SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery

Konstantin Klemmer[1], **Esther Rolf**[2], **Caleb Robinson**[3], **Lester Mackey**[1], **Marc Rußwurm**[4]
[1]Microsoft Research
[2]CU Boulder
[3]Microsoft AI for Good Research Lab
[4]Wageningen University & Research
{koklemmer, caleb.robinson, lmackey}@microsoft.com, esther.rolf@colorado.edu, marc.russwurm@wur.nl

Abstract

Geographic information is essential for modeling tasks in fields ranging from ecology to epidemiology. However, extracting relevant location characteristics can be challenging, often requiring expensive data fusion or distillation from massive global imagery datasets. To address this challenge, we introduce Satellite Contrastive Location-Image Pretraining (SatCLIP). This *global, general-purpose geographic location encoder* learns an implicit representation of locations by matching CNN and ViT inferred visual patterns of openly available satellite imagery with their geographic coordinates. The resulting SatCLIP location encoder efficiently summarizes the characteristics of any given location for convenient use in downstream tasks. In our experiments, we use SatCLIP embeddings to improve performance on nine diverse geospatial prediction tasks including temperature prediction, animal recognition, and population density estimation. SatCLIP consistently outperforms alternative location encoders and shows promise for improving geographic domain adaptation. The results demonstrate the potential of vision-location models to learn meaningful representations of our planet from the vast, varied, and largely untapped modalities of geospatial data.

Introduction

Satellite imagery has proven to be a valuable source of input data for predictive models across a wide range of real-world applications (Rolf et al. 2024), for example, interpolating missing air pollution data (Chen et al. 2019), crop yield forecasting (Tseng, Kerner, and Rolnick 2022), and agroforestry carbon stock prediction (Reiersen et al. 2022). Patterns extracted from satellite images can describe the unique characteristics of locations, by capturing their natural and built environment. Importantly, characteristics are often correlated in space: While two nearby locations are more likely to have similar features (e.g. the same land cover), two distant locations can also share location characteristics when they share similar environmental ground conditions like climate zones (Fig. 1, right).

Since the spatial patterns governing different geographic data modalities are often complex and non-linear, predictive models working with geo-data benefit from explicitly integrating intuitions for spatial and spatio-temporal dependencies (Klemmer and Neill 2021; Klemmer et al. 2022; Cole et al. 2023). Many geospatial modeling approaches in the sciences even directly leverage geographic location for improving predictions – in fields ranging from epidemiology, the Earth system sciences, to ecology (Mac Aodha, Cole, and Perona 2019; Cole et al. 2023).

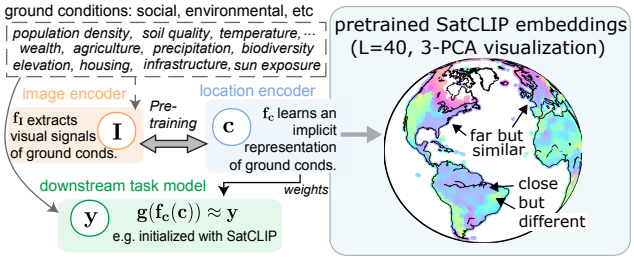


Figure 1: **Motivation for SatCLIP:** Capturing ground conditions from satellite images and transferring them into a location encoder via constrastive image-location pretaining.

But integrating geographic information into a deep learning model is not straightforward. Even though spatial coordinates are often informative, introducing them as features can amplify geographic distribution shift problems and lead to poor evaluation accuracy. This is especially of concern for the setting of geospatial domain adaptation, when data from evaluation areas (and their coordinates) are absent in the training data. On the other hand, models that ignore the spatial nature of the problem will generally perform worse in problems of spatial interpolation, where the evaluation area overlaps with the training area. This brings about an important challenge for representing Earth's diverse surface: capturing the rich information in ground conditions that computer vision models trained on satellite imagery can represent, while simultaneously respecting and integrating the spatial structures and interdependence of geospatial data.

We tackle this research problem by pretraining location encoders: models that take in latitude and longitude, and output a high dimensional embedding of any place on earth. Specifically, we leverage the location information indexing satellite images as an input to a contrastive pretraining objective that aims to match location-image pairs, similarly to the vision-language pretraining deployed in Contrastive Language Image Pretraining (CLIP) (Radford et al.

2021). Pretrained location encoders combine both pure spatial information (coordinates) and ground conditions captured by images representing the respective location (e.g. vegetation, built environment). In contrast to existing pretrained location encoders, we use uniformly sampled multispectral satellite imagery as our image input during pretraining, because we are primarily interested in training location encoders that are are useful across the entire globe, and for many possible prediction tasks.

Previous work on pretrained location encoders. Three studies have pretrained geographic location encoders, models that take as input spatial coordinates, and return learned contextual representation. Yin et al. (2019) propose GPS2Vec, a set of UTM-zone specific location encoders using geotagged Flickr images (YFCC100M) (Thomee et al. 2016) and their corresponding semantic tags for training. Geographic generalization was out of scope for this work as embeddings are only available for UTM-zones in which training data can be found. Mai et al. (2023) introduce Contrastive Spatial Pre-Training (CSP) on the iNaturalist 2018 (iNat) (Horn et al. 2018) species imagery and the Functional Map of the World (FMoW) (Christie et al. 2018) satellite image datasets. CSP is used for unsupervised pretraining and downstream prediction on the same datasets and was not conceptualized for use on other tasks. The CSP pretraining datasets, iNaturalist and FMoW, are also unevenly distributed over space, with high image densities in North America and Europe and few images outside of Western regions. Cepeda, Nayak, and Shah (2023) propose GeoCLIP, in which the authors pretrain image and location encoders using Flickr images from the MediaEval Placing Tasks 2016 (MP-16) dataset (Larson et al. 2017)–another dataset with strong overrepresentation of Western countries. GeoCLIP is developed for the task of geo-locating (natural) images and is not optimized or tested for general-purpose use.

