LESS IS MORE: SEMANTIC CURATION OF EARTH OBSERVATION DATA FOR SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth-observation (EO) foundation models are pretrained on massive satellite archives that contain substantial spatial redundancy. Yet EO pretraining data are rarely curated semantically, and are instead typically used as collected or sampled using hand-designed geographic and metadata-based rules.

We investigate the impact of data curation on self-supervised EO pretraining and introduce GEOSELECT, a training-free semantic curation pipeline tailored to EO. Starting from embeddings of a frozen EO foundation model, GEOSELECT removes acquisition-induced variability before selecting a fixed-size subset of diverse observations. On the globally distributed Sentinel-2 Major-TOM archive, pretraining on only 10% of the data yields better representations than pretraining on the full archive. Across GEO-Bench, this improves linear-probing performance by $+2.8$ points in classification and $+5.5$ mIoU in segmentation. GEOSELECT also outperforms geographic, metadata-based, and standard semantic curation baselines. These gains are robust across curation budgets, and the approach extends across sensing modalities. Our results show that semantic curation can improve EO representation quality while reducing the amount of pretraining data by an order of magnitude. We will release the curation pipeline and curated subsets.

1 INTRODUCTION

The quality of self-supervised representations depends not only on model scale, compute, and dataset size (Hoffmann et al., 2022), but also on the composition of the pretraining data (Caron et al., 2019). This has motivated large-scale filtering and data selection in language modeling (Brown et al., 2020; Xie et al., 2023), and feature-based deduplication and diversification in vision (Abbas et al., 2023; Oquab et al., 2024; Siméoni et al., 2026). Data composition is particularly important in Earth observation (EO), where spatial autocorrelation, repeated observations, and vast homogeneous regions create substantial redundancy. Yet EO pretraining datasets are still primarily constructed using hand-designed geographic, climatic, or land-cover sampling rules (Manas et al., 2021; Schmitt et al., 2019; Wang et al., 2023; Blumenstiel et al., 2025; Zhu et al., 2026).

We investigate whether feature-based semantic curation can provide a better alternative for EO pretraining. Using the near-contiguous global Sentinel-2 archive Major-TOM, we compare geographic and metadata-based heuristics with recent feature-based curation methods from vision. Each method selects a fixed-size subset of the same archive, on which we pretrain a new SSL model and evaluate its frozen representations through linear probing. This isolates the effect of data selection from that of the pretraining procedure.

Directly applying vision curation methods to EO, however, brings limited benefits. Indeed, EO embeddings encode not only scene content but also acquisition conditions, while observations are spatially structured rather than independent samples. We therefore introduce GEOSELECT, which preprocesses EO embeddings before semantic selection by removing systematic acquisition-related variability and applying spatial regularization. These inexpensive operations are performed once on cached embeddings and prevent apparent diversity caused by acquisition effects from consuming the curation budget.

Our experiments reveal substantial redundancy in large EO archives: pretraining on only 10% of Major-TOM outperforms full-archive pretraining by a large margin across 10 downstream linear-probing tasks. Hand-designed geographic and metadata heuristics already yield strong subsets, but

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

combining semantic curation with our EO-specific preprocessing performs best overall. These gains persist across data budgets and sensing modalities.

Contributions.

1. We systematically compare geographic, metadata-based, and feature-based curation strategies for self-supervised EO pretraining.
2. We introduce GEOSELECT, an EO-specific preprocessing pipeline that accounts for acquisition variability and spatial structure before semantic curation.
3. We show that curated subsets using only 10% of the archive can outperform full-data pretraining, with gains that extend across data budgets and sensing modalities.

2 RELATED WORK

Data curation for self-supervised pretraining. Curating large pretraining corpora can preserve or improve downstream performance while substantially reducing the amount of data used for training (Sorscher et al., 2022; Abbas et al., 2023; Gadre et al., 2023). In language modeling, curation ranges from manually assembled mixtures of high-quality sources (Gao et al., 2021; Laurençon et al., 2022) to automatic quality filtering and deduplication (Brown et al., 2020; Penedo et al., 2023), and distribution-aware data selection (Xie et al., 2023). In vision, SemDeDup removes semantic near-duplicates in embedding space (Abbas et al., 2023), while other approaches cluster pretrained representations to promote diversity (Sorscher et al., 2022; Tirumala et al., 2023; Vo et al., 2024). At web scale, DINOv2 combines embedding-based deduplication with retrieval from manually curated reference datasets (Oquab et al., 2024), whereas Vo et al. (2024) replace these references with hierarchical k-means and balanced sampling.

Data curation for Earth observation. Recent studies further show that pretraining performance depends strongly on which EO samples are retained, not only on geographic coverage (Purohit et al., 2025; Plekhanova et al., 2025; Kaur et al., 2026). EO pretraining datasets are typically curated using hand-designed geographic or metadata-based rules, including sampling around populated areas (Manas et al., 2021; Wang et al., 2023; Stewart et al., 2023; Astruc et al., 2026), geographically stratified sampling (Schmitt et al., 2019), balancing across land-cover classes or eco-regions (Blumenstiel et al., 2025; Jakubik et al., 2025; Szwarcman et al., 2026; Gaydon et al., 2024), and aggregating existing task-specific datasets into heterogeneous collections (Mendieta et al., 2023; Velazquez et al., 2025; Astruc et al., 2025). Major-TOM (Francis & Czerkawski, 2024) instead provides a deliberately uncurated, near-contiguous archive covering most of the Earth's landmasses, together with a random subset. Beyond these construction heuristics, Jiang et al. (2024) explore active-learning-inspired subset selection, while Perera et al. (2026) cluster hand-crafted spectral, texture, and acquisition descriptors.

More recently, feature-based curation has also received increasing attention (Adler et al., 2025; Wei et al., 2025). Hierarchical clustering of learned representations has been explored by Vo et al. (2024) and combined with semantic deduplication by Barseghyan et al. (2025), while Kerdreux et al. (2025) dynamically reclusters and prunes the training archive during SSL pretraining. In the vision-language setting, Muhtar et al. (2025) instead learn a quality model to select remote-sensing image-text pairs. In contrast, GEOSELECT performs a single offline curation pass using frozen EO embeddings, explicitly accounting for acquisition variability and spatial structure before semantic selection.

3 METHOD

The goal of GEOSELECT is to curate a fixed-size training set from a large, unfiltered archive of EO observations. Rather than training on samples drawn directly from the raw archive, we aim to retain observations that provide broad and representative coverage of its semantic content, while avoiding spurious diversity caused by acquisition conditions. The resulting subset is then used for self-supervised pretraining.

Our curation pipeline proceeds in three stages. First, we embed every observation once using a pretrained EO foundation model (Sec. 3.1). Second, we preprocess the resulting representations by correcting acquisition-induced variability (Sec. 3.2) and applying spatial regularization (Sec. 3.3).

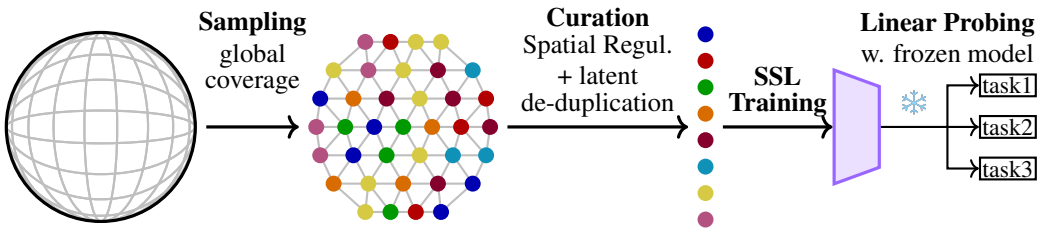


Figure 1: **Evaluation pipeline.** We curate a global EO archive, pretrain an SSL model on the selected subset, and evaluate the frozen representation by linear probing on downstream tasks.

Finally, we select a fixed number of representative observations in the processed embedding space (Sec. 3.4).

3.1 Semantic embeddings of EO samples

Let $\mathcal{X} = \{x_i\}_{i=1}^{N}$ denote the unfiltered archive of EO observations. We represent each observation using a pretrained EO foundation model ϕ,

$$f_i = \phi(x_i). \tag{1}$$

These embeddings provide a compact representation of the visual and semantic content of the archive on which curation can be performed efficiently, without repeatedly processing the original imagery.

When the output of ϕ is not already normalized, we ℓ_2-normalize the embeddings to the unit sphere. In the following, f_i always denotes the resulting normalized embedding.

