Finch 1.0: A Multimodal Foundation Model for Earth Observation

Model Card

Locamage Team
contact@locamage.com
November 18, 2026

Abstract

Finch 1.0 is a multimodal foundation model for Earth observation. It encodes imagery from 50 satellite, aerial and drone sensors at native resolution, producing 1024-dimensional embeddings that support downstream tasks with only a linear readout. Finch is pretrained on imagery alone, without labelled map data such as OpenStreetMap or land-cover and crop maps. Finch 1.0 achieves state-of-the-art results on several tasks central to food security and environmental monitoring. On CropHarvest China, it classifies crop and non-crop locations from Sentinel-1 radar alone with 70.8% accuracy, 7.1 points above the best published result (63.7, OlmoEarth v1.2 Nano) and 3.8 points above our evaluation of OlmoEarth v1.3 Base. On MADOS, the segmentation of marine debris, oil spills and other sea-surface features into 15 classes, it achieves 72.0 mIoU, compared with 68.5 for the best published model (OlmoEarth v1.1 Base). On DynamicEarthNet land-cover segmentation from daily PlanetScope imagery, a linear readout on Finch embeddings scores 32.9 mIoU, above every frozen encoder paired with a trained UNet decoder on the GEO-Bench-2 leaderboard (31.7), and on PASTIS crop mapping with a linear readout, Finch scores 50.4 mIoU against 39.1 for the AlphaEarth Foundations embeddings. Finch also generalizes to sensors absent from its training data: on hyperspectral-only NeonTree imagery from the NEON imaging spectrometer, Finch and UniverSat-B, a model pretrained on that sensor, both score 46.9 test mIoU. Across 34 core benchmark tasks from three public suites, Finch performs on par with the strongest open models, averaging 65.1 on the 18 OlmoEarth tasks (OlmoEarth v1.3 Base: 65.5) and 65.9 on the 9 UniverSat tasks that UniverSat-B reports (UniverSat-B: 66.1); with patch sizes selected on validation data, these averages rise to 65.7 and 66.4. We release one embedding dataset per core task under CC BY 4.0.

1 Introduction

Foundation models for Earth observation map raw satellite and aerial imagery to general-purpose representations, allowing practitioners to address tasks such as land-cover, crop, forest and water mapping with limited labels and a lightweight readout, without training deep models or processing raw imagery (Herzog et al., 2025; Tseng et al., 2026b; Perron et al., 2026; Brown et al., 2025). Most existing models, however, support a single sensor or a small set of sensors at fixed resolutions.

Finch 1.0 is trained on imagery from 50 satellite, aerial and drone sensors, including multispectral and synthetic aperture radar instruments, across a wide range of spatial resolutions. It accepts imagery at its native resolution and produces 1024-dimensional embeddings on a grid whose spacing depends on the input resolution and the patch size. Finch 1.0 is a proprietary model of Locamage: its weights and code are not released. The public release consists of the embeddings of the samples of the 34 core benchmark tasks (Section 5); the embedding maps in this card are demonstrations. This model card presents a comprehensive evaluation of Finch 1.0, including:

- **Comparisons with published results** on the OlmoEarth suite, the UniverSat suite and seven GEO-Bench-2 tasks, and with the AlphaEarth Foundations and TESSERA embeddings (Section 4).
- **Patch-size selection on validation data**, quantifying the gain when Finch selects its patch size for each task (Section 4.5).
- **Re-evaluations of the baselines**: we re-evaluated six published baselines with our pipeline. Their suite means were within 1.0 point of the published values. Section 4.7 reports the per-task results.
- **An evaluation on hyperspectral imagery** from a sensor absent from the training data (Section 4.8).
- **A release of one embedding dataset per core task**, keyed by the sample identifiers and splits of each benchmark (Section 5).

The remainder of this card is organized as follows. Section 2 describes our infrastructure, Section 3 the evaluation methodology and Section 4 the results. Section 5 describes the embedding release, and Section 6 discusses intended use and responsible deployment.

2 Infrastructure

Finch 1.0 was developed on an in-house software stack, summarized below.

Scheduling and compute. Training and evaluation jobs are managed by an internal scheduler built on components of the open-source Marin project (The Marin Community, 2025). It provides a single interface for submitting, monitoring and terminating jobs across Google Cloud TPU v4, TPU v5e and TPU v6e accelerators and NVIDIA H100 GPUs in multiple Google Cloud regions, and allocates jobs to available capacity, including preemptible capacity. The scheduler runs directly on Compute Engine virtual machines rather than on Kubernetes, as Google Kubernetes Engine was not available to us. Support for AWS Trainium accelerators was planned but is deferred to future work, as AWS capacity was not available for this release.

Data loading. Training data is stored in object storage within each compute region, so that every job reads its data locally. Input pipelines are built on datamage, our Rust data loader with Python bindings, which streams columnar tables from object storage with seeded shuffling, prefetching and exactly resumable read positions.

Model code. Model code is written in JAX (Bradbury et al., 2018) using jaxmage, our library of neural network layers, checkpointing utilities, optimizers and accelerator kernels. jaxmage includes a fused linear cross-entropy kernel, similar to those in Liger Kernel (Hsu et al., 2024) and Tokamax (Tokamax, 2025), which computes the loss without materializing the full logit matrix. We plan to release jaxmage as open source following this model card.

Evaluation and profiling. Evaluation runs as a separate service. Each checkpoint is evaluated with the model code stored alongside it, so that every result in this card is traceable to the code that produced it. Training and evaluation are profiled with xprof-rs (Locamage, 2026), our open-source reader for XProf profiles.

3 Evaluation Setup

3.1 Protocol

In all evaluations, the Finch encoder is frozen and only a lightweight readout is trained on its embeddings: either a k-nearest-neighbour (kNN) classifier, which has no trainable parameters, or a single linear layer, with approximately 4K to 3.4M parameters on the 34 core tasks, depending on the number of classes and the output resolution of the task. Each task uses the readout specified by its published protocol. Readout hyperparameters (pooling, learning rate and number of epochs) are selected on the validation split, and we report the test-split result of the selected configuration in the primary metric of the task. All results use a single probe seed, and we did not estimate the variability of the suite means. Where the primary metric of a task's protocol differs from the metric of its published results, we report the published metric. Unless otherwise noted, Finch uses the standard patch size of each benchmark.

3.2 Benchmarks

Our core evaluation comprises 34 benchmark tasks drawn from three suites. The OlmoEarth suite (Herzog et al., 2025; Tseng et al., 2026b) contains 18 tasks over Sentinel-2, Sentinel-1 and Landsat imagery, seven of which originate from GEO-Bench (Lacoste et al., 2023). The UniverSat suite (Perron et al., 2026) contains 11 tasks and adds HLS, NAIP and aerial imagery; Nine tasks have published UniverSat-B scores. We also report our own baseline runs on the two remaining tasks. From GEO-Bench-2 (Simumba et al., 2025), we include five high-resolution tasks spanning 0.2 m aerial imagery, WorldView-3 multispectral imagery and 3 to 4.8 m PlanetScope imagery. Appendix A describes each task and its source (Table 14).

3.3 Baselines

We compare Finch against published results. Published results do not always share our evaluation conditions; where they differ, for example in the readout, the input sources or the temporal coverage, we state the conditions alongside the comparison, without assuming which model they favour. On the OlmoEarth suite, we compare with the strongest published model of each version by 18-task mean: OlmoEarth v1.2 Base (Tseng et al., 2026b) and OlmoEarth v1.3 Base (Tseng et al., 2026a). On the UniverSat suite, we compare with UniverSat-B (Perron et al., 2026). For embedding products, we compare with AlphaEarth Foundations (Brown et al., 2025) and TESSERA (Feng et al., 2025). For GEO-Bench-2, we report the best results on the GEO-Bench-2 leaderboard (GEO-Bench-2, 2025), all of which train a UNet decoder.

3.4 Pretraining Data and Benchmark Overlap

Finch 1.0 was pretrained on imagery alone. Its pretraining used no labelled map data, such as OpenStreetMap or land-cover and crop-type maps, and no labels from any evaluation benchmark. Finch therefore learns its representations without supervision from map products or other label sources. Pretraining imagery was not filtered by evaluation location, so it may include the places and dates of benchmark samples. The NEON, Wyvern and ICEYE results accordingly establish generalization to instruments absent from the training data, not to unseen places.

4 Results

4.1 OlmoEarth Suite

Table 1 compares Finch with OlmoEarth v1.3 Base and v1.2 Base on the 18 tasks of the OlmoEarth suite, together with the best published result of any model on each task. Finch averages 65.1 over the 18 tasks, compared with 65.5 and 65.8, respectively, and achieves state-of-the-art results on two tasks. On CropHarvest China, it classifies crop and non-crop locations from Sentinel-1 radar alone with 70.8% accuracy, 7.1 points above the best published result (63.7, OlmoEarth v1.2 Nano); radar-based cropland mapping is critical in regions where persistent cloud cover hides the growing season from optical sensors. On MADOS, the 15-class segmentation of marine debris, oil spills and other sea-surface features, it achieves 72.0 mIoU against a best published result of 68.5 (OlmoEarth v1.1 Base). Finch also outperforms both OlmoEarth Base models on Cashew Plantation and ForestNet, although other models hold the best published results on these tasks. Most of the remaining gap to OlmoEarth is attributable to CropHarvest Togo and PASTIS crop segmentation.

