Towards Global, General-Purpose Pretrained Geographic Location Encoders

Konstantin Klemmer
Microsoft Research

Esther Rolf
Harvard University

Caleb Robinson
Microsoft AI for Good

Lester Mackey
Microsoft Research

Marc Rußwurm
Wageningen University

Abstract

Location information is essential for modeling tasks in climate-related fields ranging from ecology to the Earth system sciences. However, obtaining meaningful location representation is challenging and requires a model to distill semantic location information from available data, such as remote sensing imagery. To address this challenge, we introduce SatCLIP, a global, general-purpose geographic location encoder that provides vector embeddings summarizing the characteristics of a given location for convenient usage in diverse downstream tasks. We show that SatCLIP embeddings, pretrained on multi-spectral Sentinel-2 satellite data, can be used for various predictive out-of-domain tasks, including temperature prediction and animal recognition in imagery, and outperform existing competing approaches. SatCLIP embeddings also prove helpful in overcoming geographic domain shift. This demonstrates the potential of general-purpose location encoders and opens the door to learning meaningful representations of our planet from the vast, varied, and largely untapped modalities of geospatial data.

1 Introduction

Much of the world's data is geospatial. From images taken with a cellphone to the movement trajectories of taxis, different modalities live in the same geometric space: planet Earth. Geographic features, the characteristics describing any location on our planet, are commonly used in climate-related predictive modeling tasks, e.g., by using satellite imagery for crop yield prediction [9, 4]. Much of the existing work in geospatial machine learning does not directly encode spatial information as an input in the modeling process, even though this information is often readily available. Instead, indirect contextual information is used, such as images taken close to the location. New approaches aim to represent relevant features of locations *directly* in an implicit neural representation by embedding contextual information in location encoder model weights [8, 17, 12]. The location information within georeferenced images lends itself conveniently to contrastive, self-supervised pretraining objectives: given a dataset that includes satellite images and their corresponding locations, we can devise a pretraining task that matches images to locations or more precisely, image embeddings to location embeddings. This is analogous to the text-image pretraining deployed in the popular CLIP model [10].

Satellite imagery can be freely acquired for all land mass of our Planet at regular intervals, e.g. using the Sentinel-2 satellite. This enables pretraining a location encoder which, for any given location on the planet, returns contextual vectors encoding the characteristics of this location captured by satellite imagery. Location representations obtained from this encoder can be deployed in a range of different downstream tasks, as we show in this paper. We first outline the intuition for pretraining general-purpose geographic location encoders and deploying them in real-world tasks (see also

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

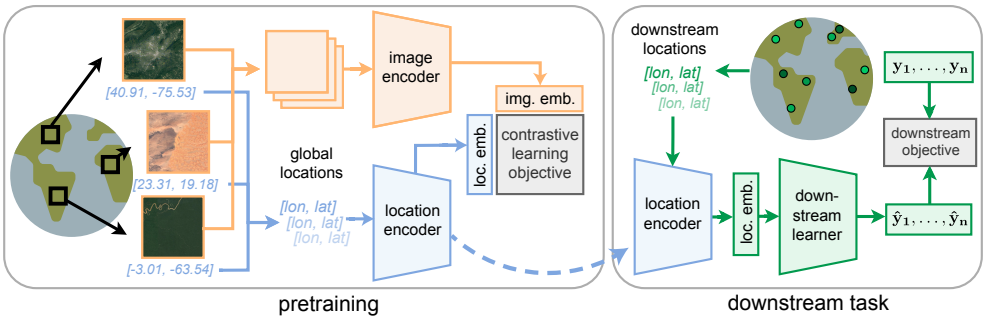


Figure 1: Pretraining (left) to downstream task deployment (right) pipeline of *general-purpose, global-coverage geographic location encoders*. Trained using large, unlabeled geographic data (e.g., satellite imagery), the learned location encoder can then be queried for contextual embeddings at downstream locations and–in combination with other relevant features–used for geospatial predictive modeling tasks.

Figure 1). We then introduce a new model from this family: **SatCLIP**, a pretrained geographic location encoder leveraging recent advances in geospatial coordinate encoding with spherical harmonics functions [12]. We train SatCLIP using a new dataset, **S2-100K**, compiled from uniformly distributed Sentinel-2 satellite imagery. Lastly, we compare **SatCLIP** to existing location encoders on different real-world tasks, highlighting how **SatCLIP** embeddings are highly effective for predictive modeling across geospatial tasks.

