AnySat: An Earth Observation Model for Any Resolutions, Scales, and Modalities

Supplementary Material

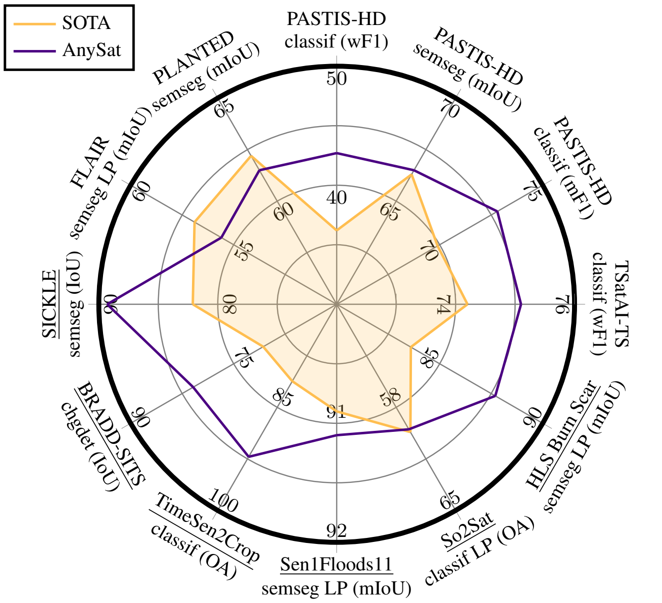

Figure A. **Overall Performance.** We underline external datasets. LP stands for Linear Probing.

In this appendix, we provide detailed results in Sec. A, an extended ablation study in Sec. B, and provide implementation details in Sec. C. Finally, we provide more details on the datasets and experiments of the main paper in Sec. D

A. Detailed Results

We provide qualitative illustrations of our predictions and detailed quantitative results for the test sets of GeoPlex.

Qualitative Results. We present qualitative illustrations in Fig. B for four segmentation tasks: PASTIS, FLAIR, SICKLE, and BraDD-S1TS. AnySat predicts precise segmentations that closely follow the extents of buildings, trees, and parcels. Notably, the predictions do not display grid artifacts despite our segmentation head being a simple linear layer applied to each subpatch. This suggests that using subpatches of small sizes (*e.g.*, 4×4 pixels for PASTIS and 10×10 pixels for FLAIR), combined with larger context through patch embeddings, is an effective strategy for producing smooth and consistent segmentation maps.

Quantitative Results. We provide in Tab. A and Tab. B the detailed performance of AnySat, with and without pretraining, and an extensive comparison with recent EO models. Pretraining on GeoPlex improves performance for smaller datasets (*e.g.*, TreeSatAI-TS, PASTIS in classification), but this effect is more limited for segmentation datasets (FLAIR, PASTIS in segmentation) or larger ones like PLANTED. We hypothesize that this is due to the quantity of available supervision; for instance, FLAIR has over 20 billion individual labels. In the case of FLAIR, the pretrained model is 0.5 points behind training from scratch, which we attribute to stochastic noise, as our performance on the validation set is on par with training from scratch: 54.7 for pretrained *vs.* 54.8 from scratch.

B. Additional Ablation

We propose an additional experiment to evaluate the impact of one of our design choices.

No Modality or Temporal Masking. In this experiment, we remove the modality and temporal masking for the student encoder during pretraining. This modification results in a slight increase in segmentation performance by $+0.4$ mIoU but a decrease in classification performance by -0.6 F1 score. These ambiguous results are similar to the effects we observed with naive patch dropping. An advantage of including modality and temporal masking is that it reduces the memory requirements during training by up to 30%. Since our goal is to train a single model on several datasets aimed to be fine-tuned for multiple tasks, we keep a unique configuration and adopt this masking strategy.

C. Implementation Details

GeoPlex. See Tab. C for more details on the composition of GeoPlex. GeoPlex is composed of five distinct datasets—TSAI-TS, PASTIS-HD, FLAIR, PLANTED, and S2NAIP-URBAN—which collectively offer a rich combination of data types, including images, time series, and various modalities. These datasets span extensive geographical areas, ranging from 180 km² to over 211,000 km², and provide a wide array of spatial resolutions (from 0.2m to 250m), temporal resolutions (from 1 to 140 time steps), and spectral resolutions (from 3 to 10 bands). The inclusion of multiple satellite and aerial platforms, such as Sentinel-1/2, Landsat 7/8/9, SPOT6/7, and NAIP, ensures a robust and varied training set.