4.2 UniverSat Suite

Table 2 compares Finch with UniverSat-B on the nine UniverSat tasks for which UniverSat-B reports results. Finch outperforms UniverSat-B on five of the nine tasks and averages 65.9 over these 9 tasks, compared with 66.1. The largest margins are on deforestation-driver classification from Landsat (ForestNet, +6.2), tree-crown segmentation in 0.1 m aerial imagery (NeonTree, +3.2) and solar-panel detection (PV4GER, +1.0); the largest deficit is on HLS burn-scar segmentation (72.2 against 81.5). On the two remaining tasks of the suite, Finch achieves 72.9 on MADOS and 53.5 on PASTIS with class 19 excluded.

Table 1: OlmoEarth suite: Finch 1.0 against the published results of OlmoEarth v1.3 Base and v1.2 Base, and the best published result of any model of any size on each task. All published results are from Table 6 of Tseng et al. (2026a), except OlmoEarth v1.1, from Table 1 of the first version of Tseng et al. (2026b). The readout column gives the readout trained on the frozen embeddings and its number of parameters for Finch and OlmoEarth; kNN has no trained parameters, and the best published results use the same readout type. MADOS is reported in mIoU, its published metric. The best score of each row is in bold.

Task	Input	Readout (Finch / OlmoEarth)	Finch 1.0	OlmoEarth v1.3 Base	OlmoEarth v1.2 Base	Best published	Best published by
BreizhCrops	S2 series	Linear, 11K / 8.5K	72.1	73.2	73.3	**79.1**	OlmoEarth v1.2 Small
CropHarvest China	S1 series	Linear, 4.1K / 3.1K	**70.8**	62.4	58.2	63.7	OlmoEarth v1.2 Nano
CropHarvest China	S2 series	Linear, 4.1K / 3.1K	74.7	77.7	76.7	**81.9**	Galileo Base
CropHarvest China	S1 + S2 series	Linear, 4.1K / 3.1K	77.1	78.3	77.7	**80.3**	OlmoEarth v1 Tiny
CropHarvest Togo	S1 series	Linear, 4.1K / 3.1K	69.3	72.9	74.8	**80.1**	OlmoEarth v1 Base
CropHarvest Togo	S2 series	Linear, 4.1K / 3.1K	79.4	82.0	81.4	**87.3**	OlmoEarth v1 Base
CropHarvest Togo	S1 + S2 series	Linear, 4.1K / 3.1K	74.8	80.1	81.0	**83.7**	OlmoEarth v1 Nano
BigEarthNet	S2	kNN	64.0	**66.1**	65.5	**66.1**	OlmoEarth v1.3 Base
Brick Kiln	S2	kNN	95.3	**96.8**	94.7	**96.8**	OlmoEarth v1.3 Base
Cashew Plantation	S2	Linear, 115K / 86K	30.6	29.1	29.6	**54.1**	DINOv3 Sat 7B
EuroSAT	S2	kNN	93.7	93.4	95.0	**96.4**	OlmoEarth v1 Large
ForestNet	Landsat 8	kNN	46.6	41.7	45.4	**52.3**	Panopticon Base
South Africa Crop Type	S2	Linear, 164K / 123K	30.6	**32.8**	32.0	**32.8**	OlmoEarth v1.3 Base
So2Sat	S2	kNN	63.4	58.2	67.3	**68.2**	OlmoEarth v1 Large
MADOS	S2	Linear, 246K / 185K	**72.0**	67.1	67.2	68.5	OlmoEarth v1.1 Base
PASTIS	S1 series	Linear, 312K / 234K	29.0	**32.6**	31.1	**32.6**	OlmoEarth v1.3 Base
PASTIS	S2 series	Linear, 312K / 234K	50.4	54.4	54.2	**57.5**	OlmoEarth v1.3 Base+P
Sen1Floods11	S1	Linear, 33K / 25K	77.8	79.3	78.9	**79.8**	OlmoEarth v1 Large
Mean			65.1	65.5	**65.8**		

Table 2: UniverSat suite: Finch 1.0 against the published results of UniverSat-B, under the UniverSat protocol (kNN for classification, linear probe for segmentation). The best score of each row is in bold. The last two tasks have no published UniverSat-B result and are not in the mean.

Task	Input	Finch 1.0	UniverSat-B	Finch readout	UniverSat-B readout
AI4SmallFarms	S2 (4 bands)	**41.8**	41.1	Linear, 35K	Linear, 3.1K
Burn Scars	HLS	72.2	**81.5**	Linear, 35K	Linear, 3.1K
Chesapeake	NAIP	63.9	**64.7**	Linear, 461K	Linear, 346K
ForestNet	Landsat 8	**47.9**	41.7	kNN	kNN
PV4GER	Aerial RGB	**93.7**	92.7	kNN	kNN
Sen1Floods11	S1	77.8	**80.1**	Linear, 133K	Linear, 3.1K
NeonTree	Aerial RGB, 0.1 m	**54.2**	51.0	Linear, 3.28M	Linear, 26K
Brick Kiln	S2	**94.7**	94.5	kNN	kNN
PASTIS	S1 + S2 series	47.3	**47.9**	Linear, 330K	Linear, 17K
Mean (9 tasks)		65.9	**66.1**		
MADOS	S2	72.9	–	Linear, 246K	–
PASTIS (class 19 ignored)	S1 + S2 series	53.5	–	Linear, 312K	–

4.3 High-Resolution Tasks

Table 3 reports Finch on the five high-resolution GEO-Bench-2 tasks, which range from 0.2 m aerial imagery to 4.8 m PlanetScope imagery. Finch uses a single linear layer on frozen embeddings, with one probe seed. The published GEO-Bench-2 results instead train a UNet decoder of roughly 15M to 20M parameters, with a feature-pyramid neck for ViT encoders, and average five runs; we list them as a reference rather than as a matched comparison. All published GEO-Bench-2 results in this card are from the leaderboard at revision `f977fd5` (GEO-Bench-2, 2025). With only a linear readout, Finch outperforms TerraMind v1 Large (Jakubik et al., 2025) with a frozen encoder and a UNet decoder on DynamicEarthNet (32.9 against 31.7), and comes within 0.9 points of DINOv3 ConvNeXt-L (Siméoni et al., 2025) with a UNet decoder on FLAIR #2.

Table 3: GEO-Bench-2 high-resolution tasks. Finch uses a single linear layer on frozen embeddings. The published columns give the best result of the GEO-Bench-2 leaderboard for each task, with a frozen encoder and a trained UNet decoder of about 15M to 20M parameters, and with full fine-tuning of the encoder and decoder. The metric is mIoU, except macro F1 for TreeSatAI.

Task	Input	Finch 1.0	Readout	Frozen + UNet	Fine-tuned
FLAIR #2	Aerial, 0.2 m	48.9	Linear, 3.41M	49.8 (DINOv3 ConvNeXt-L)	54.7 (DINOv3 ViT-L/16)
SpaceNet 2	WorldView-3	77.7	Linear, 525K	86.2 (DINOv3 ViT-L/16)	88.3 (DINOv3 ConvNeXt-L)
SpaceNet 7	PlanetScope	54.9	Linear, 525K	61.3 (Clay v1 Base)	63.3 (Clay v1 Base)
DynamicEarthNet	PlanetScope	32.9	Linear, 1.84M	31.7 (TerraMind v1 Large)	34.5 (TerraMind v1 Large)
TreeSatAI	Aerial, 0.2 m	32.8	Linear, 17K	54.1 (DINOv3 ViT-L/16)	66.4 (DINOv3 ViT-L/16)

4.4 Embedding Products

Table 4 compares Finch with published results of two embedding products: AlphaEarth Foundations (Brown et al., 2025) and TESSERA (Feng et al., 2025). On PASTIS crop mapping, Finch achieves 50.4 mIoU against 39.1. The conditions differ: Finch encodes the Sentinel-2 time series of each sample and is read out by a linear layer of 312K parameters, whereas the AlphaEarth annual embeddings summarize Sentinel-1, Sentinel-2 and additional sources over a year and are read out by a linear layer of 1.2K parameters. The comparison is therefore one of usable embedding products rather than an isolated comparison of encoders. The same table of Tseng et al. (2026a) reports 58.3 for TESSERA v2 and 58.6 for the 128-dimensional OlmoEarth v1.3 embeddings, also from annual embeddings with a linear readout; these products outperform Finch on this task under their protocol. Class handling also affects PASTIS results: on the UniverSat version of the task, Finch scores 47.3 with all classes and 53.5 when class 19 is excluded. On PASTIS-R from Sentinel-1 and Sentinel-2, TESSERA (Feng et al., 2025) reports 50.7 mIoU with a lightweight UNet head trained on its 128-dimensional annual embeddings, and the same paper reports 51.1 for the AlphaEarth Foundations embeddings with the same head, while Finch achieves 47.3 with a single linear layer; as the UNet head is the stronger readout, we list these results as a reference.

Table 4: Finch 1.0 against published results of embedding products. The PASTIS results of the products are linear probes on their annual embeddings, as reported by Tseng et al. (2026a). The PASTIS-R results train a lightweight UNet head on the TESSERA and AlphaEarth Foundations embeddings with all training labels, as reported by Feng et al. (2025); this readout is stronger than a linear layer, so these rows are a reference rather than a matched comparison.

