PLANTEUR: A Crop-Type Segmentation Benchmark for Tropical Regions

Piper Wolters **Gabriel Tseng** **Favyen Bastani** **Joseph Redmon**

Yawen Zhang **Hadrien Sablon** **Ando Shah**

Allen Institute for AI

Code: https://github.com/allenai/planteur-benchmark

Dataset: https://huggingface.co/datasets/allenai/planteur-benchmark

Contact: olmoearth@allenai.org

Abstract

We introduce **PLANTEUR**, a benchmark for crop-type mapping from satellite images in the French overseas territories. Building on the PASTIS design, PLANTEUR provides georeferenced and aligned Sentinel-2, Sentinel-1, and Landsat-8 sequences with parcel-derived labels spanning six tropical regions and an 8-class label space. We define both a standard test split and a leave-one-island-out protocol to measure cross-territory generalization, and evaluate fully supervised models, linear probes, and fine-tuned foundation models. The benchmark code and data are open source and available on GitHub and Hugging Face.

1 Introduction

Foundation models and open data are driving rapid progress in machine learning for remote sensing [6, 8, 10–12]. Benchmarks play a key role in tracking model improvement with actual applications. Good benchmarks have: high-quality labels, associated geospatial metadata (so users can fetch additional modalities/time series), and high correlation with real-world tasks. PASTIS [6] is one of the field's most trusted benchmarks because it pairs dense, high-quality parcel labels with complete geospatial metadata — but its coverage is limited to metropolitan France. Agricultural systems where mapping could be most valuable — smallholder, tropical, topographically complex — are almost entirely absent from evaluation. We extend PASTIS to six tropical regions to fill this gap.

We present PLANTEUR, a new open-access crop type segmentation benchmark. PLANTEUR complements the existing PASTIS benchmark by expanding the geographic diversity to tropical regions, while maintaining high quality labels. We use France's Land Parcel Identification System data source for crop type labels spanning six tropical regions, and align corresponding Sentinel-2, Sentinel-1, and Landsat-8 time series.

We evaluate twelve models across linear probing, fine-tuning and supervised training, on a standard split and leave-one-island-out setup, as well as with varying inputs and hyperparameters. We find that all methods we evaluated exhibit substantially lower performance on PLANTEUR compared to PASTIS. This validates our motivation for developing PLANTEUR and provides a new challenge for the modeling community to tackle.

To evaluate the effect of region transfer, we experiment with training on PASTIS samples in addition to PLANTEUR samples and find that those added samples, though different from the distribution of the PLANTEUR test set, improve performance for several models. We explore performance on unseen regions by

proposing a leave-one-island-out split and see a major drop in performance for most methods, suggesting a gap between a model performing well on a benchmark and generalizing to real world tasks.

Our contributions include 1) a dataset with permissive licenses including imagery from three sensor types, 2) geographically balanced splits, including a leave-one-island-out setup, and 3) extensive results on a variety of baseline methods, showing significant room for improvement. In Section 2 we introduce the construction and breakdown of PLANTEUR. In Section 3 we present the experimental results on PLANTEUR across a wide range of setups.

2 PLANTEUR: Crop-Type Segmentation in France's Overseas Territories

PLANTEUR is an open-access satellite image time series (SITS) semantic segmentation benchmark for crop type mapping. This dataset is publicly available at `https://huggingface.co/datasets/allenai/planteur-benchmark`.

2.1 Background: PASTIS

PASTIS [6] comprises 2,433 multi-spectral image sequences with both semantic and instance annotations. Each sequence covers a 128×128 pixel patch at 10m resolution, with between 38 and 61 Sentinel-2 acquisitions taken over a one-year period, for a total of over 2 billion annotated pixels spanning more than 4,000 km^2 across four regions of metropolitan France. This benchmark demonstrated that parcel-based crop-type segmentation from Sentinel-2 time series can be benchmarked with high label fidelity, leveraging France's Land Parcel Identification System, which claims a relative error rate in terms of surfaces under 0.3% [6]; this is because the parcel boundary and crop type data comes directly from farmers, and there are financial penalties and random audits that incentivize very accurate reporting. PASTIS is a reliable and trustworthy benchmark, but it is geographically limited to France, a single country in Europe. Fortunately, the same high quality data exists for other French regions. PLANTEUR extends this design to France's overseas territories, preserving the same annotation provenance while adding geographic and agronomic diversity through tropical crop regimes that are absent or underrepresented in metropolitan France, which we hope will be a useful complement to PASTIS.

2.2 Description

We aim to follow the setup of PASTIS for ease of adoption. PLANTEUR consists of 2,750 multi-spectral sequences, spanning September 2018 to December 2019. Each Sentinel-2 L2A series has shape $T \times 10 \times 128 \times 128$ and each Sentinel-1 RTC series $T \times 2 \times 128 \times 128$, with T between 40 and 150 observations (median 87 and 123 respectively; series are required to have at least 40 and are capped at 150). Landsat-8 is provided at $T \times 10 \times 64 \times 64$ with a co-registered panchromatic band at $T \times 1 \times 128 \times 128$, where T ranges from 28 to 61 (median 30), reflecting its 16-day revisit. Like PASTIS-R [7], we added additional sensor modalities to encourage multi-modal experiments, and extended the time range to make it easier to run with recent embedding products. The dataset spans 4,506 km^2 across Réunion, Guadeloupe, Martinique, Guyane, Mayotte, and Saint-Martin, geographically visualized in Figure 1.

Array formats. Each patch is released as per-sensor arrays with the acquisition dates recorded in a metadata geojson, following PASTIS: Sentinel-2 as $(T_{S2}, 10, 128, 128)$ uint16 (B02–B08, B8A, B11, B12), Sentinel-1 as $(T_{S1}, 2, 128, 128)$ float32 (VV, VH), and Landsat-8 as $(T_{LS}, 10, 64, 64)$ uint16 (B1–B7, B9–B11) with the panchromatic B8 held separately at $(T_{LS}, 1, 128, 128)$. Same-day acquisitions are mosaicked to a single observation, so the date lists are unique and chronological and their length matches each array's T.

Labels. Labels are derived from the French Land Parcel Identification System for the 2019 declaration year, rasterized to the patch grid at 10 m. Semantic labels are $(1, 128, 128)$ uint8.

Alongside the array formats provided, semantic labels are also released as georeferenced GeoTIFFs with one per patch, 128×128 uint8 at 10 m with nodata 255, keeping the patch's native UTM CRS and affine

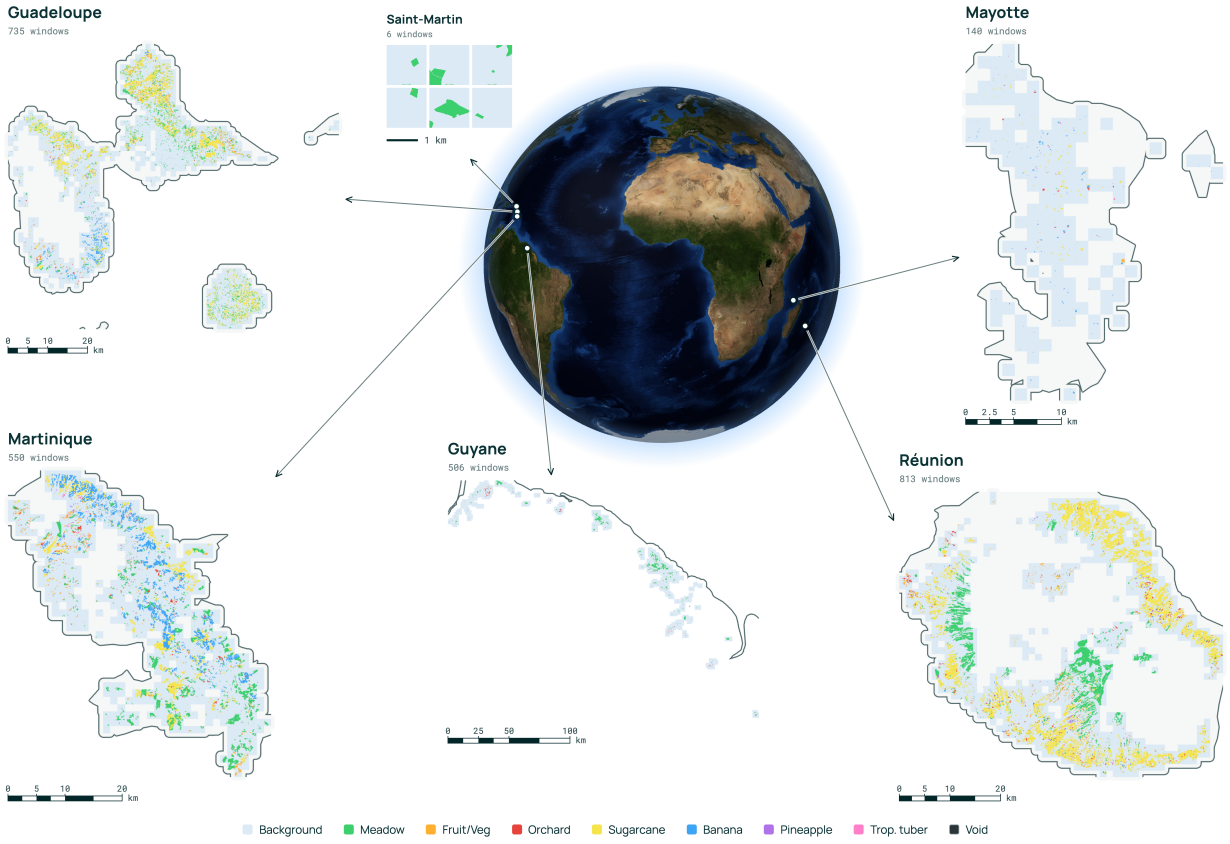


Figure 1 PLANTEUR pixels illustrated on maps of the six tropical regions included in the dataset. Note that the whitespace represents pixels that do not include labels, whether due to being non-agricultural (ex. mountains) or labels not existing in the data source.

transform. The .npy arrays are pixel-identical but carry no spatial reference, so the GeoTIFFs let users rasterize additional modalities directly onto the label grid and align external data.