Network Architecture. AnySat's architecture follows the Vision Transformer (ViT) template and has 125M learnable parameters, of which 73.6% are modality-agnostic and resolution-adaptive. The components of the model are:

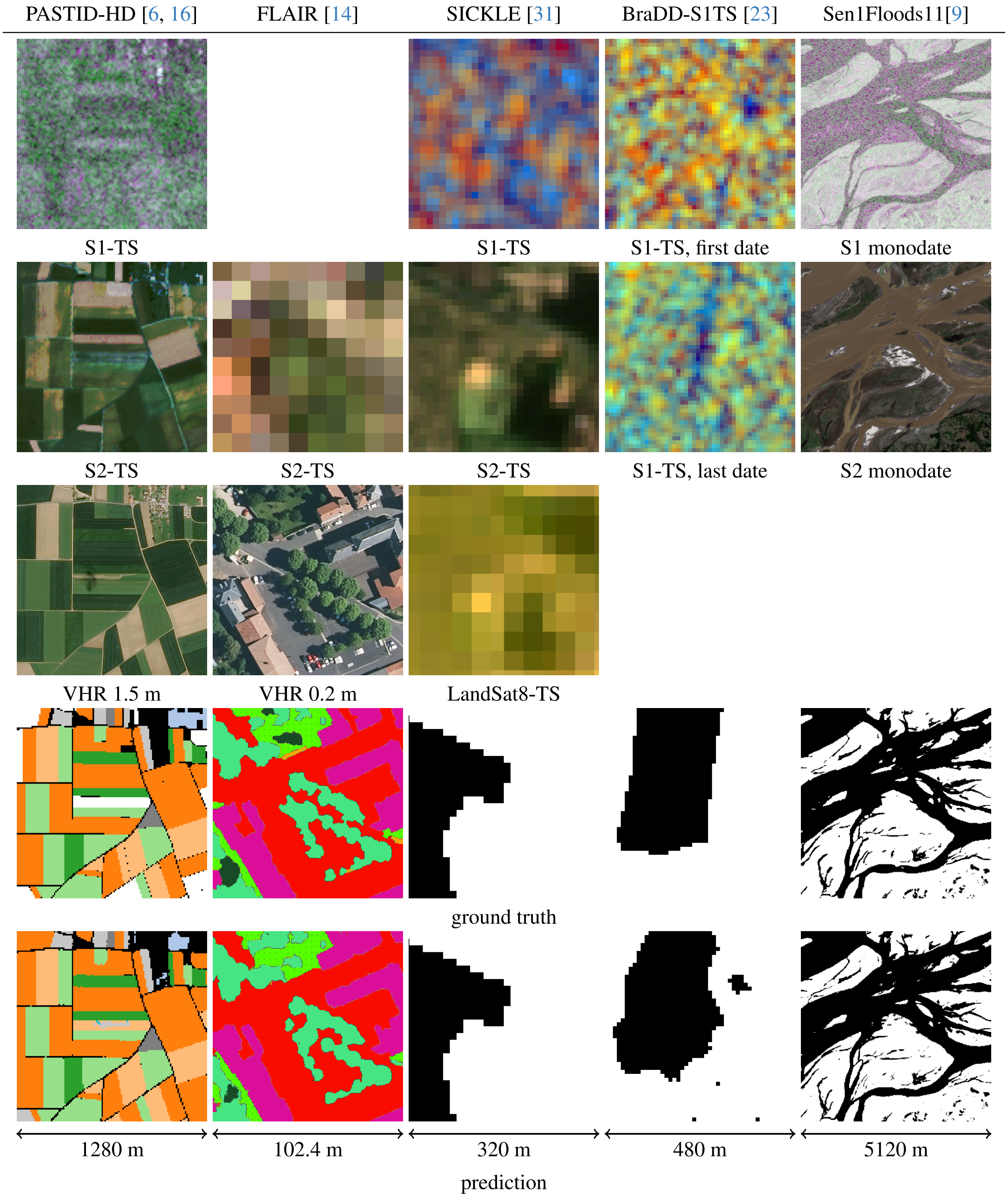

Figure B. **Illustration of Results.** We represent the inputs, predictions, and ground truth for tiles from four datasets. The colormaps are taken directly from the papers. TS: time series, a single date has been chosen. S1/2 stands for Sentinel-1/2. For PASTIS-HD, white parcels are not annotated (void label).

