Breaking Representation Barriers for Earth Observation: A Sensor-Agnostic Foundation Model

Gencer Sumbul, Devis Tuia

Environmental Computational Science and Earth Observation Laboratory (ECEO)

Supported by the European Space Agency through the Discovery and Preparation Program, and is part of the project Toward a Foundation Model for Multi-Sensor Earth Observation Data with Language Semantics.

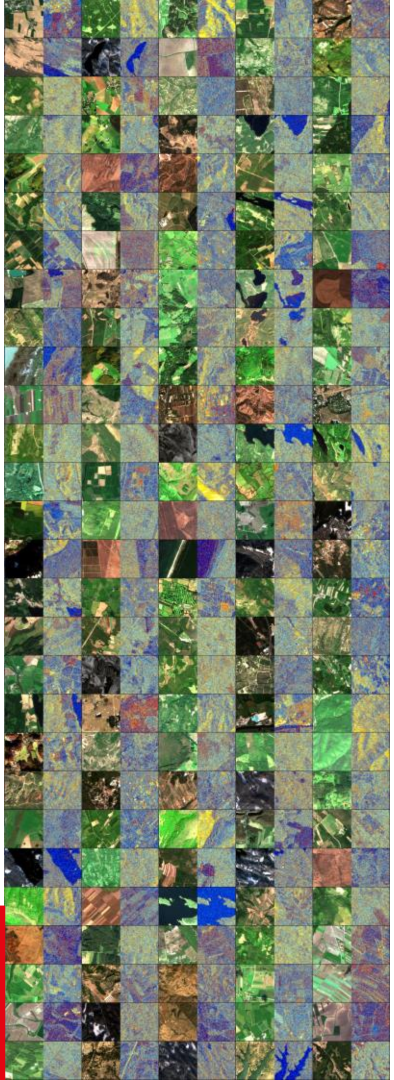
Gencer Sumbul

Foundation models (FMs) for Earth observation (EO)

- FMs for automatic and large-scale analysis of massive EO data through learning transferable image representations.

- Existing FMs for EO are either:
  - sensor-specific (e.g., Scale-MAE for RGB, SatMAE for Sentinel-2 multispectral); or
  - computationally complex (e.g., DOFA, TerraMind);
  - relying on a fixed combination of sensors (e.g., CROMA) with sensor/modality-specific efforts (e.g., AnySat, TerraMind)
  - requiring massive pretraining sets (e.g., DOFA, AnySat, TerraMind)

A significant barrier remains: the lack of unified image representations for sensor-agnostic processing of EO data.

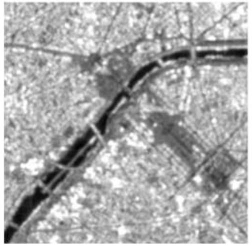


Intrinsic heterogeneity of EO imagery sensors

- The heterogeneous nature of EO imagery sensors makes achieving such a goal difficult.

RGB
3 bands
1m GSD

Multispectral
13 bands
10m GSD

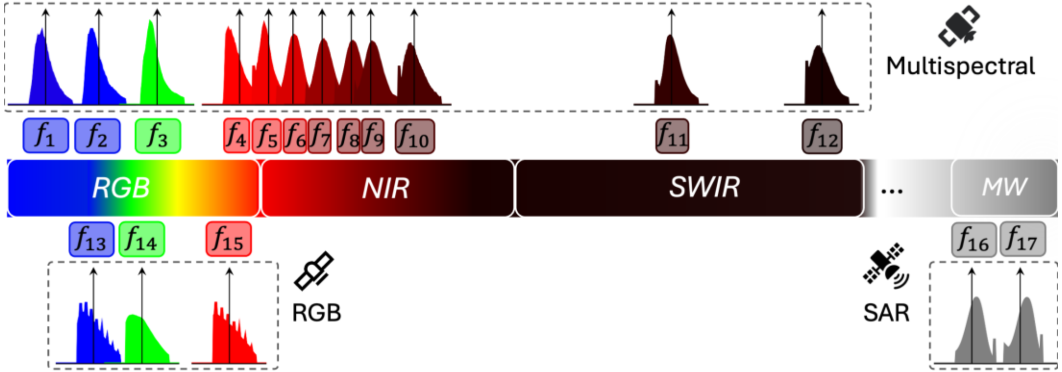


SAR
2 bands
10m GSD

- Across heterogeneous sensors, how to:
  - break the representation barriers;
  - pretrain a simple yet effective model, demanding as little data as possible;
  - enable downstream transfer using a unified model?

