Using Multiple Input Modalities Can Improve Data-Efficiency and O.O.D. Generalization for ML with Satellite Imagery

Arjun Rao [1] **Esther Rolf** [1]

Abstract

A large variety of geospatial data layers is available around the world ranging from remotely-sensed raster data like satellite imagery, digital elevation models, 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 machine learning models trained on satellite imagery **(SatML)**, however, are designed primarily for *optical* input modalities such as multi-spectral satellite imagery. To better understand the value of using other input modalities alongside optical imagery in supervised learning settings, we generate augmented versions of SatML benchmark tasks by appending additional geographic data layers to datasets spanning classification, regression, and segmentation. Using these augmented datasets, we find that fusing additional geographic inputs with optical imagery can significantly improve SatML model performance. 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 SatML models. Surprisingly, we find that hard-coded fusion strategies outperform learned variants, with interesting implications for future work.

1. Introduction

SatML models that effectively leverage the volume and diversity of data from Earth Observation (EO) satellites have the potential to translate petabyte-scale raw data into data-driven insights. Users of SatML systems need models that can integrate these 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 shifts across time, space, spectrum, and scale (Rolf et al., 2024).

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* and *out-of-sample deployment* distributions are less clear. Additional geographic inputs could inform a SatML model with structural information that may allow the model to learn geospatial image representations with fewer labeled training samples (label-efficiency); they could also require more complex (data-hungry) models to represent the various modalities of data. Additional inputs could help SatML models generalize across regions; they could also cause models to overfit to local patterns that only manifest in-sample, which could then decrease performance.

In this work, we study the label-efficiency and out-of-sample generalization capability associated with adding non-optical, contextual inputs to commonly used SatML architectures.

As outlined in Roscher et al. (2024), data-centric learning is a systematic method of algorithmic evaluation where the primary focus involves curating diverse, complete, unbiased, and relevant data for optimal model performance. We perform a *data-centric* study on the benefits and nuances of leveraging these widely available geographic input layers, complementing previous lines of model-centric research that study how to utilize multi-modal inputs for a fixed training/pretraining strategy and/or model architecture.

Our primary findings in this work are: (1) We show improvements in label-efficiency when multi-modal, auxiliary geographic inputs are fused with optical imagery on 3 SatML task-types: Multi-label land-cover classification, land cover segmentation, and tree-cover regression. (2) We find that these auxiliary geographic inputs are especially helpful when SatML models are evaluated

[1]Department of Computer Science, University of Colorado Boulder. Correspondence to: Arjun Rao <raoarjun@colorado.edu>.

Accepted to TerraBytes: Towards global datasets and models for Earth Observation Workshop at the 42nd International Conference on Machine Learning, Vancouver, Canada.

OOD through results on the spatially buffered test split of the BigEarthNetv2.0 dataset (Clasen et al., 2024), and the OOD test cities of the EnviroAtlas dataset in Austin, TX, and Durham, NC (Rolf et al., 2022). (3) Through our ablations, we find surprising results that show the ineffectiveness of finetuning SatML models arbitrarily on common benchmark tasks with these auxiliary geographic inputs.

Our contributions also include a large-scale, multi-dataset release containing modified versions of the SustainBench farmland boundary delineation dataset (Yeh et al., 2021), and the USAVars tree-cover regression dataset (Rolf et al., 2021) with additional geographic inputs georeferenced to the optical imagery. Additionally, we release the BigEarthNetv2.0 dataset (Clasen et al., 2024) with pre-computed patch-embeddings with the SatCLIP location encoder (Klemmer et al., 2025). A full list of contributed data products is shown in column "Additional Data Layers" in Table 1.

2. Prior Work

2.1. Multi-Modal SatML

Adding a non-optical context to machine learning models trained on geospatial imagery has been performed extensively in prior work. Tang et al. (2015) extracts GPS features from the Yahoo Flickr Creative Commons 100M dataset, and fuses embeddings of location information with final embeddings from a convolution-based image network. Chu et al. (2019) incorporates geolocation information into fine-grained image classification through the use of geolocation priors, introducing the computer vision community to geo-aware neural networks. 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) add a contextual input to predict future vegetation state given temporally rich satellite imagery and future weather information. Wang et al. (2020) propose an unsupervised multi-modal framework which incorporates both street view imagery and point-of-interest data to learn neighborhood embeddings in urban areas. Johnson et al. (2022); Fonte et al. (2020); Patriarca et al. (2019) introduce large-scale Sentinel-2 datasets georeferenced with OpenStreetMap (OSM) rasters (Haklay & Weber, 2008) converted to be used as a land-use-land-cover map (LULC). However, these methods, which utilize geographic data layers publicly available, intend for their usage to be restricted as ground-truth masks for land-cover classification problems.

Recently, Nedungadi et al. (2024) introduce large, multi-modal pre-training datasets built with Sentinel-2 imagery that contain several geographic modalities like ESA WorldCover (Zanaga et al., 2022) and Digital Elevation Model. 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 data-efficiency improvements when their self-supervised models are linear-probed on various downstream classification tasks. Sosa et al. (2025) utilize the Aster-DEM and the ESA-Worldcover raster produced by Nedungadi et al. (2024) as additional input to a masked autoencoder (MAE). However, a bulk of their experiments is performed with various permutations of Sentinel-2-derived multispectral modalities.

2.2. Token Fusion

Studies on Vision Transformers (ViTs) have explored the use of additional tokens to improve performance and capture more nuanced information. In Dosovitskiy et al. (2021), a *class token* (`[CLS]`) was introduced and appended to the patch embeddings, enabling the model to learn a global representation useful for classification tasks. Touvron et al. (2021) introduce a *distillation token* to facilitate knowledge transfer from a teacher model, boosting accuracy without substantially increasing computational cost. Jia et al. (2022); Wang et al. (2024) demonstrate that injecting a small set of learnable prompts into the early layers of pre-trained ViTs can effectively adapt them to new downstream tasks. Darcet et al. (2024) highlights the importance of internal "registers" in ViT architectures, arguing that specialized design choices can better accommodate these additional tokens for more robust representations.

3. Methods

Our experiments measure performance of models trained with just multi-spectral inputs and with additional geographic inputs. We consider three different fusion mechanisms that allow for SatML models to learn from these geographic auxiliary inputs. We then describe the model architectures used for each fusion mechanism, and the benchmark datasets that we train our models on.

3.1. Geographic Data Fusion

Figure 1 contains an overview of the proposed geographic input-fusion techniques used in this work. For land cover segmentation with the EnviroAtlas dataset, we fuse the original inputs (NAIP aerial imagery) with roads, waterways, and waterbody data from the OSM repository (Haklay & Weber, 2008) using the fusion method `STACK`. We compute the hand-crafted prior for the training split in Pittsburgh, and test splits in Austin and Durham using the methodology proposed in Rolf et al. (2022). The generation of the prior is denoted by $f(\cdot)$ in Figure 1, and is described in detail in appendix Appendix A.1. The resulting

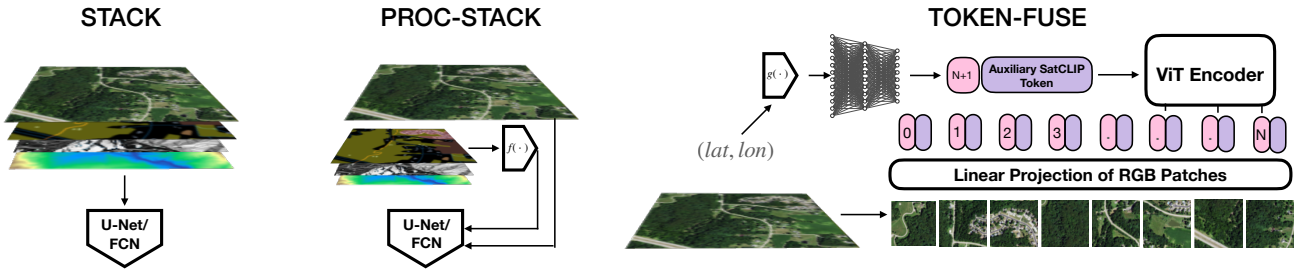

Figure 1. **Geographic data input fusion mechanisms used in this work:** `STACK` involves concatenating one or more geographic raster inputs with the optical input before passed jointly as an input to a convolution-based architecture. `PROC-STACK` passes the geographic input to a function $f(\cdot)$ before stacking the geographic data with the optical input. `TOKEN-FUSE` passes a latitude-longitude pair to a location encoder $g(\cdot)$ and uses location embeddings as an auxiliary token to a Vision Transformer (ViT). Experiments in Section 4.1 and Section 4.2 use frozen models for f and g; ablation experiments in Section 4.3 use trainable models.

prior along with the raw geographic data layers are used as input to the prior function and are fused to the SatML. The generation of the prior followed by fusion with the optical input forms our fusion method `PROC-STACK`.