2 Approach

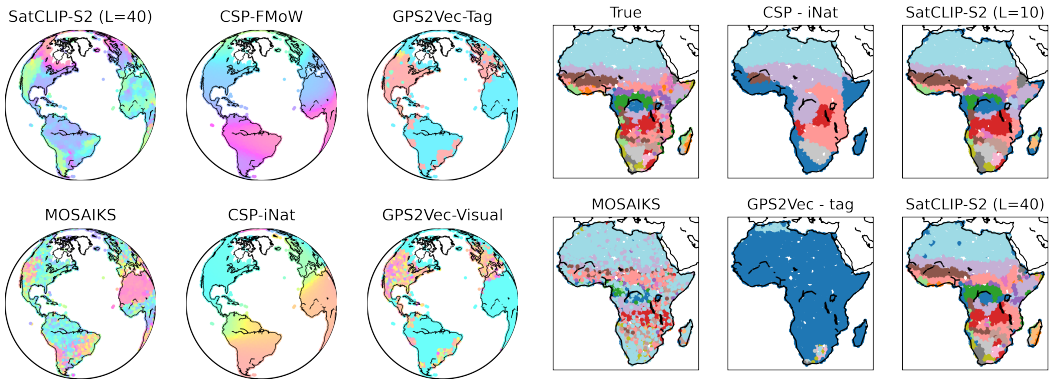
2.1 Pretrained geographic location encoders

The inputs to a geographic location encoder are longitude/latitude coordinate pairs $\mathbf{c}_i = [\lambda_i, \phi_i]$ at locations i and on the spherical surface $\mathbb{S}$. For each location i, we have corresponding contextual data, e.g. an image $\mathbf{I}_i \in \mathbb{R}^{m \times n \times c}$ with c channels. The this context describes the characteristics of that location, in the case of SatCLIP we use satellite images. We now define two encoders, a *location encoder* $\text{Enc}_{\text{loc}}(\mathbf{c}_i, \theta_{\text{loc}}) : \mathbb{S}^2 \rightarrow \mathbb{R}^d$ that takes in 2-dimensional coordinates and returns a d-dimensional latent embedding and a *context encoder* $\text{Enc}_{\text{cont}}(\mathbf{x}_i, \theta_{\text{cont}}) : \mathbb{R}^m \rightarrow \mathbb{R}^d$ that takes in images $\mathbf{I}_i$ and returns a d-dimensional latent embedding.

We can then define a loss that matches context vector embeddings and their correpsonding geolocation embeddings and contrasts them to others in a training batch. This contrastive self-supervised learning transfers location clues from the context encoder to the location encoder weights during pretraining. The goal is for the location encoder and context encoder to both capture the relevant representations of how the context data varies across space – though each model has distinctly different inputs. After pretraining, the location encoders can process *any* location coordinate. The amount to which pre-trained embeddings will represent the salient aspects of arbitrary locations will likely depend on the geographic distribution of the training data; learned representations in areas with low or no coverage in the training data rely on interpolating between the nearest available observations and may thus be less expressive.

2.2 SatCLIP

Data. The key to fulfilling the promise of *global-coverage* and *general-purpose* location encoders lies in the geographic context, we provide during training. This comes down to a fundamental question: what data best and most completely describes a location? Remotely sensed data is generally considered highly informative, as it can capture both natural and built features. We construct a new pretraining dataset, **S2-100K**, that (1) more generally represents location features and (2) is uniformly distributed across space, tackling the problem of underrepresentation of certain – especially non-Western – geographic areas. We sample $100,000$ tiles of multi-spectral Sentinel-2 satellite imagery and their associated centroid locations, uniformly distributed across global landmass.



(a) **Latent space visualization.** (b) **Ecoregion prediction.**

Figure 2: Figure 2a shows a visualization of the top-3 principal components, plotted as RGB channels, of the different embeddings on regularly distributed locations around the planet. Figure 2b shows predictions for MLPs trained using SatCLIP or comparison embeddings in a few-shot geographic adaptation setting: Locations in Africa are held-out during training (except a few examples) and constitute the test set.

Image and Location Encoder. Designing effective location encoders can be challenging: A common difficulty is to balance the low-dimensional location inputs (a 2-d vector) with the often high-dimensional context inputs (e.g., an image). This leads to substantially different sizes in the parameter space of θ_{loc} and θ_{cont} which, can lead to overfitting issues during training. To overcome this, we use Sentinel-2 pretrained ResNet18, ResNet50 and ViT16 **image encoders**, published by [15], freezing the network except for the last linear projection layer. Another challenge is the definition of an appropriate location encoder. Latitude and longitude values represent coordinates on a sphere and benefit from a positional encoding before being processed by neural networks [5, 7, 3]. We deploy a **location encoder** recently proposed by [12], which includes a positional encoding based on orthogonal spherical harmonics functions in combination with sinusoidal representation networks [Siren, 13]. This new location encoder is designed specifically to encode *global* locations. The hyperparameter L (corresponding to the number of Legendre polynomials used for harmonics calculation) controls the spatial resolution of the location encoder.

