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 PANGAEA 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 [2] 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) [5] 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 [7] 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 [16]. These encodings embed the target GSD directly into the sinusoidal positional functions as:

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

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

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 [12] evaluation protocol. In details, the pretrained encoder is frozen while a UPerNet [19] 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) [5].

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 [12] for detailed informations on the evaluation protocol.

A.4. GFLOPs and inference time calculation

We use the *fvcore* [17] 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 [4]	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 [1]	9 820 km^2	62 848	SPOT6	263	4		1.5	5 - 20m (1m)	32
			S2	39	12	✓	10		
MMEarth64 [13]	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**)

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.

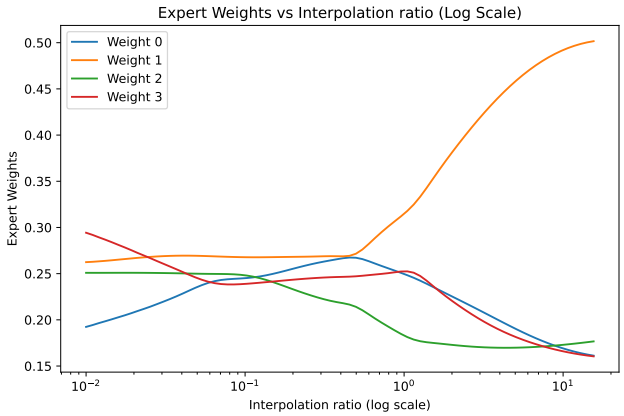


Figure A. **Specialization of the convolutional experts across interpolation ratios**. We plot the normalized weights of the four convolutional experts as a function of the interpolation ratio GSD$_m$/GSD$_{target}$.

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 $\text{GSD}_m/\text{GSD}_{target}$. While the weights vary smoothly with the scale factor, we note some clear specialization patterns:

- **Expert 1 handle upsampling regimes.** When $\text{GSD}_m/\text{GSD}_{target} > 1$ (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 $\text{GSD}_m/\text{GSD}_{target} \sim 1$ 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 $\approx$ 800 GPU-hours. This represents a considerably lower computational budget than that reported by other EO foundation models. Notably, TerraMind [9] reports a pretraining cost of about 6 days on 32 A100s GPUs for its Base version ($\approx$ 4608 GPU-hours) and over 10 days for its Large configuration ($\approx$ 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 [8] 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) [10] 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 [6] 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 [15] 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 [18] 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}

