Supplementary Material: OmniSat: Self-Supervised Modality Fusion for Earth Observation

Guillaume Astruc [1,4,3], Nicolas Gonthier [1,2], Clement Mallet [1], 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

In this appendix, we present an extended ablation study in Section A-1, details our competing methods in Section A-2, provide the classwise performance in Section A-3, and provide qualitative illustrations in Figure A-2.

A-1 Supplementary Ablations

We propose supplementary ablations to evaluate the impact of several design choices.

Table A-1: Supplementary Ablation. Performance (weighted F1) on TreeSatAI-TS of alternate VHR encoders (**a**-**c**) and masking schemes (**d**-**e**).

Experiment	All	VHR	S2	S1
Default	**74.2**	70.5	62.9	**56.7**
a linear w. random init	66.8	57.3	58.9	54.8
b ViT	70.5	**70.8**	64.0	52.6
c linear from ScaleMAE	68.9	51.2	**66.7**	52.2
d Spatial masking	73.2	70.1	63.2	54.6
e Modality masking	72.4	70.2	61.2	55.4

Alternate VHR Encoder. To train OmniSat on both VHR (0.2 m) and Sentinel (10m) images, we must embed patches of 50×50 pixels. We consider here alternative encoders to CNNs: a linear layer (Tab. A-1.**a**) and a ViT with 10×10 patches (Tab. A-1.**b**). The results suggest that 50×50 patches are too large to use linear projection. While ViTs reach slightly higher unimodal performance, CNNs allow us to bypass maxpool indices to the decoders leading to higher multimodal performances.

Using Pre-trained VHR models. Rescaling the 50×50 patches to the 224×224 resolution of ScaleMAE or SatMAE proved impractical in terms of memory. Instead, we use the pre-trained patch encoder of ScaleMAE by rescaling our 50×50 patches to 16×16, removing the infrared channel, and adding a projection layer to our token size $D = 256$ (Tab. A-1.**c**). Interestingly, this leads to a cross-modal distillation which improves the results for S2. The VHR and multimodal performance remain below OmniSat, which can be attributed to the lack of a NIR channel.

Masking Strategies. We report the results for spatially consistent masking (patches are masked for all modalities simultaneously, Tab. A-1.**d**) and modality masking (the patches of a random modality are all masked, Tab. A-1.**e**). Our random masking strategy performs better.

Relative *vs*. Absolute Positional Encoding. We evaluate the impact of replacing the relative positional encoding of tokens, based on the patch position, with an absolute position encoding, based on the position of the patches in their tile—similar to what is classically done for image processing.

With an absolute positional encoding, OmniSat reaches an F1-score of 58.4 and 73.0 when fine-tuned with 10% and 100% of the training set of TreeSatAI-TS, respectively. This is 2.7 and 1.2% below a model trained with relative positional encodings. We conclude that relative positional encodings are better suited for analyzing EO images. While the upper patches of natural images are bound to correspond to the sky, and the lower patches contain ground, no such analogy can be made for EO data, whose distribution is equivariant through small horizontal translation.

Impact of Pre-training on Monomodal Performance. We aim to determine how our multimodal pre-training scheme improves the monomodal performance (*e.g.*, +13.2% for Sentinel-2 in full supervision). We consider two mechanisms that may lead to more discriminative features: (i) multimodality allows us to train the modality combiner network $\mathcal{C}$ with more data, or (ii) our cross-modal and token-wise alignment-based losses provide a strong supervisory signal. We propose an experiment to verify which mechanism is the leading reason of our scheme's strong performance.

We pre-train OmniSat on TreeSatAI-TS in mono- and multimodal settings *with a constant amount of tokens*. More precisely, we pre-train OmniSat using *all* input tokens from the S2 modality *only*, and using *all* 3 modalities but only 33% of patches. This means that each experiment considering the same number P of input tokens. We then train a single linear layer to map these representations to class scores (linear probing) using 10 and 100% of the annotated S2 data. Finally, we evaluate the quality of these linear mappings on the test set using only the S2 modality.

The model trained with a multimodal pretext task reaches a F1-score of 44.7 for 10% and 46.3 for 100% of the training data. The model trained only with S2 performs significantly worse: 26.9 for 10% and 29.8 for 100% of data. This result suggests that the key to the efficacy of our pretraining scheme is the supervisory signal of per-patch contrastive and reconstruction objectives, rather than just increasing the number of tokens viewed by the transformer backbone.

A-2 Adapting Competing Methods

We adapt competing methods to allow them to handle single images and time series at different resolutions. We performed multiple tests for each approach and kept the configurations leading to the competing approach' highest performance.