Panoptic Segmentation Labels. We format the benchmark as closely to PASTIS as possible, including panoptic labels. Instance annotations give a per-pixel parcel id, an object-centerness heatmap and Voronoi zones, matching the PASTIS panoptic format. The focus in our experiments is on semantic segmentation and evaluating embedding models, so we leave panoptic evaluation to users interested in that task.

2.3 Annotation

Each pixel of PLANTEUR is labeled as one of the 7 crop classes or a Background, non-crop, class. The benchmark label space contains the following crop classes: Meadow, Fruit/Vegetable/Flower, Orchard, Sugarcane, Banana, Pineapple, Tropical tuber. The original label space contains additional crop types; we map any class with fewer than 40k labeled pixels to *void*. Void pixels are masked out of the training loss and excluded from evaluation metrics.

Class distributions, broken down by number of windows, parcels, and pixels, are shown in Figure 2. Sample images containing each crop type are in Figure 3.

Void mapping. The following crop types are mapped to void (ignored in the loss and dropped from the metric): Vanilla, Coffee/Cacao, Corn, Potatoes, Ylang-ylang, Leguminous fodder, Grapevine, Beet, Soybeans. In total, void accounts for 92,145 pixels (0.20% of all pixels overall; 0.37% in the test split).

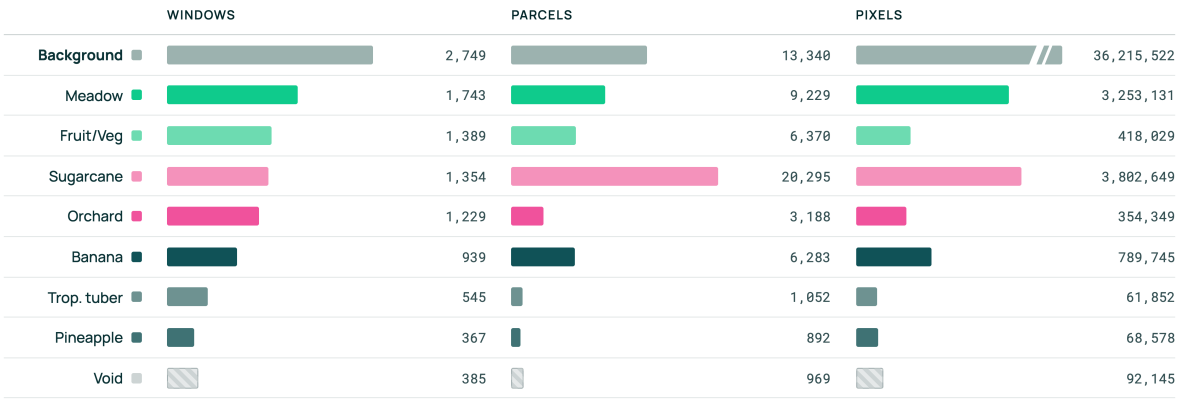


Figure 2 PLANTEUR class distribution, in terms of number of windows containing that class, number of parcels for that class, and number of pixels labeled as that class. The background class represents non-agriculture, and the void class consists of underrepresented tropical classes.

Island	test windows	test pixels	Background %	Trivial mIoU
Réunion	172	2,805,239	71.02%	8.88
Guadeloupe	69	1,130,352	85.99%	10.75
Martinique	58	950,272	84.20%	10.52
Guyane	80	1,301,548	78.83%	9.85
Mayotte	64	1,043,953	99.45%	12.43

Table 1 Per-island composition of the 443-window test split, broken down by number of windows and pixels. Note that Saint-Martin has just six windows in the train set and does not contribute to the test set.

2.4 Splits

Splits are assigned to spatial blocks, not to individual windows. Each 128×128 window is mapped to the 0.1° (≈ 11 km) latitude/longitude cell containing its centroid, and the split is assigned to the cell, so every window in a neighborhood lands in the same partition. This is necessary because the windows are densely tiled: adjacent windows observe the same landscape and crop calendar, and frequently the same parcels straddling a window boundary. Within a territory, we greedily select cells based on a hash-derived order, and assigned cumulatively until 70%, 85% and 100% of that territory's windows are covered. The 70/15/15 allocation is applied independently per territory, so every territory except Saint-Martin appears in all three partitions. This yields **1,896 train / 411 validation / 443 test** windows (68.9% / 14.9% / 16.1%); the 443-window test partition is used for all in-domain evaluation.

Because whole blocks are indivisible, realized per-territory ratios deviate from the target in inverse proportion to a territory's block count. All territories except Mayotte have **train / validation / test** ratios similar to the global ratio. Mayotte is relatively small, yielding a substantial deviation (39% train, 15% validation, 46% test). Saint-Martin is *very* small, so it just goes in the train set. We prioritize the spatial-block constraint over a closer ratio, because leakage would be a larger issue. Per-territory composition is given in Table 1.

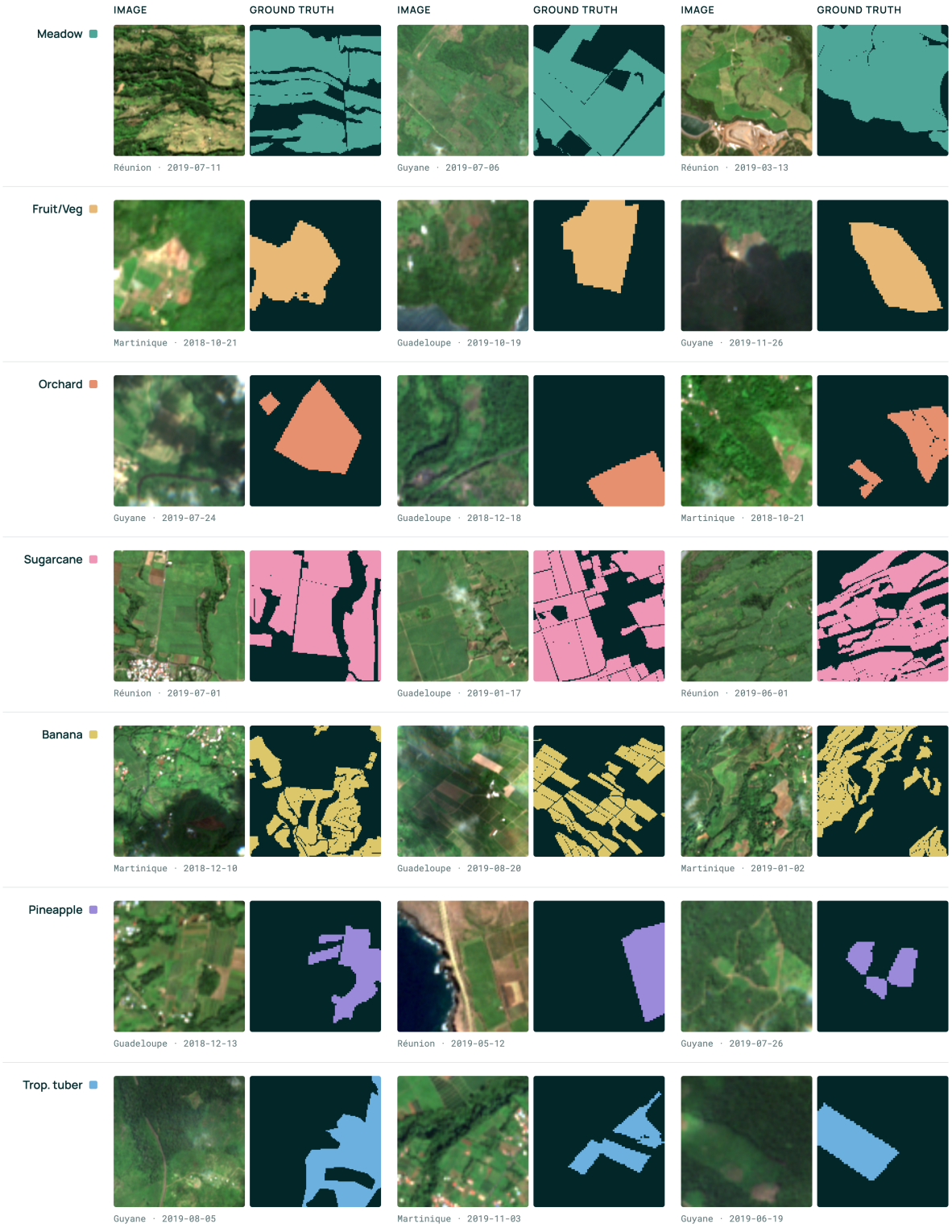


Figure 3 Example parcels for each scored crop class in PLANTEUR. Each chip is a Sentinel-2 RGB composite from a single acquisition date, with the ground-truth mask showing **just** that class' pixels.