Training setting. To train our model, we use the CLIP [10] objective, using normalized dot product similarities between image and location embeddings to compute the loss. We pretrain **SatCLIP** with S2-100K using 90% of the data, selected uniformly at random, reserving 10% as validation set to monitor overfitting. We find that SatCLIP models pretrain best with batch sizes of $8k$. We train models for 500 epochs on a single A100 GPU.

3 Experiments

Comparison Models. We compare SatCLIP embeddings to embeddings obtained by several recently proposed algorithms. *CSP* [8] is a location encoder trained similarly to SatCLIP, but using additional auxiliary losses. It deploys a sinusoidal location encoder [6] and pretrained image encoders. Two models, one pretrained on iNaturalist 2018 (iNat)–a dataset of species imagery–and one pretrained on Functional Map of the World (FMoW)–a satellite image dataset focusing on built infrastructure, are published. *GPS2Vec* [17] uses a two-layer soft encoding and trains two models on YFCC-100M, a social media imagery dataset and their corresponding semantic tags. *MOSAIKS* [11] returns random convolutional features extracted directly from the nearest gridded satellite image at a given location. Notably, competing approaches either don't provide global coverage or are trained as domain-experts. SatCLIP is the first *general-purpose, global-coverage* location encoder.

Downstream Tasks. We run experiments on five different environmental modeling tasks: (1) *Air Temperature* prediction (regression), *Elevation* prediction (regression), species classification from *iNat* imagery (classification), *Biome* prediction (classification) and *Ecoregion* prediction (Classification). For all tasks, we tune and train simple multi-layer perceptrond (MLP) that take location embeddings (or in the case of iNat additionally image embeddings) as input.

Table 1: **Downstream task performance using SatCLIP ($L = 40$) vs. baseline location embeddings.** We report average MSE and accuracy ± 1 standard deviation across 10 independently initialized MLP training runs from an unseen, randomly sampled test set.

Task ↓ Data →	SatCLIP-RN50 (S2-100K)	SatCLIP-ViT16 (S2-100K)	CSP (FMoW)	CSP (iNat)	GPS2Vec (tag)	GPS2Vec (visual)	MOSAIKS (Planet)
Regression	MSE ↓						
Air temperature	$\mathbf{0.27 \pm 0.03}$	$\mathbf{0.25 \pm 0.02}$	2.81 ± 1.11	4.71 ± 1.78	2.37 ± 0.00	2.92 ± 0.01	4.61 ± 6.05
Elevation	$\mathbf{0.15 \pm 0.00}$	$\mathbf{0.15 \pm 0.01}$	0.80 ± 0.05	1.11 ± 0.06	1.11 ± 0.01	1.17 ± 0.00	0.98 ± 0.01
Classification	% Accuracy ↑						
iNaturalist	$\mathbf{66.03 \pm 0.54}$	$\mathbf{65.98 \pm 0.61}$	56.73 ± 0.83	60.47 ± 0.56	58.78 ± 0.48	53.27 ± 0.78	56.73 ± 0.80
Biome	$\mathbf{94.41 \pm 0.14}$	$\mathbf{94.27 \pm 0.15}$	75.81 ± 1.53	73.18 ± 5.58	69.69 ± 0.06	68.29 ± 0.11	79.61 ± 0.42
Ecoregions	$\mathbf{91.67 \pm 0.15}$	$\mathbf{91.61 \pm 0.22}$	76.87 ± 1.27	78.43 ± 1.71	68.46 ± 0.06	67.26 ± 0.02	70.48 ± 0.21

3.1 Downstream Task Performance

Figure 2a plots principal component latent representations of different embeddings, highlighting that SatCLIP has global coverage over land masses, as opposed to GPS2Vec and MOSAIKS. The full predictive results from our experiments can be found in Table 1. SatCLIP is consistently the best performing pretrained location encoder. Interstingly, SatCLIP embeddings are more informative for iNat classification than a location encoder pretrained on iNat (CSP-iNat). We believe that this is due to SatCLIP embeddings being able to provide auxiliary information not contained within the iNat imagery. A second experiment test the ability of SatCLIP embeddings to overcome geographic distribution shift for Ecoregion prediction: Here, we hold out the whole continent of Africa as test set, adding only a tiny portion (1%) of points to the training set to effectively create a few-shot geographic adaptation problem. Figure 2b shows how SatCLIP is the only location encoder up to this task. This is primarily because existing location encoders are not trained on globally distributed data or don't possess the expressiveness of Sentinel-2 imagery. Overall, the results confirm that SatCLIP models trained on SK-100K data provide meaningful features to help with prediction in climate-relevant predictive modeling.