– **Multimodality.** We train methods that are not natively multimodal (PSE [5], ViT [2], DOFA [9], SatMAE, ScaleMAE) using a late-fusion scheme [6] by concatenating

the embeddings learned in each modality, as suggested by Ahlswede *et al*. [1]. For UT&T [3], initially designed for VHR images and Sentinel-2 time series, we add a branch for Sentinel-1 integration, which is identical to the Sentinel-2 branch except for the first layer.

- **Handling Temporal Data.** To evaluate image models (SatMAE, ScaleMAE, CROMA) on time series, we convert image sequences to single images by concatenating for each pixel and channel channel-wise the median observation for the four seasons: spring, summer, fall, and winter [7].
- **Handling VHR Data.** To evaluate methods designed for low-resolution images (PSE, LTAE [4]) in a multimodal setting that includes VHR images, we concatenate their final embedding to the the one of a ResNet network.
- **Scaling Models.** The considered datasets are smaller than the ones typically used to train large ViT-based models, making them prone to overfitting. To address this issue we select a ViT-Small [2] backbone for SatMAE, ScaleMAE and CROMA. For DOFA, we use a ViT-Base, the smallest pretrained model available.
- **Multi-Class Prediction.** To evaluate ViT-based models on classification experiments, we insert a linear layer that maps the embedding of the class token $\langle \texttt{CLS} \rangle$ to a vector of class scores. For the UT&T model, we compute a spatial average of the last feature map, followed by a similar linear projection.

A-3 Supplementary Results

We report the performance of different approaches for each class for the two datasets graphically in Figure A-1 and as a table in Table A-2. OmniSat is able to parse complex scenes including mixed forest, cultures, and complex urban areas. In particular, Omnisat leverage temporal dynamics to distinguish between different vegetation species.

Failure Case. We report in the bottom half of Figure A-2 hard examples from our three datasets and compare the prediction of OmniSat and other models. For the TreeSatAI-TS example, the Sentinel-2 optical time-series is highly occluded: over 80% of acquisitions are covered by clouds. Furthermore, the forest tile contains a large variety of tree species organized in densely connected canopy, making its classification particularly hard. Indeed, the texture of the images in closed forests does not bring additional discriminative information.

The example from FLAIR is a scrap yard, which is almost entirely covered by broken vehicles. Since FLAIR's annotations focus on the ground rather than transient or stationary objects, identifying the actual land cover in such scenarios is very challenging.

The image taken from PASTIS contains a mix of several different crop types, including the class *mixed cereal* which can already correspond to a parcel with various cereal types. This leads to a hard classification problem for all methods.