Aim and contributions. Existing work leaves two important gaps: understanding how location encoders generalize across downstream tasks and prioritizing global coverage and approximately equal performance of location embeddings across geographies. We address both of these challenges by introducing the first *global-coverage, general-purpose pretrained geographic location encoder* based on Satellite Contrastive Location Image Pretraining – **SatCLIP**. SatCLIP distills spatially varying visual patterns from globally-distributed satellite data into an implicit neural representation in a comparatively small and efficient neural network. This location encoder projects a latitude and longitude coordinate into a higher-dimensional vector representation that is matched with a visual vector representation from a computer vision encoder (CNN or ViT), as detailed in the next section. More generally, the proposed framework represents a step towards geographically-informed "foundation models" trained with large, unlabeled datasets, that are usable for a wide range of tasks, and extrapolate to unseen geographic areas. Our contributions can be summarized as follows:

- We develop the first task-generalizable, global-coverage location encoder–**SatCLIP**–trained on Sentinel-2 multispectral satellite imagery. We release the pretrained encoder as a `PyTorch` model. We also release our new globally distributed pretraining dataset, **S2-100K**.
- We compare **SatCLIP** with existing pretrained location encoders and other geographic feature generation approaches on nine diverse downstream tasks ranging from temperature prediction to population density estimation, highlighting improved performance.

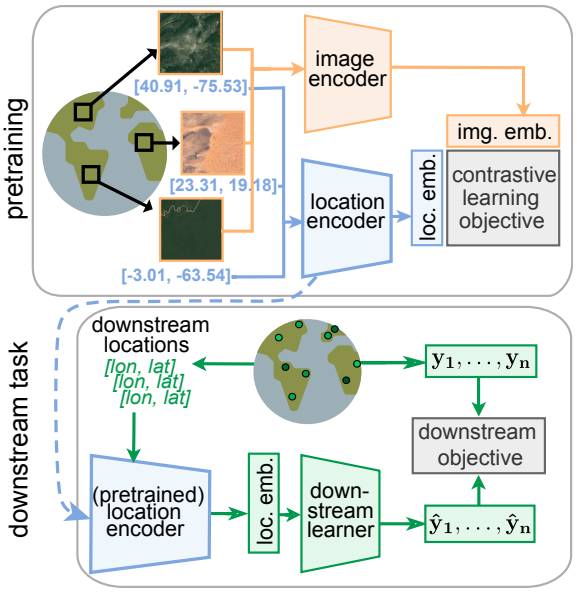


Figure 2: **The SatCLIP pretraining and deployment pipeline.** *Top:* SatCLIP pretraining through image-location matching. *Bottom:* The pretrained location encoder can then be used in downstream tasks.

Satellite Contrastive Location-Image Pretraining

With **SatCLIP**, we aim to train models that (1) provide *general purpose embeddings* and (2) are *globally representative*. In this section, we first motivate SatCLIP in light of these goals and then outline its components and training paradigm.

Various factors influence the appearance of Earth's surface, as captured in satellite images. In Fig. 1, we highlight diverse environmental and socioeconomic factors that are reflected in visual markers, such as the appearance of mountain ranges with their specific vegetation, the geometry and structure of agricultural fields, and the design of buildings. By training on the SatCLIP matching objective, the image encoder f_{I} is trained to associate an image $\mathbf{I}$ with a location $\mathbf{c}$ based on the various ground factors detectable in the images. At the same time, the location encoder f_{c} is trained to associate a given location with location-specific image characteristics. Effectively, both models learn to align the embeddings of a location and its corresponding image to maximize their similarity, as shown in the top panel of Fig. 2. Inductive biases in the location encoder model control spatial smoothness (i.e. the L hyperparameter in SatCLIP), allowing the

model to interpolate to areas where no image-location pairs are present in the training data. After successful training, location features can be extracted at arbitrary locations. These features can be used to train any downstream learner g that takes locations as input (bottom panel of Fig. 2).

There are several advantages to training downstream models with location embeddings, as opposed to raw coordinates or images extracted at downstream locations. Models trained on raw coordinates $\mathbf{c}$ solely rely upon spatial dependencies without taking any ground conditions into account, such as local elevation patterns or local climate zones. Models trained on full images $\mathbf{I}$, while able to capture ground conditions, require expensive data preprocessing (downloading images for every downstream location) and training of large vision models. This can be infeasible in many, especially low-resource settings. SatCLIP has the potential to provide the best of both worlds: location embeddings capture both spatial effects and ground conditions while also being relatively low-dimensional (our SatCLIP location embeddings are 256-dimensional vectors) and runtime efficient due to the small size of the location encoder model. This is particularly helpful in resource-constrained settings and allows fast encoding of coordinates to embedding space without GPUs.

Pretraining with the SatCLIP Objective

The inputs to a geographic location encoder are latitude/longitude coordinate pairs $\mathbf{c}_i = [\lambda_i, \phi_i]$, where λ_i is the longitude, ϕ_i is the latitude, and i indexes locations on the spherical surface $\mathbb{S}^2$. For each location i, we have a corresponding multi-spectral image $\mathbf{I}_i \in \mathbb{R}^{m \times n \times k}$ with k channels. We now define two encoders, a *location encoder* $f_{\text{c}} : \mathbb{S}^2 \to \mathbb{R}^d$ that takes in 2-dimensional coordinates $\mathbf{c}_i$ and returns a d-dimensional latent embedding and a *image encoder* $f_{\text{I}} : \mathbb{R}^{m \times n \times k} \to \mathbb{R}^d$ that takes in an image $\mathbf{I}_i$ and also returns a d-dimensional latent embedding.

