, with both the best and final model checkpoints preserved for reproducibility. After convergence, CLS embeddings and reconstructed radar amplitudes are exported in NumPy format, followed by t-SNE and K-Means analyses automatically compiled into a PDF report containing loss curves, cluster distributions, and reconstruction plots.

A.2.4. Benchmark Models

Below, we briefly describe the benchmark models used for comparison:

- **CNN baseline** [29]: A compact convolutional encoder trained on radar–optical–climate combinations to establish a non-transformer benchmark.
- **SatMAE++** [9]: A self-supervised Transformer pretrained on Sentinel-2 data using masked autoencoding for spectral reconstruction.
- **ViT-S/16 (SSL4EO)** [44]: A Vision Transformer pretrained via contrastive learning on global Sentinel-1/2 patches, serving as a strong S1–S2 baseline.
- **Presto** [42]: A lightweight multimodal Transformer leveraging S1, S2, and ERA5 inputs with coordinate embeddings for geospatial alignment.

A.2.5. Downstream Dataset

We evaluate the pretrained representations on four diverse downstream datasets that differ in spatial coverage, sensing modality, and semantic complexity:

- **Sickle[36]:** Agricultural crop classification from Sentinel-2 imagery (200 samples, 5 classes).
- **Sen1Floods11 [3]:** Flood extent detection using Sentinel-2 imagery (1,000 samples, 2 classes).
- **VIIRS Active Fire [37]:** Burned area mapping combining Sentinel-1 radar and Sentinel-2 optical data (5,100 samples, 2 classes).
- **Dynamic World Landcover** [5]: Global land cover clas-

sification from Sentinel-2, sentinel-1, climate and PlanetScope imagery with near real-time updates across 9 semantic classes (5000 samples, 9 classes).

A.3. Results

In this section, we present detailed quantitative results for each modality, complementing the global reconstruction metrics reported in Table 1 of the main paper. All evaluations follow the same inference and normalization protocols as during pretraining. Metrics are reported on held-out validation data using normalized scales: optical (S2, Planet) and climate inputs were normalized to $[0, 1]$, while radar (S1) values were computed on raw VV/VH amplitudes. Each table summarizes the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and F1 score for each class or feature. Higher R^2 and F1 values and lower MAE/MSE indicate improved reconstruction fidelity.

A.3.1. Sentinel-2 SpectralMAE

Table 7 reports per-class reconstruction metrics for the Sentinel-2 SpectralMAE, illustrating consistent performance across diverse land-cover types.

Table 7. **Per-class reconstruction metrics for Sentinel-2 SpectralMAE.** All values are computed in normalized reflectance units (0–1).

Class ID	R^2↑	F1↑	MSE↓	MAE↓
0	0.9669	0.9590	0.00029	0.0101
1	0.9708	0.9508	0.00064	0.0112
2	0.9872	0.9827	0.00090	0.0156
3	0.9661	0.9639	0.00062	0.0116
4	0.9682	0.9719	0.00119	0.0135
5	0.9855	0.9814	0.00059	0.0117
6	0.9668	0.9716	0.00124	0.0138
7	0.9807	0.9763	0.00099	0.0151
8	0.9902	0.9926	0.00124	0.0186

A.3.2. Sentinel-1 BYOL + Denoising SSL

Table 8 presents per-class results for the Sentinel-1 model trained using BYOL and denoising self-supervision. The model achieves high R^2 and F1 scores, demonstrating its ability to recover structural details from noisy radar inputs.

A.3.3. Planet SpectralMAE

Per-class results for PlanetScope imagery are shown in Table 9. Despite fewer spectral channels, the model demonstrates strong generalization with R^2 > 0.95 across most classes.

A.3.4. Climate MAE

Table 10 reports per-class results for the Climate MAE. The model attains high R^2 and F1 for temperature-related variables, while precipitation exhibits expected variability due to its higher spatiotemporal noise.

Table 8. **Per-class reconstruction metrics for Sentinel-1 BYOL+DenoisingSSL.** Metrics computed on raw VV/VH amplitude values.

