S2MAE: A Spatial-Spectral Pretraining Foundation Model for Spectral Remote Sensing Data

Xuyang Li [1,2] **Danfeng Hong** [1,2] **Jocelyn Chanussot** [3]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences
[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
[3] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK

Introduction

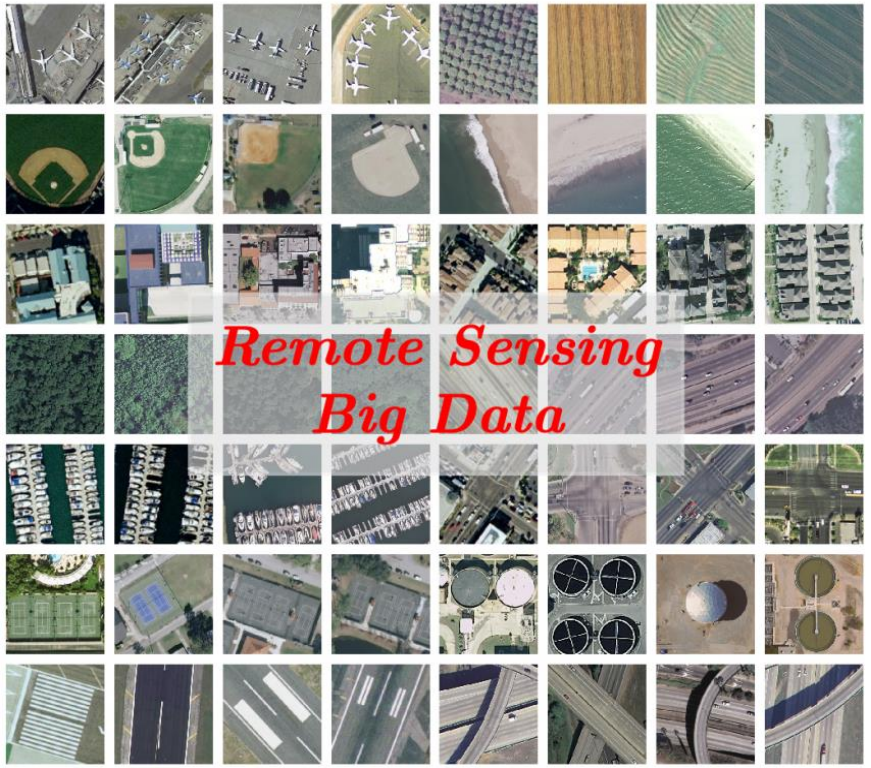
- Remote sensing (RS) data, gathered by satellites or aircraft, capture electromagnetic reflections and emissions from Earth's surface.
- The volume of data expands with the increasing number of satellite launches.

High resolution Optical Images:
WorldView, Gaofen …

Multispectral Images:
Landsat, Sentinel-2 …

Hyperspectral Images:
EnMAP …

SAR Images:
Sentinel-1 …

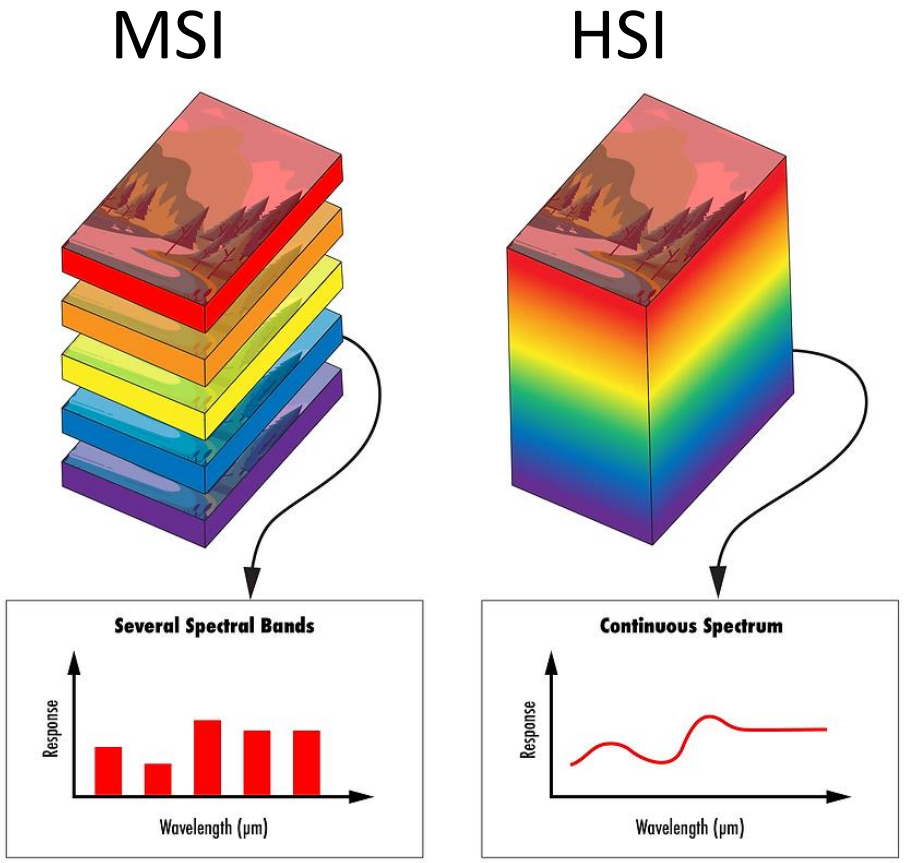


Hong, Danfeng, et al. "SpectralGPT: Spectral remote sensing foundation model." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