We train both encoders with the simple but highly effective CLIP (Radford et al. 2021) objective

$$\mathcal{L} = \frac{1}{2N}\left[\sum_{i=1}^{N}\mathcal{L}_{\text{loc}}(\mathbf{c}_i, \mathbf{I}_{1,\dots,N}) + \sum_{i=1}^{N}\mathcal{L}_{\text{img}}(\mathbf{I}_i, \mathbf{c}_{1,\dots,N})\right] \quad (1)$$

that matches each coordinate $\mathbf{c}_i$ with the corresponding image $\mathbf{I}_i$ and against all images $\mathbf{I}_{1,\dots,N}$ using

$$\mathcal{L}_{\text{loc}}(\mathbf{c}_i, \mathbf{I}_{1,\dots,N}) = -\log \frac{\exp(\langle f_{\text{c}}(\mathbf{c}_i), f_{\text{I}}(\mathbf{I}_i)\rangle/\tau)}{\sum_{j=1}^{N}\exp(\langle f_{\text{c}}(\mathbf{c}_i), f_{\text{I}}(\mathbf{I}_j)\rangle/\tau)} \quad (2)$$

and each image with the corresponding coordinate using $\mathcal{L}_{\text{img}}(\mathbf{I}_i, \mathbf{c}_{1,\dots,N})$ over a batch $(\mathbf{c}_i, \mathbf{I}_i)_{i=1}^N$ of N coordinate-image tuples. The normalized dot-product is denoted by $\langle\cdot,\cdot\rangle$, and τ is a temperature hyperparameter. This objective optimizes the weights of the location encoder f_{c} and image encoder f_{I} simultaneously to embed the feature vectors of the corresponding location $f_{\text{c}}(\mathbf{c}_i) \in \mathbb{R}^d$ and image $f_{\text{I}}(\mathbf{I}_i) \in \mathbb{R}^d$ nearby in a common d-dimensional feature space.

Encoder Architectures

Geographic **location encoders** f_{c} typically take the form $f_{\text{c}} = \text{NN}(\text{PE}(\mathbf{c}_i))$, (Mai et al. 2023) where $\text{PE}(\mathbf{c}_i)$ is a nonparametric functional positional encoding and $\text{NN}(\cdot)$ is a small neural network. The positional encodings usually have a scale hyperparameter that controls spatial smoothness of the encoding. The neural network weights encode an implicit neural representation of a signal at a specific coordinate (Cole et al. 2023). In this work, we train $\text{Siren}(\text{SH}(\mathbf{c}_i))$ location encoders proposed by Rußwurm et al. (2024), which use spherical harmonics basis (SH) functions as positional encoders and are particularly well-suited for coordinates on spherical surfaces. They are combined with sinusoidal representation networks (Siren) (Sitzmann et al. 2020) that are broadly used for implicit neural representations. The spatial smoothness of the representation is controlled by the number of Legendre polynomials L. This effectively defines the resolution of the location encoding and its capacity to learn small and large-scale geospatial patterns, with larger L corresponding to finer spatial resolution.

As **image encoder**, we need a vision model that is expressive enough to learn visual patterns from satellite images. In this work, we use ResNet18, ResNet50, (He et al. 2016) and ViT16 (Dosovitskiy et al. 2020) vision encoders pretrained with momentum-contrast (MoCo) (He et al. 2020) on Sentinel-2 satellite imagery by Wang et al. (2022). To account for the size discrepancy between the large image models and relatively smaller location encoders (for example, the image encoder of a SatCLIP ViT16 has ~ 22 million parameters, while the location encoder has ~ 1 million parameters) during training, we freeze the vision encoder except for the last linear projection layer.

SatCLIP Implementation Details

We pretrain **SatCLIP** using the S2-100K dataset (described in the following section). We use 90% of the data points, selected uniformly at random, for pretraining and reserve the remaining 10% as a validation set to monitor overfitting. During pretraining, we found that batch sizes of $8k$ help the model to learn more fine-grained representations, while too large batch sizes can prevent learning, as was also recently observed on CLIP models by Zhai et al. (2023). We train models for 500 epochs on an A100 GPU. More pretraining details can be found in the appendix.

Experimental Setup

In our experiments, we focus on three research questions. From a performance perspective, we ask *how generalizable are SatCLIP embeddings* from Sentinel-2 data, both *across a diverse range of geospatial modeling tasks* (*RQ 1*) and *across unseen geographic areas* (*RQ 2*), compared to existing pretrained location encoders and location-only prediction? We design experiments to test the performance of SatCLIP embeddings for downstream tasks, both for in-sample prediction (spatial interpolation), and for geographic domain adaptation (in which the training and test sets are separated geographically). Geographic generalization is an important aspect of performance, as distributional changes across geographic areas are a common challenge in environmental problems like species distribution modeling (Beery et al. 2022), land cover classification (Rußwurm et al. 2020), crop

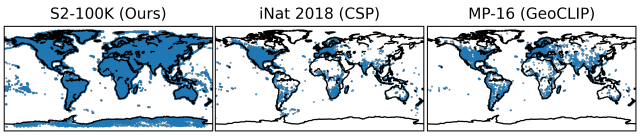


Figure 3: **Spatial distribution of the S2-100K dataset** we construct to pretrain SatCLIP, and of iNaturalist 2018 and MP-16, which are used to pretrain CSP and GeoCLIP models. iNaturalist and MP-16 heavily overrepresent North America and Europe.

type mapping (Kondmann et al. 2021). While location-based features are generally unsuitable for geographic generalization tasks, the implicit neural representation of environmental factors within the SatCLIP location encoder may provide generalizable information supporting prediction in areas with no or few labeled data points available.

