FROM PIXELS TO PATCHES: POOLING STRATEGIES FOR EARTH EMBEDDINGS*

Isaac Corley[λ], **Caleb Robinson**[γ], **Inbal Becker-Reshef**[γ], **Juan M. Lavista Ferres**[γ]
[λ]Wherobots, [γ]Microsoft AI for Good Research Lab

ABSTRACT

Geospatial foundation models increasingly expose pixel-level embedding products that can be downloaded and reused without access to the underlying encoder. In this setting, downstream tasks with patch- or region-level labels require a post-hoc aggregation step that maps dense pixel embeddings to a single representation. The default choice, mean pooling, discards within-patch variability and can underperform under spatial distribution shift. To study this setting, we introduce *EuroSAT-Embed*: 81,000 embedding GeoTIFFs derived from three foundation models: AlphaEarth, OlmoEarth, and Tessera. Using these fixed embedding products, we benchmark 11 training-free pooling methods and 2 train-set-fitted baselines under both random and geographically disjoint test splits. Richer pooling schemes reduce the geographic generalization gap by over 50% relative to mean pooling and improve accuracy by up to 6% on spatial splits. We recommend a three-tier strategy: (1) *mean* as a baseline, (2) *stats* pooling (min/max/mean/std) as the default at 4× the embedding dimension, and (3) *covariance* pooling for peak accuracy. Across all three embedding products, simple distributional statistics improve spatial-split performance over mean pooling.

1 INTRODUCTION

Geospatial foundation models (GFMs) turn satellite data time-series into embedding data products that can be reused across tasks and regions Fang et al. (2026). Recent GFMs create dense satellite imagery embeddings at *pixel* resolution, but many downstream tasks operate on objects or regions (fields, parcels, buildings, solar farms, gridded patches), requiring aggregation of pixel embeddings over polygons. Pooling is therefore an inherent step when using GFM output for these tasks, but naive methods may erase relevant signals in the embeddings.

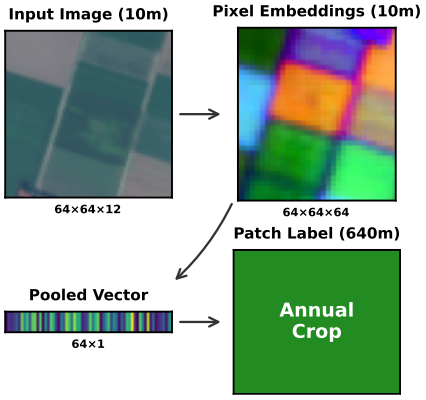

Figure 1: **Pixel-to-patch pooling.** The input-label resolution mismatch requires aggregating dense pixel embeddings (shown as PCA pseudo-RGB) to a lower resolution for downstream tasks.

This problem compounds when labels exist at a coarser-than-pixel resolution, known as the *input-label resolution mismatch* (Workman et al., 2023). For example, EuroSAT (Helber et al., 2019) contains 64×64 patches of 10 m pixels corresponding to a single land-cover label, which can be viewed as a *zonal-statistics* aggregation operation (Cressie, 2015). Unlike raw spectral bands, pixel embeddings encode semantic features that vary within a land-cover patch. This heterogeneity suggests that averaging can remove class signal that richer statistics keep. This makes pooling a design decision for geospatial applications with coarse labels and/or temporal dimensions.

*No swimming required.

Pooling is underexplored for earth embeddings[1], though it is heavily studied in related domains. In image retrieval, choices like generalized mean (GeM) pooling (Radenović et al., 2018) often matter as much as the learned encoder. In audio and video processing, statistical pooling (concatenating means and standard deviations) effectively summarizes variable-length sequences (Miech et al., 2017). Mean pooling dominates for its simplicity, but its performance under spatial shift is not well measured. Classical object-based image analysis (EO4GEO, 2026) methods like Bag of Visual Words (BoVW) (Csurka et al., 2004) aggregate hand-crafted features based on region statistics for vision tasks, but there is little work on aggregating modern pixel embeddings into patch-level representations.

In this paper, we study the post-hoc pooling step in the fixed-embedding setting. Given pixel embeddings from a released geospatial embedding product, how should a practitioner aggregate them into a patch-level representation when the encoder itself is unavailable? We benchmark 11 training-free methods and 2 train-set-fitted baselines using pixel embeddings generated by three GFMs on the common EuroSAT land-cover classification dataset Helber et al. (2019). We evaluate the effects of pooling choice with kNN and linear probes under the standard random and geographically disjoint dataset splits.

