Using Multiple Input Modalities can Improve Data-Efficiency for ML with Satellite Imagery

Arjun Rao & Esther Rolf
University of Colorado Boulder
raoarjun@colorado.edu

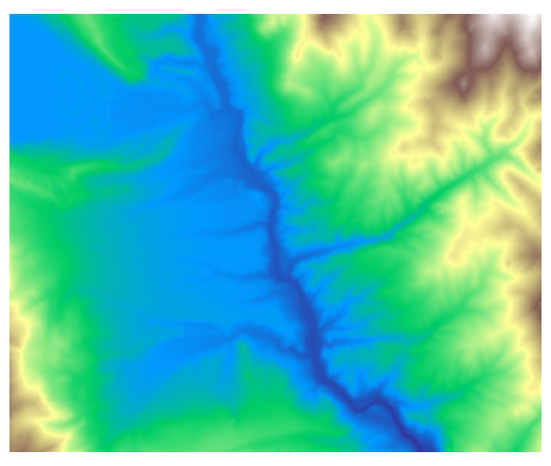
Problem 1

GeoML Models fail to Generalize OOD

Problem 2

Many (Publicly Available) Modalities
Few Labels!

Optical


DEM

OSM
(Human-Annotated)

Considerations

Additional Input Modalities can provide valuable context clues → BUT → Could Require Data-Hungry Models!

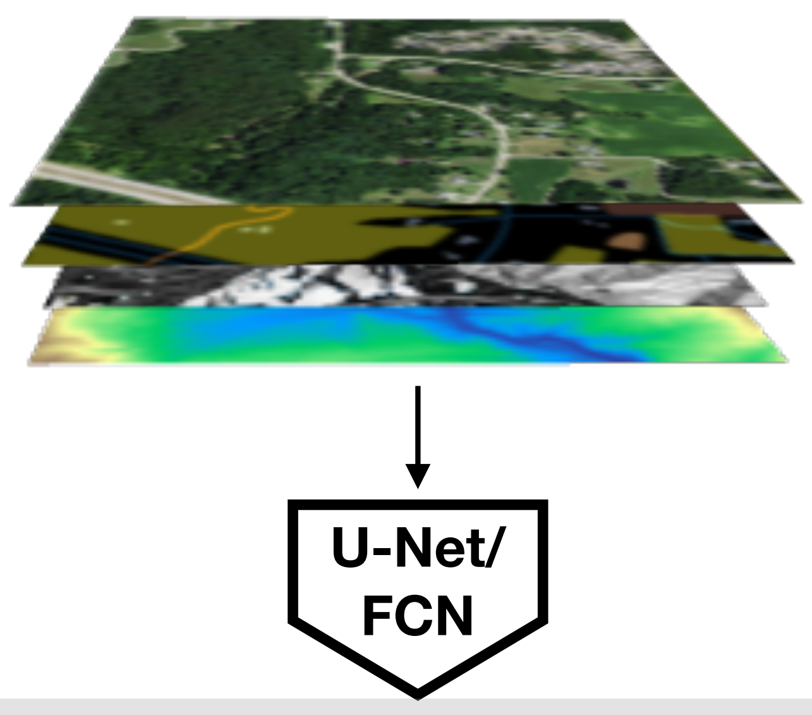
Additional Input Modalities can help OOD generalization → BUT → Could cause models to overfit to local patterns ID!

Research Goal

We study the label-efficiency and OOD generalization capability associated with adding non-optical, contextual inputs to commonly used GeoML architectures

