FOUNDATION MODELS IN REMOTE SENSING: INSIGHTS FROM MULTISPECTRAL AND HYPERSPECTRAL SELF-SUPERVISED LEARNING

Nassim Ait Ali Braham

EO Data Science, Remote Sensing Technology Institute, DLR

Data Science in Earth Observation, Technical University of Munich, Germany

DLR

Outline

1. Introduction to Self-Supervised Learning

2. SSL on Sentinel 2 data: a forest-monitoring use-case

3. SpectralEarth: Training hyperspectral foundation models at scale

4. Conclusion

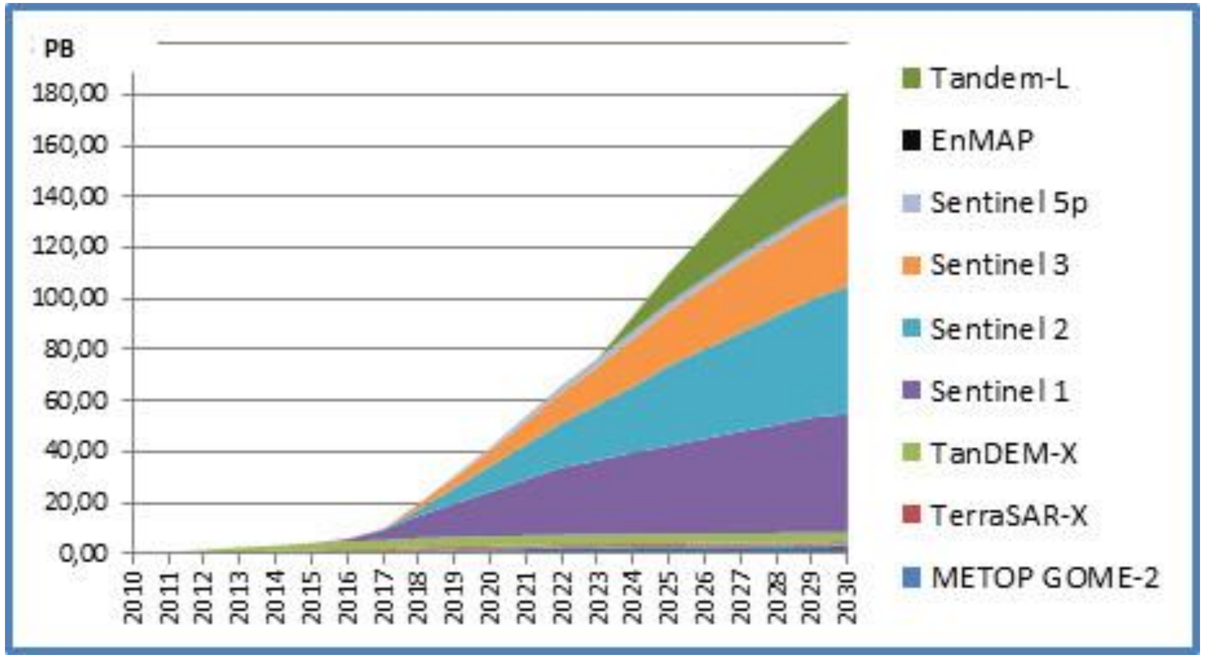
INTRODUCTION TO SELF-SUPERVISED LEARNING

Motivation: Why SSL?

- Deep Learning requires annotated data

- Labeled data is rare
  - Costly to obtain
  - Tedious annotation process

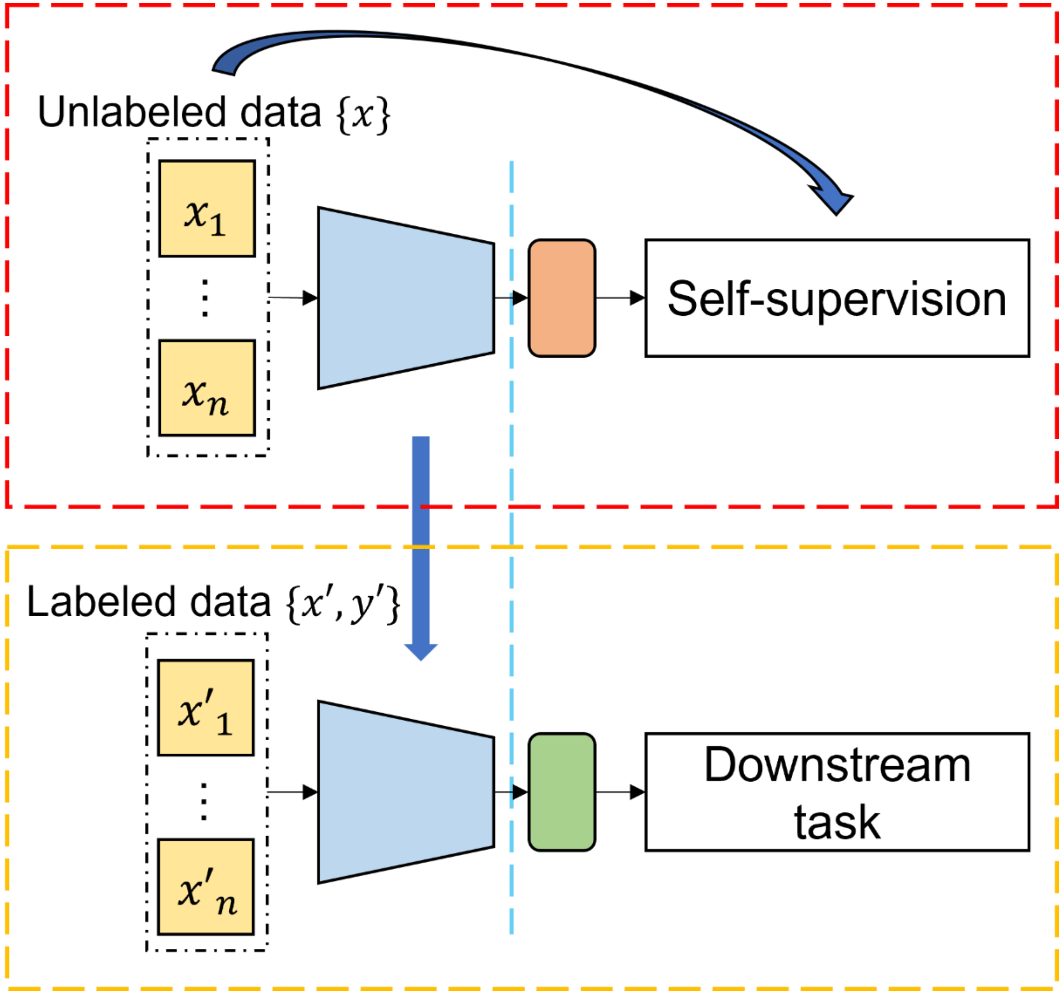
- Unlabeled data is abundant
  - Satellite archives with Petabytes of data



How to exploit unlabeled data for deep learning with RS image analysis? → **Self-Supervised Learning**

What is SSL?

- **Goal**
  - Obtain *training feedback* from the data itself
  - Learn representations in a self-supervised fashion
    - no human annotation
- **Why?**
  - A pre-trained model can be transferred to downstream tasks
  - Improve **accuracy** and **label efficiency**



