RAMEN: Resolution-Adjustable Multimodal Encoder for Earth Observation

Nicolas Houdré [1]   Diego Marcos [2]   Hugo Riffaud de Turckheim [2]   Dino Ienco [3]

Laurent Wendling [1]   Camille Kurtz [1]   Sylvain Lobry [1]

[1] Université Paris Cité, LIPADE, F-75006 Paris, France; [2] UMR TETIS, EVERGREEN, Inria, Univ. Montpellier, Montpellier, France; [3] UMR TETIS, EVERGREEN, INRAE, Inria, Univ. Montpellier, Montpellier, France

Abstract

Earth observation (EO) data spans a wide range of spatial, spectral, and temporal resolutions, from high-resolution optical imagery to low resolution multispectral products or radar time series. While recent foundation models have improved multimodal integration for learning meaningful representations, they often expect fixed input resolutions or are based on sensor-specific encoders limiting generalization across heterogeneous EO modalities. To overcome these limitations we introduce RAMEN, a resolution-adjustable multimodal encoder that learns a shared visual representation across EO data in a fully sensor-agnostic manner. RAMEN treats the modality and spatial and temporal resolutions as key input data features, enabling coherent analysis across modalities within a unified latent space. Its main methodological contribution is to define spatial resolution as a controllable output parameter, giving users direct control over the desired level of detail at inference and allowing explicit trade-offs between spatial precision and computational cost. We train a single, unified transformer encoder reconstructing masked multimodal EO data drawn from diverse sources, ensuring generalization across sensors and resolutions. Once pretrained, RAMEN transfers effectively to both known and unseen sensor configurations and outperforms larger state-of-the-art models on the community-standard PANGAEA benchmark, containing various multi-sensor and multi-resolution downstream tasks. Our code and pretrained model are available at `https://github.com/nicolashoudre/RAMEN`.

1. Introduction

Earth observation (EO) data is inherently heterogeneous. Remote sensing (RS) data ranging from aerial RGB cameras to multispectral satellites, SAR sensors, or elevation maps exhibit significant variability in terms of channel characteristics, ground sampling distance (GSD - distance between pixel centers measured on the ground), and tempo-
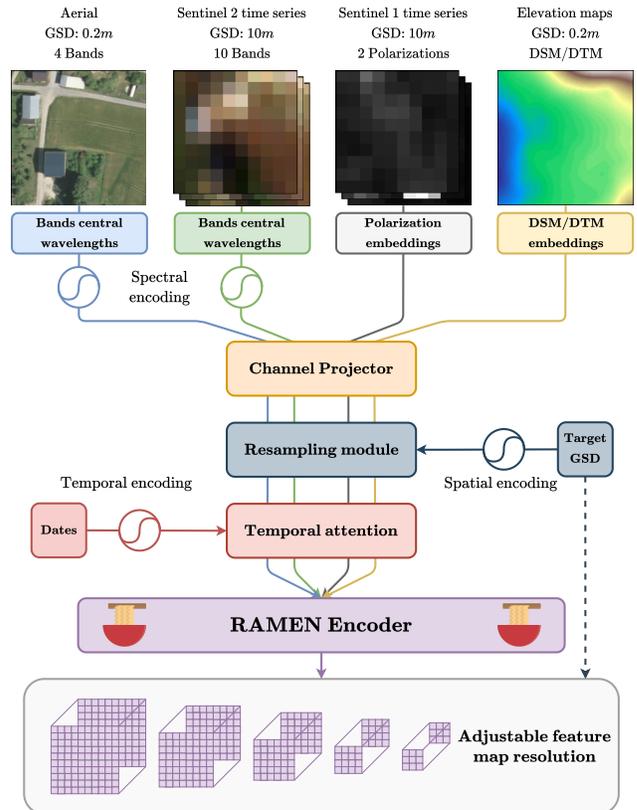

Figure 1. **Visual workflow of RAMEN.** RAMEN enables consistent adaptation across multimodal imagery via resolution-specific projection modules. Thanks to our **adjustable resampling** strategy and pretraining scheme, RAMEN allows practitioners to define the feature map spatial resolution of encoded inputs, allowing fine-grained representations and trade-offs between downstream performances and computational overhead.

ral sampling patterns. This diversity, while challenging to construct a unified model, is what enables EO data to serve a wide range of applications. Tasks such as land-cover mapping [14, 25, 39], disaster response [19, 32] or urban planning [10] rely on different combinations of spatial de-

tail, spectral information or temporal dynamics. Leveraging this large volume of heterogeneous data in a coherent manner remains a core challenge for general-purpose EO models. Several multimodal EO foundation models (FMs) tackled these issues by integrating sensor-specific encoders [7, 11, 26]. Using these models on new modalities requires changing and retraining parts of the architecture, thus limiting the generalization capabilities of such models. Recent efforts on EO FMs enabled more flexibility in terms of pretraining data [4], multimodal integration [2, 41], or spatial [23, 33] and spectral [36, 48] resolution handling. However, these models still fall short of fully accounting for the diverse characteristics of EO sensors. They often neglect differences in modal, spatial and temporal properties and produce feature representations at fixed resolutions, limiting their scalability across tasks requiring different spatial detail or computational efficiency.

To address these challenges, we propose a **R**esolution-**A**djustable **M**ultimodal **En**coder (**RAMEN**) that treats every spatio-spectro-temporal resolution aspect of modalities and project representations in a resolution-aware shared space. As illustrated in Fig. 1, this enables RAMEN to learn a unified transformer encoder on a wide combination of heterogeneous modalities from diverse datasets. Each modality, with its own channels and spatial and time resolutions, is treated separately to produce a universal multimodal representation: (i) A channel-conditioned projector based on each band's central wavelength (or radar polarization/elevation maps) projects every input to the latent space dimension; (ii) An adjustable spatial resampler module enables alignment of spatially heterogeneous inputs by mapping them to a target GSD. At pretraining time, this target GSD is chosen randomly for each modality and dataset, ensuring that the encoder learns to generalize across scales and is not biased toward a specific setting. Such a mechanism further allows practitioners to explicitly choose the desired resolution during fine-tuning or inference depending on the downstream task and computational resources available; (iii) A temporal attention module enriched with temporal positional encoding captures dependencies across time. Together, these components allow RAMEN to optimize multimodal representations that are robust to heterogeneous spatio-spectro-temporal characteristics. Leveraging these unified multimodal representations, we pretrain RAMEN using a large-scale multimodal EO corpus combining multiple datasets in a self-supervised manner using a masked image modeling [17] strategy by reconstructing randomly masked regions of each modality representation in the resolution-aware space. Besides type-aware embeddings covering three major input types (i.e. spectral imagery, radar imagery and elevation maps), all learnable parameters in our model are shared across all modalities, enabling an architecture trainable on highly heterogeneous multimodal EO datasets. Finally, RAMEN also provides users with explicit control over desired spatial resolution of the representations at inference, allowing them to balance task-specific needs with available computational resources and desired precision. We evaluate RAMEN on eight datasets from the community-standard PANGAEA [27] benchmark, covering diverse uni- and multimodal semantic segmentation tasks, and show that it delivers state-of-the-art performance, strong flexibility and robust generalization across resolutions - including those unseen during pretraining. Our contributions can be summarized as follows:

- We propose RAMEN, a resolution-adjustable encoder that can process remote sensing images of any sensors and configurations without any retraining;

- The proposed architecture allows practitioners to configure the desired spatial resolution at inference time, enabling flexible adaptation of the compute/performance trade-off to the need of the task;

- We demonstrate that when pretrained on a large collection of diverse modalities, RAMEN outperforms the state-of-the-art on standard EO benchmark datasets.

2. Related work

In this section, we highlight recent supervision and architecture advancements designed to support multimodal FMs in remote sensing, particularly dealing with multi-resolution setup and their adaptability to diverse input data.

Self-supervised learning (SSL) has emerged as a powerful paradigm to learn generalizable representations from large non-annotated datasets transferable to various downstream tasks. Among SSL approaches, masked image modelling [17] aims at reconstructing randomly masked regions of an input, encouraging the model to learn rich semantic structures without requiring specific data augmentation. The abundance of raw EO data has naturally led to the adaptation of such methods for RS. In this context, *monomodal SSL* approaches pretrain models on single sensor data, often optical imagery [7, 18, 30, 37, 38]. While effective to learn meaningful representations from large-scale EO datasets, these methods remain inherently constrained to specific assumptions on input data. To move beyond this limitation, *multimodal SSL* aims at jointly leveraging heterogeneous EO sensors (e.g. multispectral and radar imagery) through sensor-specific encoders [1, 11, 46] or projections in a unified representation space [15, 20, 22, 28, 29, 40, 44]. However, many FMs rely on common computer vision architectures not explicitly designed for the specificities of EO data such as heterogeneous modalities (e.g. multispectral, radar, elevation maps), variable resolutions or spatio-temporal dependencies [42].

Spectrally-aware FMs. While most EO foundation models process multispectral imagery as generic multi-channel in-

puts, recent works have explored incorporating the physical meaning of spectral bands into model design. DOFA [48] authors introduce a wavelength-conditioned hypernetwork that dynamically generates convolutional weights based on the central wavelength of each input band. Similarly, SMARTIES [36] or Panopticon [45] use the central wavelength of each band to respectively guide spectrum-aware projection layers or attention over channels.

Spatially-aware FMs. Several architectures have been proposed to address the varying spatial resolution of EO images. Notably, Scale-MAE [33] introduced GSD-based positional encodings, allowing the model to adapt feature extraction based on input resolution. AnySat [2] maintains spatial awareness by considering patch sizes for each modality representing real ground distances rather than pixels, but introduce modality-specific projectors reducing generalization capabilities of the model. FlexiMo [23] builds on the theory introduced by [5] to dynamically adapt both raw images and patch embedding convolution weights.

Temporally-aware FMs. Temporal information plays a central role in remote sensing, as many EO applications such as crop monitoring or disaster response rely on multi-date observations. This aspect has motivated several EO foundation models to explicitly model temporal dynamics. PRESTO [40] and Galileo [41] process each time-steps as separate token sequences, while Prithvi [38] treats spatio-temporal sequences using 3D patch embeddings coupled with 3D positional embeddings for time, height and width. AnySat employs a Lightweight Temporal Attention Encoder (LTAE) [13] to embed patches coming from temporal modalities.

Generalization. Existing multimodal EO foundation models typically handle only specific aspects of resolution heterogeneity: some focus on spectral variability (e.g. DOFA, SMARTIES), while others incorporate spatial awareness (e.g. Scale-MAE, FlexiMo) or temporal dynamics (e.g., AnySat, Galileo). Yet none address all three aspects jointly. Compared to previous works, RAMEN is the first model to provide a modality-agnostic, multi-temporal and resolution-adjustable framework.

3. RAMEN: Resolution-Adjustable Multimodal Encoder

In this section, we introduce the RAMEN architecture (see also Fig. 2). We first describe the resolution aware modules treating the channels of the different modalities, as well as spatial and temporal resolutions and their integration in a multimodal network in Sec. 3.1. Then, we detail our self-supervised pretraining scheme in Sec. 3.2.