Introduction

Spectral Imagery

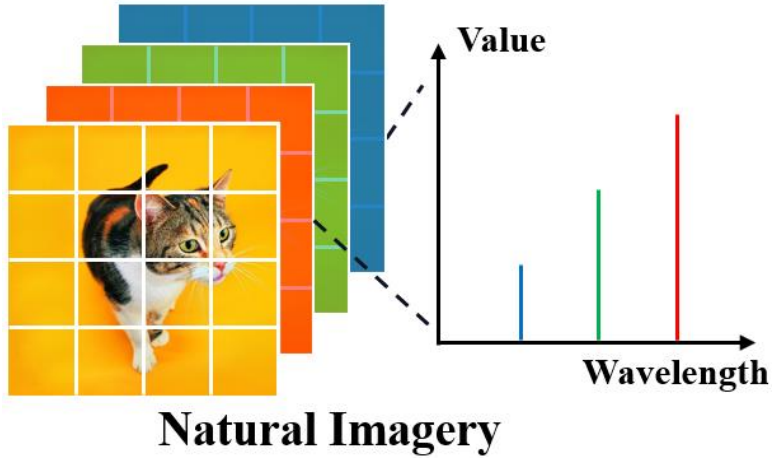
- Each surface material has a unique spectral signature.
- **Multi-spectral Imagery (MSI)**
  contain 10+ bands of wavelength ranges.
- **Hyper-spectral Imagery (HSI)**
  contains spectrum in many contiguous wavebands.
- The **narrower** the range of wavelength in a band, the **finer** its spectral resolution would be.



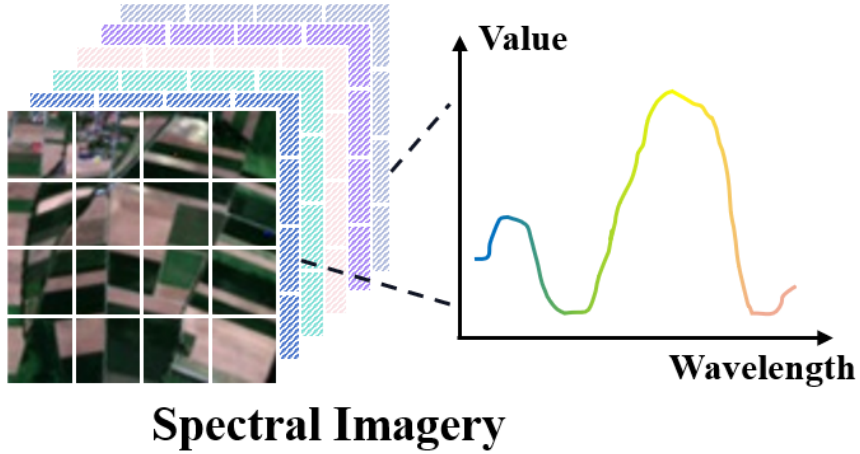
https://www.mapmyops.com/imaging-spectroscopy-exploring-hyperspectral-imaging

Motivation

- **RS Imagery *vs.* Natural Imagery**
  - Distance & Resolutions
  - Subject & Background
  - Multi-Channels & RGB Channels

