TESSERA v2: SCALING PIXEL-WISE EARTH FOUNDATION MODELS

Zhengpeng Feng[1] **Sadiq Jaffer**[1] **Ira Shokar**[2] **Jovana Knezevic**[1] **Mark Elvers**[1]
Clement Atzberger[3] **Robin Young**[1] **Aneesh Naik**[1] **Niall Robinson**[2] **Andrew Blake**[1]
David Coomes[1] **Anil Madhavapeddy**[1] **Srinivasan Keshav**[1*]
[1]University of Cambridge [2]NVIDIA [3]dClimate Labs

ABSTRACT

Pixel-wise Earth-observation (EO) foundation models are now achieving state-of-the-art performance via generated spatial embeddings. However, how these models scale and how best to spend a pretraining budget remain poorly understood. We present the largest controlled scaling study for EO to date: 395 training runs on 1,024 GH200 superchips within a fixed pixel-wise BARLOW TWINS family, each evaluated on 15 downstream tasks. We find that pretraining loss barely predicts downstream performance ($|\text{Pearson } r| < 0.2$), so selecting models by loss wastes a large share of the compute. We also find that, as the training budget grows, the encoder and the data should grow together while the projector stays fixed, which gives a simple rule for allocating compute. Using this rule, we train a family of pixel-wise models (0.5B and 1B, with a 2B model in training) and distill them into compact students for embeddings-as-data deployment. The 21-million-parameter distilled TESSERA v2-1B-M in aggregate outperforms all open and proprietary models tested, some of which are orders of magnitude larger. These students produce MATRYOSHKA representations that are inexpensive to serve: a 16-dimensional prefix keeps 92% of the full 128-dimensional performance at $1/8$ of the storage. Upon completion of training we plan to release v2 global embeddings covering 2017-2025. Together, these results give a concrete, empirically grounded recipe for scaling pixel-wise EO foundation models: train large encoders, select by downstream performance, and distil into flexible student models. All code will be released at https://github.com/ucam-eo/tessera.

1 INTRODUCTION

The main bottleneck in using Earth observation (EO) for downstream tasks is preparing the data: radiometric calibration, cloud and shadow masking, cross-sensor harmonisation, and expensive computation over raw imagery. Beyond this preparation, the ground-truth labels these tasks need are often scarce and tied to specific regions and seasons (Metcalfe et al., 2025; Hou et al., 2026). The data itself is awkward in ways natural images are not: Sentinel-2 and Sentinel-1 observe a given pixel at different, irregular cadences, cloud removes much of the optical record (Figure 1a), and the cloud-free composites most models train on remove the phenological dynamics that downstream tasks depend on (Zeng et al., 2020; Xiao et al., 2025). Lowering these barriers requires a reusable representation layer for global surface state, delivered as spatially mapped data.

Earth embeddings are vector representations of specific places and times that compress and fuse multi-source observations (Klemmer et al., 2025; Fang et al., 2026). Provided at global scale, in the 'embeddings-as-data' approach, these representations free analysts from the onerous acquisition of heterogeneous raw data, the need for remote-sensing expertise, and from user-side GPU compute. With task-specific heads, they have been shown to match, and often exceed, task-specific models.

An ideal embedding product should satisfy at least the following five desiderata. It should be *analysis-ready*, giving users a geospatial data layer they can work on directly, without processing raw imagery or running GPU inference. It should be *transferrable*, so that the representation carries across tasks

*Corresponding author.

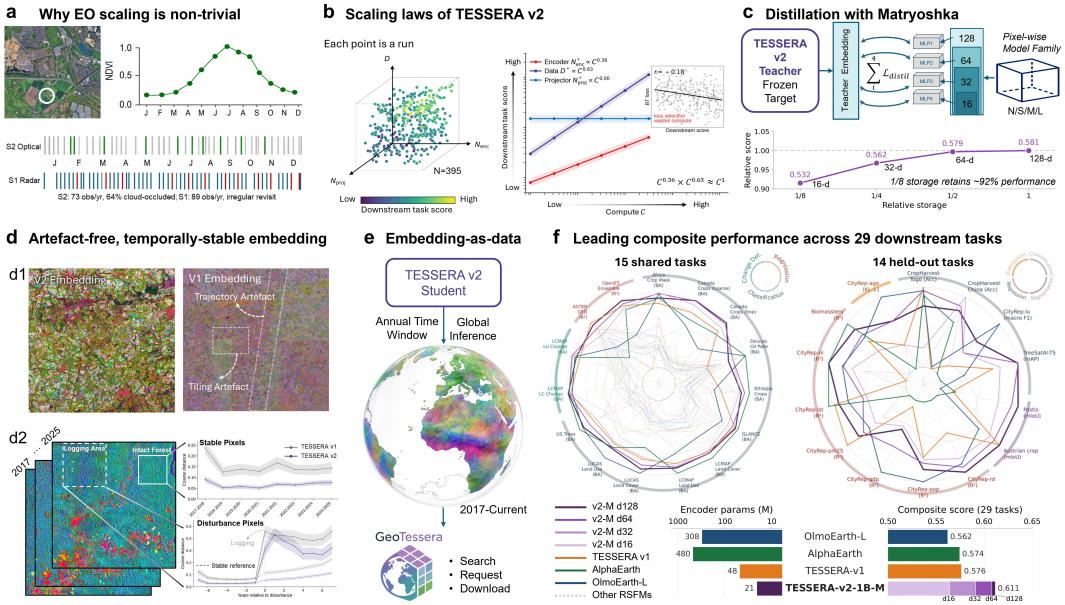

Figure 1: Overview of TESSERA v2. **(a)** Sparse, irregular Sentinel-2/Sentinel-1 sampling at one location. **(b)** 395-run downstream-driven scaling sweep and compute-optimal fits; inset: pretraining loss vs. downstream score. **(c)** MATRYOSHKA distillation of the 1 B teacher into N/S/M/L students, $d \in \{16, 32, 64, 128\}$. **(d)** Artefact removal and inter-annual stability vs. v1. **(e)** Deployment via GEOTESSERA. **(f)** Leading composite performance across the 29-task full suite (15 shared ALPHAEARTH suite tasks + 14 held-out datasets). Radars show per-task scores on the shared (left) and held-out (right) tasks; bars give the 29-task composite (TESSERA v2-1B-M, shades $d \in \{16, 32, 64, 128\}$) and encoder size. TESSERA v2 leads at 0.611, clearly ahead of all three baselines.

and workflows (classification, regression, change detection, high-resolution mapping, etc.) rather than being tuned to only one benchmark or region (Klemmer et al., 2025; Lyu et al., 2026). It should be *reproducible*, letting users inspect, compare, and extend the released artifacts rather than download opaque outputs. It should be *economical* in the resources that limit EO adoption, namely labels, training compute, inference compute, storage, I/O, and engineering time. Finally, it should offer *adaptivity*, an explicit characterisation of the trade-off between accuracy and computational (or economic) cost, so that users are not locked into a single model size or a single embedding dimensionality.