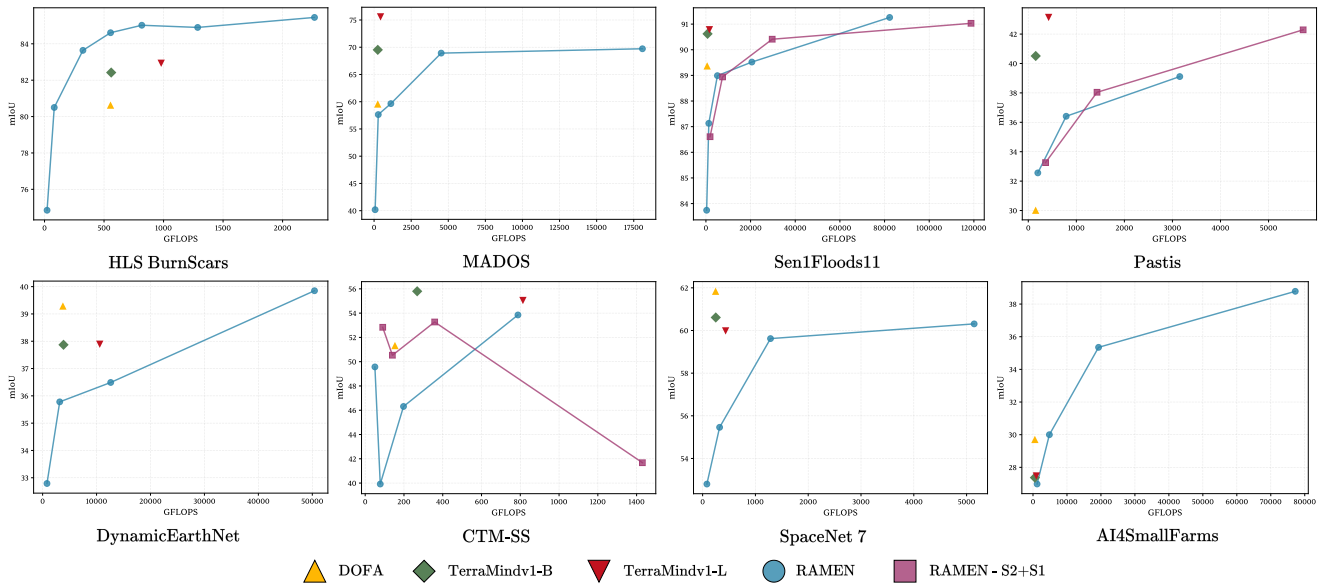


Figure B. **Compute/performance trade-off across eight downstream tasks.** We plot mIoU versus average GFLOPs per test tile for RAMEN at various target spatial resolutions compared to fixed-resolution foundation models.

using RAMEN temporal encoder and Late LTAE fusion respectively.

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

CropTypeMapping - South Sudan [11] 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 [3] 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 [14] 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}.

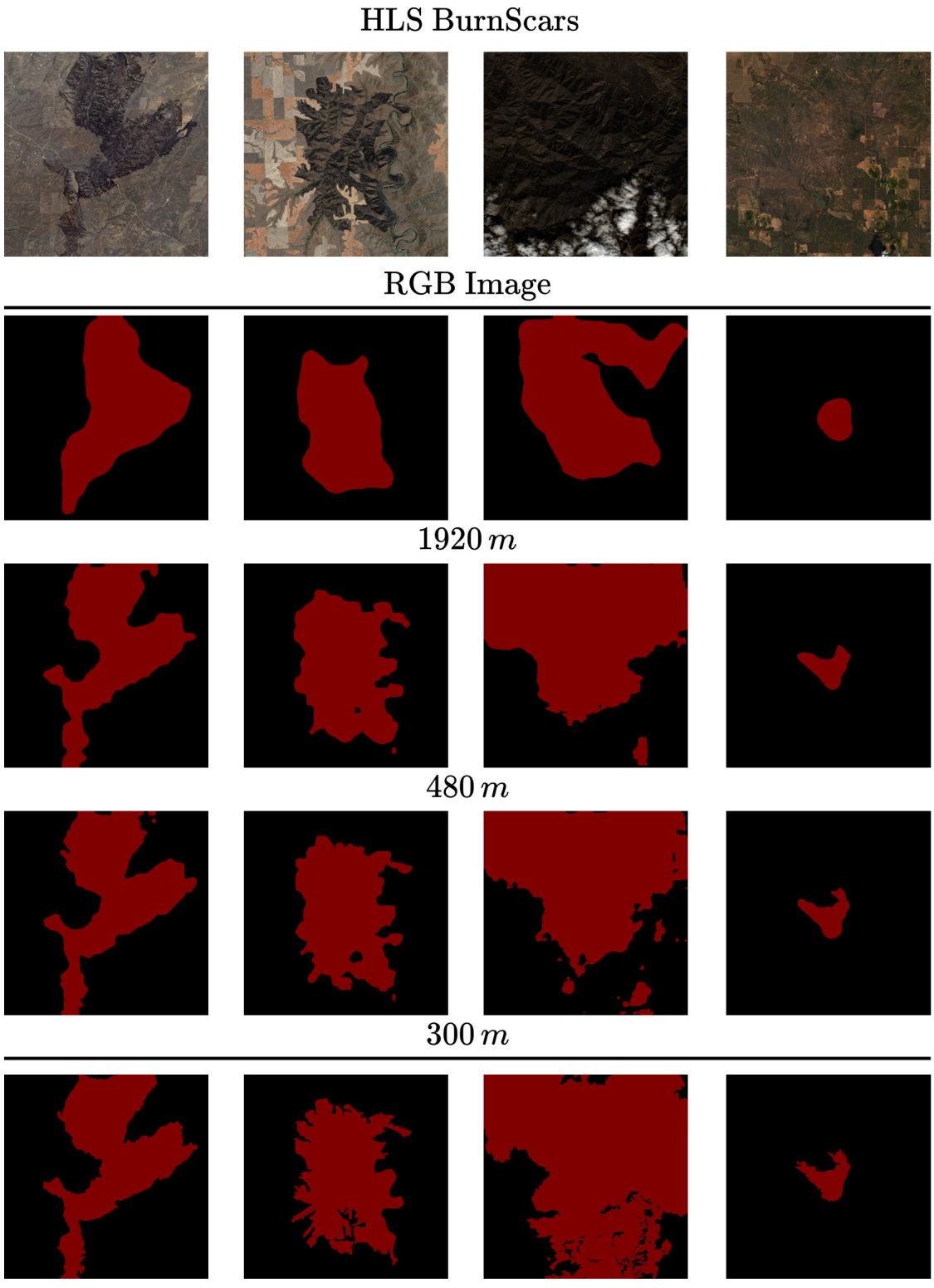
Model	Modality	Img size	GSD_m	Input size	GSD_{target}	val mIoU	test mIoU	GFLOPs	Inference time/tile (ms)
RAMEN	S2	64	10	64	80	47.26	49.57	50.05	24.58
					60	46.71	39.92	77.73	26.37
					40	53.85	46.31	197.64	31.59
					20	56.95	53.85	787.96	61.81
	S2+S1	64	10	64	80	58.19	52.83	90.26	31.83
					60	59.92	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
				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