Table A. **Model Performance on the Test Sets of GeoPlex.** For time series, we denote by 📅 when a single date has been selected, and 🗓 when seasonal medians have been concatenated in the channel dimension. AL stands for ALOS-2 and MO for MODIS. LP stands for linear probing

Model	Pre-training	Modalities			
TSAI-TS - multilabel classif.		VHR	S1	S2	wF1
AnySat (ours)	GeoPlex	✓	✓	✓	**75.1**
AnySat (ours)	None	✓	✓	✓	72.7
OmniSat [6]	TSAI-TS	✓	✓	✓	74.2
DOFA [44]	DOFA	✓	📅	📅	71.6
PSE+LTAE [15]	None		✓	✓	71.2
PSE + ResNet [6]	None	✓	📅	📅	68.1
ScaleMAE [29]	TSAI	✓		📅	62.5
SatMAE [11]	TSAI	✓		📅	61.5
CROMA [13]	TSAI	✓		📅	61.0
UT&T [14]	ImageNet	✓	✓	✓	56.7
MOSAIKS[30]	TSAI			📅	56.0
PRESTO [38]	PRESTO			📅	46.3

Model	Pre-training	Modalities					
PLANTED - classif.		S1	S2	LS	AL	MO	maF1
AnySat (ours)	GeoPlex	✓	✓	✓	✓	✓	61.5
AnySat (ours)	None	✓	✓	✓	✓	✓	61.2
ViViT [5, 27]	None	✓	✓				**62.2**
ViViT [5, 27]	None	✓	✓	✓	✓	✓	59.3

FLAIR - semantic seg		VHR	S2	mIoU
AnySat (ours)	GeoPlex	✓	✓	55.1
AnySat (ours)	None	✓	✓	55.6
UT&T [14]	ImageNet	✓	✓	**56.9**
UNet [19]	ImageNet	✓		54.7
UTAE [16]	None		✓	36.1

PASTIS-HD - multilabel classif.		VHR	S1	S2	maF1
AnySat (ours)	GeoPlex	✓	✓	✓	**72.8**
AnySat (ours)	None	✓	✓	✓	65.5
OmniSat [6]	PASTIS-HD	✓	✓	✓	69.9
CROMA [13]	PASTIS-HD		🗓	🗓	60.1
DOFA [44]	DOFA	✓	🗓	🗓	55.7
UT&T [14]	ImageNet	✓	✓	✓	53.5
UTAE [16]	None		✓	✓	46.9
ScaleMAE [29]	PASTIS-HD	✓		🗓	42.2

PASTIS-HD - semantic seg		VHR	S1	S2	OA	mIoU
AnySat (ours)	GeoPlex	✓	✓	✓	85.0	**66.5**
AnySat (ours)	None	✓	✓	✓	84.8	66.3
SkySense [18]	SkySense	✓	✓	✓	**85.9**	-
UTAE-MM [17]	None		✓	✓	84.2	66.3
TSViT [37]	None			✓	83.4	65.4
UTAE [16]	None			✓	-	63.1

PASTIS-HD - semseg LP		VHR	S1	S2	mIoU
AnySat LP (ours)	GeoPlex	✓	✓	✓	**42.7**
S12-DINO LP [26, 34]	foundation	✓	✓	✓	36.2
S12-MoCo LP [20, 34]	foundation	✓	✓	✓	34.5
S12-D2V LP [7, 34]	foundation	✓	✓	✓	34.3
SpectralGPT [21]	foundation	✓	✓	✓	35.4
Prithvi [22]	foundation	✓	✓	✓	33.9

- **Modality Projectors ϕ_m^{proj} (33M parameters for** 11 **projectors).** These modules are MLPs responsible for projecting the input data of each modality into a common feature space.
- **Spatial Transformer ϕ^{trans} (45M parameters).** Composed of three self-attention transformer blocks, this module captures the spatial relationships between subpatches for each modality and patch.
- **Modality Combiner ϕ^{comb} (49M parameters).** This module consists of three self-attention blocks followed by a cross-attention block, and merges the representations from different modalities into a unified feature vector for each patch.
- **Predictor ϕ^{pred} (29M parameters).** Exclusive to the student, this module is a single self-attention block and predicts the teacher's embeddings for the dropped patches.