After benchmarking performance, we turn to analyses that help us to further understand why SatCLIP works and what limitations it might exhibit. Ablation studies systematically vary the image encoder architecture and the scale parameters of the location encoder to understand the joint effect of each design choice. Complementing these quantitative experiments, we ask: *Do SatCLIP embeddings correspond to ground conditions contained within satellite images (like vegetation or built environment), and can they help to identify visually similar areas that are spatially far* (*RQ 3*)? Qualitative experiments shed light on what spatial relationships the pretrained SatCLIP embeddings harvest from S2-100K imagery, probing our intuition from Fig. 1.

In the remainder of this section, we detail our experimental design, including the pretaining dataset, downstream tasks, and comparison models.

Pretraining Dataset: S2-100K

To construct our pretraining dataset, **S2-100K**, we sample $100,000$ tiles of 256×256 pixel, multi-spectral (12-channel) Sentinel-2 satellite imagery and their associated centroid locations. We design the S2-100K dataset with the goals of *multi-task applicability* and *geographic generalization performance* in mind. Our dataset (1) represents general location features by using multi-spectral satellite imagery (as illustrated in Fig. 1) *and* (2) is nearly uniformly distributed across global land mass (Fig. 3, left). More details on the S2-100K dataset and sampling procedure are provided in the appendix.

In contrast to S2-100k, the pretraining datasets used in comparison methods often significantly underrepresent certain–especially non-Western–geographic areas, as they were not specifically designed to provide general-purpose embeddings. Fig. 3 illustrates the spatial coverage of S2-100K compared to the highly clustered distributions of iNaturalist, which is used as a pretraining dataset for CSP (Mai et al. 2023) and MediaEval 2016, which is used as a pretraining dataset for GeoCLIP (Cepeda, Nayak, and Shah 2023). Similar biases are exhibited by the Yahoo-Flickr Creative Commons 100 Million (YFCC100M) dataset (Thomee et al. 2016) of image-tag-location triplets (used in GPS2Vec (Yin et al. 2019)), and the Functional Map of the World (FMoW) (Christie et al. 2018) dataset, which samples satellite imagery mostly near human-built infrastructure.

Downstream Tasks

To test the general applicability of SatCLIP location embeddings, and whether they capture various ground conditions, we run experiments on a wide range of geospatial predictive modeling tasks. In all datasets, the inputs are raw latitude/longitude coordinates, which we transform into location embeddings. The nine downstream datasets we choose for evaluation span socioeconomic and environmental applications. To evaluate the degree to which location embeddings capture socioeconomic factors, we regress **Median Income** (Jia and Benson 2020), **California Housing** prices (Pace and Barry 2003), and logged **Population Density** (Rolf et al. 2021). We predict variables including **Air Temperature** (Hooker, Duveiller, and Cescatti 2018) and **Elevation** (Rolf et al. 2021) from coordinates as environmental regression objectives. We additionally classify **Biomes**, **Ecoregions** (Dinerstein et al. 2017), and compile a new country code classification task **Countries**. Lastly, we classify **iNaturalist** species (Horn et al. 2018). Here, we have additional image features extracted from an InceptionV3 model released by (Mac Aodha, Cole, and Perona 2019), which we concatenate with location embeddings during downstream training. Further details on downstream experiments can be found in the appendix.

Comparison Methods

We compare trained SatCLIP location embeddings to GPS2Vec (Yin et al. 2019), CSP (Mai et al. 2023) and GeoCLIP (Cepeda, Nayak, and Shah 2023) pretrained location embeddings. We refer to each comparison model by first stating the pretraining algorithm and then the pretraining dataset. For instance, CSP-FMoW represents CSP pretraining on FMoW dataset. Unless stated otherwise, we show results from SatCLIP models using a ViT-16 vision encoder. Like SatCLIP, GeoCLIP and CSP use the CLIP loss, with CSP adding loss terms for negative location sampling and SimCSE. GPS2Vec uses a KL divergence loss on the context (image and semantic tags) and location data.

In addition to comparing to existing location encoding methods, we compare to two different informative baselines to contextualize performance across our nine geospatial tasks. To compare to an image-only embedding, we use globally precomputed MOSAIKS (Rolf et al. 2021) features. To assess the performance improvement from the integration of contextual information over location-only prediction, we compare to downstream learners trained on raw latitude/longitude coordinates $g(\mathbf{c})$. We refer to this approach as "Identity" throughout our experiments. More details on the comparison methods can be found in the appendix.

Downstream Model Training

For all downstream tasks, we train multi-layer perceptron (MLP) models g with location embeddings and raw latitude/longitude coordinates as input to predict a (continuous

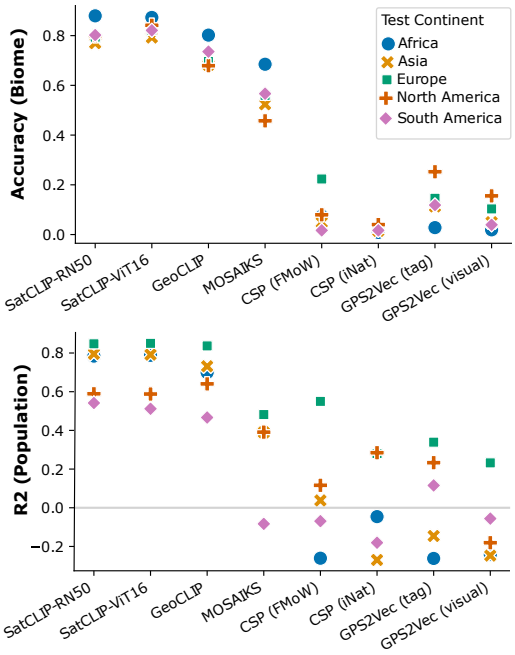


Figure 4: **Performance metrics aggregated by continent** highlight how location embeddings perform in different geographic areas for population density estimation and biome classification. $L = 40$ SatCLIP models are shown.

or discrete) outcome variable y. Regression models use a mean squared error (MSE) loss, and classification models use cross-entropy loss. Hyperparameters like learning rate, number of layers, or hidden dimensions are tuned using a random search on an independent validation set. All results are reported for an unseen test set. More details on the downstream task setups can be found in the appendix.

