

OmniSat: Self-Supervised Modality Fusion for Earth Observation

Guillaume Astruc^{1,3,4}, Nicolas Gonthier^{1,2},
Clement Mallet¹, and Loic Landrieu^{1,3}

¹ Univ Gustave Eiffel, IGN, ENSG, LASTIG, France ² IGN, France
³ LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, France ⁴ CNES, France

Abstract. The diversity and complementarity of sensors available for Earth Observations (EO) calls for developing bespoke self-supervised multimodal learning approaches. However, current multimodal EO datasets and models typically focus on a single data type, either mono-date images or time series, which limits their impact. To address this issue, we introduce OmniSat, a novel architecture able to merge diverse EO modalities into expressive features without labels by exploiting their alignment. To demonstrate the advantages of our approach, we create two new multimodal datasets by augmenting existing ones with new modalities. As demonstrated for three downstream tasks—forestry, land cover classification, and crop mapping—OmniSat can learn rich representations without supervision, leading to state-of-the-art performances in semi- and fully supervised settings. Furthermore, our multimodal pretraining scheme improves performance even when only one modality is available for inference. The code and dataset are available at <https://github.com/gastruc/OmniSat>.

Keywords: Earth observation · Multi-modality · Self-supervised learning

1 Introduction

Self-supervised multimodal learning has recently gathered significant interest within computer vision [38, 82, 102]. Earth Observation (EO) is particularly well-suited for developing and evaluating such approaches [29, 51], thanks to the large amount of open-access data captured by sensing technologies with complementary capabilities [36, 79]. Combining different sources of EO observations is crucial for several high-impact applications, including environmental [21, 80, 83] and climate monitoring [56, 97], as well as improving food security [66]. Moreover, learning with few or no labels is essential for developing regions with limited data annotation capabilities [5, 55, 62].

Despite this potential, most multimodal EO datasets and models focus on a single data type, either mono-date images or time series. This limitation prevents them from simultaneously leveraging the spatial resolution of aerial images [57, 64], the temporal and spectral resolutions of optical satellite time series [26], and the resilience of radar to weather effects [4, 65]. Additionally, existing approaches are often limited for a given set of sensors, limiting their applicability.

To address these challenges, we introduce OmniSat, a novel architecture designed for the self-supervised fusion of diverse EO data. Existing multimodal approaches often map

multiple unrelated observations from different modalities to one pivot modality [38, 82] or a shared latent space [39, 84]. In contrast, OmniSat merges multiple views of the same area from different modalities into a single representation combining the specific information of each modality [14, 41, 73].

In computer vision, obtaining finely aligned multimodal observations generally requires specialized sensors [54, 59, 67] or the computation of complex mappings between modalities [23, 75]. On the other hand, EO data can be naturally aligned with georeferencing. To leverage this property, we adapt multimodal contrastive learning [49, 72] and cross-modal masked auto-encoding techniques [43] to learn rich multimodal EO representations with a generalist fusion scheme and without annotations.

To address the scarcity of EO datasets with a diverse range of heterogeneous modalities (see Tab. 1), we enrich the TreeSatAI [2] and PASTIS-R [33, 34] datasets with new aligned modalities. This allows us to evaluate OmniSat’s ability to handle an arbitrary number of inputs with varying natures and resolutions. Our contributions can be summarized as follows:

- We introduce OmniSat, a new model that learns to combine varied sources of EO observations in a self-supervised manner, resulting in richer joint representations that capture the unique characteristics of each modality.
- We augment two EO benchmarks to create the first datasets with three modalities of different natures (very high resolution images, optical and SAR time series).
- We demonstrate that OmniSat can leverage diverse modalities to learn rich representations, establishing new states-of-the-art for tree species, crop type, and land cover classification. Furthermore, our cross modal self-supervised training scheme improves performance even when only one modality is available during inference.

2 Related Work

This section provides an overview of self-supervised and multimodal learning, emphasizing the specificities of their usage for Earth observation. Lastly, we highlight the scarcity of multimodal EO datasets with diverse data types.

Self-Supervised Learning. This technique consists in learning expressive data representations without labels by using a pretext task. This approach has been particularly successful for natural language [52] and image [70] analysis. Initially focused on discriminative tasks [37, 68, 100], recent self-supervised approaches for images can be categorized as contrastive or generative.

Contrastive methods minimize the distance between representations of paired samples, often the same image under different transformations, and maximize the distance with other samples [15, 17, 46]. More efficient methods only consider positive samples and avoid mode collapse by introducing various asymmetries [18, 42] or normalization [16]. Such approaches have been successfully adapted to EO, for which samples are paired according to their location [89] or time of acquisition [7, 63].

Generative methods reason at the level of individual token—a small portion of the input, typically a patch for images [25]. The objective is to reconstruct the masked tokens of an input image in pixel [10, 45, 95] or feature space [6]. This principle has been successfully adapted to EO analysis [20, 30, 99], and was further extended to handle multiple spatial scales [74], multimodality [29, 51], or hyperspectral observations [50, 60].

Table 1: Publicly Available Multimodal EO Datasets. We provide in parenthesis the spatial resolutions of the single-date images and labels, and the temporal resolutions of time series. S1/S2 denotes Sentinel-1 and 2. * : **modalities added in this work**.

Dataset	Modalities		Labels
	images (single date)	time series	
SpaceNet6 [81]	SAR+optical (0.5m-2m)	✗	building footprint (<1m)
TreeSatAI [2]	aerial + S1/S2 (0.2-10m)	✗	forestry (60m)
BigEarthNet [86]	S1/S2 (10m)	✗	land cover (100m)
DFC20 [76]	S1/S2 (10m)	✗	land cover (500m)
MDAS [48]	S1/S2 + hyperspectral (2.2-10m)	✗	land cover (0.25m)
DOFA [96]	NAIP + Gaofen + S1/S2 + EnMAP (1-30m)	✗	✗
PASTIS-R [33, 34]	✗	S1/S2 (30-140 / year)	agriculture (10m)
SSL4EO-S12 [92]	✗	S1/S2 (4 / year)	✗
DFC21-DSE [61]	✗	S1/S2 + LS8 (3-9/year)	human activity (500m)
MapInWild [28]	✗	S1/S2 (4 / years)	protected areas (10m)
SEN12MS-CR-TS [27]	✗	S1/S2 (30 / years)	cloud cover (10m)
MultiSenGE [93]	✗	S1/S2 (30-140 / years)	land cover (10m)
FLAIR [31]	aerial (0.2m)	S2 (20-114 / year)	land cover (0.2m)
Satlas [12]	NAIP (0.5 -2m)	S2 (8-12 / year)	various
PASTIS-HD	* SPOT 6-7 (1.5m)	S1/S2 (30-140 / year)	agriculture (10m)
TreeSatAI-TS	aerial (0.2m)	* S1/S2 (10-70 / year)	forestry (60m)

Several hybrid approaches combine the discriminative power of contrastive methods and the scalability of generative objectives for natural images [70, 101] and EO data [29]. Our proposed OmniSat model also implements both mechanisms. A key feature is that we leverage the precise alignment between different sources of EO data to contrastively match small patches of different modalities rather than entire images or time series.

Self-Supervised Multimodal Learning. Multimodal computer vision has received a lot of interest [13], notably due to the success of cross-modal pre-training [72]. Recent models align the embeddings of heterogeneous modalities such as video and sound [49], depth and images [44], text and image [3, 9], or multiple combinations of these modalities [38, 39, 82, 84].