Gencer Sumbul

SA-MAE: A Sensor-agnostic FM

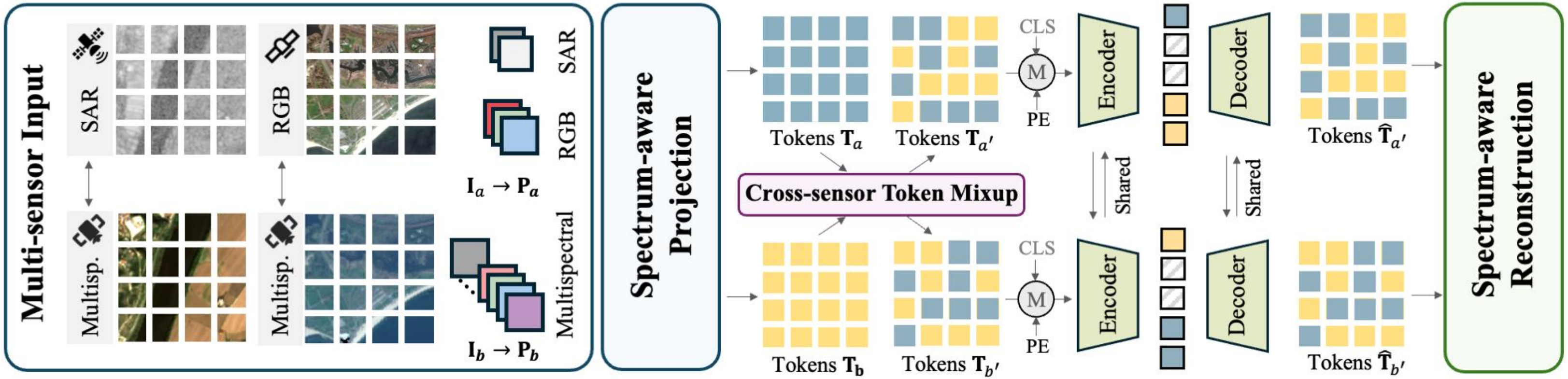


All the different sensors capture subsets of the full electromagnetic spectrum with well-defined physical properties.

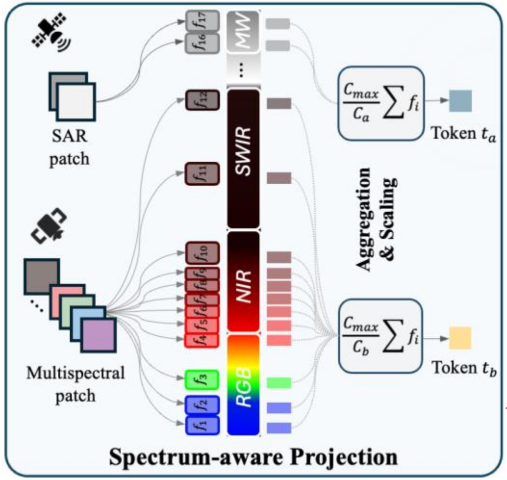
1. Unify sensor representations by projecting any sensor data into a shared and divisible space called the spectrum-aware space.
2. Pretrain a single transformer model with a self-supervised objective:
   - reconstruct randomly masked regions of the sensor-agnostic representations in the spectrum-aware space.