Handling MODIS data. In the Planted dataset [27], MODIS observations are included, but their resolution (250 meters) is larger than the entire observed tile (120 meters). We treat these observations as *context* tokens: we concatenate their ϕ^{patch} embeddings to the $|\mathbf{M}| \cdot (S/P)^2$ tokens from all other modalities. We do not add positional encoding, and this token is not included in the contrastive loss.

Optimization Parameters. To better manage our memory usage, we adapt the batch size to the size of the samples of each dataset: TreesatAI-TD: 384, PASTIS-HD: 8, FLAIR: 96, PLANTED: 2048, S2NAIP: 16. We use 8 NVIDIA H100 for experiments on GeoPlex, PLANTED and Pastis-HD , and a smaller cluster of 3 A600 for TreeSatAI-TS and FLAIR.

Beyond the changes above, all optimization parameters are shared across all datasets. We used the AdamW [24] optimizer with a learning rate of 5×10^{-5} for all our experiments (pretraining and fine-tuning). We used a `LinearWarmupCosineAnnealingLR` [1] for classification and `ReduceLROnPlateau` [2] scheduler for pretraining and segmentation.

We set he contrastive temperature γ to 0.1 to n Eq. X. We

Table B. **External Datasets.** We evaluate our pretrained model on 4 external datasets, in the fine-tuning or linear probing settings. 📅 stands for single-date observations. We report the number of trainable parameters for probing experiments.

SICKLE [31]	L8	S1	S2	mIoU
AnySat (fine-tune)	✓	✓	✓	**89.3**
AnySat (linear 6.1K)	✓	✓	✓	82.0
Unet3d [25, 31]	✓	✓	✓	82.1
UTAE [16, 31]	✓	✓	✓	51.4
BraDD-S1TS [23]		S1		mIoU
AnySat (fine-tune)		✓		**80.9**
AnySat (linear 6.1K)		✓		78.9
UTAE [16]		✓		70.7
3D-UNet [25]		✓		68.1
Conv-LSTM [33]		✓		63.7
TimeSen2Crop [42]			S2	OA
AnySat (fine-tune)			✓	**92.2**
AnySat (linear 14K)			✓	70.3
OS-CNN [36, 41]			✓	81.2
MLP+TAE [15, 40]			✓	80.9
W.LSTM [10, 32]			✓	78.2
Transformer [39]			✓	78.1
MSResNet [12]			✓	76.3
Sen1Floods11 [9]		S1	S2	mIoU
AnySat (linear 6.1K)		📅	📅	**91.1**
CROMA [13] (UperNet 47M)		📅	📅	90.9
CROMA [13] (fine-tune 350M)		📅	📅	90.9
Prithvi [22] (fine-tune 130M)		📅	📅	90.4
Prithvi [22] (UperNet 39M)		📅	📅	88.3
Prithvi2 [35] (fine-tune 630M)		📅	📅	90.4
SatlasNet [8] (UperNet 33M)		📅	📅	90.3
HLS Burn Scar [28]		HLS		mIoU
AnySat (fine-tune)		✓		**90.6**
AnySat (linear 3M)		✓		87.7
Prithvi2 [35] (fine-tune 630M)		✓		90.5
Prithvi [22] (fine-tune 130M)		✓		86.9
Prithvi [22] (UperNet 39M)		✓		83.6
CROMA [13] (UperNet 47M)		✓		82.4
DOFA [44] (UperNet 39M)		✓		80.6
So2Sat [45]		S1	S2	OA
AnySat (linear 29k)		📅	📅	59.1
DOFA [44] (linear)		📅	📅	**59.3**
CROMA [13] (linear)		📅	📅	49.2
SatMAE [11] (linear)		📅	📅	46.9

used an EMA decay of 0.996. All other hyperparameters are shared with original JEPA implementation.

Position Encodings. We describe here our scale-adaptive positional encoding which allows us to use the same encoders for different resolutions, scales, and patch size. The input tokens to the modality combiner ϕ^{comb} correspond to patches of size $P \times P$ meters, while those to the spatial transformer ϕ^{trans} represent subpatches of size $(R_m \delta_m) \times (R_m \delta_m)$ meters. Here, R_m varies per sensor modality m, and P is randomly chosen for each batch during training. To train a single scale-aware model capable of handling varying resolutions, we employ a scale-adaptive positional encoding inspired by Scale-MAE [29].