Measured against these criteria, every existing system leaves a gap. Remote-sensing foundation models (Cong et al., 2022; Reed et al., 2023; Fuller et al., 2023; Guo et al., 2024; Tseng et al., 2024; 2025; Astruc et al., 2025) learn strong representations but release models rather than products, so the preprocessing and inference burden stays with the user (Fang et al., 2026). Specialised embedding archives are useful within their scope: ESD (Chen et al., 2026) compresses 25 years of Landsat/MODIS reflectance into quantised 30 m embeddings, though its resolution, sensor scope, and demonstrated evaluation breadth stop short of the specification above. ALPHAEARTH (Brown et al., 2025) is analysis-ready at global scale but only open-output: the embeddings are public while the training procedure and model weights are not (Hou et al., 2026), which limits inspection and extension. TESSERA v1 (Feng et al., 2026) is open, pixel-wise, and label-efficient, yet it ships one fixed 128-dimensional specification, so every user inherits the same storage and I/O budget regardless of their deployment constraints. Each of these meets part of the specification. None is an open, pixel-wise, analysis-ready family in which users pick the model size and embedding dimension that fit their budget.

These product constraints reach back into training. If an embedding field is produced once, served globally, and reused by many downstream users, then pretraining compute becomes part of the product budget rather than an isolated modelling expense. The relevant question is not simply whether a

larger backbone lowers a self-supervised loss, but which allocation of compute produces the most useful representation per unit cost.

In the language and vision domains, that allocation comes from scaling laws fit to the pretraining loss (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022). The few existing EO scaling studies either entangle architecture with capacity or fit power laws to the loss and defer the downstream comparison (Dionelis et al., 2025; Wickrema et al., 2025). In EO the loss is a doubtful target, because cloud and orbital sampling dominate input variance and a redundancy-reduction objective can be minimised through invariances that carry no downstream value.

We therefore ran a downstream-driven scaling study: 395 controlled pretraining runs of pixel-wise BARLOW TWINS (Zbontar et al., 2021) encoders (the backbone family of TESSERA v1),[1] each evaluated on 15 downstream tasks. The result is a different scaling picture from loss-driven practice. Pretraining loss turns out to be a weak proxy for downstream utility, and selecting models by loss wastes roughly 254% of compute relative to downstream-driven selection (**F1**). The downstream-optimal allocation instead puts additional budget into encoder capacity and training data while the projector stays essentially fixed (**F2**). This yields a simple production rule: train a large encoder on matched data, then recover deployment efficiency through distillation. The sweep is expensive, but it is a one-off measurement of the production function of this model family: once the downstream-calibrated allocation is known, larger teachers can be trained by rule rather than by repeated loss-driven search.

TESSERA v2 follows this approach. We train a large pixel-wise Sentinel-1/2 teacher as a representation distribution rather than a deployed artifact, and distil it into a family of compact students (N/S/M/L). The students produce precomputed annual embeddings served through GEOTESSERA, so users train lightweight heads on analysis-ready data instead of running a backbone. MATRYOSHKA prefixes expose 16-, 32-, 64-, and 128-dimensional views of the same embedding, which turns storage and I/O into a user-side knob. Across a 29-task suite (the 15-task ALPHAEARTH suite plus 14 further held-out datasets), the distilled students lead every open and proprietary embedding product we compare, the smallest prefixes keep most of the full-dimensional score, and the embeddings shed the Sentinel acquisition artefacts visible in v1 (Figure 1d). This nested representation must be learned through distillation rather than by adding prefix losses during self-supervised pretraining.

In summary, our contributions are:

1. **Downstream-driven scaling for product-grade EO embeddings.** Across 395 runs evaluated on 15 downstream tasks, we show that the pretraining loss is a poor selection target (**F1**) and fit a compute-allocation rule (**F2**): additional pretraining budget should go to encoder capacity and training data, not projector size.
2. **A deployable pixel-wise embedding product family.** Guided by this rule, we train large teachers and distil them into compact N/S/M/L students whose annual Sentinel-1/2 embeddings are served as analysis-ready data through GEOTESSERA, with the best composite score and mean rank on the 15-task ALPHAEARTH suite at a deployment cost two orders of magnitude below the teacher's.
3. **Storage-adaptive MATRYOSHKA embeddings.** Each student exposes prefixes at $d \in \{16, 32, 64, 128\}$ from one embedding, an accuracy/storage knob that needs no retraining. The d=16 prefix keeps ∼92% of the d=128 score at 1/8 of the storage.

2 RELATED WORK

The present work builds on TESSERA v1 (Feng et al., 2026), which adapts BARLOW TWINS redundancy reduction to cloud-corrupted EO time series following Lisaius et al. (2024). We reuse its d-pixel formulation and pretraining recipe as the fixed model family for the scaling study.

EO foundation models. Remote-sensing foundation models pretrain spatial backbones on single-time, often cloud-filtered patches with contrastive or masked-image objectives (Mañas et al., 2021; Guo et al., 2024; Cong et al., 2022; Reed et al., 2023; Tang et al., 2023; Li et al., 2024; Noman et al., 2024; Wang et al., 2022; Wanyan et al., 2024), and have grown into broader families (Sun et al., 2023;

[1] The sweep ran on 1,024 NVIDIA GH200 superchips McIntosh-Smith et al. (2024), each pairing a Grace CPU with one H100 GPU (96 GB HBM3).