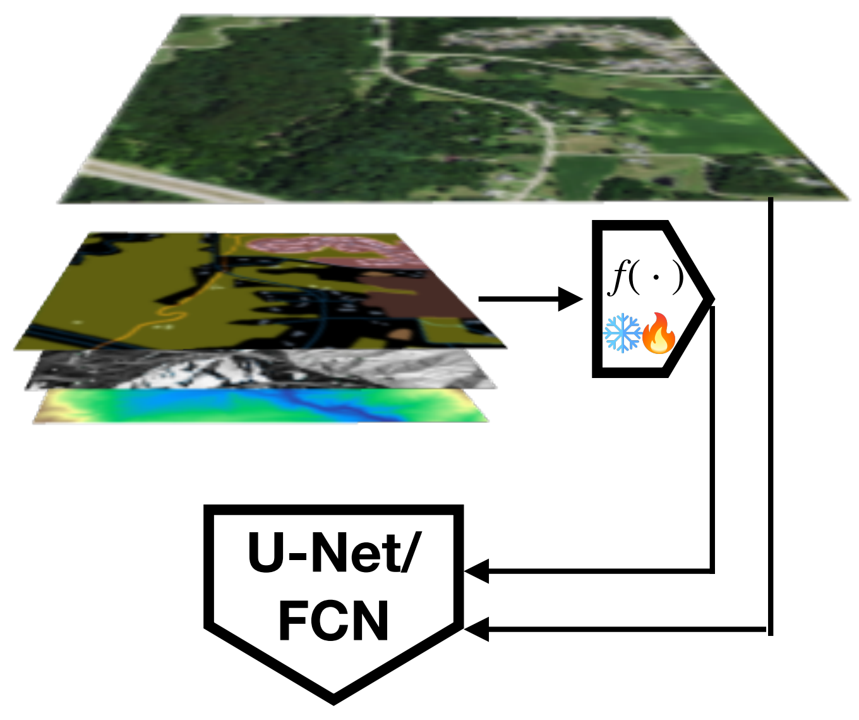
Methodology

[STACK]



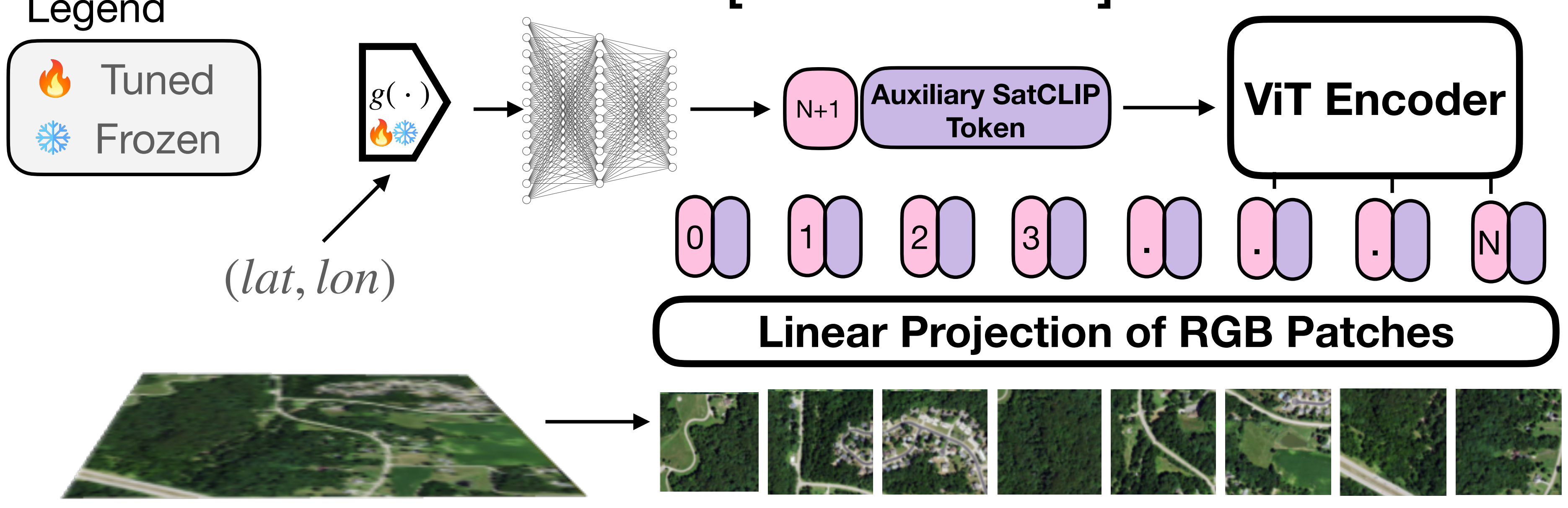
Fuse raw OSM and DEM rasters and stack below optical modality with a vanilla U-Net **[STACK]**

[PROC-STACK]