Task	Input	Finch 1.0	Product	Product result	Finch readout	Product readout
PASTIS	S2 series	50.4	AlphaEarth	39.1	Linear, 312K	Linear, 1.2K
PASTIS	S2 series	50.4	TESSERA v2	58.3	Linear, 312K	Linear
PASTIS	S2 series	50.4	OlmoEarth v1.3	**58.6**	Linear, 312K	Linear
PASTIS-R	S1 + S2 series	47.3	TESSERA	50.7	Linear, 330K	UNet head
PASTIS-R	S1 + S2 series	47.3	AlphaEarth	**51.1**	Linear, 330K	UNet head

4.5 Patch Size Selected on Validation

Finch can encode imagery at multiple patch sizes. The results above use the standard patch size of each benchmark. Here, we instead select the patch size for each task among the evaluated sizes (4, 8, 16 and the standard size) using validation performance alone, and report the test result of that selection. As shown in Table 5, this raises the averages to 65.7 on the OlmoEarth suite and 66.4 on the nine UniverSat-B tasks, against 65.8 for OlmoEarth v1.2 Base and 66.1 for UniverSat-B. Appendix B reports the result of every task at every patch size (Table 15).

Table 5: Mean test score of Finch 1.0 at the standard patch size of each benchmark and at the patch size selected on validation for each task.

Suite	Tasks	Standard patch size	Selected on validation	Best published baseline
OlmoEarth	18	65.1	65.7	**65.8** (OlmoEarth v1.2 Base)
UniverSat (graded tasks)	9	65.9	**66.4**	66.1 (UniverSat-B)
UniverSat (all tasks)	11	65.4	65.8	–
GEO-Bench-2	5	49.5	50.0	–
All tasks	34	62.9	63.4	–

4.6 Comparison with the OlmoEarth Family

Table 6 places Finch among the OlmoEarth variants listed in Table 6 of Tseng et al. (2026a) on the 18-task suite. At the standard patch size, Finch outperforms every listed OlmoEarth variant except OlmoEarth v1 Base, v1.3 Base and v1.2 Base, which lead by 0.2 to 0.7 points. With patch sizes selected on validation data, Finch outperforms every listed OlmoEarth variant except OlmoEarth v1.2 Base, which leads by 0.1 points.

Table 6: Mean test score on the 18 tasks of the OlmoEarth suite: published OlmoEarth results, all from Table 6 of Tseng et al. (2026a), and Finch 1.0.

Model	Mean	Model	Mean
OlmoEarth v1 Nano	60.5	OlmoEarth v1.2 Small	64.9
OlmoEarth v1 Tiny	62.2	OlmoEarth v1.2 Base	**65.8**
OlmoEarth v1 Base	65.3	OlmoEarth v1.3 Base	65.5
OlmoEarth v1 Large	64.2	OlmoEarth v1.3 Base+P	63.2
OlmoEarth v1.2 Nano	61.6	Finch 1.0, standard patch size	65.1
OlmoEarth v1.2 Tiny	63.0	Finch 1.0, patch size selected on validation	65.7

4.7 Baselines Under Our Evaluation

To validate our evaluation pipeline, we evaluated the released weights of OlmoEarth v1.2 (Nano, Tiny, Small and Base), OlmoEarth v1.3 Base and UniverSat-B with the same protocol as Finch. Table 7 compares these runs with the published results. Our evaluation of OlmoEarth v1.3 Base yields 65.35 against a published 65.45. The published OlmoEarth results are from Table 6 of Tseng et al. (2026a), which updates some OlmoEarth v1.2 results; every reproduction falls within 1.0 point of these means, and within 0.7 points of the means of the original OlmoEarth v1.2 publication. The comparisons in the preceding sections use published results.

Tables 8 and 9 give the result of every task for each of these runs, so that all models are compared under one evaluation pipeline, with the same data, splits, readouts and hyperparameter selection, except where Table 9 states otherwise. Under these conditions, Finch averages 65.1 on the OlmoEarth suite, against 65.4 for OlmoEarth v1.3 Base and 65.3 for OlmoEarth v1.2 Base, and achieves the best result of the six models on six of the 18 tasks. On the nine graded UniverSat tasks, Finch averages 65.9 against 65.6 for UniverSat-B and achieves the better result on five tasks. Individual tasks can differ from their published results: on CropHarvest China from Sentinel-1, our evaluation of OlmoEarth v1.3 Base yields 67.0 against a published 62.4, which is 3.8 points below Finch (70.8).

Table 7: Baselines run through our evaluation code, against their published results.

Model	Suite	Tasks	Our run	Published
OlmoEarth v1.2 Nano	OlmoEarth	18	61.3	61.6
OlmoEarth v1.2 Tiny	OlmoEarth	18	62.9	63.0
OlmoEarth v1.2 Small	OlmoEarth	18	63.9	64.9
OlmoEarth v1.2 Base	OlmoEarth	18	65.3	65.8
OlmoEarth v1.3 Base	OlmoEarth	18	65.4	65.5
UniverSat-B	UniverSat (graded tasks)	9	65.6	66.1

Table 8: OlmoEarth suite: test score of each task for Finch 1.0 and the OlmoEarth models, all evaluated by our pipeline with the same data, splits, readouts and hyperparameter selection, at the standard patch size of the benchmark. The best score of each row is in bold.

			OlmoEarth v1.2				OlmoEarth v1.3
Task	**Input**	**Finch 1.0**	**Nano**	**Tiny**	**Small**	**Base**	**Base**
BreizhCrops	S2 series	72.1	67.1	71.2	72.5	**73.5**	**73.5**
CropHarvest China	S1 series	**70.8**	63.4	58.1	63.2	61.7	67.0
CropHarvest China	S2 series	74.7	79.7	**80.0**	78.4	76.8	78.1
CropHarvest China	S1 + S2 series	77.1	**79.4**	79.0	77.5	72.7	77.5
CropHarvest Togo	S1 series	69.3	75.2	76.5	71.6	**77.8**	73.2
CropHarvest Togo	S2 series	79.4	80.7	**81.4**	78.8	78.1	81.0
CropHarvest Togo	S1 + S2 series	74.8	78.1	81.0	77.5	**83.0**	77.8
BigEarthNet	S2	64.0	59.4	62.0	64.6	65.1	**66.1**
Brick Kiln	S2	**95.3**	93.2	92.3	90.7	93.4	95.2
Cashew Plantation	S2	**30.6**	26.6	26.7	28.3	27.7	29.2
EuroSAT	S2	**93.7**	87.5	88.4	90.3	91.6	91.0
ForestNet	Landsat 8	**46.6**	41.8	40.8	45.2	43.8	41.7
South Africa Crop Type	S2	30.6	22.1	23.9	26.9	31.4	**32.3**
So2Sat	S2	63.4	55.0	58.8	65.7	**67.2**	58.0
MADOS	S2	**72.0**	57.8	63.0	61.3	67.3	68.5
PASTIS	S1 series	29.0	22.2	24.8	28.0	31.3	**32.4**
PASTIS	S2 series	50.4	35.8	45.1	50.6	54.1	**54.3**
Sen1Floods11	S1	77.8	78.8	78.8	79.3	79.3	**79.6**
Mean		65.1	61.3	62.9	63.9	65.3	**65.4**

Table 9: UniverSat suite: test score of each task for Finch 1.0 and UniverSat-B, both evaluated by our pipeline with the same data and readouts. UniverSat-B uses the input settings and splits of its published evaluation; on MADOS, these select its hyperparameters on the test split, and Finch selects on the validation split, so this row is not a matched comparison. The best score of each row is in bold. The last two tasks have no published UniverSat-B result and are not in the mean.

Task	Input	Finch 1.0	UniverSat-B
AI4SmallFarms	S2 (4 bands)	**41.8**	41.1
Burn Scars	HLS	72.2	**82.2**
Chesapeake	NAIP	63.9	**64.1**
ForestNet	Landsat 8	**47.9**	41.5
PV4GER	Aerial RGB	**93.7**	92.7
Sen1Floods11	S1	77.8	**78.0**
NeonTree	Aerial RGB, 0.1 m	**54.2**	48.7
Brick Kiln	S2	**94.7**	94.2
PASTIS	S1 + S2 series	47.3	**47.6**
Mean (9 tasks)		**65.9**	65.6
MADOS	S2	**72.9**	52.5
PASTIS (class 19 ignored)	S1 + S2 series	53.5	**54.4**

4.8 Hyperspectral Input

NeonTree (Weinstein et al., 2020; Lacoste et al., 2023) also provides imagery from the NEON imaging spectrometer, with 369 spectral bands spanning 0.38 to 2.47 µm at 1 m resolution. This sensor is absent from the training data of Finch, whereas UniverSat-B was pretrained on NEON imagery. We evaluate both models on the RGB imagery, the hyperspectral bands and their combination, with the same linear-probe protocol and data splits. Table 10 reports the results.

On the hyperspectral bands alone, Finch and UniverSat-B both score 46.9 test mIoU, although Finch has never seen the sensor. Finch outperforms UniverSat-B by 7.0 points on RGB and by 3.4 points on the combined input. For both models, adding the hyperspectral bands lowers the RGB result on this task. We hypothesize that the 1 m spectra cannot resolve the boundaries of tree crowns labelled at 0.1 m; no ablation isolates this cause. We were unable to evaluate on SpectralEarth (Braham et al., 2024), the hyperspectral benchmark on which UniverSat reports results, because the underlying EnMAP data could not be obtained for this release.