3.2 Correcting for acquisition conditions

EO imagery varies not only with the content of the observed scene, but also with the conditions under which it was acquired. Atmospheric conditions, illumination, viewing geometry, season, and sensor-specific effects can substantially alter image appearance without reflecting meaningful differences in scene content. These factors are consequently also encoded by ϕ: observations of similar scenes may lie far apart in feature space simply because they were acquired under different conditions. Such variability is useful during pretraining, where the model should be exposed to diverse acquisition conditions and can learn invariance to them. For curation, however, it is undesirable: acquisition effects can make otherwise similar samples appear artificially diverse, causing a selection method to spend part of its budget covering acquisition variability rather than scene diversity.

We address this issue by removing systematic acquisition-dependent offsets from the embedding space before sample selection. The main difficulty is that the acquisition factors driving these offsets—such as local atmospheric conditions or illumination—are not always available for each sample. We therefore exploit the acquisition structure of Sentinel-2 imagery to construct two nested groups that act as proxies for shared acquisition conditions at different spatial scales.

Satellite Pass Grouping. We associate each observation x_i to its satellite pass $\mathrm{p}(i)$. A pass corresponds to a continuous Sentinel-2 acquisition identified by its sensing start time and typically spans thousands of kilometers; in our data, the median pass covers $4{,}330\,\mathrm{km}$ and crosses five distinct Köppen climate classes. Observations from the same pass therefore share several acquisition factors, including sensor calibration, viewing geometry, and processing configuration, while covering geographically distant and semantically unrelated scenes. This makes the pass a natural grouping for estimating acquisition-specific feature shifts with limited dependence on scene content.

Spatial Block Grouping. A single pass, however, is too long to assume spatially constant acquisition conditions. In particular, atmospheric and illumination conditions can vary substantially along its trajectory. Consistently, Fig. 9 shows that acquisition-induced feature shifts remain approximately stable only over distances substantially shorter than a full pass. We therefore introduce a second, finer grouping. Each pass is partitioned spatially into $120\,\mathrm{km}$ blocks, and we denote by $\mathrm{b}(i)$ the block containing x_i. This hierarchical construction separates acquisition effects into a pass-level

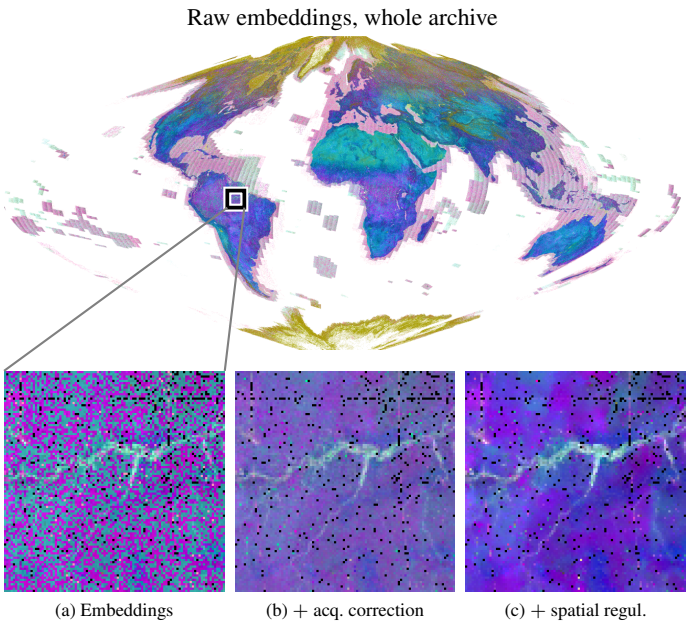


Figure 2: **Effect of the two preprocessing stages on EO embeddings.** RGB visualization of the first three principal components of the UniverSat embeddings over the full Major-TOM archive (top), with a 1000×1000 km window in the Amazon highlighted and shown below at three stages: (a) raw embeddings, (b) after acquisition correction, and (c) after spatial regularization. Black pixels denote missing acquisitions. A common PCA basis and color scale are used throughout. Acquisition correction removes large structured artifacts and reveals coherent geographic patterns, while spatial regularization further smooths residual local noise. Additional examples are shown in the Appendix.

component, capturing factors shared across the full acquisition, and a local block-level component, capturing conditions that vary along the pass. Each embedding is consequently corrected using both a pass-wide and a local acquisition offset, as described below.

Acquisition-effect correction. We model each embedding as the sum of a spatially smooth component and acquisition-dependent nuisance effects,

$$f_i = s_i + \delta^{\text{pass}}[\text{p}(i)] + \delta^{\text{block}}[\text{b}(i)] + \varepsilon_i, \tag{2}$$

where s_i captures feature variations that evolve smoothly in geographic space, δ^{pass} and δ^{block} capture offsets shared respectively within a satellite pass and within a local block of that pass, and ε_i accounts for the remaining sample-specific variability. This additive decomposition is deliberately simple: acquisition conditions are neither strictly additive nor constant within a pass or block. We use group-wise offsets as a tractable approximation to their dominant shared component, leaving finer and nonlinear effects in the residual.

We estimate these components using four iterations of Gauss-Seidel backfitting. At each iteration, s_i is estimated by Gaussian smoothing of the embeddings after removing the current acquisition effects ($\sigma_s = 6$ km), after which the pass- and block-level offsets are updated sequentially using shrinkage estimates. The latter follow from zero-centered random-effect priors, preventing poorly sampled acquisition groups from receiving large corrections. More details are provided in the Appendix.

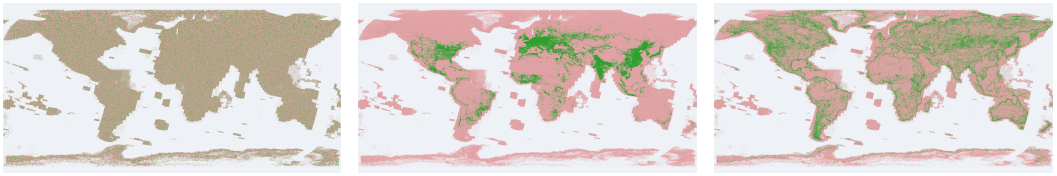
After backfitting, we remove only the estimated acquisition effects,

$$f_i^{\text{corr}} = f_i - \delta^{\text{pass}}[\text{p}(i)] - \delta^{\text{block}}[\text{b}(i)]. \tag{3}$$

The spatial smoothing is therefore used only to estimate the slowly varying geographic component; f_i^{corr} itself is not spatially averaged.

3.3 Spatial Regularization

The acquisition correction of Sec. 3.2 removes nuisance effects associated with known acquisition groups. We complement it with a weaker, metadata-free spatial prior. Nearby locations typically have correlated semantic content, while Major-TOM selects acquisitions independently for each spatial cell, such that neighboring samples can originate from different dates and acquisition conditions. Spatial smoothing can therefore reduce residual nuisance variability, but must remain sufficiently local and edge-aware to preserve genuine geographic diversity.



(a) Uniform sampling (b) Urban, top-% built-up (c) GEOSELECT

Figure 3: **Spatial distribution of selected samples.** Major-TOM samples retained at a 10% budget by (a) uniform sampling, (b) prioritization by built-up fraction, and (c) GEOSELECT. Green samples are retained and red samples are discarded.

We construct a neighborhood graph $\mathcal{E}$ using a k-d tree over 3-D Earth-centered coordinates, connecting each sample to a fixed number of neighbors within a maximum geographic radius (Tab. 5). Edges combine geographic proximity and embedding similarity,

$$w_{ij} = \exp\left(-\frac{d_{\text{hav}}(i,j)^2}{2\sigma_{\text{geo}}^2}\right) \exp\left(-\frac{1-\cos\left(f_i^{\text{corr}}, f_j^{\text{corr}}\right)}{\sigma_r}\right), \tag{4}$$

where d_{hav} is the Haversine distance, $\sigma_{\text{geo}} = 15\,\text{km}$, and σ_r controls sensitivity to representation similarity. The latter term reduces coupling across semantic boundaries, yielding a bilateral-type, edge-aware affinity.

Starting from the acquisition-corrected embeddings F^{corr}, we estimate a spatially smooth field $H = \{h_i\}_{i=1}^N$ through graph Laplacian regularization,

$$H = \arg\min_Z \quad \|Z - F^{\text{corr}}\|_F^2 + \mu \sum_{(i,j)\in\mathcal{E}} w_{ij} \|z_i - z_j\|_2^2 . \tag{5}$$

This strictly convex objective has a unique solution, which we approximate with a fixed number of Jacobi iterations. We normalize μ by the mean graph degree to reduce sensitivity to neighborhood density. Finally, rather than replacing each embedding by its spatial estimate, we conservatively shrink it toward h_i,

$$f_i^{\text{smooth}} = \text{normalize}\left((1-\lambda_i)\, f_i^{\text{corr}} + \lambda_i h_i\right), \tag{6}$$

where λ_i decreases for samples with weak spatial support, such as near coasts, islands, or gaps in coverage. Thus, unlike the acquisition correction, this stage can modify local feature differences, but the edge-aware affinity and partial shrinkage limit the loss of genuine semantic detail. Figure 2 illustrates the variations removed by the two stages.