Specifically, we release *EuroSAT-Embed* and benchmark 13 pooling methods across three fixed embedding products, two probes, and both random and spatial splits.

Contributions. We make three contributions: (1) we release *EuroSAT-Embed*, a set of three aligned pixel-level embedding datasets derived from AlphaEarth, OlmoEarth, and Tessera; (2) we provide a controlled benchmark of 11 training-free pooling methods and 2 train-set-fitted baselines across two probes and both random and spatial splits in the fixed-embedding setting; (3) we show that simple distributional pooling methods, especially *stats*, consistently improve spatial generalization over mean pooling, while *covariance* offers the strongest accuracy when higher-dimensional representations are acceptable.

2 Data and Embeddings

EuroSAT. We use the 10-class EuroSAT land-cover dataset (Helber et al., 2019) which contains 27,000 Sentinel-2 patches at 64×64 pixels with classes such as forest, residential, and agricultural land. We evaluate on both the standard random split (Neumann et al., 2019) and a spatial split (Ekim et al., 2025) that partitions samples by longitude to reduce train-test leakage via spatial autocorrelation.

EuroSAT-Embed. We construct three aligned embedding datasets from AlphaEarth (Brown et al., 2025) (64-d), OlmoEarth-Nano (Herzog et al., 2025) (128-d), and Tessera (Feng et al., 2025) (128-d). AEF and OlmoEarth embed patch context but emit *pixel-wise* vectors, whereas Tessera encodes per-pixel time-series without spatial context of adjacent pixels. We use these embedding products to study pooling behavior, not to compare the underlying GFMs.

3 Pooling Methods

We benchmark training-free operators that compress an $H \times W \times D$ embedding tensor into a d-dimensional vector. We hypothesize that class signal lies not only in the mean but in the distribution (variance and extremes). For example, a residential neighborhood might share a mean embedding over space with industrial areas, but show higher variance from mixed rooftops, vegetation, and roads. Let $X \in \mathbb{R}^{H \times W \times D}$ denote a patch with $N{=}HW$ pixels. Pooling operates per channel (e.g., $\max(X)$ returns the per-channel maximum). Note that pooling methods can increase the dimensionality by a multiplier due to concatenation of multiple statistics.

First-order statistics (D-d each): *mean* ($\mu = \frac{1}{N}\sum_i x_i$) captures the "typical" pixel but discards variation; *max* preserves extreme activations; generalized mean (*GeM*) interpolates between them (Radenović et al., 2018); *center-weighted mean* down-weights boundary pixels.

[1] A relevant exception is the recent OlmoEarth paper Herzog et al. (2025) that reports searching over mean and max temporal pooling strategies as a hyperparameter in downstream tasks.

Table 1: Linear probe accuracy across embedding sources. Cell shading indicates higher accuracy within each column (darker = higher). Best per column in **bold** and second-best in *italics*. Gap = accuracy drop from random to spatial split. *OlmoEarth-Nano variant.

	Spatial split				Random split				
Pooling	AEF	OlmoEarth*	Tessera	Avg	AEF	OlmoEarth*	Tessera	Avg	Gap↓
Std	78.6	90.6	87.8	85.7	90.7	94.8	95.5	93.6	8.0
Mean	84.0	92.3	85.5	87.3	95.5	96.5	96.1	96.0	8.8
GeM	85.0	91.4	86.9	87.8	95.6	95.9	96.5	96.0	8.2
Center-Weighted	84.9	92.4	87.0	88.1	96.2	96.7	96.3	96.4	8.3
Max	86.0	91.1	89.7	88.9	92.9	93.0	93.8	93.2	4.3
Median+IQR	83.6	93.3	91.7	89.5	94.6	96.1	96.2	95.6	6.1
Mean+Std	82.4	92.7	93.9	89.7	96.0	96.5	97.3	96.6	6.9
Percentiles	87.4	*93.8*	92.8	91.3	96.0	*96.8*	97.1	96.6	5.3
Mean+Max	88.2	93.2	93.9	91.8	*96.3*	96.1	96.7	96.4	4.6
Covariance	**91.0**	92.9	**94.4**	*92.8*	96.2	**97.3**	**97.7**	**97.1**	*4.3*
Stats	*90.6*	**94.1**	*94.3*	**93.0**	**96.4**	96.8	*97.4*	*96.8*	**3.8**
Parametric pools (fit on train set)									
PCA	80.8	91.5	79.3	83.9	95.4	95.9	95.0	95.4	11.6
BoVW	89.0	88.3	88.7	88.7	94.6	94.4	96.5	95.1	6.5