Class ID	R^2↑	F1↑	MSE↓	MAE↓
0	0.9957	0.9441	0.3470	0.3424
1	0.9910	0.9910	0.1661	0.2497
2	0.9970	0.9761	0.1825	0.2316
3	0.9915	0.9884	0.3127	0.2978
4	0.9872	0.9861	0.2942	0.2846
5	0.9895	0.9857	0.3588	0.3019
6	0.9812	0.9817	0.4438	0.3438
7	0.9954	0.9820	0.3034	0.2802
8	0.9857	0.9896	0.6324	0.4060

Table 9. **Per-class reconstruction metrics for Planet SpectralMAE.** All metrics are computed in normalized reflectance units (0–1).

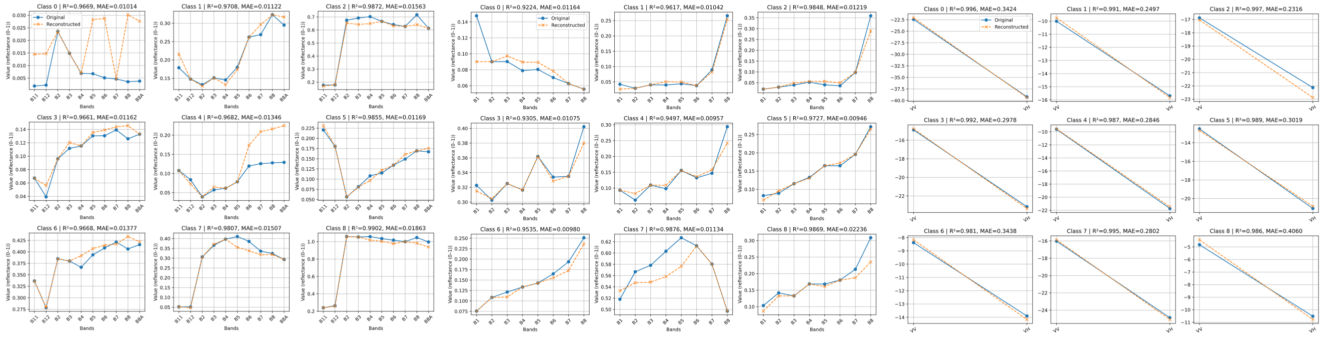
Class ID	R^2↑	F1↑	MSE↓	MAE↓
0	0.9224	0.9278	0.00076	0.0116
1	0.9617	0.9331	0.00069	0.0104
2	0.9848	0.9699	0.00077	0.0122
3	0.9305	0.9349	0.00065	0.0108
4	0.9497	0.9227	0.00051	0.0096
5	0.9727	0.9536	0.00045	0.0095
6	0.9535	0.9372	0.00049	0.0098
7	0.9876	0.9765	0.00048	0.0113
8	0.9869	0.9839	0.00167	0.0224

Table 10. **Per-class reconstruction metrics for Climate MAE.** Metrics are computed in standardized feature space.

Class ID	R^2↑	F1↑	MSE↓	MAE↓
0	0.8565	0.9865	0.0148	0.0289
1	0.7595	0.9529	0.0280	0.0540
2	0.9740	0.9769	0.0095	0.0304
3	0.8735	0.9763	0.0074	0.0186
4	0.9703	0.9776	0.0012	0.0143
5	0.9248	0.9811	0.0117	0.0376
6	0.8785	0.9655	0.0089	0.0257
7	0.9850	0.9828	0.0043	0.0300
8	0.8712	0.9964	0.0746	0.1066

These extended quantitative evaluations highlight the stability and cross-sensor robustness of the pretrained models, complementing the global metrics summarized in Table **??**.

Spectral Reconstruction Visualization. To qualitatively assess reconstruction fidelity, we visualize the original and reconstructed reflectance curves for representative samples from each land cover class across modalities. For every modality, one example per class was selected (middle index within each class), and the predicted spectral response was



(a) Sentinel-2 Spectral-MAE

(b) Planet MAE

(c) Sentinel-1 BYOL+Denoise

Figure 3. **Original vs reconstructed reflectance (per class).** Each subplot compares the true (solid) and reconstructed (dashed) reflectance curves for a representative pixel from every land cover class. The close alignment indicates accurate per-band reconstruction and strong spectral consistency across modalities.

plotted against the true reflectance values across all input bands. This allows direct comparison of band-wise reconstruction accuracy and highlights how each model preserves spectral smoothness and relative band intensities. Figure 3 illustrates these reconstructions for Sentinel-2, PlanetScope, and Sentinel-1, demonstrating strong correlation between original and reconstructed values across diverse surface types.