3.4 DIVERSITY-AWARE SUBSET SELECTION

Our goal is to select a fixed-size subset of samples that preserves the semantic diversity of the full archive $\mathcal{X}$ while removing redundant observations. Exhaustively comparing all pairs is prohibitively expensive at archive scale. Following the idea of SemDeDup (Abbas et al., 2023), we therefore first partition the archive into K semantic clusters using the spatially regularized embeddings f_i^{smooth} (Tab. 5), and perform exact pairwise comparisons only within each cluster.

Within each cluster, samples are ordered by decreasing distance to the cluster centroid, such that the most atypical observations are considered first. The centroid is used only to define this ordering and is not itself a candidate sample. For each sample i, we measure its redundancy with respect to observations in the same cluster that are farther from the centroid:

$$r_i = \max_{j:\, c(j)=c(i)\,,\, d_j > d_i} \cos\left(f_i^{\text{smooth}}, f_j^{\text{smooth}}\right), \tag{7}$$

where $c(i)$ denotes the cluster assignment of sample i and d_i its distance to the corresponding centroid. Thus, a large r_i indicates that i is closely matched by a more atypical observation and contributes little additional diversity.

Finally, we rank all observations globally by increasing redundancy score r_i and retain the first B samples, where B is the desired curation budget. This converts the local comparisons of SemDeDup into a global, budget-controlled selection rule: rather than specifying an embedding-dependent similarity threshold, we directly select the desired number of least redundant observations.

Table 1: **Downstream performance on GEO-Bench.** The first line reports the performance when pretraining on the full archive; all other entries report the difference relative to this reference. Classification uses accuracy except for BigEarthNet (μF1), while segmentation uses mIoU. GEOSELECT combines SemDeDup (SDD) on UniverSat features with spatial regularization (spatial regul.) and acquisition correction (acq. correc.), with ocean samples optionally excluded from the candidate pool. Our proposed variants are shown in bold.

Curation			Classification						Segmentation					
Criterion	Sampling	Kept	So2 Sat	BEN	Brick	Forest Net	PV4 GER	**Mean**	Cash ew	Chesa peake	SA Crop	Neon Tree	PV4 GER	**Mean**
Full dataset		100%	41.9	45.0	93.1	42.0	91.2	65.6	27.3	53.1	22.3	52.5	69.6	44.9
Official subsplit		11%	-0.2	+2.8	+0.5	+0.6	+1.4	+0.7	-1.3	+0.8	+1.3	+0.0	+0.7	+0.3
Uniform	random	10%	-1.3	+0.5	+0.5	-1.0	+1.1	-0.2	-0.4	-0.9	-0.2	-0.4	-0.9	-0.5
Europe	subset	7.2%	+1.4	+2.8	+0.0	+0.5	+0.9	+0.7	+2.2	+0.3	+1.3	-1.0	+1.1	+0.8
Urban	top-%	10%	-0.3	+4.3	-0.6	+2.0	+1.4	+0.8	+6.1	+2.5	+2.9	+0.2	+3.3	+3.0
Land cover	stratified	10%	+0.8	+1.4	-0.4	+0.6	+0.4	+0.1	+0.8	+0.0	+0.9	+0.1	+1.0	+0.6
Climate	stratified	10%	+1.0	+3.0	-0.1	+0.1	+1.4	+0.7	+1.2	+1.3	+1.3	-0.1	+1.7	+1.1
Soil	stratified	10%	+1.1	+2.7	+0.6	+2.2	+1.5	+1.4	+1.4	+1.1	+1.6	+0.2	+0.8	+1.0
UniverSat	Hier. k-means	10%	+1.1	+3.2	-0.3	+1.4	+1.1	+1.4	-0.5	+0.0	+1.5	-1.9	+0.7	+0.0
UniverSat	pHash + Hier. k-m.	10%	+0.7	+4.5	+0.6	+2.2	**+1.8**	+1.5	+1.2	+1.2	+1.3	-0.7	+1.9	+1.0
DINOv2	SemDeDup (SDD)	10%	+1.0	-0.1	+0.1	-1.3	+1.0	-0.2	-0.8	+0.2	+0.5	-0.3	+0.3	+0.0
UniverSat	SDD	10%	-1.7	+4.2	**+0.9**	+1.0	+0.1	+0.4	+4.3	+1.8	+2.3	+0.2	+2.9	+2.3
UniverSat	**Spatial Regul. + SDD**	10%	**+1.5**	**+8.1**	-0.6	+5.9	+1.0	+2.6	+8.1	+4.1	+4.6	+0.6	+6.8	+4.9
UniverSat	**Acq. correc. + SDD**	10%	+0.5	+6.9	+0.1	+3.5	+1.4	+2.0	+8.0	+3.8	+4.7	+0.4	+5.3	+4.4
UniverSat	**GEOSELECT**	10%	+0.5	**+8.1**	-0.2	+4.1	+0.8	+2.1	+8.2	+4.1	**+5.0**	**+1.0**	+7.2	+5.2
UniverSat	**GEOSELECT (no ocean)**	10%	+0.3	+7.6	+0.4	**+7.4**	+1.6	**+2.8**	**+10.3**	**+4.3**	+4.5	+0.8	**+7.7**	**+5.5**

4 EXPERIMENTS

4.1 EVALUATION PROTOCOL

We evaluate curation methods by measuring the quality of representations learned from the subsets they select. We use Major-TOM Core-S2L2A, a near-contiguous and minimally curated global archive of 2,238,687 single-date Sentinel-2 10×10 km samples covering the Earth's landmasses and coastlines. Its scale and limited prior filtering make it particularly well suited for studying data curation. Starting from this archive, we (i) select a fixed-size subset, (ii) pretrain a new model on the selected samples using a self-supervised objective, and (iii) evaluate the resulting representations by linear probing on downstream tasks.

Curation parameters. Unless stated otherwise, each method retains 10% of the archive. For curation, we consider the two pretrained representations available for the full Major-TOM archive at the time of submission: the EO-specific UniverSat (Perron et al., 2026), and vision-based DINOv2, whose features are computed from the RGB bands of Sentinel-2. These models are used only for curation and are distinct from the model subsequently pretrained on the selected samples. Only the identifiers of the selected samples are retained; no curation embeddings or processed representations are passed to pretraining.

Self-supervised pretraining. For each curated subset, we pretrain a ViT-B from scratch with MAE, using random 256×256 pixel crops drawn from the selected samples. We use AdamW with a plateau learning-rate schedule and early stopping, with identical optimization hyperparameters and number of optimizer steps across curation methods.

Downstream evaluation. We evaluate the frozen representations on 10 classification and semantic segmentation tasks from GEO-Bench (Lacoste et al., 2023). For classification, we train a linear classifier on the frozen `[CLS]` representation to predict the K classes. For segmentation, we apply a linear classifier to each frozen patch representation, predicting $P \times P \times K$ logits corresponding to K class scores for each pixel in the patch. For each task, we sweep the probe learning rate and weight decay, select the best probe on the validation set, and report test performance. Training and evaluation details are provided in the Appendix.

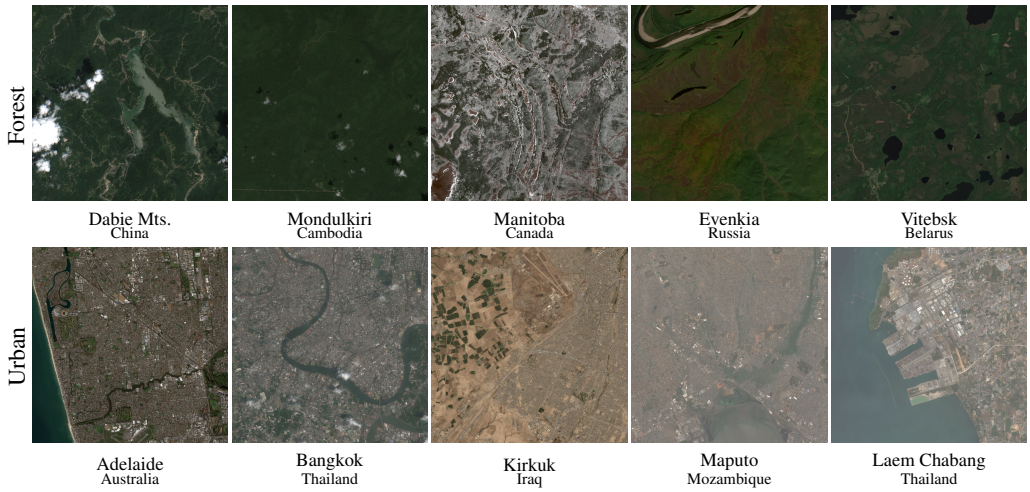


Figure 4: **Selected samples.** We consider samples retained by GEOSELECT at a 10% budget and restrict them to two land-cover groups: $> 80\%$ forest and $> 50\%$ urban cover. Within each group, we cluster the corresponding UniverSat embeddings into $k = 5$ groups and show the sample closest to each centroid. The selected representatives span different continents, biomes, and urban morphologies.