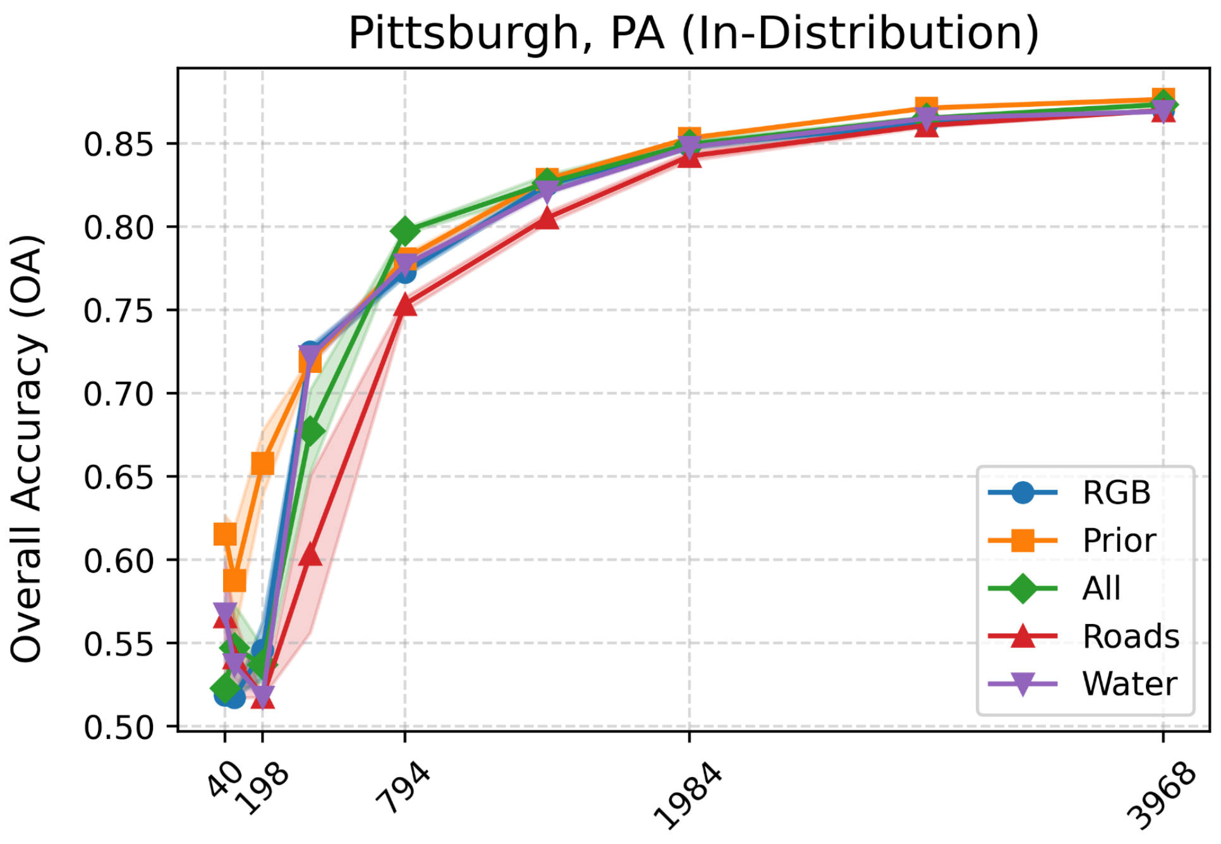
Fuse pre-generated prior from [2] generated from OSM data with a vanilla FCN **[PROC-STACK]**

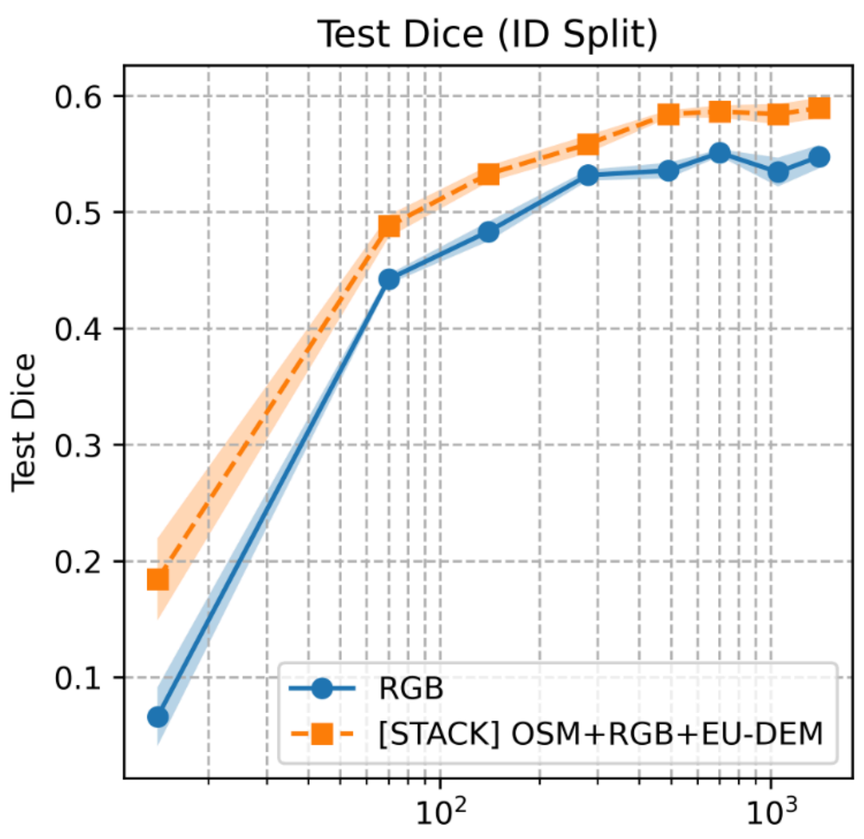
[TOKEN-FUSE]