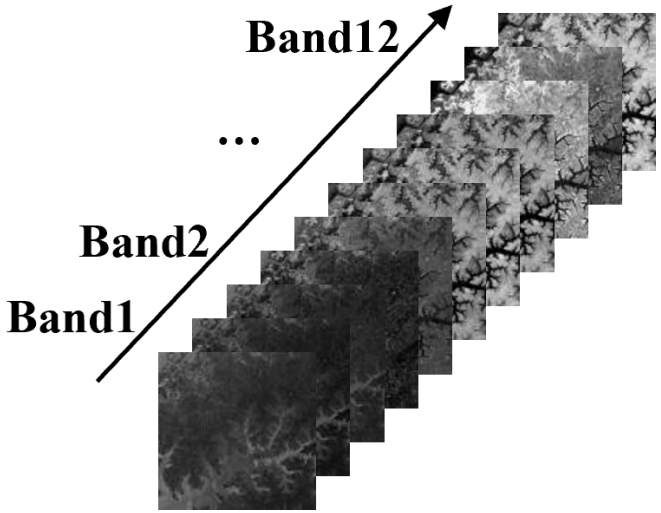


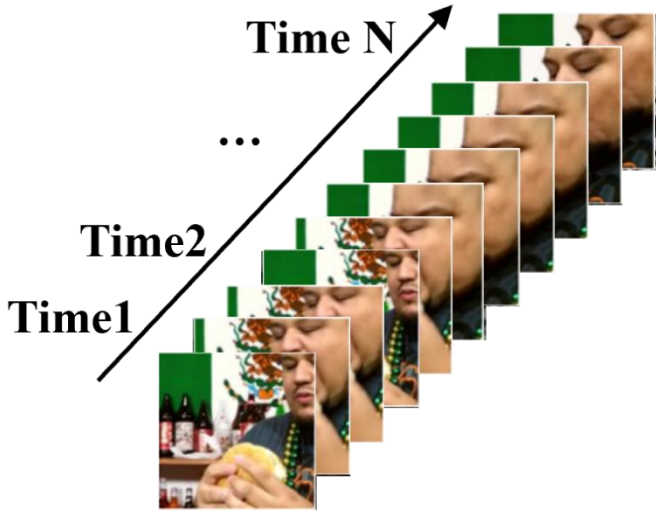
- **RS Spectral Imagery *vs.* RS RGB Imagery**
  - Multi-Channels & 3 Channels
  - Rich Spectral Information
  - Datatype



Motivation

- **Spectral Imagery *vs.* Video Data**
  - No movements in spectral imagery.
  - No subjects or background in remote sensing imagery.
  - Details are crucial for remote sensing analysis.





Motivation

- **The feature of remote sensing big data is well-suited for developing Pretraining Foundation Models (PFMs) .** 😇

- **Most RS PFMs focus on RS RGB imagery.** 😇
  - 3 Channels
  - easy to transfer methods from computer vision field

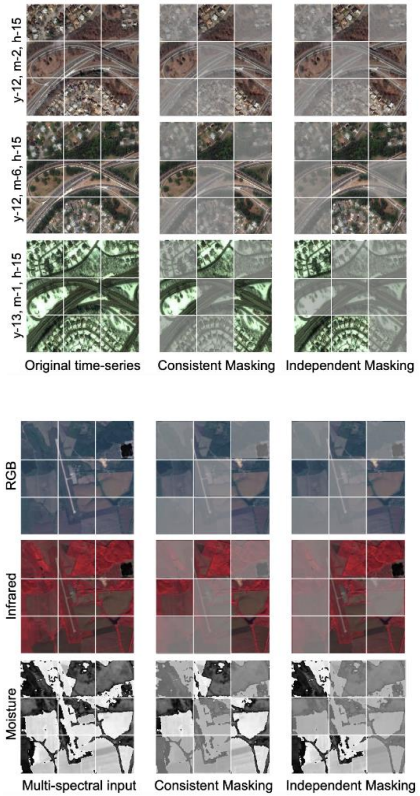
- **Few studies focus on PFMs in RS spectral imagery.** 😟
  - SatMAE