4.2 CURATION STRATEGIES

We compare GEOSELECT against alternative strategies for selecting a budget of B observations from the archive $\mathcal{X}$. Unless stated otherwise, all methods are evaluated under the same pretraining and probing protocol of Sec. 4.1.

Reference splits. We report two configurations that do not match the target budget and therefore serve as reference points rather than direct competitors. *Full dataset* uses the complete Major-TOM archive. We also experimented with proportionally longer training to account for its larger size, but validation performance consistently peaked within the optimization budget used for the curated subsets. *Official subset* uses Core-S2L2A-249k, the 249k-tile subset released with Major-TOM (Francis & Czerkawski, 2024), corresponding to approximately 11% of the full archive. This subset is constructed through systematic geographic subsampling of the Major-TOM grid and thus provides a natural reference for geographically curated pretraining data.

Geographic priors. We consider three hand-designed geographic strategies. *Uniform* partitions the globe into latitude-longitude bins and allocates the budget equally across occupied bins. *Europe*, motivated by the strong geographic transfer reported by Kaur et al. (2026), samples uniformly within 25°W–45°E and 34°–72°N. Finally, motivated by the effectiveness of urban imagery for SSL pretraining (Manas et al., 2021; Wang et al., 2023; Stewart et al., 2023), *Urban* retains the 10% samples with the largest built-up fraction according to ESA WorldCover (Zanaga et al., 2022).

Metadata stratification. We also consider curation based on geographic metadata, following the widespread use of such attributes to construct diverse EO datasets (Schmitt et al., 2019; Blumenstiel et al., 2025; Jakubik et al., 2025). We use three categorical attributes: *Climate*, defined by the Köppen–Geiger climate zones at 1 km resolution (Beck et al., 2023); *Land cover*, from ESA WorldCover (Zanaga et al., 2022); and *Soil*, given by the most probable WRB soil class from SoilGrids (Poggio et al., 2021). Each sample is assigned the majority class over its footprint.

For each attribute, we stratify the target budget across classes : classes receive equal shares whenever possible, with the allocation capped by the number of available observations and any remaining budget redistributed among the other classes. This deliberately flattens the metadata distribution, similarly in spirit to balanced-cluster curation (Vo et al., 2024), and tests whether explicitly balancing known environmental factors is sufficient to obtain a diverse and effective pretraining subset.

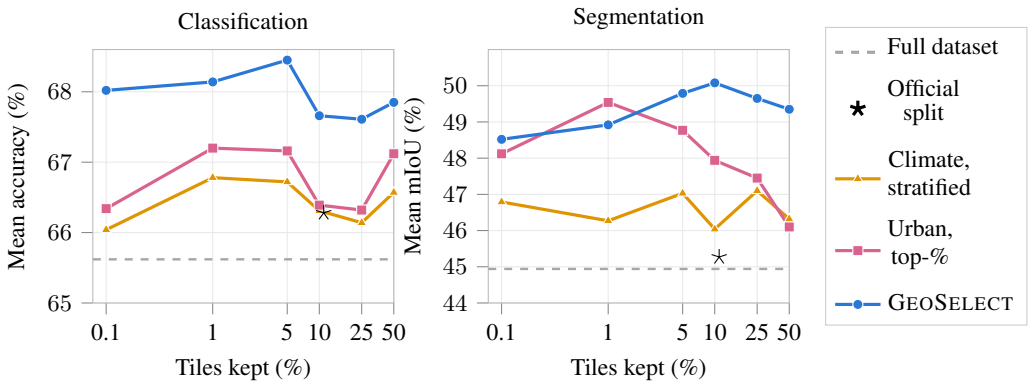


Figure 5: **Downstream performance across curation budgets.** Mean GEO-Bench classification accuracy (left) and segmentation mIoU (right) as a function of the fraction of Major-TOM samples retained for self-supervised pretraining. The dashed line denotes pretraining on the full archive and the star the official Core-S2L2A-249k subset. We compare two hand-designed curation strategies (climate stratification and urban prioritization) with GeoSelect applied to the whole archive, *i.e.*, without excluding ocean samples.

Semantic curation. We also compare against automatic curation methods operating directly on image or embedding similarity obtained with UniverSat or DINOv2. *SemDeDup* (Abbas et al., 2023) denotes our budget-controlled adaptation of the original deduplication procedure to subset selection. We additionally consider *Hierarchical k-means* (Vo et al., 2024), which recursively partitions the embedding space and samples across the resulting clusters to promote semantic coverage. Finally, where indicated, we prepend a perceptual-hash (*pHash*) deduplication step (Zauner, 2010) to remove near-identical images before applying the curation strategy.

4.3 Results and Analysis

We report downstream classification and segmentation performance in Tab. 1. Despite retaining only 10% of the original archive, most curation strategies match or outperform pretraining on the full dataset, showing that the composition of the pretraining set matters at least as much as its size. GeoSelect performs best on both evaluation tracks, improving over the full archive by $+2.8$ points on classification and $+5.5$ mIoU on segmentation.

Curation offsets a $10\times$ reduction in data. Even simple subsets preserve the performance of pretraining on the full archive: uniform sampling remains within 0.2 points on classification and 0.5 mIoU on segmentation using only 10% of the data, while most targeted strategies perform better. Hand-designed EO priors are already strong baselines: urban sampling reaches $+3.0$ mIoU on segmentation, while soil stratification gives the best metadata-based classification result at $+1.4$ points. This indicates substantial redundancy in the raw archive, while also showing that simple geographic or metadata priors can recover much of the useful diversity. These strategies retain very different parts of the archive (Fig. 3): uniform sampling thins every region evenly and urban prioritization concentrates the budget on a few densely built regions, whereas GeoSelect spreads it over all landmasses, favoring heterogeneous areas such as coastlines and river networks while strongly downsampling open water and deserts.

Embedding-based curation. We evaluate curation directly in pretrained feature spaces, without explicit geographic or metadata criteria. Hierarchical k-means with UniverSat features improves classification by $+1.4$ points but leaves segmentation unchanged; adding pHash raises the gains to $+1.5$ and $+1.0$, respectively. SemDeDup performs better overall, but depends strongly on the representation used for curation: DINOv2 features provide essentially no gain over the full archive, whereas UniverSat yields $+0.4$ points in classification and $+2.3$ mIoU in segmentation. With acquisition correction and spatial regularization, GeoSelect further improves to $+2.8$ points and $+5.5$ mIoU, exceeding the strongest hand-designed baselines by $+1.4$ and $+2.5$, respectively, and performing best on both evaluation tracks. Even within a single land-cover type, the retained samples span different continents, biomes, and urban morphologies (Fig. 4).

Table 2: **Cross-modality curation from optical to radar.** Sentinel-1 pretraining on the full archive or subsets curated from Sentinel-1 or co-registered Sentinel-2 embeddings. Evaluation uses the S1-native reBEN-S1 and Sen1Floods11 benchmarks.

Pretraining			Classification	Segmentation
Archive	Curated from	Kept	reBEN S1 μF1	Sen1 Floods11 mIoU
Sentinel-1	*(full archive)*	100%	59.6	79.0
Sentinel-1	Sentinel-1	10%	59.7	79.0
Sentinel-1	Sentinel-2	10%	60.4	79.3

Performance across curation budgets. Figure 5 shows that curation remains effective across a wide range of budgets. Remarkably, even at 0.1% of the archive, all tested strategies outperform full-archive pretraining. GEOSELECT performs best for classification across all budgets and for segmentation from 5% onward, while the Urban heuristic is particularly competitive at very small segmentation budgets. Climate stratification is comparatively insensitive to the keep ratio, with performance varying little across budgets. In contrast, the other strategies peak at intermediate budgets and degrade as more samples are retained, consistent with curation becoming less selective and progressively reintroducing redundant or less informative samples.

Cross-modality curation. We test whether GEOSELECT extends to another modality, and whether a subset curated from one modality can benefit SSL pretraining on another. The latter would indicate that curation captures properties of the underlying scene rather than modality-specific appearance.