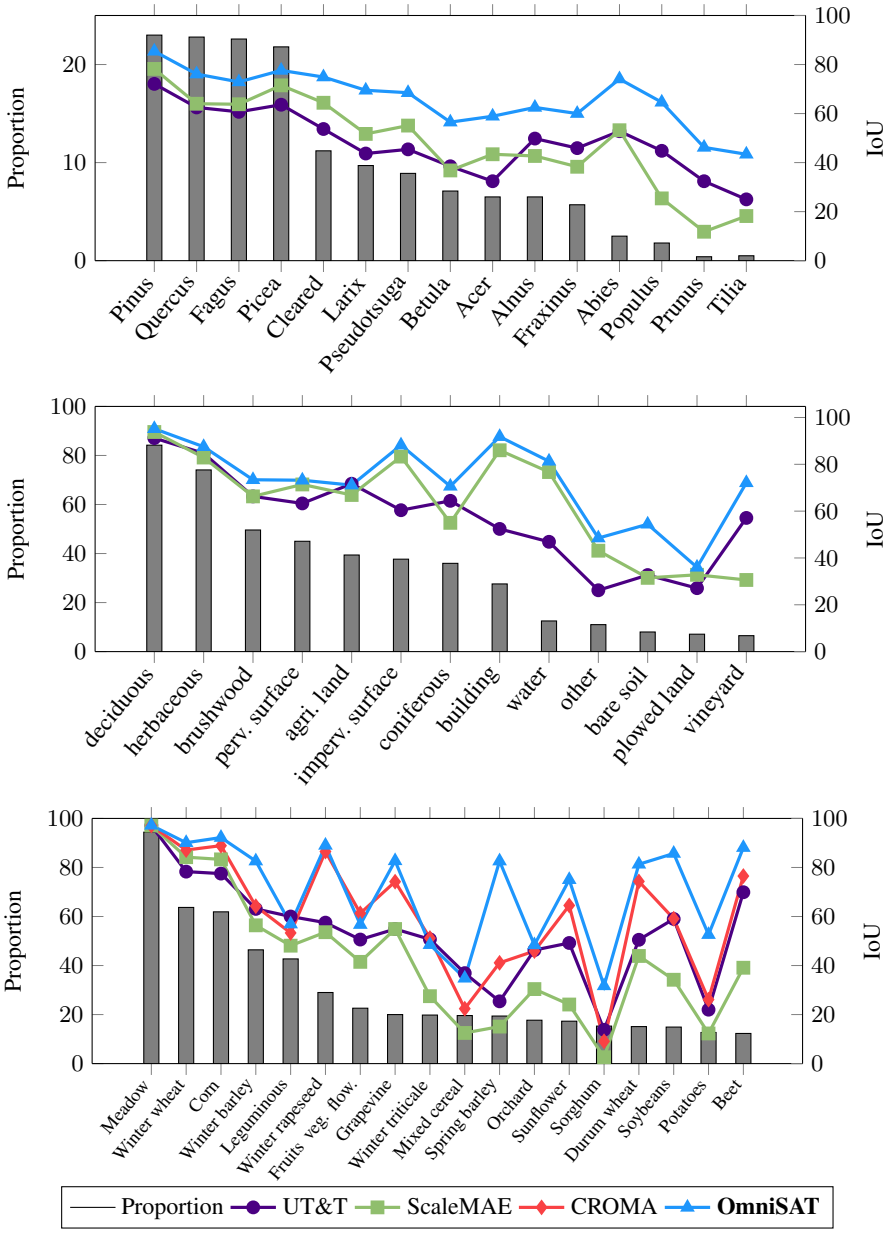


Fig. A-1: Class-Wise Performance. We plot the performance of different models for each class, sorted by decreasing frequency. OmniSat improves the performance across the board, and for rare classes in particular.

TreeSatAI-TS

Method	Macro F1	Abies	Acer	Alnus	Betula	Cleared	Fagus	Fraxinus	Larix	Picea	Pinus	Populus	Prunus	Pseudotsuga	Quercus	Tilia
Proportion in %		2.5	6.5	6.5	7.1	11.2	22.6	5.7	9.7	21.8	23.0	1.8	0.8	8.9	22.8	0.5
UT&T	48.8	52.8	43.4	49.8	36.8	53.7	60.7	45.9	43.7	63.6	72.1	44.8	32.4	45.4	62.5	25.0
Scale-MAE	47.3	53.2	32.4	42.7	38.5	64.4	63.8	38.3	51.7	71.5	78.2	25.4	11.8	55.1	64.0	18.2
OmniSat	**73.4**	**74.1**	**58.9**	**62.5**	**56.5**	**74.9**	**72.9**	**60.0**	**69.5**	**77.6**	**85.4**	**64.6**	**46.2**	**68.5**	**76.1**	**43.4**

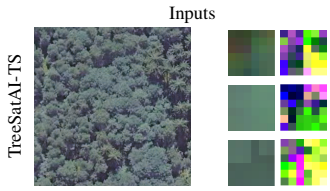
FLAIR

Method	Macro F1	building	perv. surface	imperv. surface	bare soil	water	coniferous	deciduous	brushwood	vineyard	herbaceous	agri. land	plowed land	other
Proportion in %		27.6	45.0	37.7	8.0	12.5	36.0	84.2	49.6	6.5	74.1	39.4	7.1	11.0
UT&T	57.3	52.4	63.3	60.4	32.7	46.9	64.4	91.2	66.3	57.1	84.7	**71.7**	27.1	26.2
Scale-MAE	70.0	90.1	72.0	87.1	47.1	**81.3**	65.1	**95.2**	72.3	53.7	**88.8**	70.2	**39.7**	45.7
OmniSat	**75.8**	**91.8**	**73.2**	**88.2**	**54.4**	**81.3**	**70.6**	**95.2**	**73.4**	**72.1**	87.5	71.1	36.0	**48.5**