Results

Downstream Task Performance (*RQ 1*)

Comparison across tasks. Tab. 1 shows performance across different downstream tasks and methods. SatCLIP embeddings (in either $L = 10$ or $L = 40$ configuration) achieve the best prediction scores by a large margin on all seven of the globally distributed tasks. On the two other tasks (the Cali. Housing dataset, which is limited to California, and the Median Income dataset, limited to the continental United States (US)), the GeoCLIP model trained on US-centric MP-16 data performs equal to or better than SatCLIP. We also observe that SatCLIP embeddings with higher spatial resolution ($L = 40$) perform better than coarse-grained ($L = 10$) embeddings at these two regionally constrained tasks.

Comparison across continents. Fig. 4 shows the performance of each method stratified by continent for the tasks of biome classification and population density estimation. This experiment highlights how SatCLIP outperforms prior location encoders (CSP and GPS2Vec) trained on spatially biased training data, which perform better in Europe and North America than in the underrepresented continents of Africa, Asia, and South America. GeoCLIP is closest in performance to SatCLIP for both tasks, performing similarly well to SatCLIP on the population density task, but worse across continents on the Biome classification task.

Zero/Few-Shot Geographic Adaptation (*RQ 2*)

Tab. 2 shows how our embeddings perform at geographic adaptation. For these experiments, we deploy a spatial train/test split strategy: We hold out entire continents, either Africa or Asia, as test sets and use the remaining data for model training and validation–this emulates a frequent problem of spatial data gaps in real-world applications. We highlight results on our classification tasks. We test a few-shot domain adaptation setting, where we add a small sample (1%, uniformly sampled) of test continent points to the training set, with the Countries and Ecoregions datasets, which both contain labels that are unique to a continent. For the Biome and iNat datasets, we test zero-shot adaptation by not providing any training points from the held-out test continent. In iNaturalist 2018, this means that the model will not be able to recognize any species that are endemic to the test continent (i.e., that do not live in any other continent) but only those known from training continents.

Overall, SatCLIP embeddings outperform existing location encoders at few-shot geographic adaptation by a large margin. Here, several existing location encoders even perform significantly worse than directly encoding latitude and longitude for out of sample prediction ("Identity" in Tab. 2). For both datasets, the more coarse grained $L = 10$ SatCLIP models are able to adapt better. Fig. 5 visualizes the predictions and failure modes from different methods on few-shot Ecoregion prediction in Africa. In contrast, zero-shot geographic adapation for Biome classification shows GeoCLIP and MOSAIKS embeddings outperforming SatCLIP embeddings. For iNaturalist zero-shot geographic adapation, accuracy values are extremely low and no method outperforms the simple "Identity" baseline.

While zero-shot adaptation remains a significant challenge for all methods, our results highlight the promise of SatCLIP embeddings for supporting geographic adaptation in the few-shot geographic domain adaption setting. We partially attribute this success to the ability of SatCLIP to, for a given location, identify visually similar areas spatially far (e.g. on a different continent). This can help downstream learners to adapt to new geographic areas better, resulting in improved predictive performance. We investigate this intuition further in the following section.

Location Embedding Analysis (*RQ 3*)

We now investigate qualitatively to what degree the SatCLIP embeddings have learned an implicit representation of different ground conditions in the location encoder weights. We first visualize a low-dimensional projection of the latent representations from our location encoders. Fig. 6a shows an RGB representation of the first three principal components of SatCLIP embeddings at locations around the planet. The figure highlights how embeddings learned by SatCLIP provide fine-grained representations of different locations (expressed by different colors), capturing global patterns like

Task ↓ Data →	SatCLIP$_{L=10}$ (S2-100K)	SatCLIP$_{L=40}$ (S2-100K)	CSP (iNat)	GPS2Vec (tag)	MOSAIKS (Planet)	GeoCLIP (MP-16)	Identity ($y \sim g(\mathbf{c})$)
Regression	R^2 ↑						
Air temperature	$\mathbf{0.90 \pm 0.13}$	$\mathbf{0.91 \pm 0.01}$	-0.56 ± 0.59	0.22 ± 0.00	-0.52 ± 2.00	-3.11 ± 5.24	0.82 ± 0.16
Median income	0.42 ± 0.01	$\mathbf{0.47 \pm 0.12}$	-0.01 ± 0.02	0.21 ± 0.00	0.02 ± 0.05	$\mathbf{0.50 \pm 0.01}$	-0.84 ± 0.94
Cali. housing	0.35 ± 0.04	0.57 ± 0.02	-0.00 ± 0.00	0.71 ± 0.03	0.24 ± 0.02	$\mathbf{0.75 \pm 0.01}$	0.05 ± 0.02
Elevation	0.83 ± 0.01	$\mathbf{0.88 \pm 0.00}$	0.11 ± 0.05	0.10 ± 0.00	0.21 ± 0.01	0.83 ± 0.00	0.25 ± 0.08
Population	0.79 ± 0.00	$\mathbf{0.82 \pm 0.00}$	0.36 ± 0.11	0.25 ± 0.00	0.46 ± 0.02	0.79 ± 0.00	0.46 ± 0.03
Classification	% Accuracy ↑						
Countries	94.28 ± 0.18	$\mathbf{96.00 \pm 0.14}$	82.11 ± 1.72	70.35 ± 0.06	76.16 ± 0.50	90.72 ± 0.44	82.94 ± 2.23
iNaturalist	$\mathbf{65.69 \pm 0.18}$	$\mathbf{66.22 \pm 0.40}$	60.47 ± 0.56	58.78 ± 0.48	56.73 ± 0.8	62.01 ± 0.59	60.83 ± 0.53
Biome	92.23 ± 0.26	$\mathbf{94.41 \pm 0.14}$	73.18 ± 5.58	69.69 ± 0.06	79.61 ± 0.42	89.57 ± 0.45	83.55 ± 2.43
Ecoregions	89.32 ± 0.31	$\mathbf{91.67 \pm 0.15}$	78.43 ± 1.71	68.46 ± 0.06	70.48 ± 0.21	84.65 ± 0.32	77.07 ± 2.54

Table 1: **Downstream task performance using SatCLIP (with ResNet50) vs. comparison location embeddings.** We report average test set R^2 and accuracy ± 1 standard deviation across 10 independently initialized training runs.