We use the same positional encodings in ϕ^{comb} and ϕ^{trans}. We first describe the positional encoding of a token by ϕ^{comb}. We denote by pos_x the index of the token's patch within its tile along the x-axis; similarly, pos_y along the y-axis. If the embeddings of the token have a dimension D, the positional encodings $\mu_x(\text{pos}_x, i)$ and equivalently $\mu_y(\text{pos}_y, i)$ are of size $D/2$. For $i \in [0, D/2[$ we have:

$$\mu_x(\text{pos}_x, i) = \sin\left(\frac{g}{G}\frac{\text{pos}_x}{10000^{\frac{i}{E}}} + \frac{\pi}{2}\text{mod}(i, 2)\right) , \quad \text{(A)}$$

where $g = P$ is the size in meter of the patch considered unit: patch of size for ϕ^{comb}, and G is a reference length that we set to one meter. We compute $\mu_y(\text{pos}_y, i)$ similarly, and the positional encoding is the channelwise concatenation of both vectors. The positional encoding is directly added to the embeddings.

For ϕ^{trans}, we define the positional encoding of each subpatch within its patch with the same formula, but set g to $g = R_m \delta_m$, the size of the subpatch in meter.

D. Datasets and Tasks

Here, we provide more details about the datasets used to train and evaluate AnySat and their associated tasks. See Tab. C for an overview of the datasets used in GeoPlex.

TreeSatAI-TS [3, 6]: This multimodal dataset is designed for tree species identification and consists of 50,381 tiles, each covering an area of 60×60 meters, with multi-label annotations across 20 classes. All data were collected in Germany. The dataset includes Very High Resolution (VHR) images at 0.2 m with a NIR band, Sentinel-2 time series, and Sentinel-1 time series.

PASTIS-HD [6, 17]: This crop mapping dataset supports classification, semantic segmentation, and panoptic segmentation. Each agricultural parcel is delineated at a resolution of 10 m and annotated across 18 crop types. The dataset contains 2,433 tiles with an extent of 1,280×1,280 m, including Sentinel-2 time series, Sentinel-1 time series (we use only the ascending orbit), and SPOT6 VHR imagery at 1.5 m resolution.

Table C. **Considered Datasets.** We present the detailed composition of GeoPlex, the collection of datasets used for self-supervised training, and our external evaluation datasets. For each dataset, we consider a set of acceptable patch sizes.
img: img, **t.s.**: time series: t.s. S1/2: Sentinel-1/2. † upsampled from original acquisition resolution.

Dataset	Extent	Sample Size (S)	Modalities	Resolution		
		Patch Size (P)		Spatial (R)	Temporal (T)	Spectral (C)
		GeoPlex				
TSAI-TS [3, 6]	50k × (1 img + 2 t.s.) 180 km² - 4.7 GPix	$S = 60$m $P \in \{10, 20, 30\}$m	Aerial VHR S1 S2	0.20m 10m 10m	1 10-70 10-70	4 3 10
PASTIS-HD [6, 16]	2433 × (1 img + 2 t.s.) 3986 km² - 7.5 GPix	$S = 1280$m $P \in \{40, 80, 160\}$m	SPOT6/7 S1 S2	1m† 10m 10m	1 140 38-61	4 3 10
FLAIR [14]	78k × (1 img + 1 t.s.) 815 km² - 20 GPix	$S = 102.4$m $P \in \{10, 20, 50\}$m	Aerial VHR S2	0.2m 10m	1 20-114	5 10
Planted [27]	1.3M × (5 t.s.) 33,120 km² - 3.0 GPix	$S = 120$m $P \in \{30, 60\}$m	S2 S1 Landsat 7 ALOS-2 MODIS	10m 10m 30m 30m 250m	8 8 20 4 60	10 3 3 3 7
S2NAIP-URBAN [4, 43]	515k × (1 img + 3 t.s.) 211,063 km² - 136 GPix	$S = 640$m $P \in \{40, 80, 160\}$m	NAIP S2 S1 Landsat 8/9	1.25m 10m 10m 10m†	1 16-32 2-8 4	4 10 3 8
		External datasets				
BraDD-S1TS [23]	13k × (1 t.s.) 2,995 km² - 1.2 GPix	$S = 480$m $P = 10$ m	S1	10m	20-66	10
Sickle [31]	35k × (2 t.s.) 3,584 km² - 3.6 GPix	$S = 320$m $P = 10$m	S2 Landsat 8/9	10m 10m†	13-148 8-34	10 8
TimeSen2Crop [42]	1.2M × (1 t.s.) 120 km² - 35 MPix	$S = 10$m $P = 10$m	S2	10m	29	10
Sen1floods11 [9]	4.8k × (2 img) 125,829 km² - 2.6 GPix	$S = 5120$m $P = 80$m	S2 S1	10m 10m	1 1	10 3
So2Sat [45]	400k × (2 img) 41,029 km² - 82 GPix	$S = 320$m $P = 10$m	S2 S1	10m 10m	1 1	10 3
HLS Burn Scar [28]	804 × (1 t.s.) 188,208 km² - 211 MPix	$S = 15300$m $P = 240$m	HLS	30m	1	6