SA-MAE: A Sensor-agnostic FM

1. *Spectrum-aware Image Projection*
2. *Cross-sensor Token Mixup*
3. *Spectrum-aware Image Reconstruction*
4. *Sensor-agnostic Downstream Transfer*

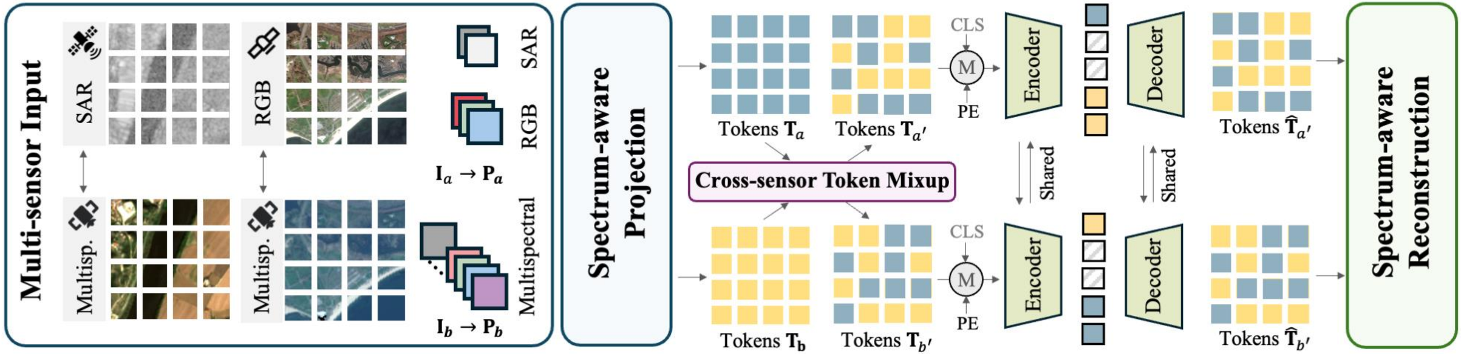


SA-MAE: A Sensor-agnostic FM



- *Spectrum-aware Image Projection*:
  - We learn spectrum-aware projections depending on the considered wavelengths.
  - Each sensor's bands are first projected using wavelength-specific projection functions, and then aggregated to obtain tokens.

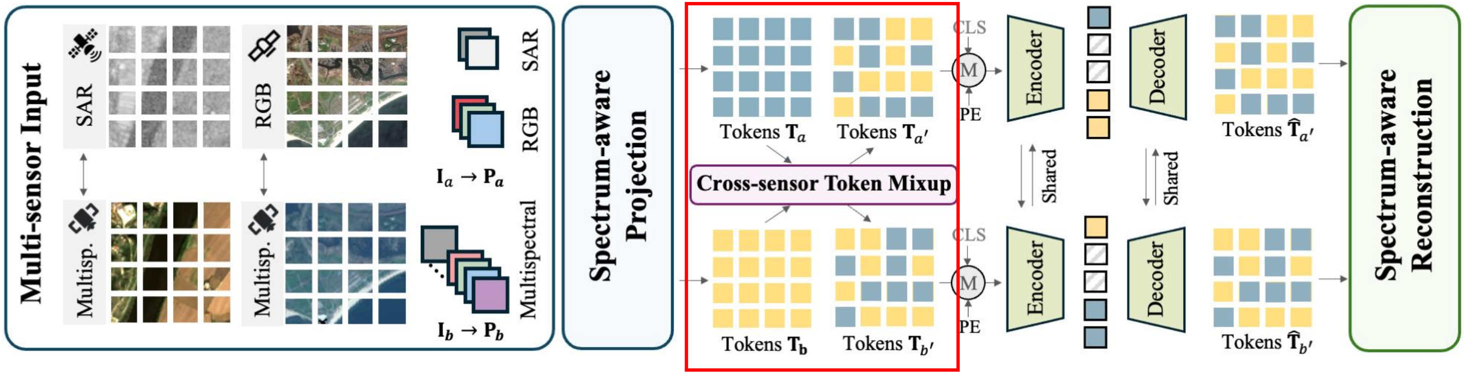
Eliminates the need for separate models and backbones for different sensors.



SA-MAE: A Sensor-agnostic FM

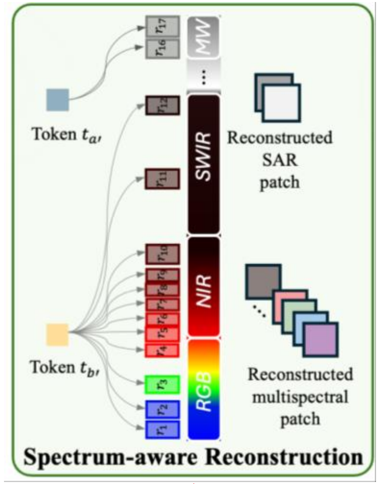
- *Cross-sensor Token Mixup:*
  1. We first use pairs of aligned images from different sensors;
  2. then exchange tokens across the images of a pair.