Wang et al., 2023; Bastani et al., 2023; Schmude et al., 2024; Szwarcman et al., 2025; Mendieta et al., 2023; Han et al., 2024; Wu et al., 2025; Luo et al., 2024; Zhu et al., 2025; Zhang et al., 2025). Multi-sensor and multi-resolution models fuse optical and SAR or ingest many sensors at once (Fuller et al., 2023; Yao et al., 2023; Xiong et al., 2024; Astruc et al., 2025; Perron et al., 2026; Tseng et al., 2025). PRESTO (Tseng et al., 2024) processes per-pixel time series, and MOSAIKS (Rolf et al., 2021) explored lightweight universal features. Two assumptions recur across these designs: natural-image scaling intuition is inherited by analogy, and the model is deployed through per-task fine-tuning of the backbone, which is compute- and label-intensive for EO users. We are not aware of a controlled study of where additional compute should go given irregular revisits, cloud occlusion, and label scarcity. Section 3 provides one for the pixel-wise BARLOW TWINS family.

Embedding products. A second line publishes precomputed, analysis-ready embeddings instead of a backbone: TESSERA v1 releases global annual 10 m pixel-wise int8 embeddings with the GEOTESSERA retrieval library (Feng et al., 2026; Madhavapeddy et al., 2026), ALPHAEARTH provides global annual 10 m embedding fields from many instruments (Brown et al., 2025), and ESD (Chen et al., 2026) compresses 25 years of Landsat/MODIS reflectance into quantized 30 m embeddings. The shared limitation is a fixed embedding specification: one dimension, one storage and I/O budget for every user, and no coordinate ordering that would let a user truncate to a smaller dimension without retraining or loss. TESSERA v2 keeps the paradigm and adds two degrees of freedom: students at four sizes, produced by scaling-law-guided distillation, and MATRYOSHKA prefixes $d \in \{16, 32, 64, 128\}$ from a single embedding without retraining. Section 5 explains why the ordering that makes these prefixes usable requires distillation rather than self-supervision alone. Evaluating embeddings across such budgets has begun to attract dedicated benchmarks (Vinge et al., 2025).

Scaling laws. Empirical scaling laws have shaped recent language and vision work (Kaplan et al., 2020; Hoffmann et al., 2022; Zhai et al., 2022), and are typically fit to a self-supervised pretraining loss on the assumption that the loss proxies downstream quality. This assumption can fail: in language modelling the mapping from pretraining loss to downstream performance is sometimes noisy or non-monotone, so a lower loss need not yield a better task model (Hu et al., 2025). The two EO studies we are aware of do not test it directly. Dionelis et al. (2025) sweep architecture, size, and data on PhilEO Bench but entangle architecture with capacity over a coarse grid, and Wickrema et al. (2025) fit peta-pixel power laws to the validation loss in a data-limited regime that, as they note, is confounded by under-trained large models, and defer the loss-vs-downstream comparison to future work. Neither isolates encoder, projector, and data under matched compute. We isolate all three and fit the compute allocation against task performance directly, evaluating every run on 15 downstream tasks.

Distillation and nested embeddings. Knowledge distillation (Hinton et al., 2015) transfers representations from a high-capacity teacher to a compact student. MATRYOSHKA representation learning (Kusupati et al., 2022) produces nested embeddings whose prefixes work at multiple dimensionalities. The two are rarely combined in EO, and rarely with attention to whether the self-supervised objective can support nested coordinates at all. Our analysis shows that naive MATRYOSHKA-BARLOW TWINS fails because redundancy-reduction objectives identify subspaces only up to rotation. Distillation against a fixed teacher supplies the ordering signal that self-supervision lacks.

3 DOWNSTREAM-DRIVEN SCALING LAWS

This section answers the allocation question: within a fixed architectural family, how should pretraining compute be split between encoder size, projector size, and training data? The experiments below describe the family we sweep, pixel-wise Sentinel-1/2 encoders pretrained with BARLOW TWINS (Zbontar et al., 2021). Note that we do **not** advance them as universal EO scaling laws.

3.1 STUDY DESIGN

Architecture choice. Performance differences attributed to size or data are only reasonable when the architecture is fixed. We therefore fix the architecture before scaling, sweeping seven structural

axes (encoder microarchitecture, projector form, Sentinel-1/Sentinel-2 fusion, temporal aggregation, cloud handling, sequence length, and at what point MATRYOSHKA nesting is introduced) one at a time and selecting by aggregated downstream score. The selected configuration is held fixed for the entire scaling sweep.

Controlled sweep. With architecture fixed, we pretrain 395 models on 1,024 H100 GPUs in an iso-FLOP-style grid over encoder size N_{enc} (16 widths, 7–278 M), projector size N_{proj} (four widths), and training data D (0.03–9,984 M d-pixels). At each compute level, the compute-optimal size is the vertex of a quadratic fit in $\log N$. The compute axis is

$$C \;=\; 12\, D\, \big(N_{\text{enc}}\, L_{\text{ref}} + N_{\text{proj}}\big), \qquad L_{\text{ref}} = 240, \tag{1}$$

where the constant $12 = 6 \times 2$ is the textbook $6ND$ factor (two FLOPs per multiply–add, threefold forward-plus-backward) times the two BARLOW TWINS augmentation views per step, and L_{ref} is a nominal annual sequence length common to all runs.

Downstream evaluation. Every pretrained model is evaluated on 15 ALPHAEARTH suite tasks drawn from 10 source datasets (classification, segmentation, change detection, regression), with chance-adjusted metrics: (balanced accuracy $- 1/K)/(1 - 1/K)$ for classification and $\max(0, R^2)$ for regression. The per-task scores average into one composite downstream score, the y-axis of Figure 2. On the same suite, TESSERA v1 scores 0.541 and ALPHAEARTH 0.560. Both are baselines in Figure 2a–c. Both baselines are full-budget production systems, whereas the sweep grid deliberately spans many small, data-limited configurations in order to trace out the compute frontier; most individual runs therefore fall below the baselines, while the upper envelope of the sweep approaches them. For every run we also record the converged BARLOW TWINS loss on held-out d-pixels, normalised so that runs with different projector widths are comparable.