FLAIR [14]: This dataset combines VHR aerial imagery at a 0.2 m resolution with Sentinel-2 time series data and comprises 77,762 tiles acquired across metropolitan France. The VHR images include five channels: RGB, near-infrared, and a normalized digital surface model derived by photogrammetry. Each VHR pixel is annotated with one of 13 land cover classes.

PLANTED [27]: The PLANTED dataset is specifically designed for tree species identification and features 1,346,662 tiles of planted forest across the world. Each tile is associated with one of 40 distinct classes. This dataset integrates imagery from five different satellites with various resolutions: Sentinel-2 (10 m), Landsat-7 (30 m), MODIS (250 m), as well as radar time series from Sentinel-1 (10 m) and ALOS-2 (30 m). The time series are temporally aggregated at various intervals—seasonally, monthly, or yearly.

S2Naip-Urban [4, 43]: This dataset includes images captured at the same locations as the S2NAIP-Urban super-resolution dataset [43], which is a subset of the extensive S2NAIP [4] dataset focused on urban areas. This split comprises 515,270 tiles, featuring imagery from NAIP at a 1.25 m resolution, Sentinel-2 and Sentinel-1 time series, and Landsat-8/9 data rescaled to a 10 m resolution. We use this dataset for pretraining only because there are no official labels and evaluations.

BraDD-S1TS [23]: BraDD-S1TS (Brazilian Deforestation Detection) is a change detection dataset comprising Sentinel-1 time series of the Amazon rainforest, aiming to segment deforested areas. It includes 13,234 tiles covering regions with varying deforestation rates, providing pixel-wise binary annotations for deforestation events occurring between the time series' first and last radar image.

Sickle [31]: SICKLE is a multimodal crop mapping dataset from India containing 34,848 tiles with Sentinel-1, Sentinel-2, and Landsat-8 time series. We use the paddy / non-paddy culture binary semantic segmentation task. As the test set has not been released by the authors, we perform our experiments on the validation set.

TimeSen2Crop [42]: TimeSen2Crop is a crop mapping dataset consisting of 1,212,224 single-pixel Sentinel-2 time series, a configuration not present in GeoPlex. It includes data from Slovenia with annotations for 16 different crop types.

Sen1floods11 [9]: Sen1Floods11 is a flood segmentation dataset featuring 4,831 pairs of Sentinel-1 and Sentinel-2 images, each annotated with dense flooded/not-flooded labels. The dataset spans diverse global regions, with each tile covering a 5120 × 5120 m area (2600 hectares) and containing a single acquisition date per sensor.

So2Sat [45]: So2Sat is a local climate zone classification dataset containing co-registered single-date Sentinel-1 and Sentinel-2 imagery across multiple cities worldwide. It comprises 400,673 image patches, each annotated with one of 17 local climate zone classes according to the LCZ scheme. An image represents a zone of size 320 × 320 m. So2Sat specifically targets urban morphology classification tasks for sustainable urban planning and climate studies.

HLS Burn Scar [28]: HLS Burn Scar is designed for post-fire burn scar detection using Harmonized Landsat-Sentinel (HLS) imagery. It contains 804 tiles covering a 15.3 × 15.3 km area 23400 hectares) at 30m resolution and covering multiple wildfire events across diverse ecosystems in the United States.