3 Experiments & Results

3.1 Models

We benchmark three method families on PLANTEUR, spanning supervised models that performed well on PASTIS [5, 6], linear probes – released embedding products [2–4] and frozen encoders of remote sensing foundation models [1, 13, 14, 16, 17], and fine-tuned foundation models [1, 3, 13–15, 17].

Supervised from scratch. Baselines are trained on the PLANTEUR training split with no pretraining, using the code released for these models.

- **U-TAE** [6]: U-TAE is a deep learning architecture, originally proposed alongside PASTIS, that combines 2D convolutional layers with a temporal self-attention mechanism to perform highly accurate, date-agnostic spatio-temporal segmentation on satellite image time series datasets. We use the original PASTIS recipe (1.08M parameters, Adam 10^{-3}, batch 4, 100 epochs) for experiments on PLANTEUR.
- **TSViT** [5]: TSViT is a fully-attentional, transformer-based architecture, that was a prominent baseline model for PASTIS. The model operates on 24×24 tiles with 12-step sequences. We use the original PASTIS recipe (1.66M parameters, AdamW 10^{-3} with a cosine schedule and 10 warm-up epochs, batch 16, 100 epochs) for experiments on PLANTEUR.

Both of these models struggled with class imbalance, showing poor performance on underrepresented classes, so we experiment with different loss functions in Appendix 5.1. Experiments in the main paper use the original recipes and losses.

Linear probe. We evaluate two linear-probe setups.

- **Released embeddings.** For released annual embedding products, we fit a linear head on the precomputed embeddings: 50 epochs, one embedding per pixel (patch size = 1) over a 16×16 window. The probe learning rate is selected from an 8-point sweep (10^{-4} to 0.5) on the validation set. At most, the learning rate affects mIoU by 2.3 points.
  - **AEF** [2]. AlphaEarth Foundations (AEF) is Google DeepMind's pixel-wise embedding model that distills multi-sensor satellite observations into 64-dimensional annual embeddings at 10m resolution. We use their supplied embeddings, read from precomputed rasters.
  - **Tessera v1**[3]. Tessera v1 is a dual-encoder model over Sentinel-2 and Sentinel-1. We use the released precomputed 128-dimensional embeddings.
  - **Tessera v1.1**[3]. Tessera v1.1 embeddings improve upon v1 with fewer tiling artifacts, better temporal stability, and a wider encoder. We use the released precomputed 128-dimensional embeddings.
- **Frozen encoders.** Across all models we fit a linear head on features from a frozen encoder, which is never updated: 50 epochs, mean pooling, a 16×16 window, and a per-pixel readout obtained by bilinearly upsampling each encoder's token grid to the label resolution. The probe learning rate is selected per model from the same 8-point sweep (10^{-4} to 0.5). Patch size follows each model's native setting and is stated per model below.
  - **AnySat**[1]. AnySat is a multi-modal JEPA-style encoder that takes in various combinations of sensors and spatial resolutions through a scale-aware patch embedding. We use a patch size of 4, window size of 16, and input up to 96 Sentinel-2 observations in 2019.
  - **Galileo**[14]. Galileo is a multi-modal masked-autoencoder trained with paired global and local objectives over stacked modalities. We use a patch size of 4, window size of 16, and input 12 Sentinel-2 monthly mosaics from 2019.
  - **Presto**[13]. Presto is a lightweight transformer pretrained on pixel time series, encoding each pixel's temporal sequence independently. We use a patch size of 1, window size of 16, and input 12 Sentinel-2 monthly mosaics from 2019.
  - **Tessera v2** [4]. Tessera v2 is a pixel-wise Earth foundation model producing 128-dimensional annual

embeddings for 2019 at 10m. Since there are no released precomputed embeddings, we run the released v2 inference code to generate embeddings using all Sentinel-2 and Sentinel-1 observations in 2019, and keep the encoder frozen.

- **Copernicus-FM**[17]. Copernicus-FM is a ViT that is sensor-agnostic across spectral inputs. We resize tiles to 224×224 with bilinear interpolation to match the native pretraining setup; with patch size 16 and window size 128 this yields a 14×14 token grid per tile. It is a single-image encoder with no temporal axis so we encode each of the 12 monthly mosaics for 2019 independently and mean-pool the results.
- **OlmoEarth v1.3**[16]. OlmoEarth v1.3 is the OlmoEarth model specifically built for embeddings. It is a d=768 encoder with a distilled linear student, and we read it out three ways — through the distilled 64-dimensional and 128-dimensional projections, and at full register width with no projection — reported as separate rows throughout. We use a patch size of 1, window size of 16, and 12 monthly mosaics for 2019 as input.

Fine-tuned.

- **AnySat, Galileo, Presto, Tessera v1, Copernicus-FM**[1, 3, 13, 14, 17]. We also evaluate the performance of fine-tuning these foundation models. We use a window size of 128 and patch size of 16 for Copernicus-FM, a window size of 64 and patch size of 4 for OlmoEarth v1.2, Galileo, and AnySat and a window size of 16 and patch size of 1 for the pixelwise models, Presto and Tessera v1. The same inputs are used for fine-tuning as described for linear probes.
- **OlmoEarth v1.2** [15]. OlmoEarth v1.2 is an update of the original OlmoEarth model [9]. We fine-tune the model with patch size 4 and window size 64 (unless otherwise stated) for 20 epochs at 10^{-4} with batch size 8. We fine-tune OlmoEarth v1.2 deliberately: v1.2 is the better fine-tuning target, whereas we released v1.3 as an embeddings-specific model.

3.2 Protocols

Each protocol isolates a different axis of difficulty.

- **In-domain (443-window test split).** Train and test on splits of all six territories within PLANTEUR, except Saint-Martin which is only in the train split. Measures achievable accuracy when labels exist for the target region.
- **Balanced In-domain.** Train on a subset of the in-domain train split, where the number of Background samples equals the total number of crop samples, to reduce the class imbalance challenge. Test on the in-domain test split. This is a less practical setting than the in-domain split, but allows comparing the effectiveness of different satellite image encoders.
- **Leave-one-island-out (LOIO).** Train on four territories, test on the fifth. Measures transfer to a region absent from training.
- **Few-shot.** Train on X labeled pixels per PLANTEUR class, $X \in \{10, 25, 100, 1000\}$, and evaluate on the PLANTEUR test split. Measures performance given a small number of in-domain labels.
- **Few-shot transfer.** Train on some number of PASTIS samples plus X labeled pixels per PLANTEUR class, $X \in \{10, 25, 100, 1000\}$, and evaluate on the PLANTEUR test split. We experiment with using 1) *all PASTIS*, which uses all PASTIS samples and adds X PLANTEUR pixels per class, and 2) *balanced*, which draws X pixels of every class from both PASTIS and PLANTEUR, including overlapping classes. Measures how far a small number of local labels closes the gap left by cross-region transfer.

All evaluation uses mIoU over the eight scored classes (mIoU-8); void is excluded from both the loss and the metric. For folds where a class is absent from the ground truth, we include it in the mean with IoU set to 0 to keep results comparable across territories. Overall pixel accuracy is reported alongside, but this metric can be deceiving due to the Background class being the majority class.

Family		Model	mIoU	acc.
Linear probe	*Released embeddings*	AEF	27.71	89.51
		Tessera v1.1	32.78	**89.84**
		Tessera v1	**33.13**	89.77
	Frozen encoders	AnySat	28.79	86.59
		Galileo	34.71	89.91
		Presto	16.25	80.15
		Copernicus-FM	26.46	86.74
		Tessera v2	34.24	90.26
		OlmoEarth v1.3 (64d)	35.17	90.55
		OlmoEarth v1.3 (128d)	35.39	90.51
		OlmoEarth v1.3 (768d)	**36.73**	**91.40**
Supervised		U-TAE	35.82	90.62
		TSViT	36.33	90.27
		AnySat	38.35	90.58
		Galileo	38.25	91.27
		Presto	23.69	87.36
		Tessera v1	35.77	89.97
		Copernicus-FM	32.54	89.21
		OlmoEarth v1.2 FT	**41.91**	**91.91**
Trivial (all Background)			10.07	80.56

Table 2 Results on the PLANTEUR 443-window test split. Best within each block in bold, best overall highlighted.