We pretrain identical MAE models on Sentinel-1 using either the full archive, a subset curated from Sentinel-1 embeddings obtained by applying UniverSat to Major-TOM Core-S1RTC, or a subset curated from co-registered Sentinel-2 embeddings, and evaluate them on the S1-native reBEN-S1 (Clasen et al., 2025) and Sen1Floods11 benchmarks (Tab. 2). Sentinel-1-based curation matches the full archive using only 10% of the data, while Sentinel-2-based curation slightly improves both tasks, reaching 60.4 micro-F1 and 79.3 mIoU. This suggests that optical embeddings can identify informative locations for radar pretraining, supporting a modality-independent semantic component in the curation criterion.

Efficiency. Curation adds little overhead compared with SSL pretraining. On the full 2.24M-sample archive, acquisition correction takes under 20 min on a 24-core 2.4 GHz CPU, while spatial regularization, clustering, and pruning take under 2 min.

All experiments use matched numbers of optimizer steps across subsets and the full archive; in our setting, curation therefore improves representation quality at matched pretraining compute rather than reducing compute. It does, however, reduce the amount of data that must be stored, transferred, and repeatedly accessed during pretraining by up to an order of magnitude. The resulting keep-lists can also be reused across training runs, amortizing the one-time curation cost.

5 CONCLUSION

We systematically investigated how large EO archives should be curated for self-supervised pretraining, comparing geographic, metadata-based, and semantic selection strategies under a common evaluation protocol. We further introduced GEOSELECT, an EO-specific semantic curation method that corrects acquisition-induced variability before removing redundant observations in embedding space. On Major-TOM, models pretrained on only 10% of the archive match or outperform models trained on all available data, while consistently improving over alternative curation strategies. Beyond the performance gains, reducing the effective pretraining corpus by an order of magnitude substantially lowers storage and data-transfer requirements, making large-scale EO pretraining easier to distribute and reuse. Given the rapidly growing number and scale of EO foundation models, our results suggest that principled data curation is a promising and largely underexplored direction for improving both the efficiency and quality of future pretraining datasets.

Ethics Statement

This work uses only publicly available Earth observation archives (Major-TOM, GEO-Bench) and does not involve human subjects, personal data, or crowd-sourced annotation. Sentinel-1/-2 imagery is acquired at a spatial resolution (10 m) that does not enable identification of individuals. As with other Earth observation foundation models, the learned representations could in principle be applied to dual-use tasks such as surveillance or military reconnaissance; our curation method does not enable or prevent any specific applications.

Reproducibility Statement

We describe the curation pipeline in detail in Sec. 3, with the acquisition-correction and spatial-regularization derivations expanded in Secs. C and C.1. Full pretraining and curation hyperparameters are reported in Tabs. 4 and 5 (Sec. D), covering the MAE recipe and every stage of the curation pipeline, so that a reader can reproduce the reported keep-lists and downstream results without consulting the code. All datasets used (Major-TOM Core-S2L2A/Core-S1RTC for pretraining and the GEO-Bench suite for evaluation) are publicly available, as are the precomputed UniverSat embeddings used for curation, released alongside Major-TOM. We will release the curation pipeline and curated subsets upon publication.

LLM Usage Disclosure

Generative AI disclosure. We used generative AI tools for: providing feedback on research methodology and experimental design; implementing methods; and interpreting results.

We did not use generative AI tools for: generating synthetic datasets; developing theoretical models or conceptual frameworks; proposing or refining hypotheses; assisting with translation; cleaning or reformatting datasets; supporting qualitative or thematic data analysis; formulating mathematical claims; providing critical ingredients for proving mathematical claims; or assisting in the writing of proofs.

Additionally, we used generative AI tools for: creating or modifying scientific figures; suggesting experimental parameters; creating or editing software code; identifying relevant literature; editing the manuscript to improve readability; and proposing titles or keywords.

We reviewed all AI-assisted work. References were checked manually, the manuscript was extensively rewritten and edited by the authors, all claims and analyses were verified, and AI-assisted code and figures were manually inspected and corrected where necessary. We take full responsibility for the final content of this work, including text, claims, code, figures, and other artifacts produced with the aid of generative AI.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR Workshop MRL*, 2023. 1, 2, 5, 8

Mathis Jürgen Adler, Leonard W. Hackel, Gencer Sumbul, and Begüm Demir. Redundancy-aware pretraining of vision-language foundation models in remote sensing. In *IGARSS*, 2025. 2

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: One Earth observation model for many resolutions, scales, and modalities. In *CVPR*, 2025. 2

Guillaume Astruc, Eduard Trulls, Jan Hosang, Loic Landrieu, and Paul-Edouard Sarlin. UniGeoCLIP: Unified geospatial contrastive learning. In *CVPR Workshop EarthVision*, 2026. 2

Alvard Barseghyan, Ani Vanyan, Hakob Tamazyan, Evan Shelhamer, and Hrant Khachatrian. Less is more? Data specialization for self-supervised remote sensing models. In *ICML workshop TerraBytes*, 2025. 2

Hylke E. Beck, Tim R. McVicar, Noemi Vergopolan, Alexis Berg, Nicholas J. Lutsko, Ambroise Dufour, Zhenzhong Zeng, Xin Jiang, Albert I. J. M. van Dijk, and Diego G. Miralles. High-resolution (1 km) Köppen–Geiger maps for 1901–2099 based on constrained CMIP6 projections. *Scientific Data*, 2023. 7

Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabé-Moreno, and Nicolas Longépé. TerraMesh: A planetary mosaic of multimodal Earth observation data. In *CVPR Workshop EarthVision*, 2025. 1, 2, 7

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1, 2

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *ICCV*, 2019. 1

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reBEN: Refined BigEarthNet dataset for remote sensing image analysis. In *IGARSS*, 2025. 9

Alistair Francis and Mikolaj Czerkawski. Major TOM: Expandable datasets for Earth observation. In *IGARSS*, 2024. 2, 7

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, et al. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS Datasets and Benchmarks*, 2023. 2

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800GB dataset of diverse text for language modeling. *arXiv:2101.00027*, 2021. 2

Charles Gaydon, Michel Daab, and Floryne Roche. Fractal: An ultra-large-scale aerial lidar dataset for 3D semantic segmentation of diverse landscapes. *arXiv:2405.04634*, 2024. 2

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. In *NeurIPS*, 2022. 1

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. TerraMind: Large-scale generative multimodality for Earth observation. In *ICCV*, 2025. 2, 7

Xuemei Jiang, Linus Scheibenreif, and Damian Borth. Less is more: Active self-supervised learning in remote sensing. In *IGARSS*. IEEE, 2024. 2

Amandeep Kaur, Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. Pre-train where? Investigating how pretraining data diversity impacts geospatial foundation model performance. In *CVPR Workshop EarthVision*, 2026. 2, 7

Thomas Kerdreux, Alexandre Tuel, Quentin Febvre, Alexis Mouche, and Bertrand Chapron. Efficient self-supervised learning for Earth observation via dynamic dataset curation. In *CVPR Workshop MORSE*, 2025. 2

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. GEO-Bench: Toward foundation models for Earth monitoring. In *NeurIPS Datasets and Benchmarks*, 2023. 6

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The BigScience ROOTS corpus: A 1.6 TB composite multilingual dataset. In *NeurIPS Datasets and Benchmarks*, 2022. 2

Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, 2021. 1, 2, 7

Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *ICCV*, 2023. 2

Dilxat Muhtar, Enzhuo Zhang, Zhenshi Li, Feng Gu, Yanglangxing He, Pengfeng Xiao, and Xueliang Zhang. Quality-driven curation of remote sensing vision-language data via learned scoring models. In *NeurIPS*, 2025. 2

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024. 1, 2

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *NeurIPS Datasets and Benchmarks*, 2023. 2

Amal Shehan Perera, Chandi Witharana, Elias Manos, Michael Pimenta, and Anna K. Liljedahl. Clustering guided domain-specific pretrained foundation model for very high-resolution Arctic remote sensing. *arXiv:2605.30467*, 2026. 2

Yohann Perron, Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. UniverSat: Resolution- and modality-agnostic transformers for Earth observation. *arXiv:2606.23503*, 2026. 6

Elena Plekhanova, Damien Robert, Johannes Dollinger, Emilia Arens, Philipp Brun, Jan Dirk Wegner, and Niklaus E. Zimmermann. SSL4Eco: A global seasonal dataset for geospatial foundation models in ecology. In *CVPR Workshop EarthVision*, 2025. 2

Laura Poggio, Luis M. de Sousa, Niels H. Batjes, Gerard B. M. Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 2021. 7

Mirali Purohit, Gedeon Muhawenayo, Esther Rolf, and Hannah Kerner. How does the spatial distribution of pre-training data affect geospatial foundation models? In *AAAI Workshop Good Data for Generative AI*, 2025. 2