Table 10: NeonTree tree-crown segmentation from RGB, hyperspectral (HSI) and both inputs, mIoU on the validation and test splits. Both models use the same linear-probe protocol and data splits; the number of its parameters is given in parentheses, and depends on the embedding width and on the number of label pixels under each embedding. Finch encodes at patch sizes 8 and 16 of the finest input; the HSI alone has no score at patch size 16, because its 40 px images do not divide into whole cells of the 400 px labels.

Input	**Finch 1.0, patch 8** val / test	**Finch 1.0, patch 16** val / test	**UniverSat-B** val / test
RGB, 0.1 m	54.0 / 54.4 (131K)	55.2 / **55.4** (525K)	47.6 / 48.4 (25K)
HSI, 369 bands, 1 m	46.4 / **46.9** (13.1M)	–	47.2 / **46.9** (2.46M)
RGB + HSI	49.4 / 49.6 (131K)	50.5 / **51.3** (525K)	46.8 / 47.9 (25K)

4.9 Additional GEO-Bench-2 Tasks

We further evaluate Finch on two GEO-Bench-2 tasks that test capabilities outside the core suites: agricultural field-boundary segmentation across 14 countries from two Sentinel-2 acquisitions per season (Fields of the World, Kerner et al., 2025), and cloud and cloud-shadow segmentation in Sentinel-2 imagery (CloudSEN12+, Aybar et al., 2024). Table 11 reports the results with a single linear layer on frozen embeddings, one probe seed and the patch size selected on validation data; the published results train a UNet decoder and average five runs. On CloudSEN12+, Finch outperforms every frozen encoder paired with a trained UNet decoder on the GEO-Bench-2 leaderboard. On Fields of the World, Finch trails the best frozen encoder with a UNet decoder by 3.0 points, as a linear readout on 80 m cells cannot fully resolve field boundaries that are one to two pixels wide.

Table 11: Additional GEO-Bench-2 tasks, mIoU on the test split. Finch uses a single linear layer on frozen embeddings at patch size 8, selected on validation data. The published columns give the best result of the GEO-Bench-2 leaderboard with a frozen encoder and a trained UNet decoder of about 15M to 20M parameters, and with full fine-tuning.

Task	**Input**	**Finch 1.0**	**Finch readout**	**Frozen + UNet**	**Fine-tuned**
Fields of the World	S2, 2 dates	57.3	Linear, 197K	60.3 (Clay v1 Base)	60.8 (Clay v1 Base)
CloudSEN12+	S2	**63.7**	Linear, 262K	63.5 (SSL4EO-S12 ResNet-50)	70.2 (DINOv3 ConvNeXt-L)

4.10 Results by Task Group

Table 12 gives the average Finch result over the tasks of each group. The groups combine different metrics (accuracy, F1 and mIoU) across different datasets, so these averages summarize the benchmarks and are not comparable across groups.

4.11 Embedding Maps

Figure 1 visualizes Finch 1.0 embeddings for two locations in the southern United States, computed from public NAIP aerial imagery at 0.6 m and Sentinel-2 and Sentinel-1 imagery at 10 m. For each input, we project the embeddings onto

Table 12: Mean test score of Finch 1.0 by task group, at the standard patch size.

Task group	Tasks	Finch 1.0
All tasks	34	62.9
Classification and multi-label	16	71.9
Segmentation	18	54.9
Sentinel-2, Sentinel-1, Landsat and HLS inputs	26	64.6
Aerial and commercial high-resolution inputs	8	57.4

their first three principal components and display them as red, green and blue, so that regions the model represents similarly share a colour. Without any supervision, the maps delineate roads, forest stands, clearings and field boundaries, and the Sentinel-2 and Sentinel-1 maps of each location recover the same principal structures, although one sensor is optical and the other is radar. Appendix C presents larger maps from six imagery sources, two of which are absent from the training data.

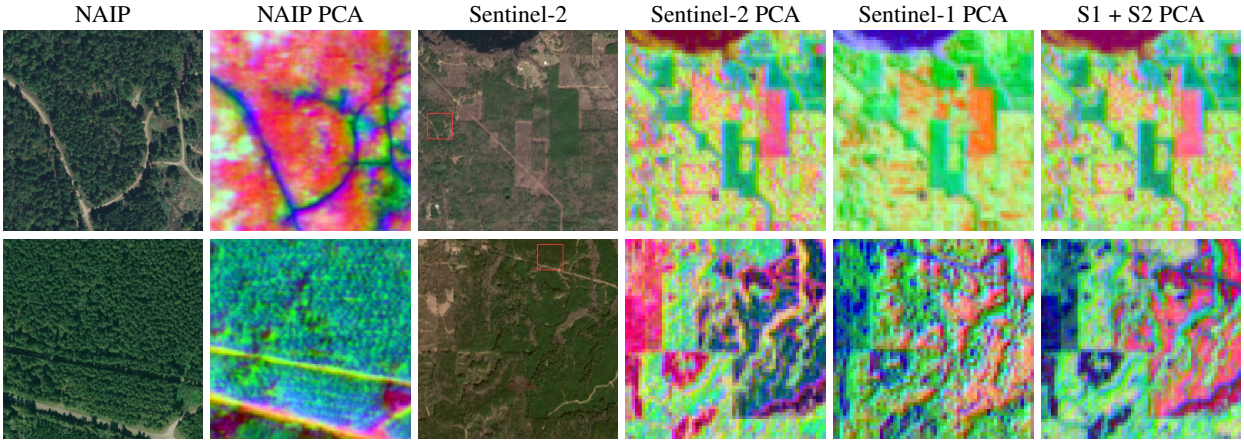


Figure 1: PCA maps of Finch 1.0 embeddings. Top: northwest Louisiana, 2020. Bottom: southwest Mississippi, 2018. The NAIP views are about 300 m across, and the Sentinel views are 2.4 km across; the red square on each Sentinel-2 image marks the NAIP view. Each map shows the first three principal components of the embeddings as red, green and blue, at the resolution of the embeddings. The principal components are fitted separately for each panel, so colours are not comparable across panels. NAIP: USDA Farm Production and Conservation Business Center (public domain). Contains modified Copernicus Sentinel data 2018 and 2020.

5 Released Embeddings

We release one Finch 1.0 embedding dataset per core task, covering every sample of the train, validation and test splits of each task. The release is limited to these samples and does not include embeddings of other locations. Each embedding is keyed by the sample identifier and split of its benchmark, allowing users to join it with the labels distributed with the original dataset. The released files support re-evaluation of the 34 core tasks at their benchmark-standard patch sizes with the corresponding readout protocol, as well as the evaluation of new readouts and metrics. The embeddings are licensed under CC BY 4.0. Table 13 summarizes the release.

Generation. The embeddings were produced with the same code that produced the results in this card. The frozen encoder of Finch 1.0 encodes every sample at the standard patch size of its benchmark, with the location of the sample where the dataset provides one. For each task, we release the embedding that its reported result uses: a single 1024-dimensional vector per sample for classification and multi-label tasks, and a grid of 1024-dimensional embeddings per sample for segmentation tasks. We read back each released dataset and verified that it is bit-identical to the embeddings used in the original evaluation. We then ran the readout protocols of this card again on the released files. This changed test results by up to 1.5 points on the six CropHarvest tasks, by 0.6 points on PV4GER and by 0.4 points

on ForestNet (OlmoEarth suite). All other results were within 0.1 points of the reported results. The embeddings are stored in float16, the precision in which the encoder outputs them.

Table 13: Contents of the embedding release.

Property	Value
Content	One embedding dataset per core task of Table 14, covering every sample, at the standard patch size of each benchmark, and of no other places
Keys	Split, the sample identifier of the original dataset, and the tile index for tasks whose protocol divides a sample into tiles
Embeddings	For each task, the embedding that its reported result uses: one pooled vector per sample for classification and multi-label tasks, and the embedding grid per sample for segmentation tasks. Bit-identical to the embeddings used in evaluation. Observed score differences are reported in Section 5
Values	1024 dimensions, float16
Format	One Lance dataset per task, with the keys as columns and each embedding as a fixed-size float16 vector column, readable with `pylance`, PyArrow and DuckDB
Host	`https://embeddings.locamage.com`, one dataset per task at `<benchmark>/<task>.lance`
Terms of use	CC BY 4.0, as for the AlphaEarth Foundations embeddings. The source imagery and labels remain under the licences of their datasets
Manifest	`https://embeddings.locamage.com/manifest.json` lists every file of each task with its size and SHA-256 checksum

Access. Each task is a Lance dataset with the columns `split` (`train`, `val` or `test`), `sample` (the sample identifier of the original dataset), `tile` (0 for an untiled sample) and `embedding`. The schema metadata gives the `kind` of the embedding (`pooled` or `grid`) and its `shape` (`[1024]`, or `[h, w, 1024]` in row-major order for a grid). The full release is 207.6 GB. The following example downloads one small task (EuroSAT, 8 MB), verifies each file against the manifest and loads its training embeddings; the same code reads a segmentation task, such as `olmoearth/mados`, as a grid of embeddings per sample:

```python
import hashlib, json, pathlib
import lance, numpy as np, requests

BASE, TASK = "https://embeddings.locamage.com", "olmoearth/m-eurosat"
response = requests.get(f"{BASE}/manifest.json")
response.raise_for_status()
manifest = response.json()
for entry in manifest["tasks"][TASK]["files"]:
    response = requests.get(f"{BASE}/{entry['path']}")
    response.raise_for_status()
    data = response.content
    assert hashlib.sha256(data).hexdigest() == entry["sha256"]
    path = pathlib.Path("finch-1.0") / entry["path"]
    path.parent.mkdir(parents=True, exist_ok=True)
    path.write_bytes(data)

dataset = lance.dataset(f"finch-1.0/{TASK}.lance")
shape = json.loads(dataset.schema.metadata[b"shape"])
table = dataset.to_table(filter="split = 'train'")
embeddings = np.stack(table["embedding"].to_numpy(zero_copy_only=False))
embeddings = embeddings.reshape(-1, *shape)
```

6 Safety, Security, and Responsibility

6.1 Intended Use

Finch embeddings are intended for land-cover, crop, forest and wetland mapping with limited labels, change detection, similarity search, and research that requires consistent features across sensors. They are not intended for safety-critical decisions without local validation, or for the identification or tracking of individuals.