Multimodal learning also has a long history in EO [58, 71, 98] due to the large variety and complementarity of sensors [36, 79]. However, recent transformer-based architectures [90] for EO are often limited to one type of modality, be it a single image [20, 74] or time-series [34, 87]. For example, CROMA [29] and PRESTO [88] are specifically designed for paired optical and radar observations, but cannot handle Very High Resolution (VHR) data. USat [51] considers images with different resolutions, but only takes a single date within a time series. UT&T [31] can natively take single and multi-date observations of different modalities, but cannot be easily pre-trained in a self-supervised manner since it relies on convolutions and an ad-hoc late fusion scheme.

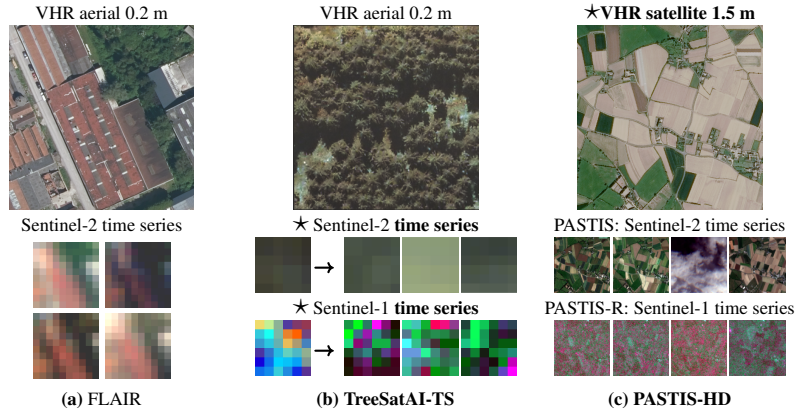


Fig. 1: Datasets. We represent three tiles from the considered multilabel classification datasets: FLAIR (a), TreeSatAI-TS (b) and PASTIS-HD (c). TreeSatAI-TS is a new dataset built by replacing the single-date Sentinel-1 and 2 images of TreeSatAI [2] by year-long time series. PASTIS-HD (c) adds VHR satellite images to PASTIS-R [34]. \star : **modalities added in this work**.

Multimodal EO Datasets. As reported in Table 1, many multimodal EO datasets use Sentinel-1 [11] and 2 [26] data for applications ranging from land cover to forestry analysis and fire detection. We also note that most multimodal datasets only contain data of one type: mono-date image or time series. Several datasets (BigEarthNet [86], DFC20 [76], MDAS [48]) select a single date from time series. However, single Sentinel-1 and 2 acquisitions can be significantly affected by rain and cloud cover, respectively. Furthermore, capturing the temporal dynamics is crucial to characterize the phenology of vegetation [91],

FLAIR [31] is the first multimodal EO dataset to propose both very high spatial resolution ($\leq 2\text{m}$) and high temporal resolution (> 4 images/year). Satlas [12] combines Sentinel-2 time series and for 5% to tiles (continental US), very high definition NAIP images. The functional map of the World [19] integrates observations from various sensors, but most areas are only observed with one sensor. Two other datasets contain time series and single images from multiple sources, but were not available at the time of writing this article: IARPA-SMART [40] and DOFA [96].

To showcase how OmniSat can consume an arbitrary number of modalities with different spatial, spectral, and temporal resolutions, we selected two commonly used EO benchmarks, TreeSatAI [2] and PASTIS-R [34], whose focus on crop type mapping and forestry differs from the land cover analysis of FLAIR. We added new modalities to these datasets to reach three distinct data types: VHR aerial images, optical time series, and SAR time series. See Fig. 1 for an illustration, and Sec. 4.1 for more details on how we extended these datasets.

3 Method

We consider a tile x observed through a set \mathbf{M} of M distinct sensors or modalities. The goal of the OmniSat model is to learn in a self-supervised fashion to combine

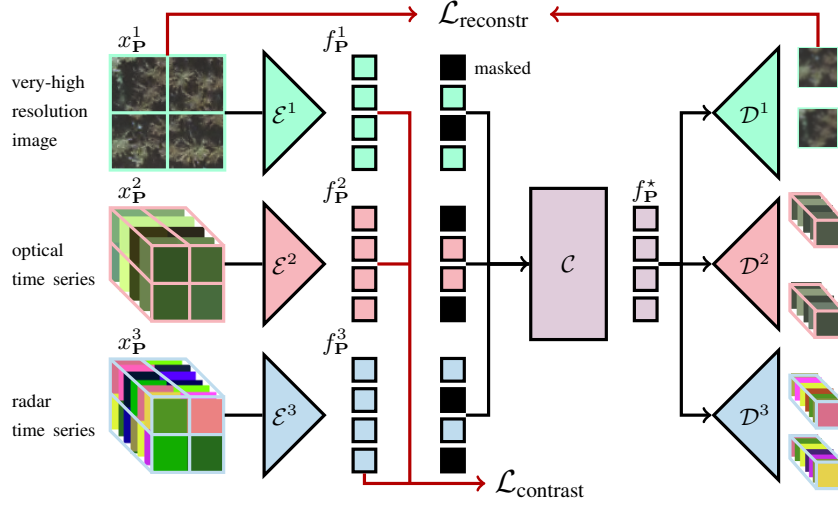


Fig. 2: OmniSat Architecture. We illustrate OmniSat for $M = 3$ modalities, and a tile split into $P = 4$ patches. The $M \times P$ input tokens $x_{\mathbf{P}}^{\mathbf{M}}$ are encoded by M modality-specific encoders $\mathcal{E}^{\mathbf{M}}$, yielding the token representations $f_{\mathbf{P}}^{\mathbf{M}}$. The module \mathcal{C} combines them into multimodal patch representations $f_{\mathbf{P}}^*$. The token embeddings $f_{\mathbf{P}}^{\mathbf{M}}$ are supervised by a contrastive cross-modal objective. We also use a reconstruction objective: the masked multimodal representations $f_{\mathbf{P}}^*$ are decoded by modality-specific networks $\mathcal{D}^{\mathbf{M}}$ to reconstruct their corresponding inputs in $x_{\mathbf{P}}^{\mathbf{M}}$.

all modalities \mathbf{M} into a multimodal representation f^* . We first provide details about OmniSat’s architecture in Sec. 3.1. We then explain our training scheme, which consists of a cross-modal contrastive objective (Sec. 3.2) and a multimodal masked encoding task (Sec. 3.3). Finally, we present the implementation details in Sec. 3.4. The overall method is represented in Fig. 2.

3.1 Architecture

This section presents the tokenization process, the structures of the encoder and decoder for each modality, and the architecture of the modality combiner network.

Multimodal Tokenization. All available modalities are spatially aligned through georeferencing. This allows us to divide the tile x into a set \mathbf{P} of P non-overlapping patches consistently across all modalities: $x_p^{\mathbf{M}} = \{x_p^m\}_{m \in \mathbf{M}}$ corresponds to M distinct views of the same patch p with different modalities. Each modality m takes its values in a space Ω^m such that $x_p^m \in \Omega^m$. We index tokens with pairs (m, p) , defined for each modality m and patch p , for a total of $M \times P$ tokens.

Time series from Sentinel satellites may experience registration errors spanning several meters, complicating their precise alignment with high-resolution imagery. However, using temporal sequences of satellite data mitigates these errors as aggregation over time tends to balance out misalignments.