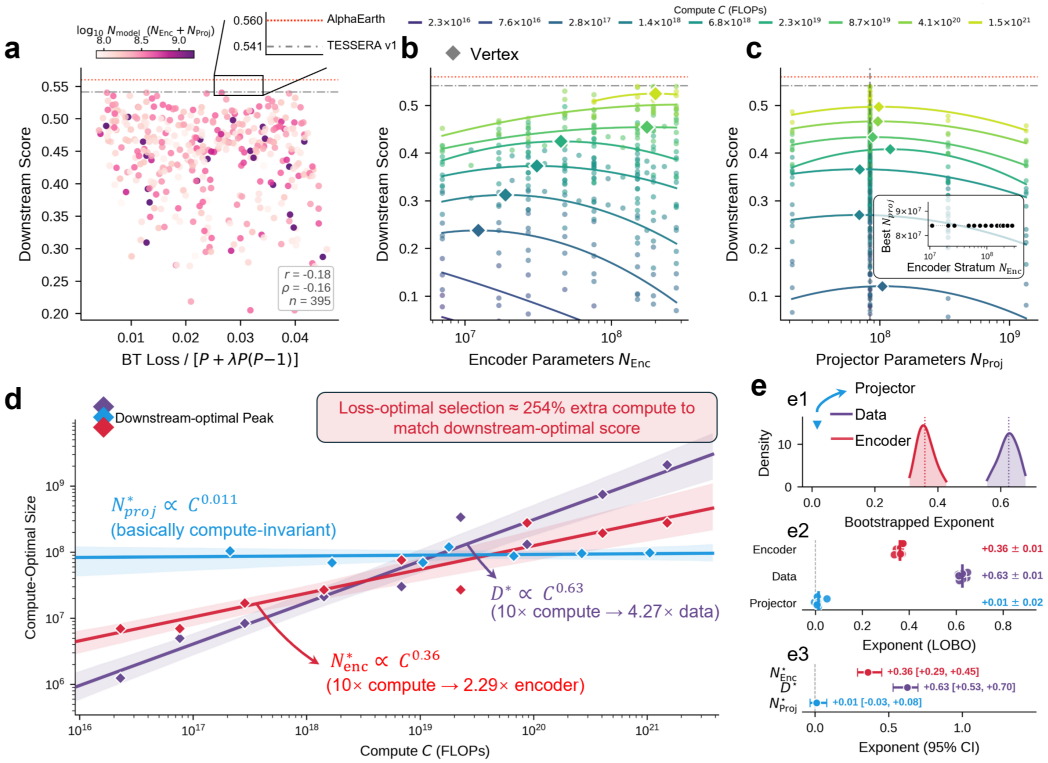

Figure 2: Downstream-driven scaling laws (395 runs; composite score over 15 tasks). **(a)** Pretraining loss vs. downstream score, coloured by model size. **(b, c)** Iso-FLOP parabolas for encoder and projector; vertices mark compute-optimal sizes per compute bucket. **(d)** Power-law fits to the vertices, $N^{\star}_{\text{enc}} \propto C^{0.36}$, $D^{\star} \propto C^{0.63}$, $N^{\star}_{\text{proj}} \propto C^{0.00}$; dashed line is the loss-optimal encoder fit. **(e1–e3)** Bootstrap and leave-one-bucket-out stability of the three exponents.

3.2 FINDING 1: THE PRETRAINING LOSS IS NOT A GOOD PREDICTOR OF DOWNSTREAM TASK PERFORMANCE

Across the 395 runs, the converged BARLOW TWINS loss and the composite downstream score are nearly independent (Figure 2a; Pearson $r = -0.18$, Spearman $\rho = -0.16$). Clouds and orbital sampling dominate the per-sample variance without being downstream-relevant, and a redundancy-reduction objective can be minimised through invariances orthogonal to the physical processes downstream tasks depend on. The consequence is quantitative, not just statistical: fitting separate power laws through loss-selected and score-selected bucket peaks, loss-based selection needs roughly 254% more compute to reach the same downstream score. EO scaling laws must therefore be fit against downstream metrics. **F1** prices the alternative.

3.3 FINDING 2: ENCODER SIZE AND DATA REQUIREMENTS SCALE WITH COMPUTE; THE PROJECTOR DOES NOT

We group runs into nine iso-FLOP buckets and fit, within each, a quadratic in $\log N$ whose vertex is the compute-optimal size at that budget. Encoder vertices shift right as compute grows (Figure 2b), while projector vertices stack along a vertical line (Figure 2c). Power-law fits through the vertices (Figure 2d) give $N^\star_{\text{enc}} \propto C^{0.36}$ (95% CI $[+0.29, +0.45]$), $D^\star \propto C^{0.63}$ ($[+0.53, +0.70]$), and $N^\star_{\text{proj}} \propto C^{0.00}$ ($[-0.03, +0.08]$). Two observations. First, $0.36 + 0.63 \approx 1$ (precisely 0.99), recovering the $C \propto N{\cdot}D$ balance of Hoffmann et al. (2022) without it being assumed by the fitting procedure, an internal consistency check on the sweep. Second, the projector exponent is statistically indistinguishable from zero: the compute-optimal projector size does not grow with compute. Encoder capacity, not projector capacity, is the load-bearing scaling axis.

From findings to a recipe. Since encoder capacity and data are the axes that absorb compute and the projector is compute-invariant, the compute-optimal use of a large budget is one oversized encoder trained on correspondingly more data, with the projector held near its optimum as a disposable training scaffold. Deployment flexibility, in model size and embedding dimension, is then recovered through distillation. Section 4 trains the large teacher at the upper end of the swept regime. Section 5 distils it into a compact student family.

4 A PIXEL-WISE TEMPORAL TEACHER

This section focuses on pretraining using a single large encoder. Figure 3 summarises the design. Relative to TESSERA v1 (Feng et al., 2026), TESSERA v2 adds multi-scale temporal pretraining, adaptive full-observation inference, a unified all-Transformer architecture with cross-modal fusion, and scaling-law-guided distillation into MATRYOSHKA students (Section 5). We discuss these innovations in more detail next.

Inputs. A d-pixel at location (i, j) is the time series of all Sentinel-1 and Sentinel-2 observations at that $10\,\text{m}$ pixel over one year, with a binary mask marking valid timesteps (cloud-free for Sentinel-2, present for Sentinel-1) (Feng et al., 2026). This preserves the full temporal phenology while tolerating cloud occlusion and/or irregular revisits.

Architecture. The teacher is a $1\,B$ dual-branch pixel-wise encoder (Figure 3a). Each modality branch linearly embeds its valid observations, adds a sinusoidal day-of-year positional encoding, runs a four-layer Transformer, and aggregates the variable-length sequence by learned attention pooling. A two-layer fusion Transformer then combines the two modality tokens into one embedding $\boldsymbol{t} \in \mathbb{R}^{d_T}$, $d_T = 768$, followed by a final affine-free LayerNorm (no learnable scale or shift). A batch-normalised projector is also used, but only during pretraining. d_T deliberately exceeds the student's 128: this leaves room for compression, and the prefix heads of Section 5 map student prefixes into the teacher's space. The encoder is larger than any point in the Section 3 sweep (which tops out at 278 M); following the recipe, we extrapolate its $C^{0.36}$ encoder law to the $1\,B$ teacher, whose training compute is measured on the same axis (Equation (1)).