Mitigates the bias specific to sensor/spectra combinations.

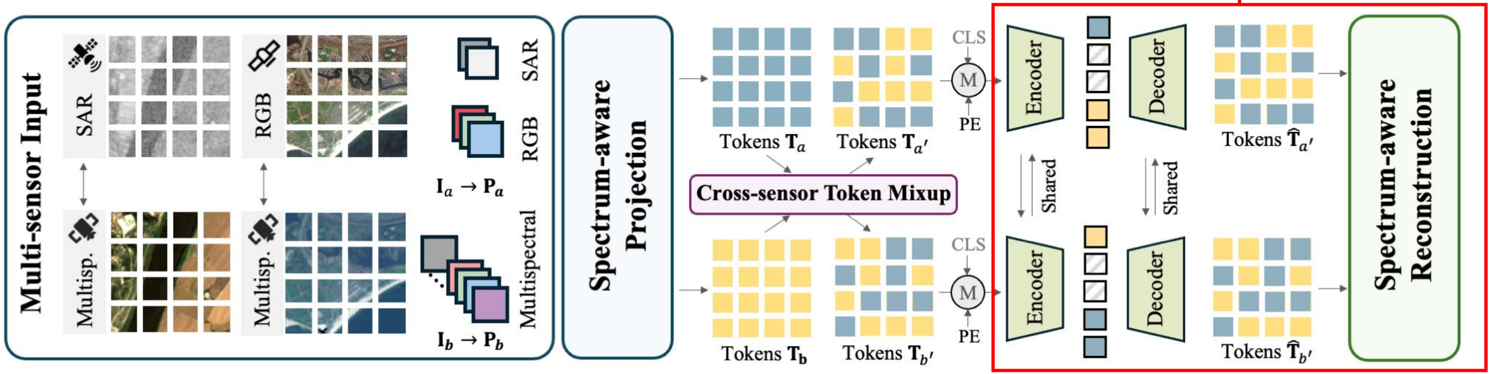


SA-MAE: A Sensor-agnostic FM

- *Spectrum-aware Image Reconstruction*:
  1. We feed the cross-sensor mixed embeddings into a standard encoder-decoder based transformer with masked tokens.
  2. We reproject the decoded images back to the original spectral bands through spectrum-aware remapping functions.



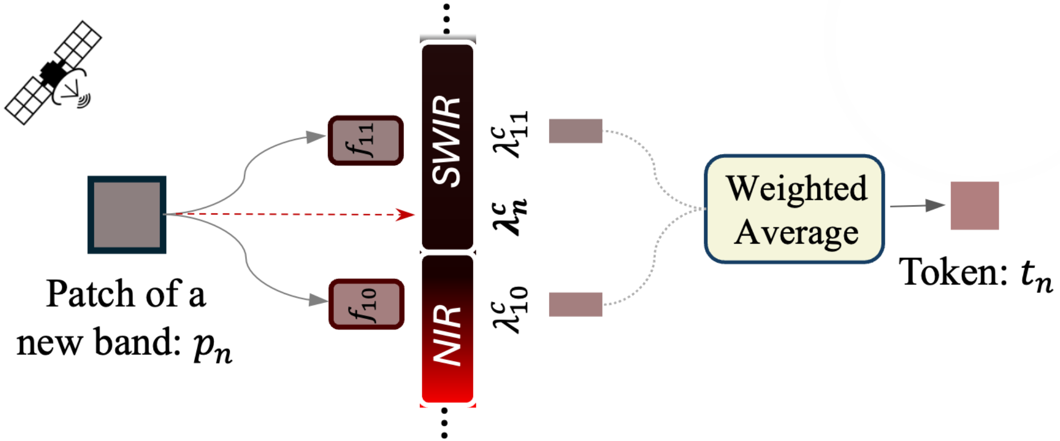
Effectively scales into larger models with more data.



