USING MULTIPLE INPUT MODALITIES CAN IMPROVE DATA-EFFICIENCY FOR ML WITH SATELLITE IMAGERY

Arjun Rao
Department of Computer Science
University of Colorado Boulder
raoarjun@colorado.edu

Esther Rolf
Department of Computer Science
University of Colorado Boulder
esther.rolf@colorado.edu

ABSTRACT

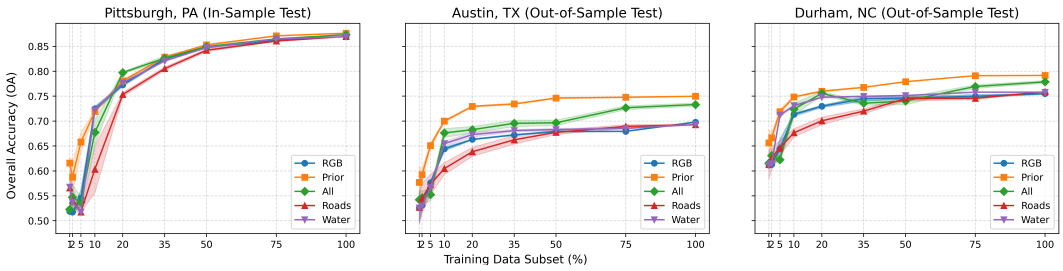
A large corpus of diverse geospatial data layers is available around the world ranging from remotely-sensed raster data like satellite imagery, digital elevation maps, predicted land cover maps, and human-annotated data, to data derived from environmental sensors such as air temperature or wind speed data. A large majority of geospatial machine learning **(GeoML)** models, however, are designed primarily for *optical* input modalities such as multi-spectral satellite imagery. We show improved GeoML model performance on three benchmark datasets spanning classification and segmentation tasks when geospatial inputs are fused with optical input imagery. Benefits are largest in settings where labeled data are limited and in geographic out-of-sample settings, suggesting that multi-modal inputs may be especially valuable for data-efficiency and out-of-sample performance of GeoML models.

1 INTRODUCTION

Users of GeoML systems need agile models that can integrate the vast arrays of publicly available geographic data into a cohesive representation of the world, allowing for accurate predictions even with limited training data, or when faced with covariate shift across time, space, spectrum, and scale (Rolf et al., 2024). However, a large subset of GeoML models have historically *underutilized* these geographic and analytic inputs in the training stage. **In this work, we study the label-efficiency and out-of-sample generalization capability associated with adding non-optical, contextual inputs to commonly used GeoML architectures.**

While including additional input layers is clearly likely to increase performance for in-sample prediction with ample training data, the effects of adding additional input layers in settings with limited label data or out-of-sample deployment distributions are less clear. Additional geographic inputs can inform a GeoML model with structural information that may allow the model to learn geospatial image representations with fewer labeled training samples (label-efficiency). They can also require more complex (data-hungry) models to represent the various modalities of data. Additional inputs could help GeoML models generalize across regions. Or, they could cause models to overfit to local patterns that only manifest in-sample, which could then decrease performance. While prior work established that additional inputs can increase GeoML model performance in a variety of settings, they did not study how this benefit differs across (i) amount of label data, and (ii) in-sample and out-of-sample prediction settings. We aim to study how, in general, fusing geographical, contextual clues affects model performance in these important settings.

Prior Work. Adding a non-optical context to machine learning models trained on geospatial imagery has been performed extensively in prior work. Tang et al. (2015) and Chu et al. (2019) add GPS coordinates from the Yahoo Flickr Creative Commons dataset containing 100 million geotagged images in the natural domain. Location embeddings are concatenated as additional features to the final embedding from a Convolutional Neural Network (CNN). Mac Aodha et al. (2019) performed fine-grained image classification with a location, time, and photographer prior to differentiate between similar classes that are spatially disparate. Benson et al. (2024); Wang et al. (2020) use geospatial data layers for task-specific downstream applications such as vegetation health prediction and learning embeddings of neighborhood embedding-learning. Recently, Nedungadi et al.



(a) Data-efficiency of 5-layer FCN trained with geographic input layers on the EnviroAtlas dataset. Shaded regions report standard error over 5 random seeds.