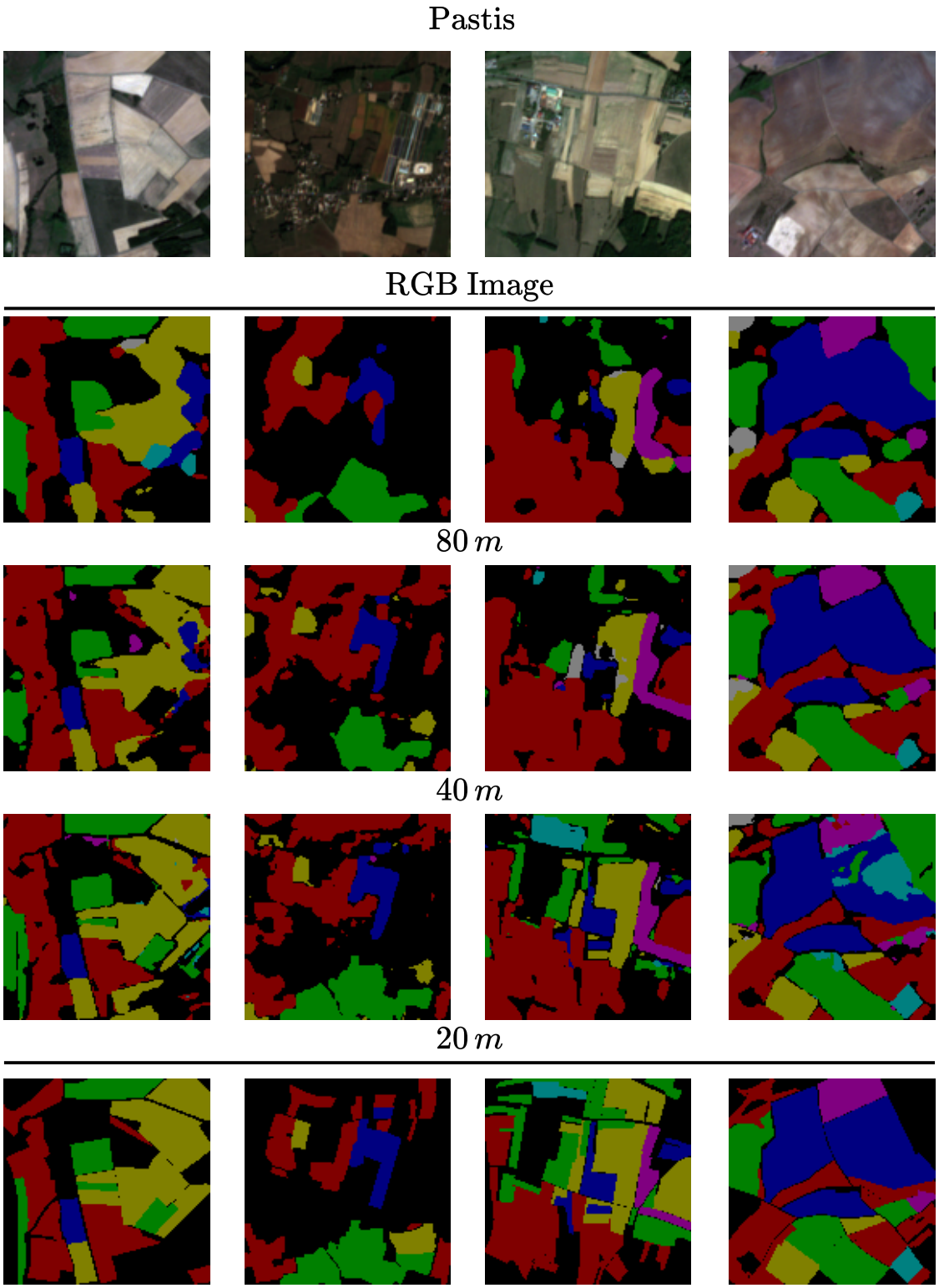
Table N. **AI4SmallFarms performances.** We report **only test** mIoU (no validation set provided), along with GFLOPs and inference time per tile for RAMEN evaluated at different GSD_{target}.