Objective and training. Similar to TESSERA v1, we train using BARLOW TWINS over two temporally subsampled views, with an additional mix-up consistency regulariser (Bandara et al.,

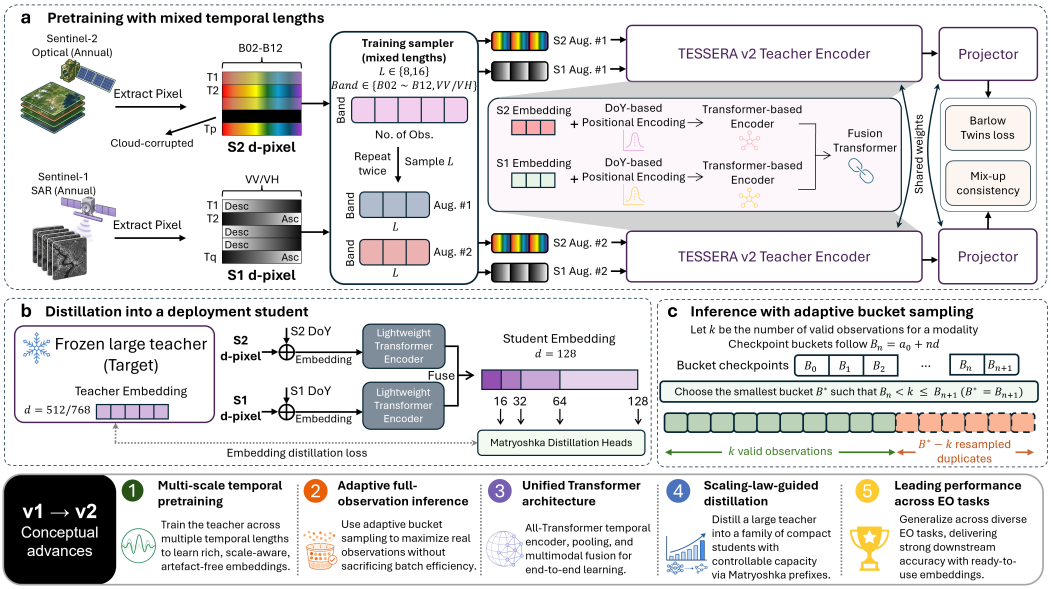

Figure 3: TESSERA v2 architecture. **(a)** Pretraining: two views per d-pixel at random length $L \in \{8, 16\}$, per-modality Transformers with day-of-year encoding, cross-modal fusion, and BARLOW TWINS + mix-up. **(b)** Distillation: MATRYOSHKA prefix heads at $d \in \{16, 32, 64, 128\}$ reconstruct the frozen teacher embedding. **(c)** Inference: each pixel's k valid observations are packed into the smallest bucket $B^\star \geq k$, with residual slots filled by midpoint resampling.

2023) and global shuffling of d-pixels across tiles. Unlike v1's single fixed sample length, at every step the view length is a random $L \sim \mathrm{Uniform}\{8, 16\}$, so each view is a sparse random subsample of the year that forces the encoder to recover annual phenology from few observations. At inference we instead pass *all* valid observations (Section 4): the encoder is given strictly more evidence about the same phenology, each placed at its true day-of-year and aggregated by length-agnostic attention pooling, rather than a longer out-of-distribution input. Whole modalities are also dropped with small probability, which doubles as the training signal for the inference case of a pixel with no valid observations in one modality (e.g. a persistently clouded pixel with zero Sentinel-2).

The teacher is pretrained for a single epoch over $\sim$4.2 billion d-pixels at a global batch size of 131,072 on 512 GPUs under FSDP in `bf16`.

The teacher as a distillation target. Given its large size, we expect most users cannot run a 1 B pixel-wise encoder over global Sentinel-1/2 even once, let alone on a recurring basis. The scaling laws say *where* to spend pretraining compute: on a large encoder (Section 3). What users can afford to *serve* is something much smaller, set by the recurring cost of global inference. The teacher is therefore the right model to train but the wrong one to deploy, so we treat the frozen teacher $\mathcal{T}_\theta : P_{i,j} \mapsto \boldsymbol{t}$ as a fixed *representation distribution* (the distribution of target embeddings it induces over d-pixels) and obtain every deployed artifact by distilling against it (Section 5).

Adaptive bucket sampling at inference. The number of valid observations k varies from a handful in heavily clouded regions to roughly a hundred in clean ones, yet batched inference wants fixed-length inputs. Where TESSERA v1 sampled a fixed $L = 40$ timesteps—discarding observations when $k > L$ and bluntly duplicating when $k < L$—TESSERA v2 packs each pixel into the smallest bucket from the ladder $\{16, 32, 48, \ldots\}$ that fits all k observations, filling the residual slots by midpoint resampling across the year (Figure 3c). No observation is discarded, and a batch still partitions into a few fixed-length groups that run in parallel.

5 DISTILLING A DEPLOYABLE STUDENT FAMILY

We expect users to use our *embedding as data*: consuming precomputed annual pixel embeddings with lightweight task heads, without running a backbone (Feng et al., 2026; Brown et al., 2025). The teacher, however, is too computationally expensive to support this, since running a complex teacher for all land areas on earth would be extremely expensive. Instead, we adopt the well-known idea of distillation to reduce the cost of inference.

Inference cost for one global, annual, 10 m Sentinel-1/2 pass follows a near-linear power law in encoder parameters, $\mathrm{GPUy}(N) \approx 0.041\, N^{1.03}$ H100-years for N in millions. The $1\,B$ teacher costs roughly 50 H100-years per global pass. In contrast, the students cost 0.04 to 2 H100-years, two orders of magnitude lower. Note that, even with lower inference costs, embedding users still need to download and pay for storage in proportion to embedding dimension. We therefore provide MATRYOSHKA prefixes to reduce this cost.

Student family. We distil the teacher into four students of the same architectural form at different capacities: TESSERA v2-1B-L ($44\,M$, for provider-side global inference), M ($21\,M$, a balanced default), S ($7\,M$, low-resource), and N ($1\,M$, edge and on-device). Each emits a 128-dimensional embedding and differs only in backbone width and depth. Distillation uses $\sim$200 million d-pixels on 64 H100 GPUs.