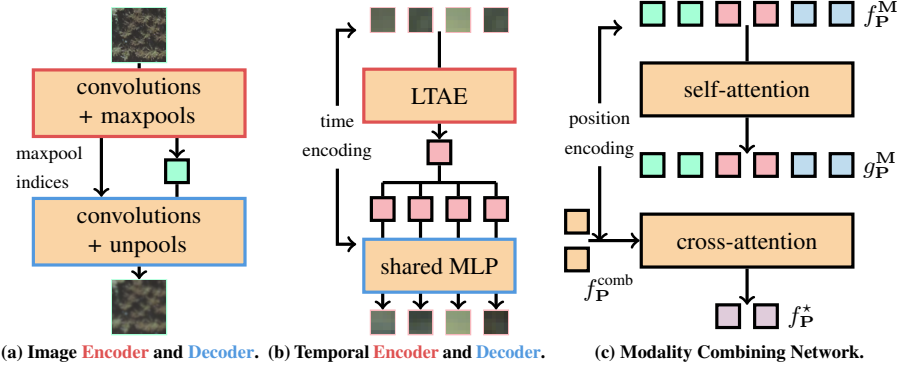


Fig. 3: OmniSat Architecture. OmniSat is composed of dedicated patch encoders for image (a) and time series b, here represented for a length of $L = 4$ time stamps. The modality combining module \mathcal{C} is depicted in (c) with $P = 2$ and $M = 3$. Elements colored in orange are learned networks or parameters.

Encoder-Decoder for Images. We split image tiles into small square patches: $\Omega^{\text{img}} = \mathbb{R}^{C \times W \times W}$ with W the size of the patches in pixels and C the number of channels. As shown in Fig. 3a, we encode these inputs with a sequence of convolutions and max-pool layers until the spatial dimension is fully collapsed. Decoding involves a symmetric sequence of convolutions and un-pooling layers. Contrary to existing masked auto-encoders, we pass the pooling indices from the encoder’s max-pooling to the decoder’s un-pooling in the manner of SegNet [8]. This dispenses the encoder from learning the intra-patch spatial configuration. This allows the image encoder to focus on the radiometric information, which may be more relevant depending on the application.

Encoder-Decoder for Time Series. Each temporal patch is represented by L sequential observations with C channels: $\Omega^{\text{TS}} = \mathbb{R}^{C \times L}$, each associated with a time stamp. We encode the temporal patches using a Lightweight Temporal Attention Encoder (LTAE) model [32], an efficient network for geospatial time series processing. We decode vector representations into time series by repeating the vector L times across the temporal dimension, adding a temporal encoding for each time step, and using an MLP to map the results to size C . See Fig. 3b for an illustration.

Optical time series are notoriously affected by clouds [85]. This may affect the validity of the reconstruction task: the decoder cannot know which observations are cloudy, making the reconstruction objective unpredictable. To circumvent this issue, we use the temporal attention maps of the encoder’s LTAE to select dates to reconstruct: cloudless observations are more informative and should have a higher attention score [78]. We only consider in the reconstruction loss $\mathcal{L}_{\text{reconstr}}$ the top 25% dates in terms of the LTAE’s attention maps.

Modality Combining Network. The modality combining network \mathcal{C} , represented in Fig. 3c, takes the $M \times P$ token embeddings f_P^M , some of whom can potentially be masked. We equip each token with a Euclidean relative positional encoding [94], calculated based on their patch’s position $\{r(p, q) \mid (p, q) \in \mathbf{P}^2\}$, allowing each token to

selectively consider its spatial surroundings. As most EO data are captured from above (satellite or aerial), their distribution is invariant by horizontal translation, making this choice of encoding preferable to an absolute position encoding.

The modality combining module \mathcal{C} starts with a series of B residual self-attention blocks connecting all tokens across modality. We then perform cross-attention between the resulting token embeddings $g_{\mathbf{P}}^{\mathbf{M}} \in \mathbb{R}^{d \times M \times P}$ and P copies $f_{\mathbf{P}}^{\text{comb}}$ of a modality combining token $f^{\text{comb}} \in \mathbb{R}^d$ learned as a free parameter. Each copy of f_p^{comb} is spatially located at the patch p for the relative positional encoding r . The module \mathcal{C} outputs P multimodal encodings $f_{\mathbf{P}}^*$ combining all available modalities for each patch:

$$g_{\mathbf{P}}^{\mathbf{M}} = \text{self-attention}(f_{\mathbf{P}}^{\mathbf{M}}; r) \quad (1)$$

$$f_{\mathbf{P}}^* = \text{cross-attention}(f_{\mathbf{P}}^{\text{comb}}, g_{\mathbf{P}}^{\mathbf{M}}; r) . \quad (2)$$

3.2 Contrastive Objective

We denote by f_p^m the d -dimensional encodings of the input patch x_p^m given by their dedicated encoders. We propose to supervise the embeddings f_p^m with a contrastive objective encouraging spatial consistency *across modalities*. Indeed, while each modality captures distinct characteristics of p , all encodings f_p^m share the same latent variable: the semantic content of the patch.

In practice, we want f_p^m to be closer to f_p^n for $n \neq m$, than to f_q^n for other patches $q \neq p$. We define \mathbf{B} as the set of patches within the current batch of observations. We adapt the classic InfoNCE loss [69] to our setting with two main differences, illustrated in Fig. 4. (i) Each token (m, p) has $M - 1$ positive matches: the tokens corresponding to the same patch p but viewed in another modality $n \neq m$; and (ii) as EO observations are generally spatially regular, nearby patches may be visually indistinguishable. Therefore, we exclude from the negative matches of (m, p) all tokens in modality m and which are too close to p . To this end, we remove the set $T(m, p)$ of tokens with modality m and whose patches are in the same tile as p . Our loss function $\mathcal{L}_{\text{contrast}}$ is defined as such:

$$\mathcal{L}_{\text{contrast}} = \frac{1}{M|\mathbf{B}|} \sum_{(m,p) \in \mathbf{M} \times \mathbf{B}} \log \left(\frac{\sum_{n \neq m} \exp(\langle f_p^m, f_p^n \rangle / \gamma)}{\sum_{(n,q) \in \mathbf{M} \times \mathbf{B} \setminus T(m,p)} \exp(\langle f_p^m, f_q^n \rangle / \gamma)} \right), \quad (3)$$

with γ a temperature parameter, and $\langle \cdot, \cdot \rangle$ the scalar product in \mathbb{R}^d . This function, specifically designed for geospatial data, allows us to contrast individual patches across modalities, which is not typically feasible for natural images. However, as the contrastive objective aligns multimodal representations, the patch encoders may be encouraged to overlook the distinct attributes of their respective modality. Instead, they may focus only on features shared by all modalities, *i.e.*, their *common denominator*. To ensure that encoders also capture modality-specific information, we incorporate a reconstruction objective, detailed in Sec. 3.3.

3.3 Multimodal Reconstruction Objective

During training, we mask a fraction of tokens $\mathbf{K} \subset \mathbf{M} \times \mathbf{P}$ and replace their embeddings with a learned vector $f^{\text{mask}} \in \mathbb{R}^d$. Note that the masking can differ across modalities,

		Tile ₁						Tile ₂					
		m_1		m_2		m_3		m_1		m_2		m_3	
Tile ₁	m_1	p_1	p_2	p_1	p_2	p_1	p_2	q_1	q_2	q_1	q_2	q_1	q_2
		o	o	+	-	+	-	-	-	-	-	-	-
	m_2	+	-	o	+	-	+	-	-	-	-	-	-
	m_3	+	-	+	-	o	o	-	-	-	-	-	-
		-	+	-	+	o	o	-	-	-	-	-	-
	m_3	-	+	-	+	o	o	-	-	-	-	-	-
Tile ₂	m_1	-	-	-	-	-	-	o	o	+	-	+	-
		-	-	-	-	-	-	o	o	-	+	+	+
	m_2	-	-	-	-	-	-	+	-	o	o	+	-
	m_3	-	-	-	-	-	-	-	+	o	-	+	+
		-	-	-	-	-	-	-	+	-	+	o	o
	m_3	-	-	-	-	-	-	-	+	-	+	o	o

Fig. 4: Contrastive Loss. We represent the token matching matrix for two tiles Tile₁ and Tile₂ viewed across 3 modalities m_1 , m_2 , and m_3 . Tile₁ is composed of the patches p_1 and p_2 , while Tile₂ comprises q_1 and q_2 . In contrast to classic approaches which ignore the diagonal and assign each sample with a single positive match, our loss defines operates at the patch level, considers multiple positives per token, and excludes tokens in a block-diagonal fashion.