Model	S2	S1	S1+S2	S2+LS	S1+S2+LS
OlmoEarth v1.3 (64d)	**35.17**	24.14	34.94	34.90	35.01
OlmoEarth v1.3 (128d)	35.39	23.60	35.77	35.74	**36.02**
OlmoEarth v1.3 (768d)	36.73	25.73	37.52	37.36	**37.75**
OlmoEarth v1.2 FT	**41.91**	31.08	40.57	41.09	40.08

Table 3 OlmoEarth varying inputs on the 443-window PLANTEUR test split. Best input combination per row in bold.

3.3 In-domain results (443-window test split)

Table 2 reports results on the PLANTEUR 443-window test split. Fine-tuning OlmoEarth v1.2 is strongest at 41.91 mIoU, ahead of both supervised baselines (TSViT 36.33, U-TAE 35.82). Among frozen encoders of comparable embedding dimensionality, OlmoEarth v1.3 (128d) is best, with Tessera v2 close behind (35.39 vs. 34.24, 1.2 points).

Monthly mosaics vs. all observations. We evaluate both models that input monthly mosaics (e.g. Galileo, OlmoEarth) and models that input all observations (e.g. AnySat, Tessera). These models cannot be compared apples-to-apples due to compute cost, so we plot performance as a function of the compute cost in Figure 4. This shows that OlmoEarth v1.2 is the best model at a reasonable cost, compared to AnySat and Tessera v1 which have lower performance and a higher compute cost.

Do additional input modalities improve performance? For the OlmoEarth variants, we experiment with varying the input modalities. Results for both OlmoEarth v1.3 probes and OlmoEarth v1.2 fine-tuning with Sentinel-2, Sentinel-1, and Landsat-8 input combinations are reported in Table 3. Sentinel-1 or Landsat-8 combined with Sentinel-2 gives a slight increase (768d: +0.8 S1+S2, +0.6 S2+LS, +1.0 S1+S2+LS; 128d: +0.4, +0.4, +0.6 mIoU points) for the linear probes on OlmoEarth v1.3 128-dimensional and 768-dimensional embeddings, but doesn't improve for the 64-dimensional embeddings or OlmoEarth v1.2.

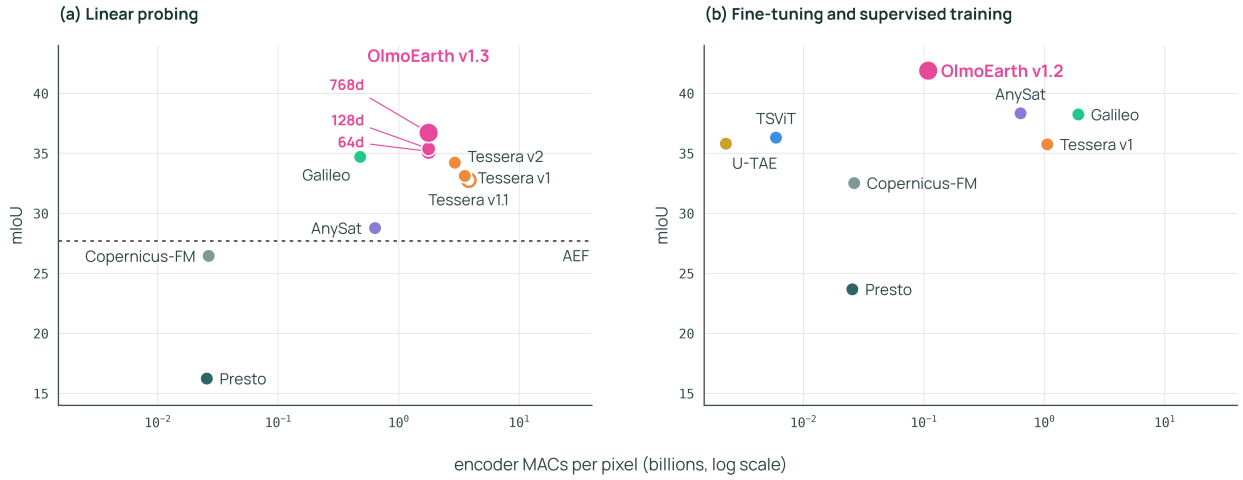


Figure 4 MACs vs. mIoU on PLANTEUR standard test set, separated into two plots for linear probe and supervised methods. The reported MACs represent compute cost of forward pass, so released embedding products wouldn't bear this cost. The dotted line marks AEF, a closed-weight embedding product for which encoder MACs are unavailable.

Model	$w = 4$	$w = 8$	$w = 16$	$w = 32$	$w = 64$
OlmoEarth v1.3 (64d)	33.21	32.37	35.17	33.69	**37.56**
OlmoEarth v1.3 (128d)	34.69	33.78	35.39	36.17	**37.64**
OlmoEarth v1.3 (768d)	34.45	35.84	36.73	37.81	**38.09**
OlmoEarth v1.2 FT	34.40	38.41	39.81	40.65	**41.91**

Table 4 Spatial-context ablation on the 443-window PLANTEUR test split. Best window size per row in bold.

How important is temporal context? Results for varying the temporal sequence length for the OlmoEarth v1.3 linear probes and OlmoEarth v1.2 fine-tuning are shown in Figure 5. Across all three OlmoEarth v1.3 embedding dimension sizes, performance increases by 11.6–12.8 mIoU points from one to twelve monthly mosaics, with the largest jump in improvement between a sequence length of one and three.

Does window size matter? We vary the window size for the frozen encoder linear probe results with all three dimensionality variants of OlmoEarth v1.3. Table 4 shows that a larger window size of 64 increases mIoU by 1.4–2.4 points over the $w = 16$ setting of Table 2, for both the distilled read-outs and the full width. Fine-tuning OlmoEarth v1.2 shows the same trend (34.40 at $w = 4$ to 41.91 at $w = 64$).

3.4 Balanced In-domain

The train / validation / test split used in Section 3.3 has majority Background samples, and this class imbalance is likely the reason for models over-predicting Background. This also means that employing specialized training strategies for imbalanced datasets can provide substantial performance gains — however, this is not the focus of our benchmark. Thus, we create a class-balanced training set to better distinguish the effectiveness of different remote sensing foundation models. We use the same test set to keep the evaluation realistic.

Confusion matrices for models trained on both the imbalanced and balanced sets are shown in Figure 8 in the appendix. These matrices and the results in Table 5 show an improvement (but not a *huge* improvement) for all methods, except for U-TAE, when using the balanced train set instead of the baseline train split. There are still a lot of pixels predicted as Background rather than the ground truth crop, suggesting that the benchmark's difficulty goes beyond class imbalance.

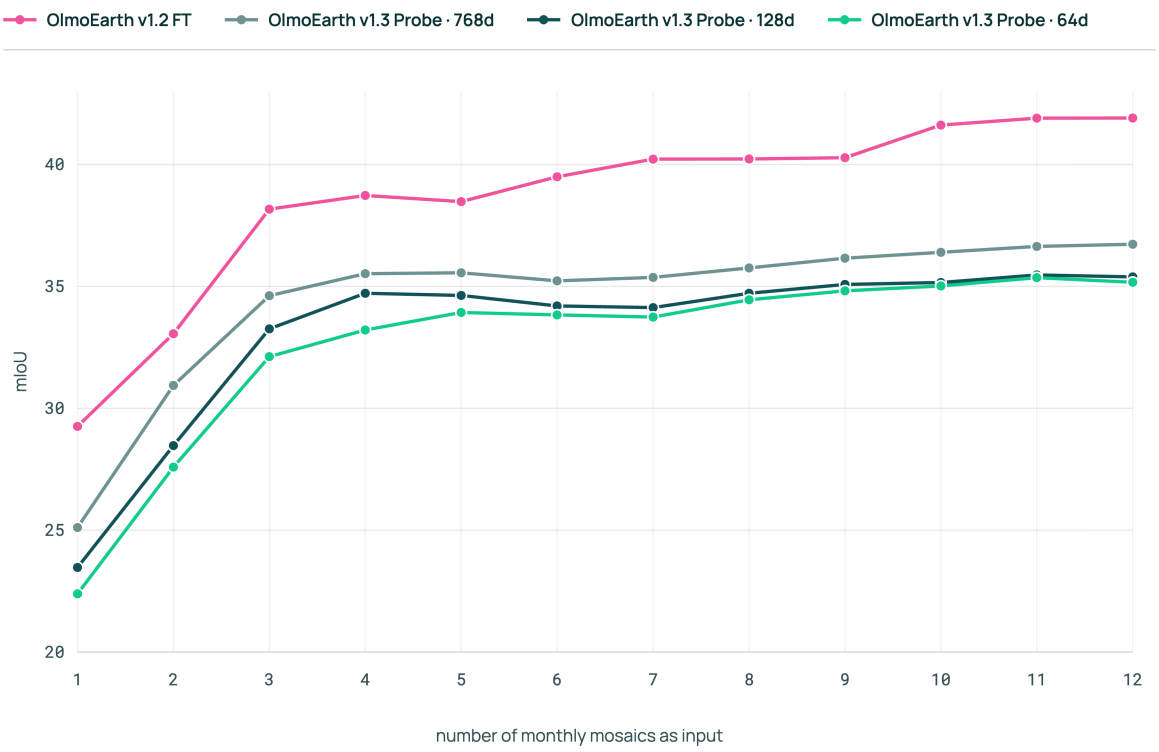


Figure 5 Temporal sequence-length ablation for OlmoEarth v1.3 linear probes and OlmoEarth v1.2 finetuning: mIoU-8 as a function of the number of monthly mosaics provided as input. Performance improves with a larger sequence length, though the rate of increase is largest between a sequence length of one and three for all methods.