Distributional statistics (D-d to $5D$-d): *std* captures pixel variability; *mean+std* ($2D$-d) and *mean+max* ($2D$-d) concatenate complementary signals; *stats* ($4D$-d) stacks min/max/mean/std; *percentiles* ($5D$-d) computes five summary quantiles; *median+IQR* ($2D$-d) provides robust location and spread estimates; *covariance* extracts the upper triangle of the pixel covariance matrix ($D(D+1)/2$-d).

Parametric methods (fit after embeddings are produced): *PCA* projects mean-pooled embeddings to lower dimensions; *Bag of Visual Words (BoVW)* (Csurka et al., 2004) clusters pixel embeddings with mini-batch k-means and represents each patch as a normalized histogram of cluster assignments. These still operate in the post-hoc setting, but unlike the training-free pooling operators above, they fit an additional summarization stage on the training split.

4 Experimental Setup

We evaluate two probes: kNN (k=5, cosine distance) and multinomial logistic regression, selecting regularization C for each pooling method by 3-fold cross-validation on the training split only. kNN probes geometry; linear probes test linear separability. We standardize before linear probing. We evaluate on both random and geographically disjoint (spatial) splits. Results are reported in Tables 1, 2.

5 Results

Distributional statistics improve generalization. For linear probes, *stats* pooling reaches 91–94% on spatial splits (avg 93.0%) versus 84–92% for *mean* (avg 87.3%), a +6% gain (Table 1). *Covariance* pooling achieves the highest accuracy on two of three encoders (AEF: 91.0%, Tessera: 94.4%) at $D(D+1)/2$ dimensions. On random splits, differences narrow to ∼1% as spatial leakage boosts all methods.

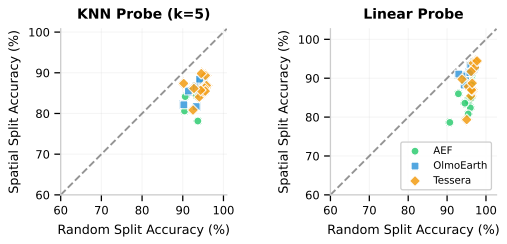

Figure 2: **Random vs spatial accuracy.** Points near the diagonal (where random = spatial) have smaller generalization gaps.

kNN results (Table 2) show consistent trends: *stats* leads on spatial splits (avg 88.0%) while *center-weighted mean* leads on random splits (avg 95.3%). Second-order statistics (*covari-*

Table 2: KNN accuracy (k=5) across embedding sources. Cell shading indicates higher accuracy within each column (darker = higher). Best per column in **bold** and second-best in *italics*. Gap = accuracy drop from random to spatial split. *OlmoEarth-Nano variant.

	Spatial split				Random split				
Pooling	AEF	OlmoEarth*	Tessera	Avg	AEF	OlmoEarth*	Tessera	Avg	Gap↓
Std	78.1	81.7	84.0	81.3	93.7	93.3	94.1	93.7	12.4
Max	80.6	85.5	80.8	82.3	90.4	91.4	92.5	91.4	9.1
GeM	85.8	86.6	85.4	85.9	94.6	93.3	95.1	94.3	8.4
Median+IQR	86.1	86.0	85.7	85.9	94.1	94.0	94.4	94.1	8.2
Mean	85.6	87.8	84.7	86.0	94.6	94.8	94.9	94.8	8.7
Covariance	84.5	84.8	**89.8**	86.4	93.4	94.3	94.5	94.1	7.7
Center-Weighted	86.1	87.8	85.5	86.5	**95.4**	**95.0**	95.5	**95.3**	8.8
Mean+Max	85.4	87.9	86.9	86.7	94.7	94.4	**95.8**	94.9	8.2
Percentiles	*86.8*	87.5	86.2	86.8	94.6	94.9	95.4	95.0	8.1
Mean+Std	86.5	*88.3*	86.7	*87.2*	94.9	*94.9*	95.3	*95.0*	7.9
Stats	**87.0**	87.6	*89.3*	**88.0**	*95.0*	94.4	*95.6*	95.0	7.0
Parametric pools (fit on train set)									
BoVW	84.1	82.1	87.4	84.6	90.6	90.3	90.2	90.4	**5.8**
PCA	85.4	**88.4**	86.1	86.7	93.5	94.1	92.7	93.5	*6.8*