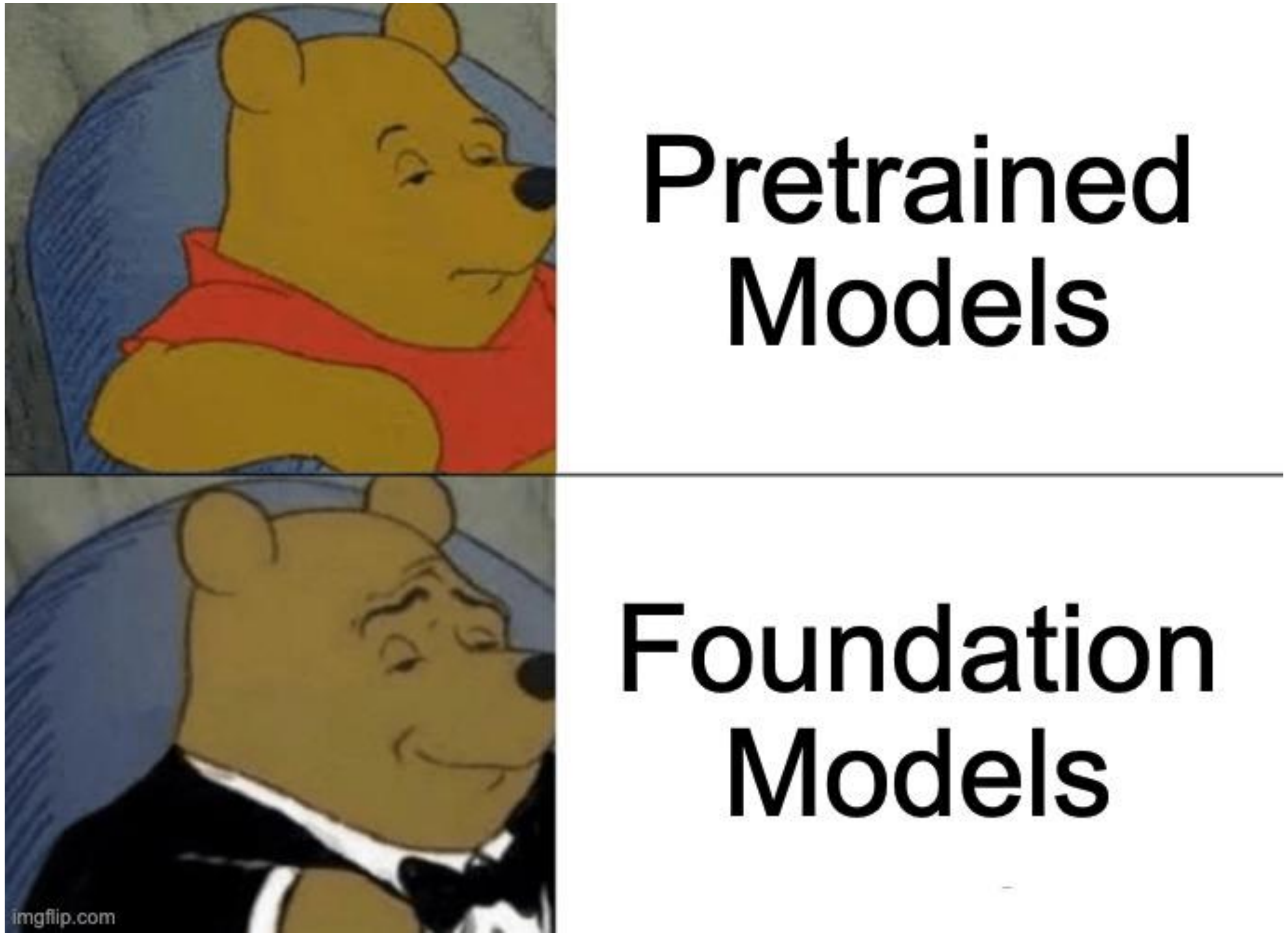
Overview of Self Supervised Learning

Foundation Models

- Foundation models, latest buzzword in the AI sphere

- Foundation models = Big Architecture + SSL algorithm + a lot of data

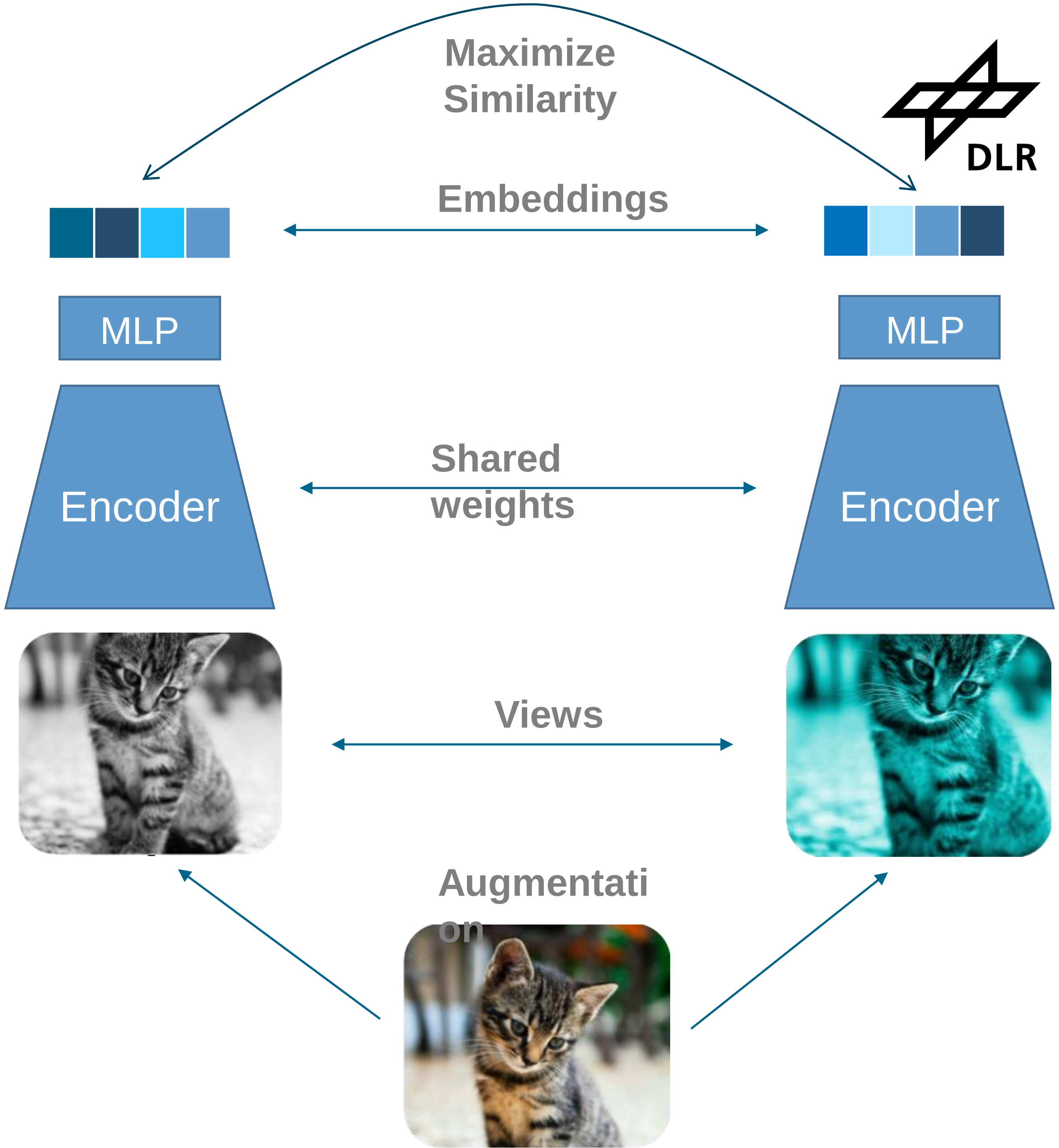
- SSL algorithms
  - Contrastive methods
  - Masked Image Modeling
  - …



Jia-Bin Huang auf X: „Making pretrained models cool again! https://t.co/puJA3zUJzG“ / X

Contrastive Learning

- **General idea**
  - Siamese architecture with shared parameters
  - Similar images (views) are generated using **data augmentation**
  - Enforce **invariance** to the augmentations
- Problem: a **constant** function is invariant (collapse)
- Mitigating collapse
  - Negative sampling: MoCo, SimCLR
  - Clustering: SwAV
  - Knowledge distillation: BYOL, SimSiam, DINO
  - Redundancy reduction: BarlowTwins, VICReg

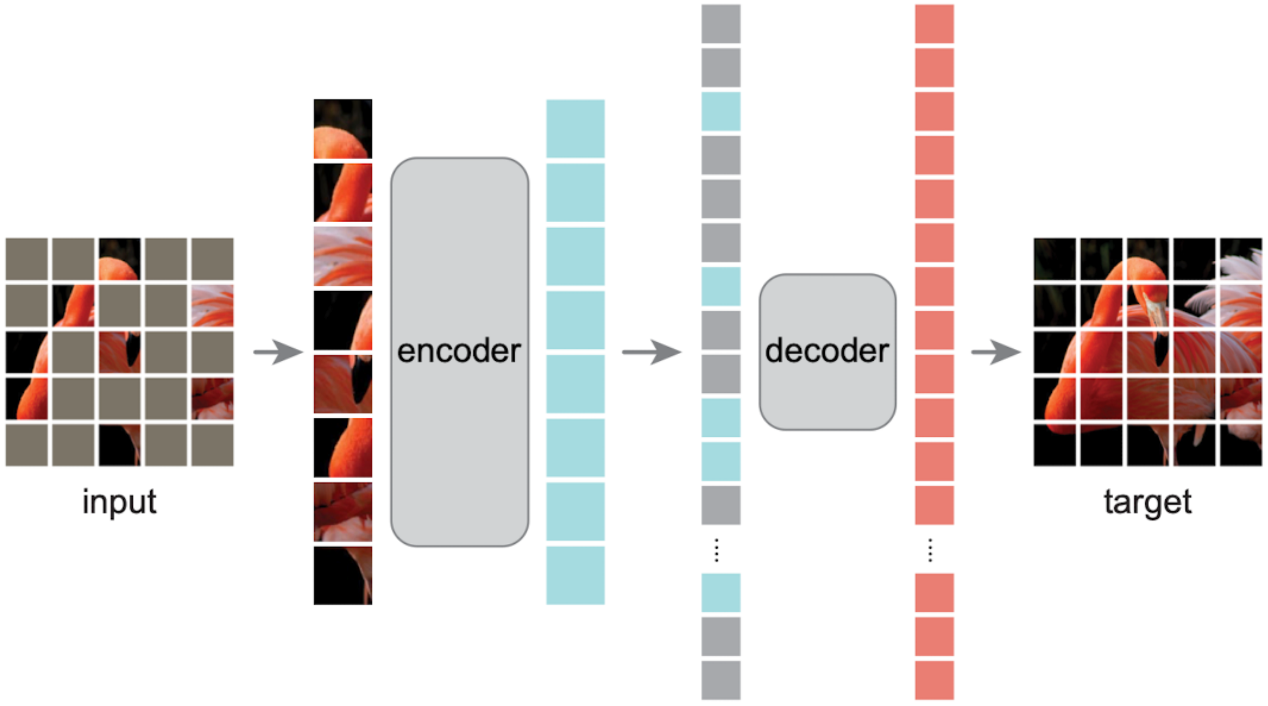


Masked Image Modeling

- **General idea**
  - Predict missing patches from visible ones
  - Typically high masking ratio ($\sim 75\%$)

- **Prediction targets**
  - Raw pixels: MAE
  - Hand-crafted features: MaskFeat
  - Visual tokens: BEiT
  - Latent representations: data2vec

- Generally used with Transformer backbones



A Schematic overview of Masked Autoencoders*

*He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Contrastive vs Masked Image Modeling

- **Contrastive Learning**
  - \+ Highly semantic features, great for classification tasks
  - \+ Architecture agnostic
  - \+ Competitive results on ImageNet
  - \+ Can require a large batch size
  - \+ Requires having good augmentations
  - \+ Special care for negative samples/collapse

- **Masked Image Modeling**
  - \+ Conceptually simple, no positive/negative pairs
  - \+ Masking generally reduces pre-training time
  - \+ Competitive results on ImageNet
  - – Requires Transformer backbone
  - – Lower-level features => requires fine-tuning, poor linear performance

Ongoing efforts to combine the benefits of both approaches

SSL in RS

- A lot of research happening in the field
- > 100 foundation model papers in the past few years
- Predominantly for multispectral and high resolution RGB imagery
- Little work in the hyperspectral domain
- A trend towards multi-sensor foundation models