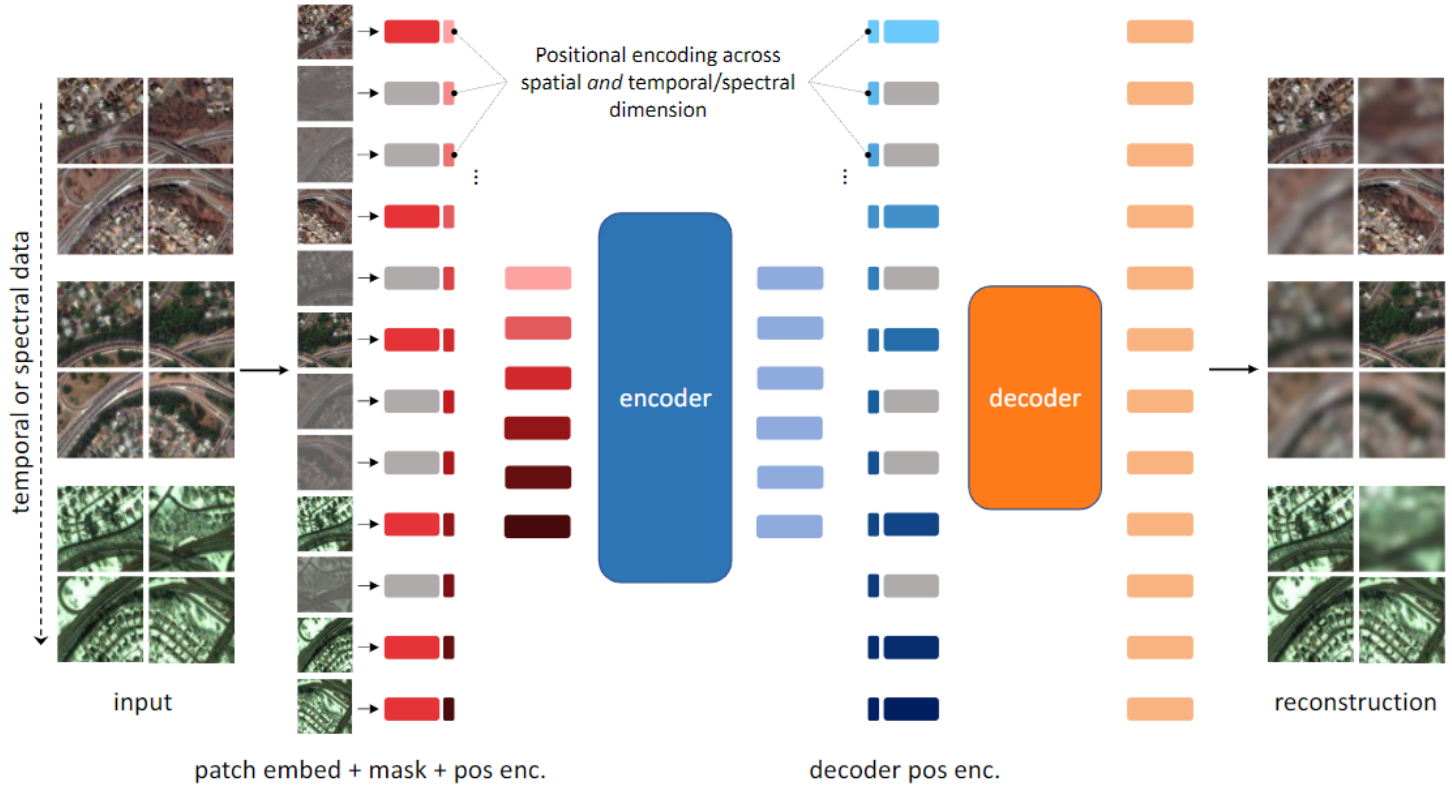
Motivation

SatMAE

Temporal RGB Data

Spectral Data

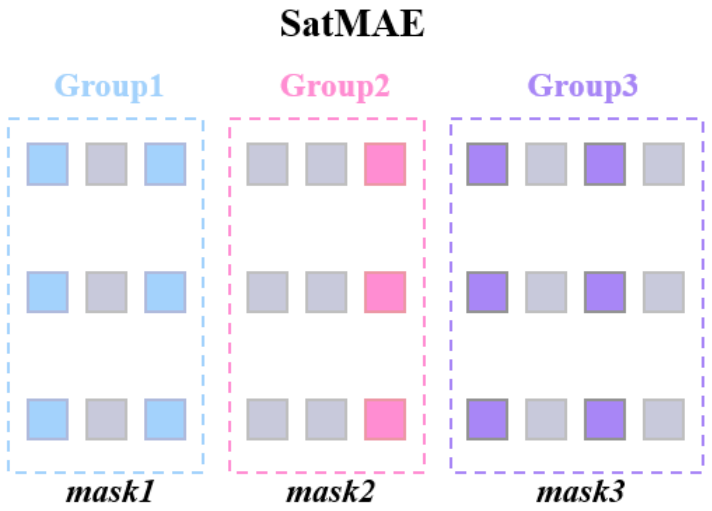


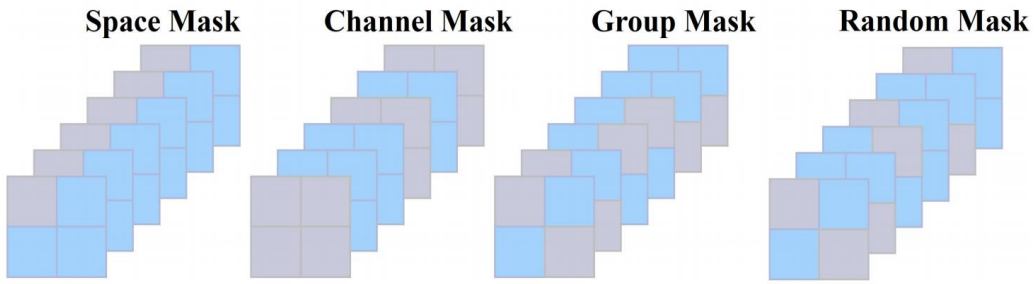


Cong, Yezhen, et al. "Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery." Advances in Neural Information Processing Systems 35 (2022): 197-211.

