UniverSat: Resolution- and Modality-Agnostic Transformers for Earth Observation

Yohann Perron[*,1,5] **Guillaume Astruc**[*,1,2,4] **Nicolas Gonthier**[2,3]
Clément Mallet[2] **Loic Landrieu**[1,2]

[1]LIGM, Ecole Nationale des Ponts et Chaussées, IP Paris, Univ Gustave Eiffel, CNRS
[2]LASTIG, Univ Gustave Eiffel, IGN, ENSG, [3]IGN [4]CNES [5]EFEO

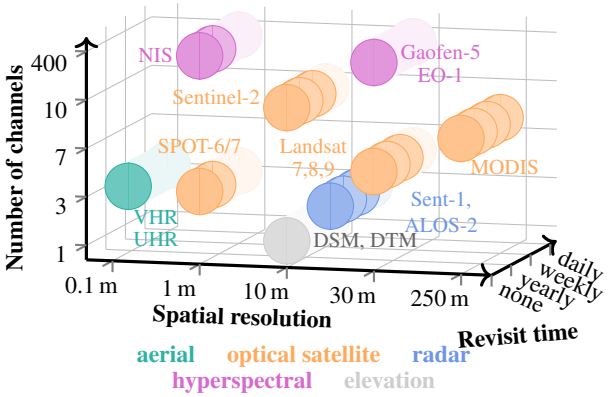


Figure 1: **One Model, Many Sensors.** A single UNIVERSAT is trained jointly on 13 sensors from 7 datasets, with wide variations in spatial resolution, channel count, and revisit frequency. The resulting model generalizes to unseen sensors within this gamut *without input resampling*.

Abstract

Vision Transformers (ViT) dominate computer vision. However, their reliance on rigid patch projectors hinders transfer to Earth Observation (EO), where input modalities, scales, and resolutions vary widely. We introduce *UniverSat*, a ViT-style backbone built around a *Universal Patch Encoder* that maps patches from arbitrary spatial, spectral, and temporal resolutions, and from both optical and non-optical sensors, into a shared embedding space with a shared set of weights. This enables training a single model on heterogeneous multimodal corpora via self-supervision, yielding robust, sensor-agnostic spatial features. We validate this approach with strong results across classification and segmentation on standard EO benchmarks from GeoBench, PANGEABench, and SpectralEarth. Our code and models are available at `github.com/gastruc/UniverSat`.

1 Introduction

Vision Transformers (ViTs) are the de facto backbone for visual representation learning, owing to their versatility and compatibility with self-supervised learning (SSL) objectives: masking [1], contrastive learning [2], predictive methods [3], and clustering-based approaches [4, 5]. This paradigm has been successfully applied to Earth Observation (EO) [6, 7, 8, 9, 10]. However, EO is more than "just vision": it spans heterogeneous *modalities* (optical, radar, LiDAR, elevation), *scales* (from hectares to tens of km^2), and *resolutions*: spatial (cm–km), temporal (single acquisitions to dense time series), and spectral (1–400 bands). This diversity is a strength only with an architecture able to exploit it.

Most EO foundation models adopt the ViT architecture with minimal adaptation, implicitly fixing band counts, temporal sampling, and patch size. When operating on a narrow set of widely available modalities (*e.g.*, Sentinel or aerial RGB imagery), this design is appropriate and largely sufficient.

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: **Flexible Multimodal EO Foundation Models.** For each model, we list the training modalities, whether unseen spatial/temporal/spectral configurations are handled *without input resampling*, and the feature-map granularity. UNIVERSAT supports the broadest modality mix, handles unseen configurations, and offers flexible output resolution—all with a single set of weights.

	training modalities					supports unseen resolution			output resolution
	📷	🎥	🌐	⛰	📊	spat.	temp.	spec.	
TerraMind [9]	1		1	1		✓			■
OmniSat [12]	2	2	1	1			✓		■
EarthView [23]	3		1	1			✓		■
Panopticon [13]	**5**		1		1			✓	■
Galileo [6]		1	1	1		✓	✓		■
AnySat [7]	2	**4**	**2**	1		✓	✓		▪
DOFA [16]	3		1		1	✓		✓	■
FoMo [24]	4	2	1	1			✓	✓	■
Ramen [17]	1	1	1	1		✓	✓	✓	◪
UniverSat (ours)	4	**4**	**2**	**2**	**3**	✓	✓	✓	◪

📷 : optical snapshot
🎥 : optical time series
🌐 : radar
⛰ : elevation
📊 : hyperspectral

output resolutions:
■ : image
▪ : patch
· : pixel
◪ : selectable

However, EO practitioners often combine sensors with widely varying pixel sizes, spectral and temporal densities, and spatial extents. Under these conditions, a rigid architecture becomes limiting and often requires aggressive resampling, band selection/aggregation, or modality-specific processing. While recent work improves robustness across scales [11, 12], datasets [7, 13], and spectral resolution [14, 15, 16, 17], most approaches remain constrained to specific modality sets and rely on modality-specific projectors. We argue that this rigidity, inherited from fixed ViT-style patch projectors, limits the generality of EO foundation models.

We introduce UNIVERSAT, a ViT-like model tailored to EO by replacing fixed patch projectors with a *Universal Patch Encoder* (UPE). The UPE uses linear-complexity axial cross-attention to embed patches of arbitrary spatial, spectral, and temporal dimensions into a shared latent space. Integrated into a transformer operating over spatialized tokens, this design provides four key advantages:

- **Modality-Agnostic.** A single set of weights processes many modalities combinations and arbitrary resolutions without input resampling or channel filtering.
- **Resolution-flexible.** The spatial resolution of the output feature map is specified at inference time and decoupled from the input patch size.
- **Granular.** A skip connection preserves fine spatial details beyond patch-level embeddings.
- **SSL-compatible.** The architecture retains the core ViT structure, enabling training with standard self-supervised learning objectives.

We train UNIVERSAT with a self-supervised objective that extends recent latent masked modeling [18, 19] to multimodal and multitemporal EO data. A single model is trained jointly on 7 datasets spanning 13 sensors across four modalities (optical, hyperspectral, radar, elevation), covering 0.1–300 m Ground Sampling Distance (GSD), 1–150 timestamps, and 1–396 spectral channels. Despite its flexibility and ability to incorporate unseen sensor configurations, UNIVERSAT remains highly competitive on standard benchmarks, as demonstrated through linear and convolutional probing on GeoBench [20], PANGEABench [21], and challenging hyperspectral benchmarks [22]. Our main contributions are as follows:

- A unified ViT-like architecture for EO that processes heterogeneous sensors without modality-specific projectors or preprocessing;
- A multimodal self-supervised training framework tailored to heterogeneous EO data;
- Competitive performance across a broad range of datasets and tasks, from high-resolution RGB imagery to challenging radar time series and hyperspectral benchmarks;
- Demonstrated generalization to unseen sensors (EnMAP, PlanetScope) and modality combinations.

2 Related work

We provide an overview of the state-of-the-art in flexible, multimodal foundation models for Earth Observation (EO); a synthesis is presented in Tab. 1, and an extended list in the appendix.

Self-supervised Learning for EO. Early EO SSL was predominantly contrastive [25, 26, 27, 28, 29]. Subsequent work explored generative masking in the spirit of MAE [8, 11, 30, 31], and predictive objectives [6, 7]. Several methods incorporate EO priors—spectral [8, 15], temporal [26, 32],

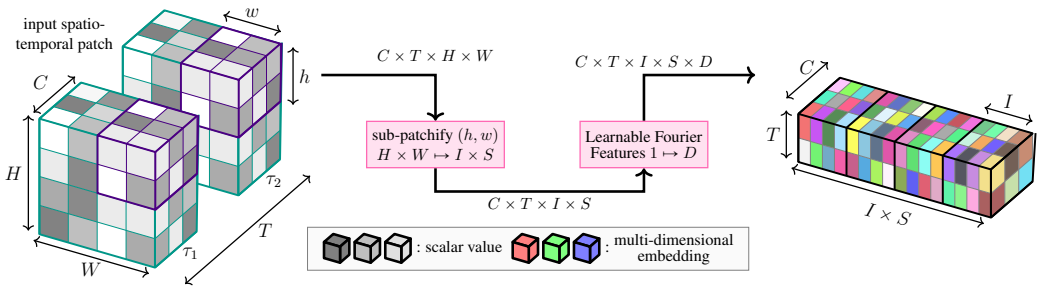


Figure 2: **Patch Formatting.** An input patch of size $C \times T \times H \times W$ is converted to a tensor of size $C \times T \times I \times S$ with S sub-patch of $I = h\,w$ pixels. We then lift all scalar values to dimension D with learned Fourier features.

and spatial [11, 25]–either in the loss or in the encoder. Nevertheless, many pipelines remain monomodal/monotemporal, which limits their generality. We also note that some SSL settings include semantic products as inputs, effectively moving toward semi-supervision [6, 10, 33, 34, 35].

Multimodal SSL for EO. Cross-modal pretraining leverages geo-registration: different sensors observe the same scene in a common reference frame. Masked cross-modal reconstruction [12, 36, 37] and cross-sensor contrast [28, 38, 39] have both been effective. Teacher–student hybrids add stability at scale [10, 35, 40]. In practice, most studies focus on Sentinel-1/2 time-series and single-date inputs [14, 38]. Some models also apply to elevation [7, 9, 12, 24, 41] and hyperspectral data [13, 16].

Flexible Models. Many EO backbones already support variable-length time series [7, 10, 36, 40]. Spectral variability is typically handled with token-per-band strategies [14, 24], band grouping [6, 8, 42], or continuous wavelength encodings and dynamic weights [13, 16, 15, 17]. Spatial generalization at inference is addressed through feature pyramids [16, 43], scale-aware positional encodings [11], or FlexiViT-style encoders [6, 17, 19, 44]; nevertheless, many pipelines still resample inputs, inflating data volume by orders of magnitude. A persistent bottleneck is the *fixed patch projector*: it enforces rigid input formats (monotemporal/fixed patch size) or modality-specific encoders [7, 6], and typically requires retraining for new configurations. *Atomizer* [45] and *Ramen* [17] are conceptually close to our work with per-pixel, per-band "atoms", but spatially resample its inputs to accommodate varying resolutions. To the best of our knowledge, UNIVERSAT is the first EO foundation model to jointly support flexibility along the spatial, spectral, and temporal axes without input resampling, see Tab. 1.