3.1. Architecture

A central challenge in RS is that different sensors exhibit differences along three axes: *modality* (e.g. RGB or multi-

spectral optical bands, radar polarizations or digital elevation maps), *spatial* (e.g. 0.2 m aerial imagery vs. 10-30 m satellite acquisitions) and *temporal* (e.g. sparsely acquired aerial imagery or regular satellite revisits). Therefore, direct multimodal alignment or concatenation is often infeasible due to the heterogeneous dimensions.

To unify such heterogeneous data, RAMEN introduces three modules encoding each of these input axes (modality, spatial and temporal) in a shared latent space while preserving their physical meaning.

Formally, we consider a set of geospatially aligned images $(x_1, \ldots, x_M)$ observed through M modalities. Each input image consists of a tensor $x_m \in \mathbb{R}^{T_m \times C_m \times H_m \times W_m}$ where m, T_m, C_m, H_m, W_m represent the considered modality, number of temporal observations, number of channels, height and width, respectively.

Channel-conditioned projector. EO sensors differ, not only in the number of channels captured, but also in their physical interpretation. To unify these heterogeneous channel configurations, RAMEN adopts channel-wise embeddings carrying each band's physical meaning. For optical and multispectral modalities, we follow the DOFA [48] formulation and embed the central wavelength of each channel through a sinusoidal positional encoding [43]:

$$\text{PE}(\lambda_m^i, 2k) = \sin \frac{\lambda_m^i}{10000^{2k/D}} \; ; \qquad (1)$$

$$\text{PE}(\lambda_m^i, 2k+1) = \cos \frac{\lambda_m^i}{10000^{2k/D}} \; , \qquad (2)$$

where λ_m^i is the central wavelength of channel i in nanometers and D the latent space dimension.

For non-optical modalities, we use a specific learned embedding, also of size D, corresponding to each of the available channels in the considered datasets: the four transmit/receive polarizations for SAR images (VV, VH, HH and HV) in both ascending and descending orbits, as well as the digital surface (DSM), terrain (DTM) and slope models for elevation data.

These encodings are concatenated and processed by a lightweight multi-layer perceptron (MLP) to obtain a channel-wise projection matrix $M_m \in \mathbb{R}^{C_m \times D}$ that maps raw inputs to the latent space as:

$$x_m^S(t, d, h, w) = \sum_{c=1}^{C_m} x_m(t, c, h, w) M_m(c, d) \, , \qquad (3)$$

where $x_m^S \in \mathbb{R}^{T_m \times D \times H_m \times W_m}$.

Spatial resampler. To unify features originating from sensors with different native spatial resolutions, we design an *adjustable spatial resampler* that maps each spectrally projected features $x_m^S \in \mathbb{R}^{T_m \times D \times H_m \times W_m}$ (acquired at a ground sampling distance GSD_m) to a user-defined target resolution GSD_{target}. The GSD_{target} can be defined at
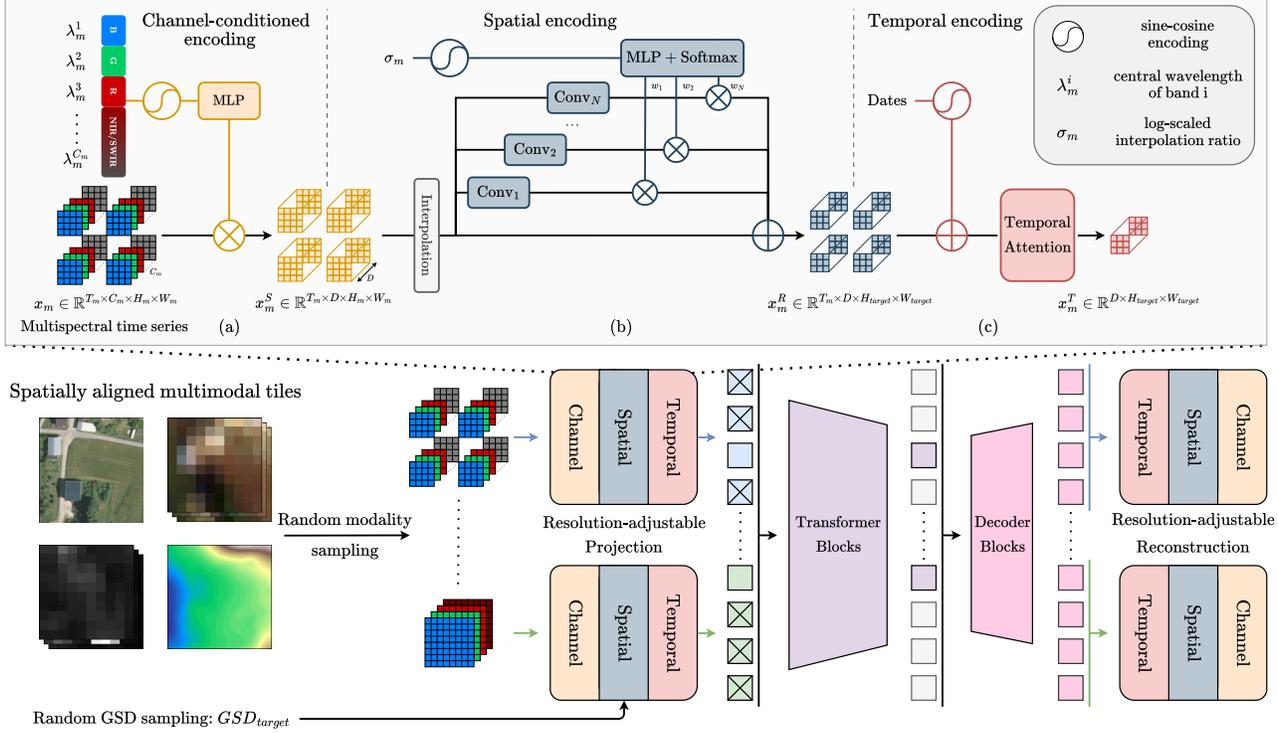
Figure 2. **Architecture of RAMEN.** At each iteration, a subset of modalities and a target ground sampling distance (GSD) are sampled. Each selected modality is projected into a shared latent space through three resolution-specific modules: (a) A **channel-conditioned projector** that embeds the physical meaning of each channel; (b) An **adjustable spatial resampler** maps features to the user-defined GSD_{target} with scale-aware convolutional adaptation; (c) A **temporal attention module** treats time-series enriched with the day of acquisition encoding. We adopt a masked image modeling pretraining scheme, where we reconstruct each modality at its native spectral, spatial and temporal resolution with inverted modules.

inference time based on the downstream task of interest. Because sensors can differ by orders of magnitude in spatial resolution, we introduce a *mixture of N_{conv} convolutions* mechanism that dynamically adapts interpolated images based on the interpolation ratio between the input and target resolutions.

Let $\sigma_m = \log(\text{GSD}_m/\text{GSD}_{target})$ denote the log-scaled interpolation ratio. This scalar quantifies how much the spatial resolution of modality m differs from the target resolution in the log-space, ensuring both down/upsampling are treated symmetrically. Each convolution operator, or expert, is implemented as a single 1×1 convolution. Conditioning their mixture on σ_m enables the model to select how to refine interpolated features depending on the magnitude and direction of the scale change. This provides a lightweight yet expressive mechanism to correct feature statistics after interpolation without altering spatial structure.

Given $\mathcal{I}_{\sigma_m}$ a bilinear interpolation operator parameterized by σ_m, and Conv_n the nth expert convolution, the final spatially aligned representation $x_m^R \in$

$\mathbb{R}^{T_m \times D \times H_{target} \times W_{target}}$ is defined as:

$$x_m^R = \mathcal{I}_{\sigma_m}(x_m^S) + \sum_{n=1}^{N_{conv}} w_n \text{Conv}_n(\mathcal{I}_{\sigma_m}(x_m^S)), \quad (4)$$

where w_n is the weight assigned to the nth expert, $H_{target} = \exp(\sigma_m)H_m$ and $W_{target} = \exp(\sigma_m)W_m$.

Analogously to spectral encodings, we embed σ_m into a D-dimensional vector using sinusoidal positional encoding. The resulting embedding is passed through a MLP followed by a softmax layer to produce normalized weights:

$$(w_1, \ldots, w_{N_{conv}}), \text{with} \sum_{n=1}^{N_{conv}} w_n = 1. \quad (5)$$

This formulation enables continuous interpolation across diverse spatial resolutions while refining spatial details through scale-aware convolutional adaptation.

Temporal attention. To process time-series and model temporal dependencies, we employ a Lightweight Temporal Attention Encoder (LTAE) [13]. To preserve temporal continuity, a sine-cosine positional encoding based on the day

of acquisition is added to each timestamp. Self-attention along the temporal axis is then applied to spectrally and spatially projected features x_m^R, resulting in a temporally aggregated representation:

$$x_m^T = \text{LTAE}(x_m^R), \text{with } x_m^T \in \mathbb{R}^{D \times H_{target} \times W_{target}} . \quad (6)$$

Multimodal network. After temporal aggregation, we obtain a set of feature maps $\{x_m^T\}_{m=1}^M$, each representing a different modality projected into the shared resolution-aware latent space. Each spatial position is treated as an individual token, such that one pixel at ground sampling distance GSD_{target} corresponds to one D-dimensional embedding.

Similarly to Scale-MAE [33], we enrich these tokens with GSD-based positional encodings that convey information about the target spatial resolution GSD_{target}. We note that this approach combined with our adjustable spatial resampler and interpolation ratio encoding provides implicit knowledge about the original input spatial resolution as well.

All modality-specific tokens are then flattened along the sequence dimension and concatenated to produce the final multimodal sequence $Z \in \mathbb{R}^{N \times D}$, where $N = M \cdot H_{\text{target}} \cdot W_{\text{target}}$ denotes the total number of tokens across modalities. These multimodal tokens are jointly processed by a shared Transformer [43] encoder, enabling cross-modal reasoning and global spatio-spectro-temporal interactions without requiring modality-specific branches.

3.2. Self-supervised training

RAMEN is trained using a masked image modeling through masked autoencoders (MAE) objective [17] applied to the unified multimodal token sequence Z. The goal is to reconstruct all modalities at their native spatial, channel, and temporal resolutions, using only a subset of visible tokens in the latent space. This forces the model to learn resolution-consistent and modality-agnostic representations.

Encoding and decoding. We apply random masking with ratio R uniformly across all tokens in the multimodal sequence Z. Visible tokens and a learnable [CLS] token are processed by a ViT encoder in the unified resolution-aware latent space. Following MAE formulation, the decoder receives the full set of tokens composed of: (i) the encoded visible tokens, (ii) shared mask tokens leveraged by positional encodings.