Fuse projected SatCLIP [1] embedding as auxiliary token with patch tokens to a ViT-B, S **[TOKEN-FUSE]**

Key Result 1: Multi-Modal Inputs Aid Label-Efficiency ID!



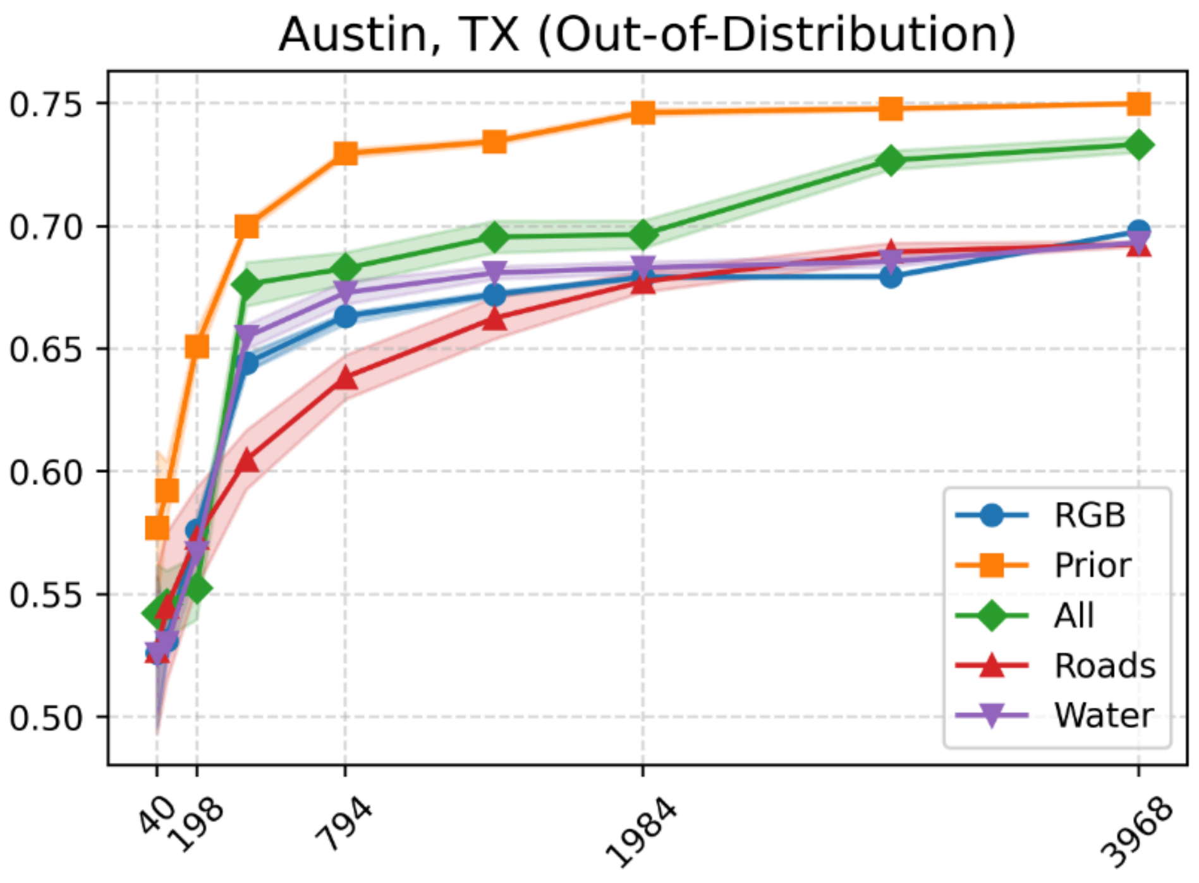


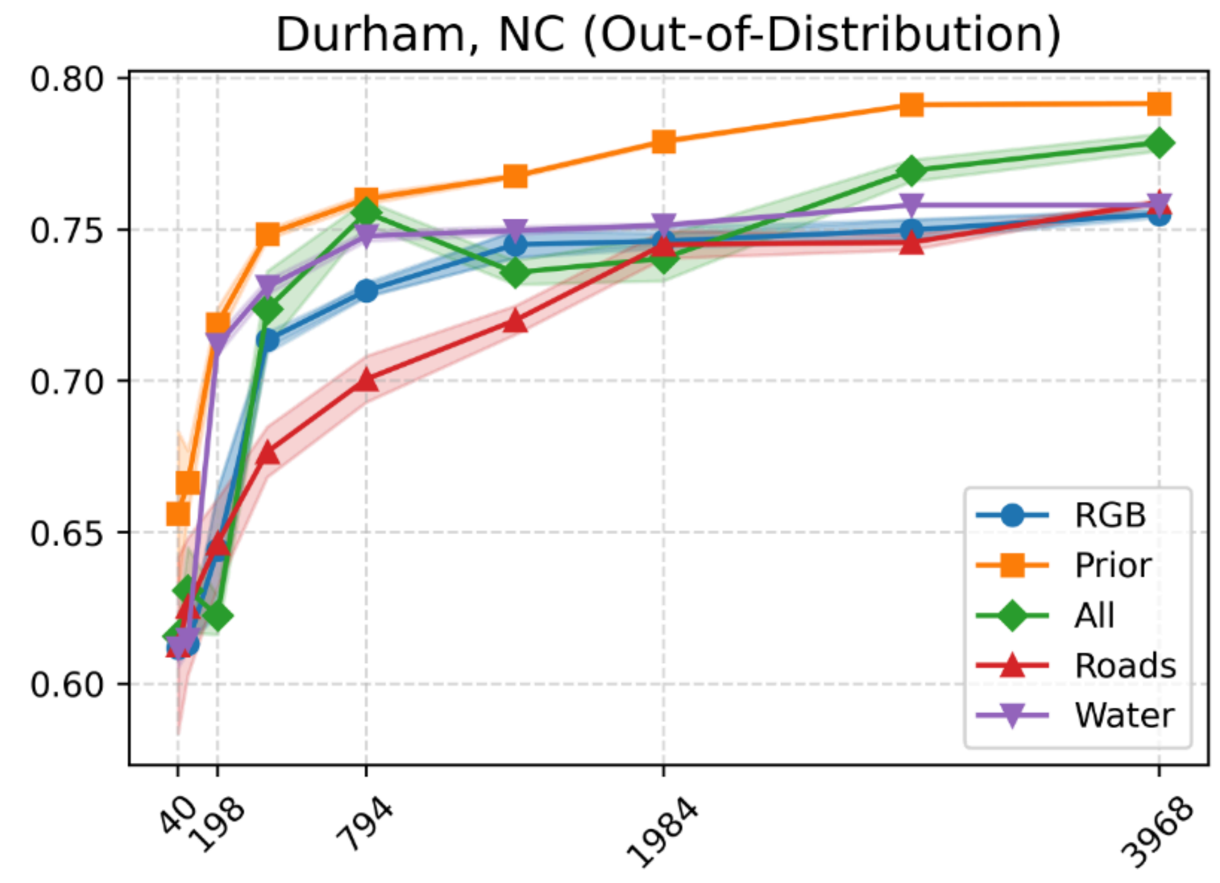
Between 100-700 training samples:

9.3% Improvement in test OA with EnviroAtlas using [PROC-STACK]

8.1% Improvement in test Dice with SustainBench Field Delineation using [STACK]

Key Result 2: Multi-Modal Inputs Aid OOD Generalization!