Model	baseline	balanced
OlmoEarth v1.3 (64d)	35.17	36.77
OlmoEarth v1.3 (128d)	35.39	38.50
OlmoEarth v1.3 (768d)	**36.73**	**40.57**
Tessera v2	34.24	37.49
AEF	27.71	28.72
U-TAE	35.82	33.98
TSViT	36.33	37.95
OlmoEarth v1.2 FT	**41.91**	**42.03**

Table 5 Baseline vs. balanced, where the balanced setup reduces the number of Background samples to equal the total number of crop samples, evaluated on PLANTEUR 443-window test split. Best within each block in bold, best overall highlighted. The difference in performance between columns is not large, suggesting that the task difficulty is beyond dealing with class imbalance.

3.5 Leave-one-island-out (LOIO)

Held-out-territory results are in Table 6. The best mean across folds is 24.93 (OlmoEarth v1.3 128-dimensional probe), closely followed by OlmoEarth v1.2 fine-tuned (24.31, 0.6 mIoU points lower).

LOIO compared to in-domain. This setup is difficult for all models. Every method achieves roughly 55–70% of its in-domain performance: likely due to the crop type distribution varying widely across the different islands and some islands only having a subset of the crop classes in PLANTEUR.

Model	Réunion	Guadeloupe	Martinique	Guyane	Mayotte	mean
OlmoEarth v1.3 (64d)	22.70	28.63	29.94	13.35	**12.82**	21.49
OlmoEarth v1.3 (128d)	**30.36**	31.17	**35.17**	**15.52**	12.42	**24.93**
OlmoEarth v1.3 (768d)	13.06	**33.02**	34.19	14.10	12.42	21.36
Tessera v2	12.09	28.79	29.67	12.36	12.44	19.07
AEF	9.91	23.08	20.77	12.83	12.42	15.80
U-TAE	**22.34**	28.43	30.25	14.52	**12.52**	21.61
TSViT	18.72	28.33	33.86	12.23	12.37	21.10
OlmoEarth v1.2 FT	22.32	**32.16**	**37.20**	**17.47**	12.42	**24.31**
Trivial (all Background)	9.33	10.17	10.18	10.33	12.42	10.49

Table 6 PLANTEUR leave-one-island-out (LOIO) results. Mean is across the 5 folds. Best within each block in bold, best overall highlighted.

Model	X = 10	X = 25	X = 100	X = 1000
OlmoEarth v1.3 (64d)	21.46	21.04	25.23	27.67
OlmoEarth v1.3 (128d)	**21.82**	**21.64**	26.96	28.89
OlmoEarth v1.3 (768d)	20.74	20.99	**27.56**	**29.67**
Tessera v2	16.52	16.97	22.29	26.56
AEF	15.27	17.64	21.59	24.18
U-TAE	**6.60**	6.70	11.11	21.69
TSViT	5.15	6.90	9.15	19.92
OlmoEarth v1.2 FT	3.03	**11.69**	**28.55**	**31.05**
Trivial (all Background)	10.07			

Table 7 Few-shot on PLANTEUR labels only: train on X labeled PLANTEUR pixels per class, evaluate on the 443-window test split. Best within each block in bold, best overall highlighted.

Interpreting Mayotte. Mayotte is majority Background, so results on this fold do not hold much weight. We do not throw this region out though, because a small region with very few positive samples is a realistic application. None of the reported methods have a meaningful performance on this island.

3.6 Few-shot: PLANTEUR labels only

We test few shot learning on PLANTEUR by training on X pixels per class, $X \in \{10, 25, 100, 1000\}$, and evaluating on the PLANTEUR test set, see Table 7. For the supervised models, we use random cropping around the labeled pixels to try to minimize overfitting.

Performance with *very* few examples. At X = 10, the models have just eighty labeled pixels, and there is a wide range of performance. OlmoEarth v1.3 (128d) outperforms AEF and Tessera v2 by more than 5 mIoU points, and the best supervised model (U-TAE) by 15.2 points. With just 10 or 25 examples, the supervised models overfit and perform similar to predicting Background everywhere. There are likely better training recipes that would improve performance for supervised models.

Comparing OlmoEarth v1.3 embeddings (linear probe) and OlmoEarth v1.2 (fine-tuning). In the few-shot setting, OlmoEarth v1.3 (linear probe) outperforms OlmoEarth v1.2 (default fine-tuning recipe + random cropping) by 18.8 and 10.0 mIoU points at X = 10 and X = 25. With 100 or 1000 samples OlmoEarth v1.2 only outperforms OlmoEarth v1.3 probes by 1.0–1.4 points. This suggests that linear probing on embeddings is a strong option when labeled data is limited.

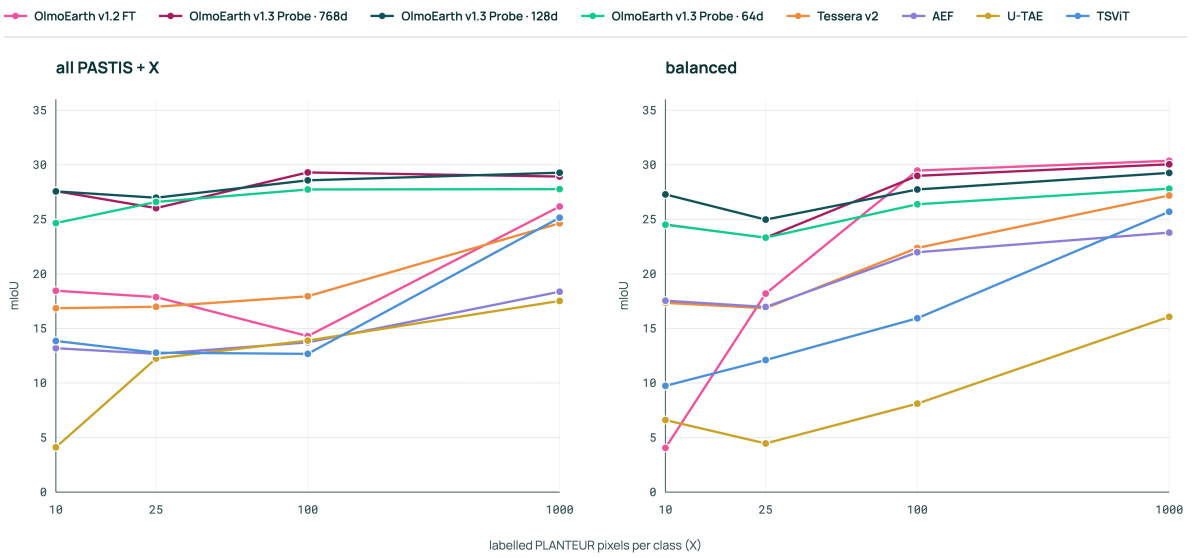


Figure 6 PLANTEUR few-shot transfer: mIoU-8 against X labeled PLANTEUR pixels per class, on top of PASTIS.

3.7 Few-shot transfer: adding PASTIS

Next, we see if adding PASTIS data to the training distribution, using the same PLANTEUR budget as Section 3.6, can narrow the remaining gap between using the whole training split and the few-shot setting. Like Section 3.6, we apply random cropping to the supervised runs. Results for all models are shown in Figure 6, which visualizes the results when using X PLANTEUR pixels per class, *all PASTIS*, which uses all PASTIS samples and adds X PLANTEUR pixels per class, and *balanced*, which draws X pixels of every class from both PASTIS and PLANTEUR, including overlapping classes.