Held-out Test Continent	SatCLIP$_{L=10}$ (S2-100K)	SatCLIP$_{L=40}$ (S2-100K)	CSP (iNat)	GPS2Vec (tag)	MOSAIKS (Planet)	GeoCLIP (MP-16)	Identity ($y \sim g(\mathbf{c})$)
Asia							
Countries† % Acc. ↑	$\mathbf{36.90 \pm 4.32}$	19.17 ± 2.82	1.28 ± 0.01	1.12 ± 0.00	1.56 ± 0.47	23.12 ± 2.50	1.24 ± 0.12
Ecoregions†	$\mathbf{21.02 \pm 1.09}$	10.86 ± 1.19	1.41 ± 0.14	1.49 ± 0.03	1.36 ± 0.10	6.65 ± 1.03	1.52 ± 0.47
iNaturalist*	19.60 ± 0.78	$\mathbf{20.91 \pm 0.77}$	$\mathbf{21.49 \pm 0.85}$	17.52 ± 0.38	16.14 ± 0.42	$\mathbf{20.94 \pm 0.38}$	$\mathbf{21.08 \pm 0.69}$
Biome*	25.89 ± 2.79	16.44 ± 1.21	3.00 ± 2.60	1.76 ± 0.04	$\mathbf{37.81 \pm 4.47}$	$\mathbf{31.67 \pm 1.91}$	6.24 ± 2.71
Africa							
Countries† % Acc. ↑	$\mathbf{30.65 \pm 4.23}$	10.22 ± 1.62	0.45 ± 0.04	0.47 ± 0.01	0.48 ± 0.00	10.32 ± 2.75	2.74 ± 2.52
Ecoregions†	$\mathbf{32.03 \pm 1.19}$	12.91 ± 1.63	0.94 ± 0.04	0.88 ± 0.01	0.92 ± 0.12	12.41 ± 2.20	7.72 ± 3.93
iNaturalist*	$\mathbf{9.53 \pm 0.57}$	6.23 ± 0.47	8.65 ± 0.52	7.47 ± 0.53	5.18 ± 0.38	7.69 ± 0.30	$\mathbf{9.96 \pm 0.33}$
Biome*	35.72 ± 5.48	12.34 ± 1.75	1.09 ± 0.48	1.29 ± 0.04	$\mathbf{49.86 \pm 1.57}$	28.28 ± 3.06	1.46 ± 0.67

Table 2: **Zero-shot ($*$) and few-shot ($\dagger$) geographic adaptation capabilities of SatCLIP (with ViT16 vision encoder) vs. baseline location embeddings to new geographic areas.** We report average test set accuracy in % ± 1 standard deviation across 10 independently initialized fine-tuning runs.

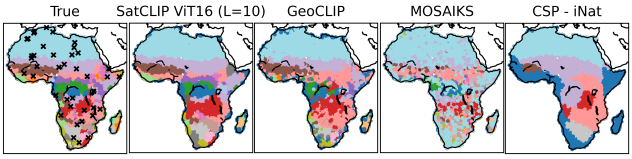


Figure 5: **Geographic adaptation: Africa Ecoregion predictions**. SatCLIP with $L = 10$ maps Ecoregions in Africa closest to the ground truth, followed by GeoCLIP. MOSAIKS predictions are too fine-grained, while CSP-iNat is too coarse. The "$\times$" symbols in the "True" panel show the sparse training locations for few-shot learning in Africa (on average 480km apart from their nearest neighbor).

climate zones. The figure also highlights the different spatial smoothness and change-over-space of $L = 10$ and $L = 40$ SatCLIP embeddings and the impact of the vision encoder.

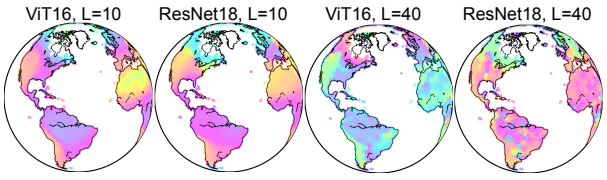
Next, we examine similarities between embeddings of different locations measured in the cosine distance between the embedding of a location $f_{\text{c}}(\mathbf{c})$ with respect to a reference location $f_{\text{c}}(\mathbf{c}_\star)$. Fig. 6b shows the similarity of a grid of locations, i.e., the map with respect to reference locations in the Congo Basin and on the east coast of North America on the right panel. The reference locations are marked by a star $\star$ on the map. SatCLIP location embeddings show high similarity between the Congo Basin location and other areas close to the Equator, particularly the Amazon and Indonesia (red areas on the left panel). In comparison, embeddings of the North American location are similar to areas in Europe or northern China that are similarly population-dense and industrialized.

Effect of Encoder Design on Performance

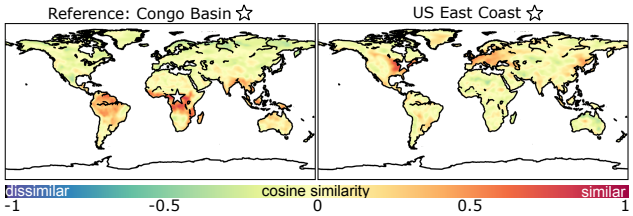
Lastly, we report what effect the vision encoder type and spatial resolution of the location encoder (L hyperparameter) have on the downstream accuracy. The choice of image encoder for pretraining (ViT-16, ResNet-18, ResNet-50) appears to only marginally affect results (performance differences of $< 1\%$). In contrast, different location encoder resolutions (scale parameters L) have a greater effect: the $L = 40$ location encoders tend to be better for interpolation tasks, as shown in Tab. 3. $L = 10$ location encoders are better for the zero- and few-shot adaptation experiments.