For the farmland-parcel delineation task with the SustainBench dataset, and the socioeconomic regression task with the USAVars dataset, we use OSM raster layers that contain all the geographic data layers used for the EnviroAtlas dataset, with the addition of several new land-use and land-cover classes that are roughly relevant to the task. These additional raster layers include high-level biome information such as forests, wetlands, or urban-type terrain. Output Geodataframes are pre-processed to RGB space. We apply a smoothing kernel ($\sigma = 1.0$) to remove sharp edges and features from the API response. A complete list of raster inputs queried for the USAVars dataset is detailed in appendix Figure 9. Additionally, we pull a digital elevation map (DEM) from the Continental Europe Digital Terrain Model available as part of the OpenTopography API. The DEM raster, originally available at a 20m GSD, is resized to the Sentinel-2 RGB spatial resolution of 10m/px. Unlike the OSM rasters, the DEM is passed as raw input with fusion mechanism `STACK`.

To be comparable to previous benchmark results, we use a fully convolutional network (FCN) for the EnviroAtlas Rolf et al. (2022) Dataset, a U-Net (Ronneberger et al., 2015b) for the SustainBench-field-delineation dataset Yeh et al. (2021), and a ResNet50 He et al. (2015) for the regression task proposed in the USAVars dataset Rolf et al. (2021).

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., 2025), 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 8 and Section 4.3.

3.2. Models

Convolutional Architectures: In this work, we use simple, widely-used convolutional neural networks when trained on data fused with fusion mechanisms `STACK` and `PROC-STACK`. We choose simple architectures over specialized SatML model architectures because we are primarily interested in comparing different data settings and fusion strategies. We choose models to be consistent with model architectures used in prior work. For experiments on the EnviroAtlas dataset, we use a 5-layer FCN. For segmentation on the SustainBench field-boundary delineation, we use a U-Net (Ronneberger et al., 2015b) with identical architectural setup and hyperparameters as Aung et al. (2020) to allow for consistency when comparing results. For regression on the USAVars tree-cover dataset, we use a vanilla ResNet50 (He et al., 2015) with randomly initialized weights.

Vision Transformers (ViTs): Vision Transformers (ViTs) (Dosovitskiy et al., 2021) utilize the transformer architecture proposed in (Vaswani et al., 2017). Input images are decomposed into a sequence of small, non-overlapping patches which are mapped to embeddings (tokens) with a linear-layer projection. Unlike (Cong et al., 2022; Reed

Dataset	Task Description	Multispectral Input	Model	Additional Data Layers	OOD?
SustainBench (Yeh et al., 2021)	Farmland boundary delineation	Sentinel-2 RGB	U-Net	OSM rasters†, EU-DEM†	✗
EnviroAtlas (Rolf et al., 2022)	Land-cover segmentation	NAIP RGB + NIR	FCN	Prior (Rolf et al., 2022), OSM rasters	✓
BigEarthNetv2.0 (Clasen et al., 2024)	Land-cover classification	Sentinel-2 (10 bands)	ViT	SatCLIP (Klemmer et al., 2025) embeddings†	✓
USAVars (Rolf et al., 2021)	Tree-cover regression	NAIP RGB + NIR	ResNet-50	OSM rasters†	✗

Table 1. **Experimental framework and source tasks used in this work:** We test fusion mechanisms `STACK` and `STACK-PROC` on the EnviroAtlas (Rolf et al., 2022), SustainBench (Yeh et al., 2021), and the USAVars (Rolf et al., 2021) benchmark datasets. We test fusion mechanism `TOKEN-FUSE` on the BigEarthNetv2.0 (Clasen et al., 2024) classification dataset. Labels queried that form OSM rasters are shown in appendix Figure 9. † denotes geographic data layers released with this work (aligned with the benchmark datasets).

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 with learnable positional encodings.

Learned location encoders: Location encoders in SatML help models interpolate to new geographic regions by incorporating terrain and environmental signals given a (lat, lon) pair. SatCLIP (Klemmer et al., 2025) builds on GeoCLIP (Vivanco Cepeda et al., 2023), CSP (Mai et al., 2023), and GPS2Vec (Yin et al., 2019) by integrating a CLIP-inspired (Radford et al., 2021) contrastive learning framework specifically designed for satellite imagery from the Sentinel-2 EO satellite. SatCLIP's location encoder, which can be used out-of-the-box, accurately captures terrain, environmental, and socioeconomic signals (Klemmer et al., 2025). Unlike the previously used convolutional architectures that accept a rasterized input of geographic data projected to the correct Coordinate Reference System (CRS), models trained with the SatCLIP location encoder accept embeddings as an auxiliary token.

3.3. Datasets

We conduct experiments using 4 benchmark datasets in ML for remote sensing. These datasets cover different prediction tasks, multi-spectral input sources, and additional data layers used. Table 1 presents an overview of the datasets and additional layers used. *All additional geographic data layers, georeferenced with benchmark datasets, are available as a hosted dataset at* https://huggingface.co/datasets/arjunrao2000/geolayers. We release our code that allows for training models on our datasets at https://github.com/arjunarao619/geolayers-terrabytes.

BigEarthNet (Classification): 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). Unlike the original BigEarthNet dataset in Sumbul et al. (2019), BigEarthNetv2.0 Clasen et al. (2024) constructs a training, validation, and test split by using a grid-based split assignment algorithm. Validation and test areas-of-sampling are not within the geographic extent of the training area-of-sampling, ensuring no data-leakage. Thus, our results reported on the BigEarthNetv2.0 dataset can be considered an out-of-sample validation and test.

EnviroAtlas (Land Cover Segmentation): The EnviroAtlas dataset (compiled by Rolf et al. (2022) and composed of data from Pickard et al. (2015)) consists of high-resolution (1m) land cover maps derived from NAIP aerial imagery. In this dataset, coarse land-cover maps from the National Land Cover Database (NLCD) are aligned with buildings, road networks, water bodies, and waterways from public sources such as the OSM project (Haklay & Weber, 2008). The "prior" data layer constructed in Rolf et al. (2022) is a (hand-coded) fusion of NLCD data with OSM data, in the form of `PROC-STACK`. EnviroAtlas's train split only covers the Pittsburgh region. We use the provided out-of-sample validation and test datasets in Austin and Durham and in-distribution validation and test datasets in Pittsburgh.

SustainBench (Field Boundary Delineation) The SustainBench benchmark 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 the field-delineation task which 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 corre-

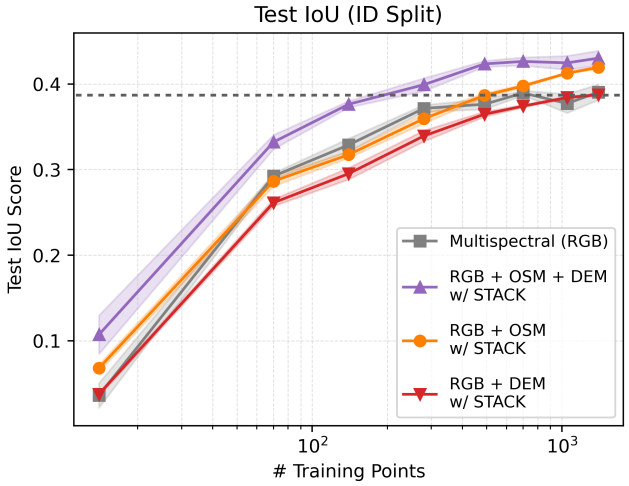

Figure 2. **Performance and label-efficiency of a U-Net trained on SustainBench's Farmland Boundary Delineation Dataset.** We use the standard ID split as benchmarked on in (Aung et al., 2020). Label efficiency and out-of-distribution performance reported as IoU scores averaged over five random seeds. OSM and EU-DEM-aided models match RGB-only model's best score with 221 training images (total = 1573 images).