Resolution-adjustable image reconstruction. The obtained decoded representations are projected back to their modality-specific resolutions via dedicated *reconstruction* modules that mirror their corresponding projection components introduced in Sec. 3.1. Reconstruction proceeds in three steps: (i) Temporal expansion: for each modality m, its sequence is expanded back to the temporal dimension T_m. Each time step is enriched with day of acquisition encoding and fed into a single transformer block; (ii) Spatial

resampling: features are mapped from their GSD_{target} spatial resolution back to GSD_m with a spatial resampler module; (iii) Channel reconstruction: modality-specific channels are recovered from the embedding dimension via the transpose of the channel-wise projection matrix produced by the channel-conditioned projector.

We supervise the model using the Mean Squared Error (MSE) loss between the original masked pixels x_m^{masked} and their reconstructed equivalent $\hat{x}_m^{masked}$, yielding the final reconstruction loss:

$$\mathcal{L} = \frac{1}{M} \sum_{m=1}^M \frac{\left(\hat{x}_m^{masked} - x_m^{masked}\right)^2}{H_m W_m} . \quad (7)$$

Generalizing on heterogeneous modalities across datasets. RAMEN is designed to flexibly encode arbitrary combinations of multimodal inputs regardless of their spatio-spectro-temporal characteristics. To fully leverage this resolution adjustable encoding during pretraining, we employ a *random multimodal sampling strategy* on a large heterogeneous corpus of multimodal EO datasets.

At each pretraining iteration:

1. a dataset is randomly sampled from the multimodal corpus;
2. a subset of modalities available in that dataset is randomly selected;
3. a target GSD_{target} is drawn from a dataset-specific range of resolutions;
4. all sampled modalities are projected into the shared latent space using the modules described in Sec. 3.1.

This stochastic strategy has two desirable effects. First, it exposes RAMEN to a continuously varying combination of sensors and spatio-spectro-temporal resolutions, encouraging the model to learn modality-agnostic and resolution-consistent representations. Second, it substantially improves computational efficiency: high resolution multimodal sequences that are the most memory- and time-demanding are only sampled occasionally, while the model learns to adapt and generalize to the full spectrum of available resolutions and combinations.

4. Experiments

4.1. Pretraining data

To pretrain RAMEN on a wide range of sensors exhibiting varied spatio-spectro-temporal resolutions, we assemble a large-scale EO corpus combining three complementary datasets:

- **FLAIR-HUB** [12]: A multi-sensor land-cover dataset with very high resolution RGB-NIR imagery, 10 bands Sentinel-2 (S2) time series, VV/VH Sentinel-1 (S1) time series and elevation (DSM/DTM) maps covering France. The target GSD is uniformly sampled in the range 3-20 m at 1 m interval;

Table 1. **PANGAEA [27] benchmark results on 8 EO downstream tasks**. The best model per column is highlighted in bold, the second best is underscored. We indicate unimodal datasets with *. Encoders are frozen for pretrained models, while U-Net and ViT baselines are trained from scratch for each specific task.

Model	Model size	BurnSr* [19]	MADOS* [21]	PASTIS [14]	Sen1Fl11 [32]	DEN* [39]	CTM-SS [25]	SN7* [10]	AI4Farms* [31]	Avg. mIoU	Avg. Rank
U-Net baseline [35]	-	<u>84.51</u>	54.79	31.60	**91.42**	<u>39.46</u>	47.57	<u>62.09</u>	**46.34**	57.22	4.25
ViT baseline [9]	Base	81.58	48.19	38.53	87.66	36.83	44.08	52.57	38.37	53.48	10.88
CROMA [11]	Large	82.42	67.55	32.32	90.89	38.29	49.38	59.28	25.65	55.72	6.50
DOFA [48]	Base	80.63	59.58	30.02	89.37	39.29	51.33	61.84	27.07	54.89	7.50
GFM-Swin [28]	Base	76.90	64.71	21.24	72.60	34.09	46.98	60.89	27.19	50.57	11.00
Prithvi 1.0 100M [38]	Base	83.62	49.98	33.93	90.37	27.86	43.07	56.54	26.86	51.53	10.88
RemoteCLIP [24]	Base	76.59	60.00	18.23	74.26	31.78	52.05	57.76	25.12	49.47	12.63
SatlasNet [4]	Base	79.96	55.86	17.51	90.30	36.31	46.97	61.88	25.13	51.74	10.63
Scale-MAE [33]	Large	76.68	57.32	24.55	74.13	35.11	25.42	**62.96**	21.47	47.21	12.75
SpectralGPT [18]	Base	80.47	57.99	35.44	89.07	37.85	46.95	58.86	26.75	54.17	9.50
S.-S12-MoCo [16]	Small	81.58	51.76	34.49	89.26	35.44	48.58	57.64	25.38	53.02	10.50
S.-S12-DINO [6]	Small	81.72	49.37	36.18	88.61	34.81	48.66	56.47	25.62	52.68	11.00
S.-S12-MAE [17]	Small	81.91	49.90	32.03	87.79	34.08	45.80	57.13	24.69	51.67	12.88
S.-S12-Data2Vec [3]	Small	81.91	44.36	34.32	88.15	35.90	54.03	58.23	24.23	52.64	10.63
TerraMindv1-B [20]	Base	82.42	69.52	40.51	90.62	37.87	**55.80**	60.61	28.12	58.18	4.25
TerraMindv1-L [20]	Large	82.93	**75.57**	**43.13**	90.78	37.89	<u>55.04</u>	59.98	27.47	<u>59.10</u>	<u>3.75</u>
SMARTIES-B [36]	Base	82.80				38.50		62.20			
RAMEN (ours)	Base	**85.02**	<u>69.72</u>	<u>42.29</u>	<u>91.03</u>	**39.85**	53.27	60.31	<u>38.78</u>	**60.03**	**2.63**

- **WorldStrat** [8]: A global collection covering 10000 km^2 of matched high resolution RGB-NIR imagery and low resolution 12 bands S2 time-series stratified to all type of land-use across the world. To reduce memory constraints, the original SPOT6 (1052x1052px) and S2 (156x156px) tiles are divided in four non-overlapping tiles during pre-training. The target GSD is uniformly sampled in the range 5-20 m at 1 m interval;

- **MMEarth64** [29]: A large corpus of 1.2 million locations distributed around the world combining 13 bands S2 imagery, 8 bands S1 spanning all available ascending/descending polarizations and elevation (DSM/slope) maps. For computational efficiency, we sample 60% of all available locations, stratified by biome, as our pretraining set. The target GSD is uniformly sampled in the range 20-100 m at 10 m interval.

To mitigate cross-sensors and cross-datasets distribution shifts, all inputs are standardized with per-channel mean and standard deviation values.

4.2. Experimental setup

Pretraining hyperparameters. We pretrain RAMEN using a ViT-Base backbone [9] as our encoder and follow the same architectural parameters as the vanilla MAE whenever possible. We set the number of convolutional experts to $N_{conv} = 4$ for our main experiment and masking ratio R to 75%. In detail, RAMEN is trained for 100 epochs on a cluster of 16 H100 GPUs using the AdamW optimizer and a base learning rate of $1.5e-4$. The learning rate is warmed up for 20 epochs and decayed with a cosine schedule. To fit memory requirements, batch size is tailored to the size of input tiles and target GSD range for each dataset: FLAIR-HUB uses a batch size of 64, WorldStrat has a batch size of 32 and MMEarth uses a batch size of 512.

Evaluation protocol. We evaluate RAMEN across 8 downstream datasets from the PANGAEA benchmark [27]. It covers a diverse set of unimodal, multimodal and multi-temporal EO semantic segmentation tasks including land-cover mapping, flood detection, or wildfire detection. These datasets span a wide range of modalities (aerial, multispectral, SAR), spatial resolutions (0.2m to 30m), and temporal dynamics (single-date and time-series), providing a rigorous test of model generalization capabilities. Following TerraMind [20], we exclude xView2 and BioMassters tasks due to non-reproducible results, and omit FiveBillionPixels as the custom dataset and splits used are not available.

For fair comparison, we follow the standardized PANGAEA evaluation protocol on all datasets. During experiments, the pretrained RAMEN encoder is frozen, and only a UPerNet [47] decoder is finetuned on the downstream tasks. To assess the effect of controllable feature map resolution, we conduct extensive experiments varying the target GSD at inference, examining the trade-off between spatial precision, computational cost, and task performance. We report the results obtained by maximizing the mIoU on the validation set of each downstream task. For large input images, we use the PANGAEA protocol in which random cropping is applied during training to fit the model's memory constraints. Multimodal inputs are fused by concatenating features from all modalities along the embedding dimension.

4.3. Results and analysis

Tab. 1 compares RAMEN against state-of-the-art EO foundation models and baselines across eight diverse semantic segmentation tasks from PANGAEA. RAMEN sets new state-of-the-art performance, achieving the best average mIoU of **60.03** and the best average ranking of **2.63**. Notably, RAMEN surpasses TerraMindv1-L [20] (59.10 average mIoU) while using a lighter ViT-Base encoder. This illustrates that our resolution-adjustable design and multi-
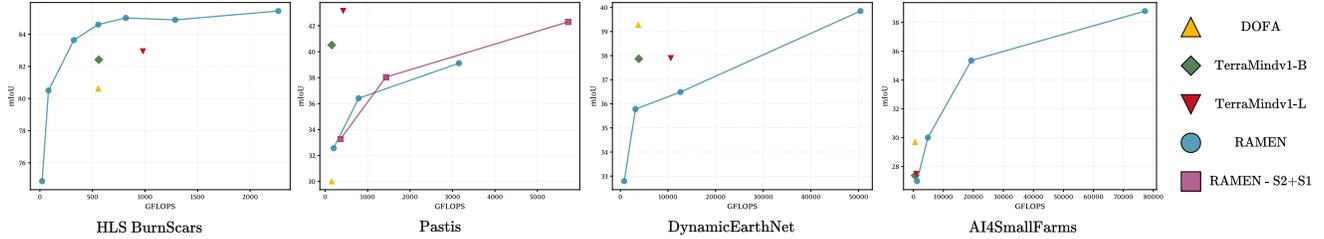
Figure 3. **Compute/performance trade-off across four downstream tasks.** We plot mIoU versus average GFLOPs per test tile for RAMEN at various target spatial resolutions compared to fixed-resolution foundation models.

Table 2. **Effects of adjusting GSD$_{target}$ on performances.** We report the **validation** mIoU obtained on HLS BurnScars and MADOS datasets for different GSD$_{target}$ values.

	HLS BurnScars			MADOS			
Modality	GSD$_m$	GSD$_{target}$	mIoU	Modality	GSD$_m$	GSD$_{target}$	mIoU
HLS	30	360	87.07	S2	10	80	57.09
		300	**88.44**			40	67.05
		240	88.30			20	76.27
		180	88.09			10	**78.07**

Table 3. **Multimodal fusion performances.** We report the best **validation** mIoU across three tasks in both uni- and multimodal setup.

Modalities	Sen1Floods11	Pastis	CTM-SS	Avg mIoU
S2	89.96	40.99	57.20	62.72
S2+S1	**91.20**	**44.25**	**57.35**	**64.27**

modal pretraining strategy enable cross-sensors generalization. Beyond overall performance, RAMEN stands out for its consistency across all tasks. It ranks in the top 2 models on 6 out of 8 tasks, highlighting its robustness to diverse EO modalities. We argue that this consistency is enabled by RAMEN ability to adapt its resolution to the specific needs of each task.