6.2 Privacy and Misuse

The released embeddings cover only samples from public benchmark datasets whose imagery is already publicly available. Similarity search over Finch embeddings can retrieve locations that resemble a given site, including sensitive sites. Readouts trained on few labels may be confidently wrong, and we recommend validation against local data before deployment. Satellite coverage, label availability and benchmark locations are unevenly distributed across regions; we report results only where benchmarks provide labels and make no claims about performance elsewhere.

7 Discussion

Finch 1.0 is a multimodal foundation model trained on satellite, aerial and drone imagery from 50 sensors. It achieves state-of-the-art results on crop and non-crop classification from Sentinel-1 radar alone (CropHarvest China) and on the segmentation of marine debris, oil spills and other sea-surface features (MADOS), and with a single linear layer it outperforms every frozen encoder paired with a trained UNet decoder on DynamicEarthNet land-cover segmentation. With only a kNN classifier or a linear layer on frozen embeddings, Finch performs within 0.7 points of the strongest published OlmoEarth and UniverSat models on their own suites; with patch sizes selected on validation data, it exceeds UniverSat-B and comes within 0.1 points of OlmoEarth v1.2 Base. Under our evaluation pipeline, it averages within 0.3 points of OlmoEarth v1.3 Base and above UniverSat-B. With linear readouts, it outperforms the AlphaEarth Foundations embeddings by 11.3 points on PASTIS under the conditions stated in Section 4, and on hyperspectral-only NeonTree it scores 46.9 test mIoU, the same as UniverSat-B, which was pretrained on that sensor. We release one embedding dataset per core task, so that others can run the corresponding protocols again or evaluate new readouts.

Contributor and Acknowledgments

Contributor

Pinak Paliwal

Acknowledgments

I thank my father, my mother and my sister for their constant guidance and support over the last 19 years. We thank a16z Speedrun Alpha and Bryan Kim of Mido Capital for their support, and Google's TPU Research Cloud and TPU Builder Program for the compute that made this work possible. We also thank the teams behind OlmoEarth, UniverSat, GEO-Bench, GEO-Bench-2, AlphaEarth Foundations, DINOv3 and TerraMind for publishing their models and results, and the creators of every evaluation dataset in Table 14.

References

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*, 2022:1–22, 2022.

Cesar Aybar, Lesly Bautista, David Montero, Julio Contreras, Daryl Ayala, Fernando Prudencio, Jhomira Loja, Luis Ysuhuaylas, Fernando Herrera, Karen Gonzales, et al. CloudSEN12+: The largest dataset of expert-labeled pixels for cloud and cloud shadow detection in Sentinel-2. *Data in Brief*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 210–211, 2020.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: Composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *arXiv preprint arXiv:2408.08447*, 2024.

Christopher F. Brown, Michal R. Kazmierski, Valerie J. Pasquarella, William J. Rucklidge, et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Zhengpeng Feng et al. TESSERA: Temporal embeddings of surface spectra for earth representation and analysis. *arXiv preprint arXiv:2506.20380*, 2025.

Anatol Garioud, Apolline De Wit, Marc Poupée, Marion Valette, Sébastien Giordano, and Boris Wattrelos. FLAIR #2: Textural and temporal information for semantic segmentation from multi-source optical imagery. *arXiv preprint arXiv:2305.14467*, 2023.

GEO-Bench-2. GEO-Bench-2 leaderboard. https://huggingface.co/spaces/aialliance/GEO-Bench-2-Leaderboard, revision f977fd5149aad5ba0fdcd9e25d57b858fd7cd13d, 2025. Accessed November 2026.

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*, 2019.

Henry Herzog, Favyen Bastani, Yawen Zhang, et al. OlmoEarth: Stable latent image modeling for multimodal Earth observation. *arXiv preprint arXiv:2511.13655*, 2025.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, et al. Liger Kernel: Efficient Triton kernels for LLM training. *arXiv preprint arXiv:2410.10989*, 2024.

Jeremy Irvin, Hao Sheng, Neel Ramachandran, Sonja Johnson-Yu, Sharon Zhou, Kyle Story, Rose Rustowicz, Cooper Elsworth, Kemen Austin, and Andrew Y. Ng. ForestNet: Classifying drivers of deforestation in Indonesia using deep learning on satellite imagery. *arXiv preprint arXiv:2011.05479*, 2020.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, et al. TerraMind: Large-scale generative multimodality for Earth observation. *arXiv preprint arXiv:2504.11171*, 2025.

Z. Jin, C. Lin, C. Weigl, J. Obarowski, and D. Hale. Smallholder cashew plantations in Benin, version 1.0. Radiant MLHub, 2021.

Hannah Kerner, Snehal Chaudhari, Aninda Ghosh, Caleb Robinson, Adeel Ahmad, Eddie Choi, Nathan Jacobs, Chris Holmes, Matthias Mohr, Rahul Dodhia, et al. Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025.

Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, et al. GEO-Bench: Toward foundation models for Earth monitoring. *arXiv preprint arXiv:2306.03831*, 2023.

Jihyeon Lee, Nina R. Brooks, Fahim Tajwar, Marshall Burke, Stefano Ermon, David B. Lobell, Debashish Biswas, and Stephen P. Luby. Scalable deep learning to identify brick kilns and aid regulatory capacity. *Proceedings of the National Academy of Sciences*, 118(17):e2018863118, 2021.

Locamage. xprof-rs: A Rust reader for XProf profiles, 2026. URL https://github.com/Locamage/xprof-rs.

Kevin Mayer, Benjamin Rausch, Marie-Louise Arlt, Gunther Gust, Zhecheng Wang, Dirk Neumann, and Ram Rajagopal. 3D-PV-Locator: Large-scale detection of rooftop-mounted photovoltaic systems in 3D. *Applied Energy*, 310:118469, 2022.

Yohann Perron, Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. UniverSat: Resolution- and modality-agnostic transformers for Earth observation. *arXiv preprint arXiv:2606.23503*, 2026.

Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. AI4SmallFarms: A dataset for crop field delineation in Southeast Asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 2023.

Christopher Phillips, Sujit Roy, Kumar Ankur, and Rahul Ramachandran. HLS foundation burn scars dataset, 2023.

Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019.

Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. BreizhCrops: A satellite time series dataset for crop type identification. In *Proceedings of the International Conference on Machine Learning Time Series Workshop*, 2019.

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.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Naomi Simumba, Nils Lehmann, Paolo Fraccaro, Hamed Alemohammad, et al. GEO-Bench-2: From performance to capability, rethinking evaluation in geospatial AI. *arXiv preprint arXiv:2511.15658*, 2025.

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

The Marin Community. Marin: An open lab for building foundation models. https://github.com/marin-community/marin, 2025. Started at the Stanford Center for Research on Foundation Models.

Tokamax. Tokamax: A GPU and TPU kernel library, 2025. URL https://github.com/openxla/tokamax.

Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022.

Gabriel Tseng, Ivan Zvonkov, Catherine Lilian Nakalembe, and Hannah Kerner. CropHarvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Gabriel Tseng, Favyen Bastani, Patrick Alan Johnson, et al. OlmoEarth v1.3: Embeddings. *arXiv preprint arXiv:2610.32294*, 2026a.

Gabriel Tseng, Yawen Zhang, Favyen Bastani, Henry Herzog, Joseph Redmon, et al. OlmoEarth v1.2: A more efficient family of OlmoEarth models. *arXiv preprint arXiv:2605.20804*, 2026b.

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2021.

Ben G. Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, Aditya Singh, Sarah J. Graves, and Ethan White. NEON crowns: A remote sensing derived dataset of 100 million individual tree crowns. *bioRxiv*, 2020.

Western Cape Department of Agriculture. Crop type classification dataset for Western Cape, South Africa. Radiant MLHub, 2021.

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, et al. So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, 2020.

A Evaluation Details

Table 14 describes each evaluation task, including its content, source, input, primary metric and readout.

Table 14: Description of each task. The metric is the primary metric of the task, and the readout is the readout of its published protocol.

