

SLED: Scalable Location Encoding via Distillation

Kevin Lane¹, Zhongying Wang¹, Esther Rolf¹, Morteza Karimzadeh¹

¹University of Colorado Boulder

Kevin.Lane@colorado.edu, Zhongying.Wang@colorado.edu, Esther.Rolf@colorado.edu, Karimzadeh@colorado.edu

Abstract

The plethora of readily available geospatial data offers exciting opportunities to learn high quality representations of the planet, but the sheer size of the Earth Observations (EO), differing modalities, and different sensor types pose significant challenges in doing so. Location encoders have emerged as an efficient way of compressing EOs into location-specific embeddings. However, current state-of-the-art location encoders rely on computationally expensive CLIP-style frameworks that require large batch sizes in the 16K–32K range, suffer from false negative samples, and scale poorly with additional modalities. We introduce the Scalable Location Encoder via Distillation (SLED), a distillation-based location encoder that uses geospatial location as a binding modality to pretrain location encoders with any modality of geospatial data. The resulting location encoder framework is lightweight, modular, and can flexibly incorporate multiple modes, while eliminating the need for spatiotemporal coregistration of samples. SLED is performant with batch sizes as small as 128, enabling pretraining at a fraction of the runtime and compute costs of current state-of-the-art models. We demonstrate our approach by pretraining unimodal and multimodal SLED models on Sentinel-1, Sentinel-2, and Landsat imagery. We show that both unimodal and multimodal SLED models keep pace with or outperform existing approaches on a diverse set of 19 human-centric benchmark tasks and explore the benefits of using additional modes in pretraining.

Code & Pip Install — <https://github.com/geohai/sled>

Datasets & Weights — <https://huggingface.co/geohai>

Introduction

Models capable of learning general purpose representations of Earth Observations (EO) have become increasingly popular (Mai et al. 2025). Embeddings generated from one such geospatial foundation model (GeoFM) can be leveraged on a variety of downstream geospatial tasks (Rolf et al. 2021), including flood plain mapping (Bonafilia et al. 2020), air pollution monitoring (Karimzadeh, Wang, and Crooks 2025), land cover classification (Christie et al. 2018), and many others (Lacoste et al. 2023).

Despite the efficacy of GeoFM embeddings, using them can be challenging. In order to generate embeddings, users have to first acquire and process data of a particular modality—or modalities, in the case of multimodal GeoFMs such as Astruc et al. (2025); Danish et al. (2025); Fuller, Millard, and

Green (2023); Jakubik et al. (2025); Xiong et al. (2024)—and run inference on them. For perspective, the popular Sentinel-2 satellite constellation produces 1.7 TB of data per day (Sudmanns et al. 2020) and has been in orbit for over a decade. The sheer size means that downloading data and passing samples through pretrained GeoFM models often requires significant time, storage, and computational resources.

Given these limitations, a type of GeoFM called a location encoder has emerged (Mai et al. 2023; Klemmer et al. 2025b; Vivanco Cepeda, Nayak, and Shah 2023; Dollinger et al. 2025), which treats geospatial position, in the form of latitude and longitude, as its own modality. These models train an encoder to learn representations of geospatial position by feeding the location encoder latitude and longitude, then contrasting the resulting location embeddings with EO embeddings associated with that location extracted from pretrained encoders. Thus, location encoders learn how to parametrically generate representations of *any* geospatial location, and are capable of interpolating to spatial gaps within their training samples (Klemmer et al. 2025b). Location encoders do not require (but optionally allow the appendage of) an input image (or other modes) at the time of inference (Klemmer et al. 2025b; Vivanco Cepeda, Nayak, and Shah 2023; Mai et al. 2023; Astruc et al. 2026; Liu et al. 2025). Given the inherent geospatial nature of many tasks, location embeddings can be easily leveraged in existing workflows and can dramatically improve performance on classification and regression tasks alike (Wu et al. 2024; Karimzadeh, Wang, and Crooks 2025).

While existing location encoders represent a significant step forward in accessible generation of embeddings, they generally remain computationally expensive to train. Almost all existing location encoders (Klemmer et al. 2025b; Mai et al. 2023; Vivanco Cepeda, Nayak, and Shah 2023; Astruc et al. 2026; Liu et al. 2025) rely on the InfoNCE loss (Radford et al. 2021) between location embeddings and modality-specific embeddings. This requires large batch sizes (Klemmer et al. 2025b), often in the range of 16k–32K, which can create compute bottlenecks. InfoNCE loss also suffers from false negatives, where similar samples within a batch are treated the same as disparate samples. Given the inherent similarity of spatially-nearby locations (Waters 2017), this is a particularly undesirable property for location encoders. Addressing false negatives remains an active area of research for

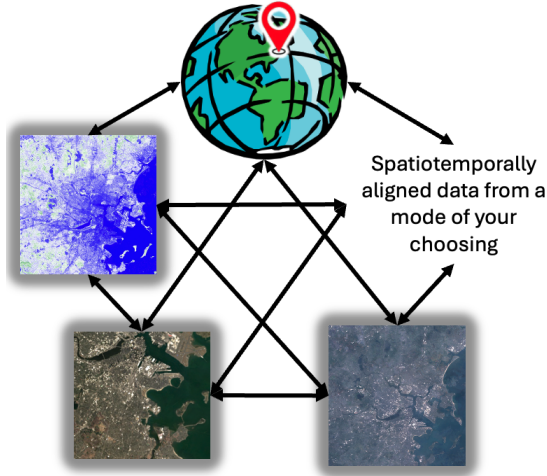


Figure 1: Multimodal location encoding without location as a binding modality. All modes must be spatiotemporally coregistered per sample. The sample above is of Boston, MA, USA, with samples from Sentinel-1, Sentinel-2, and Landsat 9 in Fall, 2024.

CLIP-style models (Auh, Cho, and Kim 2026; Huynh et al. 2022).

Despite their promising qualities, existing location encoders either do not leverage multimodal geospatial data or do so in a way that constrains their training dataset. Many current location encoders are trained with a single modality of geospatial data, perhaps due to the computational cost of adding additional modalities and the required batch size for contrastive training. Those that train against more than one mode do so via contrasting samples across all data, as seen in Figure 1. This requires costly cross-modal coregistration for every sample and excludes any training samples that do not have all modes of information available at a given location. It also introduces subjectivity into the training process, as samples are rarely perfectly spatiotemporally aligned across modes and thus require some spatiotemporal buffering (Stewart et al. 2023), but the degree to which that is done varies from dataset to dataset.

We propose a novel distillation-style framework, the Scalable Location Encoder via Distillation (SLED), for distilling location embeddings from (vision) GeoFMs. Our framework’s computational complexity scales linearly with batch size, avoids false negative issues present in existing models, and exhibits strong performance with batch sizes as small as 128. We also take inspiration from ImageBind (Girdhar et al. 2023) and (Klemmer et al. 2025a) to treat location as a binding modality when training multimodal versions of SLED. This allows us to leverage new modalities while avoiding the subjective and costly process of spatiotemporally coregistering samples from different modalities, as demonstrated in both Figure 1 and Figure 2.

Our core contributions are as follows:

1. We demonstrate that it is possible to distill unimodal and multimodal EOs into performant location embeddings via

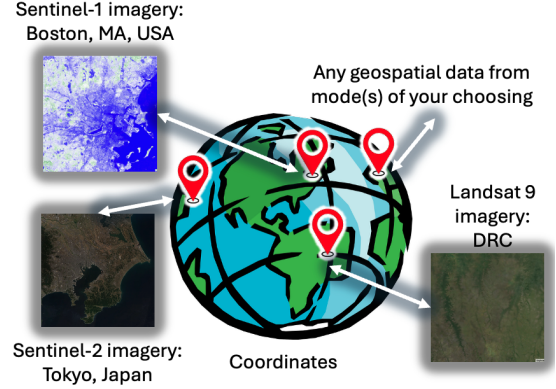


Figure 2: Multimodal location encoding with location as a binding modality. We leverage the same satellite sensors present in Fig 1, but can do so via independent samples. Note the lack of requirement for spatiotemporal coregistration between modes.

Mean Squared Error (MSE) Loss with batch sizes as small as 128 samples.

2. We show that SLED’s distillation framework speeds up pretraining by up to 47x when compared to pretraining times for existing state-of-the-art (SOTA) location encoders.
3. We make pretrained versions of SLED and our pretraining framework available via a pip module `sled-geo`, which is designed to work with `pytorch-lightning`.
4. We release a location encoder benchmark dataset of climate-related variables and a separate uniform-at-random Sentinel-1 location encoder pretraining dataset.

Related Work

Among the various GeoFMs developed, two broad categories have emerged:

1. Foundation models trained to generate embeddings from one or more of their geospatial training modalities, such as Sentinel-2. Examples include TerraFM (Danish et al. 2025), FG-MAE (Wang et al. 2024), OmniSat (Astruc et al. 2025), DOFA (Xiong et al. 2024), DeCUR (Wang et al. 2023), CROMA (Fuller, Millard, and Green 2023), and GFM (Mendieta et al. 2023). These models require access to the input modality (or modalities) at downstream inference, follow a myriad of different designs, and have been surveyed in (Lane and Karimzadeh 2026) and (Lu et al. 2025).
2. Location encoder models trained to generate embeddings directly from latitude and longitude, without requiring access to the original input modality. Prior work typically use contrastive learning against geospatial representations derived from data at those locations. Examples include CSP (Mai et al. 2023), SatCLIP (Klemmer et al.