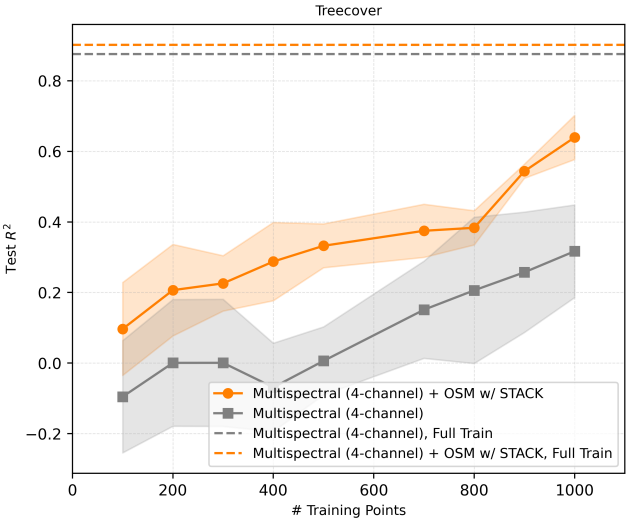

Figure 3. **Performance and data-efficiency of a ResNet50 trained with and without an OSM raster data layer as `STACK` on the USAVars treecover regression task.** Dashed lines show the performance of each input set using the full dataset (100,000 points) as training data.

sponding to an approximately 5 km^2 surface area covered per image.

USAVars Tree-cover (Regression):
The USAVars dataset proposed in Rolf et al. (2021) comprises approximately 100,000 pairs of NAIP aerial imagery cropped to a spatial extent of 1-sq-km per image containing real-valued labels of tree-cover, population density. We pull rasters of several land-cover and infrastructure-related classes from OSM (Haklay & Weber, 2008) as a geographic input, aligned to the RGB layers. Our final set of labels cover broad biome-related land-cover classes such as waterbodies, forests, and buildings with fine-grained labels covering sub-categories of biomes. A complete list of labels pulled from OSM are shown in appendix Figure 9.

4. Results

Across all four SatML benchmark datasets covering tasks in semantic segmentation and multi-label classification for land cover, field boundary delineation, and regression, we found that adding contextual, geographic inputs improves model performance, with largest gains in settings with limited label data (Section 4.1) and out-of-distribution test sets (Section 4.2). Ablation experiments (Section 4.3) provide evidence that fine-tuning encoders aided by geographic input layers does not necessarily help in these critical settings.

4.1. Geographic inputs can aid data-efficiency

The benefit of additional geographic data inputs on data-efficiency of SatML models can be seen in all four experimental settings and all three fusion mechanisms.

From Figure 2, we see performance improvements with low amounts of training data when using the `STACK` approach to fuse additional raster layers. A U-Net trained with an OSM and DEM raster layer using fusion mechanism `STACK` exhibits an 8.1% test dice score improvement in-sample when trained on between 1-5% of training data on the SustainBench field-boundary delineation dataset, compared to a 4.1% improvement when using the full training dataset. From appendix Table 7, we find that these performance improvements hold with most commonly used SatML segmentation model architectures introduced over the past five years. From Figure 3, stacking OSM raster layers as input to a ResNet-50 for the USAVars tree-cover regression task improves R^2 by 0.162 points when trained on between 60 to 250 training images. This performance improvement reduces to a 0.026 improvement in R^2 when the full 68,000 image training dataset is used.

From Table 3, we find that a prior generated and fused with `PROC-STACK` improves in-distribution test accuracy of land-cover segmentation on the EnviroAtlas dataset (Pickard et al., 2015) by 9.3% when trained on between 1 to 5% of the training dataset, compared to a 0.6% improvement when trained with the full training dataset. When the raw data-layers used to generate the prior in Rolf et al.

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

Table 2. **Comparison of a ViT's Average Precision and multi-label F1 score (Macro-averaged) on the BigEarthNetv2.0 test split with and without a SatCLIP location encoder auxiliary token.** BigEarthNetv2.0 (Sumbul et al., 2019) consists of $549,488$ Sentinel-2 image tiles. Ablation includes linear probing a pre-trained SatCLIP location encoder (Right). Mean results over five random seeds. Unless specified, all results report $\leq 0.1\%$ standard error.

(2022) are fused with fusion mechanism `STACK` before training, data-efficiency improvements drop to approximately 2% over ten random seeds for this range $(1-5\%)$ of training data, still an improvement.

On the SustainBench field-boundary delineation and the USAVars tree-cover regression datasets, we note that largest gains in label-efficiency are observed with training dataset sizes of 100-700 images, which we observe to be the low-data-regime where geographic input layers consistently outperform models trained on optical modalities. For example, on the USAVars tree-cover regression task, we observe a diminished gap in the test R^2 metric as we scale from 700 training samples ($\Delta_{R^2} = 0.36$) to 1400 training samples ($\Delta_{R^2} = 0.08$).

We also note that not *all* geographic inputs/combinations of these inputs improve label-efficiency and OOD performance when fused with the SatML model using the fusion mechanisms introduced in Figure 1. In Figure 4, we note that a road-map raster worsens performance compared to standard, multispectral-only training. Similarly, from Figure 2, concatenating a single DEM raster to optical imagery for a field-boundary delineation task on the SustainBench dataset hurts performance in these settings.

4.2. Geographic inputs can aid out-of-distribution performance

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 low-label settings (here $<$800 training points) for all test cities in the EnviroAtlas dataset, Figure 4 also shows an improvement in overall test accuracy across all amounts of training data for the out-of-distribution test cities in different states (Austin, TX and Durham, NC). We observe a 4.12% improvement in the overall accuracy with the prior geographic data layer using `PROC-STACK` and a 2.03% improvement when the raw raster data layers used to generate the prior are fused with `STACK`. Unlike the ID test set (Pittsburgh), the gains in performance in the OOD settings do not appear to diminish with more training samples, as the OOD performance curves remain significantly separated across settings, even using 100% of the training data.

From Table 2, performance improvements on the BigEarthNetv2.0 dataset with the auxiliary SatCLIP token fused with `TOKEN-FUSE` also hold over all training data subsets. This reflects OOD performance as the BigEarthNetv2.0 validation and test splits use a spatial buffering approach (Clasen et al., 2024). For a ViT-B, we observe a 3.1% improvement in the multi-label F1 metric, and a 2.5% improvement in the multi-label average precision metric. Interestingly, for a ViT-S, this improvement in out-of-sample accuracy across all data subsets drops to a 2% improvement in average precision and a 1% improvement in the multi-label F1 metric. We hypothesize that this difference in performance can possibly be attributed to the reduced *model expressivity* of ViT-S that prevents it from fully exploiting the SatCLIP auxiliary token (embedding size of 384 vs 768).

4.3. Finetuning geographic-input aided SatML models can hurt label-efficiency and OOD performance

To determine if geographic inputs that are learned during training aid label efficiency and out-of-sample generalization of SatML models on commonly used benchmark datasets, we conduct ablation studies for the fusion mechanisms `TOKEN-FUSE` and `PROC-STACK`. In sections Sections 4.1 and 4.2, we freeze the intermediate modules $f(\cdot)$ in `PROC-STACK` and $g(\cdot)$ in `TOKEN-FUSE` ($f(\cdot)$ and $g(\cdot)$ from Figure 1). In Sections 4.3.1 and 4.3.2, we finetune these modules jointly with the SatML model.