+	positive match
-	negative match
o	ignored

and some patches may be entirely masked. All tokens are then processed by the modality combining network \mathcal{C} , which outputs P multimodal embeddings $f_{\mathbf{P}}^*$:

$$f_{\mathbf{P}}^* = \mathcal{C} \left(\{f_p^m\}_{(m,p) \notin \mathbf{K}} \cup \{f^{\text{mask}}\}_{(m,p) \in \mathbf{K}} \right). \quad (4)$$

To encourage the patch embeddings $f_{\mathbf{P}}^*$ to capture information from all modalities, we build a multimodal reconstruction objective. We denote by $\mathcal{D}^m : \mathbb{R}^d \mapsto \Omega^m$ the dedicated decoder of each modality m and write the reconstruction loss as:

$$\mathcal{L}_{\text{reconstr}} = \frac{1}{|\mathbf{K}|} \sum_{(m,p) \in \mathbf{K}} \frac{1}{\dim(\Omega^m)} \|\mathcal{D}^m(f_p^*) - x_p^m\|^2, \quad (5)$$

with $\dim(\Omega^m)$ the dimension of Ω^m . The total loss is the sum of $\mathcal{L}_{\text{reconstr}}$ and $\mathcal{L}_{\text{contrast}}$.

3.4 Implementation Details

We detail here the specific parameters chosen in all our experiments.

Tokenization. We split each tile along a regular spatial grid to produce a set of non-overlapping patches \mathbf{P} consistent across all modalities. For TreeSat and FLAIR, we use a 10×10 m grid, meaning that the VHR input tokens are small image patches of size 50×50 with 0.2 m per pixel. The patches of Sentinel observations with a resolution of 10m are single-pixel temporal sequences of spectral measurements. For PASTIS-HD, we use a 40×40 m grid, meaning that the VHR patches are of size 40×40 with 1.0 m per pixel. The patches of Sentinel observations [26] are 4×4 image time series which we spatially flatten before encoding.

Hyperparameters. To show the versatility of OmniSat, we use the same configuration throughout all experiments. The embedding size is $d = 256$, resulting in image encoders and decoders with 3.6M and 1.8M parameters, 403k and 96k for optical time series, and 402k and 95k for radar time series. The modality combiner module is composed of $B = 6$ residual self-attention blocks and a single cross-attention block, for a total of 3.6M parameters. We train our model on 3 A6000 GPUs with a batch size of 128

multimodal tiles per GPU and set the contrastive temperature γ to 0.1. We train our model with the ADAM optimizer [53], with a learning rate of 10^{-4} for pretraining and 2×10^{-5} for fine-tuning, and a ReduceLROnPlateau scheduler [1] with a patience of 10 epochs and a decay rate of 0.1. When re-implementing competing methods, we use the hyperparameters of their open-source repository.

4 Experiments

We evaluate OmniSat’s performance across three multimodal datasets, including two new datasets introduced in this work, and presented in Sec. 4.1. We outline our experimental protocol and our adaptation of competing methods in Sec. 4.2. We then present our quantitative results and analysis in Sec. 4.3 and conduct an ablation study in Sec. 4.4.

4.1 Datasets

We consider three multimodal datasets: FLAIR [31], and the augmented TreeSatAI-TS [2] and PASTIS-HD [33, 34]. See Fig. 1 for an illustration of these two last datasets.

TreeSatAI-TS: TreeSatAI [2] is a multimodal dataset for tree species identification, containing 50,381 tiles of 60×60 m with multi-label annotations for 20 classes and all taken in Germany. Each tile is associated with a very high resolution RGB and near-infrared (NIR) image (0.2 m pixel resolution), a single Sentinel-2 multi-spectral image (10 m per pixel resolution, 10 bands), and a single Sentinel-1 radar image (10 m per pixel resolution, 3 bands: two polarization channels and their ratio).

Motivated by the fact that fine-grained vegetation discrimination relies heavily on temporal dynamics information [91], we introduce TreeSatAI-TS¹. This extended version uses open-source data to add Sentinel-1 and Sentinel-2 time series to each tile, spanning the closest available year to the VHR observation for Sentinel-2. Note that due to the weather patterns and position of the area of interest with respect to Sentinel-2’s orbit, the optical time series is particularly irregular and occluded, with up to 50% of acquisitions being non-exploitable. Despite this challenge, we included the raw observations without pre-processing, whereas TreeSatAI’s single-date images have been manually selected.

PASTIS-HD: The PASTIS dataset [33], is designed for semantic and panoptic segmentation of agricultural parcels using Sentinel-2 time series and covers 18 crop types across 2,433 image time series with dimensions of 1280×1280 m. Each series contains between 38 and 61 observations with 10 spectral bands. PASTIS-R [34] adds the corresponding Sentinel-1 radar time series. We only used the ascendent time series of Sentinel-1 for our training and evaluation, for a total of 169,587 radar images with three bands.

To enhance the spatial resolution and utility of PASTIS, we introduce PASTIS-HD², which integrates contemporary VHR satellite images (SPOT 6-7 [24]). We apply orthorectification and pansharpening, resample the resulting images to a 1m resolution, and finally convert them to 8 bits. We follow the protocol of Irvin *et al.* [51] to use the

¹The dataset is available at <https://huggingface.co/datasets/IGNF/TreeSatAI-Time-Series>.

²The dataset is available at <https://huggingface.co/datasets/IGNF/PASTIS-HD>.

dense annotations for a multi-label classification task: each patch is associated with the labels of all of its pixels. This conversion allows us to evaluate all methods in the same setting and configuration as TreeSatAI.

FLAIR. The FLAIR dataset [31] combines VHR aerial images with time series data. It comprises 77,762 aerial tiles (512×512 pixels, 0.2 m resolution) with five channels (RGB, near-infrared, and a normalized digital surface model) taken in France, alongside corresponding Sentinel-2 time series (10 m resolution, 10 spectral bands, 20 to 114 observations per year). We apply the same processing as PASTIS to use the dense annotation for a multi-label classification task.

4.2 Experimental Setting

This section details our experimental protocol and our adaption of competing algorithms.

Evaluation Protocol. All experiments follow a similar setting:

- **Pre-training (optional).** Methods that support self-supervised pre-training (OmniSat, SatMAE [20], ScaleMAE [74], CROMA [29]) are pre-trained for up to 250 epochs on the entire training set without access to labels.
- **Fine-Tuning.** We propose two settings for fine-tuning:
 - **Fully Supervised Fine-Tuning.** We train the resulting models using all the labels in the training set.
 - **Semi-Supervised Fine-Tuning.** We use a portion of 10% or 20% of the training set, stratified by the distribution of classes, to fine-tune the models. For models without pre-training, this corresponds to supervision in the low-data regime.
- **Unimodal and Multimodal Evaluation.** We evaluate all methods using each available modality independently and combining all supported modalities.