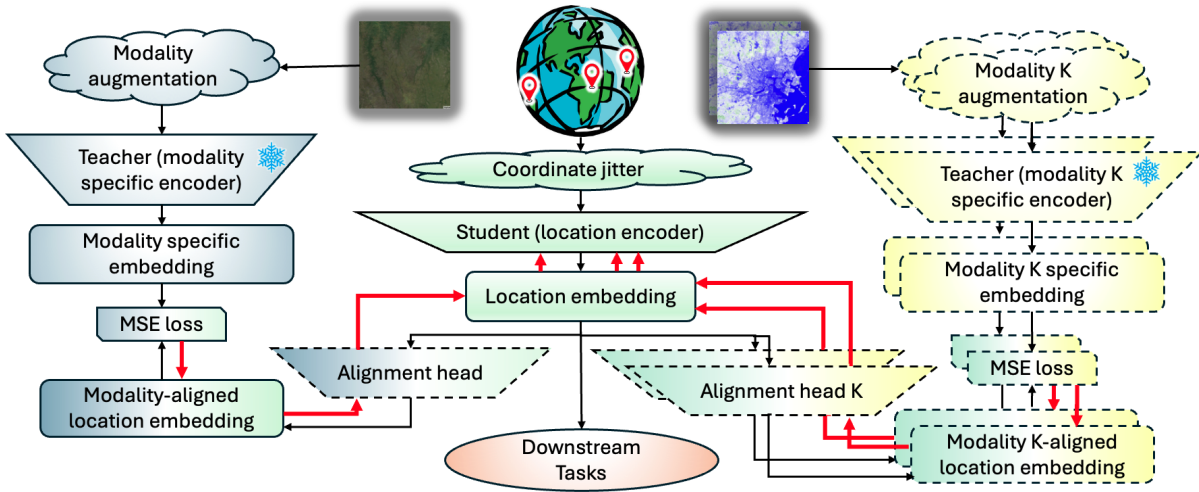


Figure 3: Diagram of SLED’s training framework with 1 to K modes. The current model outlines when $K = 3$. Components outlined with a dashed line are only present when training with 2 or more modes. For unimodal training, no alignment head is present. The red arrow in bold denotes back-propagation.

2025b), GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023), GAIR (Liu et al. 2025), CLIMPLICIT (Dollinger et al. 2025), and UniGeoCLIP (Astruc et al. 2026). Of these, only GAIR (Liu et al. 2025) and UniGeoCLIP (Astruc et al. 2026) are pretrained with multiple data modalities. Both rely on InfoNCE loss and require spatiotemporal alignment of their training modalities. CLIMPLICIT (Dollinger et al. 2025) is the only model present that does not rely on InfoNCE loss, rather using MSE loss against a vector of climatic variables.

The first type of model requires relatively complete spatial coverage of input modalities to generate embeddings indexed to latitude and longitude, as in MOSAICS (Rolf et al. 2021), AlphaEarth (Brown et al. 2025), and OlmoEarth (Herzog et al. 2025). In contrast, location encoders learn spatially continuous representations for *any* location without requiring a database of indexed embeddings wherever input modalities were available. This is partially enabled by the positional encoding component of the location encoder.

Positional encodings map geographic coordinates to (learnable) high-dimensional representations (Mai et al. 2022). Recent work builds on concepts of periodic activation functions (Sitzmann et al. 2020) to enable learning high-frequency spatial signals, by incorporating Earth-aware basis functions, such as spherical harmonics, in conjunction with sinusoidal neural networks to produce continuous, multi-scale embeddings of geographic coordinates (Rußwurm et al. 2024). While they use similar loss functions, existing work on pretrained location encoders vary in their position encoding strategy. SatCLIP (Klemmer et al. 2025b) utilizes spherical harmonics (Rußwurm et al. 2024), while GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023), GAIR (Liu et al. 2025), and UniGeoCLIP (Astruc et al. 2026) use some form of Random Fourier Features (RFF) (Tancik et al. 2020).

Location encoders differ in training modalities as well. SatCLIP was trained on spatially uniform distributed Sentinel-2

imagery, while GeoCLIP was trained on ground-level Flickr imagery that is primarily available in North America and Europe. GAIR (Liu et al. 2025) incorporated Local Implicit Image Function (Chen, Liu, and Wang 2021) to fuse embeddings from point-based ground level imagery (Hou et al. 2024) into large footprints of Sentinel-2 imagery. UniGeoCLIP (Astruc et al. 2026) performs multimodal training between location and 4 other modes, but does so on proprietary data. The multimodal contrastive methodology for GAIR and UniGeoCLIP can broadly be described by Figure 1.

There is also evidence to suggest that concatenation of representations from complementary models, such as SatCLIP and AlphaEarth, improves generalizable downstream performance when compared to using one model’s representations (van der Plas et al. 2026). However, the optimal manner of fusing representations together and choosing complementary models is not yet well understood.

SLED represents the first location encoder to learn location embeddings via distillation and the first to do so while using location as a binding modality, which enables a new way to fuse complementary EO embeddings.

Methodology

Framework

A schematic overview of our framework is given in Figure 3. When training with a single modality, we simply compute MSE loss between our location embeddings and modality-specific embeddings from a frozen teacher. When training with more than one modality, we take inspiration from ImageBind (Girdhar et al. 2023) and treat position as the single *binding modality*. We introduce K parallel linear layers on top of the position encoder to serve as alignment heads, where K is the number of modes. These alignment heads serve two main purposes: (1) They decouple the size of the position encoder’s latent space from the sizes of the teacher encoders’

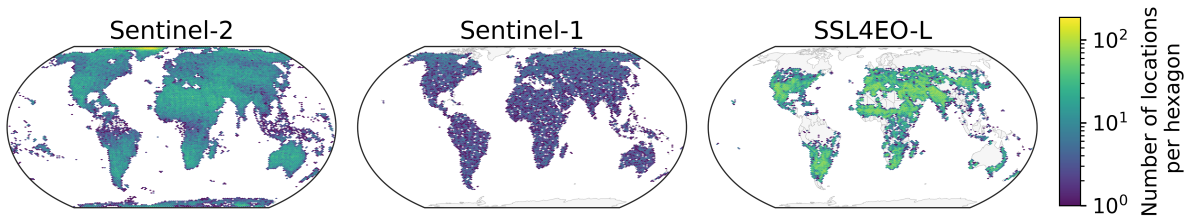


Figure 4: Spatial density of the pretraining datasets used in this work.

latent spaces, improving model flexibility, (2) They improve model convergence while avoiding embedding collapse.

With each additional data modality, an additional frozen modality-specific teacher and trainable alignment head are added to the model. Each sample has modality-specific data, latitude, and longitude. When training, all samples are fed through the online student location encoder and their corresponding modality-specific teacher encoder. In this manner, every sample fires the student encoder and its respective modality-specific teacher encoder.

For the teacher encoders, we use the ViT-L version of TerraFM (Danish et al. 2025) for Sentinel-2, the ViT-L version of FG-MAE (Wang et al. 2024) for Sentinel-1, and the ViT-B version of MoCo (He et al. 2020) pretrained on Landsat imagery for Landsat 8/9. FG-MAE and MoCo are obtained through the TorchGeo python package (Stewart et al. 2025). Additional details may be found in the supplemental materials.

Data

Figure 4 shows the spatial distribution of our training datasets. In each modality, a sample consists of an image and a corresponding latitude and longitude. SLED readily allows for incorporating multiple modalities. When training SLED for our experiments, we distill information from up to three distinct modalities of satellite imagery:

- Sentinel-2 level 2 A: 100K samples of 12-band, 256×256 resolution multi-spectral imagery from the S2-100K dataset originally generated by (Klemmer et al. 2025b).
- Landsat 8/9 OLI SR: 250K samples of 7-band, 224×224 multi-spectral imagery from the SSL4EO-L dataset (Stewart et al. 2023). The SSL4EO-L dataset contains 250K locations, each with 4 images corresponding to the seasons. In order to avoid temporal overlap, we randomly choose one image per location to use in our dataset.
- Sentinel-1: 20K samples of 224×224 resolution C-band Synthetic Aperture Radar (SAR) imagery with VH/VV polarization sampled in uniform at random (UAR) spatial distribution on landmasses. We created this dataset via the Microsoft Planetary STAC API and have made it available for use.

We selected these three freely-available satellite data modalities as widely used representative EO that capture complementary information across spectral, spatial, and sensing modalities. Optical multi-spectral Sentinel-2 and Landsat satellites provide globally consistent observations,

and Sentinel-1 provides all-weather microwave measurements, allowing for experimentation for various tasks. It is worth noting, however, that available benchmarks, including those that we use in this paper, are primarily based on tasks where optical imagery excels.

Pretraining Specifications

To isolate the impact of multimodal training, all models are pretrained on 100K samples in total, meaning in multimodal training, only subsets of each satellite dataset are used to keep the total training samples constant. We train with early stopping, randomly selecting 10% of our training dataset to use as a validation set, with a minimum validation loss improvement of 0.001 and a patience of 10 epochs. We limit pretraining to 150 epochs. When pretraining with multiple modes, the distribution of our validation set is proportional to the overall distribution of our training dataset (i.e., if we pre-train with 50k Sentinel-2 images and 50k Landsat-9 images, our validation set will contain 5k samples from Sentinel-2 and 5k samples from Landsat-9).

We use an equal number of samples per mode when able. This means that when training with Sentinel-2 and Landsat, each mode has 50K random samples apiece. When training with Sentinel-1, Sentinel-2, and Landsat, we employ a 20K/40K/40K split. We take MSELoss per mode, then backpropagate loss through the appropriate alignment head and location encoder. As such, the overall loss for the location encoder can be thought of as a weighted average of the MSELoss per modality.