SSL ON SENTINEL 2 DATA: A FOREST-MONITORING USE-CASE

Evoland

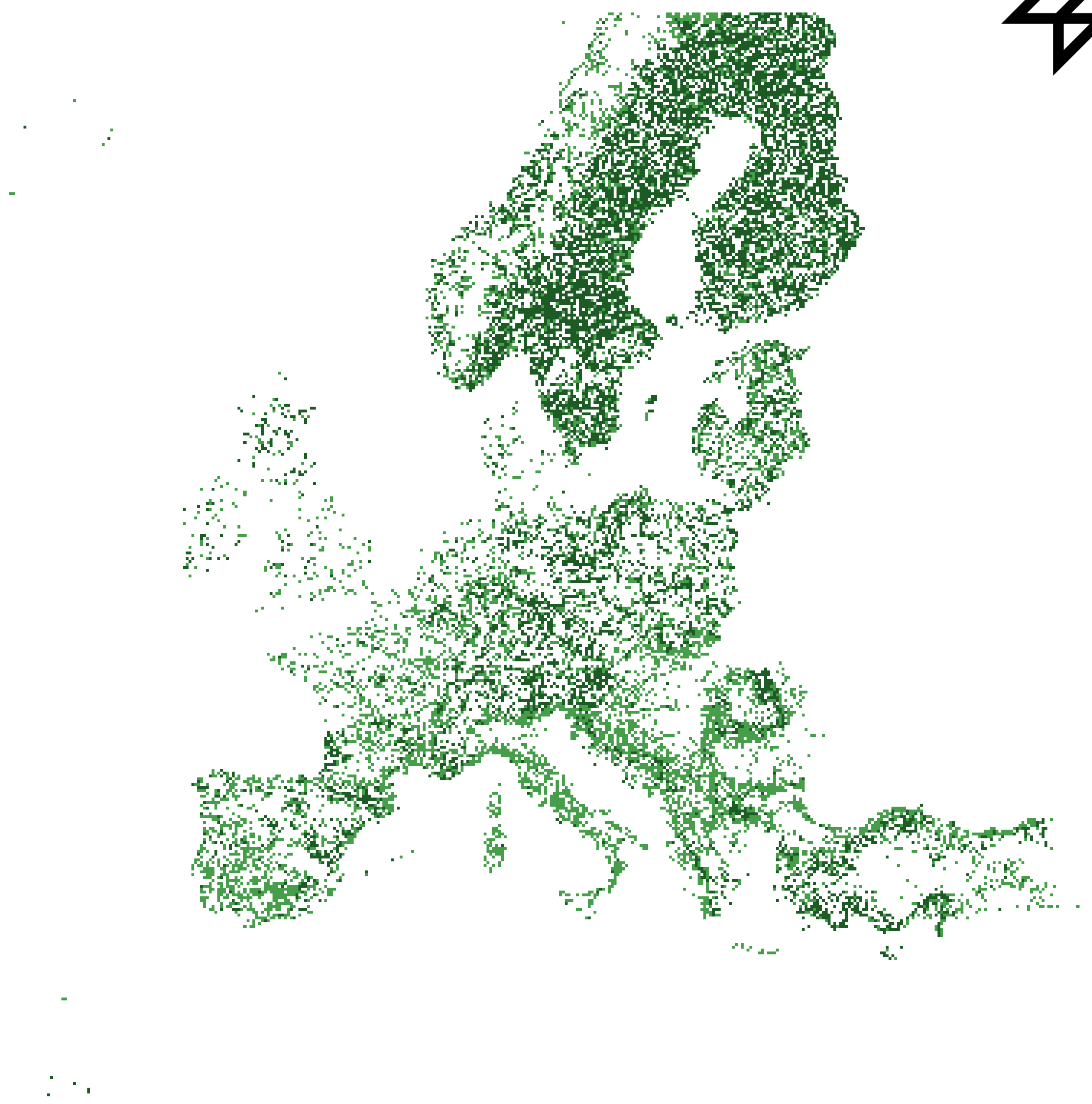
- Goals
  - Improve/extend existing Copernicus Land Monitoring Service products
  - Leverage ML for land surface continuous monitoring
  - Application to agriculture, forest, water, urban and general land-cover

Evoland

LAND MONITORING EVOLUTION

Evoland: Forest Use Case

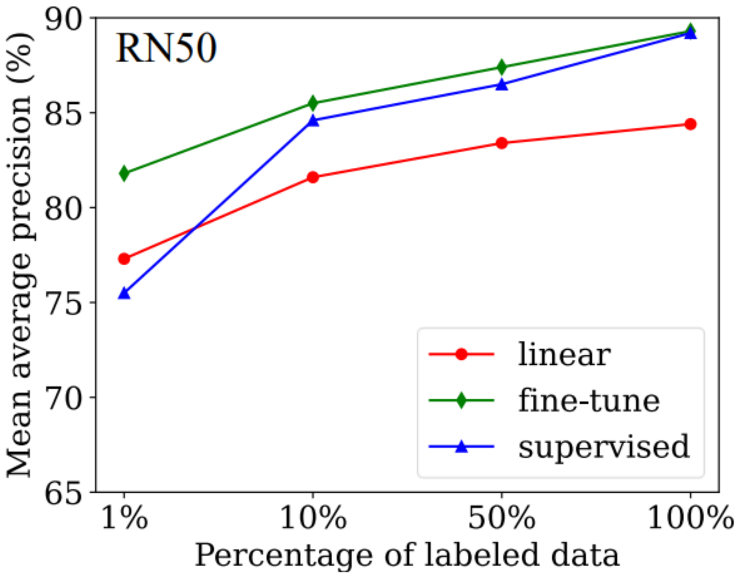
- **Goal:** Increase temporal frequency for forest monitoring
- **Input:** Single Sentinel 2 timestamp
- **Output:** Binary tree masks, tree density, forest disturbance



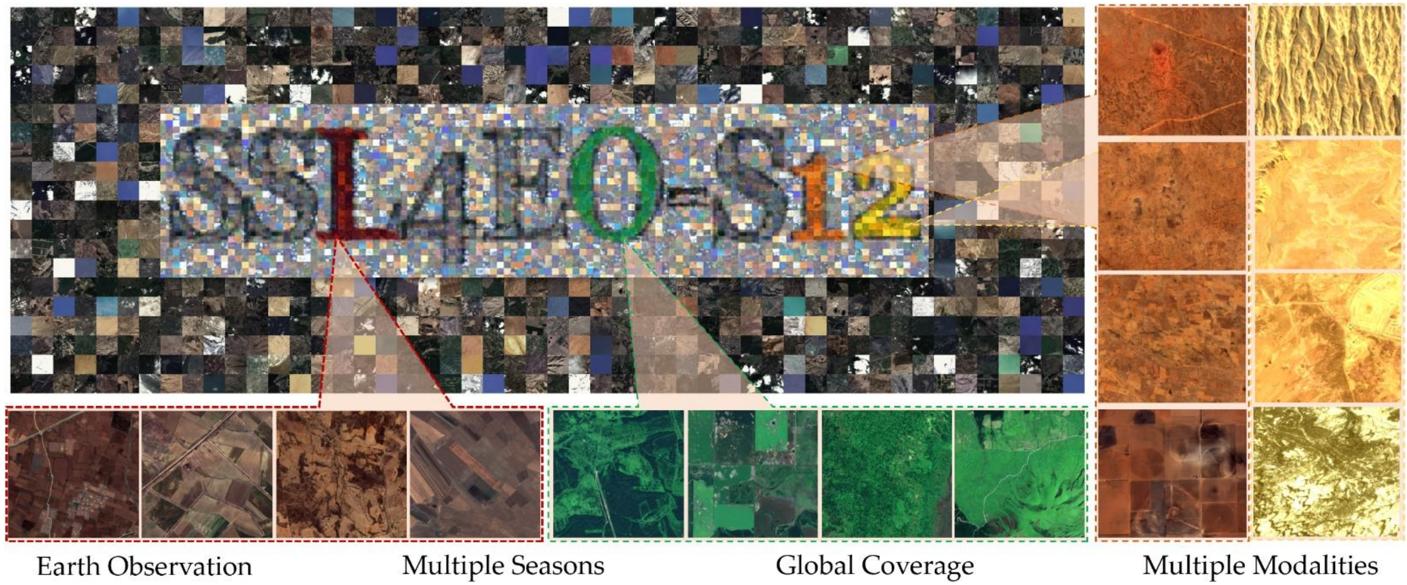
From [Dominant Leaf Type 2018 — Copernicus Land Monitoring Service]()

SSL4EO-S12

- ~250,000 S2-S1 patches
- 264x264 pixels
- 1.5TB of data
- 4 timestamps per location



Results on BigEarthNet: Pre-training improves performance and label efficiency



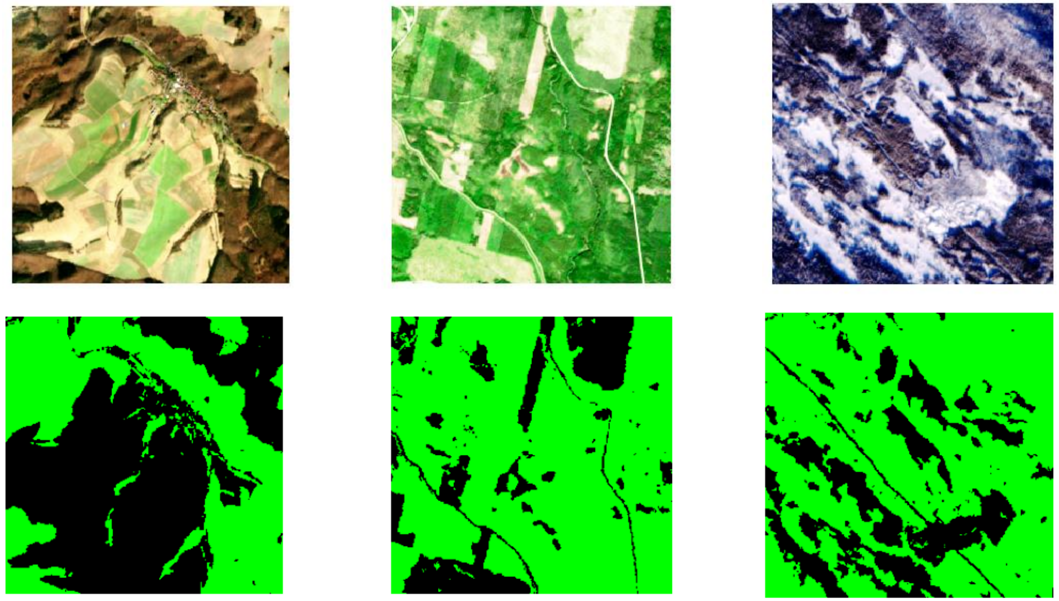
Wang, Y., Braham, N. A. A., Xiong, Z., Liu, C., Albrecht, C. M., & Zhu, X. X. (2023). SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, *11*(3), 98-106.