Adapting Competing Approaches. We report the performance of several methods taken from the literature on our considered datasets: LightGBM [2], PRESTO [88], and MOSAIKS [77]. However, few existing methods can operate on single- and multi-date data at the same time. To ensure a fair evaluation of competing approaches, we modify various state-of-the-art models to handle a broader combination of modalities. We provide details on these changes in the appendix.




4.3 Numerical Experiments and Analysis

In this section, we report our model’s performance and efficiency compared to other approaches across the considered datasets and propose our analysis.

TreeSatAI-TS. Tab. 2 presents the performance of different models on TreeSatAI and TreeSatAI-TS. We report several key observations:

- **Benefit of Time Series.** For the original TreeSatAI dataset with single-date Sentinel-1/2 observations, none of the pre-training schemes significantly improve performance beyond simple baselines such as ResNet, PSE, or MLP, even in a semi-supervised setting. In particular, single-date S1 observations yield low performance for all methods (below 20 F1-score), emphasizing the need to use the entire time series.

Table 2: Performance on TreeSatAI-TS. We report the weighted F1 for multi-label tree species classification on TreeSatAI (TSAI) and our extended TreeSatAI-TS (TSAI-TS) dataset when fine-tuning with 10% and 100% of training labels. The first line of the table is the modality used for evaluation. We distinguish methods that are **best for one modality** within a dataset, **best in a dataset** across all modalities, and the **best overall** performance. *: late feature fusion with a ResNet pre-trained on ImageNet. : Foundation model trained on extensive external data. †: model evaluated on this dataset for the first time.

Model	pre- training	All		Sentinel-1		Sentinel-2		VHR Image	
		10%	100%	10%	100%	10%	100%	10%	100%
Evaluated on TreeSatAI: single date for Sentinel-1 and Sentinel-2									
† PSE [35]	None	47.2*	68.1*	11.5	14.6	48.5	58.3	-	-
† ViT [25]	None	42.7	57.1	8.7	17.5	39.8	57.3	36.7	51.7
MLP [2]	None	42.6*	71.5*	3.4	10.1	22.1	52.0	-	-
ResNet [2]	ImageNet	-	-	-	-	-	-	58.8	70.1
LightGBM [2]	ImageNet	-	54.3*	-	11.9	-	48.2	-	44.0
PRESTO [88]		-	-	-	19.8	-	46.3	-	-
† DOFA [96]		59.5	71.6	11.6	19.3	48.2	57.0	51.6	67.5
MOSAICS [22, 77]	TSAI	-	-	-	-	-	56.0	-	-
† CROMA [29]	TSAI	49.6	61.0	10.1	12.7	47.8	55.7	-	-
† SatMAE [20]	TSAI	46.1	61.5	-	-	40.3	49.7	44.1	61.4
† ScaleMAE [74]	TSAI	47.6	62.5	-	-	46.7	55.2	46.9	63.6
OmniSat (ours)	TSAI	56.2	70.4	5.3	6.4	48.4	57.1	52.8	68.9
Evaluated on TreeSatAI-TS: Sentinel-1 and Sentinel-2 time series spanning one year									
† PSE+LTAE [35]	None	59.4*	71.2*	42.6	52.4	44.0	57.2	-	-
† UT&T [31]	ImageNet	43.8	56.7	42.3	55.2	41.5	57.0	44.3	55.9
† DOFA [96]		41.8	71.3	0.0	0.0	25.0	39.4	51.6	67.5
† Scale-MAE [74]	TSAI-TS	44.1	60.4	-	-	11.0	31.5	46.9	63.6
OmniSat (ours)	None	52.2	73.3	31.6	55.9	33.9	49.7	51.4	71.0
OmniSat (ours)	TSAI-TS	61.1	74.2	48.2	56.7	51.4	62.9	58.3	70.5

OmniSat exhibits significantly improved results on TreeSatAI-TS, with or without pretraining. Image models struggles to extract meaningful features temporally aggregated temporal observations, while OmniSat learn rich dynamic features.


The foundation model DOFA [96], with 111M parameters and a large closed-source training set, outperforms all models when evaluated on single-date modalities. However, OmniSat reaches higher performances on TreeSatAI-TS with only 10 million parameters, which we attribute to its ability to leverage temporal modalities.

- **Benefits of Multimodality.** When using all modalities, OmniSat outperforms all competing methods by a margin of 3% F1-score. The multimodal performance of OmniSat and CROMA, which learn to combine data sources, is strictly superior to the F1-score of their best modality by 3.7% and 5.3% points, respectively. Conversely, the performance of methods that rely on late-fusion (SatMAE, ScaleMAE,

Table 3: Performance on PASTIS-HD. We report the macro-averaged F1-score for crop-type multi-class classification on the PASTIS-HD dataset. We distinguish methods that are **best for one modality**, **best in a dataset** across all modalities. *: late feature fusion with a ResNet. †: model evaluated on this dataset for the first time.

Model	pre-trained	All		Sentinel-1		Sentinel-2		VHR image	
		20%	100%	20%	100%	20%	100%	20%	100%
† UTAE [33, 34]	None	36.8*	46.9*	20.1	40.7	32.7	37.6	-	-
† ResNet50 [47]	ImageNet	-	-	-	-	-	-	57.6	59.3
† UT&T [31]	ImageNet	54.2	53.5	58.8	62.8	54.9	61.3	51.1	49.8
† DOFA [96]		53.7	55.7	36.7	41.5	50.8	53.4	47.9	54.8
† Scale-MAE [74]	PASTIS-HD	42.0	42.2	-	-	41.2	46.1	48.8	51.9
† CROMA [29]	PASTIS-HD	57.5	60.1	55.3	56.1	53.0	56.7	-	-
OmniSat (ours)	No	42.0	59.1	58.2	60.2	51.7	60.1	47.3	52.8
OmniSat (ours)	PASTIS-HD	62.6	69.9	60.8	69.0	61.8	70.8	54.6	59.3

Table 4: Performance on FLAIR. We report the macro-averaged F1-score for land cover multi-class classification on the FLAIR dataset. We distinguish methods that are **best for one modality** and **best in a dataset**. †: model evaluated on this dataset for the first time.

Model	pre-trained	All		Sentinel-2		VHR Image	
		10%	100%	10%	100%	10%	100%
† UT&T [31]	ImageNet	44.2	48.8	57.4	62.0	58.9	65.5
† DOFA [96]		70.6	74.9	57.0	61.0	66.8	72.1
† ScaleMAE [74]	FLAIR	63.1	70.0	52.5	61.0	61.2	67.3
OmniSat (ours)	No	62.5	70.0	56.1	65.4	64.7	71.5
OmniSat (ours)	FLAIR	60.6	73.4	56.8	65.4	65.2	71.6

ViT) is comparable to their best modality. This demonstrates the value of learning to combine information from different sources end-to-end.

- **Benefits of Cross-Modal Pre-Training.** With access to all modalities, our self-supervised pre-training improves by 0.9% point the F1-score of the model fine-tuned on 100% of labels, compared to not pre-training, and 8.9% when using only 10% of labels. This shows that our pre-training leads to more expressive multimodal features. Interestingly, when performing inference with Sentinel-2 time series alone, the performance increase linked to the pre-training becomes 13.2% with 100% labels and 17.5% with 10%. This illustrates that our self-supervised pre-training scheme improves the features learned by each encoder despite not relying on annotated data.