We leverage the image augmentations used in SatCLIP’s pretraining: random flip, random crop, and Gaussian blur. For the latitude and longitude, we employ ~ 1 km of coordinate jitter. We pretrain with the RFF (Tancik et al. 2020) position encoder that was developed for GeoCLIP. Ablations with the SirenNet (Rußwurm et al. 2024) position encoder backbone can be found in our supplemental materials. We train on a NVIDIA RTX A6000 GPU using a batch size of 128, a learning rate of 0.001, and an AdamW optimizer.

Benchmarking

Unless explicitly stated, all benchmarks are done on a frozen location encoder (after pretraining as described above) with a trainable linear probe. We use early stopping while training the linear probe, requiring an improvement of 0.001 in validation loss over 10 epochs in order to continue training. Training is capped at 100 epochs. We pretrain 5 randomly seeded runs of each SLED model. For pre-existing SOTA

Table 1: SLED results compared to current state-of-the-art models on 4 distinct benchmark suites. SLED uses RFF for position encoding. Reporting average results $\pm$ standard deviation over 5 randomly seeded runs. Best results are in **bold**, second best are *italicized*. Aggregate results per benchmark suite are at the top of each sub-section

Benchmark	SatCLIP	GeoCLIP	SLED	SLED	SLED
Model Details					
Training Modalities	S2	Flickr	S2	S2+LS	S2+LS+S1
Trainable Parameters	1M	10M	14.1M	14.1M	14.1M
Number of samples	100K	4.7M	100K	100K	100K
Batch size	16K	512	128	128	128
Loss style	InfoNCE	InfoNCE	MSE	MSE	MSE
Hours per epoch	<i>3.344</i>	4.912	0.105	0.105	0.105
CHELSEA Regression (R^2)					
Aggregate Results	0.693 ± 0.017	0.738 ± 0.016	<i>0.810 $\pm$ 0.023</i>	0.818 $\pm$ 0.017	0.807 ± 0.021
Frost Change Frequency	0.651 ± 0.008	0.658 ± 0.023	<i>0.768 $\pm$ 0.017</i>	0.780 $\pm$ 0.020	0.747 ± 0.017
Grow Season Length	0.770 ± 0.021	0.779 ± 0.010	0.863 ± 0.018	<i>0.868 $\pm$ 0.003</i>	0.879 $\pm$ 0.011
Grow Season First	0.554 ± 0.009	0.555 ± 0.007	<i>0.680 $\pm$ 0.015</i>	0.690 $\pm$ 0.009	0.671 ± 0.020
Isothermality	0.757 ± 0.013	0.847 ± 0.006	0.941 ± 0.004	0.945 $\pm$ 0.005	<i>0.943 $\pm$ 0.005</i>
Precipitation Annual	0.570 ± 0.016	0.572 ± 0.019	0.694 $\pm$ 0.027	<i>0.693 $\pm$ 0.027</i>	0.666 ± 0.033
Precipitation Seasonality	0.621 ± 0.018	0.731 ± 0.012	0.728 ± 0.029	0.754 $\pm$ 0.014	<i>0.735 $\pm$ 0.014</i>
Snow Cover Days	0.805 ± 0.004	0.869 ± 0.010	0.961 $\pm$ 0.001	<i>0.959 $\pm$ 0.004</i>	0.958 ± 0.003
Temperature Seasonality	0.815 ± 0.005	0.867 $\pm$ 0.002	0.841 ± 0.013	<i>0.859 $\pm$ 0.006</i>	0.858 ± 0.006
TorchSpatial Classification (Acc@3)					
Aggregate Results	0.888 ± 0.001	0.890 $\pm$ 0.001	0.888 ± 0.001	0.890 $\pm$ 0.001	0.890 $\pm$ 0.001
iNat2018	0.904 ± 0.001	<i>0.907 $\pm$ 0.000</i>	0.901 ± 0.001	0.908 $\pm$ 0.001	0.906 ± 0.001
Flickr	0.842 ± 0.002	0.844 $\pm$ 0.002	0.841 ± 0.002	0.844 $\pm$ 0.001	0.843 ± 0.002
fMoW	0.917 ± 0.001	0.921 $\pm$ 0.000	0.921 $\pm$ 0.001	0.918 ± 0.001	0.920 ± 0.001
TorchSpatial Regression (R^2)					
Aggregate Results	0.409 ± 0.001	0.503 $\pm$ 0.001	0.478 ± 0.003	<i>0.502 $\pm$ 0.002</i>	0.492 ± 0.001
Forest Cover	0.675 $\pm$ 0.000	<i>0.674 $\pm$ 0.000</i>	0.634 ± 0.000	0.639 ± 0.000	0.628 ± 0.000
Elevation	0.295 ± 0.000	0.560 ± 0.000	0.562 ± 0.000	0.641 $\pm$ 0.000	<i>0.621 $\pm$ 0.000</i>
Nightlights	0.085 ± 0.002	0.116 $\pm$ 0.001	0.091 ± 0.003	<i>0.100 $\pm$ 0.003</i>	0.092 ± 0.001
Population Density	0.580 ± 0.001	0.661 $\pm$ 0.001	0.624 ± 0.006	<i>0.629 $\pm$ 0.004</i>	0.626 ± 0.002
SustainBench Regression (R^2)					
Aggregate Results	-0.154 ± 0.099	0.115 $\pm$ 0.003	0.014 ± 0.082	0.044 ± 0.046	<i>0.077 $\pm$ 0.040</i>
Asset Index	-0.041 ± 0.101	0.217 $\pm$ 0.001	0.082 ± 0.101	0.119 ± 0.042	<i>0.161 $\pm$ 0.022</i>
Sanitation Index	-0.193 ± 0.100	0.099 $\pm$ 0.002	0.015 ± 0.073	0.059 ± 0.032	<i>0.085 $\pm$ 0.054</i>
Water Index	-0.141 ± 0.100	0.036 $\pm$ 0.007	-0.047 ± 0.079	0.014 ± 0.047	<i>0.029 $\pm$ 0.043</i>
Women Education	-0.241 ± 0.135	0.107 $\pm$ 0.001	0.005 ± 0.108	-0.016 ± 0.076	<i>0.031 $\pm$ 0.051</i>

models, all evaluations are done with 5 randomly seeded runs of the same pretrained model. All benchmarking is done via simple linear probing in order to explore the latent space with minimal hyperparameter tuning.

When performing regression, we normalize regression targets to the range of $[0,1]$ using a linear transformation with the lowest and highest available labels. The only exception is for the TorchSpatial NightLights task, where we use a log transform on the labels instead, as inspired by (Klemmer et al. 2025b). These transformations change the scale of the regression targets, but not the distribution.

We evaluate on eight uniformly-at-random (UAR) sampled climate-related variables from the CHELSA-TraCE21k-centennial dataset (Karger et al. 2023), four regression on social indices from SustainBench (Yeh et al. 2021), four regression tasks from TorchSpatial (Wu et al. 2024), and three classification tasks with fused image embeddings from TorchSpatial (Wu et al. 2024). Additional details can be found in the supplemental materials.

We include GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023) and SatCLIP (Klemmer et al. 2025b) as established benchmarks for location encoders. While GAIR (Liu et al.

2025) and UniGeoCLIP (Astruc et al. 2026) would be intriguing baselines to test against, their model weights have not yet been made publicly available.

Downstream Performance

On CHELSA climate regression tasks, all SLED models (regardless of training modality) outperform other location encoders on 7 out of 8 tasks, and ranks second in temperature seasonality (Table 1). In TorchSpatial classification tasks, SLED is on par with other location encoders. Across TorchSpatial’s regression tasks, we see more variance. SLED lags behind existing location encoders in forest cover, but exhibits a sizable boost on elevation mapping, particularly with multimodal versions of SLED.

On the remaining SustainBench regression benchmarks, SLED outperforms SatCLIP by a significant margin, but falls short of GeoCLIP. We believe that this is largely due to these tasks being more observable from ground-level images compared to satellite imagery. Whereas SatCLIP and SLED rely on satellite imagery, GeoCLIP is trained on ground-level images. However, SLED’s performance gain over SatCLIP is noteworthy and further demonstrates the potential of distillation over contrastive learning for location encoding.

We also see the relative training expense in Table 1. SatCLIP’s large batch size and GeoCLIP’s large dataset size both act as training bottlenecks in a way that SLED is unhindered by, further contextualizing SLED’s performance.

Modality Ablations

While pretraining with purely Sentinel-2 imagery results in strong performance for SLED as seen in Table 1, pretraining on a mixture of Sentinel-2 and Landsat imagery further improves metrics, particularly on TorchSpatial regression and SustainBench tasks.

However, in several cases, performance decreases with the inclusion of Sentinel-1 data (when comparing S2+LS vs. S2+LS+S1 variants), perhaps due to fewer channels in Sentinel-1 imagery, which over land has 2 "bands", i.e., polarization VV and VH, compared to 7 spectral bands for Landsat and 12 bands for Sentinel-2. It is also possible that this decrease is a result of comparatively better multi-spectral (MS) imagery encoders compared to SAR encoders. Optimal strategies for encoding MS imagery have been explored substantially more compared to SAR data (Lane and Karimzadeh 2026). Despite this performance dip in some cases, the inclusion of Sentinel-1 provides a boost in model performance on the SustainBench task, while still producing comparable results on CHELSA and TorchSpatial’s classification task.

Qualitative Embedding Analysis

For all modality ablations of SLED, SLED learns meaningful geospatial representations (Figure 5). Embeddings of proximate areas and similar (but distant) biomes are more similar, while the variations within specific regions indicate that SLED is also capable of learning region-specific spatial heterogeneity. The map is also striking in encoding higher elevations, as seen in southwestern China (the Qinghai-Tibet Plateau) and western South America (the Andes).