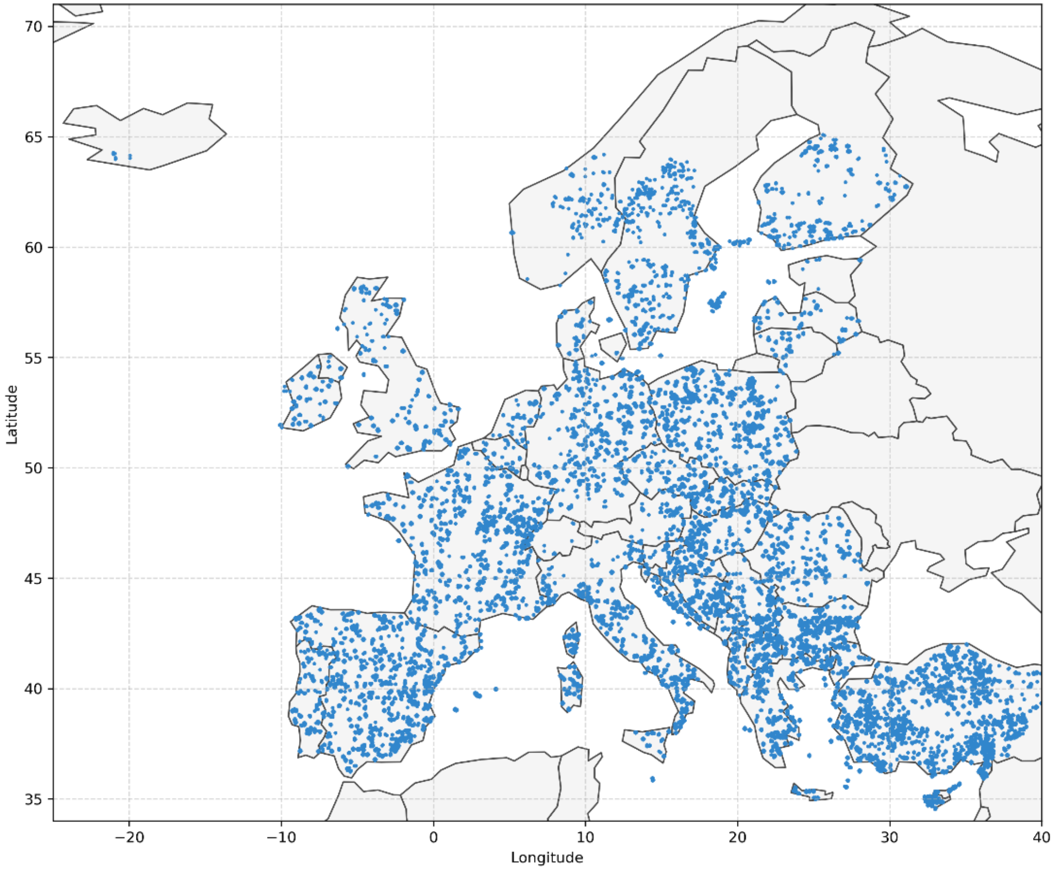
SSL4EO-EU-Forests

- **~16,000** locations
- **4 seasons**
- Sentinel 2 images, HLR 2018 mask

S2 Images

HLR Masks





Geographical distribution of the SSL4EO-EU-Forest dataset

Initial Results

- Pre-training consistently improves the results
- ResNet-50 does not improve upon ResNet-18
- Similar performance for ViT and ResNet
- UNet never gets old

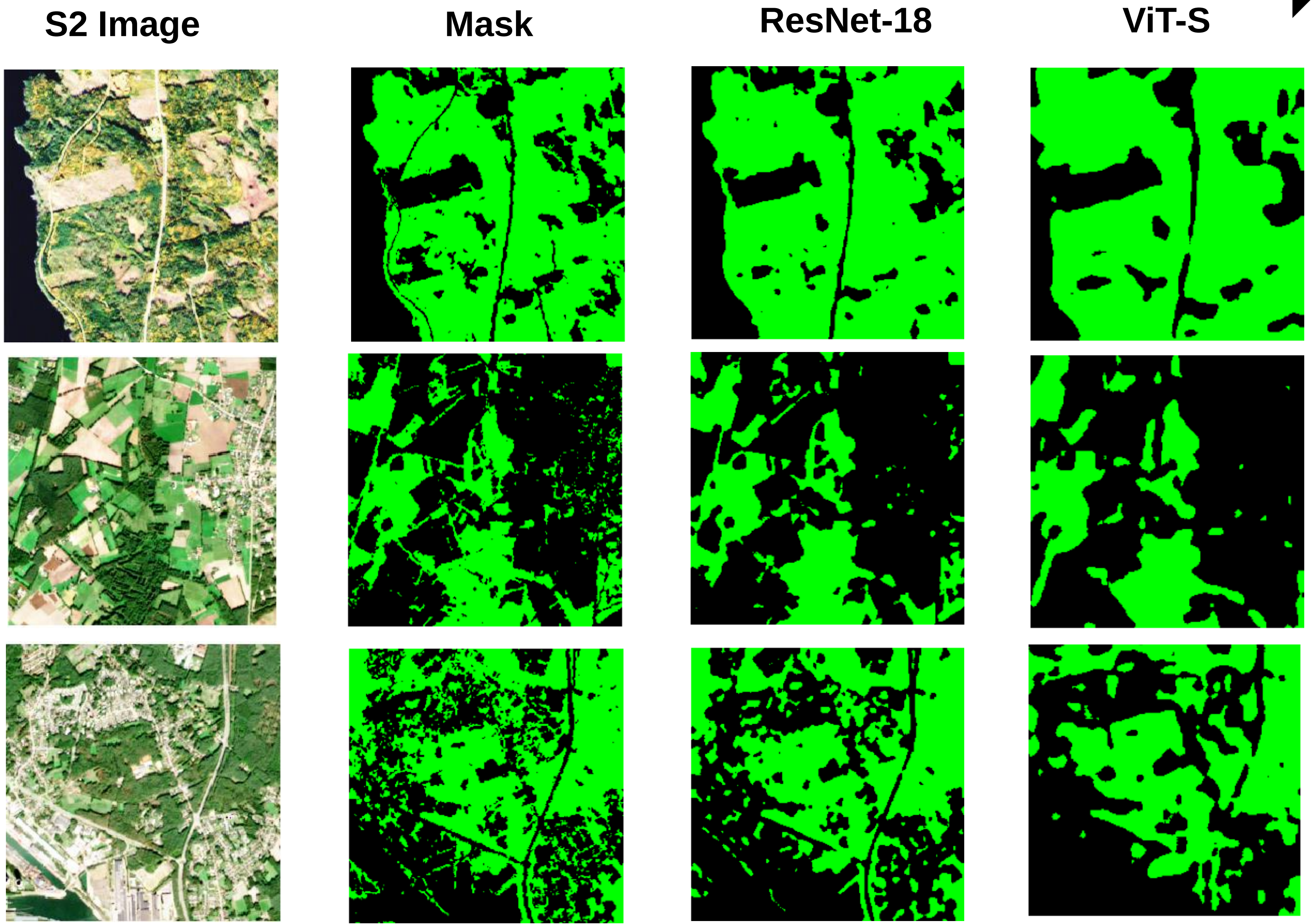
Segmentation Protocol	Encoder	Pre-training Weights	Overall Accuracy	Mean IoU
UNet	ResNet-18	Random	85.58	75.19
		MoCo	**88.03**	**78.61**
		DINO	88.72	79.72
	ResNet-50	Random	85.69	74.97
		MoCo	**88.68**	**79.66**
		DINO	88.18	78.85
DeepLabV3+	ResNet-18	Random	84.89	73.95
		MoCo	87.37	77.58
		DINO	**87.82**	**78.29**
	ResNet-50	Random	84.73	73.65
		MoCo	**88.14**	**78.80**
		DINO	87.59	77.92
UpConv	ViT-S	Random	86.35	76.03
		MoCo	87.38	77.59
		DINO	**88.57**	**77.49**

Fine-tuning results after 100 epochs

Qualitative assessment

Loss of fine-grained features!

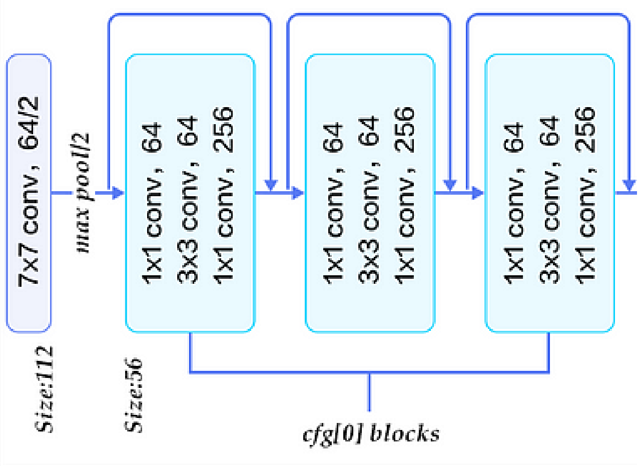
Similar scores for ResNet-18 and ViT-S, different visual appearance



Improving details preservation

- **Architecture:** ResNet Stem layer downscales the image by a factor of 4

Remove the pooling and set stride to 1
Introduce a stride of 2 in the 1st residual block