Experiments on PASTIS-HD. The analysis of the performance of various models on PASTIS-HD is reported in Tab. 3, and is consistent with the ones of TreeSatAI-TS. First, by learning to combine all modalities despite their different resolutions, OmniSat achieves state-of-the-art results on this benchmark. Second, our cross-modal pretraining

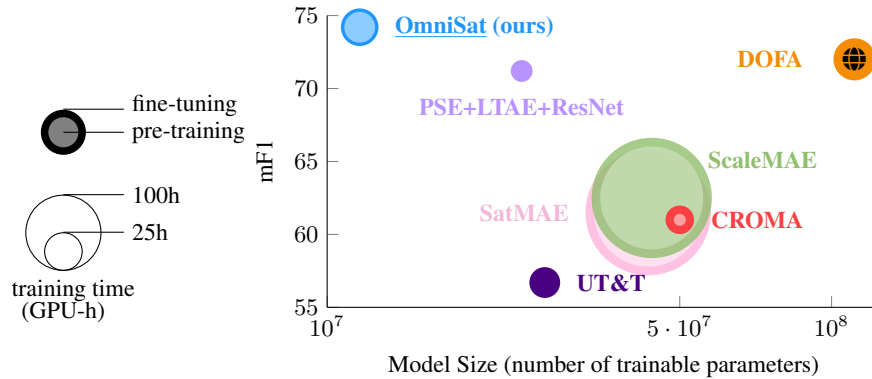


Fig. 5: Efficiency. We report the best performance of different models between TreeSatAI and TreeSatAI-TS, with pre-training and fine-tuning using 100% of labels. The area of the markers is proportional to the training time, broken down in pre-training and fine-tuning when applicable.

significantly improves OmniSat’s performance in the multimodal (+10.8 pF1-score with 100% of training label) and all single-modality settings (8.8 points for Sentinel-1, 10.7 for Sentinel-2, and 6.5 for the VHR images).

Experiments on FLAIR. We report in Tab. 4 the results on the bimodal FLAIR dataset for multilabel classification. OmniSat outperforms the much larger ScaleMAE [74] and UT&T [31] models with 100% of labels and both modalities by 3.4%. Our pre-training scheme had a smaller impact than for the TreeSatAI-TS experiment. We attribute this to the fact that only two modalities are available, which decreases the supervisory power of our cross-modal contrastive objective and our multimodal reconstruction loss. This highlights a limitation of OmniSat: the model needs to be pre-trained on a modality-rich dataset to achieve its best performance.

Efficiency Evaluation. We plot in Fig. 5 the best performance between TreeSatAI and TreeSatAI-TS for different models according to their size and training time. OmniSat is more compact, faster to train, and performs better than all evaluated models, including the DOFA foundation model. The highly-specialized combination of PSE, LTAE, and ResNet is a strong contender, outperforming significantly larger models with generic encoding-decoding schemes.

4.4 Ablation Study

In this section, we report the results of several experiments evaluating the impact and validity of our main design choices, see Tab. 5.

a) Encoder/Decoder Architecture. We propose several improvements to the standard image encoder-decoder scheme used in computer vision to accommodate the specificities of EO data. In particular, passing the max-pool indices from the image patch encoder to its decoder allows the learned representation to focus on characterizing the spectral signature instead of fine-grained spatial information, and leads to a performance increase of 0.7% in the full supervision setting.

Table 5: Ablation Study. We present the impact of several design choices on the TreeSatAI-TS dataset, measured in terms of macro-averaged F1-score.

Experiment	10%	100%	Experiment	10%	100%
OmniSat	61.1	74.2	b) no contrastive loss	55.6	73.4
a) no index bypass	57.5	73.5	b) naive contrastive loss	57.8	72.2
a) no date filtering	58.2	71.6	b) no reconstruction loss	59.0	72.2

As clouds frequently obstruct optical time series, we use a unsupervised date-filtering scheme to reconstruct only meaningful acquisitions. This approach leads to a significant improvement of 3.6%, showcasing the benefit of developing modality-aware approaches for EO.

b) Role of Loss Functions. When training without contrastive loss, we observe a decrease in performance of 0.8% in the fully supervised regime, and a more pronounced drop of 5.5% in the semi-supervised regime. This demonstrates how learning consistent encoding across encoders facilitates their subsequent fusion. Interestingly, when implementing a naive contrastive loss that considers all negative examples from the batch, the decrease is greater than simply removing this loss (2% in full supervision). This strategy may introduce indistinguishable negative examples and perturb the learning process.

We also remove the reconstruction loss, meaning that only the encoders are learned contrastively during pre-training. This results in a drop of 2% F1-score point, illustrating the importance of pre-training the transformer \mathcal{C} alongside its encoders.

Limitations. All datasets used in our experiments are based in Europe, primarily due to the availability of open-access annotations. This regional focus prevents us from evaluating our model’s performance in tropical and developing countries, which present unique challenges in terms of label provision, heterogeneity, and complex classes.

A limitation of our pre-training scheme is its dependence on a sufficient number of aligned modalities, as illustrated by its moderate impact on the bimodal FLAIR dataset.

5 Conclusion

We introduced OmniSat, a new architecture for the self-supervised modality fusion of Earth Observation (EO) data from multiple sources. To facilitate its evaluation, we augmented two existing datasets with new modalities of different natures and resolutions. We experimentally showed that leveraging diverse modalities with a flexible model improves the model’s performance in both fully and semi-supervised settings. Moreover, our training scheme can exploit the spatial alignment of multiple modalities to improve our model’s unimodal performance. Finally, we proposed several improvements to leverage the unique structure of EO data in the architecture of our model, such as automatic date filtering for reconstructing time series. We hope that our promising results and new datasets will encourage the computer vision community to consider EO data as a playing field for evaluating and developing novel self-supervised multimodal algorithms.

Acknowledgements

This work was supported by ANR project READY3D ANR-19-CE23-0007, and was granted access to the HPC resources of IDRIS under the allocations AD011014719 and AD011014286R1 made by GENCI. We thank Anatol Garioud and Sébastien Giordano for their help on the creation of TreeSatAI-TS and PASTIS-HD datasets. The SPOT images are opendata thanks to the Dataterra Dinamis initiative in the case of the "[Couverture France DINAMIS](#)" program. We thank Jordi Inglada for inspiring discussions and valuable feedback.

References

1. PyTorch: ReduceLROnPlateau. [org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html#torch.optim.lr_scheduler.ReduceLROnPlateau](https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html#torch.optim.lr_scheduler.ReduceLROnPlateau), accessed: 2024-02-29 [9](#)
2. Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., Kleinschmit, B.: TreeSatAI Benchmark Archive: A multi-sensor, multi-label dataset for tree species classification in remote sensing. *Earth System Science Data Discussions* (2022) [2](#), [3](#), [4](#), [9](#), [10](#), [11](#)
3. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: A visual language model for few-shot learning. In: *NeurIPS* (2022) [3](#)
4. Amitrano, D., Di Martino, G., Guida, R., Iervolino, P., Iodice, A., Papa, M.N., Riccio, D., Ruello, G.: Earth environmental monitoring using multi-temporal synthetic aperture radar: A critical review of selected applications. *Remote Sensing* (2021) [1](#)
5. Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., Friedl, L.: Earth observation in service of the 2030 agenda for sustainable development. *Geo-spatial Information Science* (2017) [1](#)
6. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: *CVPR* (2023) [2](#)
7. Ayush, K., Uzket, B., Meng, C., Tanmay, K., Burke, M., Lobell, D., Ermon, S.: Geography-aware self-supervised learning. In: *ICCV* (2021) [2](#)
8. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI* (2017) [6](#)
9. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: *ICML* (2022) [3](#)
10. Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. In: *ICLR* (2021) [2](#)
11. Bao, X., Zhang, R., Lv, J., Wu, R., Zhang, H., Chen, J., Zhang, B., Ouyang, X., Liu, G.: Vegetation descriptors from Sentinel-1 SAR data for crop growth monitoring. *ISPRS Journal of Photogrammetry and Remote Sensing* (2023) [4](#)
12. Bastani, F., Wolters, P., Gupta, R., Ferdinando, J., Kembhavi, A.: SatlasPretrain: A large-scale dataset for remote sensing image understanding. In: *ICCV* (2023) [3](#), [4](#)
13. Bayouddh, K., Knani, R., Hamdaoui, F., Mtibaa, A.: A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer* (2022) [3](#)