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

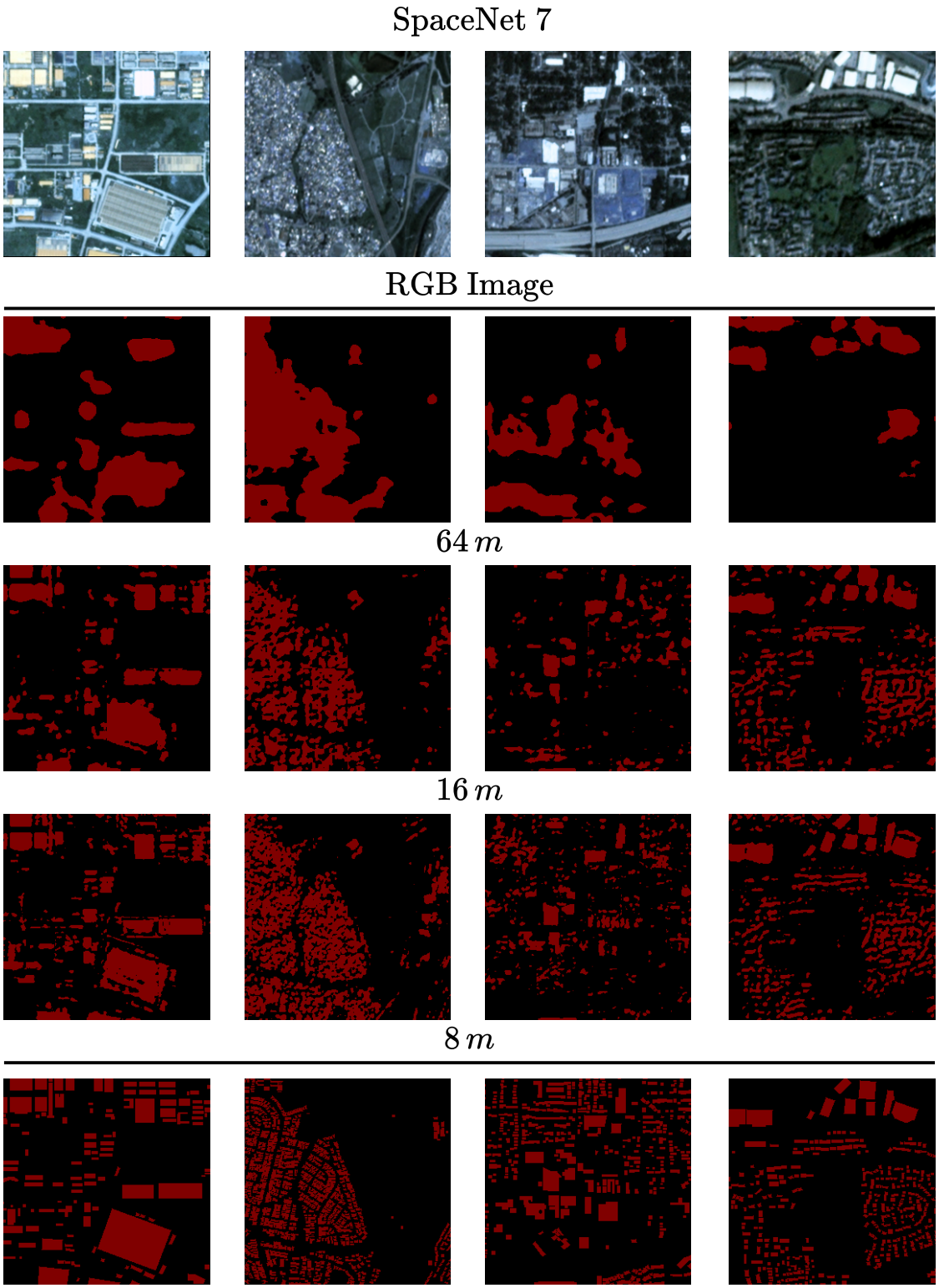


(a) HLS BurnScars

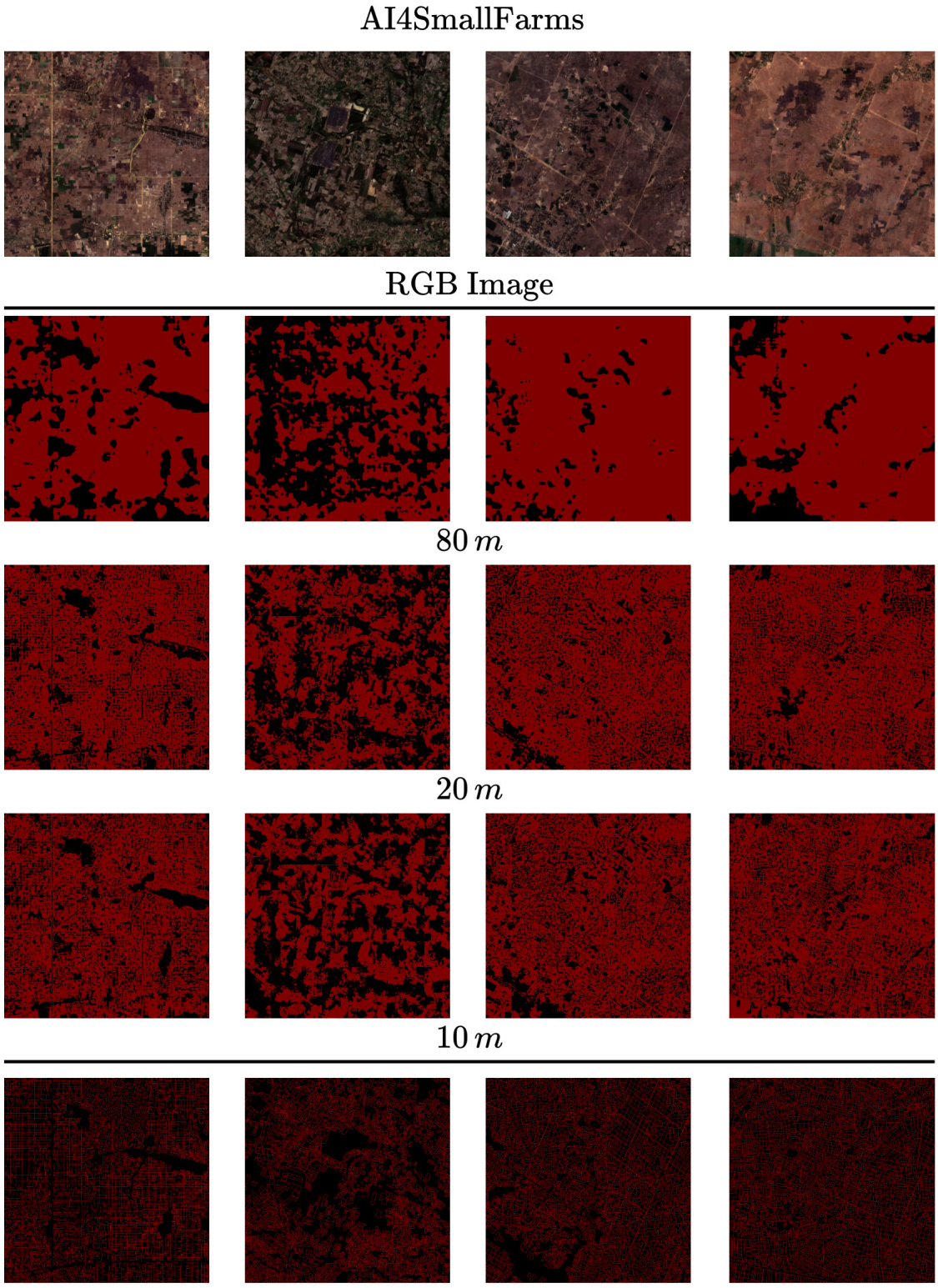
Figure C. **Illustrative prediction results across datasets.** Segmentation maps produced at various coarse to fine GSD_{target}.



(b) Pastis



(c) SpaceNet 7



(d) AI4SmallFarms

References

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

[2] 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. 1

[3] 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. 4

[4] 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. 2

[5] 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. 1

[6] 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. 3

[7] 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. 1

[8] 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. 3

[9] 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. 3

[10] 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. 3

[11] 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. 4

[12] 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. 1

[13] 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

[14] 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. 4

[15] 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. 3

[16] 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. 1

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

[18] 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. 3

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