Subset (%)	Pittsburgh			Austin			Durham		
	RGB	Prior	All	RGB	Prior	All	RGB	Prior	All
1%	0.51	**0.61**	0.52	0.53 ± 0.03	**0.58 ± 0.03**	0.54 ± 0.02	0.61 ± 0.00	**0.66 ± 0.03**	0.62 ± 0.00
2%	0.51	**0.58**	0.54	0.53 ± 0.01	**0.59 ± 0.01**	0.55 ± 0.01	0.61 ± 0.00	**0.67 ± 0.01**	0.63 ± 0.01
5%	0.54	**0.65**	0.55	0.58 ± 0.00	**0.65 ± 0.01**	0.55 ± 0.01	0.64 ± 0.02	**0.72 ± 0.01**	0.62 ± 0.01

(b) Performance of EnviroAtlas prior, `all` versus RGB input with 1%, 2%, and 5% of input training data.

Figure 1: **Performance and data-efficiency of a Fully-Convolutional Network on the EnviroAtlas Land Cover Segmentation Dataset.** Test subsets in Austin and Durham are out-of-sample test splits. Modality "All" refers to a 4-band data product with stacked and geo-referenced road, building, waterway, and water body footprints. Results averaged over 10 random seeds. $1\times$ standard error of Pittsburgh reported $\leq 1e-3$ over 10 random seeds.

(2024) introduce large, multi-modal pre-training datasets built with Sentinel-2 imagery that contains several geographic modalities like the ESA WorldCover dataset (Zanaga et al., 2022) and Digital Elevation Maps. Although MMEarth (Nedungadi et al., 2024) is pre-trained on these modalities, it is only used to predict the modalities given a Sentinel-2 RGB image as input; nonetheless, they find that pretraining on multiple modalities has large gains for data-efficiency.

2 METHODS

Our experiments measure performance of models trained with just multispectral input and with additional geographic inputs. To measure data-efficiency, we train models on subsets (simple random samples) of the input dataset, where samples vary across random seeds during training. We conduct our experiments on three benchmark datasets in ML for remote sensing.

The **BigEarthNetv2.0** dataset (Sumbul et al., 2019; Clasen et al., 2024) is a multi-label classification task that consists of approximately 550,000 pairs of Sentinel-2 image patches, paired with ground labels of over 19 land cover classes. Our models input 10 Sentinel-2 bands to ensure consistency with benchmark results reported in Clasen et al. (2024). Second, the **EnviroAtlas** dataset (Pickard et al., 2015) consists of high-resolution (1m GSD) land cover maps derived from NAIP imagery; we use the 5-class land cover segmentation task established in Rolf et al. (2022). Third, we use the farmland parcel delineation dataset introduced as part of the **SustainBench** suite of datasets (Yeh et al., 2021) – particularly the field boundary segmentation task benchmarked in Aung et al. (2020).

We choose the model architectures and additional geographic layers that make sense for each task. For the two segmentation tasks, we fuse the original inputs (NAIP aerial imagery for EnviroAtlas and Sentinel-2 RGB layers for SustainBench-Field-Delineation) with roads, waterways, and waterbodies data from the Open Street Maps (OSM) repository (Haklay & Weber, 2008), following the methodology in (Rolf et al., 2022). For the EnviroAtlas task, we use the pre-fused *Prior* layer from Rolf et al. (2022), constructed to represent an uncertain probability distribution on the class values for each pixel. Since all of these auxiliary layers are raster data, we simply stack them as additional channels to the models during training and testing. To be comparable to previous benchmark results, we use a fully convolutional network for the EnviroAtlas Dataset, and a U-Net (Ronneberger et al., 2015) for the SustainBench-field-delineation dataset.