Adapting resolution to the task. A key advantage of RAMEN is the ability to adjust the feature map resolution during inference through control of the target GSD. While increasing resolution generally leads to better performance, a crucial finding is that optimal GSD$_{target}$ is task-dependent and does not always correspond to the finest resolution. Tab. 2 illustrates how varying GSD$_{target}$ impacts performances on two representative tasks:

- On HLS BurnScars [19], a wildfire segmentation task, coarser resolutions yield higher accuracy, as burned areas span large homogeneous regions;
- On MADOS [21], a marine pollutants segmentation task, finer resolutions significantly improve performances, as the model can resolve smaller details.

This adjustable mechanism enables task-aware control over the trade-off between spatial precision and efficiency. In practice, it allows a single model to handle a wide array of real-world applications, from rapid disaster response at coarse scales (e.g. wildfire or flood detection) to fine-grained monitoring (e.g. agriculture delineation, urban mapping, pollutant detection).

Compute/performance trade-off. Fig. 3 illustrates the compute/performance trade-off of RAMEN on four representative datasets, showing how resolution adjustability enables direct control over the balance between accuracy and computational cost. As a transformer-based archi-

tecture, RAMEN exhibits quadratic complexity with the number of tokens, causing GFLOPs to increase rapidly at finer resolutions. Yet across tasks RAMEN remains competitive with fixed-resolution baselines at comparable computational budgets — achieving 85.02 mIoU at 817 GFLOPs on BurnScars (versus TerraMind-L 82.93 mIoU at 980 GFLOPs) and reaching 35.78 mIoU at 3170 GFLOPs on DynamicEarthNet compared to DOFA (39.29 mIoU at 3760 GFLOPs) and TerraMind-L (37.89 mIoU at 10591 GFLOPs). Practitioners can select coarser GSDs to reduce costs — on Pastis, 33.26 mIoU at 359 GFLOPs achieves ∼ 80% of peak performance with ∼ 7.4× faster inference. Notably, accessing finer resolutions enables RAMEN to surpass fixed-resolution limits: on AI4SmallFarms, RAMEN achieves 38.78 mIoU where all foundation models plateau below ∼ 30 mIoU, demonstrating that resolution adjustability enables qualitatively better performance on detail-critical tasks.

Multimodal fusion. RAMEN unified framework naturally handles multimodal inputs by projecting them in a unified resolution-aware latent space. Tab. 3 demonstrates the effectiveness of this approach across three tasks requiring multimodal integration. Incorporating Sentinel-1 SAR data alongside Sentinel-2 imagery consistently improves performances across all evaluated tasks. By learning shared representations that embed physical properties of each modality, RAMEN enables effective multimodal fusion without requiring modality-specific architectural adaptations.

4.4. Ablations

Wavelength encoding. To validate that incorporating physical information improves over treating spectral bands as generic channels, we progressively refine the encoded central wavelengths from coarse approximations to true

Table 4. **Ablation studies on RAMEN's resolution-aware modules.**

(a) **Impact of central wavelength channel encoding on performance.** Progressively refining the encoded central wavelength from coarse approximations to true Sentinel-2 wavelengths improves mIoU on the Sen1Floods11 dataset. **Bold** values denote the actual Sentinel-2 band central wavelengths.

| Strategy | Encoded band central wavelengths (nm) | | | | | | | | | | | | | mIoU |
	CA	R	G	B	RE1	RE2	RE3	NIR-B	NIR-N	H$_2$O	SWIR1	SWIR2	SWIR3	
RGB	490	**490**	**560**	**665**	665	665	665	665	665	665	665	665	665	79.14
RGB-NIR	**440**	**490**	**560**	**665**	865	865	865	865	**865**	865	865	865	865	79.82
Key groups	**440**	**490**	**560**	**665**	740	**740**	740	865	**865**	865	1610	**1610**	1610	82.24
Full	**440**	**490**	**560**	**665**	705	740	783	842	865	945	1373	1610	2200	**83.74**

(b) **Effects of adjustable resampling.** We report the mIoU on the Sen1Floods11 dataset at four coarse to fine GSD$_{target}$.

GSD$_m$	GSD$_{target}$	Naive	False encoding	Adjustable
10	320	77.67	78.63	**78.84**
	160	80.84	83.23	**83.74**
	80	85.45	86.02	**87.13**
	40	88.30	88.98	**88.99**

(c) **Temporal integration performances.** We compare RAMEN with multi-temporal models on three tasks. Late fusion indicates that each timestep is processed independently before aggregation via a LTAE [13].

Model	Pastis	DEN	CTM-SS	Avg mIoU
Prithvi	**33.93**	27.86	43.07	34.95
SatlasNet	17.51	**36.31**	46.97	33.59
RAMEN	28.07	33.16	**53.01**	**38.08**
RAMEN-Late fusion	42.29	39.85	53.27	45.14

Sentinel-2 central wavelengths. We test four encoding strategies with increasing spectral precision: from encoding only RGB channels while collapsing others to red wavelength, to progressively distinguishing RGB-NIR, key spectral groups (Coastal Aerosol, RGB, Red-Edge, NIR, SWIR), and finally using true central wavelengths for all 13 bands.

Tab. 4a shows results on the Sen1Floods11 dataset using the 13 Sentinel-2 bands. We observe consistent improvements with increasing spectral precision, with a total gain of +4.6 mIoU from the RGB-only true encoding to full-precision. This improvement demonstrates that incorporating the physical meaning of spectral bands rather than treating them as abstract feature channels enables the model to learn more semantically meaningful representations.

Spatial resampling module. To evaluate the benefit of our adjustable spatial resampling strategy, we compare three variants of RAMEN differing only in how we handle resampling of the input data to the target GSD at inference. (1) **Naive resampling**: inputs are directly bilinearly interpolated without convolutional adaptation; (2) **False encoding**: inputs are resampled but a false interpolation ratio (set to 1) is provided to the spatial encoding. This prevents our mixture of convolutional experts from adapting to the true interpolation scale; (3) **Adjustable resampling**: inputs are resampled with the true interpolation ratio encoded (i.e. using our proposed spatial resampler).

As shown in Tab. 4b, our adjustable resampling strategy consistently improves mIoU across all target GSDs, demonstrating that explicitly encoding the interpolation ratio allows the mixture of convolutional experts to adapt effectively to varying resolutions. The benefit of adjustable resampling becomes increasingly pronounced as the target GSD moves further from the original GSD.

Temporal attention. To ensure fair comparison with non-temporal models that handle time-series via learned late fusion during fine-tuning, we evaluate RAMEN under both our native temporal attention and late fusion configuration. Tab. 4c shows results across three tasks with diverse temporal dynamics — from first 6 days of every month for DEN to evenly distributed over years samples for Pastis. Among multi-temporal models, Prithvi achieves 34.95 average mIoU and SatlasNet 33.59, while RAMEN reaches 38.14. When using late fusion, performance gains are highly task-dependent: CTM-SS improves minimally (+0.26 mIoU), while Pastis and DEN gain substantially (+14.02 and +6.69 respectively). These results show that a late fusion provides additional gains depending on temporal dynamics inherent to each dataset, at the cost of fine-tuning an additional LTAE module. However, RAMEN temporal handling, without the need to fine-tune an additional module, outperforms other specialized temporal models.

5. Conclusion

We presented RAMEN, a resolution-adjustable encoder for EO that enables practitioners to dynamically control feature map resolution at inference time to match task-specific requirements. To the best of our knowledge, RAMEN is the first EO foundation model which can simultaneously learn modality-agnostic, multi-temporal and resolution-adjustable representations. Once pretrained, our model sets new state-of-the-art performances across multiple tasks and modalities and outperforms larger models. This framework provides a scalable foundation for general-purpose EO models capable of adapting to diverse sensor configurations and application needs.

Acknowledgement

References

[1] Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. OmniSat: Self-supervised modality fusion for earth observation. In *ECCV, Procs.*, pages 409–427, 2024. 2

[2] Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loïc Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *CVPR, Procs.*, pages 19530–19540, 2025. 2, 3

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML, Procs.*, pages 1298–1312. PMLR, 2022. 6

[4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *ICCV, Procs*, pages 16726–16736, 2023. 2, 6

[5] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. FlexiViT: One Model for All Patch Sizes. In *CVPR, Procs.*, pages 14496–14506, 2023. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV, Procs.*, pages 9650–9660, 2021. 6

[7] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery. In *NeurIPS, Procs.*, 2022. 2

[8] Julien Cornebise, Ivan Orsolic, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset - with application to super-resolution. In *NeurIPS, Procs.*, 2022. 6, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR, Procs.*, 2021. 6, 1

[10] Adam Van Etten, Daniel Hogan, Jesus Martinez-Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *CVPR, Procs.*, pages 6398–6407, 2021. 1, 6, 4

[11] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: remote sensing representations with contrastive radar-optical masked autoencoders. In *NeurIPS, Procs.*, 2023. 2, 6

[12] Anatol Garioud, Sébastien Giordano, Nicolas David, and Nicolas Gonthier. FLAIR-HUB: Large-scale multimodal dataset for land cover and crop mapping. *CoRR*, abs/2506.07080, 2025. 5, 2

[13] Vivien Sainte Fare Garnot and Loïc Landrieu. Lightweight temporal self-attention for classifying satellite images time series. In *ECML-PKDD Workshop, Procs.*, pages 171–181, 2020. 3, 4, 8, 1

[14] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *CVPR, Procs.*, pages 4872–4881, 2021. 1, 6, 3

[15] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, Kang Wu, Dingxiang Hu, Huimei He, Jian Wang, Jingdong Chen, Ming Yang, Yongjun Zhang, and Yansheng Li. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *CVPR, Procs.*, pages 27662–27673, 2024. 2

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR, Procs.*, pages 9729–9738, 2020. 6

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR, Procs.*, pages 15979–15988, 2022. 2, 5, 6, 1

[18] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jón Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5227–5244, 2024. 2, 6

[19] Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *CoRR*, abs/2310.18660, 2023. 1, 6, 7, 3

[20] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabé-Moreno, and Nicolas Longépé. TerraMind: Large-scale generative multimodality for earth observation. In *ICCV, Procs.*, 2025. 2, 6, 3

[21] Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in sentinel-2 imagery. *ISPRS-J. Photogramm. Remote Sens.*, 210:39–54, 2024. 6, 7, 3

[22] Antoine Labatie, Michael Vaccaro, Nina Lardiere, Anatol Garioud, and Nicolas Gonthier. MAESTRO: masked autoencoders for multimodal, multitemporal, and multispectral earth observation data. *CoRR*, abs/2508.10894, 2025. 2

[23] Xuyang Li, Chenyu Li, Pedram Ghamisi, and Danfeng Hong. FlexiMo: A Flexible Remote Sensing Foundation Model. *CoRR*, abs/2503.23844, 2025. 2, 3