Benchmark	Description	Details
OlmoEarth suite (Herzog et al., 2025; Tseng et al., 2026b)		
BreizhCrops	Crop type classification of parcels in Brittany, France, 9 classes (Rußwurm et al., 2019).	Monthly Sentinel-2 series. Accuracy, linear probe.
CropHarvest China	Crop or non-crop classification of points in China, 2 classes (Tseng et al., 2021).	Monthly Sentinel-1, Sentinel-2, or both. Accuracy, linear probe.
CropHarvest Togo	Crop or non-crop classification of points in Togo, 2 classes (Tseng et al., 2021).	Monthly Sentinel-1, Sentinel-2, or both. Accuracy, linear probe.
BigEarthNet	Multi-label land cover classification of patches in Europe, 43 classes (Sumbul et al., 2019); GEO-Bench version (Lacoste et al., 2023).	Sentinel-2. Micro F1, kNN.
Brick Kiln	Detection of brick kilns, 2 classes (Lee et al., 2021); GEO-Bench version.	Sentinel-2. Accuracy, kNN.
Cashew Plantation	Segmentation of cashew plantations and land cover in Benin, 7 classes (Jin et al., 2021); GEO-Bench version.	Sentinel-2. mIoU, linear probe.
EuroSAT	Land use and land cover classification in Europe, 10 classes (Helber et al., 2019); GEO-Bench version.	Sentinel-2. Accuracy, kNN.
ForestNet	Classification of deforestation drivers in Indonesia, 12 classes (Irvin et al., 2020); GEO-Bench version.	Landsat 8. Accuracy, kNN.
South Africa Crop Type	Crop type segmentation in the Western Cape, South Africa, 10 classes (Western Cape Department of Agriculture, 2021); GEO-Bench version.	Sentinel-2. mIoU, linear probe.
So2Sat	Local climate zone classification, 17 classes (Zhu et al., 2020); GEO-Bench version.	Sentinel-2. Accuracy, kNN.
MADOS	Segmentation of marine debris, oil spills and sea surface features, 15 classes (Kikaki et al., 2024).	Sentinel-2. mIoU (the published metric; the protocol's primary metric is micro F1), linear probe.
PASTIS	Crop type segmentation of parcels in France, 19 classes (Sainte Fare Garnot and Landrieu, 2021).	Monthly Sentinel-1 or Sentinel-2 series. mIoU, linear probe.

Continued on next page

Benchmark	Description	Details
Sen1Floods11	Segmentation of flood water, 2 classes (Bonafilia et al., 2020).	Sentinel-1. mIoU, linear probe.
UniverSat suite (Perron et al., 2026)		
AI4SmallFarms	Crop field delineation in Southeast Asian smallholder farms, 2 classes (Persello et al., 2023).	Sentinel-2, 4 bands. mIoU, linear probe.
Burn Scars	Segmentation of burn scars in the contiguous United States, 2 classes (Phillips et al., 2023).	HLS. mIoU, linear probe.
NeonTree	Segmentation of tree crowns, 2 classes (Weinstein et al., 2020); GEO-Bench version.	Aerial RGB at 0.1 m. mIoU, linear probe.
Brick Kiln	Detection of brick kilns, 2 classes (Lee et al., 2021); GEO-Bench version.	Sentinel-2. Accuracy, kNN.
Chesapeake	Land cover segmentation in the Chesapeake Bay area, 7 classes (Robinson et al., 2019); GEO-Bench version.	NAIP. mIoU, linear probe.
ForestNet	Classification of deforestation drivers in Indonesia, 12 classes (Irvin et al., 2020); GEO-Bench version.	Landsat 8. Accuracy, kNN.
PV4GER	Detection of rooftop solar panels in Germany, 2 classes (Mayer et al., 2022); GEO-Bench version.	Aerial RGB. Accuracy, kNN.
MADOS	Segmentation of marine debris, oil spills and sea surface features, 15 classes (Kikaki et al., 2024).	Sentinel-2. mIoU, linear probe.
PASTIS	Crop type segmentation of parcels in France, 20 classes (Sainte Fare Garnot and Landrieu, 2021).	Sentinel-2 and Sentinel-1 series. mIoU, linear probe.
PASTIS (class 19 ignored)	As PASTIS, with class 19 ignored, 19 classes (Sainte Fare Garnot and Landrieu, 2021).	Sentinel-2 and Sentinel-1 series. mIoU, linear probe.
Sen1Floods11	Segmentation of flood water, 2 classes (Bonafilia et al., 2020).	Sentinel-1. mIoU, linear probe.
GEO-Bench-2 suite (Simumba et al., 2025)		
FLAIR #2	Land cover segmentation in France, 13 classes (Garioud et al., 2023).	Aerial imagery at 0.2 m. mIoU, linear probe.
SpaceNet 2	Building footprint segmentation in Las Vegas, Paris, Shanghai and Khartoum, 2 classes (Van Etten et al., 2018).	WorldView-3 at 0.3 m. mIoU, linear probe.
SpaceNet 7	Building segmentation in monthly mosaics of 101 places, 2 classes (Van Etten et al., 2021).	PlanetScope at about 4.8 m. mIoU, linear probe.
DynamicEarthNet	Land cover segmentation from daily imagery, 7 classes (Toker et al., 2022).	PlanetScope at 3 m, 6 daily frames. mIoU, linear probe.
TreeSatAI	Multi-label tree species classification in Germany, 15 classes (Ahlswede et al., 2022).	Aerial imagery at 0.2 m. Macro F1, linear probe.

B Scores at Each Patch Size

Table 15 reports the test result of Finch 1.0 on each task at patch sizes 4, 8 and 16, at the standard patch size of the benchmark, and at the patch size selected on the validation split. A dash denotes a patch size that was not evaluated.

Table 15: Test score of Finch 1.0 at each patch size. "Selected" is the patch size with the best validation score, and the last column is its test score.

Suite	Task	4	8	16	Standard	Selected	Test at selected
OlmoEarth	BreizhCrops	72.1	72.2	72.1	72.1	16	72.1
	CropHarvest China, S1	70.8	72.1	69.9	70.8	16	69.9
	CropHarvest China, S2	72.0	72.6	74.2	74.7	8	72.6
	CropHarvest China, S1 + S2	77.1	76.9	77.1	77.1	16	77.1
	CropHarvest Togo, S1	69.9	69.6	70.3	69.3	16	70.3

Continued on next page

Suite	Task	4	8	16	Standard	Selected	Test at selected
	CropHarvest Togo, S2	79.4	75.8	78.4	79.4	8	75.8
	CropHarvest Togo, S1 + S2	78.8	75.8	74.8	74.8	4	78.8
	BigEarthNet	64.0	63.7	63.0	64.0	4	64.0
	Brick Kiln	95.3	93.4	92.0	95.3	4	95.3
	Cashew Plantation	30.6	32.7	43.2	30.6	16	43.2
	EuroSAT	93.7	93.9	92.0	93.7	4	93.7
	ForestNet	46.6	46.6	45.8	46.6	16	45.8
	South Africa Crop Type	30.6	30.8	29.2	30.6	4	30.6
	So2Sat	63.4	57.9	51.8	63.4	4	63.4
	MADOS	72.0	65.2	50.0	72.0	4	72.0
	PASTIS, S1	29.0	26.8	25.2	29.0	4	29.0
	PASTIS, S2	50.4	46.6	39.6	50.4	4	50.4
	Sen1Floods11	77.8	77.6	77.2	77.8	4	77.8
UniverSat	AI4SmallFarms	41.8	41.1	40.3	41.8	8	41.1
	Burn Scars	72.2	74.2	74.4	72.2	8	74.2
	NeonTree	56.5	55.0	54.8	54.2	4	56.5
	Brick Kiln	96.7	93.7	89.9	94.7	4	96.7
	Chesapeake	64.5	63.9	61.6	63.9	4	64.5
	ForestNet	46.1	46.6	45.3	47.9	16	45.3
	PV4GER	93.9	94.2	94.4	93.7	8	94.2
	MADOS	72.9	67.8	54.7	72.9	4	72.9
	PASTIS	47.3	44.3	37.6	47.3	4	47.3
	PASTIS (class 19 ignored)	53.5	50.0	42.4	53.5	4	53.5
	Sen1Floods11	80.3	77.8	76.9	77.8	8	77.8
GEO-Bench-2	FLAIR #2	–	47.7	48.9	48.9	16	48.9
	SpaceNet 2	–	–	77.7	77.7	16	77.7
	SpaceNet 7	–	57.4	54.9	54.9	8	57.4
	DynamicEarthNet	–	32.4	32.9	32.9	16	32.9
	TreeSatAI	29.6	33.2	32.8	32.8	16	32.8

C Embedding Maps of Open Imagery

This appendix presents Finch 1.0 embedding maps of open imagery from seven sensor and location pairs. Two of the sensors are absent from the training data: the Wyvern Dragonette hyperspectral imager (23 bands) and the ICEYE X-band synthetic aperture radar at 0.25 m. Finch encodes both without fine-tuning.

Each image is read at its native resolution, without resampling. Large scenes are encoded as tiles, each with the location of its centre, and the tile embeddings are assembled into a single map. Each map displays the first three principal components of the 1024-dimensional embeddings of one scene as red, green and blue, so that regions the model represents similarly share a colour. The principal components are fitted separately for each scene, so colours are not comparable across maps. Areas outside a scene are shown in white. Table 16 lists the location and dates of each map.

Table 16: The maps of this appendix. The Delta is the Sacramento–San Joaquin Delta, California. A date range gives the first and last of four frames. "Embeddings" is the size of the embedding map.