Michael Schmitt, Lloyd Haydn Hughes, Chunping Qiu, and Xiao Xiang Zhu. SEN12MS – A curated dataset of georeferenced multi-spectral Sentinel-1/2 imagery for deep learning and data fusion. *ISPRS Annals*, 2019. 1, 2, 7

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *TMLR*, 2026. 1

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *NeurIPS*, 2022. 2

Adam J. Stewart, Nils Lehmann, Isaac A. Corley, Yi Wang, Yi-Chia Chang, Nassim Ait Ali Braham, Shradha Sehgal, Caleb Robinson, and Arindam Banerjee. SSL4EO-L: Datasets and foundation models for Landsat imagery. In *NeurIPS Datasets and Benchmarks*, 2023. 2, 7

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. 2

Kushal Tirumala, Dániel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *NeurIPS Datasets and Benchmarks*, 2023. 2

Diego Velazquez, Pau Rodriguez, Sergio Alonso, Josep M Gonfaus, Jordi Gonzalez, Gerardo Richarte, Javier Marin, Yoshua Bengio, and Alexandre Lacoste. EarthView: A large scale remote sensing dataset for self-supervision. In *WACV Workshops*, 2025. 2

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, et al. Automatic data curation for self-supervised learning: A clustering-based approach. *TMLR*, 2024. 2, 7, 8

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 2023. 1, 2, 7

Fan Wei, Runmin Dong, Yushan Lai, Yixiang Yang, Zhaoyang Luo, Jinxiao Zhang, Miao Yang, Shuai Yuan, Jiyao Zhao, Bin Luo, and Haohuan Fu. RS-Prune: Training-free data pruning at high ratios for efficient remote sensing diffusion foundation models. *arXiv:2512.23239*, 2025. 2

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. In *NeurIPS*, 2023. 1, 2

Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino. ESA WorldCover 10 m 2021 v200. Zenodo, 2022. 7

Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg, 2010. 8

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J. Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of Earth foundation models. *Communications Earth & Environment*, 2026. 1

Less Is More: Semantic Curation of Earth Observation Data for Self-Supervised Learning

Supplementary Material

In this Appendix, we provide additional analyses of whether semantic curation merely recovers metadata-based priors (Sec. A). We also include further qualitative examples (Sec. B), empirical motivation for the acquisition-correction scheme (Sec. C), its detailed derivation (Sec. C.1), and full implementation details (Sec. D).

A Does semantic curation align with metadata heuristics?

We ask whether embedding-based curation simply recovers a combination of standard EO heuristics, such as balancing geography, climate, land cover, or urbanization. We therefore measure the association between each method's keep/discard decision and several sample attributes (Tab. 3). For categorical attributes, we report Cramér's V; for continuous built-up fraction, we use point-biserial correlation. A selector that retains the same fraction of samples in every class has $V = 0$, whereas stronger class-dependent retention yields larger values.

As expected, metadata-based strategies are strongly associated with the attribute they explicitly target, with Cramér's V between 0.41 and 0.53 for the stratified criteria; the deterministic Europe baseline reaches $V = 1$ for the Europe indicator. By contrast, all embedding-based methods exhibit similarly weak associations with the tested metadata priors. For GeoSelect, these are 0.14 for climate, 0.14 for soil, 0.10 for land cover, and 0.01 for Europe.

Crucially, the different embedding-based variants have very similar association profiles despite large differences in downstream performance. Acquisition correction and spatial regularization therefore improve curation along dimensions not captured by these coarse metadata priors, rather than merely finding a better combination of them. Figure 6 supports the same conclusion: semantic curation leaves the marginal soil and land-cover distributions, and most climate classes, essentially unchanged. The main exceptions are arid hot-desert and polar ice-cap/frost climates, which are strongly downsampled by embedding-based curation; these classes are visually highly homogeneous, making their reduction consistent with semantic redundancy rather than explicit climate balancing.

B Additional qualitative examples

Impact of Corrections and Regularization. Figure 2 illustrates the effect of the two correction stages of Sec. 3 at one location. Figure 7 extends this analysis to six additional regions spanning diverse landscapes, including boreal wetlands, river deltas, island coasts, tropical forest, and savanna. Across all regions, acquisition correction removes the dominant structured variability and reveals geographically coherent patterns such as coastlines, river networks, and relief, while spatial regularization produces a smaller refinement.

Sentinel-1 embeddings. Figure 8 visualizes the Sentinel-1 embedding field used in Sec. 4.3. Compared with Sentinel-2, its leading components are more strongly dominated by the water–land contrast, while acquisition effects appear as large rectangular structures aligned with the sampling geometry.

C Acquisition-group analysis

Figure 9 examines the acquisition structure motivating Sec. 3.2. For each sample, we compute the semivariance between its embedding and those of neighboring samples at increasing geographic distances, separately for pairs from the same and different Sentinel-2 passes. Nearby samples are expected to have similar content, and therefore low semivariance, especially over open water where scene variability is limited. Instead, different-pass pairs exhibit high semivariance even at short

Criterion	Sampling	Kept	Geo. cells	Climate	Soil	Land cover	Built-up frac.	Europe
Uniform	stratified	10%	**0.41**	0.23	0.08	0.10	-0.01	0.01
Climate	stratified	10%	0.35	**0.48**	0.14	0.14	0.05	0.04
Soil	stratified	10%	0.30	0.15	**0.49**	0.12	0.02	0.01
Land cover	stratified	10%	0.30	0.19	0.14	**0.53**	0.07	0.01
Urban	top-%	10%	0.58	0.43	0.30	0.44	**0.43**	0.19
Europe	stratified	7.2%	0.98	0.50	0.27	0.19	0.08	**1.00**
DINOv2	SemDeDup	10%	0.21	0.15	0.14	0.23	-0.03	0.01
UniverSat	SemDeDup	10%	0.23	0.12	0.13	0.11	0.07	0.01
UniverSat	SemDeDup + spatial regul.	10%	0.23	0.13	0.13	0.12	0.07	0.01
UniverSat	SemDeDup + acq. correction	10%	0.23	0.13	0.13	0.09	0.05	0.01
UniverSat	GeoSelect	10%	0.23	0.14	0.14	0.10	0.06	0.01

Table 3: **Association between curation decisions and metadata attributes.** For categorical attributes, we report Cramér's V between the binary keep/discard decision and the sample attribute; the built-up fraction column reports point-biserial correlation. Higher values indicate stronger dependence of the selection on that attribute. **Bold** marks the attribute explicitly targeted by each metadata-based strategy.

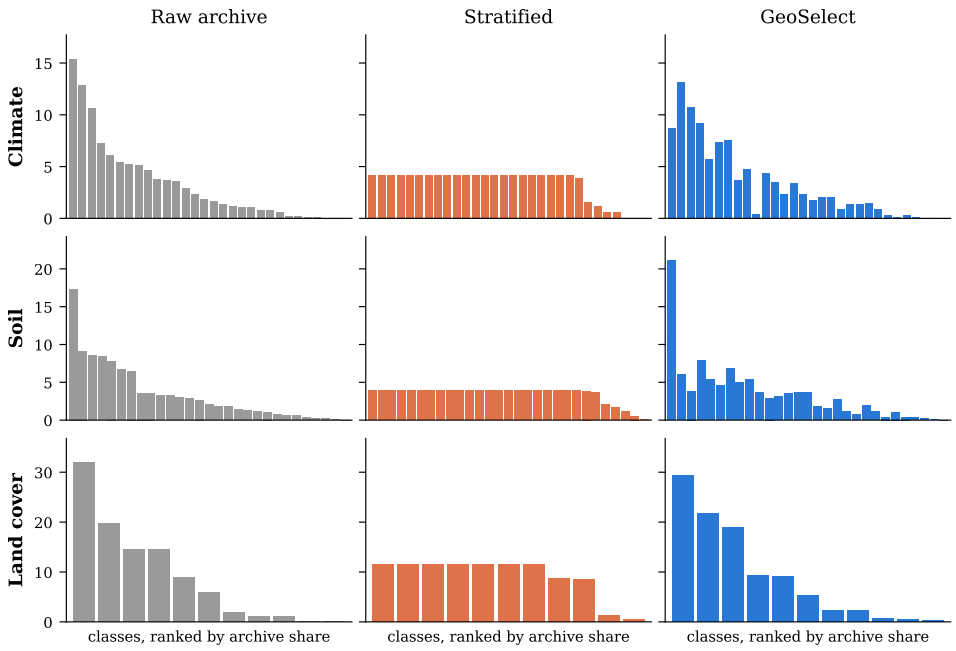
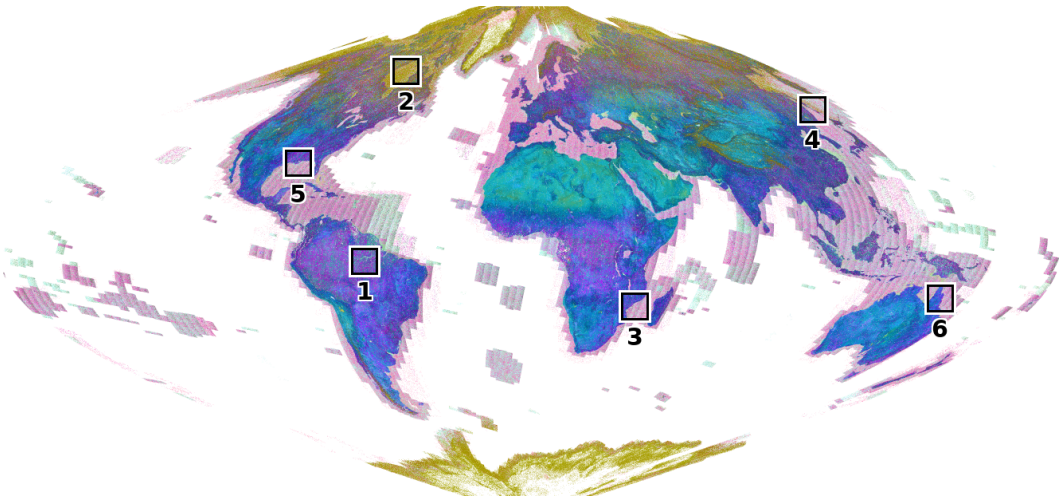