[24] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–16, 2024. 6

[25] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *CVPR Workshops, Procs.*, pages 75–82, 2019. 1, 6, 4

[26] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i-Nieto, David Vázquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV, Procs.*, pages 9394–9403, 2021. 2

[27] Valerio Marsocci, Yuru Jia, Georges Le Bellier, Dávid Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *CoRR*, abs/2412.04204, 2024. 2, 6, 1

[28] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen, and Mu Li. GFM: building geospatial foundation models via continual pretraining. *CoRR*, abs/2302.04476, 2023. 2, 6

[29] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge J. Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *ECCV, Procs.*, pages 164–182, 2024. 2, 6

[30] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman H. Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *CVPR, Procs.*, pages 27811–27819, 2024. 2

[31] Claudio Persello, Jeroen Grift, Fan Xinyan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andy Nelson. AI4SmallholderFarms: A large-scale data set for crop field delineation in smallholder farms in southeast asia. *IEEE Geosci. Remote Sens. Lett.*, 2023. 6, 4

[32] Clément Rambour, Nicolas Audebert, E Koeniguer, Bertrand Le Saux, M Crucianu, and Mihai Datcu. Flood detection in time series of optical and SAR images. *ISPRS Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, 43:1343–1346, 2020. 1, 6, 3

[33] Colorado J. Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *ICCV, Procs.*, pages 4065–4076, 2023. 2, 3, 5, 6, 1

[34] FAIR (META research). fvcore 0.1.5. https://github.com/facebookresearch/fvcore, 2025. 1

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI, Procs.*, pages 234–241, 2015. 6

[36] Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: spectrum-aware multi-sensor auto-encoder for remote sensing images. In *ICCV, Procs.*, 2025. 2, 3, 6

[37] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Trans. Geosci. Remote. Sens.*, 61:1–22, 2023. 2

[38] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Porsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, João Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, Srija Chakraborty, Sizhe Wang, Ankur Kumar, Myscon Truong, Denys Godwin, Hyunho Lee, Chia-Yu Hsu, Ata Akbari Asanjan, Besart Mujeci, Trevor Keenan, Paulo Arevalo, Wenwen Li, Hamed Alemohammad, Pontus Olofsson, Christopher Hain, Robert Kennedy, Bianca Zadrozny, Gabriele Cavallaro, Campbell D. Watson, Manil Maskey, Rahul Ramachandran, and Juan Bernabé-Moreno. Prithvi-EO-2.0: A versatile multi-temporal foundation model for earth observation applications. *CoRR*, abs/2412.02732, 2024. 2, 3, 6

[39] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *CVPR, Procs.*, pages 21158–21167, 2022. 1, 6, 3

[40] Gabriel Tseng, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *CoRR*, abs/2304.14065, 2023. 2, 3

[41] Gabriel Tseng, Anthony Fuller, Marlena Reil, Henry Herzog, Patrick Beukema, Favyen Bastani, James R. Green, Evan Shelhamer, Hannah Kerner, and David Rolnick. Galileo: Learning global and local features in pretrained remote sensing models. *CoRR*, abs/2502.09356, 2025. 2, 3

[42] Devis Tuia, Konrad Schindler, Begüm Demir, Gustau Camps-Valls, Xiao Xiang Zhu, Mrinalini Kochupillai, Saso Dzeroski, Jan N. van Rijn, Holger H. Hoos, Fabio Del Frate, Mihai Datcu, Jorge-Arnulfo Quiané-Ruiz, Volker Markl, Bertrand Le Saux, and Rochelle Schneider. Artificial intelligence to advance earth observation: a perspective. *CoRR*, abs/2305.08413, 2023. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS, Procs.*, 2017. 3, 5

[44] Diego A. Velázquez, Pau Rodríguez López, Sergio Alonso, Josep M. Gonfaus, Jordi Gonzàlez, Gerardo Richarte, Javier Marin, Yoshua Bengio, and Alexandre Lacoste. EarthView: A large scale remote sensing dataset for self-supervision. In *WACV Workshops, Procs.*, pages 1138–1147, 2025. 2

[45] Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J. Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *CVPR Workshops, Procs.*, pages 2204–2214, 2025. 3

[46] Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *ECCV, Procs.*, pages 286–303, 2024. 2

[47] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified Perceptual Parsing for Scene Understanding. In *ECCV, Procs.*, pages 432–448, 2018. 6, 1

[48] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J. Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired foundation model for observing the earth crossing modalities. *CoRR*, abs/2403.15356, 2024. 2, 3, 6

RAMEN: Resolution-Adjustable Multimodal Encoder for Earth Observation

Supplementary Material

In the supplementary, we provide implementation details about RAMEN pretraining and evaluation (Sec. A), additional analysis of RAMEN architecture and efficiency (Sec. B), extended results on all downstream tasks in PAN-GAEA benchmark (Sec. C) and qualitative examples of RAMEN representations at various GSD_{target} (Sec. D).

A. Implementation details

A.1. RAMEN architecture

RAMEN follows the Vision Transformer architecture [9] and has $98.5M$ learnable parameters, all shared across sensors except for the channel-conditioned projectors separated across three modality types (Optical, radar and DEM):

- **Channel-conditioned projectors** ($3 \times 2.4M$ **parameters**). Each channel-conditioned projector consists of a MLP processing channel-conditioned encodings to produce a channel-wise projection matrix;
- **Spatial resampler** ($2.5M$ **parameters**). Composed of 4 1×1 convolution experts. A MLP and softmax layer process log scaled interpolation ratio encoding to produce normalized weights $\{w_n\}_{n=1}^4$;
- **Temporal encoder** ($3.7M$ **parameters**). A lightweight temporal attention encoder (LTAE) [13] processes time-series inputs. Sinusoidal encoding based on day of acquisition is added to each timestep before temporal aggregation;
- **Encoder blocks** ($85.1M$ **parameters**). We follow the standard ViT-Base architecture composed of 12 self-attention blocks with 12 heads. The embedding dimension D is set to 768.

Decoder blocks and reconstruction ($33.0M$ **parameters**). We follow the MAE [17] framework for the decoder architecture. In details, the decoder has an embedding dimension of 512 and is composed of 8 self-attention blocks with 16 heads. Resolution-adjustable reconstruction modules are similar to their corresponding projection ones, except for the temporal reconstruction composed of one self-attention block processing feature map expanded and enriched with day-of-acquisition encoding independently.

GSD-based positional encoding. To ensure coherent processing across arbitrary target resolutions, RAMEN incorporates GSD-based positional encodings following Scale-MAE [33]. These encodings embed the target GSD directly into the sinusoidal positional functions as:

$$\text{GSDPE}(\text{pos}_x, 2k) = \sin\left(\frac{GSD_{target}}{G} \frac{\text{pos}_x}{10000^{2k/D}}\right);$$
(A)

$$\text{GSDPE}(\text{pos}_y, 2k+1) = \cos\left(\frac{GSD_{target}}{G} \frac{\text{pos}_y}{10000^{2k/D}}\right),$$
(B)

where G is a reference length set to one.

This formulation is essential for our resolution-adjustable behavior: while RAMEN can ingest any number of tokens depending on GSD_{target}, the positional encoding ties the representation to a consistent physical scale across coarse and fine resolutions. This enables the encoder to maintain spatial coherence and interpret feature maps in a resolution-aware manner.

A.2. RAMEN pretraining corpus

RAMEN pretraining data covers a wide range of heterogeneous EO modalities, incorporating diverse modal, spatial and temporal resolutions. See Tab. A for more details on RAMEN pretraining corpus characteristics.

A.3. PANGAEA evaluation protocol

For all downstream tasks, we follow the standardized PANGAEA [27] evaluation protocol. In details, the pretrained encoder is frozen while a UPerNet [47] decoder is fine-tuned for 80 epochs. On all tasks, AdamW optimizer is used with a base learning rate of $1e-4$, weight decay of 0.05 and batch size of 8. The learning rate is decayed $10\times$ after 60% and 90% of the total steps. On multi-temporal tasks, non-temporal models and RAMEN - Late fusion process each timestep independently before aggregation with a lightweight temporal attention encoder (LTAE) [13].

To process large input tiles, random cropping is applied during training to match the encoder expected input size. For evaluation, a sliding window inference strategy is employed, dividing the image in evenly distributed smaller crops. While RAMEN can natively handle any input size, we restricted input size on large tiles and high resolutions experiments to match memory and time-processing constraints. Extensive results for diverse input size and resolutions can be found in Sec. C. We refer the readers to PANGAEA [27] for detailed informations on the evaluation protocol.

A.4. GFLOPs and inference time calculation

We use the *fvcore* [34] library to compute GFLOPs estimation of processing one input tile for each dataset. We

Table A. **RAMEN pretraining corpus.** We combine three large-scale multimodal EO datasets covering diverse spectral, spatial and temporal resolutions. The image size is expressed in pixels for a square image. During pretraining, the target GSD is randomly sampled from a dataset-specific range.

Dataset	Spatial extent	Num tiles	Modality	Image size	Bands	Time-series	GSD (m)	Target GSD range (interval)	Batch size (pretraining)
FLAIR-HUB [12]	2 528 km^2	241 100	Aerial VHR	512	4		0.2	3 - 20m (1m)	64
			S2	10	10	✓	10		
			S1	10	2	✓	10		
			DEM	512	2		0.2		
WorldStrat [8]	9 820 km^2	62 848	SPOT6	263	4		1.5	5 - 20m (1m)	32
			S2	39	12	✓	10		
MMEarth64 [29]	460 800 km^2	720 000	S2	64	13		10	20 - 100m (10m)	512
			S1	64	8		10		
			DEM	64	2		10		

Table B. **Unseen sensor performances.** We report the mIoU and rank (across all PANGAEA models) on three tasks using sensors unseen during pretraining. RAMEN achieves the highest average mIoU and rank, highlighting the robustness of its modality-agnostic design and its ability to transfer to new sensor configurations.

Model	mIoU (Rank)			Avg. mIoU (Rank)
	BurnSr	DEN	SN7	
CROMA	82.42 (5)	38.29 (4)	59.28 (9)	60.00 (6.00)
DOFA	80.63 (12)	39.29 (3)	**61.84** (4)	60.59 (6.33)
Terramindv1-B	82.42 (5)	37.87 (6)	60.61 (6)	60.30 (5.66)
Terramindv1-L	82.93 (4)	37.89 (5)	59.98 (8)	60.27 (5.66)
RAMEN	**85.02** (1)	**39.85** (1)	60.31 (7)	**61.73 (3.00)**

Figure A. **Specialization of the convolutional experts across interpolation ratios**. We plot the normalized weights of the four convolutional experts as a function of the interpolation ratio $\mathrm{GSD}_m/\mathrm{GSD}_{target}$.

include for these estimations the processing time of the encoder and UPerNet decoder. Because of PANGAEA sliding window inference strategy, we multiply GFLOPs obtained for one crop by the number of cropped inputs necessary to produce the final segmentation map. Average inference time per tile is computed over 10 steps. All results of GFLOPs and inference time per task for model, input size and GSD_{target} are reported in Sec. C.