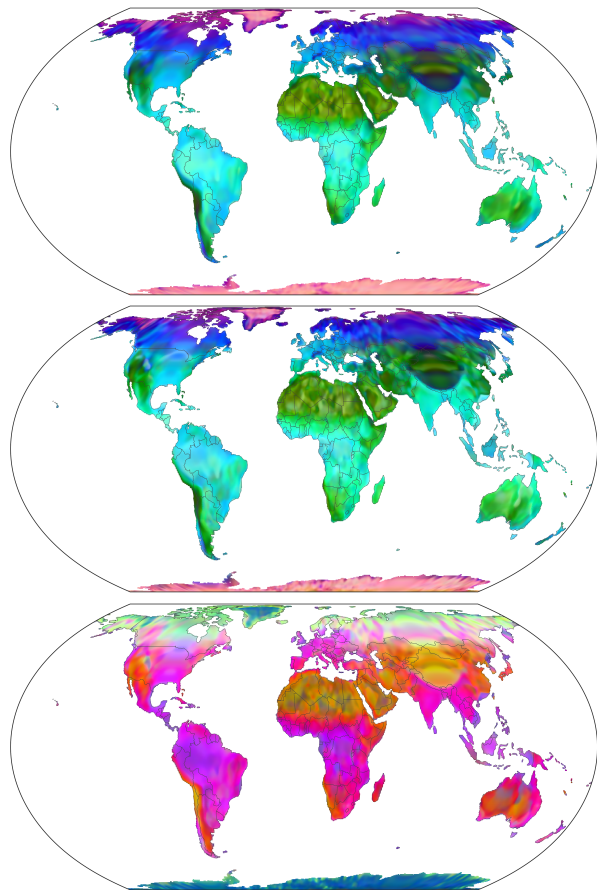


Figure 5: Equal earth projection of first three principal components of SLED embeddings. From top to bottom, models are SLED trained on S2, SLED trained on S2+LS and SLED trained on S2+LS+S1. Given the underlying PCA calculations, colors should not be compared from map to map.

We also visualize how embeddings change with the addition of new modalities. We use Centered Kernel Alignment (CKA), a latent space similarity metric that is agnostic to model initialization (Cortes, Mohri, and Rostamizadeh 2012). Because CKA needs to be computed over sets of embeddings rather than individual samples, we generate embeddings on a 0.5° resolution grid and apply a 3×3 sliding window over the resulting embedding maps to define local neighborhoods for CKA analysis. Due to land boundaries and ocean cells, not every grid location contains valid embeddings across all models being compared. To ensure reliable estimates, we only compute CKA scores for center cells with at least three valid spatial locations containing embeddings from both models within the corresponding window.

Within our CKA visualizations, it is notable to see how SLED embeddings change from training with Sentinel-2 (which has UAR distribution) to training with Sentinel-2 and Landsat 8/9 (which is from the SSL4EO-L dataset clustered around population centers) in Figure 6. We see embeddings in more densely populated areas (particularly in the UK, northeast USA, northern South America, southern

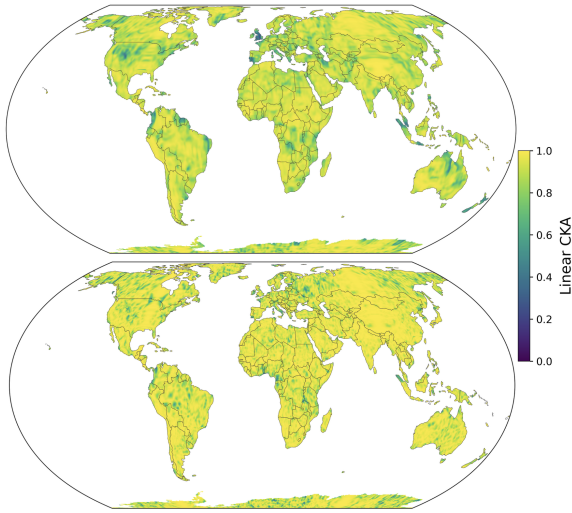


Figure 6: Spatially mapped Centered Kernel Analysis (CKA) of SLED embeddings trained on different modalities. Top: CKA between SLED trained on S2 compared with SLED trained on S2/LS. Bottom: CKA between SLED trained on S2+LS compared with SLED trained on S2+S1+LS.

Africa, and eastern Australia) change the most dramatically. This demonstrates SLED’s ability to leverage modalities with spatially disparate distributions.

Figure 6 also demonstrates how adding a notably different training modality (in this case, Sentinel-1 SAR compared with optical imagery from Sentinel-2 and Landsat 8/9) picks up key features that may be missed by other modalities. We see that embeddings trained with Sentinel-1/Sentinel-2/Landsat differ from optical-only embeddings primarily over water, including the Great Lakes in North America, Lake Victoria in Africa, and bodies of water in eastern Europe. We include additional CKA and PCA figures in our supplemental materials.

Discussion

Our results in Table 1 demonstrate SLED’s ability to capture signals meaningful to a variety of downstream tasks. We also see in Figure 5 that SLED is capable of producing geospatially meaningful embeddings, which are the first of its kind to come from a location encoder model trained via distillation. More importantly, we believe that SLED will accelerate research in the field, particularly by facilitating the addition of more modes, as each mode only requires a frozen teacher and a simple linear alignment layer. By leveraging geospatial location as a binding modality, SLED only requires the modality-specific teacher and the location in memory at any given time, regardless of the number of modalities used in training. While the number of frozen parameters will rise with the addition of each mode, the number of trainable parameters needed in memory will remain relatively static. Mixture of Experts (MoE)-style models serve as a useful precedent here, as they separate performance from computation requirements by limiting the number of trainable param-

eters necessary to have in memory at any given time (Gamal, Afifa, and Gamal 2025). As such, we think SLED could further facilitate multimodal training for location encoders.

The teacher-based formulation of SLED enables flexible choices in how modalities are incorporated. We currently train with a teacher per modality, but the rise of multimodal remote sensing encoders means that a single multimodal teacher, such as TerraFM (Danish et al. 2025), DOFA (Xiong et al. 2024), OmniSat (Astruc et al. 2025), could serve as the teacher for a number of remotely sensed modes, cutting down the number of frozen parameters required in the model. Conversely, one could also leverage multiple different teachers for a single mode. Our modality ablations investigate the common case where each additional mode is introduced through a dedicated teacher, but the framework naturally supports decoupling teacher and modality scaling. Recent work shows that adding an additional teacher for distillation to existing pretraining frameworks for satellite *image* encoders improves performance (Mendieta et al. 2023), suggesting that additional teachers serve as regularizers in pretraining. Therefore, scaling teacher diversity alongside modality diversity is a promising direction for multimodal location encoding.

Our results suggest that the addition of modalities generally improves model performance, with the largest gains from information-rich multi-spectral imagery. Across several tasks, multimodal variants of SLED outperform single-mode models. The modality ablations in Table 1 and latent space analysis in Figure 6 further demonstrate that SLED can incorporate additional modalities while maintaining the expressive structure learned from the original Sentinel-2 modality. Although the benefits of adding Sentinel-1 SAR are less consistent than we expected, this may be caused by fewer channels in Sentinel-1 SAR imagery when compared to Sentinel-2 and Landsat MS imagery. The value of incorporating SAR may also become more apparent as SAR-specific benchmark tasks become available.

It is also worth addressing the metrics observed on the SustainBench suite. GeoCLIP performs best on these tasks, which we hypothesize is due to the benchmark containing tasks with stronger relationships to observable landscape characteristics. However, all models perform relatively poorly on this benchmark, where targets are primarily derived from survey-based measurements.

Conclusion

In this paper, we presented SLED, a distillation-based framework for location encoding. Our unimodal experiments demonstrate that distillation is a viable strategy for training location encoders more quickly with smaller batch sizes. We further showed how this framework extends to multimodal learning by using geospatial location as a binding modality, eliminating the need for costly spatiotemporal alignment (i.e., coregistration). Both unimodal and multimodal SLED achieve competitive performance with SOTA location encoders while providing a scalable, efficient approach for multimodal geospatial representation learning.

Acknowledgments

This research was supported by the CU Boulder Research & Innovation Office.

Many thanks to those in the GeoHAI Lab and Rolf Lab for their kindness and support.