t-SNE Visualization of Latent Representations. To better understand the structure of the learned embeddings, we visualize the 2D t-SNE projections of the latent representations obtained from each pretrained encoder. A subset of 10,000 embeddings per modality was sampled and clustered using k-means with $K = 9$ to reveal semantic separability among different land cover types. The resulting t-SNE scatter plots show how the models organize samples with similar physical or spectral characteristics into coherent clusters, highlighting the discriminative power of the self-supervised embeddings. Figure 4 compares the t-SNE projections across Sentinel-2, PlanetScope, and Sentinel-1 modalities.

A.4. Embedding Reconstruction on Different Land Cover Types

To assess the representational fidelity of the learned embeddings, we reconstruct imagery from embeddings for different land cover types-*urban*, *vegetation*, *water*, and *snow* regions-using sentinel-2, sentinel-1 and Planet specific model.

PCA-based Reconstruction Approach. Following the approach proposed in the *TESSERA* paper, we perform a dimensionality reduction of the learned embedding vectors via Principal Component Analysis (PCA). The first three principal components (PC1, PC2, and PC3) are mapped to the RGB color channels to form a pseudo-RGB visualization. This enables interpretable spatial visualization of the embedding structure without retraining the decoder. Comparing these reconstructions with ground truth inputs provides insight into the degree to which spectral-spatial information is preserved in the latent space.

A.4.1. Sentinel-2 (S2) Embedding Reconstruction

A qualitative overview of the reconstructed Sentinel-2 (S2) embeddings can be observed in Figure 5

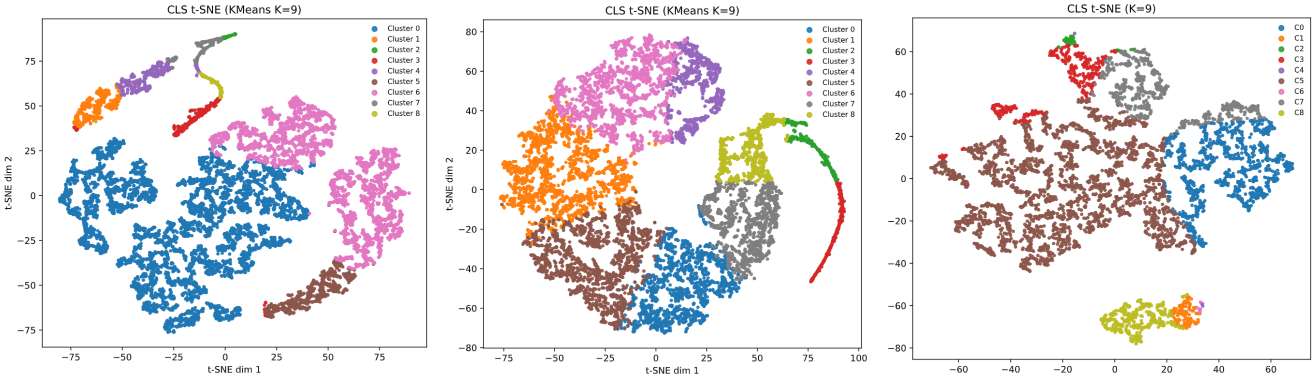
A.4.2. Sentinel-1 (S1) Embedding Reconstruction

The results of the Sentinel-1 (S1) embedding reconstruction are illustrated in Figure 6.