B. Additional analysis

B.1. Generalization to unseen sensors

Tab. B reports performance on three tasks featuring sensors unseen during pretraining. RAMEN achieves the highest average mIoU and rank across the benchmark, notably outperforming all other foundation models on BurnScars and DynamicEarthNet, two tasks exhibiting widely different sensor configurations (HLS - 6 multispectral channels/30m GSD and PlanetFusion - RGB-NIR/3m GSD respectively).

These results validate the sensor-agnostic design of RAMEN. The channel-conditioned projector adapts to per-channel characteristics, while our resolution-adjustable framework provides a consistent interface across modalities and spatial scales. As a result, RAMEN transfers effectively to new sensors without requiring sensor-specific pretraining
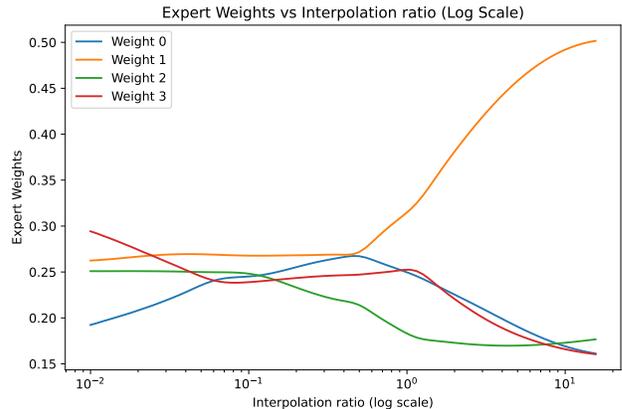
or architectural adaptation, demonstrating strong robustness on heterogeneous EO data.

B.2. Spatial resampling analysis

In standard Vision Transformer architectures, input images are divided into fixed-size patches using strided convolutions (e.g., 16×16 convolution with stride 16). However, this method does not let us explicitly control the resolution of the resulting feature map, nor does it allow the model to understand the ground sampling distance associated with each pixel.

Our spatial resampling module takes a different approach through a two-stage process. First, bilinear interpolation geometrically aligns features to the target GSD, establishing consistent spatial structure across heterogeneous inputs. This interpolation handles all spatial resampling. Second, a mixture of 1×1 convolutional experts refines the interpolated features through scale-specific transformations. The use of 1×1 kernels for our convolution experts is central to our design philosophy: they operate independently on each

Table C. **Effects of adjustable resampling.** We report the validation mIoU for coarse to fine GSD_{target}.

Task	GSD_m	GSD_{target}	Naive	False encoding	Adjustable
HLS BurnScars	30	1920	75.88	76.42	**77.14**
		960	79.70	81.87	**82.45**
		480	83.98	86.31	**87.07**
		300	86.51	87.64	**88.44**
Pastis (S2)	10	80	29.42	32.92	**32.97**
		40	32.22	37.85	**38.02**
		20	37.35	40.66	**40.99**

spatial location, adjusting channel statistics without introducing any spatial dependencies. After interpolation has established the spatial structure, the experts focus purely on feature-space refinement and adapting channel responses based on the interpolation ratio.

Expert specialization across interpolation scales. Fig. A reveals how the four convolutional experts specialize across different interpolation ratios GSD_m/GSD_{target}. While the weights vary smoothly with the scale factor, we note some clear specialization patterns:

- **Expert 1 handle upsampling regimes.** When $GSD_m/GSD_{target} > 1$ (i.e., when RAMEN increases spatial resolution), Expert 1 consistently receives the highest weight. This indicates a learned specialization for refining interpolated coarse inputs;
- **Balanced mixture near identity and weak downsampling.** Around $GSD_m/GSD_{target} \sim 1$ down to 0.1, the weights of the experts remain relatively even. In this regime, the model appears to rely on a mixture of all transformations rather than favoring any single convolution;
- **Expert 3 dominates under strong downsampling.** For heavy downsampling ratios (up to $10e - 2$), the weights gradually shift toward another layer (Expert 3), while Expert 0 become less prominent. This suggests that certain transformations are more suitable when the input has been heavily compressed and requires coarser adjustments.

Additional ablations on adjustable resampling. Tab. C reports adjustable resampling effect compared with n aive and false encoding strategy on two additional tasks, with an **average gain of 0.48 mIoU** across the three evaluated tasks, demonstrating the added benefit of our scale-conditioned refinement of interpolated features.

B.3. Pretraining efficiency

RAMEN was pretrained for 100 epochs in approximately 50 hours on a cluster of 16 H100 GPUs cluster, corresponding to ≈ 800 GPU-hours. This represents a considerably lower computational budget than that reported by other EO foundation models. Notably, TerraMind [20] reports a pretraining cost of about 6 days on 32 A100s GPUs for its Base version (≈ 4608 GPU-hours) and over 10 days for its Large

configuration (≈ 7680 GPU-hours).

C. Detailed results

We present in this section extended results across various GSD_{target} on all downstream tasks. Fig. B illustrates the compute/performance trade-off on the eight downstream tasks considered.

C.1. HLS BurnScars (BurnSr)

HLS BurnScars [19] is a post-fire burn scar segmentation task using Harmonized Landsat-Sentinel (HLS) imagery. It contains 804 tiles of size 512×512 at $GSD = 30m$, with six multispectral bands (RGB-NIR-SWIR1-SWIR2).

Tab. D presents the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}.

C.2. MADOS

Marine Debris and Oil Spill (MADOS) [21] is a marine pollutants segmentation task using Sentinel-2 data. It contains 2803 tiles of size 240×240 at $GSD = 10m$, with eleven multispectral bands.

Tab. E presents the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}.

C.3. Pastis

Pastis [14] is a semantic segmentation task of agricultural parcels using multi-temporal Sentinel-2 and Sentinel-1 data. It contains 2433 tiles of size 128×128 at $GSD = 10m$, with ten multispectral bands and two polarizations (VV/VH) respectively. Following PANGAEA benchmark setup, 6 evenly distributed over time captures are selected.

Tab. F and Tab. G detail the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target} using RAMEN temporal encoder and Late LTAE fusion respectively.

C.4. Sen1Floods11 (Sen1Fl11)

Sen1Floods11 [32] is a global flood mapping segmentation task using Sentinel-2 and Sentinel-1 data. It contains 4831 tiles of size 512×512 at $GSD = 10m$, with 13 multispectral bands and two polarizations (VV/VH) respectively.

Tab. H presents the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}.

C.5. DynamicEarthNet (DYN)

DynamicEarthNet [39] is a semantic segmentation task using multi-temporal at daily observations PlanetFusion data. It spans 75 areas of interest of size 1024×1024 at $GSD = 3m$, with RGB-NIR bands. Following PANGAEA benchmark setup, the 6 first captures of every month are selected.

Tab. I and Tab. J detail the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}
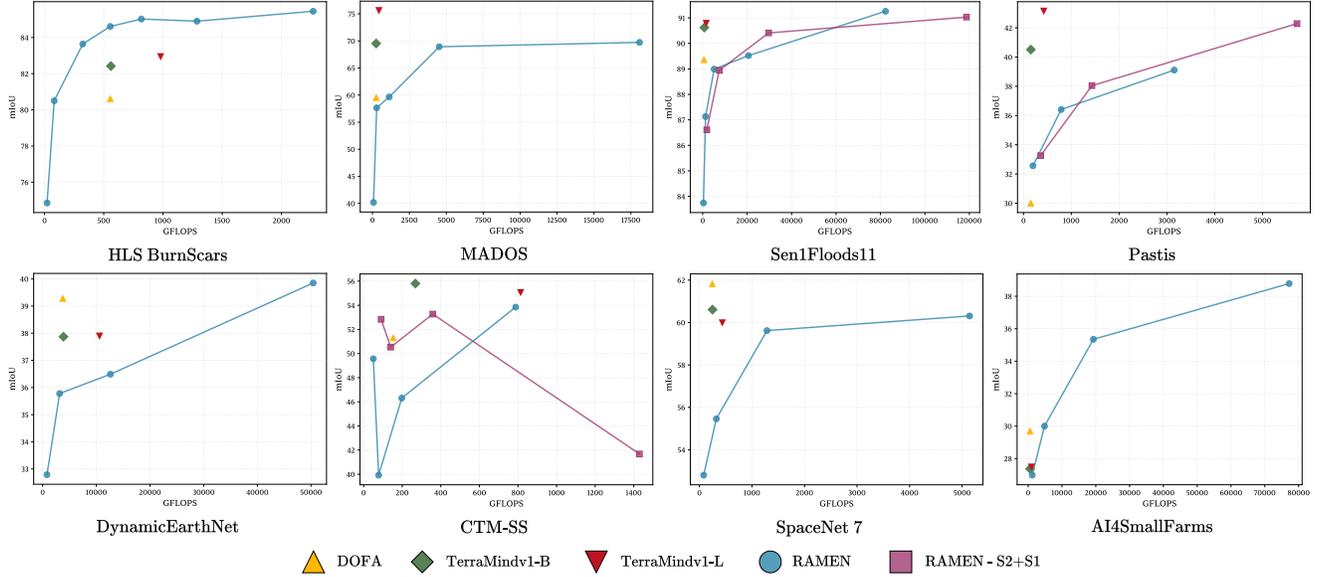
Figure B. **Compute/performance trade-off across eight downstream tasks.** We plot mIoU versus average GFLOPs per test tile for RAMEN at various target spatial resolutions compared to fixed-resolution foundation models.

using RAMEN temporal encoder and Late LTAE fusion respectively.

C.6. CropTypeMapping - South Sudan (CTM-SS)

CropTypeMapping - South Sudan [25] is a semantic segmentation task of agricultural parcels in underrepresented regions using multi-temporal Sentinel-2 and Sentinel-1 data. It contains 837 tiles of size 64×64 at GSD $= 10m$, with ten multispectral bands and two polarizations (VV/VH) respectively.

Tab. K and Tab. L detail the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target} using RAMEN temporal encoder and Late LTAE fusion respectively.

C.7. SpaceNet 7 (SN7)

SpaceNet 7 [10] is a urban semantic segmentation task using PlanetScope data. It contains 15973 tiles of size 256×256 at GSD $= 3m$, with RGB-NIR bands.

Tab. M presents the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}.

C.8. AI4SmallFarms (AI4Farms)

AI4SmallFarms [31] is an agricultural delineation semantic segmentation task using Sentinel-2 data. It contains 62 tiles of size 496×496 at GSD $= 10m$, with RGB-NIR bands. As no validation set is given in PANGAEA benchmark, we computed test scores on the last checkpoint (epoch 80) after training.

Tab. N presents the obtained results, GFLOPs estimation and inference time per tile across various GSD_{target}.

D. Qualitative results

We present in Fig. C a set of qualitative examples of the predicted segmentation maps at various GSD_{target} values across four downstream tasks. These results showcase how increasing spatial resolution enables RAMEN to produce fine-grained representations on pixel-level segmentation tasks (e.g. AI4SmallFarms).