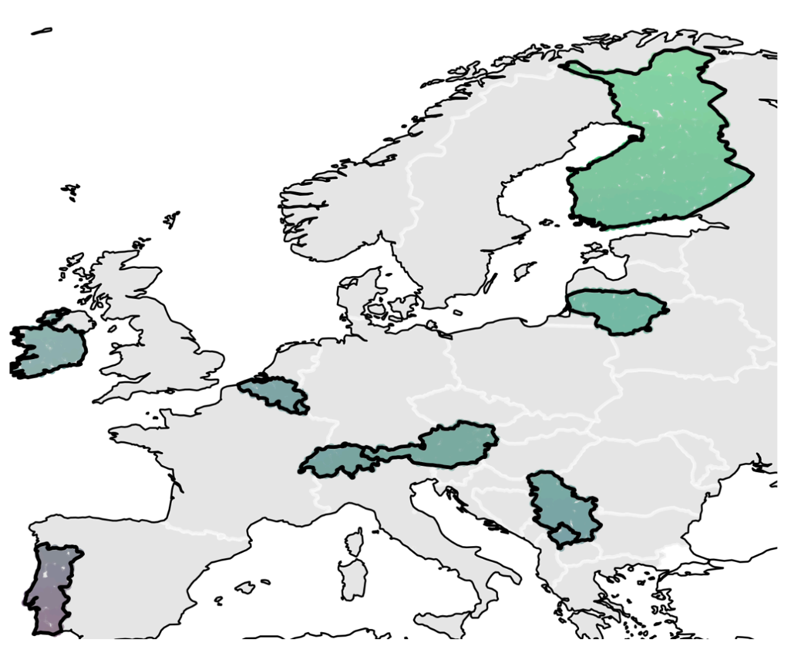


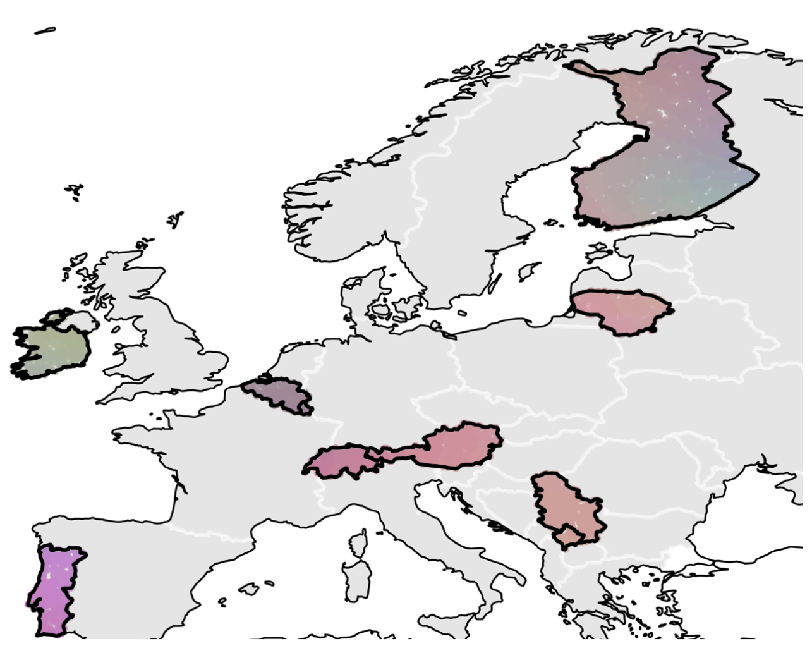
When Evaluated OOD:

4.1% Improvement with the Prior [2] on EnviroAtlas across OOD cities

3.1% Improvement on test F1 with an auxiliary SatCLIP token on BigEarthNetv2.0 using [TOKEN-FUSE]

⚠ Arbitrarily learned inputs can hurt GeoML OOD and Label-Efficiency!





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: Avg Prec/F1 with Frozen (F) vs Register [3] vs Fine-Tuned (FT) SatCLIP auxiliary token on BigEarthNetv2.0

Finding: Learned embeddings when [TOKEN-FUSE] when fine-tuned become highly localized to countries covered in train split; global context of multi-modal input is lost!

References

[1] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. SatCLIP: Global, general-purpose location embeddings with satellite imagery. AAAI 2025.

[2] Esther Rolf, Nikolay Malkin, Alexandros Graikos, Ana Jojic, Caleb Robinson, and Nebojsa Jojic. Resolving label uncertainty with implicit posterior models. UAI 2022

[3] Timoth´ee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. ICLR 2024