References

- Astruc, G.; Gonthier, N.; Mallet, C.; and Landrieu, L. 2025. OmniSat: Self-supervised Modality Fusion for Earth Observation. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 409–427. Cham: Springer Nature Switzerland. ISBN 978-3-031-73390-1.
- Astruc, G.; Trulls, E.; Hosang, J.; Landrieu, L.; and Sarlin, P.-E. 2026. UniGeoCLIP: Unified Geospatial Contrastive Learning. *ArXiv:2604.11668 [cs]*.
- Auh, J.; Cho, C.; and Kim, S.-T. 2026. Mitigating False Negatives in Contrastive Learning via Progressive Hyperparameter Adjustment. *IEEE Access*, 14: 10276–10285.
- Bonafilia, D.; Tellman, B.; Anderson, T.; and Issenberg, E. 2020. Sen1Floods11: a georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 835–845. ISSN: 2160-7516.
- Brown, C. F.; Kazmierski, M. R.; Pasquarella, V. J.; Rucklidge, W. J.; Samsikova, M.; Zhang, C.; Shelhamer, E.; Lahera, E.; Wiles, O.; Ilyushchenko, S.; et al. 2025. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*.
- Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8628–8638.
- Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional Map of the World. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6172–6180. IEEE.
- Cortes, C.; Mohri, M.; and Rostamizadeh, A. 2012. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13: 795–828.
- Danish, M. S.; Munir, M. A.; Shah, S. R. A.; Khan, M. H.; Anwer, R. M.; Laaksonen, J.; Khan, F. S.; and Khan, S. 2025. TerraFM: A scalable foundation model for unified multisensor earth observation. *arXiv preprint arXiv:2506.06281*.
- Dinerstein, E.; Olson, D.; Joshi, A.; Vynne, C.; Burgess, N. D.; Wikramanayake, E.; Hahn, N.; Palminteri, S.; Hedao, P.; Noss, R.; et al. 2017. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6): 534–545.
- Dollinger, J.; Robert, D.; Plekhanova, E.; Drees, L.; and Wegner, J. D. 2025. CLIMPLICIT: Climatic implicit embeddings for global ecological tasks. *arXiv preprint arXiv:2504.05089*.
- Fuller, A.; Millard, K.; and Green, J. 2023. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36: 5506–5538.
- Gamal, F.; Afifa, T.; and Gamal, F. 2025. MoE at Scale: From Modular Design to Deployment in Large-Scale Machine Learning Systems. *Preprints:2025041313*.
- Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15180–15190.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.
- Herzog, H.; Bastani, F.; Zhang, Y.; Tseng, G.; Redmon, J.; Sablon, H.; Park, R.; Morrison, J.; Buraczynski, A.; Farley, K.; et al. 2025. OlmoEarth: Stable Latent Image Modeling for Multimodal Earth Observation. *arXiv preprint arXiv:2511.13655*.
- Hou, Y.; Quintana, M.; Khomiakov, M.; Yap, W.; Ouyang, J.; Ito, K.; Wang, Z.; Zhao, T.; and Biljecki, F. 2024. Global Streetscapes — A comprehensive dataset of 10 million street-level images across 688 cities for urban science and analytics. *ISPRS Journal of Photogrammetry and Remote Sensing*, 215: 216–238.
- Huynh, T.; Kornblith, S.; Walter, M. R.; Maire, M.; and Khademi, M. 2022. Boosting contrastive self-supervised learning with false negative cancellation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2785–2795.
- Jakubik, J.; Yang, F.; Blumenstiel, B.; Scheurer, E.; Sedona, R.; Maurogiovanni, S.; Bosmans, J.; Dionelis, N.; Marsocci, V.; Kopp, N.; et al. 2025. Terramind: Large-scale generative multimodality for earth observation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7383–7394.
- Karger, D. N.; Nobis, M. P.; Normand, S.; Graham, C. H.; and Zimmermann, N. E. 2023. CHELSA-TraCE21k—high-resolution (1 km) downscaled transient temperature and precipitation data since the Last Glacial Maximum. *Climate of the Past*, 19(2): 439–456.
- Karimzadeh, M.; Wang, Z.; and Crooks, J. L. 2025. Performance and generalizability impacts of incorporating location encoders into deep learning for dynamic $PM_{2.5}$ estimation. *GIScience & Remote Sensing*, 62(1): 2594797.
- Klemmer, K.; Rolf, E.; Camps-Valls, G.; Czerkawski, M.; Ermon, S.; Francis, A.; Jacobs, N.; Kerner, H.; Mackey, L.; Mai, G.; et al. 2025a. Earth Embeddings: Towards AI-centric Representations of our Planet.
- Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; and Rußwurm, M. 2025b. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4): 4347–4355.
- Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; and Rubel, F. 2006. World map of the Köppen-Geiger climate classification updated.

- Kusupati, A.; Bhatt, G.; Rege, A.; Wallingford, M.; Sinha, A.; Ramanujan, V.; Howard-Snyder, W.; Chen, K.; Kakade, S.; Jain, P.; and Farhadi, A. 2024. Matryoshka Representation Learning. *ArXiv:2205.13147* [cs].
- Lacoste, A.; Lehmann, N.; Rodriguez, P.; Sherwin, E.; Kerner, H.; Lütjens, B.; Irvin, J.; Dao, D.; Alemohammad, H.; Drouin, A.; et al. 2023. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093.
- Lane, K.; and Karimzadeh, M. 2026. A Genealogy of Foundation Models in Remote Sensing. *ACM Transactions on Spatial Algorithms and Systems*.
- Liu, Z.; Zhang, F.; Jiao, J.; Lao, N.; and Mai, G. 2025. GAIR: Improving Multimodal Geo-Foundation Model with Geo-Aligned Implicit Representations. *ArXiv:2503.16683* [cs].
- Lu, S.; Guo, J.; Zimmer-Dauphinee, J. R.; Nieusma, J. M.; Wang, X.; VanValkenburgh, P.; Wernke, S. A.; and Huo, Y. 2025. Vision Foundation Models in Remote Sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 13(3): 190–215.
- Mai, G.; Janowicz, K.; Hu, Y.; Gao, S.; Yan, B.; Zhu, R.; Cai, L.; and Lao, N. 2022. A review of location encoding for GeoAI: methods and applications. *International Journal of Geographical Information Science*, 36(4): 639–673.
- Mai, G.; Lao, N.; He, Y.; Song, J.; and Ermon, S. 2023. CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. In *Proceedings of the 40th International Conference on Machine Learning*, 23498–23515. PMLR.
- Mai, G.; Xie, Y.; Jia, X.; Lao, N.; Rao, J.; Zhu, Q.; Liu, Z.; Chiang, Y.-Y.; and Jiao, J. 2025. Towards the next generation of Geospatial Artificial Intelligence. *International Journal of Applied Earth Observation and Geoinformation*, 136: 104368.
- Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; and Chen, C. 2023. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16806–16816.
- Neuhof, G.; Ollmann, T.; Rota Bulò, S.; and Kotschieder, P. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, 4990–4999.
- of Washington University, M. V. R. L. 2024. RSHF: Remote sensing pretrained models easy loading using huggingface. <https://github.com/mvrl/rshf>. Accessed: 2025-11-25.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Rolf, E.; Proctor, J.; Carleton, T.; Bolliger, I.; Shankar, V.; Ishihara, M.; Recht, B.; and Hsiang, S. 2021. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications*, 12(1): 4392.
- Roy, O.; and Vetterli, M. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, 606–610.
- Rußwurm, M.; Klemmer, K.; Rolf, E.; Zbinden, R.; and Tuia, D. 2024. Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks. In *12th International Conference on Learning Representations (ICLR 2024)*.
- Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.
- Spoto, F.; Sy, O.; Laberinti, P.; Martimort, P.; Fernandez, V.; Colin, O.; Hoersch, B.; and Meygret, A. 2012. Overview Of Sentinel-2. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, 1707–1710.
- Stewart, A.; Lehmann, N.; Corley, I.; Wang, Y.; Chang, Y.-C.; Ait Ali Braham, N. A.; Sehgal, S.; Robinson, C.; and Banerjee, A. 2023. SSL4EO-L: Datasets and foundation models for Landat imagery. *Advances in Neural Information Processing Systems*, 36: 59787–59807.
- Stewart, A. J.; Robinson, C.; Corley, I. A.; Ortiz, A.; Lavista Ferres, J. M.; and Banerjee, A. 2025. TorchGeo: Deep Learning With Geospatial Data. *ACM Transactions on Spatial Algorithms and Systems*, 11(4): 1–28.
- Sudmanns, M.; Tiede, D.; Augustin, H.; and Lang, S. 2020. Assessing global Sentinel-2 coverage dynamics and data availability for operational Earth observation (EO) applications using the EO-Compass. *International Journal of Digital Earth*, 13(7): 768–784. <https://doi.org/10.1080/17538947.2019.1572799>.
- Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; and Ng, R. 2020. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33: 7537–7547.
- Tang, K.; Paluri, M.; Fei-Fei, L.; Fergus, R.; and Bourdev, L. 2015. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, 1008–1016.
- van der Plas, T. L.; Bakermans, J. J.; Nedungadi, V.; Tijūnaitytė, G.; Rußwurm, M.; and Athanasiadis, I. N. 2026. Better Together: Evaluating the Complementarity of Earth Embedding Models. *arXiv preprint arXiv:2605.18667*.
- Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.
- Vivanco Cepeda, V.; Nayak, G. K.; and Shah, M. 2023. Geo-CLIP: CLIP-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *Advances in Neural Information Processing Systems*, 36: 8690–8701.
- Wang, Y.; Albrecht, C. M.; Braham, N. A. A. A.; Liu, C.; Xiong, Z.; and Zhu, X. X. 2023. DeCUR: decoupling common & unique representations for multimodal self-supervision.
- Wang, Y.; Hernández, H. H.; Albrecht, C. M.; and Zhu, X. X. 2024. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of*

Selected Topics in Applied Earth Observations and Remote Sensing, 18: 321–336.

Waters, N. 2017. Tobler’s first law of geography. *The international encyclopedia of geography*, 1–13.

Wu, N.; Cao, Q.; Wang, Z.; Liu, Z.; Qi, Y.; Zhang, J.; Ni, J.; Yao, X.; Ma, H.; Mu, L.; et al. 2024. TorchSpatial: A location encoding framework and benchmark for spatial representation learning. *Advances in Neural Information Processing Systems*, 37: 81437–81460.

Xiong, Z.; Wang, Y.; Zhang, F.; Stewart, A. J.; Hanna, J.; Borth, D.; Papoutsis, I.; Saux, B. L.; Camps-Valls, G.; and Zhu, X. X. 2024. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*.

Yeh, C.; Meng, C.; Wang, S.; Driscoll, A.; Rozi, E.; Liu, P.; Lee, J.; Burke, M.; Lobell, D. B.; and Ermon, S. 2021. SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Supplemental Materials

Technical Details

Pretraining When training SLED on Sentinel-2 imagery, we use all 100K samples from the S2-100K dataset and use a frozen TerraFM (Danish et al. 2025) encoder to encode the images. We feed the latitude and longitude through a trainable position encoder, then compute a simple MSELoss between the position embeddings and the image embeddings. The resulting loss is backpropagated through the position encoder. We can think of the frozen TerraFM encoder as the teacher and the position encoder as the student in this distillation setup. Since TerraFM embeds 768 dimensions, our position encoder also encodes in 768 dimensions. We utilize TerraFM since it trains on Sentinel-2’s ground-level L2A imagery (Spoto et al. 2012), which is the data product available in the S2-100K dataset (Klemmer et al. 2025b). Most other Sentinel-2 encoders train on the raw Sentinel-2 L1C data (Lane and Karimzadeh 2026).