Motivation

The Limitations of SatMAE



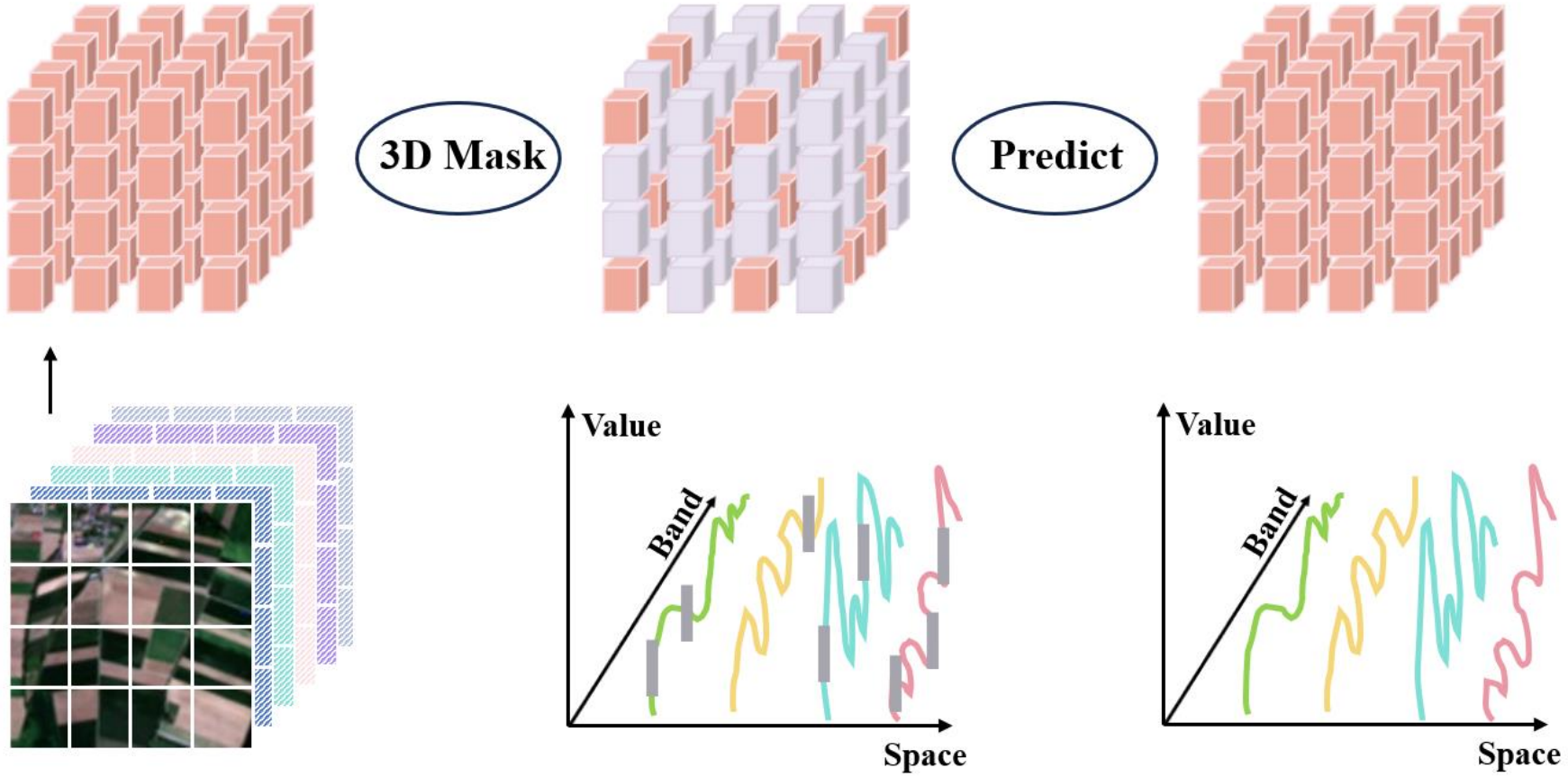


- Inappropriate interaction between groups
- Limited band combinations in grouping
- Extra inductive bias

Can MAE exploit local spectral continuity in spectral data with variable band counts to learn strong representations and reduce inductive bias?

Method

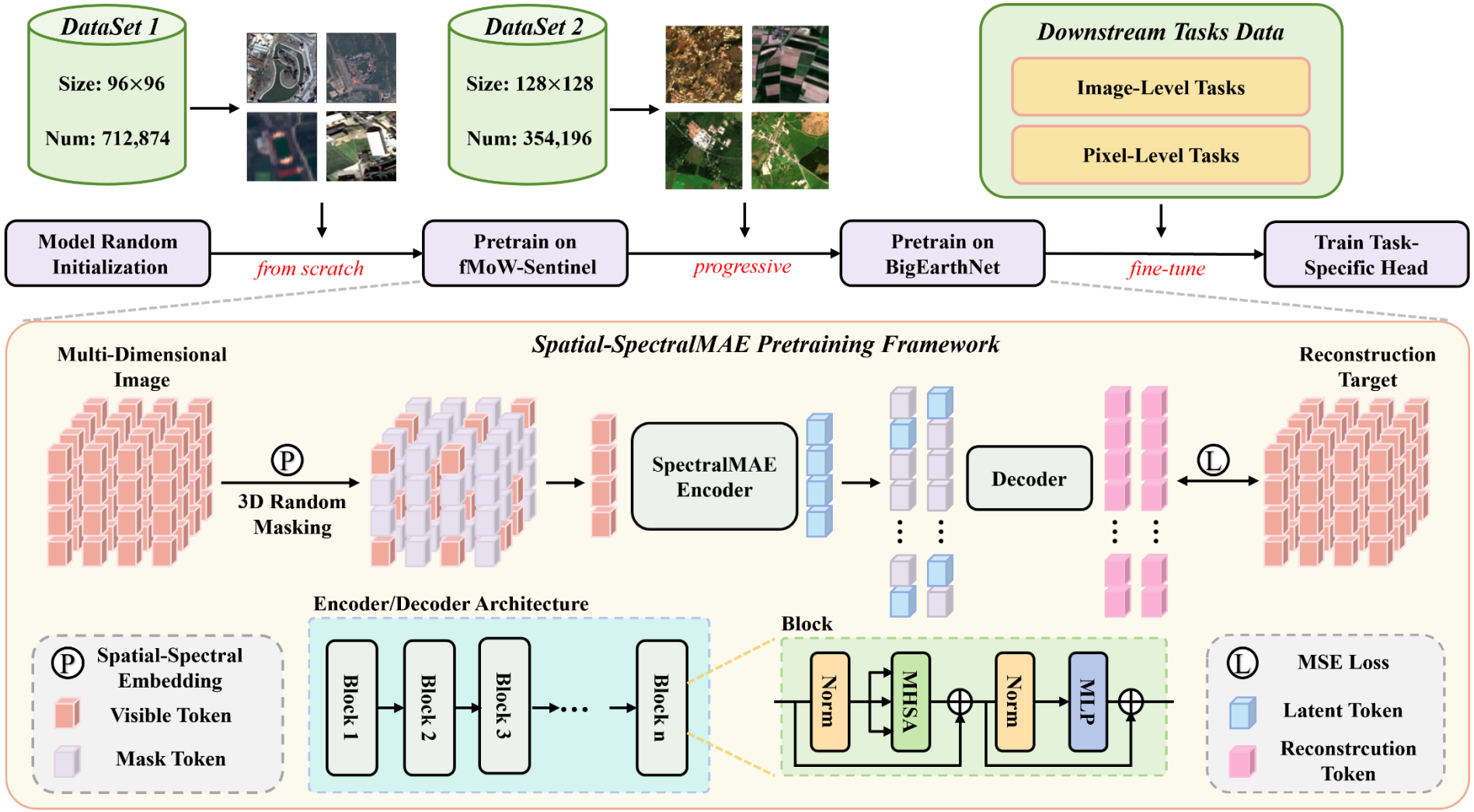
3D Masked Autoencoders (MAE)