Figure 6: **Metadata distributions under different curation strategies.** Class frequencies for climate, soil, and land cover under the raw archive, the corresponding stratified baseline, and GeoSelect, all at a 10% budget. Classes are ordered by their frequency in the raw archive. Explicit stratification flattens the distributions by construction, whereas GeoSelect remains close to the original marginals.

distances, whereas same-pass pairs are substantially more similar: at 10 km, semivariance is 92% lower over all samples and 95% lower over open water. This shows that pass identity accounts for a large component of local embedding variability. Within-pass semivariance nevertheless increases with distance, motivating the finer spatial blocks used in our correction.



Raw embeddings, whole archive; numbered boxes mark the six windows below.

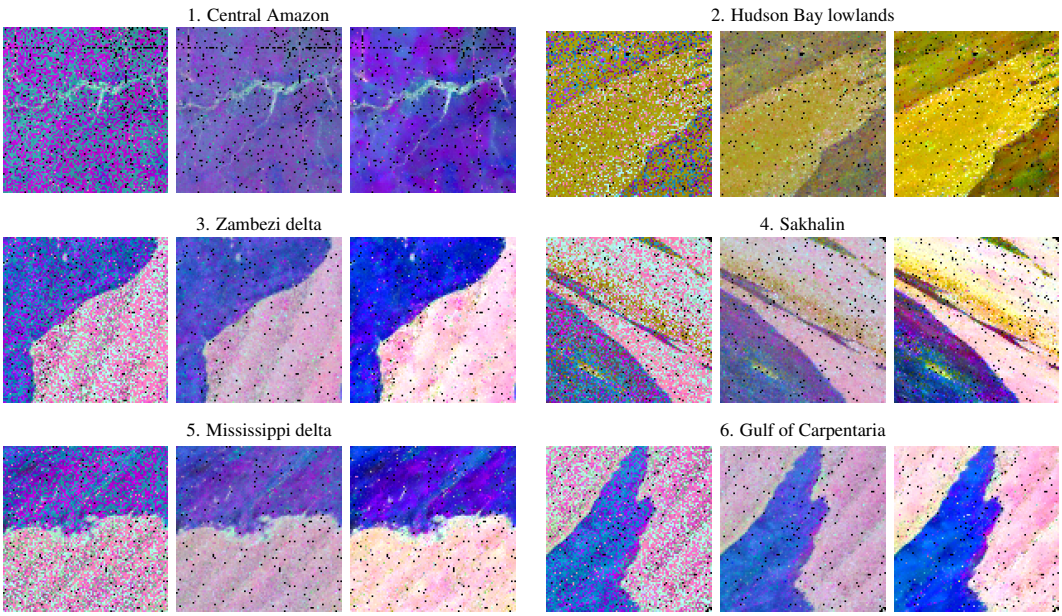


Figure 7: **Acquisition and spatial correction across diverse locations.** Within each triple: (left) raw UniverSat embeddings, (middle) after acquisition correction, and (right) after spatial regularization. Each panel shows a 1000×1000 km window of the Major-TOM grid, colored in RGB by the three leading principal components. All panels share the PCA basis and color stretch of Fig. 2. Acquisition correction reveals coastlines, river networks, and relief that are largely obscured by speckle in the raw embeddings, consistently across locations; spatial regularization then acts on the finer-scale variation remaining between acquisition bands.

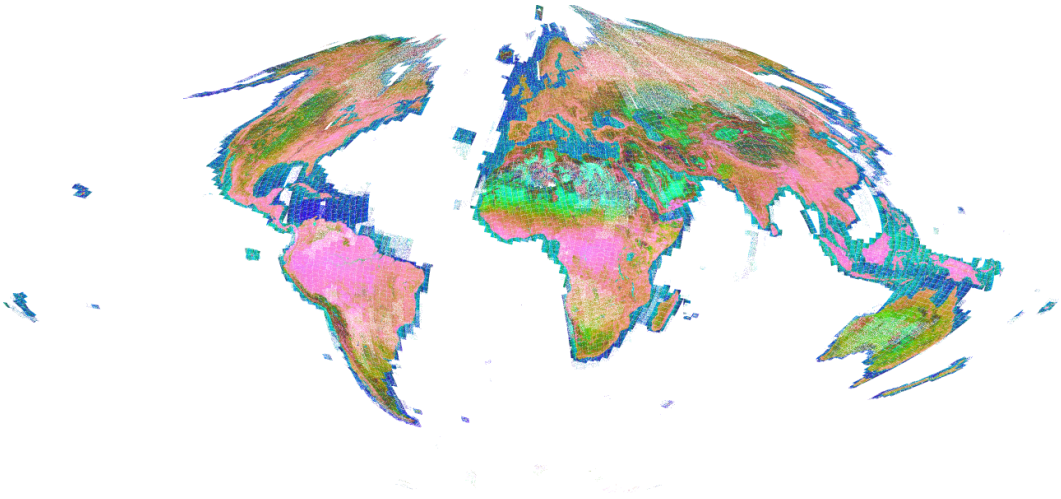


Figure 8: **Global Sentinel-1 embedding field.** RGB visualization of the first three principal components of frozen UniverSat embeddings over Major-TOM Core-S1RTC (1,458,492 samples), mapped on the Major-TOM grid using the same procedure as in Fig. 2.

C.1 Acquisition-effect estimation

We provide here the detailed estimation procedure for the acquisition-effect model of Eq. (2). We model the pass- and block-level offsets as zero-centered random effects,

$$\delta^{\mathrm{pass}}[p] \sim \mathcal{N}\left(0, \sigma_{\mathrm{p}}^2 I\right), \qquad \delta^{\mathrm{block}}[b] \sim \mathcal{N}\left(0, \sigma_{\mathrm{b}}^2 I\right), \qquad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 I\right). \tag{8}$$

These priors induce shrinkage toward zero for acquisition groups supported by few observations.

We jointly estimate the spatially smooth component s_i together with δ^{pass} and δ^{block} using Gauss–Seidel backfitting. Acquisition offsets are initialized to zero. At iteration $t+1$, we first update s_i by smoothing the embedding field after removing the current acquisition effects,

$$s_i^{(t+1)} = \frac{\sum_j a_{ij}\left(f_j - \delta^{\mathrm{pass},(t)}[\mathrm{p}(j)] - \delta^{\mathrm{block},(t)}[\mathrm{b}(j)]\right)}{\sum_j a_{ij}}, \tag{9}$$

$$a_{ij} = \exp\left(-\frac{d\left(\ell_i, \ell_j\right)^2}{2\sigma_s^2}\right), \qquad \sigma_s = 6\,\mathrm{km}, \tag{10}$$

where ℓ_i denotes the geographic location of observation i. Thus, s_i represents only the low-frequency component of the embedding field under the imposed spatial smoothness prior; it is not intended to isolate semantic content.