A.4.3. PlanetScope Embedding Reconstruction

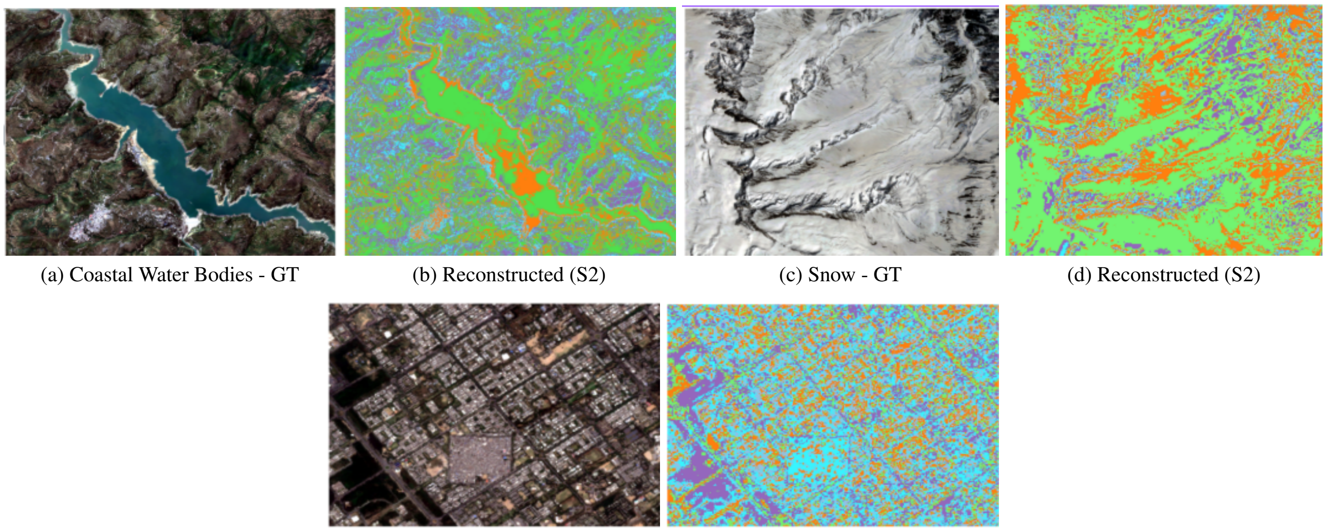
The results of the PlanetScope embedding reconstruction are illustrated in Figure 7.

A.4.4. Summary of Key Innovations



(a) Sentinel-2 Spectral-MAE (b) Planet MAE (c) Sentinel-1 BYOL+Denoise

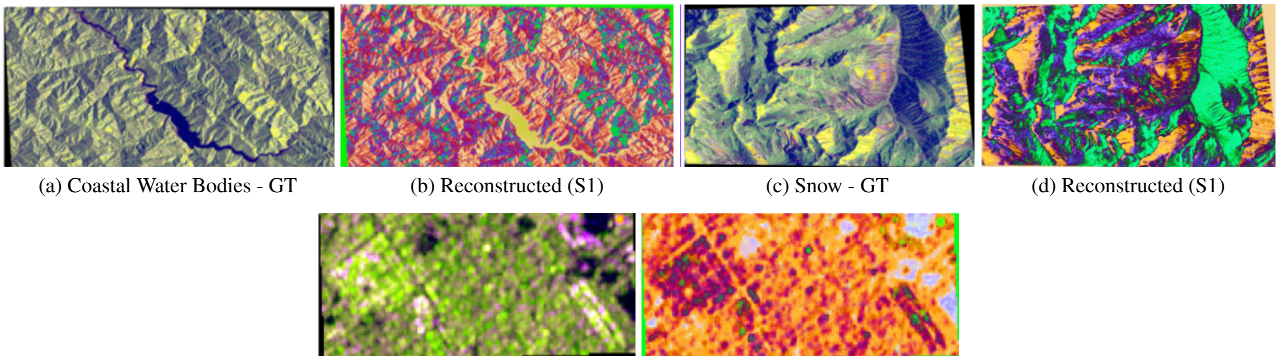
Figure 4. **t-SNE visualization of modality-specific embeddings.** Each plot shows the two-dimensional t-SNE projection of 10,000 embeddings, color-coded by K-means cluster ID ($K = 9$). The clear clustering patterns indicate that the learned latent spaces effectively capture modality-relevant semantic structure and inter-class relationships.