Integrating local spectral continuity and spatial invariance via small tensor cubes.

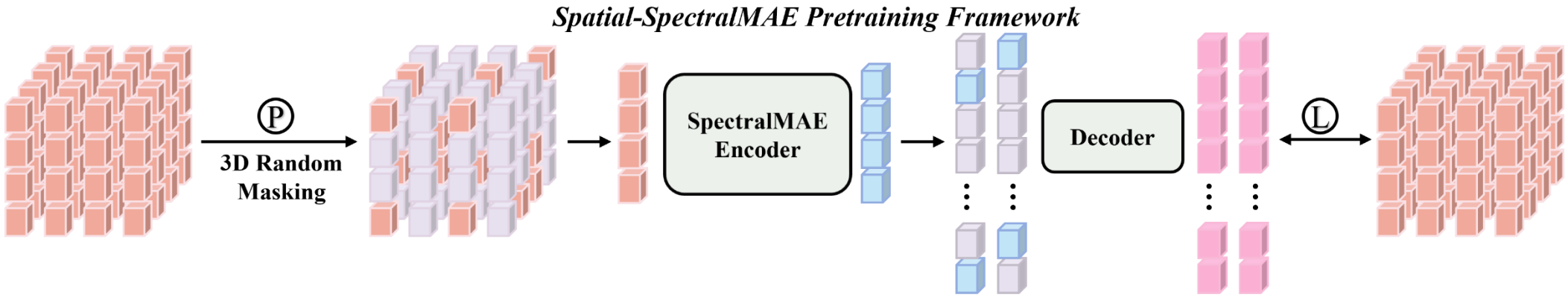
Method

Spatial-Spectral Masked Autoencoders (S2MAE)

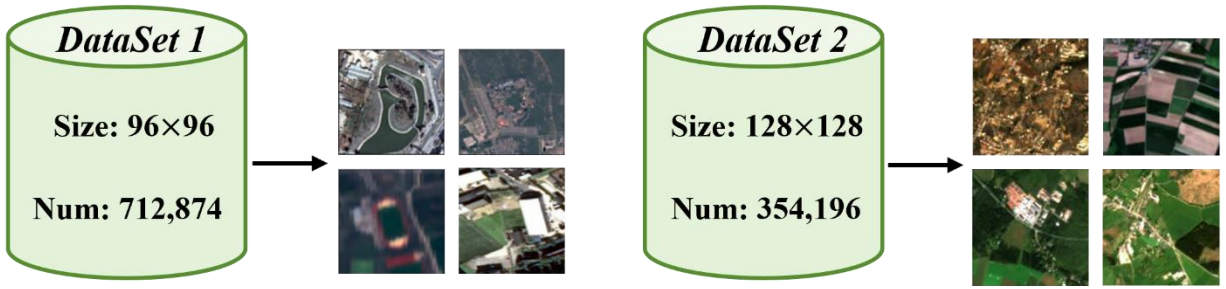


Method

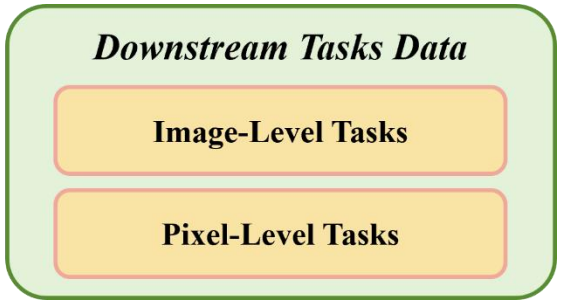
3D Spatial-Spectral Masked Image Modeling