Gencer Sumbul

SA-MAE: A Sensor-agnostic FM

- *Sensor-agnostic Downstream Transfer*: Thanks to the spectrum-aware image projection, the resulting encoder can easily generalize to different sensors by using:
  - either the existing projection layers (when available) or
  - adapting them for unseen sensors by interpolation.

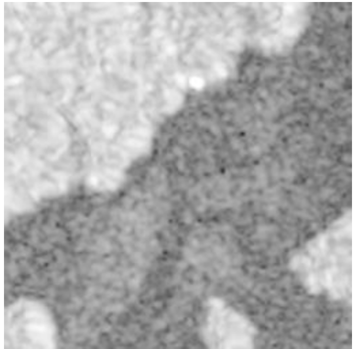


Allows downstream transfer to any EO sensor

Experimental Setup

- *Pretraining:*
  - 120K paired images from the submeter fMoW-RGB dataset and its Sentinel-2 counterpart fMoW-S2; and
  - 376K paired images from the BigEarthNet-MM dataset, including Sentinel-1 and Sentinel-2 images.
  - We pretrained two models based on ViT-B and ViT-L backbones, each for 300 epochs.
    - ViT-B model has 116.3M parameters, 4.8M more than MAE
    - ViT-L model has 334.8M parameters, 5.9M more than MAE

- *Downstream transfer* on diverse inputs and tasks:
  - Single/multi-modal single/multi-label image scene classification with variable scale ratios
  - Semantic segmentation with zero-shot sensor transfer
  - Few-shot classification



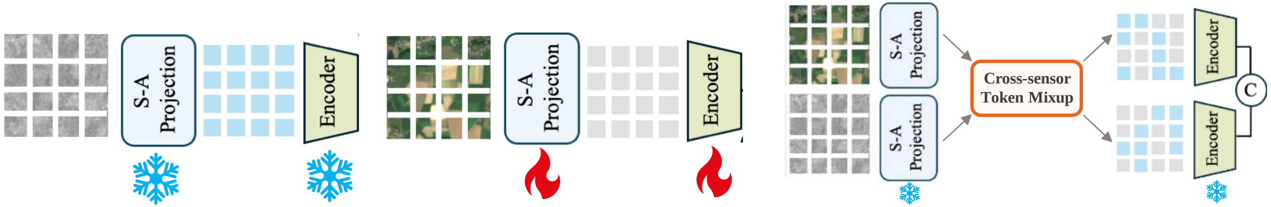
Experimental Results (Multispectral, Radar, Multi-Modal)

Model	Backbone		BigEarthNet-MM 10%		
			BEN-S1 (LP)	BEN-S2 (FT)	BEN-MM (LP)
SatMAE (S2)	ViT-B		❌	85.9	❌
GFM	Swin-B		❌	86.3	❌
SatLas	Swin-B		❌	82.8	❌
I-JEPA	ViT-B		❌	85.9	❌
SpectralGPT	ViT-B		❌	85.6	❌
S2MAE	ViT-B		❌	85.6	❌
msGFM	Swin-B		67.5	86.8	-
SA-MAE (Ours)	ViT-B		**78.9**	**86.9**	**85.4**
	Backbone	S2 Pretraining Data			
SatMAE (S2)	ViT-L	713K	❌	82.1	❌
CROMA	ViT-B (x2)	1M	79.8	87.6	85.2
SpectralGPT	ViT-L	713K	❌	86.9	❌
S2MAE	ViT-L	713K	❌	86.5	❌
SatMAE++ (S2)	ViT-L	713K	❌	85.1	❌
SA-MAE (Ours)	ViT-L	**248K**	**80.5**	**87.7**	**86.7**

BigEarthNet-MM multi-label scene classification results (mAP)

❌ indicates the methods that are not applicable

linear-probing (LP) and finetuning (FT) are applied with 10% of the training set



Experimental Results (Multispectral)