4 Discussion and Conclusion

We present SatCLIP, a location encoder model trained via a geographically-aware self-supervised contrastive learning loss. SatCLIP is pretrained using *global*, publicly available Sentinel-2 satellite images, achieves high predictive performance on climate-related geospatial tasks and can help overcome geographic distribution shift. The SatCLIP model we build and evaluate here is an instantiation of a broader paradigm of contrastively pretrained location encoders, which offers many possible modifications and adaptations–in terms of both the data and models used. This paradigm extends naturally to using varied data sources and modalities as input to a context encoder. For example, adding administrative data as layers appended to multi-spectral satellite image bands might more comprehensively represent different geospatial phenomena. On the modeling side, advancements might focus on designing model architectures to achieve joint compatibility of location and context encoders for more efficient learning.

This work towards general-purpose pre-trained location encoders has the potential to coalesce global data into a succinct representation of any location. To this end, *global, general-purpose location encoders* can be useful for practitioners that want to leverage the spatial variability trends of remotely sensed data in their modeling problem, but cannot easily access this data, e.g. due to resource constraints (in a similar way to [11]). SatCLIP can help in exactly this case: embeddings capturing a locations characteristics can be queried at any given location on the planet for use in downstream applications. We make our S2-100K dataset and pretrained location encoders available at `github.com/microsoft/satclip`.

Acknowledgements

Esther Rolf is supported by the Harvard Data Science Initiative and and the Center for Research on Computation and Society.

References

[1] Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. pp. 6172–6180, 2018. URL https://www.digitalglobe.com/resources/.

[2] 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. The iNaturalist species classification and detection dataset. pp. 8769–8778, 2018. ISBN 5,986561,767. URL www.inaturalist.org.

[3] Klemmer, K., Safir, N. S., and Neill, D. B. Positional encoder graph neural networks for geographic data. pp. 1379–1389. PMLR, 4 2023. URL https://proceedings.mlr.press/v206/klemmer23a.html.

[4] Lobell, D. B., Thau, D., Seifert, C., Engle, E., and Little, B. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164:324–333, 2015.

[5] Mac Aodha, O., Cole, E., and Perona, P. Presence-only geographical priors for fine-grained image classification. In *ICCV*, October 2019.

[6] Mai, G., Janowicz, K., Cai, L., Zhu, R., Regalia, B., Yan, B., Shi, M., and Lao, N. Sekge: A location-aware knowledge graph embedding model for geographic question answering and spatial semantic lifting. *Transactions in GIS*, 24:623–655, 6 2020. ISSN 1467-9671. doi: 10.1111/TGIS.12629. URL https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12629https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12629https://onlinelibrary.wiley.com/doi/10.1111/tgis.12629.

[7] Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., and Lao, N. Multi-scale representation learning for spatial feature distributions using grid cells. 2 2020. URL http://arxiv.org/abs/2003.00824.

[8] Mai, G., Lao, N., He, Y., Song, J., and Ermon, S. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. 5 2023. URL https://arxiv.org/abs/2305.01118v2.

[9] Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., and Athanasiadis, I. N. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021.

[10] 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. Learning transferable visual models from natural language supervision. pp. 8748–8763. PMLR, 7 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[11] Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature Communications 2021 12:1*, 12:1–11, 7 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-24638-z. URL https://www.nature.com/articles/s41467-021-24638-z.

[12] Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., and Tuia, D. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *Under submission.*, 2023.

[13] Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., and Wetzstein, G. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33: 7462–7473, 2020.

[14] Thomee, B., Elizalde, B., Shamma, D. A., Ni, K., Friedland, G., Poland, D., Borth, D., , and Li, L. J. YFCC100M. *Communications of the ACM*, 59:64–73, 1 2016. ISSN 15577317. doi: 10.1145/2812802. URL https://dl.acm.org/doi/10.1145/2812802.

[15] Wang, Y., Ait, N., Braham, A., Xiong, Z., Liu, C., Albrecht, C. M., and Zhu, X. X. SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in Earth observation. 11 2022. URL https://arxiv.org/abs/2211.07044v2.

[16] Wang, Y., Albrecht, C. M., Braham, N. A. A., Mou, L., and Zhu, X. X. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10:213–247, 6 2022. ISSN 21686831. doi: 10.1109/MGRS.2022.3198244. URL https://arxiv.org/abs/2206.13188v2.

[17] Yin, Y., Liu, Z., Zhang, Y., Wang, S., Shah, R. R., and Zimmermann, R. GPS2Vec: Towards generating worldwide GPS embeddings. pp. 416–419. Association for Computing Machinery, 11 2019. ISBN 9781450369091. doi: 10.1145/3347146.3359067. URL https://dl.acm.org/doi/10.1145/3347146.3359067.