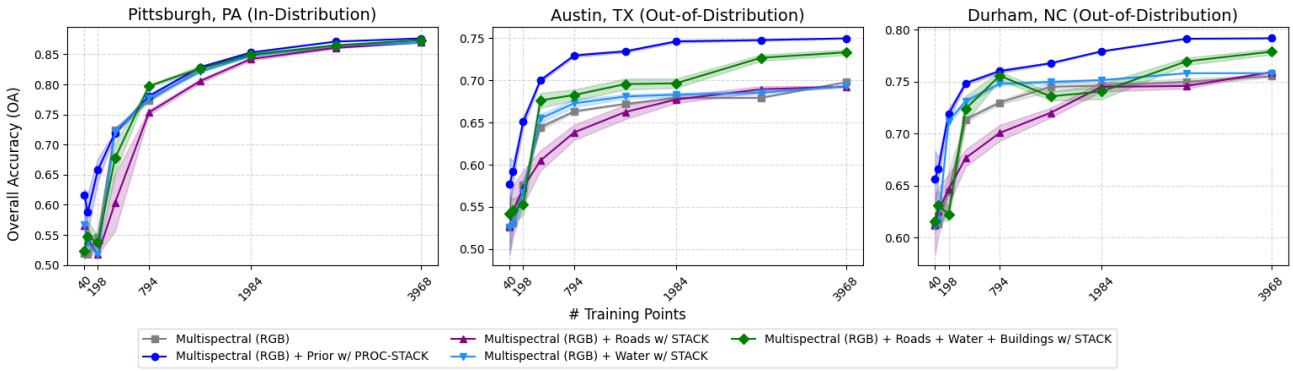

Figure 4. **Performance and label-efficiency of a FCN on the EnviroAtlas Land Cover Segmentation Dataset with `STACK` and `PROC-STACK` geographic input fusion**. Austin and Durham are out-of-sample test splits. Results averaged over 10 random seeds. $1\times$ standard error of Pittsburgh reported $\leq 1e^{-3}$ over 10 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	$\mathbf{0.58 \pm 0.03}$	0.54 ± 0.02	0.61 ± 0.00	$\mathbf{0.66 \pm 0.03}$	0.62 ± 0.00
2%	0.51	**0.58**	0.54	0.53 ± 0.01	$\mathbf{0.59 \pm 0.01}$	0.55 ± 0.01	0.61 ± 0.00	$\mathbf{0.67 \pm 0.01}$	0.63 ± 0.01
5%	0.54	**0.65**	0.55	0.58 ± 0.00	$\mathbf{0.65 \pm 0.01}$	0.55 ± 0.01	0.64 ± 0.02	$\mathbf{0.72 \pm 0.01}$	0.62 ± 0.01

Table 3. **Performance of EnviroAtlas prior, all raster inputs versus RGB input with 1%, 2%, and 5% of input training data.**

Sub%	PROC-STACK, FCN_{out}		RGB Only	STACK
	1	3		
1%	**40.8/26.3**	35.7/22.7	5.5/2.9	26.5/10.1
5%	**49.5/33.7**	47.0/31.5	45.7/29.9	47.0/31.6
10%	52.7/36.6	53.0/36.9	49.0/32.6	**54.7/37.9**
20%	55.2/38.9	54.7/38.5	53.8/37.7	**57.3/39.9**
35%	56.8/40.5	55.9/39.7	54.6/38.4	**59.3/42.5**
50%	57.1/40.9	57.0/40.7	56.3/38.7	**60.3/42.8**
75%	58.1/41.8	58.5/42.3	56.9/40.3	**60.7/43.9**
100%	59.9/43.6	59.3/43.1	57.9/39.5	**61.4/42.9**

Table 4. **Test Dice/IoU score when OSM and EU-DEM rasters are fused via a trainable FCN with `PROC-STACK` on the SustainBench field-boundary delineation task.** We allow the intermediate FCN to output (FCN_{out}) 1 and 3-channel raster outputs. Averaged over 5 random seeds. 100% corresponds to 1572 total training points. RGB, `STACK` reported from Figure 2.

Sub%	F SatCLIP	Register Token	FT SatCLIP
1%	**46.3/36.1**	45.1/33.2	45.4/34.7
2%	**55.6/45.9**	50.3/40.5	53.2/42.8
5%	62.7/54.1	61.6/53.9	**63.5/56.2**
20%	**66.8/60.6**	65.3/59.8	65.3/59.1
50%	**70.1/64.7**	68.1/60.9	67.1/60.1
100%	**70.3/65.2**	66.5/59.6	66.0/59.1

Table 5. **Average Precision (Macro)/ Multi-Label F1 score with Frozen (`F`) vs Register vs Fine-tuned (`FT`) SatCLIP auxiliary token on the BigEarthNetv2.0 dataset**. Results with a register token are reported with the addition of one register token to a ViT-B. 100% corresponds to $\approx 430,000$ image patches. Results averaged over 5 random seeds.

4.3.1. Learned compression with PROC-STACK

To understand when a compressed embedding of geographic rasters can confer similar results as using all as input, we design a trainable PROC-STACK fusion mechanism used to train a U-Net on the SustainBench field boundary delineation task. In this approach, we pass both the DEM (1 channel) and OSM (19 channels) geographic data layers to a trainable FCN architecture. Outputs from the FCN are stacked with the original optical input and passed to the U-Net, and both models are trained simultaneously[1].

Label efficiency on the SustainBench field boundary delineation dataset is shown in Table 4. The fusion mechanism PROC-STACK on learned, compressed inputs is not competitive with a simple STACK of the pre-processed, original rasters. Interestingly, we observe significantly improved la-

[1]To accommodate for the increased number of trainable parameters, we increase the number of epochs the models are trained on and allow for convergence.

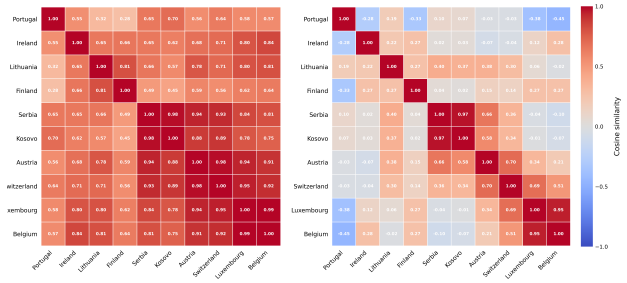

Figure 5. **Pairwise cosine similarity of SatCLIP embeddings used to form auxiliary ViT token: Frozen (Left) vs Fine-Tuned (Right).** On the BigEarthNetv2.0 land cover classification task, the fine-tuned SatCLIP token exhibits significantly greater pairwise disagreement between countries compared to the frozen token in countries covered by the train-split of BigEarthNetv2.0.

bel efficiency of the trained PROC−STACK ablation model between subsets 1% and 5%. These label efficiency improvements, however, do not hold across all subsets.

4.3.2. Fine-tuning location encoders in TOKEN−FUSE

For classification on the BigEarthNetv2.0 dataset, instead of using a frozen SatCLIP encoder with a learnable linear projection layer (as in Section 4.2), we now allow for the SatCLIP model to be trainable given the original pretrained SatCLIP location encoder weights.

We find that the label-efficiency and out-of-sample performance degrade when the SatCLIP weights are learnable during training (Table 5). Figure 5 shows that fine-tuning the SatCLIP model in this fashion leads to embeddings that are highly localized within various countries covered by the BigEarthNetv2.0 dataset. This suggests that the augmented ViT may be overfitting to the auxiliary SatCLIP token, leading to lower test set performance when the SatCLIP model is trainable. Furthermore, overfitting is particularly likely considering that the trainable weights of the SatCLIP location encoder span 360k parameters – significantly higher than other image tokens input to the ViT. [2]

To understand the performance discrepancy between the fine-tuned and frozen location encoders in the TOKEN−FUSE strategy, we compare performance of our auxiliary SatCLIP token against a generic (non-geospatial) learnable register token as a baseline. First introduced in (Darcet et al., 2024), register tokens are randomly initialized, fully-trainable prefix tokens. Register tokens capture high-norm "outlier" artifacts that hold significantly lower local-patch information. ViTs aided with registers show improvements only when trained with sufficiently large numbers of trainable parameters (ViT-B, ViT-L, ViT-H) over long training durations. We choose a ViT-B (86M trainable parameters) with identical hyperparameters as experiments that produced results in Table 2.