When training on Landsat imagery, we use a frozen ViT-Base MoCo (He et al. 2020) encoder, made available through TorchGeo (Stewart et al. 2025), trained on Landsat OLI SR imagery as the teacher model. We randomly sample the SSL4EO-L dataset to obtain our 100K samples. Relatively few GeoFMs are trained only on Landsat data (Lane and Karimzadeh 2026). We leverage this MoCo model as a way to have a unimodal teacher for Landsat, as opposed to a teacher that will generate multimodal embeddings, such as TerraMind (Jakubik et al. 2025) or OmniSat (Astruc et al. 2025).

When training on Sentinel-1 imagery, we use the 20k samples from our custom Sentinel-1 dataset, which is partially available in our code supplement. We use the FG-MAE (Wang et al. 2024) for encoding SAR imagery, as we found the feature-guided reconstruction targets a particularly compelling SSL task for generating robust representations of Sentinel-1 imagery.

When training multimodal SLED, we use PyTorch Lightning’s CombinedDataLoader class to load in several modality-specific dataloaders and employ the max size combination strategy. This means that every batch will contain K modes until we have exhausted all the training samples for one mode. Then every batch will contain K-1 modes, and so on. Our training code can be found in the code supplement; the ReadMe.md file should contain appropriate details regarding the repository.

Benchmark Suites We evaluate on 8 distinct climate-related variables as regression targets. This benchmark consists of 10K locations, sampled in UAR across all landmasses on Earth. The regression targets are annual temperature, annual precipitation, temperature seasonality, precipitation seasonality, isothermality, snow cover days, frost change frequency, first day of growing season, and growing season length. All variables in question are taken from the CHELSA-TraCE21k-centennial dataset (Karger et al. 2023), which contains 1km-resolution maps of all variables. All tasks use a random train/val/test split of 60%/20%/20%. This dataset can be found in our code supplement.

We also benchmark against 4 regression tasks (forest cover, elevation, nightlights, and population density) from TorchSpatial (Wu et al. 2024). Lastly, we include an additional 3 classification tasks, also with fused image embeddings, from TorchSpatial (Wu et al. 2024): iNat2018 (Van Horn et al. 2018), YFCC (Tang et al. 2015), and Functional Map of the World (Christie et al. 2018). Specifically, we concatenate the pretrained Inception-V3 image features from the benchmark with location embeddings, then pass the combined vector through a final classification layer. Lastly, we include 4 regression tasks from SustainBench (Yeh et al. 2021): asset index, water index, sanitation index, and women education index. While SustainBench makes ground-level Mapillary (Neuhof et al. 2017) imagery, composite Landsat imagery, and nightlights imagery available for each sample, we simply regress against the latitude and longitude to benchmark models. This accounts for the difference seen between the results reported in (Yeh et al. 2021) and our paper.

For all benchmarks, we use a simple linear layer as our classification head. This differs from other works (Klemmer et al. 2025b; Wu et al. 2024), which have used a trainable MLP as their projection head, and accounts for the discrepancy between our results and theirs. We use a linear layer as a means to avoid costly hyperparameter tuning and ensure that benchmarking is a function of the latent space rather than a specific classification head architecture.

Ablations

Position Encoders We trained additional models with SirenNet (Sitzmann et al. 2020), Spherical Harmonics (Rußwurm et al. 2024) and UniGeoCLIP’s multi-scale RFF encoder (Astruc et al. 2026) as a position encoder, but the results lagged behind the RFF encoder used in GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023), often dramatically. Given the lack of performance, we ceased using compute resources to evaluate non-RFF models. However, we do have results available for 5 randomly seeded runs across

Table 2: Ablations between SLED location encoder using Spherical Harmonics, Random Fourier Features, SirenNet, and UniGeoCLIP’s multi-scale RFF encoder. Examining models trained with Sentinel-1 (S1), Sentinel-2 (S2), and Landsat 8/9 (LS). Results contain mean plus/minus standard deviation across 5 randomly seeded runs. Best results are in **bold** and second-best are in *italics*.

Modalities	CHELSA (R^2)	SustainBench (R^2)	eRank (dim)
Random Fourier Features			
S2	<i>0.81 ± 0.02</i>	0.01 ± 0.08	35.4
S2+LS	0.82 ± 0.02	<i>0.04 ± 0.05</i>	48.9
S2+LS+S1	0.81 ± 0.02	0.08 ± 0.04	45.7
Spherical Harmonics			
S2	−0.53 ± 0.42	−2.03 ± 0.96	22.4
S2+LS	−0.59 ± 0.28	−1.66 ± 0.52	42.1
S2+LS+S1	−0.75 ± 0.37	−1.94 ± 0.79	40.4
SirenNet			
S2+LS+S1	−0.15 ± 0.23	−1.79 ± 0.48	<i>48.9</i>
UniGeoCLIP			
S2+LS+S1	0.73 ± 0.02	−1.07 ± 0.55	57.8

all CHELSA and SustainBench tasks. When training SLED with Spherical Harmonics, we use 10 Legendre Polynomials. All variations of SLED encode in 768 dimensions. For RFF, SirenNet, and Spherical Harmonics, we use the default constructors available through the RSHF python package (of Washington University 2024). The details for our encoder variations for SLED are available here in Supplement table 2. All models training on more than one mode use the alignment head method detailed in the main paper. All benchmarking is done via linear probing.

We find that the RFF encoder used in GeoCLIP (Vivanco Cepeda, Nayak, and Shah 2023) is by far the best performing position encoder option. The UniGeoCLIP multi-scale RFF encoder comes reasonably close on the CHELSA suite, but falls behind in the SustainBench suite. We also include the effective rank (eRank) (Roy and Vetterli 2007) of each model as a way to quickly capture how many dimensions in our latent space have meaningful variation. In other words, eRank provides a useful tool for demonstrating how much of our latent space is effectively being used with each approach. We see how the addition of Landsat as a modality results in a meaningful improvement in eRank for both RFF and Spherical Harmonics versions of SLED. We also see eRank dip slightly with the inclusion of Sentinel-1, which we theorize is due to the relatively fewer number of bands in SAR imagery when compared to multispectral imagery from Sentinel-2 or Landsat.

Interestingly, we see eRanks that are comparable to (or better than) RFF versions of SLED for SLED using SphericalHarmonics, SirenNet, and UniGeoCLIP. This suggests that, despite their currently poor downstream performance,

Table 3: Examining cross-modal alignment options performance. All models train with Sentinel-1, Sentinel-2, and Landsat 8/9, with a 20/40/40 training dataset split. Options are Projection Head (Project.), Matryoshka embedding (Matryo.), or Alignment Heads (Align.). Best results are in **bold**, second best are *italicized*.

Method	Biome (acc)	EcoRegion (acc)	Temp (R^2)
Random Fourier Features			
Project.	0.72 ± 0.00	0.74 ± 0.00	−0.01 ± 0.00
Matryo.	<i>0.76 ± 0.01</i>	<i>0.74 ± 0.01</i>	<i>0.79 ± 0.78</i>
Align.	0.77 ± 0.00	0.76 ± 0.00	0.88 ± 0.01
Spherical Harmonics			
Project.	0.71 ± 0.01	0.72 ± 0.01	−0.02 ± 0.01
Matryo.	0.75 ± 0.02	0.63 ± 0.04	−0.07 ± 0.22
Align.	0.74 ± 0.02	0.56 ± 0.06	−0.46 ± 0.32

there is untapped potential in leveraging additional position encoding strategies for distillation-style location encoders.

Cross-Modal Alignment We explored a variety of options for how to calculate loss across multimodal teachers with different latent space sizes. Given that GeoFMs have any number of different embedding dimensions, ensuring that SLED could easily incorporate new, potentially disparate GeoFMs, was important.

We examined 3 options for aligning latent spaces before calculating loss:

1. Projection head: adding a trainable projection head atop each modality specific teacher. This projection head would be a linear layer that projected the teacher to the same number of dimensions as SLED’s latent space.
2. Matryoshka: a form of Matryoshka embeddings (Kusupati et al. 2024), where we perform MSE loss on the first N available dimensions, where N is the minimum of the number of SLED’s dimensions in the latent space and number of the teacher’s dimensions in the latent space.
3. Alignment heads: the alignment head method described in the main paper. This differs from projection heads in that the alignment heads sit atop the location encoder, *not* the modality specific teachers. This means that the loss for our alignment heads is propagated within the context of the rest of our trainable parameters in the location encoder.

We examine these methods on three benchmarks: Koppen biome classification (Kottek et al. 2006), EcoRegion classification (Dinerstein et al. 2017), and temperature regression from CHELSA. Results are averaged over 3 runs and can be found in Supplement Table 3. We see that projection heads and Matryoshka loss both lag behind our preferred approach of using alignment heads.

Additional Figures

We provide additional map visualizations in the hopes that these can provide qualitative understanding of SLED. In Figure 7, we see embedding cosine similarity compared to a single point of interest. In both instances, we see the surrounding region and other similar biomes around the world highlighted as the most similar.

The following pages include CKA and PCA visualizations already seen in the main paper, but blown up in size for convenient reference.