Progressive Pretraining across Datasets



Finetune



Experiments

Single-Label Classification

Method	Pretrained Dataset	Acc. (%)
ResNet50[12]	ImageNet-1k	96.72
SeCo[27]	SeCo	97.23
ViT[8]	From scratch.	98.73
ViT[8]	ImageNet-22k	98.91
SatMAE[4]	fMoW-S2	99.09
S2MAE	fMoW-S2	99.16
S2MAE*	fMoW-S2+BigEarthNet	**99.19**

EuroSAT
10 Classes
Scene Classification

Multi-Label Classification

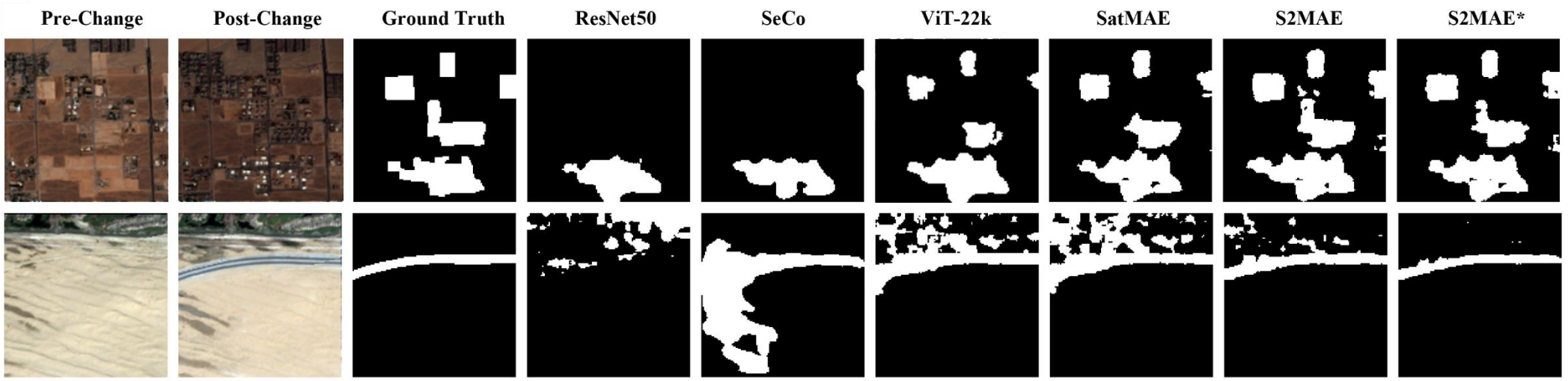
Method	Pretrained Dataset	mAP
ResNet50[12]	ImageNet-1k	80.06
ViT[8]	From scratch.	80.15
SeCo[27]	SeCo	82.82
ViT[8]	ImageNet-22k	84.67
SatMAE[4]	fMoW-S2	84.93
S2MAE	fMoW-S2	85.59
S2MAE*	fMoW-S2+BigEarthNet	**87.41**

BigEarthNet
19 Classes
Land Cover Classification

The results underscore the superior generalization capabilities of S2MAE.