From Table 5, we find that registers do not improve label efficiency and out-of-sample performance of a ViT-B trained on the BigEarthNetv2.0 dataset compared to a frozen SatCLIP location encoder. We find that adding additional register tokens up to 3 tokens doesn't significantly alter this result. Interestingly, both a register token and a fine-tuned SatCLIP token outperform a vanilla ViT-B when trained on between 1% to 20% of training data, but perform worse than a vanilla ViT in the large-data (50%, 100%) regime.

5. Experimental Takeaways

Takeaway 1: Auxiliary geographic inputs improve performance in low-data settings. In Section 4.1, we find notable performance improvements in low-data settings with an auxiliary OSM and DEM geographic input layer (0.08 IoU on SustainBench, 9.3% OA on EnviroAtlas, 0.162 R^2 improvement on USAVars). On the SustainBench field boundary delineation task, a U-Net trained with an OSM and EU-DEM raster matches the test IoU of an RGB-only model with only 224 training samples (compared to 1573 training samples for the RGB-only model).

Takeaway 2: Auxiliary geographic inputs improve performance OOD. From Section 4.2, we find that these geographic layers are especially helpful when evaluated on OOD splits of the benchmark datasets: 4.12% improvement in EnviroAtlas's OOD cities, 3.1% improvement on BigEarthNetv2.0's spatially-buffered test splits.

Takeaway 3: Finetuning SatML models aided by auxiliary geographic inputs *can* hurt performance. Surprisingly, when we allow the intermediate module in PROC−STACK (denoted by $f(\cdot)$ in Figure 1) to be trainable and act as a geographic input compression module, test IoU scores drop, on average, by 4.1% on the Table 4) test set. Higher performance drops occur as the expressivity of the intermediate FCN is increased from 1 to 3 output channels. From Section 4.3.2, we find that allowing a SatCLIP encoder to be jointly trained with the SatML classification model causes the model to overfit (Figure 5), hurting label efficiency and OOD performance in the BigEarthNet task.

Limitations and future work: In Figures 2 to 4 and 6 and Tables 2 and 3, we use geographic data-layers that make sense for the downstream task. As we are primarily interested in potential benefits of using additional data layers, we restrict the scope of the study only to these geographic input layers and do not train on a larger corpus of raster and scalar inputs. Here, we use fusion mech-

[2] Addition of layer-normalization to the SatCLIP token doesn't significantly alter performance, label-efficiency, and OOD generalization.

anisms `STACK`, `PROC-STACK` for convolutional models and `TOKEN-FUSE` for ViTs since they involve minimal modifications to the source architectures; future work will examine more sophisticated fusion mechanisms.

Acknowledgements

A majority of training runs conducted in this work were run on an NVIDIA Grace-Hopper (GH200) GPU node provided by the University of Colorado Boulder's high performance computing system Alpine. We thank Brandon Reyes and the RC computing team at CU Boulder for allowing access to this resource. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, Colorado State University, and the National Science Foundation (award 2201538).

OpenStreetMap is open data, licensed under the Open Data Commons Open Database License by the OpenStreetMap Foundation (OSMF).

DEM data in this work is derived from services provided by the OpenTopography Facility with support from the National Science Foundation under NSF Award Numbers 2410799, 2410800 & 2410801 (Hengl et al., 2020).

We thank Dr. Caleb Robinson for invaluable feedback during the writing stage of this work. We also thank the anonymous reviewers for their comments and suggestions.

Impact Statement

By lowering annotation costs and delivering consistent accuracy when models cross regional, temporal or sensor boundaries, our approach can democratize high-impact Earth-observation applications such as crop monitoring, disaster assessment and biodiversity mapping for organizations with limited resources. Because the fusion layers are lightweight and the best results come from *frozen* tokens using *pretrained* encoders, our work avoids the large training footprints typical of foundation-model fine-tuning, mitigating energy use relative to existing alternatives. However, the work also surfaces risks: uneven coverage or quality in auxiliary datasets (e.g., OSM) could entrench geographic biases, and fine-tuning the location encoder can cause severe overfitting to local patterns. Practitioners should therefore audit input-layer availability and monitor model generalization before deployment in safety- or equity-critical settings.

References

Aung, H. L., Uzkent, B., Burke, M., Lobell, D., and Ermon, S. 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.

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

Benson, V., Robin, C., Requena-Mesa, C., Alonso, L., Carvalhais, N., Cortés, J., Gao, Z., Linscheid, N., Weynants, M., and Reichstein, M. Multi-modal learning for geospatial vegetation forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27788–27799, 2024.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., Leung, T., and Adam, H. Geo-aware networks for fine-grained recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Clasen, K. N., Hackel, L. W., Burgert, T., Sumbul, G., Demir, B., and Markl, V. reBEN: Refined BigEarthNet dataset for Remote Sensing Image Analysis. *CoRR*, abs/2407.03653, 2024. URL https://doi.org/10.48550/arXiv.2407.03653.

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D., and Ermon, S. SatMAE: Pretraining transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.11929. arXiv:2010.11929.

Fan, T., Wang, X., Cheng, M., and Tao, D. Ma-net: Multiscale attention network for liver and tumor segmentation. *IEEE Access*, 8:179683–179691, 2020. doi: 10.1109/ACCESS.2020.3025372.

Fonte, C. C., Patriarca, J., Jesus, I., and Duarte, D. Automatic extraction and filtering of openstreetmap data to generate training datasets for land use land cover classification. *Remote Sensing*, 12(20):3428, 2020.

Haklay, M. and Weber, P. OpenStreetMap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008.

He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Hengl, T., Leal Parente, L., Krizan, J., and Bonannella, C. Continental europe digital terrain model at 30 m resolution based on gedi, icesat-2, aw3d, glo-30, eudem, merit dem and background layers. *Version Dataset v3. 0. Zenodo*, 2020.

Hou, Y., Liu, Z., Zhang, T., and Li, Y. C-unet: Complement unet for remote sensing road extraction. *Sensors*, 21(6): 2153, 2021. doi: 10.3390/s21062153.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

Johnson, N., Treible, W., and Crispell, D. Opensentinelmap: A large-scale land use dataset using OpenStreetMap and Sentinel-2 imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1333–1341, 2022.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., and Rußwurm, M. SatCLIP: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI conference on artificial intelligence*, 2025.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

Mac Aodha, O., Cole, E., and Perona, P. Presence-only Geographical Priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9596–9606, 2019.

Mai, G., Lao, N., He, Y., Song, J., and Ermon, S. Csp: Self-supervised contrastive spatial pre-training for geospatial-visual representations. In *International Conference on Machine Learning*, pp. 23498–23515. PMLR, 2023.

Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., and Lang, N. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Patriarca, J., Fonte, C., Estima, J., de Almeida, J.-P., and Cardoso, A. Automatic conversion of OSM data into LULC maps: comparing FOSS4G based approaches towards an enhanced performance. *Open Geospatial Data, Software and Standards*, 4:1–19, 2019.

Pickard, B. R., Daniel, J., Mehaffey, M., Jackson, L. E., and Neale, A. EnviroAtlas: A new geospatial tool to foster ecosystem services science and resource management. *Ecosystem Services*, 14:45–55, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., and Darrell, T. 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.

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*, 12(1):4392, 2021.

Rolf, E., Malkin, N., Graikos, A., Jojic, A., Robinson, C., and Jojic, N. Resolving label uncertainty with implicit posterior models. In Cussens, J. and Zhang, K. (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1707–1717. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/rolf22a.html.

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

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015a.

Ronneberger, O., Fischer, P., and Brox, T. 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, 2015b.

Roscher, R., Russwurm, M., Gevaert, C., Kampffmeyer, M., Dos Santos, J. A., Vakalopoulou, M., Hänsch, R., Hansen, S., Nogueira, K., Prexl, J., and Tuia, D. Better, not just more: Data-centric machine learning for earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):335–355, 2024. doi: 10.1109/MGRS.2024.3470986.

Sosa, J., Rukhovich, D., Kacem, A., and Aouada, D. Multimae meets earth observation: Pre-training multi-modal multi-task masked autoencoders for earth observation tasks, 2025. URL https://arxiv.org/abs/2505.14951.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. 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.