Place	Sensor	Centre	Dates	Pixel	Embeddings
Berkeley, CA	NAIP	37.872, −122.259	2024-05-20	0.6 m	625 × 625
Delta, CA	Sentinel-2 L2A	38.007, −121.587	2026-07-17 to 11-11	10 m	600 × 600
	Sentinel-1 RTC	38.007, −121.587	2026-11-05 to 11-12	10 m	600 × 600
	Landsat 8/9 L2	38.007, −121.587	2026-06-07 to 10-13	30 m	200 × 200
	Wyvern Dragonette	38.007, −121.587	2025-05-05	5.26 m	3125 × 1573
Black Rock, NV	ICEYE X35	40.786, −119.204	2024-08-30	0.25 m	1250 × 1250
Sydney, AU	ICEYE X53	−33.853, 151.214	2025-12-23	0.25 m	1250 × 1250

NAIP, 0.6 m



Finch 1.0 embeddings

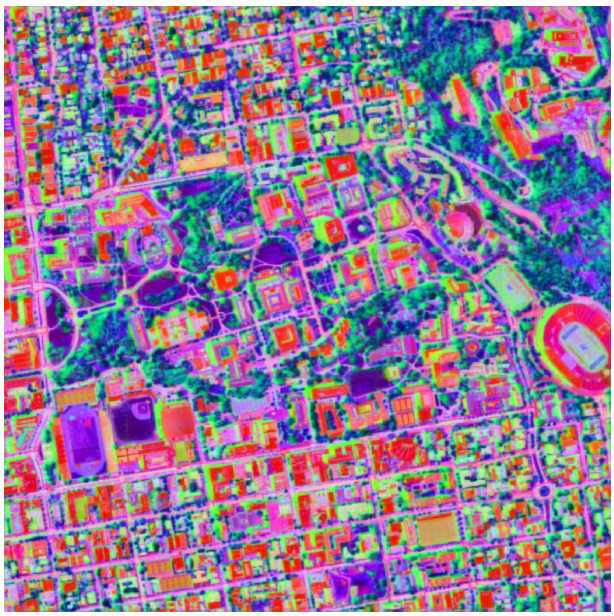


Figure 2: University of California, Berkeley, 1.5 km across, from NAIP imagery of 2024-05-20 at 0.6 m. Each embedding covers 2.4 m. The map separates roofs, trees, lawns, roads, sports fields and the stadium. NAIP: USDA Farm Production and Conservation Business Center (public domain).

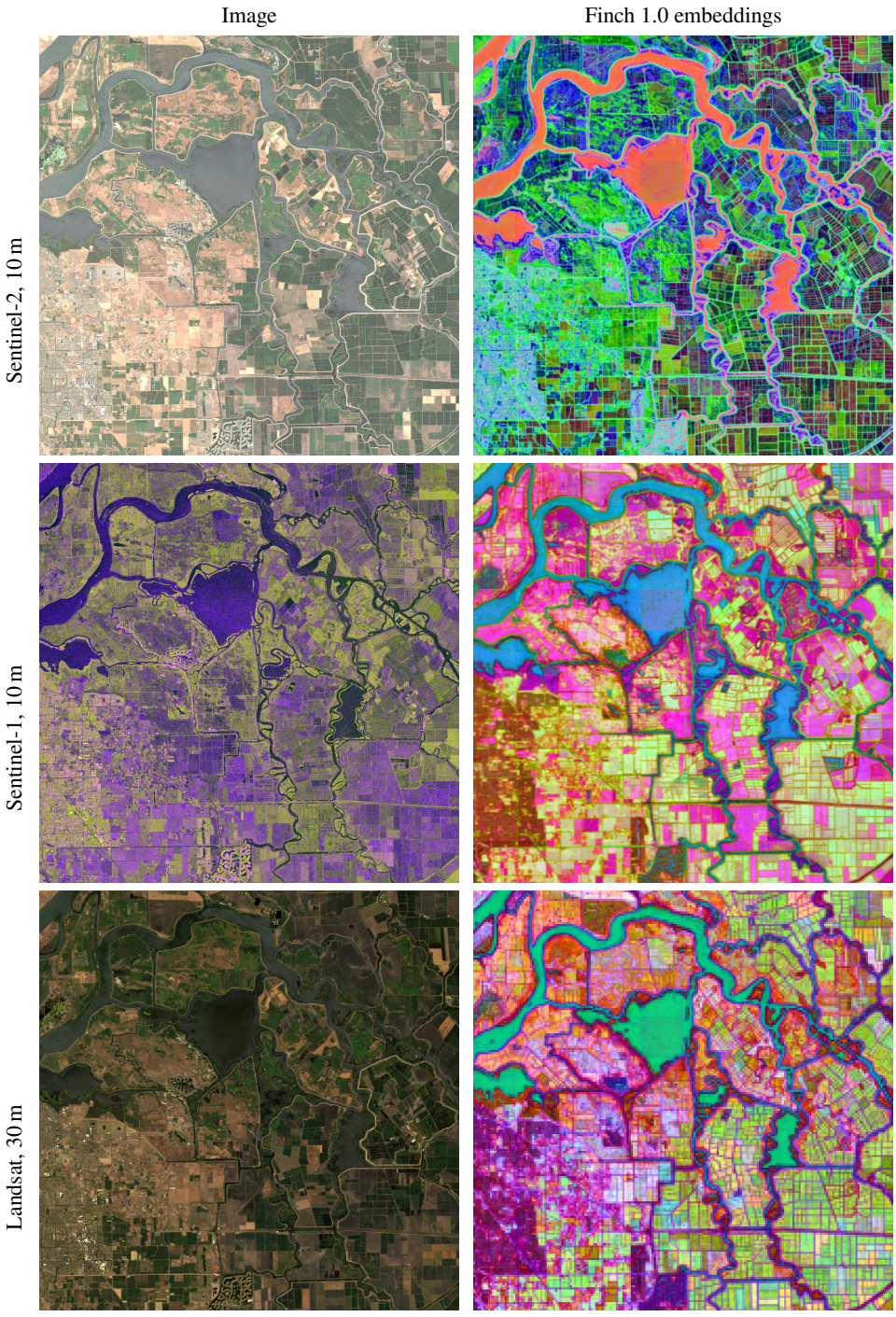


Figure 3: The Sacramento–San Joaquin Delta, California, 24 km across, from Sentinel-2 (true colour), Sentinel-1 (VV, VH and their difference) and Landsat 8/9 (true colour). Each map encodes four frames of 2026; each image shows the first frame. Each embedding covers 40 m for Sentinel and 120 m for Landsat. The three maps show the same channels, islands, wetlands, fields and towns. Contains modified Copernicus Sentinel data 2026. Landsat imagery courtesy of the U.S. Geological Survey.

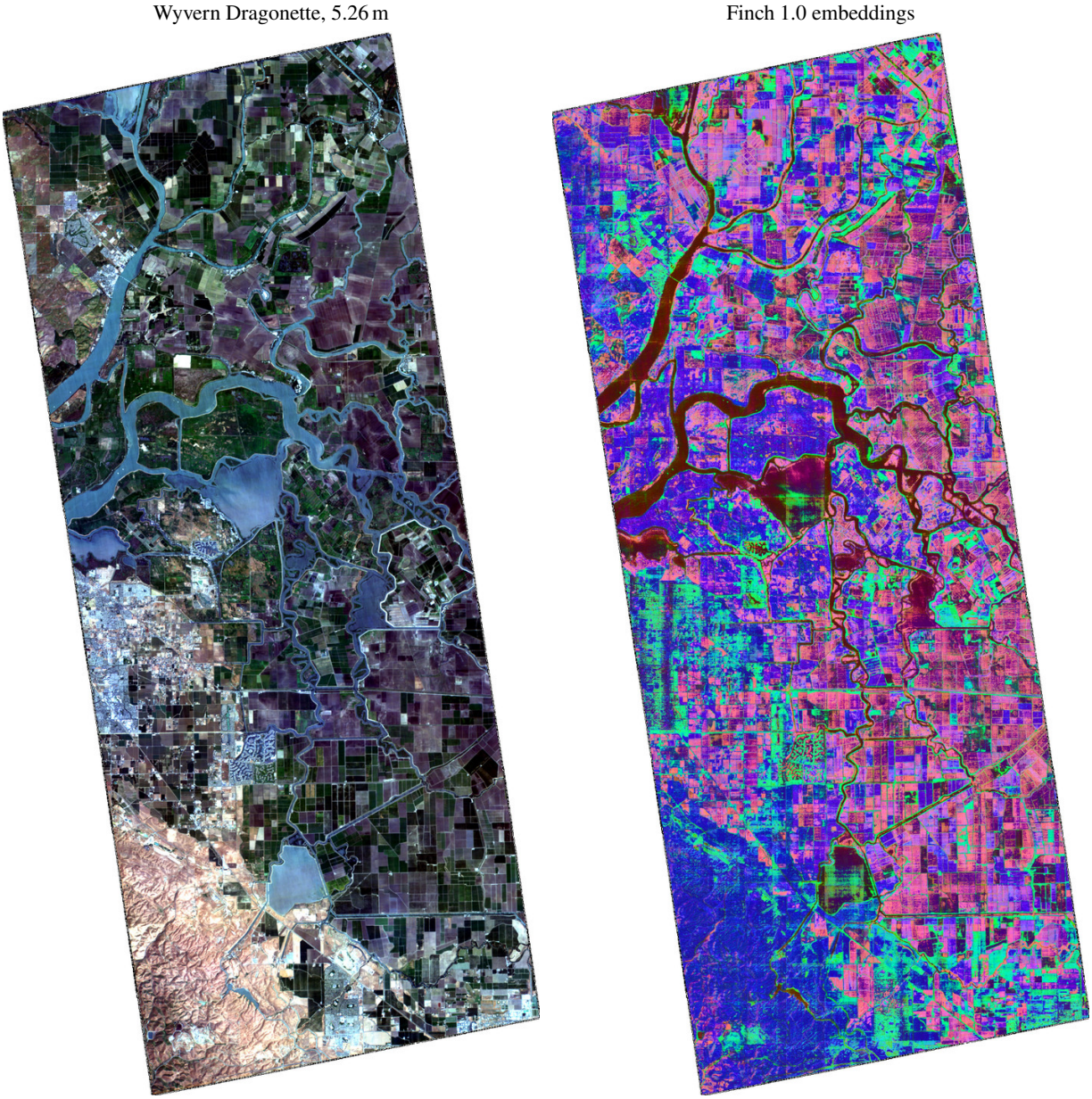


Figure 4: A full Wyvern Dragonette hyperspectral scene of 23 bands over the Sacramento–San Joaquin Delta, 2025-05-05, about 33 by 66 km. Wyvern is not in the training data. The image shows the bands nearest to red, green and blue. Each embedding covers 21 m. © 2025 Wyvern Inc. All Rights Reserved.