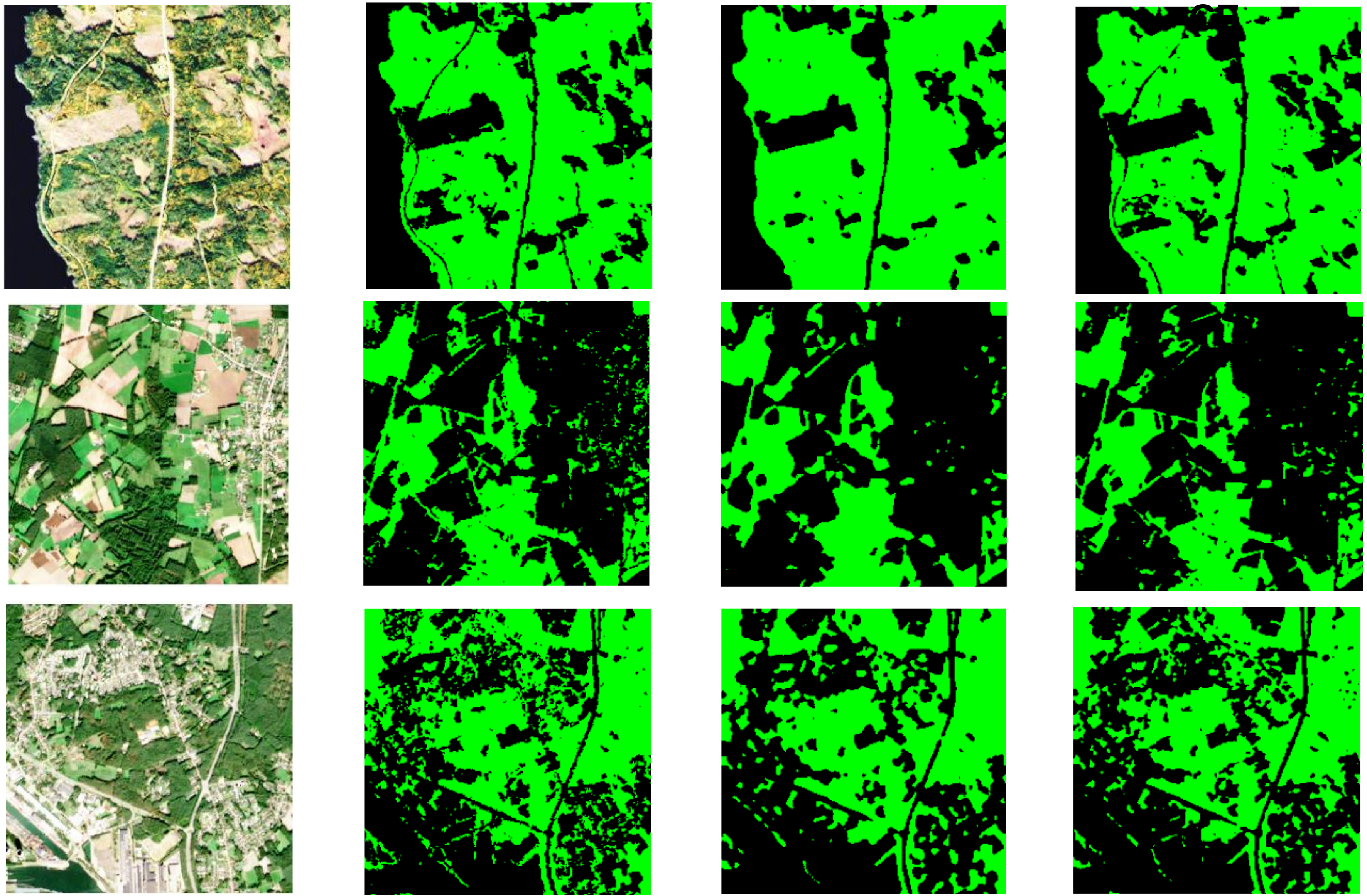


- **Loss function:** Fine-grained features are diluted in the cross-entropy loss

Put a higher weight on the boundary pixels of the mask in the loss

Improved Results

S2 Image | Mask | ResNet-18 | Custom ResNet-18 + Weighted



Refined outputs! Yet, no significant change in mIoU/accuracy

How Practical are Foundation Models?

Advantages

- Strong generalization capabilities
- Little to no fine-tuning needed, works out of the box
- Label efficiency
- Cool branding

Limitations

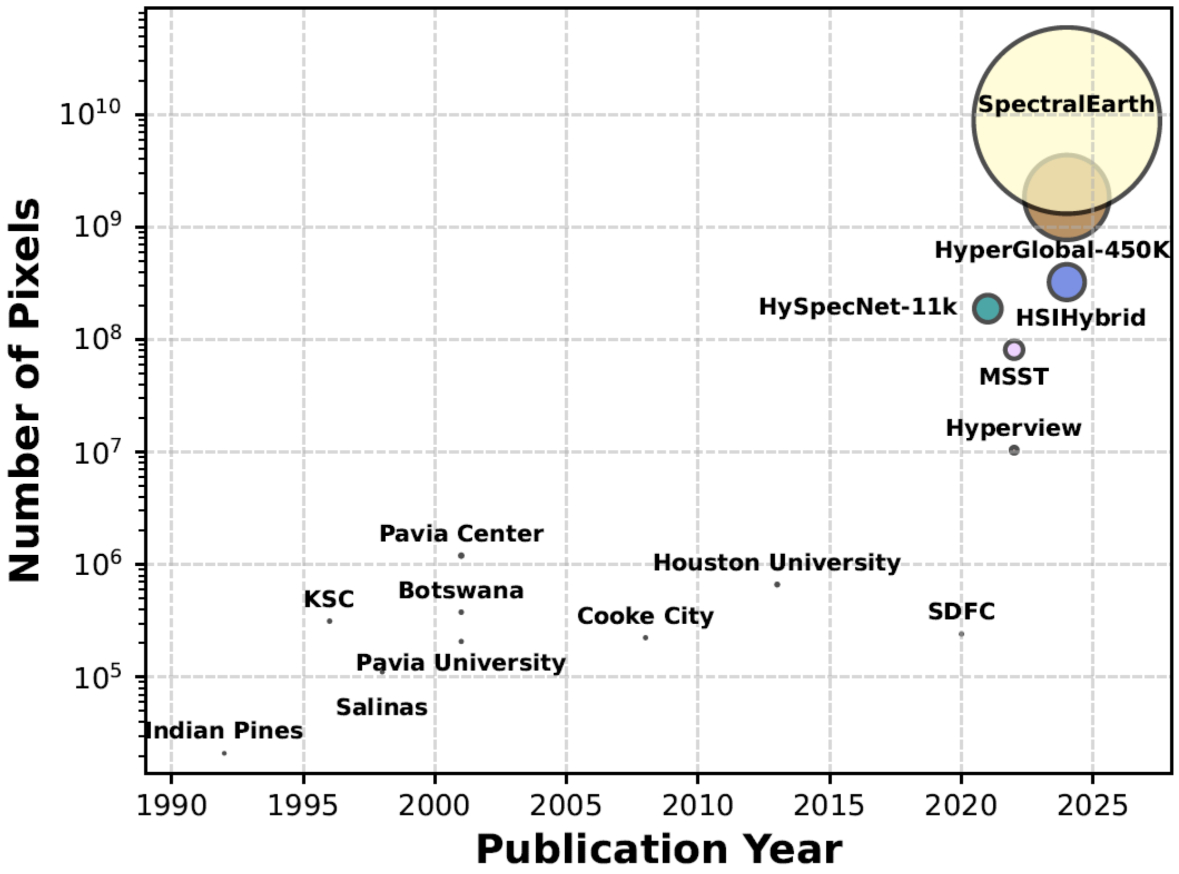
- High inference cost
- High memory cost
- Good in many tasks, not necessarily the best in any
- ViT limitations for pixel-level tasks
- Still requires some labels

What can we do to make SSL/foundation models more useful for real-world applications?

SPECTRALEARTH: TRAINING HYPERSPECTRAL FOUNDATION MODELS AT SCALE

Motivation

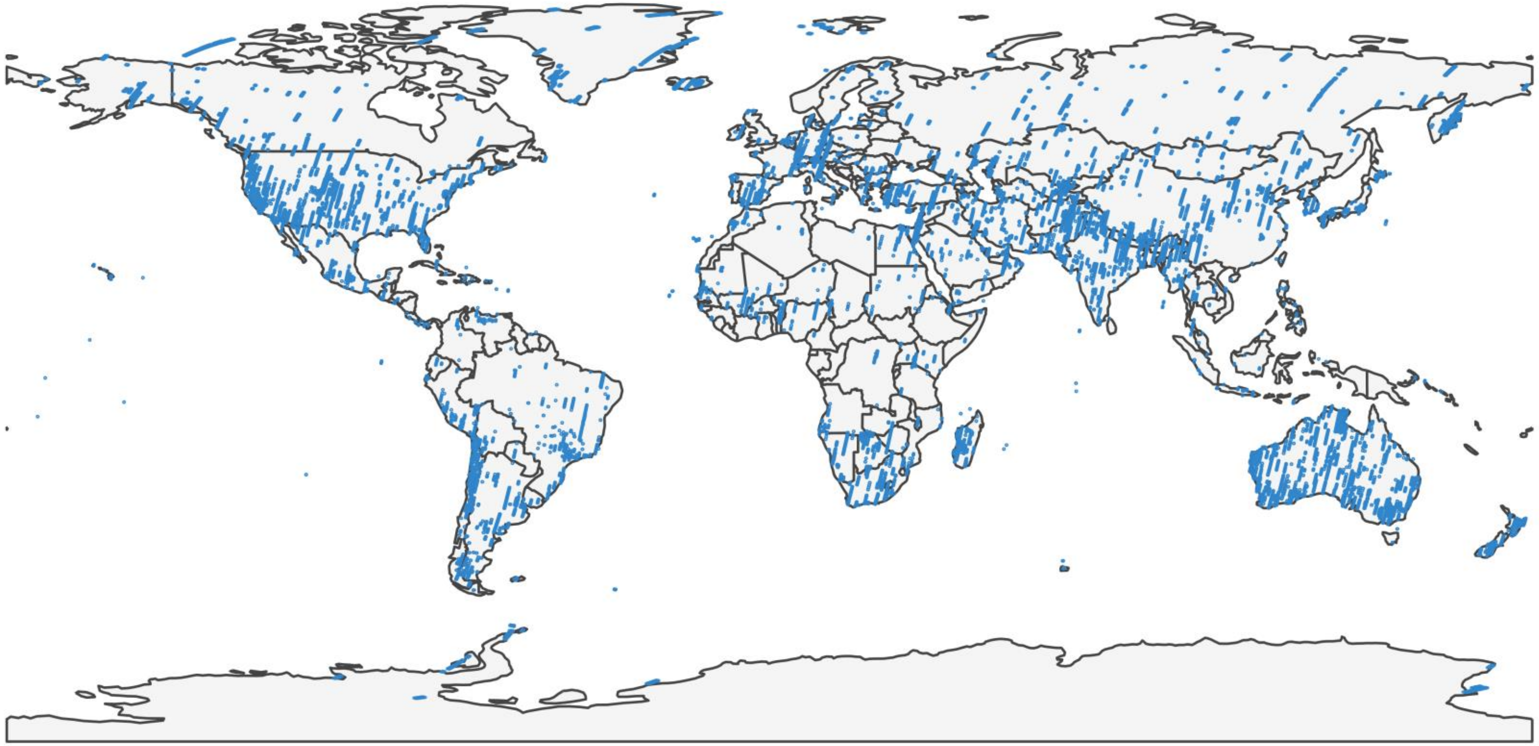
- A lot of research on foundation models for **MSI**: SatMAE, ScaleMAE, Prithvi, DOFA, SkySense, etc.
- Less research on foundation models in **HSI**
- **No suitable dataset for pre-training hyperspectral foundation models**
- **Contribution: SpectralEarth** a globally distributed dataset, pre-trained models and benchmark



https://doi.org/10.48550/arXiv.2408.08447

SpectralEarth: A large-scale HSI dataset

- Based on *EnMAP* imagery
- 30m resolution, 202 bands
- **~538,974** patches, 128x128 pixels.
  - **~415,153** unique locations
  - **~73,000** locations with > 1 timestamp
  - Sampled from **11,636** tiles
- **~3.3 TB** of data
- Mostly cloud free