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., and Bourdev, L. Improving image classification with location context. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1008–1016, 2015.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image transformers & distillation through attention. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., and Łukasz Kaiser. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vivanco Cepeda, V., Nayak, G. K., and Shah, M. Geoclip: Clip-inspired alignment between locations and images for effective worldwide geo-localization. *Advances in Neural Information Processing Systems*, 36:8690–8701, 2023.

Wang, Y., Cheng, L., Fang, C., Zhang, D., Duan, M., and Wang, M. Revisiting the Power of Prompt for Visual Tuning. In *ICML*, 2024. URL https://openreview.net/forum?id=2Y93PtAqCl.

Wang, Z., Yu, J., Wu, Z., Zhang, R., Mao, J., Li, L., Feng, Z., and Yin, J. 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.

Weng, L., Xu, Y., Xia, M., Zhang, Y., Liu, J., and Xu, Y. Water areas segmentation from remote sensing images using a separable residual segnet network. *ISPRS International Journal of Geo-Information*, 9(4):256, 2020. doi: 10.3390/ijgi9040256.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

Yeh, C., Meng, C., Wang, S., Driscoll, A., Rozi, E., Liu, P., Lee, J., Burke, M., Lobell, D. B., and Ermon, S. 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.

Yin, Y., Liu, Z., Zhang, Y., Wang, S., Shah, R. R., and Zimmermann, R. GPS2Vec: Towards generating worldwide GPS embeddings. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 416–419, 2019.

Yuan, W., Wang, J., and Xu, W. Shift pooling pspnet: Rethinking pspnet for building extraction in remote sensing images from entire local feature pooling. *Remote Sensing*, 14(19):4889, 2022. doi: 10.3390/rs14194889.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., et al. ESA WorldCover 10 m 2021 v200. 2022.

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, 2017.

A. Experimental Setup

A.1. FCN on EnviroAtlas Land Cover Segmentation with `STACK`, `PROC-STACK`:

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 $1e-3$ are fixed across all training data subsets and random seeds reported in Figure 4. We fix the lower bound learning rate to $1e-7$. Table 6 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).

Hand-crafted prior generation process. In our PROC-STACK experiments, the hand-crafted prior $f(x_i) \equiv p_i(\ell)$ is constructed exactly as in (Rolf et al., 2022) ("Coarse data in weakly supervised segmentation", §3), using the NLCD 30 m land-cover map to induce per-pixel beliefs over our four high-resolution classes. Concretely, we first compute the empirical co-occurrence matrix

$$P(\ell \mid c) \;=\; \frac{\big|\{\text{high-res label} = \ell,\ \text{NLCD class} = c\}\big|}{\sum_{\ell'} \big|\{\text{high-res label} = \ell',\ \text{NLCD class} = c\}\big|}$$

from a held-out set of aligned NAIP+NLCD+Land Cover tiles. Then, for each pixel i with NLCD class c_i, we set

$$p_i(\ell) \;=\; P(\ell \mid c_i)$$

and apply a small Gaussian blur ($\sigma = 1$ pixel) to smooth block artifacts. In PROC-STACK mode, we further enrich this prior with binary auxiliary masks (roads, buildings, waterways): for each feature j, we define

$$M_j(i) = \begin{cases} 1, & \text{if feature } j \text{ lies within a 10 m radius of pixel } i, \\ 0, & \text{otherwise,} \end{cases}$$

and boost the corresponding class by adding a fixed weight w_j to $p_i(\ell = j)$. Finally, we re-normalize $p_i(\ell)$ so that $\sum_\ell p_i(\ell) = 1$. This yields a spatially varying, hand-crafted prior that both encodes coarse NLCD statistics and injects domain knowledge via auxiliary GIS layers, as required by the PROC-STACK formulation.

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

Table 6. Training epochs scaled by subset size for all label-efficiency experiments.

A.2. ViT on BigEarthNetv2.0 Multi-label classification with `TOKEN-FUSE`

Our experiments with the Vision Transformer (ViT) use a ViT-Base and a ViT-Small (86M and 22M trainable parameters) with a fixed patch size of 8. All ViTs are randomly initialized for a fixed random seed. We prepend a learnable location token $x_{\text{loc}} \in \mathbb{R}^D$ to the input sequence in addition to a class token $x_{\text{cls}} \in \mathbb{R}^D$ and N patch tokens $x^{(i)}_{\text{patch}} \in \mathbb{R}^D$. The token sequence is given by

$$X_{\text{tokens}} = \Big[x_{\text{cls}};\ x_{\text{loc}};\ x^{(1)}_{\text{patch}}, \ldots, x^{(N)}_{\text{patch}}\Big] \in \mathbb{R}^{(N+2)\times D},$$

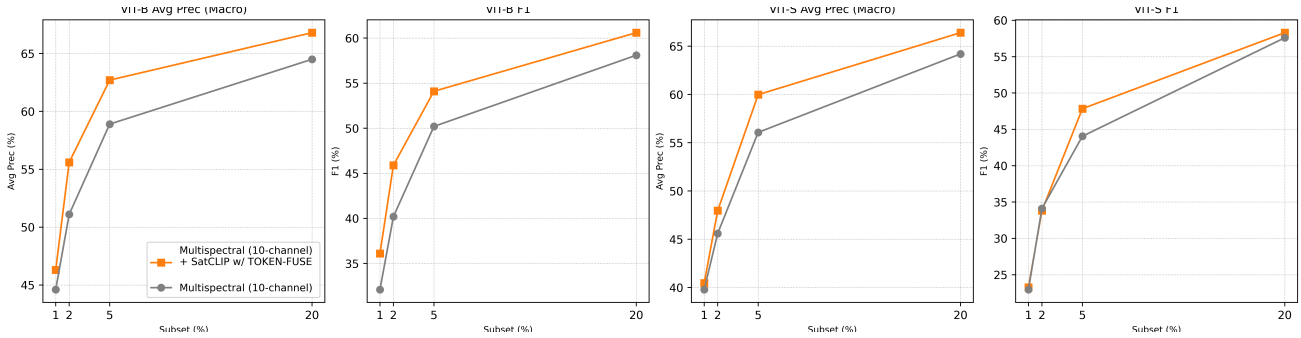

Figure 6. **Label efficiency of a ViT trained with an auxiliary SatCLIP token. Left:** ViT-Base (86M trainable parameters). SatCLIP linear projection layer mapped to embedding dimension of 768. **Right:** ViT-small (22M trainable params), SatCLIP linear projection layer mapped to embedding dimension of 384.

We add corresponding learnable positional embeddings

$$E_{\text{pos}} = \left[e_{\text{cls}};\ e_{\text{loc}};\ e^{(1)}_{\text{patch}}, \ldots, e^{(N)}_{\text{patch}}\right].$$

Our final sequence is $z_0 = X_{\text{tokens}} + E_{\text{pos}}$. With the addition of the auxiliary SatCLIP token with `TOKEN-FUSE`, our sequence length is increased by one and allows the model to jointly encode class and location information.

Why SatCLIP? SatCLIP is currently the only location encoder in previous work that is pre-trained on Sentinel-2 satellite imagery, hence making it a suitable candidate for our experiments that primarily train, validate, and test on geospatial satellite imagery. Future work will incorporate the label-efficiency and out-of-sample performance for SatML models trained with newer location encoders that are pre-trained with satellite or geospatial imagery.

Our experiments on the BigEarthNetv2.0 dataset use a batch size of 700 and run for 15 epochs (5 warmup epochs) at a base learning rate of $5e-4$. We use a dropout rate of 0.15 to prevent overfitting across all settings (Finetuned SatCLIP (`FT`), Frozen SatCLIP (`F`), and Register token). We record macro and micro-averaged average precision, recall, and F1 score in addition to class-wise accuracies. Figure 6 shows label-efficiency results (similar to Table 2) of a frozen SatCLIP auxiliary token with a learnable linear projection layer on the BigEarthNet2.0 dataset.