3 Method

We introduce a method to process and merge heterogeneous Earth observations into dense feature maps, independently of their respective resolutions. At its core is a *Universal Patch Encoder* (UPE) that projects patches of arbitrary shape into a shared embedding space (Sec. 3.1). Built on top of UPE, UNIVERSAT aggregates multimodal observations into dense spatial representations (Sec. 3.3), and is trained in a self-supervised manner on a heterogeneous multimodal corpus (Sec. 3.3). Each component addresses one consequence of EO heterogeneity: the UPE maps patches of any shape to a fixed width, with axial attention keeping this tractable; cross-modal contrast aligns the UPE outputs of co-registered sensors; ACA fuses modalities at the patch level; a spatial Transformer models interactions between patches; masked prediction on fixed random targets trains the whole encoder with targets of uniform size across sensors; and a sub-patch skip connection restores fine spatial detail for dense prediction.

3.1 Universal Patch Encoder

The Universal Patch Encoder (UPE) replaces the fixed patch projector of a standard ViT. Its goal is to produce a fixed-width embedding from a patch acquired by any sensor or modality, regardless of its spatial, spectral, or temporal resolution. To achieve this, UPE represents patches as atomic tokens and progressively collapses their dimensions using axial cross-attention [46].

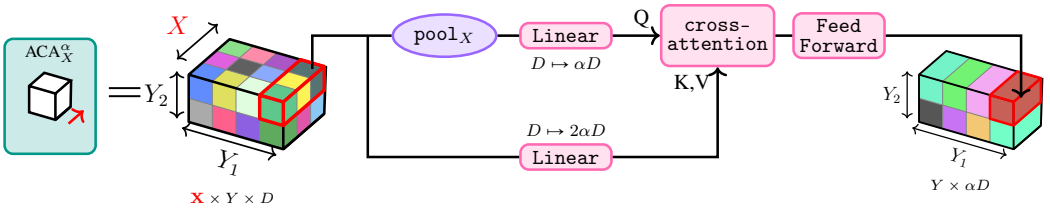


Figure 3: **Axial Cross-Attention.** Given an input of dimension $X \times Y$ tokens of dimension D (Y can be several axes), the ACA^{α}_{X} module collapses target dimension X and expands the feature dimension by α. We first pool along the target dimension X and generate queries Q for the remaining indices. We then compute keys K and values V of dimension αD for all tokens and perform cross-attention along dimension X, broadcasting along all other indices. Finally, a feed-forward network FF yields an array of size Y tokens of dimension αD.

Atomic formatting. We consider a patch $x \in \mathbb{R}^{C \times T \times H \times W}$, with C channels, T timestamps, and spatial extent $H \times W$. Channels may encode heterogeneous measurements across modalities, including spectral bands (optical), polarization responses (radar), or elevation products (*e.g.*, DSM/DTM). We factorize the spatial grid into $S = HW/(hw)$ subpatches of size $I = hw$ pixels, and reshape x as $C \times T \times I \times S$. Each scalar value $x_{c,t,i,s}$ is then lifted to dimension D using Learnable Fourier Features (LFF, see Appendix for details), yielding

$$e = \text{LFF}(x), \quad \text{dim: } C \times T \times I \times S \times D \,. \tag{1}$$

We refer to these D-dimensional vectors as *atomic tokens* [45]. Note that each atom is already a scalar (one value for one pixel, channel, and timestamp): LFF *expands* it rather than compressing a vector. LFF is cheap enough to apply to every atom while being nonlinear, which lets the network form value-selective responses (*e.g.*, band-pass or threshold-like filters on reflectance, backscatter, or elevation); a linear scalar embedding performs worse (Tab. 5).

Axial Cross-Attention (ACA). Directly projecting e with an MLP is impractical, as C, T, I, and S vary widely across sensors. Applying full self-attention over all atomic tokens would also be prohibitively expensive, since each axis can reach the hundreds. Instead, we aggregate tokens one axis at a time using axial cross-attention [46] (see Fig. 3).

Given a tensor $t \in \mathbb{R}^{X \times Y \times D}$, where X is the axis to collapse and Y groups the remaining dimensions, the module ACA^{α}_{X} produces a representation of size $Y \times \alpha D$ by attending over X only. Concretely, for each index in Y, a query is formed from pooled features, while keys and values are computed from all elements along X. Attention is then applied along axis X, followed by a feed-forward network. Restricting attention to a single axis ensures linear complexity in the number of atomic tokens. Axis-specific metadata is injected before attention with dedicated positional encodings representing the nature of the modality and the characteristics of the atom: wavelength, polarization ratio, time in year, etc. Full details are provided in the appendix.

Sequential dimension collapse. Starting from a patch of shape $C \times T \times I \times S \times D$, we progressively aggregate information by collapsing one axis at a time using ACA. We follow a fixed order: pixel within each sub-patch (I), channel (C), time (T), and finally sub-patch within the patch (S), gradually reducing the tensor while increasing the feature dimension. Collapsing I first keeps the cost manageable for high-resolution inputs, while keeping S until the end provides the sub-patch embeddings used for dense prediction. Collapsing channels before time lets temporal attention operate on spectrally contextualized features, which helps with atmospheric and illumination variations; the reverse order degrades performance on long time series (Tab. 5):

$$\text{UPE}(x) = (f,\ f^{\text{sub}}) \quad \text{with} \quad f^{\text{sub}} = \text{ACA}^{2}_{T}\big(\text{ACA}^{2}_{C}\big(\text{ACA}^{1}_{I}(e)\big)\big)\,, \quad f = \text{ACA}^{2}_{S}\big(f^{\text{sub}}\big)\,. \tag{2}$$

The UPE produces: (i) a global patch embedding $f \in \mathbb{R}^{8D}$, and (ii) sub-patch embeddings $f^{\text{sub}} \in \mathbb{R}^{S \times 4D}$, which are later used in a high-resolution skip connection. Both are then linearly projected to the width of the spatial Transformer. The expansion factors above correspond to UNIVERSAT-Tiny; UNIVERSAT-B also expands the pixel axis (ACA^{2}_{I}), see the appendix.

3.2 UNIVERSAT

We consider a tile partitioned into non-overlapping patches $\mathbf{P}$. Each patch $p \in \mathbf{P}$ is observed by a subset of modalities $\mathbf{M}$. In practice, $\mathbf{M}$ may include multiple sensors within the same modality

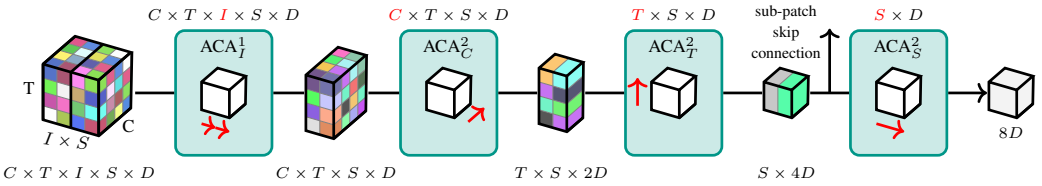


Figure 4: **Universal Patch Encoder.** The input of this module is a patch with C channels, T time stamps, S sub-patch of I pixels, and a feature dimension of D. We use our proposed Axial Cross-Attention module (ACA) to sequentially collapse the pixel, channel, time, and sub-patch dimensions while simultaneously increasing the embedding dimension. This results in a single vector of dimension $8D$ and sub-patch embeddings of dimension $S \times 4D$.

(*e.g.*, Sentinel-2 and SPOT for optical imagery), and we use the term *modality* to refer to these inputs generically. We denote by x_p^m the observation of patch p under modality m. Our objective is to produce a dense multimodal feature map $\{z_p\}_{p \in \mathbf{P}}$ at a *user-specified* spatial resolution. As illustrated in Fig. 5, UNIVERSAT follows a ViT-like backbone but differs in four key aspects: (i) patches of arbitrary spatial, spectral, and temporal resolutions are embedded using a shared UPE, (ii) co-registered modalities are fused via axial cross-attention, (iii) fine spatial details are preserved through a sub-token skip connection, (iv) the output resolution is specified at inference time.

Patch-level modality fusion. For each patch p and modality m, we compute $f_p^m,\ f_p^{m,\text{sub}} = \text{UPE}(x_p^m)$, where f_p^m is the patch embedding and $f_p^{m,\text{sub}}$ are the corresponding sub-patch embeddings. For a fixed patch p, the modality-specific embeddings $\{f_p^m\}_{m \in \mathbf{M}}$ are stacked along the modality axis and fused using an axial cross-attention module that collapses the modality dimension:

$$f_p = \text{ACA}_M^1\left(\{f_p^m\}_{m \in \mathbf{M}}\right). \tag{3}$$

This yields a single multimodal embedding per patch.

Spatial Transformer. The multimodal patch embeddings $f_{\mathbf{P}} = \{f_p\}_{p \in \mathbf{P}}$ are then processed by B gated Transformer blocks [47], denoted $\text{Trans}^1, \cdots, \text{Trans}^B$:

$$g_{\mathbf{P}} = \text{Trans}^B \circ \cdots \circ \text{Trans}^1\left(f_{\mathbf{P}}\right), \tag{4}$$

yielding refined patch embeddings g. We use RoPE positional encodings over patch centers, scaled by patch size [11], and include four register tokens [48].