MATRYOSHKA distillation objective. For each prefix length $k \in \mathcal{K} = \{16, 32, 64, 128\}$ we attach a linear head $h_k : \mathbb{R}^k \to \mathbb{R}^{d_T}$, used only during distillation, and train the student against the frozen teacher embedding $\boldsymbol{t}$ with

$$\mathcal{L}_{\mathrm{DIST}}(\boldsymbol{s}, \boldsymbol{t}) \;=\; \sum_{k \in \mathcal{K}} \Big(1 - \cos\big(h_k(\boldsymbol{s}_{1:k}),\; \boldsymbol{t}\big)\Big). \tag{2}$$

Each prefix is supervised to reconstruct the *full* teacher embedding, which makes it a rate–distortion code for $\boldsymbol{t}$ rather than a copy of the teacher's first k coordinates. At inference the user takes any prefix $\boldsymbol{s}_{1:k}$ directly. Because the teacher is frozen, we run it over the distillation corpus once and cache the targets, so distillation compute is spent almost entirely on the student.

Nesting through distillation, not pretraining. We found that the obvious alternative, adding prefix-BARLOW TWINS losses during self-supervised pretraining unexpectedly fails. This is because the redundancy-reduction objective identifies a representation subspace only up to rotation, so prefix losses break the coordinate symmetry through gradient imbalance rather than through a signal about axes' information content. Distillation against a fixed target, however, supplies this important signal.

5.1 BENCHMARK RESULTS

We evaluate in two stages. First, on the 15-task ALPHAEARTH suite (classification, segmentation, regression, and change detection) we compare against the two directly comparable embedding-as-data systems, ALPHAEARTH (Brown et al., 2025) and TESSERA v1 (Feng et al., 2026), plus PRESTO (Tseng et al., 2024), OlmoEarth (Herzog et al., 2026), MOSAIKS (Rolf et al., 2021), and a suite of RSFMs and generic backbones (Fuller et al., 2023; Guo et al., 2024; Reed et al., 2023; Tseng et al., 2025; Astruc et al., 2025; Dosovitskiy et al., 2021). Second, we take the four strongest of these on the ALPHAEARTH suite (TESSERA v2, TESSERA v1, ALPHAEARTH, and OlmoEarth) and run a *held-out* evaluation on 14 further datasets that played no part in development, covering classification, segmentation, regression, and distribution prediction; the two stages together form a 29-task full suite. Every task uses a fixed lightweight head: a two-layer MLP for pixel-wise tasks and a simple CNN ($<$2M parameters) for patch-level tasks.

Headline results. On the full 29-task suite TESSERA v2-1B-M has the best composite score of any system, 0.611, against 0.576 for TESSERA v1, 0.574 for ALPHAEARTH, and 0.562 for OlmoEarth-L, and it is the smallest of the four (Figure 1f). On the 15 ALPHAEARTH suite tasks alone (Figure 4c) TESSERA v2-1B-L leads at 0.584 (M, S, N at 0.581, 0.570, 0.558), above ALPHAEARTH (0.560) and TESSERA v1 (0.541); it also has the best mean rank (2.4 at d=128 vs. 3.8 and 4.9; Figure 4b) and is best or second-best on nearly all tasks. The ordering depends on the task set: ALPHAEARTH

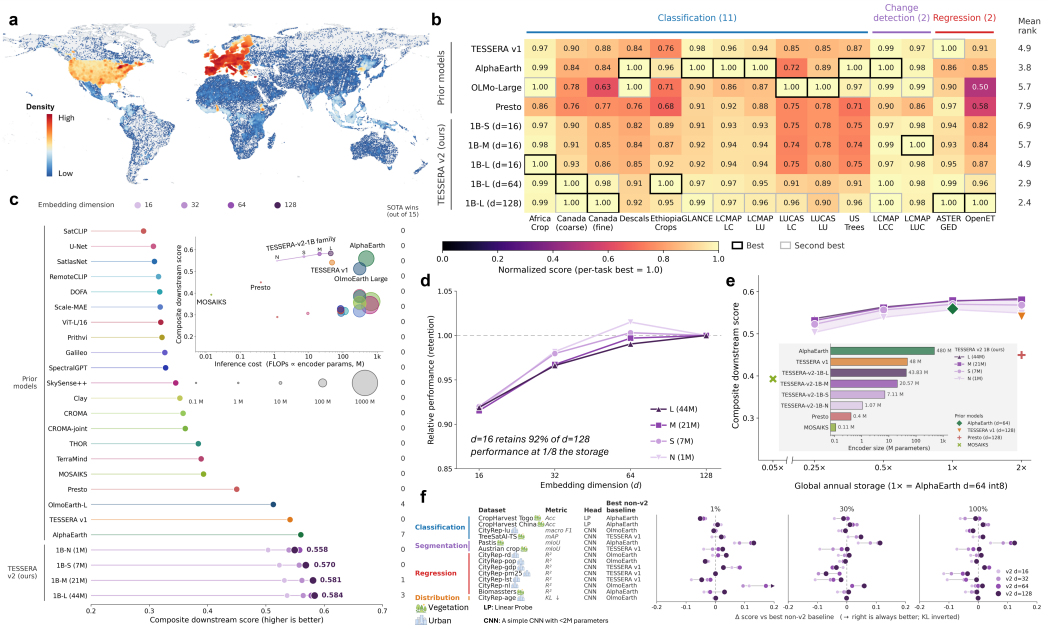

Figure 4: ALPHAEARTH suite results (15 tasks) and held-out generalisation (14 datasets). **(a)** Geographic density of all downstream labels. **(b)** Per-task heatmap with mean rank. **(c)** Composite score per model; the four markers per TESSERA v2 row are $d \in \{16, 32, 64, 128\}$. **(d)** Score vs. prefix dimension: d=16 retains $\sim$92% of the d=128 composite at $1/8$ of the storage. **(e)** Score vs. global annual storage, relative to ALPHAEARTH at d=64 int8. **(f)** Label efficiency on the 14 held-out datasets (vegetation, urban) at $1/30/100\%$ of the labels: Δ score vs. the best non-v2 baseline.

edges TESSERA v1 on its own suite but falls behind it across the full 29 tasks, while TESSERA v2 leads on both.

Graceful degradation across students. Across the four students, d=16 performs at $\sim$92% of the d=128 composite at $1/8$ of the storage, with d=32 and d=64 at $\sim$97% and $\sim$99% (Figure 4d). Performance is best on land-cover and change-detection tasks and worst on fine-class-count and regression tasks, for which d=64 is the default. Regarding storage use, every student is on the Pareto frontier of score versus storage and encoder parameters (Figure 4e): at $0.25\times$ the storage of ALPHAEARTH, TESSERA v2-1B-S already matches it.