A.3. U-Net on SustainBench Field Boundary Delineation with `STACK`

Our standard U-Net setup consists of 4 downsampling blocks, a bottleneck, and corresponding upsampling blocks with skip connections. Input images are georeferenced with a pre-processed OSM raster and are stored as an HDF5 dataset with 7 total channels. We use a random crop, horizontal, and vertical flip augmentation during training and a center crop for evaluation. The model is trained for 20 epochs with a batch size of 48 at a learning rate of 1×10^{-4}. A learning rate scheduler cognizant of validation loss plateaus is used (factor 0.5, patience 5). We record the Dice coefficient, and the IoU score.

Ablations with model architectures: We conduct a broad survey of commonly used SatML model architectures for semantic segmentation tasks from published work spanning 2020 to 2025. We find that most commonly used segmentation architectures include:

- Fully Convolutional Networks (FCN) (Long et al., 2015)
- U-Net (Ronneberger et al., 2015a; Hou et al., 2021)
- SegNet (Badrinarayanan et al., 2017; Weng et al., 2020)
- PSPNet (Zhao et al., 2017; Yuan et al., 2022)
- DeepLabv3+ (Chen et al., 2018)

- SegFormer (Xie et al., 2021)
- MA-Net (Fan et al., 2020)

We choose 4 commonly used segmentation model architectures from the list above, and perform the label-efficiency experiments similar to Section 4.1 with and without an auxiliary geographic input of an OSM and EU-DEM raster layer. From Table 7, we see that our performance improvements hold consistently over all data subsets with the auxiliary geographic input.

A.4. ResNet50 on USAVars Regression with `STACK`

Our generated USAVars dataset comprises images with 7 channels and corresponding scalar labels. A custom `HDF5Dataset` class is used to load the data. For 7-channel inputs, the first four channels are normalized to $[0, 1]$ by division by 255, while channels 4–6 are scaled from the original categorical values returned from the OSM API to the RGB space. Random cropping (to an image size of 256), horizontal, and vertical flips are applied during training, while a center crop is used for validation and testing. To accommodate inputs with 4 or 7 channels, the initial convolutional layer of ResNet50 is re-initialized accordingly. The final fully-connected layer is replaced with a linear layer outputting a single value for regression. We use a base learning rate of $1e-4$ with a batch size of 512. We train the model for 20 epochs. All experiments are seeded for reproducibility and results are reported over five random seeds. We record the mean squared error loss and the R^2 score.

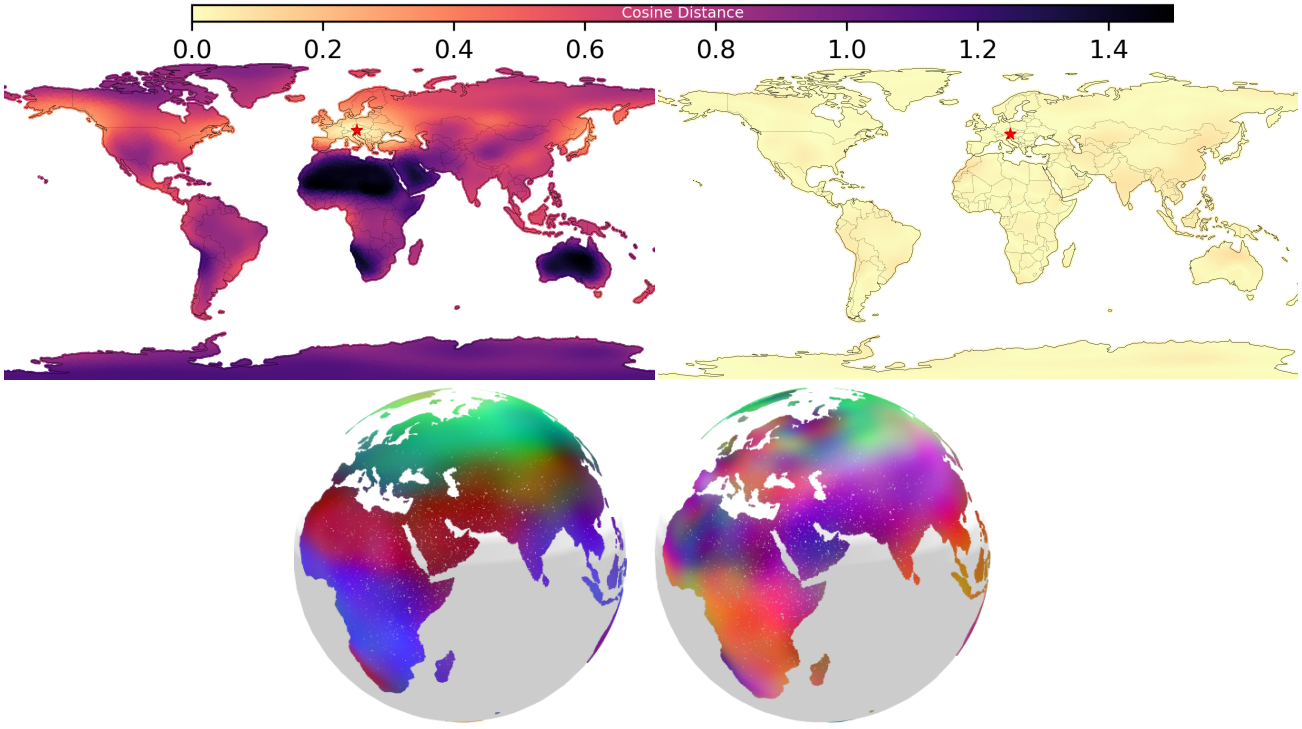

Figure 7. **Qualitative Result: Frozen `F` vs Finetuned `FT` SatCLIP auxiliary token** [Top-left] Cosine distance of standard SatCLIP embeddings to a fixed reference point in Austria. [Top-Right] Absolute difference between cosine distances between our `F` SatCLIP location encoder + trained linear projection layer and original SatCLIP location encoder cosine distances. [Bottom] Global PCA embeddings of `F` vs `FT` SatCLIP auxiliary token with `TOKEN-FUSE`

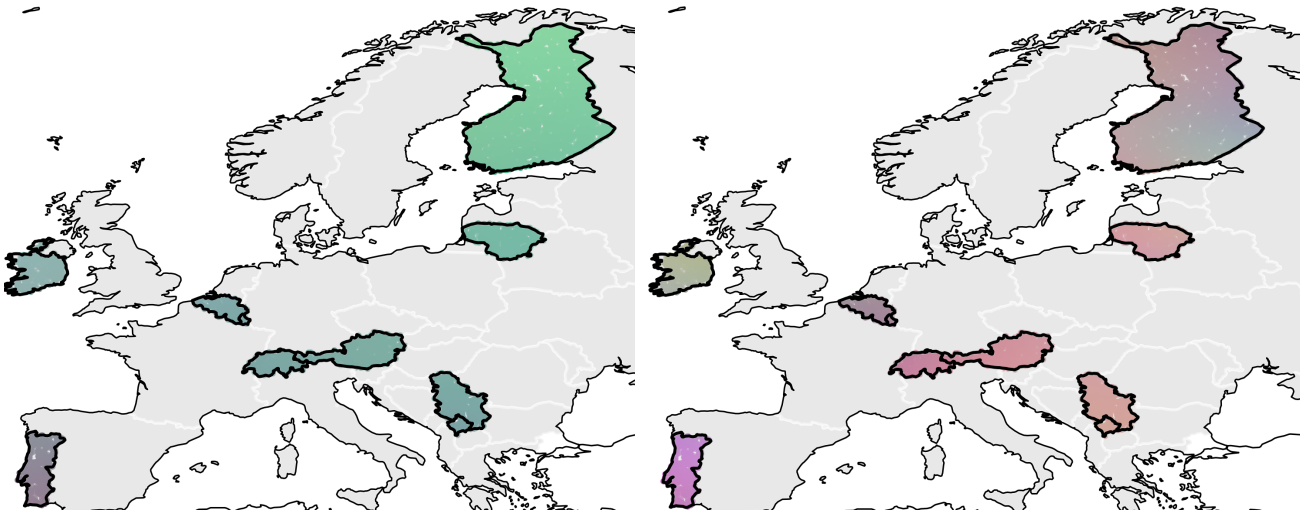

Figure 8. **Qualitative result: Frozen vs Finetuned SatCLIP auxiliary ViT token on the BigEarthNetv2.0 land-cover classification task:** Maps: PCA embeddings of the SatCLIP tokens: frozen (left) vs finetuned (right) on 10 European countries covered by the BigEarthNetv2.0 dataset.