14. Benedetti, P., Ienco, D., Gaetano, R., Ose, K., Pensa, R.G., Dupuy, S.: M³-fusion: A deep learning architecture for multiscale multimodal multitemporal satellite data fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2018) 2
15. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: *NeurIPS* (2020) 2
16. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *ICCV* (2021) 2
17. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *ICML* (2020) 2
18. Chen, X., He, K.: Exploring simple siamese representation learning. In: *CVPR* (2021) 2
19. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: *CVPR* (2018) 4
20. Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., Ermon, S.: SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In: *NeurIPS* (2022) 2, 3, 10, 11
21. Coppin, P., Lambin, E., Jonckheere, I., Muys, B.: Digital change detection methods in natural ecosystem monitoring: A review. *Analysis of multi-temporal remote sensing images* (2002) 1
22. Corley, I., Robinson, C., Dodhia, R., Ferres, J.M.L., Najafirad, P.: Revisiting pre-trained remote sensing model benchmarks: Resizing and normalization matters. *arXiv preprint arXiv:2305.13456* (2023) 11
23. Dai, A., Nießner, M.: 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In: *ECCV* (2018) 2
24. DataTerra Dinamis: Diffusion OpenData Dinamis, <https://dinamis.data-terra.org/opendata/>, accessed: 2023-12-15 9
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *ICLR* (2020) 2, 11
26. Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P.: Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment* (2012) 1, 4, 8
27. Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X.: SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE TGRS* (2022) 3
28. Ekim, B., Stomberg, T.T., Roscher, R., Schmitt, M.: MapInWild: A remote sensing dataset to address the question of what makes nature wild. *IEEE Geoscience and Remote Sensing Magazine* (2023) 3
29. Fuller, A., Millard, K., Green, J.R.: CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In: *NeurIPS* (2023) 1, 2, 3, 10, 11, 12
30. Gao, Y., Sun, X., Liu, C.: A general self-supervised framework for remote sensing image classification. *Remote Sensing* (2022) 2
31. Garioud, A., Gonthier, N., Landrieu, L., De Wit, A., Valette, M., Poupée, M., Giordano, S., Wattrelos, B.: FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. In: *NeurIPS Dataset and Benchmark* (2023) 3, 4, 9, 10, 11, 12, 13
32. Garnot, V.S.F., Landrieu, L.: Lightweight temporal self-attention for classifying satellite images time series. In: *Advanced Analytics and Learning on Temporal Data: ECML PKDD Workshop* (2020) 6
33. Garnot, V.S.F., Landrieu, L.: Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: *ICCV* (2021) 2, 3, 9, 12

34. Garnot, V.S.F., Landrieu, L., Chehata, N.: Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing* (2022) [2](#), [3](#), [4](#), [9](#), [12](#)
35. Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N.: Satellite image time series classification with pixel-set encoders and temporal self-attention. In: *CVPR* (2020) [11](#)
36. Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., et al.: Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine* (2019) [1](#), [3](#)
37. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: *ICLR* (2018) [2](#)
38. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: *CVPR* (2023) [1](#), [2](#), [3](#)
39. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: A single model for many visual modalities. In: *CVPR* (2022) [2](#), [3](#)
40. Goldberg, H.R., Ratto, C.R., Banerjee, A., Kelbaugh, M.T., Giglio, M., Vermote, E.F.: Automated global-scale detection and characterization of anthropogenic activity using multi-source satellite-based remote sensing imagery. In: *Geospatial Informatics XIII*. SPIE (2023) [4](#)
41. Greenwell, C., Crall, J., Purri, M., Dana, K., Jacobs, N., Hadzic, A., Workman, S., Leotta, M.: WATCH: Wide-area terrestrial change hypercube. In: *WACV* (2024) [2](#)
42. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In: *NeurIPS* (2020) [2](#)
43. Hackstein, J., Sumbul, G., Clasen, K.N., Demir, B.: Exploring masked autoencoders for sensor-agnostic image retrieval in remote sensing. *arXiv preprint arXiv:2401.07782* (2024) [2](#)
44. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: *ACCV* (2017) [3](#)
45. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: *CVPR* (2022) [2](#)
46. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: *CVPR* (2020) [2](#)
47. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR* (2016) [12](#)
48. Hu, J., Liu, R., Hong, D., Camero, A., Yao, J., Schneider, M., Kurz, F., Segl, K., Zhu, X.X.: MDAS: A new multimodal benchmark dataset for remote sensing. *Earth System Science Data Discussions* (2022) [3](#), [4](#)
49. Huang, P.Y., Sharma, V., Xu, H., Ryali, C., Fan, H., Li, Y., Li, S.W., Ghosh, G., Malik, J., Feichtenhofer, C.: MAViL: Masked audio-video learners. In: *NeurIPS* (2023) [2](#), [3](#)
50. Ibanez, D., Fernandez-Beltran, R., Pla, F., Yokoya, N.: Masked auto-encoding spectral-spatial transformer for hyperspectral image classification. *IEEE TGRS* (2022) [2](#)
51. Irvin, J., Tao, L., Zhou, J., Ma, Y., Nashold, L., Liu, B., Ng, A.Y.: USat: A unified self-supervised encoder for multi-sensor satellite imagery. *arXiv preprint arXiv:2312.02199* (2023) [1](#), [2](#), [3](#), [9](#)
52. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: *NAACL* (2019) [2](#)
53. Kingma, D.P., Ba, J.: ADAM: A method for stochastic optimization. *ICLR* (2015) [9](#)
54. Krispel, G., Opitz, M., Waltner, G., Possegger, H., Bischof, H.: Fuseseg: LiDAR point cloud segmentation fusing multi-modal data. In: *WACV* (2020) [2](#)