Discussion

In reference to RQ1, our results show that *SatCLIP models can provide useful information for a wide range of downstream tasks*. SatCLIP outperforms other location encoding approaches on globally distributed downstream tasks, while



(a) **Latent space visualization**: First 3 principal components from the 256-dimensional SatCLIP embedding as RGB for four models with different L values and vision encoders. PCs are calculated separately for each globe; colors across globes are incomparable.



(b) **Location similarities learned by SatCLIP** give insights into areas with similar visual features. Highlighted are similarities between location embeddings for a Congo Basin location (left) and a US East Coast location (right)–marked by a star–and the rest of the world. SatCLIP associates the Congo Basin with other Equatorial locations, e.g, in the Amazon, and associates the US East Coast with other densely populated areas, e.g., in Central Europe.

Figure 6: **Analysis of SatCLIP location embeddings**.

	Countries	iNaturalist	Biomes	Ecoregions
ViT16$_{L=10}$	93.97 ± 0.30	65.69 ± 0.50	92.07 ± 0.22	89.53 ± 0.28
ResNet18	93.92 ± 0.30	65.56 ± 0.29	92.10 ± 0.23	89.57 ± 0.23
ResNet50	94.28 ± 0.18	65.50 ± 0.43	92.23 ± 0.26	89.32 ± 0.31
ViT16$_{L=40}$	95.77 ± 0.14	65.98 ± 0.61	94.27 ± 0.15	91.61 ± 0.22
ResNet18	95.92 ± 0.10	66.40 ± 0.49	94.33 ± 0.10	91.53 ± 0.15
ResNet50	96.00 ± 0.14	66.03 ± 0.54	94.41 ± 0.14	91.67 ± 0.15

Table 3: Comparisons of different vision encoders (% accuracy) at classification tasks (comp. Tab. 1). Full results can be found in the appendix.

for regional datasets such as Cali. Housing, other methods like GeoCLIP are competitive. SatCLIP embeddings perform well across continents and are less prone to geographic bias in comparison to other methods like GPS2Vec or CSP, for which performance tends to degrade outside of Europe or North America.

Regarding *RQ2*, our results indicate that the *transfer of spatial patterns in Sentinel-2 imagery into the SatCLIP location encoder enables some degree of generalization across geographic areas*. SatCLIP is by far the best of all models tested under conditions of few-shot geographic domain adaptation and improves on two informative baselines: a satellite image-only embedding precomputed globally ("MOSAIKS") and a location-only ("Identity") baseline method. This highlights one of the key strengths of SatCLIP: its ability to globally measure location similarities based on visual markers. This allows downstream models trained on SatCLIP embeddings to learn patterns in sparse areas, as they can leverage visually similar areas around the world.

Qualitative analyses support these interpretations of our results and confirm the intuition motivating *RQ3*. Our similarity analysis of location embeddings shows that spatially far locations like in South America (Amazon rainforest) and Africa (Congo Basin) are embedded nearby in the SatCLIP embedding space due to the visual similarity of satellite images from these locations. This can explain why SatCLIP models can sometimes generalize to unseen geographic areas with no or little training data, as demonstrated in our geographic domain generalization experiments.

Additional experiments show that the downstream performance of our current SatCLIP is more affected by changes to the location encoder scale factor than the exact vision architecture used for the image encoder. Lastly, our experiments also highlight the continued difficulty of zero-shot geographic adaptation, where SatCLIP does not provide an improvement over existing methods.

Conclusion

We present SatCLIP, a method that learns an implicit neural representation of visual patterns on the globe by matching satellite images and their respective coordinates using a contrastive location-image pretraining objective. Experiments show that SatCLIP is effective for global prediction tasks spanning social and environmental domains, for both interpolation and out-of-sample geographic prediction, and compared to existing location encoders, image-only and location-only prediction. Methodologically, the two key innovations of SatCLIP are its use of a pretraining dataset that is *uniformly distributed across the globe* and its deployment of a Siren(SH) location encoder, well-suited for learning global-scale data representations.

The effectiveness of SatCLIP despite its relative simplicity in implementation (a single contrastive loss on $100,000$ openly available satellite images) suggests several avenues for expansion. First, the image encoder we utilize in SatCLIP can be seen as a special case of a more general *context* encoder that may integrate other georeferenced data modalities like audio from acoustic sensors or text from geolocated social media posts for multi-source geospatial learning. Second, the current SatCLIP pre-trained weights have limited spatial scales, dictated by the L parameter of the location encoder. For extremely high resolution or local phenomena, it will be important to further study the effect of different choices of the location encoder and its use for downstream learning. Third, there is the potential to expand the SatCLIP training framework to encode locations in both time and space. Here, we effectively marginalize over all time points in the S2-100k dataset (which is sampled over two years and thus includes seasonal differences in images) but did not directly embed time in a space-time encoder of e.g., the form $f(lat, lon, time)$. Research in these directions would require separate innovations (on e.g.model architecture), and would need a new set pretraining data and/or downstream tasks for evaluation. We seek to tackle these extensions in future work. Code for SatCLIP pretraining and downstream experiments as well as the S2-100K dataset is available at https://github.com/microsoft/satclip. To access the full appendix of our paper, including additional details and results, please refer to the `arXiv` version (Klemmer et al. 2023).

Acknowledgments

ER was supported by the Harvard Data Science Initiative, the Center for Research on Computation and Microsoft.