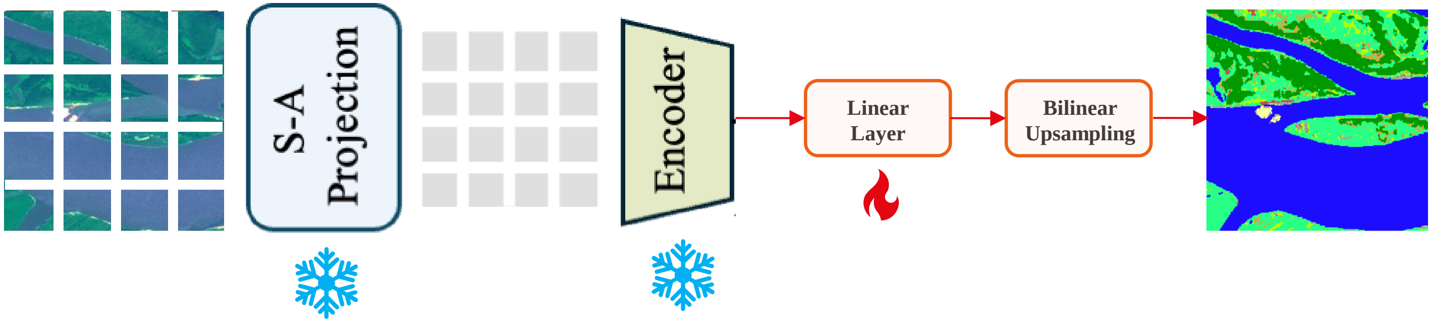
Model	Backbone	Linear Probing	Finetuning
SeCO	ResNet-18	-	93.1
GASSL	ResNet-18	-	89.5
SeCO	ResNet-50	95.6	97.2
CACo	ResNet-50	95.9	-
SatMAE (S2)	ViT-B	96.6	99.2
I-JEPA	ViT-B	95.6	99.2
SpectralGPT	ViT-B	-	99.2
S2MAE	ViT-B	-	99.2
SA-MAE (Ours)	ViT-B	**98.4**	**99.4**
SatMAE (S2)	ViT-L	97.7	99.0
SatMAE (RGB)	ViT-L	93.0	95.7
CROMA	ViT-B (x2)	97.6	99.2
SatMAE++ (S2)	ViT-L	-	99.0
SA-MAE (Ours)	ViT-L	**98.9**	**99.6**

Top-1 accuracy (%) on EuroSAT for scene classification under linear probing and finetuning.

Experimental Results (Semantic Segmentation)

Model	Backbone	mIoU
I-JEPA	ViT-B	36.7
SatMAE (S2)	ViT-B	45.5
CROMA	ViT-B	46.6
SA-MAE (Ours)	ViT-B	**47.9**

Semantic segmentation on DFC2020 dataset with frozen backbone finetuning



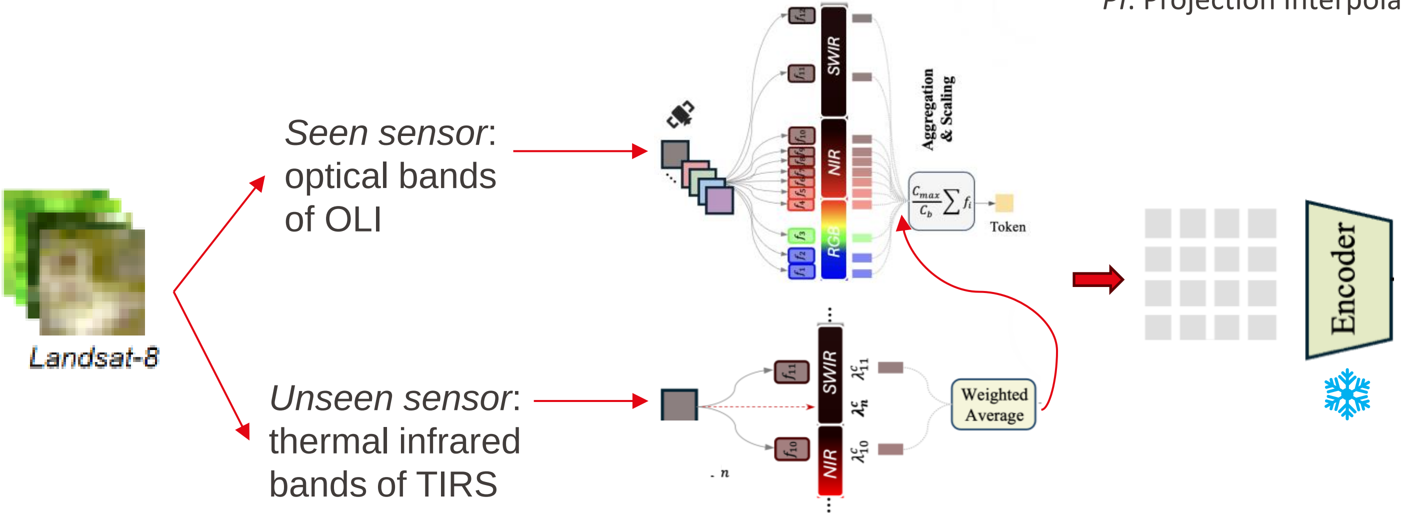
Experimental Results (Unseen Sensor, Segmentation)