Any-Resolution Prediction. We can specify a target ground sampling distance (GSD), defining a new output spatial grid $\mathbf{O}$. The patch embeddings are resampled to this grid via bilinear interpolation:

$$g_{\mathbf{O}} = \text{Bilinear}\left(g_{\mathbf{P}},\ \mathbf{P} \to \mathbf{O}\right). \tag{5}$$

To recover fine spatial details lost during patch-level aggregation, each target token attends to all sub-patch embeddings via cross-attention CA with a residual connection:

$$z_{\mathbf{O}} = g_{\mathbf{O}} + \text{CA}\left(g_{\mathbf{O}}, \{f_p^{m,\text{sub}}\}_{p \in \mathbf{P},\, m \in \mathbf{M}}\right). \tag{6}$$

3.3 Training

We exploit spatial alignment across modalities to train UNIVERSAT in a self-supervised manner on multiple heterogeneous datasets. Our objective combines (i) cross-modal contrast at the patch level and (ii) latent multimodal masked modeling (LM_3).

Scale Augmentation and Masking. We train with aggressive input dropping to improve robustness across scales and sensor configurations. For each tile, we sample the input patch size and target output resolution from dataset-specific presets, then apply a compositional masking operator that drops modalities, timestamps, channels, and patches. This produces a visible subset of modalities $\mathbf{M}^\star$ and patches $\mathbf{P}^\star$, while masked patches $\mathbf{D} = \mathbf{P} \backslash \mathbf{P}^\star$ are used for masked prediction. Overall, approximately 90% of input atoms are removed, reducing computation and encouraging invariance across modalities, temporal sampling, spectral density, and spatial scale. Sampling details are given in the Appendix.

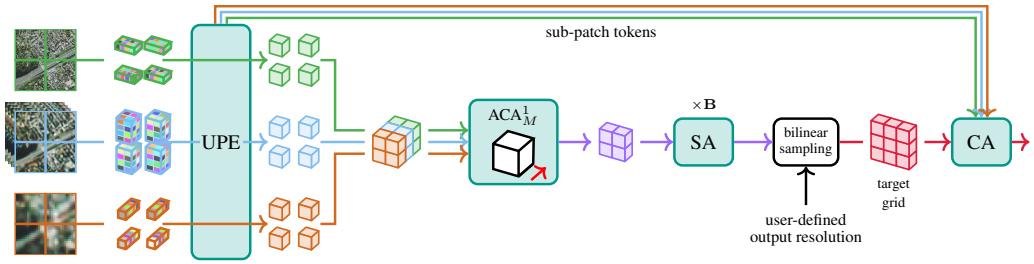


Figure 5: **UNIVERSAT.** A tile is observed by multiple sensors of arbitrary modality and resolution. Inputs are *patchified* and embedded by a shared Universal Patch Encoder (UPE). The resulting tokens are stacked along the modality axis and collapsed to one token per patch via Axial Cross-Attention (ACA). We then apply B self-attention (SA) blocks and resample the resulting feature map to the target resolution. Finally, the token attend high-resolution sub-token embeddings via cross-attention (CA) to recover fine spatial details.

Cross-modal contrast. Different modalities observing the same patch share a common latent variable: their content [12, 28]. We encourage modality-invariant representations by applying a batch-wise, multi-positive contrastive loss $\mathcal{L}_{\text{con}}$ [7] to the UPE embeddings $\{f_p^m\}_{p\in\mathbf{P}^\star,\, m\in\mathbf{M}^\star}$ of visible patches. Embeddings of the same patch in different modalities are positives, embeddings of different patches in different modalities are negatives, and pairs from different patches in the same modality are excluded, as they are trivially separable.

Linear Multimodal Masked Modeling (LM$_3$). We extend latent masked image modeling [49] to multimodal and multitemporal EO data by predicting representations of masked patches at selected timestamps. Following [19], we define targets in a latent space using frozen random projections. Concretely, for each modality m, we use a randomly initialized linear projection ϕ_m^{rand} that maps monotemporal input patches to a fixed-size target space and remains frozen during training.

A key challenge is to balance spatial and temporal supervision. Using all timestamps may lead the model to exploit trivial temporal cues, while ignoring time degrades representation quality. We therefore sample a small set of timestamps per tile and assign each masked patch $p \in \mathbf{D}$ a target time $\tau(p)$. Given the visible multimodal patch embeddings $\{z_q\}_{q\in\mathbf{P}^\star}$, a predictor ϕ^{pred} infers representations for masked patches conditioned on $\tau(p)$:

$$\left\{z_p^{\tau(p)}\right\}_{p\in\mathbf{D}} = \phi^{\text{pred}}\left(\{z_q\}_{q\in\mathbf{P}^\star}\right). \tag{7}$$

Modality-specific heads ϕ_m^{head} then map these embeddings to the target space, supervised with a contrastive objective in which the target of the same patch, modality, and timestamp is the only positive, and the targets of all other masked patches are negatives:

$$\mathcal{L}_{\text{LM}^3} = \sum_{m\in\mathbf{M}}\sum_{p\in\mathbf{D}} \text{InfoNCE}\left(\phi_m^{\text{head}}\left(z_p^{\tau(p)}\right), \phi_m^{\text{rand}}\left(x_p^m[\tau(p)]\right)\right). \tag{8}$$

Although the prediction heads ϕ_m^{head} are modality-specific, they all take the same multimodal patch representation z_p as input. Solving LM$_3$ therefore requires z_p to preserve information from all modalities in a shared latent space. As the modality-specific heads are used only during pretraining, the model itself remains modality agnostic. Finally, since the targets are random projections and not learned, the objective avoids collapse. See Appendix for more details. Random projections give targets of the same size for sensors whose raw patches differ in dimension by orders of magnitude, and InfoNCE is less sensitive than a regression loss to the widely varying scale of these targets; both choices clearly outperform their alternatives (Tab. 5). The two losses are complementary: $\mathcal{L}_{\text{con}}$ only supervises the UPE and aligns the content shared across modalities, whereas $\mathcal{L}_{\text{LM}^3}$ trains the full encoder and requires z_p to retain modality-specific information.

Final Loss. The final loss writes: $\mathcal{L} = \mathcal{L}_{\text{LM}^3} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$, where λ_{con} is a non-negative hyperparameter.

4 Experiments

We first describe the experimental setting and our multi-dataset training corpus (Sec. 4.1). We then report probing experiments demonstrating how the embeddings learned by UNIVERSAT transfer to

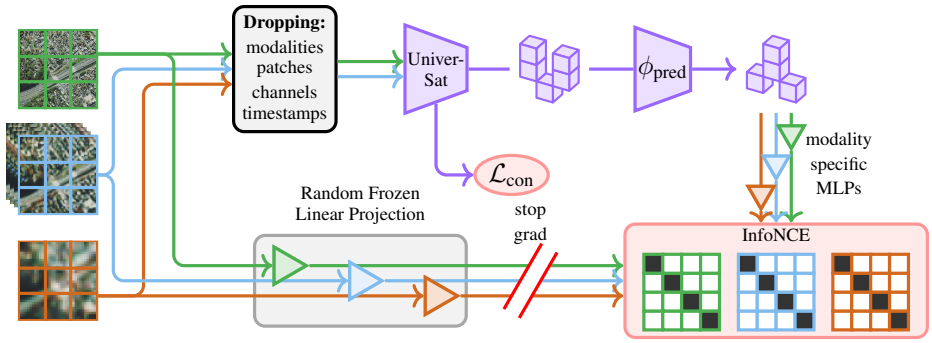


Figure 6: **Training Scheme.** We give a UNIVERSAT network a heavily masked version of the input patches. We apply a cross-modal contrastive loss to harmonize the output of its Universal Patch Encoders. We then give the embedded patches that were not masked to a decoder network, which tries to predict the values of random projections of the raw input patches. We use a batch-wise InfoNCE loss on each modality separately.

diverse and challenging EO tasks (Sec. 4.2). Finally, we provide an ablation study analyzing the impact of key architectural and training choices (Sec. 4.3).

4.1 Experimental setting

Training Set. We train a single UNIVERSAT model on a composite corpus of seven heterogeneous datasets: FLAIR-Hub [50], PASTIS-HD [51, 12], TSAI-TS [52, 12], Planted [53], S2NAIP-Urban [54, 55], HyperGlobal [56], and the NEON subset of EarthView [23]. Note that folds 4 and 5 of PASTIS, used for validation and testing in our benchmarks, are excluded from the unsupervised training set, even though no labels from PASTIS are used during pretraining. An overview is provided in Fig. 7, with full details in the appendix. The sensors span a wide range of acquisition conditions: spatial resolutions from 0.1 to 300 m, temporal depth from 1 to 140 images per year, and spectral width from 1 to 396 channels, with tile extents ranging from 0.4 to 600 ha. To our knowledge, UNIVERSAT is the first EO foundation model jointly trained on such a broad set of 13 sensors and modalities, including some that are rarely incorporated into large-scale pretraining, such as very-high-resolution hyperspectral time series.

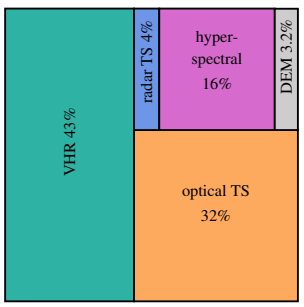
4.2 Benchmarks

We evaluate our model on 16 datasets from GeoBench [20], PangaeaBench [21], and SpectralEarth [22]. UNIVERSAT remains competitive with less versatile approaches in standard settings, while also delivering strong performance on more specialized tasks such as hyperspectral analysis.

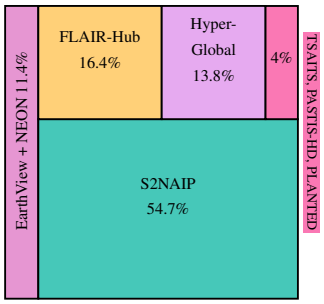
Probing Experiments. In Tab. 2, we report results in a strict probing setting (kNN and linear probing only), following the protocol of Tseng et al. [6]. Our model achieves strong performance across a wide range of datasets, including state-of-the-art results on BrickKiln [57] and Sen1Flood11 [58], despite being significantly more general than competing approaches, which are typically restricted to specific sensors (*e.g.*, Sentinel-1/2 or Landsat) or single timestamps.