Given the updated smooth component, we update the pass-level effects,

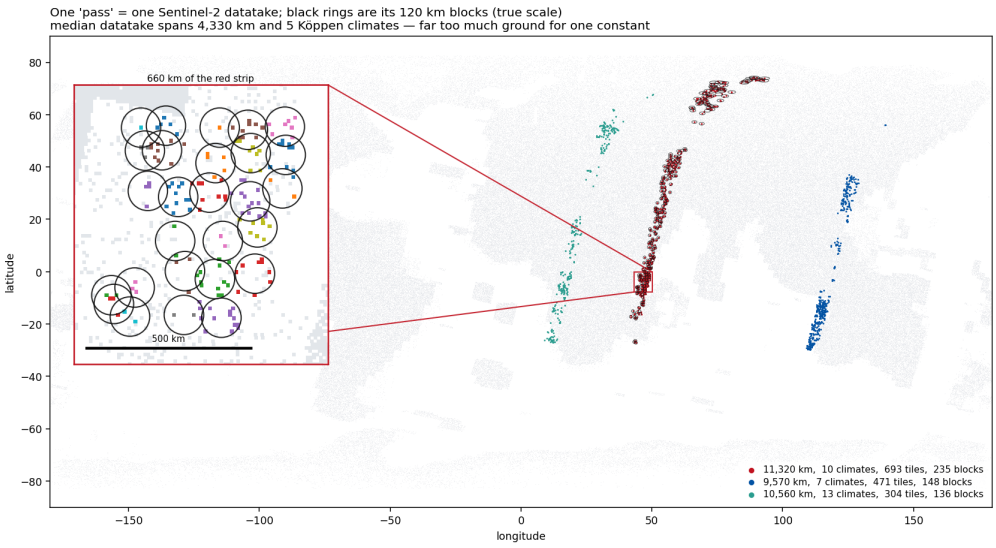
$$\delta^{\mathrm{pass},(t+1)}[p] = \frac{1}{n_p + \sigma_\varepsilon^2/\sigma_{\mathrm{p}}^2} \sum_{i:\,\mathrm{p}(i)=p} \left(f_i - s_i^{(t+1)} - \delta^{\mathrm{block},(t)}[\mathrm{b}(i)]\right), \tag{11}$$

followed by the block-level effects,

$$\delta^{\mathrm{block},(t+1)}[b] = \frac{1}{n_b + \sigma_\varepsilon^2/\sigma_{\mathrm{b}}^2} \sum_{i:\,\mathrm{b}(i)=b} \left(f_i - s_i^{(t+1)} - \delta^{\mathrm{pass},(t+1)}[\mathrm{p}(i)]\right), \tag{12}$$

where n_p and n_b denote the number of observations in pass p and block b, respectively. The use of the newly updated pass effect in Eq. (12), and of both updated acquisition effects in the subsequent smooth-component update, yields a Gauss–Seidel-style backfitting scheme.

The shrinkage factors arise directly from the Gaussian random-effect model: well-supported groups approach their empirical residual mean, whereas poorly-supported groups are pulled toward zero.



(a) Three example passes and their 120 km blocks

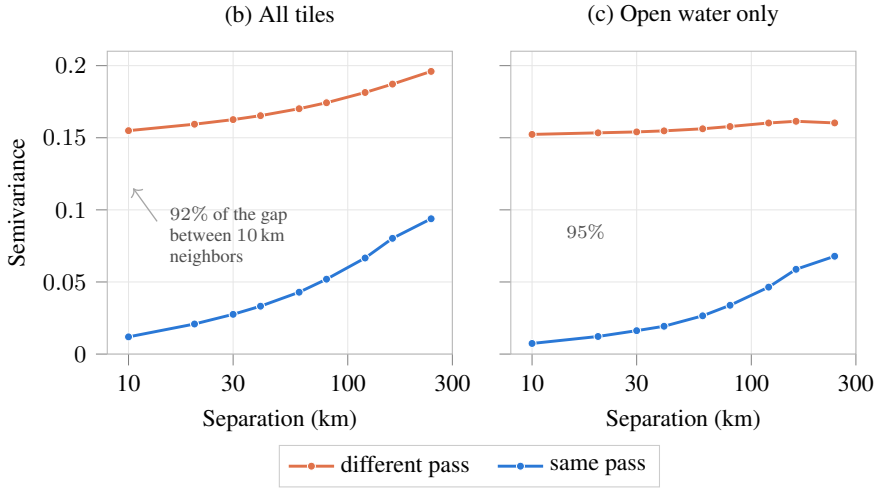


Figure 9: **Embedding variability across acquisition groups.** (a) Example Sentinel-2 passes on the Major-TOM grid, with their 120 km blocks shown in the inset. (b,c) Semivariance of UniverSat embeddings as a function of geographic separation for sample pairs from the same or different passes, over (b) all samples and (c) open-water samples. Same-pass and different-pass pairs are compared at matched geographic separations.

The variance terms σ^2_ε, σ^2_{p}, and σ^2_{b} are estimated from the data and introduce no additional curation hyperparameters. We use four backfitting iterations, after which the estimates are stable in practice.

The final corrected embedding is obtained by subtracting the estimated acquisition effects from the original embedding,

$$f_i^{\mathrm{corr}} = f_i - \delta^{\mathrm{pass}}[\mathrm{p}(i)] - \delta^{\mathrm{block}}[\mathrm{b}(i)]. \tag{13}$$

Importantly, the Gaussian smoothing is used only to estimate the spatially smooth component needed during backfitting; it is not applied directly to f_i^{corr}.

D Implementation Details

We report the full MAE pretraining configuration in Tab. 4. The curation pipeline of Sec. 3 has its own configuration, reported in Tab. 5: it is run once per archive and is independent of everything trained on its output.

Hyperparameter	Value
Backbone (ViT-Base MAE)	
Patch size	16
Embedding dim	768
Depth / heads	12 / 12
MLP ratio	4.0
Registers	4
Drop-path	0.1
Decoder depth	6
Masking	
Spatial ratio	0.8
Spectral drop	0.1
Optimization	
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	0.05
Adam ϵ	10^{-5}
Scheduler	ReduceLROnPlateau (factor 0.1, patience 10)
Epochs	200 (early stop, patience 20)
Global batch size	256
Precision	bfloat16
Seed	32
Data / evaluation	
Input size	256×256
Channels (S2 / S1)	12 / 3
Tokens per tile	256
Linear-probe interval	every 5 epochs

Table 4: **MAE pretraining hyperparameters** for the model trained on the (de)duplicated MajorTOM data. Values from the `MajorTOM_MAE` configuration.

Why the shrinkage is adaptive. Equation (6) shrinks each embedding toward its spatial estimate h_i by a per-tile weight λ_i that decreases with local neighbor support ($\kappa = 4$ in Tab. 5), rather than by a single constant. We verified this choice with an internal ablation that replaces it with a flat $\lambda = 0.7$ for every tile with at least one neighbor ($\kappa = 0$): under the same 10% budget, recipe, and checkpoint (ocean included, as in Tab. 5), this costs -0.98 points on classification and -1.04 mIoU on segmentation. Coasts, islands, and coverage gaps are exactly where spatial support is weakest, so a flat λ over-smooths them; the adaptive term in Eq. (6) is what keeps the correction conservative there.

Stage / hyperparameter	Value
Embedding (Sec. 3.1)	
Scorer	UniverSat, frozen
Bands	10 S2 (B02–B08, B8A, B11, B12)
Input	whole tile, 1068×1068 at 10 m
Patch size	120 m
Pooling	mean over spatial tokens
Embedding dim	768, ℓ_2-normalized
Storage	fp16 memmap, 3.2 GiB for 2.24 M tiles
Clustering (Sec. 3.4)	
Algorithm	k-means, fp16 assign / fp32 similarity
Clusters	11,193
Iterations	20, on all 2.24 M tiles
Acquisition correction (Sec. 3.2)	
Grouping	sensing timestamp $\times$ 120 km block
Levels	pass and block-within-pass, fitted jointly
Shrinkage	empirical Bayes, $n/(n + \hat{\sigma}_\varepsilon^2/\hat{\sigma}_a^2)$
Geography estimate s	Gaussian, $\sigma = 6$ km on the grid
Backfitting iterations	4
Spatial regularization (Sec. 3.3)	
Neighbors	17 within 25 km
Field estimator	Gaussian MRF, 15 Jacobi sweeps
Prior strength μ	1
Geographic kernel	Gaussian, $\sigma = 15$ km
Range term	on
Shrinkage λ	0.7
Support constant κ	4, $\lambda_i = \lambda\, n_i/(n_i + \kappa)$, n_i neighbors
Deduplication	
Candidate pool	whole archive
Priority order	decreasing distance to centroid
Similarity	cosine on f^{smooth} (Eq. (7))
Budget $\rightarrow$ threshold	quantile of r_i
Seed	42

Table 5: **Curation configuration.** The pipeline of Sec. 3, run once per archive. Clustering, atypicality ordering, and the redundancy score all operate on the same spatially regularized embeddings f^{smooth}.