Subset (%)	w/ SatCLIP Aux. Token				Vanilla ViT				SatCLIP (F1)
	Vit-B Avg Prec	Vit-B F1	Vit-S Avg Prec	Vit-S F1	Vit-B Avg Prec	Vit-B F1	Vit-S Avg Prec	Vit-S F1	
1%	**46.3**	**36.1**	40.45	23.27 ± 1.27	44.6	32.1	39.78	22.95 ± 1.31	15.9
2%	**55.6**	**45.9**	47.96	33.82 ± 1.10	51.1	40.2	45.60	34.11	14.1
5%	**62.7**	**54.1**	59.98	47.84 ± 2.08	58.9	50.2	56.07	44.05 ± 1.13	10.1
20%	**66.8**	**60.6**	66.4	58.3	64.5	58.1	64.2	57.6	12.5
50%	**70.1**	**64.7**	70.1	64.3	68.7	63.5	69.2	63.7	21.7
100%	70.3	65.2	**70.8**	**65.4**	69.5	64.1	70.1	64.5	23.2

Table 1: **Average Precision (Macro)/Multi-Label F1 score with Frozen (`F`) vs Register vs Fine-tuned (`FT`) SatCLIP auxiliary token on the BigEarthNetv2.0 test split with a location encoder.** Results averaged over five random seeds. Unless specified, all results report $\leq 0.1\%$ standard deviation.

For the BigEarthNetv2.0 image-level multi-label classification task we use vision transformer (ViT, Vit-B/8, Vit-S/8) architectures. To the Sentinel-2 input, we fuse general purpose global SatCLIP location embeddings (Klemmer et al., 2023), which distill socioeconomic and environmental signals in satellite imagery into a pretrained location encoder $g(\text{lat,lon})$ with output dimension 256. Embeddings from SatCLIP's location encoder are passed as an auxiliary token to the ViT's encoder along with image tokens. We add a linear layer to SatCLIP's location encoder that maps the 256-dimensional SatCLIP embeddings to the desired sequence length expected by the Vit-S/ViT-B. The auxiliary SatCLIP token is assigned a positional encoding of $N + 1$ where N is the total number of encoder tokens excluding the classification token. For our main experiments, the parameters within the SatCLIP model $g(\text{lat,lon})$ are frozen; we experiment with unfreezing these weights in Figure 3.

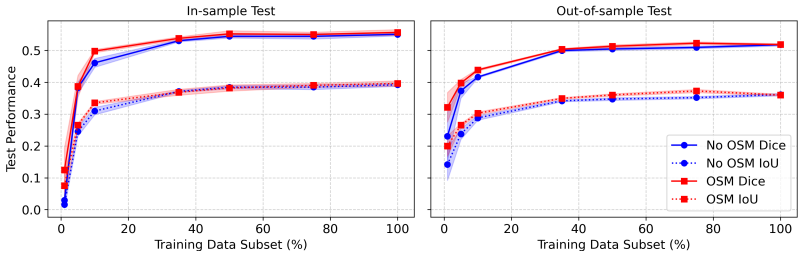
To test performance across in-domain (ID) and out-of-domain (OOD), we use a combination of the existing train/test splits in each dataset and add data splits when needed. Since the BigEarthNetv2.0 uses a geographically buffered train/val/test split, the benchmark results already test OOD performance to some degree. The EnviroAtlas task from Rolf et al. (2022) tests both ID performance and OOD performance, since the training set is restricted to Pittsburgh, PA, but there are test sets in Pittsburgh (ID), Durham, NC (OOD), and Austin, TX (OOD). For the SustainBench task, we use the benchmark ID test split, and generate a new spatially buffered split (see Appendix Figure 4) to test OOD performance.

3 Results and Discussion

We found that adding contextual, geographic inputs improves model performance, with largest gains in settings with limited label data across all 3 tasks. In table 1, a vision transformer trained with an auxiliary SatCLIP token (ResNet18, $L = 10$) improves multi-label classification average precision on the BigEarthNetv2.0 dataset, on average, by 3.3% when trained on between 1 to 5% of the training dataset. The token-aided ViT also exhibits a 4.5% improvement in multi-label F1 score. Note that with increasing training data, the improvement in average precision and F1 score drops to 1.6% , highlighting the label-efficiency associated with passing an auxiliary multi-modal token to the ViT. We observe similar improvements in data efficiency when a FCN is trained on EnviroAtlas's land-cover classification dataset (Figure 1).