On PangaeaBench (Tab. 3), unlike prior work using large decoder heads (*e.g.*, UperNet [61]), we exploit UNIVERSAT's dense embeddings and perform semantic segmentation with a simple linear probe. Despite using 3700–5000× fewer supervised parameters, our model remains competitive and reaches state-of-the-art performance on PASTIS-R [51] and AI4Farms [62]. In particular, it outperforms RAMEN [17], which also uses a shared encoder but resamples inputs to a common format, on two of the four datasets. Under full end-to-end fine-tuning, UNIVERSAT trains stably and reaches state-of-the-art performance on Sen1Floods11 (Sec. A.4.1).

Importantly, we separate unseen *configurations* of known sensor families—mono-temporal Sentinel-1/2, reduced band subsets, and derived products such as HLS—from genuinely *unseen sensors*, absent from pretraining in any form: PlanetScope and EnMAP. UNIVERSAT is robust in both regimes: on DynamicEarthNet (PlanetScope), its 9K linear probe reaches 37.1 mIoU, above a fully supervised ViT (36.8) and close to UperNet decoders using 4400× more supervised parameters.



(a) Atoms across modalities.



(b) Atoms across datasets.

	S	T	C	Gå
NAIP	1.2	1	4	540
aerial UHR	0.2	1	3	271
SPOT6/7	1.5	1	4	20.0
UAV UHR	0.1	3	3	44.0
Sentinel-2	10	20+	10	3534
Landsat-7/8/9	30	4-20	3-8	69.2
MODIS	250	60	7	0.6
Sentinel-1	10	100+	3	960
ALOS2	30	4	3	0.3
EO1	30	1	175	143
NIS	0.1	3	396	173
Gaofen-5	30	1	150	151
DSM	30	1	1	71.2
nDEM	0.2	1	1	63.0

(c) Training modalities.

Figure 7: **Training Datasets.** Distribution of atoms (å), defined as one pixel, one band, and one timestamp [45], across modalities (Fig. 7a) and datasets (Fig. 7b). Figure 7c summarizes the supported sensors, reporting their typical spatial resolution (S, in meters), temporal depth (T, in images per year), number of channels (C), and total number of atoms (Gå).

Hyperspectral Data. We evaluate our model on the SpectralEarth benchmark [22] (Tab. 4), which consists of multiple tasks based on EnMAP hyperspectral imagery (up to 500 bands), following their probing protocol. Compared to DOFA [16], a foundation model trained on EnMAP, our approach achieves consistently better performance across all tasks, despite not being trained on EnMAP.

We also compare to specialized hyperspectral models. Our method outperforms SpatSIGMA [56] and approaches the performance of SpectralEarth-L, a model specifically designed for EnMAP and trained with self-supervision on the evaluation data itself. These results highlight that, beyond its versatility, our model remains highly competitive in specialized regimes such as hyperspectral analysis.

Feature Maps. We visualize feature maps from different models in Fig. 8 using PCA projections. Thanks to its controllable output resolution, UNIVERSAT produces higher-resolution embeddings that preserve fine spatial structures (*e.g.*, field boundaries and roads), compared to fixed-resolution models. As observed by Ye et al. [63], several models exhibit patterns consistent with *positional collapse* [64], where dominant PCA components are largely driven by positional encodings.

4.3 Ablation Study

We assess in Tab. 5 the impact of our main design choices in a simplified setting. For all variants, we train a Tiny model (width 192, $D = 48$) on the full training corpus and evaluate it on one classification benchmark (m-Brick-Kiln) and two semantic segmentation benchmarks (Sen1Floods11 and PASTIS).

		classification			semantic segmentation			
	unseen config	m-brick-kiln	m-pv4ger	m-forest net	PASTIS-R	Sen1 Floods11	m-chesa peake	m-Neon Tree
	unseen sensor							RGB aerial
	Modality	S2	S2	L8	S1+S2	S1	aerial	
	Metric	Acc.	Acc.	Acc.	mIoU	mIoU	mIoU	mIoU
SSL	DINOv2 [5]	-	-	-	-	-	64.0	59.1
SSL	DINOv3 7B [59]	91.3	-	48.0	-	-	-	-
SSL	CROMA-L [38]	91.7	94.5	-	44.4	78.8	-	-
SSL	CopernicusFM [60]	85.9	-	-	-	77.6	32.2	-
SSL	AnySat [7]	84.5	90.3	34.0	46.2	77.8	61.7	49.9
SSL	DOFA [16]	-	95.8	-	13.4	77.4	59.2	55.4
SSL	Panopticon [13]	92.9	96.7	52.3	-	-	60.8	50.4
SSL	Satlas [54]	83.0	-	36.9	28.0	72.9	-	-
SSL	**UniverSat-B (ours)**	94.5	92.7	41.7	47.9	80.1	64.7	51.0
SSup	Galileo-B [6]	91.1	93.1	-	39.2	79.4	-	-
SSup	TerraMind [9]	91.9	-	-	-	78.7	-	-
SSup	OlmoEarth-L [19]	93.4	-	41.6	51.8	79.8	-	-

Table 2: **Probing Experiment.** We evaluate our model on several classification and segmentation datasets using kNN for classification and linear probing for segmentation. SSL designates methods trained with only sensor observation, and SSup models that use labels as supervision or training modalities. Results of competing methods are taken from the literature [6, 13, 19]. When values conflict, we take the highest. Top 3 performance are highlighted. 📅: single time stamp.

Table 3: **Probing with Decoders.** We evaluate our model on datasets from the Pangaea Benchmark using a linear probe, whereas other models use a heavyweight UperNet decoder. ViT-B *sup.* is a fully supervised baseline trained end-to-end [21]. 📅: single time stamp. unseen config., unseen sensor.

Modality	Probe param	PASTIS-R S1+S2	BurnScar HLS📅	DynEN PlanetScope	AI4Farms S2📅RGBNiR
ViT-B *sup.* [21]	125M	38.5	81.6	36.8	38.4
Satlas [54]	33M	17.5	80.0	36.3	25.1
CROMA [38]	47M	32.3	82.4	38.3	25.7
DOFA + [16]	47M	40.3	86.5	41.1	29.9
Ramen [17]	40M	42.3	85.0	39.9	38.8
TerraMind [9]	47M	43.1	82.9	37.9	27.5
UniverSat-B	9K	47.9	81.5	37.1	41.1

A **UPE.** We replace the UPE with $|\mathbf{M}|$ modality-specific MLP projectors (*UPE↦MLPs*), as in a standard ViT. This ties the model to sensors seen during training, prevents arbitrary patch sizes, and reduces the number of tokens processed by each encoder. It leads to a marked performance drop, while increasing the parameter count by 58% and removing the ability to handle unseen sensors. However, this variant is more efficient, with a $2\times$ reduction in training time. Replacing the LFF atom embedding with a linear scalar embedding (*LFF↦Linear*) degrades performance, supporting a nonlinear embedding of atoms. Reversing the aggregation order to collapse time before channels (*Time before channels*) mostly affects PASTIS, which has long time series. Finally, training only on the Sentinel-2 part of the corpus (*S2 only training*) performs worse, even on the Sentinel-2-only Brick-Kiln: multimodal training causes no negative transfer and instead provides complementary supervision.

B **UniverSat architecture.** We remove the skip connection that enables cross-attention to sub-patch embeddings (*No skip connection*). This degrades performance on all three benchmarks, most strongly on Brick-Kiln (-3.9), showing that the skip connection helps preserve the fine-grained spatial features lost during patch-level aggregation.
We then disable resolution control by keeping the same patch size throughout the network (*Fixed Output Res.*). This degrades performance on all three benchmarks, most notably on PASTIS (-3.7) and Brick-Kiln (-2.3), showing that decoupling the output resolution from the input patch size benefits both segmentation and classification. Finally, we replace ACA-based modality fusion with either late fusion (*Late fusion*), where each modality is processed independently and the resulting embeddings are averaged, or mid-fusion (*Average Pooling*), where ACA_M is replaced by average pooling over modalities right after the UPE. Both cause a small performance drop, showing that ACA is an effective fusion mechanism. It adds little cost, as it operates on patch embeddings rather than atoms, whereas late fusion requires one pass per modality.

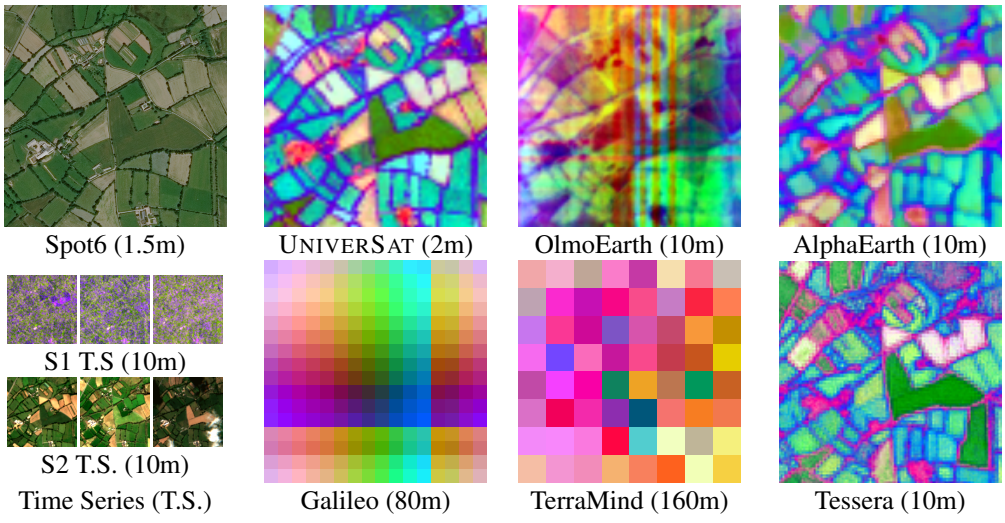
C **Training objective.** We remove the cross-modal contrastive loss applied to patch embeddings directly after the UPE ($\lambda_{\text{con}} = 0$). This substantially decreases performance, especially for multimodal segmentation. We hypothesize that this auxiliary supervision stabilizes UPE training and encourages stronger alignment of sub-patch representations across modalities. We also evaluate variants of our LM_3. Replacing the contrastive InfoNCE loss with a simple L2 loss in the frozen projection space results in a significant decrease in performance. Replacing LM_3 with a traditional MAE objective in pixel space also incurs a significant, but lesser, loss of performance. LM_3 is thus both easier to implement for massively multimodal training and

unseen sensor all datasets use EnMAP	Free param.	Trained w. EnMAP	Segmentation (mIoU) Cdl	Nlcd	Euro crops	Tree map	BD Foret	BnetD	Class. (F1) Corine
SpectralEarth-L [22]	1.2M	✔	72.7	42.7	62.1	42.6	71.1	44.0	74.0
SpatSIGMA-B [56]	900K		63.5	36.0	50.7	37.0	58.2	38.6	68.2
DOFA-L [16]	1.2M	✔	60.6	33.4	46.1	34.7	51.3	37.4	67.5
UniverSat SSL	900K		**64.7**	**39.0**	**55.3**	**38.2**	**63.3**	**43.9**	**73.6**
UniverSat SSL w. linear	9K		61.4	34.6	52.8	35.3	57.5	40.3	71.5