B. Qualitative Result: `TOKEN-FUSE`

To qualitatively evaluate the quality of embeddings learned by our linear projection layer, which is responsible for mapping the 256-dimensional SatCLIP embeddings to the token size expected by the ViT, we calculate the disagreement of this learned layer with a standard SatCLIP location encoder. With a fixed reference SatCLIP embedding in Austria (E_{Austria}), we calculate the cosine distance between SatCLIP embeddings of 200,000 global, randomly sampled SatCLIP embeddings with E_{Austria}. The disagreement of our learned linear projection layer is calculated by repeating the same procedure after passing standard SatCLIP embeddings through the learned linear projection layer before calculating the cosine distance. Figure 7 [top-right] shows that our learned linear projection layer successfully maps SatCLIP embeddings to the SatCLIP auxiliary token without a significant disagreement from original embeddings. Figure 7[Bottom] also shows a PCA visualization of a frozen (`F`) vs finetuned (`FT`) SatCLIP auxiliary token's embeddings mapped to RGB space. Figure 8 surprisingly shows these PCA embeddings for countries covered by the train-split of the BigEarthNetv2.0 dataset (Clasen et al., 2024). These results support our observation in Figure 8 that show that a finetuned SatCLIP token with `TOKEN-FUSE` learns high-resolution, arbitrary information compared to a frozen token.

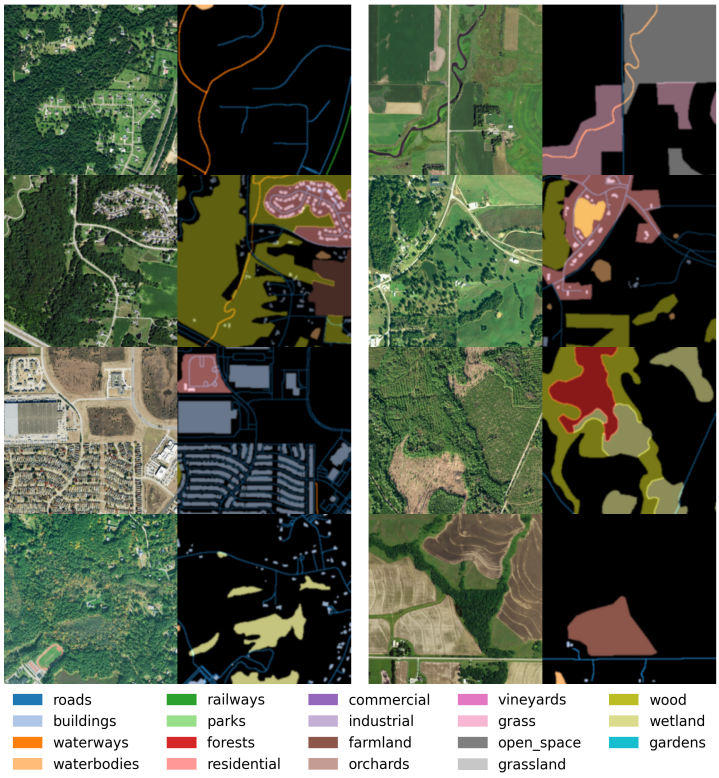

Figure 9. **NAIP Imagery from USAVars (Rolf et al., 2021) georeferenced with our OpenStreetMaps (OSM) raster geographic data-layer:** OSM products are smoothed with a Gaussian Kernel and pre-processed to RGB space.

Subset	Model	RGB		RGB + OSM + EU-DEM w/ `STACK`	
		IoU	Dice	IoU	Dice
0.01	deeplabv3+	0.163 ±0.046	0.261 ±0.079	0.180 ±0.051	0.287 ±0.072
	fpn	0.133 ±0.040	0.222 ±0.063	0.135 ±0.042	0.224 ±0.065
	pspnet	0.125 ±0.043	0.207 ±0.074	0.132 ±0.048	0.221 ±0.066
	unetpp	0.251 ±0.006	0.397 ±0.008	0.259 ±0.007	0.407 ±0.009
0.05	deeplabv3+	0.293 ±0.009	0.445 ±0.012	**0.311 ±0.006**	0.465 ±0.008
	fpn	0.114 ±0.024	0.197 ±0.036	0.120 ±0.021	0.207 ±0.032
	pspnet	0.156 ±0.025	0.262 ±0.038	0.148 ±0.036	0.246 ±0.054
	unetpp	0.318 ±0.007	0.475 ±0.008	0.333 ±0.008	0.491 ±0.009
0.10	deeplabv3+	0.317 ±0.004	0.472 ±0.005	**0.333 ±0.007**	**0.490 ±0.009**
	fpn	0.123 ±0.012	0.212 ±0.019	0.139 ±0.010	0.238 ±0.015
	pspnet	0.157 ±0.007	0.264 ±0.011	0.169 ±0.016	0.281 ±0.025
	unetpp	0.363 ±0.004	0.524 ±0.004	**0.377 ±0.004**	**0.538 ±0.004**
0.20	deeplabv3+	0.326 ±0.003	0.482 ±0.003	**0.343 ±0.004**	**0.500 ±0.004**
	fpn	0.193 ±0.008	0.314 ±0.011	**0.213 ±0.004**	**0.342 ±0.006**
	pspnet	0.153 ±0.002	0.258 ±0.005	0.154 ±0.002	0.258 ±0.003
	unetpp	0.385 ±0.003	0.548 ±0.004	**0.405 ±0.002**	**0.567 ±0.002**
0.35	deeplabv3+	0.343 ±0.004	0.501 ±0.004	**0.360 ±0.003**	**0.520 ±0.003**
	fpn	0.241 ±0.010	0.377 ±0.012	**0.266 ±0.011**	**0.409 ±0.014**
	pspnet	0.163 ±0.006	0.272 ±0.009	0.164 ±0.004	0.273 ±0.005
	unetpp	0.391 ±0.002	0.554 ±0.002	**0.415 ±0.003**	**0.578 ±0.003**
0.50	deeplabv3+	0.353 ±0.003	0.513 ±0.003	**0.363 ±0.003**	**0.522 ±0.003**
	fpn	0.253 ±0.005	0.393 ±0.007	**0.285 ±0.004**	**0.433 ±0.005**
	pspnet	0.167 ±0.008	0.278 ±0.011	0.174 ±0.002	0.289 ±0.002
	unetpp	0.398 ±0.001	0.561 ±0.001	**0.420 ±0.002**	**0.582 ±0.002**
0.75	deeplabv3+	0.358 ±0.001	0.518 ±0.001	**0.383 ±0.002**	**0.545 ±0.002**
	fpn	0.285 ±0.007	0.433 ±0.008	**0.315 ±0.004**	**0.468 ±0.005**
	pspnet	0.187 ±0.007	0.306 ±0.010	0.194 ±0.007	0.317 ±0.010
	unetpp	0.402 ±0.002	0.564 ±0.002	**0.430 ±0.001**	**0.593 ±0.001**
1.00	deeplabv3+	0.368 ±0.003	0.529 ±0.003	**0.390 ±0.004**	**0.553 ±0.004**
	fpn	0.311 ±0.003	0.464 ±0.003	**0.335 ±0.004**	**0.491 ±0.004**
	pspnet	0.199 ±0.005	0.323 ±0.007	0.204 ±0.003	0.330 ±0.004
	unetpp	0.408 ±0.002	0.571 ±0.002	**0.436 ±0.001**	**0.598 ±0.001**

Table 7. **Performance and Label Efficiency of commonly used SatML semantic segmentation model architectures on the SustainBench field boundary delineation dataset with and without an OSM and EU-DEM auxiliary geographic data layer:** Model choices are informed by a surveying the SatML segmentation model literature spanning five years. Test Dice and IoU scores reported based on a hold-out validation set. Results averaged over 3 random seeds. Bolded numbers indicate best model performance for a given data subset.