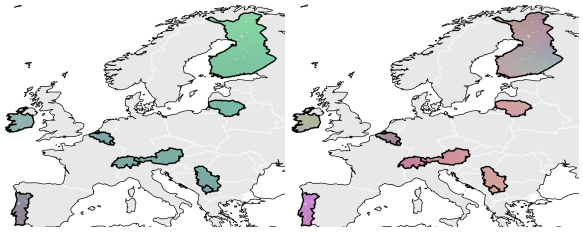
We also found that fusing additional geographic input layers to remotely sensed imagery can significantly aid geographic domain generalization. While the value of additional input layers is clear in the low-label regime for all test cities in the EnviroAtlas dataset, Figure 1 shows an improvement in overall test accuracy in out-of-distribution test cities (Austin and Durham) across all amounts of training data. From Figure 2, we find the performance improvements with low data holds over our proposed spatially-buffered test split of the SustainBench-field-delineation dataset. A U-Net trained with an OSM raster layer exhibits a 5% dice-score improvement in-sample and a 4.6% improvement out-of-sample when trained on between 1-10$\%$ of training data. Note that the test split in the BigEarthnetv2.0 dataset is also spatially buffered from the training data.

To determine if arbitrary tokens aid GeoML model performance on the BigEarthNet multi-label classification task, we allow for the pre-initialized SatCLIP model $g(\text{lat,lon})$ to be trainable as a token —



Subset (%)	In-sample Split		Out-of-sample Split	
	RGB	RGB + OSM	RGB	RGB + OSM
1%	2.96	**12.47**	23.07	**32.17**
5%	38.27	**39.72**	37.30	**39.77**
10%	46.12	**49.21**	41.64	**43.88**
35%	53.05	**53.91**	49.98	**51.02**

Figure 2: **Performance and label-efficiency of a U-Net trained on SustainBench's Farmland Parcel Delineation Dataset.** Label efficiency and out-of-sample performance reported as Dice and IoU scores averaged over 5 random seeds. Bottom: Test dice score for subsets 1% to 35%.



Sub%	F SatCLIP	FT SatCLIP
1%	**46.3/36.1**	45.4/34.7
2%	**55.6/45.9**	53.2/42.8
5%	62.7/54.1	**63.5/56.2**
20%	**66.8/60.6**	65.3/59.1
50%	**70.1/64.7**	67.1/60.1
100%	**70.3/65.2**	66.0/59.1

Figure 3: **Frozen (F) vs Fine-Tuned (FT) SatCLIP auxiliary ViT token on the BigEarthNetv2.0 land-cover classification task:** Maps: PCA embeddings of the SatCLIP tokens: frozen (left) vs finetuned (right). Table: Label-efficiency results on the BigEarthNetv2.0 classification task. Average precision/multi-label F1 score (macro-averaged) reported.

effectively, finetuning the SatCLIP token on an image-classification pseudo-task. From Figure 3, we find that the label-efficiency and out-of-sample performance are worse when the SatCLIP weights are not frozen during training. Figure 3 shows that the fine-tuning SatCLIP tokens leads to embeddings that are highly localized within various countries covered by the BigEarthNetv2.0 dataset. It is likely that this results in the augmented ViT overfitting to the auxiliary SatCLIP token leading to lower test set performance.

4 Conclusion

Improvements in label-efficiency and out-of-distribution GeoML performance directly translate to applicability for real-world, downstream applications in climate and ecological monitoring. Here, we found that fusing additional geographic input layers into GeoML models resulted in better label-efficiency and out-of-sample performance, compared to models that only used multispectral input. To get these improvements, we used simple and effective multi-modal geospatial data layer fusion methodologies – input stacking for raster layers, and adding auxiliary multi-modal tokens to a ViT for location embeddings. Given the simplicity of the architectures used in this work (5-layer FCN and Vanilla ViT-B), we hypothesize that the results presented are merely a lower bound in data-efficiency improvements with multi-modal inputs.

These experimental settings can be extended to several benchmark datasets with minimal modifications to the training procedure. Auxiliary tokens, for example, can be generated from any input modality source and fused with a ViT's image tokens. Future work will also perform a comprehensive analysis of interpretability of models augmented with auxiliary geographic tokens and study which and when additional geographic inputs improve data efficiency and out-of-distribution performance of GeoML models.