(a) Coastal Water Bodies - GT (b) Reconstructed (S2) (c) Snow - GT (d) Reconstructed (S2)

(e) Urban - GT (f) Reconstructed (S2)

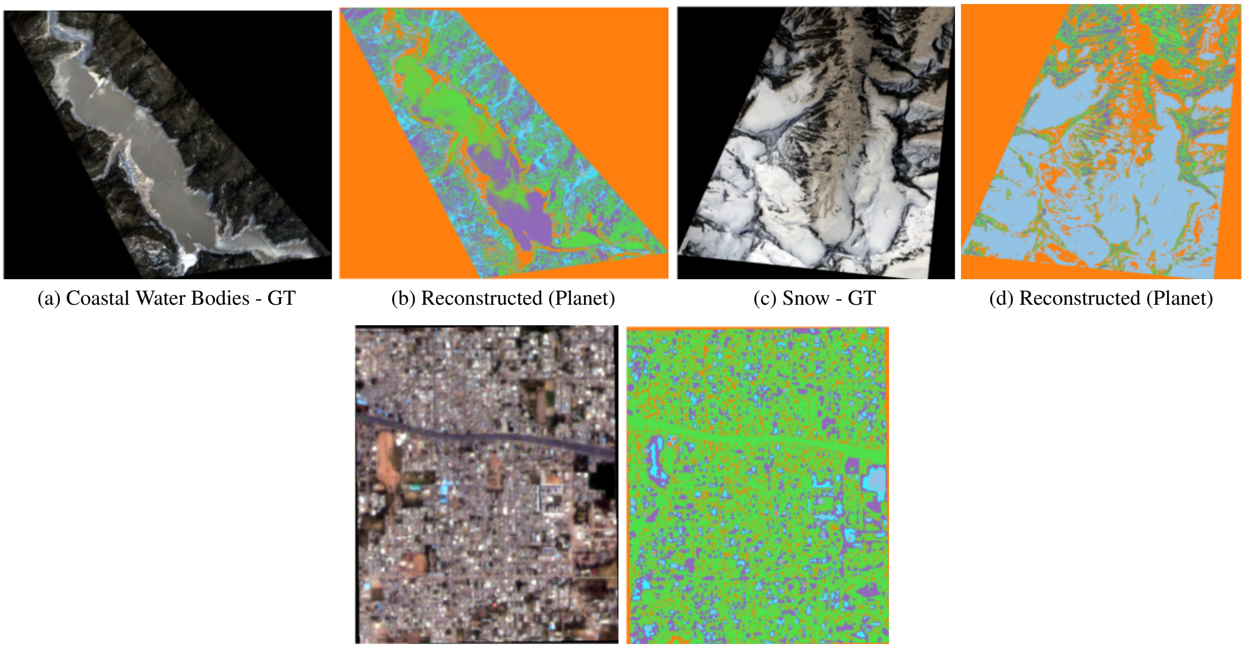
Figure 5. Sentinel-2 embedding reconstructions for different land cover types using PCA-mapped embeddings. Ground truth (GT) vs reconstructed representations are shown side-by-side.



(a) Coastal Water Bodies - GT (b) Reconstructed (S1) (c) Snow - GT (d) Reconstructed (S1)

(e) Urban - GT (f) Reconstructed (S1)

Figure 6. Sentinel-1 embedding reconstructions for different land cover types using PCA-mapped embeddings. Ground truth (GT) vs reconstructed representations are shown side-by-side.



(a) Coastal Water Bodies - GT (b) Reconstructed (Planet) (c) Snow - GT (d) Reconstructed (Planet)

(e) Urban - GT (f) Reconstructed (Planet)

Figure 7. Planet embedding reconstructions for different land cover types using PCA-mapped embeddings. Ground truth (GT) vs reconstructed representations are shown side-by-side.

Table 11. **Summary of Key Innovations.** SPEAR introduces architectural innovations for pixel-level multi-modal spectral fusion in self-supervised Earth observation.

#	**Novel Component**	**Description**	**Impact**
1	Pixel-wise SSL	Learns per-pixel latent spectra	Fine-grained precision
2	λ–$\Delta\lambda$ Meta-Embedding	Wavelength-aware spectral tokens	Sensor-agnostic reasoning
3	Dual SSL (MAE + BYOL)	Joint reconstruction + denoising	Noise robustness
4	Sensor & GSD Encoding	Sensor identity + ground sampling encoding	Cross-domain fusion
5	Pixel-level Fusion Transformer	Aligns optical, radar, and climate embeddings	Robust to missing sensors
6	Unified SSL → Downstream	Single encoder reused for all stages	Efficient training + inference
7	Modality Dropout Robustness	Invariance to missing or corrupted inputs	Stable multi-sensor deployment
8	EO-wide Pretraining (S1+S2+Planet+ERA5)	Multi-physics, multi-modal representation learning	High transferability