References

[1] Lightning: LinearWarmupCosineAnnealingLR. https://lightning-flash.readthedocs.io/en/stable/api/generated/flash.core.optimizers.LinearWarmupCosineAnnealingLR.html. Accessed: 2024-11-20. 3

[2] PyTorch: ReduceLROnPlateau. org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html#torch.optim.lr_scheduler.ReduceLROnPlateau. Accessed: 2024-02-29. 3

[3] 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 Discussions*, 2022. 4, 5

[4] allenai.org. AI2-S2-NAIP. https://huggingface.co/datasets/allenai/s2-naip, 2024. [Online; accessed 01-Sept-2024]. 5

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *CVPR*, 2021. 3

[6] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *ECCV*, 2024. 2, 3, 4, 5

[7] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022. 3

[8] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, 2023. 4

[9] 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. 2, 4, 5, 6

[10] Lorenzo Bruzzone and Sebastiano B Serpico. Classification of imbalanced remote-sensing data by neural networks. *Pattern recognition letters*, 1997. 4

[11] 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. 3, 4

[12] Bo Dang and Yansheng Li. MSResNet: Multiscale residual network via self-supervised learning for water-

body detection in remote sensing imagery. *Remote Sensing*, 2021. 4

[13] Anthony Fuller, Koreen Millard, and James R Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *NeurIPS*, 2023. 3, 4

[14] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, and Boris Wattrelos. FLAIR: A country-scale land cover semantic segmentation dataset from multi-source optical imagery. In *NeurIPS Dataset and Benchmark*, 2023. 2, 3, 5

[15] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *Advanced Analytics and Learning on Temporal Data: ECML PKDD Workshop*, 2020. 3, 4

[16] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *ICCV*, 2021. 2, 3, 4, 5

[17] Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. 3, 4

[18] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, et al. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *CVPR*, 2024. 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[21] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Pedram Ghamisi, Naoto Yokoya, Hao Li, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral remote sensing foundation model. *TPAMI*, 2024. 3

[22] 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. *URL https://arxiv.org/abs/2310.18660*. 3, 4

[23] Kaan Karaman, V Sainte Fare Garnot, and Jan Dirk Wegner. Deforestation detection in the Amazon with Sentinel-1 SAR image time series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2023. 2, 4, 5, 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. 3

[25] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *CVPR Workshop EarthVision*, 2019. 4

[26] 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. *TLMR*, 2023. 3

[27] 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. *International Geoscience and Remote Sensing Symposium*, 2024. 3, 5

[28] Christopher Phillips, Sujit Roy, Kumar Ankur, and Rahul Ramachandran. HLS foundation burnscars dataset, 2023. 4, 5, 6

[29] 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. 3, 4

[30] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 2021. 3

[31] Depanshu Sani, Sandeep Mahato, Sourabh Saini, Harsh Kumar Agarwal, Charu Chandra Devshali, Saket Anand, Gaurav Arora, and Thiagarajan Jayaraman. SICKLE: A multi-sensor satellite imagery dataset annotated with multiple key cropping parameters. In *WACV*, 2024. 2, 4, 5, 6

[32] Hochreiter Sepp and Schmidhuber Jürgen. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 2012. 4

[33] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-Chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 4

[34] Adam Stewart, Nils Lehmann, Isaac Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. SSSL4EO-l: Datasets and foundation models for Landsat imagery. *NeurIPS*, 36, 2024. 3

[35] 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 preprint arXiv:2412.02732*, 2024. 4

[36] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Michael Blumenstein, and Jing Jiang. Omni-scale CNNs: A simple and effective kernel size configuration for time series classification. In *ICLR*, 2021. 4

[37] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. ViTs for SITS: Vision transformers for satellite image time series. In *CVPR*, 2023. 3

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

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[40] Elliot Vincent, Jean Ponce, and Mathieu Aubry. Satellite image time series semantic change detection: Novel architecture and analysis of domain shift. *arXiv preprint arXiv:2407.07616*, 2024. 4

[41] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008. 4

[42] Giulio Weikmann, Claudia Paris, and Lorenzo Bruzzone. Timesen2crop: A million labeled samples dataset of Sentinel 2 image time series for crop-type classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021. 4, 5, 6

[43] Piper Wolters, Favyen Bastani, and Aniruddha Kembhavi. Zooming out on zooming in: Advancing super-resolution for remote sensing, 2023. 5

[44] 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 foundation model for observing the Earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2024. 3, 4

[45] Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Häberle, Yuansheng Hua, Rong Huang, et al. So2Sat LCZ42: A benchmark dataset for global local climate zones classification. *arXiv preprint arXiv:1912.12171*, 2019. 4, 5, 6