All PASTIS versus balanced. When the full PASTIS training set is added to the X labeled PLANTEUR pixels per class, the PLANTEUR supervision is a tiny fraction of the training signal: at most $8X$ pixels, against millions of labeled PASTIS pixels. The four tropical classes (sugarcane, banana, pineapple, and tropical tuber) appear only in PLANTEUR, so most models never learn to predict them at small X, and their mIoU comes almost entirely from the classes the two datasets share. Their scores then reflect how well PASTIS-trained features happen to transfer to those shared classes rather than how many PLANTEUR labels were added, which explains the sometimes non-monotonic trends in the *all PASTIS* rows.

When there is a more equal distribution of crops between PASTIS and PLANTEUR, the tropical classes carry real weight and so performance generally rises from $X = 100$ to $X = 1000$, though most methods dip between $X = 10$ and $X = 25$.

Do PASTIS samples improve performance on PLANTEUR when limited samples are available? Compared to the few-shot results in Table 7, all OlmoEarth v1.3 variants see an improvement when adding PASTIS samples in both the *all PASTIS* and *balanced* settings, with a consistent boost for most values of X. Tessera v2 and AEF do not see the same increase; the *balanced* barely changes them, and *all PASTIS* lowers them (-4.3 and -7.9 at $X = 100$).

For OlmoEarth v1.2 fine-tuning, adding PASTIS samples is important when X is small, but doesn't matter when X is 100 or 1000, likely meaning that the performance on shared crop classes between PASTIS and PLANTEUR improve, while performance on tropical classes drops.

4 Conclusion

We introduce PLANTEUR, a new crop-type mapping segmentation benchmark, meant to expand the community's evaluations to tropical regions, while maintaining the high-quality and accuracy of the ground truth labels. As the results show, this is a difficult benchmark compared to PASTIS. Results across models are not saturated, so there is room to improve. We hope the community tackles this benchmark with new ideas and methods!

Although PLANTEUR expands data coverage to tropical crops and regions, while maintaining high accuracy in parcel boundaries and crop types, a limitation of France's Land Parcel Identification System is completeness – participation of farmers in reporting their data to this system is only mandatory if they participate in the EU Common Agricultural Policy, so there are still many small-scale farms that don't participate in this program and thus do not report to this system. This is evident in the original data analysis, where we had to throw out crops like coffee and vanilla because the number of samples was way too low. Also, this program combines some tropical fields into generic or vague categories (e.g., mango and lychee fields are lumped into the orchards category). There is also a chance of crop change in between seasons that is not accounted for, or differences in growing seasons (e.g. Réunion is in the southern hemisphere so the growing seasons are out of sync with the exported imagery). In future work, it is critical to source additional high quality crop labels that expand the distribution to more and more small-scale crops.

References

[1] Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: An Earth observation model for any resolutions, scales, and modalities. *arXiv preprint arXiv:2412.14123*, 2024.

[2] Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, Noel Gorelick, Lihui Lydia Zhang, Sophia Alj, Emily Schechter, Sean Askay, Oliver Guinan, Rebecca Moore, Alexis Boukouvalas, and Pushmeet Kohli. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data, 2025.

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

[4] Z. Feng, S. Jaffer, I. Shokar, J. Knezevic, M. Elvers, C. Atzberger, R. Young, A. Naik, N. Robinson, A. Blake, D. Coomes, A. Madhavapeddy, and S. Keshav. Tessera v2: Scaling pixel-wise earth foundation models, 2026.

[5] Michael Garn et al. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[6] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[7] Vivien Sainte Fare Garnot, Loïc Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *CoRR*, abs/2112.07558, 2021.

[8] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[9] Henry Herzog, Favyen Bastani, Yawen Zhang, Gabriel Tseng, Joseph Redmon, Hadrien Sablon, Ryan Park, Jacob Morrison, Alexandra Buraczynski, Karen Farley, et al. Olmoearth: Stable latent image modeling for multimodal earth observation. In *CVPR*, 2026.

[10] Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Crastici, Benjamin Segler, Hannah Matcham, Aaron Courville, David David, Nicolas Longepe, Alexandre Gava, et al. Geo-bench: Toward foundation models for earth monitoring. In *Advances in Neural Information Processing Systems 36 (NeurIPS Datasets and Benchmarks Track)*, pages 51080–51093. Curran Associates, Inc., 2023.

[11] Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, Geeth De Mel, Salman Khan, Manil Maskey, Nicolas Longépé, Xiao Xiang Zhu, Hannah Kerner, et al. Geo-bench-2: From performance to capability, rethinking evaluation in geospatial ai. *Transactions on Machine Learning Research*, 2026.

[12] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019.

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

[14] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global & local features of many remote sensing modalities. In *Forty-second International Conference on Machine Learning*, 2025.

[15] Gabriel Tseng, Yawen Zhang, Favyen Bastani, Henry Herzog, Joseph Redmon, Hadrien Sablon, Piper Wolters, Ando Shah, Patrick Alan Johnson, Christopher Wilhelm, and Patrick Beukema. Olmoearth v1.2: A more efficient family of olmoearth models. *arXiv preprint arXiv:2605.20804*, 2026.

[16] Gabriel Tseng, Yawen Zhang, Favyen Bastani, Joseph Redmon, Hadrien Sablon, Piper Wolters, Ando Shah, and Patrick Johnson. Olmoearth v1.3: Embeddings. 2026.

[17] Yi Wang, Zhitong Xiong, Chenying Liu, Adam J Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, et al. Towards a unified copernicus foundation model for earth vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

5 Appendix

5.1 Ablations

5.1.1 Class weighting and focal loss

Table 8 compares class weighting and focal loss against the default losses of Table 2. Inverse-frequency weighting hurts every model, by 2.8–9.4 mIoU points. The other variants change the supervised baselines by about one point or less: U-TAE is best with inverse-square-root weighting (35.82 → 36.85) and TSViT with focal loss alone (36.33 → 36.45). OlmoEarth v1.2 is best with its default loss and loses 1.2–9.4 points under every alternative. The exception is the OlmoEarth v1.3 (768d) linear probe, which gains 4.5 points (36.73 → 41.20) with inverse-square-root weighting. Because no single loss helps consistently, we keep the default losses for all models.

5.1.2 Few-Shot Plus Unseen Region Transfer

Table 9 shows the results for the few-shot PLANTEUR plus unseen region transfer experiments where we add PASTIS samples to the training set before evaluating on the PLANTEUR test set. These results correspond to Figure 6 in the main paper.

5.1.3 Unseen Region Transfer

As the proposed benchmark is a regional extension of an existing one, it is interesting to ask whether there can be transfer between the two. We fine-tune OlmoEarth v1.3 and OlmoEarth v1.2 on PASTIS, then fit a linear probe on PLANTEUR and evaluate on the PLANTEUR test split. There are four overlapping classes between PASTIS and PLANTEUR and thus four unseen (aside from training the linear probe) classes introduced in PLANTEUR. The results, shown in Table 10 are not impressive, which is why we emphasize the positive results of few-shot transfer, as discussed in Section 3.7, where adding just a few examples of your region of interest to a much larger data source, can result in strong performance.

5.1.4 LOIO Ablations

We repeat a few of the ablations mentioned in Section 3.3 for the LOIO protocol.

Do additional input modalities improve performance in this setup? We again run experiments with varying input modalities for OlmoEarth v1.3 and OlmoEarth v1.2. Input modality does not have a consistent effect on LOIO results as shown in Table 11.

What is the optimal window size for LOIO? We run experiments to vary the spatial context for the OlmoEarth v1.3 models. Results in Table 12 show a similar trend to the in-domain setting, with a larger window size improving performance for most of the held out islands.

5.1.5 Overall Accuracies

In the main paper we mostly report mIoU as it is the prioritized metric. Overall accuracy can also tell us useful information, so we include Tables 13, 14, 15, 16, and 17 of overall accuracies for each experimental setup in the main paper.

5.2 Additional Visuals

Sample Predictions. Figure 7 shows samples of Sentinel-2 image, ground truth mask, and model predictions, specifically sampled to display a few of the different tropical classes.

Confusion Matrices. Figure 8 shows confusion matrices for OlmoEarth v1.3, Tessera v2, AEF, and OlmoEarth v1.2 for the standard setting and the balanced setting, where the training data contained equal samples of background class as all crop classes.

Model	Regime	baseline	inv	invsqrt	focal γ = 2	invsqrt+focal
U-TAE	supervised	35.82	28.82	**36.85**	36.51	36.68
TSViT	supervised	36.33	27.55	35.76	**36.45**	34.48
OlmoEarth v1.2	fine-tune	**41.91**	32.50	40.69	40.20	39.92
OlmoEarth v1.3 (768d)	linear probe	36.73	33.93	**41.20**	36.02	40.12

Table 8 Class-weighting and focal-loss ablations on PLANTEUR (443-window test split), mIoU-8. Best per row in bold.