55. Kuffer, M., Thomson, D.R., Boo, G., Mahabir, R., Grippa, T., Vanhuysse, S., Engstrom, R., Ndugwa, R., Makau, J., Darin, E., et al.: The role of Earth observation in an integrated deprived area mapping “system” for low-to-middle income countries. *Remote sensing* (2020) [1](#)
56. Lacoste, A., Sherwin, E.D., Kerner, H., Alemohammad, H., Lütjens, B., Irvin, J., Dao, D., Chang, A., Gunturkun, M., Drouin, A., et al.: Toward foundation models for Earth monitoring: Proposal for a climate change benchmark. *arXiv preprint arXiv:2112.00570* (2021) [1](#)
57. Li, D., Tong, Q., Li, R., Gong, J., Zhang, L.: Current issues in high-resolution Earth observation technology. *Science China Earth sciences* (2012) [1](#)
58. Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., Chanussot, J.: Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation* **112**, 102926 (2022) [3](#)
59. Liao, Y., Xie, J., Geiger, A.: KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE TPAMI* (2022) [2](#)
60. Liu, Y., Li, X., Hua, Z., Xia, C., Zhao, L.: A band selection method with masked convolutional autoencoder for hyperspectral image. *IEEE Geoscience and Remote Sensing Letters* (2022) [2](#)
61. Ma, Y., Li, Y., Feng, K., Xia, Y., Huang, Q., Zhang, H., Prieur, C., Licciardi, G., Malha, H., Chanussot, J., et al.: The outcome of the 2021 IEEE GRSS data fusion contest-Track DSE: Detection of settlements without electricity. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2021) [3](#)
62. Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., Gao, S., Liu, T., Cong, G., Hu, Y., et al.: On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798* (2023) [1](#)
63. Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., Rodriguez, P.: Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In: *ICCV* (2021) [2](#)
64. Manfreda, S., McCabe, M.F., Miller, P.E., Lucas, R., Pajuelo Madrigal, V., Mallinis, G., Ben Dor, E., Helman, D., Estes, L., Ciraolo, G., et al.: On the use of unmanned aerial systems for environmental monitoring. *Remote sensing* p. 641 (2018) [1](#)
65. Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., Papathanassiou, K.P.: A tutorial on synthetic aperture radar. *IEEE Geoscience and Remote Sensing Magazine* (2013) [1](#)
66. Nakalembe, C.: Urgent and critical need for sub-saharan african countries to invest in Earth observation-based agricultural early warning and monitoring systems. *Environmental Research Letters* (2020) [1](#)
67. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: *ECCV* (2012) [2](#)
68. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: *ECCV* (2016) [2](#)
69. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018) [7](#)
70. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. *TLMR* (2023) [2](#), [3](#)
71. Pohl, C., Van Genderen, J.L.: Multisensor image fusion in remote sensing: Concepts, methods and applications. *International journal of remote sensing* (1998) [3](#)
72. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *ICML* (2021) [2](#), [3](#)

73. Recasens, A., Lin, J., Carreira, J., Jaegle, D., Wang, L., Alayrac, J.b., Luc, P., Miech, A., Smaira, L., Hemsley, R., et al.: Zorro: The masked multimodal transformer. arXiv preprint arXiv:2301.09595 (2023) [2](#)
74. Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In: ICCV (2023) [2](#), [3](#), [10](#), [11](#), [12](#), [13](#)
75. Robert, D., Vallet, B., Landrieu, L.: Learning multi-view aggregation in the wild for large-scale 3D semantic segmentation. In: CVPR (2022) [2](#)
76. Robinson, C., Malkin, K., Jojic, N., Chen, H., Qin, R., Xiao, C., Schmitt, M., Ghamisi, P., Hänsch, R., Yokoya, N.: Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2021) [3](#), [4](#)
77. Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., Hsiang, S.: A generalizable and accessible approach to machine learning with global satellite imagery. Nature communications (2021) [10](#), [11](#)
78. Rußwurm, M., Körner, M.: Self-attention for raw optical satellite time series classification. ISPRS Journal of Photogrammetry and Remote Sensing (2020) [6](#)
79. Schmitt, M., Zhu, X.X.: Data fusion and remote sensing: An ever-growing relationship. IEEE Geoscience and Remote Sensing Magazine (2016) [1](#), [3](#)
80. Secades, C., O'Connor, B., Brown, C., Walpole, M., et al.: Earth observation for biodiversity monitoring: A review of current approaches and future opportunities for tracking progress towards the aichi biodiversity targets. CBD technical series (2014) [1](#)
81. Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T., et al.: Spacenet 6: Multi-sensor all weather mapping dataset. In: CVPR Workshop EarthVision (2020) [3](#)
82. Shukor, M., Dancette, C., Rame, A., Cord, M.: UnIVAL: Unified model for image, video, audio and language tasks. TMLR (2023) [1](#), [2](#), [3](#)
83. Skidmore, A.K., Coops, N.C., Neinavaz, E., Ali, A., Schaepman, M.E., Paganini, M., Kissling, W.D., Vihervaara, P., Darvishzadeh, R., Feilhauer, H., et al.: Priority list of biodiversity metrics to observe from space. Nature Ecology & Evolution (2021) [1](#)
84. Srivastava, S., Sharma, G.: OmniVec: Learning robust representations with cross modal sharing. In: WACV (2024) [2](#), [3](#)
85. Sudmanns, M., Tiede, D., Augustin, H., Lang, S.: Assessing global Sentinel-2 coverage dynamics and data availability for operational Earth observation (EO) applications using the EO-Compass. International Journal of Digital Earth (2019) [6](#)
86. Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V.: BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. IEEE Geoscience and Remote Sensing Magazine (2021) [3](#), [4](#)
87. Tarasiou, M., Chavez, E., Zafeiriou, S.: ViTs for SITS: Vision transformers for satellite image time series. In: CVPR (2023) [3](#)
88. Tseng, G., Zvonkov, I., Purohit, M., Rolnick, D., Kerner, H.: Lightweight, pre-trained transformers for remote sensing timeseries. arXiv preprint arXiv:2304.14065 (2023) [3](#), [10](#), [11](#)
89. Tseng, W.H., Lê, H.Â., Boulch, A., Lefèvre, S., Tiede, D.: CROCO: Cross-modal contrastive learning for localization of Earth observation data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2022) [2](#)
90. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) [3](#)

91. Vrieling, A., Meroni, M., Darvishzadeh, R., Skidmore, A.K., Wang, T., Zurita-Milla, R., Oosterbeek, K., O'Connor, B., Paganini, M.: Vegetation phenology from Sentinel-2 and field cameras for a Dutch barrier island. *Remote sensing of environment* (2018) [4](#), [9](#)
92. Wang, Y., Braham, N.A.A., Xiong, Z., Liu, C., Albrecht, C.M., Zhu, X.X.: SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in Earth observation. *IEEE Geoscience and Remote Sensing Magazine* (2023) [3](#)
93. Wenger, R., Puissant, A., Weber, J., Idoumghar, L., Forestier, G.: MultiSenGE: A multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (2022) [3](#)
94. Wu, K., Peng, H., Chen, M., Fu, J., Chao, H.: Rethinking and improving relative position encoding for vision transformer. In: *ICCV* (2021) [6](#)
95. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: SimMim: A simple framework for masked image modeling. In: *CVPR* (2022) [2](#)
96. Xiong, Z., Wang, Y., Zhang, F., Stewart, A.J., Hanna, J., Borth, D., Papoutsis, I., Saux, B.L., Camps-Valls, G., Zhu, X.X.: Neural plasticity-inspired foundation model for observing the Earth crossing modalities. *arXiv preprint arXiv:2403.15356* (2024) [3](#), [4](#), [11](#), [12](#)
97. Yang, J., Gong, P., Fu, R., Zhang, M., Chen, J., Liang, S., Xu, B., Shi, J., Dickinson, R.: The role of satellite remote sensing in climate change studies. *Nature climate change* (2013) [1](#)
98. Yang, M.Y., Landrieu, L., Tuia, D., Toth, C.: Multi-modal learning in photogrammetry and remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* (2021) [3](#)
99. Yuan, Y., Lin, L., Liu, Q., Hang, R., Zhou, Z.G.: SITS-Former: A pre-trained spatio-spectral-temporal representation model for sentinel-2 time series classification. *International Journal of Applied Earth Observation and Geoinformation* (2022) [2](#)
100. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: *ECCV* (2016) [2](#)
101. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: Image BERT pre-training with online tokenizer. In: *ICLR* (2022) [3](#)
102. Zong, Y., Mac Aodha, O., Hospedales, T.: Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008* (2023) [1](#)