REFERENCES

Han Lin Aung, Burak Uzkent, Marshall Burke, David Lobell, and Stefano Ermon. Farm parcel delineation using spatio-temporal convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 76–77, 2020.

Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27788–27799, 2024.

Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Kai Norman Clasen, Leonard Hackel, Tom Burgert, Gencer Sumbul, Begüm Demir, and Volker Markl. reben: Refined bigearthnet dataset for remote sensing image analysis. *arXiv preprint arXiv:2407.03653*, 2024.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. *arXiv preprint arXiv:2311.17179*, 2023.

Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning, 2024. URL https://arxiv.org/abs/2405.02771.

Brian R Pickard, Jessica Daniel, Megan Mehaffey, Laura E Jackson, and Anne Neale. Enviroatlas: A new geospatial tool to foster ecosystem services science and resource management. *Ecosystem Services*, 14:45–55, 2015.

Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4088–4099, 2023.

Esther Rolf, Nikolay Malkin, Alexandros Graikos, Ana Jojic, Caleb Robinson, and Nebojsa Jojic. Resolving label uncertainty with implicit posterior models. *arXiv preprint arXiv:2202.14000*, 2022.

Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Position: Mission critical–satellite data is a distinct modality in machine learning. In *Forty-first International Conference on Machine Learning*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5901–5904. IEEE, 2019.

Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pp. 1008–1016, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomex, and Łukasz Kaiser. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Zhecheng Wang, Jinhua Yu, Ziwei Wu, Rui Zhang, Jiuchuan Mao, Liang Li, Zhiyong Feng, and Jun Yin. Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2068–2076. ACM, 2020.

Christopher Yeh, Chenlin Meng, Sherrie Wang, Anne Driscoll, Erik Rozi, Patrick Liu, Jihyeon Lee, Marshall Burke, David B. Lobell, and Stefano Ermon. 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)*, 2021. URL `https://openreview.net/forum?id=5HR3vCylqD`.

Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, et al. Esa worldcover 10 m 2021 v200. 2022.

A Experimental Setup

FCN on EnviroAtlas Land Cover Segmentation: We train on EnviroAtlas' train images in Pittsburgh, PA on a 5-layer Fully Convolutional Network with 64 filters and an output smoothing of 10^{-2}. A batch size of 128 and a learning rate of $1e-3$ is fixed across all training data subsets and random seeds reported in Figure 1. We fix the lower bound learning rate to $1e-7$. Table 3 reports the number of training epochs each data-efficient FCN is trained on. Note that FCNs trained on 1% of EnviroAtlas' training data for 700 epochs trigger our early-stopping logic between epoch 200-300. We use TorchGeo's `RandomGeoSampler` with an input image size of 128. Our test dataset uses TorchGeo's `GridGeoSampler` with an input image size of 256 and a stride of 512 to avoid overlapping image patches. Our multi-modal inputs include a Road, water, waterway, and waterbody footprint from (Haklay & Weber, 2008).

ViT on BigEarthNet Multi-label Classification:

Vision Transformers (ViTs) (Dosovitskiy (2020)) utilize the transformer architecture proposed in (Vaswani et al., 2017) for image classification tasks. Input images are decomposed into a sequence of small, non-overlapping patches $X = (x_1, \ldots, x_N)$ where N denotes the total number of input patches. N Input patches are mapped to embeddings (tokens) with a linear-layer projection. For multi-label classification with the BigEarthNet dataset (Sumbul et al., 2019), we add an additional classification token denoted by `[CLS]` with a positional encoding of 0, and an auxiliary token containing SatCLIP location encoder embeddings with a positional encoding of $N+1$. A pre-trained SatCLIP location encoder efficiently summarizes terrain and geographic information by encoding visual similarities of spatially distant environments. `[CLS]` is a learnable additional token introduced to capture label information. Unlike (Cong et al., 2022; Reed et al., 2023) that use various versions of sinusoidal positional encodings that are sensitive to Ground Sampling Distance (GSD) and temporal information, we augment image patches $X = (x_1, \ldots, x_N)$ with scalar positional encodings $(1, 2, \ldots, N)$. We do not observe a significant performance improvement by using sinusoidal positional encodings commonly used in Masked Autoencoders (MAEs), and attribute this to the large spatial redundancies present in most satellite-image datasets.