Geographical distribution of SpectralEarth

Creating the dataset

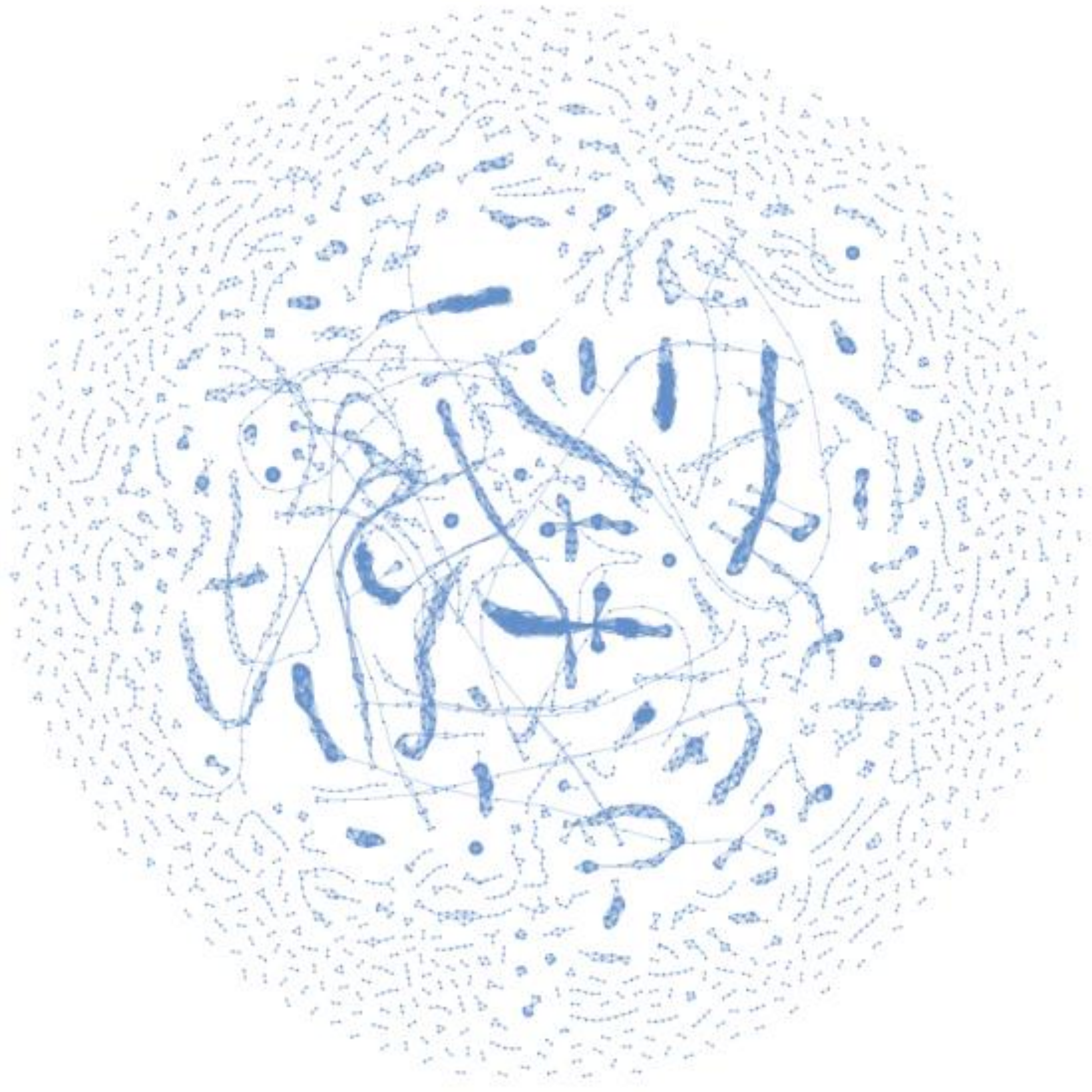
- Input: **~11K** EnMAP tiles

- Ideally, we want to maximize the # of patches with temporal positives

- The longer the time series, the better

=> Prioritize the **areas of overlaps**, prioritize areas with **higher degrees of overlap**

- More costly than I initially expected
  - Some tiles have degree > 30



A graph representing EnMAP tiles overlaps: nodes are tiles, two nodes are connected iff the two tiles overlap

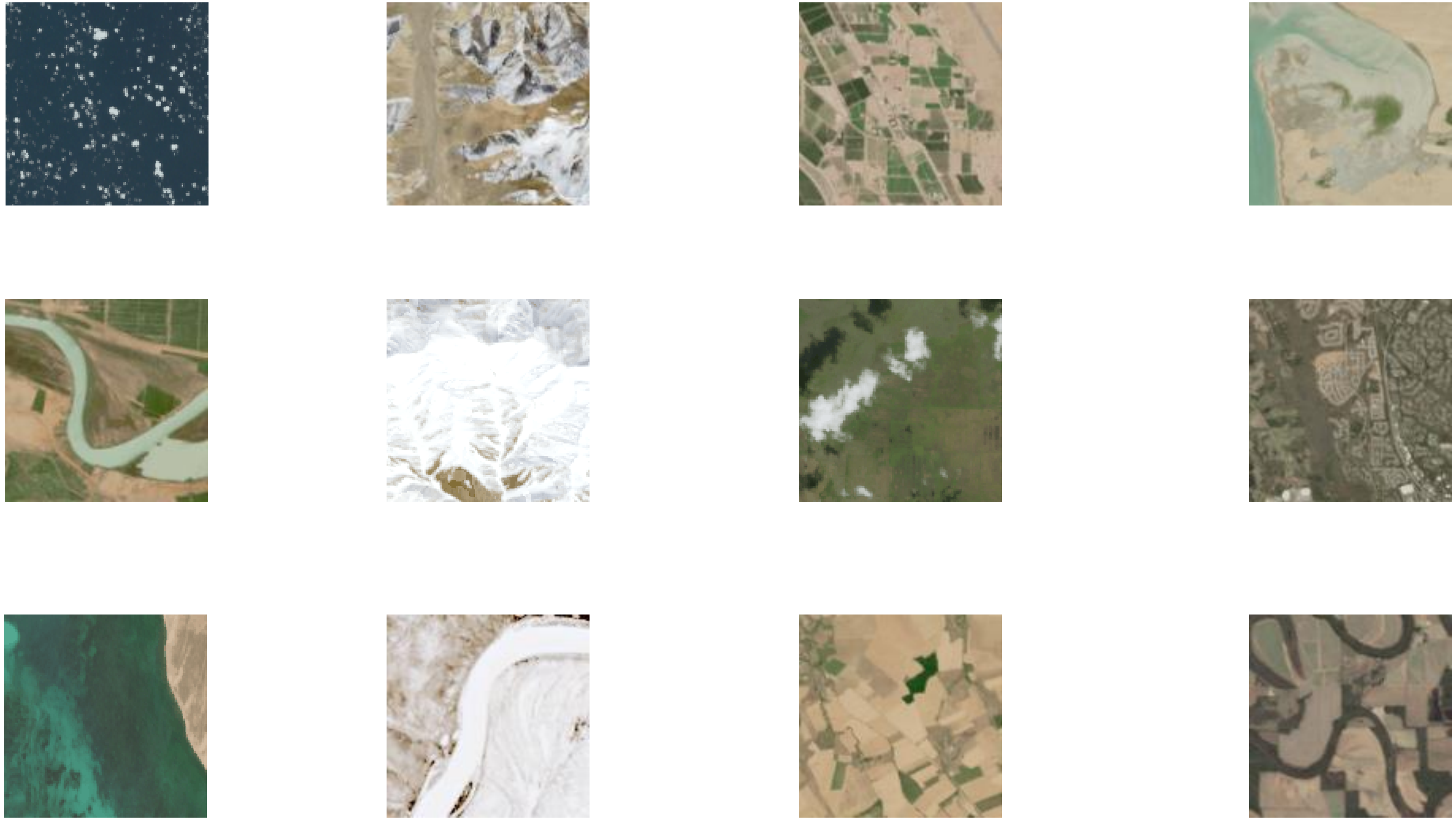
Patchifying the data

- Simple pipeline, but a lot of nasty details
- Annoying details: NaN values, duplicate tiles, projections…
- A lot of time optimizing the script: reducing # combinations, avoiding redundant computation, more efficient overlap checking, reducing I/O, parallelizing the script over connected components…