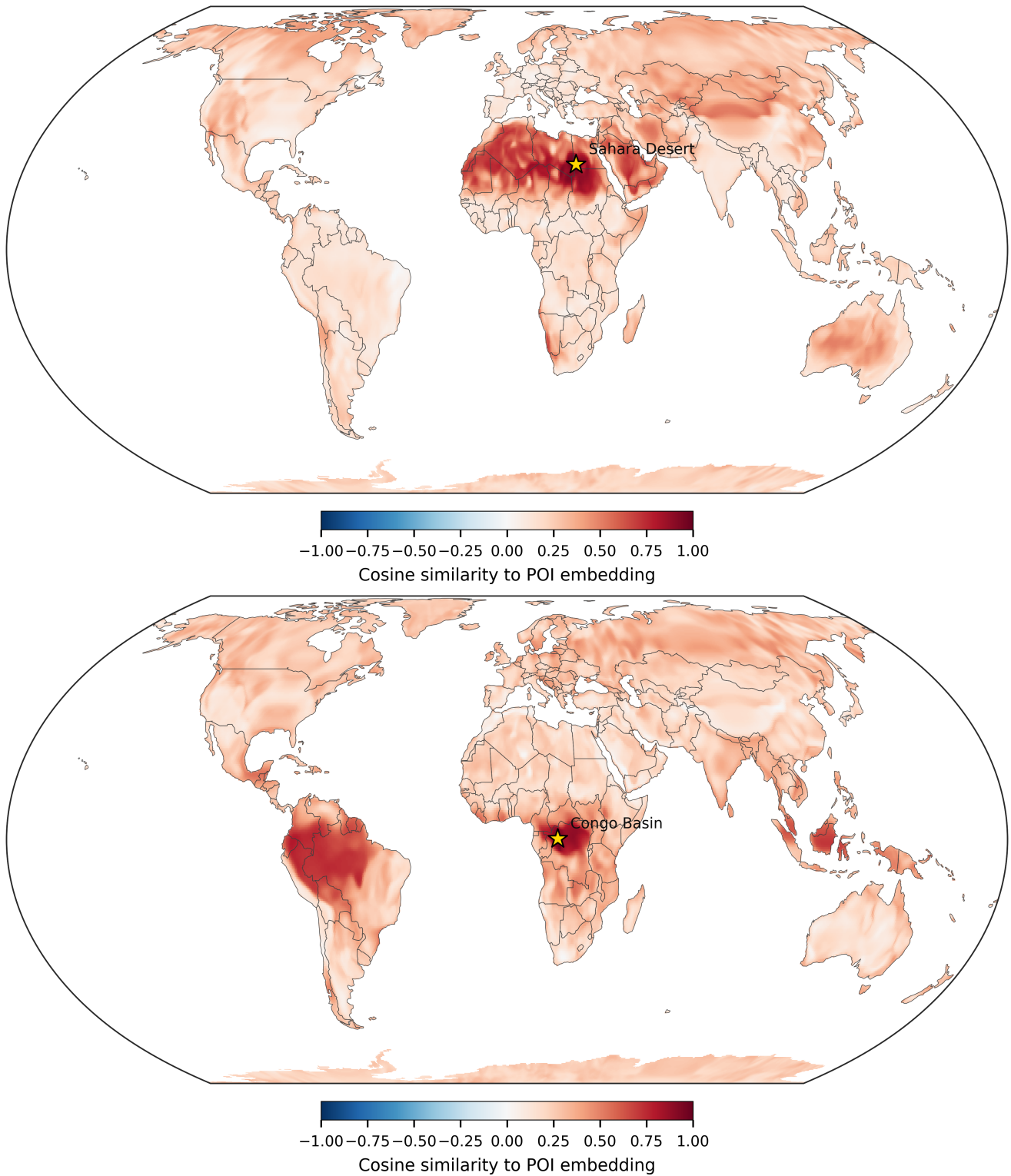


Figure 7: Spatial patterns of cosine similarity between global location embeddings and selected point-of-interest (POI) embeddings. The top figure shows a POI located in the Sahara Desert, while the bottom shows a POI located in the Congo Basin. The resulting similarity maps show that SLED captures geographically and environmentally meaningful relationships, with high similarity concentrated around regions sharing similar spatial or environmental characteristics with each reference location.

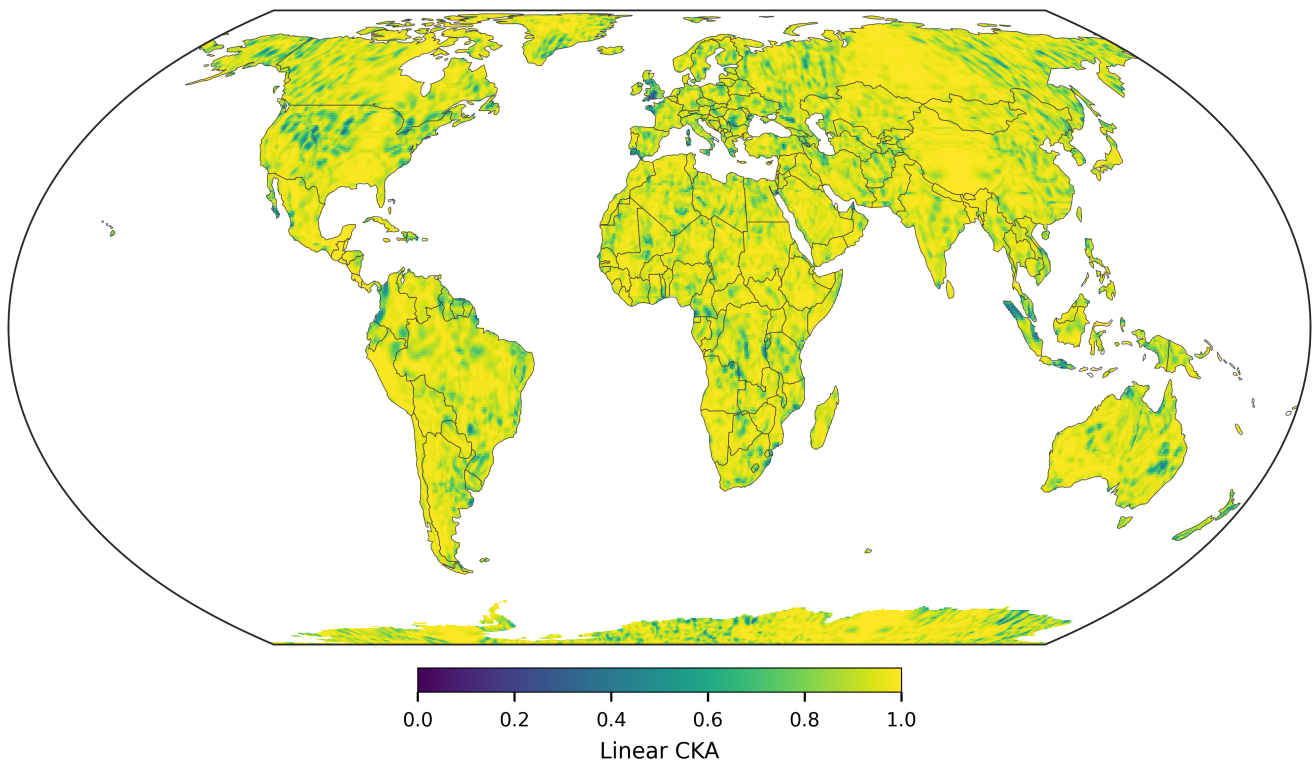


Figure 8: Blown-up version of CKA analysis of Sentinel-2 SLED compared with Sentinel-2/Landsat SLED.

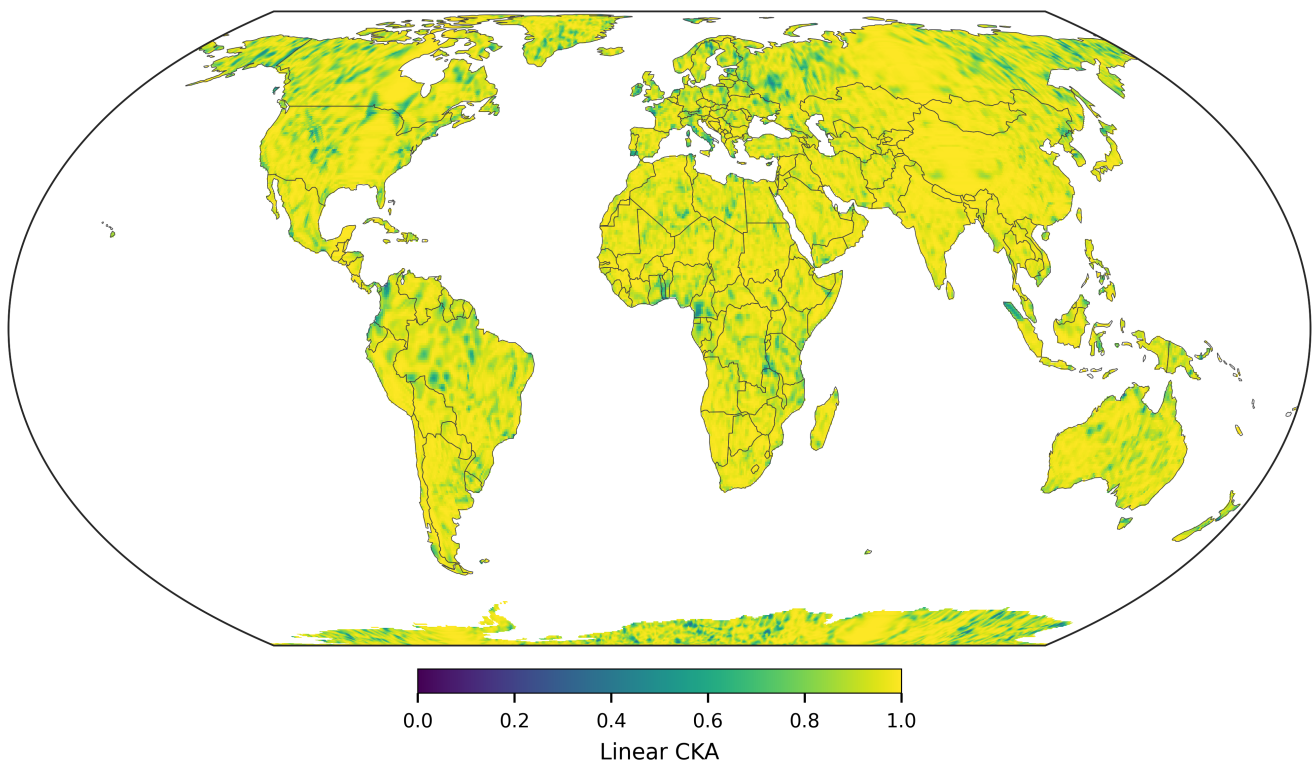


Figure 9: Blown up CKA analysis of Sentinel-2/Landsat SLED compared with Sentinel-1/Sentinel-2/Landsat SLED.