References

Beery, S.; Wu, G.; Edwards, T.; Pavetic, F.; Majewski, B.; Mukherjee, S.; Chan, S.; Morgan, J.; Rathod, V.; and Huang, J. 2022. The Auto Arborist Dataset: A Large-Scale Benchmark for Multiview Urban Forest Monitoring Under Domain Shift. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 21294–21307.

Cepeda, V. V.; Nayak, G. K.; and Shah, M. 2023. GeoCLIP: Clip-Inspired Alignment between Locations and Images for Effective Worldwide Geo-localization. *arXiv preprint arXiv:2309.16020*.

Chen, C.; Li, K.; Teo, S. G.; Zou, X.; Wang, K.; Wang, J.; and Zeng, Z. 2019. Gated residual recurrent graph neural networks for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 485–492. AAAI Press. ISBN 9781577358091.

Christie, G.; Fendley, N.; Wilson, J.; and Mukherjee, R. 2018. Functional Map of the World. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 6172–6180.

Cole, E.; Horn, G. V.; Lange, C.; Shepard, A.; Leary, P.; Perona, P.; Loarie, S.; and Mac Aodha, O. 2023. Spatial Implicit Neural Representations for Global-Scale Species Mapping. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 6320–6342. PMLR.

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.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

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.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. volume 2016-Decem, 770–778. ISBN 9781467388504.

Hooker, J.; Duveiller, G.; and Cescatti, A. 2018. Data descriptor: A global dataset of air temperature derived from satellite remote sensing and weather stations. *Scientific Data*, 5: 1–11.

Horn, G. V.; Aodha, O. M.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; Belongie, S.; Google, C. .; and Tech, C. 2018. The iNaturalist Species Classification and Detection Dataset. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 8769–8778. ISBN 5,986561,767.

Jia, J.; and Benson, A. R. 2020. Residual Correlation in Graph Neural Network Regression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 588–598. Association for Computing Machinery. ISBN 9781450379984.

Klemmer, K.; and Neill, D. B. 2021. Auxiliary-task learning for geographic data with autoregressive embeddings. In *SIGSPATIAL: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*.

Klemmer, K.; Rolf, E.; Robinson, C.; Mackey, L.; and Rußwurm, M. 2023. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*.

Klemmer, K.; Xu, T.; Acciaio, B.; and Neill, D. B. 2022. SPATE-GAN: Improved Generative Modeling of Dynamic Spatio-Temporal Patterns with an Autoregressive Embedding Loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 4523–4531.

Kondmann, L.; Toker, A.; Rußwurm, M.; Camero, A.; Peressuti, D.; Milcinski, G.; Mathieu, P.-P.; Longépé, N.; Davis, T.; Marchisio, G.; et al. 2021. DENETHOR: The DynamicEarthNET dataset for Harmonized, inter-Operable, analysis-Ready, daily crop monitoring from space. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Larson, M.; Soleymani, M.; Gravier, G.; Ionescu, B.; and Jones, G. J. 2017. The Benchmarking Initiative for Multimedia Evaluation: MediaEval 2016. *IEEE Multimedia*, 24: 93–96.

Mac Aodha, O.; Cole, E.; and Perona, P. 2019. Presence-Only Geographical Priors for Fine-Grained Image Classification. In *ICCV*.

Mai, G.; Lao, N.; He, Y.; Song, J.; and Ermon, S. 2023. CSP: Self-Supervised Contrastive Spatial Pre-Training for Geospatial-Visual Representations. *arXiv preprint arXiv:2305.01118*.

Pace, R. K.; and Barry, R. 2003. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33: 291–297.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings in the International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Reiersen, G.; Dao, D.; Lütjens, B.; Klemmer, K.; Amara, K.; Steinegger, A.; Zhang, C.; and Zhu, X. 2022. ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 12119–12125.

Rolf, E.; Klemmer, K.; Robinson, C.; and Kerner, H. 2024. Position: Mission Critical – Satellite Data is a Distinct Modality in Machine Learning. In *Forty-first International Conference on Machine Learning*.

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 2021 12:1*, 12: 1–11.

Rußwurm, M.; Klemmer, K.; Rolf, E.; Zbinden, R.; and Tuia, D. 2024. Geographic Location Encoding with Spherical Harmonics and Sinusoidal Representation Networks. *Proceedings in the International Conference on Learning Representations (ICLR)*.

Rußwurm, M.; Wang, S.; Korner, M.; and Lobell, D. 2020. Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, 200–201.

Sitzmann, V.; Martel, J. N. P.; Bergman, A. W.; Lindell, D. B.; and Wetzstein, G. 2020. Implicit Neural Representations with Periodic Activation Functions. *Advances in Neural Information Processing Systems*, 33: 7462–7473.

Thomee, B.; Elizalde, B.; Shamma, D. A.; Ni, K.; Friedland, G.; Poland, D.; Borth, D.; and Li, L. J. 2016. YFCC100M. *Communications of the ACM*, 59: 64–73.

Tseng, G.; Kerner, H.; and Rolnick, D. 2022. TIML: Task-informed meta-learning for agriculture. *arXiv preprint arXiv:2202.02124*.

Wang, Y.; Ait, N.; Braham, A.; Xiong, Z.; Liu, C.; Albrecht, C. M.; and Zhu, X. X. 2022. SSL4EO-S12: A Large-Scale Multi-Modal, Multi-Temporal Dataset for Self-Supervised Learning in Earth Observation.

Yin, Y.; Liu, Z.; Zhang, Y.; Wang, S.; Shah, R. R.; and Zimmermann, R. 2019. GPS2Vec: Towards generating worldwide GPS embeddings. In *SIGSPATIAL: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, 416–419. Association for Computing Machinery. ISBN 9781450369091.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L.; and Deepmind, G. 2023. Sigmoid Loss for Language Image Pre-Training. In *ICCV*.