SustainBench Field Boundary Segmentation The SustainBench suite of benchmarks proposed in Yeh et al. (2021) contains a collection of 15 benchmark tasks in machine learning for remote sensing spanning 7 United Nations' sustainable development goals (SDGs). We use SustainBench's

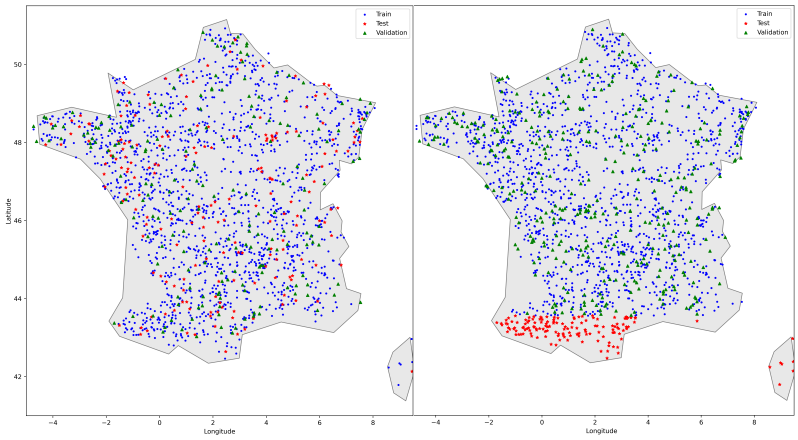


Figure 4: **Proposed train-validation-test split of the SustainBench-field-delineation dataset.** Original train-validation-test split (left) and proposed spatially-buffered split proposed (right) used in Figure 2

field-delineation task which targets the "No Hunger" SDG. SustainBench-Field-Delination consists of Sentinel-2 imagery in France in 2017. Each input image is at a 10m ground-sampling distance and has a size of 224×224 pixels corresponding to an approximately 5 kilometer surface area covered per image. We train a U-Net (Ronneberger et al., 2015) on simple random samples of the original training dataset and a spatially-buffered test split shown in Figure 4. The spatially-buffered test split is created to allow an in-sample train and validation set, and a test split that only spans southern France. This is implemented by creating a latitude threshold that is modified to collect a test-dataset size comparable to the original test-dataset size in Yeh et al. (2021). A brief algorithm that creates the spatially buffered test split is detailed in algorithm 1. We pull rasters of roads, water, waterbodies, and buildings from OSM (Haklay & Weber, 2008) given the input image's bounding co-ordinates. We train a U-Net with an identical learning rate of $1e-4$ for 20 epochs scaled linearly according to Table 3 (Note: Subset size of 1 is trained for 20 epochs which implies subset size of 0.01 is trained for 2000 epochs). We report test dice and IoU segmentation scores for the field boundary segmentation task based on the best in-sample validation performance. Dice score (or F1 score) is defined as $Dice = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$, and the Intersection over Union (IoU) score.

Algorithm 1 Create spatially-buffered train-validation-test split for SustainBench-Field-Delineation dataset

Input: Train, Validation, Test CSV Files with (lat, lon) co-ordinates per image, random seed.
Output: New CSV files for train-validation-test.
$D \leftarrow$ COMBINE(train, val, test)
for all Image $s \in D$ **do**
    $lat(s) \leftarrow$ COMPUTECENTERLATITUDE(s)
end for
Determine threshold T^* so that $\texttt{Count}\{s \in D : lat(s) < T^*\}$ meets the desired test size.
$Test \leftarrow \{s \in D \mid lat(s) < T^*\}$
$Pool \leftarrow D \setminus Test$
$(Train, Val) \leftarrow$ RANDOMSPLIT($Pool$, train-validation-ratio, random seed)
SAVE DATASET(Train, Val, Test)

Subset Size	Training Epochs
100%	7
75%	9
50%	14
35%	20
20%	35
10%	70
5%	140
2%	350
1%	700

Table 3: Training epochs scaled by subset size for all label-efficiency experiments.