Table 4: **Hyperspectral performance.** We evaluate UNIVERSAT on the SpectralEarth benchmark, composed of multiple tasks using EnMAP hyperspectral data. We compare against specialized hyperspectral models (top) and general foundation models supporting hyperspectral inputs (bottom). Despite no training on EnMAP, UNIVERSAT surpasses the state-of-the-art DOFA foundation model, and remains competitive even when using only a linear probe instead of a convolutional decoder.

Table 5: **Ablation Study.** We evaluate the impact of our key design choices in a simplified setting.

		classification m-Brick-Kiln	segmentation Sen1flood11	segmentation Pastis
	Best Configuration	**92.4**	**78.4**	**36.4**
A	UPE↦ MLPs	91.9	77.6	21.5
A	LFF↦ Linear atom emb.	88.8	77.8	34.9
A	Time before channels	91.2	76.8	25.2
A	S2 only training	87.0	-	29.3
B	No skip connection	88.5	76.9	34.0
B	Fixed Output Res.	90.1	77.5	32.7
B	Late modality fusion	-	-	32.6
B	Mid-fusion: Average Pooling	92.1	76.8	32.7
C	No cross-modal contrast	91.3	76.8	27.9
C	InfoNCE↦ L2	69.8	69.7	10.7
C	Pixel MAE	77.5	74.4	12.7



(a) Input (b) Multimodal Models (c) Embedding Fields

Figure 8: **Embedding Visualization.** Embeddings of a PASTIS multimodal test tile (1.6 km^2) are projected using PCA, with the top three components mapped to RGB; colors are lightly harmonized across images for visualization.

more effective. Removing $\mathcal{L}_{\text{LM}^3}$ is not a meaningful ablation, as only the UPE would then be supervised.

Limitations and Impact. UNIVERSAT trades specialization for generality: in standard settings (*e.g.*, VHR RGB or mono-temporal Sentinel-2), modality-specific models may be more accurate or efficient. Its computational overhead is moderate, and it is even faster on long time series, but memory use is typically higher (Sec. A.5.1): this is the cost of supporting heterogeneous inputs with one architecture. Our evidence for unseen sensors covers EnMAP and PlanetScope, whose characteristics lie within the pretraining gamut (roughly 0.2–250 m GSD, up to 140 observations per year, and 1–500 channels); we do not claim generalization to arbitrary sensors far outside it, such as UAV imagery or very dense temporal streams. Axial aggregation assumes co-registered inputs: under strong misalignment across timestamps or modalities, the UPE may blur fine details, which standard co-registration can mitigate. It is also less seamless for non-optical sensors, which require learning a small modality encoding vector alongside the probe. Beyond technical aspects, and as with any EO model, UNIVERSAT may also enable large-scale monitoring capabilities, raising concerns around surveillance or misuse.

5 Conclusion

We introduced UNIVERSAT, a transformer architecture for multimodal Earth Observation that replaces fixed patch projectors with a Universal Patch Encoder, enabling a single model to process multimodal inputs of any resolutions. Through self-supervised training across 13 sensors and 7 heterogeneous datasets, UNIVERSAT learns modality- and resolution-agnostic representations that transfer effectively to a wide range of downstream EO tasks. Our results demonstrate that a single, shared backbone remains competitive with modality-specific foundation models, and surpasses them in several settings, while offering far greater input flexibility.

Acknowledgments and Disclosure of Funding

This work was supported by Hi! PARIS (ANR-23-IACL-0005), the ANR project SHARP (ANR-23-PEIA-0008) in the context of the PEPR IA, the French Space Agency PhD program (CNES), and the French National Institute of Geographic and Forest Information (IGN). This work was granted access to the HPC resources of IDRIS under the allocations AD011014719R2, AD011014781R2, and A0200116178 made by GENCI. We thank Jordi Inglada for inspiring discussions and valuable feedback.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 1

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[5] 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. *TMLR*, 2024. 1, 8

[6] 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. In *ICML*, 2025. 1, 2, 3, 7, 8, 19, 21

[7] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *CVPR*, 2025. 1, 2, 3, 6, 8, 19, 21

[8] 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. In *NeurIPS*, 2022. 1, 2, 3, 21

[9] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for Earth observation. *arXiv:2504.11171*, 2025. 1, 2, 3, 8, 9, 19, 21

[10] 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:2507.22291*, 2025. 1, 3, 21

[11] 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 *ICCV*, 2023. 2, 3, 5, 16, 21

[12] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for Earth observation. In *ECCV*, 2024. 2, 3, 6, 7, 21

[13] 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. *arXiv:2503.10845*, 2025. 2, 3, 8, 19, 21

[14] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. *arXiv:2408.11000*, 2024. 2, 3

[15] Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *ICCV*, 2025. 2, 3, 21

[16] 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:2403.15356*, 2024. 2, 3, 8, 9, 21

[17] Nicolas Houdré, Diego Marcos, Hugo Riffaud de Turckheim, Dino Ienco, Laurent Wendling, Camille Kurtz, and Sylvain Lobry. RAMEN: Resolution-adjustable multimodal encoder for Earth observation. *arXiv:2512.05025*, 2025. 2, 3, 7, 9, 21

[18] Kun Yi, Yixiao Ge, Xiaotong Li, Shusheng Yang, Dian Li, Jianping Wu, Ying Shan, and Xiaohu Qie. Masked image modeling with denoising contrast. In *ICLR*, 2023. 2

[19] Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. OlmoEarth: Stable latent image modeling for multimodal Earth observation. In *CVPR*, 2026. 2, 3, 6, 8, 17, 19

[20] 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. In *NeurIPS*, 2023. 2, 7

[21] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv:2412.04204*, 2024. 2, 7, 9

[22] 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. 2, 7, 8, 9

[23] Diego Velazquez, Pau Rodriguez, 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 *WACV Workshops*, 2025. 2, 7, 21

[24] Nikolaos Ioannis Bountos, Arthur Ouaknine, Ioannis Papoutsis, and David Rolnick. FoMo: Multi-modal, multi-scale and multi-task remote sensing foundation models for forest monitoring. In *AAAI*, 2025. 2, 3, 21

[25] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *ICCV*, 2021. 2, 3

[26] Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, 2021. 2

[27] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, 2023. 2

[28] Wei-Hsin Tseng, Hoàng-Ân Lê, Alexandre Boulch, Sébastien Lefèvre, and Dirk Tiede. CROCO: Cross-modal contrastive learning for localization of Earth observation data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2022. 2, 3, 6

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

[30] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *CVPR*, 2024. 2

[31] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *ICCV*, 2023. 2

[32] Iris Dumeur, Silvia Valero, and Jordi Inglada. Self-supervised spatio-temporal representation learning of satellite image time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2

[33] Gabriel Tseng, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv:2304.14065*, 2023. 3, 21

[34] 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. *arXiv:2506.06281*, 2025. 3, 21

[35] Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *CVPR*, 2025. 3

[36] Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. MAESTRO: Masked autoencoders for multimodal, multitemporal, and multispectral earth observation data. In *WACV*, 2026. 3, 21

[37] Yi Wang, Conrad Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiaoxiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 2022. 3

[38] Anthony Fuller, Koreen Millard, and James R Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *NeurIPS*, 2023. 3, 8, 9, 19, 21

[39] Guillaume Astruc, Eduard Trulls, Jan Hosang, Loic Landrieu, and Paul-Edouard Sarlin. UniGeoCLIP: Unified geospatial contrastive learning. In *CVPR Workshop EarthVision*, 2026. 3

[40] 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 *ICCV*, 2025. 3, 21

[41] Ibrahim Fayad, Max Zimmer, Martin Schwartz, Fabian Gieseke, Philippe Ciais, Gabriel Belouze, Sarah Brood, Aurélien de Truchis, and Alexandre d'Aspremont. DUNIA: Pixel-sized embeddings via cross-modal alignment for earth observation applications. In *ICML*, 2025. 3, 21

[42] Jeremy Irvin, Lucas Tao, Joanne Zhou, Yuntao Ma, Langston Nashold, Benjamin Liu, and Andrew Y Ng. USat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv:2312.02199*, 2023. 3

[43] Manuel Weber and Carly Beneke. PyViT-FUSE: A foundation model for multi-sensor earth observation data. *arXiv:2504.18770*, 2025. 3

[44] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. FlexiViT: One model for all patch sizes. In *CVPR*, 2023. 3, 17

[45] Hugo Riffaud de Turckheim, Sylvain Lobry, Roberto Interdonato, and Diego Marcos. Atomizer: Generalizing to new modalities by breaking satellite images down to a set of scalars. In *BMVC*, 2025. 3, 4, 8, 21

[46] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv:1912.12180*, 2019. 3, 4, 16

[47] Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, et al. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *NeurIPS*, 2025. 5