Held-out generalisation. The 14 held-out datasets split into a vegetation group (tree species, crop and parcel segmentation, biomass) and an urban group (land use, road density, population, GDP, nighttime lights, pollution, surface temperature, demographics). Across $1/30/100\%$ label budgets TESSERA v2 beats the best non-v2 baseline on most of them, with the widest margins in the low-label regime (Figure 4f).

Embedding perceptual quality and temporal stability. v1 embeddings often show along-track striping and tile-seam discontinuities aligned with Sentinel-2/Sentinel-1 acquisition geometry. In contrast, v2 embeddings have far fewer artefacts while preserving geographic structure. We also note that, as long as land cover is stable, consecutive-year cosine distances for the same pixel are markedly lower for v2 than v1. Moreover, after a temporary disturbance v2 returns to its prior baseline where v1 stays noisy (Figure 1d).

6 DISCUSSION

In language and vision models, the pretraining loss tracks downstream quality closely enough that scaling laws can be fit directly to it. EO breaks this assumption: selecting models by the loss wastes

roughly a factor of three in compute (**F1**). Future EO scaling studies should therefore budget for downstream evaluation rather than trust in the loss.

Once selection is downstream-task driven, **F2** gives the compute budget allocation rule we use, which is to spend the budget on encoder capacity and matched data, holding the projector fixed, and recovering deployable encoders by distillation. Treating the teacher as a frozen representation distribution then lets a single pretraining run amortise across four student sizes and, within each, four embedding dimensions.

Distillation, not pretraining, is what makes the embedding dimension a usable control "knob." Self-supervision fixes only the subspace the embedding spans, leaving its coordinates unordered; distillation against the frozen teacher imposes an order on them using a MATRYOSHKA loss target, so that each prefix is a usable lower-dimensional code.

Limitations. Our work suffers from a few limitations. First, the scaling laws are empirical and apply only to pixel-wise Sentinel-1/2 encoders, one self-supervised objective, and one 15-task evaluation suite. Hence, our analysis of budget allocation is specific to this set of tasks. Second, the use of an expensive teacher model requires substantial computation. Finally, our benchmarks are drawn from well-studied regions, so generalisation to under-represented climates and unseen seasons remains to be evaluated.

7 CONCLUSION

We present the design, implementation, and analysis of the second generation of Tessera foundation model. Our principal research question is how to allocate a compute budget across the encoder and projector. A 395-run downstream-driven scaling study found that pretraining loss is a weak predictor of downstream performance, encoder size and data scale together with compute, and projector size does not. Following this rule, we trained a $1\,B$ pixel-wise teacher and distilled it into embedding-as-data students that lead the 15-task ALPHAEARTH suite; the $d{=}16$ MATRYOSHKA prefix keeps $\sim$92% of the $d{=}128$ score at $1/8$ of the storage. Finally, we found that MATRYOSHKA-style nested embeddings cannot be learned by naive self-supervised prefix losses. Instead, we use distillation to turn prefix learning into supervised ordered compression, a technique we hope carries over to other EO embedding products.

REPRODUCIBILITY STATEMENT

We will release training code, the controlled scaling-study sweeps, distilled pixel-wise student checkpoints, the AlphaEarth-suite evaluation harness, and a frozen GEOTESSERA-style embedding-as-data product. Hyperparameters, dataset splits, and compute estimates are documented alongside the public code release.

ETHICS STATEMENT

TESSERA v2 is a generic representation model for publicly available Sentinel-1/2 imagery. We follow standard responsible-release practices for publicly available foundation models.

ACKNOWLEDGMENT

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. Omnisat: Self-supervised modality fusion for earth observation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities, 2025. URL https://arxiv.org/abs/2412.14123.

Wele Gedara Chaminda Bandara, Celso M. de Melo, and Vishal M. Patel. Guarding Barlow Twins against overfitting with mixed samples. *arXiv preprint arXiv:2312.02151*, 2023.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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. URL https://arxiv.org/abs/2507.22291.

Shuang Chen, Jie Wang, Shuai Yuan, Jiayang Li, Yu Xia, Yuanhong Liao, Junbo Wei, Jincheng Yuan, Xiaoqing Xu, Xiaolin Zhu, Peng Zhu, Hongsheng Zhang, Yuyu Zhou, Haohuan Fu, Huabing Huang, Bin Chen, Fan Dai, and Peng Gong. Democratizing planetary-scale analysis: An ultra-lightweight Earth embedding database for accurate and flexible global land monitoring. *Earth System Science Data Discussions*, 2026. doi: 10.5194/essd-2026-57. URL https://essd.copernicus.org/preprints/essd-2026-57/. Preprint.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Nikolaos Dionelis, Riccardo Musto, Jente Bosmans, Simone Sarti, Giancarlo Paoletti, Peter Naylor, Valerio Marsocci, Sébastien Lefèvre, Bertrand Le Saux, and Nicolas Longépé. Scaling laws for geospatial foundation models: A case study on PhilEO bench. *arXiv preprint arXiv:2506.14765*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Heng Fang, Adam J. Stewart, Isaac Corley, Xiao Xiang Zhu, and Hossein Azizpour. Earth embeddings as products: Taxonomy, ecosystem, and standardized access. *arXiv preprint arXiv:2601.13134*, 2026.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C Lisaius, Markus Immitzer, Toby Jackson, James Ball, et al. Tessera: Temporal embeddings of surface spectra for earth representation and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34818–34831, 2026.

Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:5506–5538, 2023.

Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27662–27673, 2024. doi: 10.1109/CVPR52733.2024.02613.

Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27852–27862, 2024. doi: 10.1109/CVPR52733.2024.02631.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34806–34817, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shuyang Hou, Haoyue Jiao, Ziqi Liu, Lutong Xie, Guanyu Chen, Shaowen Wu, Zhangyan Xu, Zengjie Wang, Shaoqing Tang, Yaxian Qing, Jianyuan Liang, Xuefeng Guan, and Huayi Wu. Alphaearth foundations (aef) in earth observation: A systematic review of applications and practices. *Preprints.org*, 2026. doi: 10.20944/preprints202605.0981.v1.