Experiments

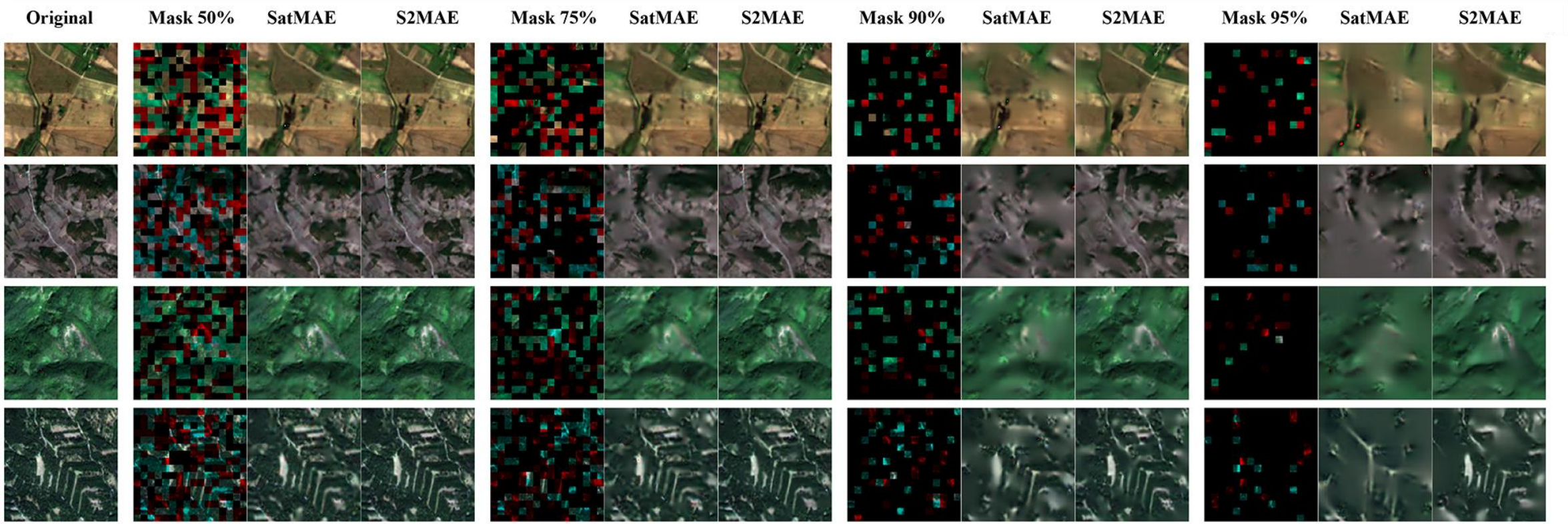
Change Detection



Method	Pretrained Dataset	Precision	Recall	F1
ResNet50[12]	ImageNet-1k	**65.42**	38.86	48.10
SeCo[27]	SeCo	57.71	49.23	49.82
ViT[8]	From scratch.	56.71	47.52	51.71
ViT[8]	ImageNet-22k	52.09	52.37	52.23
SatMAE[4]	fMoW-S2	55.18	50.54	52.76
S2MAE	fMoW-S2	53.89	55.87	53.28
S2MAE*	fMoW-S2+BigEarthNet	54.90	**56.81**	**54.26**

Experiments

Reconstruction Visualization



Experiments

Ablation Stuides

Embedding	Method	mAP
Vanilla	sin-cos	85.57
Spatialspectral	sin-cos	**85.62**
Spatialspectral	learnable	85.59

(a) Positional Embedding

Blocks	mAP
2	84.91
4	**85.59**
8	85.47

(b) Decoder Depth

Case	Scale	mAP
Without norm	-	84.86
Normalization	[0,1]	**85.59**
Standardization	[-1,1]	85.52

(c) Reconstruction Target

Init. Weights	Patch Size	mAP
Random	16	70.68
S2MAE	16	78.42
Random	8	80.15
S2MAE	8	**85.59**

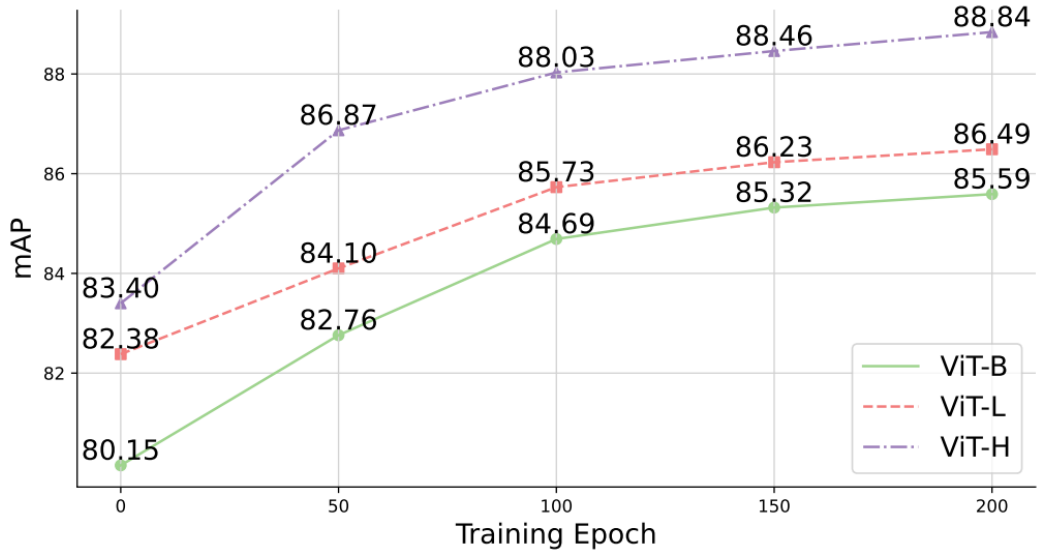
(d) Patch Size

Ratio	mAP
25%	82.81
50%	84.32
75%	84.94
90%	**85.59**

(e) Masking Ratio

Pretrained Dataset	mAP
From scratch.	80.15
BigEarthNet	83.11
fMoW-S2	85.59
fMoW-S2+BigEarthNet	**87.41**

(f) Data Scale



Conclusion

- **Summary**
  - We introduced S2MAE, an MAE extension for spectral RS imagery pretraining.
  - S2MAE incorporates a 3D transformer architecture, employing a random masking strategy and integrating learnable spectral-spatial embeddings.

- **Key Observations**
  - For highly redundant spectral images, a high masking ratio (90%) during pretraining is very important.
  - The masking strategy needs to align with the properties of the spectral images.
  - Progressive pretraining on different datasets can enhance the model's performance.

Future Works

- **How to capture longer spectral sequences?** 
  - S2MAE utilizes a 3D masking strategy to only capture the local spectral consistency.
  - Focusing on the reconstruction of information in the spectral sequence dimension may yield richer representations.

- **Focus on self-supervised methods using multimodal RS data.** 
  - Fusing data from various satellites and aircraft is crucial for building a foundational model in the remote sensing field.