[48] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 5

[49] Yibing Wei, Abhinav Gupta, and Pedro Morgado. Towards latent masked image modeling for self-supervised visual representation learning. In *ECCV*, 2024. 6, 17

[50] Anatol Garioud, Sébastien Giordano, Nicolas David, and Nicolas Gonthier. FLAIR-HUB : Large-scale multimodal dataset for land cover and crop mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2026. 7

[51] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, 2021. 7

[52] Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Förster, Florencia Arias, Jörn Hees, Begüm Demir, and Birgit Kleinschmit. TreeSatAI Benchmark Archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data*, 15:681–695, 2023. 7

[53] Luis Miguel Pazos-Outón, Cristina Nader Vasconcelos, Anton Raichuk, Anurag Arnab, Dan Morris, and Maxim Neumann. Planted: A dataset for planted forest identification from multi-satellite time series. In *IGARSS*, 2024. 7

[54] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, 2023. 7, 8, 9

[55] Piper Wolters, Favyen Bastani, and Aniruddha Kembhavi. Zooming out on zooming in: Advancing super-resolution for remote sensing. *arXiv:2311.18082*, 2023. 7

[56] Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *TPAMI*, 47(8):6427–6444, 2025. 7, 8, 9

[57] Jihyeon Lee, Nina R Brooks, Fahim Tajwar, Marshall Burke, Stefano Ermon, David B Lobell, Debashish Biswas, and Stephen P Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 2021. 7

[58] 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 *CVPR Workshop EarthVision*, 2020. 7

[59] 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:2508.10104*, 2025. 8, 21

[60] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for Earth vision. *arXiv:2503.11849*, 2025. 8

[61] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 7

[62] Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. AI4SmallFarms: A dataset for crop field delineation in southeast asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 2023. 7

[63] Dingqi Ye, Daniel Kiv, Wei Hu, Jimeng Shi, and Shaowen Wang. Any model, any place, any time: Get remote sensing foundation model embeddings on demand. *arXiv:2602.23678*, 2026. 8

[64] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *arXiv:2502.08769*, 2025. 8

[65] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable Fourier features for multi-dimensional spatial positional encoding. *NeurIPS*, 2021. 15

[66] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024. 15

[67] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, 2024. 15, 16

[68] Johannes Jakubik, S Roy, CE Phillips, P Fraccaro, D Godwin, B Zadrozny, D Szwarcman, C Gomes, G Nyirjesy, B Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arxiv:2310.18660*. 19

[69] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Thorsteinn 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 multi-temporal foundation model for earth observation applications. *arXiv:2412.02732*, 2024. 19, 21

UniverSat: Resolution- and Modality-Agnostic Transformers for Earth Observation

In this appendix, we provide details on compute cost and licensing (Sec. A.1), additional methodological insights into the Universal Patch Encoder (UPE) (Sec. A.2), and reproducibility details (Sec. A.3). We then report additional fine-tuning experiments (Sec. A.4), discuss computational cost (Sec. A.5.1) and information compression (Sec. A.5.2), and provide an exhaustive description of the training datasets with an extended comparison to competing methods (Sec. A.6).

A.1 Compute Cost and License

License Code is released under the MIT License at https://github.com/gastruc/UniverSat, allowing unrestricted use, modification, and distribution. Pretrained weights are available on the Hugging Face Hub at https://huggingface.co/g-astruc/UniverSat. All datasets used in this work are publicly available; we adhere to their respective licenses and cite them appropriately in the paper and supplementary material.

Compute Resources All experiments were conducted on H100 GPUs. Training UniverSat in the self-supervised setting requires approximately 240 GPU-h. Linear probing experiments are lightweight, ranging from 1 to 10 GPU-h per dataset, for a total of 60 GPU-h. The full project, including ablations and preliminary experiments, required approximately 30K GPU-h. Using standard estimates of carbon intensity computed with CodeCarbon[1], this corresponds to an approximate carbon footprint of 31 tons CO_2.

A.2 Additional Methodological Details

We provide exact derivations of the Learnable Fourier Features, Metadata Encoding, and Axial Cross-Attention modules.

A.2.1 Learnable Fourier Features (LFF)

We use Fourier features followed by a learned projection, in the spirit of Learnable Fourier Features (LFF) [65], to lift scalar values (*e.g.*, radiance or wavelength) to a D-dimensional embedding $\mathrm{LFF} : \mathbb{R} \to \mathbb{R}^D$. For a scalar $u \in \mathbb{R}$,

$$\mathrm{LFF}(u) = W \left[\sqrt{2} \cos(\omega_k u + \phi_k) \right]_{k=1}^{D/4}, \tag{9}$$

where the frequencies $\omega_k \sim \mathcal{N}(0, (2\pi\beta)^2)$ and phases $\phi_k \sim \mathcal{U}(0, 2\pi)$ are sampled at initialization and kept frozen, and $W \in \mathbb{R}^{D \times D/4}$ is a learned linear projection. Following EDM2 [66], the $\sqrt{2}$ factor yields unit-variance features. We use a bandwidth $\beta = 2.5$ for atom values and $\beta = 1$ for wavelengths. When applied to a tensor, LFF modules are applied element-wise.

A.2.2 Metadata Encoding

We consider a tensor $t \in \mathbb{R}^{X \times Y \times D}$, where X is the axis to collapse, D the feature dimension, and Y aggregates all remaining dimensions. Patch x is associated with its ground sampling distance (GSD) ρ, its acquisition times $\tau = \{\tau_t\}_{t=1}^{T}$, and per-channel descriptors $b = \{b_c\}_{c=1}^{C}$: wavelength for optical data, categorical identifier otherwise. We add to t some axis-specific metadata encodings, depending on the dimension X being collapsed:

- **Channel (C):** optical channels receive a dedicated LFF embedding of their wavelength (in µm). Non-optical channels (*e.g.*, radar polarizations or elevation products) cannot be described by a single scalar and are instead mapped to learned lookup embeddings.
- **Time (T):** we apply 1D RoPE [67] to the day of the year of each acquisition. For multi-year series, the position is the day of the year plus 365 times the number of years elapsed since the start of the sequence, which emphasizes seasonality while still distinguishing years.

[1] https://codecarbon.io/

- **Space (I and S):** we apply 2D RoPE relative positional encodings [67] at both spatial aggregation stages (pixels within sub-patches, then sub-patches within patches), with positions scaled by the GSD ρ following Scale-MAE [11].

Additive metadata embeddings are added to t before computing keys and values, while RoPE is applied as a rotary transformation to the projected query and key vectors. This makes the ACA module both modality-aware and geometry-aware.

A.2.3 Axial Cross-Attention (ACA)

Directly projecting e with an MLP is impractical as $C \times T \times I \times S \times D$ varies significantly across sensors. Using self-attention would be too expensive as the channel, pixel, and temporal dimensions can each reach the hundreds. Instead, we collapse one axis at a time with axial cross-attention [46], see Fig. 3.

We consider an Axial Cross-Attention module ACA_X^α where X is the dimension to collapse, and α is the feature expansion factor. Let t be a tensor of dimension $X \times Y \times D$ with X the target dimension, D the feature dimension, and Y representing all remaining dimensions. For example, to collapse dimension I of a tensor of dimension $C \times T \times I \times S \times D$, we would have $X = I$ and $Y = C \times T \times S$.

We first compute keys, queries, and values with RMSNorm (N) as follows:

$$K, V = \psi_{\text{k}}(\text{N}(t)), \psi_{\text{v}}(\text{N}(t)) \qquad \text{dim: } X \times Y \times \alpha D \tag{10}$$

$$Q = \psi_{\text{q}}\big(\big[\text{mean}_X(\text{N}(t)),\ \max_X(\text{N}(t))\big]\big) \qquad \text{dim: } Y \times \alpha D\ , \tag{11}$$

where ψ_{k}, ψ_{v} are linear layers mapping D to the expanded feature axis αD, mean_X and $\max_X$ pool along the X axis, $[\cdot, \cdot]$ denotes concatenation, and $\psi_{\text{q}} : 2D \mapsto \alpha D$. Queries and keys are further normalized per head (QK-norm). We perform gated cross-attention along axis X only, then apply a residual feed-forward network FF with RMSNorm:

$$\text{ACA}_X^\alpha(t) = \text{FF}\left(\text{pool}_X^\alpha(t) + \gamma_{CA}\, \text{softmax}_X\left(\frac{Q \otimes_D K}{\sqrt{\alpha D}}\right) \otimes_X V\right)\ , \tag{12}$$

where softmax_X is on axis X, pool_X^α is a max pooling along X for $\alpha = 1$ and the concatenation of mean and max pooling for $\alpha = 2$, and γ_{CA} is a LayerScale parameter initialized at 10^{-6} (the residual branch of FF uses the same scheme). We denote by $\otimes_D$ the vector product in axis D with broadcasting on dimension $X \times Y$, and $\otimes_X$ the product in axis X with broadcasting on dimension $Y \times D$:

$$[Q \otimes_D K]_{x,y} = \sum_d Q_{y,d} K_{x,y,d} \qquad [A \otimes_X V]_{y,d} = \sum_x A_{x,y} V_{x,y,d}\ . \tag{13}$$

Since each query attends only along axis X, the cost is *linear* in the number of atomic tokens in each patch.

A.2.4 Scale Augmentation and Masking

We consider a multimodal tile $x = \{x_p^m\}_{p \in \mathbf{P},\, m \in \mathbf{M}}$ composed of patches $\mathbf{P}$ observed in modalities $\mathbf{M}$. We randomly sample both the input patch size and the target output resolution from dataset-specific presets, encouraging robustness to spatial scale. We then apply a compositional masking operator:

- **Modality dropping.** Randomly drop modalities for the entire tile with probability 30%, ensuring that at least one remains. The retained modalities are denoted $\mathbf{M}^\star$.
- **Time dropping.** For time-series modalities, randomly drop 50% of timestamps and, with probability 10%, retain a single timestamp to emulate mono-date observations.
- **Channel dropping.** For each retained modality with N channels, drop a fraction $\psi(N) = 1/(1 + \sqrt{5/N})$, so that spectrally dense sensors are thinned more aggressively. Dropping is applied in two contiguous channel blocks.
- **Patch dropping.** Drop a random subset $\mathbf{D} \subset \mathbf{P}$ of 75% of the patches, and denote the visible patches by $\mathbf{P}^\star = \mathbf{P} \backslash \mathbf{D}$.