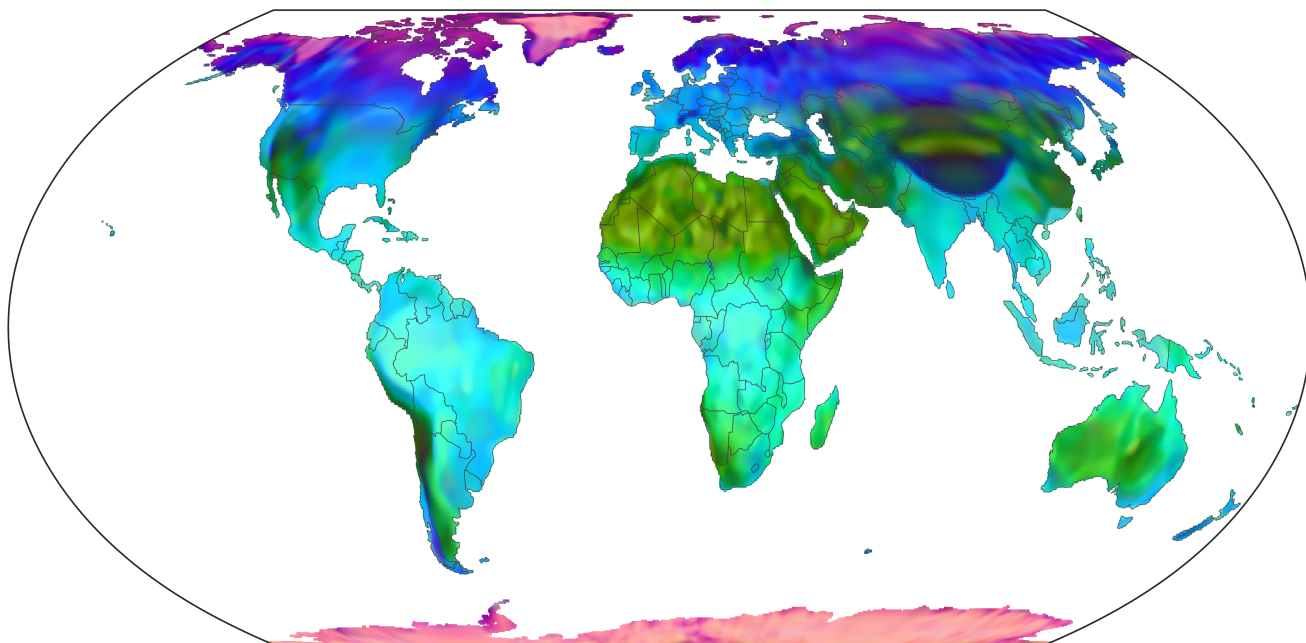


Figure 10: Blown-up version of first three principal components visualization of SLED embeddings trained on Sentinel-2.

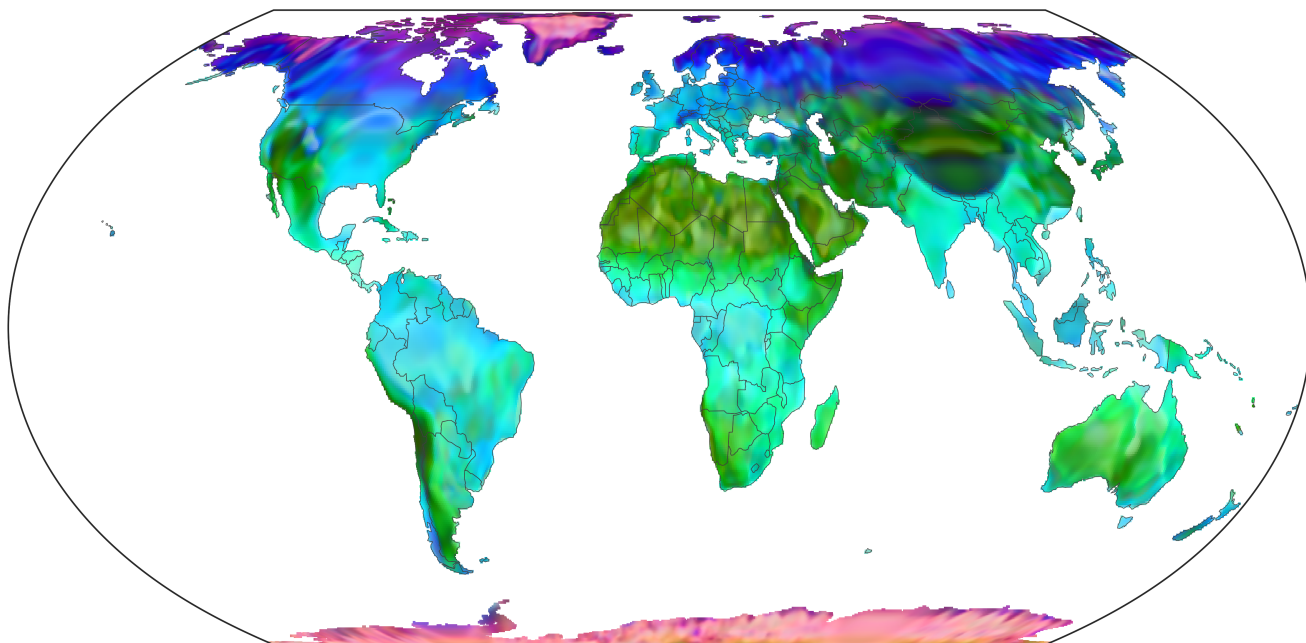


Figure 11: Blown-up version of first three principal components visualization of SLED embeddings trained on Sentinel-2 and Landsat 8/9.

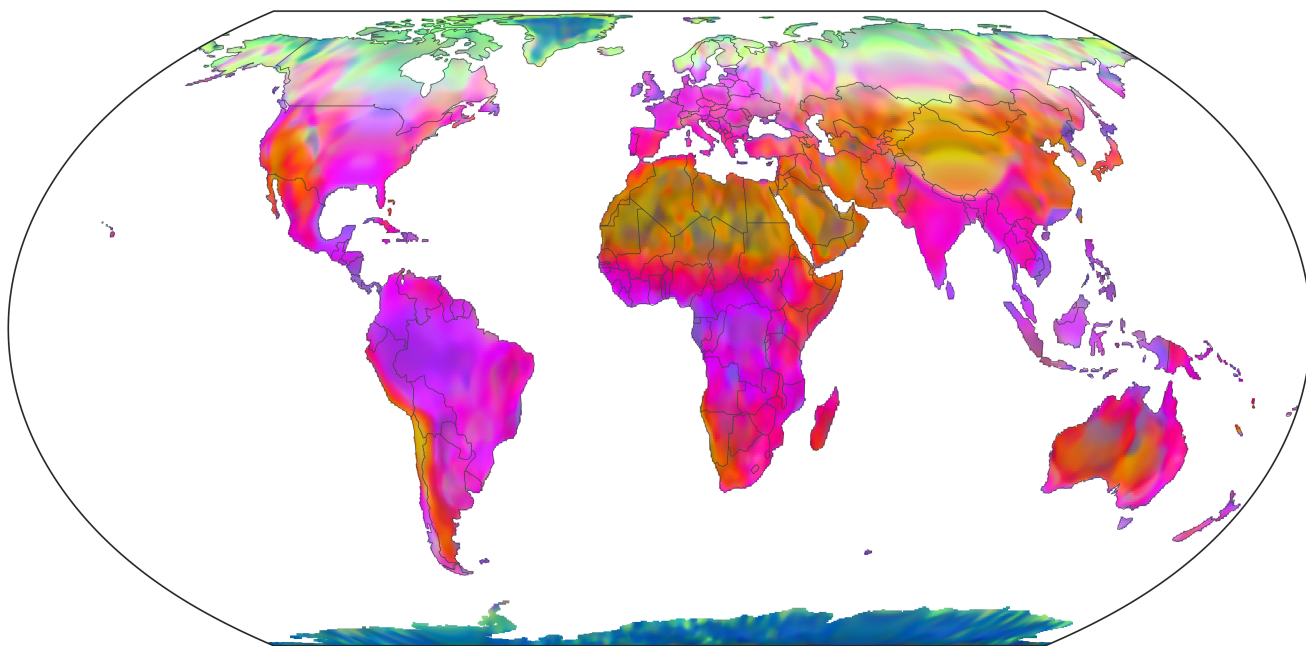


Figure 12: Blown-up version of first three principal components visualization of SLED embeddings trained on Sentinel-2, Landsat 8/9, and Sentinel-1. To be clear, the colors of PCA are individualized per map. Colors in this figure should not be compared with 11 or 10