Table D. **HLS BurnScars (BurnSr) performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	512	30	512	1920	77.14	74.85	21.41	6.87
				960	82.45	80.50	81.67	7.96
				480	87.07	83.64	322.68	14.75
				360	87.54	84.61	554.91	21.13
				300	**88.44**	85.02	817.25	30.31
				240	88.30	84.90	1286.74	51.84
				180	88.09	85.45	2268.13	112.67
DOFA						80.63	554.51	29.38
TerraMind-B						82.42	560.12	27.37
TerraMind-L						82.93	980.06	53.41

Table E. **MADOS performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	240	10	240	160	46.59	40.18	71.16	7.13
				80	57.09	57.64	283.15	13.38
			120	40	67.05	59.65	1131.15	37.02
			60	20	76.27	68.92	4523.13	139.95
				10	**78.07**	69.72	18084.79	600.00
DOFA						59.58	247.83	13.67
TerraMind-B						69.52	249.04	12.83
TerraMind-L						75.57	435.68	24.44

Table F. **Pastis with RAMEN temporal encoder performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Modality	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	S2	128	10	128	80	24.07	23.58	86.26	10.30
					40	26.51	25.93	342.38	17.35
				64	20	23.46	23.14	1367.25	52.78
	S2+S1	128	10	128	40	28.88	28.27	504.89	29.38
				64	20	**29.62**	28.07	2016.87	96.98
DOFA						30.02		153.52	27.16
TerraMind-B						40.51		154.94	21.88
TerraMind-L						43.13		423.70	41.66

Table G. **Pastis with late LTAE fusion performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Modality	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	S2	128	10	128	80	32.97	32.56	198.24	32.54
					40	38.02	36.41	788.66	63.30
				64	20	40.99	39.11	3152.37	177.77
	S2+S1	128	10	128	80	35.51	33.26	358.91	50.17
					40	40.03	38.04	1430.76	112.62
				64	20	**44.25**	42.29	5720.32	371.49
DOFA							30.02	153.52	27.16
TerraMind-B							40.51	154.94	21.88
TerraMind-L							43.13	423.70	41.66

Table H. **Sen1Floods11 performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Modality	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	S2	512	10	512	160	82.32	83.74	324.11	14.68
				256	80	86.09	87.13	1288.57	42.12
				128	40	88.70	88.99	5146.42	153.71
				64	20	89.82	89.52	20577.85	609.96
				64	10	89.96	91.26	82277.51	2850.12
	S2+S1	512	10	256	80	86.43	86.61	1858.47	76.08
				128	40	88.94	88.94	7424.81	280.00
				64	20	90.45	90.41	29690.20	1119.95
				64	10	**91.20**	91.03	118721.00	7539.63
DOFA							89.37	559.72	29.92
TerraMind-B							90.62	729.93	37.11
TerraMind-L							90.78	1565.03	85.63

Table I. **DynamicEarthNet with RAMEN temporal encoder performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	1024	3	512	96	28.21	32.26	361.18	35.46
				48	29.72	33.52	1385.32	72.75
			256	24	31.08	31.23	5483.24	216.42
				12	**31.71**	33.16	21869.41	870.45
DOFA						39.29	3760.22	230.98
TerraMind-B						37.87	3873.00	205.79
TerraMind-L						37.89	10591.81	625.38

Table J. **DynamicEarthNet with late LTAE fusion performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	1024	3	512	96	27.64	32.79	808.81	70.16
				48	30.84	35.78	3170.16	197.36
			256	24	31.74	36.49	12622.58	658.30
				12	**32.19**	39.85	50404.06	3658.17
DOFA						39.29	3760.22	230.98
TerraMind-B						37.87	3873.00	205.79
TerraMind-L						37.89	10591.81	625.38

Table K. **CropTypeMapping- South Sudan with RAMEN temporal encoder performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Modality	Image size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPS	Inference time/tile (ms)
RAMEN	S2	64	10	64	80	46.81	50.69	21.65	8.68
					60	45.89	33.03	33.65	8.64
					40	54.13	46.21	85.66	9.78
					20	**57.20**	53.01	341.68	17.03
	S2+S1	64	10	64	80	50.48	45.09	31.86	12.37
					60	47.80	47.93	49.57	13.00
					40	47.25	51.92	126.31	14.98
					20	51.00	44.82	504.08	28.77
DOFA							51.33	153.49	27.71
TerraMind-B							55.80	268.12	24.08
TerraMind-L							55.04	813.65	66.19

Table L. **CropTypeMapping - South Sudan with late LTAE fusion performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Modality	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	S2	64	10	64	80	47.26	49.57	50.05	24.58
					60	46.71	39.92	77.73	26.37
					40	53.85	46.31	197.64	31.59
					20	56.95	53.85	787.96	61.81
	S2+S1	64	10	64	80	58.19	52.83	90.26	31.83
					60	59.92	50.53	140.50	34.93
					40	**62.22**	53.27	358.17	47.22
					20	57.35	41.68	1429.94	109.66
DOFA							51.33	153.49	27.71
TerraMind-B							55.80	268.12	24.08
TerraMind-L							55.04	813.65	66.19

Table M. **SpaceNet 7 performances.** We report validation and test mIoU, along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	256	4	256	64	50.63	52.80	80.60	7.40
			256	32	53.93	55.46	321.61	13.93
			128	16	58.69	59.62	1285.97	41.42
			64	8	**59.48**	60.31	5143.45	152.53
DOFA						61.84	245.67	13.75
TerraMind-B						60.61	248.94	12.91
TerraMind-L						59.98	435.47	24.42

Table N. **AI4SmallFarms performances.** We report **only test** mIoU (no validation set provided), along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

Model	Img size	GSD_m	Input size	GSD_{target}	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	496	10	496	80	26.98	1207.16	48.39
			248	40	30.00	4826.39	168.77
			124	20	35.35	19303.31	658.78
			64	10	**38.78**	77210.97	2617.79
DOFA					27.07	553.33	29.71
TerraMind-B					28.12	560.11	27.37
TerraMind-L					27.47	980.05	27.47

HLS BurnScars

RGB Image

1920 m

480 m

300 m

Ground Truth

(a) HLS BurnScars
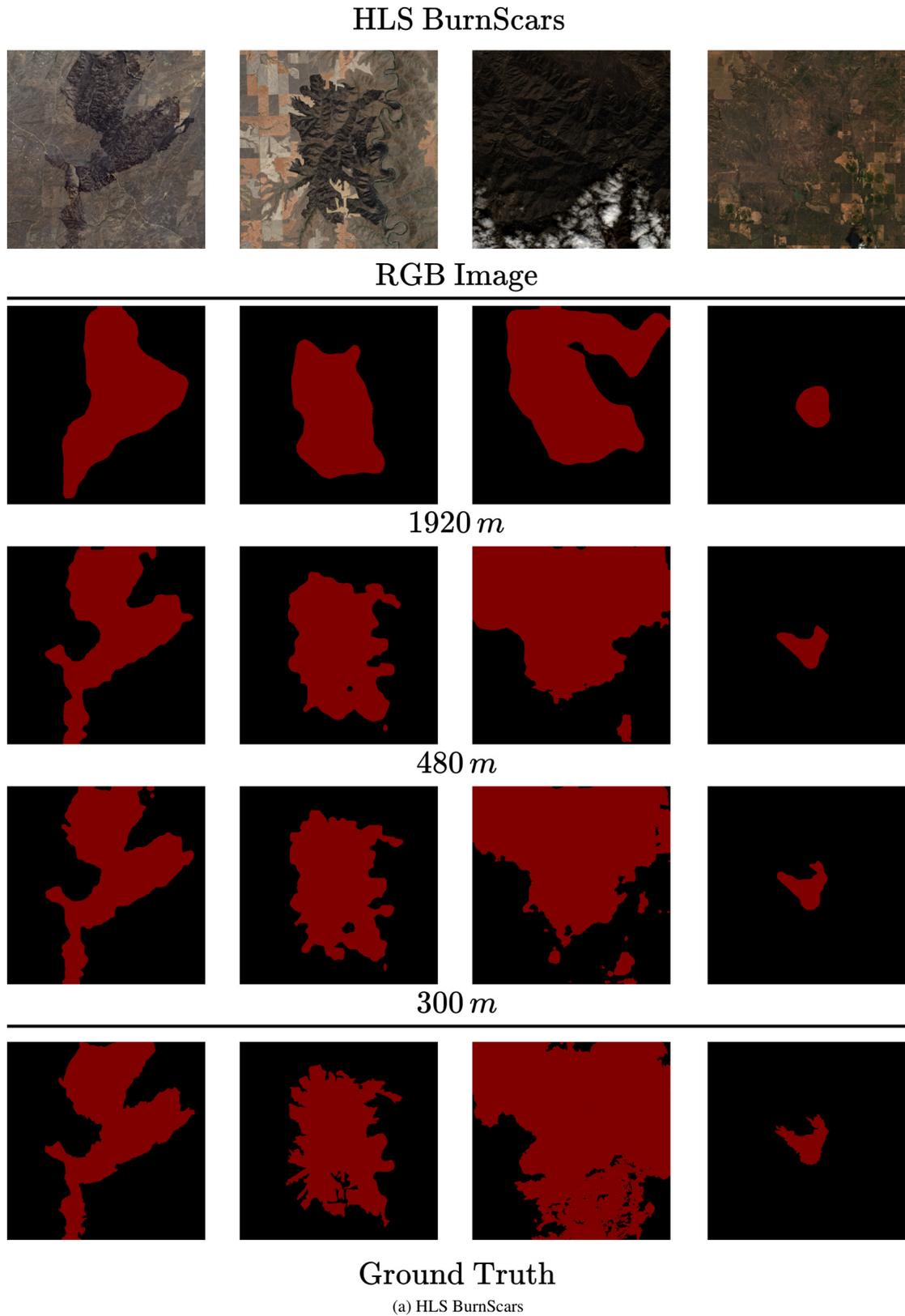
Figure C. **Illustrative prediction results across datasets.** Segmentation maps produced at various coarse to fine GSD_{target}.
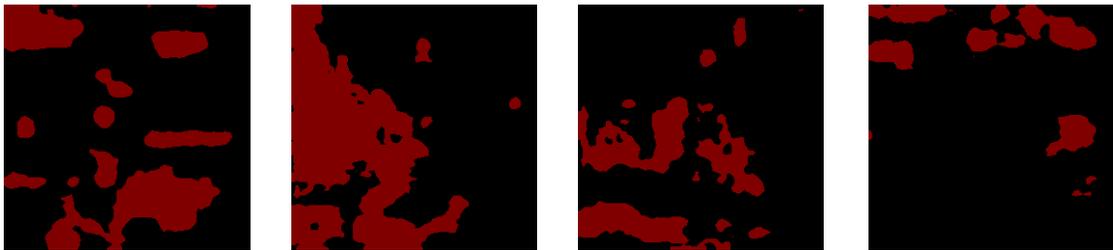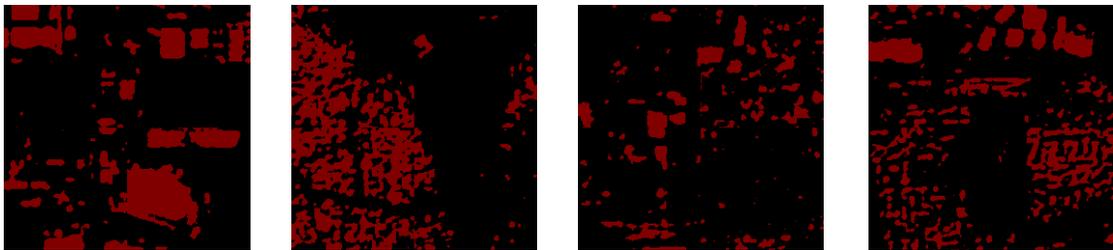
Pastis
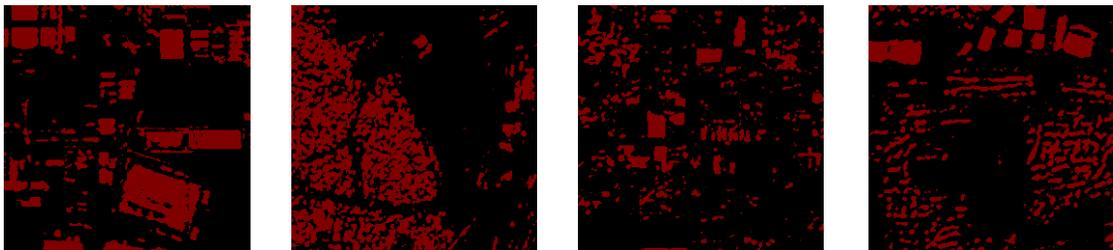

