OmniSat: Self-Supervised Modality Fusion for Earth Observation

Guillaume Astruc[1, 2, 3] Loic Landrieu[1, 2] Nicolas Gonthier[2] Clement Mallet[2]
[1]LIGM, ENPC [2]LASTIG, Univ Gustave Eiffel, IGN, ENSG [3]CNES

1- Motivation

Context: Earth Observation modalities/sensors are:

- **Diverse** and capturing **complementary data characteristics**
- Available in **large quantity** • Spatially **aligned**

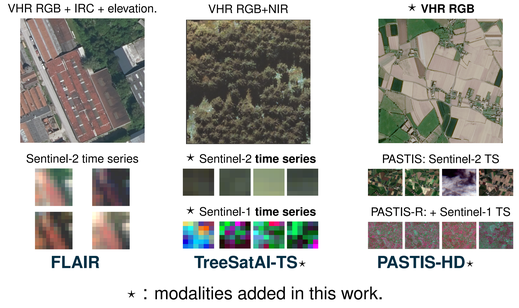
Intuition:

- All observations share a variable: the **latent semantics**
- Unimodal representations should be **spatially consistent**
- Multimodal representations should keep **sensor-specific information**

2- Datasets

Dataset	Modalities: images (single date)	Modalities: time series	Labels
SpaceNet6	SAR+optical (0.5m-2m)	✗	building footprint (<1m)
TreeSatAI	aerial + S1/S2 (0.2-10m)	✗	forestry (60m)
BigEarthNet	S1/S2 (10m)	✗	land cover (100m)
MDAS	S1/S2 + hyperspectral (2.2-10m)	✗	land cover (0.25m)
DOFA	NAIP + Gaofen + S1/S2 + EnMAP (1-30m)	✗	✗
PASTIS-R	✗	S1/S2 (30-140 / year)	agriculture (10m)
SSL4EO-S12	✗	S1/S2 (4 / year)	✗
DFC21-DSE	✗	S1/S2 + LS8 (3-9/year)	human activity (500m)
MultiSenGE	✗	S1/S2 (30-140 / years)	land cover (10m)
FLAIR	aerial (0.2m)	S2 (20-114 / year)	land cover (0.2m)
Satlas	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)

Current datasets are either unimodal or monotemporal



⋆ : modalities added in this work.

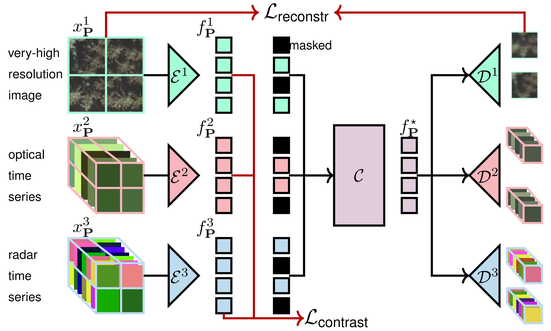
3- OmniSat Model

Method Objectives:

- Handle multiple modalities of different nature and spatial/temporal/spectral resolutions
- Learn rich **multimodal representations** without labels
- Use **multimodal alignment** as (weak) supervision

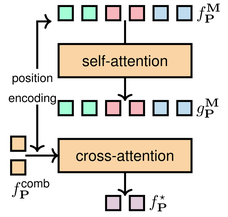
Contrastive loss:
Unimodal representations are spatially consistent

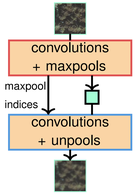
Reconstruction loss:
Omnimodal representations can reconstruct all inputs



4- Architecture Details

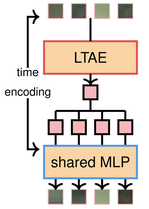
Modality Combining Network $\mathcal{C}$: Self-attention between patches of different modalities + cross-attention with omnimodal tokens (one per patch)



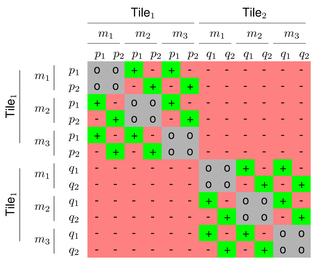


High-Resolution Images Decoders $\mathcal{D}^1$: Pass maxpooling indices from encoder to decoder ⇒ no need to learn intra-patch geometry.

Temporal Decoders $\mathcal{D}^2, \mathcal{D}^3$: Only reconstruct acquisitions with the most temporal attention in the encoder (top 25%) ⇒ ignore noisy images, *e.g.* cloud cover.



5- Contrastive Loss

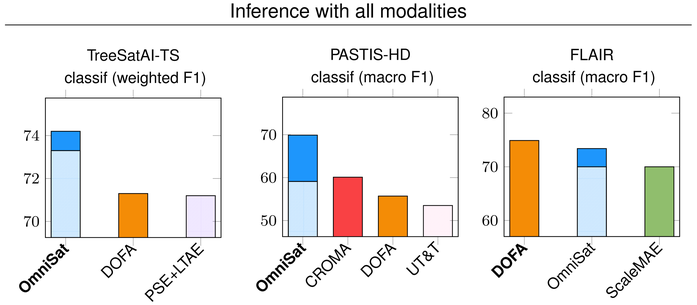


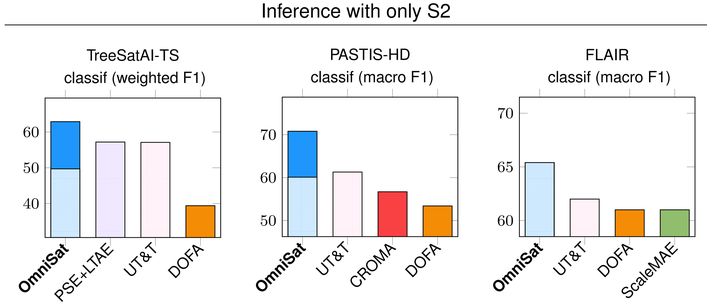
Tiles: Tile$_1$, Tile$_2$
Patches: Tile$_1$: p_1, p_2
Tile$_2$: q_1, q_2
Modalities: m_1, m_2, m_3

- **Positive match**: same patch, different modality
- **Negative match**: different patch, different modality
- **Ignored**: same tile, same modality

6- Results

- State-of-the-art on 2 EO datasets
- Diverse modalities lead to **richer representations**
- Multimodal self-supervized pre-training improves downstream task performance
- Pre-training helps **even when only one modality is available** for inference





trained from scratch

pre-trained and fine-tuned

7- Efficiency (TreeSatAI-TS)