Overall, approximately 90% of input atoms are removed.

A.2.5 Linear Multimodal Masked Modeling (LM_3)

We adapt the LMIM-Lite objective [49, 19] to a multimodal and multitemporal setting. The goal is to predict representations of masked patches in a fixed latent space defined by frozen random projections.

Random Projection Targets For each modality m, we define a frozen linear projection applied to single-time-step slices:

$$\phi_m^{\text{rand}} : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^D. \tag{14}$$

For optical modalities, varying patch sizes are handled via kernel interpolation, following FlexiViT [44]. For SAR data, where spatial resampling is less straightforward, we use separate projections for each patch size. The projections ϕ_m^{rand} are randomly initialized at the beginning of training and kept frozen thereafter.

Temporal Sampling A key challenge is to balance spatial and temporal training signals. Using all timestamps may lead the model to exploit trivial temporal cues (*e.g.*, acquisition dates) instead of spatial content, while ignoring temporal variation degrades representation quality. To address this, we select a small set of $K = 4$ timestamps per tile, sampled at random among those with a cloud density below 30% when cloud masks are available. Each masked patch $p \in \mathbf{D}$ is assigned a target timestamp $\tau(p)$ from this set.

Prediction We define a predictor network ϕ^{pred}, composed of a stack of self-attention blocks. It takes as input the visible patch embeddings $\{z_q\}_{q\in\mathbf{P}^\star}$ together with learned mask tokens for masked patches $p \in \mathbf{D}$. We apply spatial RoPE positional encodings to all tokens, and inject absolute temporal encodings into masked tokens to represent their assigned timestamps $\tau(p)$. The predictor produces representations for masked patches at their target times:

$$\{z_p^{\tau(p)}\}_{p\in\mathbf{D}} = \phi^{\text{pred}}\big(\{z_q\}_{q\in\mathbf{P}^\star}\big). \tag{15}$$

Modality-specific linear heads $\{\phi_m^{\text{head}}\}_{m\in\mathbf{M}}$ map these embeddings to the target space:

$$h_p^{m,\tau(p)} = \phi_m^{\text{head}}\big(z_p^{\tau(p)}\big). \tag{16}$$

Loss We define a contrastive loss between predicted embeddings and random projection targets of the masked inputs at the corresponding timestamps, using temperature τ_{LM^3}:

$$\mathcal{L}_{\text{LM}^3} = -\sum_{m\in\mathbf{M}}\sum_{p\in\mathbf{D}} \log \frac{\exp\left(\frac{1}{\tau_{\text{LM}^3}}\left\langle h_p^{m,\tau(p)}, \phi_m^{\text{rand}}\big(x_p^m[\tau(p)]\big)\right\rangle\right)}{\sum_{q\in\mathbf{D}}\exp\left(\frac{1}{\tau_{\text{LM}^3}}\left\langle h_p^{m,\tau(p)}, \phi_m^{\text{rand}}\big(x_q^m[\tau(q)]\big)\right\rangle\right)}. \tag{17}$$

This objective combines masked prediction in latent space with a contrastive training signal. It avoids collapse without requiring momentum teachers, while coupling information across modalities and time.

A.3 Reproducibility Details

We summarize the implementation and evaluation details required to reproduce the experiments.

A.3.1 Model Configurations

Table A.1: Model configuration details required for reproducing UNIVERSAT-Tiny and UNIVERSAT-B.

Item	UNIVERSAT-Tiny	UNIVERSAT-B
Usage	Ablation study in Tab. 5	Main benchmark tables
Atomic embedding size D	48	96
UPE patch embedding	192	768
UPE expansion α (I, C, T, S)	1, 2, 2, 2	2, 2, 2, 2
UPE sequence	Pixel, Channel, Time, Sub-patch	
Spatial Transformer blocks	SA×6	SA×12
Attention heads: UPE + encoder	8	12
Attention heads: predictor	8	12
FFN expansion ratio	4	
Normalization	Pre-norm RMSNorm, QK-norm	
Gating (MLP and Attention)	True	
QKV bias (ACA, CA and SA)	False	
Dropout	dropout 0.0; attention dropout 0.0	
Stochastic depth	stochastic depth 0.0	
Register tokens	4	
UPE register tokens	2 (time axis), 4 (sub-patch axis)	
Predictor depth	depth 8	
Predictor time subset size	4	
Predictor register	4	
Predictor dimensions	192	768
Random targets	Frozen linear projection; FlexiViT kernel at 32×32 or lower (SAR: one per patch size)	
Prediction heads ϕ_m^{head}	Linear, one per modality	

A.3.2 Self-Supervised Pretraining

Table A.2: Self-supervised pretraining hyperparameters and optimization details.

Item	Value / description
Supervision	Self-supervised only; no labels are used during pretraining.
Time	Full pretraining requires approximately 240 GPU-h.
GPUs	2 nodes × 4 GPUs per node = 8 GPUs
GPU memory	H100 80GB
Precision	bf16 mixed precision
Optimizer type	AdamW
Optimizer LR	5×10^{-4}
Optimizer weight decay	10^{-3}
Optimizer epsilon and betas	$\epsilon = 10^{-5}$, betas use PyTorch AdamW defaults
LR schedule	10 warmup epochs, then cosine annealing to 10^{-6}
Gradient accumulation	2
Training length	100k steps
Final loss	$\mathcal{L}_{\text{LM}^3} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$.
Loss weights	$\lambda_{\text{con}} = 1.0$
LM^3 temperature τ_{LM^3}	0.1
Cross-modal contrast temperature	0.2

Table A.3: Modality-level preprocessing configuration: native resolution, number of input bands or descriptors, sub-patch factor, and datasets using each modality.

Modality	Resolution (m)	Bands	Sub-patch size (px)	Datasets
spotRGBN	1.6	4	2	FLAIR-Hub.
aerialflair	0.2	4	8	FLAIR-Hub.
s2flair	12.8	10	1	FLAIR-Hub.
s1flair	12.8	3	1	FLAIR-Hub.
dem	0.2	2	10	FLAIR-Hub.
spot	1	3	10	PASTIS-HD.
s2	10	10	1	PASTIS-HD, Planted, TSAI-TS, S2NAIP-Urban.
s1	10	3	1	PASTIS-HD, Planted, TSAI-TS, S2NAIP-Urban.
l7	30	6	1	Planted.
alos	30	3	1	Planted.
modis	250	7	1	Planted.
aerial	0.2	4	10	TSAI-TS.
naip	1.25	4	10	S2NAIP-Urban.
l8	10	11	1	S2NAIP-Urban.
EO1	30	175	1	HyperGlobal.
GF5	30	150	1	HyperGlobal.
enmap	30	202	1	SpectralEarth (evaluation only).
rgbneon	0.1	3	20	EarthView.
ndemneon	1	1	2	EarthView.
neon	1	369	2	EarthView.

Table A.4: Dataset-level spatial configuration: image size from the config, input and target output scales, and modalities used for each dataset.

Dataset	Image size ($\times$10 m)	Input scales ($\times$10 m)	Output scales ($\times$10 m)	Dataset weight	Modalities
FLAIR-Hub	10.24	1.28, 2.56	0.64, 1.28	2	spotRGBN, aerialflair, s2flair, s1flair, dem.
PASTIS-HD	128	4, 8, 16	2, 4	2	spot, s2, s1.
Planted	12	3	3	1	s2, s1, l7, alos, modis.
TSAI-TS	6	2, 1	1	1	aerial, s2, s1.
S2NAIP-Urban	64	16, 4, 8	2, 4	4	naip, l8, s2, s1.
HyperGlobal	192	12, 6	6, 3	2	EO1, GF5.
EarthView	6.4	0.8, 1.6	0.4, 0.8	4	rgbneon, ndemneon, neon.

A.4 Additional Experiments

A.4.1 End-to-End Fine-Tuning

Our main protocol uses frozen-backbone probing, following Galileo [6] and OlmoEarth [19]. To verify that the UPE does not cause optimization instability, we also fine-tune UNIVERSAT-B end-to-end on Sen1Floods11 and BurnScars. The model trains stably and reaches strong performance, setting a new state of the art on Sen1Floods11 and approaching the best models on BurnScars (Tab. A.5).

Table A.5: **End-to-end fine-tuning.** mIoU of fully fine-tuned models.

Model	Sen1Floods11	BurnScars
AnySat [7]	77.4	**90.6**
RemoteCLIP	–	65.1
Prithvi [68]	–	86.9
Prithvi v2 [69]	–	90.5
Panopticon [13]	79.1	–
CROMA [38]	79.4	–
TerraMind [9]	79.5	–
OlmoEarth [19]	79.8	–
UNIVERSAT-B, linear probing	80.1	81.5
UNIVERSAT-B, fine-tuning	**80.2**	89.6

Data	Timesteps	Model	Inference (samples/s)		Training (samples/s)		Memory (GB VRAM)	
S2	40	MultiMLP baseline	2.1	-	6.9	-	4.6	-
		UniverSat	2.0	×0.95	2.7	×0.39	12.2	×2.7
S1	140	MultiMLP baseline	0.21	-	1.8	-	11.8	-
		UniverSat	1.7	×8.1	1.8	×1.0	12.7	×1.1
Aerial 20cm	-	MultiMLP baseline	130	-	26	-	1.8	-
		UniverSat	79	×0.61	14	×0.54	3.7	×2.1

Table A.6: **Computational cost.** Throughput and memory of UNIVERSAT and of a ViT-style baseline with one linear projector per modality and patch size (MultiMLP), measured on an A100 GPU. Ratios are relative to the baseline.