Algorithm 1 Temporal Views Extraction

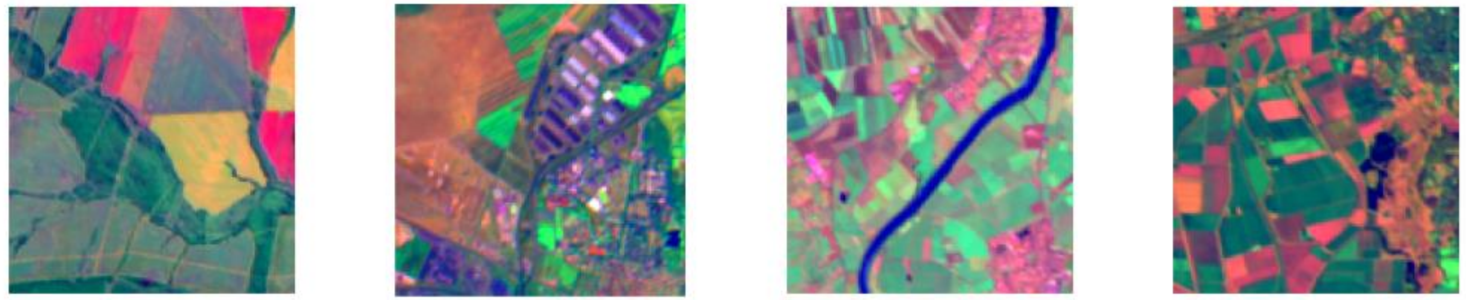
```
procedure MAIN PROCEDURE
    tiles ← EnMAPData
    overlap_graph ← GETOVERLAPS(tiles)
    R_tree ← K_0        ▷ empty tree, for SpectralEarth patches
    for tile in tiles do
        combs ← BUILDCOMBINATIONS(tile, overlap_graph)
        for tile_subset in combs do
            intersection ← INTERSECTION(tile_subset)
            patches ← PATCHIFY(intersection)
            UPDATE(R_tree, patches)
        end for
    end for
end procedure

function BUILDCOMBINATIONS(tile, overlap_graph)
    combinations ← GETEDGES(tile, overlap_graph)[2]
    for subset size n in [3, 4, . . . ] do
        get n-tuples from (n-1)-tuples in combinations
        compute intersections of all n-tuples
        keep largest n-tuples by area
        if no valid n-tuple found then
            break
        end if
        add n-tuples to combinations
    end for
    return combinations
end function
```

Samples from SpectralEarth



Downstream Tasks

- Paired EnMAP imagery with Land Cover and Crop Type products
  - **CORINE:** Multi-label land cover classification
  - **CDL:** Crop type segmentation
  - **NLCD:** Land cover segmentation



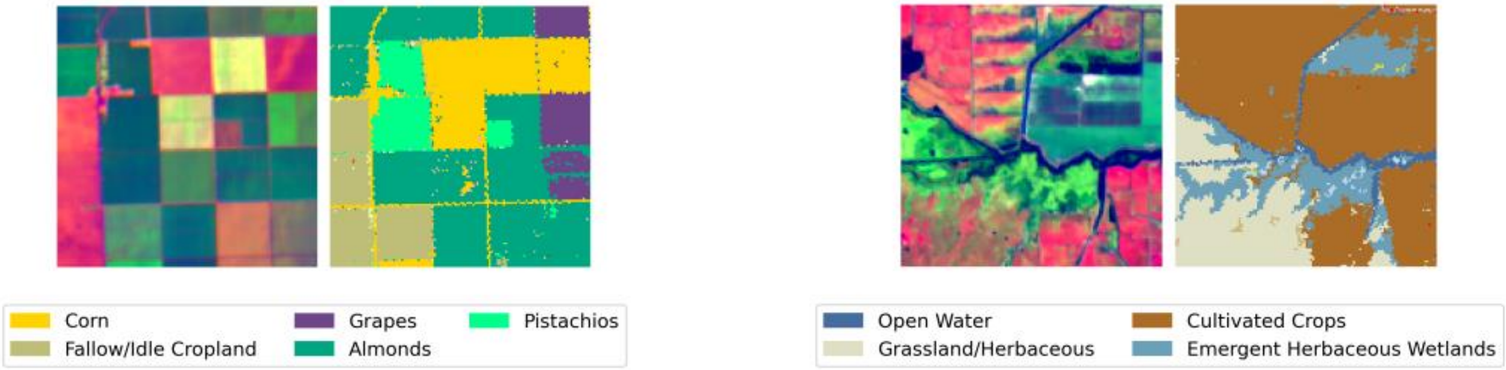
(a) **Classes**: Arable land, Coniferous forest, Moors, heathland, sclerophyllous vegetation.

(b) **Classes**: Urban fabric, Industrial units, Arable land, Natural grassland, sparsely vegetated areas.

(c) **Classes**: Urban fabric, Arable land, Permanent crops, Complex cultivation patterns, Inland waters.

(d) **Classes**: Urban fabric, Arable land, Complex cultivation patterns, Coniferous forest.

Figure 4. Sample pseudo-RGB images of the curated EnMAP-CORINE multi-label classification benchmark.



(a) EnMAP-CDL

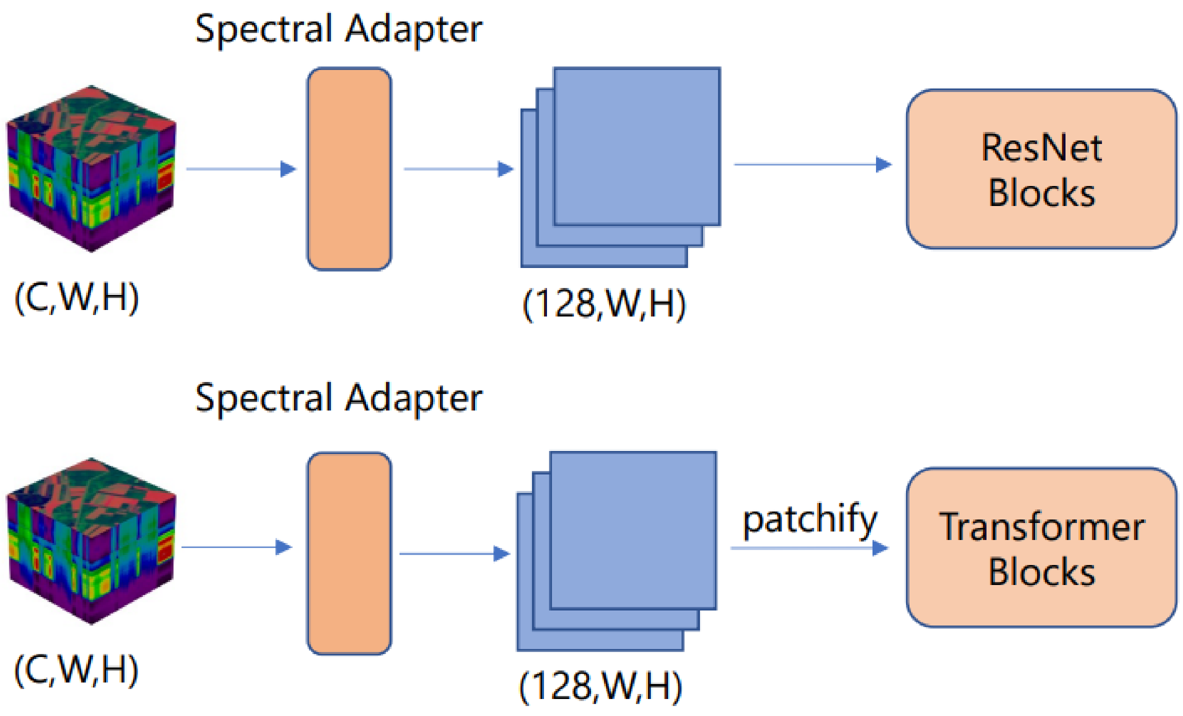
(b) EnMAP-NLCD

SpectralEarth downstream tasks

Models

- **Network Architectures**
  - Simple variation of classical CNN and Vision Transformer architecutres
  - 1D convolutions to extract spectral features
  - Models ranging from **22M** to **1.1B** parameters

- **3** SSL Algorithms
- **> 10** pre-trained models



Backbone architectures

Results: Comparing SSL Algorithms

- **DINO** and **MoCo** perform well in **frozen encoder** evaluation
  - Little benefit when fine-tuning
- **MAE** is competitive in segmentation tasks, and improves fine-tuning performance
- **ConvNets** are not out of the game

Evaluation protocol (# of trainable params.)	Init Weights	EnMAP-CORINE (F1 Score)		EnMAP-CDL (mIoU)		EnMAP-NLCD (mIoU)	
		Spec. RN50	Spec. ViT-S	Spec. RN50	Spec. ViT-S	Spec. RN50	Spec. ViT-S
Frozen Encoder (0)	Random	70.53	70.42	44.72	46.52	35.86	36.85
	MoCo-V2	73.97	73.60	**51.66**	50.37	**41.98**	39.85
	DINO	**76.64**	**75.06**	51.53	51.01	41.77	40.31
	MAE	–	72.72	–	**51.37**	–	**41.17**
Full Fine-tuning (>20M)	Random	78.31	77.78	57.53	55.07	**48.18**	45.95
	MoCo-V2	**78.57**	78.40	**58.10**	55.84	48.09	45.78
	DINO	77.98	78.34	57.77	55.70	47.75	45.71
	MAE	–	**78.66**	–	**57.66**	–	**47.82**
Fine-tune Adapter (56K)	MoCo-V2	76.27	76.12	**55.36**	54.37	**44.66**	43.40
	DINO	**78.43**	**77.95**	55.26	53.50	44.41	43.00
	MAE	–	76.80	–	**54.61**	–	**43.92**