Model	Arm	X = 10	X = 25	X = 100	X = 1000
OlmoEarth v1.3 (64d)	all PASTIS	24.66	26.60	27.74	27.77
OlmoEarth v1.3 (64d)	balanced	24.51	23.33	26.38	27.81
OlmoEarth v1.3 (128d)	all PASTIS	27.57	**26.98**	28.58	29.28
OlmoEarth v1.3 (128d)	balanced	27.29	24.98	27.74	29.26
OlmoEarth v1.3 (768d)	all PASTIS	**27.59**	26.03	**29.30**	28.92
OlmoEarth v1.3 (768d)	balanced	24.53	23.32	28.99	**30.05**
Tessera v2	all PASTIS	16.86	16.99	17.95	24.64
Tessera v2	balanced	17.35	16.86	22.38	27.19
AEF	all PASTIS	13.19	12.67	13.73	18.37
AEF	balanced	17.55	16.97	21.99	23.79
U-TAE	all PASTIS	4.11	12.25	13.89	17.52
U-TAE	balanced	6.61	4.46	8.11	16.07
TSViT	all PASTIS	13.86	12.78	12.67	25.15
TSViT	balanced	9.74	12.11	15.94	25.70
OlmoEarth v1.2 FT	all PASTIS	**18.46**	17.88	14.30	26.18
OlmoEarth v1.2 FT	balanced	4.05	**18.20**	**29.47**	**30.37**
Trivial	–	10.07			

Table 9 Unseen-region transfer: PASTIS plus X labeled PLANTEUR pixels per class. *all PASTIS* keeps the full PASTIS training set; *balanced* draws X pixels of every class from both sources. Best within each block in bold, best overall highlighted.

Model	mIoU-8	trop-4	acc.
OlmoEarth v1.3 (64d)	26.92	22.86	89.06
OlmoEarth v1.3 (128d)	31.73	31.29	89.24
OlmoEarth v1.3 (768d)	35.82	33.99	90.73
OlmoEarth v1.2 FT	37.83	35.65	90.13

Table 10 Unseen transfer: encoders fine-tuned on PASTIS, then linear-probed on the PLANTEUR 443-window test split. The table shows mIoU-8 and mIoU over just the four tropical classes absent from PASTIS (Sugarcane, Banana, Pineapple, Tropical tuber; trop-4). Each cell is the best of an 8-point learning-rate sweep selected on validation.

Model	Inputs	Réunion	Guadeloupe	Martinique	Guyane	Mayotte	mean
OlmoEarth v1.3 (64d)	S2	22.70	28.63	**29.94**	13.35	12.82	**21.49**
OlmoEarth v1.3 (64d)	S1	14.16	20.99	22.31	11.87	12.41	16.35
OlmoEarth v1.3 (64d)	S1+S2	20.70	**29.02**	29.42	**14.85**	12.61	21.32
OlmoEarth v1.3 (64d)	S2+LS	**22.99**	28.48	28.92	13.58	12.79	21.35
OlmoEarth v1.3 (64d)	S1+S2+LS	21.31	28.85	29.11	14.01	12.59	21.17
OlmoEarth v1.3 (128d)	S2	**30.36**	**31.17**	35.17	15.52	12.42	**24.93**
OlmoEarth v1.3 (128d)	S1	12.92	21.99	20.87	11.56	12.38	15.94
OlmoEarth v1.3 (128d)	S1+S2	19.06	31.00	**35.27**	**15.74**	12.42	22.70
OlmoEarth v1.3 (128d)	S2+LS	29.12	30.67	34.16	14.78	12.42	24.23
OlmoEarth v1.3 (128d)	S1+S2+LS	20.04	30.70	34.39	15.42	12.42	22.59
OlmoEarth v1.3 (768d)	S2	13.06	**33.02**	**34.19**	14.10	12.42	21.36
OlmoEarth v1.3 (768d)	S1	12.08	21.86	23.50	11.31	12.35	16.22
OlmoEarth v1.3 (768d)	S1+S2	14.50	32.50	33.94	**15.11**	12.42	**21.69**
OlmoEarth v1.3 (768d)	S2+LS	**14.98**	32.78	33.19	13.74	12.42	21.42
OlmoEarth v1.3 (768d)	S1+S2+LS	14.92	32.50	33.19	14.42	12.42	21.49
OlmoEarth v1.2 FT	S2	22.32	**32.16**	**37.20**	**17.47**	12.42	**24.31**
OlmoEarth v1.2 FT	S1	17.57	24.83	26.59	12.94	12.33	18.85
OlmoEarth v1.2 FT	S1+S2	27.11	29.33	33.89	16.47	12.43	23.85
OlmoEarth v1.2 FT	S2+LS	26.22	28.07	32.01	15.85	12.42	22.91
OlmoEarth v1.2 FT	S1+S2+LS	**27.81**	29.24	34.01	16.07	12.44	23.91

Table 11 Leave-one-island-out (LOIO) input ablations for OlmoEarth. Best mIoU-8 within each read-out block in bold, best overall highlighted, except for the Mayotte column which is not meaningful.

Probe	w	Réunion	Guadeloupe	Martinique	Guyane	Mayotte	mean
OlmoEarth v1.3 (64d)	4	20.57	29.29	26.78	11.70	12.40	20.15
OlmoEarth v1.3 (64d)	8	19.14	28.80	27.78	12.69	12.52	20.19
OlmoEarth v1.3 (64d)	16	22.70	28.63	29.94	13.35	12.82	**21.49**
OlmoEarth v1.3 (128d)	4	20.25	29.48	28.10	13.43	12.42	20.74
OlmoEarth v1.3 (128d)	8	17.91	29.88	30.51	12.60	12.53	20.69
OlmoEarth v1.3 (128d)	16	30.36	31.17	35.17	15.52	12.42	**24.93**
OlmoEarth v1.3 (768d)	4	12.99	30.77	29.21	12.77	12.42	19.63
OlmoEarth v1.3 (768d)	8	16.21	32.56	30.47	13.51	12.42	21.03
OlmoEarth v1.3 (768d)	16	13.06	33.02	34.19	14.10	12.42	**21.36**

Table 12 Spatial-context ablation under LOIO protocol. Mean mIoU-8 is across the 5 folds. Best mean per section in bold.

Model	S2	S1	S1+S2	S2+LS	S1+S2+LS
OlmoEarth v1.3 (64d)	**90.55**	84.17	90.21	90.47	90.26
OlmoEarth v1.3 (128d)	90.51	83.16	**90.64**	90.50	90.61
OlmoEarth v1.3 (768d)	91.40	87.29	**91.47**	91.39	**91.47**
OlmoEarth v1.2 FT	**91.91**	89.45	91.46	91.48	91.61

Table 13 OlmoEarth varying inputs on the 443-window PLANTEUR test split, overall pixel accuracy (%). Same runs as Table 3. Best input combination per row in bold.

Model	w = 4	w = 8	w = 16	w = 32	w = 64
OlmoEarth v1.3 (64d)	89.85	89.93	**90.55**	89.97	89.55
OlmoEarth v1.3 (128d)	90.28	90.40	90.51	91.06	**91.32**
OlmoEarth v1.3 (768d)	90.55	91.03	91.40	91.71	**91.91**
OlmoEarth v1.2 FT	89.03	90.15	90.67	91.39	**91.91**

Table 14 Spatial-context ablation on the 443-window PLANTEUR test split, overall pixel accuracy (%). Same runs as Table 4. Best window size per row in bold.

Model	baseline	balanced
OlmoEarth v1.3 (64d)	90.55	89.18
OlmoEarth v1.3 (128d)	90.51	89.58
OlmoEarth v1.3 (768d)	**91.40**	**89.88**
Tessera v2	90.26	89.70
AEF	89.51	87.73
U-TAE	90.62	89.86
TSViT	90.27	88.88
OlmoEarth v1.2 FT	**91.91**	**90.66**

Table 15 Baseline vs. balanced, overall pixel accuracy (%), where balanced reduces the number of Background samples to equal the total number of crop samples, evaluated on the PLANTEUR 443-window test split. Same runs as Table 5. Best within each block in bold, best overall highlighted.

Model	X = 10	X = 25	X = 100	X = 1000
OlmoEarth v1.3 (64d)	65.35	58.40	68.50	70.60
OlmoEarth v1.3 (128d)	**70.02**	**63.12**	**69.59**	70.81
OlmoEarth v1.3 (768d)	57.76	53.16	66.01	68.32
Tessera v2	41.07	45.27	58.67	70.54
AEF	46.87	49.15	63.93	**72.53**
U-TAE	**44.59**	28.79	39.58	65.95
TSViT	34.31	**44.71**	**68.03**	61.54
OlmoEarth v1.2 FT	9.90	18.08	67.89	**74.00**
Trivial (all Background)	80.56			

Table 16 Few-shot on PLANTEUR labels only, overall pixel accuracy (%): train on X labeled PLANTEUR pixels per class with no PASTIS data, evaluate on the 443-window test split. Same runs as Table 7. Best within each block in bold, best overall highlighted.