PASTIS-HD

Method	Macro F1	Meadow	Soft winter wheat	Corn	Winter barley	Winter rapeseed	Spring barley	Sunflower	Grapevine	Beet	Winter triticale	Winter durum wheat	Fruits, vegetables, flowers	Potatoes	Leguminous fodder	Soybeans	Orchard	Mixed cereal	Sorghum
Proportion in %		94.4	63.7	61.9	46.4	29.0	19.4	17.3	20.0	12.3	19.8	15.1	22.6	12.7	42.7	14.9	17.7	19.6	15.3
UT&T	53.5	97.1	78.3	77.5	63.1	57.5	25.4	49.2	54.9	69.9	50.7	50.5	50.6	22.0	**60.0**	59.0	46.3	**36.9**	13.8
CROMA	60.1	97.1	87.1	88.9	64.2	86.5	41.1	64.5	74.2	76.5	**51.2**	74.3	**61.3**	26.1	53.3	59.1	45.9	22.4	9.1
Scale-MAE	42.2	97.1	84.2	83.3	56.4	53.6	15.1	24.1	54.9	39.1	27.5	43.9	41.5	12.3	48.1	34.2	30.4	12.5	2.6
OmniSat	**69.9**	**97.2**	**90.1**	**92.2**	**82.6**	**89.0**	**64.7**	**75.0**	**82.7**	**88.1**	48.5	**81.3**	56.7	**52.6**	56.9	**85.7**	**48.4**	34.8	**31.7**

Table A-2: Class-Wise Performance. We report the F1-score for each class for TreeSatAI-TS, FLAIR, and PASTIS-HD for multilabel classification. We also report the unweighted class-averaged F1-score (Macro-F1). We can observe that OmniSat outperforms UT&T [3] and Scale-MAE [8] on nearly all classes for both datasets. In particular, we observe the most significant gains for classes with discriminative temporal dynamics, such as broadleaf tree species and the vineyards class.



	Inputs	Ground truth	**OmniSat**	UT&T [3]	Scale-MAE [8]
TreeSatAI-TS		- Picea - Betula - Alnus - Quercus	- Picea - Betula - Alnus - ✗	- Picea - Betula - Alnus - ✗ - Pinus	- Picea - ✗ - ✗ - ✗
FLAIR		- building - pervious surf. - impervious surf. - deciduous - brushwood - herbaceous - agricultural - vineyard	- building - pervious surf. - impervious surf. - deciduous - brushwood - herbaceous - agricultural - vineyard	- building - pervious surf. - impervious surf. - deciduous - brushwood - herbaceous - agricultural - ✗	- building - pervious surf. - impervious surf. - deciduous - brushwood - herbaceous - ✗ - ✗
PASTIS-HD		- Meadow - Soft winter wheat - Corn - Winter rapeseed - Beet	- Meadow - Soft winter wheat - Corn - Winter rapeseed - Beet	- Meadow - ✗ - ✗ - ✗ - ✗ - Potatoes	- Meadow - ✗ - ✗ - ✗ - ✗ - Sunflower - Grapevine
TreeSatAI-TS		- Quercus - Acer - Alnus - Larix	- Quercus - ✗ - ✗ - ✗ - Abies	- ✗ - ✗ - ✗ - ✗ - Abies - Betula	- ✗ - ✗ - ✗ - ✗ - Picea
FLAIR		- decideous - herbaceous - water - pervious surf. - bare soil	- decideous - herbaceous - water - ✗ - ✗ - building - imperv. surf. - brushwood	- decideous - herbaceous - ✗ - pervious surf. - ✗ - building - imperv. surf. - brushwood - coniferous	- decideous - herbaceous - ✗ - pervious surf. - ✗ - building - imperv. surf. - brushwood - coniferous - other
PASTIS-HD		- Meadow - Winter wheat - Corn - Potatoes - Mixed cereal	- Meadow - Winter wheat - Corn - Potatoes - ✗ - Winter barley - Wint. rapeseed - Legum. fodder	- Meadow - ✗ - ✗ - ✗ - ✗ - Spring barley - Orchard - Legum. fodder - Durum wheat - Fruits, veg..	- Meadow - Winter wheat - Corn - ✗ - ✗ - Winter barley - Wint. rapeseed - Winter triticale

Fig. A-2: Qualitative Results. We report predictions of OmniSat and two competing models on tiles from our datasets, including a failure case (bottom). OmniSat can detect classes with recognizable temporal dynamics such as agricultural lands or mixed forest areas with both coniferous and deciduous trees. Other methods, and in particular ScaleMAE, struggle to detect these classes.

References

1. 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) 3
2. 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, 3
3. 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, 5, 6
4. 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) 3
5. 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) 2
6. Hong, D., Gao, L., Yokoya, N., Yao, J., Chanussot, J., Du, Q., Zhang, B.: More diverse means better: Multimodal deep learning meets remote-sensing imagery classification. IEEE TGRS (2020) 2
7. Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A.: Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters (2017) 3
8. 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) 5, 6
9. 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) 2