Model	Training	mIoU	Accuracy	F1 Score
U-Net 2D	Scratch	47.7	69.7	62.7
DeepLap V3+	Scratch	48.5	71.2	63.2
SA-MAE (w/o PI)	Frozen	35.4	55.8	50.6
SA-MAE (ViT-B)	Frozen	**50.2**	**75.5**	**63.7**

Zero-shot sensor transfer for crop-type segmentation on SICKLE.

Frozen: a segmentation head is finetuned with frozen backbone.

PI: Projection Interpolation



interpolation to unseen spectrum ranges via weighted averaging

Experimental Results (VHR RGB)

Model	Backbone	WHU-RS19	UCMerced
SatMAE (RGB)	ViT-L	69.9	69.7
Scale-MAE	ViT-L	79.5	75.0
Cross-Scale MAE	ViT-L	79.8	74.5
SA-MAE (Ours)	ViT-L	**80.4**	**77.0**

Average kNN classification accuracy with different scale ratios (100%, 50%, 25%, 12.5%)

<table>
<tr><th>Model</th><th>Backbone</th><th>Top-1 Accuracy</th><th>VHR RGB Pretraining Data Size</th></tr>
<tr><td>MAE</td><td>ViT-L</td><td>93.3</td><td rowspan="5">364K</td></tr>
<tr><td>SatMAE (RGB)</td><td>ViT-L</td><td>94.8</td></tr>
<tr><td>MCMAE</td><td>ViT-B (x2)</td><td>95.0</td></tr>
<tr><td>Scale-MAE</td><td>ViT-L</td><td>95.7</td></tr>
<tr><td>SatMAE++ (RGB)</td><td>ViT-L</td><td>97.5</td></tr>
<tr><td>SA-MAE (Ours)</td><td>ViT-L</td><td>95.8</td><td>60K</td></tr>
</table>

Top-1 accuracy (%) of finetuning on RESISC-45 for scene classification.

Experimental Results (Few-shot Classification)

Model	Number of Parameters	Pretraining Data Size	Accuracy
CLIP-ViT-B/16	152M	100M	39.7
Prithvi v1.0	100M	0.75M	46.9
Prithvi v2.0	300M	16.8M	47.5
SA-MAE (ViT-B)	116M	**0.5M**	**52.6**
TerraMindv1-B	700M	64M	57.5
TerraMindv1-L	900M	64M	56.6

Full-way 1-shot classification on image features of EuroSAT dataset over 200 runs

Support Set
Full-way: 10 classes
1-shot: one image per class

Query Set

Credit: Helber et. al, 2019.

Conclusion

- SA-MAE breaks representation barriers across EO sensors by:
  - projecting diverse sensory data into shared spectrum-aware space; and
  - pretraining with masked data modelling and cross-sensor token mixup.

- This leverages synergies between sensors characterized by different spectral properties, while eliminating the need for isolated efforts in training sensor-specific models with a high pretraining data efficiency.

- Toward unified multi-sensor EO:
  - extensions to the temporal domain with spatial-resolution aware projections;
  - deeper analysis on any sensor downstream transfer; and
  - scaling to more sensors and more data.

- Stay tuned for model weights, code, paper, and more!

Interested in pursuing a PhD on Multi-Modal Foundation Models for EO?

SCAN ME