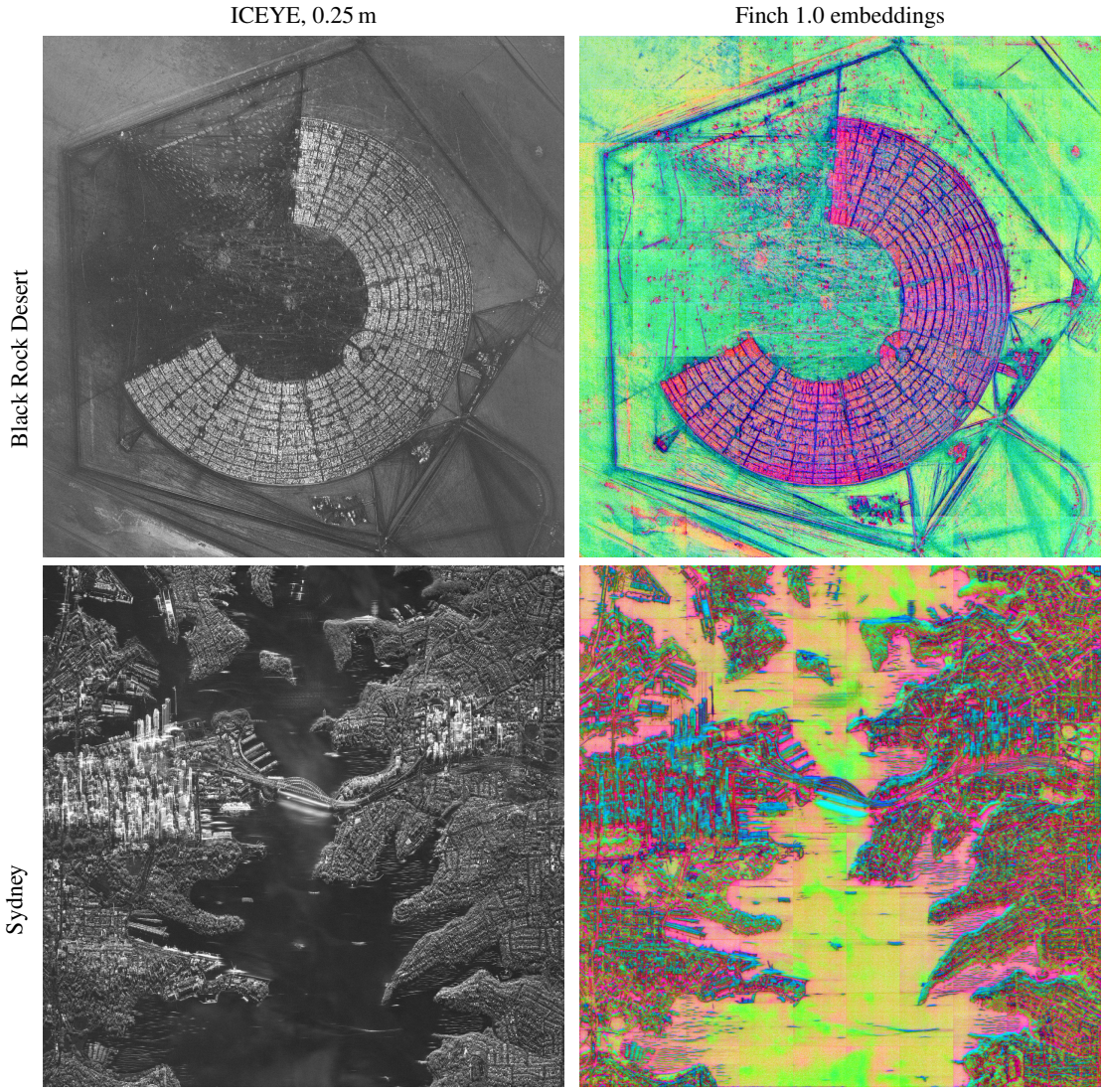
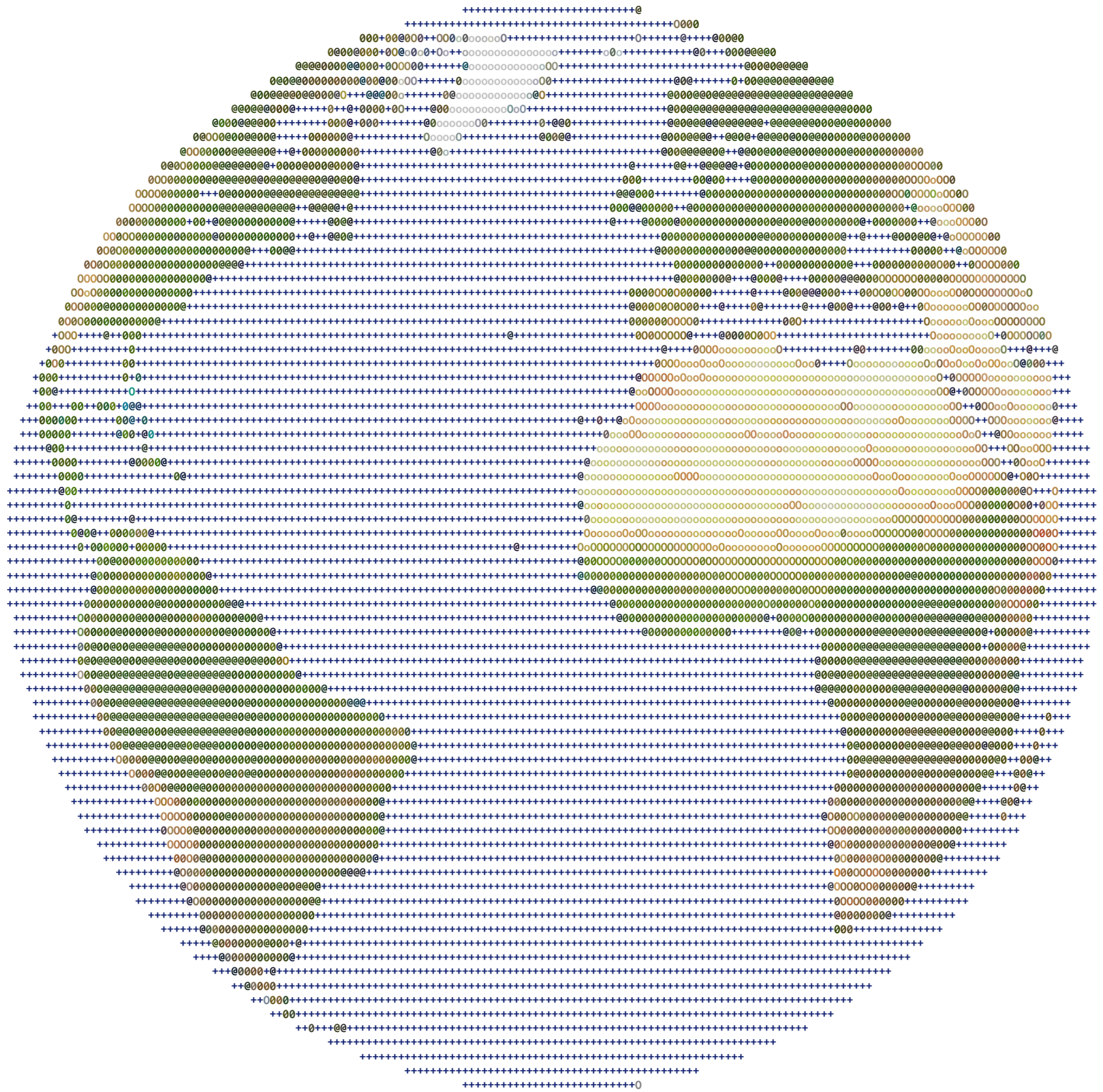


Figure 5: Two full ICEYE X-band radar scenes at 0.25 m, 5 km across: Black Rock City in the Black Rock Desert, Nevada, 2024-08-30 (top), and Sydney Harbour, Australia, 2025-12-23 (bottom). ICEYE is not in the training data. The images show the VV amplitude in decibels. Each embedding covers 4 m. The maps separate the city blocks from the open desert, and the water, the city centre, the Harbour Bridge, the suburbs and the bushland. ICEYE Open Data, © ICEYE, licensed under CC BY 4.0.



Rendered from NASA Blue Marble imagery.