Model	Réunion	Guadeloupe	Martinique	Guyane	Mayotte	mean
OlmoEarth v1.3 (64d)	83.84	88.41	87.85	83.35	98.97	88.48
OlmoEarth v1.3 (128d)	**87.55**	87.28	88.56	**84.50**	**99.39**	**89.46**
OlmoEarth v1.3 (768d)	74.81	**88.86**	**88.82**	82.92	**99.39**	86.96
Tessera v2	75.95	87.02	88.24	83.05	99.26	86.70
AEF	74.31	85.90	85.28	80.02	99.37	84.98
U-TAE	**83.66**	87.51	88.35	83.71	97.78	88.20
TSViT	80.73	85.82	88.07	81.84	98.46	86.98
OlmoEarth v1.2 FT	82.70	**89.27**	**90.19**	**85.31**	99.32	**89.36**
Trivial (all Background)	74.64	81.36	81.44	82.64	99.36	83.89

Table 17 PLANTEUR leave-one-island-out (LOIO), Sentinel-2 inputs: pixel accuracy (%). Mean is across the 5 folds. Best within each block in bold, best overall highlighted.

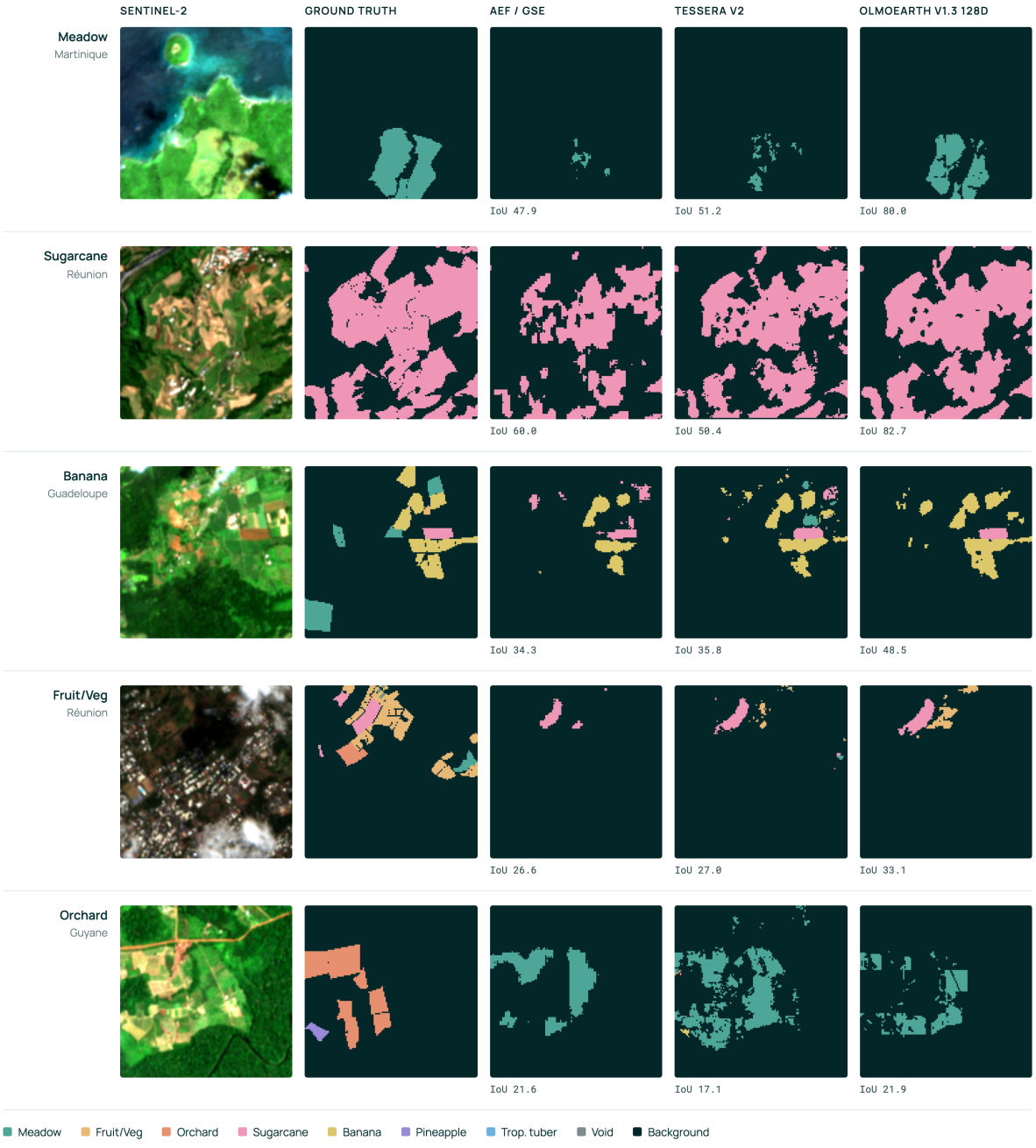


Figure 7 Qualitative PLANTEUR examples, with each row showing a Sentinel-2 mosaic, corresponding ground truth labels, and predicted mask from linear probing AEF, Tessera v2, and OlmoEarth v1.3.

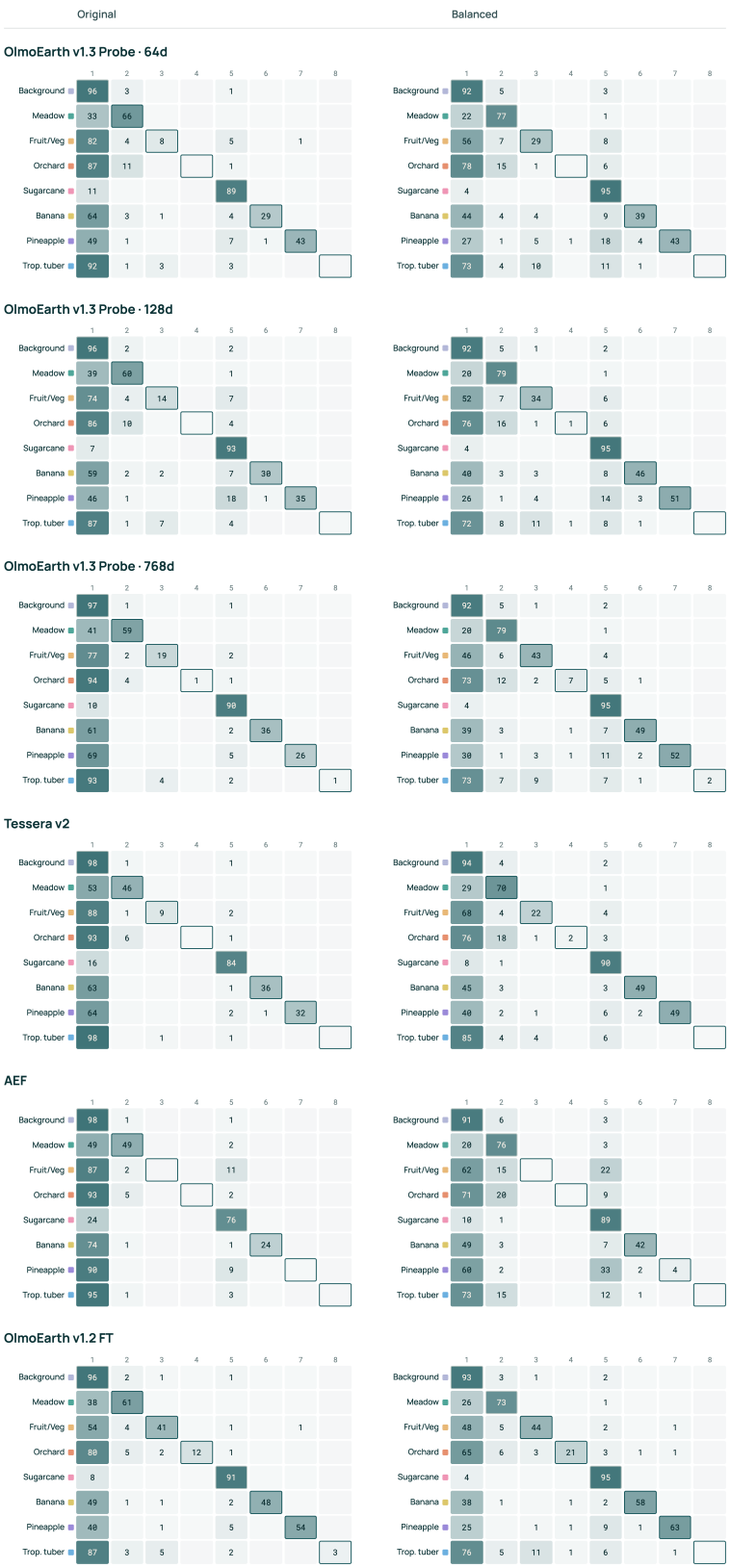


Figure 8 Confusion matrices for methods trained on the imbalanced and balanced train set and evaluated on the 443-window PLANTEUR test split. Rows are ground-truth classes, columns predictions, row-normalized; darker cells carry more mass. Off-diagonal mass sits almost entirely in the Background column.