Michael Hu et al. Scaling laws are unreliable for downstream tasks: A reality check. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2507.00885.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Konstantin Klemmer, Esther Rolf, Marc Russwurm, Gustau Camps-Valls, Mikolaj Czerkawski, Stefano Ermon, Alistair Francis, Nathan Jacobs, Hannah Rae Kerner, Lester Mackey, et al. Earth embeddings: Towards ai-centric representations of our planet. *EarthArXiv preprint*, 2025. doi: 10.31223/X5HX9S.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24088–24097, 2024.

Madeline C. Lisaius, Andrew Blake, Srinivasan Keshav, and Clement Atzberger. Using Barlow Twins to create representations from cloud-corrupted remote sensing time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:13162–13168, 2024.

Junyuan Liu, Xinglei Wang, Zichao Zeng, Jiazhuang Feng, Quan Qin, Ilya Ilyankou, Guangsheng Dong, and Tao Cheng. CITYREP: A unified benchmark for urban representations across cities, tasks, and modalities. *arXiv preprint arXiv:2605.26036*, 2026.

Junwei Luo, Zhen Pang, Yongjun Zhang, Tingzhu Wang, Linlin Wang, Bo Dang, Jiangwei Lao, Jian Wang, Jingdong Chen, Yihua Tan, et al. SkySenseGPT: A fine-grained instruction tuning dataset and model for remote sensing vision-language understanding. *arXiv preprint arXiv:2406.10100*, 2024.

Jienan Lyu, Miao Yang, Jinchen Cai, Yiwen Hu, Guanyi Lu, Junhao Qiu, and Runmin Dong. Structure-semantic decoupled modulation of global geospatial embeddings for high-resolution remote sensing mapping. *arXiv preprint arXiv:2604.19591*, 2026.

Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9394–9403, 2021. doi: 10.1109/ICCV48922.2021.00928.

Anil Madhavapeddy, Frank Feng, Robin Young, Janne Mäyrä, Sadiq Jaffer, Olli Niemitalo, Srinivasan Keshav, and David Allsopp. ucam-eo/geotessera, 2026. URL https://github.com/ucam-eo/geotessera.

Simon McIntosh-Smith, Sadaf Alam, and Christopher Woods. Isambard-ai: a leadership-class supercomputer optimised specifically for artificial intelligence. In *Proceedings of the Cray User Group*, pp. 44–54. 2024.

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16760–16770, 2023. doi: 10.1109/ICCV51070.2023.01541.

Daniel B. Metcalfe, Emily Anders, Hanna Axén, et al. Gaps in tropical science from unrepresentative distribution of sampling and citation across natural terrestrial environments. *Nature Communications*, 2025. doi: 10.1038/s41467-025-67617-4.

Andrea Nascetti, Ritu Yadav, Kirill Brodt, Qixun Qu, Hongwei Fan, Yuri Shendryk, Isha Shah, and Christine Chung. Biomassters: A benchmark dataset for forest biomass estimation using multi-modal satellite time-series. *Advances in Neural Information Processing Systems*, 36:20409–20420, 2023.

Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27811–27819, 2024. doi: 10.1109/CVPR52733.2024.02627.

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.

PyTorch Team. `torch.utils.flop_counter.FlopCounterMode`: per-operator flop accounting in pytorch. https://pytorch.org/docs/stable/torch.utils.flop_counter.html, 2024. Accessed 2026.

Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4065–4076, 2023. doi: 10.1109/ICCV51070.2023.00378.

Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1):4392, 2021.

Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022. doi: https://doi.org/10.1016/j.isprsjprs.2022.03.012.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, et al. Prithvi-WxC: Foundation model for weather and climate, 2024. URL https://arxiv.org/abs/2409.13598.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Eli Gislason, Benedikt Blumenstiel, Rinki Ghosal, et al. Prithvi-EO-2.0: A versatile multi-temporal foundation model for earth observation applications, 2025. URL https://arxiv.org/abs/2412.02732.

Maofeng Tang, Andrei Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale MAE: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:20054–20066, 2023.

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 (Round 2)*, 2021. URL https://openreview.net/forum?id=JtjzUXPEaCu.

Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries, 2024. URL https://arxiv.org/abs/2304.14065.

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, 2025. URL https://arxiv.org/abs/2502.09356.

Rikard Vinge, Isabelle Wittmann, Jannik Schneider, Michael Marszalek, Luis Gilch, Thomas Brunschwiler, and Conrad M. Albrecht. Neuco-bench: A novel benchmark framework for neural embeddings in earth observation, 2025. arXiv:2510.17914.

Yi Wang, Conrad M. Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint SAR-optical representation learning, 2022. URL https://arxiv.org/abs/2204.05381.

Yuelei Wang, Ting Zhang, Liangjin Zhao, Lin Hu, Zhechao Wang, Ziqing Niu, Peirui Cheng, Kaiqiang Chen, Xuan Zeng, Zhirui Wang, Hongqi Wang, and Xian Sun. RingMo-lite: A remote sensing multi-task lightweight network with CNN-transformer hybrid framework, 2023. URL https://arxiv.org/abs/2309.09003.

Xinye Wanyan, Sachith Seneviratne, Shuchang Shen, and Michael Kirley. Extending global-local view alignment for self-supervised learning with remote sensing imagery, 2024. URL https://arxiv.org/abs/2303.06670.

Charith Wickrema, Eliza Mace, Hunter Brown, Heidys Cabrera, Nick Krall, Matthew O'Neill, Shivangi Sarkar, Lowell Weissman, Eric Hughes, and Guido Zarrella. Scaling remote sensing foundation models: Data domain tradeoffs at the peta-scale. In *Applied Imagery Pattern Recognition Workshop*, pp. 132–153. Springer, 2025.

Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, et al. A semantic-enhanced multi-modal remote sensing foundation model for earth observation. *Nature Machine Intelligence*, pp. 1–15, 2025.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2025.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, Changshuo Chen, Jiaqi Yu, Xian Sun, and Kun Fu. RingMo-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023. doi: 10.1109/TGRS.2023.3316166.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Linglin Zeng, Brian D. Wardlow, Daxiang Xiang, Shun Hu, and Deren Li. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sensing of Environment*, 237:111511, 2020.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yingying Zhang, Lixiang Ru, Kang Wu, Lei Yu, Lei Liang, Yansheng Li, and Jingdong Chen. SkySense V2: A unified foundation model for multi-modal remote sensing. *arXiv preprint arXiv:2507.13812*, 2025.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. SkySense-O: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14733–14744, 2025.