Results: Large Vision Transformers

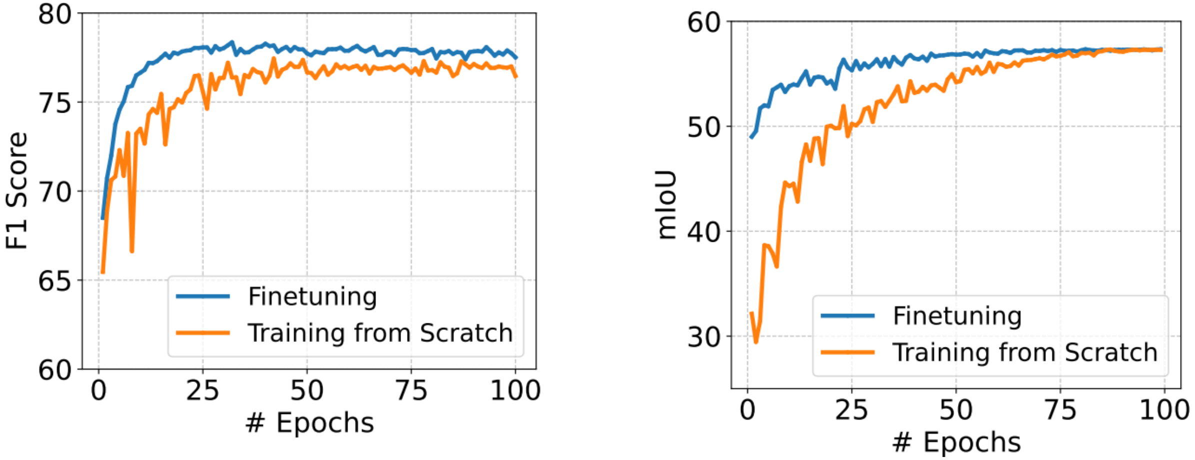
- **MAE** with large ViTs always improves the results
  - Fine-tuning the Spectral Adapter sometimes outperforms training from scratch

- Modest improvements from increasing model size
  - **Large** ViTs require **very large** datasets

Evaluation Protocol	EnMAP-CORINE (F1 Score)				EnMAP-CDL (mIoU)				EnMAP-NLCD (mIoU)			
	B	L	H	g	B	L	H	g	B	L	H	g
Training from Scratch	76.99	77.24	**77.28**	76.85	54.79	54.50	**54.83**	54.74	**45.96**	45.62	45.53	45.58
Frozen Encoder	74.72	75.07	**76.06**	75.33	51.20	53.14	**53.19**	52.77	40.52	42.88	**43.32**	42.63
Full Fine-tuning	79.05	79.18	**79.80**	78.38	57.70	**58.19**	58.06	57.86	48.10	**48.37**	48.28	48.08
Fine-tune Adapter	77.09	77.79	**77.94**	77.86	54.79	**55.35**	55.14	54.90	43.97	44.46	44.67	**44.73**

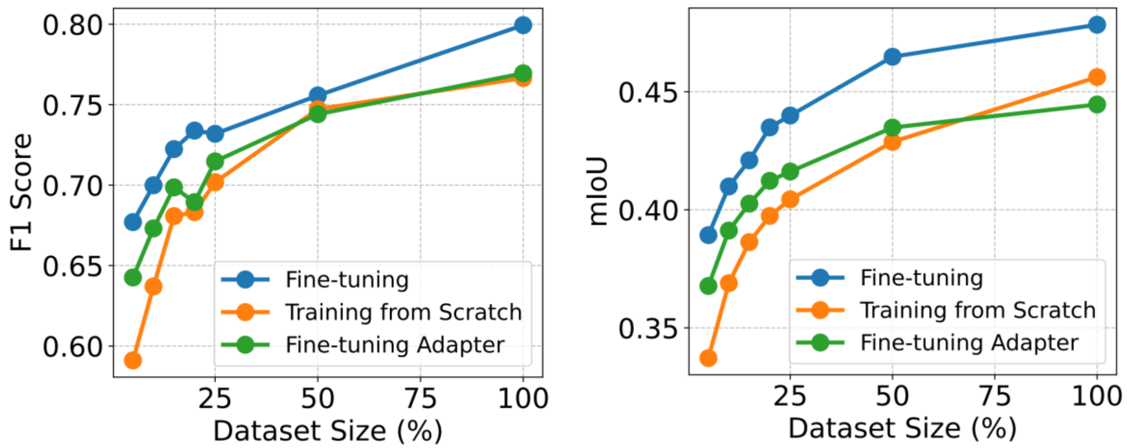
Results: Efficient Training

Pre-trained models converge faster when fine-tuned



Convergence speed: EnMAP-CORINE and EnMAP-CDL

Pre-trained models help when labels are scarse

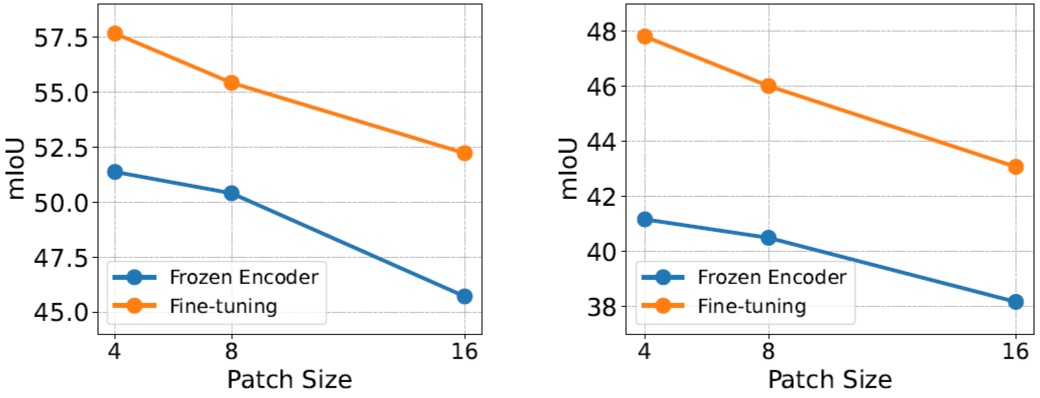


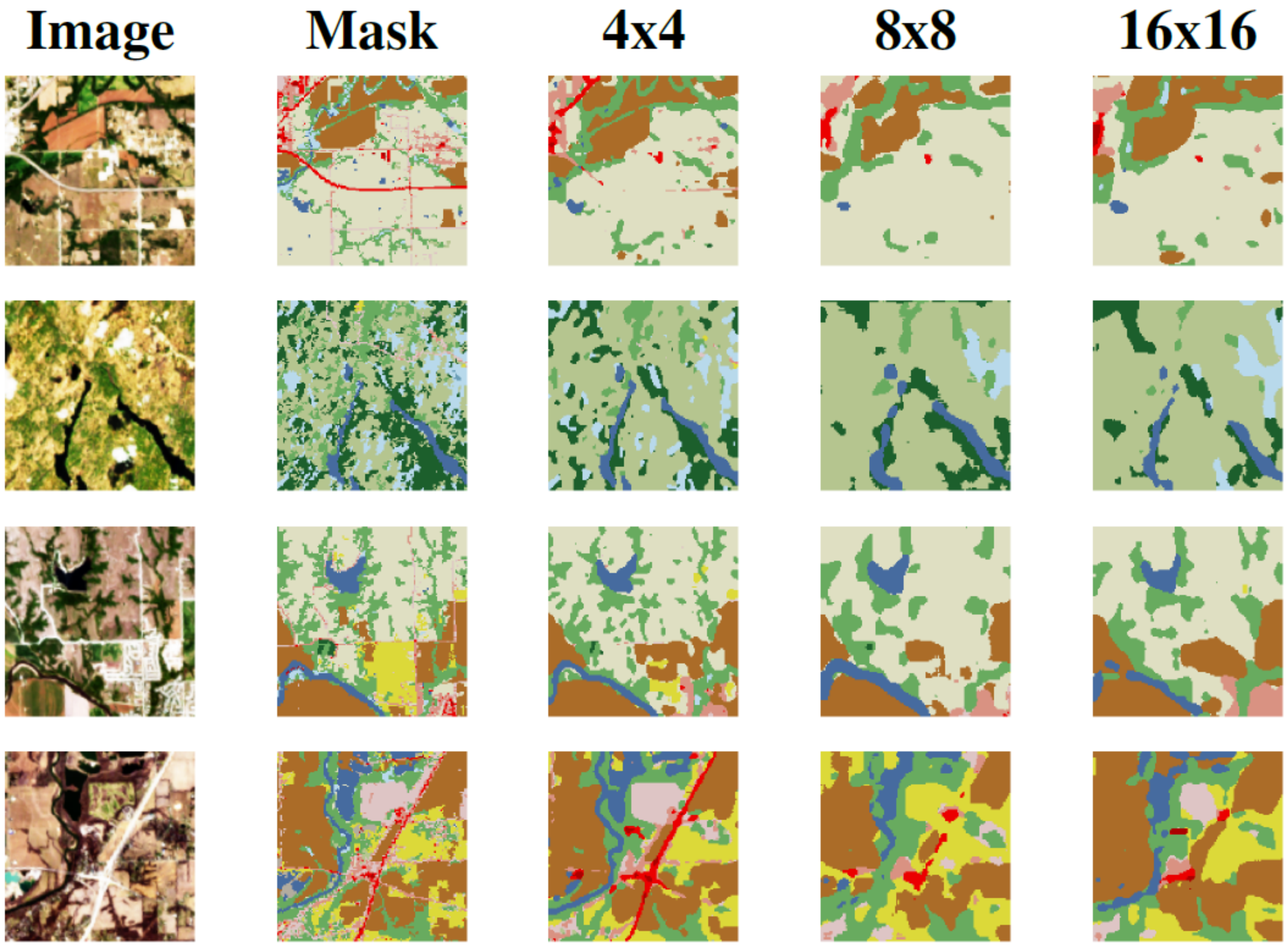
Limited labels setting: EnMAP-CORINE and EnMAP-CDL

ViT Patch Size: An Important Hyperparameter

- Tokens representing smaller patches help preserve finer spatial and spectral details

Frozen encoder eval with varying patch size





Future Directions

- Explore more complex backbone architectures
- Extend the set of pre-training algorithms
- ***SpectralEarth-MM***
  - Pair SpectralEarth with other sensors (Sentinel 2, Sentinel 1, Landsat 8)
  - Investigate multi-sensor pre-training => exploit complementarity of different sensors

Dataset available through EOC Geoservice
https://geoservice.dlr.de/web/datasets/enmap_spectralearth

CONCLUSION

Some Open Questions

- What can we do to make SSL/foundation models more useful for real-world applications? Could model distillation help?
- Specialized models vs. Foundation models, when to resort to each?
- What evaluation protocols are most relevant for evaluating foundation models? Frozen encoder? Full fine-tuning? Partial fine-tuning?
- Are we getting the full picture from benchmark tables? E.g., models with similar mIoU can behave differently
- How far should we chase the ultimate foundation model that can process any sensor (even unseen ones)? What is the right balance between fitting a sensor well and generalizing to as many sensors as possible?

Questions?