RGB Image
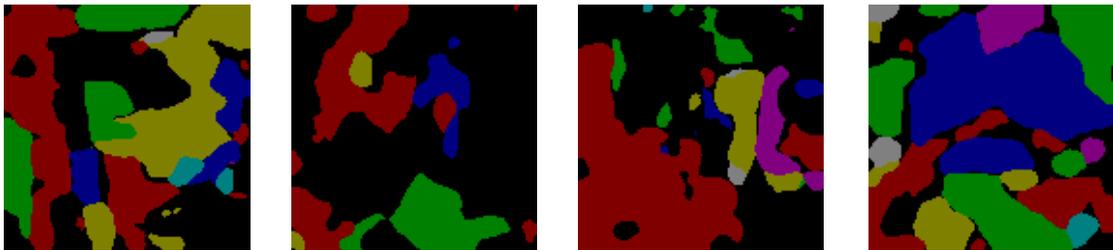
80 m
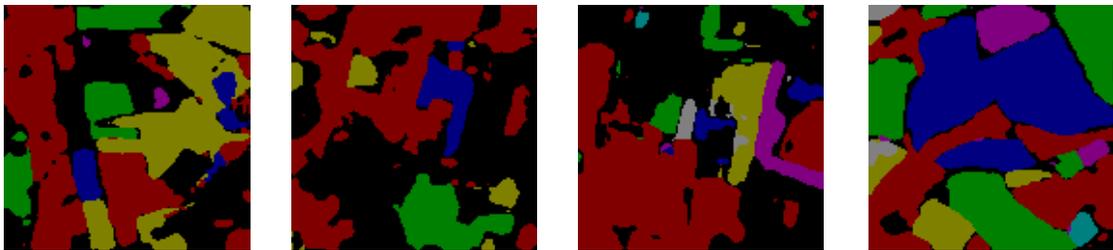
40 m
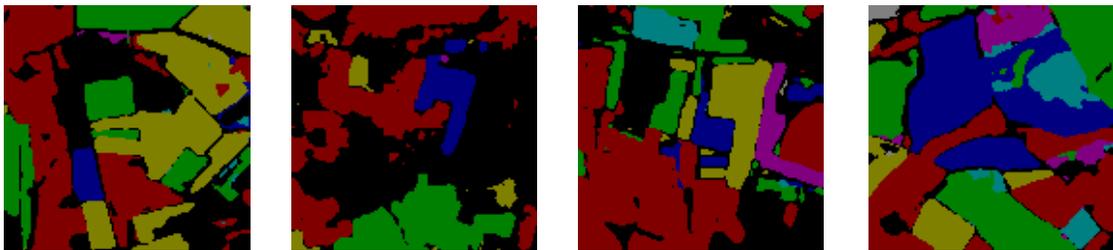
20 m
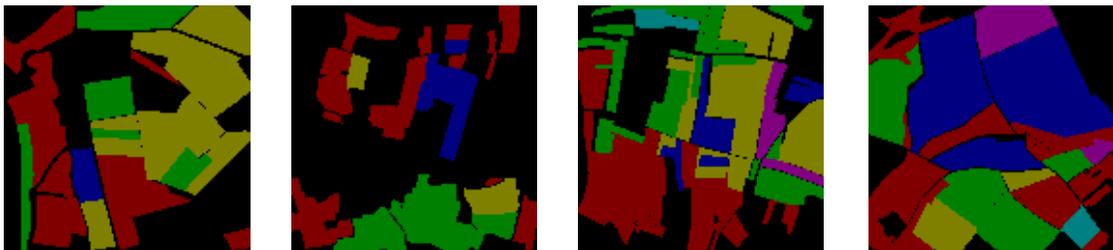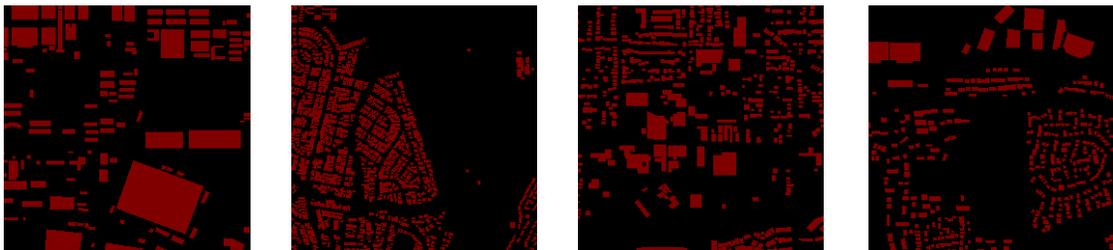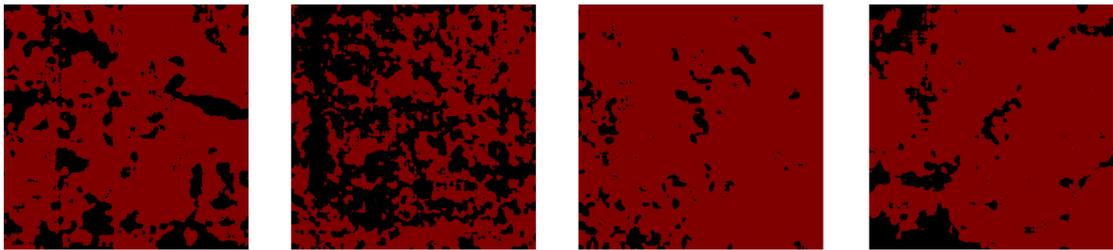
Ground Truth

(b) Pastis

SpaceNet 7
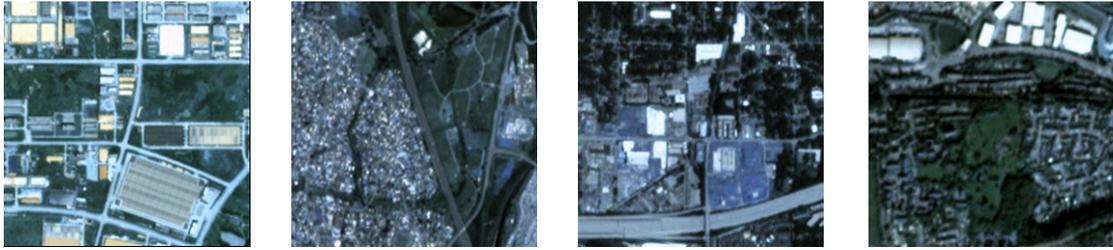
RGB Image

64 m
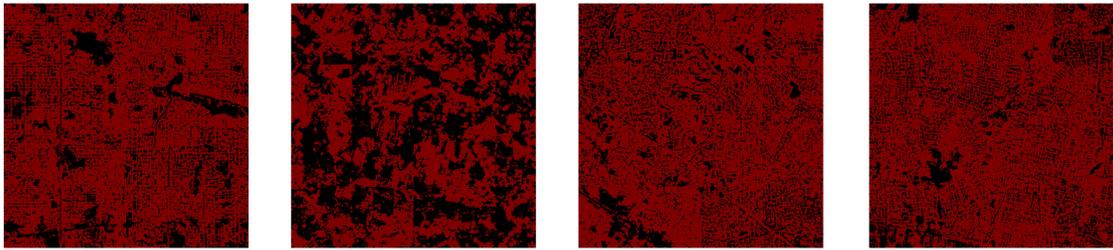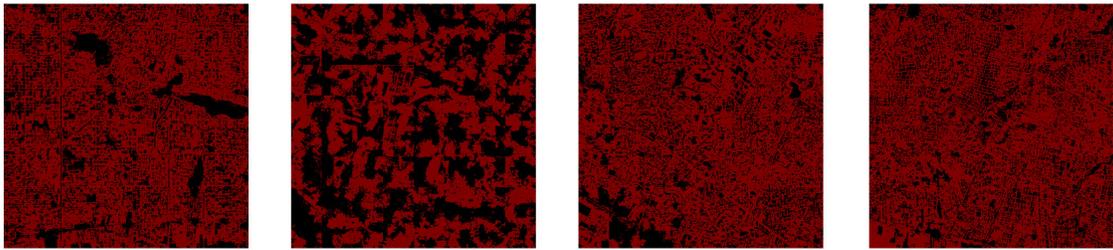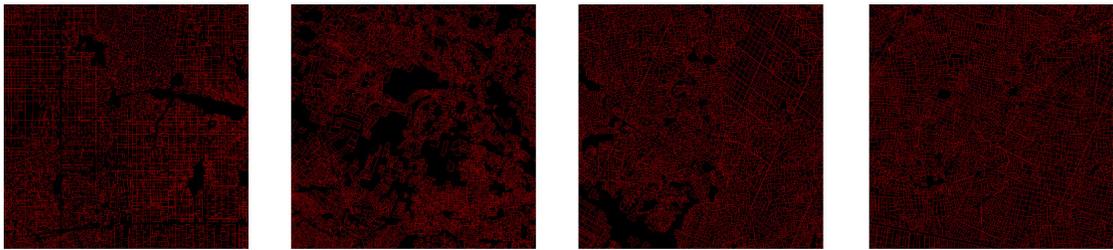
16 m

8 m

Ground Truth

(c) SpaceNet 7

AI4SmallFarms

RGB Image

80 m

20 m

10 m

Ground Truth

(d) AI4SmallFarms