ance) help kNN less than linear probes, suggesting that they benefit linear classification more than similarity search.

Generalization gaps. Figure 2 shows methods diverge under distribution shift. *Mean* pooling drops 8.8 pp (random to spatial); *stats* drops only 3.8 pp—a 57% reduction in gap. *Covariance* (4.3 pp) and *mean+max* (4.6 pp) also exhibit small gaps while maintaining high accuracy.

6 Discussion

Recommendation. For practitioners working with fixed pixel-level embedding products, pooling should be treated as a first-order design choice rather than a default preprocessing step. *Mean* pooling ($1\times$) is a useful baseline but leaves significant accuracy on the table under spatial shift. When a modest increase in representation size is acceptable, *stats* pooling ($4\times$) is the strongest overall default across encoders and probes on spatial splits. When representation size is less constrained and maximum accuracy is the priority, *covariance* pooling ($D(D{+}1)/2$-d) is often strongest.

Accuracy and robustness are aligned. We do not observe an accuracy–robustness tradeoff: methods with the highest spatial-split accuracy also tend to have the smallest random-to-spatial generalization gaps. In particular, *stats*, *covariance*, and *mean+max* are both more accurate and more robust than *mean*.

Why distributional statistics help. *Mean* pooling collapses each patch to a first-moment summary, discarding intra-patch heterogeneity; distributional pooling preserves variability that appears useful under spatial shift.

Limitations. We evaluate on a single benchmark (EuroSAT, 10 classes) with three embedding sources; broader coverage across datasets such as BigEarthNet and fMoW, and multi-label settings, would strengthen conclusions. We focus on training-free pooling; learned methods such as attention pooling or adaptive GeM may improve further but require task-specific training (Touvron et al., 2022). Our conclusions should be interpreted as guidance for post-hoc pooling of fixed embedding products rather than for end-to-end learned aggregation with encoder access. On EuroSAT-Embed, pooling choice materially changes spatial-split performance, and simple distributional summaries are a strong default.

References

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon

Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pp. 1–2. Prague, 2004.

Burak Ekim, Girmaw Abebe Tadesse, Caleb Robinson, Gilles Hacheme, Michael Schmitt, Rahul Dodhia, and Juan M Lavista Ferres. Distribution shifts at scale: Out-of-distribution detection in earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2265–2274, 2025.

EO4GEO. Object-based image analysis (obia) introduction. `https://eo4geocourses.github.io/PLUS_OBIA-Introduction`, 2026. Accessed 2026-01-28.

Heng Fang, Adam J. Stewart, Isaac Corley, Xiao Xiang Zhu, and Hossein Azizpour. Earth embeddings as products: Taxonomy, ecosystem, and standardized access, 2026. URL `https://arxiv.org/abs/2601.13134`.

Zhengpeng Feng, Clement Atzberger, Sadiq Jaffer, Jovana Knezevic, Silja Sormunen, Robin Young, Madeline C. Lisaius, Markus Immitzer, Toby Jackson, James Ball, David A. Coomes, Anil Madhavapeddy, Andrew Blake, and Srinivasan Keshav. Tessera: Temporal embeddings of surface spectra for earth representation and analysis, 2025. URL `https://arxiv.org/abs/2506.20380`.

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

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. *arXiv preprint arXiv:2511.13655*, 2025.

Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017.

Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain representation learning for remote sensing. *arXiv preprint arXiv:1911.06721*, 2019.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, and Hervé Jégou. Augmenting convolutional networks with attention-based aggregation. In *International Conference on Machine Learning*, pp. 21668–21680. PMLR, 2022.

Scott Workman, Armin Hadzic, and M Usman Rafique. Handling image and label resolution mismatch in remote sensing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3709–3718, 2023.