A.5 Limitations

A.5.1 Computational Efficiency

To quantify UniverSat's overhead, we compare it to a minimal ViT-style baseline in which the UPE is replaced by independent linear projectors for each modality, patch size, and resolution, resulting in a token for each patch, modality, and timestep. This baseline is particularly computationally efficient but also highly impractical: it requires a separate projector for every input modality and patch size (42 projectors), cannot be applied to unseen configurations, and performs substantially worse, *e.g.*, by approximately 15 points on PASTIS (*UPE↦→MLPs* in Tab. 5). This suggests that the cheaper interface fails to learn shared geospatial representations across heterogeneous inputs.

We now report throughput and memory for representative input types on A100, using images of 128×128 for S1/S2 time series with patch size 4 and images of 512×512 with patch size 64 for aerial photography.

UniverSat is slower and more memory-intensive for short time series and high-resolution aerial imagery, but its overhead remains moderate relative to the added flexibility. Memory usage is generally higher, but can be controlled through the input patch size and target output resolution.

These results clarify the intended trade-off. UniverSat is not designed to be cheaper than a specialized encoder optimized for one fixed input format. Instead, it provides a single native-resolution interface that avoids resampling, supports unseen configurations, and yields stronger representations across heterogeneous EO inputs at a measurable computational cost.

A.5.2 UPE Compression of Information

Due to its flexibility, the UPE maps heterogeneous patches to fixed-width representations which can lead to discarding some raw information. It is a representation encoder, not an invertible reconstruction model. However, it does not collapse the complete patch into a single vector: it outputs both a global patch embedding and a set of sub-patch embeddings used by the high-resolution skip connection.

The effective compression rate depends on the physical patch extent and on the spatial, spectral, and temporal density of the sensor. Across our input configurations, it ranges from an expansion of 0.02× for sparse ALOS inputs to a compression of 369× for the dense NEON hyperspectral time series, with a geometric average of 1.56× when removing NEON. Importantly, the patch extent is selected at runtime rather than fixed by the architecture, allowing the user to control this trade-off.

Such compression is not unique to UPE: a standard ViT also maps each input patch to a fixed-width token. The main difference is that UPE supports heterogeneous input shapes with shared weights and preserves sub-patch embeddings, which mitigate spatial information loss and retain fine-scale features for dense prediction.

The largest compression rates occur for very-high-resolution and hyperspectral inputs, where adjacent pixels or spectral bands are often strongly redundant. Our strong frozen-feature performance on hyperspectral benchmarks suggests that the UPE preserves substantial task-relevant information even in this regime, although we do not claim that it retains every raw measurement.

A.6 Data and Models

We provide here an extended comparison with recent models (Tab. A.7) and the datasets used to train UNIVERSAT (Tab. A.8).

	training modalities						supports unseen resolution			feature resolution
	📷	🎥	📡	⛰	📊	✒	spat.	temp.	spec.	
DINOv3 [59]	1									■
SatMAE [8]		1								■
CROMA [38]	1		1							■
DOFA [16]	3		1		1		✓		✓	■
Atomizer [45]	1								✓	■
Prithvi v2 [69]		1								■
DUNIA [41]		1		1					✓	■
MAESTRO [36]	1	1	1					✓		■
SkySense v2 [40]	1	1	1					✓		■
PRESTO [33]		1	1	1		**6**		✓		▪
EarthView [23]	3		1	1				✓		■
Galileo [6]		1	1	1		**6**	✓	✓		■
Smarties [15]	1		1						✓	■
TerraMind [9]	1		1	1	1		✓			■
Scale-MAE [11]	1						✓			■
TerraFM [34]	1		1		2					■
Panopticon [13]	**5**		1		1				✓	■
AlphaEarth [10]		3	1	2		5		✓		▪
OmniSat [12]	2	2	1	1				✓		■
FoMo [24]	4	2	1	1				✓	✓	■
Ramen [17]	1	1	1	1			✓	✓	✓	▞
AnySat [7]	2	4	**2**	1			✓	✓		▪
UniverSat	4	**4**	**2**	**2**	3		✓	✓	✓	▞

📷/🎥 : optical snapshot/ time series 📡 : radar ⛰ : elevation 📊 : hyperspectral ✒ : labels
■ : image ▣ : patch ▪ : pixel ▞ : any resolutions

Table A.7: **Multimodal EO Foundation Models.** For each model, we list the training modalities, whether unseen spatial/temporal/spectral configurations are handled *without input resampling*, and the feature-map granularity. UNIVERSAT supports the broadest modality mix, handles unseen configurations, and offers flexible output resolution—all with a single set of weights.

Dataset	Labels	Extent	Modalities	Resolution			# atoms
				Spat. m	Temp. img/yr	Spect. chan.	Gå
TSAI-TS	Tree species classification multilabel	50K 60×60 m	Aerial VHR	0.20	1	4	18
			S1	10	10-70	3	0.2
			S2	10	10-70	10	0.7
PASTIS-HD	Crop type segmentation dense	2433 1280×1280 m	SPOT6/7	1m†	1	4	16
			S1	10	140	3	17
			S2	10	38-61	10	18
FLAIR-Hub	Land cover segmentation	241K 102.4×102.4 m	Aerial VHR	0.2	1	4	253
			DSM+nDSM	0.2	1	1	63
			SPOT6	1.5	1	4	3.9
			S1	10	0-122	2	9.1
			S2	10	20-146	10	21
Planted	Tree species classification	1.3M 120×120 m	S2	10	8	10	15
			S1	10	8	3	4.5
			Landsat 7	30	20	3	1.2
			ALOS-2	30	4	3	0.3
			MODIS	250	60	7	0.6
S2NAIP-Urban	Land cover segmentation dense	515K 640×640 m	NAIP	1.25	1	4	540
			S2	10	16-32	10	506
			S1	10	2-8	3	38
			Landsat 8/9	10m†	4	8	68
HyperGlobal		450K 1920×1920 m	Gaofen-5	30	1	150	151
			EO 1	30	1	175	143
EarthView		35K 64 ×64 m	UHR aerial RGB	0.1	1	3	44
			NIS	1	3	396	173
			LiDAR	1	3	1	0.4

Table A.8: **Pretraining Datasets Overview.** Summary of the seven heterogeneous datasets used to train UNIVERSAT. The table details the semantic labels, spatial extent, and sensor modalities for each dataset, along with their respective spatial (m), temporal (images/year), and spectral (channels) resolutions, and the total volume of atomic tokens (Gå).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We make two main claims. First, we make the ViT architecture more flexible by replacing its patch projector with a more flexible Universal Patch Encoder (UPE). The increased flexibility is validated by our ability to train a single model on heterogeneous EO data across modalities (radar, optical, elevation models, hyperspectral), spatial resolutions (0.1–250 m GSD), spectral dimensions (1–400 channels), and temporal depths (1–150+ timestamps). Second, we claim that this flexibility enables training a single model on multiple datasets and sensors, yielding versatile representations. This is supported by strong performance on standard benchmarks (e.g., Table 2) and generalization to unseen sensors, where the model remains competitive with specialized approaches (e.g., Table 4) .

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss limitations in the main text. In particular, we note that the generality of UNIVERSAT introduces efficiency and performance trade-offs compared to modality-specific architectures in homogeneous settings (e.g., VHR RGB or mono-temporal optical data). We also clarify that the model is best suited for heterogeneous multimodal scenarios, where its flexibility outweighs these costs. More broadly, we discuss how performance depends on data diversity and input configuration, and outline the scope in which our claims hold.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical results or formal guarantees; all contributions are architectural and empirical.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed description of the architecture, training pipeline, and evaluation protocols in the main paper and supplementary material, including dataset composition, preprocessing, and key hyperparameters. We also specify the linear probing setup used for all evaluations. Code and pretrained models are publicly released at https://github.com/gastruc/UniverSat and https://huggingface.co/g-astruc/UniverSat.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide detailed implementation instructions, dataset descriptions, and training hyperparameters in the supplementary material. Code, configuration files, and pretrained models are publicly released at https://github.com/gastruc/UniverSat and https://huggingface.co/g-astruc/UniverSat, enabling full reproduction of our experimental results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper describes the datasets, data splits, model architecture, self-supervised pretraining objective, and evaluation protocols in the main text. Detailed training procedures, hyperparameters, and optimization settings are provided in the supplementary material, ensuring a complete understanding of the experimental setup.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars. Training the model is computationally intensive and repeated runs are impractical. Evaluation is performed via linear probing with fixed protocols, which introduces negligible stochasticity and yields stable results across runs.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the type of hardware used (H100 GPUs), training duration, and compute requirements for both pretraining and probing experiments. We also provide an estimate of the total compute used for the project, including ablations, as well as an approximate carbon footprint.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. All datasets used are publicly available and widely used in the community, and no sensitive or personally identifiable information is processed.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both positive and negative societal impacts. On the positive side, the proposed approach can improve the use of foundation models in scientific domains such as Earth observation, supporting applications in environmental monitoring and scientific analysis. On the negative side, we highlight potential risks including misuse for large-scale monitoring or surveillance, as well as biases introduced through metadata and heterogeneous data sources, and discuss these considerations in the limitations section.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce datasets or models with specific high-risk characteristics requiring dedicated safeguards beyond standard practice. It relies on publicly available datasets and proposes a general-purpose architecture similar in nature to existing vision models. Potential misuse (e.g., large-scale monitoring) is inherent to EO models and is discussed in the limitations section, but does not require controlled release mechanisms specific to this work.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and external resources used in this work are publicly available and properly cited. We adhere to their respective licenses and terms of use, and provide references and access details in the paper and supplementary material.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a new model and training framework, which are fully described in the paper and supplementary material. Code, pretrained models, and documentation (including usage instructions and configuration details) are released under the MIT License at https://github.com/gastruc/UniverSat and https://huggingface.co/g-astruc/UniverSat.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or experiments with human subjects. All datasets used are publicly available and were collected by their original authors.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All datasets used are publicly available and no new data collection was performed.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for editing and polishing the manuscript, and do not contribute to the methodology, experiments, or results.